

Détection de communautés dynamiques dans les réseaux évolutifs de développeurs au sein de Chrome

G. Caporossi, S. Perron, F. Barès, A. Correc

G-2021-60

November 2021

La collection *Les Cahiers du GERAD* est constituée des travaux de recherche menés par nos membres. La plupart de ces documents de travail a été soumis à des revues avec comité de révision. Lorsqu'un document est accepté et publié, le pdf original est retiré si c'est nécessaire et un lien vers l'article publié est ajouté.

The series *Les Cahiers du GERAD* consists of working papers carried out by our members. Most of these pre-prints have been submitted to peer-reviewed journals. When accepted and published, if necessary, the original pdf is removed and a link to the published article is added.

Citation suggérée : G. Caporossi, S. Perron, F. Barès, A. Correc (Novembre 2021). Détection de communautés dynamiques dans les réseaux évolutifs de développeurs au sein de Chrome, Rapport technique, Les Cahiers du GERAD G- 2021-60, GERAD, HEC Montréal, Canada.

Suggested citation: G. Caporossi, S. Perron, F. Barès, A. Correc (November 2021). Détection de communautés dynamiques dans les réseaux évolutifs de développeurs au sein de Chrome, Technical report, Les Cahiers du GERAD G-2021-60, GERAD, HEC Montréal, Canada.

Avant de citer ce rapport technique, veuillez visiter notre site Web (<https://www.gerad.ca/fr/papers/G-2021-60>) afin de mettre à jour vos données de référence, s'il a été publié dans une revue scientifique.

Before citing this technical report, please visit our website (<https://www.gerad.ca/en/papers/G-2021-60>) to update your reference data, if it has been published in a scientific journal.

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2021
– Bibliothèque et Archives Canada, 2021

Legal deposit – Bibliothèque et Archives nationales du Québec, 2021
– Library and Archives Canada, 2021

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

Tél. : 514 340-6053
Télec. : 514 340-5665
info@gerad.ca
www.gerad.ca

Détection de communautés dynamiques dans les réseaux évolutifs de développeurs au sein de Chrome

Gilles Caporossi ^{a, b}

Sylvain Perron ^{a, b}

Franck Barès ^c

Anaïs Correc ^d

^a GERAD, Montréal (Qc), Canada, H3T 1J4

^b Department of Decision Sciences, HEC Montréal, Montréal (Qc), Canada, H3T 2A7

^c Department of Entrepreneurship and Innovation, HEC Montréal, Montréal (Qc), Canada, H3T 2A7

^d Department of Management Sciences and Quantitative Methods, HEC Montréal, Montréal (Qc), Canada, H3T 2A7

franck.bares@hec

gilles.caporossi@gerad.ca

November 2021

Les Cahiers du GERAD

G–2021–60

Copyright © 2021 GERAD, Caporossi, Perron, Barès, Correc

Les textes publiés dans la série des rapports de recherche *Les Cahiers du GERAD* n'engagent que la responsabilité de leurs auteurs. Les auteurs conservent leur droit d'auteur et leurs droits moraux sur leurs publications et les utilisateurs s'engagent à reconnaître et respecter les exigences légales associées à ces droits. Ainsi, les utilisateurs:

- Peuvent télécharger et imprimer une copie de toute publication du portail public aux fins d'étude ou de recherche privée;
- Ne peuvent pas distribuer le matériel ou l'utiliser pour une activité à but lucratif ou pour un gain commercial;
- Peuvent distribuer gratuitement l'URL identifiant la publication.

Si vous pensez que ce document enfreint le droit d'auteur, contactez-nous en fournissant des détails. Nous supprimerons immédiatement l'accès au travail et enquêterons sur votre demande.

The authors are exclusively responsible for the content of their research papers published in the series *Les Cahiers du GERAD*. Copyright and moral rights for the publications are retained by the authors and the users must commit themselves to recognize and abide the legal requirements associated with these rights. Thus, users:

- May download and print one copy of any publication from the public portal for the purpose of private study or research;
- May not further distribute the material or use it for any profit-making activity or commercial gain;
- May freely distribute the URL identifying the publication.

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Abstract : The time-stamped database provided by Google traces the code contributions to Chrome browser on the period from September 2008 to January 2014. It describes the participation of 1045 developers in 46 offices in 22 countries with some 324 860 changelists (CLs).

An individual who contributes to a CL can be done either as “submitter”, “approver”, “Comment” “Comment unsolicited”, “Disapprover”, “reviewer” or “reviewer silent”. We thus have a social network evolving over nearly six years in which the participation of developers in a CL is a collaborative relationship.

The network can be illustrated by a graph, that is to say a set of points called vertices connected by edge segments. In this case, participation in the same CL results in the existence of edges between the vertices representing the parties involved.

The quantitative study of the database proceeds according to two approaches : descriptive statistical analysis with the objective of drawing the dynamic profile of individuals ; and, secondly, an automatic classification which aims to involve developers in sub-groups, also called communities more strongly interconnected than with the rest of the network.

In graph theory, the classification is equivalent to community detection in a graph and is a difficult combinatorial problem, even more so if every relationship is subject to temporal information.

The detectable communities vary over time. In addition, for a given community structure, quantitative measures provide an understanding of the internal organization of the group and measure the effectiveness of the collaborative work.

In light of our analysis, it appears that there are several profiles in the activity for an individual, characterized in particular by experience ; the association of developers can be interpreted in part by membership in an office ; and the effectiveness of a community is based on the structure and composition.

Keywords : Google-Chrome browser, business support-HR, teaming, social network analysis, Big Data

Résumé : Cette étude exploratoire est originale dans le champ de l’accompagnement. A partir de la base de données horodatées fournie par Google, notre étude retrace le processus de mises à jour du navigateur Chrome sur la période allant de septembre 2008 à janvier 2014. Cette base décrit la participation temporelle et volontaire de 1045 développeurs répartis dans 46 bureaux sur 22 pays à quelques 324 860 Changelists (notés CL) ou mises à jour du logiciel. Un individu qui contribue à un CL peut le faire à titre de soit “submitter”, “approver”, “commenter,” “commenter unsolicited”, “disapprover”, “reviewer” ou “reviewer silent”. Nous avons ainsi un réseau social évoluant sur près de six ans dans lequel la participation des développeurs à un CL commun constitue une relation de collaboration. Le réseau peut être illustré par un graphe, c’est-à-dire un ensemble de points appelés sommets reliés ou non entre eux par des segments dits arêtes. Dans ce cas, la participation à un même CL se traduit par l’existence d’arêtes entre les sommets représentant les parties impliquées.

L’étude quantitative de la base de données procède selon deux approches : soit, dans un premier temps, une analyse statistique descriptive dont l’objectif est de dessiner le profil dynamique des individus ; et, dans un second temps, une classification automatique dont le but est d’associer les développeurs en sous-groupes, aussi appelés communautés, plus fortement connectés entre eux qu’avec le reste du réseau. En théorie des graphes, la classification est équivalente à la détection de communautés dans un graphe et constitue un problème combinatoire difficile, plus encore si chaque relation est assortie d’une

information temporelle. Les communautés détectables sur de plus ou moins longues périodes subissent ainsi des modifications dont la chronologie peut être racontée. De plus, à partir d'une structure communautaire déterminée, des mesures quantitatives permettent de comprendre l'organisation interne du groupe et de mesurer l'efficacité du travail de collaboration.

À la lumière de notre analyse, il apparaît qu'il existe plusieurs profils à l'activité d'un individu, caractérisés en particulier par l'expérience ; que l'association des développeurs peut être interprétée en partie par l'affiliation à un bureau ; que l'efficacité d'une communauté est fonction de la structure et sa composition.

Mots clés : Google-navigateur Chrome, accompagnement managérial, équipes (composition, évolution, efficacité), analyse de réseaux sociaux, méga données

1 Méthodologie de l'approche quantitative ¹

La modélisation sous forme de graphe, c'est-à-dire la représentation d'un ensemble de points appelés sommets reliés ou non entre eux par des segments appelés arêtes, est couramment utilisée pour comprendre les relations entre des individus dans des réseaux complexes naturels et artificiels tels que les réseaux sociaux [2], neuraux [3], biologiques [4], de citation [5], de collaboration [6], d'information tels que le *World Wide Web* [7] et d'autres. Ici, le terme individu désigne une *entité* unique du réseau - non pas nécessairement une *personne* - laquelle est représentée par un nœud. Les arêtes liant deux sommets expriment une relation soit binaire (la relation existe ou non), dirigée (la relation n'est pas réciproque), pondérée (la relation est plus ou moins forte) ou encore temporelle (la relation perdure sur une certaine période). L'étude de tels graphes a mené à la découverte de propriétés intéressantes comme la distribution des degrés des nœuds, c'est-à-dire le nombre de liens incidents, selon une loi de puissance [8] ; ou encore l'existence de groupes d'individus aux caractéristiques similaires et plus densément connectés entre eux qu'avec le reste du graphe, dits communautés [9], à partir desquelles il est possible d'identifier les rôles qu'y occupent les individus. En ajoutant une composante temporelle, la classification des sommets dans un contexte dynamique peut révéler l'évolution des communautés et permettre une compréhension approfondie de l'historique du réseau.

1.1 Notations

Soit (u, v) une arête non orientée reliant les sommets u et v ; $G = G(V, E, W)$ un graphe pondéré, non orienté et simple, soit sans boucles ni liens multiples, représentant un réseau où V est l'ensemble des nœuds et E celui des arêtes ; $d(u) = d_u$ le degré, i.e. la somme des poids des arêtes non orientées dont une extrémité touche le sommet u . Soit $\mathcal{C} = \{C_1, C_2, \dots, C_k\}$ une collection de sous-ensembles de V telle que le sous-graphe $C_i \in \mathcal{C}$ est une communauté de G . Dans le cas de communautés disjointes, $C_i \cap C_j = \emptyset$ pour tout $C_i \neq C_j$, alors que dans le cas de communautés avec recouvrement l'ensemble $C_i \cap C_j$ peut être non vide pour $C_i \neq C_j$. Dans les deux cas, un sommet peut être non-affilié, c'est-à-dire $u \in G$ mais $u \notin C_i, \forall i$. Alors, $d_{C_i}^{in}(u)$ est le degré interne de u dans la communauté C_i , soit la somme des degrés des arêtes entre u et v pour $v \in C_i$; et $d_{C_i}^{out}(u)$ son degré externe, soit la somme des degrés des arêtes entre u et v pour $v \notin C_i$. En contexte dynamique, soit $\mathcal{G}$ un graphe évoluant dans le temps tel que $\mathcal{G} = (G^{(0)}, G^{(1)}, \dots, G^{(t_{max})})$ est une séquence d'instantanés où $G^{(t)}$ est l'état du graphe au temps t ; $\Delta G^{(t)} = (\Delta V^{(t)}, \Delta E^{(t)})$ les changements apportés au réseau entre les états t et $t+1$ et tel que $G^{(t+1)} = G^{(t)} + \Delta G^{(t)}$; $C_i^{(t)}$ la i -ème communauté de l'ensemble des communautés $\mathcal{C}^{(t)}$ détectées dans l'instantané $G^{(t)}$.

1.2 Définitions

Méta-communauté Communauté elle-même formée de plus petites communautés.

Graphe biparti Un graphe est dit biparti si l'ensemble de ses sommets V peut être divisé en deux sous-ensembles disjoints, dits classes, tels qu'il n'existe aucun lien entre les sommets de la même classe.

Partition, répartition Soit $\mathcal{C} = \{C_1, \dots, C_{N_c}\}$, l'ensemble des communautés d'un graphe G . $\mathcal{C}$ est une partition de G si $C_i \cap C_j = \emptyset$ pour $C_i \neq C_j$. $\mathcal{C}$ est une répartition de G si $C_i \cap C_j \neq \emptyset$ pour $C_i \neq C_j$. Il peut exister un ou des sommets qui ne sont affiliés à aucune communauté.

Instantané On désigne par *instantané* l'état du réseau au temps t en tant que l'agrégat d'événements survenus entre les temps $t-1$ et t .

Réseau évolutif Un réseau évolutif ² est un ensemble d'individus inter-connectés dont la structure des relations évolue assez lentement dans le temps.

1. Cette section est extraite, parfois littéralement, du document de Correc [1].

2. *evolving network*

Lien intra-communautés, lien inter-communautés Une arête connectant deux sommets $u, v \in V$ tels que $u \in C_i$ et $v \in C_i$ pour un certain i est dit *intra-communautés*. Une arête connectant deux sommets $u, v \in V$ tels que $u \in C_i$ et $v \in C_j$ avec $C_i \neq C_j$ est dit *inter-communautés*.

Communauté Il n'existe pas de définition rigoureuse de la notion de communauté. Une définition généralement admise est qu'il s'agit d'un ensemble de nœuds plus densément connectés entre eux qu'avec le reste du graphe.

Modularité La **modularité** de Girvan et Newman [10] est la plus répandue et la plus populaire des mesures de qualité de la partition statique d'un graphe en classes. Basée sur l'idée qu'un graphe aléatoire ne devrait démontrer de structure communautaire, elle compare la densité de liens d'un sous-graphe à celle obtenue si les sommets du graphe avaient été connectés de façon aléatoire. Elle se mesure de la façon suivante

$$mod(\mathcal{C}) = \frac{1}{2m} \sum_{u,v \in V} [A_{uv} - P_{uv}] \delta(u, v), \quad (1)$$

où P_{uv} est le nombre attendu d'arêtes entre u et v dans le graphe aléatoire et le delta de Kronecker $\delta(u, v)$ vaut 1 lorsque les sommets u et v appartiennent à la même communauté, 0 sinon. La modularité globale peut être généralisée aux graphes pondérés [11], bipartis [12, 13], ou encore au recouvrement de communautés [14, 15, 16].

Homogénéité, hétérogénéité Soit j l'origine géographique commune d'un sous-groupe d'individus classés dans la communauté C_i et s_{ij} la part originaire de j dans la communauté C_i . Alors,

$$\Upsilon_i = 1 - \sum_j s_{ij}^2$$

exprime la probabilité que deux individus choisis au hasard proviennent d'origines distinctes. Une communauté parfaitement homogène a un indice de 0. Plus la communauté est hétérogène, en particulier plus les portions qui la composent sont petites, plus l'indice se rapproche de 1. La moyenne sur toutes les communautés $C_i \in \mathcal{C}$ est utilisée pour définir l'homogénéité globale d'une partition (ou répartition) $\mathcal{C}$.

1.3 État de l'art

La détection de communautés dans des graphes captive de nombreux scientifiques depuis les vingt dernières années (voir Porter et al. [17] ou Fortunato et al. [18] pour des revues complètes de la littérature en contexte statique ; voir Bilgin et Yener [19], Aynaud et al. [20], Cazabet [21], Aggarwal et Subbian [22], Hartmann et al. [23] pour des revues partielles de la littérature en contexte dynamique). L'idée de classer en communautés des individus d'un réseau est dérivée du problème classique de partitionnement d'un graphe en k ensembles de taille similaire. Dans l'espoir d'étudier des graphes réels, la limite imposée par l'ignorance du paramètre k a mené Girvan et Newman [9] à poser le problème en termes de détection automatique de communautés sans connaissance préalable du réseau. Il existe un nombre imposant de méthodes de détection de communautés dans un graphe statique allant de l'optimisation de la modularité par programmation linéaire [24, 25], par recherche à voisinage variable [26], par recuit simulé [27], par optimisation spectrale [28, 29], par optimisation extrémale [30], par algorithme hiérarchique [31], agglomératif [32] ou glouton [33] ; aux marches aléatoires [34, 35, 36] ; à la percolation de cliques [37] ; ou encore à la propagation d'étiquettes (*label propagation*) [38, 39]. Notons plus particulièrement les algorithmes suivants :

- **Louvain** : L'algorithme agglomératif d'optimisation de la modularité de Blondel et al. [33] considère initialement chaque sommet comme une communauté de singleton. À la première itération, chacun des sommets est associé à celui de ses voisins occasionnant le plus grand gain

positif de modularité. Ces groupes forment dès lors des *super*-nœuds sur lesquels le même procédé est appliqué, à la seule différence que le gain de modularité est calculé localement sur le graphe initial. L'algorithme s'arrête lorsqu'il devient impossible d'augmenter la valeur de la modularité et retourne un dendrogramme représentant les niveaux hiérarchiques de la solution.

- **Infomap** : Rosvall et Bergstrom [35, 36] utilisent la théorie de l'information pour déduire la structure d'un réseau. Ainsi, un flot d'information aléatoire sur un graphe avec une structure modulaire sous-jacente aura tendance à rester coincé à l'intérieur d'une même communauté, alors qu'à l'opposé, il glissera librement dans un graphe aléatoire. Un identifiant est attribué à chacune des structures importantes du graphe (tel qu'il l'est fait pour une ville sur une carte routière, par exemple) ainsi qu'aux sommets qu'elles contiennent. Les noms des sommets sont réutilisés à l'intérieur des structures différentes permettant une compression de l'information plus grande que si chacun des nœuds avait eu un identifiant unique (comme le nom d'une rue qui peut être identique dans deux villes différentes). La détection de communautés dans un graphe se réduit donc au problème d'encodage minimal : soit celui de déduire une partition qui minimise la description d'une marche aléatoire sur un graphe.
- **OSLOM** : La méthode de Lancichinetti et al. [40] est basée sur l'optimisation locale d'une fonction d'ajustement exprimant à quel point les modules sont statistiquement significatifs par rapport aux fluctuations aléatoires. Dans un premier temps, une communauté C_i est prise au hasard et élargie progressivement par ajout ou retrait de nœuds selon la probabilité qu'ils ont d'être affiliés au même groupe par rapport à un graphe aléatoire. **OSLOM** mesure la probabilité que le nombre de liens *intra*-communauté pour un nœud u soit plus ou moins grand que le nombre d'arêtes entre u et les nœuds de C_i dans le modèle aléatoire correspondant. Dans ce cas, le sommet est considéré statistiquement significatif et ajouté à C_i . De plus, la structure interne de la communauté est validée de telle sorte que les nœuds non significatifs en soient retirés. Dans un second temps, **OSLOM** détermine s'il y a lieu d'unir des communautés entre elles. Ce processus est répété aléatoirement sur chacun des modules et aboutit à une répartition statistiquement significative des sommets.

La façon la plus simple d'intégrer la composante temporelle consiste à appliquer un quelconque algorithme statique indépendamment sur chacun des instantanés : il s'agit alors de résoudre un problème d'appariement entre les apparitions temporelles d'une même entité soit par la taille de leur intersection [41] ; par le suivi de représentants [42] ; ou par des mesures de similitudes [43].

Il est aussi possible d'exploiter l'information temporelle associée au réseau et ainsi de comprendre l'évolution d'une communauté sur plusieurs périodes au travers des étapes de son cycle de vie - naissance, croissance, contraction, mort, fusion, division - dont on doit le formalisme à Palla et al. [44] ou Asur et al. [45]. À cet effet, Chakrabarti et al. [46] ont défini le premier *cadre évolutif*³ prenant en compte à la fois la qualité statique et la qualité séquentielle de sorte que la partition idéale des sommets en communautés est un compromis entre la structure modulaire inhérente au réseau selon l'information immédiate et celle de la période précédente. Plusieurs adaptations de ce cadre existent et sont résolues par algorithme agglomératif hiérarchique et des k -moyennes [46, 47] ou génétique [48] ; par regroupement spectral [49, 50, 51] ; par des modèles génératifs probabilistes [52, 53] ; par des méthodes de densité-et-particule [54] ; etc. Pour pallier la contrainte combinatoire de recalculer la structure modulaire du graphe sur chaque instantané, d'autres ont proposé de mettre à jour localement les parties du graphe affectées par des modifications récentes, par exemple en initialisant un algorithme statique avec la solution de la période précédente [55, 56]. Dans le même ordre d'idée, **OSLOM** permet de passer en paramètre la solution de l'instantané précédent [40].

En supposant une structure modulaire $\mathcal{C} = \{C_1, C_2, \dots, C_k\}$ produite par un quelconque algorithme, les sommets peuvent alors être classés selon leur connectivité. Gimera et Amaral [57, 58]

déterminent le rôle d'un sommet u par la paire (z_u, P_u) avec z_u son degré *intra*-communauté

$$z_u = \frac{d_{C_i}^{in}(u) - \bar{d}_{C_i}^{in}}{\sigma_{C_i}^{in}},$$

où $\bar{d}_{C_i}^{in}$ et $\sigma_{C_i}^{in}$ sont la moyenne et l'écart-type des degrés internes de tous les sommets de C_i ; et P_u son coefficient de participation

$$P_u = 1 - \sum_{i=1}^{|C|} \left(\frac{d_{C_i}^{in}(u)}{K_{C_i}} \right),$$

ou K_i est le degré total de C_i . Pour $z_u \geq 2,5$, un sommet fortement connecté est dit *hub* soit *provincial* ($P_u \geq 0,3$), *connecteur* ($0,3 < P_u < 0,75$) ou *global* ($P_u > 0,75$). Pour $z_u < 2,5$, un sommet faiblement connecté est dit *non-hub* soit *ultrapériphérique* ($P_u \approx 0$), *périphérique* ($P_u < 0,625$), *connecteur* ($0,625 < P_u < 0,8$) ou *global* ($P_u > 0,8$). Une communauté sans *hub* est complètement décentralisée.

2 Données

La base de données est dérivée du site <http://src.chromium.org> et contient la liste de toutes les contributions au logiciel ouvert ⁴ Chrome. Elle contient 1 607 632 lignes de la forme

#person, issue, type, office, date

tel que

- *#person* est un code anonyme de trois lettres attribué à l'un des 1045 programmeurs participant occasionnellement et de façon volontaire au développement du logiciel;
- *issue* est un code numérique attribué à l'une des 324 860 *ChangeLists* (noté CL) ou mises à jour du logiciel;
- *type* est la description du rôle de l'individu parmi sept rôles possibles et classés en actions individuelles (*submitter*) ou interactives (*approver*, *commenter*, *commenter unsolicited*, *disapprove*, *reviewer*, *reviewer silent*);
- *office* identifie le lieu géographique où se trouve l'employé parmi 46 bureaux dans 22 pays;
- *date* est la date à laquelle l'entrée est enregistrée à l'intérieur d'un intervalle de 1974 jours.

Il existe des bureaux affectés à une spécificité de **Chrome** comme les fonctionnalités réseau, la sécurité ou le graphisme; autant qu'il existe des équipes inter-bureau ou des individus isolés. Ainsi, rien ne présuppose que la provenance géographique implique une nécessaire association entre les programmeurs.

Un CL comprend en moyenne 4,95 actions ou entrées (maximum 80, minimum 1) pas nécessairement attribuées à des programmeurs distincts, car un individu peut intervenir autant qu'il le veut sur un même CL dans autant de rôles différents. Un individu a fait en moyenne 1538,4 actions sur toute la période (maximum 40 300, minimum 10), soit environ une par jour ouvrable.

Les données couvrent la période de septembre 2008 à janvier 2014. Leur nombre atteint un sommet vers le printemps 2013; il décroît lentement depuis comme illustré à la figure 1a. Une interprétation possible est l'atteinte de la maturité de **Chrome**. Le portrait est le même lorsque les employés de US-MTV, le siège social, sont retirés parce que représentant une part disproportionnée de l'ensemble des programmeurs (42 %), voir la figure 1b. Le nombre d'individus actifs suit la même évolution, comme illustré à la figure 2, ce qui correspond à l'embauche de programmeurs additionnels pour composer avec la taille grandissante du logiciel **Chrome**; et la moyenne d'actions par employé est globalement stable quoique très variable comme illustré à la figure 3.

4. *Open source*

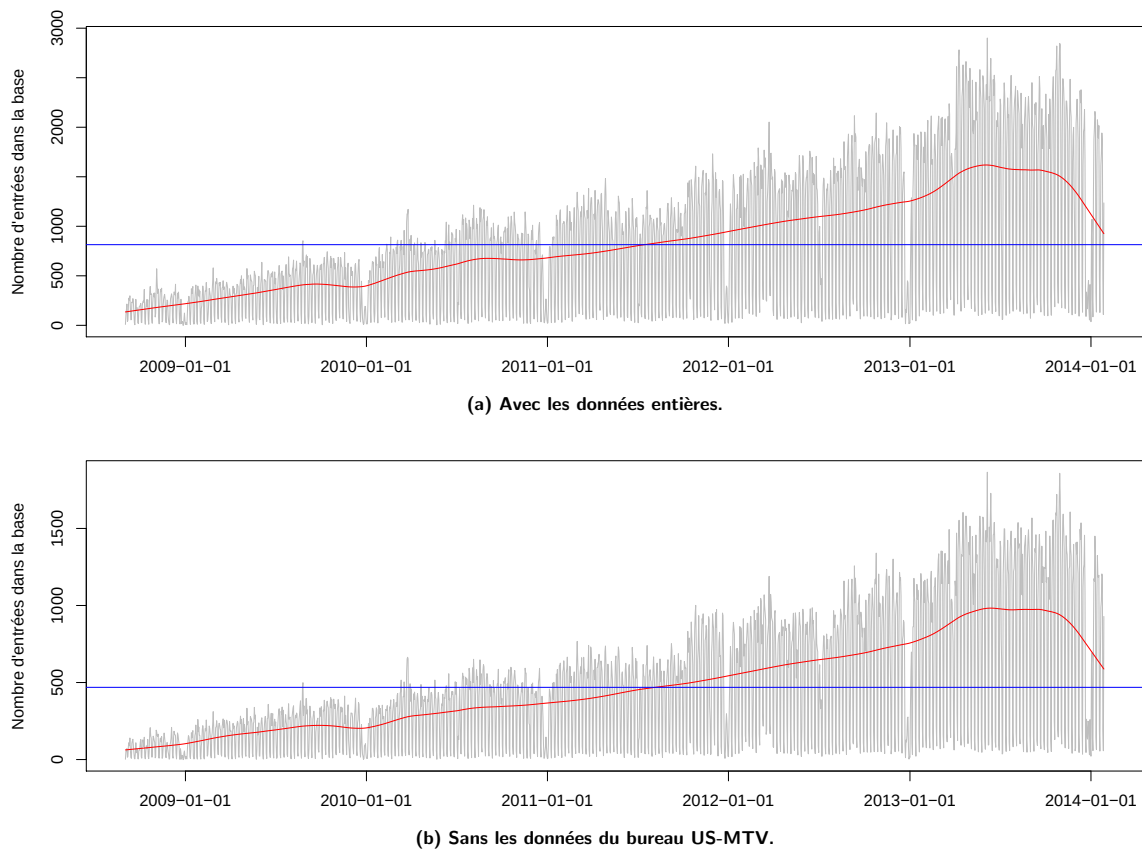


Figure 1 – Évolution du nombre d'entrées correspondant aux opérations de mise à jour de Chrome de septembre 2008 à janvier 2014.

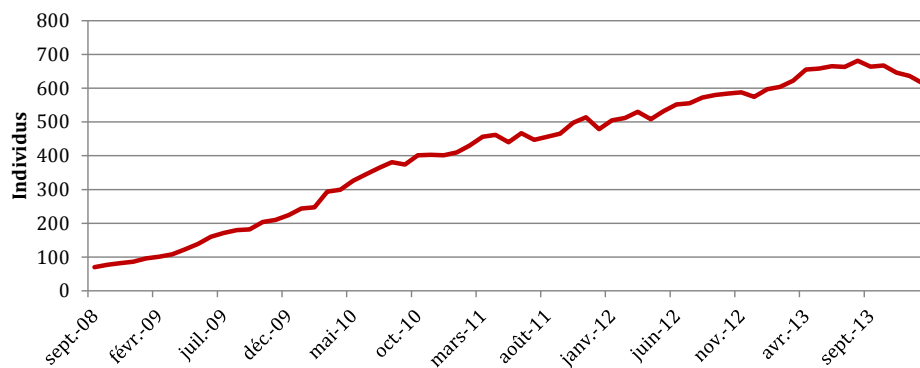


Figure 2 – Évolution du nombre de participants aux mise à jour de Chrome agrégé par mois de septembre 2008 à janvier 2014.

2.1 Analyse descriptive

Cette section s'intéresse au portrait de l'activité des programmeurs impliqués dans le développement de **Google Chrome** d'une part au niveau individuel et d'autre part au niveau collectif, lorsque regroupés par bureau. L'analyse est basée sur l'entièreté des données fournies par nos partenaires chez **Google**. La moyenne d'actions par individus - le nombre d'entrées dans la base de données associées à

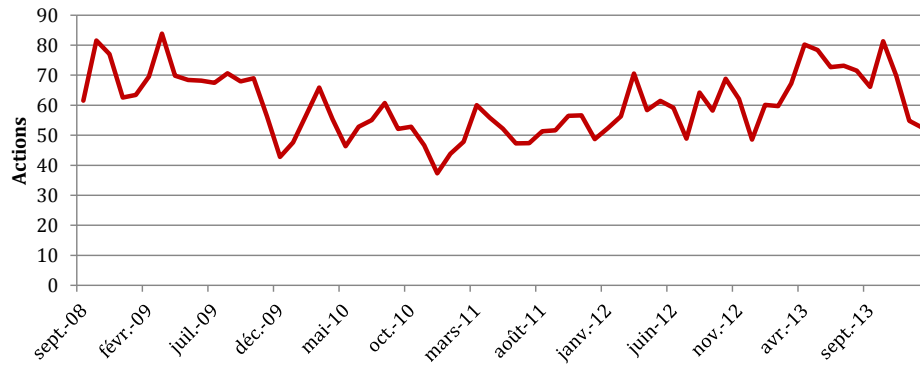


Figure 3 – Évolution de la moyenne d'action par individu de septembre 2008 à janvier 2014.

un des 1045 employés - est peu représentative du portrait réel du comportement des employés. En fait, l'activité de chaque individu pourrait plausiblement suivre une loi de puissance, distribution souvent observée dans la nature, selon Albert et Barabási [59] ou Clauset et al. [60], où une quantité x obéit à une loi de puissance si sa probabilité de distribution est de la forme

$$p(x) \propto x^{-\alpha},$$

où α est la constante de proportionnalité généralement dans l'intervalle $(2, 3)$. Cette hypothèse est soutenue par le graphe 4 représentant la distribution cumulative $P(x)$ et son ajustement par maximisation de la vraisemblance à une loi de puissance avec $\alpha = 2,92$. Ainsi, la grande majorité (95 %) des individus pose moins de 6000 actions $\sim$ quatre par jour ; 20 individus ont plus de 10 000 actions $\sim$ plus de sept/jour ; trois individus ont plus de 20 000 actions $\sim$ plus de quatorze par jour. En d'autres mots, le 20 % supérieur est responsable de 72 % de l'activité totale.

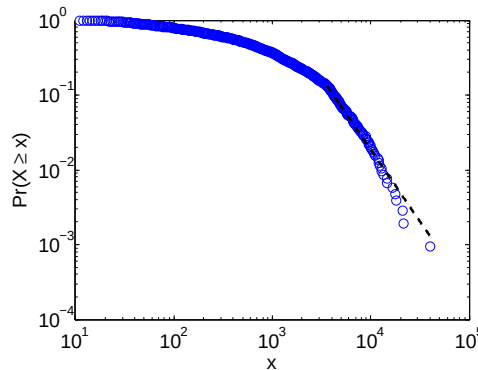


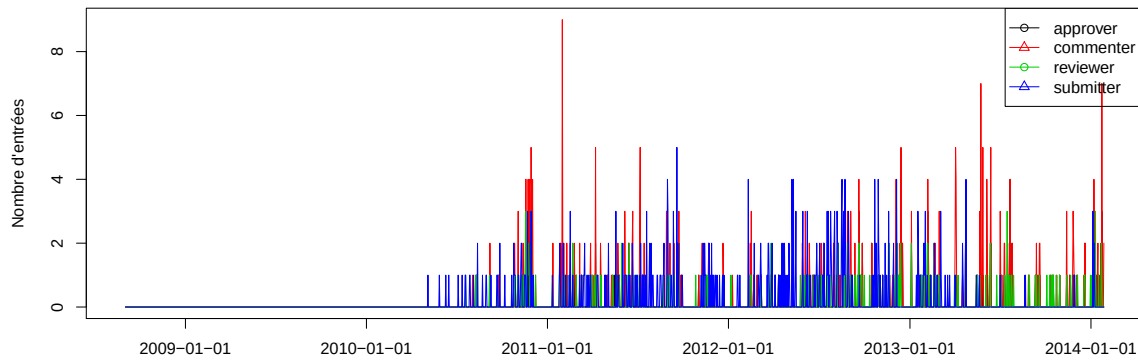
Figure 4 – Distribution cumulative $P(x)$, pour x = activité des individus enregistrée dans la base de données, et son ajustement par maximisation de la vraisemblance à une loi de puissance. $\alpha = 2,92$.

La nature de la participation par des programmeurs du premier décile, soit les individus peu actifs, est très différente de celle du décile supérieur, soit les individus très actifs, de sorte que se dessinent deux profils appelés pour la cause *novice* et *expérimenté*. Le *novice* est plus porté à un comportement individuel qu'à un interactif, à l'inverse de l'*expérimenté*. Il semble qu'un individu dont la participation est très récente ou sporadique s'implique individuellement sans réelle collaboration alors qu'un individu très actif, très expérimenté aura plus tendance à regarder le travail des autres ou à collaborer. Plus encore, la période d'activité est positivement corrélée à la propension à jouer les rôles *approuver*, *commenter*, *reviewer* mais négativement corrélée à la propension à *soumettre* ; c'est-à-dire

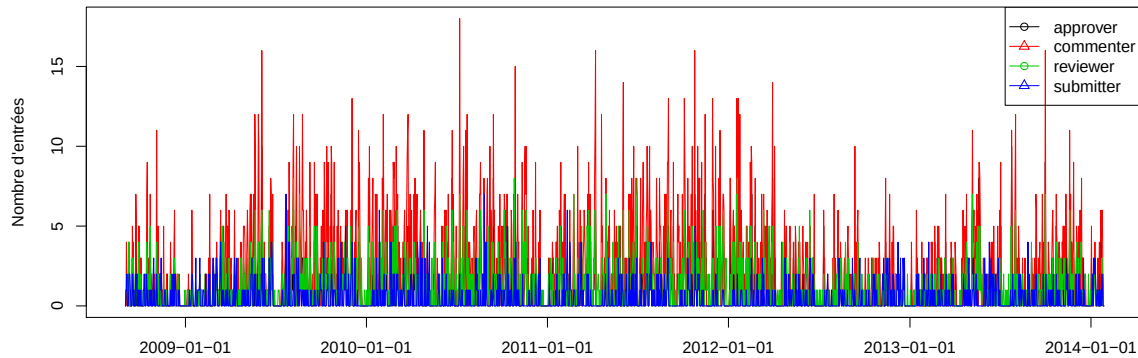
que les individus dont l'activité est récente sont plus portés à être les initiateurs d'une mise à jour qu'à tout autre rôle à l'inverse des individus dont l'activité remonte à plus loin (voir les détails des corrélations dans le tableau 1 ci-joint). Les figures 5a et 5b illustrent ces remarques par deux exemples d'activités d'individus respectivement *novice* et *expérimenté*. Notons que nos contacts chez **Google** ont indiqué qu'il serait possible de trouver dans la nature du travail pour le logiciel **Chrome** des éléments corroborant nos observations.

Table 1 – Coefficients de corrélation entre la période d'activité et la propension à jouer l'un des sept rôles possibles pour un individu.

Rôle	Coefficients de corrélation
<i>approver</i>	0,4776
<i>commenter</i>	0,4418
<i>reviewer</i>	0,4624
<i>commenter unsolicited</i>	-0,1138
<i>disapprover</i>	0,0357
<i>reviewer silent</i>	-0,0443
<i>submitter</i>	-0,4929



(a) Portrait d'un programmeur moins expérimenté. Les soumissions sont plus fréquentes au début de la période d'activité.



(b) Portrait d'un programmeur très expérimenté. Les soumissions sont moins fréquentes sur toute la période d'activité.

Figure 5 – Évolution du nombre d'entrées en tant que submitter, approuver, commenter, reviewer septembre 2008 à janvier 2014.

Par ailleurs, chaque CL fait intervenir un certain nombre d'employés de sorte que 90 % des CL se font avec un, deux ou trois individus. Il pourrait exister un plafond au nombre de participants au-delà duquel la collaboration devient difficile : la question pourrait être étudiée.

À l'échelle mésoscopique, il existe une énorme disparité dans le nombre d'employés par bureau ce qui pourrait fausser toute comparaison en fonction de l'appartenance géographique. D'une part, l'activité sur tout l'horizon temporel à l'intérieur de chacun des bureaux ne suit pas la même courbe,

en particulier certains bureaux participent sporadiquement (US-SMO, CN-PEK, CZ-PRG, FI-OUL, etc.), d'autres ont réduit leur participation (IL-TLV, IN-HYD, US-RES, US-SBO, etc.). Notons aussi que US-DAL, US-SBO, US-SMO, CA-TOR, entre autres, montrent un pic d'activité à l'automne 2010. Enfin US-PDX, US-PIT, IE-DUB, US-ARB, KR-SEO n'ont entamé leur participation qu'après 2011. D'autre part, il n'existe aucune corrélation entre le nombre d'individus par bureau et l'activité ou le nombre de collaborateurs, de sorte que les individus ne sont pas restreints par le fait d'avoir peu de collègues du même bureau. Enfin, comme illustré à la figure 6, pour 49 % des CL, les participants proviennent du même bureau; pour 41 % des CL, les participants proviennent d'exactly deux bureaux; pour 11 % des CL, les participants proviennent de trois ou plus bureaux différents; en particulier, 99 % des CL se font à l'intérieur de un, deux ou trois bureaux. Sans les programmeurs de US-MTV : pour 70 % des CL, les participants proviennent du même bureau; pour 26 % des CL, les participants proviennent d'exactly deux bureaux; pour 3 % des CL, les participants proviennent de trois ou plus bureaux différents; en particulier, 99 % des CL se font à l'intérieur de un, deux ou trois bureaux. En conséquence, US-MTV est en quelque sorte un liant entre les différents bureaux et il est probable que les communautés aient une composante géographique.

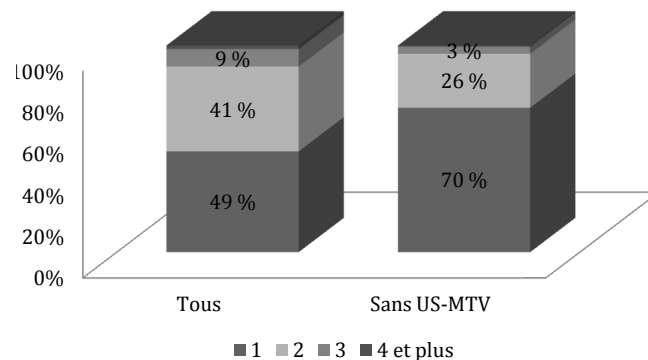


Figure 6 – Parts des CL dont les participants proviennent de un, deux, trois bureaux ou plus.

2.2 Modélisation

Un graphe est construit à partir du jeu de données dans lequel les arêtes entre les sommets expriment la collaboration à un CL. Les sections qui suivent proposent trois modélisations aux propriétés distinctes sur lesquelles seront détectées les communautés. Dans tous les cas, les relations sont considérées comme réciproques et de nature équivalente peu importe le rôle des participants; par conséquent, les arêtes sont non-orientées et initialement non-pondérées.

Modélisation en graphe biparti Dans un graphe biparti, les sommets peuvent appartenir à deux classes différentes; dans le cas qui nous intéresse, il s'agit des classes *issue* et *# person* telles qu'il n'existe de liens entre deux individus ou entre deux mises à jour. Le poids d'une arête est d'au plus un, de sorte que le graphe biparti est non pondéré. La modélisation biparti a l'avantage de conserver l'information sur la nature du lien entre les individus une fois les communautés détectées, c'est-à-dire la possibilité de retrouver au travers de quelles mises à jour certains employés ont collaboré et ainsi faciliter l'analyse et la catégorisation *a posteriori* des communautés détectées.

Projection de la modélisation en graphe biparti La projection du graphe biparti consiste à éliminer les sommets *issue*. Soient u un sommet d'une classe à retirer et v et w des sommets de l'autre classe et tels qu'il existe un lien entre u et v et entre u et w dans la modélisation biparti du graphe; la projection est telle qu'un lien est créé entre v et w , puis u est éliminé. Dans la version pondérée du graphe, si deux individus ont collaboré à plus d'une reprise à la même mise à jour, plusieurs arêtes

unitaires existent entre les sommets les représentant. Ils sont alors agrégés en un seul lien dont le poids est la somme des arêtes unitaires.

Modélisation en graphe étoilé La modélisation de la sous-section précédente ne fait aucune différence dans les rôles adoptés par les employés dans leur participation à une mise à jour. En particulier, elle accorde la même importance dans la relation entre l'initiateur de la mise à jour et tout autre individu qu'entre individus *approver*, *commenter*, *commenter unsolicited*, *disapprover*, *reviewer*, *reviewer silent*. Cependant, rien ne confirme l'existence d'un rapport quelconque entre ces rôles accessoires, alors qu'un rapport semble bien exister entre l'auteur de la mise à jour et tout autre individu y ayant participé. Dans la modélisation en graphe étoilé, pour chaque mise à jour, des arêtes unitaires sont affectées uniquement entre le sommet représentant l'initiateur et tout autre sommet représentant un participant. Dans la version pondérée du graphe, les liens incidents aux deux mêmes sommets sont agrégés en une seule arête dont le poids est la somme des arêtes unitaires.

Instantanés de graphe La création d'un graphe instantané $G^{(t)}$ modélisé selon l'un des trois cas - soient biparti, étoilé ou projeté - ne consiste qu'à partitionner la base de données selon la date de chacune des entrées à l'intérieur d'un intervalle $(t - 1, t]$. Ici, le jeu de données horodatés sur 1974 jours est découpé en onze intervalles couvrant des périodes de 182 jours ($\sim$ six mois) chacun. Le choix de la période repose sur la nécessité d'un nombre d'intervalles d'une part suffisamment grand pour comprendre la chronologie des événements sur les communautés et pour différencier la solution de celle de la partition statique; et d'autre part suffisamment petit pour ne pas représenter une contrainte computationnelle. Notons que les arêtes et sommets ne sont pas cumulés d'un intervalle de temps à l'autre et qu'il s'agit bien d'instantanés disjoints. Dans le cas contraire, des nœuds pourraient survivre sur tout l'horizon temporel alors qu'ils n'apparaissent que sporadiquement, faussant ainsi l'interprétation des communautés détectées. De plus, la partition finale serait dans les faits équivalente à la partition statique et beaucoup d'information sur l'évolution des communautés serait alors diluée dans une solution intertemporelle.

2.3 Ébauche des résultats

Les données sont nettoyées, de sorte à éliminer les CL ne faisant intervenir qu'un seul individu, et modélisées sous forme de graphes biparti, projeté et étoilé. Les trois réseaux sont de natures assez différentes, mais, dans tous les cas, les graphes sont connexes même lorsque découpés en tranches temporelles.

En analysant de nombreux réseaux sociaux, Albert et Barabási [8] ont observé que 1) les sommets d'un réseau de collaboration suivent une loi de puissance avec une traîne plus épaisse; 2) il existe des sommets avec un très grand degré. En fait, Barabási et al. [61, 62] argumentent que les interactions sociales sont caractérisées par des crues d'occurrences suivies de longues périodes d'inactivité de sorte qu'elles peuvent être modélisées plus fidèlement par une distribution en loi de puissance avec une traîne épaisse. Les répartitions des degrés pour chacune des modélisations pourraient suivre une loi de puissance, voir l'ajustement des données à cette distribution dans les figures 7a, 7b et 7c, mais aussi une loi log-normale (les observations pour des grandes valeurs de x , soit la queue de la série, semblent bien moins s'ajuster à la distribution théorique de la loi de puissance). La majorité des sommets ont un faible degré, c'est-à-dire que la majorité des individus sont peu connectés et que les traînes épaisses démontrent l'existence de sommets au degré très élevé, c'est-à-dire d'individus très connectés. Dans les faits, les données semblent dépeindre une situation quelque part entre la règle générale⁵ du 1 % (1 % des participants contribuent activement, 9 % sporadiquement, 90 % très rarement) et le principe de Pareto [63] (20 % des participants sont responsables de 80 % des contributions).

Nous avons testé une dizaine d'algorithmes de natures différentes, mais au final seuls les structures modulaires de **Louvain**, **OSLOM** et **Infomap** sur les onze instantanés avec appariement ont été re-

5. Rule of thumb

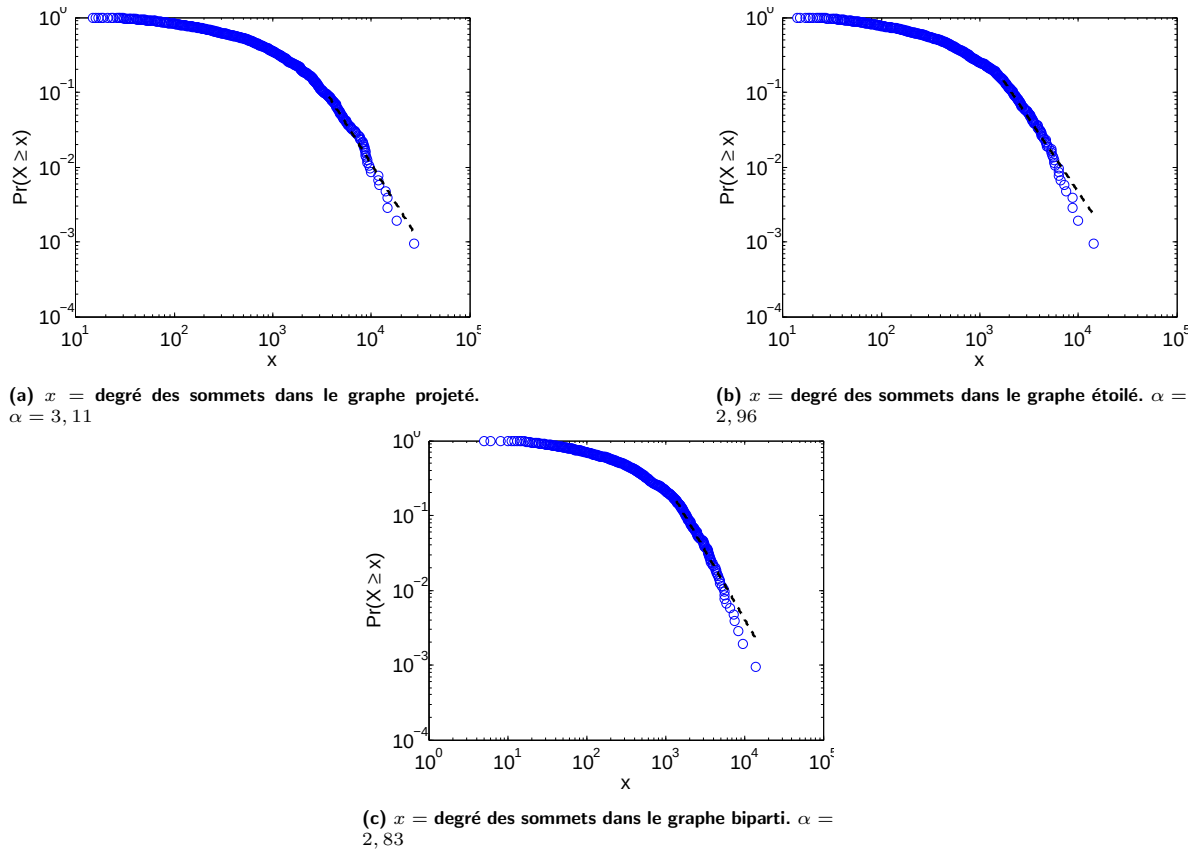


Figure 7 – Distribution cumulative $P(x)$ et son ajustement par maximisation de la vraisemblance à une loi de puissance.

tenus et les communautés de chaque intervalle ont été appariées avec l'indice de similitude de Takaffoli et al. [43], voir [1] pour plus de détails. Les résultats exposés sont basés sur un consensus déduit des multiples itérations des trois algorithmes sur les trois modélisations possibles. En fait, l'évaluation quantitative des solutions est un enjeu majeur de la détection de communautés : il n'existe aucune mesure capable de départager les résultats autrement qu'en disposant au préalable de la vérité topologique du réseau étudié et, bien évidemment, nous ne connaissons pas la véritable structure communautaire des employés chez **Google Chrome**. Nous sommes dès lors forcés de chercher des confirmations d'événements observés dans une solution parmi toutes les autres. Ainsi, chaque algorithme raconte une histoire et bien que sa trame narrative soit la même pour tous - il s'agirait de la vérité inhérente à la base de données - ses détails peuvent varier.

Nous constatons alors que le choix des collaborations n'est pas aléatoire :

- il existe des communautés observables sur des périodes plus ou moins longues ;
- le nombre de communauté croît avec l'augmentation du nombre de participants ;
- les communautés sont plus clairement définies avec le temps ;
- certaines communautés ont une interprétation locale ou régionale et d'autres non ;
- la majorité des communautés sont formées de sous-groupes de programmeurs provenant d'un petit nombre de bureau ;
- certaines communautés sont formées de programmeurs d'un seul et même bureau ;
- avec l'expérience, un employé devient plus important à l'intérieur de sa communauté et/ou en tant que lien avec les autres communautés (connecteur) ;
- certains individus sont clairement classés, d'autres le sont avec ambiguïté ou peuvent appartenir à plusieurs groupes simultanément ;

- lorsque l’affiliation multiple est considérée, la part moyenne d’une communauté en provenance de US-MTV augmente, semblant appuyer le lien de centralisation ou de dépendance entre US-MTV et les autres bureaux.

Plus en détail, il semble que les employés se regroupent par bureau, ce qui n’est pas si surprenant étant donné que 49 % des CL se règlent à l’interne. La majorité des programmeurs de n’importe laquelle des communautés est dans une même communauté que des membres de US-MTV, mais il existe des communautés d’employés de US-MTV exclusivement. Autre point, les programmeurs de US-PDX et US-PIT sont toujours ensemble dans une communauté décentralisée et isolée apparue fin 2011 (en rose à la figure 8) et ce jusque la fin de l’horizon temporel. Aussi, il existe une communauté formée d’individus initialement de DK-AAR laquelle perdure sur les six années (en rouge à la figure 8), mais évolue de sorte que se joignent à elle des membres de RU-STP et DE-MUC vers fin 2010, qui finiront par rejoindre ou former d’autres communautés. Ainsi, DK-AAR collabore avec peu d’autres bureaux, en particulier des bureaux nord-américains et les 23 programmeurs de DK-AAR sont presque toujours regroupés dans une même communauté assez centralisée. Elle l’est cependant de moins en moins avec le temps. Autre exemple, les employés de US-SEA et US-KIR forment ensemble plusieurs sous-groupes (en orange et vert à la figure 8). Certains sont des communautés à part entière alors que d’autres sont incorporés à d’autres communautés.

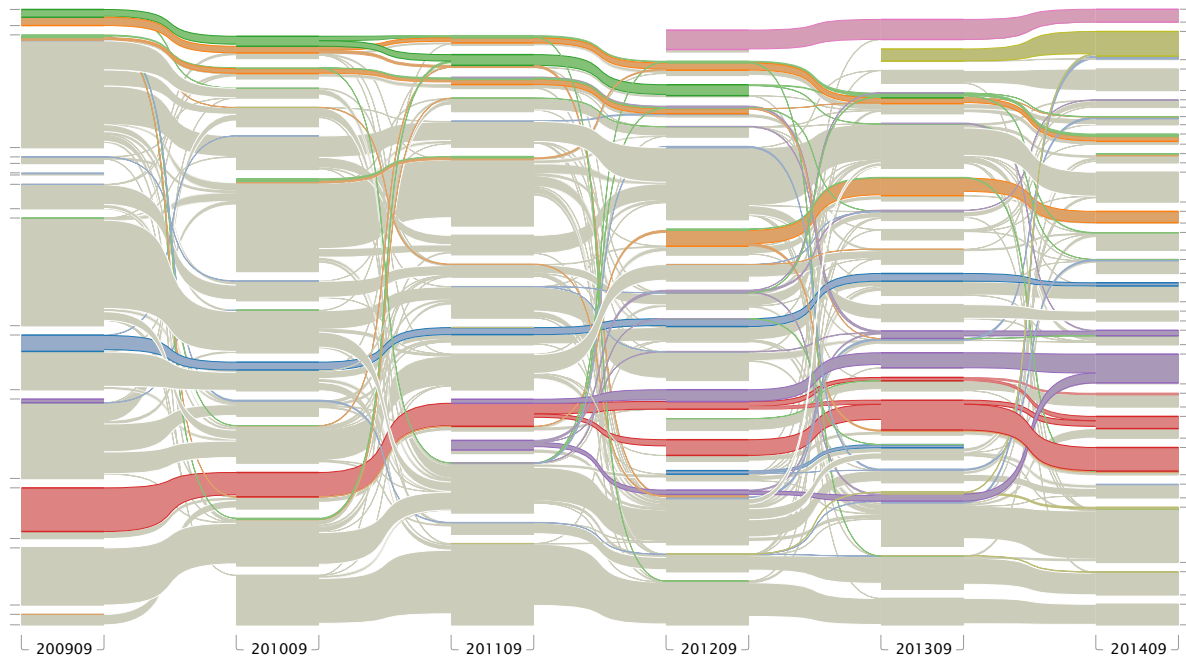


Figure 8 – Diagramme alluvial représentant l’évolution des communautés. Il est possible de suivre l’évolution de certaines communautés par leur apparition, fusion, division, etc. Mentionnons qu’il s’agit d’une vue partielle du réseau et que la largeur des bandes n’est pas proportionnelle à la taille des communautés mais au volume de flot. US-SEA : orange. US-KIR : vert. DK-AAR : rouge. RU-STP : violet. US-PIT/PDX : rose. CA-MON : bleu.

À partir des partitions produites par les différents algorithmes, nous tentons de déterminer les facteurs d’efficacité en termes de participation aux CL d’une communauté. Ainsi, un groupe efficace doit avoir une moyenne d’actions par individu supérieure à la moyenne sur toutes les communautés. Nous tentons d’établir des corrélations entre l’intensité de l’activité d’une communauté et d’autres mesures : la taille, la composition, l’homogénéité, la part de US-MTV, la présence de *hub*, la nature des *hub* ou *non-hub*, etc.

Table 2 – Coefficients de corrélation entre différentes quantités à partir des structures modulaires produites par trois algorithmes dans les trois modélisations.

	Activité	Homogénéité
Taille	0,0612	
Homogénéité	-0,0104	
Proportion de US-MTV	-0,2066	-0,2586
Proportion de <i>hub</i>	-0,1854	-0,3284
- <i>provincial</i>	-0,0818	-0,0810
- <i>connecteur</i>	-0,1897	-0,0876
- <i>global</i>	-0,1062	-0,2930
Proportion de <i>non-hub</i>	0,1854	0,3284
- <i>ultrapériphérique</i>	0,2840	0,0388
- <i>périphérique</i>	0,0312	-0,0158
- <i>connecteur</i>	-0,2280	0,2417
- <i>global</i>	-0,1432	0,0777
Ratio <i>provincial</i> / <i>connecteur</i>	0,2832	-0,0232

Nous arrivons aux observations suivantes

- Il ne semble pas y avoir de lien entre l’homogénéité d’une communauté et l’intensité de son activité ;
- Les communautés sont de moins en moins actives sur la première moitié, puis de plus en plus actives sur la seconde moitié de la période observée, comme illustré à la figure 9 ;
- Plus la part de US-MTV est grande, moins la communauté semble active ;
- Plus la communauté est décentralisée, plus la communauté est active ;
- Plus les *hubs* sont régionaux versus globaux, plus la communauté est active ;
- Si l’affiliation multiple est considérée (donc les communautés ont une proportion plus grande en provenance de US-MTV), plus les communautés sont grandes, plus elles sont actives ;
- Plus la communauté est homogène, plus elle est centralisée.

Il apparaîtrait que les groupes plus efficaces seraient plus petits, plus homogènes, décentralisés et moins en contact avec US-MTV. Cependant, la réciproque est fautive : il existe des communautés répondant exactement à ces critères et qui sont pourtant peu efficaces.

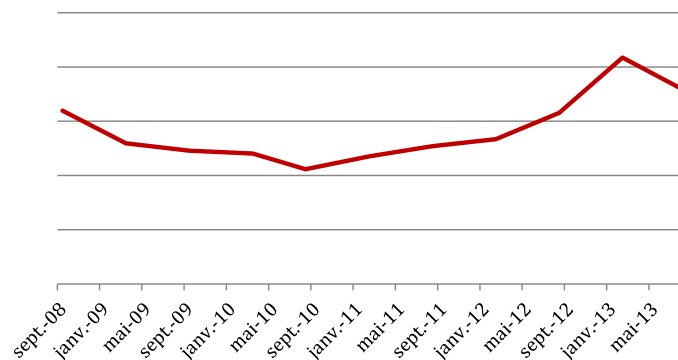


Figure 9 – Activité moyenne des communautés.

3 Conclusion et travaux futurs

Après consultation auprès des partenaires chez **Google Chrome**, nous avons soulevé certaines limites du jeu de données fourni. D’une part, l’affiliation d’un individu à un bureau est celle au moment où la base a été compilée, non pas celle au moment où l’entrée est enregistrée dans la base. Par

conséquent, un employé d'un bureau i pourrait être classé dans une communauté au temps t avec d'autres d'un bureau j et apparaître comme valeur aberrante, alors qu'il faisait, à t , partie de j et qu'il a par la suite déménagé vers i . D'autre part, nous cherchons à établir l'efficacité d'une communauté, cependant le nombre de CL n'est peut-être pas le meilleur indicateur de performance si leur niveau de difficulté diffère. Par exemple, un CL pourrait prendre l'équivalent de dix autres CL en termes de temps et pourtant être comptabilisé avec la même valeur. Nous sommes en attente d'une base de données révisée dans laquelle les points ci-dessus seront pris en compte. Si les solutions calculées à partir de la base de données originale restent bonnes, la correction des quelques erreurs d'affiliations et l'intégration de métadonnées permettront à notre recherche de s'ouvrir sur de nouvelles avenues prometteuses. En particulier, si nous tentons de dessiner le portrait de l'organisation optimale d'une communauté de collaboration depuis sa formation du point de vue du profil de chacun de ses participants, de sa structure interne ou de ses liens avec les autres communautés, il est primordial d'établir des critères d'efficacité robustes.

Références

- [1] A. Correc, Comparaison d'algorithmes de détection de communautés dynamiques dans les réseaux évolutifs, Master's thesis, HEC Montréal (2015).
- [2] S. Wasserman, Social network analysis : Methods and applications, Vol. 8, Cambridge university press, 1994.
- [3] E. Bullmore, O. Sporns, Complex brain networks : graph theoretical analysis of structural and functional systems, *Nature Reviews Neuroscience* 10 (3) (2009) 186–198.
- [4] J. A. Dunne, R. J. Williams, N. D. Martinez, Food-web structure and network theory : the role of connectance and size, *Proceedings of the National Academy of Sciences* 99 (20) (2002) 12917–12922.
- [5] D. J. de Solla Price, D. J. de Solla Price, D. J. de Solla Price, D. J. de Solla Price, Little science, big science... and beyond, Columbia University Press New York, 1986.
- [6] R. De Castro, J. W. Grossman, Famous trails to paul erdős, *The Mathematical Intelligencer* 21 (3) (1999) 51–53.
- [7] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph structure in the web, *Computer networks* 33 (1) (2000) 309–320.
- [8] R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, *Reviews of modern physics* 74 (1) (2002) 47.
- [9] M. Girvan, M. E. Newman, Community structure in social and biological networks, *Proceedings of the National Academy of Sciences* 99 (12) (2002) 7821–7826.
- [10] M. E. Newman, M. Girvan, Finding and evaluating community structure in networks, *Physical review E* 69 (2) (2004) 026113.
- [11] M. E. Newman, Analysis of weighted networks, *Physical Review E* 70 (5) (2004) 056131.
- [12] M. J. Barber, Modularity and community detection in bipartite networks, *Physical Review E* 76 (6) (2007) 066102.
- [13] R. Guimerà, M. Sales-Pardo, L. A. N. Amaral, Module identification in bipartite and directed networks, *Physical Review E* 76 (3) (2007) 036102.
- [14] V. Nicosia, G. Mangioni, V. Carchiolo, M. Malgeri, Extending the definition of modularity to directed graphs with overlapping communities, *Journal of Statistical Mechanics : Theory and Experiment* 2009 (03) (2009) P03024.
- [15] D. Chen, M. Shang, Z. Lv, Y. Fu, Detecting overlapping communities of weighted networks via a local algorithm, *Physica A : Statistical Mechanics and its Applications* 389 (19) (2010) 4177–4187.
- [16] H. Shen, X. Cheng, K. Cai, M.-B. Hu, Detect overlapping and hierarchical community structure in networks, *Physica A : Statistical Mechanics and its Applications* 388 (8) (2009) 1706–1712.
- [17] M. A. Porter, J.-P. Onnela, P. J. Mucha, Communities in networks, *Notices of the AMS* 56 (9) (2009) 1082–1097.
- [18] S. Fortunato, Community detection in graphs, *Physics Reports* 486 (3) (2010) 75–174.

- [19] C. C. Bilgin, B. Yener, Dynamic network evolution : Models, clustering, anomaly detection, IEEE Networks (2006).
- [20] T. Aynaud, E. Fleury, J.-L. Guillaume, Q. Wang, Communities in evolving networks : Definitions, detection, and analysis techniques, in : Dynamics On and Of Complex Networks, Volume 2, Springer, 2013, pp. 159–200.
- [21] R. Cazabet, Détection de communautés dynamiques dans des réseaux temporels, Ph.D. thesis, Université Paul Sabatier-Toulouse III (2013).
- [22] C. Aggarwal, K. Subbian, Evolutionary network analysis : A survey, ACM Computing Surveys (CSUR) 47 (1) (2014) 10.
- [23] T. Hartmann, A. Kappes, D. Wagner, Clustering evolving networks, arXiv preprint arXiv :1401.3516 (2014).
- [24] D. Aloise, S. Cafieri, G. Caporossi, P. Hansen, S. Perron, L. Liberti, Column generation algorithms for exact modularity maximization in networks, Physical Review E 82 (4) (2010) 046112.
- [25] G. Agarwal, D. Kempe, Modularity-maximizing graph communities via mathematical programming, The European Physical Journal B-Condensed Matter and Complex Systems 66 (3) (2008) 409–418.
- [26] S. Cafieri, P. Hansen, L. Liberti, Edge ratio and community structure in networks, Physical Review E 81 (2) (2010) 026105.
- [27] R. Guimerà, M. Sales-Pardo, L. A. N. Amaral, Modularity from fluctuations in random graphs and complex networks, Physical Review E 70 (2) (2004) 025101.
- [28] M. E. Newman, Finding community structure in networks using the eigenvectors of matrices, Physical review E 74 (3) (2006) 036104.
- [29] J. Ruan, W. Zhang, Identifying network communities with a high resolution, Physical Review E 77 (1) (2008) 016104.
- [30] J. Duch, A. Arenas, Community detection in complex networks using extremal optimization, Physical review E 72 (2) (2005) 027104.
- [31] M. E. Newman, Fast algorithm for detecting community structure in networks, Physical review E 69 (6) (2004) 066133.
- [32] A. Clauset, M. E. Newman, C. Moore, Finding community structure in very large networks, Physical review E 70 (6) (2004) 066111.
- [33] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, Journal of Statistical Mechanics : Theory and Experiment 2008 (10) (2008) P10008.
- [34] P. Pons, M. Latapy, Computing communities in large networks using random walks, in : Computer and Information Sciences-ISCIS 2005, Springer, 2005, pp. 284–293.
- [35] M. Rosvall, C. T. Bergstrom, Maps of random walks on complex networks reveal community structure, Proceedings of the National Academy of Sciences 105 (4) (2008) 1118–1123.
- [36] M. Rosvall, D. Axelsson, C. T. Bergstrom, The map equation, The European Physical Journal-Special Topics 178 (1) (2009) 13–23.
- [37] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (7043) (2005) 814–818.
- [38] U. N. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, Physical Review E 76 (3) (2007) 036106.
- [39] S. Gregory, Finding overlapping communities in networks by label propagation, New Journal of Physics 12 (10) (2010) 103018.
- [40] A. Lancichinetti, F. Radicchi, J. J. Ramasco, S. Fortunato, Finding statistically significant communities in networks, PloS one 6 (4) (2011) e18961.
- [41] M. Spiliopoulou, I. Ntoutsis, Y. Theodoridis, R. Schult, Monic : modeling and monitoring cluster transitions, in : Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2006, pp. 706–711.
- [42] Y. Wang, B. Wu, N. Du, Community evolution of social network : feature, algorithm and model, arXiv preprint arXiv :0804.4356 (2008).
- [43] M. Takaffoli, F. Sangi, J. Fagnan, O. R. Zäiane, Community evolution mining in dynamic social networks, Procedia-Social and Behavioral Sciences 22 (2011) 49–58.

- [44] G. Palla, A.-L. Barabási, T. Vicsek, Quantifying social group evolution, *Nature* 446 (7136) (2007) 664–667.
- [45] S. Asur, S. Parthasarathy, D. Ucar, An event-based framework for characterizing the evolutionary behavior of interaction graphs, *ACM Transactions on Knowledge Discovery from Data (TKDD)* 3 (4) (2009) 16.
- [46] D. Chakrabarti, R. Kumar, A. Tomkins, Evolutionary clustering, in : *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2006, pp. 554–560.
- [47] K. S. Xu, M. Kliger, A. O. Hero III, Adaptive evolutionary clustering, *Data Mining and Knowledge Discovery* 28 (2) (2014) 304–336.
- [48] F. Folino, C. Pizzuti, A multiobjective and evolutionary clustering method for dynamic networks, in : *Advances in Social Networks Analysis and Mining (ASONAM)*, 2010 International Conference on, IEEE, 2010, pp. 256–263.
- [49] Y. Chi, X. Song, D. Zhou, K. Hino, B. L. Tseng, Evolutionary spectral clustering by incorporating temporal smoothness, in : *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2007, pp. 153–162.
- [50] S. X. Yu, J. Shi, Multiclass spectral clustering, in : *Computer Vision*, 2003. *Proceedings. Ninth IEEE International Conference on*, IEEE, 2003, pp. 313–319.
- [51] K. S. Xu, M. Kliger, A. O. Hero III, Tracking communities in dynamic social networks, in : *Social Computing, Behavioral-Cultural Modeling and Prediction*, Springer, 2011, pp. 219–226.
- [52] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, B. L. Tseng, Analyzing communities and their evolutions in dynamic social networks, *ACM Transactions on Knowledge Discovery from Data (TKDD)* 3 (2) (2009) 8.
- [53] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, B. L. Tseng, Facetnet : a framework for analyzing communities and their evolutions in dynamic networks, in : *Proceedings of the 17th international conference on World Wide Web*, ACM, 2008, pp. 685–694.
- [54] M.-S. Kim, J. Han, A particle-and-density based evolutionary clustering method for dynamic networks, *Proceedings of the VLDB Endowment* 2 (1) (2009) 622–633.
- [55] T. Aynaud, 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*, IEEE, 2010, pp. 513–519.
- [56] D. Greene, D. Doyle, P. Cunningham, Tracking the evolution of communities in dynamic social networks, in : *Advances in Social Networks Analysis and Mining (ASONAM)*, 2010 International Conference on, IEEE, 2010, pp. 176–183.
- [57] R. Guimera, L. A. N. Amaral, Functional cartography of complex metabolic networks, *Nature* 433 (7028) (2005) 895–900.
- [58] R. Guimera, L. A. N. Amaral, Cartography of complex networks : modules and universal roles, *Journal of Statistical Mechanics : Theory and Experiment* 2005 (02) (2005) P02001.
- [59] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, *science* 286 (5439) (1999) 509–512.
- [60] A. Clauset, C. R. Shalizi, M. E. Newman, Power-law distributions in empirical data, *SIAM review* 51 (4) (2009) 661–703.
- [61] A.-L. Barabasi, The origin of bursts and heavy tails in human dynamics, *Nature* 435 (7039) (2005) 207–211.
- [62] A. Vázquez, J. G. Oliveira, Z. Dezső, K.-I. Goh, I. Kondor, A.-L. Barabási, Modeling bursts and heavy tails in human dynamics, *Physical Review E* 73 (3) (2006) 036127.
- [63] V. Pareto, *Manuale di economia politica*, Vol. 13, Societa Editrice, 1906.