THÈSE DE DOCTORAT DE
L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité

INFORMATIQUE

Présentée par

LAMIA BENAMARA

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Dynamique des graphes de terrain : caractérisation et étude du biais lié à la mesure

Thèse dirigée par
Clémence MAGNIEN

Soutenance : le 29 Novembre 2011

Jury :

Rapporteurs :	Eric FLEURY	-	Professeur, ENS de Lyon
	Thomas NOËL	-	Professeur, Université de Strasbourg
Examinateurs :	Pierre BORGNAT	-	Chargé de recherche CNRS, ENS de Lyon
	Alessandra CARBONE	-	Professeur, UPMC
	Bertrand DUCOURTHIAL	-	Professeur, Université de Compiègne
Directrice :	Clémence MAGNIEN	-	Chargée de recherche CNRS, UPMC

Remerciements

Une thèse est le fruit d'un travail collectif. C'est la raison pour laquelle je tiens à remercier ici toutes les personnes m'ayant aidée tout au long de mon parcours.

Tout d'abord, je remercie vivement mon encadrante et directrice de thèse Clémence Magnien pour m'avoir donné la chance de réaliser ma thèse dans les conditions les plus favorables. Je la remercie pour m'avoir guidée, conseillée, soutenue et pour avoir été disponible tout au long de ma thèse.

J'aimerais aussi exprimer toute ma gratitude à Éric Fleury et Thomas Noël pour avoir accepté d'être rapporteurs de cette thèse. Je remercie aussi vivement Pierre Borgnat, Alessandra Carbone et Bertrand Ducourthial qui m'ont fait l'honneur de faire partie de mon jury de thèse.

Je tiens à remercier également tous les membres de l'équipe "Complex Networks" du Lip6, en particulier Mathieu Latapy grâce à qui j'ai pu rejoindre l'équipe, Jean-loup Guillaume qui a accepté de relire ma thèse, Bénédicte et Fabien pour leur soutien et leur aide. Je remercie également tous les thésards et les post-doctorants qui m'ont accompagnée pendant ces années et grâce à qui j'ai travaillé dans une bonne ambiance. Merci à eux pour leur gentillesse et les nombreuses discussions partagées.

Merci aussi à tous les membres du personnel du laboratoire, qui m'ont facilité les tâches administratives, en particulier Véronique Varenne pour sa gentillesse et son efficacité.

Merci à toutes mes amies pour leur soutien et leur écoute, je citerais particulièrement Nadjet et Anissa avec qui j'ai partagé des bons moments au Lip6, Nawel et Soumia qui ont toujours été présentes pour moi.

Enfin, je tiens à exprimer ma profonde affection et mes plus chaleureux remerciements à ma famille, en particulier mes chers parents Mohand et Fazia qui m'ont toujours soutenue et qui m'ont fait confiance. Leur amour est pour beaucoup dans ce travail, je leur dédie spécialement cette thèse. Merci à mon frère Djamel qui m'a tellement aidée et encouragée à continuer mes études, merci aussi à tous mes autres frères et sœurs : Hakima, Samia, Souhila, Saida, Nadir et Aziz, mes beaux frères et belles sœurs ainsi que toute ma belle famille. Je ne pourrais pas oublier mes adorables neveux et nièces pour leur tendresse et leur don de me redonner le sourire en toutes circonstances.

Pour terminer, je voudrais remercier du fond du cœur mon mari Mehenna, qui a été toujours présent pour m'écouter, me conseiller, me soutenir et qui a su me supporter et m'encourager surtout dans la phase de rédaction.

Table des matières

Chapitre 1

Introduction

Dans beaucoup de contextes pratiques des graphes apparaissent, comme par exemple en informatique avec les graphes issus du Web (ensemble de pages Web et liens entre elles), la topologie de l'internet (routeurs et liens entre eux, ...), les réseaux d'échanges (courriers électroniques, fichiers, ...), les échanges pair-à-pair, ... On peut également citer plusieurs autres exemples dans d'autres disciplines : les réseaux sociaux (relations entre individus ou groupes d'individus), les réseaux biologiques (interactions protéiques, topologie du cerveau, ...) ou les réseaux linguistiques (réseaux de synonymie, réseaux de co-occurrences, ...). Ces graphes apparaissant dans des contextes pratiques sont regroupés sous l'appellation graphes de terrain (*complex networks* en anglais).

Il a été montré récemment que la plupart de ces graphes, bien qu'ils soient issus de contextes différents, ont des propriétés statistiques en commun qui ont fait l'objet de nombreuses études [WS98, BA99, Str01, New03]. Ils sont généralement de grande taille mais possèdent aussi une densité faible (deux nœuds choisis au hasard sont liés avec une probabilité très faible), une distance moyenne faible (il existe un chemin court entre presque toutes les paires de nœuds), une distribution de degrés hétérogène (il existe un nombre non négligeable de sommets possédant un très fort degré par rapport à une majorité de sommets possédant un très faible degré, et tous les comportements intermédiaires) et une densité locale forte (les liens entre les sommets qui sont proches sont beaucoup plus probables que les liens entre sommets éloignés). Ces propriétés constituent aujourd'hui un ensemble de référence, considéré comme fondamental, souvent complété par diverses autres propriétés selon le cas d'étude.

Depuis lors, l'étude de ce type de graphes a connu un essor très important, et un grand nombre de travaux ont introduit un ensemble d'outils pour l'analyse et la description de

ces graphes, voir par exemple [BS03, AB02].

Outre le fait que la plupart des graphes de terrain ont des propriétés statistiques communes, il est apparu que de très nombreuses questions qui se posent sur ces graphes sont en fait très générales. Ces problématiques peuvent se répartir en plusieurs grandes familles [Lat07]. Parmi ces problématiques, la *mesure* consiste en une opération permettant d'acquérir des informations sur les nœuds et liens présents dans le graphe. Cette mesure fournit généralement une vision partielle et biaisée de l'objet réel. La *métrologie* a donc pour but d'étudier ce biais, de tenter de le corriger et/ou proposer des méthodes capables de capturer certaines propriétés de manière fiable.

Le fait que les graphes de terrain ont une très grande taille (de l'ordre de centaines de milliers, voire de millions ou de milliards de nœuds et de liens) rend impossible la compréhension de leur structure par une observation directe. L'*analyse* a pour objet de décrire cette structure, par l'introduction de propriétés statistiques et/ou structurelles qui résument l'information et synthétisent de façon pertinente les principales caractéristiques de ces graphes. On peut citer également les problématiques liées à la *modélisation*, *l'algorithmique* et les phénomènes ayant lieu sur des graphes de terrain (par exemple des phénomènes de diffusion).

Jusqu'à récemment ces objets étaient principalement étudiés sous un angle statique, c'est-à-dire comme un instantané du graphe obtenu à un instant donné. Or, la plupart de ces graphes sont en réalité des graphes dynamiques. Cette dynamique peut apparaître d'une façon différente selon les contextes : réseaux sociaux dans lesquels des connexions entre individus apparaissent et disparaissent au cours du temps, graphes du web dans lesquels des pages sont créées ou supprimées, internet où les routeurs, les AS et/ou les liens entre eux sont créés et supprimés, etc.

Prendre en compte la dynamique de ces graphes est important, tant pour les aspects fondamentaux que pour les applications correspondantes. Il y a donc un fort besoin de développer des outils et méthodes pour l'analyse de la dynamique de graphes et ce dans de nombreux domaines : informatique, biologie, sociologie, économie, etc.

L'intérêt porté à l'étude des graphes dynamiques est encore récent, les travaux effectués sur ce domaine sont peu nombreux, et beaucoup reste à faire. Il a été montré que les problématiques de la mesure, la métrologie et l'analyse sont toujours pertinentes dans le cas dynamique, et que de nouvelles questions apparaissent également [Mag10] : comment décrire la dynamique ? quelles sont les échelles de temps pertinentes ? comment définir une notion de normalité ? peut-on détecter des anomalies ? comment modéliser la dynamique ?, etc.

Plusieurs travaux ont abordé ces questions [ADG07, NBW06, AB02]. On peut citer

par exemple des méthodes permettant de manipuler des graphes dynamiques [KKK00, BXFJ03], le dessin de graphes dynamiques [FT08,CGD04] ou la détection d'événements [STF06, KNRT05,HLM10].

La détection ou le suivi de communautés est également une problématique qui a reçu beaucoup d'intérêt ces dernières années. La très faible densité globale des grands graphes de terrain couplée à leur forte densité locale révèlent la présence de communautés (groupes de nœuds très denses mais avec peu de liens entre les groupes eux-mêmes). La détection de communautés est un problème difficile et de nombreuses méthodes heuristiques ont été proposées dans le cadre des réseaux statiques [For10, BGLL08]. Plusieurs études ont également été faites pour intégrer la dynamique dans la détection de communautés, par exemple en cherchant à détecter des décompositions différentes à plusieurs instants puis en essayant de suivre les communautés entre ces multiples décompositions [HKKS04,PBV07]. Dans [AG10], les auteurs proposent une méthode pour trouver une partition unique dans les graphes dynamiques qui soit toujours de bonne qualité sur une période donnée.

La prédiction de liens fait aussi partie des problèmes de recherche dans l'analyse des réseaux dynamiques. Étant donné un instantané d'un graphe considéré à un moment donné, la prédiction de liens consiste à prédire les liens qui vont probablement apparaître dans le futur. Plusieurs travaux ont étudié ce problème. La plupart d'entre eux sont basés sur des mesures de similarité entre les nœuds [NK03,Hua06]. Dans [HCSZ06, OHS05, WSP07, TAB09], les auteurs ajoutent plusieurs mesures non-topologique basées sur les attributs de nœud et utilisent un algorithme d'apprentissage supervisé pour faire la prédiction des liens. Certains travaux s'adressent au sujet de la prédiction lien dans les réseaux bipartis : dans [HLC05], les auteurs adaptent certaines mesures topologiques utilisées dans les graphes classiques. Enfin, dans [AML11], les auteurs définissent la notion de liens internes dans les graphes bipartis et proposent une méthode de prédiction basée sur ces liens.

Plusieurs autres travaux traitant la description des réseaux dynamiques se sont intéressés à des cas particuliers, comme par exemple les réseaux de contacts définis par la proximité entre les individus [HCS+05, SBF+08, CMRM07, PC11], les échanges pair-à-pair [LBGL05, LBFM09, JLGLB04, SZR07], la topologie de l'internet [LMO08, OZZ07, MOVL09, PPG04], les réseaux biologiques [BLA+04, PIL05], les réseaux de citations d'articles [LCSN07], et plusieurs types de réseaux sociaux en ligne [CR09, SP09, TRW09].

Dans ce contexte, l'objectif de cette thèse est de contribuer au développement d'outils et de méthodes génériques pour l'analyse de la dynamique des graphes de terrain tout en prenant en compte le biais lié à la mesure. Il est possible d'obtenir des résultats intéressants en étudiant l'évolution de propriétés existantes pour les graphes statiques au fil du temps. Cependant, ceci n'est pas suffisant et il est nécessaire d'introduire des notions prenant intrinsèquement en compte la dynamique des graphes étudiés, permettant par exemple de

formaliser des notions intuitives comme le taux d'apparition / de disparition des nœuds et des liens, ou leur durée de vie, mais également des propriétés plus complexes.

Nous nous sommes intéressés particulièrement à la question de la métrologie, c'est-à-dire l'étude du biais induit par la mesure. En effet, caractériser de manière fiable la dynamique d'un graphe est une tâche difficile pour différentes raisons. Par exemple le fait que la période de mesure est finie nous empêche d'observer certains événements (ceux qui apparaissent avant ou après la période de mesure), ce qui induit un biais dans les observations [RLA00, SGG03].

Notre but consiste donc à proposer des méthodes génériques éliminant, ou du moins réduisant les différents biais qui peuvent exister afin de caractériser précisément la dynamique des graphes étudiés. Dans un deuxième temps, nous cherchons également à définir des propriétés pertinentes pour décrire cette dynamique.

Ce mémoire est organisé de la manière suivante : nous présentons dans le chapitre 2 les différentes méthodes permettant de comparer des distributions, que nous utilisons dans la suite de ce mémoire. Dans le chapitre 3, nous nous intéressons au biais induit dans l'étude d'un graphe dynamique par le fait que la fenêtre d'observation soit par définition finie. Nous avons introduit une nouvelle méthodologie qui permet de caractériser une propriété dans un système dynamique. Nous illustrons sa pertinence en l'appliquant à l'étude de plusieurs propriétés dans un grand système P2P, en utilisant deux jeux de données différents. Les résultats de ce chapitre ont été partiellement publiés dans [BM10, BM11]. Dans le chapitre 4, nous nous intéressons à la caractérisation de la dynamique dans les réseaux de contacts humains. Nous analysons différents jeux de données avec l'objectif de décrire à la fois le comportement global de ces systèmes et de détecter si certains nœuds ont des comportements spécifiques. Enfin, nous présentons nos conclusions et perspectives dans le chapitre 5.

Chapitre 2

Interpréter et comparer des distributions

Contents

2.1 Introduction

Après une collecte de données, on dispose le plus souvent de données numériques brutes présentées sous la forme d'une série de valeurs qui sont rarement parlantes. Il est donc nécessaire de trouver une représentation adéquate qui permet de mieux comprendre ces données et de faire ressortir des informations utiles.

Dans ce chapitre nous allons nous intéresser à l'étude de la distribution (fréquence d'apparition de chaque valeur dans une série de données) obtenue à partir de l'étude d'une propriété donnée. Le but est d'abord de pouvoir comprendre le comportement de cette distribution afin de voir par exemple si elle suit un certain comportement spécifique, si l'ensemble des valeurs sont homogènes, hétérogènes, etc. Ensuite, nous explorerons les différentes méthodes permettant de comparer deux ou plusieurs distributions, qu'il s'agisse de méthodes reposant sur l'aspect visuel de ces distributions ou de méthodes plus formelles reposant sur des tests et des calculs statistiques.

2.2 Comprendre une distribution

Supposons que l'on étudie un ensemble de valeurs $x_1, x_2, ..., x_n$. La distribution correspondant à cet échantillon correspond à : $(x_1,\ n_1)$, $(x_2,\ n_2)$, ..., $(x_k,\ n_k)$ où n_i est le nombre de fois que la valeur x_i a été observée dans cet ensemble de données. Par exemple si l'on dispose de la série d'observations suivante :

1, 30, 1, 100, 500, 1 000, 1, 30, 200, 3 600, 740, 3 000, 1, 500, 900, 500, 2 300, 3 000, 1,

1500, 500, 100, 30, 1 500, 900, 200, 3 000

alors la distribution de ces valeurs est :

(1, 5), (30, 3), (100, 2), (200, 2), (500, 4), (740, 1), (900, 2), (1 000, 1), (1 500, 2),

(2 300, 1), (3 000, 3), (3 600, 1),

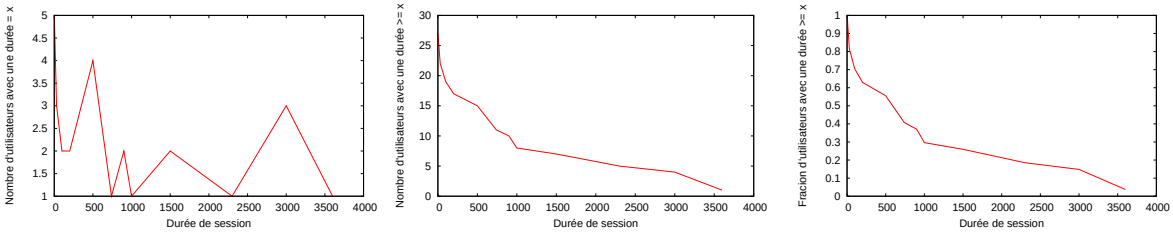
ce qui signifie qu'on a observé 5 fois la valeur 1, 3 fois la valeur 30, 2 fois la valeur 100, ... La représentation graphique correspondant à cette distribution est présentée dans la figure 2.1 (a).

Il peut être également utile de calculer la distribution cumulative inverse, notée P^c, qui permet de lisser la courbe et donc de mieux voir les tendances. Pour chaque valeur k, P^c_k représente le nombre valeurs observées qui sont supérieures ou égales à k. Pour l'exemple précédent, la distribution cumulative inverse correspondante sera :

(1, 27), (30, 22), (100, 19), (200, 17), (500, 15), (740, 11), (900, 10), (1 000, 8)

$$(1\,500,\ 7),\ (2\,300,\ 5)\ (3\,000,\ 4),\ (3\,600,\ 1)$$

ce qui signifie qu'il y a 27 observations qui ont une valeur supérieure ou égale à 1, parmi lesquelles 22 ont une valeur supérieur ou égale à 30, ... et une seule valeur qui est supérieure ou égale à 3600. Cette distribution cumulative inverse est présentée dans la figure 2.1 (b).



(a) Distribution. (b) Distribution cumulative inverse. (c) Distribution cumulative inverse normalisée.

FIG. 2.1 – Exemples de distributions.

Une distribution donnée est dépendante du nombre total de valeurs qu'elle représente. Une solution pour y remédier est de normaliser les valeurs de n_k en les divisant par le nombre total d'observations. Ceci permet notamment de comparer des distributions provenant d'échantillons de taille différente.

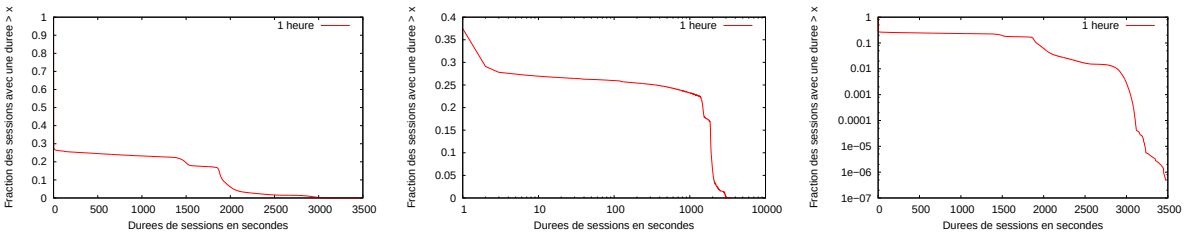
Pour l'exemple précédent, la normalisation revient à diviser chaque valeur de n_k par 27 (nombre total d'observations). Cela revient à dire que 100% des observations ont une valeur supérieure ou égale à 1, parmi lesquelles 81% ont une valeur supérieur ou égale à 30, etc. La représentation graphique correspondante est présentée dans la figure 2.1 (c).

2.2.1 Observation visuelle d'une distribution

L'étude visuelle d'une distribution peut apporter beaucoup d'informations : symétrie, nombre de pics, inclinaison, etc. Elle permet aussi d'observer des caractéristiques spécifiques comme par exemple des parties des distributions où il n'y a pas d'observations ou encore des valeurs extrêmes. Dans ce qui suit, nous allons détailler les points importants qui peuvent ressortir lors de l'étude visuelle d'une distribution donnée.

2.2.1.1 Choix des échelles

Afin de capturer le plus d'informations possibles à partir d'une représentation graphique d'une distribution, il est important de la représenter sous différentes échelles. Nous allons illustrer ceci dans la figure 2.2 qui présente la même distribution sous différentes échelles.



(a) Échelle lin-lin. (b) Échelle log-lin. (c) Échelle lin-log.

FIG. 2.2 – Distribution cumulative inverse des durées de sessions des utilisateurs dans un système P2P observé pendant une heure.

La figure 2.2 (a) correspond à la représentation en échelle linéaire sur les deux axes. On peut observer le comportement global de la distribution, mais on remarque qu'il existe certaines parties dans cette courbe sur lesquelles nous ne sommes pas capable de dire grand chose, notamment les observations avec des petites valeurs sur l'axe des x qui sont écrasées par la présence de valeurs très grandes. On à l'impression qu'il y a plus de 70% d'observations qui ont une valeur égale à 0. Dans le but d'étaler cette partie, nous utilisons une échelle logarithmique sur l'axe des x (figure 2.2 (b)) qui permet de donner une vision plus claire sur l'intervalle compris entre 0 et 100. Ici, on peut constater qu'il y a environ 60% des observations ayant une valeur 0 (qu'on ne voit pas à cause de l'échelle logarithmique sur l'axe des x), environ 10% des observations ayant une valeur comprise entre 1 et 2 et une très petite fraction (moins de 2%) ayant une valeur comprise entre 2 et 10. Ceci donne une vision plus précise que celle que l'on a eu dans la première représentation.

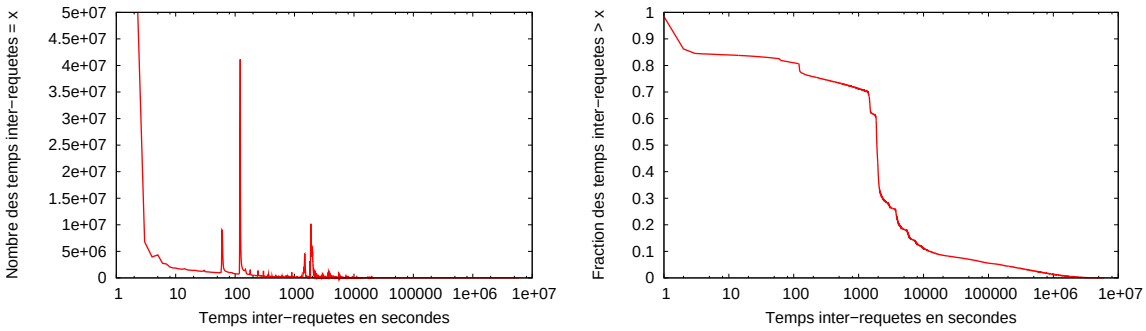
On n'arrive pas non plus a bien observer dans la première figure les petites valeurs sur l'axe des y. De même, pour étaler cette partie et mieux voir le comportement de ces observations, nous utilisons une échelle logarithmique sur l'axe des y que l'on présente dans la figure 2.2 (c). On peut voir qu'on a plus de précision sur cette partie d'observations avec des très petites fractions sur l'axe des y et de très grandes valeurs sur l'axe des x.

Comparer différentes échelles est également utile pour distinguer entre plusieurs types de distributions, en particulier les distributions hétérogènes (par exemple les lois de puissance) qui sont plus ou moins proches de droites en échelle doublement logarithmique.

En conclusion, pour pouvoir étudier une distribution d'une manière correcte, complète et précise, il est important de la représenter sous toutes les échelles possibles et de comparer les résultats obtenus.

2.2.1.2 Distribution standard/cumulative

Nous avons déjà montré l'intérêt d'une distribution cumulative inverse par rapport à une distribution standard, mais cela ne veut pas forcément dire qu'il faut toujours privilégier l'une par rapport à l'autre. Dans certains cas, les deux représentations ont du sens, et c'est le cas dans l'exemple que nous présentons dans la figure 2.3.



(a) Distribution. (b) Distribution cumulative inverse.

FIG. 2.3 – Distribution des temps inter-requêtes, pour un système P2P.

La distribution standard présentée dans la figure 2.3 (a) montre des pics très clairs aux environs des valeurs 60, 120, etc. Ces pics constituent une caractéristique importante de cette distribution.

Afin de lisser la courbe et de mieux voir les tendances générales, nous présentons la distribution cumulative inverse dans la figure 2.3 (b). Dans cette courbe, nous pouvons faire d'autres observations. Par exemple, on voit qu'il y a une très forte densité de valeurs comprises entre 1 000 et 10 000 secondes, ce qui n'est pas du tout clair sur la figure 2.3 (a).

Au final, nous pouvons conclure que chacune des deux représentation peut fournir des informations importantes qui sont complémentaires, et qu'aucune des deux isolément ne permet d'avoir une vision complète de la distribution étudiée.

2.2.1.3 Valeurs extrêmes

Dans l'analyse de certaines distributions, nous remarquons parfois certaines valeurs qui s'écartent de la plupart des autres valeurs. Par exemple, dans la distribution présentée dans la figure 2.2 (c), nous pouvons remarquer une très petite fraction de valeurs allant jusqu'à 3 500 secondes, alors que la majorité des observations ne dépassent pas les 1 000 secondes. On appelle ce type de valeurs des valeurs extrêmes.

On peut également considérer comme extrêmes les valeurs les plus petites. Contrairement aux valeurs extrêmes très élevées qu'on ne s'attend pas forcément à observer, la

présence de valeurs très petites est la plupart du temps tout a fait normale. Elles peuvent cependant avoir un comportement différent qui peut les rendre spécifiques. Dans la figure 2.2 (b), on peut voir qu'une grande fraction d'observations (plus de 60%) ont une valeur égale à 0. Ce type d'observations mérite de l'attention et une étude plus approfondie. Par exemple, il serait intéressant d'étudier la distribution sans prendre en compte la valeur 0, pour mieux analyser le comportement des autres valeurs.

Il existe plusieurs types de valeurs extrêmes. Dans certains cas, il s'agit de valeurs aberrantes. Il s'agit du cas où ce type d'observations sont considérées comme étant des erreurs dues aux mesures ou à d'autres raisons. Barnett et Lewis [BL96] utilisent le terme de valeur suspecte pour décrire une valeur douteuse mais qui n'est pas jugée aberrante, tandis que le terme de valeur aberrante correspond a une valeur extrême qui est statistiquement discordante. Une valeur suspecte est donc une valeur moins extrême qu'une valeur aberrante dans cette terminologie. Dans notre cas, nous allons utiliser le terme de valeur extrême pour désigner des observations qui semblent dévier de façon marquée de l'ensemble des autres observations, autrement dit des observations qui ne sont pas en harmonie avec le reste des observations. Tenter une caractérisation plus détaillée de ces valeurs dépasse le cadre de nos travaux.

2.2.2 Paramètres statistiques d'une distribution

Afin d'avoir une idée plus précise sur une distribution, deux types de mesure peuvent être utiles : la mesure de la position centrale des observations et la mesure de leur dispersion ou de leur variabilité, c'est-à-dire la mesure de la répartition des observations autour de cette position centrale.

La moyenne représente la mesure la plus courante de tendance centrale des observations. Quand les données sont présentées sous forme d'une distribution normalisée avec k valeurs différentes observées $(x_1, x_2, ...x_k)$, la moyenne s'exprime en fonction des fréquences relatives :

$$\bar{x} = \sum_{i=1}^{k} f_i x_i,$$

avec $f_i = \frac{n_i}{n}$ et $\sum_{i=1}^{k} f_i = 1$.

L'écart-type (racine carrée de la variance) représente la mesure de variabilité la plus courante. Il mesure la dispersion de chaque observation autour de la moyenne. L'écart-type

d'une distribution est donné par la formule suivante :

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2}.$$

L'écart-type est nul lorsque toutes les valeurs sont identiques ; sa valeur serait en revanche maximale si les résultats se répartissaient dans les mêmes proportions aux deux extrémités de l'échelle de mesure.

En raison de ses liens étroits avec la moyenne, l'écart-type peut être très influencé si cette dernière n'est pas représentative de la majorité des valeurs, si la distribution est hétérogène en particulier. Il peut être aussi influencé par les valeurs extrêmes ; une seule de ces valeurs pourrait avoir une grande influence sur les résultats. Couplé à l'analyse de la représentation graphique d'une distribution, l'écart-type peut être considéré comme un bon indicateur d'existence de ce type de valeurs.

Généralement, plus les valeurs sont largement distribuées, plus l'écart-type est élevé. Il n'est cependant pas toujours facile d'évaluer l'importance que doit avoir l'écart-type pour que l'on estime que les données sont largement dispersées. L'importance de l'écart-type dépend en effet aussi de l'importance de la valeur moyenne de l'ensemble des données. Lorsque l'on mesure quelque chose en millions, l'écart-type sera naturellement plus grand que lorsque l'on mesure le poids d'un ensemble de personnes par exemple.

2.3 Comparaison de distributions

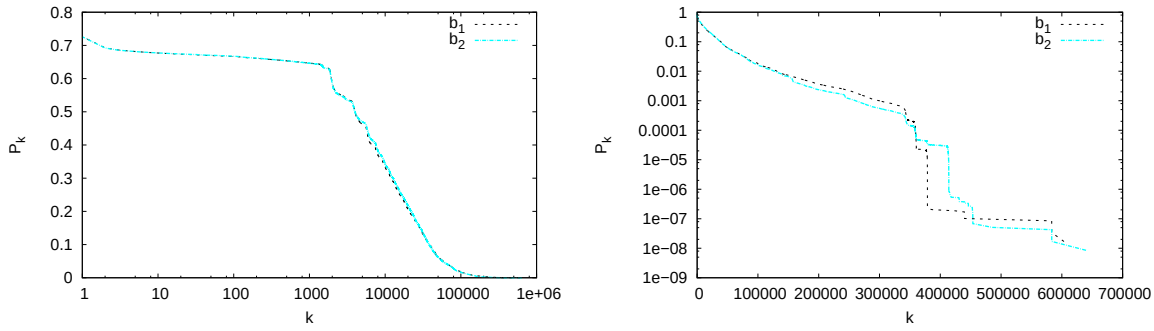
Dans certains contextes, il est très important de pouvoir comparer deux ou plusieurs distributions. Quand on s'intéresse à l'étude de la dynamique d'une propriété donnée dans un système, pouvoir comparer différentes distributions obtenues en prenant des échelles de temps différentes est important. Cela permet d'étudier l'évolution de cette propriété et de voir par exemple à quel point cette propriété est stable. Il existe plusieurs méthodes qui permettent de faire cette comparaison, certaines reposant sur une étude visuelle, d'autres, plus formelles, reposant sur des tests statistiques. Nous allons présenter ces méthodes en nous appuyant sur des exemples tirés des analyses que nous présentons dans le chapitre suivant.

La première méthode consiste à comparer les représentations graphiques de ces différentes distributions. Il est cependant très important de choisir la bonne représentation graphique, notamment les échelles des axes utilisées pour cette représentation ainsi que la normalisation des distributions. Nous avons déjà vu dans la section 2.2 que cela a un impact fort sur l'étude visuelle que l'on peut faire. Nous allons voir que ces choix peuvent

aussi influencer fortement les observations qu'on obtient. La deuxième méthode consiste à comparer les distributions en utilisant différents tests statistiques qui permettent d'avoir des observations plus précises et des résultats plus formels. Nous verrons également qu'ils permettent d'étudier comment une séquence de distributions dépendant d'un paramètre évolue en fonction de ce paramètre.

2.3.1 Choix des échelles

Afin de montrer le rôle du choix des échelles, nous présentons dans la figure 2.4 deux distributions (cumulatives inverses) représentées sous différentes échelles.



(a) Échelle log-lin.

(b) Échelle lin-log.

FIG. 2.4 – Distribution cumulative inverse des durées de sessions des utilisateurs pour deux différentes fenêtres d'observation b_1 et b_2 dans une capture de données d'un système P2P.

Sur la figure 2.4 (a), nous utilisons une échelle logarithmique sur l'axe des x et une échelle linéaire sur l'axe des y. Avec cette représentation, les deux distributions semblent pratiquement identiques. Par contre, dans le cas où l'on prend une échelle linéaire sur l'axe des x et logarithmique sur l'axe des y (figure 2.4 (b)), les distributions semblent très différentes. Cette différence n'est cependant présente que pour une petite fraction de valeurs (pour des valeurs de $y < 0.01$).

Au-delà du fait que cette deuxième représentation montre des différences que l'on ne voit pas avec la première, elle permet aussi de nous montrer un autre point important. Les valeurs qui diffèrent entre les deux distributions, que l'on peut bien distinguer après environ une valeur de 100 000 s, correspondent aux valeurs vues après le coude de la figure 2.4 (a) et sont significativement plus rares que les valeurs inférieures à ce coude. En effet plus de 99% des observations ont une valeur inférieure à 100 000 s, et moins de 1% ont une valeur qui peut aller jusqu'à 600 000 s. Il s'agit donc de valeurs extrêmes.

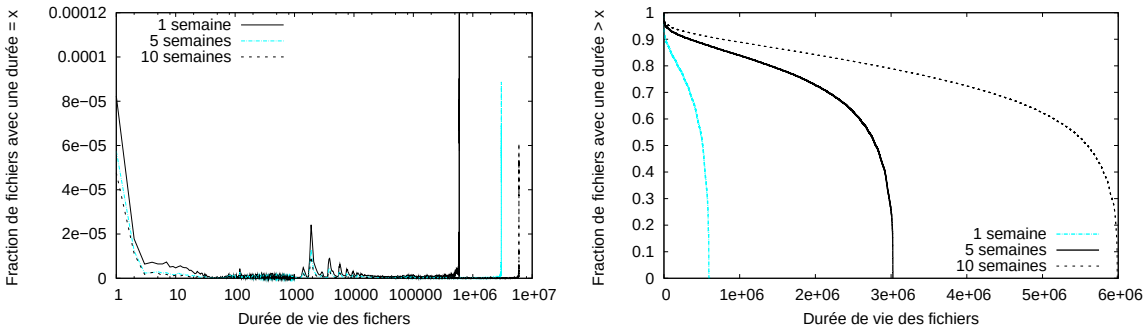
Au final, chacune des représentations nous rapporte des informations intéressantes et

complémentaires. Il est donc important de toujours tester différentes représentations avec différentes échelles pour être vigilant et pour obtenir le plus d'informations possible.

2.3.2 Normalisation des distributions

Dans certains cas, les distributions que l'on veut comparer dépendent d'un certain paramètre. Ceci est le cas par exemple lorsque l'on s'intéresse à l'étude de l'évolution d'une propriété donnée en fonction de la taille de la fenêtre d'observation. Pour comparer ces distributions, il peut être utile de normaliser leurs valeurs par ce paramètre.

Nous allons illustrer cette méthode sur le cas de la durée de vie des fichiers dans un système P2P. Nous présentons dans la figure 2.5 la séquence de distributions (standards et cumulatives inverses) des durées de vie de fichiers observées pendant $1, 5$ et 10 semaines.



(a) Distributions.

(b) Distributions cumulatives inverses.

FIG. 2.5 – Distributions des durées de vie des fichiers dans un système P2P, observées pendant $l = 1$, 5 et 10 semaines.

Dans cette figure, nous pouvons observer que les distributions évoluent considérablement avec l : plus on observe le système longtemps, plus les valeurs observées pour les durées de vie des fichiers ont tendance à être élevées. Afin de mieux voir la façon dont cette propriété évolue en fonction de la taille de la fenêtre d'observation l, nous appliquons une normalisation par rapport à l. La première chose à faire est de choisir l'unité de normalisation. Ensuite, il faut diviser les valeurs de l'axe des x de chaque distribution par l en prenant en compte l'unité de normalisation choisie. Enfin, afin d'obtenir des distribution normalisées correctement, il faut aussi multiplier les valeurs de l'axe des y de chaque distribution par la même valeur. Pour l'exemple présenté dans la figure 2.6, nous avons choisi une unité de normalisation de 1 semaine. Pour la normalisation, la distribution correspondant à 1 semaine est donc inchangée. On divise les valeurs de l'axe des x de la

distribution correspondant à 5 semaines (respectivement à 10 semaines) par 5 (respectivement par 10), et on multiplie les valeurs de l'axe des y de la distribution correspondant à 5 semaines (respectivement à 10 semaines) par 5 (respectivement par 10). Les distributions ainsi renormalisées sont présentées dans la figure 2.6. Notons qu'une telle normalisation ne peut pas s'appliquer sur les distributions cumulatives.

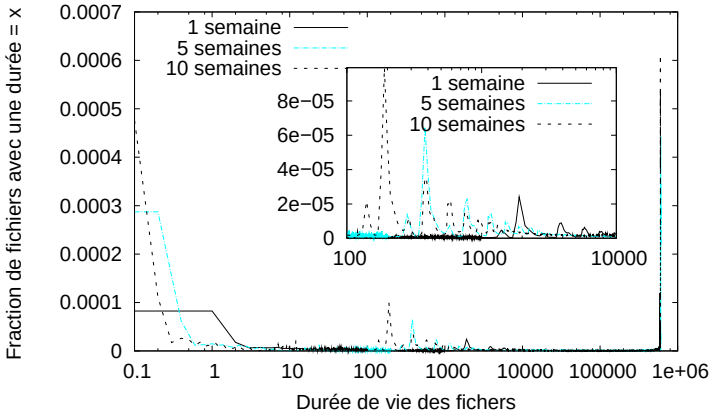


FIG. 2.6 – Distributions des durées de vie des fichiers observées pendant l = 1,5 et 10 semaines, normalisées par rapport à l. Encart : zoom sur une partie de ces distributions.

On voit que dans ce cas, les pics observés pour les valeurs maximales des distributions coïncident, ce qui n'était pas le cas dans la figure 2.5 (a). Cependant, les pics intermédiaires, qui étaient confondus ne le sont plus.

2.3.3 Tests statistiques

2.3.3.1 La distance de Kolmogorov-Smirnov

La distance de *Kolmogorov-Smirnov*, ou K-S distance [CLR67], est utilisée pour mesurer à quel point deux distributions sont proches l'une de l'autre. Plus précisément, la K-S distance compare deux distributions cumulatives normalisées (qu'elles soient complémentaires ou pas). Elle mesure l'écart (ou la distance) maximal entre les deux distributions cumulatives. Si on considère deux distributions cumulatives P^c et Q^c, alors la K-S distance est égale à :

$$KS(P,Q) = \max_k |P_k^c - Q_k^c|.$$

La valeur de la K-S distance est toujours inférieure à 1, et plus elle est proche de 0, plus les deux distributions sont similaires.

2.3.3.2 La distance de Monge-Kantorovich

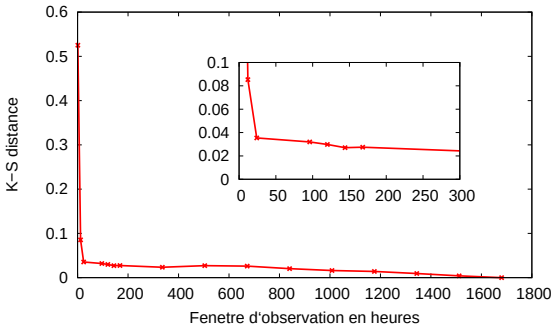
Une question importante soulevée par la K-S distance est de savoir si les distributions que l'on compare diffèrent par la valeur correspondante sur un grand intervalle de valeurs, ou seulement sur un seul point. Afin de répondre à cette question, nous étudions la distance de *Monge-Kantorovich*, ou M-K distance [GKT09], qui calcule la moyenne de l'écart entre deux distributions cumulatives :

$$MK(P,Q) = (\sum_k |P_k^c - Q_k^c|)/k_{\max}.$$

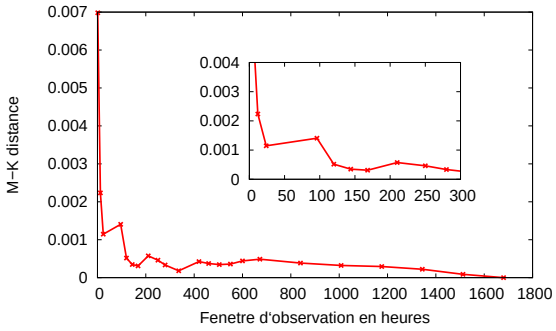
Deux distributions qui ne diffèrent que par un seul point auront donc une K-S distance élevée mais une M-K distance petite.

2.3.4 Séquence de distributions

Pour étudier comment une famille ordonnée de distributions $(P(1), P(2), ...P(n))$ évolue, on peut utiliser les indicateurs que l'on a défini auparavant. Afin d'étudier si les distributions se stabilisent, on calcule à chaque fois la K-S distance (respectivement la M-K distance) entre les distributions $P(k)$ et $P(n)$, où P_n est la dernière distribution de la famille. Nous traçons ensuite ces valeurs en fonction de k.



(a) KS$(P(k), P(n))$ en fonction de k.



(b) MK$(P(k), P(n))$ en fonction de k.

FIG. 2.7 – Étude de l'évolution de P (durées de sessions des utilisateurs) avec la K-S et la M-K distance.

À titre d'illustration, nous présentons dans la figure 2.7 les résultats obtenus en calculant la K-S et la M-K distance sur une séquence de distributions obtenues en étudiant les durées de sessions des utilisateurs dans un système P2P, en fonction du temps pendant lequel on observe le système.

Par exemple, dans la figure 2.7 (a) qui correspond aux résultats obtenus pour la K-S distance, le point de coordonnées $(168, 0.03)$ correspond à la valeur de la K-S distance obtenue entre la distribution correspondant à une longueur $l = 168$ heures (1 semaine) et celle correspondant à la plus grande fenêtre qui est $l = 1680$ heures (10 semaines). Dans ce cas, nous pouvons observer que les valeurs ont tendance à être très élevées au début, puis à diminuer rapidement pour atteindre de très petites valeurs. Après cela, les valeurs ont tendance à se stabiliser.

La figure 2.7 (b) présente les résultats obtenus pour la M-K distance. Nous pouvons observer que le comportement est légèrement différent par rapport à celui de la K-S distance : les valeurs observées ont tendance à diminuer (avec des fluctuations), jusqu'à ce que la période de mesure atteigne approximativement 150 heures. Après cela, la valeur de la M-K distance devient très petite : cela montre que les distributions correspondantes sont très proches les unes des autres.

Enfin, pour mieux comprendre comment la distribution d'une propriété donnée évolue, nous pouvons calculer aussi sa moyenne et son écart-type en fonction de la longueur de la fenêtre de mesure, en suivant l'analyse faite dans [WAL04]. Cela permet soit de confirmer les résultats obtenues avec la K-S et la M-K distance dans le cas où on obtient un comportement similaire, soit d'obtenir d'autres conclusions dans le cas ou le comportement observé est différent.

2.4 Conclusion

Pour comparer une ou plusieurs distributions, il est nécessaire d'utiliser plusieurs méthodes afin de comprendre et d'interpréter les différentes observations que l'on peut obtenir. La première approche consiste à étudier visuellement les distributions, ce qui permet d'obtenir une première intuition et d'acquérir des informations que l'on peut confirmer avec plus de précision en utilisant des méthodes formelles et statistiques. Ces méthodes sont complémentaires et permettent d'obtenir une vision globale et complète de la distribution que l'on souhaite étudier.

Chapitre 3

Biais causé par la durée de la mesure

Contents

3.1 Introduction

La compréhension de la dynamique d'un graphe de terrain est une question clé. Cependant, caractériser de façon précise cette dynamique est une tâche difficile. En particulier, le fait que la fenêtre d'observation soit par définition finie induit un biais dans les observations. Bien qu'intuitivement ce biais tende à diminuer lorsque l'on augmente la période d'observation, dans la pratique, il est difficile de le quantifier et de savoir s'il est négligeable ou non.

En particulier, une fenêtre d'observation trop petite peut ne pas être représentative du comportement du système global. Par exemple, il est clair que mesurer l'activité dans un système P2P pendant une heure n'est pas suffisant pour capturer totalement la dynamique des utilisateurs, à cause de la variation de l'activité jour/nuit par exemple. Cependant, on ne sait pas *a priori* si un jour, ou deux, ou une semaine, est une période suffisamment longue.

Dans ce chapitre, nous introduisons une nouvelle méthodologie qui permet de caractériser rigoureusement les propriétés de systèmes dynamiques réels. Cette méthodologie est différente et complémentaire d'autres méthodologies existantes dans la littérature [SR06, SGG03, GT99]. Elle a deux avantages principaux :

- elle permet de déterminer si la longueur de la période d'observation est suffisante pour une caractérisation rigoureuse d'une propriété donnée ;
- elle peut être appliquée à n'importe quelle propriété caractérisant un graphe de terrain dynamique.

Pour illustrer sa pertinence, nous avons appliqué cette méthodologie à l'étude de plusieurs propriétés dans un système pair-à-pair. Nous utilisons deux jeux de données différents qui fournissent des informations complémentaires.

Dans la section 3.2, nous présentons notre méthodologie et les données utilisées. De la section 3.3 à la 3.7 nous appliquons notre méthodologie à l'étude de plusieurs propriétés. Enfin, nous présentons nos conclusions et nos perspectives dans la section 3.9.

3.2 Méthodologie et données

3.2.1 Méthodologie

Supposons que l'on commence l'observation d'un graphe dynamique à l'instant t, pour une durée l. On note $W_{t,l}$ cette fenêtre d'observation.

Lors de la caractérisation de la dynamique d'un graphe à partir de l'observation de $W_{t,l}$ on est face à deux problèmes principaux. Premièrement, l doit être suffisamment longue

pour que $W_{t,l}$ soit *représentative*. Par exemple, il est clair qu'il est difficile voire impossible de caractériser avec précision l'activité dans un système P2P après l'avoir observé pendant seulement une heure : cela ne permet même pas d'observer la variation de l'activité en fonction de l'heure de la journée. Deuxièmement, même si $W_{t,l}$ est représentative, le fait que l soit *finie* induit toujours un biais dans les observations. En effet, on ne peut pas capturer les événements survenus avant t ou après $t+l$, ce qui nous empêche de caractériser avec précision certaines quantités (par exemple, les durées de sessions des utilisateurs, ou la corrélation entre des événements). Un point important est que plus la fenêtre d'observation est grande, plus le biais induit sera petit.

Notre méthodologie traite ces deux problèmes en même temps. Intuitivement, elle vise à déterminer si la période d'observation $W_{t,l}$ est suffisamment longue pour caractériser une propriété P donnée, c'est-à-dire si le biais induit par le fait que la période d'observation est finie sur P est négligeable. Il est évident que si la fenêtre d'observation $W_{t,l}$ est assez longue, alors, si l'on utilise une fenêtre d'observation plus grande de taille $l+x$, la propriété observée sera la même : $P(W_{t,l}) = P(W_{t,l+x})$.

Afin de pouvoir décider quand une fenêtre d'observation est suffisamment longue, on utilise des fenêtres de taille croissante W_{t,l_1},W_{t,l_2}, ..., W_{t,l_n} $(l_1 < l_2 < ... < l_n)$. En étudiant la façon dont la propriété observée $P(W_{t,l_1})$,$P(W_{t,l_2})$, ...$P(W_{t,l_n})$ évolue en fonction de l, nous pouvons déterminer si elle est correctement évaluée ou non : si elle fluctue ou varie fortement avec l'augmentation de l, alors P n'est certainement pas évaluée avec précision. En effet, une fenêtre d'observation plus ou moins longue aurait donné une valeur différente. Par contre, si P a tendance à se stabiliser avec l'augmentation de la taille de la fenêtre l, alors elle est probablement évaluée précisément.

Enfin, un point important est que la caractérisation d'une propriété P n'a de sens que si le système est stationnaire, c'est à dire si P n'évolue pas pendant que la mesure progresse. Cependant, dans le cas où le système n'est pas stationnaire, notre méthodologie ne sera pas en mesure de fournir une caractérisation : la propriété observée P ne va pas se stabiliser quand on fera augmenter la taille de la fenêtre d'observation l. Si elle devient stable, cela signifie à la fois que $W_{t,l}$ est suffisamment longue, et que P est stationnaire [1].

Remarquons qu'en fonction de la propriété étudiée, d'autres types de biais peuvent apparaître, voir par exemple [SR06]. Dans notre contexte, certains biais proviennent de l'identification des utilisateurs et de leurs sessions. Nous faisons de notre mieux pour y faire face de manière rigoureuse, comme nous le détaillons dans les sections suivantes. Cependant, nous insistons sur le fait que notre but ici n'est pas de traiter tous les types de

[1] Remarquons que le système peut être stationnaire par rapport à une propriété donnée P et pas par rapport à une autre P' ; dans un tel cas, notre méthodologie va permettre de caractériser P mais pas P'.

biais en même temps, mais de montrer le rôle joué par la taille de la fenêtre d'observation.

Ici, la plupart des propriétés que nous étudions sont des distributions. Pour étudier la façon dont une distribution observée P évolue en fonction de la taille de la fenêtre d'observation, nous utiliserons les méthodes présentées dans le chapitre 2.

Notons que nous prenons t comme étant le début de notre période de mesure, donc nous fixons $t = 0$ dans la suite. Toutefois, dans la section 3.3.2, nous présentons un exemple illustrant l'influence du moment de début de la mesure sur nos observations.

3.2.2 Données

Nous avons utilisé deux jeux de données : le premier provient de [ALM09], et consiste en une capture du trafic UDP sur un serveur *eDonkey*. Il s'agit de l'ensemble des requêtes effectuées par les utilisateurs et de l'ensemble des réponses du serveur. Il existe deux types de requêtes. Les premières sont de la forme suivante :

$$T \;\; IP \;\; L,$$

où T est le moment auquel la requête a eu lieu, IP est l'adresse IP (anonymisée) de l'utilisateur qui a émis cette requête et L est une liste de mots clefs décrivant le fichier recherché. La réponse du serveur est de la forme suivante :

$$T \;\; IP \;\; (F_1, S_1) \;\; (F_2, S_2) \;\; ... \;\; (F_n, S_n),$$

où IP est l'adresse IP de l'utilisateur qui reçoit cette réponse et (F_1, S_1) (F_2, S_2) $...(F_n, S_n)$ est une liste d'identifiants de fichiers correspondant aux mots-clés, avec un fournisseur pour chaque fichier.
Le deuxième type de requêtes est de la forme suivante :

$$T \;\; IP \;\; F_1 \;\; F_2 \;\; ... \;\; F_k,$$

où F_1 F_2 ... F_k est la liste des identifiants de fichiers que l'utilisateur veut télécharger. La réponse du serveur pour ce type de requêtes est de la forme suivante :

$$T \;\; IP \;\; (F_1, S_{11}...S_{1n_1}) \;\; (F_2, S_{21}...S_{2n_2}) \;\; ... \;\; (F_k, S_{k1}...S_{kn_k}),$$

où $S_{i1}...S_{in_i}$ est une liste de fournisseurs pour le fichier F_i.

La mesure que nous utilisons ici a duré 10 semaines, ce qui représente 1 milliard de messages, 89 millions de pairs (adresses IP distinctes) et 275 millions de fichiers.

Le deuxième ensemble de données consiste en une capture de connexions et de déconnexions de pairs sur un serveur *eDonkey* [LBFM09]. Les informations de connexion et de déconnexion permettent de connaître avec précision les durées des sessions des utilisateurs. Cependant, une petite fraction des sessions présente certains problèmes :

- certaines sessions n'ont pas de fin dans notre jeu de données, probablement parce que la mesure s'est arrêtée avant la déconnexion de l'utilisateur ;
- certaines sessions d'un même utilisateur sont imbriquées les unes dans les autres ; par exemple, on observe deux connexions consécutives suivies de deux déconnexions consécutives. Dans ce cas, il n'est pas possible de savoir quelle déconnexion correspond à quelle connexion et donc on ne peut pas connaitre la durée de session.

Nous n'avons pas pris en compte ces deux types de sessions pour nos analyses (elles représentent environ 2% de toutes les sessions). Ce jeu de données contient au final plus de 200 millions de connexions effectuées par plus de 14 millions de pairs, sur une période de 27 jours.

Les deux jeux de données sont complémentaires : le premier ne donne pas les moments de connexion et de déconnexion des utilisateurs, et le second ne contient pas d'informations sur les requêtes. Dans la suite, nous allons nommer le premier jeu de données les *données requêtes* et le second les *données logins*.

3.3 Durées de sessions des utilisateurs – *données requêtes*

Dans cette section, nous étudions la distribution S des durées de sessions, dans les *données requêtes*. Nous avons vu que les durées de sessions ne sont pas directement disponibles dans ce jeu de données (voir la section 3.2.2), d'où la nécessité de les inférer à partir de l'étude des requêtes effectuées par un utilisateur. Nous allons d'abord détailler ce processus avant de passer à l'étude de la distribution des durées de sessions proprement dite.

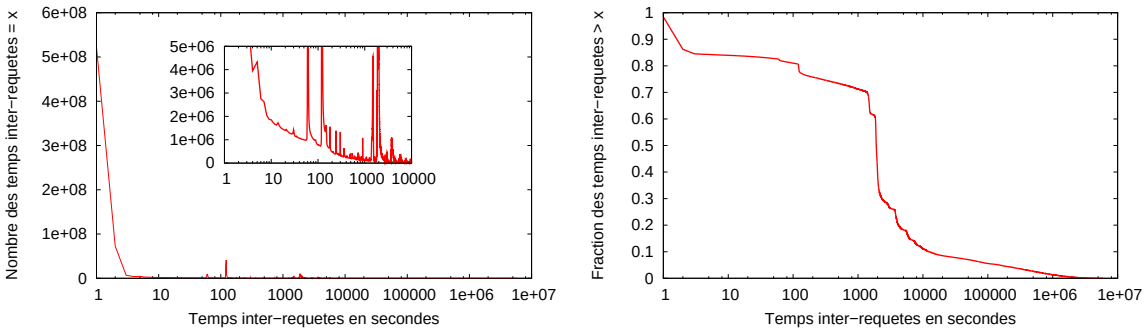
3.3.1 Identification des utilisateurs et des sessions

L'identification des utilisateurs dans nos données n'est pas une question triviale. Nous avons en effet seulement accès aux adresses IP et aux ports utilisés par les ordinateurs à partir desquels les requêtes sont effectuées. À un moment donné, un ordinateur est identifié par une adresse, mais cela peut changer et en général on n'est pas capable de détecter qu'un même ordinateur a deux adresses différentes (à cause d'une allocation d'adresses dynamiques, par exemple) et/ou que deux ordinateurs différents utilisent la même adresse

(parce qu'ils sont derrière le même NAT par exemple). Rajoutons à cela le fait qu'un même utilisateur peut utiliser plusieurs ordinateurs et que plusieurs utilisateurs peuvent utiliser le même ordinateur, et l'identification des utilisateurs devient encore plus difficile. En l'absence d'une méthode satisfaisante, on dispose de deux solutions naturelles : la première consiste à considérer qu'un utilisateur correspond à une adresse IP et l'autre consiste à considérer qu'un utilisateur correspond au couple d'informations composé de l'adresse IP et du port UDP utilisé.

Nous avons effectué des analyses en utilisant chacune de ces deux définitions. Nous présentons d'abord une analyse détaillée des résultats obtenus en utilisant la première définition, qui permet de capturer les sessions de manière pertinente comme nous le verrons ci-dessous. Ensuite, nous présentons aussi quelques résultats obtenus en utilisant la deuxième définition. En général, nous verrons que dans les deux cas, les résultats sont très similaires avec de petites différences que nous illustrons dans la suite.

Pour inférer les sessions d'un utilisateur donné, nous étudions le temps qui sépare deux requêtes consécutives. Il est naturel de considérer que deux requêtes consécutives effectuées par une même adresse IP appartiennent à la même session si le temps qui les sépare est relativement court, et appartiennent à deux sessions différentes s'il est long. Le problème qui se pose est donc de trouver un seuil approprié pour pouvoir distinguer ces deux cas. Pour étudier cela, on considère la distribution des temps inter-requêtes présentée dans la figure 3.1 (on présente à la fois la distribution (a) et la distribution cumulative inverse (b)).



(a) Distribution.

(b) Distribution cumulative inverse.

FIG. 3.1 – Distribution des temps inter-requêtes, pour les *données requêtes*.

Dans la distribution, on observe clairement des pics à 60 secondes et aux multiples de cette valeur (120, 240, 300, 900, ...). Ces pics (que l'on peut voir plus clairement dans l'encart) indiquent que même si les utilisateurs décident quelles requêtes ils effectuent et

quand ils les effectuent, il y a une forte influence du protocole sur les données observées : la plupart des applications client effectuent automatiquement des requêtes périodiques. Bien que ces pics deviennent plus petits après 1 800 secondes, un zoom sur la courbe (non présenté ici) montre qu'ils sont clairement définis pour des valeurs allant jusqu'à au moins 20 000 secondes.

Afin de lisser la courbe, nous considérons la distribution cumulative inverse (figure 3.1 (b)). On remarque qu'il y a une forte densité de valeurs comprises entre 1 000 secondes et une valeur légèrement inférieure à 10 000 secondes (la pente de la distribution est forte dans cette région). Une telle densité indique un temps inter-requête normal pour une session et cela aurait donc peu de sens de choisir un seuil dans cet intervalle ou avant. Par conséquent, le seuil doit être supérieur ou égal à 10 000 secondes.

Pour étudier l'importance des pics dans la distribution, nous avons calculé, pour une même fenêtre de mesure, les distributions des durées de sessions obtenues avec deux seuils différents, le premier choisi juste avant un pic et le second juste après. Nous avons fait une comparaison entre ces deux distributions et nous n'avons observé aucune différence significative.

Dans ce qui suit, nous avons choisi d'utiliser un seuil de $t = 10\,800$ secondes, soit 3 heures. Par conséquent, si un même utilisateur envoie deux requêtes successives séparées de moins de trois heures, ces requêtes vont appartenir à la même session, sinon, elles vont appartenir à deux sessions différentes [2].

3.3.2 Caractérisation des durées de sessions

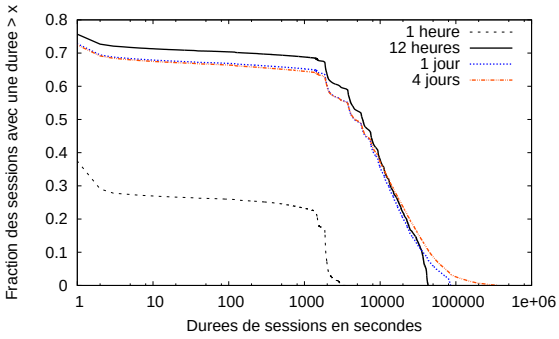
Nous appliquons maintenant notre méthodologie à l'étude des distributions des durées de sessions, en étudiant $S(W_{0,l})$ pour différentes valeurs de l. Tout d'abord, nous avons observé que ces distributions sont très irrégulières et qu'elles présentent des pics et des vallées assez clairs qui sont liés aux pics qu'on a observé dans la distribution des temps inter-requêtes, voir la figure 3.1 (a). Différentes longueurs et positions de fenêtres d'observation donnent des observations similaires. Par conséquent, nous allons considérer les distributions cumulatives inverses, pour lisser les irrégularités.

Nous allons discuter les résultats obtenus en considérant les deux définitions possibles d'un utilisateur présentées dans la section 3.3.1.

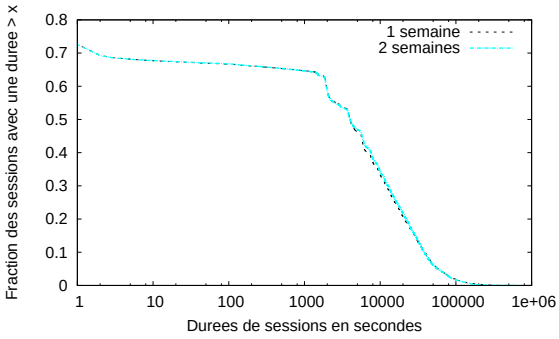
[2] Une étude détaillée des durées de sessions en considérant d'autres valeurs de seuil serait sans doute utile. Cependant, notre objectif ici est d'illustrer notre méthodologie et de montrer que nous pouvons obtenir des observations intéressantes sur les caractéristiques des durées de sessions. D'autres seuils mènent à des résultats similaires à ceux que nous présentons ici.

3.3.2.1 Adresses IP

Dans cette partie, nous considérons qu'un utilisateur correspond à une adresse IP.



(a) l = 1h, 12h, 1 jour et 4 jours.



(b) $l = 1$ et 2 semaines (les deux distributions sont presque superposées).

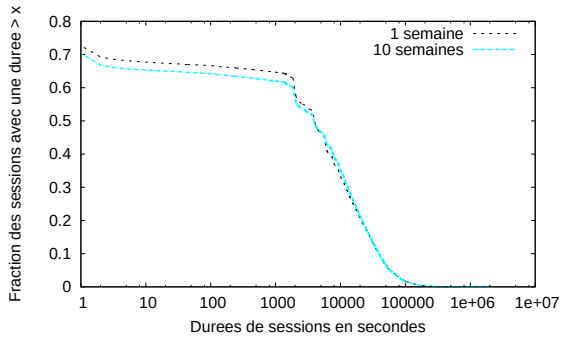
FIG. 3.2 – Distribution cumulative inverse de $S(W_{0,l})$ pour différentes longueurs de fenêtres d'observation l, pour les *données requêtes*. Un utilisateur correspond à une adresse IP.

La figure 3.2 présente la distribution cumulative inverse de $S(W_{0,l})$ pour différentes valeurs de l, allant de $l = 1$ heure jusqu'à $l = 2$ semaines. On peut remarquer que les distributions sont décalées verticalement. Ceci s'explique par le fait que les fractions des sessions avec une durée nulle (qui correspondent aux utilisateurs qui font une seule requête) ne sont pas les mêmes [3].

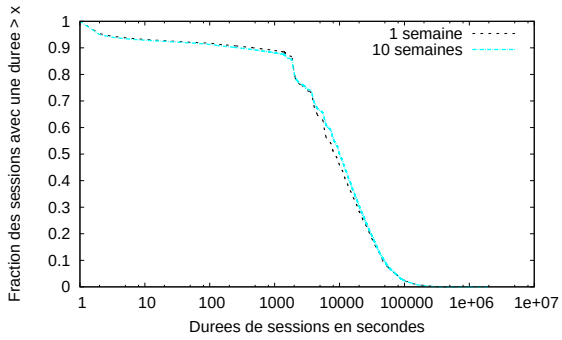
Les formes de ces distributions sont néanmoins similaires, avec une petite fraction de sessions de durée inférieure à 2 000 secondes, et une forme à peu près linéaire entre 2 000 et 100 000 secondes. Cependant, pour des fenêtres d'observation de longueur $l \leq 1$ jour, les distributions s'arrêtent brusquement. Ceci n'est plus le cas quand $l \geq 4$ jours : la queue de la distribution s'aplatit, après un coude situé aux environs de 100 000 secondes ($\sim$ 28 heures), et l'on observe une petite fraction de valeurs extrêmes après ce coude. Pour des fenêtres d'observation de longueur supérieure à 4 jours, la forme de la distribution semble ne plus évoluer : la figure 3.2 (b) montre que les distributions obtenues pour $l = 1$ semaine et $l = 2$ semaines sont très similaires et ont la même forme que celle obtenue pour $l = 4$ jours.

Cependant, lorsque l augmente à nouveau, on observe une légère différence entre les distributions obtenues. La figure 3.3 (a) présente $S(W_{0,l})$ pour $l = 1$ semaine et $l = 10$ semaines. On observe un petit écart entre ces distributions, causé par la fraction de sessions

[3] Comme l'axe des x est en échelle logarithmique, le point $(0, 1)$ qui appartient à toutes ces distributions n'apparaît pas.



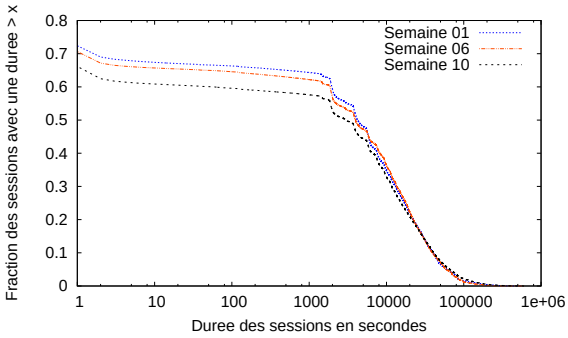
(a) Distribution normalisée.



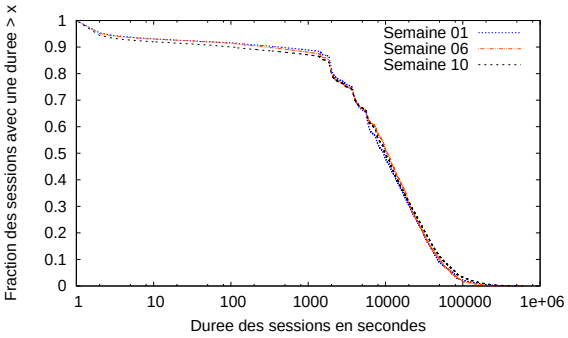
(b) Distribution normalisée par les valeurs supérieures à 0.

FIG. 3.3 – Distributions cumulatives inverses de $S(W_{0,l})$ pour des fenêtres d'observation de longueurs $l = 1$ semaine et $l = 10$ semaines, pour les *données requêtes*. Un utilisateur correspond à une adresse IP.

de durée nulle (qu'on ne voit pas sur la figure à cause de l'échelle logarithmique sur l'axe des x) : lorsque l'on normalise ces distributions par le nombre de sessions de durée supérieure à 0 (figure 3.3 (b)), cet écart disparaît. Ceci montre que la forme de la distribution n'évolue plus, bien que la fraction de sessions de longueur nulle, elle, varie.



(a) Distribution normalisée.



(b) Distribution normalisée par les valeurs supérieures à 0.

FIG. 3.4 – Distributions cumulatives inverses de $S(W_{t_i,l})$ pour une fenêtre d'observation de longueur $l = 1$ semaine avec des points de départs $t_1 = 0$, $t_2 = 6$ semaines et $t_3 = 10$ semaines, pour les *données requêtes*. Un utilisateur correspond à une adresse IP.

Afin d'étudier l'influence de l'instant de début de la période d'observation, nous considérons des fenêtres $W_{t_1,l}$ et $W_{t_2,l}$ de même longueur mais avec des débuts différents. Nous présentons dans la figure 3.4 les distributions cumulatives inverses de $S(W_{t_i,l})$ pour des

fenêtres d'observation de longueur $l = 1$ semaine avec 3 instants de départs t_i différents : le début de la première semaine, le début de la sixième semaine et le début de la dixième semaine.

Nous observons qu'en général $S_0(W_{t_1,l}) \neq S_0(W_{t_2,l})$. Comme on peut l'observer sur la figure 3.4 (a), cette différence est due au décalage vertical entre les distributions qui est causé par la fraction de sessions de durée nulle qui diffère entre ces distributions. Ceci montre que la fraction $S_0(W_{t,l})$ de sessions de durée nulle dépend à la fois de t et de l, mais que la forme générale de la distribution, lorsque cette fraction n'est pas prise en compte (voir la figure 3.4 (b)), ne change pas.

Visuellement, nous avons vu que les distributions semblent ne plus évoluer une fois que la longueur de la fenêtre d'observation a atteint 4 jours. Cependant, il faut considérer les observations visuelles avec prudence. En effet, la figure 3.5 montre les distributions pour $l = 1$ semaine et $l = 2$ semaines, mais avec une échelle linéaire sur l'axe des x et une échelle logarithmique sur l'axe des y.

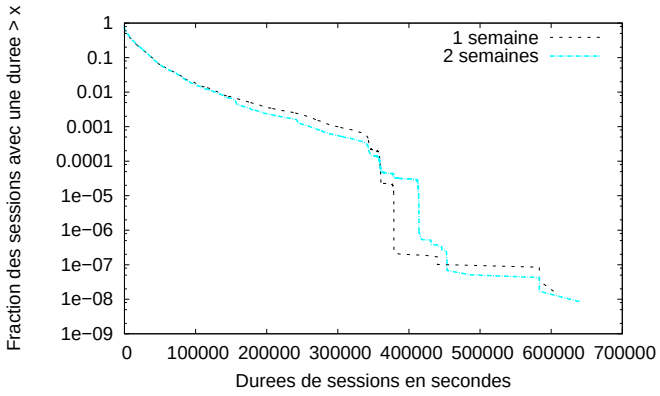
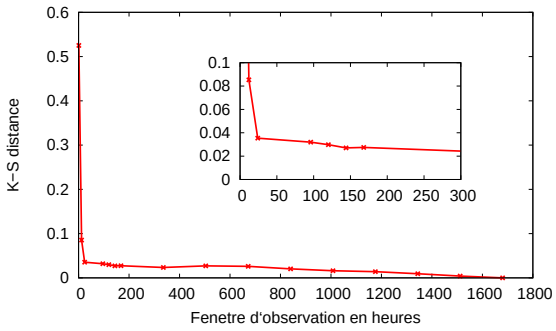
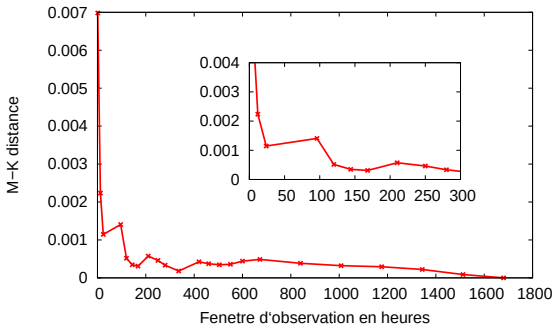


FIG. 3.5 – Distribution cumulative inverse de $S(W_{0,l})$ pour des fenêtres d'observation de longueurs $l = 1$ semaine et $l = 2$ semaines en échelle lin-log, pour les *données requêtes*. Un utilisateur correspond à une adresse IP.

À première vue, les distributions semblent très différentes l'une de l'autre. Cependant, on peut constater qu'elles sont similaires pour au moins 99% des valeurs. Elles ne diffèrent que pour des valeurs supérieures à environ 150 000 secondes, qui sont les valeurs vues après le coude de la figure 3.2 (b), et sont significativement plus rares que les valeurs inférieures à ce coude. Pour cette raison, on les considère comme étant des valeurs *extrêmes*. Le fait que les valeurs extrêmes changent quand l augmente montre qu'elles ne peuvent pas être caractérisées par notre méthodologie. Leur étude constitue une perspective intéressante.



(a) KS($S(W_{0,l})$, $S(W_{0,l_{max}})$) en fonction de l.



(b) MK($S(W_{0,l})$, $S(W_{0,l_{max}})$) en fonction de l.

FIG. 3.6 – Étude de l'évolution de $S(W_{0,l})$ avec la K-S et la M-K distance, pour les *données requêtes*.

Nous allons étudier maintenant l'évolution des distributions avec la K-S et la M-K distance ainsi que nous l'avons présenté dans le chapitre 2. La figure 4.8 (a) présente, en fonction de l, la K-S distance entre la distribution observée avec une fenêtre de longueur l et celle observée avec la plus grande fenêtre dont nous disposons, l_{max}, c'est-à-dire KS($S(W_{0,l})$, $S(W_{0,l_{max}})$). Les premières valeurs observées sont assez élevées et décroissent rapidement, jusqu'à 4% pour une fenêtre d'observation correspondant à $l = 24$ heures. Après cela, les valeurs ont tendance à décroître plus ou moins linéairement. Ceci montre clairement qu'une fenêtre d'observation de moins de 24 heures n'est pas représentative. Cependant, on ne sait pas si une valeur de 4% est assez petite pour considérer que les distributions sont similaires ou pas. De plus, la forme linéaire ne correspond pas à une valeur qui fluctue avant de devenir stable. Cette courbe ne permet donc pas de décider quand est-ce que la fenêtre d'observation devient suffisamment longue, ni même de savoir si cela se produit pendant la mesure. Par conséquent, on ne peut pas tirer de conclusion d'après la K-S distance.

Nous présentons la comparaison avec la M-K distance dans la figure 4.8 (b) : nous calculons MK($S(W_{0,l})$, $S(W_{0,l_{max}})$) en fonction de l. Nous pouvons observer que le comportement est différent par rapport à celui de la K-S distance : les valeurs observées ont tendance à diminuer (avec des fluctuations), jusqu'à ce que la période de mesure atteigne approximativement 150 heures (6 jours et 6 heures). Après cela, la valeur de la M-K distance devient très petite : cela montre que les distributions correspondantes sont très proches les unes des autres.

Pour finir, nous calculons l'écart type et la moyenne de $S(W_{0,l})$ en fonction de l, présenté dans la figure 3.7. On peut voir que la moyenne se stabilise quand l atteint environ 1 semaine, au même moment que la M-K distance. Cela confirme qu'une fenêtre d'observation d'une semaine est suffisamment longue pour caractériser avec précision une distribution.

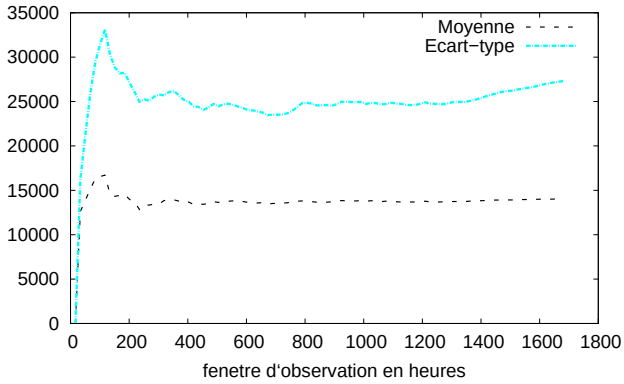


FIG. 3.7 – Moyenne et écart-type de $S(W_{0,l})$, en fonction de l, pour les *données requêtes*.

Cependant, l'écart-type ne semble pas se stabiliser [4] quand on fait augmenter la longueur de la fenêtre d'observation, ce qui confirme que la distribution ne peut pas être caractérisée dans sa totalité. Ceci est cohérent avec la distinction entre la partie normale de la distribution et les valeurs extrêmes. En effet, les valeurs extrêmes sont très grandes et donc ont un impact fort sur l'écart-type. Le fait qu'elles ne peuvent pas être caractérisées fait varier l'écart type, alors que la moyenne est stable car la partie normale de la distribution est caractérisée.

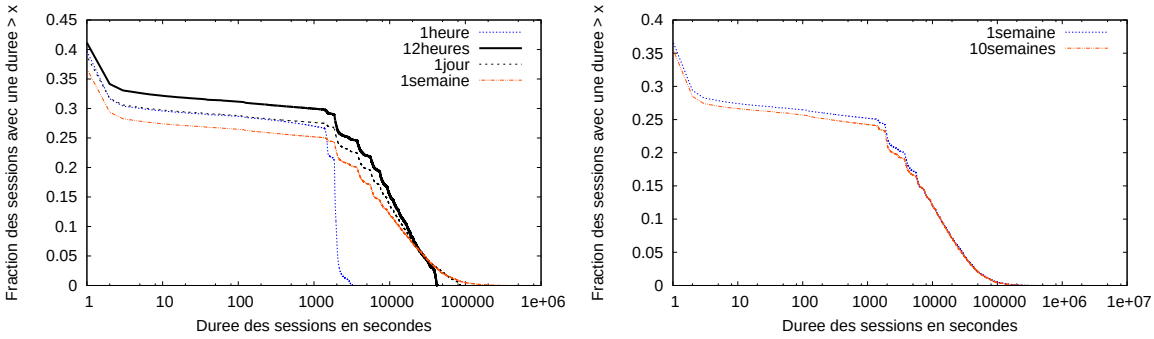
Cela confirme l'intuition obtenue par l'étude visuelle des distributions : une fois que la longueur de la fenêtre d'observation atteint une semaine, la partie normale de la distribution des durées de sessions cesse d'évoluer. Cela signifie deux choses. Premièrement, cette distribution est stationnaire sur des échelles de temps de l'ordre de la totalité de la durée de mesure, et donc cela a du sens de la caractériser. Deuxièmement, une fenêtre d'observation d'une semaine est suffisamment longue pour que l'on puisse la caractériser avec précision. Cependant, les valeurs extrêmes de cette distribution ne peuvent pas être caractérisées par notre méthodologie.

3.3.2.2 Adresses IP et ports UDP

Dans cette partie, nous allons considérer qu'un utilisateur correspond au couple d'informations composé de l'adresse IP et du port UDP utilisé, c'est-à-dire qu'une même adresse IP pour laquelle des requêtes sont effectuées avec deux ports différents sera considérée comme correspondant à deux utilisateurs différents.

[4] Remarquons que si on avait arrêté la mesure à 1200 heures, on aurait l'impression qu'il se stabilise, d'où l'importance d'avoir une fenêtre d'observation assez grande.

Les résultats obtenus coïncident avec ceux que nous avons montré dans la section précédente. Pour cette raison, nous allons présenter seulement quelques courbes dans le but de comparer les deux cas et de montrer à quel point les résultats sont similaires.



(a) l = 1h, 12h, 1 jour et 4 jours. (b) l = 1 et 10 semaines.

FIG. 3.8 – Distribution cumulative inverse de $S(W_{0,l})$ pour différentes longueurs de fenêtres d'observation l, pour les *données requêtes*. Un utilisateur correspond à une adresse IP + un port UDP.

La figure 3.8 présente la distribution cumulative inverse de $S(W_{0,l})$ pour différentes valeurs de l, allant d'une heure à 10 semaines. On remarque que la forme de ces distributions est presque identique à celle des distributions observées dans la figure 3.2. Cependant, on peut remarquer aussi certaines différences. Notamment la fraction de sessions de durée nulle qui est beaucoup plus importante dans ce cas. En effet, pour la distribution correspondant à l = 1 jour, cette fraction dépasse les 60% (voir figure 3.8 (a)) alors que pour la même distribution observée dans la figure 3.2, cette fraction est inférieure à 30%. Ceci s'explique par le fait qu'une seule session de durée supérieure à 0 correspondant à une adresse IP donnée peut correspondre à plusieurs sessions de durée nulle ayant des ports différents (à cause d'un changement du port UDP utilisé par l'adresse IP).

Nous observons aussi le même type de décalage vertical entre les différentes distributions qui est dû au fait que la fraction des sessions de durée nulle change d'une distribution à une autre. En effet, la représentation des deux distributions montrées dans la figure 3.8 (b) en ne prenant en compte que les sessions de durée supérieure à zéro (voir figure 3.9) donne des distributions très similaires et presque superposées.

En conclusion, les observations obtenues avec un autre choix de définition d'un utilisateur permettent encore de confirmer la fiabilité de nos résultats et de montrer que la caractérisation de cette propriété n'est pas dépendante d'un certain choix de paramètres.

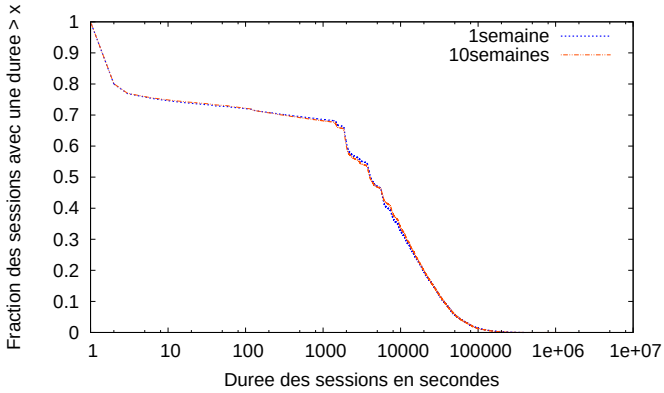
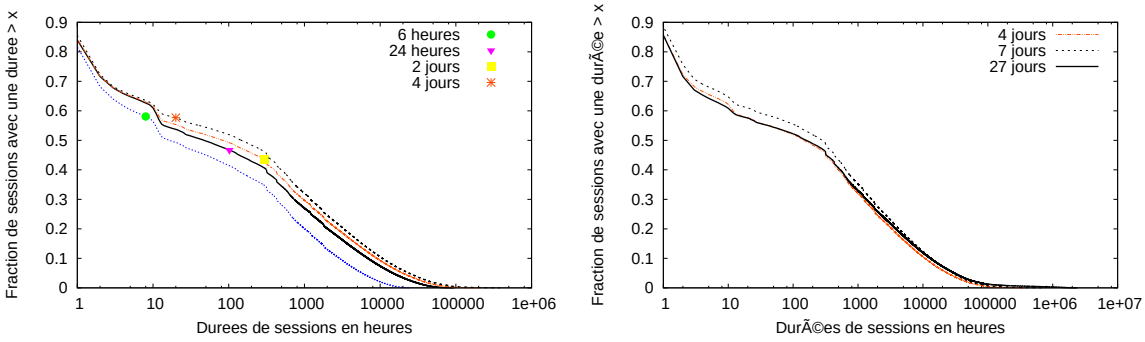


FIG. 3.9 – Distribution cumulative inverse normalisée par les valeurs supérieures à 0 de $S(W_{0,l})$ pour $l = 1$ semaine et $l = 10$ semaines, pour les *données requêtes*. Un utilisateur correspond à une adresse IP + un port UDP.

3.4 Durées de sessions des utilisateurs – *données logins*

Dans cette section, nous nous intéressons à l'étude des distributions des durées de sessions S, dans les *données logins*.



(a) $l = 6$ heures à 4 jours

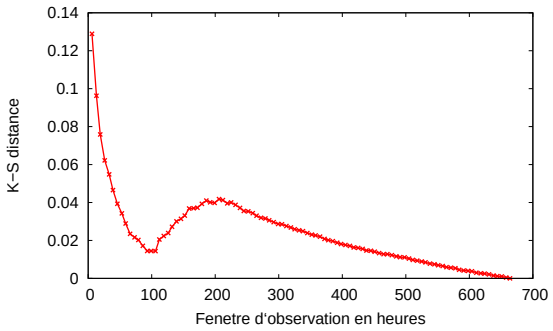
(b) $l = 4$ jours à 27 jours

FIG. 3.10 – Distribution cumulative inverse de $S(W_{0,l})$ pour différentes longueurs de fenêtres d'observation l, pour les *données logins*.

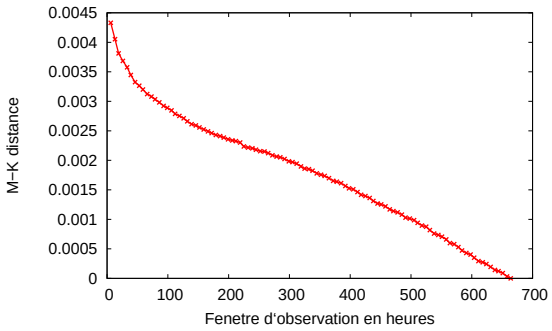
La figure 3.10 présente la distribution cumulative inverse de $S(W_{0,l})$ pour différentes valeurs de l, allant de 6 heures jusqu'à 27 jours. On observe que la forme de ces distributions est similaire et que plus l augmente, plus ces distributions se rapprochent les unes des autres : dans la figure 3.10 (a), on observe que la distribution correspondant à $l = 6$ heures est un peu différente des autres distributions. Pour des longueurs allant de $l = 1$ jour jusqu'à 4 jours, les distributions deviennent proches. Quand on augmente la valeur de l à

7 et 27 jours (voir figure 3.10 (b)), les distributions restent proches, mais on observe aussi que la distribution correspondant à $l = 4$ jours est plus proche de celle de 27 jours que de celle de 7 jours. Cela donne l'impression que ces distributions fluctuent.

Afin d'avoir une meilleure intuition, nous comparons ces distributions avec la K-S et la M-K distance. La figure 3.11 (a) présente KS($S(W_{0,l})$, $S(W_{0,l_{max}})$) en fonction de l. On peut voir qu'au début, les valeurs sont plutôt grandes, puis elles diminuent rapidement pour atteindre environ 2% pour une fenêtre d'observation correspondant à $l = 4$ jours. Après cela, ces valeurs augmentent légèrement jusqu'à $l = 7$ jours, ce est cohérent avec nos observations de la figure 3.10 (b). Après $l = 200$ heures, les valeurs ont tendance à diminuer linéairement.



(a) KS($S(W_{0,l})$, $S(W_{0,l_{max}})$) en fonction de l.



(b) MK($S(W_{0,l})$, $S(W_{0,l_{max}})$) en fonction de l.

FIG. 3.11 – Étude de l'évolution de $S(W_{0,l})$ avec la K-S et la M-K distance, pour les *données logins*.

Lorsque l'on compare les mêmes distributions en utilisant la M-K distance (figure 3.11 (b)), on n'observe pas le même phénomène. Les valeurs obtenues ont tendance à décroître linéairement, ce qui signifie que les distributions changent à un taux à peu près constant. Ceci montre que, bien que les distributions correspondant à $l = 4$ et 27 jours semblent visuellement proches et qu'elles aient une K-S distance relativement petite, elles ne sont pas aussi proches que cela. En effet, une observation plus détaillée des distributions a montré que la distance entre elles n'est pas très grande mais elle est présente pour un grand intervalle de valeurs de l'axe des x. La distance entre les distributions correspondant à 7 et 27 jours n'est grande que pour les petites valeurs de l'axe des x, ce qui fait que visuellement elles semblent moins proches.

Pour finir, nous avons calculé également la moyenne et l'écart-type de $S(W_{0,l})$ en fonction de l, que nous présentons dans la figure 3.12. On peut observer que les valeurs obtenues, tant celles de la moyenne que celles de l'écart-type, augmentent linéairement avec la longueur de la fenêtre d'observation. Ceci est cohérent avec les observations obtenues

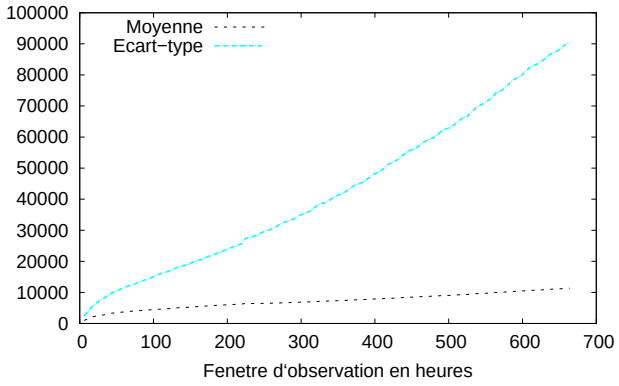


FIG. 3.12 – Moyenne et écart-type de $S(W_{0,l})$ en fonction de l, pour les *données logins*.

avec la M-K distance.

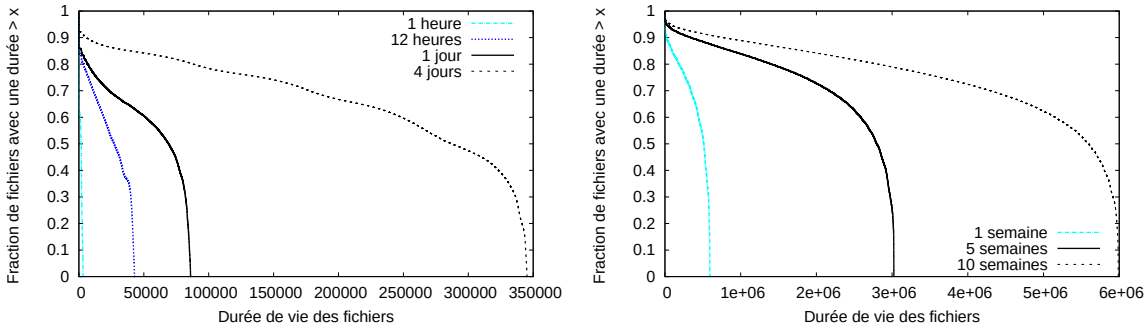
Visuellement, nous avons vu que les distributions $S(W_{0,l})$ sont proches les unes des autres dès que l est suffisamment grand. Cependant, les analyses numériques montrent qu'elles évoluent plus ou moins linéairement avec l. Par conséquent, nous ne sommes pas capables de faire une caractérisation complète de cette propriété car varier la longueur de la période d'observation fait légèrement varier la distribution observée. Cependant, nous avons confiance dans le fait que la forme générale de la distribution est celle que l'on a observé.

3.5 Durées de vie des fichiers

Nous allons maintenant nous intéresser à l'étude de la distribution des durées de vie des fichiers, que l'on note F. Les informations concernant les fichiers ne sont disponibles que dans les *données requêtes*. Dans cette section et le reste de ce chapitre, nous ne considérons que les fichiers pour lesquels il existe au moins un fournisseur, car plusieurs fichiers dans ce jeu de données sont recherchés mais ne sont jamais fournis. Ce sont des fichiers qui n'existent pas dans le système, du moins pendant la mesure, et nous ne les prenons donc pas en compte.

Il y a deux façons possibles de définir la durée de vie d'un fichier. La première est la même que pour les durées de sessions des utilisateurs, et consiste à considérer qu'un fichier n'est pas présent dans le système si l'intervalle de temps entre deux requêtes consécutives pour ce fichier est supérieur à un seuil donné. Dans le deuxième cas, la durée de vie d'un fichier donné est définie par l'intervalle de temps entre la première et la dernière requête

pour ce fichier. On estime que la notion de seuil n'est pas nécessairement pertinente dans ce cas : on s'attend à ce que les fichiers soient plus stables dans le système que les utilisateurs, et que le fait qu'il n'y a plus de requêtes sur un fichier pendant un certain temps ne signifie pas forcément qu'il n'est plus présent dans le système. Nous avons cependant étudié chacune des deux définitions. Dans les deux cas, cette propriété ne se stabilise pas ce qui indique que l'on ne peut pas la caractériser de manière fiable. Nous présentons ici les résultats obtenus en considérant la seconde définition, car ils conduisent à des observations intéressantes.



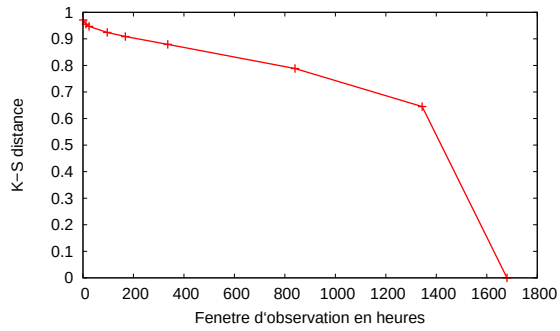
(a) $l = 1$ heure, 12 heures, 1 jour et 4 jours.

(b) $l = 1$, 5 et $l = 10$ semaines.

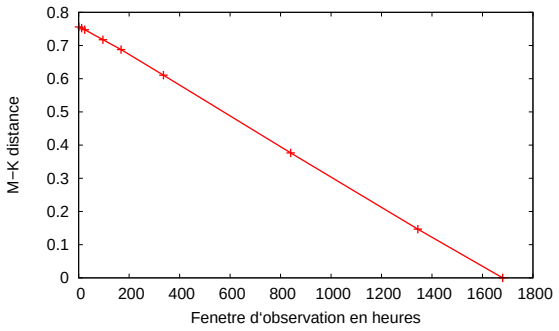
FIG. 3.13 – Distributions cumulatives inverses de $F(W_{0,l})$ pour différentes longueurs de fenêtres d'observation l.

La figure 3.13 (a) présente les distributions cumulatives inverses de $F(W_{0,l})$ pour différentes valeurs de l, de $l = 1$ heure jusqu'à $l = 4$ jours. Nous pouvons voir que la forme des différentes distributions évolue considérablement avec l. Nous obtenons les mêmes observations avec l'étude des distributions correspondant à $l = 1$, 5 et 10 semaines (figure 3.13 (b)). On observe que, plus la fenêtre d'observation est grande, plus les valeurs des durées de vie des fichiers ont tendance à l'être aussi : ceci peut être expliqué par le fait que certains fichiers existent dans le système pour de très longues périodes de temps, et donc leurs durée de vie observée augmente avec la longueur de la fenêtre d'observation.

Afin de confirmer ces observations de manière plus formelle, nous comparons les distributions avec la K-S et la M-K distance. La figure 3.14 (a) présente KS($F(W_{0,l})$, $F(W_{0,l_{max}})$) en fonction de l. On peut voir que les valeurs obtenues sont très élevées et varient quand on augmente l : pour une fenêtre d'observation correspondant à $l = 1344$ heure (8 semaines), la valeur de la K-S distance est toujours supérieure à 60%. Nous comparons également les mêmes distributions avec la M-K distance et nous étudions MK($F(W_{0,l})$, $F(W_{0,l_{max}})$) en fonction de l (figure 3.14 (b)). Les résultats montrent le même comportement que pour la K-S distance : les valeurs observées ont tendance a décroître linéairement et sont aussi très élevées.



(a) KS($F(W_{0,l})$, $F(W_{0,l_{max}})$) en fonction de l.



(b) MK($F(W_{0,l})$, $F(W_{0,l_{max}})$) en fonction de l.

FIG. 3.14 – Étude de l'évolution de $F(W_{0,l})$ avec la K-S et la M-K distance.

Enfin, nous présentons dans la figure 3.15 la moyenne et l'écart-type des distributions $F(W_{0,l})$, en fonction de l. Remarquons que les deux évoluent de manière continue avec l'augmentation de la longueur de la fenêtre d'observation. Ces observations sont cohérentes avec la figure 3.13 et montrent que, plus la fenêtre d'observation est grande, plus les durées de vie observées des fichiers le sont.

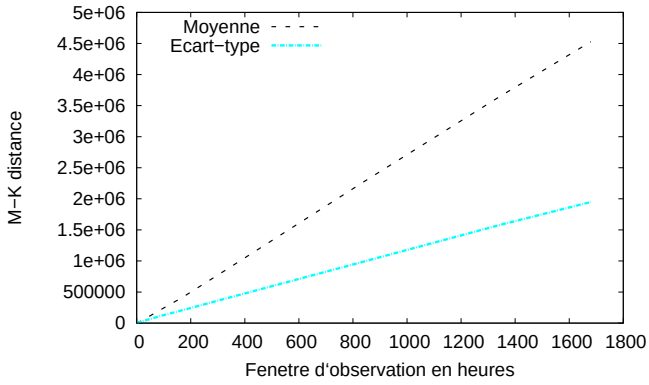
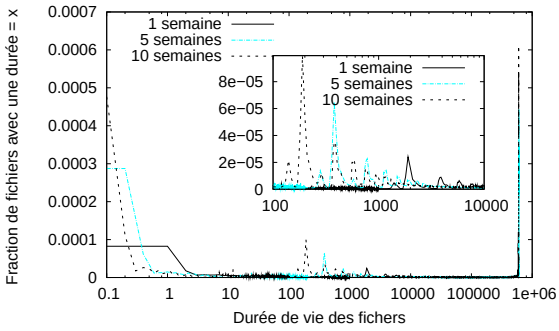


FIG. 3.15 – Moyenne et écart-type de $F(W_{0,l})$, en fonction de l.

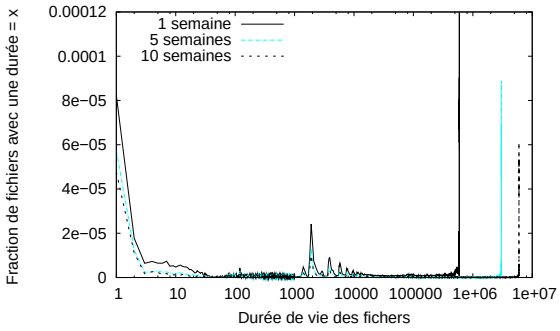
Afin de vérifier si cette propriété évolue réellement d'une façon linéaire avec la longueur de la fenêtre d'observation l, nous étudions la manière dont les distributions classiques (non pas les cumulatives inverses) évoluent en fonction de l, et ceci en les normalisant par rapport à l ainsi que nous l'avons présenté dans le chapitre 2. Pour ce faire, on divise les valeurs de l'axe des x de la distribution $F(W_{0,l})$ par la longueur de la fenêtre d'observation

l. Pour obtenir des distributions normalisées, nous devons aussi multiplier les valeurs de l'axe des y par l.

Nous présentons les distributions normalisées obtenues dans la figure 3.16 (a), pour $l = 1, 5$ et 10 semaines. Afin de mieux comprendre ces courbes, nous présentons les distributions classiques (c'est-à-dire non normalisées) dans la figure 3.16 (b). Nous pouvons observer plusieurs choses.



(a) Distributions normalisées par rapport à l.



(b) Distributions classiques.

FIG. 3.16 – Distributions de $F(W_{0,l})$ pour des longueurs de fenêtres d'observations de $l = 1$, 5 et 10 semaines.

Premièrement, toutes les distributions normalisées présentent des pics aux valeurs maximales possibles (604 800 secondes = 1 semaine, qui est l'unité de normalisation pour cette courbe). Cela correspond au fait qu'une fraction de fichiers relativement importante a une durée de vie égale à la longueur de la fenêtre d'observation, comme on peut l'observer dans la figure 3.16 (b).

Deuxièmement, les distributions normalisées présentent quelques pics intermédiaires, qui ne sont pas situés aux mêmes valeur x pour les différentes distributions. Ceci est dû au fait que les distributions non normalisées (figure 3.16 (b)) présentent des pics qui coïncident. Nous avons déjà observé le même phénomène pour les durées de sessions des utilisateurs dans la section 3.3, dû au fait que certains clients envoient des requêtes périodiques, voir figure 3.1 (a). Lorsque les distributions sont normalisées, ces pics se décalent et donc ne coïncident plus.

Comme la K-S et la M-K distance ne peuvent être calculées que pour des distributions cumulatives, ainsi que nous l'avons expliqué dans la section 2.3.3, il n'est pas possible de les calculer pour les distributions montrées dans la figure 3.16 (il n'existe pas de méthode naturelle pour calculer la cumulative d'une distribution normalisée de cette façon). Par conséquent, nous avons seulement étudié la moyenne et l'écart-type des distributions normalisées, et nous les présentons dans la figure 3.17. Nous observons qu'après quelques

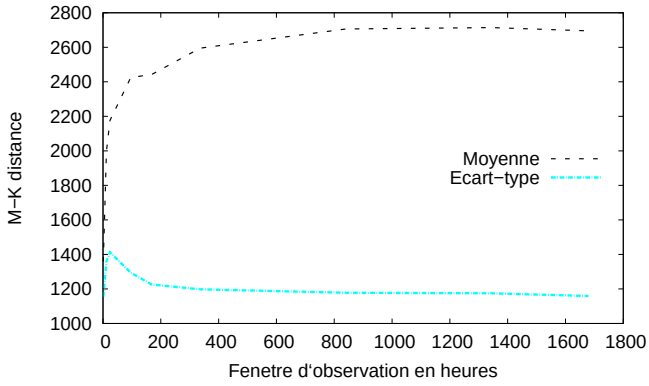


FIG. 3.17 – Moyenne et écart-type (pour les distributions normalisées par rapport à l) de $F(W_{0,l})$, en fonction de l.

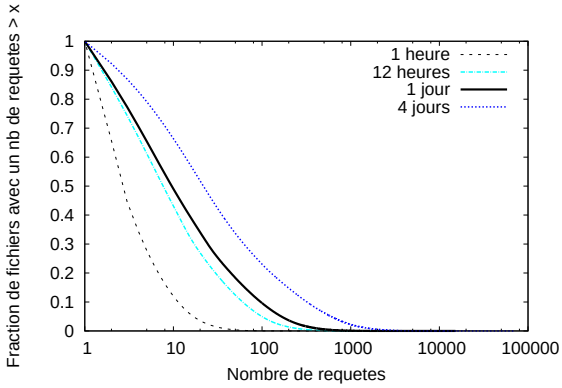
fluctuations au début, chacun des deux se stabilise (l'écart-type se stabilise plus rapidement que la moyenne). Il est intéressant de noter que le fait que la moyenne et l'écart-type se stabilisent ne veut pas forcément dire que les distributions correspondantes se stabilisent aussi.

Pour conclure, cette propriété ne peut pas être caractérisée dans notre mesure. Les distributions non normalisées évoluent d'une façon linéaire avec la longueur de la fenêtre d'observation. Normaliser les distributions par rapport à la longueur de la fenêtre d'observation montre que cette évolution n'est pas régulière, même s'il est possible de caractériser leur moyenne et écart-type. Une question importante qui reste ouverte est de savoir si cette propriété pourrait être caractérisée avec une mesure de plus de 10 semaines, ou si c'est la propriété elle même qui n'est pas stationnaire.

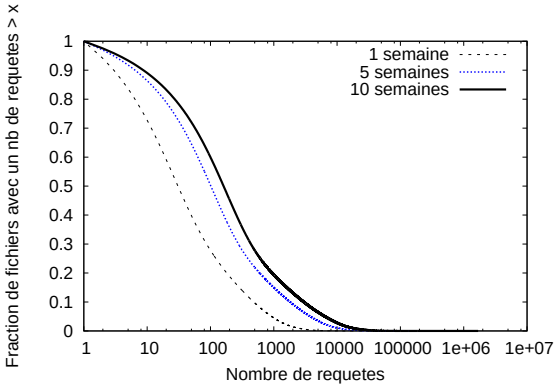
3.6 Nombre de requêtes par fichier

Nous allons maintenant étudier les distributions du nombre de requêtes par fichier Q, dans les *données requêtes*.

La figure 3.18 présente les distributions cumulatives inverses de $Q(W_{0,l})$ pour différentes valeurs de l, allant de 1 heure jusqu'à 10 semaines. Nous pouvons voir que les différentes distributions partagent certaines propriétés communes : globalement, on observe une forme linéaire au début de chaque distribution, ce qui montre qu'il y a une grande fraction de fichiers avec un petit nombre de requêtes (cette fraction diminue quand l augmente). Par contre, la queue de ces distributions a tendance à s'aplatir ce qui signifie qu'il y a une petite



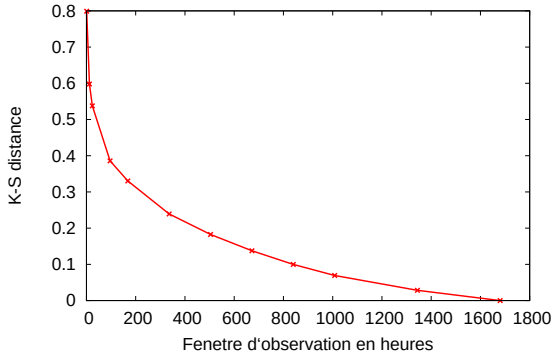
(a) $l = 1$ heure, 12 heures, 1 jour et 4 jours.



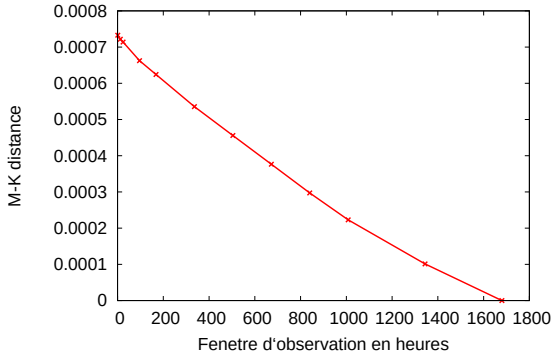
(b) $l = 1$, 5 et $l = 10$ semaines.

FIG. 3.18 – Distributions cumulatives inverses de $Q(W_{0,l})$ pour différentes longueurs de fenêtres d'observations l.

fraction de fichiers avec un très grand nombre de requêtes. Nous observons aussi que les distributions évoluent significativement avec l : le nombre de requêtes par fichier augmente avec la longueur de la fenêtre d'observation.



(a) KS($Q(W_{0,l})$, $Q(W_{0,l_{max}})$) en fonction de l.



(b) MK($Q(W_{0,l})$, $Q(W_{0,l_{max}})$) en fonction de l.

FIG. 3.19 – Étude de l'évolution de $Q(W_{0,l})$ avec la K-S et la M-K distance.

Nous confirmons avec la K-S et la M-K distance que cette propriété ne se stabilise pas. La figure 3.19 (a) présente KS($Q(W_{0,l})$, $Q(W_{0,l_{max}})$) en fonction de l. Tout d'abord, on observe bien que les valeurs obtenues sont très grandes : elles commencent à environ 80% pour une fenêtre d'observation correspondant à $l = 12$ heures, pour atteindre environ 35% pour $l = 1$ semaine. Après cela, les valeurs ont tendance a décroître linéairement. La M-K distance (présentée dans la figure 3.19 (b)) suit presque le même comportement, sauf que

les valeurs ont tendance à décroître d'une façon plus linéaire. Ces observations sont tout à fait cohérentes avec celles faites sur la figure 3.18.

Dans la figure 3.20, nous présentons la moyenne et l'écart-type de $Q(W_{0,l})$ en fonction de l. On peut clairement voir que les valeurs obtenues pour les deux, comme prévu, ont tendance a augmenter linéairement avec la longueur de la fenêtre d'observation.

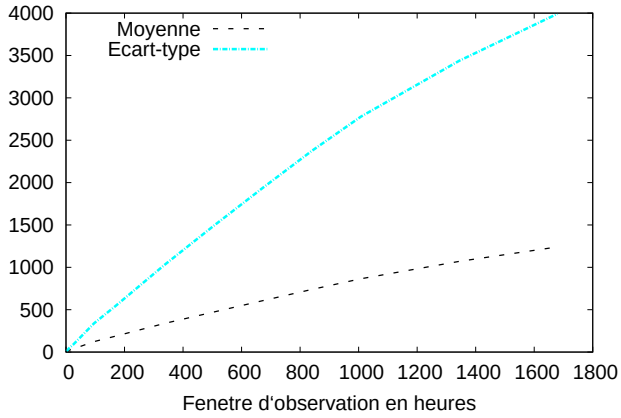
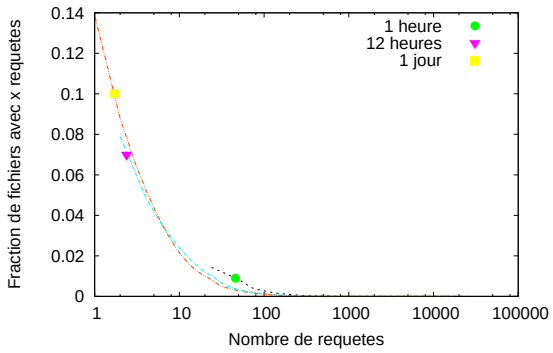
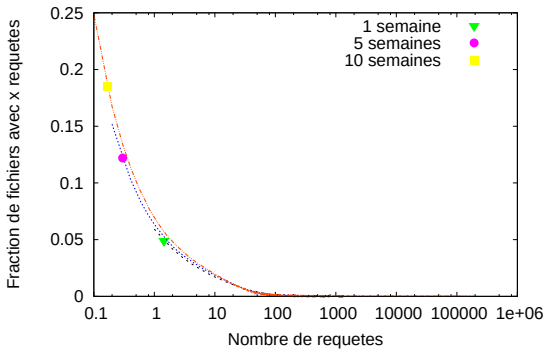


FIG. 3.20 – Moyenne et écart-type de $Q(W_{0,l})$, en fonction de l.



(a) $l = 1$ heure, 12 heures et 1 jour.



(b) $l = 1$, 5 et 10 semaines.

FIG. 3.21 – Distributions de $Q(W_{0,l})$ pour différentes fenêtres d'observations l, normalisées par rapport à l.

Comme on l'a observé précédemment pour les durées de vie des fichiers, dans la section 3.5, les distributions semblent évoluer linéairement avec la longueur de la fenêtre d'observation. Afin de vérifier cette intuition, nous étudions les distributions normalisées par rapport à la longueur de la fenêtre d'observation. On effectue cette normalisation de la

même façon que dans la section précédente, c'est-à-dire que l'on divise les valeurs de l'axe des x par l et que l'on multiplie celles de l'axe des y par l. Les distributions obtenues sont présentées dans la figure 3.21. On peut observer que ces distributions coïncident, ce qui signifie qu'elles évoluent linéairement avec la longueur de la fenêtre d'observation.

Ceci est confirmé par le calcul de la moyenne et l'écart-type des distributions normalisées que nous présentons dans la figure 3.22. Nous observons que les valeurs obtenues pour la moyenne et l'écart type suivent le même comportement : au début, elles ont tendance à diminuer rapidement, puis se stabilisent une fois que l atteint environ 1 semaine. Notons que l'écart-type décroît légèrement avec l, ce qui indique que la proportion des valeurs très grandes (après le coude de la figure 3.18) a tendance a diminuer. Savoir s'il deviendra complètement stable avec une fenêtre d'observation plus longue reste une question ouverte.

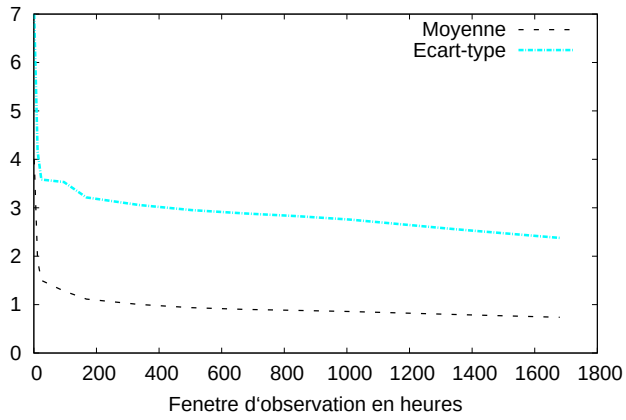


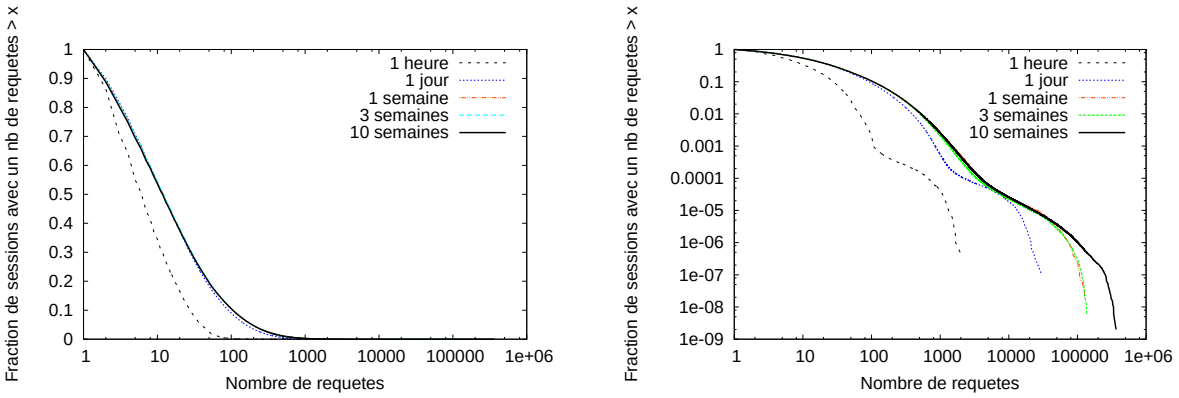
FIG. 3.22 – Moyenne et écart-type (pour les distributions normalisées par rapport à l) de $Q(W_{0,l})$, en fonction de l.

Enfin, nous pouvons conclure que les distributions du nombre de requêtes par fichier évoluent quand la longueur de la fenêtre d'observation l augmente. Néanmoins, l'étude de ces distributions en les normalisant par l montre que cette évolution est linéaire, ce qui signifie que nous sommes capables de caractériser cette propriété, au sens où nous pouvons prédire la distribution pour une valeur de l supérieure à la période de mesure.

3.7 Nombre de requêtes par sessions

Dans cette section, nous étudions la distribution du nombre de requêtes par session G, dans les *données requêtes*. Nous considérons la même définition de sessions que celle

de la section 3.3.1, et nous étudions le nombre de requêtes effectuées par l'utilisateur correspondant dans chaque session.



(a) Échelle log-lin. (b) Échelle log-log.

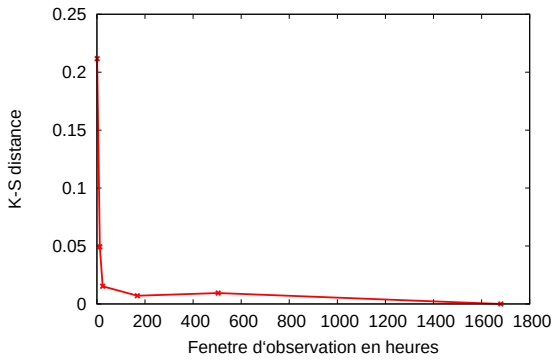
FIG. 3.23 – Distributions cumulatives inverses de $G(W_{0,l})$ pour différentes longueurs de fenêtres d'observations l.

La figure 3.23 présente la distribution cumulative inverse de $G(W_{0,l})$ pour différentes valeurs de l, à partir de $l = 1$ heure jusqu'à $l = 10$ semaines. Dans la figure 3.23 (a), on présente ces distributions en échelle logarithmique sur l'axe des x et en échelle linéaire sur l'axe des y. On peut observer que la forme des distributions est très similaire, avec une grande fraction de sessions ayant peu de requêtes et une petite fraction de sessions avec plus de 1 000 requêtes. Nous observons que pour une fenêtre d'observation de longueur supérieure à 1 jour, les distributions sont presque totalement superposées et ne semblent plus évoluer quand l augmente.

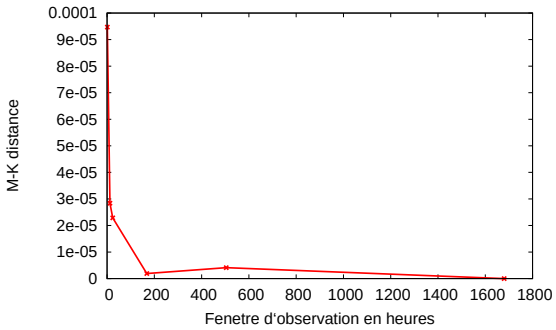
Lorsque l'on compare les mêmes distributions mais en utilisant une échelle logarithmique sur les deux axes (figure 3.23 (b)), on observe visuellement qu'elles semblent plus différentes. Cependant, nous pouvons observer que les distributions correspondant à $l = 1$, 3 et 10 semaines sont similaires pour plus de 99% des valeurs. Elles diffèrent seulement pour les valeurs supérieures à 1 000, qui sont après le coude de la figure 3.23 (a).

La figure 3.24 présente $\mathrm{KS}(G(W_{0,l}), G(W_{0,l_{max}}))$ et $\mathrm{MK}(G(W_{0,l}), G(W_{0,l_{max}}))$ en fonction de l. Nous remarquons que les deux suivent le même comportement : les premières valeurs sont élevées et diminuent rapidement jusqu'à $l = 24$ heures. Après cela, elles diminuent légèrement et ont tendance à se stabiliser après 1 semaine. Cela montre que les distributions correspondantes sont très proches les unes aux autres, ce qui est assez cohérent avec les observations obtenues à partir de la figure 3.23.

Dans la figure 3.25, nous présentons la moyenne et l'écart-type de $G(W_{0,l})$ en fonction de l. Nous observons que la valeur de la moyenne augmente légèrement au début et devient



(a) KS($G(W_{0,l})$, $G(W_{0,l_{max}})$) en fonction de l.



(b) MK($G(W_{0,l})$, $G(W_{0,l_{max}})$) en fonction de l.

FIG. 3.24 – Étude de l'évolution de $G(W_{0,l})$ avec la K-S et la M-K distance.

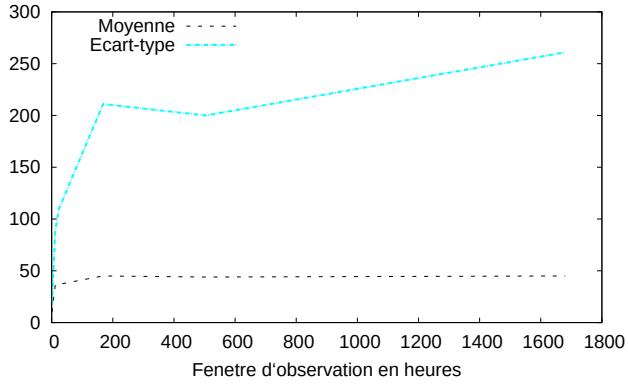


FIG. 3.25 – Moyenne et écart-type de $G(W_{0,l})$, en fonction de l.

stable une fois que l atteint 1 semaine, en même temps que la K-S et la M-K distance. Ceci montre qu'une fenêtre d'observation d'une semaine est suffisamment longue pour caractériser la forme de la distribution de cette propriété. Cependant, l'écart-type n'a pas tendance à se stabiliser quand la longueur de la fenêtre d'observation augmente. Ceci peut être expliqué par la présence de très grandes valeurs (supérieures à 1 000) observées dans la figure 3.23.

Pour finir, nous pouvons constater que cette propriété a un comportement très similaire à celui de la première propriété que nous avons étudiée (durées de sessions des utilisateurs, section 3.3). Nous distinguons deux parties dans la distribution : la première correspond à la grande fraction de sessions avec moins de 1 000 requêtes, que nous sommes capables de caractériser. La deuxième correspond à la petite fraction des valeurs extrêmes qui ne sont

pas caractérisées par notre méthodologie.

3.8 État de l'art

Lorsque l'on a effectué des mesures sur un graphe de terrain, les données obtenues sont incomplètes : il est absolument impossible dans l'immense majorité de capturer tous les nœuds et liens. La métrologie à donc pour but d'étudier ce biais et de tenter de le corriger en évaluant la représentativité de l'échantillon obtenu et d'en extraire des informations pertinentes.

Plusieurs travaux ont également abordé la question de la méthodologie dans un réseau dynamique et les biais possibles qui apparaissent dans ce contexte. Par exemple, le fait que l'on n'est capable d'observer un système que pour une période de temps finie nous empêche d'observer certains événements (ceux qui apparaissent avant ou après la période de mesure), ce qui induit un biais dans les observations.

L'étude de ce type de biais à été surtout reconnu pour la dynamicité (caractérisée par les arrivées et les départs) des utilisateurs, dans les systèmes P2P [BSV03,RLA00,SGG03, WYL07,SR06,SRD+09], mais aussi dans d'autres contextes [LM08,KW06].

Willinger *et al.* [WAL04] se sont intéressés, dans le contexte des flux IP, à la question de savoir si la fenêtre d'observation est suffisamment longue pour caractériser certaines propriétés dynamiques. Ils ont étudié l'écart-type de la distribution des tailles de flux en fonction de la longueur de mesure, et indiquent que le fait qu'il ne se stabilise pas signifie que les échantillons peuvent provenir d'une distribution avec une variance infinie. Ceci peut à son tour rendre difficile l'ajustement des propriétés observées avec un modèle.

Le fait qu'il est seulement possible de capturer les tailles de sessions qui commencent et finissent pendant la période de mesure, fait qu'il est moins probable d'observer les sessions longues et donc crée un biais vers les sessions courtes. Pour supprimer ce biais, la méthode *create-based* [RLA00, SGG03] propose de diviser la fenêtre de mesure de taille T en 2 moitiés, et de ne considérer que les sessions qui commencent dans la première moitié et qui ont une durée inférieure à $T/2$. Cela permet d'avoir une estimation non biaisée des sessions ayant une longueur inférieur à $T/2$.

Cette méthodologie est complémentaire à la notre, qui ne supprime pas formellement le biais, mais qui permet d'avoir des observations sur la forme de la distribution même pour des valeurs supérieures à $T/2$. Par ailleurs, nos observations ont montré que si une fenêtre de mesure est trop courte, la méthode *create-based* ne sera pas capable dans certains cas de fournir une estimation non biaisée. Un dernier point important par rapport à cette méthode est qu'elle n'est applicable que sur des propriétés pour lesquelles la notion de session peut être définie, ce qui n'est pas toujours le cas. Par exemple, il n'est pas possible de l'appliquer

à l'étude du nombre de requêtes par fichier, contrairement à notre méthodologie.

Par ailleurs, Wang *et al.* [WYL07] affirment que la méthode *create-based* est biaisée quand les données sont obtenus par des captures périodiques d'instantanés, parce que les événements courts peuvent être manqués ou mal observés. Ils proposent un nouvel algorithme d'échantillonnage nommé RIDE (ResIDual-based Estimator) qui permet de mesurer la distribution des durée de sessions avec une grande précision et qui nécessite une faible fréquence d'échantillonnage.

Stutzbach *et al.* [SRD+09] étudient les questions qui se posent lorsque le système complet n'est pas connu, et que les informations sur les nœuds et les liens sont obtenus par une procédure d'échantillonnage. Ils se sont intéressés au problème du choix aléatoire et non biaisé de sommets dans un graphe dynamique, au moyen de marches aléatoire modifiées.

Enfin, le biais induit par le fait que la période de mesure est finie n'est pas le seul qui apparait dans notre contexte. Stutzbach et Rejaie [SR06] ont étudié différents aspects de la dynamique dans 3 différentes classes des réseaux P2P (Gnutella, Kad et BitTorrent). Ils ont analysé rigoureusement les différents types de biais qui peuvent influencer une telle étude, et ont présenté une liste de ceux qui ont été identifiés, qui inclut les problèmes liés à l'identification précise des utilisateurs que nous avons abordés dans ce chapitre.

3.9 Conclusion

Dans ce chapitre, nous avons introduit une méthodologie qui permet de savoir quand le biais induit par le fait que la période d'observation est finie devient négligeable dans les systèmes dynamiques. Nous avons illustré sa pertinence en l'appliquant à l'étude de plusieurs propriétés dans un grand système P2P, en utilisant deux jeux de données différents.

Cela a mené à plusieurs conclusions principales :

- si un système est observé pour une période de temps qui est trop courte, il n'est pas possible d'obtenir une évaluation précise de ses propriétés, ce qui montre la pertinence de notre méthodologie ;
- dans un même système, il est possible de caractériser certaines propriétés, mais pas d'autres. C'est le cas par exemple des *données requêtes*, dans lesquelles il est possible de caractériser avec précision la distribution des durées de sessions, mais pas celle des durées de vie des fichiers. Cela montre qu'il n'y a pas d'échelle de temps absolue qui soit pertinente pour étudier un système dans son ensemble, mais que chaque propriété doit être étudiée indépendamment. Ceci est confirmé par le fait que, pour les propriétés que nous avons été capables de caractériser, la longueur minimale de la fenêtre d'observation nécessaire n'est pas exactement la même. Notre méthodologie ne permet pas de savoir si les propriétés que nous n'avons pas été capables de caractériser

ne sont pas stationnaires, ou si des mesures plus longues seraient nécessaires pour les caractériser ;

- le degré avec lequel nous sommes capables de caractériser les propriétés d'un système varie : dans certains cas, nous sommes capables de caractériser la totalité de la distribution, et dans d'autres cas nous pouvons caractériser la distribution à l'exception de quelques valeurs extrêmes ; dans d'autres cas encore, nous avons une idée sur la forme globale de la distribution, mais nous n'avons pas confiance en ses propriétés numériques exactes. Savoir à quel point on peut avoir confiance en une propriété donnée est d'une importance cruciale pour l'étude de n'importe quel système.

Enfin, l'avantage principal de notre méthodologie est qu'elle est générique et peut être appliquée à n'importe quelle propriété dans n'importe quel système et permet de savoir en quelles propriétés on peut avoir confiance ou pas.

Une direction intéressante pour étendre ce travail serait d'étudier des modèles des différentes propriétés que nous avons étudiées. Cela nous permettrait d'avoir une meilleure intuition sur les phénomènes étudiés, et de confirmer formellement nos résultats. Ceci pourrait également fournir des bornes formelles pour la longueur minimale de la fenêtre d'observation nécessaire pour caractériser une propriété donnée avec une certaine précision.

Nous avons montré que notre méthodologie permet de distinguer le comportement différent de certaines valeurs, dites valeurs extrêmes. Trouver une méthode formelle qui permet de distinguer ce type de valeurs est également une perspective intéressante.

Chapitre 4

Caractérisation des réseaux de contacts entre personnes

Contents

4.1 Introduction

L'utilisation massive d'appareils mobiles depuis plusieurs années à amené la communauté scientifique à étudier les interactions entre des entités en mouvement, connus sous le nom de réseaux de contacts. De tels réseaux sont définis par des ensembles de nœuds représentant les personnes et des ensembles de liens représentant la proximité entre ces personnes. Pour pouvoir mesurer cette proximité, les personnes sont munies de capteurs ou d'appareils mobiles qui envoient périodiquement des messages de présence et enregistrent les messages de présence des autres nœuds qui sont à proximité.

La durée d'un contact entre deux nœuds est définie par l'intervalle de temps entre le moment où ces deux nœuds deviennent suffisamment proches pour que l'un d'eux puisse détecter l'autre, et le moment où aucun des deux ne détecte plus l'autre. Ce type de réseaux est dynamique du fait que ces contacts apparaissent et disparaissent continuellement.

Ce sujet a reçu beaucoup d'attention ces dernières années. Les principales propriétés qui ont été étudiées sont les temps de contact et d'inter-contact entre les nœuds [CMRM07, HCS+05, CE07, SBF+08, TLB+09, CLF07].

Passarella *et al.* [PC11] ont remis en cause l'approche consistant à supposer que la distribution globale des temps de contacts ou d'inter-contacts est représentative des distributions obtenues pour chaque nœud individuellement. Ils ont montré que la distribution globale n'est justement pas représentative des distributions obtenues pour chaque nœud individuellement, et que de plus, se baser sur le comportement de l'ensemble des nœuds peut conduire à des conclusions complètement fausses sur les propriétés du réseau.

Fleury *et al.* [FGRS07] proposent une analyse plus avancée de la structure évolutive de ce type de réseaux, et montrent que la seule caractérisation des durées de contacts et d'inter-contacts au travers d'une loi de puissance n'est pas suffisante pour capturer l'évolution du réseau.

Friggeri *et al.* [FCF+11] ont étudié le biais dans les durées de contacts observées causé par le fait que certains capteurs n'arrivent pas a détecter les autres à certains moments, et ont proposé une méthode de reconstruction.

Le but global de ce chapitre est de fournir des outils pour décrire la dynamique de tels réseaux. Pour atteindre ce but, nous allons analyser différents jeux de données afin de comparer nos observations. Nous allons nous intéresser aux questions suivantes : est-ce que le comportement observé change en fonction de la période d'observation et en fonction du jeu de données ? À quel point le comportement des nœuds affecte-t-il le comportement du système ? Est-on capable de détecter des comportements différents qui ne sont pas liés à la dynamique du système mais à la dynamique intrinsèque d'un nœud donné ? Y a-t-il des nœuds qui ont un comportement différent ? Est-ce qu'il y a des moments où certain nœuds

ont un comportement globalement différent ? Est-ce qu'il y a des nœuds qui changent de comportement au fil du temps ?

4.2 Jeux de données utilisés

4.2.1 Données Rollernet

Ces données ont été collectées lors d'une randonnée roller organisée à Paris [TLB+09]. Cet ensemble de données se compose des enregistrements des contacts entre des capteurs iMote répartis entre 62 participants. Chaque capteur iMote effectue des scans réguliers (toutes les 15 secondes) et enregistre les adresses MAC des périphériques aux alentours qui ont répondu. La durée totale de cette randonnée est d'environ trois heures. Elle est composée de deux sessions de 80 minutes, séparées par une pause de 20 minutes. Ce jeu de données contient environ 60 000 contacts. Dans la suite, nous ne prenons en considération que les contacts d'une durée strictement supérieure à 0.

4.2.2 Données Infocom

Dans ces données, des capteurs iMote ont été distribués à 41 étudiants participant à un *workshop* de la conférence Infocom 2005 qui a duré 4 jours [CHC+07]. Chaque capteur iMote envoie des scans chaque 120 secondes (chaque scan dure 5 secondes) et enregistre les réponses reçues. Ce jeu de données contient environ 30 000 contacts. Dans la suite, nous ne prenons en considération que les contacts d'une durée strictement supérieure à 0.

4.2.3 Données PNAS

Il s'agit d'une expérience visant à mesurer le réseau social d'une école pendant une journée [SKL+10]. Des capteurs ont été confiés à environ 800 personnes dont 655 étudiants, 73 enseignants, 55 employés et 5 autres personnes. Chaque capteur envoie des scans toutes les 20 secondes et enregistre ceux qu'il reçoit. La mesure a duré une journée. Ce jeu de données contient environ 760 000 contacts.

Les trois jeux de données contiennent l'ensemble des contacts, formattés de la manière suivante :

$$ID1,\ ID2,\ T_d,\ T_f,$$

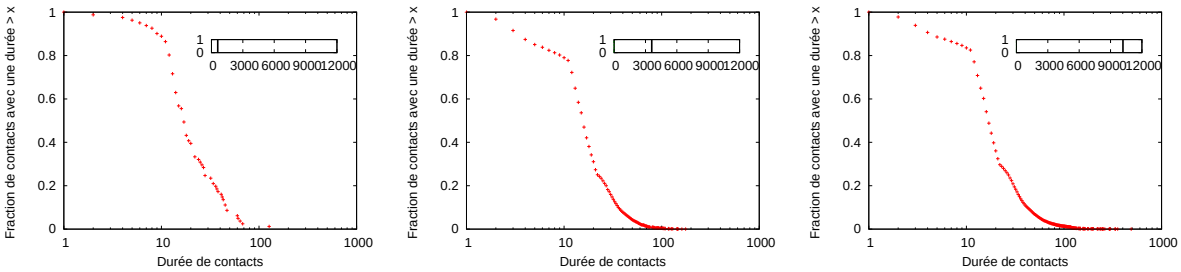
où $ID1$ et $ID2$ correspondent aux identifiants des capteurs entre lesquels le contact a eu lieu, et T_d et T_f correspondent aux temps du début et de la fin de ce contact.

4.3 Impact de la période d'observation

Pour commencer l'étude de ces jeux de données, nous étudions l'impact de la durée de la mesure sur la distribution des durées de contacts (notée C) et des durées d'inter-contacts (notée IC). Pour cela, nous appliquons la méthodologie présentée dans le chapitre précédent.

4.3.1 Distributions des durées de contacts

La durée d'un contact correspond à l'intervalle entre T_d et T_f : $T_f - T_d$. Dans ce qui suit, nous allons présenter les résultats obtenus sur les trois jeux de données.



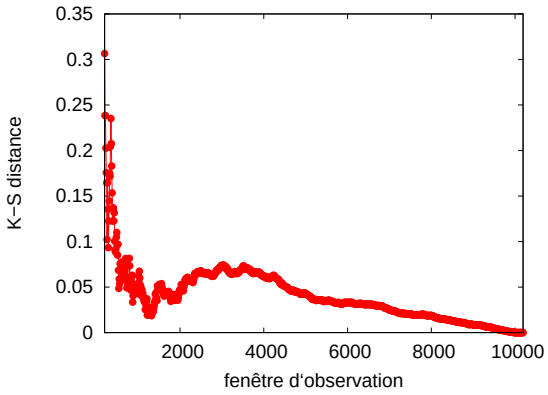
(a) 10 minutes (b) 1 heure (c) 3 heures

FIG. 4.1 – Distribution cumulative inverse des durées de contacts pour différentes longueurs de fenêtres d'observation, pour les données Rollernet.

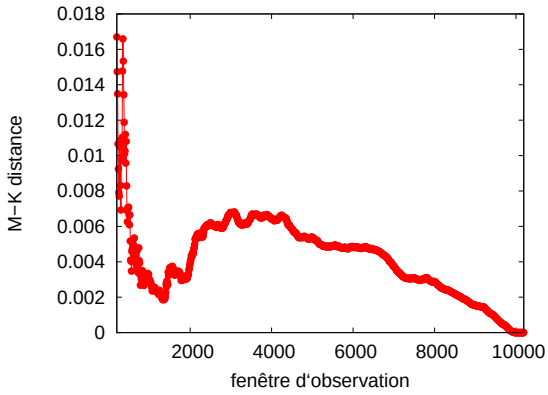
La figure 4.1 présente les distributions cumulatives inverses de C pour différentes fenêtres d'observation allant de 10 minutes à 3 heures, pour les données Rollernet. Le rectangle présenté dans chaque courbe représente la période totale (en secondes) de la mesure (3 heure pour Rollernet), et la ligne verticale représente la durée de la fenêtre d'observation correspondant à la distribution représentée.

Nous pouvons voir que les formes de ces distributions sont presque identiques, avec une très grande fraction (environ 80%) de contacts qui ont une durée comprise entre 10 et 100 secondes et une petite fraction (environs 20%) de contacts avec une durée inférieure à 10 secondes. Nous observons aussi une très petite fraction de contacts avec une durée supérieure à 100 secondes, qui dépend de la taille de la fenêtre d'observation. Pour une fenêtre de 3 heures, la durée maximale peut aller jusqu'à environ 500 secondes, alors que pour une fenêtre d'une heure, elle ne dépasse pas les 200 secondes. Notons que les observations obtenues pour les contacts courts peuvent être biaisées à cause de la granularité (fréquence à laquelle les paquets sont envoyés) qui est égale à 20 secondes pour ces données. En effet, un contact qui dure moins de 20 secondes peut ne pas être capturé s'il

débute après le moment où le premier message est émis et se termine avant le moment où le deuxième message est émis.



(a) KS($C(W_{0,l})$, $C(W_{0,l_{max}})$) en fonction de l.



(b) MK($C(W_{0,l})$, $C(W_{0,l_{max}})$) en fonction de l.

FIG. 4.2 – Étude de l'évolution des durées de contacts $C(W_{0,l})$ avec la K-S et la M-K distance, pour les données Rollernet.

Étudions maintenant l'évolution de ces distributions avec la K-S et la M-K distance. La figure 4.2 (a) présente KS($C(W_{0,l})$, $C(W_{0,l_{max}})$) en fonction de l. Les premières valeurs observées sont assez élevées et ont tendance à décroître rapidement. Après une fenêtre correspondant à un peu moins de 2 000 secondes, les valeurs augmentent légèrement pour finir par décroître d'une manière presque linéaire. Ces résultats ne permettent pas de savoir si les distributions se stabilisent ou pas. Nous présentons la comparaison avec la M-K distance dans la figure 4.2 (b). Nous pouvons observer que le comportement est très similaire à celui de la K-S distance, sauf que les valeurs sont beaucoup plus petites (ce qui est normal puisque la M-K distance correspond à la moyenne des distances alors que la K-S distance correspond à la distance maximale).

Nous avons également calculé la moyenne et l'écart-type de $C(W_{0,l})$ en fonction de l, que nous présentons dans la figure 4.3. On peut observer que les valeurs obtenues pour la moyenne et l'écart-type suivent le même comportement : elles ont tendance à augmenter rapidement au début, puis diminuent jusqu'à une fenêtre d'observation d'environ 1 heure. Après 1 heure, les valeurs augmentent à nouveau légèrement et ne se stabilisent pas à la fin de la mesure.

Nous pouvons donc conclure qu'au-delà du fait que les distributions ont une forme similaire et un comportement similaire pour une grande fraction de valeurs, le fait qu'une partie des valeurs dépendent toujours de la fenêtre d'observation fait que cette propriété ne se stabilise pas complètement, du moins au bout des 3 heures de mesure de Rollernet.

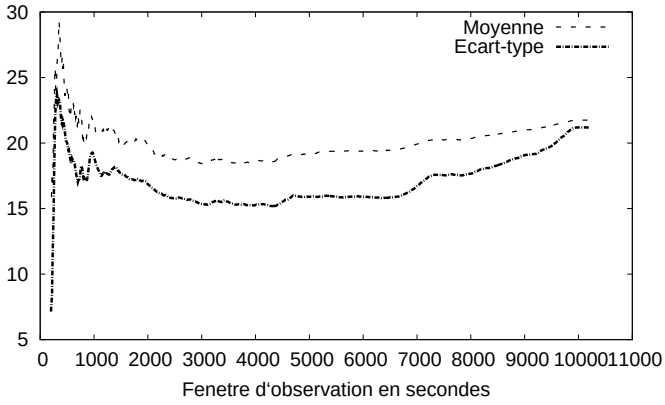
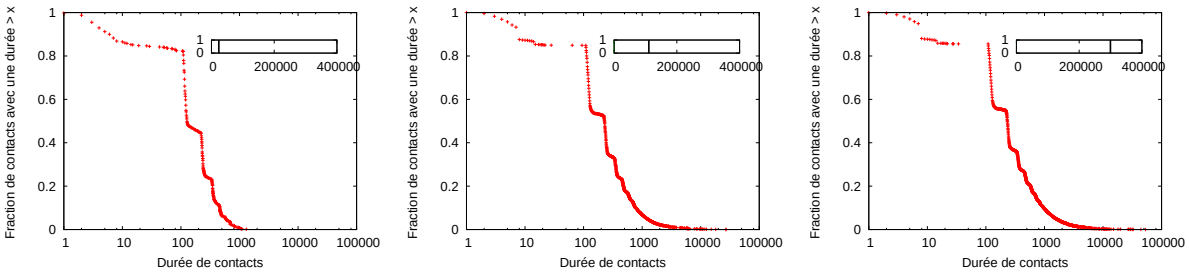


FIG. 4.3 – Moyenne et écart-type de $C(W_{0,l})$, en fonction de l, pour les données Rollernet.

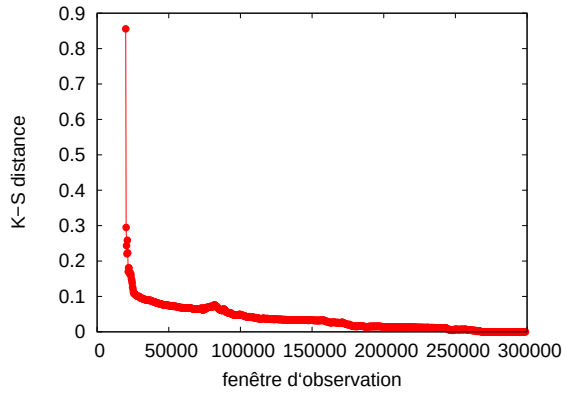


(a) 1 heure (b) 1 jour (c) 3 jours

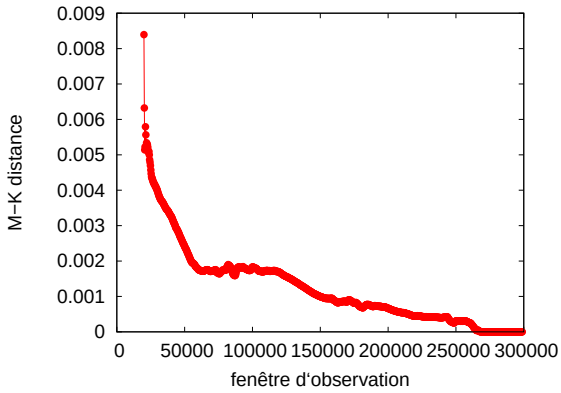
FIG. 4.4 – Distribution cumulative inverse des durées de contacts pour différentes longueurs de fenêtres d'observation, pour les données Infocom.

Nous allons maintenant présenter la même étude pour les données Infocom et PNAS. La figure 4.4 présente les distributions cumulatives inverses de C pour différentes fenêtres d'observations allant de 1 heure à 3 jours, pour les données Infocom. Nous observons un comportement très similaire à celui de Rollernet : une forme similaire, une grande fraction de valeurs comprises entre 100 et 1 000 secondes et une petite fraction qui dépendent de la fenêtre d'observation.

Cependant, les valeurs obtenues pour la K-S et la M-K distance suivent un comportement différent que celui observé dans Rollernet. Les valeurs de la K-S distance (figure 4.5 (a)) commencent par être très élevées au début, puis diminuent rapidement pour atteindre 1% pour une fenêtre d'observation d'environ 25 000 secondes. Après cela, les valeurs ont tendance à diminuer lentement jusqu'à la fin de la mesure. Pour la M-K distance (figure 4.5 (b)), les valeurs ont tendance à avoir le même comportement au début que



(a) KS($C(W_{0,l})$, $C(W_{0,l_{max}})$) en fonction de l.



(b) MK($C(W_{0,l})$, $C(W_{0,l_{max}})$) en fonction de l.

FIG. 4.5 – Étude de l'évolution des durées de contacts $C(W_{0,l})$ avec la K-S et la M-K distance, pour les données Infocom.

pour la K-S distance, bien qu'elles diminuent d'une façon plus remarquable et ne finissent par se stabiliser qu'à la toute fin de la mesure.

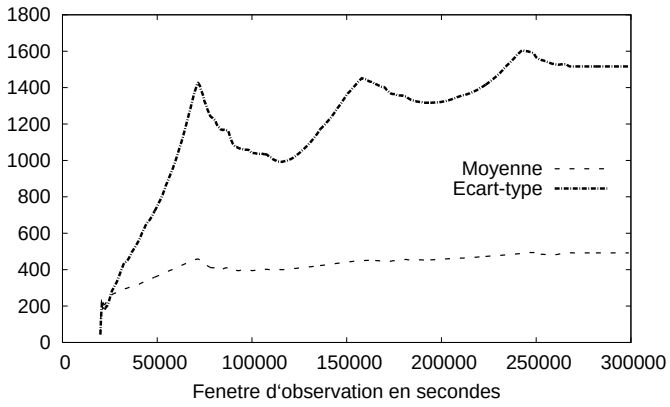
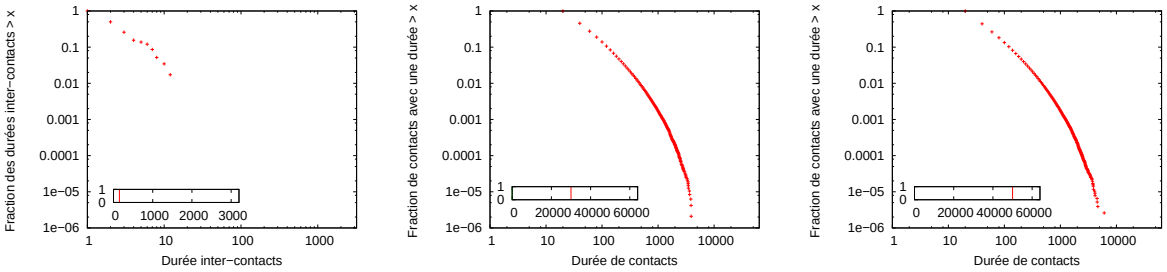


FIG. 4.6 – Moyenne et écart-type de $C(W_{0,l})$, en fonction de l, pour les données Infocom.

Dans la figure 4.6, nous présentons l'évolution de la moyenne et de l'écart-type pour les données Infocom. Les valeurs de la moyenne augmentent légèrement au début, puis elles commencent à se stabiliser à partir d'une fenêtre de 75 000 secondes. Par contre, les valeurs de l'écart-type ont un comportement très différent. Elles ont tendance à fluctuer pendant toute la mesure, et on observe trois pics clairs aux environs de 75 000, 160 000 et 250 000 secondes. En réalité, ces pics correspondent à des moments particuliers des trois jours du

workshop, à savoir aux environ de 8 heures du matin de chaque jour[1]. L'existence de ces pics est due au fait qu'il y a quelques contacts très longs la nuit (entre les participants qui partagent une même chambre ou qui sont dans des chambres voisines). Ces contacts sont donc très éloignés de la moyenne de la durée des contacts qui est relativement petite, ce qui fait que l'écart-type devient très grand. Après 8 heures du matin, les valeurs de l'écart-type diminuent, car les participants créent d'autres contacts avec d'autres participants lors des sessions du *workshop*.

À partir de toutes ces observations, nous pouvons conclure que bien que les différentes distributions des durées de contacts aient une forme similaire, les résultats obtenus pour les différents test statistiques montrent qu'on ne peut pas complètement caractériser cette propriété pour les données Infocom. Par ailleurs, la présence des contacts très longs s'explique par l'existence d'un phénomène jour/nuit, ce qui est tout à fait normal dans ce type de mesures.



(a) 3 000 secondes (b) 30 000 secondes (c) 50 000 secondes

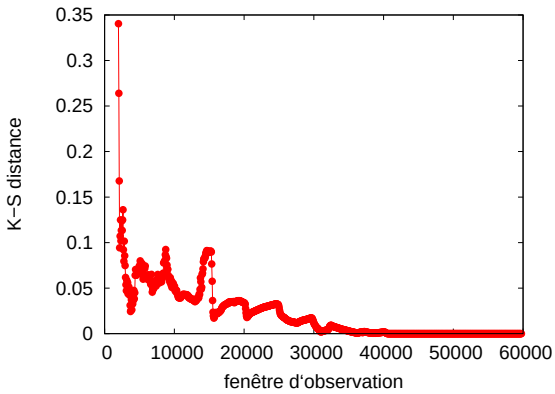
FIG. 4.7 – Distribution cumulative inverse des durées de contacts pour différentes longueurs de fenêtres d'observation, pour les données PNAS.

La figure 4.7 présente les distributions cumulatives inverses de C pour différentes fenêtres d'observations allant de 3 000 secondes à 50 000 secondes, dans les données PNAS.

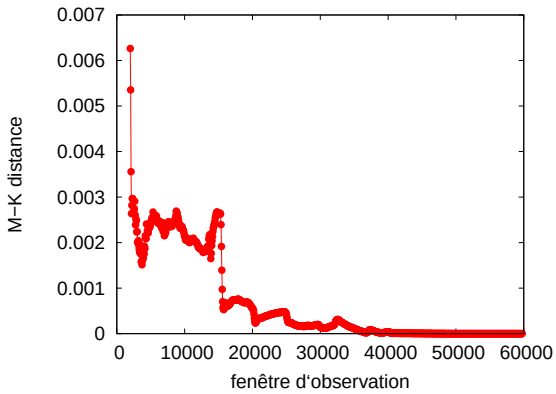
Pour une fenêtre d'observation de 3 000 secondes (figure 4.7 (a)), nous pouvons constater que la durée maximale d'un contact ne dépasse pas les 15 secondes. Ceci est dû au fait qu'on a très peu de contacts au début de la mesure (avant 8 heures du matin). Pour des fenêtres d'observation plus grandes, nous observons un comportement assez similaire, avec une très grande fraction (environ 90%) de contacts ayant une durée inférieure à 100 secondes. Nous observons aussi une petite fraction de contacts longs qui peuvent aller jusqu'à 4 000 secondes (plus d'une heure) pour une fenêtre d'observation de 30 000 secondes, et jusqu'à 6 000 secondes pour une d'observation de 50 000 secondes. Cette partie de la distribution

[1] On n'observe que 3 pics alors que le *workshop* a duré 4 jours. Ceci est dû au fait que les capteurs n'ont été distribués que vers midi du premier jour.

évolue donc quand on change la taille de la fenêtre d'observation. Comme les contacts longs sont rares, elle dépend aussi probablement du moment où est située la fenêtre.



(a) KS($C(W_{0,l})$, $C(W_{0,l_{max}})$) en fonction de l.



(b) MK($C(W_{0,l})$, $C(W_{0,l_{max}})$) en fonction de l.

FIG. 4.8 – Étude de l'évolution des durées de contacts $C(W_{0,l})$ avec la K-S et la M-K distance, pour les données PNAS.

Nous présentons l'étude de l'évolution de ces distribution avec la K-S et la M-K distance dans la figure 4.8. Nous pouvons constater que le comportement des valeurs obtenues est très similaire pour ces deux mesures, avec des grandes fluctuations au début de la mesure et des fluctuations moins visibles à partir d'une fenêtre d'observation de 15 000 secondes. La courbe devient quasiment plate à partir d'une fenêtre d'environ 30 000 secondes, ce qui est dû au fait qu'il y a très peu de contacts après cette période. La première intuition qu'on obtient de ces résultats est qu'on est capable de caractériser des durées des contacts dans les données PNAS.

Afin de confirmer cette intuition, nous étudions également l'évolution de la moyenne et de l'écart-type. Les résultats obtenus (figure 4.9) sont parfaitement cohérents avec ceux de la K-S et M-K distance. Les valeurs de la moyenne et de l'écart-type ont tendance a fluctuer au début de la mesure, puis commencent a se stabiliser à partir d'une fenêtre d'observation d'environ 15 000 secondes. Ceci montre bien que l'on peut caractériser les durées des contacts dans ce jeu de données.

4.3.2 Distributions des durées d'inter-contacts

La durée d'inter-contact correspond à l'intervalle de temps entre le T_f d'un contact et T_d du contact suivant entre la même paire de nœuds. Autrement dit, la durée inter-contact représente le temps pendant lequel il n'y a pas de contact entre une paire de nœuds donnée.

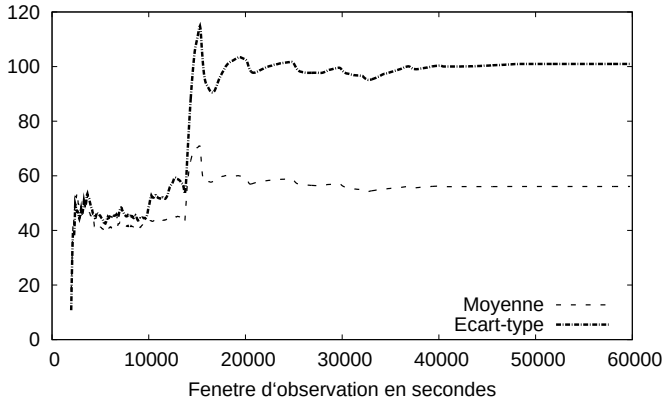
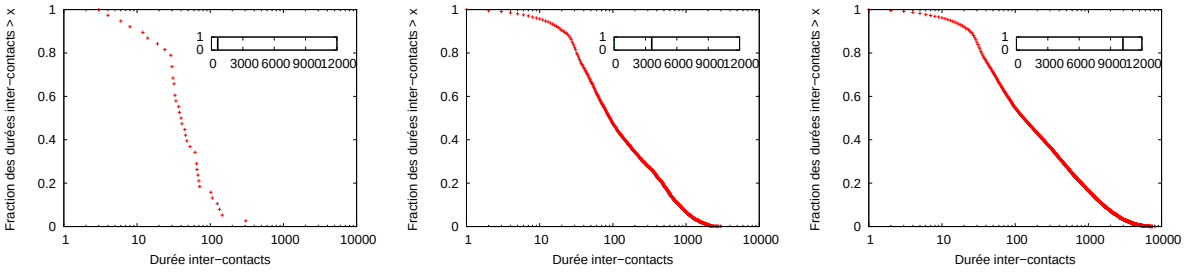


FIG. 4.9 – Moyenne et écart-type de $C(W_{0,l})$, en fonction de l, pour les données PNAS.

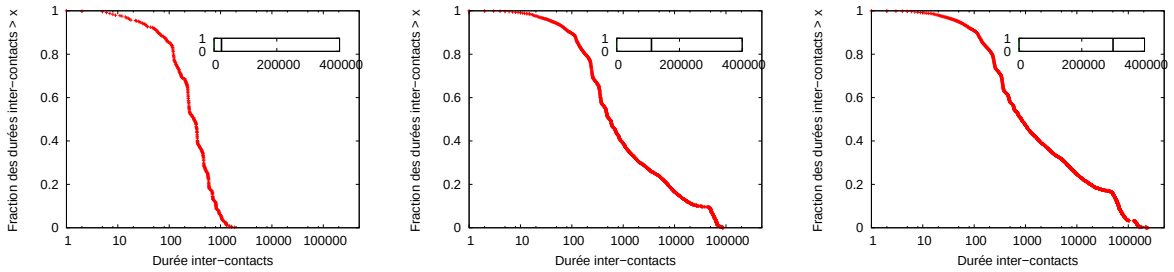


(a) 10 minutes (b) 1 heure (c) 3 heures

FIG. 4.10 – Distribution cumulative inverse des durées d'inter-contacts pour différentes longueurs de fenêtres d'observation, pour les données Rollernet.

La figure 4.10 présente les distributions cumulatives inverses de IC pour différentes fenêtres d'observations allant de 10 minutes à 3 heures, pour les données Rollernet. Nous pouvons voir que la forme globale de ces distributions est similaire et ressemble à celle des distributions obtenues pour les durées de contacts (figure 4.1), mais avec des durées plus grandes. Nous observons que les durées d'inter-contacts peuvent aller jusqu'à la durée totale de la fenêtre d'observation. En effet, pour une fenêtre d'observation d'une heure (figure 4.10 (b)), la durée d'inter-contact maximale dépasse les 3 000 secondes, et pour une fenêtre de 3 heures (figure 4.10 (c)), cette durée dépasse les 8 000 secondes. Ceci montre que certaines paires de nœuds ne sont pas en contact pendant presque toute la durée de la mesure (les contacts qu'ils ont sont situés vers le début et la fin de la mesure).

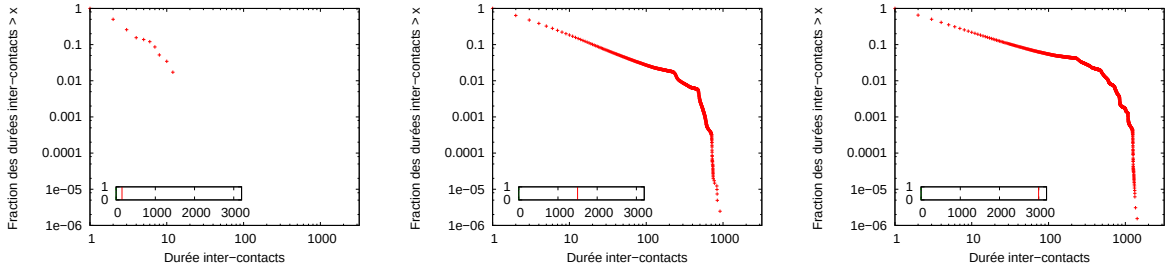
Pour les données Infocom (figure 4.11), nous observons aussi que la forme des différentes distributions est assez similaire. Elle diffère de celle observée pour les durées de contacts



(a) 1 heure (b) 1 jour (c) 3 jours

FIG. 4.11 – Distribution cumulative inverse des durées d'inter-contacts pour différentes longueurs de fenêtres d'observation, pour les données Infocom.

(figure 4.4) : il y a une grande fraction de grandes valeurs. Nous observons aussi le même phénomène que pour Rollernet : on a des durées d'inter-contacts qui sont presque égales à la durée totale de la fenêtre d'observation.



(a) 3 000 secondes (b) 30 000 secondes (c) 50 000 secondes

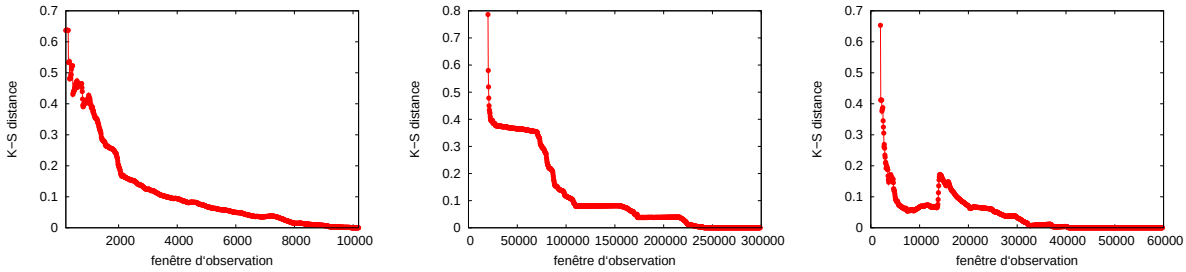
FIG. 4.12 – Distribution cumulative inverse des durées d'inter-contacts pour différentes longueurs de fenêtres d'observation, pour les données PNAS.

Enfin, pour les données PNAS (figure 4.12), on a le même type d'observations que celles obtenues pour les deux premiers jeux de données : des distributions ayant des formes similaires avec des valeurs de durées d'inter-contacts qui dépendent de la fenêtre d'observation.

Dans la figure 4.13, nous présentons l'étude de l'évolution des durées d'inter-contacts avec la K-S distance pour les trois jeux de données. Nous pouvons observer que les observations sont similaires pour les différents jeux de données. Elles commencent par être très élevées puis diminuent lentement et n'ont pas tendance à se stabiliser.

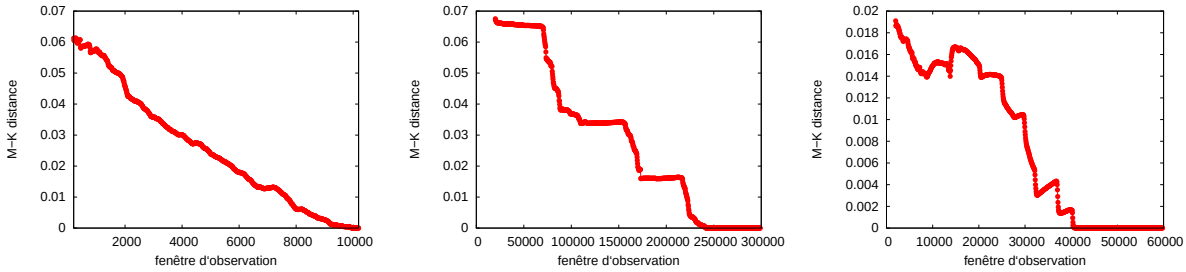
En étudiant l'évolution des distributions avec la M-K distance (figure 4.14), nous observons également le même type de comportement. Ceci montre qu'on ne peut pas caractériser les durées d'inter-contacts dans ces jeux de données.

Nous confirmons ce résultat avec l'étude de la moyenne et de l'écart-type des durées



(a) Données Rollernet. (b) Données Infocom. (c) Données PNAS.

FIG. 4.13 – Étude de l'évolution des durées d'inter-contacts $IC(W_{0,l})$ avec la K-S distance.



(a) Données Rollernet. (b) Données Infocom. (c) Données PNAS.

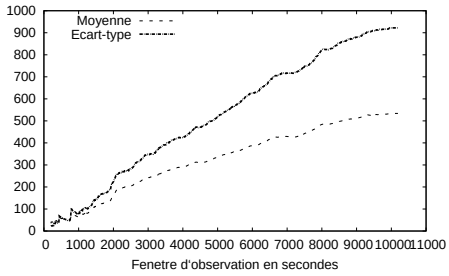
FIG. 4.14 – Étude de l'évolution des durées d'inter-contacts $IC(W_{0,l})$ avec la M-K distance.

d'inter-contacts dans les trois jeux de données (figure 4.15). Comme on s'y attendait, les valeurs de la moyenne et l'écart-type ont tendance à augmenter linéairement avec la durée de la fenêtre d'observation, pour les trois jeux de données.

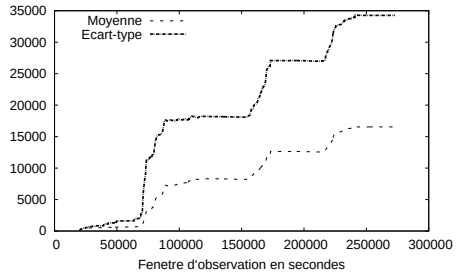
En conclusion, pour les trois jeux de données, on constate que, bien que la forme des distributions soit similaire lorsque l'on augmente la largeur de la fenêtre d'observation, elles ne se stabilisent pas et on ne peut donc pas les caractériser.

4.4 Étude du comportement des nœuds

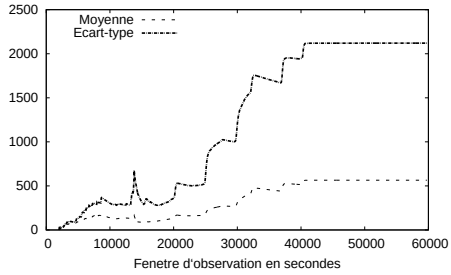
Après avoir étudié quelques propriétés générales pour nos jeux de données, nous tentons maintenant d'étudier la dynamique du réseau, en nous intéressant au comportement des nœuds.



(a) Données Rollernet.



(b) Données Infocom.



(c) Données PNAS.

FIG. 4.15 – Moyenne et écart-type de $IC(W_{0,l})$, en fonction de l.

4.4.1 Comportement global

Pour tenter de caractériser le comportement des nœuds dans le réseau, nous étudions la corrélation entre la moyenne et l'écart-type de la durée des contacts pour chaque nœud du réseau. Cela nous permet de voir si les différents nœuds ont tendance a avoir le même comportement ou des comportements différents.

4.4.1.1 Données Rollernet

La figure 4.16 représente la moyenne et l'écart-type des durées de contacts pour chaque nœud, dans les données Rollernet. Chaque point est associé à une légende indiquant le numéro du nœud correspondant. Par exemple le nœud 22 a une moyenne égale à 18 secondes et un écart-type d'environ 30 secondes. Nous pouvons observer que les différents nœuds sont plus ou moins dispersés et qu'il n'y a pas une forte corrélation entre la moyenne et l'écart-type. Toutefois, la majorité des nœuds ont une moyenne comprise entre 18 et 26 secondes, et un écart-type compris entre 15 et 30 secondes. Nous pouvons distinguer certains nœuds qui s'écartent de ce comportement. Par exemple, le nœud 4 a une moyenne relativement petite et un grand écart-type. Ceci s'explique par le fait que la majorité de ses

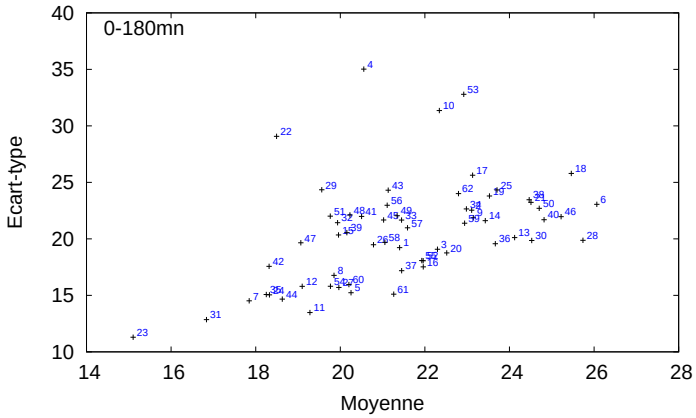


FIG. 4.16 – Moyenne et écart-type des durées de contacts pour chaque nœud, données Rollernet.

contacts sont courts, donc il a une moyenne faible, mais qu'il a également quelques contacts très long (jusqu'à 500 secondes) ce qui fait que son écart-type est grand. Un autre exemple est celui du nœud 23 qui a une petite moyenne et un petit écart-type. Ceci s'explique par le fait que tous ses contacts ont une durée courte (aucun ne dépasse les 70 secondes).

4.4.1.2 Données Infocom

La figure 4.17 représente la moyenne et l'écart-type des durées de contacts pour chaque nœud, dans les données Infocom. Dans ce cas, la corrélation entre la moyenne et l'écart-type est plus forte que pour les données Rollernet. Nous observons aussi un comportement assez similaire à celui de Rollernet. La majorité des nœuds ont une moyenne inférieure à 800 secondes et un écart-type qui ne dépasse pas les 2 500 secondes. Cependant, il existe quelques nœuds qui ont un comportement différent, notamment les nœuds 7, 18, 20 et 33 qui ont des valeurs de moyenne et d'écart-type très grandes.

4.4.1.3 Données PNAS

La figure 4.18 présente la moyenne et l'écart-type des durées de contacts pour chaque nœud, pour les données PNAS.

La figure 4.18 (a) correspond aux résultats obtenus pendant toute la mesure (de 6h à 20h), tandis que la figure 4.18 (b) correspond aux résultats obtenus pendant la période active de la journée, c'est à dire de 8h jusqu'à 16h (ce qui correspond à l'intervalle de temps entre 10 800 et 43 200 secondes).

Dans la figure 4.18 (a), nous pouvons observer que le nœud 781 a une très grande

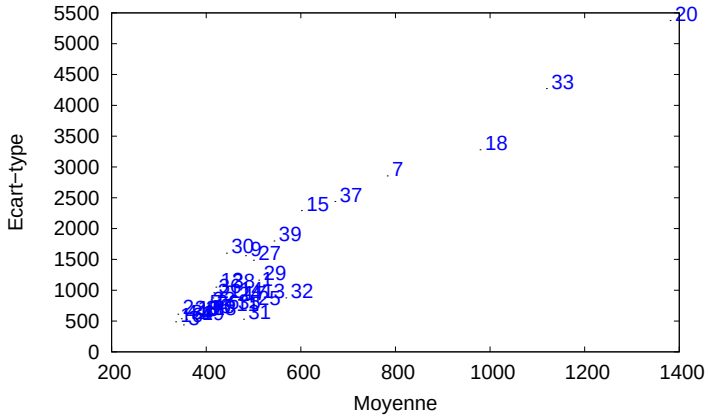
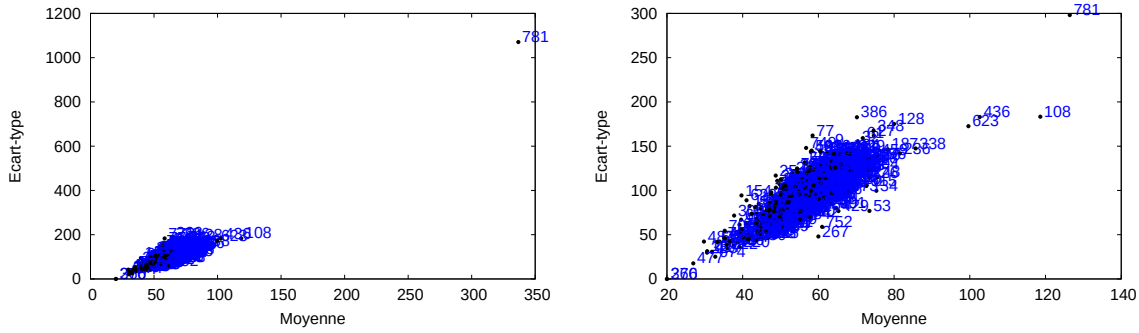


FIG. 4.17 – Moyenne et écart-type des durées de contacts pour chaque nœud, données Infocom.



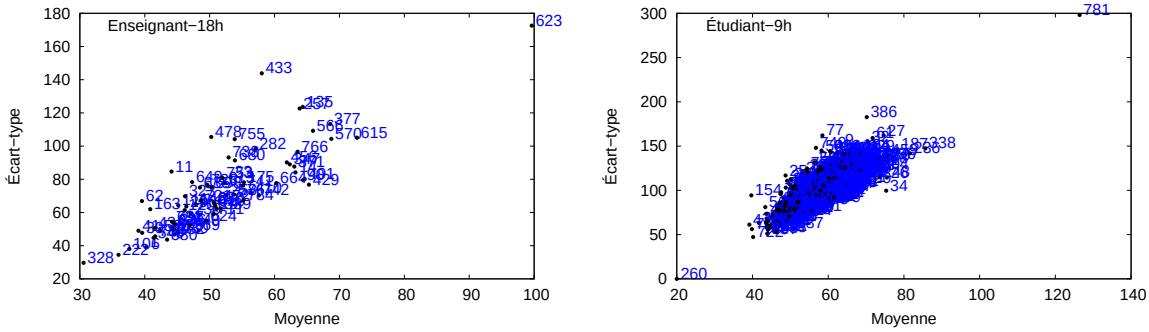
(a) 6 h à 20 h

(b) 8 h à 16 h

FIG. 4.18 – Moyenne et écart-type des durées de contacts pour chaque nœud, données PNAS.

moyenne et aussi un très grand écart-type. Ce comportement change quand on observe la figure 4.18 (b). Le nœud 781 se distingue toujours par rapport aux autres, mais pas avec le même écart. Ce nœud a un comportement spécifique pour deux raisons : premièrement, le premier contact qu'il a est aux environs de 15 h et il continue d'apparaitre jusqu'à 18h ce qui explique le changement de son comportement entre les deux courbes. Deuxièmement, il a quelques contacts très longs qui peuvent aller jusqu'à 2 heures ce qui explique le fait qu'il a un grand écart-type. D'autres nœuds ont un comportement particulier, comme le nœud 108 et le nœud 623. Un autre point qu'on peut observer est qu'il y a une très grande corrélation entre la moyenne et l'écart-type. Les nœuds sont presque alignés sur une ligne droite (plus la moyenne est grande, plus l'écart-type l'est aussi).

Comme nous l'avons déjà mentionné, dans les données PNAS il y a 4 catégories de



(a) Enseignants (b) Étudiants entre 8 h et 16 h

FIG. 4.19 – Moyenne et écart-type des durées de contacts pour chaque nœud, données PNAS.

personnes : enseignants, étudiants, employés et quelques autre personnes. Dans le but de voir si le comportement d'une catégorie donnée suit le comportement général des nœuds, ou a un comportement spécifique, nous avons refait la même courbe en ne prenant en considération que les enseignants et les étudiants. Les résultats obtenus sont présentés dans la figure 4.19. Nous pouvons observer que pour ces deux catégories de personnes, le comportement des nœuds est assez similaire au comportement global. Le comportement obtenu pour les employés (nous ne présentons pas la courbe) est aussi similaire au comportement global (et la dernière catégorie ne contient que 5 nœuds).

4.4.2 Évolution en fonction de la fenêtre d'observation

Afin d'étudier à quel point les observations que nous avons faites sont stables, nous appliquons notre méthodologie à la propriété étudiée. Cela revient à calculer la moyenne et l'écart-type de la durée des contacts de chaque nœud, en augmentant la durée de la fenêtre d'observation. On peut considérer la suite de courbes obtenues comme une vidéo. Les figures que nous présentons dans cette section sont extraites de cette vidéo.

4.4.2.1 Données Rollernet

Dans la figure 4.20, nous présentons le nombre de contacts ainsi que le nombre de contacts distincts (dans le cas ou une paire de nœuds a plusieurs contacts dans la fenêtre d'observation, nous n'en comptons qu'un) observés dans une fenêtre d'observation. Les deux lignes verticales représentent respectivement le début et la fin de cette fenêtre d'observation. Dans cet exemple, le temps de début correspond au début de la mesure et le temps de fin correspond à 10 000 secondes. La ligne de début reste donc toujours dans la même position,

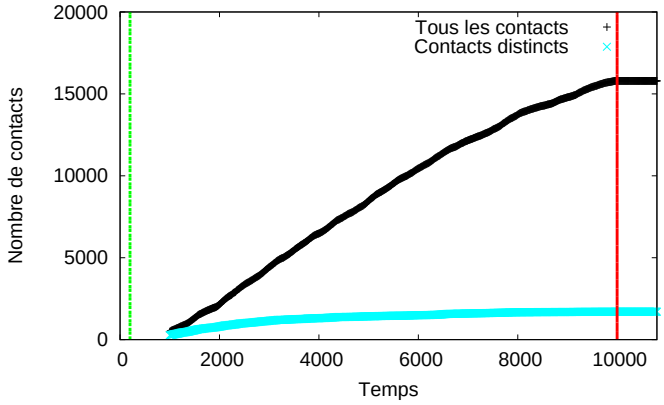
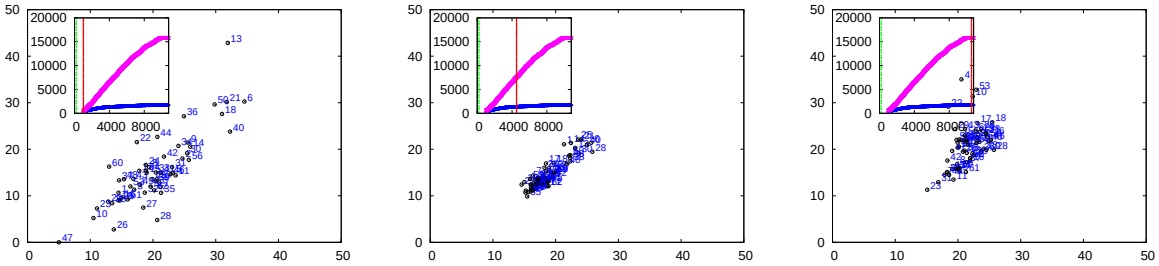


FIG. 4.20 – Nombre de contacts au fil du temps, Rollernet.

tandis que la deuxième se décalera au fil du temps vers la droite. Nous allons utiliser cette courbe comme une indication dans les figures que nous allons présenter dans la suite, pour étudier comment le comportement de la moyenne et de l'écart-type de la durée des contacts des nœuds évolue en fonction du nombre de contacts.



(a) fenêtre de 15 minutes (b) fenêtre de 30 minutes (c) fenêtre de 3 heures

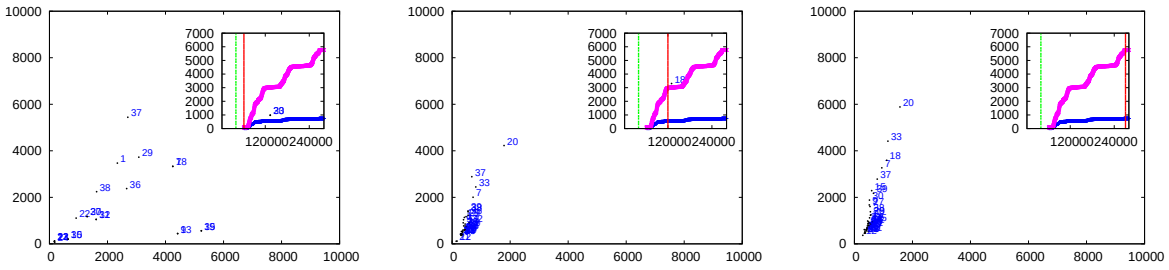
FIG. 4.21 – Évolution des corrélations moyenne/écart-type des durées des contacts, données Rollernet

Nous remarquons que plus on agrandit la fenêtre d'observation, plus les comportements des nœuds sont homogènes. En effet, dans la figure 4.21 (a) qui correspond à une petite fenêtre d'observation d'environ 15 minutes, nous observons un comportement hétérogène avec une moyenne et un écart-type qui varient beaucoup entre les différents nœuds : par exemple le nœud 47 a une moyenne assez petite et un écart-type presque nul, contrairement au nœud 13 qui a une moyenne plus élevée et un grand écart-type. Pour une fenêtre d'observation d'environ 70 minutes (figure 4.21 (b)), le comportement observé est plus homogène. Les nœuds ont des valeurs de moyenne et d'écart-type assez proches et on n'observe plus de nœuds ayant un comportement très différent des autres. Enfin, pour une fenêtre d'observa-

tion de 180 minutes qui correspond à la totalité de la mesure (figure 4.21 (c)), on observe à nouveau un comportement légèrement hétérogène avec quelques nœuds qui se distinguent (d'une façon moins claire que dans la figure 4.21 (a)) par rapport aux autres, notamment les nœuds 4 et 53. Plus généralement, la vidéo [2] permet d'affiner ces constatations et nous observons une forte hétérogénéité pour des fenêtres d'observation très petites. À partir de 25 minutes, le comportement devient de moins en moins hétérogène. Nous observons à nouveau une légère hétérogénéité quand la fenêtre dépasse les 100 minutes.

On constate aussi que les résultats obtenus dépendent du moment de début de notre observation : si on ne prend pas en compte la première partie de la mesure, les valeurs obtenues sont plus homogènes.

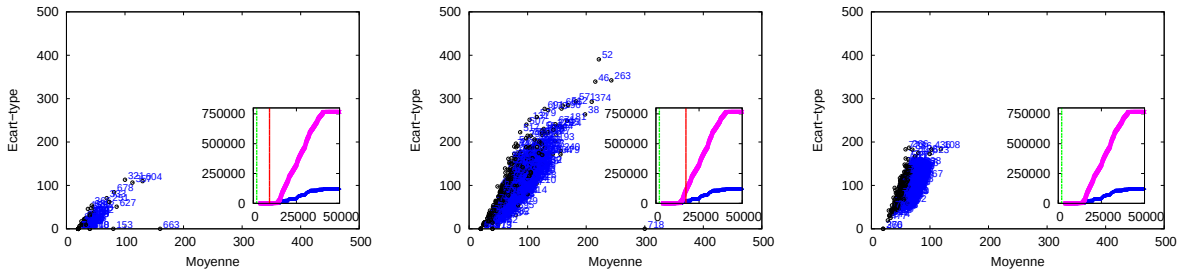
4.4.2.2 Données Infocom



(a) fenêtre de 6 heures (b) fenêtre de 24 heures (c) fenêtre de 72 heures

FIG. 4.22 – Évolution des corrélations moyenne/écart-type des durées des contacts, données Infocom.

La figure 4.22 présente l'évolution de la moyenne et de l'écart-type de la durée des contacts des nœuds au fil du temps, dans les données Infocom. La vidéo correspondant à cette séquence de courbes est disponible sur le lien [3]. Le comportement observé est très similaire à celui de Rollernet. De manière générale, plus la fenêtre d'observation est grande, plus le comportement des nœuds est homogène.
On observe de plus une sensibilité au moment où la fenêtre d'observation commence, mais dans le sens inverse de ce que nous avons observé pour Rollernet. Pour une fenêtre d'observation de 6 heures commençant après le début de la mesure, les nœuds ont un comportement plus homogène que ce qu'on observe dans la figure 4.22 (a).



(a) fenêtre de 6 000 secondes (b) fenêtre de 14 000 secondes (c) fenêtre de 50 000 secondes

FIG. 4.23 – Évolution des corrélations moyenne/écart-type des durées des contacts, PNAS

4.4.2.3 Données PNAS

La figure 4.23 présente l'évolution de la moyenne et de l'écart-type de la durée des contacts des nœuds au fil du temps, dans les données PNAS. La vidéo correspondant à cette séquence de courbes est disponible sur le lien [4]. Pour une fenêtre d'observation de 6 000 secondes (figure 4.23 (a)), nous observons qu'on a très peu de contacts, car cette fenêtre correspond à la période entre 6h et 8h du matin et donc il y a très peu de personnes qui sont arrivées à l'établissement. Quand on augmente la taille de la fenêtre d'observation (figure 4.23 (b) et (c)), nous observons également le même type de comportement que celui observé pour Rollernet et Infocom : plus la fenêtre d'observation est grande, plus le comportement des nœuds est homogène.

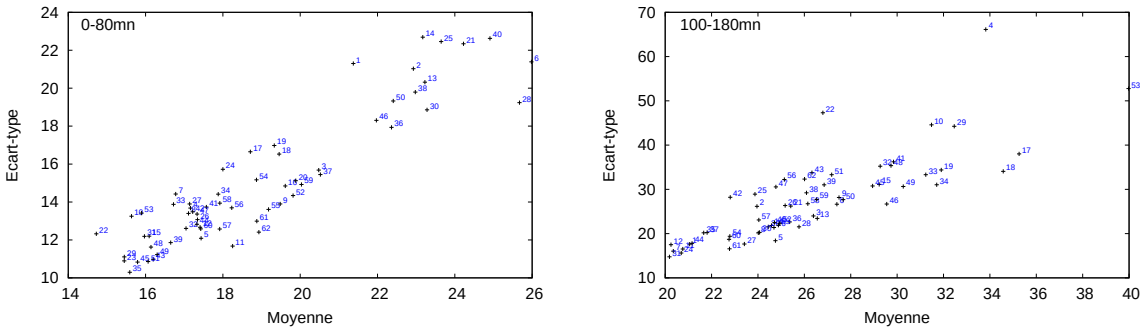
4.5 Variation du comportement au fil du temps

Dans la section précédente, nous avons pu constater qu'il y a une influence de la taille de la fenêtre d'observation sur le comportement observé du réseau. Ici, nous allons étudier la façon dont le comportement observé peut changer lorsque l'on fait varier non pas la taille mais le début de la fenêtre d'observation. Dans ce qui suit, nous présentons les résultats obtenus sur les 3 jeux de données.

4.5.1 Données Rollernet

Nous commençons par étudier la moyenne et l'écart-type de la durée des contacts pour tous les nœuds pendant les périodes situées avant et après la pause qui a eu lieu au milieu

[2]disponible sur le lien : `http://www-rp.lip6.fr/~benamara/Rollernet2.html`.
[3]`http://www-rp.lip6.fr/~benamara/infocom2.html`.
[4]`http://www-rp.lip6.fr/~benamara/PNAS2.html`.



(a) Première partie de la mesure (b) Deuxième partie de la mesure

FIG. 4.24 – Moyenne et écart-type des durées de contacts pour chaque nœud, données Rollernet.

de la mesure. Les résultats obtenus sont présentés dans la figure 4.24. Nous pouvons voir que certains nœuds ont un comportement différent entre la première partie de la mesure (de 0 à 80 minutes) et la seconde partie (de 100 à 180 minutes). Par exemple, le nœud 4 a un petit écart-type (14 secondes) dans la première partie mais un grand écart-type (66 secondes) dans la seconde partie ainsi que dans la mesure complète. Nous remarquons aussi que certains nœuds (1 et 40 par exemple) se distinguent dans la première partie, alors que ce n'est le cas ni dans la deuxième partie ni dans la mesure complète. Ceci montre que le comportement des nœuds peut changer au cours de la mesure. Nous allons étudier ceci plus en détail.

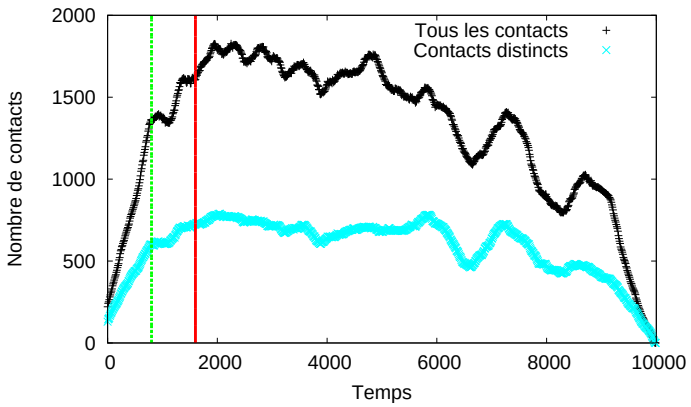
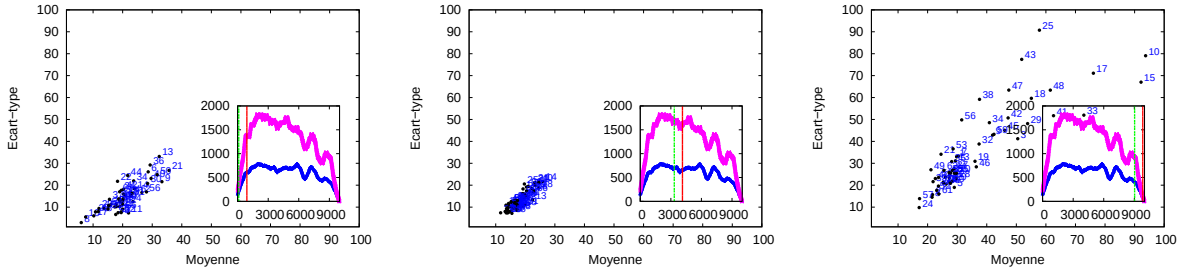


FIG. 4.25 – Nombre de contacts au fil du temps pour une fenêtre d'observation de 800 secondes débutant au temps x, données Rollernet.

Dans la figure 4.25, nous présentons le nombre total de contacts ainsi que le nombre

de contacts distincts présents dans une fenêtre d'observation d'une taille donnée (dans cet exemple, 800 secondes). Les deux lignes verticales représentent respectivement le début et la fin de la fenêtre d'observation. Lorsque l'on fait varier la fenêtre d'observation, ces lignes verticale se déplacent en même temps vers la droite. Nous allons utiliser cette courbe comme une indication dans les figures que nous allons présenter dans la suite, pour étudier les liens entre le comportement de la moyenne et de l'écart-type de la durée des contacts et celui du nombre de contacts.



(a) début de la mesure (b) milieu de la mesure (c) fin de la mesure

FIG. 4.26 – Moyenne et écart-type de la durée des contacts des nœuds à différents moments, pour une fenêtre d'observation de 800 secondes, données Rollernet.

La figure 4.26 représente une séquence de figures extraites de la vidéo[5]. Dans cette figure, nous observons tous les contacts dans une fenêtre d'observation glissante d'une durée de 800 secondes, représentée par les lignes verticales dans l'encart.

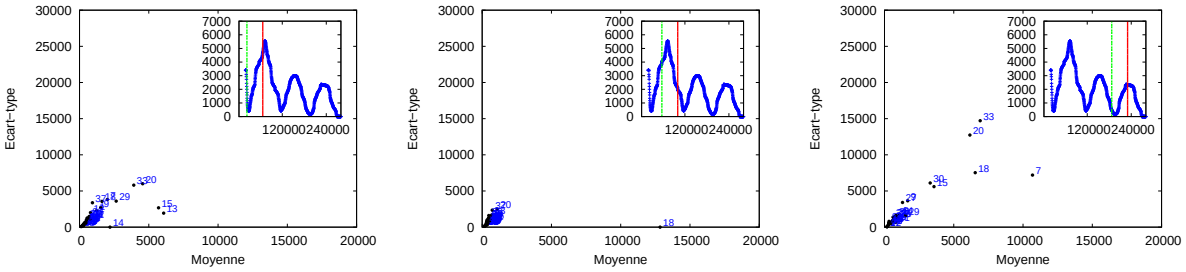
Tout d'abord, nous observons qu'il y a une corrélation entre le nombre de contacts vus dans une fenêtre donnée et les valeurs de la moyenne et l'écart-type des durées de contacts des nœuds. En effet, nous observons au début de la mesure un comportement hétérogène, avec quelques grandes valeurs de la moyenne et l'écart-type (figure 4.26 (a)). Ceci correspond à un moment où il y a peu de contacts. Quand la fenêtre d'observation est dans une période où il y a un nombre de contacts élevé (figure 4.26 (b)), nous observons que les valeurs de la moyenne et de l'écart-type sont plus homogènes. À la fin de la mesure (figure 4.26 (c)), le nombre de contacts décroît et nous observons une variabilité de valeurs avec une présence de certaines valeurs de la moyenne et/ou de l'écart-type très grandes. Nous observons aussi que certains nœuds ont des comportements différents de la majorité. Par exemple, dans la figure 4.26 (c), le nœud 10 a une grande moyenne et un grand écart-type, ce qui signifie que la majorité de ses contacts sont longs, mais qu'il a aussi quelques contacts courts.

[5]disponible sur le lien : `http://www-rp.lip6.fr/~benamara/Rollernet1.html`.

Un autre point qu'on peut observer dans ce jeu de données, est qu'il existe certains nœuds qui se distinguent à certains moments mais pas lorsque l'on prend en compte toute la mesure (figure 4.16). C'est le cas par exemple du nœud 25 qu'on peut distinguer vers la fin de la mesure (figure 4.26 (c)).

La taille de la fenêtre d'observation joue également un rôle : quand on prend une fenêtre d'observation de 1500 secondes, les observations obtenues sont moins hétérogènes que celles obtenues pour une fenêtre de 800 secondes. Pour une fenêtre d'observation de 2500 secondes, elles sont encore plus homogènes, mais nous n'observons pas une très grande différence avec celles obtenues pour 1500 secondes. Ceci montre l'intérêt de trouver un compromis entre une fenêtre trop petite qui donne des observations très hétérogènes mais permet une observation fine de la dynamique, et une fenêtre trop grande dans laquelle on perd la résolution temporelle.

4.5.2 Données Infocom



(a) début de la mesure (b) milieu de la mesure (c) fin de la mesure

FIG. 4.27 – Moyenne et écart-type de la durée des contacts des nœuds à différents moments, pour une fenêtre de 12 heures, données Infocom.

La figure 4.27 représente une séquences de figures extraites de la vidéo[6]. Dans cette figure, nous observons tous les contacts dans une fenêtre d'observation glissante d'une durée de 12 heures.

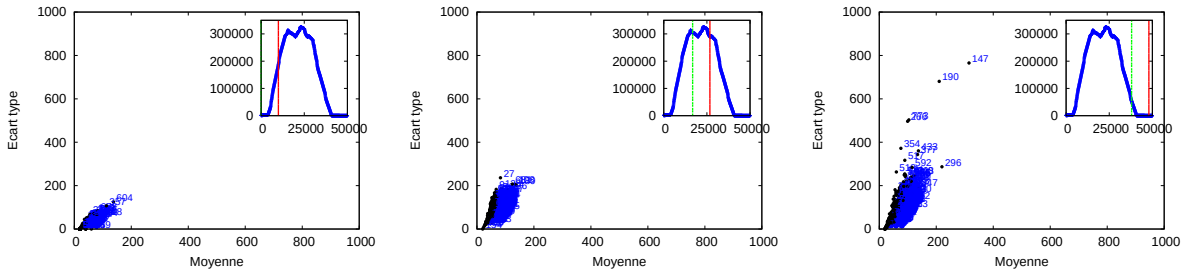
Dans ce jeu de données, nous observons aussi une forte dépendance entre le comportement des nœuds et le nombre de contacts ayant lieu pendant la fenêtre d'observation. Les comportements sont plus homogènes lorsque la fenêtre d'observation contient beaucoup de contacts. On observe également clairement un effet jour/nuit dans ce jeu de données, avec très peu de contacts observés pendant la nuit. De même que pour Rollernet, quand

[6]disponible sur le lien : `http://www-rp.lip6.fr/~benamara/infocom1.html`.

on fait augmenter la durée de la fenêtre d'observation, le comportement des nœuds devient plus homogène. Ceci soulève à nouveau la question du compromis entre stabilité des observations et finesse de la granularité d'observation.

Enfin, certains nœuds ont un comportement particulier. Par exemple, on distingue bien le comportement du nœud 18 (de coordonnés $(13\,000, 0)$ sur la figure 4.27 (b)) : il alterne entre des périodes où la moyenne est stable et l'écart-type varie[7] et des périodes où la moyenne varie et l'écart-type est nul[8], et d'autres périodes où les deux varient au même temps. La même observation qu'on a fait sur Rollernet est valable aussi sur ce jeu de donnée : certains nœuds, comme les nœuds 15 et 21, se distinguent à certains moments de la vidéo mais pas si l'on considère toute la mesure (figure 4.17).

4.5.3 Données PNAS



(a) début de la mesure (b) milieu de la mesure (c) fin de la mesure

FIG. 4.28 – Moyenne et écart-type de la durée des contacts des nœuds à différents moments, pour une fenêtre de 3 heures, données PNAS.

La figure 4.28 représente une séquences de figures extraites de la vidéo[9]. Dans cette figure, nous observons tous les contacts dans une fenêtre d'observation glissante d'une durée de 3 heures. Au début de la mesure (figure 4.28 (a)), nous constatons qu'il y a peu de nœuds, car peu de personnes sont arrivées à l'établissement. Au milieu de la mesure, le nombre de contacts (voir l'encart) est maximal et nous observons que le comportement des

[7] ceci arrive quand il a deux contacts assez longs tels que le temps de début (respectivement le temps de la fin) du premier contact est inférieur au temps de début (respectivement le temps de la fin) du deuxième contact. Quand la fenêtre d'observation que l'on considère commence avant le deuxième contact et finit après le premier, la somme des durées de ces deux contacts est toujours la même, ce qui fait que la moyenne est stable, mais l'écart-type change en fonction de la différence entre la durée des deux contacts.

[8] ceci est dû au fait que pendant ces périodes, le nœud 18 a un seul contact long (pendant la nuit), ce qui fait qu'il a un écart-type nul, mais en fonction de la fenêtre d'observation que l'on considère, la durée observée du contact change, ce qui fait que la moyenne varie.

[9] disponible sur le lien : `http://www-rp.lip6.fr/~benamara/PNAS1.html`.

nœuds est homogène. Cette période correspond probablement au moment du repas. Vers la fin de la mesure (figure 4.28 (c)), on observe très peu de contacts, et des nœuds ont un comportement assez différent (la corrélation entre le nombre de contacts et le comportement des nœuds est donc très visible également dans ce jeu de données).
Nous observons aussi certains nœuds qui se distinguent, notamment à la fin de la mesure, comme le cas du nœud 190 et le nœud 147, qui sont des nœuds qu'on avait pas repérés dans la figure 4.18.

Les différentes observations que nous avons obtenues à partir des trois jeux de données étudiés nous permettent de constater que certains comportements sont communs à tous les jeux de données. En particulier, il y a une forte corrélation entre le nombre de contacts et l'homogénéité du comportement des nœuds. Il y a également le compromis entre une fenêtre d'observation trop petite qui accentue l'hétérogénéité et une fenêtre trop grande qui ne permet pas d'observer la dynamique avec autant de finesse. Enfin, le comportement de certains nœuds se distingue de celui des autres à certains moments de la mesure.

4.6 Comportement spécifique des nœuds

Nous avons identifié dans les sections précédentes des nœuds qui se comportent d'une façon différente de la majorité des autres nœuds. Afin de les étudier plus en détail, nous étudions comment leurs contacts se répartissent dans le temps. Ceci permet de voir à quel point ces nœuds sont spécifiques et de voir si c'est le cas seulement à des moments particuliers ou s'ils ont un comportement spécifique durant toute la mesure.

Les figures que nous présentons ci-dessous représentent tous les contacts d'un nœud donné. Pour chaque contact, nous plaçons un point dont les coordonnées sont son moment d'apparition (axe des x) et sa durée (axe des y). Chaque point est associé à une légende indiquant le numéro du nœud avec qui le nœud étudié est en contact.

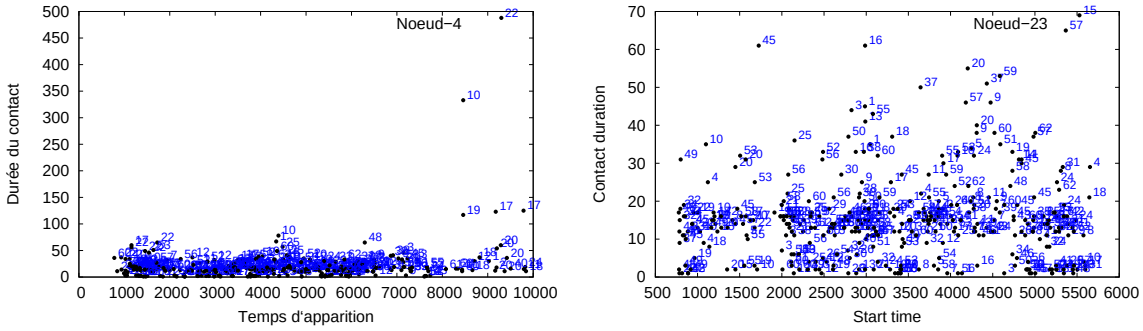
4.6.1 Données Rollernet

Dans les sections précédentes, nous avons observé certains nœuds qui ont un comportement très spécifique par rapport aux autres, notamment le nœud 4 qui a une petite moyenne et un grand écart-type, et le nœud 23 qui a une petite moyenne et un petit écart-type (voir la figure 4.16). Dans la figure 4.29, nous présentons tous les contacts des nœuds 4 et 23.

Pour le nœud 4 (figure 4.29 (a)), la majorité des contacts ont une durée relativement petite (moins de 100 secondes). Il a cependant quelques contacts très longs vers la fin de la mesure, notamment un contact avec le nœud 22 qui dure presque 500 secondes. Cela explique le fait qu'il a un grand écart-type, et aussi que son comportement n'est pas le

même entre la première et la deuxième partie de la mesure (figure 4.24). En effet, ses contacts longs n'apparaissent que vers la fin de la mesure, ce qui fait que dans la première partie il a un petit écart-type, alors que dans la deuxième il a un grand écart-type.

Le comportement du nœud 23 (figure 4.29 (b)) est homogène pendant toute la mesure (aucun de ses contacts ne dépasse une durée de 70 secondes) et on n'observe pas de contacts qui se distinguent significativement des autres, ce qui explique le fait que son écart-type soit petit.



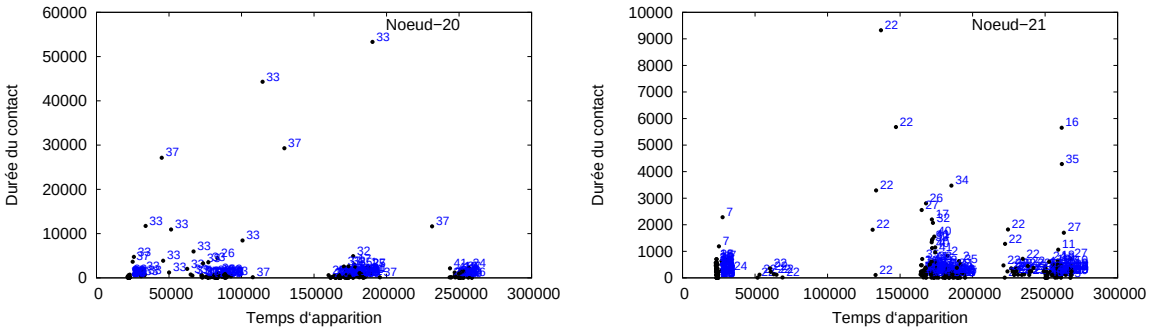
(a) Nœud 4 (b) Nœud 23

FIG. 4.29 – Durées de tous les contacts des nœuds 4 et 23 en fonction de leur temps d'apparition, données Rollernet.

4.6.2 Données Infocom

Dans ce jeu de données, nous avons observé que plusieurs nœuds se distinguent par rapport aux autres. Certains se distinguent par la moyenne et/ou l'écart-type de la durée de tous leurs contacts (figure 4.17), comme le nœud 20, d'autres ont été observés à certains moments spécifiques, comme le nœud 21.

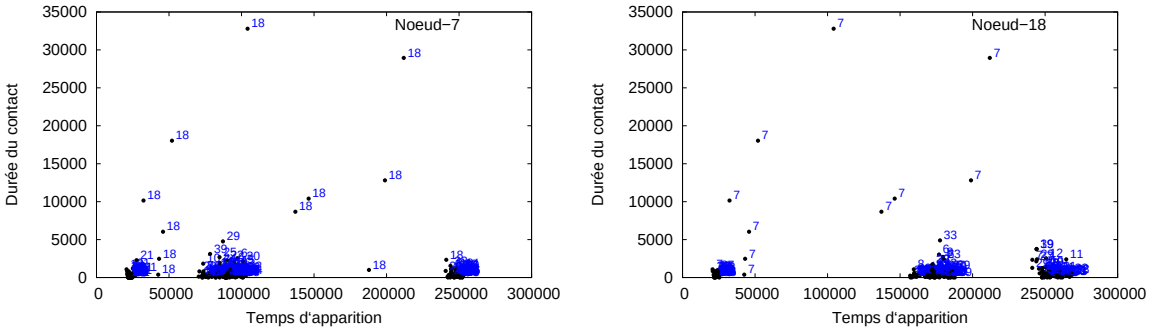
Nous étudions les contacts de ces nœuds au fil du temps. Ceux du nœud 20 sont présentés dans la la figure 4.30 (a). Nous pouvons observer que ce nœud a des contacts relativement courts (ne dépassant pas les 2 heures) regroupés à certains moments de la journée, correspondant généralement soit aux sessions du *workshop* soit aux heures des repas. Par ailleurs il a quelques contacts très longs avec les nœuds 33 et 37 qui peuvent aller jusqu'à 15 heures (55 000 secondes). Ces nœuds correspondent probablement aux personnes avec qui le nœud 20 partage sa chambre ou qui sont dans une chambre voisine, vu que certains moments d'apparition de ces contacts correspondent à des heures tardives de la journée, ou durant la nuit. La majorité des contacts du nœud 21 (figure 4.30 (b)) sont très courts (moins de 30 minutes) et sont regroupés aux mêmes moments que ceux du nœud 20 (pour les raisons déjà



(a) Nœud 20 (b) Nœud 21

FIG. 4.30 – Durées de tous les contacts des nœuds 20 et 21 en fonction de leur temps d'apparition, données Infocom.

citées : sessions du *workshop* et heures des repas). Cependant il a aussi quelques contacts dispersés qui sont assez longs (mais ne dépassent pas 3 heures), notamment avec le nœud 22 avec qui il a eu plusieurs contacts successifs aux environs des 140 000 secondes. Ceci correspond au moment où on a observé (à partir de la vidéo présentée dans la section 4.5.2) que ce nœud avait un comportement spécifique. Il est probable que ces contacts successifs correspondent à un seul contact long qui apparaît comme plusieurs contacts distincts suite à quelques pertes de paquets.



(a) Nœud 7 (b) Nœud 18

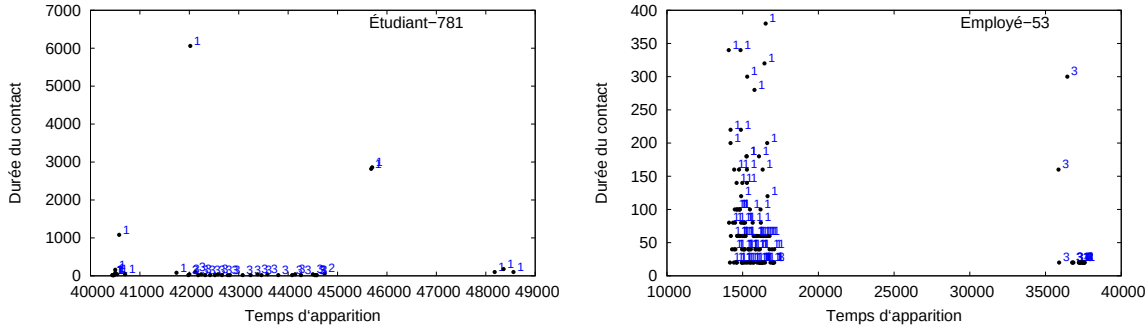
FIG. 4.31 – Durées de tous les contacts des nœuds 7 et 18 en fonction de leur temps d'apparition, données Infocom.

Dans la figure 4.31, nous présentons les mêmes courbes pour les nœuds 7 et 18. Ces deux nœuds ont un comportement très similaire. Ils ont en commun plusieurs contacts longs (jusqu'à 10h), probablement parce qu'ils partagent la même chambre. Ils ont aussi chacun trois périodes durant lesquelles ils ont de nombreux contacts courts, deux aux mêmes

moments (vers 30 000 et 250 000 secondes) et un autre différent (ils ont probablement assisté à des sessions *workshop* différentes).

4.6.3 Données PNAS

Pour ce jeu de données, nous allons présenter le même type de courbes que pour les premiers jeux de données. La seule modification (pour une raison de clarté de courbes) est que la légende indique le rôle du nœud avec qui le nœud étudié est en contact (1 : étudiant, 2 : enseignant, 3 : employé et 4 : autre) et non plus son identifiant.



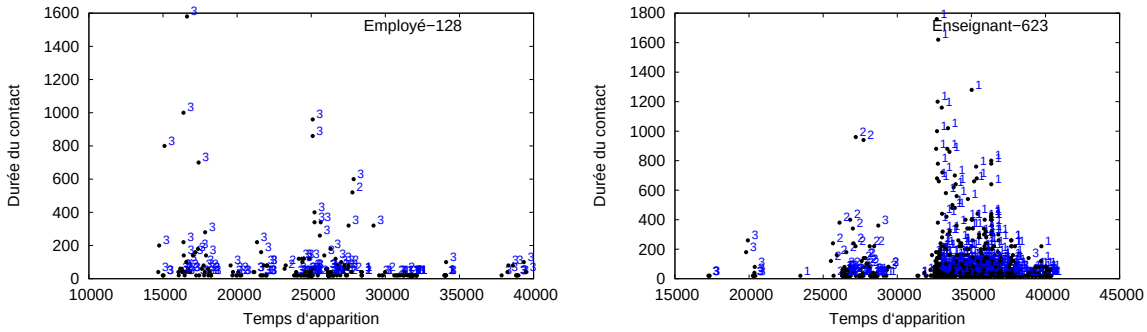
(a) Nœud 781 (b) Nœud 53

FIG. 4.32 – Durées de tous les contacts des nœuds 781 et 53 en fonction de leur temps d'apparition, données PNAS.

Dans la figure 4.32 (a) nous présentons les contacts du nœud 781 qui est un étudiant ayant un comportement très spécifique, vu qu'il n'apparaît qu'à la fin de la mesure, dans une période creuse (après 16 heures) ou la majorité des personnes ont déjà quitté l'établissement (sur la figure, l'axe des x est limité aux moments où le nœud a des contacts). Nous observons que ce nœud a quelques contacts très courts avec quelques étudiants, employés et un enseignant, mais il a aussi 3 contacts très longs (jusqu'à 6 000 secondes) avec des étudiants. C'est la présence de ce type de contacts qui explique que ce nœud a une moyenne de la durée de ses contacts assez grande et un écart-type très grand (comme on l'a observé dans la figure 4.18). Le nœud 53 (figure 4.32 (b)) qui est un employé a un comportement assez spécifique également. Il a principalement deux groupes de contacts. Le premier groupe apparaît au début de la journée avec des étudiants. Les durées de ces contacts varient de 10 jusqu'à 400 secondes. Le deuxième groupe de contacts apparaît vers la fin de la journée avec quelques autres employés.

Le nœud 128 (figure 4.33 (a)) est aussi un employé, qui a un comportement différent de celui du nœud 53. Il a des contacts presque tout au long de la mesure, et la majorité de

ces contacts sont avec d'autre employés. Les durées de ses contacts sont assez courts et le plus long d'entre eux ne dépasse pas les 30 minutes.



(a) Nœud 128 (b) Nœud 623

FIG. 4.33 – Durées de tous les contacts des nœuds 128 et 623 en fonction de leur temps d'apparition, données PNAS.

Enfin, nous présentons les contacts du nœud 623 (figure 4.33 (b)) qui est un enseignant ayant un comportement assez spécifique (on l'a distingué déjà dans la figure 4.18 car il a une moyenne et un écart-type assez grands). Ses contacts sont assez courts et se regroupent principalement à deux moments. Le premier correspond à l'heure du repas, où la majorité de ses contacts sont avec des enseignants et des employés. Le deuxième groupe de contacts qu'il a correspond probablement à un cours vu que la majorité de ses contacts sont avec des étudiants.

4.7 Conclusion

Dans ce chapitre, nous avons présenté une étude de la dynamique des réseaux de contacts entre personnes en nous basant sur l'analyse de trois jeux de données différents. Le but de cette étude n'est pas de faire séparément l'analyse de chaque jeu de donnée ou d'essayer de comprendre comment se comporte chaque réseau indépendamment des autres. Au contraire, le fait que ces réseaux proviennent de contextes différents fait que les étudier ensemble permet de voir quelles analyses permettent d'isoler uniquement des caractéristiques d'un réseau particulier, et lesquelles ont un caractère plus général.

Nous avons introduit différentes méthodes et techniques pour caractériser le comportement global d'un système donné, ainsi que les éventuels comportements spécifiques qu'on peut observer. Ces différentes techniques s'avèrent intéressantes pour comprendre de tels systèmes et pouvoir caractériser leur dynamique.

Nous avons présenté une étude qui articule le point de vue global du réseau et le point de vue de ses nœuds individuellement. En étudiant le comportement global du réseau, nous avons observé qu'il n'est pas stable mais varie au fil du temps. Ceci est principalement capturé par le nombre de contacts au fil du temps, et avait déjà été constaté [TLB+09].

D'autre part, cette étude a fait ressortir certains nœuds qui ont un comportement significativement différent de celui des autres nœuds, soit pendant toute la mesure, soit pendant une période particulière. Une analyse plus détaillée de ces nœuds permet de caractériser leur comportement plus finement.

Nous avons également mis en évidence l'importance que la procédure de mesure a sur les observations, en identifiant clairement des cas où un seul contact apparaît comme plusieurs contacts car des paquets n'ont pas été reçus. Ceci montre l'importance de la question de la métrologie et d'études de cette question telle que celle que nous avons présenté dans le chapitre 3.

Chapitre 5

Conclusion et perspectives

Nous avons abordé dans cette thèse la problématique de la caractérisation de la dynamique des graphes de terrain tout en prenant en compte le biais lié à la mesure, en nous appuyant sur des cas concrets de graphes dynamiques. Nos contributions se sont orientées dans deux directions.

Nous nous somme tout d'abord intéressés à l'étude du biais dans l'observation de la dynamique induit par le fait que la période d'observation est finie. Nous avons proposé une nouvelle méthodologie qui permet de déterminer si la longueur de la période d'observation est suffisante pour une caractérisation rigoureuse d'une propriété donnée. Cette méthodologie est générique et peut être appliquée à n'importe quelle propriété caractérisant un graphe de terrain dynamique. Pour démontrer la pertinence de notre méthodologie, nous l'avons appliquée à l'étude de différentes propriétés dans un système P2P, en utilisant deux jeux de données différents qui nous ont permis de faire des observations instructives.

Nous avons montré que le fait d'observer un système pour une période de temps qui est trop courte ne permet pas de caractériser avec précision ses propriétés. Nous avons également montré que dans un même système, il est possible de caractériser certaines propriétés, mais pas d'autres. Ceci montre qu'il n'y a pas d'échelle de temps absolue qui soit pertinente pour étudier un système dans son ensemble, mais que chaque propriété doit être étudiée indépendamment.

Notre deuxième contribution consiste en une approche pour étudier des graphes dynamiques. Nous avons cherché à la fois à caractériser la dynamique globale du système, et à identifier les éventuels nœuds ayant un comportement particulier. Nous avons étudié plusieurs jeux de données issus de réseaux de contacts entre personnes et nous avons montré que chaque jeu de données a ses particularités. Nous avons également constaté que certaines

caractéristiques sont partagées par tous les jeux de données. En particulier, la dynamique globale du réseau change en fonction de la période d'observation et le comportement de certains nœuds diffère du comportement global du système. Nous avons également mis en évidence l'importance que la procédure de mesure a sur les observations, en identifiant clairement des cas où un seul contact apparaît comme plusieurs contacts distincts à cause d'une perte de paquets. Ceci montre l'importance de la question de la métrologie abordée dans notre première contribution, et la complémentarité entre les deux approches.

Nos travaux ouvrent plusieurs perspectives. Tout d'abord, il semble important de compléter avec des simulations nos travaux concernant l'impact de la durée de la mesure sur les propriétés observées. Ceci permettrait de tester de manière systématique l'impact de la fenêtre de mesure sur plusieurs propriétés ayant des distributions différentes, et d'obtenir un jeu de test empirique fiable, auquel il serait possible de se ramener pour interpréter les observations faites en pratique. À plus long terme, ceci ouvre la voie à une modélisation de ce problème, qui permettrait d'obtenir des résultats formels.

Le fait que la période de mesure soit finie n'est pas la seule cause de biais pour les propriétés dynamiques. Plusieurs autres causes de biais ont été identifiées, comme par exemple le fait qu'étudier la dynamique en capturant des échantillons périodiquement peut empêcher d'observer certains événements courts [WYL07].

Dans le cas de réseaux de contacts capturés par des capteurs, les auteurs de [FCF+11] ont étudié le biais introduit par le système de mesure causé par la perte de certains paquets et proposent une méthode de reconstruction qui permet de retrouver les propriétés originales du système. Nous avons explicitement identifié des cas où certains contacts longs apparaissent comme plusieurs contacts courts suite à des pertes de paquets. Il semble donc important de coupler notre méthodologie avec d'autres méthodologies visant à supprimer d'autres types de biais.

Enfin, dans de nombreux systèmes, en particulier dans le cas de l'internet, il est connu que la procédure de mesure peut introduire un biais structurel même si le système n'évolue pas avec le temps. Quelques méthodes ont été proposées pour remédier à ce biais ou à le détecter, voir par exemple [LM08,LBCX03,SRD+09]. Nous estimons donc qu'il serait intéressant de combiner des méthodologies remédiant au biais dynamique avec des méthodologies remédiant au biais structurel, afin de capturer les propriétés des systèmes tels que l'internet, ainsi que leurs dynamique, de manière fiable.

L'étude que nous avons faite sur les réseaux de contacts est un premier pas vers la mise au point d'un ensemble d'outils permettant de décrire précisément un réseau dynamique. Une suite naturelle de ces travaux consiste donc à étudier d'autres aspects de ces réseaux. On peut en particulier songer au fait que les nœuds sont probablement groupés en commu-

nautés qui évoluent au fil du temps. Détecter ces communautés est une question importante à la fois pour la compréhension de la dynamique du réseau et pour les applications, comme la conception de protocoles de communication tirant parti de cette structure par exemple.

Étudier d'autres types de graphes va permettre de tester la généralité des notions que nous avons introduites. En particulier, certaines des propriétés que nous avons étudiées vont être pertinentes pour tous les cas de réseaux dynamiques. Par exemple, l'étude de la durée de vie des liens d'un nœud est probablement une notion pertinente pour la plupart des cas. Cependant, toutes les techniques que nous avons utilisées ne sont pas directement transposables au cas général, car nous nous sommes appuyés sur le fait que l'ensemble des nœuds est fixe. Elles ne permettent donc pas de prendre en compte les apparitions ou les disparitions des nœuds.

Nous avons distingué dans les réseaux que nous avons étudiés certains nœuds ayant un comportement spécifique. Dans ce contexte, une perspective intéressante serait d'étudier à quel point ce type de nœuds joue un rôle sur le comportement global du système. En particulier, on peut supposer que des nœuds vont pouvoir être plus (ou moins) efficaces que la moyenne dans une diffusion de messages. Les méthodes que nous avons introduites seraient donc dans ce cas un moyen simple d'identifier des nœuds importants pour les applications.

Dans les réseaux de contacts, chaque contact est caractérisé par la paire de nœuds qui sont en contact, sa durée et le moment où il débute. Tous ces paramètres ont un impact sur la dynamique du réseau. Pour comprendre cet impact, une direction possible est d'étudier des modèles qui reproduisent certains, mais pas tous, de ces paramètres. On peut imaginer plusieurs variantes pour ces modèles : par exemple, conserver pour chaque nœud son nombre de contacts et ses voisins, mais effectuer une permutation aléatoire sur la durée des contacts. Une autre variante possible serait de conserver pour chaque nœud son nombre de contacts et leurs durées, mais effectuer une permutation aléatoire sur le moment de début de ses contacts. La comparaison des résultats obtenus en utilisant les permutations aléatoires avec ceux obtenus en utilisant les mesures réelles va permettre d'avoir une meilleure intuition sur les différentes causes des comportements observés. À plus long terme, il y a un fort besoin de proposer des modèles pertinents pour les graphes dynamiques. Ceci serait un premier pas pour la proposition de modèles réalistes reproduisant des propriétés dynamiques observées.

Bibliographie

[AB02] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics 74*, 47, 2002.

[ADG07] Juan A. Almendral and Albert Díaz-Guilera. Dynamical and spectral properties of complex networks. *New Journal of Physics*, 9(6) :187+, June 2007.

[AG10] Thomas Aynaud and Jean-Loup Guillaume. Long range community detection. In *LAWDN - Latin-American Workshop on Dynamic Networks*, page 4 p., Buenos Aires, Argentine, 2010. INTECIN - Facultad de Ingeniería (U.B.A.) - I.T.B.A.

[ALM09] Frédéric Aidouni, Matthieu Latapy, and Clémence Magnien. Ten weeks in the life of an *eDonkey* server. In *Proceedings of HotP2P'09*, 2009.

[AML11] Oussama Allali, Clémence Magnien, and Matthieu Latapy. Link prediction in bipartite graphs using internal links and weighted projection. In *Proceedings of the third International Workshop on Network Science for Communication Networks (Netscicom 2011), In Conjuction with IEEE Infocom.*, 2011.

[BA99] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science (New York, N.Y.)*, 286(5439) :509–512, October 1999.

[BGLL08] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of community hierarchies in large networks. *CoRR*, abs/0803.0476, 2008.

[BL96] V. Barnett and T. Lewis. Outliers in statistical data, 3rd edition. *International Journal of Forecasting*, 12(1) :175–176, March 1996.

[BLA+04] Madan M. Babu, Nicholas M. Luscombe, L. Aravind, Mark Gerstein, and Sarah A. Teichmann. Structure and evolution of transcriptional regulatory networks. *Current opinion in structural biology*, 14(3) :283–291, June 2004.

[BM10] Lamia Benamara and Clémence Magnien. Estimating properties in dynamic systems : the case of churn in p2p networks. In *Proceedings of International Workshop of Network Science for Communication Networks (NetSciCom 2010), in conjuction with IEEE Infocom*, 2010.

[BM11] Lamia Benamara and Clémence Magnien. Removing bias due to finite measurement of dynamic networks. *ALGOTEL'11, Cap Estérel*, 2011.

[BS03] Stefan Bornholdt and Heinz Georg Schuster, editors. *Handbook of Graphs and Networks : From the Genome to the Internet*. John Wiley & Sons, Inc., New York, NY, USA, 2003.

[BSV03] Ranjita Bhagwan, Stefan Savage, and Geoffrey M. Voelker. Understanding availability. In *IPTPS*, pages 256–267, 2003.

[BXFJ03] Binh Bui-Xuan, Afonso Ferreira, and Aubin Jarry. Evolving graphs and least cost journeys in dynamic networks. In *Proceedings of Modeling and Optimization in Mobile, Ad-Hoc and Wireless Networks (WiOpt'03)*, pages 141–150. INRIA Press, March 2003.

[CE07] Aaron Clauset and Nathan Eagle. Persistence and periodicity in a dynamic proximity network. *DIMACS Workshop*, 2007.

[CGD04] Mathias Pohl Carsten Görg, Peter Birke and Stephan Diehl. Dynamic graph drawing of sequences of orthogonal and hierarchical graphs. In *Graph Drawing*, pages 228–238, 2004.

[CHC+07] Augustin Chaintreau, Pan Hui, Jon Crowcroft, Christophe Diot, Richard Gass, and James Scott. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing*, 6 :606–620, June 2007.

[CLF07] Vania Conan, Jeremie Leguay, and Timur Friedman. Characterizing pairwise inter-contact patterns in delay tolerant networks. In *Autonomics*, page 19, 2007.

[CLR67] I. M. Chakravarti, R. G. Laha, and J. Roy. *Handbook of Methods of Applied Statistics*, volume I. John Wiley and Sons, USE, 1967.

[CMRM07] Roberta Calegari, Mirco Musolesi, Franco Raimondi, and Cecilia Mascolo. CTG : a connectivity trace generator for testing the performance of opportunistic mobile systems. In *ESEC-FSE '07 : Proceedings of the the 6th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering*, pages 415–424, New York, NY, USA, September 2007. ACM.

[CR09] Jean-Philippe Cointet and Camille Roth. Socio-semantic dynamics in a blog network. In *CSE (4)*, pages 114–121, 2009.

[FCF+11] Adrien Friggeri, Guillaume Chelius, Eric Fleury, Antoine Fraboulet, France Mentré, and Jean-Christophe Lucet. Reconstructing social interactions using an unreliable wireless sensor network. *Computer Communications*, 34 (5), 2011.

[FGRS07] Eric Fleury, Jean-Loup Guillaume, Céline Robardet, and Antoine Scherrer. Analysis of dynamic sensor networks : Power law then what ? In *Proceedings of the Second International Conference on COMmunication System softWAre and MiddlewaRE (COMSWARE 2007), January 7-12, 2007, Bangalore, India.* IEEE, 2007.

[For10] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486 :75–174, 2010.

[FT08] Yaniv Frishman and Ayellet Tal. Online dynamic graph drawing. *IEEE Transactions on Visualization and Computer Graphics*, 14(4) :727–740, 2008.

[GKT09] Tryphon T. Georgiou, Johan Karlsson, and Mir Shahrouz Takyar. Metrics for power spectra : An axiomatic approach. *IEEE Transactions on Signal Processing*, 57(3) :859–867, 2009.

[GT99] Matthias Grossglauser and David N. C. Tse. A framework for robust measurement-based admission control. *IEEE/ACM Trans. Netw.*, 7(3) :293–309, 1999.

[HCS+05] Pan Hui, Augustin Chaintreau, James Scott, Richard Gass, Jon Crowcroft, and Christophe Diot. Pocket switched networks and human mobility in conference environments. In *Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, WDTN '05, pages 244–251, Philadelphia, PA, USA, August 2005. ACM Press.

[HCSZ06] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. Link prediction using supervised learning. In *Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*, 2006.

[HKKS04] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. Tracking evolving communities in large linked networks. *Proc Natl Acad Sci U S A*, 101 Suppl 1 :5249–5253, April 2004.

[HLC05] Zan Huang, Xin Li, and Hsinchun Chen. Link prediction approach to collaborative filtering. In *Proceedings of the Joint Conference on Digital Libraries (JCDL05). ACM*, pages 7–11, 2005.

[HLM10] Assia Hamzaoui, Matthieu Latapy, and Clémence Magnien. Detecting events in the dynamics of ego-centered measurements of the internet topology. In *WiOpt*, pages 505–512, 2010.

[Hua06] Zan Huan. Link prediction based on graph topology : The predictive value of the generalized clustering coefficient. In *Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (LinkKDD2006)*, August 2006.

[JLGLB04] Matthieu Latapy Jean-Loup Guillaume and Stevens Le-Blond. Statistical analysis of a p2p query graph based on degrees and their time-evolution. In *IWDC*, pages 126–137, 2004.

[KKK00] David Kempe, Jon Kleinberg, and Amit Kumar. Connectivity and inference problems for temporal networks. In *STOC'00*, pages 504–513, 2000.

[KNRT05] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. On the bursty evolution of blogspace. *World Wide Web*, 8 :159–178, June 2005.

[KW06] Gueorgi Kossinets and Duncan J. Watts. Empirical analysis of an evolving social network. *Science*, 311(5757) :88–90, January 2006.

[Lat07] Matthieu Latapy. Grands graphes de terrain - mesure et métrologie, analyse, modélisation, algorithmique. *Mémoire d'habilitation à diriger les recherches, UPMC*, 2007.

[LBCX03] A. Lakhina, J. Byers, M. Crovella, and P. Xie. Sampling biases in IP topology measurements. In *IEEE INFOCOM*, 2003.

[LBFM09] Stevens Le-Blond, Fabrice Le Fessant, and Erwan Le Merrer. Finding good partners in availability-aware P2P networks. In *SSS*, pages 472–484, 2009.

[LBGL05] Stevens Le-Blond, Jean-loup Guillaume, and Matthieu Latapy. Clustering in p2p exchanges and consequences on performances. In *IPTPS*, volume 3640, 2005.

[LCSN07] E. A. Leicht, G. Clarkson, K. Shedden, and Newman. Large-scale structure of time evolving citation networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 59(1) :75–83, 2007.

[LM08] Matthieu Latapy and Clémence Magnien. Complex network measurements : Estimating the relevance of observed properties. In *INFOCOM*, pages 1660–1668, 2008.

[LMO08] Matthieu Latapy, Clémence Magnien, and Frédéric Ouédraogo. A radar for the internet. In *Proceedings of ADN'08 : 1st International Workshop on Analysis*

of Dynamic Networks, in conjonction with IEEE ICDM 2008, pages 901–908, 2008.

[Mag10] Clémence Magnien. Intégrer mesure, métrologie et analyse pour l'étude des graphes de terrain dynamiques. *Mémoire d'habilitation à diriger les recherches, UPMC*, 2010.

[MOVL09] Clémence Magnien, Frédéric Ouédraogo, Guillaume Valadon, and Matthieu Latapy. Fast dynamics in internet topology : Observations and first explanations. In *2009 Fourth International Conference on Internet Monitoring and Protection*, pages 137–142. IEEE, 2009.

[NBW06] Mark E. J. Newman, Albert L. Barabási, and Duncan J. Watts, editors. *The Structure and Dynamics of Networks*. Princeton University Press, 2006.

[New03] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2) :167–256, 2003.

[NK03] David L. Nowell and Jon Kleinberg. The link prediction problem for social networks. In *Proceedings of the twelfth international conference on Information and knowledge management*, CIKM '03, pages 556–559, New York, NY, USA, 2003. ACM.

[OHS05] Joshua O'Madadhain, Jon Hutchins, and Padhraic Smyth. Prediction and ranking algorithms for event-based network data. *SIGKDD Explor. Newsl.*, 7(2) :23–30, December 2005.

[OZZ07] Ricardo V. Oliveira, Beichuan Zhang, and Lixia Zhang. Observing the evolution of internet as topology. In *Proceedings of the 2007 conference on Applications, technologies, architectures, and protocols for computer communications*, SIGCOMM '07, pages 313–324, New York, NY, USA, 2007. ACM.

[PBV07] G. Palla, Albert L. Barabási, and T. Vicsek. Quantifying social group evolution. *Nature*, 446 :664–667, April 2007.

[PC11] Andrea Passarella and Marco Conti. Characterising aggregate inter-contact times in heterogeneous opportunistic networks. In *Networking (2)*, pages 301–313, 2011.

[PIL05] Robert J. Prill, Pablo A. Iglesias, and Andre Levchenko. Dynamic properties of network motifs contribute to biological network organization. *PLoS Biol*, 3(11) :e343+, October 2005.

[PPG04] Seung T. Park, David M. Pennock, and Lee C. Giles. Comparing static and dynamic measurements and models of the internet's AS topology. In *INFOCOM*, 2004.

[RLA00] D. Roselli, J. R. Lorch, and T. E. Anderson. A comparison of file system workloads. In *Proc. of USENIX Annual Technical Conference*, 2000.

[SBF+08] Antoine Scherrer, Pierre Borgnat, Eric Fleury, Jean-Loup Guillaume, and Céline Robardet. Description and simulation of dynamic mobility networks. *Computer Networks*, 52(15) :2842–2858, 2008.

[SGG03] Stefan Saroiu, Krishna P. Gummadi, and Steven D. Gribble. Measuring and analyzing the characteristics of napster and gnutella hosts. *Multimedia Systems*, 9 :170–184, 2003.

[SKL+10] Marcel Salathé, Maria Kazandjieva, Jung W. Lee, Philip Levis, Marcus W. Feldman, and James H. Jones. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*, 107(51) :22020–22025, December 2010.

[SP09] Alina Stoica and Christophe Prieur. Structure of neighborhoods in a large social network. In *CSE (4)*, pages 26–33, 2009.

[SR06] Daniel Stutzbach and Reza Rejaie. Understanding churn in peer-to-peer networks. In *Internet Measurement Conference*, pages 189–202, 2006.

[SRD+09] D. Stutzbach, R. Rejaie, N.G. Duffield, S. Sen, and W. Willinger. On unbiased sampling for unstructured peer-to-peer networks. *IEEE/ACM Transactions on Networking*, 17(2), 2009.

[STF06] Jimeng Sun, Dacheng Tao, and Christos Faloutsos. Beyond streams and graphs : dynamic tensor analysis. In *KDD '06 : Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 374–383, New York, NY, USA, 2006. ACM Press.

[Str01] Steven H. Strogatz. Exploring complex networks. *Nature*, 410(6825) :268–276, March 2001.

[SZR07] Daniel Stutzbach, Shanyu Zhao, and Reza Rejaie. Characterizing files in the modern gnutella network. *Multimedia Syst.*, 13(1) :35–50, 2007.

[TAB09] Tomasz Tylenda, Ralitsa Angelova, and Srikanta Bedathur. Towards time-aware link prediction in evolving social networks. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*, SNA-KDD '09, pages 9 :1–9 :10, New York, NY, USA, 2009. ACM.

[TLB+09] P. U. Tournoux, J. Leguay, F. Benbadis, V. Conan, M. Dias de Amorim, and J. Whitbeck. The accordion phenomenon : Analysis, characterization, and impact on DTN routing. In *IEEE INFOCOM 2009 - The 28th Conference on Computer Communications*, pages 1116–1124. IEEE, April 2009.

[TRW09] Mojtaba Torkjazi, Reza Rejaie, and Walter Willinger. Hot today, gone tomorrow : On the migration of myspace users. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN'09)*, 2009.

[WAL04] Walter Willinger, David Alderson, and Lun Li. A pragmatic approach to dealing with high-variability in network measurements. In *Internet Measurement Conference*, pages 88–100, 2004.

[WS98] D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393 :440–442, 1998.

[WSP07] Chao Wang, Venu Satuluri, and Srinivasan Parthasarathy. Local probabilistic models for link prediction. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, pages 322–331, Washington, DC, USA, 2007. IEEE Computer Society.

[WYL07] Xiaoming Wang, Zhongmei Yao, and Dmitri Loguinov. Residual-based measurement of peer and link lifetimes in gnutella networks. In *INFOCOM*, pages 391–399, 2007.