Détection d'anomalies dans les flots de liens : combiner les caractéristiques structurelles et temporelles

Audrey Wilmet

THÈSE DE DOCTORAT DE SORBONNE UNIVERSITÉ

Spécialité Informatique

École doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par

Audrey WILMET

Pour obtenir le grade de

DOCTEUR de SORBONNE UNIVERSITÉ

Détection d'anomalies dans les flots de liens

Combiner les caractéristiques structurelles et temporelles

Soutenue le (22/23) juillet 2019 devant le jury composé de :

Rapporteurs :	Jean-Philippe Cointet	Professeur, SciencesPo Paris
	Bertrand Jouve	Directeur de Recherche, CNRS
Examinateurs :	Éric Fleury	Professeur, INRIA
	Clémence Magnien	Directrice de Recherche, CNRS
Directeurs :	Matthieu Latapy	Directeur de recherche, CNRS
	Robin Lamarche-Perrin	Chargé de recherche, CNRS

Table des matières

Introduction

Enjeux et difficultés

Les capacités à générer et à collecter des données ont fortement augmenté avec l'informatisation de notre société. Cela conduit aujourd'hui à un ensemble de données provenant de sources différentes extrêmement volumineux que l'on appelle le *Big data*. Dans de nombreuses situations, ces données résultent de la mesure des interactions entre plusieurs millions d'entités (appelées *nœuds*) au cours du temps. C'est le cas, par exemple, des appels téléphoniques, des échanges d'e-mails, des transferts d'argent, des contacts entre individus, du trafic IP, des achats en ligne, et bien d'autres encore. La **détection d'anomalies dans ces interactions temporelles** représente un enjeu majeur. Elle permet, d'une part, une meilleure compréhension de leur **comportement normal**, c'est à dire de la façon dont les interactions s'organisent, et d'autre part, de mettre en évidence des **informations cruciales** sur les caractéristiques anormales des systèmes étudiés. Parmi les exemples cités ci-dessus, elle permettrait notamment de détecter des spams, des publicités, des virus, des périodes de sur-sollicitation du réseau, des fraudes bancaires, des attaques contre les services en lignes, ou encore des individus ayant des comportements inattendus.

Une anomalie peut être définie comme étant « *une observation qui s'écarte tellement des autres observations qu'elle suscite des soupçons quant au fait qu'elle ait été générée par un mécanisme différent* » [74]. [1] Cette définition reste très abstraite et ne prend son sens qu'une fois appliquée à un contexte particulier. Pour l'illustrer et en cerner les principales difficultés, nous prenons l'exemple de la détection d'anomalies dans l'ensemble des températures quotidiennes à Juliaca, au Pérou, de décembre 2008 à janvier 2019 [2]. On note cette série temporelle

$$\mathcal{D} = \{\mathrm{Te}_t \,:\, t \in \{1, \cdots, 3682\}\},$$

où t est un jour de la période et Te_t est la température moyenne le jour t.

L'évolution de la température moyenne au cours du temps est montrée en Figure 1. En Figure 2, on trace les distributions des températures mesurées d'une part sur l'ensemble des jours, et d'autre part sur l'ensemble des jours appartenant aux mois de juillet et de novembre. Le comportement normal des observations peut être caractérisé par une distribution gaussienne ayant pour paramètres la moyenne des températures m et leur

1. « *an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism* » [74].

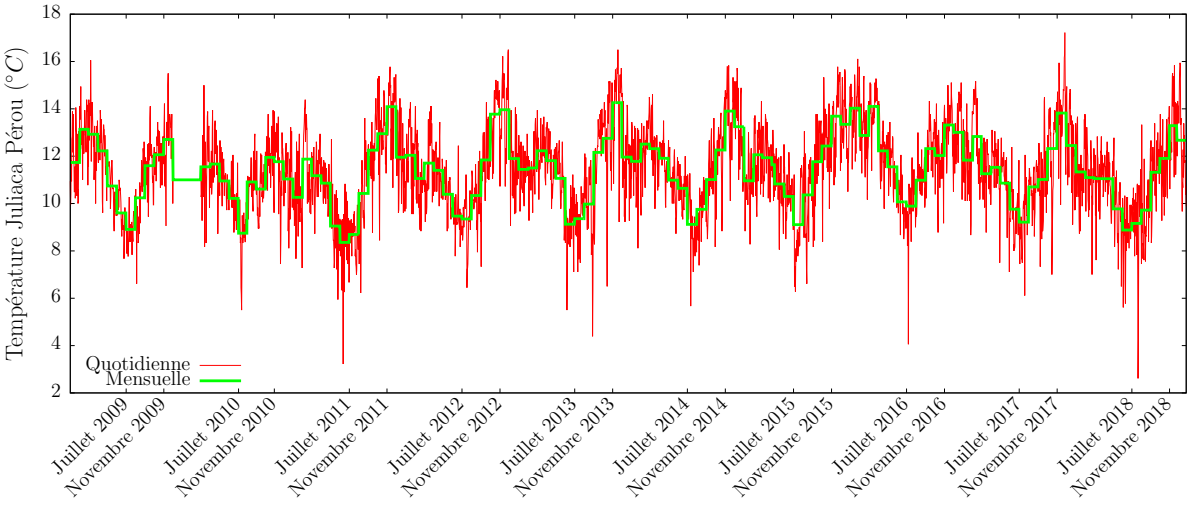


FIGURE 1 – **Évolution des températures quotidiennes et mensuelles à Juliaca au Pérou de décembre 2008 à janvier 2019.**

a) De décembre 2008 à janvier 2019

b) Sur les mois de juillet et novembre

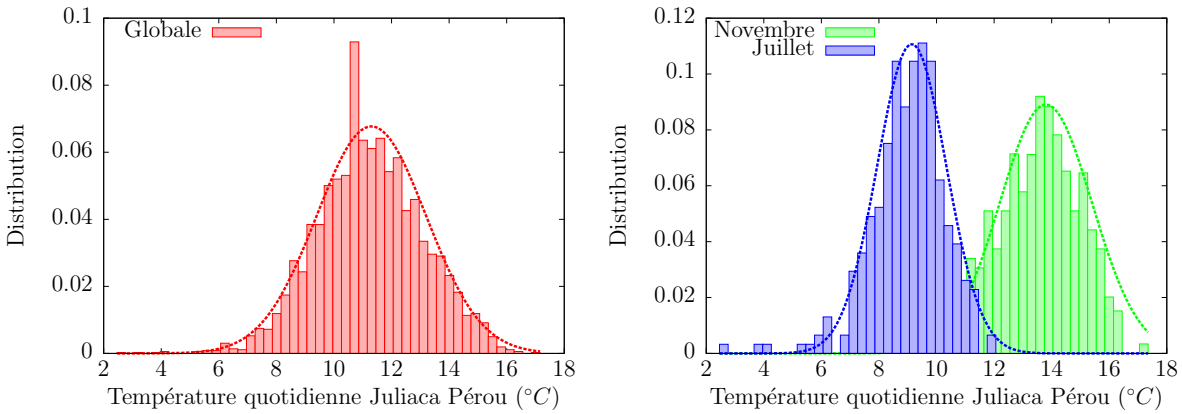


FIGURE 2 – **Distributions des températures quotidiennes à Juliaca au Pérou** – a) L'ensemble des températures peut être caractérisé par une distribution gaussienne de moyenne $m_G = 11.3$ et d'écart-type $\sigma_G = 1.9$. b) Le mois de novembre (*resp.* juillet) est le plus chaud (*resp.* froid) de l'année. Les températures sont caractérisées par une distribution gaussienne centrée en $m_N = 13.8°C$ et d'écart-type $\sigma_N = 1.6$ (*resp.* $m_J = 9.14°C$ et $\sigma_J = 1.2$).

écart-type σ.

Dans un premier temps, on considère que les anomalies sont les observations s'écartant de plus de trois écarts-types de la moyenne. Elles peuvent être classées en deux catégories : les anomalies globales et les anomalies locales ou contextuelles [72]. L'observation d'une température de $2.6°C$ est une anomalie globale, car elle dévie significativement du reste des observations. Une température de $17°C$ en juillet, cependant, est une anomalie locale. En effet, l'anormalité de cette observation dépend du contexte : si on s'était à la place restreint au sous-ensemble de jours appartenant au mois de novembre, cette observation aurait été normale.

D'autre part, on remarque sur la Figure 1 que la température est stable sur une période anormalement longue entre le mois de décembre 2009 et le mois de février 2010. Ici, c'est la répétition de la même observation qui est anormale, et non sa valeur, ou le contexte : on dit dans ce cas que l'ensemble des températures de $11°C$ de décembre 2009 à février 2010 est une anomalie collective [72]. Cette dernière est visible au niveau du pic dépassant du modèle gaussien dans la distribution globale sur la Figure 2 a. Cependant, pour la détecter, il faut considérer un autre critère d'anormalité, par exemple, l'écart entre la probabilité empirique et celle du modèle attendu.

Plus généralement, le comportement normal de la série temporelle peut être caractérisé par d'autres indicateurs qu'un modèle probabiliste attendu. On peut par exemple caractériser son aspect périodique, en calculant les corrélations entre les observations [147], ou transformer la série temporelle à l'aide d'une transformation de Fourier et repérer des anomalies comme étant des observations s'écartant des fréquences typiques [82, 99]. On peut aussi adopter un autre point de vue et considérer les anomalies comme étant des valeurs s'écartant d'une prévision réalisée à l'aide de modèles prédictifs auto-régressifs à moyenne mobile (ARMA) dans le cas stationnaire [119], ou de modèles auto-régressifs intégrés à moyenne mobile, périodiques (SARIMA) [73], ou non-périodiques (ARIMA) [106, 112], dans le cas non-stationnaire.

Ainsi, bien qu'élémentaire, cet exemple met en évidence l'aspect non trivial du problème de la recherche d'anomalies dans un ensemble d'observations. Notamment, les anomalies dépendent de nombreux paramètres comme le **contexte** considéré, ici l'ensemble des jours ou le sous-ensemble des jours appartenant au mois de juillet ou novembre. Elles dépendent également du **comportement normal** inféré des observations, qui peut être basé sur un modèle probabiliste, sur leur tendance, leur dispersion et leur périodicité, ou encore sur un modèle prédictif. Finalement, elles dépendent du choix du **critère de déviation**, qui peut être la déviation significative d'une observation par rapport à la moyenne ou, plus généralement, son écart par rapport au modèle attendu.

De façon générale, **la complexité de la recherche d'anomalies va de pair avec la complexité des données**. Ainsi, en plus des difficultés précédentes, la recherche d'anomalies dans des interactions temporelles s'accompagne des difficultés suivantes :
(1) Les interactions temporelles ne se présentent pas sous la forme d'observations numériques, comme c'est le cas pour les séries temporelles, mais sous la forme d'un ensemble

de triplets (t, u, v) indiquant que les nœuds u et v ont interagit au temps t. Il est donc nécessaire, dans un premier temps, de leur **appliquer une mesure quantitative, appelée *propriété*, de façon à en extraire un ensemble d'observations numériques** pouvant être analysé. Les anomalies détectées dépendent donc directement de cette propriété faisant du choix de cette dernière une difficulté à part entière.
(2) Dans les interactions temporelles, **le contexte peut être choisi de nombreuses façons différentes** : par rapport au temps, à la structure des interactions, ou à une combinaison des deux. Par exemple, si l'on cherche à caractériser l'état d'un nœud au temps t, on peut prendre en compte son comportement passé, ou le comportement des nœuds auxquels il est lié au temps t (appelé son *voisinage*), ou encore, le comportement passé de l'ensemble des nœuds. Ainsi, le choix du contexte constitue également une difficulté à laquelle nous devons faire face.
(3) Finalement, **caractériser le comportement normal de millions de nœuds interagissant au cours du temps** constitue également un challenge. La variation temporelle du comportement de chaque nœud d'une part, et la diversité des comportements des nœuds d'autre part, mènent à des distributions hétérogènes. Or, dans ces distributions, la déviation d'une observation n'implique pas forcément son anormalité, ce qui nous empêche d'utiliser la moyenne et l'écart-type comme estimateurs des distributions.

Problématique

Les interactions temporelles ont été intensément étudiées ces dernières années. Du fait de leur dualité temps/structure, les premières méthodes de détection d'anomalies reposent soit sur le traitement du signal, soit sur la théorie des graphes. Cependant, les propriétés définies dans ces formalismes mènent à une perte d'information induite par la réduction des interactions à leurs aspects temporels ou structurels. La détection d'anomalies dans les interactions temporelles nécessite donc l'utilisation de nouveaux outils adaptés à leur traitement. Le formalisme des **flots de liens**, où un flot de liens est une série de triplets (t, u, v) indiquant qu'une interaction a eu lieu entre u et v à l'instant t, présente une alternative à ces méthodes en permettant d'**étudier conjointement la dynamique et la structure des interactions**. Dans cette thèse, nous explorons l'apport de ce formalisme à la détection d'anomalies. Plus précisément, nous tentons de répondre à la problématique suivante :

« ***Comment identifier des sous-ensembles d'interactions anormales dans un flot de liens ?*** »

Contributions et organisation du manuscript

Les travaux effectués dans le cadre de cette thèse contribuent à la détection d'anomalies dans un flot de liens selon plusieurs aspects. Chacune des contributions est apportée en réponse aux trois difficultés précédemment relevées :
(1) Grâce aux propriétés définies dans le formalisme des flots de liens, nous **caractérisons le comportement des interactions de manière précise**, ce qui nous donne accès à des **anomalies subtiles, qui peuvent être anormales par rapport au temps, vis-à-vis de leur structure, ou les deux simultanément**.

(2) Nous proposons un outil méthodologique permettant de définir la **notion d'anomalie dans un flot de liens**. Ce dernier permet de mettre en évidence différents types d'anomalies, impliquant à la fois différentes entités (des nœuds, des instants, des relations particulières entre une paire de nœuds, etc.) et différents contextes. Cette méthode donne accès à des anomalies pertinentes et variées ce qui permet d'obtenir une compréhension plus complète de la façon dont s'organisent les interactions.
(3) Finalement, nous proposons un outil méthodologique pour détecter des anomalies dans une distribution hétérogène en **tirant profit de l'homogénéité temporelle du comportement global des interactions**.

Dans le **Chapitre 1** nous passons en revue l'état de l'art sur la détection d'anomalies dans des interactions temporelles. Dans un premier temps, nous considérons la détection d'anomalies dans les graphes, afin d'illustrer les difficultés engendrées par la prise en compte de données structurelles. Nous nous intéressons ensuite aux différentes techniques mises en œuvre pour y intégrer la dynamique.
Tout long de cette thèse, nous appliquons nos outils méthodologiques à cinq jeux de données différents : trois traces de trafic IP et deux ensembles d'échanges sur Twitter (retweets) liés à la politique. Dans le **Chapitre 2**, nous décrivons leur intérêt pour la détection d'anomalies, leurs caractéristiques et leur modélisation en flots de liens. Nous introduisons également le degré instantané des nœuds dans un flot de liens. Nous montrons en l'appliquant sur les jeux de données que, bien que pouvant paraître basique à première vue, la recherche d'anomalies à l'aide d'une telle propriété est loin d'être triviale.
Le **Chapitre 3** est consacré à la contribution (2). Nous concevons une série d'étapes à réaliser de façon systématique lors de la recherche d'anomalies dans un flot de liens, d'une part pour déterminer précisément le type d'anomalies recherché, d'autre part pour construire des contextes pertinents menant à des anomalies d'intérêt. Une première partie de ce travail a été publiée [154], une autre est en cours de soumission [155].
Le **Chapitre 4** est consacré à la contribution (3). Nous montrons qu'il est possible d'exploiter l'homogénéité temporelle pour détecter sans perte de précision des anomalies dans une séquence de distributions hétérogènes. Une première partie de ce travail a été publiée [156], une autre est en cours de révisions mineures [157].
Finalement, nous établissons un bilan du travail accompli dans le **Chapitre 5**. Nous discutons également d'une généralisation de nos méthodes ainsi que des perspectives de recherches auxquelles mènent nos travaux, notamment dans le domaine de la prédiction de liens et de la modélisation de flots de liens sans anomalies.

Chapitre 1

État de l'art et Positionnement

Sommaire

Résumé : Dans ce chapitre, nous passons en revue l'état de l'art sur la détection des anomalies dans les interactions temporelles. Nous tentons de donner un aperçu de l'étendue des méthodes existantes tant au niveau de la représentation utilisée pour modéliser les données, des entités détectées, des propriétés employées pour caractériser leurs comportements, que des techniques utilisées pour détecter les anomalies. Nous comparons ces méthodes en présentant leurs avantages et leurs inconvénients, puis positionnons les travaux effectués dans cette thèse selon les critères cités précédemment.

On note $\mathcal{R}$ la *représentation mathématique des données*. À l'inverse des séries pour les données temporelles et des graphes pour les données d'interactions statiques, les interactions temporelles n'admettent pas une représentation unique universellement adoptée. Certains chercheurs les étudient à l'aide d'une séquence temporelle de graphes statiques, d'autres, à l'aide de traitement de signal sur graphes ou de graphes augmentés qui intègrent l'information temporelle, d'autres encore, les étudient en tant que flots de liens. Toutes ces représentations ne sont pas équivalentes, la première par exemple, introduit une perte d'information en agrégeant les interactions dans une même fenêtre de temps. Ainsi, les méthodes de détection d'anomalies dans les interactions temporelles se démarquent

premièrement de par la représentation choisie pour les modéliser.

Elles se différencient également de par l'élément de la représentation $\mathcal{R}$ qui est sous étude. Dans la suite, on désignera ces éléments par le terme *entités*. Par exemple, dans une série temporelle, une entité peut être un instant t, ou un intervalle de temps ; dans un graphe, une entité peut être un sous-ensemble de nœuds, de liens ou un sous-graphe $G' \subseteq G$.

De façon générale, on formule le problème de la recherche d'anomalies dans des données d'interactions (temporelles ou non) comme étant la combinaison de deux étapes [8]. La première consiste à mesurer d propriétés sur un ensemble d'entités spécifique à la représentation $\mathcal{R}$, noté $S(\mathcal{R})$, de façon à générer un ensemble d'observations représenté par un vecteur de nombre réels à d dimensions. La deuxième étape consiste à appliquer sur ce vecteur un détecteur d'anomalie qui, en fonction des valeurs attendues considérées, affecte une décision sur le caractère anormal des observations. Formellement,

$$\begin{aligned} 1^{\text{ère}} \text{ étape} \quad &: \quad S(\mathcal{R}) \longrightarrow \left[\mathbb{R}^d\right]_{obs} \\ 2^{\text{nde}} \text{ étape} \quad &: \quad \{\left[\mathbb{R}^d\right]_{obs}, \left[\mathbb{R}^d\right]_{exp}\} \longrightarrow \{0,1\} \end{aligned}$$

où $\left[\mathbb{R}^d\right]_{obs}$ (*resp.* $\left[\mathbb{R}^d\right]_{exp}$) est l'ensemble des vecteurs de nombres réels de dimension d auquel appartient l'ensemble des valeurs observées (*resp.* attendues). Si le résultat du détecteur est 1, dans ce cas, on dit que l'entité correspondante est anormale vis-à-vis de l'ensemble de propriétés et de valeurs attendues. À la place d'une décision binaire, l'étape de détection peut fournir un score d'anormalité $s_a \in [0,1]$.

En jouant à la fois sur les ensembles d'entités, de propriétés, de valeurs attendues et sur le critère d'anormalité considérés, les anomalies peuvent être définies de nombreuses façon différentes. Dans ce chapitre, nous commençons par illustrer cette difficulté par la présentation de méthodes utilisées dans le cadre des graphes (Section 1.1). Nous montrons ensuite que l'ajout de la dimension temporelle acroît significativement le nombre de définitions possibles. Dans un premiers temps, on passe en revue les méthodes de détection dans une séquence de graphes statiques (Section 1.2). On s'intéresse ensuite aux méthodes qui n'induisent pas de perte d'informations (Section 1.3). Dans chaque section, nous classons les méthodes en fonction du type de propriétés sur lesquelles elles se basent afin de caractériser une anomalie, puis du type d'entités qu'elles considèrent. D'autres classifications sont envisageables, notamment en fonction de l'outil utilisé pour établir un comportement normal comme par exemple, l'utilisation des valeurs extrêmes, de modèles probabilistes, de relations de proximité, ou encore de la théorie de l'information [8]. Finalement, nous positionnons les travaux effectués dans le cadre de cette thèse vis-à-vis de l'état de l'art en Section 1.4.

1.1 Structure

Les données d'interactions regroupent les relations entre une collection d'objets V, appelés nœuds. Ces données peuvent être représentées sous la forme d'un graphe $G(V,E)$ où $E \subseteq V \times V$ est l'ensemble des liens symbolisant la relation entre les paires de nœuds :

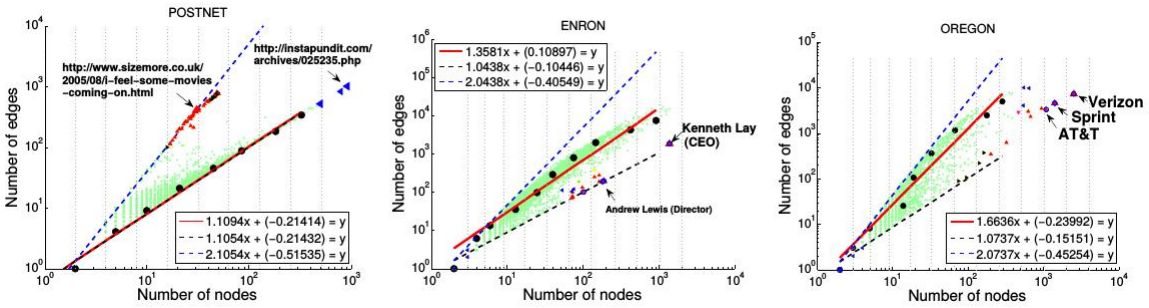


FIGURE 1.1 – **Relation entre le nombre de nœuds n_i et le nombre de liens e_i dans trois graphes différents** – Les points verts sont les observations sur l'ensemble des nœuds. La droite rouge correspond à l'ajustement par la méthode des moindres carrés des valeurs médianes (cercles noirs). Elle correspond au modèle attendu. Les droites pointillées noires et bleues correspondent aux pentes extrêmes $\alpha = 1$ et $\alpha = 2$, associées aux nœuds dont la structure du voisinage est une étoile ou une clique. Les points rouges et bleus correspondent aux nœuds anormaux. Ces images sont extraites de l'article de Akoglu et al. [13].

si $(u, v) \in E$ alors les nœuds u et v sont liés. Dans un graphe, les anomalies peuvent être associées à des nœuds anormaux, des liens anormaux ou des sous-graphes anormaux.

Par exemple, Akoglu et al. [13] trouvent des nœuds anormaux vis-à-vis de la structure de leur voisinage. Pour cela, ils caractérisent l'ego-réseau (*egonet*) de chaque nœud i en mesurant son degré, n_i, le nombre de liens entre ses voisins e_i, le poids total des liens entre ses voisins w_i, et la valeur propre principale de son sous graphe induit, λ_i. Ils organisent ensuite ces quatre propriétés en paires qui respectent les relations attendues suivantes :

$$\begin{cases} e_i \propto n_i^{\alpha} \\ w_i \propto e_i^{\beta} \\ \lambda_i \propto w_i^{\gamma} \end{cases}$$

où $\alpha \in [1, 2]$, $\beta \geq 1$ et $\gamma \in [0.5, 1]$. Finalement, ils détectent des nœuds anormaux parmi ceux dont les observations s'écartent significativement du modèle attendu (voir Figure 1.1). Dans la même idée, Sun et al. [137] et Moonesinghe et al. [109] trouvent des nœuds anormaux vis-à-vis de leur connectivité, calculée à partir d'une marche aléatoire stationnaire.

Noble et al. [113] et Eberle et al. [48] trouvent des sous-structures anormales par la rareté de leurs observations dans le graphe. Pour ce faire, ils mesurent et comparent la longueur de description minimale du graphe avant et après le retrait de cette sous-structure. Dans la continuité de cette méthode, Chakrabarti et al. [32] considèrent un graphe partitionné en communautés et trouvent des liens anormaux vis-à-vis de cette structure globale en comparant, pour chaque communauté, la longueur de compression avant et après le retrait d'un lien.

Parmi six méthodes citées ci-dessus, trois ont pour but de détecter des nœuds anormaux, deux, des sous-graphes anormaux, et une, des liens anormaux. Concernant les trois premières, l'anormalité des nœuds est basée sur leurs connectivités. Or, si la première considère uniquement l'égo-réseau des nœuds, les deux autres considèrent leurs centralités dans le graphe. Ainsi, ces quelques exemples donnent une idée de la quantité de façons différentes dont peuvent être définies les anomalies dans des données d'interactions. Les anomalies dépendent non seulement de l'ensemble d'observations, du modèle attendu, et du critère de déviation, mais également du type d'entités considéré et de la propriété mesurée pour caractériser leur comportement.

1.2 Temps et structure : séquence de graphes

Les données d'interactions temporelles regroupent les interactions entre un ensemble de nœuds au cours du temps sur une période $T = [\alpha, \beta]$,

$$\mathcal{D} = \{(t, u, v) \,:\, t \in T \text{ et } u, v \in V\},$$

tel que $(t, u, v) \in \mathcal{D}$ indique que u a interagi avec v à l'instant t. Dans ces données, les entités sont plus variées et les propriétés, de par l'ajout du temps, sont beaucoup plus nombreuses et diversifiées que dans le cas des interactions statiques, décuplant le nombre de définitions possibles d'une anomalie. La représentation la plus communément adoptée pour analyser les interactions temporelles est une séquence de $k \in \mathbb{N}$ graphes statiques :

$$\mathcal{R} = \{G_i \,:\, i \in \{0, \cdots, k-1\}\}$$

où G_i, appelé instantané i, est le graphe contenant toutes les interactions qui se sont produites entre les instants $t_i = \alpha + i\Delta$ et $t_{i+1} = \alpha + (i+1)\Delta$ avec $\Delta = \frac{\beta - \alpha}{k} \in \mathbb{R}_+^*$. Formellement,

$$G_i = (V_i, E_i), \text{ tel que } \begin{cases} V_i &= \{u_i \,:\, (t, u_i, v_i) \in \mathcal{D} \text{ et } t \in [t_i, t_{i+1}[\} \\ &\quad \cup \{v_i \,:\, (t, u_i, v_i) \in \mathcal{D} \text{ et } t \in [t_i, t_{i+1}[\}, \\ E_i &= \{(u_i, v_i) \,:\, (t, u_i, v_i) \in \mathcal{D}\,,\, t \in [t_i, t_{i+1}[\text{ et } u_i, v_i \in V_i\}, \end{cases}$$

où u_i désigne le nœud u dans l'instantané i. Dans cette représentation, les anomalies peuvent être associées à un instantané anormal, mais aussi à un nœud, un lien, ou un sous-graphe anormal dans un instantané G_i donné. De plus, une anomalie collective peut être associée à un nœud, un lien ou un sous-graphe si les observations qui leur correspondent sont anormales sur l'ensemble des instantanés. Dans ce cas, on dit que l'entité correspondante est globalement anormale.

Le principe de la détection d'anomalies dans une séquence de graphes est d'utiliser des propriétés définies dans la théorie des graphes afin de caractériser l'ensemble d'entités considéré sur chaque instantané, puis de détecter des anomalies dans les séries temporelles résultantes (voir Figure 1.2) [123, 14]. On classe les méthodes en quatre catégories, celles basées sur la mesure d'une distance (Section 1.2.1), la décomposition de matrices représentatives (Section 1.2.2), la compression des interactions (Section 1.2.3) et l'évolution des communautés (Section 1.2.4).

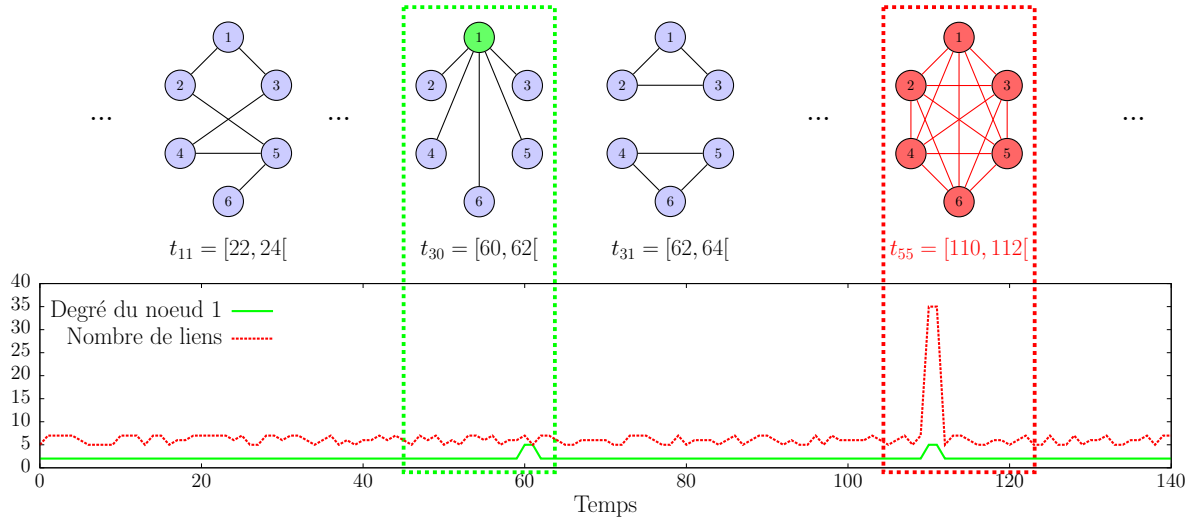


FIGURE 1.2 – **Détection d'anomalies dans une séquence de graphes** – Le nombre de liens est calculé dans chaque instantané. L'instantané G_{55} est anormal car il a un nombre de liens anormalement élevé par rapport aux nombres de liens habituellement observés dans les autres instantanés. De même, le nœud 1 est anormal vis-à-vis de son degré dans les instantanés G_{30} et G_{55}.

1.2.1 Mesure de la distance entre deux instantanés consécutifs

Il est attendu que des graphes similaires aient des propriétés similaires. De ce fait, de nombreuses méthodes sont basées sur la mesure d'une distance. Plus précisément, dans ces travaux, les auteurs caractérisent chaque graphe avec un ensemble de propriétés, puis calculent la distance entre les résultats observés sur deux instantanés consécutifs G_i et G_{i+1} de façon à ce que l'ensemble des distances à chaque pas de temps forme l'ensemble des valeurs observées. Finalement, ils analysent la série temporelle des distances et déduisent des instantanés anormaux par la présence d'une distance anormale. En appliquant ce procédé à un autre ensemble d'entités, certains auteurs détectent également des nœuds, des liens ou des sous-graphes anormaux dans un instantané.

Instantanés anormaux

Pincombe et al. [119] et Papadimitriou et al. [115] introduisent un ensemble de distances calculées à partir de propriétés de graphes élémentaires. Dans chacun des cas, les auteurs utilisent une seule propriété pour caractériser le graphe :
– le diamètre : la distance correspond à la différence des diamètres de G_i et de G_{i+1}.
– le nombre de liens et de nœuds : les distances *MCS Edge* et *MCS Vertex* qui sont fonctions du ratio entre le nombre de liens (*resp.* nœuds) du graphe commun maximal entre G_i et G_{i+1} et le nombre maximal de liens (*resp.* nœuds) dans G_i ou G_{i+1}.
– le poids des liens : les distances *Weight*, *MCS Weight* et entropique. Par exemple, la première est une fonction de la somme sur tous les liens de la différence entre le poids d'un lien (u, v) dans G_i et G_{i+1}, normalisée par son poids maximal dans G_i ou G_{i+1}.
– les valeurs et vecteurs propres : les distances modale et spectrale et la *Vector similarity*.

Par exemple, la dernière correspond à la distance euclidienne entre les vecteurs propres principaux de G_i et G_{i+1}.
Les auteurs définissent également les distances *Graph Edit* et *Median Graph Edit* qui capturent le nombre d'opérations nécessaires pour rendre G_i (ou un graphe médian basé sur un ensemble de graphes antérieurs) isomorphe à G_{i+1}. En plus de ces distances basées sur la structure de chaque graphe, ils utilisent l'index de Jaccard et le coefficient de corrélation de Spearman afin de quantifier la similarité de leurs nœuds ou liens. Masuda et al. [103] appliquent ces mesures à la détection d'états dans une séquence de graphes. Dans la continuité de ces méthodes, Latapy et al. [92] tentent d'identifier des propriétés présentant une distribution homogène avec anomalies. Notamment, ils détectent des instantanés anormaux dans des mesures de trafic IP obtenues à partir d'un seul moniteur, en calculant le nombre de nouveaux nœuds apparaissant d'un instantané à l'autre.

Berlingerio et al. [3] créent la méthode NetSimile, similaire, à la différence que la propriété est mesurée sur les ego-réseaux de chaque nœud au lieu du graphe G_i. Elle correspond souvent au degré du nœud, son coefficient de clustering, ou au degré moyen de ses voisins. Ensuite, cette propriété est agrégée en un vecteur, $\mathbf{q_{G_i}}$, dont les composantes contiennent les m moments de la distribution des observations résultant de la mesure sur chaque nœud du graphe : la médiane, la moyenne, la variance et les coefficients d'asymétrie et de kurtosis. Finalement, deux vecteurs, $\mathbf{q_{G_i}}$ et $\mathbf{q_{G_{i+1}}}$, sont comparés avec la distance de Canberra

$$d_i(G_i, G_{i+1}) = \sum_{j=1}^{m} \frac{|q^j_{G_i} - q^j_{G_{i+1}}|}{|q^j_{G_i}| + |q^j_{G_{i+1}}|} .$$

Saxena et al. [129] extraient également des caractéristiques locales du graphe. La différence est qu'ils incluent en plus sa structure globale en décomposant le graphe en k-cores. Pour chaque G_i, leur algorithme calcule deux distributions. La première, p_{G_i}, regroupe les fractions de nœuds dans chaque core. La seconde, u_G, regroupe pour chaque paire de core, les fractions des liens qui les lient. Finalement, ils calculent les distances de Jensen-Shannon entre p_{G_i} et $p_{G_{i+1}}$ d'une part, et u_{G_i} et $u_{G_{i+1}}$ d'autre part, puis obtiennent la similarité entre G_i et G_{i+1} en prenant la moyenne de ces deux distances :

$$d_i(G_i, G_{i+1}) = \frac{1}{2}\left[JS(p_{G_i}, p_{G_{i+1}}) + JS(u_{G_i}, u_{G_{i+1}})\right] .$$

Macindoe et al. [100] quantifient également les différences entre deux graphes en prenant en compte plusieurs échelles différentes. Pour un rayon r donné, les auteurs considèrent l'ensemble des sous-graphes induits par les r-voisinages de chaque nœuds. Ils attribuent ensuite un score à chaque sous-graphe à partir de trois propriétés :
– le *Leadership* (L), qui est la moyenne des différences entre le degré maximal et le degré de tous les autres nœuds ;
– le *Bonding* (B), qui calcule le ratio entre le nombre de chemins de longueur 3 (triangles) et le nombre de chemins de longueurs 2 ;
– la *Diversity* (D) qui est définie comme étant le nombre de liens qui ne partagent pas de sommets communs.
Finalement, la distance entre deux instantanés G_i et G_{i+1} est calculée avec la métrique de Wasserstein entre la distribution du score LBD sur G_i et celle sur G_{i+1}.

Nœuds anormaux

Gupta et al. [62] trouvent des paires de nœuds anormales parmi celles présentant une variation de longueur de plus court chemin maximale. Ils calculent, pour chaque instantané consécutif et chaque paire de nœuds (u, v), la différence entre la longueur du chemin l dans G_i et celle dans G_{i+1} :

$$d\left((u_i, v_i), (u_{i+1}, v_{i+1})\right) = l(u_i, v_i) - l(u_{i+1}, v_{i+1}) \ .$$

Sous-graphes anormaux

Mongiovi et al. [107] détectent des sous-graphes globalement anormaux en partant du principe qu'un sous-graphe est anormal s'il est constitué de liens anormaux. Dans chaque instantané, ils attribuent un score d'anormalité aux liens qui est fonction de la différence entre le poids du lien dans G_{i+1} et celui dans l'instantané précédent G_i : plus la différence est élevée, plus le score est élevé. Enfin, ils détectent des sous-graphes anormaux en recherchant des régions significativement anormales (SARs) parmi les ensembles de liens ayant un score d'anormalité agrégé élevé.

L'ensemble des méthodes basées sur la mesure d'une distance utilisent des propriétés de graphes relativement basiques. Elles se différencient par l'ensemble d'entités considéré : les ego-réseaux, les instantanés, les paires de nœuds au cours du temps, etc., par la propriété utilisée pour les caractériser et par leur définition de la distance. L'ensemble de ces éléments détermine le type d'anomalies mis en évidence.

1.2.2 Décomposition de matrices représentatives

Une séquence de graphes peut être représentée à l'aide d'un tenseur de dimension 3, comportant deux dimensions encodant les relations entre les paires de nœuds et une dimension encodant le temps. Les méthodes de détection à l'aide de tenseurs reposent sur leur décomposition en valeurs singulières (SVD). En jouant sur la construction de la matrice représentative d'un instantané, ces méthodes sont très flexibles et permettent de trouver des anomalies non triviales.

Instantanés anormaux

Sun et al. [138] introduisent une méthode de décomposition de tenseur, appelée Compact Matrix Decomposition (CMD) qui améliore la méthode SVD, notamment sur le fait qu'elle est moins coûteuse en temps et en mémoire. Ils analysent l'évolution de l'erreur de reconstruction de chaque instantané après leur décomposition : si elle change significativement à un instant donné, alors l'instantané correspondant est identifié comme étant anormal. Cette méthode peut également être appliquée à des blocs du tenseur et permettre de détecter des nœuds ou des sous-graphes anormaux dans des instantanés [83, 131].

D'autre part, Ide et al. [78] extraient les vecteurs propres principaux des matrices d'adjacence associées aux instantanés. Ces vecteurs, appelés vecteurs d'activité, sont ensuite regroupés dans une fenêtre temporelle sous la forme d'une matrice. En décomposant cette

matrice, les auteurs obtiennent un nouveau vecteur, appelé vecteur d'activité passée. Finalement, un instantané est identifié comme anormal lorsque le cosinus de l'angle entre son vecteur d'activité et le vecteur d'activité passée dépasse un certain seuil.

Hirose et al. [75] et Ishibashi et al. [81] utilisent une méthode basée sur ces travaux. Leur contribution repose sur le fait qu'ils construisent des matrices de corrélation à partir de la similarité entre deux nœuds au lieu de matrices d'adjacence.

Nœuds anormaux

Akoglu et al. [12] détectent des instantanés anormaux en mesurant une propriété locale sur l'ensemble des nœuds de façon similaire à Berlingerio et al. [3]. Leur contribution est d'identifier quels sont les nœuds responsables de cette anormalité. Leur méthode s'inspire des travaux de Ide et al. [78] : ils construisent une matrice d'activité passée. Pour cela, ils calculent le coefficient de corrélation de Pearson entre la série temporelle des valeurs associées à un nœud et celle associée à un autre nœud sur une fenêtre temporelle donnée. Ils répètent cette opération pour toutes les paires de nœuds de façon à obtenir une matrice de corrélation associée à la fenêtre temporelle, c'est la matrice d'activité passée. Ensuite, ils extraient son vecteur propre principal dont la composante j représente le degré de similarité du $j^{\text{ème}}$ nœud sur la fenêtre temporelle considérée. Finalement, ils identifient les nœuds responsables parmi ceux ayant un degré de similarité faible.

Par rapport aux méthodes basées sur les distances entre deux instantanés, le principe des méthodes basées sur la décomposition de matrices représentatives est plus complexe. Notamment, de par la construction d'un tenseur caractéristique de l'activité des nœuds sur une fenêtre de temps, elles ont une vision du temps plus globale et intégrative. Les méthodes appartenant à cette catégorie se différencient par la formation du tenseur à la fois au niveau de la propriété choisie pour caractériser les paires de nœuds (corrélation ou nombre d'interaction) et au niveau des paramètres choisis pour caractériser l'activité passée.

1.2.3 Compression des interactions

On distingue deux types de méthodes basées sur la compression. Les premières consistent à compresser la matrice d'adjacence d'un instantané en une chaîne binaire de façon à minimiser l'entropie de la chaîne, notamment en exploitant la répétition et la régularité de certaines structures dans le graphe. Les secondes consistent à transformer les instantanés en un sketch qui est une représentation compressée du graphe construite à partir d'un ensemble de propriétés.

Instantanés anormaux

Sun et al. [136], dans leur méthode GraphScope, considèrent que deux instantanés consécutifs similaires peuvent être groupés dans une même chaîne, ou segment, sans augmenter considérablement le coût de l'encodage. Ainsi, lorsque l'ajout d'un graphe augmente considérablement la longueur de description, ce dernier est identifié comme anormal

et n'est pas ajouté à la chaîne.

Eswaran et al. [50] détectent des instantanés pour lesquels ils observent la brusque apparition ou disparition de sous-graphes denses à l'aide de l'algorithme SPOTLIGHT. Ce dernier agrège chaque instantané G_i en un sketch $v(G_i)$ qui est un vecteur à n composantes. Premièrement, l'algorithme extrait n sous-graphes formés par un ensemble de nœuds sources et un ensemble de nœuds destinations sélectionnés au hasard dans l'ensemble des nœuds. Ensuite, il remplit les n composantes du sketch $v(G_i)$ en calculant la somme des liens de chaque sous-graphe. Pour chaque instantané, les auteurs réalisent M itérations de l'algorithme de sketching. Le score d'anormalité d'un graphe G_i sur l'itération j est calculé en prenant la moyenne des carrés des distances Euclidienne entre le sketch $v_j(G_i)$ et les sketchs des autres instantanés :

$$s_j(G_i) = \frac{1}{k}\sum_{l=0}^{k-1} d(v_j(G_i), v_j(G_l)) \quad \text{où} \quad d(\mathbf{x}, \mathbf{y}) = \sum_{l=1}^{n}(x_l - y_l)^2.$$

Finalement, le score d'anormalité global de G_i est obtenu en prenant la moyenne de $s_j(G_i)$ sur l'ensemble des itérations de l'algorithme de sketching.

Dans les deux cas, la compression peut être appliquée à des sous ensembles du graphe et permettre ainsi de détecter d'autres types d'entités anormales.

1.2.4 Évolution des communautés

L'ensemble des communautés (partition) forme une vue d'ensemble du graphe. Il est par conséquent pertinent de comparer la structure des communautés sur les différents instantanés et d'analyser leurs évolutions afin de détecter des anomalies. Ces dernières peuvent consister en un nœud ou un sous-graphe qui change de communauté, en l'apparition, la dissociation ou la fusion de communautés, ou encore la variation de la qualité de la découpe sur certains instantanés.

Instantanés anormaux

Duan et al. [47] quantifient la qualité de découpe en communautés en calculant le ratio entre la similarité des nœuds intra-cluster et celle des nœuds inter-cluster. Si par exemple l'instantané G_{i+1} a un nombre de liens deux fois supérieur à celui de l'instantané G_i, cela affecte profondément la structure des communautés et la qualité du découpage, ce qui indique la présence d'un évènement anormal.

Nœuds anormaux

Les nœuds appartenant à la même communauté sont supposés avoir des comportements similaires et par conséquent, évoluer de façon similaire. Si, par exemple, dans l'instantané G_{i+1}, un nœud établi un nombre considérable de nouvelles connexions, on s'attend à ce que les nœuds de la même communauté en fassent de même. Si ce n'est pas le cas, ce nœud est considéré anormal. Gupta et al [65] tirent profit de ce principe en calculant la variation de la probabilité d'appartenance à une communauté pour chaque nœud

entre deux instantanés consécutifs. Les nœuds pour lesquels la variation dévie significativement de la variation moyenne des nœuds de la communauté sont considérés anormaux.

Ils généralisent cette méthode à plusieurs instantanés en modélisant l'évolution normale du comportement des communautés au fil du temps et en évaluant la probabilité qu'un nœud dévie de ce comportement normal [64].

Sous-graphes anormaux

Chen et al. [35] proposent de détecter des communautés anormales selon six critères : la communauté s'agrandit ou se rétrécit, elle fusionne avec d'autres communautés ou se scinde en plusieurs sous-communautés, ou enfin, la communauté apparaît ou disparaît. Leur méthode repose sur un ensemble de règles simples définies à partir des communautés précédant ou succédant à une communauté donnée. Les méthodes développées par Aggarwal et al. [9] et Gupta et al. [63] utilisent les mêmes principes.

Araujo et al. [17] s'intéressent à un autre problème et proposent un algorithme permettant de détecter des communautés éphémères qui apparaissent et disparaissent périodiquement. Ils décomposent le tenseur représentatif de la séquence de graphes avec la méthode PARAFAC [27], qui est une généralisation de la méthode PCA aux tenseurs. De cette façon, ils obtiennent une approximation des communautés du graphe. Ensuite, ils compressent ces communautés : les communautés périodiques du fait de la répétition de motifs réguliers sont alors détectées parmi celles ayant le coût MDL le moins élevé.

Les méthodes basées sur l'étude de l'évolution des communautés représentent une grande partie des méthodes développées pour la détection d'anomalies dans des séquences de graphes. Tout comme les méthodes basées sur la décomposition de matrices représentatives, elles sont très flexibles. En effet, l'anormalité d'une entité dépend de la partition des graphes. De ce fait, elle peut être définie de multiples façons selon l'objectif visé.

Les méthodes présentées ci-dessus illustrent bien les innombrables façons de définir des anomalies dans des interactions temporelles en variant les entités, les propriétés, le comportement attendu et la déviation au comportement attendu considérés. Cette difficulté est atténuée, dans la représentation en séquence de graphes, par l'avantage que cette représentation détient, de pouvoir utiliser des méthodes de détection d'anomalies préexistantes, issues des graphes et des séries temporelles.
Cet avantage, cependant, se fait au détriment de la précision. En effet, cette représentation induit une perte d'information. En agrégeant les interactions sur une fenêtre de temps, l'ordre temporel d'arrivée des liens dans un même instantané est perdu, rendant cette représentation dépendante du choix de la taille des fenêtres Δ : trop petite, elle conduit à des instantanés triviaux, trop grande, elle entraîne d'importantes pertes d'information sur la dynamique.
Ainsi, la détection d'anomalies pâtit de la représentation des interactions temporelles en séquence de graphes, la perte d'information limitant la précision temporelle à laquelle les anomalies sont identifiées. Pour trouver des anomalies subtiles à la

fois dynamiquement et structurellement, certains auteurs étudient conjointement plusieurs séquences de graphes en utilisant des tailles de fenêtres temporelles Δ différentes [66]. D'autres encore, introduisent des séquences de graphes causaux dans lesquels deux nœuds sont liés à l'instant t s'ils interagissent dans l'intervalle $[t, t + \delta]$, où $\delta \neq \Delta$ [108, 152].

1.3 Temps et structure sans perte d'information

Une autre solution consiste à introduire de nouvelles représentations, plus fidèles aux données d'origine et qui n'induisent pas de perte d'information. Dans ce but, certains auteurs introduisent des graphes augmentés qui sont des graphes construits de façon à ce que les interactions encodent également la dimension temporelle (Section 1.3.1), d'autres adoptent le point de vue inverse et introduisent du traitement du signal sur graphes (Section 1.3.2), d'autres encore voient les interactions comme un flot de liens (Section 1.3.3).

1.3.1 Graphes augmentés

On distingue deux types de graphes augmentés. Dans le premier cas, ce sont les liens qui intègrent l'information temporelle, dans le second cas, ce sont les interactions.

Liens étiquetés

Casteigts et al. [31] introduisent des graphes dans lesquels les liens sont étiquetés avec leurs instants d'occurrence.

Batagelj et al. [23] et Praprotnik et al. [120] définissent dans cette représentation des équivalents temporels aux notions de degré, coefficient de clustering, centralité de proximité et centralité d'intermédiarité d'une part, et des valeurs propres et vecteurs propres d'autre part. Ils appliquent leur méthode à un graphe d'occurrence de mots, où deux mots sont liés s'ils sont cités dans le même document. Leurs propriétés temporelles leur permettent de détecter l'apparition ainsi que l'émergence de termes anormaux au cours du temps.

Cependant, leurs travaux illustrent le fait qu'étendre les propriétés des graphes statiques à une telle représentation sans perdre d'information est complexe : des notions aussi basiques que le degré requièrent l'introduction de nombreux outils et formalismes.

Liens structurels et liens temporels

Kostakos et al. [90] et Wehmuth et al. [151] choisissent également de représenter les interactions temporelles sous la forme d'un graphe unique. Ils créent un graphe dans lequel un nœud est un couple (t, v), avec $t \in T$ et $v \in V$, et dans lequel le nœud (t_i, u) est lié au nœud (t_j, v) soit s'ils interagissent, c'est-à-dire si $t_i = t_j = t$ et $(t, u, v) \in \mathcal{D}$, soit

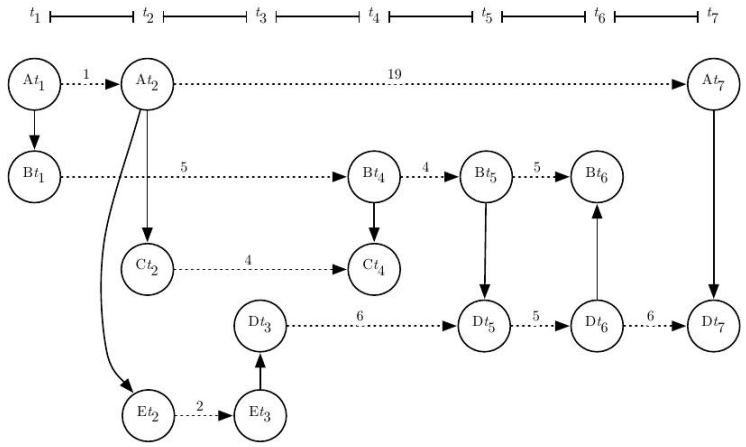


FIGURE 1.3 – **Interactions temporelles dans un graphe augmenté** – Cette image est extraite de l'article de Kostakos et al. [90]. Le lien entre le nœud At_2 et le nœud Ct_2 indique que A et C ont interagit à l'instant t_2. Le poids des liens indique la distance temporelle entre la paire de nœuds correspondante. Par exemple, le poids entre les nœuds At_2 et At_7 est de 19 jours.

s'ils sont contigus dans le temps, c'est-à-dire si $u = v$ et $t_j > t_i$ (voir la Figure 1.3).

Costa et al. [44] et Takaguchi et al. [139] définissent plusieurs mesures de centralité dans ce contexte. Dans les deux cas, ils réussissent à isoler des nœuds centraux à des instants particuliers. Par exemple, Costa et al. [44] s'intéressent à la centralité dans le cas de la diffusion d'un processus dynamique. Ils définissent le temps de couverture temporel d'un nœud u au temps t_i comme étant le nombre de pas de temps nécessaires pour qu'une diffusion commençant de u à t_i atteigne une fraction donnée τ des nœuds du graphe. Ils détectent alors des couples nœuds-temps anormaux parmi ceux présentant un temps de couverture anormalement bas.

Cependant, dans cette représentation les liens ont deux natures différentes : en plus d'augmenter le nombre de liens sous étude, cette caractéristique implique une difficulté supplémentaire pour la détection d'anomalie (nouvelles entités, propriétés ne s'appliquant pas sur l'ensemble des liens).

> Bien qu'elles n'induisent pas de perte d'information, les représentations s'appuyant sur la construction de graphes augmentés, de part leur complexité, sont beaucoup moins utilisées pour la recherche d'anomalies que celles utilisant une séquence de graphes.

1.3.2 Signal sur graphes

Le domaine de recherche concernant l'étude de signaux sur graphes est très récent. Les premiers travaux, apparus en 2012, consistent à étudier des signaux sur l'ensemble des nœuds d'un graphe statique [132, 130]. À chaque nœud u est associé un signal temporel $\{x_t^u, t \in T\}$ (voir la Figure 1.4). L'ensemble des nœuds du graphe forme alors une collec-

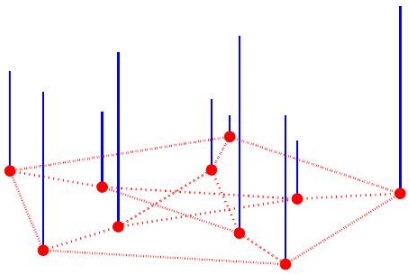


FIGURE 1.4 – **Signal sur graphe** – Cette image est extraite de l'article de Shuman et al. [132]. Elle illustre le fait que chaque nœud du graphe est associé à un signal.

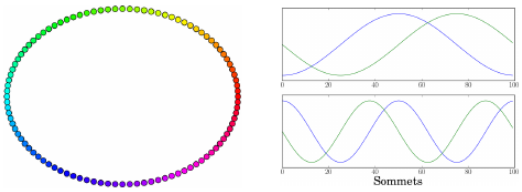


FIGURE 1.5 – **Signaux correspondants à un graphe cyclique** – Cette image est extraite de l'article de Hamon et al. [68]. Les signaux sont obtenus après une transformation de Fourier sur le graphe. Ils correspondent aux quatre premiers modes fréquentiels.

tion de signaux $X \in \mathbb{R}^{N \times T}$ qui est analysée en tenant compte des relations entre les nœuds.

Shimada et al [130] et Hamon et al. [68] montrent par exemple que la représentation d'un graphe en séries temporelles peut être réalisée à l'aide de méthodes déterministes permettant de relier les profils de fréquence des signaux aux structures du graphe. Par exemple, un graphe cyclique peut être vu comme une collection de séries temporelles périodiques : chaque nœud représente un instant, chaque instant est régulièrement espacé, et deux instants successifs correspondent à deux nœuds liés dans le graphe (voir Figure 1.5).

Formellement, soit le graphe $G = (V, E)$ tel que $|V| = N$. Le signal temporel de G est représenté par une matrice $X \in R^{N \times T}$, dans laquelle la composante X_{ut} est la valeur du signal du nœud u à l'instant t. Afin de caractériser les propriétés spectrales de X, Loukas et. al. [98] introduisent la *joint temporal and graph Fourier transform* (JFT) telle que

$$JFT(X, G) = \varphi_G X \varphi_T$$

où $\varphi_G \in R^{N \times N}$ est la matrice propre principale d'une matrice représentative de G (par exemple sa matrice laplacienne, ou d'adjacence, ou de corrélation) et où φ_T est la matrice de transformation de Fourier. Ainsi définie, cette transformation rend compte à la fois des propriétés temporelles des signaux sur les nœuds et des propriétés structurelles des relations entre les nœuds.

Cette représentation peut être appliquée aux séquences de graphes en codant la dynamique des liens à l'intérieur d'un même instantané dans le signal des nœuds, ce qui permet d'étudier les interactions sans perte d'information.

Hamon et. al. [70] s'appuient sur cette représentation pour extraire des sous graphes ayant une structure anormale. À l'aide d'une analyse spectrale dans le domaine fréquentiel, ils détectent des anomalies à des fréquences particulières. Ensuite, ils associent ces fréquences à des structures spécifiques en repassant dans le domaine du graphe (voir la Figure 1.6). Ils appliquent ces travaux dans le cas du graphe temporel issu du système de vélos en libre-service à Lyon, où un nœud est une station et où deux stations sont reliées si au moins un utilisateur a effectué le trajet correspondant [67]. Ils réussissent

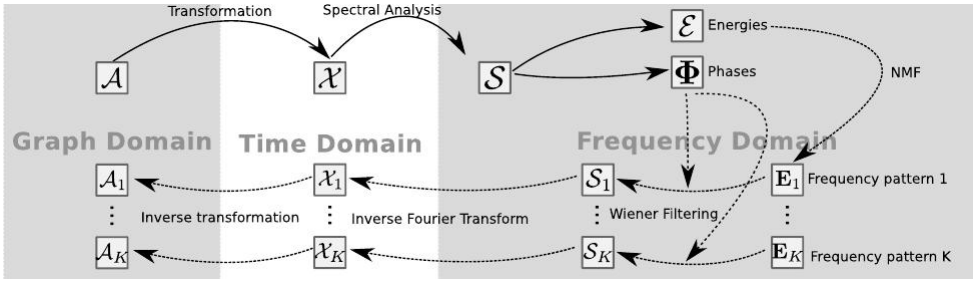


FIGURE 1.6 – **Identification de structures particulière à l'aide du signal sur graphe** – Cette image est extraite de l'article de Hamon et al. [70]. La première étape consiste à extraire un signal pour chaque nœud du graphe. Elle permet de passer du domaine du graphe au domaine temporel. La deuxième étape consiste à passer dans le domaine spectral à l'aide d'une transformation de Fourier. Des fréquences particulières sont identifiées grâce à une analyse dans le domaine spectral. Finalement, les transformations inverses permettent d'identifier les structures correspondant à ces fréquences dans le domaine du graphe.

à identifier des structures particulières indiquant des zones géographiques notablement visitées durant certaines heures.

Ainsi, cette représentation permet d'étudier les interactions temporelles en tirant profit des outils d'analyse développés dans le cadre de la théorie des graphes et du traitement du signal. La nouveauté de cette approche explique le fait qu'elle n'ait pas encore été beaucoup appliquée à la détection d'anomalies. Néanmoins, comme en attestent le nombre de travaux récents sur le sujet, elle éveille un grand intérêt dans la communauté de chercheurs en analyse des signaux [71, 69, 118, 55, 25].

1.3.3 Flots de liens

Finalement, un autre point de vue, récent lui aussi, consiste à traiter les interactions temporelles directement comme un flot de liens. Cette vision a été introduite pour la première fois par Holme et al. [76], puis formalisée depuis peu par Latapy et al. [93].

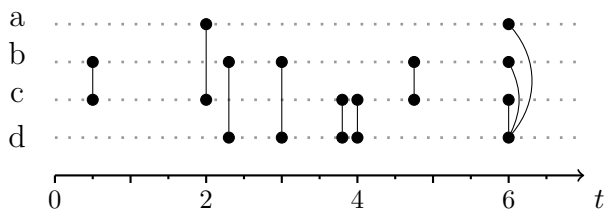


FIGURE 1.7 – **Exemple d'un flot de liens** – $L = (T, V, E)$ tel que $T = [0, 7[$, $V = \{a, b, c, d\}$, $E = \{(0.5, b, c), (2, a, c), (2.3, b, d), (3, b, d), (3.8, c, d), (4, c, d), (4.75, b, c), (6, c, d), (6, b, d), (6, a, d)\}$. Par exemple, b interagit avec c à $t = 0.5$.

Formellement, un flot de liens $L = (T, V, E)$ est défini par un intervalle de temps $T \subset \mathbb{R}$, un ensemble de nœuds V et un ensemble de liens $E \subseteq T \times V \times V$, où $(t, u, v) \in E$, indique que les nœuds u et v ont interagi au temps t (voir Figure 1.7).

Latapy et al. [93] définissent dans ce formalisme les notions de degré, de coefficient de clustering, de centralités, de chemins temporels, etc., et montrent que, dans le cas où tous les liens du flot sont présents tout le temps, ses propriétés sont équivalentes à celles du graphe formé par l'agrégation de tous les liens du flot. Ils montrent aussi que de nombreuses relations, valables dans les graphes, le sont également dans les flots de liens. Par exemple, la densité du flot de liens L est telle que

$$\delta(L) = \frac{2|E|}{|V|(|V|-1)} .$$

Bien que récemment introduite, cette représentation a déjà été mise au profit de la détection d'anomalies.

Nœuds temps anormaux

Par exemple, Yu et al. [161] détectent des couples nœuds-temps anormaux dans un flot de lien. Pour ce faire, ils caractérisent le voisinage de chaque nœud u par la matrice $M_u = A_u^T A_u$, où A_u est la matrice d'adjacence du voisinage de u. Ensuite, ils calculent pour chaque nœud les vecteurs propres principaux des matrices M et étudient leurs évolutions au cours du temps. Finalement, ils trouvent des nœuds-temps anormaux parmi ceux subissant un changement de leur structure locale et dont l'amplitude ou la direction du vecteur est modifiée.

Liens anormaux

Manzoor et al. [102] utilisent une technique similaire dans le sens où ils caractérisent également le voisinage de chaque nœud au cours du temps. Plus précisément, ils stockent le flot de liens dans un sketch construit à partir des caractéristiques des ego-réseaux de chaque nœud. Ensuite, ils comparent le sketch avant et après l'arrivée d'un lien dans le flot : si la différence entre les deux sketchs est significative, ce lien est anormal.

Ranshous et al. [122] recherchent des liens qui présentent des caractéristiques structurelles anormales dans un flot de liens. Ils ont également recours aux sketchs. Pour les construire, ils se basent sur les travaux de Cormode et al. [42] : ils compressent le flot de liens à un instant donné par une approximation de la fréquence des liens et des nœuds à l'aide d'un échantillonnage des liens à cet instant. À partir de ces sketchs, ils attribuent un score de normalité à chaque lien (t, u, v) en fonction du comportement attendu des nœuds u et v au temps t. Ce dernier est la combinaison de trois scores : le score d'échantillonnage, qui est le nombre d'occurrences du lien (u, v) sur le nombre total de liens observés, le score d'attachement préférentiel, qui prend en compte l'hétérogénéité des nœuds, et le score d'homophilie, qui prend en compte la similarité du voisinage de u et v avec une propriété semblable au coefficient de Jaccard. Finalement, les liens anormaux sont ceux qui possèdent un score de normalité en-dessous d'un seuil donné.

Eswaran et al. [49] s'appuient également sur une approximation des liens du flot. Ils attribuent un score à chaque nouveau lien entrant dans le flot en s'appuyant sur un sous-flot L' échantillonné à partir de liens antérieurs. Si le nouveau lien connecte des sous-ensembles peu connectés de L', il est alors considéré comme anormal (voir Figure 1.8).

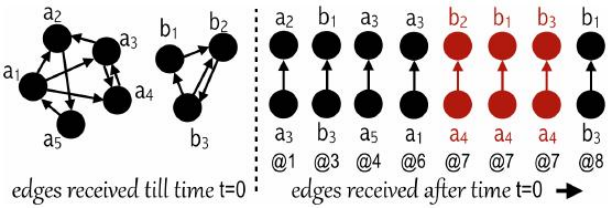


FIGURE 1.8 – **Liens anormaux dans un flot de liens** – Cette image est extraite de l'article de Eswaran et al. [49]. Les liens en rouge sont anormaux car ils connectent les communautés des nœuds a et b, précédemment disjointes.

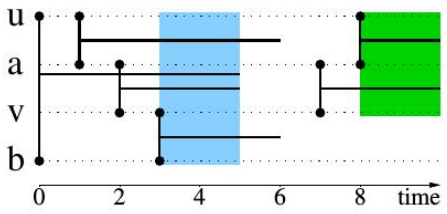


FIGURE 1.9 – **Cliques bipartites dans un flot de liens** – Cette image est extraite de l'article de Viard et al. [145]. En bleu, $C_1 = ([3,5],\{u,v\},\{a,b\})$, en vert $C_2 = ([8,10],\{u,v\},\{a\})$.

Sous-flots anormaux

Finalement, Viard et al. [145] recherchent des cliques biparties anormales en utilisant la définition des cliques dans un flot de liens fournie par Latapy et al. [93] (voir Figure 1.9). Pour éviter d'avoir à énumérer l'ensemble des cliques, ils ont également recours à un échantillonnage. Ils construisent itérativement une clique bipartie en parcourant les nœuds du flot et en ajoutant un nœud à la clique de l'itération i s'il est lié à la totalité des nœuds de la clique de l'itération $i-1$. Ils répètent cette opération de nombreuses fois et obtiennent un ensemble de cliques échantillonnées parmi lesquelles ils détectent des cliques anormales à l'aide de données étiquetées.

Les travaux présentés ci-dessus sont pour la plupart très récent (après 2016). À notre connaissance, ce sont les seules contributions à la détection d'anomalies dans un flot de liens. Dans chacun des cas, les anomalies détectées sont très précises : elles indiquent un lien, un nœud ou un ensemble de nœuds anormal en précisant exactement les instants durant lequel son comportement est anormal.

1.4 Positionnement

De nombreuses études ont été effectuées afin de détecter des anomalies dans des interactions temporelles. Une grande majorité de celles-ci profitent des méthodes développées dans le cadre des graphes statiques et intègrent l'information temporelle en représentant les interactions par une séquence de graphes. Or, si l'agrégation des interactions dans une fenêtre de temps donne un aperçu des structures sous-jacentes, elle ne fournit pas l'ordre d'apparition et de disparition des liens dans un même instantané. Compte tenu de l'importance de cette dynamique, par exemple dans des domaines tels que la propagation épidémique ou les réseaux de communication, de nombreux efforts ont été réalisés pour permettre à ces méthodes de mieux pendre en compte l'aspect temporel.

Dans cette thèse, nous choisissons de représenter les interactions par un flot de liens. La simplicité du formalisme, le fait qu'il considère le temps et la structure de façon équivalente, et qu'il permette de mettre en évidence des anomalies subtiles et précises en fait une représentation parfaitement adaptée à la détection d'anomalies dans des données d'interactions temporelles.

Comme nous l'avons vu dans la Section 1.3.3, l'état de l'art dans cette branche est encore très peu développé. De plus, les méthodes utilisant cette représentation ont recours à des approximations comme l'échantillonnage de liens ou la construction de sketchs. Ainsi, le but de cette thèse est d'apporter une contribution à ce domaine émergeant. Plus précisément, nous proposerons des solutions aux difficultés liées à cette représentation, notamment, comment définir une anomalie pertinente dans ce contexte (Chapitre 3), ou encore, comment détecter des anomalies et identifier précisément quelles entités du flot en sont à l'origine, en particulier lorsque ces entités sont très hétérogènes (Chapitre 4).

Chapitre 2

Notre approche

Sommaire

Résumé : Nous considérons deux types de jeux de données dans lesquels la détection d'anomalies représente un intérêt réel : des traces de trafic IP et des ensembles de retweets liés à la politique. Ces jeux de données se présentent sous la forme de triplets (t, u, v) indiquant que les nœuds u et v ont interagi à l'instant t. Ces interactions sont aisément modélisées par des flots de liens ponctuels. Afin de mieux rendre compte de leur dynamique, on les transforme en flots de liens avec durée. Nous menons ensuite une analyse préliminaire du degré instantané des nœuds dans les flots de liens avec durée, puis présentons une revue de l'état de l'art des approches considérées pour la détection d'anomalies dans ces données. Ceci nous permet de dégager deux problématiques auxquelles nous tenterons de répondre dans les chapitres suivants.

Les outils méthodologiques que nous proposons dans cette thèse s'appliquent aux flots de liens et donc aux interactions temporelles en général. Afin d'évaluer leur pertinence, nous conduisons des expériences sur deux types de jeux de données : du trafic IP, dans lequel une interaction désigne un échange de paquets entre deux machines, et des données provenant de Twitter, dans lesquelles une interaction indique qu'un utilisateur a retweeté le tweet d'un autre utilisateur. Ces jeux de données regroupent plusieurs millions d'interactions et ont des durées variables allant de 15 minutes à un jour pour le trafic IP, jusqu'à un mois pour Twitter.

La recherche d'anomalies a un intérêt crucial dans les deux cas. Dans le trafic IP, les attaques contre les services en ligne, les réseaux et les systèmes d'information, ainsi que les usurpations d'identité ont des coûts annuels estimés à des milliards d'euros. Ils sont la cause de nombreuses faillites et ont également des conséquences sur la fiabilité des services et la confiance des utilisateurs [5]. Dans ce contexte, il semble donc crucial de concevoir des méthodes et des outils permettant de détecter et de lutter contre ces attaques et ces programmes malveillants. Sur Twitter, la diffusion des tweets affecte les idées et les opinions des utilisateurs. Même si ces derniers ne représentent qu'une faible proportion de la population, les sujets d'actualité émergeant sur Twitter sont relayés par les médias traditionnels et peuvent ainsi atteindre un public très large. Or, si de telles tendances résultent souvent de la réaction de tous les utilisateurs à des événements extérieurs, elles peuvent également provenir de l'activité intense d'un petit groupe de personnes et induire en erreur les autres utilisateurs sur l'importance de certains sujets. La recherche d'anomalies semble donc cruciale dans ce cas également : elle permettrait une meilleur compréhension de la façon dont s'organisent les interactions pour mener à de tels évènements.

Dans ce chapitre, nous commençons par décrire en détail la façon dont les interactions sont modélisées à l'aide des flots de liens. Nous formalisons ensuite le degré instantané des nœuds, utilisé dans la suite de cette thèse pour détecter des anomalies. Ensuite, nous décrivons les jeux de données utilisés et analysons leurs comportements vis-à-vis du degré des nœuds au cours du temps. Finalement, nous positionnons l'approche adoptée dans cette thèse vis-à-vis de l'état de l'art sur la détection d'anomalies dans ce type de données.

2.1 Formalisme des flots de liens

Le formalisme des flots de liens est une généralisation de la théorie des graphes. De ce fait, lorsque le flot n'a pas de dynamique, il est équivalent à un graphe classique et ses propriétés sont identiques à celles de ce graphe. Les relations qui existent entre les différentes propriétés du graphe restent également valables pour les propriétés du flot [93].

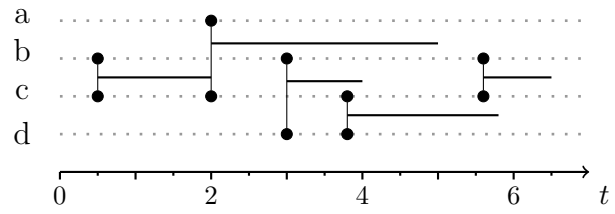
Formellement, un flot de liens L est un triplet (T, V, E) où $T = [\alpha, \omega] \subset \mathbb{R}$ est un intervalle de temps, V un ensemble de nœuds et $E \subseteq T \times V \times V$ un ensemble de liens. Si $(t, u, v) \in E$, alors les nœuds u et v interagissent au temps t. D'autre part, si $(t, u, v) \in E$ pour tout $t \in [a, b]$, alors les nœuds u et v interagissent du temps a au temps b et on note $[a, b] \times (u, v) \subseteq E$. On considère dans cette thèse que les liens du flot ne sont pas orientés.

Le flot de liens L induit un graphe classique $G(L) = (V, E(L))$ défini par $E(L) = \{(u, v) : (t, u, v) \in E\}$. De la même manière, on peut considérer la série temporelle décrivant l'ensemble du flot telle que $f(t) = |\{(u, v) : (t, u, v) \in E\}|$.

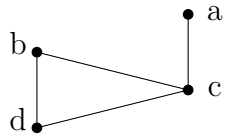
2.1.1 Exemple de flot de liens avec durée

La Figure 2.1 montre un exemple de flot de liens ainsi que son graphe statique induit et la série temporelle associée. Le flot de liens $L = (T, V, E)$ représenté est tel que :
– $T = [0, 7[$,
– $V = \{a, b, c, d\}$,
– $E = ([0.5, 2[\cup [5.5, 6.5[) \times (b, d) \cup [3, 4[\times (c, d) \cup [3.8, 5.8[\times (b, c) \cup [2, 5[\times (a, c)$.

a) Flot de liens



b) Graphe statique



c) Série temporelle

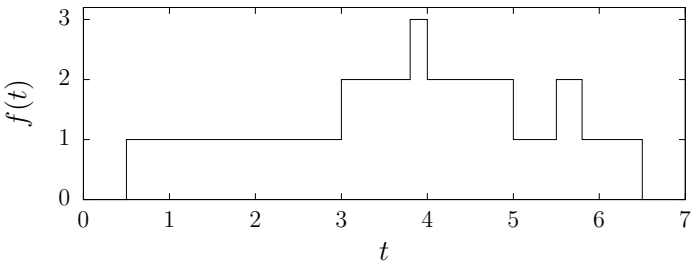


FIGURE 2.1 – **Flot de liens avec durée** – a) Dans le flot de liens, les liens apparaissent et disparaissent au cours du temps. Par exemple, b interagit avec c de $t = 0.5$ à $t = 2$. b) Le graphe induit est formé par l'agrégation de tous les liens. Par exemple, b est lié à c mais pas à a étant donné qu'il n'a pas interagi avec lui dans le flot de liens. c) La série temporelle donne le nombre de liens au cours du temps. Par exemple, $f(3.9) = 3$ car a interagit avec c, et d avec b et c à cet instant.

2.1.2 Exemple de flot de liens ponctuels

La Figure 2.2 montre un deuxième exemple de flot de liens, dans lequel les liens sont ponctuels, ainsi que son graphe statique induit et la série temporelle associée. Le flot de liens $L = (T, V, E)$ représenté est tel que :
– $T = [0, 7[$,
– $V = \{a, b, c, d\}$,
– $E = \{(0.5, b, c), (2, a, c), (2.3, b, d), (3, b, d), (3.8, c, d), (4, c, d), (4.75, b, c), (6, c, d), (6, b, d), (6, a, d)\}$.

En pratique, les instruments de mesure utilisés pour collecter les données ne sont généralement pas suffisamment précis pour enregistrer plusieurs interactions ayant lieu

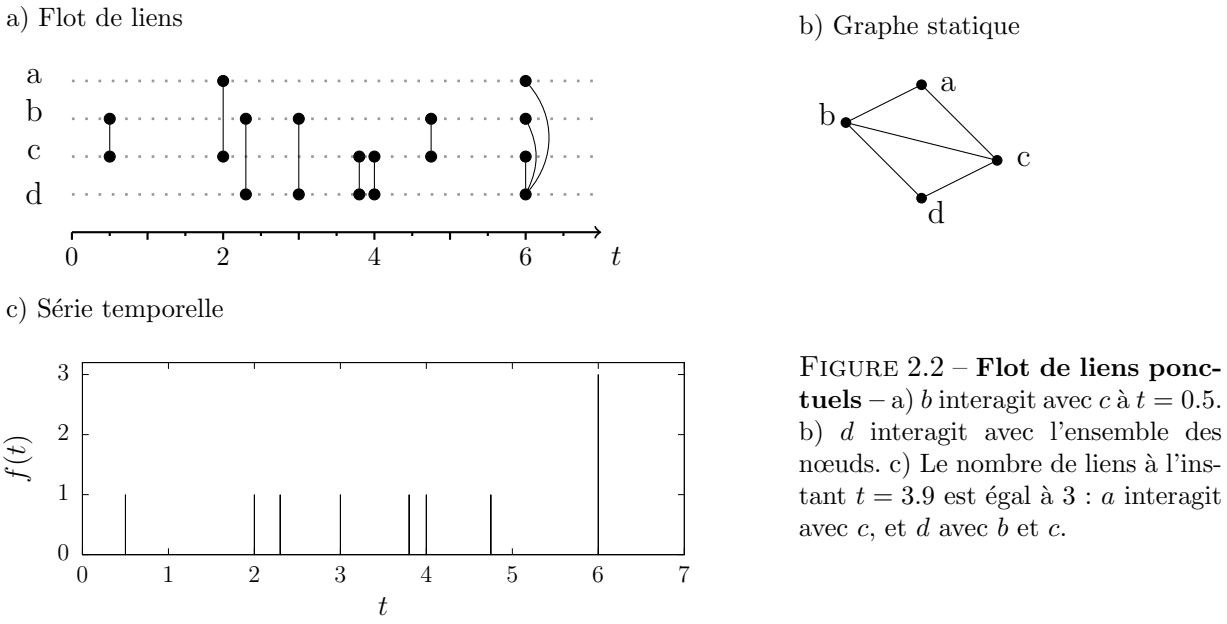


FIGURE 2.2 – **Flot de liens ponctuels** – a) b interagit avec c à $t = 0.5$. b) d interagit avec l'ensemble des nœuds. c) Le nombre de liens à l'instant $t = 3.9$ est égal à 3 : a interagit avec c, et d avec b et c.

au même instant, comme c'est le cas dans l'exemple pour $t = 6$. Ainsi, lorsque les données sont collectées à partir d'un capteur unique, leur granularité temporelle est fixée par la fréquence d'échantillonnage du capteur. Dans cette situation, la série temporelle résultante consiste en une somme de fonctions Delta de Kronecker telle que :

$$f(t) = \sum_{t',(t',u,v)\in E} \delta_{t,t'} \quad \text{où} \quad \delta_{t,t'} = \begin{cases} 1 & \text{si } t = t', \\ 0 & \text{sinon,} \end{cases}$$

ce qui limite sévèrement les informations pouvant être extraites de son analyse.

Dans ce cas, une alternative est de considérer une fenêtre temporelle et d'agréger les interactions ayant lieu à l'intérieur de celle-ci. La série temporelle s'applique alors sur les fenêtres de longueur $\tau \in [0, \omega - \alpha]$, notées $T_i = [i\tau, (i+1)\tau[$ pour tout $i \in \{\frac{\alpha}{\tau}, \ldots, \frac{\omega}{\tau}\}$, telle que

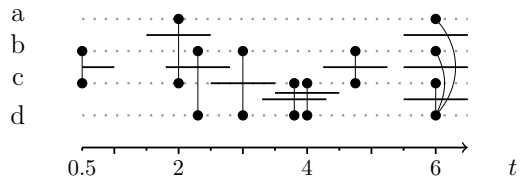
$$f(T_i) = |\{(u, v) : (t, u, v) \in E \text{ et } t \in T_i\}|.$$

Cependant, comme nous l'avons vu dans le Chapitre 1, en plus de perdre des informations sur la structure de par la transformation du flot en série temporelle, cette approche réduit la précision concernant la dynamique des interactions.

Une autre approche consiste alors à transformer le flot de liens ponctuels en un flot de liens avec durée à l'aide d'un paramètre Δ [144]. Dans ce flot, noté L_Δ, si un lien existe entre deux nœuds dans un intervalle de temps de durée égale à Δ dans L, on considère que les deux nœuds sont liés de façon continue pendant cet intervalle dans L_Δ. Soit $[t - \frac{\Delta}{2}, t + \frac{\Delta}{2}] \subseteq T$ cet intervalle, les conditions aux bords imposent $\alpha + \frac{\Delta}{2} \leqslant t \leqslant \omega - \frac{\Delta}{2}$ [144, 93]. Étant donné $L = (T, V, E)$ et une durée $\Delta \in [0, \omega - \alpha]$, on définit alors $L_\Delta = (T_\Delta, V, E_\Delta)$ tel que

$$T_\Delta = \left[\alpha + \frac{\Delta}{2}, \omega - \frac{\Delta}{2}\right] \quad \text{et} \quad E_\Delta = \left\{\left[t - \frac{\Delta}{2}, t + \frac{\Delta}{2}\right] \times (u, v) \ : \ (t, u, v) \in E \text{ et } t \in T_\Delta\right\}.$$

a) Ajout d'une durée aux liens : L'_1



b) Simplification des chevauchements : L_1

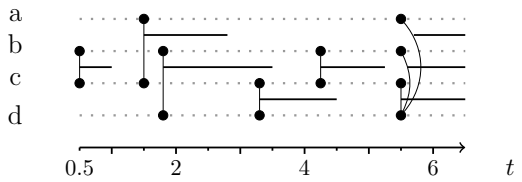
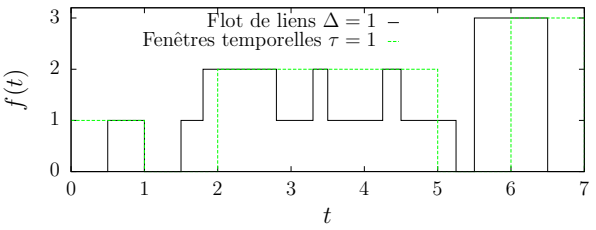


FIGURE 2.3 – **Transformation d'un flot de liens ponctuels L en un flot de liens avec durée L_Δ avec $\Delta = 1$** – a) Étant donné $E = \{(t, u, v)\}$, on crée des liens allant de $t - \frac{1}{2}$ à $t + \frac{1}{2}$ dans l'intervalle $T_\Delta = [0.5, 6.5]$. b) L_1 est obtenu en simplifiant les chevauchements de liens identiques. c) Comparaison des séries temporelles obtenues à partir de fenêtres temporelles de longueur $\tau = 1$ et L_1.

c) Série temporelle



Autrement dit, deux nœuds sont liés à l'instant t' dans L_Δ si et seulement s'ils sont liés dans L à l'instant t tel que $t' \in \left[t - \frac{\Delta}{2}, t + \frac{\Delta}{2}\right] \cap T_\Delta$. Cette opération est illustrée dans les Figures 2.3.a et 2.3.b. [1]

Dans la Figure 2.3.c, on compare les séries temporelles obtenues par l'agrégation des liens dans des fenêtres temporelles ($\tau = 1$) et le flot de liens avec durée ($\Delta = 1$). Alors que la première est constante sur l'intervalle $[2, 5]$, celle induite de L_Δ permet d'inférer la disparition d'un lien à $t = 2.8$, $t = 3.5$ et $t = 4.5$ et l'apparition d'un lien à $t = 3.3$ et $t = 4.25$.

2.2 Degré instantané des nœuds

Le degré du nœud v au temps t, noté $d(t, v)$, est le nombre de nœuds distincts avec lesquels v interagit au temps t :

$$d(t, v) = |\{u \,:\, (t, u, v) \in E\}|.$$

Le profil de degré d'un nœud v est la fonction qui à tous t associe le degré de v à cet instant. La Figure 2.4 montre le profil de degré du nœud c dans le flot de liens de la Figure 2.1.a.

De cette définition, on définit le degré $d(v)$ du nœud v, et le degré $d(t)$ à l'instant t, tels que

$$d(v) = \frac{1}{|T|} \int_t d(t, v)\, dt \qquad \text{et} \qquad d(t) = \frac{1}{|V|} \sum_{v \in V} d(t, v) \quad .$$

Le degré du flot de liens est finalement obtenu en prenant la moyenne du degré des nœuds,

1. Le graphe induit ne change pas.

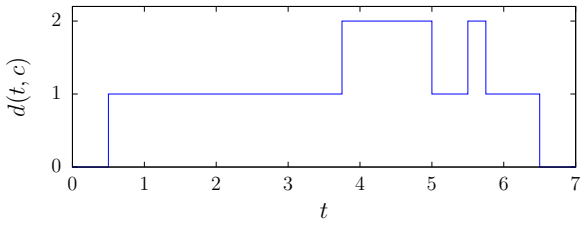


FIGURE 2.4 – **Profil de degré du nœud c dans le flot de liens de la Figure 2.1.a** – Par exemple, c a un degré égal à 2 de $t = 5.5$ à $t = 5.75$: il est lié aux nœuds b et d sur cet intervalle de temps.

ou de façon équivalente, la moyenne temporelle du degré au temps t :

$$d(L) = \frac{1}{|V|} \sum_v d(v) = \frac{1}{|T|} \int_t d(t)\, dt\,.$$

La relation entre la densité (voir Section 1.3.3) et le degré du flot de liens est la même que celle existant dans les graphes. En particulier,

$$\begin{cases} d(L) = \frac{1}{|V|} \sum_v d(v) = \frac{2|E|}{|V|} \\ \delta(L) = \frac{2|E|}{|V|(|V|-1)} \end{cases} \qquad \text{d'où} \quad \delta(L) = \frac{d(L)}{|V|-1}\,.$$

Dans cette thèse, nous caractérisons les flots de liens principalement à l'aide du degré instantané des nœuds. Bien qu'élémentaire, cette propriété donne des informations quantitatives sur leur voisinage au cours du temps. De ce fait, elle permet une première exploration de la façon dont les interactions s'organisent dans le temps et parmi les nœuds.

2.3 Trafic IP

Les données disponibles pour l'analyse du trafic IP consistent en une capture du trafic en un point du réseau Internet. Un routeur est configuré pour la capture et conserve un enregistrement de toutes les entêtes de paquets le traversant. En résulte une séquence, indiquant notamment que deux machines u et v, identifiées par une adresse IP, ont interagi par le biais du routeur à l'instant t (voir Figure 2.5).

```
1372136401.396783 IPv4, length 1468: 194.183.224.211.80 > 210.127.236.73.54770: tcp 1402
1372136401.396817 IPv4, length 92: 176.40.88.51.53 > 22.46.243.156.60382: UDP, length 50
1372136401.396848 IPv4, length 66: 57.33.62.99.443 > 176.170.45.198.53380: tcp 0
1372136401.396940 IPv4, length 1514: 57.33.63.114.80 > 176.183.76.225.39868: tcp 1460
1372136401.397069 IPv4, length 1514: 57.33.63.114.80 > 176.183.76.225.39868: tcp 1460
1372136401.397083 IPv4, length 66: 195.104.197.228.52434 > 32.111.15.242.8080: tcp 0
1372136401.397089 IPv4, length 1238: 181.236.109.73.50137 > 26.254.234.253.443: tcp 1172
1372136401.397111 IPv4, length 66: 211.39.74.71.45515 > 194.183.224.211.80: tcp 0
```

FIGURE 2.5 – **Exemple de trafic IP** – Chaque ligne représente un paquet. Le premier champ est l'instant de capture et les $5^{\text{ème}}$ et $7^{\text{ème}}$ champs sont les adresses IP. Les traces de trafic contiennent également des informations sur les ports et le protocole utilisés ainsi que la taille du paquet.

2.3.1 Description des données

Nous utilisons des traces de trafic IP fournies par le groupe MAWI, *Measurement and Analysis on the WIDE Internet*, où WIDE est un réseau universitaire japonais [87]. Ces traces de trafic IP sont capturées depuis un point de mesure sur un lien trans-Pacifique entre le réseau WIDE et les fournisseurs d'accès internet en amont. Chaque jour, 15 minutes de trafic sont collectées, anonymisées et rendues publiques. Certaines traces plus longues allant de 24 heures à 96 heures ont également été capturées dans le cadre du projet « *A Day in the Life of the Internet* » [1]. Étant donné la mesure effectuée, nous n'avons qu'une vision partielle des échanges. En effet, les interactions sont biparties : nous n'avons pas accès aux interactions entre deux machines du réseau WIDE ni deux machines du réseau Internet avec lesquelles WIDE interagit.

Jeux de données

Nous utilisons trois jeux de données différents :

– Le premier est une trace de trafic IP d'une heure du 25 juin 2013 de 00 : 00 à 1 : 00 (UTC+9). On note cette trace par un ensemble $\mathcal{D}_1$ de triplets tel que $(t, u, v) \in \mathcal{D}_1$ indique que les adresses IP u et v ont échangé au moins un paquet au temps t. L'ensemble $\mathcal{D}_1$ contient 83 386 538 triplets impliquant 1 157 540 adresses IP différentes.

– Le deuxième est une trace longue d'une journée du 25 juin 2013 à 00 : 00 au 26 juin 2013 à 00 : 00 (UTC+9). L'ensemble $\mathcal{D}_2$ contient 2 196 079 591 triplets impliquant 15 390 238 adresses IP différentes. [2]

– Le dernier est une trace de 15 minutes datant du 3 novembre 2018 de 14 : 00 à 14 : 15 (UTC+9). L'ensemble $\mathcal{D}_3$ contient 64 913 871 triplets impliquant 16 453 608 adresses IP différentes.

Ces trois jeux de données comportent plusieurs millions d'interactions. Leur précision temporelle est de l'ordre de la microseconde.

Modélisation en flot de liens

Les ensembles de triplets $\mathcal{D}_1$, $\mathcal{D}_2$ et $\mathcal{D}_3$ constituent des ensembles de liens ponctuels dans lesquels les nœuds sont les adresses IP impliquées dans $\mathcal{D}_i$ et où deux nœuds sont liés à l'instant t si $(t, u, v) \in \mathcal{D}_i$.

Afin d'étudier le degré des nœuds, il est nécessaire de transformer les flots de liens ponctuels en flots de liens avec durée. On considère alors que deux nœuds sont liés entre le temps t_1 et le temps t_2 s'ils ont échangé au moins un paquet toutes les secondes dans cet intervalle de temps ($\Delta = 1s$). On note le flot de liens résultant $\mathcal{L}_i$. Par exemple,

2. Les jeux de données $\mathcal{D}_1$ et $\mathcal{D}_2$ ne correspondent pas aux données brutes : ils sont obtenus après suppression d'une adresse IP scannant l'espace IPv4 dans le cadre du projet académique USC ANT.

$\mathcal{L}_1 = (T, V, E)$ est tel que

$$E = \left\{ \left[t - \frac{1}{2}, t + \frac{1}{2} \right] \times (u, v) \, : \, (t, u, v) \in \mathcal{D}_1 \text{ et } t \in T \right\},$$

avec $T = \left[0 + \frac{1}{2}, 3600 - \frac{1}{2}\right] = [0.5, 3599.5]$. Dans ce cas, les interactions s'apparentent à des flux IP dans lesquelles deux adresses IP sont en communication durant une période de temps, plutôt qu'à des envois ponctuels de paquets. D'autres choix concernant la durée des liens Δ peuvent être effectués. On pourrait même définir une valeur de Δ différente pour chaque paire (u, v) ou chaque lien (t, u, v), en utilisant des connaissances externes sur les flux IP [38], ou encore, utiliser une valeur de Δ qui change dans le temps de façon similaire à ce que font Léo et al. dans leurs travaux [94]. Ceci dépasse toutefois le cadre de cette thèse, où nous considérons toujours $\Delta = 1s$ dans le cas du trafic IP. Enfin, il est à noter qu'étant donné la mesure du trafic, les flots de liens sont bipartis.

2.3.2 Degré instantané des nœuds

Dans cette sous-section on s'intéresse au degré dans le flot de liens $\mathcal{L}_1$ du 25 juin 2013 de $0h$ à $1h$.

La Figure 2.6.a montre les profils de degré de trois adresses IP différentes. L'adresse ip1 interagit avec plus de 5 000 adresses IP à plusieurs reprises au cours de la trace ; ip2 n'interagit jamais avec plus d'une adresse IP à la fois ; finalement, ip3 est caractérisée par un degré moyen fluctuant autour de 300, avec un pic d'activité aux alentours de la $450^{\text{ème}}$ seconde.

La Figure 2.6.b montre la distribution (*resp.* la fonction de répartition complémentaire) du degré sur l'ensemble des entités nœuds-temps $(t, v) \in T \times V$. Un point de cette distribution en $d(t, v) = k$, indique la probabilité lorsque que l'on choisi un nœud-temps (t, v) au hasard de manière uniforme parmi les noeuds-temps actifs [3], que ce dernier ait un degré égal à k (*resp.* supérieur à k). On remarque que cette distribution est très hétérogène, c'est-à-dire que, d'une part, le degré instantané des nœuds prend des valeurs différentes de plusieurs ordres de grandeur, et d'autre part, que, globalement, plus $d(t, v)$ est élevé, plus sa probabilité est faible.

La Figure 2.6.c montre l'évolution du degré moyen au cours du temps. On remarque qu'à l'exception de quelques pics d'activité, son profil est globalement constant. La Figure 2.6.d montre sa distribution. Dans cette dernière, un point $d(t) = x$ indique la probabilité lorsque l'on tire un instant $t \in T$ au hasard de manière uniforme, que le degré moyen à cet instant soit égal à x. Comme attendu au vu de son évolution temporelle, la distribution du degré sur le temps est homogène avec anomalies, c'est-à-dire que la plupart des valeurs fluctuent autour d'une moyenne tandis qu'une très faible proportion en dévie significativement [92].

On réalise des observations similaires concernant les flots de liens $\mathcal{L}_2$ et $\mathcal{L}_3$. Le degré au cours du temps de $\mathcal{L}_3$ présente en plus des variations liées au rythme circadien.

3. Un nœud actif à l'instant t est un nœud interagissant avec au moins un autre nœud à cet instant.

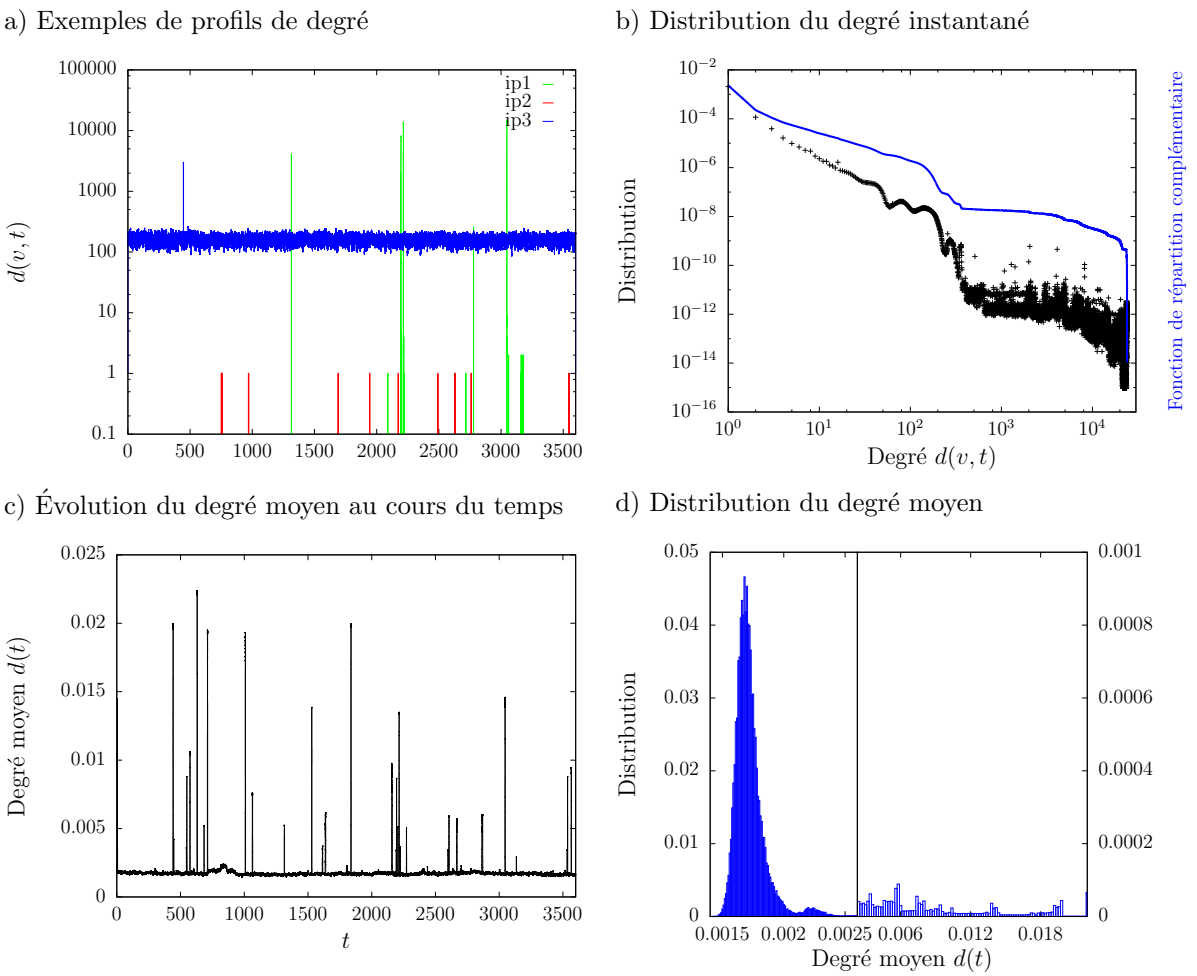


FIGURE 2.6 – **Degré dans la trace de trafic IP du 25 juin 2013 de** $0h$ **à** $1h$ **(UTC+9)** – Dans la figure d., la distribution est représentée avec deux graduations distinctes sur les axes des abscisses et des ordonnées en fonction que l'on soit situé dans le comportement normal ou anormal.

2.3.3 Comparaison : MAWILab

Les traces de trafic IP du groupe MAWI ont été intensément analysées dans le cadre du projet WIDE, notamment au travers des travaux de Borgnat et al. [26], Fontugne et al. [53, 52] et Cho et al. [37]. Les méthodes développées dans ces articles visent à détecter des anomalies dans le trafic IP à partir de décompositions en ondelettes et de transformations en sketches de séries temporelles (voir Section 1.2.3). Ils fournissent à partir de ces travaux une base de données publique, MAWILab, qui répertorie les anomalies dans les traces de trafic de 15 minutes relevées quotidiennement.

Dans cette base de données, une anomalie consiste en une adresse IP et une période de temps durant laquelle l'adresse IP est évaluée comme anormale. De plus, chaque anomalie est associée à une étiquette la classant parmi les catégories suivantes [105] :
- déni de service point à point : deux adresses IP s'échangent un grand nombre de paquets ;
- déni de service distribué : plusieurs sources envoient un grand nombre de paquets à une destination ;

- scan réseau : une adresse IP scanne un réseau de plusieurs adresses IP destination ;
- scan de ports : une adresse IP balaye plusieurs ports d'une même destination ;
- Point multipoint : trafic normal de routeur ;
- Flot alpha : trafic pair à pair normal ;
- Autres : concerne par exemple les pannes de serveur.

Dans le cadre du trafic IP, nous pouvons donc comparer nos résultats aux anomalies répertoriées dans cette base. Cependant, il est nécessaire de prendre du recul vis-à-vis de cette dernière : la base d'anomalies MAWILab constitue une vérité établie et non une vérité de terrain sur le trafic. Ainsi, certains événements peuvent avoir été omis par tous les détecteurs (faux négatifs), alors que d'autres rapportés par MAWILab peuvent ne pas être des événements (faux positifs).

2.4 Ensemble de retweets sur Twitter

Twitter fournit gratuitement deux APIs pour la collecte de tweets. La première est l'API *follow*. Pour chaque utilisateur u spécifié, les données contiennent : les tweets créés par u, les tweets retweetés par u ainsi que les réponses et les retweets des tweets créés par u. La deuxième est l'API *track* qui permet de capturer des tweets mentionnant une expression donnée. Ces APIs comportent des limites sur la taille du corpus qui peut être collecté et ne donnent accès qu'à un échantillon de la totalité des tweets. Étant donné que nous nous intéressons aux interactions entre les utilisateurs, nous considérons uniquement les retweets.

2.4.1 Description des données

Nous utilisons un sous-ensemble des retweets collectés dans le cadre du projet *Politoscope* [4]. Dans ce projet, Gaumont et al. [57] étudient la formation et l'évolution des communautés politiques tout au long de la campagne présidentielle française de 2017. À cette fin, ils rassemblent des données relatives à la politique sur une période de plus d'un an : ils suivent près de 3 700 acteurs politiques français (par exemple, des politiciens, des mairies ou des partis politiques) grâce à l'API *follow*, ainsi qu'une liste de mots-clés et de hashtags relatifs à la campagne présidentielle (par exemple, « *primaires2016* » et « *présidentielle2017* ») grâce à l'API *track*.

Jeux de données

Nous utilisons deux jeux de données différents :

– Le premier contient l'ensemble des retweets durant le mois d'août 2016. On le désigne par un ensemble $\mathcal{D}_4$ de triplets tel que $(t, s, a) \in \mathcal{D}_4$ indique que le diffuseur s (*spreader* en anglais) a retweeté un tweet de l'auteur a à l'instant t, où le tweet correspondant contient des mots clés liés à la politique, où a appartient à l'ensemble d'acteurs politiques répertoriés dans le projet *Politoscope*. Il contient 1 142 004 retweets et implique 211 155 utilisateurs différents.

– Le deuxième est identique au premier à la différence qu'il contient en plus une information sémantique sur le contenu du tweet retweeté. On le désigne par un ensemble $\mathcal{D}_5$ de quadruplets tel que $(t, s, a, k) \in \mathcal{D}_5$ indique que s a retweeté un tweet écrit par a et contenant le hashtag k à l'instant t. Il contient 30 057 hashtags différents.

Ces deux jeux de données contiennent également plus d'un million d'interactions. La précision temporelle est ici de l'ordre de la seconde.

L'utilisation de données Twitter pour la recherche soulève un certain nombre de défis éthiques [10, 150, 101]. Le fait que l'API de Twitter ne donne pas accès à l'entièreté des tweets et que Twitter soit une plateforme ouverte ne sont pas des justifications suffisantes pour collecter, analyser et classifier des profils d'utilisateurs à des fins de recherche, en particulier dans le cas de données sensibles liées à l'orientation politique des utilisateurs. De ce fait, nous portons une attention particulière au respect de la privacité : tout au long de cette thèse, les noms d'utilisateurs ne sont mentionnés que lorsqu'ils correspondent à des comptes Twitter officiels de politiciens ou d'organisations publiques comme des mairies, des journaux ou des émissions. Dans le cas contraire, ils sont désignés par les termes génériques *usern*, où n est un entier permettant de différencier les utilisateurs anonymisés. De plus, nous prenons garde à afficher uniquement des données agrégées de façon à rendre l'identification d'individus impossible.

Modélisation en flot de liens

Étant donné que la fréquence d'interaction dans Twitter est moins élevée que celle dans le trafic IP, on modélise les interactions par des flots de liens avec durée $\mathcal{L}_4$ et $\mathcal{L}_5$ en choisissant $\Delta = 30\text{min}$. Dans ce cas, l'activité d'un utilisateur s'apparente à une session de navigation plutôt qu'à des utilisations ponctuelles. Ici également, d'autres choix concernant le paramètre Δ peuvent être effectués. Notamment, on pourrait se servir des travaux de Lim et al. [96] sur la caractérisation des sessions de navigation à l'aide des historiques de connexions indiquant les heures de début et de fin d'une session. Enfin, contrairement aux traces de trafic IP, les flots de liens ne sont pas bipartis : les utilisateurs peuvent à la fois retweeter et être retweeté. De plus, dans le cas de $\mathcal{D}_5$, le flot de liens est étiqueté par les hashtags.

2.4.2 Degré instantané des nœuds

Dans cette sous-section on s'intéresse au degré dans $\mathcal{L}_4$.

La Figure 2.7.a montre le profil de degré de trois utilisateurs. Alors que user1 et user3 sont actifs sur toute la durée du mois, user2 ne devient actif qu'à partir du 24 août. On remarque également, pour chacun des utilisateurs, qu'ils sont actifs par pics d'activité comme c'est souvent le cas dans les interactions humaines [85].

La Figure 2.7.b montre la distribution du degré sur l'ensemble des nœuds-temps. Bien qu'étant plus lisse, elle est hétérogène comme celle observée dans le cas du trafic IP.

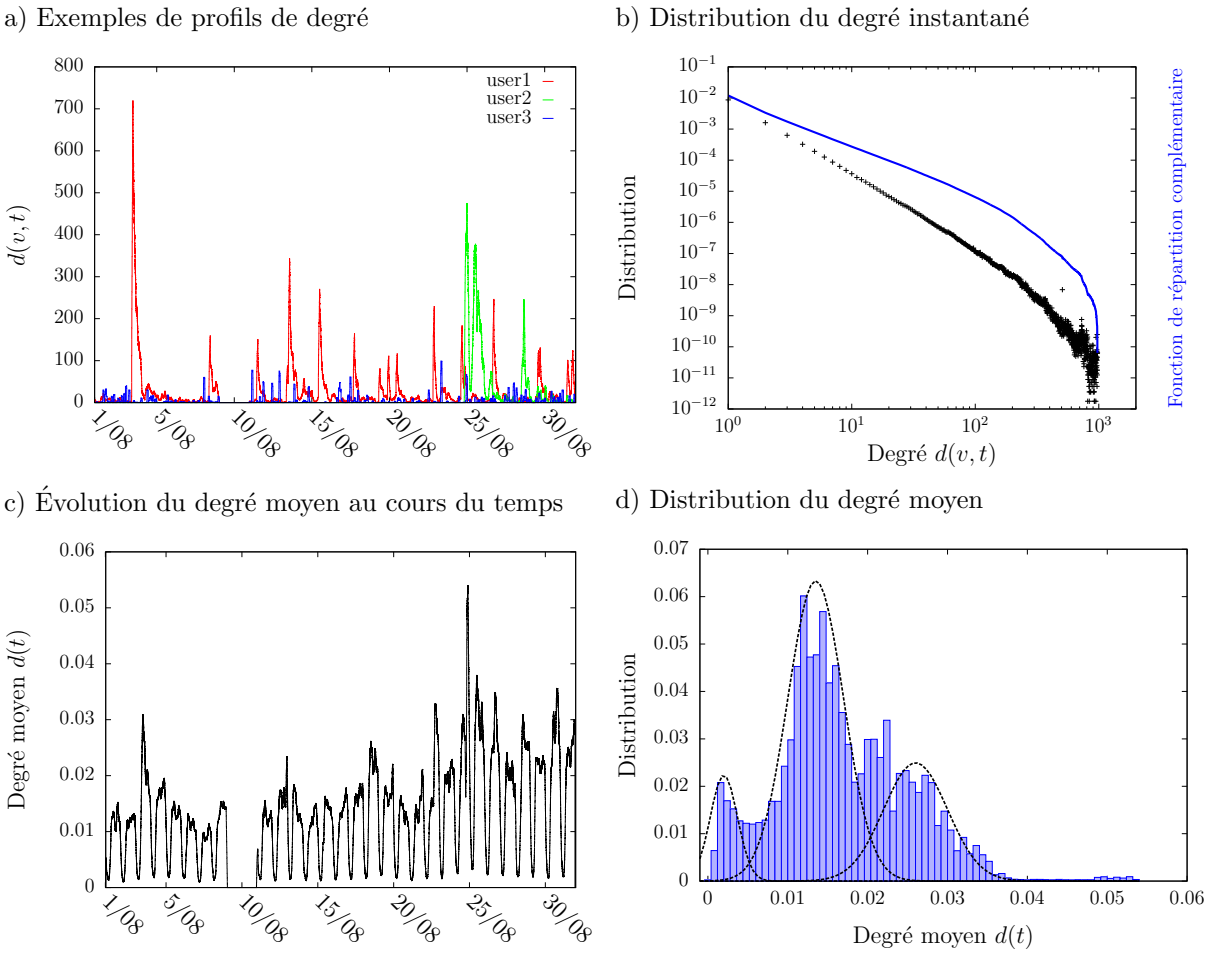


FIGURE 2.7 – **Degré dans l'ensemble de retweets du mois d'août 2016.**

L'évolution du degré moyen au cours du temps sur la Figure 2.7.c par contre, diffère de celle de la trace de trafic IP $\mathcal{D}_1$. Comme on le montrera en détail dans le Chapitre 3, on distingue trois comportements distincts (voir Figure 2.7.d) : une périodicité liée à la nuit et au jour, ainsi qu'une augmentation globale du degré sur la fin du mois. En raison d'une panne de serveur du mardi 9 au jeudi 11 août, aucune activité n'est observée pendant cette période.

2.4.3 Comparaison : évènements médiatiques

Les ensembles de retweets que nous considérons ont été intensément analysés dans le cadre du projet *Politoscope* dans les travaux de Gaumont et al. [57]. L'ensemble des méthodes développées dans cet article vise principalement à décrire et quantifier l'activité, les interactions et les particularités sémantiques des différentes communautés politiques. Ainsi, leur analyse n'est pas focalisée sur la détection d'anomalies.

Nous nous limitons dans cette thèse à comparer les anomalies que nous détectons avec les évènements politiques qui ont été médiatisés en août 2016. Notamment, nous effectuons une validation exploratoire des résultats en regardant, pour chaque anomalie, quels

sont les évènements médiatiques liés à la campagne présidentielle qui y sont associés au travers d'articles de presse trouvés sur le Web.

Nous utilisons de plus un jeu de données collecté par l'Institut national de l'audiovisuel répertoriant tous les programmes de télévision et de radio du mois d'août 2016. Pour chacune des émissions, il donne accès à son titre, son heure de diffusion, ainsi qu'aux noms des présentateurs et des invités y participant. Ce jeu de données nous permet de valider des anomalies liées à l'apparition de politiciens dans les médias télévisés et radiophoniques, et de confirmer des corrélations entre des hashtags liés à une émission et un politicien ou une actualité.

2.5 Détection d'anomalies dans le trafic IP et Twitter

Dans cette section, nous passons en revue les différentes approches utilisées pour la détection d'anomalies dans Twitter et le trafic IP. Nous nous positionnons ensuite par rapport à cet état de l'art en termes de point de vue adopté et de technique utilisée.

2.5.1 Détection d'anomalies dans le trafic IP

Le problème de la détection d'anomalies dans le trafic IP suscite un grand intérêt parmi les chercheurs et est abordé de différentes façons, notamment en fonction de la définition des anomalies considérée et des techniques utilisées.

Certains s'intéressent à la détection d'*instants* anormaux. Iliofotou et al. [80] introduisent les *Traffic Dispersion Graphs* (TDG) où un TDG est un graphe $G = (V, E)$ dans lequel les nœuds $u, v \in V$ sont des adresses IP et où un lien $(u, v) \in E$ indique que l'interaction entre u et v correspond à une règle donnée, par exemple, « les deux adresses se sont échangées un paquet », ou encore « au moins trois paquets TCP ont été échangés sur le port 53 ». Dans des travaux ultérieurs, ils détectent des instants anormaux en étudiant la dynamique d'une séquence de graphes TDG [79].

D'autres s'intéressent aux *adresses IP* anormales. Par exemple, Xu et al. [160] modélisent le trafic sous forme de graphe biparti et étudient la projection monomode du graphe pour en extraire des nœuds ayant des comportements anormaux par rapport aux autres nœuds du même ensemble.

Asai et al. [20] détectent des séquences d'*interactions* anormales. Ils incluent les informations temporelles dans les interactions d'un graphe statique plutôt que d'étudier une séquence de graphes. Ils introduisent dans ce but des graphes causaux (*Traffic Causality Graphs*) dans lesquels un nœud est une interaction et où deux nœuds sont liés s'ils ont une relation de cause à effet.

Enfin, Latapy et al. [92] détectent à la fois des *instants et des nœuds-temps* anormaux. Ils calculent un ensemble de propriétés décrivant la dynamique des graphes ou des nœuds dans une séquence de graphes, comme par exemple le nombre de nouveaux nœuds apparaissant à chaque pas de temps. Parmi ces propriétés, ils rejettent celles ayant une

distribution hétérogène et sélectionnent celles ayant une distribution homogène avec anomalies de façon à en extraire des observations anormales.

Mise à part les techniques reposant sur la *théorie des graphes* comme celles citées ci-dessus, d'autres méthodes analysent le trafic IP en tant que *signal*. Par exemple, Barford et al. [22] détectent des flux IP anormaux à des fréquences caractéristiques après avoir décomposé le signal en ondelettes. Borgnat et al. [26] caractérisent le comportement normal du trafic avec des sketches en utilisant les traces MAWI quotidiennes de 15 minutes sur une durée de 7 ans. Ils identifient ainsi plusieurs types d'anomalies et en observent les évolutions au cours des années. D'autres méthodes utilisent l'*analyse en composantes principales* (PCA) et identifient des anomalies parmi les points résidus [91, 127, 84]. Enfin, une autre catégorie de méthodes, plus générale, détecte des anomalies par *apprentissage automatique* du comportement normal [153, 51, 104, 30]. Par exemple, Casas et al. [30] détectent des flux IP anormaux à l'aide d'un algorithme de détection non supervisé (UNADA). Ils détectent d'abord des fenêtres temporelles anormales à partir de séries temporelles extraites des flux IP, puis y identifient des anomalies en utilisant des algorithmes de partitionnement.

2.5.2 Détection d'anomalies dans Twitter

Les différentes approches employées pour la détection d'anomalies sur Twitter peuvent être classées selon les mêmes critères à savoir, le type d'entités considéré et la technique employée.

Certains chercheurs considèrent les anomalies comme étant des événements du monde réel se déroulant à un endroit et à un moment donnés. Par exemple, Sakaki et al. [128] et Bruns et al. [28] identifient des *mots clés* spécifiques à un événement extérieur (par exemple « *feu* » et « *pompiers* » pour un incendie) puis détectent des anomalies en surveillant l'évolution temporelle de l'utilisation de ces mots dans les tweets. Il existe également des méthodes basées sur le regroupement de tweets. Dans ces approches, les auteurs déduisent à partir d'horodatages, de géolocalisations et du contenu des tweets une similitude entre chaque paire de tweets et trouvent des événements réels dans des groupes de tweets similaires [46, 95, 149].

D'autres chercheurs recherchent des *utilisateurs* ayant des comportements anormaux selon différents critères. Varol et al. [142] détectent des bots au moyen d'une technique d'apprentissage automatique supervisé. Ils extraient des propriétés liées aux activités des utilisateurs au cours du temps, aux réseaux d'amitié des utilisateurs ainsi qu'aux contenus des tweets, puis ils les utilisent pour identifier des bots à l'aide d'un jeu de données étiqueté. Stieglitz et al. [135] identifient les utilisateurs influents en étudiant la corrélation entre le vocabulaire qu'ils utilisent dans les tweets et le nombre de fois qu'ils sont retweetés. Ribeiro et al. [126] détectent des utilisateurs haineux. Ils commencent par classer les utilisateurs avec une méthode basée sur le lexique, puis montrent que les utilisateurs haineux diffèrent des utilisateurs normaux en termes d'activité et de structure de réseau.

Enfin, d'autres travaux visent à trouver des *relations* privilégiées entre les utilisa-

teurs. Parmi ceux-ci, les travaux de Wong et al. [158] l'appliquent à l'opinion politique en combinant une analyse du nombre de retweets entre deux utilisateurs et une analyse de sentiments des tweets retweetés.

Les méthodes citées ci-dessus utilisent des techniques d'*exploration de textes*. D'autres méthodes sont basées sur le *volume des interactions*. Par exemple, Chavoshi et al. [34] et Chierichetti et al. [36] utilisent une technique similaire à celle de Varol et al. [142], mais exploitent uniquement les activités des utilisateurs au travers de leurs nombres de tweets et de retweets. Grasland et al. [60] détectent plusieurs types d'anomalies à l'aide d'une analyse quantitative du nombre de rewteets dans le cadre de la couverture médiatique des pays dans les journaux au cours du temps, notamment, les pays ayant bénéficié d'une grande attention, les semaines où un pays donné a été le plus présent dans l'actualité, ou encore les couples semaines-pays qui ont bénéficié de la plus forte attention de la part d'un média donné.

Une autre alternative aux techniques d'exploration de textes consiste à utiliser des *propriétés de graphes*. Song et al. [134] et Bild et al. [24], par exemple, identifient les spammeurs en temps réel à l'aide d'une mesure de distance et de connectivité entre les utilisateurs dans le graphe d'amitié dirigé (*followers*) d'une part, et le graphe des retweets d'autre part. Avec le graphe des retweets également, Ten et al. [140] détectent les tendances en examinant les modifications de la taille et de la densité du plus grand composant connecté. Enfin, l'approche de Coletto et al. [40] combine une analyse du graphe d'amitié et du graphe des retweets pour identifier des controverses dans les discussions.

2.5.3 Positionnement et approche

Les avantages de la modélisation en flots de liens sur la modélisation en séquence de graphes ou en séries temporelles pour l'analyse des interactions ont déjà été discutés dans le Chapitre 1. Dans cette section, nous nous concentrons sur l'approche que nous adoptons par rapport à celles adoptées dans les autres travaux.

Les méthodes pré-existantes de détection d'anomalies dans Twitter et dans le trafic IP utilisent différentes méthodes pour détecter différents types d'anomalies. Elles considèrent souvent un seul contexte selon lequel les anomalies sont définies. Au contraire, nous souhaitons traiter les différents types d'anomalies de manière unifiée et considérer différents points de vue par rapport auxquels une observation peut être anormale. De cette façon, nous souhaitons fournir une approche systématique rendant d'une part la recherche d'anomalies dans un flot de liens plus aisée, et d'autre part, visant à ce que l'ensemble des circonstances par rapport auquel une observation est anormale soit explicite. Concernant ce dernier point, nous attacherons une importance particulière à la détermination du contexte dans lequel les anomalies sont définies afin de pouvoir au mieux interpréter et valider nos résultats [89].

Au niveau de la technique utilisée pour détecter des anomalies, nous centrons notre recherche sur la détection d'anomalies statistiques. En pratique, nous nous basons sur les travaux de Latapy et al. [92] : nous tentons de trouver des propriétés et des contextes

vis-à-vis desquels les observations se distribuent de façon homogène avec anomalies (voir Figure 2.8). La raison de ce choix est multiple.

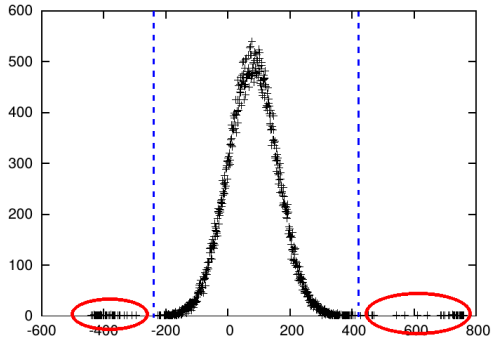


FIGURE 2.8 – **Distribution homogène avec anomalies** – Les observations anormales, entourées en rouge, sont celles déviant significativement de la moyenne.

Premièrement, nous ne souhaitons pas d'une technique dont les résultats sont très sensibles aux paramètres qu'elle implique, comme par exemple, les méthodes PCA, dont les anomalies détectées dépendent du nombre de composantes principales, ou les méthodes par apprentissage automatique supervisé qui dépendent de la qualité du jeu de données étiqueté, ou encore, les méthodes d'ajustements de distributions hétérogènes qui sont sujettes à débat comme nous le verrons en détails dans le Chapitre 4.

Deuxièmement, détecter des anomalies dans ce type de distributions est naturel : le comportement normal est bien défini et modélisé par une loi normale $P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ dans laquelle la plupart des valeurs sont réparties autour d'une moyenne μ avec un écart type σ, et les anomalies sont les valeurs qui dévient de la moyenne de plus de trois écarts-types [33, 72]. De façon rigoureuse, pour déterminer si une distribution donnée est homogène avec anomalies ou non, on effectue les étapes suivantes [61] :
(1) On supprime de manière itérative les anomalies de la distribution avec la règle $m \pm 3\sigma$, de façon à obtenir, en dernière itération, une moyenne et un écart-type non biaisés par les anomalies ;
(2) On ajuste la distribution résultante avec la loi normale en utilisant le maximum de vraisemblance (MLE) ;
(3) On évalue la qualité de l'ajustement avec le test de Kolmogorov-Smirnov (KS) qui calcule le maximum de la différence entre les fonctions de répartition de l'ajustement et de la distribution empirique.

Nous pensons que ce positionnement est adapté à une première approche de la détection d'anomalies dans les flots de liens : il parait essentiel de fournir des outils méthodologiques à la fois simples et intuitifs qui dressent un cadre rigoureux dans lequel les anomalies sont définies. Des approches plus complexes pourront par la suite être explorées (voir par exemple les travaux de Motulsky et al. [110]).

2.6 Conclusion

Les traces de trafic IP et l'ensemble de retweets sur Twitter, bien qu'étant des jeux de données *a priori* très différents, présentent des similitudes. Du fait de la diversité à la fois

des nœuds et du comportement de chacun des nœuds au cours du temps, leur degré instantané se distribue de façon hétérogène sur l'ensemble des nœuds-temps. Au contraire, une fois agrégé sur les nœuds, le degré moyen au cours du temps $d(t)$ est globalement stable lorsque la trace est suffisamment courte, et présente des régularités liées au rythme circadien lorsque cette dernière est longue d'un jour ou plus.

Au cours de ce chapitre, nous avons identifié deux problématiques concernant la recherche d'anomalies dans un flot de liens. La première, d'ordre général, résulte d'un manque dans l'état de l'art dans lequel la plupart des travaux antérieurs se limitent à la recherche d'un seul type d'anomalies, selon un unique point de vue. Elle consiste à traiter les différents types d'anomalies de façon unifiée et à préciser les circonstances exactes dans lesquelles elles se manifestent. La deuxième découle de l'étude préliminaire du degré des nœuds au cours du temps qui révèle la présence d'une multitude de comportements très différents les uns des autres. Elle consiste à inférer un comportement normal à partir de distributions hétérogènes. Nous apportons des éléments de réponse à ces questions dans les Chapitres 3 et 4.

Chapitre 3

Anomalies contextuelles dans les interactions temporelles

Sommaire

Résumé : Nous proposons une méthode permettant d'explorer systématiquement des millions d'interactions afin d'en extraire différents types d'anomalies de façon unifiée. Pour cela, nous introduisons la notion de contexte comme étant l'ensemble des circonstances par rapport auxquelles une entité est anormale. Par l'utilisation de contextes variés, nous montrons que notre méthode permet de mettre en évidence des anomalies pertinentes selon plusieurs dimensions. Dans le cadre de la communication politique sur Twitter, nous identifions des heures anormales en observant leur nombre de retweets, puis déterminons si leur anormalité est due à un ou plusieurs auteurs singulièrement retweetés, ou au contraire résulte d'un phénomène global. En poursuivant ce raisonnement, nous identifions des groupes de diffuseurs ainsi que des hashtags anormaux liés à ces évènements. Dans le trafic IP, nous identifions par la même procédure des secondes anormales puis les adresses IP qui en sont à l'origine. Nous montrons également que la modélisation des interactions temporelles par un flot de liens permet d'obtenir une meilleure précision temporelle mais complexifie la détection d'anomalies.

Dans un flot de liens, on peut rechercher un nœud ou un instant globalement anormal, mais aussi, un nœud se comportant anormalement à un instant particulier. De même, on peut vouloir détecter une paire de nœuds anormale mais aussi une interaction inhabituelle entre deux nœuds à un instant donné. Ainsi, les anomalies dépendent premièrement du type d'*entités* recherché. Par ailleurs, si l'entité choisie est un nœud-temps $(t, v) \in T \times V$, v peut être désigné comme anormal à l'instant t si par exemple son degré ou la structure de son voisinage sont anormaux, mais aussi si son degré varie de façon inhabituelle ou si la structure de son voisinage change. Les anomalies dépendent donc également de la *propriété* choisie afin de caractériser le comportement des entités. Enfin, si la propriété choisie est le degré de v au temps t, (t, v) peut être anormal vis-à-vis de tous les autres nœuds à cet instant, ou par rapport à l'activité passée de v, mais aussi, par rapport à une activité type modélisée par l'activité passée de l'ensemble des nœuds. L'anormalité d'une observation dépend donc aussi d'un *comportement attendu* auquel elle est comparée.

Dans ce chapitre, on introduit la notion de *contexte* comme étant l'ensemble des éléments, ou circonstances, formant le cadre d'une anomalie. Comme illustré ci-dessus, une anomalie peut être définie de nombreuses façons différentes dans un flot de liens tant le nombre de variables dont dépend le contexte est élevé. Soit X un ensemble d'entités du flot de liens L, on définit formellement un contexte C comme étant l'ensemble des éléments suivants :
- un ensemble de valeurs observées $\mathcal{O} = \{f(x),\, x \in X\}$, où f est la propriété mesurée ;
- un ensemble de valeurs attendues $\mathcal{E} = \{f_{exp}(x),\, x \in X\}$;
- un ensemble de valeurs de déviation $\mathcal{D} = \{d(f(x), f_{exp}(x)),\, x \in X\}$ entre les valeurs observées et les valeurs attendues.

Étant donné C, une entité anormale $x^* \in X$ est une entité pour laquelle la valeur de déviation $d(f(x^*), f_{exp}(x^*))$ est significativement plus élevée que la plupart des autres valeurs de déviation.

Dans la littérature, Liu et al. [97] formalisent également la notion de contexte mais

dans le cas de données multidimensionnelles $\{\mathbf{x_i} \in \mathbb{R}^M \mid i \in [1, N]\}$. Ils partitionnent les observations en clusters en considérant la distance Euclidienne, modélisent leur comportement avec un modèle linéaire, puis définissent le contexte vis-à-vis duquel une observation $\mathbf{x_i}$ est anormale par l'ensemble $C = \{\mathcal{A}, d(\mathbf{x_i}), \mathcal{C}_i\}$, où $\mathcal{A}$ est l'ensemble des paramètres du modèle, $d(\mathbf{x_i})$ le score d'anormalité de $\mathbf{x_i}$, et $\mathcal{C}_i$ le cluster auquel appartient $\mathbf{x_i}$. Ainsi, leur définition du contexte est similaire à la notre : une valeur observée est comparée à une valeur attendue fixée par un modèle et un cluster, et leur déviation est quantifiée à l'aide d'un score. À la différence, comme souligné précédemment, l'ensemble des valeurs observées et attendues pouvant être considérées dans un flot de liens est beaucoup plus important, ce qui permet la formation de contextes plus variés et plus complexes.

Construire des contextes variés et pertinents possède plusieurs avantages. Premièrement, cela permet de détecter de façon méthodique et organisée des anomalies d'intérêt vis-à-vis d'une application donnée. Deuxièmement, cela donne la possibilité de détecter plusieurs types d'entités anormales et, pour un même type d'entités, de chercher des anomalies selon différents points de vue. Les travaux de Gao et al. [56] illustrent cet intérêt : ils partitionnent un graphe statique, trouvent des nœuds anormaux dans le contexte de chaque communauté, puis montrent qu'en prenant en compte un contexte global, ils ne détectent pas les mêmes anomalies. Enfin, la construction d'un contexte adapté peut permettre de mieux cibler le comportement normal par l'obtention de distributions homogènes avec anomalies.

Dans ce chapitre, nous concevons un ensemble d'étapes permettant de façonner divers contextes. De ce fait, les outils méthodologiques que nous proposons ont pour objectifs de faciliter et de systématiser la recherche d'anomalies dans un flot de liens. Cependant, dans un premier temps, nous réalisons une série de choix visant à faciliter l'explication de notre méthode. Notamment, nous la formalisons dans le cadre d'interactions agrégées dans des fenêtres temporelles. Cela nous permet d'une part de limiter le nombre de degrés de liberté, et d'autre part, de représenter les valeurs observées sous la forme de cubes de données afin d'illustrer visuellement les opérations effectuées pour construire un contexte. De plus, nous choisissons une propriété basique, la quantité d'interactions, qui consiste à compter le nombre d'interactions d'une entité. Dans ce chapitre, une première partie est dédiée à l'explication de notre méthode (Sections 3.1 et 3.2). Tout au long cette explication, nous fournissons des exemples basés sur le jeu de données $\mathcal{D}_4$ provenant de Twitter (voir Section 2.4.1) : étant donné son étendue temporelle d'un mois, il présente des avantages pour l'illustration des contextes grâce à la présence de rythmes circadien et hebdomadaire. Dans une seconde partie, nous appliquons notre méthode à l'analyse de la communication politique sur Twitter, avec les jeux de données $\mathcal{D}_4$ et $\mathcal{D}_5$, et à la détection d'anomalies dans du trafic IP avec le jeu de données $\mathcal{D}_1$ (Sections 3.3 et 3.4). Afin d'illustrer au mieux les possibilités offertes par notre méthode nous considérons exceptionnellement que les liens sont orientés dans les interactions temporelles $\mathcal{D}_4$ et $\mathcal{D}_5$. En Section 3.5, nous présentons quelques perspectives de recherche. Une fois le principe de notre méthode assimilé, son application aux flots de liens et à des propriétés plus complexes se fait sans difficultés, et ce sera l'objectif de la Section 3.6. Finalement, nous concluons le chapitre en Section 3.7.

3.1 Formalisme des cubes de données

Nous illustrons le formalisme des cubes de données avec l'ensemble de retweets du jeu de données Politoscope $\mathcal{D}_4$. Une interaction $(t, s, a) \in \mathcal{D}_4$ indique que l'utilisateur s, appelé diffuseur, a retweeté un tweet posté par l'utilisateur a, appelé auteur, à l'instant t. Dans cette section, nous définissons formellement les cubes de données ainsi que les opérations qu'il est possible de réaliser pour manipuler les interactions.

3.1.1 Définition d'un cube de données

Un cube de données est un terme général utilisé pour faire référence à un tableau de valeurs multidimensionnel (tenseur) [72]. Étant donné n dimensions caractérisées par n ensembles $X_1, ..., X_n$, un cube de données de dimension n est noté $\mathcal{C}_n(X, f)$ où $X = X_1 \times ... \times X_n$ est le produit cartésien des n ensembles et f une propriété qui associe chaque n-uplet à une valeur dans l'ensemble $\mathbb{R}$ (ou $\mathbb{N}$) :

$$\begin{array}{rccc} f: & X & \longrightarrow & \mathbb{R} \\ (x_1, ..., x_n) & & \longmapsto & f(x_1, ..., x_n) \end{array}$$

Dans ce qui suit, les n-uplets sont également appelées cellules du cube et notés x tel que $x = (x_1, ..., x_n) \in X$.

Les dimensions représentent les ensembles d'entités que l'on souhaite étudier. Dans un premier temps, nous pouvons considérer trois dimensions : les diffuseurs, notés S, les auteurs, notés A, et le temps, noté T. De plus, nous pouvons organiser les éléments d'une dimension en sous-dimensions. Par exemple, la dimension temporelle peut être organisée en fonction de la granularité temporelle. Dans notre cas, nous la divisons en deux sous-dimensions, les jours, notés D, et les heures du jour, notées H, tel que $t = (d, h)$ indique l'heure h du jour d, avec $(d, h) \in D \times H$. Alors que l'ensemble des jours D dépend de l'étendue du jeu données, H est l'ensemble des heures du jour tel que $H = \{0, \cdots, 23\}$.

La propriété est une mesure numérique qui fournit les quantités en fonction desquelles nous souhaitons analyser les entités. En tant que première approche, nous considérons la quantité d'interactions, notée v. Elle donne le nombre de retweets pour toute combinaison des quatre dimensions. Dans le cube $\mathcal{C}_4(D \times H \times S \times A, v)$, v donne le nombre de fois que s a retweeté a durant l'heure h du jour d (voir Figure 3.1) :

$$v: \ D \times H \times S \times A \longrightarrow \mathbb{N}\,.$$

Dans la suite, nous appelons le cube $\mathcal{C}_4(D \times H \times S \times A, v)$, cube de base, et le notons $\mathcal{C}_{base}$. À partir de ce cube, il est possible d'extraire plusieurs types d'information à l'aide de diverses opérations. Nous en discutons dans la section suivante.

3.1.2 Opérations sur le cube de données

Nous pouvons explorer les interactions à travers trois opérations appelées agrégation, expansion et filtrage.

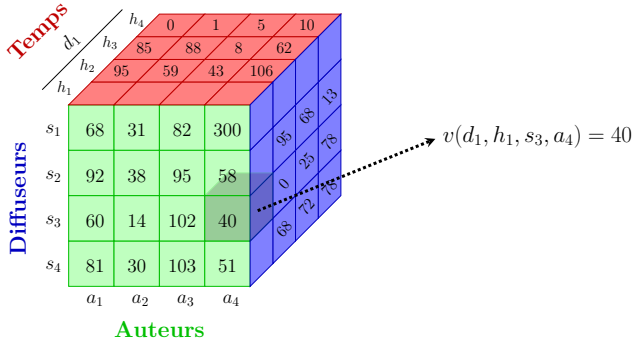


FIGURE 3.1 – **Cube** $\mathcal{C}_4(D \times H \times S \times A, v)$ – Le cube contient des informations locales sur la quantité d'interactions v. Par exemple, la cellule grise indique que s_3 a retweeté a_4 40 fois le jour d_1 à l'heure h_1.

L'agrégation est l'opération qui consiste à considérer les interactions à un niveau plus global que celui de départ. Avec le cube de données $\mathcal{C}_n(X, f)$, l'opération d'agrégation sur la dimension X_i mène au cube de données de dimension $n-1$, $\mathcal{C}_{n-1}(X', f)$ où $X' = X_1 \times ... \times X_{i-1} \times X_{i+1} \times ... \times X_n$. Formellement, une dimension X_i est agrégée en sommant les valeurs prises par la propriété sur tous les éléments $x_i \in X_i$. Nous indiquons par un $\cdot$ la dimension agrégée par rapport à f. Ainsi, $\mathcal{C}_{n-1}(X', f)$ est constitué de cellules de dimension $n-1$ notées $x' = (x_1, ..., x_{i-1}, \cdot, x_{i+1}, ..., x_n) \in X'$ où

$$f(x') = \sum_{x_i \in X_i} f(x) \quad .$$

Par exemple, on peut agréger la dimension des heures du jour de telle sorte que

$$v(d, \cdot, s, a) = \sum_{h \in H} v(d, h, s, a)$$

donne le nombre total de fois que s a retweeté a pendant le jour d.

Nous pouvons également agréger les interactions selon une partition P_i de la dimension X_i. Étant donné un cube de données $\mathcal{C}_n(X, f)$, agréger sur P_i conduit à un cube de données $\mathcal{C}_n(X', f)$ avec $X' = X_1 \times ... \times X_{i-1} \times P_i \times X_{i+1} \times ... \times X_n$. Ce cube est constitué de cellules à n dimensions notées $x' = (x_1, ..., x_{i-1}, C_k, x_{i+1}, ..., x_n) \in X'$, avec $C_k \in P_i$, tel que

$$f(x') = \sum_{x_i \in C_k} f(x).$$

Par exemple, on peut agréger selon la partition des heures $P_H = \{H_N, H_D\}$, où H_N est l'ensemble des heures de la nuit et H_D l'ensemble des heures du jour tel que

$$v(d, H_N, s, a) = \sum_{h \in H_N} v(d, h, s, a)$$

dans $\mathcal{C}_4(D \times P_H \times S \times A)$, donne le nombre total de fois que s a retweeté a pendant les heures de la nuit le jour d.

L'expansion est l'opération inverse : elle consiste à considérer des informations à un niveau plus local en introduisant des dimensions supplémentaires. Avec un cube de données $\mathcal{C}_n(X, f)$, l'opération d'expansion sur la dimension X_{n+1} conduit à un cube de données de dimension $n+1$, $\mathcal{C}_{n+1}(X', f)$ où $X' = X \times X_{n+1}$.

Finalement, *le filtrage* est l'opération qui consiste à se concentrer sur un sous-ensemble spécifique des interactions. Ainsi, filtrer le cube $\mathcal{C}_n(X, f)$ conduit à un sous-cube $\mathcal{C}_n(X', f)$ où $X' = X'_1 \times ... \times X'_n$ avec $X'_1 \subseteq X_1, ..., X'_n \subseteq X_n$.

Il est également possible de combiner plusieurs opérations. Par exemple, on peut agréger le cube de données sur la partition des heures $\mathcal{C}_4(D \times P_H \times S \times A, v)$, puis le filtrer de façon à se concentrer sur les diffuseurs qui retweetent anormalement les auteurs pendant la nuit. Il est important de faire la distinction entre le cube de données résultant $\mathcal{C}_4(D \times \{H_N\} \times S \times A, v)$ et le cube de données $\mathcal{C}_4(D \times H_N \times S \times A, v)$: dans le premier cas, une cellule (d, H_N, s, a) indique le nombre total de fois où s a retweeté a durant la nuit du jour d, alors que dans le second cas, une cellule (d, h, s, a) donne le nombre de fois où s a retweeté a au cours de l'heure (d, h) où $h \in H_N$ est heure de la nuit.

3.1.3 Ensemble de cubes de données

La Figure 3.2 montre un ensemble de cubes de données pouvant être obtenus avec les dimensions : diffuseurs, auteurs et temps. Il montre également comment naviguer d'un cube à l'autre grâce aux opérations précédemment définies. Quand on s'éloigne du cube de base, on accède à des informations de plus en plus agrégées. Par exemple [1] :
- dans le cube de base $\mathcal{C}_4(D \times H \times S \times A, v)$, la valeur $v(d_1, h_4, s_1, a_2) = 9$ associée à la cellule $x = (d_1, h_4, s_1, a_2)$ signifie que s_1 a retweeté a_2 9 fois le jour d_1 durant l'heure h_4 ;
- dans le cube $(temps, auteurs)$, $\mathcal{C}_3(D \times H \times A, v)$, la valeur $v(d_1, h_4, \cdot, a_2) = 1\,288$ associée à la cellule $x = (d_1, h_4, a_2)$ signifie que a_2 a été retweeté 1 288 fois le jour d_1 durant l'heure h_4 ;
- dans le cube $(auteurs)$, $\mathcal{C}_1(A, v)$, la valeur $v(\cdot, \cdot, \cdot, a_2) = 29\,362$ associée à la cellule $x = a_2$ signifie que a_2 a été retweeté 29 292 fois ;
- dans le cube $\mathcal{C}_0(\cdot, v)$, la valeur $v(\cdot, \cdot, \cdot, \cdot) = 1\,142\,004$ associée à la cellule $x = (\cdot, \cdot, \cdot, \cdot)$ signifie que le nombre total de retweets est égal à 1 142 004.

En haut à gauche, on se concentre sur deux diffuseurs, notés s_1 et s_2, en filtrant le cube de base. Dans le cube résultant $\mathcal{C}_4(D \times H \times \{s_1, s_2\} \times A, v)$, la valeur $v(d_1, h_4, s_1, a_{10}) = 258$ indique que s_1 a retweeté a_{10} 258 fois le jour d_1 durant l'heure h_4. Ensuite, on agrège la dimension temporelle. Dans le cube résultant $\mathcal{C}_2(\{s_1, s_2\} \times A, v)$, la valeur $v(s_2, a_8) = 3\,000$ indique que s_2 a retweeté a_8 3 000 fois sur l'ensemble du corpus. Ces opérations permettent ainsi d'étudier les interactions globales de s_1 et s_2 avec l'ensemble des auteurs.

Enfin, en haut à droite, on agrège le cube de base sur la partition $P_A = \{C_1, C_2, \cdots\}$, correspondant aux communautés des auteurs. Dans le cube résultant $\mathcal{C}_4(D \times H \times S \times P_A, v)$, la valeur $v(d_3, h_{10}, s_3, C_5) = 50$ indique que s_3 a retweeté des auteurs de la communauté

1. On considère ici une granularité temporelle en heures tel que $T = D \times H$.

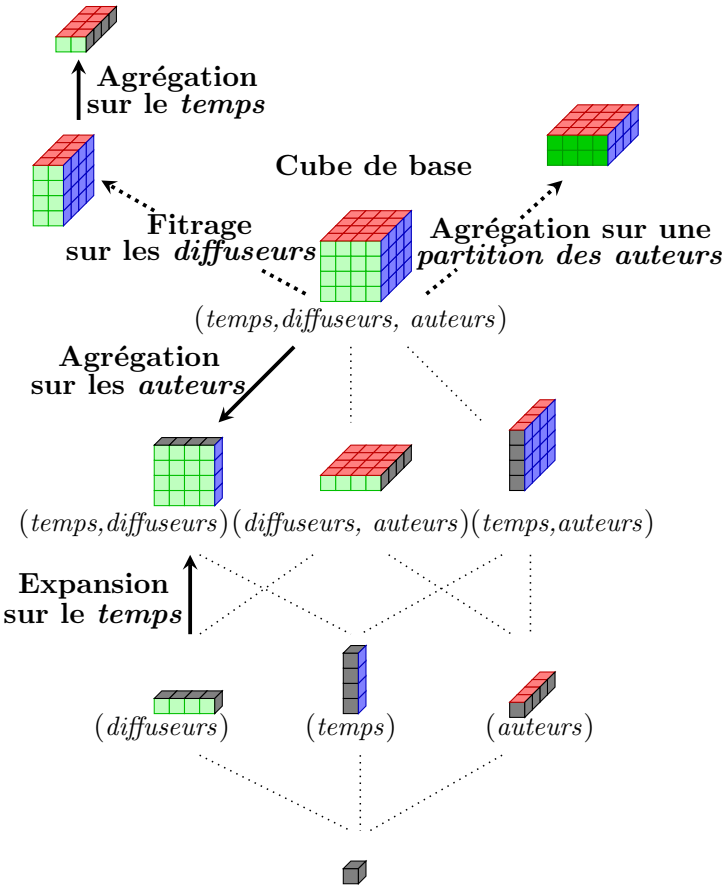


FIGURE 3.2 – **Ensemble de cubes de données obtenus en considérant les dimensions : diffuseurs, auteurs et temps.**

C_5 au total 50 fois le jour d_3 durant l'heure h_{10}. Cette opération peut permettre d'étudier les tendances politiques.

3.2 Construction de contextes

Nous nous attachons maintenant à trouver des cellules anormales dans un cube de données, c'est-à-dire des entités $x \in X$ pour lesquelles l'observation $f(x)$ est inattendue. Pour ce faire, nous construisons des contextes plus ou moins élaborés en jouant avec les valeurs observées, les valeurs attendues et les valeurs de déviation considérées.

3.2.1 Valeurs observées

La première étape consiste à choisir le cube de données correspondant au type de comportements irréguliers que l'on recherche. Ce cube, noté $\mathcal{C}_{obs}$, constitue l'ensemble des entités et des valeurs observées.

Par exemple, on peut vouloir trouver des auteurs anormaux à différentes heures. Pour cela, on choisit le cube de données agrégé sur les diffuseurs tel que $\mathcal{C}_{obs} = \mathcal{C}_3(D \times H \times A, v)$. On peut également s'intéresser aux auteurs anormaux uniquement pendant les heures de la nuit. Dans ce cas, on considère le cube de données agrégé et filtré $\mathcal{C}_{obs} = \mathcal{C}_3(D \times H_N \times A, v)$.

Dans le premier cas, on compare toutes les entités du même type, $(d, h, a) \in D \times H \times A$: on est dans un *contexte global*. Au contraire, lorsque l'on considère uniquement un sous-ensemble de toutes les entités, comme dans le deuxième exemple avec $(d, h, a) \in D \times H_N \times A$, on est dans un *contexte local*.

3.2.2 Valeurs attendues

Une fois que l'ensemble des valeurs observées est fixé, nous construisons un modèle de comportement attendu basé sur la combinaison de plusieurs autres cubes de données notés $\mathcal{C}_m(X', f)$ et appelés *cubes de données de comparaison*. Pour que le contexte soit pertinent, ceux-ci doivent provenir de l'agrégation de $\mathcal{C}_{obs} = \mathcal{C}_n(X, f)$ sur une ou plusieurs dimensions. Par conséquent, $n > m$ et $X = X' \times Y$ où Y est le produit cartésien des dimensions agrégées. Dans ce qui suit, nous construisons trois types de valeurs attendues menant au contexte basique, aux contextes agrégatifs et multi-agrégatifs.

Contexte basique

Lors de la recherche de cellules anormales dans un cube de données $\mathcal{C}_n(X, f)$, le contexte le plus simple que l'on puisse choisir est celui dans lequel la valeur attendue est une constante, identique pour chaque cellule. Nous l'appelons le *contexte basique*. Le modèle de comportement attendu correspond à la situation dans laquelle les interactions sont uniformément réparties sur les cellules. Dans ce cas, le cube de données de comparaison est $\mathcal{C}_0(\cdot, f)$ et la valeur attendue est le nombre moyen d'interactions par cellule :

$$f_{exp}(x) = \frac{f(\cdot)}{|X|} .$$

Par exemple, dans le cube de données $\mathcal{C}_3(D \times H \times A, v)$*, une cellule anormale* $x^* = (d^*, h^*, a^*)$ *indique que pendant l'heure* h^* *du jour* d^**, l'auteur* a^* *a été retweeté un nombre anormal de fois par rapport au nombre moyen de fois qu'un auteur est retweeté durant une heure,* $v_{exp}(d, h, a) = \frac{v(\cdot,\cdot,\cdot,\cdot)}{|D \times H \times A|}$.

Contextes agrégés

Pour trouver des anomalies plus subtiles et plus locales, les valeurs attendues doivent être spécifiques à chaque cellule. Le processus est le même que dans le contexte basique, excepté que le cube de comparaison considéré $\mathcal{C}_m(X', f)$ n'est pas agrégé sur toutes les dimensions de X : $X = X' \times Y$ avec $Y \neq X$ et

$$f_{exp}(x', y) = f_{exp}(x) = \frac{f(x')}{|Y|} .$$

Définie comme telle, la valeur attendue est la valeur que l'on observerait si toutes les interactions dans X' étaient réparties de manière homogène sur les dimensions Y. Nous appelons ces contextes, *contextes agrégés.*

Par exemple, dans le cube de données $\mathcal{C}_3(D \times H \times A, v)$ *par rapport au cube de données* $\mathcal{C}_2(D \times H, v)$*, les valeurs attendues sont*

$$v_{exp}(d, h, a) = \frac{v(d, h, \cdot, \cdot)}{|A|} ,$$

et une cellule anormale $x^* = (d^*, h^*, a^*)$ *indique une déviation entre le nombre de retweets reçus par* a^* *pendant l'heure* (d^*, h^*) *et celui qui aurait dû être observé si tous les auteurs avaient reçu le même nombre de retweets pendant l'heure* (d^*, h^*) .

Contextes multi-agrégatifs

Les contextes agrégés supposent que les interactions sont réparties de manière homogène sur les dimensions Y. Il est possible de créer des contextes qui différencient la répartition des interactions en fonction de l'activité de chaque cellule. Nous les appelons *contextes multi-agrégatifs.* Contrairement aux deux autres, ils nécessitent plusieurs cubes de données de comparaison. Il n'existe pas de formules génériques : le nombre et les types de cubes de comparaison, ainsi que les valeurs attendues, dépendent de l'application considérée.

Si nous reprenons l'exemple précédent, nous pouvons considérer les valeurs attendues suivantes :

$$v_{exp}(d, h, a) = v(d, h, \cdot, \cdot) \times \frac{v(\cdot, \cdot, \cdot, a)}{v(\cdot, \cdot, \cdot, \cdot)} .$$

De cette façon, on s'attend à ce que le nombre de retweets pendant (d, h) *soit réparti entre les auteurs proportionnellement à leur activité moyenne sur tout le corpus. Nous pouvons également ajouter des informations sur l'activité des auteurs à des heures spécifiques et*

considérer les cubes $\mathcal{C}_2(D \times H, v)$, $\mathcal{C}_2(H \times A, v)$ *et* $\mathcal{C}_1(H, v)$, *tel que*

$$v_{exp}(d, h, a) = v(d, h, \cdot, \cdot) \times \frac{v(\cdot, h, \cdot, a)}{v(\cdot, h, \cdot, \cdot)} .$$

Dans ce contexte, une cellule anormale $x^* = (d^*, h^*, a^*)$ *indique une déviation entre le nombre de retweets reçus par* a^* *pendant l'heure* h^* *du jour* d^* *et celui qui aurait dû être observé si* a^* *avait été retweeté de la façon dont il l'est habituellement pendant l'heure* h^* *les autres jours.*

Chacun de ces contextes peut être global ou local en fonction de l'ensemble de valeurs observées choisi dans $\mathcal{C}_{obs}$.

3.2.3 Valeurs de déviation

Enfin, pour chaque cellule x du cube $\mathcal{C}_{obs}$, nous mesurons l'écart entre sa valeur observée $f(x)$ et sa valeur attendue $f_{exp}(x)$. Nous utilisons deux fonctions de déviation différentes : le ratio et la déviation de Poisson.

Le *ratio* entre une valeur observée et une valeur attendue est défini tel que

$$d_r(f(x), f_{exp}(x)) = \frac{f(x)}{f_{exp}(x)} .$$

Le ratio $d_r = 2$ indique que la valeur observée est deux fois plus grande que la valeur attendue. Ainsi, cette fonction de déviation ne fait pas de distinction entre $f(x) = 2$ et $f_{exp}(x) = 1$, d'une part, et $f(x) = 2000$ et $f_{exp}(x) = 1000$, d'autre part. Pour prendre en compte la significativité d'une valeur de déviation, nous introduisons la *déviation de Poisson*. En effet, dans les cas où la propriété f consiste à compter le nombre d'interactions au cours d'une période donnée, comme la quantité d'interactions v, elle peut être modélisée par un processus de Poisson d'intensité f_{exp} [60], tel que

$$\forall k \in \mathbb{N},\ \Pr(v(x) = k) = \frac{f_{exp}(x)^k e^{-f_{exp}(x)}}{k!} .$$

Dans ce cas, la déviation de Poisson d_p peut être définie comme suit. Si $f(x) \leq f_{exp}(x)$, on calcule la probabilité d'observer une valeur inférieure à $f(x)$ sachant que l'on aurait dû observer $f_{exp}(x)$ en moyenne. Cette probabilité est donnée par la fonction de répartition d'une distribution de Poisson de paramètre $f_{exp}(x)$, on la note $\Pr(\text{Poiss}(f_{exp}(x)) \leqslant f(x))$. Finalement, par symétrie, on définit d_p tel que :

$$d_p(f(x), f_{exp}(x)) = \begin{cases} \log(\Pr(\text{Poiss}(f_{exp}(x)) \leqslant f(x))) \text{ si } f(x) \leq f_{exp}(x), \\ -\log(\Pr(\text{Poiss}(f_{exp}(x)) > f(x))) \text{ si } f(x) > f_{exp}(x), \end{cases}$$

où le logarithme est calculé afin d'avoir une plage de valeurs plus étendue.

Dans les deux cas, on s'attend à ce que la plupart des valeurs observées soient similaires aux valeurs attendues correspondantes. Par conséquent, on s'attend à ce que les valeurs

attendues des cellules normales fluctuent autour d'une moyenne : $\bar{d}_r = 1$ pour le ratio et $\bar{d}_p = 0$ pour la déviation de Poisson. Les cellules anormales, au contraire, correspondent aux valeurs de déviation significativement éloignées de la moyenne. Dans la suite, nous utilisons l'hypothèse classique selon laquelle une valeur est anormale si sa distance par rapport à la moyenne dépasse trois fois l'écart type [33, 72].

3.2.4 Exemples

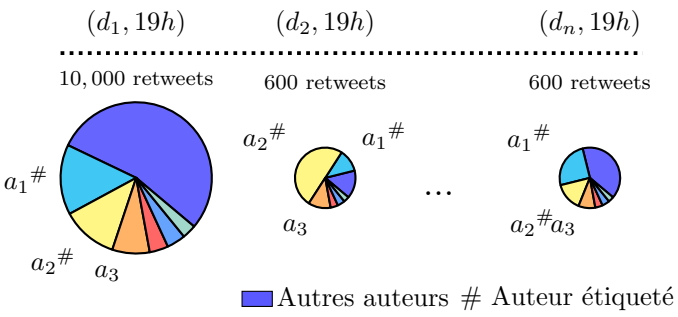


FIGURE 3.3 – **Différents contextes mènent à différentes anomalies** – Le nombre de retweets par heure est réparti entre les auteurs et représenté sous forme de diagrammes circulaires. Par exemple, l'auteur influent $a_1^{\#}$ a été retweeté 72 fois (12% de 600) à $19h$ le jour d_2.

Afin d'illustrer l'utilisation des différents contextes, nous présentons un exemple dans lequel on recherche des auteurs anormaux à différentes heures. La Figure 3.3 montre que considérer différents contextes mène à différentes anomalies. Par exemple :
- le triplet $(d_1, 19h, a_1)$ est anormal dans le contexte basique global : a_1 a été retweeté 1500 fois à $19h$ le jour d_1 (15% de 10000), ce qui est beaucoup plus que ce que l'ont été tous les autres triplets.
- Le triplet $(d_2, 19h, a_2)$ est anormal dans le contexte agrégatif global : sa proportion de retweets, égale à 50%, est supérieure à celles observées pour tous les autres triplets.
- Le triplet $(d_n, 19h, a_1)$ est anormal dans le contexte multi-agrégatif global : la déviation de l'activité de a_1 par rapport à son activité habituelle à $19h$ est supérieure aux déviations observées pour tous les autres triplets.
- Le triplet $(d_2, 19h, a_3)$ est anormal dans le contexte agrégatif local : sa proportion de retweets est supérieure à celle des autres triplets (d, h, a) dans lesquels a n'est pas un auteur étiqueté #.

Comme le montre cet exemple et comme nous le verrons en pratique dans les sections suivantes, notre approche, qui consiste à combiner des cubes de données pour créer différents contextes, conduit à de nombreux types d'anomalies, ce qui nous permet d'analyser les interactions temporelles sous différentes perspectives. Une implémentation de notre méthode est disponible à l'adresse `https://github.com/Lamarche-Perrin/data.cube`.

3.3 Application à la communication politique sur Twitter

Nous appliquons notre méthode à l'étude de la communication politique sur Twitter. Dans cette section, nous présentons une étude de cas qui, basée sur les évènements trouvés dans la dimension temporelle, recherche les causes possibles de leur émergence en

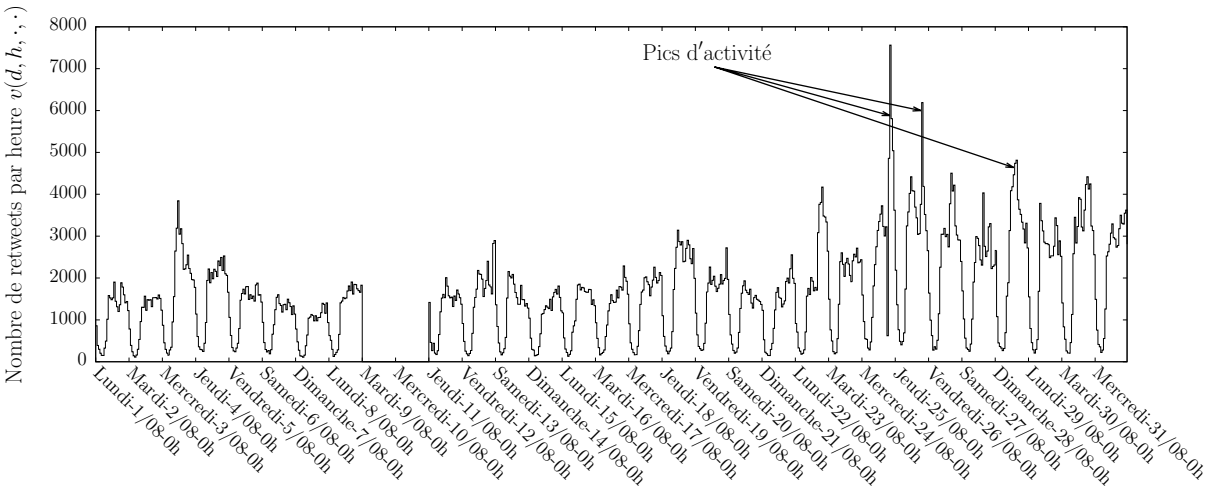


FIGURE 3.4 – **Nombre de retweets par heure sur le mois d'août 2016** – Les trois pics d'activité correspondent à des évènements médiatiques : 24/08 : interview de N. Sarkozy au journal télévisé ; 25/08 : meeting politique de N. Sarkozy ; 28/08 : meeting politique de A. Juppé.

explorant d'autres dimensions. Nous commençons par considérer les trois dimensions : diffuseurs, auteurs et temps. Ensuite, nous incluons une information sémantique en ajoutant la dimension hashtag.

3.3.1 Évènements

Nous définissons un évènement $e = ((d_1^*, h_1^*), \cdots, (d_n^*, h_n^*)) \in \mathcal{E}$ comme un ensemble d'heures anormales consécutives. Par commodité, on le note $e = (d^*, h_1^*\text{-}h_n^*)$ lorsque toutes les heures s'étendent sur le même jour d^*.

La Figure 3.4 montre l'évolution du nombre de retweets par heure. On peut distinguer trois comportements distincts :
- les heures de la nuit, caractérisées par un nombre de retweets fluctuant autour de 350,
- les heures du jour du 1er au 23 août, caractérisées par un nombre plus élevé de retweets oscillant autour de 1 700,
- les heures du jour du 24 au 31 août, caractérisées par une augmentation globale du nombre de retweets oscillant autour de 2 900.

Contexte basique

Tout d'abord, nous recherchons des évènements dans le contexte basique. Les ensembles d'entités et les valeurs observées sont fournis par le cube de données $\mathcal{C}_2(D \times H, v)$. Les valeurs attendues sont définies de telle sorte que

$$v^b_{exp}(d, h) = \frac{v(\cdot, \cdot, \cdot, \cdot)}{|D \times H|} .$$

La Figure 3.5.a montre la distribution des valeurs de déviation en considérant une déviation basée sur le ratio. On trouve sept heures anormales menant aux trois évènements suivants :

$$\mathcal{E} = \{(24, 20h\text{-}22h), (25, 19h), (28, 14h\text{-}15h)\} .$$

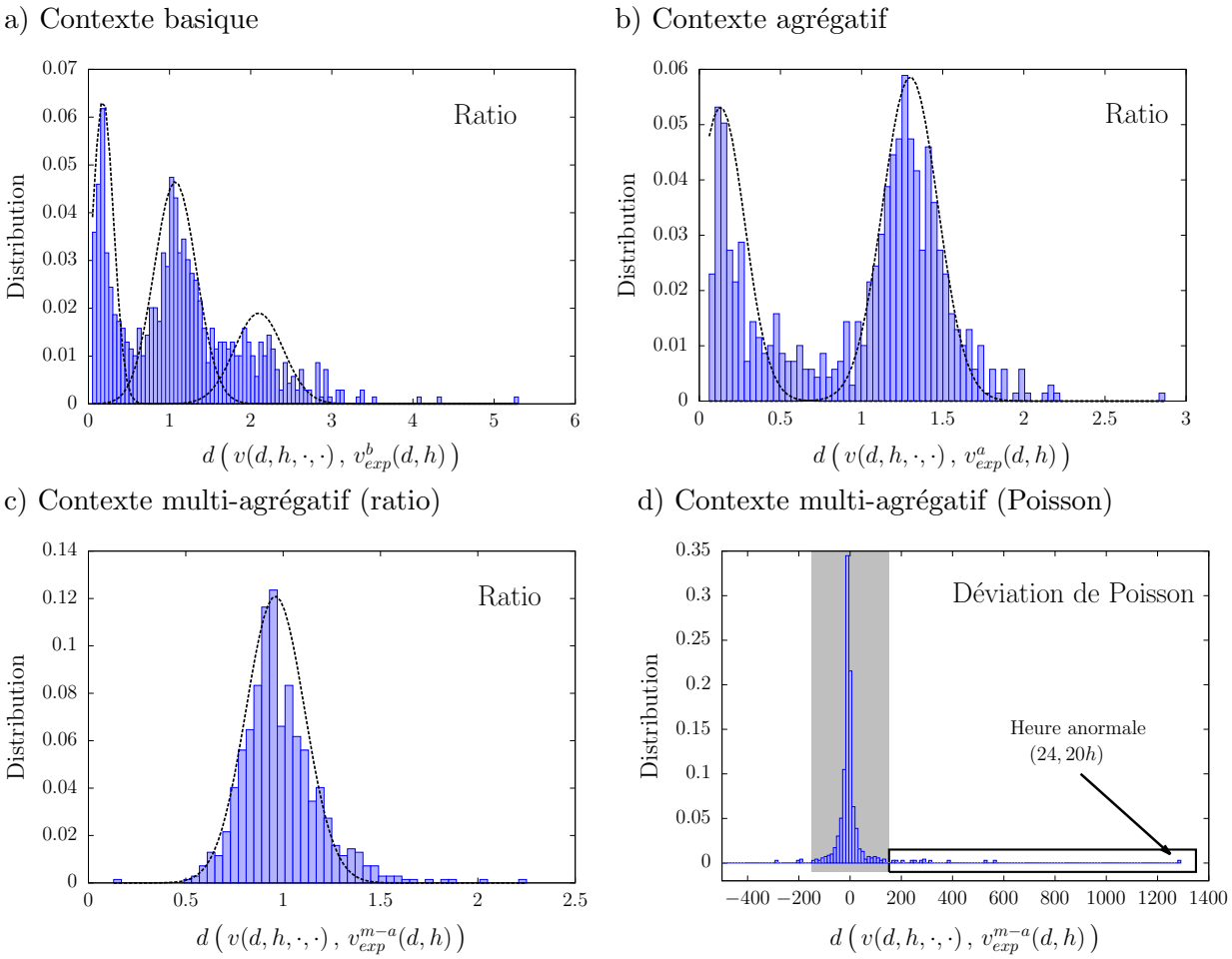


FIGURE 3.5 – **Valeurs de déviation des heures dans les contextes basique, agrégatif et multi-agrégatif** – a) Les distributions correspondent à trois comportements distincts : les heures nocturnes ($\bar{d}^1_r = 0.2$), les heures du jour du 1^{er} août au 24 août ($\bar{d}^2_r = 1,1$) et les heures du jour à partir du 24 août ($\bar{d}^3_r = 2.1$). b) On observe deux comportements : les heures nocturnes ($\bar{d}^1_r = 0,13$) et les heures du jour ($\bar{d}^2_r = 1,3$). c) Le contexte multi-agrégatif normalise les comportements. d) La plupart des valeurs de déviation sont centrées sur $\bar{d}_p = 0$ (zone grise), ce qui signifie qu'elles sont susceptibles d'être générées par un processus de Poisson d'intensité $v^{m-a}_{exp}(d,h)$.

On remarque que ces heures correspondent aux trois pics d'activité de la Figure 3.4. Par conséquent, ce contexte ne met pas en évidence les anomalies locales mais uniquement les anomalies globales s'écartant du reste des observations. Ainsi, les anomalies détectées dans ce contexte sont biaisées par les rythmes circadien et hebdomadaire ce qui nous empêche de détecter des heures nocturnes anormales ou des heures anormales appartenant à la première partie du mois.

Contexte agrégatif

Pour prendre en compte l'augmentation globale du nombre de retweets au cours du mois, nous devons utiliser un contexte agrégatif dans lequel les valeurs attendues intègrent l'activité globale de la journée. Cette dernière est fournie par le cube de données $\mathcal{C}_1(D, v)$

de telle sorte que

$$v^a_{exp}(d,h) = \frac{v(d,\cdot,\cdot,\cdot)}{|H|} .$$

De cette façon, les valeurs de déviation sont indépendantes des variations quotidiennes du nombre de retweets. C'est ce que nous observons sur la Figure 3.5.b. On trouve 10 heures anormales. Parmi celles-ci, six heures font partie de la première période du mois : le 3 à $11h$, le 12 à $23h$, le 21 à $21h$, et le 22 de $17h$ à $20h$. Cependant, les anomalies sont encore biaisées par les rythmes circadiens qui nous empêchent de détecter des heures nocturnes anormales.

Contexte multi-agrégatif

Pour résoudre ce problème, nous utilisons un contexte multi-agrégatif dans lequel on ajoute les informations agrégées relatives à l'activité typique par heure, fournies par les cubes de données $\mathcal{C}_1(H,v)$ et $\mathcal{C}_0(\cdot,v)$:

$$v^{m-a}_{exp}(d,h) = v(d,\cdot,\cdot,\cdot) \times \frac{v(\cdot,h,\cdot,\cdot)}{v(\cdot,\cdot,\cdot,\cdot)} .$$

Les Figures 3.5.c et 3.5.d montre les distributions des valeurs de déviations en considérant le ratio et la déviation de Poisson. En prenant en compte la significativité des écarts avec la déviation de Poisson, on trouve 40 heures anormales. Parmi celles-ci, plusieurs sont adjacentes, ce qui conduit à 17 évènements distincts (voir le Tableau 3.1).

Par exemple, l'heure $(11, 0h)$ est anormale. Cela signifie qu'en moyenne, à $0h$, on s'attend à observer $v(\cdot, 0h, \cdot, \cdot)/v(\cdot,\cdot,\cdot,\cdot) = 3.16\%$ du nombre total de retweets de la journée. Par conséquent, durant l'heure $(11, 0h)$, on s'attend à observer $v(11,\cdot,\cdot,\cdot) \times 3.16\% = 909$ retweets. Cependant, on observe $1,418$ retweets dans $\mathcal{C}_2(D \times H, v)$. Cet écart par rapport à la valeur attendue est beaucoup plus important que ceux observés pour la plupart des heures $(d,h) \in D \times H$. Par conséquent, $(11, 0h)$ est une heure anormale dans ce contexte multi-agrégatif.

Dans le Tableau 3.1, on remarque plusieurs heures ayant une activité généralement faible comme les heures de la nuit par exemple. Ce dernier résultat montre que l'utilisation de contextes plus sophistiqués conduit à des anomalies plus subtiles.

3.3.2 Auteurs anormaux pendant les évènements

Nous cherchons maintenant à déterminer si un évènement est dû à des auteurs spécifiques qui ont été anormalement retweetés ou, au contraire, résulte d'un phénomène plus global dans lequel on observe une augmentation globale de l'activité.

Nous utilisons un contexte local et multi-agrégatif. Les valeurs observées sont fournies par le cube de données filtré et agrégé $\mathcal{C}_3(\{e\} \times A, v)$, où $e \in \mathcal{E}$ est un évènement anormal. Une cellule (e,a) dans ce cube indique le nombre total de fois que l'auteur a a été retweeté au cours de l'évènement e. De cette façon, nous nous concentrons sur la manière dont les

Évènements	Auteurs anormaux	Évènements médiatiques
$(3, 10h$ - $13h)$	plusieurs	Intervention de la police dans une église
$(11, 0h)$	marseille	Incendie à Marseille
$(11, 3h)$	FrancoisFillon	Inconnu
$((12, 22h), \cdots, (13, 1h))$	fhollande	Victoire olympique de la France
$(13, 9h)$	aucun	Inconnu
$(19, 22h)$	aucun	Victoire olympique de la France
$(21, 21h)$	aucun	Victoire olympique de la France
$(22, 16h$ - $22h)$	plusieurs	N. Sarkozy annonce sa candidature aux présidentielles
$(23, 7h$ - $8h)$	aucun	Inconnu
$(24, 20h$ - $22h)$	plusieurs	Interview de N. Sarkozy au journal télévisé
$(25, 19h)$	NicolasSarkozy	Meeting politique de N. Sarkozy
$(26, 15h$ - $18h)$	plusieurs	Conseil d'état sur le port du burkini
$(27, 15h)$	alainjuppe	Meeting politique de A. Juppé
$(28, 0h)$	plusieurs	Interview de N. Kosciusko-Morizet dans une émission télévisée
$(28, 13h$ - $15h)$	JLMelenchon	Meeting politique de J-L. Mélenchon
$(29, 7h$ - $9h)$	NicolasSarkozy	Interview de N. Sarkozy dans une émission radio
$(30, 17h$ - $18h)$	aucun	Démission de E. Macron du gouvernement

TABLE 3.1 – **Liste des évènements et des auteurs anormaux avec leurs évènements médiatiques associés.**

interactions sont organisées entre les auteurs au sein de chaque évènement.

Nous procédons de la même manière pour obtenir les valeurs attendues. Au lieu de se restreindre à l'ensemble des auteurs lors de l'évènement e, on considère l'ensemble des auteurs au cours de chacune des périodes horaires correspondant à e sur l'ensemble des jours. On note cet ensemble d'heures $H_e = \{h^* \in H \,|\, (d^*, h^*) \in e\}$. Cela revient à considérer le cube de données $\mathcal{C}_3(D \times P_H \times A, v)$, agrégé sur la partition de H, $P_H = \{H_e\}$. Les opérations effectuées pour passer du cube d'origine $\mathcal{C}_3(D \times H \times A, v)$ au cube de données $\mathcal{C}_3(D \times \{H_e\} \times A, v)$ sont représentées sur la Figure 3.7.

Enfin, les valeurs attendues sont définies en utilisant les cubes de données de comparaison $\mathcal{C}_2(\{H_e\} \times A, v)$ et $\mathcal{C}_1(\{H_e\}, v)$, obtenus par agrégation de $\mathcal{C}_3(D \times \{H_e\} \times A, v)$ et du cube de données $\mathcal{C}_2(\{e\}, v)$, obtenu par agrégation et filtrage de $\mathcal{C}_3(D \times \{H_e\} \times A, v)$:

$$v_{exp}(e, a) = v(e, \cdot, \cdot) \times \frac{v(\cdot, H_e, \cdot, a)}{v(\cdot, H_e, \cdot, \cdot)} \,,$$

où $v(e, \cdot, \cdot) = \sum_{(d^*, h^*) \in e} v(d^*, h^*, \cdot, \cdot)$, est le nombre des retweets observés pendant e ; $v(\cdot, H_e, \cdot, a)$ est le nombre total de retweets reçus par l'auteur a pendant les heures de H_e ; et $v(\cdot, H_e, \cdot, \cdot)$ est le nombre total de retweets observés pendant H_e.

Selon ce contexte, un couple $(e, a^*) \in \{e\} \times A$ est anormal lorsqu'il existe un écart significatif entre le nombre de retweets reçus par a pendant e, et le nombre de retweets que a est censé recevoir, en moyenne, au cours de la période correspondante les autres jours. Dans la suite, on décrit les situations rencontrées à travers trois exemples spécifiques.

Un auteur principal

La Figure 3.6.a montre la distribution des valeurs de déviation pour l'évènement $e = (29, 7h\text{-}9h)$. La plupart des observations $d \in \mathcal{D}$ suivent une distribution gaussienne centrée sur $\bar{d_p} = 0$. On détecte 14 valeurs anormales. Parmi celles-ci, celle correspondant au triplet $(29, 7h\text{-}9h, \text{NicolasSarkozy})$ s'écarte considérablement des autres. En effet, dans le contexte considéré, on s'attend à ce que NicolasSarkozy constitue

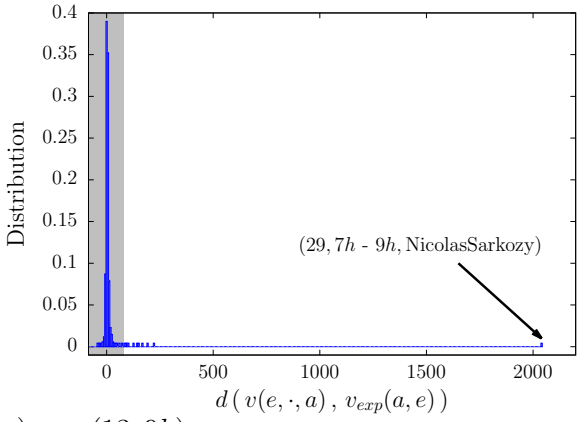
$$\frac{v(\cdot, \{7h, 8h, 9h\}, \cdot, \text{NicolasSarkozy})}{v(\cdot, \{7h, 8h, 9h\}, \cdot, \cdot)} = 2,2\%$$

des retweets observés de $7h$ à $9h$. Ainsi, le 29 août on s'attend à ce qu'il soit retweeté $v((29, 7h\text{-}9h), \cdot, \cdot) \times 2.2\% = 194$ fois de $7h$ à $9h$. Or, il a été retweeté 1644 fois, ce qui explique sa valeur de déviation élevée.
Le Tableau 3.1 répertorie les évènements ayant une distribution similaire. Dans la plupart des cas, on observe que l'évènement médiatique correspondant est centré sur l'auteur principal, comme dans le cas de meetings politiques par exemple.

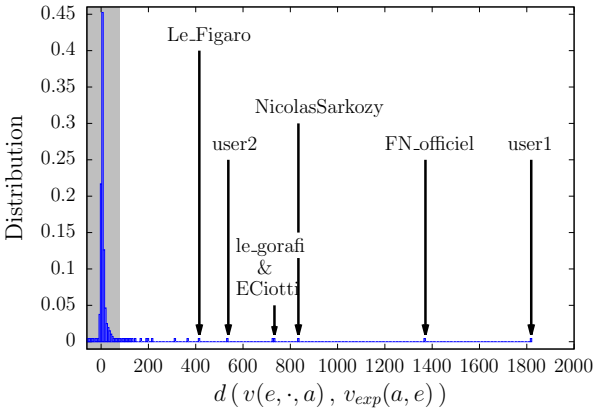
Plusieurs auteurs principaux

La Figure 3.6.b montre la distribution des valeurs de déviation pour l'évènement $e = (22, 16h\text{-}22h)$. Une fois de plus, la plupart des observations $d \in \mathcal{D}$ suivent une distribution

a) $e = (29, 7h \text{ - } 9h)$



b) $e = (22, 16h \text{ - } 22h)$



c) $e = (13, 9h)$

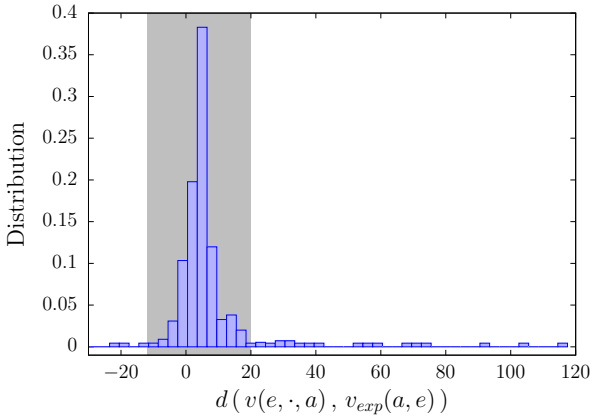


FIGURE 3.6 – **Auteurs anormaux pendant les évènements** – a) On remarque que *NicolasSarkozy* est probablement responsable de l'évènement $e = (29, 7h \text{ - } 9h)$ car son activité dévie considérablement de son activité habituelle. b) La cause de cet évènement est multiple, on remarque que plusieurs auteurs, principalement des particuliers et des politiciens de droite, sont plus retweetés qu'habituellement. c) La distribution est plus homogène faisant de cet évènement un phénomène plus global.

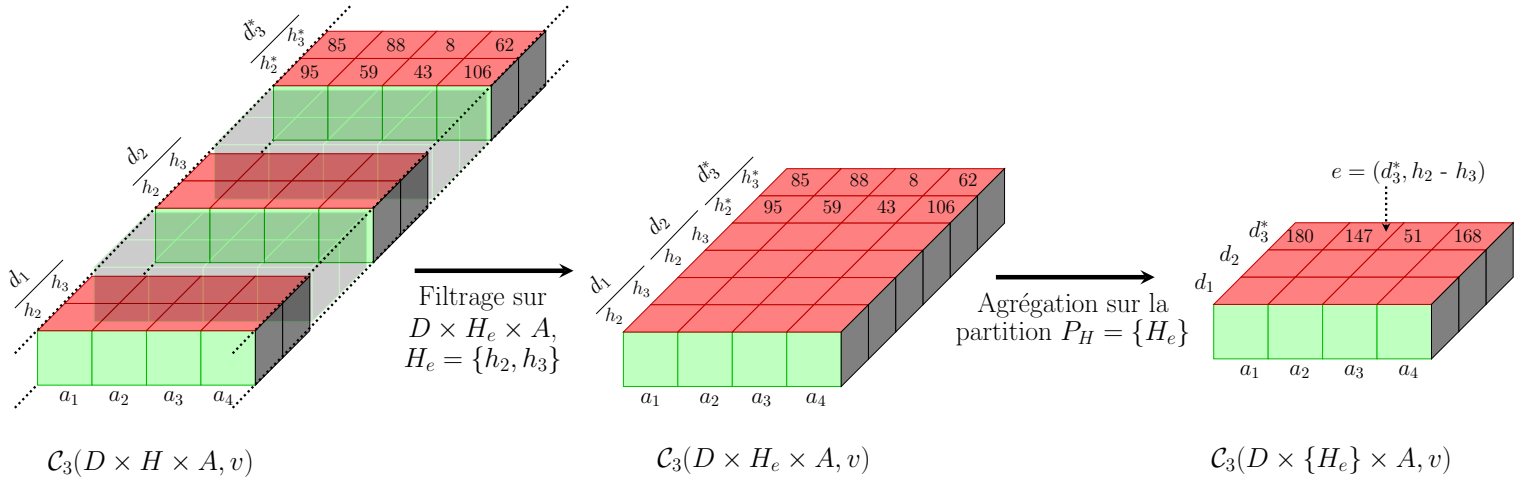


FIGURE 3.7 – **Contexte multi-agrégatif local permettant de se focaliser sur les auteurs lors des évènements** – Pour rechercher les causes possibles de l'émergence de l'évènement $e = (d_3^*, h_2^*\text{-}h_3^*)$, on caractérise les comportements habituels des auteurs pendant les périodes correspondantes sur l'ensemble des jours. Par exemple, $v(e, a_1) = 180$ indique que l'auteur a_1 a été retweeté 180 fois de h_2 à h_3 le jour d_3.

gaussienne centrée sur $\bar{d_p} = 0$. On détecte 42 anomalies. Parmi celles-ci, plusieurs valeurs s'écartent significativement des autres indiquant plusieurs auteurs principaux.
Ainsi, ces évènements ne sont pas dus à un seul auteur populaire, mais à plusieurs auteurs considérablement retweetés. Par exemple, ils peuvent être la conséquence de la réaction de ces auteurs à un fait extérieur envers lequel ils ont un intérêt. C'est ce que nous observons dans le Tableau 3.1 : les évènements médiatiques liés aux évènements ayant des distributions similaires sont souvent révélateurs de situations envers lesquelles les auteurs principaux réagissent sans y être directement liés. Par exemple, l'évènement du 3 août, correspondant à l'intervention de la police dans une église, et celui du 26 août, sur le port du burkini, sont des évènements médiatiques intensément repris par les politiciens de droite et d'extrême droite.

Pas d'auteurs principaux

La Figure 3.6.c affiche la distribution des valeurs de déviation pour l'évènement $e = (13, 9h)$. Contrairement aux exemples précédents, nous voyons que les valeurs sont distribuées de manière plus homogène et réparties sur une plage plus petite.
L'absence d'anomalies significatives montre que ces évènements sont des phénomènes plus globaux que les précédents : ils émergent car de nombreux auteurs sont retweetés au lieu de quelques-uns, intensément. Cela suggère qu'ils proviennent de la réaction d'une multitude d'auteurs à un fait d'actualité général. C'est le cas, par exemple, des deux victoires olympiques de la France les 19 et 21 août (voir Tableau 3.1).

Cette étude, centrée sur les auteurs, nous permet de mieux comprendre l'origine des événements : ils peuvent être dus à un seul auteur, ou plusieurs, ou aucun en particulier.

3.3.3 Diffuseurs anormaux pendant les évènements

Parmi les trois cas précédents, nous nous concentrons sur les évènements générés par un seul auteur principal. En particulier, nous cherchons à déterminer si leur émergence est due à un grand nombre de diffuseurs, ou au contraire, s'ils se manifestent uniquement à cause d'un nombre restreint de diffuseurs qui les retweetent anormalement.

Pour ce faire, nous procédons comme dans la section précédente et étudions localement les interactions dans le cube de données filtré $\mathcal{C}_3(\{e\} \times S \times \{a^*\}, v)$, où a^* est l'auteur anormal principal correspondant à l'évènement e. Une cellule (e, s, a^*) dans ce cube donne le nombre total de fois que s a retweeté a^* pendant e. De cette façon, on se concentre sur la manière dont chacun des diffuseurs retweete a^* pendant l'évènement.

Les valeurs attendues sont définies à partir du cube de données $\mathcal{C}_4(D \times \{H_e\} \times S \times \{a^*\}, v)$, à l'aide des cubes de données de comparaison $\mathcal{C}_3(\{H_e\} \times S \times \{a^*\}, v)$ et $\mathcal{C}_2(\{H_e\} \times \{a^*\}, v)$, obtenu par agrégation, et $\mathcal{C}_3(\{e\} \times S \times \{a^*\}, v)$, obtenu par agrégation et filtrage :

$$v_{exp}(e, s, a^*) = v(e, \cdot, a^*) \times \frac{v(\cdot, H_e, s, a^*)}{v(\cdot, H_e, \cdot, a^*)},$$

où $v(e, \cdot, a^*)$ est le nombre total de retweets reçus par a^* pendant e ; $v(\cdot, H_e, s, a^*)$ est le nombre total de fois que s a retweeté a durant les heures appartenant à H_e ; et $v(\cdot, H_e, \cdot, a^*)$

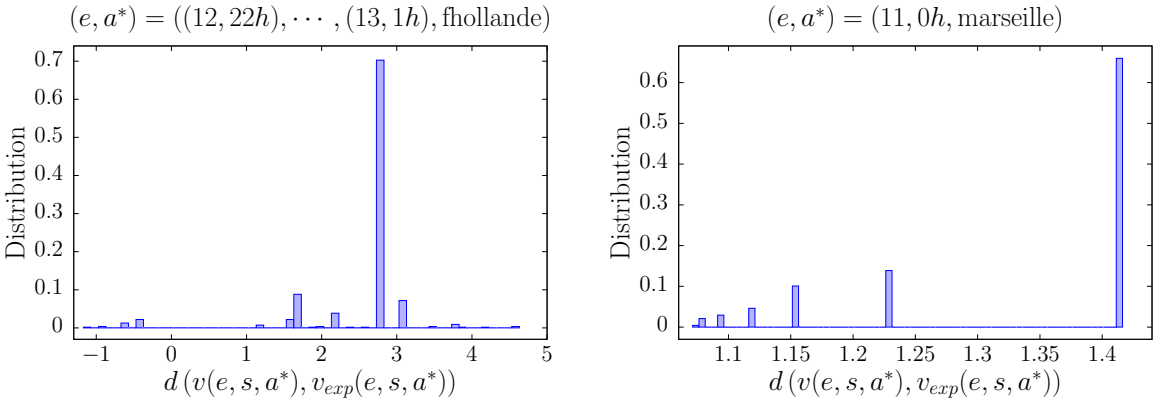


FIGURE 3.8 – **Distribution des valeurs de déviation dans le cas où tous les diffuseurs se comportent normalement** – Les barres de l'histogramme pour les valeurs supérieures (*resp.* inférieures) à 0 correspondent aux diffuseurs qui retweetent a^* pendant e plus (*resp.* moins) qu'ils ne le font habituellement. Par exemple, la valeur positive extrême pour fhollande correspond à un diffuseur qui n'a jamais retweeté fhollande de $22h$ à $1h$, sauf six fois pendant l'évènement. La valeur négative extrême correspond à un diffuseur qui l'a retweeté une fois pendant l'évènement, alors qu'il l'a retweeté 7 fois au total durant cette période.

est le nombre total de retweets reçus par a pendant H_e.

Selon ce contexte, un triplet $(e, s, a^*) \in \{e\} \times S \times \{a^*\}$ est anormal si la déviation entre le nombre de fois que s a retweeté a pendant e, et le nombre de fois que s est censé retweeter a, en moyenne, au cours de la période correspondante les autres jours est significative. De même, trois situations se présentent.

Phénomène global

Pour les évènements $((12, 22h), \cdots, (13, 1h), \text{fhollande})$ et $(11, 0h, \text{marseille})$, nous observons des distributions pour lesquelles l'étendue des valeurs de déviation est très petite (voir Figure 3.8). Dans le premier cas, nous observons 22 valeurs de déviation différentes. De plus, 90% des triplets (e, s, a^*) ont une déviation égale à 1.7, 2.2, 2.8 ou 3.1. Pour marseille, on fait les mêmes observations : il n'y a que 7 valeurs de déviation différentes, parmi lesquelles 90% des triplets sont répartis entre les valeurs suivantes : 1.41, 1.23 et 1.16 (voir Figure 3.8). On décrit quelque-uns des comportements correspondants dans le Tableau 3.2.
Ces distributions montrent un nombre limité de comportements : aucun diffuseur n'a une activité très différente de celle des autres. Ainsi, les émergences de fhollande et marseille sont des phénomènes globaux dans lesquels un grand nombre de diffuseurs les retweetent.

Groupe de militants en ligne

La Figure 3.9 montre la distribution des valeurs de déviation pour les évènements $(25, 19h, \text{NicolasSarkozy})$, $(27, 15h, \text{alainjuppe})$, $(28, 13h\text{-}15h, \text{JLMelenchon})$ et $(29, 7h\text{-}9h, \text{NicolasSarkozy})$. La plupart des observations $d_p \in \mathcal{D}$ suivent une distribution gaussienne centrée sur une moyenne $\bar{d}_p$. Contrairement aux distributions précédentes, $\bar{d}_p$ varie de $1,6$

Évènement	$((12, 22h), \cdots, (13, 1h))$				$(11, 0h)$		
Auteur anormal	fhollande				marseille		
Valeur de déviation	1.7	2.2	2.8	3.1	1.41	1.23	1.16
% de diffuseurs	9	4	70	7	66	14	10
Nombre de retweets durant e	1	2	1	2	1	2	3
Nombre total de retweets de h_i à h_j	2	3	0	0	0	0	0

TABLE 3.2 – **Comportements les plus probables dans le cas où tous les diffuseurs se comportent normalement** – Dans les deux cas, la valeur de déviation la plus probable correspond aux diffuseurs qui retweetent a^* une seule fois au cours de la période. Pour marseille, nous observons que plus le nombre de retweets pendant e est grand, plus la valeur de déviation est petite. Cela est dû à la déviation de Poisson qui prend en compte la significativité de l'écart entre la valeur observée et celle attendue.

Évènement	$(25, 19h)$	$(27, 15h)$	$(28, 13h$ - $15h)$	$(29, 7h$ - $9h)$
Auteur anormal	NicolasSarkozy	alainjuppe	JLMelenchon	NicolasSarkozy
% de diffuseurs anormaux	2.7	6.7	4.5	6
% de retweet	14	40	37	25

TABLE 3.3 – **Groupe de diffuseurs influents** – Nous observons qu'une faible proportion de diffuseurs constitue en fait une part importante de l'ensemble des retweets reçus par l'auteur principal lors de l'évènement. Par exemple, pour $(27, 15h, \text{alainjuppe})$, on détecte 19 diffuseurs anormaux ($6,7\%$ des diffuseurs). Ensemble, ils ont retweeté alainjuppe 513 fois à $15h$, soit 40% de ses retweets à cette heure.

à $2,3$. Ce changement indique que globalement, les diffuseurs ont une activité plus élevée que leur activité habituelle, ce qui explique en partie l'émergence de l'auteur principal a^*. On détecte des anomalies négatives et positives. Les anomalies négatives indiquent des diffuseurs qui retweetent a^* moins de fois qu'ils ne le sont supposés. En tant que tels, ils n'influencent pas l'émergence de a^*. Au contraire, les anomalies positives, qui sont des diffuseurs plus actifs qu'à leur habitude, jouent un rôle clé dans l'importance de a^* pendant e. C'est ce que nous observons dans le Tableau 3.3. Pour chacun des évènements, nous remarquons qu'un petit groupe de diffuseurs retweete a^* considérablement et représente une proportion non négligeable du nombre total de retweets. Au sein de ce groupe, plusieurs diffuseurs retweetent a^* plus de 50 fois durant l'évènement. Même s'ils ne représentent qu'une très faible proportion de l'ensemble des diffuseurs, ils sont une cause majeure de l'émergence de a^* pendant e.

Un seul activiste

L'évènement $(11, 3h, \text{FrancoisFillon})$ constitue un cas extrême de la situation précédente. Le groupe de diffuseurs anormaux est constitué d'un seul utilisateur qui retweete FrançoisFillon 73 fois à $3h$. Ainsi, l'émergence de François Fillon le 11 à $3h$ est uniquement due à ce diffuseur qui constitue à lui seul 100% de ses retweets.

Ici encore, l'analyse locale des diffuseurs nous amène à constater que certains évè-

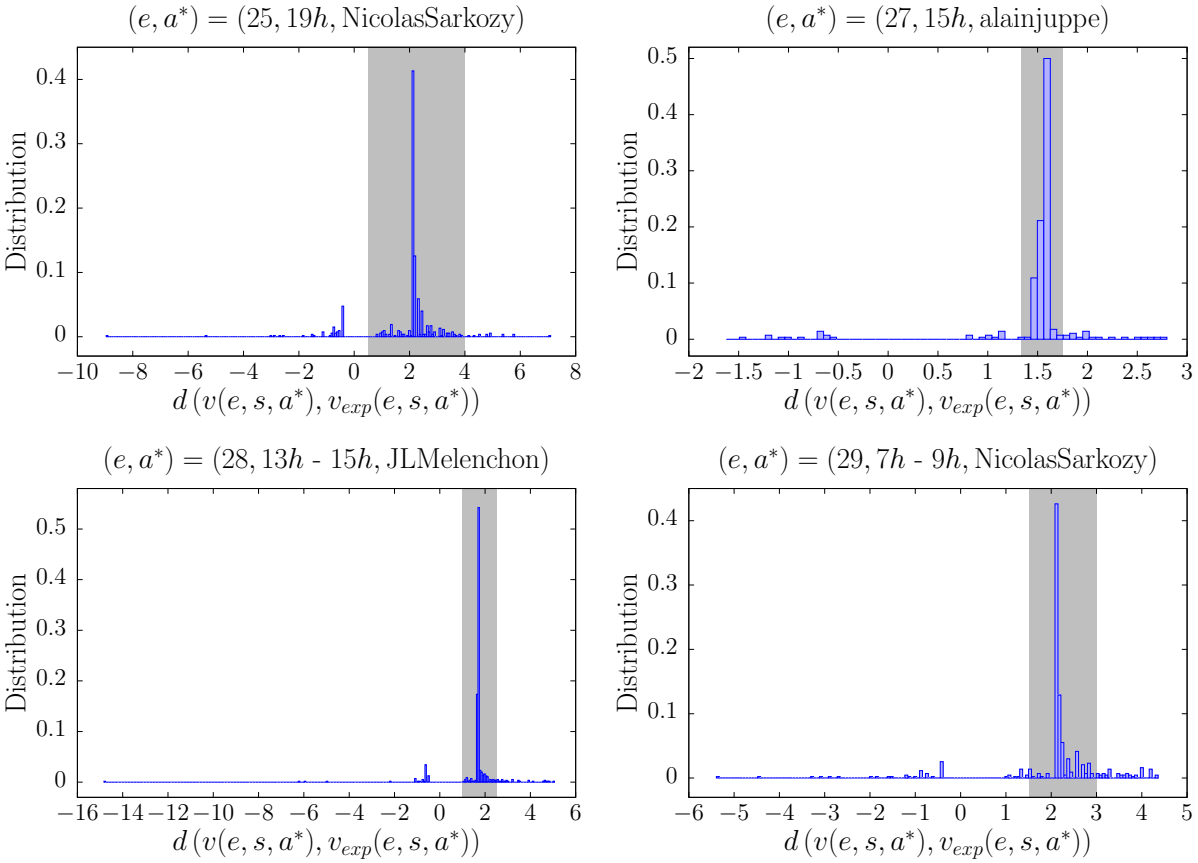


FIGURE 3.9 – **Un groupe de diffuseurs se comporte anormalement** – Dans chaque distribution, on observe des comportements similaires. La plupart des diffuseurs retweetent a^* une ou deux fois pendant e alors qu'ils ne retweetent généralement pas a^* à cette heure de la journée. Ces comportements sont inhabituels mais ne différent pas de manière significative. Ils sont représentés par la gaussienne de moyenne $\bar{d}_p$ comprise entre 1.6 et 2.3. Les diffuseurs qui ont l'habitude de retweeter a^* à cette heure de la journée ont des valeurs de déviation proches de 0 s'ils retweetent comme ils en ont l'habitude, négatifs, s'ils retweetent moins, ou positifs, s'ils retweetent plus.

nements sont des phénomènes plus globaux que d'autres. En particulier, l'émergence de certains auteurs est due en partie à un petit groupe de diffuseurs qui les retweete majoritairement, ce qui pourrait induire en erreur les autres utilisateurs sur la significativité de ces auteurs. Ainsi, cette analyse met en évidence des informations cruciales qui devraient être prises en compte afin évaluer la pertinence d'un évènement.

3.3.4 Hashtags anormaux

Il est possible d'obtenir des informations supplémentaires sur les évènements précédents en ajoutant une dimension basée sur le contenu des tweets à l'aide des hashtags. Dans cette section, nous appliquons notre méthode en considérant les quatre dimensions : diffuseurs, auteurs, hashtags et temps du jeu de données $\mathcal{D}_5$. Tout d'abord, nous recherchons des heures durant lesquelles certains hashtags sont anormalement retweetés, puis, nous établissons une corrélation avec les évènements précédemment détectés.

On s'intéresse aux triplets anormaux (d^*, h^*, k^*) dans le cube de données $\mathcal{C}_3(D \times H \times K, v)$. Étant donné la nature éphémère des hashtags, nous utilisons des valeurs attendues légèrement différentes des précédentes. Cette fois, nous prenons en compte l'activité attendue pendant l'heure h et nous l'ajustons avec le nombre de hashtags k retweetés le jour d :

$$v_{exp}(d, h, k) = v(d, \cdot, \cdot, \cdot, k) \times \frac{v(\cdot, h, \cdot, \cdot, \cdot)}{v(\cdot, \cdot, \cdot, \cdot, \cdot)}.$$

De cette façon, on ne suppose pas que le nombre de hashtags observés à heures fixes est constant. Selon ce contexte, un triplet (d^*, h^*, k^*) est anormal lorsqu'il y a une déviation significative entre le nombre de retweets contenant le hashtag k^* pendant (d^*, h^*) et le nombre de hashtags k^* qui seraient retweetés le jour d s'ils étaient répartis entre les heures proportionnellement à leurs activités moyennes.

On détecte 225 triplets anormaux (d^*, h^*, k^*), dont 114 hashtags différents (en ignorant les différences de casse et d'accents). Parmi les 225 triplets anormaux, 43% correspondent à un évènement anormal rencontré précédemment. Les Tableaux 3.4, 3.5 et 3.6 listent les hashtags anormaux en fonction de l'évènement correspondant, pour les évènements avec respectivement un, plusieurs et aucun auteur principal.

Premièrement, on remarque qu'un évènement est souvent associé à un slogan politique ainsi qu'une émission de radio ou de télévision. Dans ce cas, il y a trois situations possibles : soit l'émission reçoit un invité politique, soit l'émission parle d'une actualité associée à un ou plusieurs politiciens, ou au contraire, l'émission et le slogan politique ne sont pas corrélés (par exemple, dans le cas où plusieurs évènements d'actualité se produisent au cours de la même période).

On remarque également que les évènements des Tableaux 3.5 et 3.6 sont toujours associés à un terme général, indépendant d'un slogan politique ou d'une émission. Comme le suggère l'analyse sur les auteurs anormaux, cela montre que l'évènement correspondant résulte de la réaction à un fait externe. Par exemple, l'anormalité des hashtags « *Rio2016* » est liée à la réaction globale des utilisateurs aux victoires olympiques de la

Évènements	((12, 22h), ..., (13, 1h))	(25, 19h)	(27, 15h)	(28, 13h - 15h)	(29, 7h - 9h)
Hashtags anormaux	judo rio2016 fra espritbleu	**Slogan de campagne :** toutpourlafrance **Lieu :** chateaurenard	**Slogan de campagne :** 3moispourgagner	**Slogan de campagne :** benoithamon2017 lagauchepourgagner insoumis28aout **Émission télé/radio :** LeGrandJury	**Slogan de campagne :** toutpourlafrance **Émission télé/radio :** rtlmatin télématin bourdindirect invitépol

TABLE 3.4 – **Hashtags anormaux correspondant aux évènements avec un auteur principal.**

Évènements	(3, 10h - 13h)	(22, 16h - 22h)	(24, 20h - 22h)	(26, 15h - 18h)	(28, 0h)
Hashtags anormaux	sainterita *(nom d'une église)*	sarkozy **Slogan de campagne :** toutpourlafrance **Émission télé/radio :** clubdelapresse, e1soir	sarko **Slogan de campagne :** toutpourlafrance **Émission télé :** ns20h	burkini conseildetat **Émission télé/radio :** BFMTV	salafisme **Émission télé :** ONPC

TABLE 3.5 – **Hashtags anormaux correspondant aux évènements avec plusieurs auteurs principaux.**

Évènements	(13, 9h)	(19, 22h)	(21, 21h)	(23, 7h-8h)	(30, 17h-18h)
Hashtags anormaux	⁄	rio2016	rio2016 boxe	⁄	macron

TABLE 3.6 – **Hashtags anormaux correspondant aux évènements sans auteur principal.**

France. De même, le hashtag « sainterita » est lié à la réaction des utilisateurs à une intervention de la police dans une église. D'autre part, les évènements $(22, 16h\text{-}22h)$ et $(24, 20h\text{-}22h)$, attachés aux hashtags « *Sarkozy* » et « *Sarko* », suggèrent l'existence d'une discussion à propos de Nicolas Sarkozy décorrélée des discussions engagées par les tweets officiels et les hashtags diffusés par son équipe. Le 22 août, notamment, les utilisateurs réagissent à l'annonce de la candidature de Nicolas Sarkozy à la présidence : cet évènement correspond à la première utilisation du hashtag « *ToutpourLaFrance* », qui est son slogan de campagne.

Nous observons un autre fait intéressant : le 28 août entre $13h$ et $15h$, on détecte le slogan de campagne de JLMelenchon, « *insoumis28aout* », ce qui est attendu étant donné que JLMelenchon est l'auteur principal correspondant à cet évènement. Cependant, on détecte également les slogans de campagne de benoithamon, un autre politicien, « *benoithamon2017* » et « *LaGauchePourGagner* », ce qui est inattendu, car il n'apparaît pas comme un auteur principal dans l'étude précédente.

Enfin, on remarque que les évènements $(11, 0h)$, $(11, 3h)$, $(13, 9h)$ et $(23, 7h\text{-}8h)$ ne sont associés à aucun hashtag anormal. Cela est dû au fait que l'analyse effectuée dans cette sous-section est globale. Avec une analyse locale des hashtags anormaux, centrée sur les évènements, comme précédemment avec les auteurs dans la sous-section 3.3.2, nous parvenons à identifier les contenus sémantiques des évènements correspondants. Par exemple, lors de l'évènement $(13, 9h)$, nous identifions les hashtags anormaux « *etatdurgence* », « *cazeneuve* » et « *islamigration* », qui font référence à une mesure prise le même jour par le ministre de l'Intérieur Bernard Cazeneuve.

> Grâce à notre méthode, nous avons détecté des évènements anormaux, indépendamment de l'activité du jour ou de l'heure considérée. Ensuite, nous avons effectué une analyse locale sur chacun de ces évènements, en utilisant de nombreux contextes différents, plus ou moins filtrés ou agrégés. Cela nous a permis de comprendre leur émergence. Par exemple, nous avons appris que le 11 août à $3h$, un diffuseur unique retweete intensément François Fillon ; que du 12 à $22h$ au 13 à $1h$, de nombreux diffuseurs ont retweeté une seule fois fhollande, au sujet d'une victoire olympique de la France en judo ; ou que le 27 août à $15h$, un petit groupe de diffuseurs est en grande partie responsable de l'émergence d'alainjuppe lors de son meeting politique.

3.4 Application à la détection d'anomalies dans du trafic IP

Notre méthode s'applique aux interactions temporelles en général. On peut par conséquent l'utiliser également dans le but de détecter des anomalies dans du trafic IP. On l'illustre dans cette section à l'aide de la trace $\mathcal{D}_1$ du 25 juin 2013 de 0h à 1h (UTC +9). Une interaction $(t, u, v) \in \mathcal{D}_1$ indique que les adresses IP u et v se sont échangées un paquet à l'instant t. Après avoir illustré notre méthode dans le cas d'interactions orientées, nous nous replaçons, dans cette section, dans le cas plus général d'interactions non-orientées par soucis de simplification. Comme précédemment, nous présentons une étude de cas

qui, basée sur les évènements détectés dans la dimension temporelle, recherche les causes possibles de leur émergence en explorant les autres dimensions. Étant donné sa faible étendue temporelle, la trace de trafic IP n'est pas soumise au rythme circadien. De ce fait, les contextes agrégatifs sont suffisants pour mettre en valeur les anomalies.

3.4.1 Dimensions et propriété

Nous étudions les données selon les adresses IP U, les adresses IP V, et le temps T. De plus, on divise la dimension temporelle en deux sous-dimensions, les minutes, notées M, et les secondes dans une minute, notées S, tel que $t = (m, s)$ indique la seconde s de la minute m, avec $(m, s) \in M \times S$. Ici, on a $M = \{0, \cdots, 59\}$ et $S = \{0, \cdots, 59\}$.

On utilise la même propriété que précédemment, à savoir la quantité d'interactions. Dans le cas du trafic IP, elle correspond au nombre de paquets échangés. On la note p afin d'éviter sa confusion avec l'adresse IP v. Dans le cube de base $\mathcal{C}_{base} = \mathcal{C}_4(M \times S \times U \times V, p)$, p donne le nombre de paquets échangés entre les adresses IP u et v durant la seconde s de la minute m. Par exemple $p(25, 58, u, v) = 30$ indique que les adresses IP u et v se sont échangées 30 paquets entre 0h25min58s et 0h25min59s. Par agrégation, $p(\cdot, \cdot, m, s)$ indique le nombre total de paquets échangé durant la seconde (m, s) (voir Figure 3.10). Étant donnés des liens non-orientés, la propriété p est symétrique telle que $p(m, s, u, \cdot) = p(m, s, \cdot, u)$ et $p(m, s, u, v) = p(m, s, v, u)$.

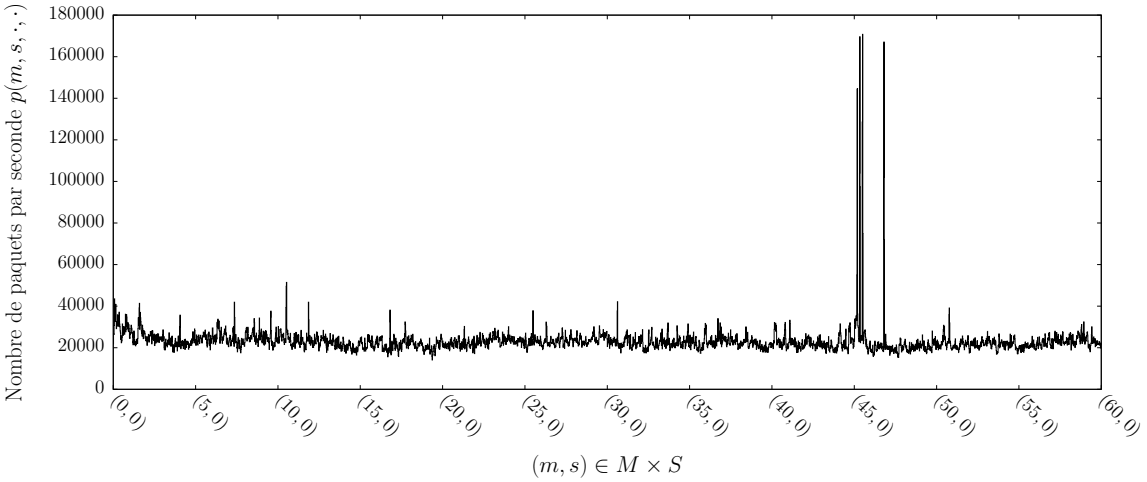


FIGURE 3.10 – **Nombre de paquets par seconde le 25 juin 2013 de 0h à 1h.**

3.4.2 Évènements

Un évènement $e = ((m_1^*, s_1^*), \cdots, (m_n^*, s_n^*)) \in \mathcal{E}$ est un ensemble de secondes anormales consécutives. On le note $e = (m^*, s_1^*\text{-}s_n^*)$ lorsque toutes les secondes s'étendent sur la même minute m^*. Les valeurs attendues sont les suivantes :

$$p_{exp}(m, s) = \frac{p(m, \cdot, \cdot, \cdot)}{|S|} \ .$$

La Figure 3.11 montre la distribution des valeurs de déviation en considérant la déviation de Poisson. On détecte 18 évènements (voir Tableau 3.7). En comparaison avec Twitter, on remarque que les déviations entre les valeurs observées et les valeurs attendues des secondes anormales sont beaucoup plus importantes. Par exemple, la valeur de déviation extrême égale à 177 523 sur la seconde $(46, 48)$ a une valeur observée de 167 064 pour une valeur attendue de 24 610.

En mettant en correspondance le Tableau 3.7 avec la Figure 3.10, on voit qu'en plus des anomalies globales correspondant aux pics d'activités autour de la $45^{\text{ème}}$ minute, on détecte des anomalies plus subtiles, moins facilement identifiables visuellement.

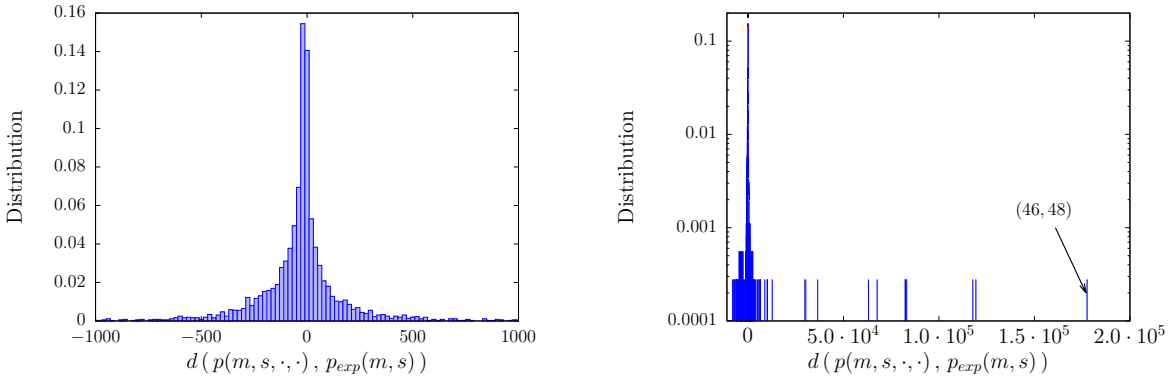


FIGURE 3.11 – **Secondes anormales dans le contexte agrégatif** – (Gauche) Zoom sur les valeurs de déviation correspondant aux observations normales. (Droite) L'axe des ordonnées est en échelle logarithmique afin de mettre en évidence les anomalies. La valeur de déviation extrême correspond à la seconde $(46, 48)$.

3.4.3 Adresses IP anormales pendant les évènements

On recherche localement si une ou plusieurs adresses IP sont responsables de l'émergence des évènements précédemment identifiés.

Les valeurs attendues sont les nombres de paquets que l'on observerait si les paquets échangés durant l'évènement e étaient distribués parmi les adresses IP proportionnellement à leurs activités globales :

$$v_{exp}(e, u) = v(e, \cdot, \cdot) \times \frac{v(\cdot, \cdot, u, \cdot)}{v(\cdot, \cdot, \cdot, \cdot)} \,.$$

La Figure 3.12 montre la distribution des valeurs de déviation pour les évènements sur les secondes $(25, 29)$, $(46, 47$ - $49)$ et $(1, 34$ - $35)$.

On observe que les évènements sont liés à une seule adresse IP (67%), deux adresses IP (22%) ou plus de deux adresses IP (11%) (voir Tableau 3.7). Contrairement aux retweets sur Twitter, le cas pour lequel l'évènement est un phénomène global où l'ensemble des nœuds a une activité plus élevée ne se présente pas.

	Évènements	IP anormales	Paires d'IP anormales	Anomalies réseau
1	$(0, 4)$	595	/	Point multipoint
2	$(1, 34$ - $35)$	820 ; 761 ;933 ; 922 ; 920 ; 438	(920,438) ;(920,933) (920,922) ;(820,761)	DoS Heavy hitter
3	$(4, 3)$	742	(584,742)	DoS
4	$(7, 21$ - $22)$	379	/	Point multipoint
5	$(9, 33$ - $34)$	603	/	Ntscan
6	$(10, 29$ - $31)$	619 ; 392 ; 944	(944,392)	Heavy hitter Point multipoint
7	$(11, 51)$	103	/	Point multipoint
8	$(16, 48$ - $49)$	595	/	Point multipoint
9	$(17, 43)$	274	/	Point multipoint
10	$(25, 29)$	141	/	Point multipoint
11	$(30, 36$ - $37)$	859	/	Point multipoint
12	$(36, 43)$	809	(809,920)	DoS
13	$(41, 4$ - $5)$	197	(584,197)	DoS
14	$(45, 9$ - $11)$	888 ; 449	(888,449)	Heavy hitter
15	$(45, 19$ - $21)$	888 ; 449	(888,449)	Heavy hitter
16	$(45, 29$ - $31)$	888 ; 449	(888,449)	Heavy hitter
17	$(46, 47$ - $49)$	888 ; 449	(888,449)	Heavy hitter
18	$(50, 45$ - $46)$	094	/	Ntscan

TABLE 3.7 – **Liste des évènements ainsi que des adresses IP, paires d'adresses IP et anomalies réseau associées** – Les adresses IP anormales sont répertoriées à l'aide d'un identifiant unique (simplifié par clarté dans la description). On remarque ici qu'à l'exception de 595, 888 et 449, les adresses IP apparaissent une seule fois. Les évènements sont étiquetés en fonction de leurs signatures à l'aide de la taxonomie proposée par Mazel et al. [105].

La remarque précédente selon laquelle les anomalies sont beaucoup plus marquées et significatives que dans Twitter s'applique également dans cette situation. À titre d'exemple, l'adresse IP 141 est censée échanger 7 paquets lors de l'évènement de la seconde $(25, 29)$, or elle en échange 13 813, ce qui explique sa valeur de déviation extrême (voir Figure 3.12).

3.4.4 Paires d'adresses IP anormales pendant les évènements

Dans le Tableau 3.7, 12 évènements sur 18 ont une adresse IP principale, notée u^*, échangeant un nombre de paquets significativement plus élevé que celui attendu durant l'évènement. L'anormalité de ces adresses IP peut provenir de deux schémas d'échange différentes :
(1) u^* échange un grand nombre de paquets avec un grand nombre d'autres adresses IP. Dans ce cas, les adresses IP avec lesquelles u^* interagit n'échangent que quelques paquets et gardent par conséquent une activité normale.
(2) u^* échange un grand nombre de paquets avec une seule autre adresse IP, v, ayant une activité importante. Dans ce cas, le nombre de paquets échangés entre les deux adresses IP n'est pas suffisant pour faire basculer le comportement normal de v vers un compor-

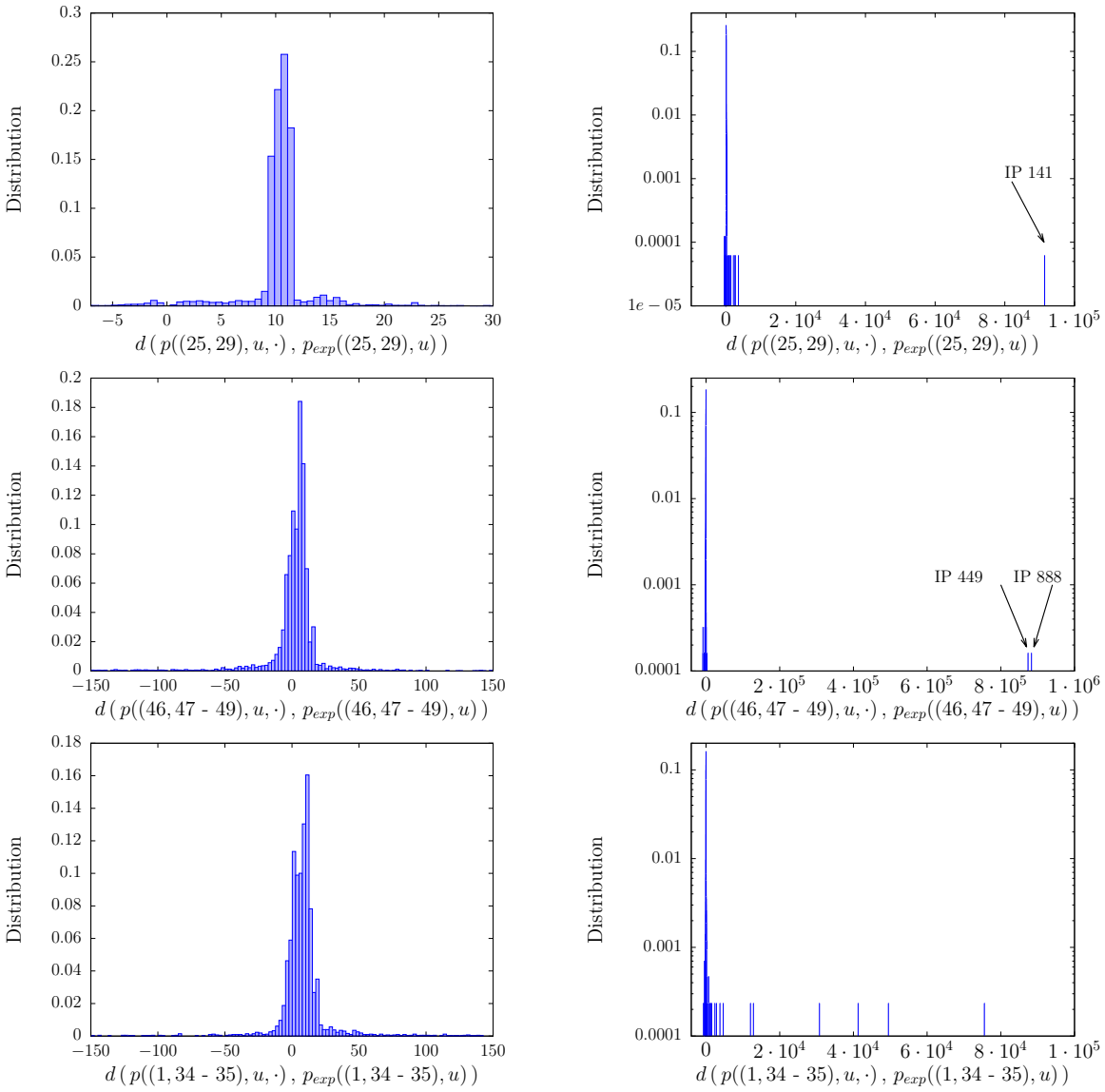


FIGURE 3.12 – **Adresses IP anormales durant les évènements dans un contexte agrégatif local** – Tandis que les figures de gauche montrent un zoom sur la partie centrale normale de la distribution, les figures de droite ont leur ordonnée en échelle logarithmique afin de mettre en évidence les anomalies. On distingue trois situations différentes : pour l'évènement $(25, 29)$, une seule adresse IP est anormale, pour l'évènement $(46, 47-49)$, on distingue deux adresses IP pour lesquelles les valeurs observées dévient significativement des valeurs attendues, enfin, lors de l'évènement $(1, 34$ - $35)$, on en distingue huit.

tement anormal, ce qui explique l'observation d'une seule IP anormale au lieu de deux [2] (cf. lignes 3, 12 et 13 du Tableau 3.7).

Afin d'obtenir plus de précisions sur ces évènements, on s'intéresse aux paires d'adresses IP anormales tel que

$$v_{exp}(e, u, v) = v(e, \cdot, \cdot) \times \frac{v(\cdot, \cdot, u, v)}{v(\cdot, \cdot, \cdot, \cdot)} .$$

Dans le cas (2), contrairement au cas (1), on s'attend à observer une paire (u^*, v) significativement anormale.

C'est ce que l'on observe pour les évènements sur les secondes $(4, 3)$, $(36, 43)$ et $(41, 4$ - $5)$. Pour chacun d'eux, on remarque sur la Figure 3.13.a, la présence d'une paire dont la valeur de déviation est significativement plus élevée que celle des autres paires. Sur la Figure 3.13.c, on voit que cette situation correspond au cas (2) : l'adresse IP principale u^* interagit avec une IP ayant une activité élevée. Pour les autres évènements, on est dans la situation (1) : la distribution des valeurs de déviation est plus homogène et les paires anormales ne sont pas corrélées avec u^* (voir Figure 3.13.b).

L'identification des paires anormales nous permet également d'approfondir notre connaissance des autres évènements (voir Tableau 3.7). Notamment, on remarque que l'évènement de la seconde $(1, 34$ - $35)$ est la combinaison de deux évènements : d'une part l'adresse IP 920 échange un grand nombre de paquets avec les adresses IP 438, 933 et 922, et d'autre part, la paire (820,761) agit anormalement. On observe la même chose concernant l'évènement de la $10^{\text{ème}}$ minute. Finalement, on voit sur la Figure 3.14 que les évènements se déroulant entre la $45^{\text{ème}}$ et la $46^{\text{ème}}$ minute sont uniquement dus à l'échange d'un grand nombre de paquets entre les adresses IP 888 et 449.

Cette étude locale des évènements nous permet ainsi d'identifier quelles adresses IP sont responsables de l'augmentation du nombre de paquets. En plus de cette information quantitative, l'analyse des paires d'IP anormales nous donne une information sur la structure des interactions. Notamment, on repère des évènements induits par l'échange d'un grand nombre de paquets entre une adresse IP et un grand nombre d'autres adresses IP (voir Figure 3.15.a), entre une adresse IP et un nombre restreint d'autres adresses IP (voir Figure 3.15.b), ou entre deux adresses IP (voir Figure 3.15.c).

3.4.5 Classification des anomalies

Contrairement aux retweets politiques sur Twitter, les traces de trafic IP ne nous permettent pas de valider nos résultats par comparaison aux évènements médiatiques. Dans cette sous-section, nous classons les anomalies détectées en termes d'anomalies réseau à l'aide de la taxonomie proposés par Mazel et al. [105]. Dans cette dernière, les anomalies réseau sont associées à une étiquette en fonction d'ensembles de règles (signatures) qui caractérisent leur trafic. Les résultats sont listés dans le Tableau 3.7.

2. Ce raisonnement s'étend à un petit groupe d'adresses IP ayant une activité élevée.

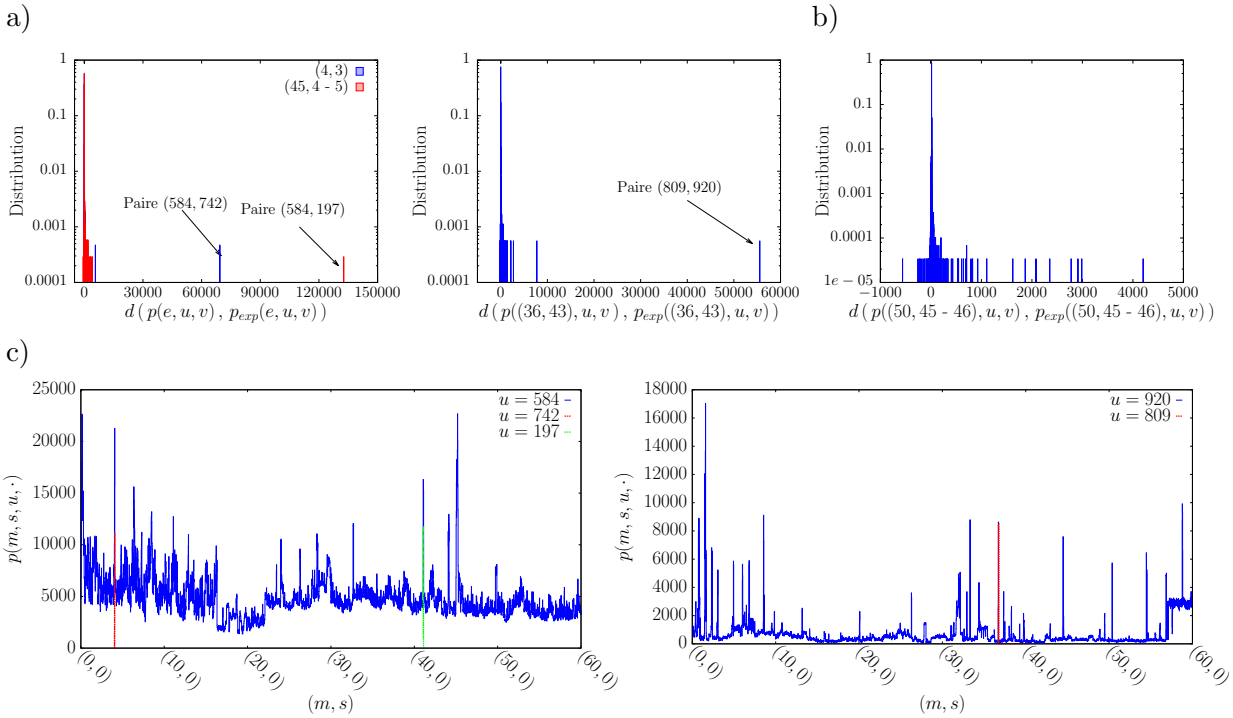


FIGURE 3.13 – **Paires d'adresses IP anormales pendant les évènements dans un contexte agrégatif local** – a) Les distributions indiquent que l'évènement est lié à l'échange de paquets entre une paire d'adresses IP. b) La distribution des valeurs de déviation est plus homogène, l'étendue des valeurs est plus petite, montrant que l'évènement de la minute 50 est principalement dû à l'activité de l'adresse IP 094, déviant significativement, plutôt qu'à l'activité des paires. c) Prises séparément, les adresses IP 584 et 920 ne sont pas anormales étant donné leur activité élevée.

a) Paires d'IP anormales durant les évènements

b) Nombre de paquets par seconde de la paire (888,449)

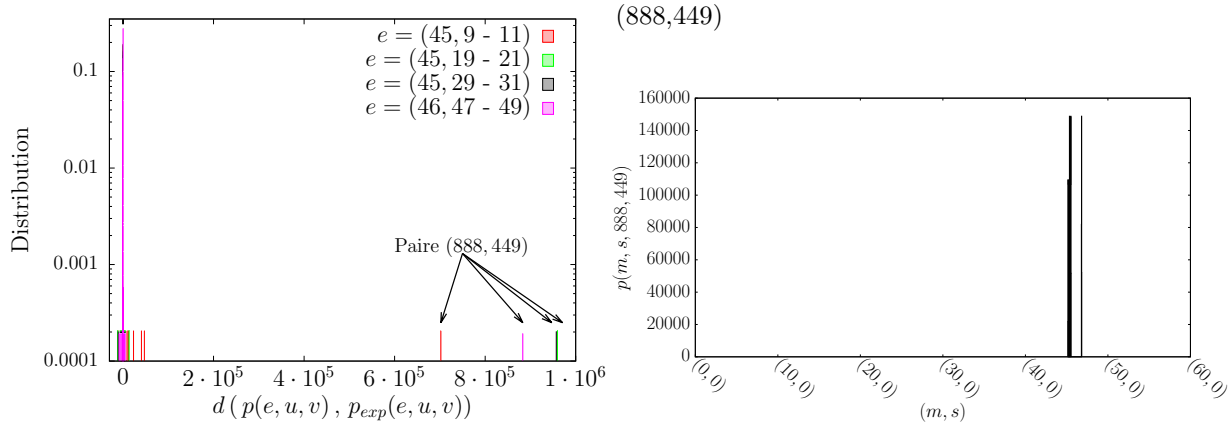


FIGURE 3.14 – **Les évènements entre la $45^{\text{ème}}$ et la $46^{\text{ème}}$ minutes sont dus à la paire d'adresses IP (888,449)** – Lors de l'évènement $(46, 47 - 49)$, la paire est censée s'échanger 4 643 paquets. On voit sur l'évolution du nombre de paquets échangés qu'elle en échange en réalité 141 314 d'où son anormalité durant cet évènement.

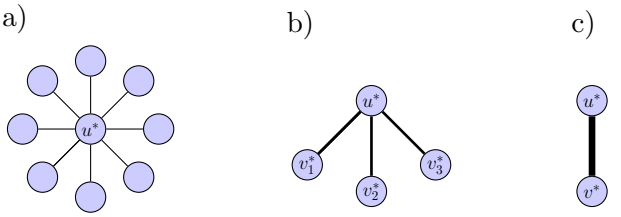


FIGURE 3.15 – **Identification de structures particulières grâce à l'étude des paires d'adresses IP** – La largeur des liens est proportionnelle au nombre de paquets envoyés entre les adresses IP correspondantes. a) Une adresse IP principale. b) Nombre restreint d'adresses IP anormales. c) Paire d'IP anormale.

Les anomalies impliquant une paire d'adresses IP (u^*, v^*) et pour lesquelles ni u^* ni v^* n'ont d'autres voisins, correspondent, par définition, à des attaques de type *heavy hitter* caractérisées par une adresse IP source et une adresse IP destination s'envoyant un nombre de paquets supérieur à 1000 avec une fréquence élevée.

Les dénis de service (DoS) sont caractérisés par un nombre d'adresses IP source inférieur à 20, un nombre d'adresses IP destination inférieur à 5 ainsi qu'un grand nombre de paquets et une fréquence d'interaction élevée. Les anomalies correspondant à cette signature impliquent les paires d'adresses IP (u^*, v^*) pour lesquelles soit u^* soit v^* possède un petit nombre un nombre de voisins. C'est ce que l'on observe pour les paires contenant les IP 920 et 584.

Finalement, les anomalies point à multipoint concernent les communications comportant une adresse IP source et plus de 20 adresses IP destination. Ces anomalies sont celles pour lesquelles la structure du voisinage est représenté par la Figure 3.15.a. Parmi ces anomalies, on identifie des scans de réseau lorsque les adresses IP destinations appartiennent au même réseau.

L'analyse du nombre de paquets échangés dans les trois dimensions U, V et T à l'aide de notre méthode nous a permis de détecter des anomalies pertinentes vis-à-vis de la surveillance du trafic IP. La précision temporelle de nos résultats peut être augmentée en considérant une granularité temporelle plus fine que celle des secondes. Cette étude de cas montre ainsi que notre méthode s'applique à toutes sortes d'applications, si tant est que l'on dispose d'un ensemble d'interactions temporelles, de dimensions selon lesquelles étudier les interactions et d'une propriété permettant d'analyser les relations entre les dimensions.

Nous avons appliqué notre méthode à l'analyse de la communication politique sur Twitter et à la détection d'anomalies dans du trafic IP en utilisant la quantité d'interactions. Dans les deux cas, nous avons montré que notre méthode met en évidence des évènements et permet de faire des hypothèses concernant les causes de leur émergence. Sur Twitter, nous avons détecté des auteurs anormalement retweetés, des groupes de diffuseurs très actifs et les sujets d'actualité correspondant aux évènements anormaux. Nous avons ainsi montré que notre méthode met en évidence des informations cruciales à prendre en compte pour évaluer la fiabilité d'un évènement sur Twitter. Dans le trafic IP, nous avons identifié plusieurs struc-

tures anormales correspondant à des attaques connues et répertoriées dans l'état de l'art, en particulier, des paires d'adresses IP qui s'échangent des milliers de paquets et des adresses IP qui échangent des paquets avec des milliers d'autres adresses IP.

3.5 Autres applications

Les études de cas des sections 3.3 et 3.4 ne présentent qu'une petite partie de l'étendue des possibilités offertes par notre méthode. En jouant à la fois sur les ensembles d'entités et d'observations considérés, ainsi que sur le type de comportement attendu et la propriété utilisée, il est possible de construire de nombreux autres contextes. Dans cette section, nous présentons en détails deux perspectives de nos travaux dans le jeu de donnée $\mathcal{D}_5$: l'analyse des réactions des utilisateurs aux émissions de télévision via Twitter, et l'étude de la dynamique des sujets ainsi que la prédiction des liens utilisateur-sujet.

3.5.1 Caractérisation de l'utilisation du second écran

Les observations faites lors de la détection d'anomalies dans les interactions (d, h, s, a, k) de Twitter montrent l'omniprésence des hashtags k liés aux médias au sein de chaque évènement. Par conséquent, il serait intéressant de décrire plus précisément les interactions entre les utilisateurs et les émissions de télévision via Twitter. Dans l'état de l'art, cette utilisation particulière des réseaux sociaux est regroupée sour le terme d'*utilisation du second écran* [141, 45].

La caractérisation de l'utilisation du second écran est un domaine d'étude très récent. Le terme *second écran* fait référence à un écran connecté à internet, comme un smartphone ou un ordinateur portable, que les gens utilisent pour commenter les programmes télévisés sur les réseaux sociaux tout en les regardant. Dans le cadre de cette étude, il est intéressant d'analyser les différences entre ce qui est dit dans le programme télévisé et les discussions qui en découlent sur Twitter. Cela a été appliqué dans de nombreuses situations, en particulier pour suivre les évènements sportifs [43] et les débats politiques [58, 54, 59]. Dans ce qui suit, nous analysons l'utilisation du second écran à l'aide de notre méthode lors de l'apparition de Nicolas Sarkozy au journal télévisé pour le lancement de sa campagne le 24 août de $20h$ à $22h$. Cette approche est nouvelle dans la mesure où les études précédentes consistent souvent soit en une comparaison manuelle entre le contenu des tweets et un enregistrement des discussions de l'émission, soit en une étude de l'audience télévisée et du nombre de tweets observés au cours du temps.

Premièrement, nous nous concentrons sur les auteurs anormaux à $20h$, $21h$ et $22h$ en utilisant les mêmes valeurs attendues que dans la section 3.3.2. La Figure 3.16 montre la distribution de l'ensemble des valeurs de déviation pour $e_1 = (24, 20h)$, $e_2 = (24, 21h)$ et $e_3 = (24, 22h)$. À $20h$, il y a deux auteurs principaux : NicolasSarkozy et TTpourlaFrance, le slogan de son parti. À $21h$, une autre situation se produit. L'ensemble des valeurs est plus homogène : il y a plus de valeurs aberrantes, mais moins significatives. Parmi celles-ci, nous identifions de nombreux journalistes ainsi que des hommes politiques de droite sou-

tenant Nicolas Sarkozy. Nous remarquons que certains utilisateurs anonymes, qui ne sont liés ni à un journal ni à un parti politique, commencent à apparaître parmi des auteurs anormaux. Enfin, à $22h$, la plage de valeurs est encore plus réduite, ce qui signifie que l'évènement observé n'est pas le résultat d'une focalisation sur un nombre limité d'auteurs, mais un phénomène global durant lequel tout le monde retweete. Parmi les nœuds anormaux, on identifie uniquement des journalistes et des utilisateurs anonymes. Par conséquent, plus le temps passe, plus les distributions sont homogènes, ce qui montre que l'évènement devient un phénomène global au fur et à mesure que l'information se propage.

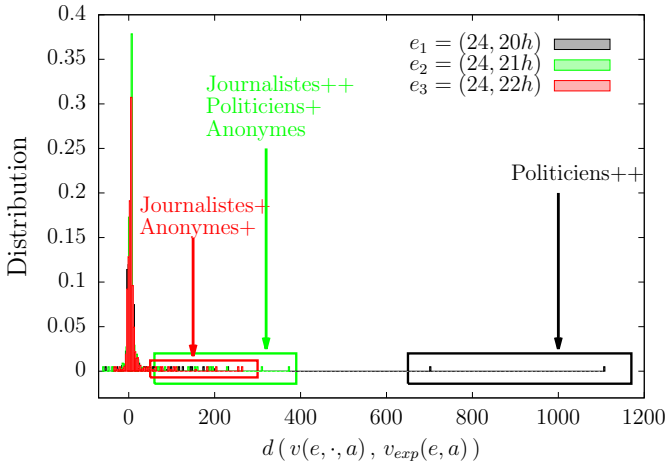


FIGURE 3.16 – **Évolution des distributions sur les auteurs anormaux le 24 août de $20h$ à $22h$.**

Cette analyse montre que l'interview de Nicolas Sarkozy à la télévision est reprise par les utilisateurs sur les réseaux sociaux. Nous nous concentrons maintenant sur les hashtags anormaux pour analyser l'évolution de la discussion. Nous observons des distributions similaires. À $20h$, les deux hashtags « *ns20h* » et « *toutpourlafrance* » se démarquent significativement. À $21h$ et $22h$, les distributions sont plus homogènes. Les deux hashtags précédents, publiés par l'équipe de Nicolas Sarkozy, sont encore anormaux à $21h$, mais redeviennent normaux à $22h$. Les autres hashtags ne sont anormaux que de $21h$ à $22h$ ou de $22h$ à $23h$. Parmi ceux-ci, on trouve des termes utilisés par Nicolas Sarkozy lors de l'interview, par exemple « *chomage* ». Enfin, on observe une évolution des hashtags faisant référence à un même sujet : à $21h$, « *hollande* », puis à $22h$, « *hollandedemission* » ; ou « *schengen* » à $21h$, puis « *stopschengen* » à $22h$; ou encore « *burkini* » de $20h$ à $22h$, puis « *bikini* » à partir de $22h$.

Cette analyse préliminaire peut être approfondie. Par exemple, lorsque l'on étudie les hashtags anormaux, on pourrait utiliser des contextes locaux, restreints aux journalistes, à l'équipe politique de Nicolas Sarkozy ou à des utilisateurs indépendants, afin d'analyser les hashtags que chacune de ces communautés propage. Nous pourrions également nous concentrer sur l'évolution des hashtags appartenant à un même sujet et voir s'ils sont retweetés par la même communauté de diffuseurs.

3.5.2 Prédiction des liens utilisateur-sujet

Cette dernière question suscite beaucoup d'intérêt parmi les chercheurs : beaucoup s'intéressent à la dynamique des sujets et en particulier à la prédiction des liens utilisateur-sujet. Ainsi, une autre perspective de nos travaux consiste à se concentrer sur l'aspect sémantique des interactions. La première difficulté consiste à trouver un ensemble des termes formant un sujet, c'est-à-dire un contenu sémantique cohérent. Certains chercheurs caractérisent un sujet à partir d'un ensemble de hashtags dont les évolutions temporelles sont similaires [116], ou à partir de groupes de hashtags fortement associés dans les tweets [29]. D'autres utilisent des techniques d'analyse de texte pour déduire un sujet de l'intégralité du texte contenu dans les tweets, plutôt que d'utiliser uniquement les hashtags [159]. Pour prédire les liens utilisateur-sujet, la plupart des chercheurs utilisent des techniques d'apprentissage automatique pour l'analyse des sentiments [124, 133, 41, 125]. On en trouve également utilisant des méthodes basées sur le lexique [114]. Dans ce qui suit, nous proposons une nouvelle approche consistant à rechercher des sujets parmi les hashtags anormalement retweetés.

Nous possédons uniquement la structure des retweets (d, h, s, a, k). Afin d'identifier les sujets à partir de ces données, nous tirons parti du fait que les utilisateurs sont engagés dans une cause, en particulier dans le cas de la communication politique. Autrement dit, un auteur aura tendance à souvent publier des tweets liés à une cause et les diffuseurs attachés à cette cause auront tendance à les retweeter intensément. Nous définissons alors un sujet comme étant un ensemble de hashtags intensément retweetés par les mêmes diffuseurs et pour lesquels un groupe d'auteurs est intensément retweeté.

Formellement, soit $K_N \subseteq K$ un ensemble de N hashtags. Premièrement, pour chaque hashtag $k_i \in K_N$, nous recherchons localement les diffuseurs anormaux associés à k_i en fonction des valeurs attendues suivantes

$$v_{exp}(d, h, s, k_i) = v(d, h, \cdot, \cdot, k_i) \times \frac{v(\cdot, h, \cdot, s, \cdot)}{v(\cdot, h, \cdot, \cdot, \cdot)} \quad .$$

On obtient alors un groupe de diffuseurs anormaux noté $S^*_{k_i}$ tel que $s \in S^*_{k_i}$ est un diffuseur qui retweete le hashtag k_i anormalement durant une heure donnée, étant donné son activité habituelle à cette heure de la journée. Après avoir exécuté cette étape sur tous les hashtags, nous définissons le groupe de diffuseurs associés à K_N comme étant l'ensemble des diffuseurs anormaux communs à tous les hashtags de l'ensemble : $S^*_{K_N} = \bigcap_{i=1}^{i=N} S^*_{k_i}$. Nous procédons de manière symétrique pour trouver l'ensemble des auteurs anormaux liés à K_N, notés $A^*_{K_N}$. Étant donné l'ensemble de diffuseurs et d'auteurs anormaux liés à K_N, ont dit alors que K_N est un sujet si $S^*_{K_N}$ et $A^*_{K_N}$ ne sont pas vides (voir Figure 3.17). La raison pour laquelle nous ne nous intéressons qu'aux auteurs et aux diffuseurs anormaux est que ces utilisateurs veulent incontestablement propager le sujet.

Avec $N = 3$ et en considérant l'ensemble des triplets obtenus à partir des 114 hashtags anormaux identifiés précédemment, nous trouvons 876 sujets. Par exemple, on identifie le sujet $K_3 = \{$*chateaurenard, ns20h, toutpourlafrance*$\}$, qui possède 4 auteurs anormaux

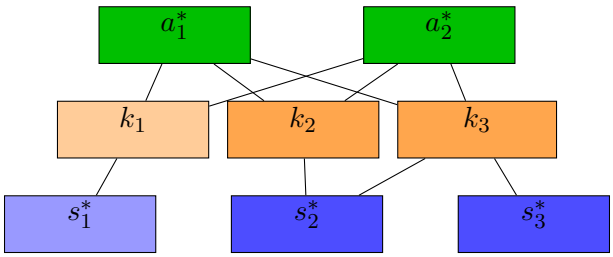


FIGURE 3.17 – **Formation des sujets à partir des hashtags** – Pour $K_3 = \{k_1, k_2, k_2\}$, $S^*_{k_1} = \{s^*_1\}$, $S^*_{k_2} = \{s^*_2\}$, et $S^*_{k_3} = \{s^*_2, s^*_3\}$. Par conséquent, $S^*_{K_3} = \emptyset$ et K_3 ne constitue pas un sujet. D'autre part, $K_2 = \{k_2, k_2\}$ est un sujet étant donné que $A^*_{K_2} = \{a^*_1, a^*_2, a^*_3\}$ et $S^*_{K_2} = \{s^*_2\}$.

appartenant au même parti politique,

$$A^*_{K_3} = \{\text{ GilAverous, LArribage, NicolasSarkozy, TTpourlaFrance}\},$$

et 48 diffuseurs anormaux ; le sujet $K'_3 = \{$*3moispourgagner, legrandrdv,uemedef2016*$\}$ associé à un auteur anormal, alainjuppe, et à un groupe de 18 diffuseurs anormaux ; et le sujet $K''_3 = \{$*boxe, judo, rio2016*$\}$ associé à 7 auteurs anormaux d'origines différentes, et seulement 3 diffuseurs anormaux.[3] La Figure 3.18 montre l'évolution temporelle de chaque hashtag dans chaque sujet. On constate que les hashtags appartenant au même sujet n'ont pas nécessairement la même dynamique.

À partir de cet ensemble de sujets, on peut déduire des communautés d'utilisateurs en fonction de la similarité des sujets qu'ils ont l'habitude de retweeter (ou par rapport auxquels ils sont retweetés). Nous abordons maintenant le problème de la prédiction des liens utilisateur-sujet. Plus précisément, nous voulons prédire le nombre d'interactions entre le diffuseur s, appartenant à la communauté c_s, et le sujet K_N au cours de l'heure (d, h). La prédiction de lien est intimement liée à la détection de liens anormaux. En effet, si la détection de quadruplés anormaux (d, h, s, K_N) est basée sur la mesure de l'écart entre une valeur observée $v(d, h, s, K_N)$ et sa valeur attendue $v_{exp}(d, h, s, K_N)$, la prédiction de lien se concentre sur la description du comportement normal et est donc basée uniquement sur les valeurs attendues. Par exemple, on pourrait prédire le nombre d'interactions entre s et K_N pendant (d, h) tel que

$$v_{exp}(d, h, s, K_N) = \underbrace{\frac{v(\cdot, \cdot, c_s, \cdot, K_N)}{v(\cdot, \cdot, \cdot, \cdot, K_N)}}_{(1)} \times \underbrace{\frac{v(\cdot, h, s, \cdot, \cdot)}{v(\cdot, h, c_s, \cdot, \cdot)}}_{(2)} \times \underbrace{\frac{v(d, h, \cdot, \cdot, K_N)}{|D|}}_{(3)}$$

où (1) prend en compte l'activité de la communauté de s vers le sujet K_N, (2) l'activité de s au sein de sa communauté durant l'heure h de la journée, et (3) le nombre attendu de retweets de K_N pendant l'heure h du jour d.

Cette analyse préliminaire peut également être approfondie. Notamment, on peut améliorer la prédiction en prenant en compte le comportement des auteurs que c_s a l'habitude

3. Il est à noter que dans ce cas, on ne trouve que 3 diffuseurs anormaux puisque les évènements liés au sport sont généralement des évènements homogènes qui ne présentent pas de groupes de diffuseurs militants.

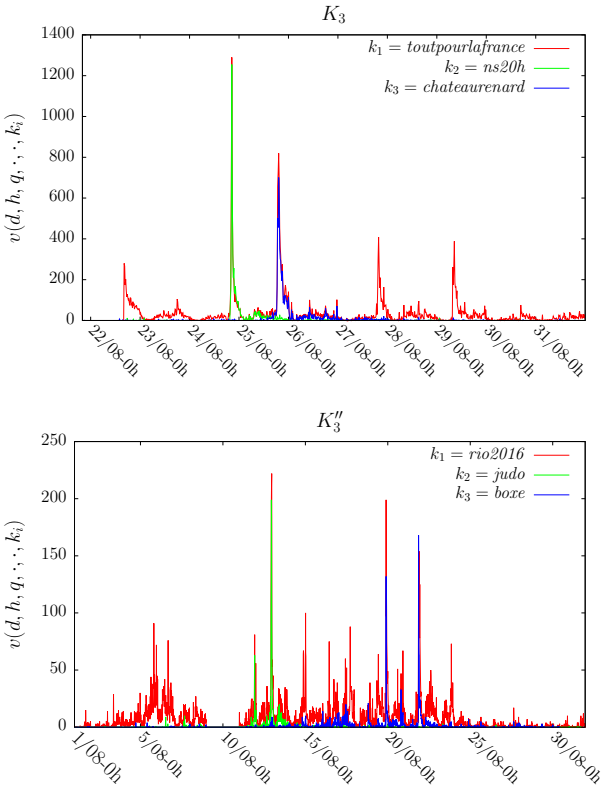


FIGURE 3.18 – **Evolution du nombre de retweets contenant le hashtag k_i pour trois sujets différents** – Pour obtenir une meilleure précision, nous avons tracé le nombre de retweets contenant le hashtag k_i par trimestre q (*quarter* en anglais). Nous voyons que la dynamique des hashtags dans un même sujet peut être décorrélée comme dans K'_3, ou corrélée comme « *ns20h* » et « *chateaurenard* » avec « *toutpoutlafrance* » dans K_3.

de retweeter envers le sujet K_N. D'autre part, si K_N est un nouveau sujet, on peut imaginer remplacer l'activité du sujet K_N par l'activité moyenne d'un ensemble de sujets qui lui sont liés.

Au travers de ces deux exemples, nous avons montré que notre méthode est utile dans de nombreuses études et applications empiriques. De manière complémentaire, ces applications induisent des informations et des questionnement nécessaires pour créer des contextes de plus en plus complexes et pertinents et ainsi tirer parti de l'étendue des possibilités offertes par notre méthode.

3.6 Application au degré dans un flot de liens

Dans les sections précédentes, nous avons analysé les interactions temporelles $(t, u, v) \in T \times V \times V$ selon leur dimensions structurelles V et temporelle T à l'aide de la quantité d'interactions. Nous avons coupé l'axe temporel en fenêtres de temps (d'une heure sur Twitter et d'une seconde dans le trafic IP), puis agrégé les interactions dans chacune d'entre elles. Comme nous l'avons vu dans le Chapitre 2 dans le cas des flots de liens ponctuels, cette approche réduit la précision concernant la dynamique des interactions. Pour pallier à ce problème, nous appliquons notre méthode aux flots de liens avec durée. Dans ces derniers, deux nœuds sont liés l'un à l'autre de t_1 à t_2 s'ils interagissent au

moins une fois tous les Δ dans cet intervalle de temps [4] (voir Section 2.1.2). Selon cette modélisation, l'utilisation de la quantité d'interactions n'est pas judicieuse : le nombre d'interactions entre deux nœuds à l'instant t est égal à 1 s'ils interagissent à cet instant, 0 sinon. De ce fait, nous utilisons le degré instantané des nœuds. Les dimensions sont l'ensemble des nœuds V et le temps T. Le temps étant continu dans les flots de liens, la représentation visuelle des opérations sur les cubes de donnée n'est pas envisageable dans ce cas. Néanmoins, la méthode de construction de contextes reste identique. Dans cette section, nous recherchons des nœuds-temps $(t, v) \in T \times V$ anormaux dans les contextes basique et agrégatifs.

3.6.1 Contexte basique

Afin d'obtenir un premier aperçu des nœuds-temps globalement anormaux, nous utilisons le contexte basique. Dans ce contexte, le degré attendu du nœud v au temps t est le degré du flot de liens,

$$d_{exp}(t, v) = \frac{1}{|T \times V|} \sum_{v' \in V} \int_{t' \in T} d(t', v')\, dt' \,.$$

Les distributions des valeurs de déviation pour Twitter et le trafic IP sont illustrées dans la Figure 3.19.

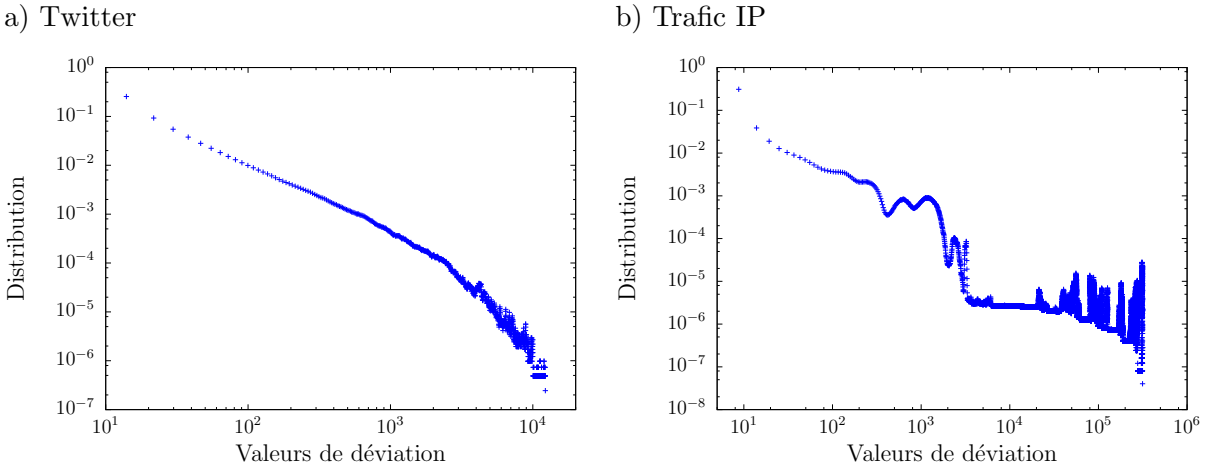


FIGURE 3.19 – **Distributions des valeurs de déviation du degré instantané des nœuds dans le contexte basique global.**

Dans les deux cas, on observe que les distributions sont très hétérogènes : la plage des valeurs de déviation s'étend sur plusieurs ordres de grandeur et plus la valeur de déviation est élevée, plus sa probabilité est basse. Étant données les distributions du degré instantané des nœuds dans $\mathcal{L}_4$ et $\mathcal{L}_1$, tracées dans les Sections 2.3.2 et 2.4.2, ce résultat était envisageable, le degré étant simplement divisé par une constante.

4. Dans Twitter, $\delta = 30min$, dans le trafic IP, $\delta = 1s$.

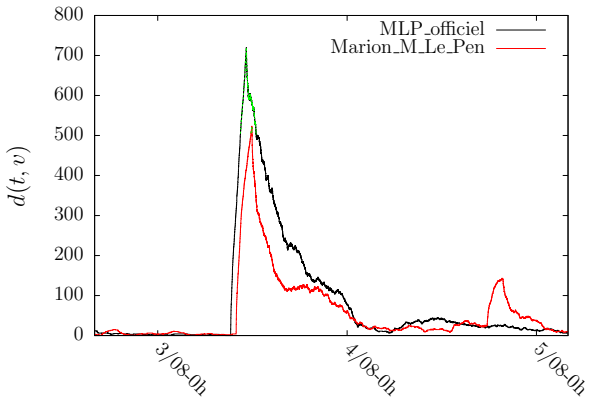
Dans ces distributions, nous détectons des nœuds-temps anormaux parmi les valeurs extrêmes. [5] Dans les retweets de Twitter, on remarque que les anomalies détectées correspondent aux évènements centrés sur un ou plusieurs auteurs principaux et ayant un nombre restreint de diffuseurs anormaux significatifs. Par exemple, les anomalies dont la valeur de déviation est supérieure à 6 000 correspondent à NicolasSarkozy, qui a un degré anormal le 24 août lors de son interview, MLP_officiel et Marion_M_Le_Pen, qui ont un degré anormal le 3 août lors de l'intervention de la police dans l'église Sainte Rita, ou encore fhollande, qui a un degré anormal le 12 août lors de la victoire olympique de la France. On détecte également un utilisateur qui reçoit une attention considérable lors de l'interview de NicolasSarkozy le 24 août. On trace leurs profils de degré dans les Figures 3.20.a, 3.20.b et 3.20.c en marquant les intervalles sur lesquels ils sont détectés anormaux. On remarque que la meilleure résolution temporelle met en évidence la dynamique des interactions.

Bien qu'anormaux du point de vue de la quantité d'interactions, alainjuppe et JLMelenchon lors de leurs meetings politiques les 27 et 28 août, et NicolasSarkozy, lors de son interview à la radio le 29 août, n'apparaissent pas parmi les valeurs de déviation extrêmes du point de vue du degré. Cela est dû au fait que le nombre de retweets qu'ils reçoivent lors de ces évènements, étant donné le groupe de militants en ligne, est distribué parmi un plus petit nombre de voisins comme le confirme le Tableau 3.3.

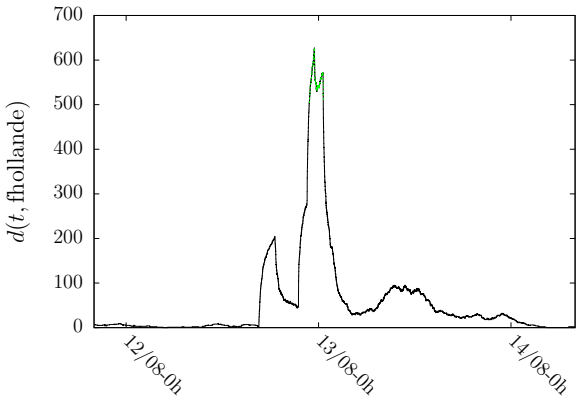
On fait les mêmes remarques concernant le trafic IP (voir Figure 3.20.d). Parmi les valeurs de déviation les plus élevées (supérieures à 10 000) on détecte des adresses IP identifiées précédemment avec la quantité d'interactions, en particulier toutes celles dont la structure du voisinage correspond à la structure 3.15.a. On en détecte de plus 7 nouvelles. Le fait qu'elles aient été détectées avec le degré dans le flot de liens avec durée et non la quantité d'interactions s'explique de deux façons différentes :
– la première provient du fait que le degré et la quantité d'interactions fournissent des informations différentes sur les interactions : alors que le degré quantifie la diversité du voisinage d'un nœud-temps, la quantité d'interactions quantifie son nombre d'interaction. Ainsi, lorsqu'un nœud envoie un seul paquet à de nombreux autres nœuds, son degré et sa quantité d'interactions sont identiques. Or, au regard des autres quantités d'interactions, cette dernière peut paraître normale, à l'inverse du degré.
– La deuxième raison est liée à l'agrégation dans une fenêtre de temps fixe. Par exemple, considérons parmi ces adresses, l'adresse IP 379 (profil de degré bleu dans la Figure 3.20.d). Cette dernière est détectée anormale avec le degré de $t_1 = 441.47$ à $t_2 = 442.67$. Quand on s'intéresse aux liens ponctuels, on remarque que les interactions sont regroupées sur la période allant de 441.78 à 442.37. La quantité d'interactions de l'IP 379 sur cette période est de 21 713. Or, sur les secondes 441 et 442, elle est de 9 255 et 12 458, respectivement. Ainsi, le fait que l'intervalle de temps anormal se situe à cheval entre deux secondes divise pratiquement de moitié le nombre d'interactions effectivement observées, rendant ces secondes normales du point de vue de la quantité d'interactions par seconde.

5. Nous discutons de ce choix dans la conclusion (Section 3.7).

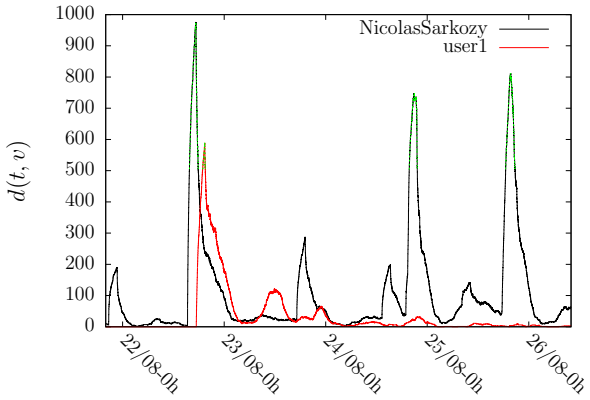
a) Twitter : exemple 1



b) Twitter : exemple 2



c) Twitter : exemple 3



d) Trafic IP

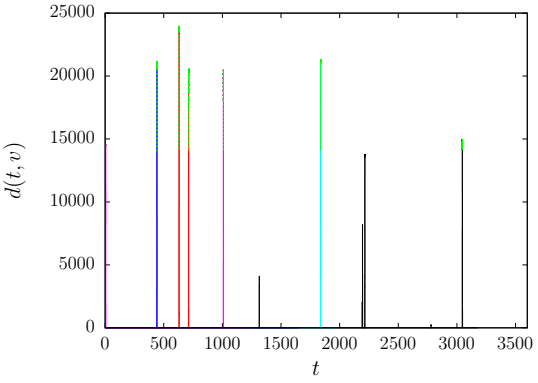


FIGURE 3.20 – **Profil de degré des nœuds anormaux dans le contexte basique global** – Les intervalles durant lesquels les nœuds sont anormaux sont surlignés en pointillés verts. a,c) On remarque que la dynamique du degré de Marion_Le_Pen suit à 42 minutes près celle du degré de MLP_officiel. De même, le degré de user1 augmente 2 heures et 6 minutes après l'augmentation du degré de NicolasSarkozy. Cela indique soit que les deux auteurs réagissent à un même évènement médiatique extérieur soit que le deuxième auteur réagit à ce qu'annonce le premier.

Cette anomalie pourrait être détectée avec la quantité d'interactions des liens ponctuels en considérant une fenêtre temporelle d'agrégation plus grande, ou en prenant une fenêtre temporelle mobile. Cependant, chaque anomalie ayant une dynamique propre, à moins d'itérer la méthode en considérant plusieurs tailles de fenêtres, ces deux solutions mènent à une perte de précision, si la fenêtre est trop grande, ou une diminution du nombre d'anomalies détectées, si la fenêtre est trop petite. Dans le flot de liens avec durée, le degré des nœuds est mis à jour dès qu'ils changent de voisinages, de ce fait, les anomalies détectées sont beaucoup moins sensibles à la durée des liens δ.

3.6.2 Contexte agrégatif

Dans le contexte précédent, les valeurs observées sont comparées à une valeur attendue unique quelque soit le nœud et l'instant sous étude. Avec un contexte agrégatif, on peut considérer le degré d'un nœud-temps vis-à-vis du degré à l'instant correspondant ou vis-à-vis du degré du nœud.

Dans le contexte agrégatif sur les nœuds, la valeur attendue est le degré à l'instant t,

$$d_{exp}(t, v) = \frac{1}{|V|} \sum_{v' \in V} d(t, v') \, .$$

Le calcul des valeurs de déviation dans ce contexte s'avère être très long, en particulier lorsque la précision temporelle est fine. En effet, dès lors que le degré d'un nœud change, la valeur attendue à l'instant correspondant doit être mise à jour. De ce fait, chaque nœud possède à chaque instant une valeur attendue différente.

Dans le trafic IP, à l'exception de quelques pics d'activités, le degré au cours du temps se distribue homogènement (voir Section 2.3.2). Ainsi, étant donné l'hétérogénéité des nœuds-temps vis-à-vis du degré, la distribution des valeurs de déviation est similaire à la précédente et met en évidence les mêmes anomalies. Dans le cas de Twitter, le degré au cours du temps présente les mêmes variations quotidiennes et hebdomadaires que la quantité d'interactions (voir Section 2.4.2). Ainsi, en plus des anomalies détectées dans le contexte basique, les valeurs de déviation extrêmes mettent en évidence des utilisateurs ayant un degré parmi les plus élevés durant des périodes de faible activité (voir Figure 3.21).

Dans le contexte agrégatif sur le temps, la valeur attendue est le degré du nœud v,

$$d_{exp}(t, v) = \frac{1}{|T|} \int_{t' \in T} d(t', v) \, dt' \, .$$

On remarque sur la Figure 3.22, que considérer le degré instantané des nœuds relativement à leur activité habituelle normalise la distribution des valeurs de déviation. En effet, pour qu'une valeur de déviation soit élevée, il faut que la différence entre la valeur attendue et la valeur observée soit significative. Par conséquent, plus le degré moyen d'un nœud est

a) Distribution des valeurs de déviation b) Profil de degré d'un utilisateur anormal

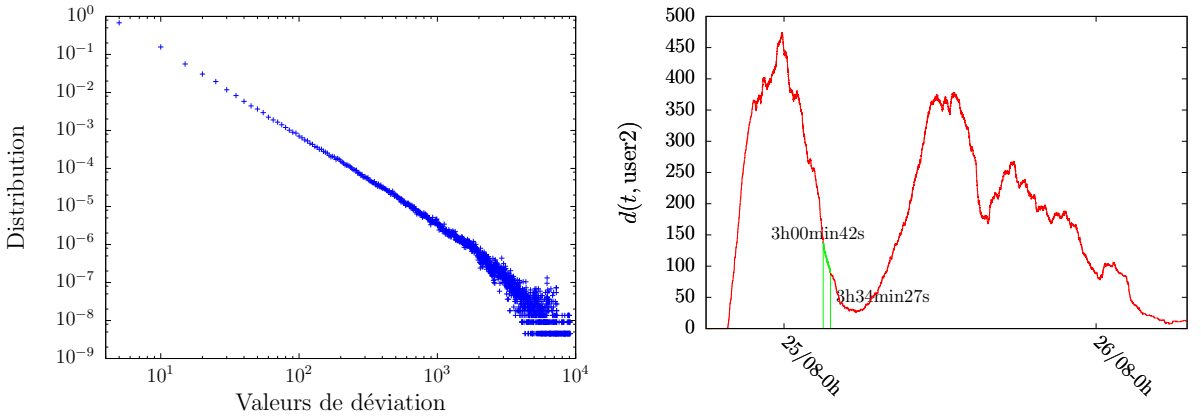


FIGURE 3.21 – **Anomalies dans le contexte agrégatif sur les nœuds sur Twitter** – a) La distribution des valeurs de déviation est hétérogène. b) Elle met en évidence des utilisateurs ayant un degré anormal durant des périodes de faible activité. Dans l'exemple user2 est anormal le 24 août de 3h00min42s à 3h34min47s (en pointillés verts).

élevé, plus sa différence avec le degré instantané doit être importante afin que la valeur de déviation correspondante soit élevée. Inversement, pour obtenir une valeur de déviation identique, plus le degré moyen d'un nœud est faible moins la différence doit être élevée. De ce fait, le contexte agrégatif permet d'homogénéiser les comportements des nœuds ayant des degrés d'ordres de grandeur différents.

a) Trafic IP b) Twitter

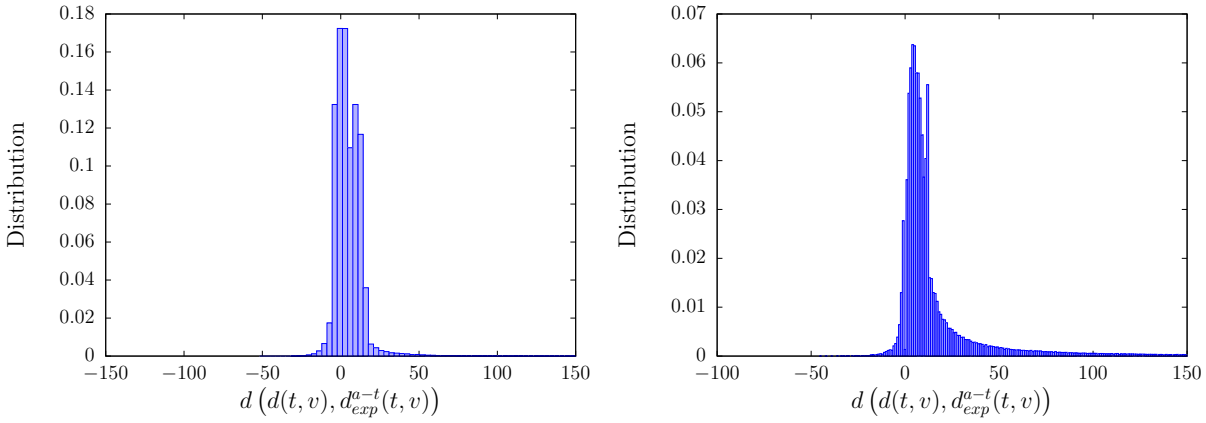


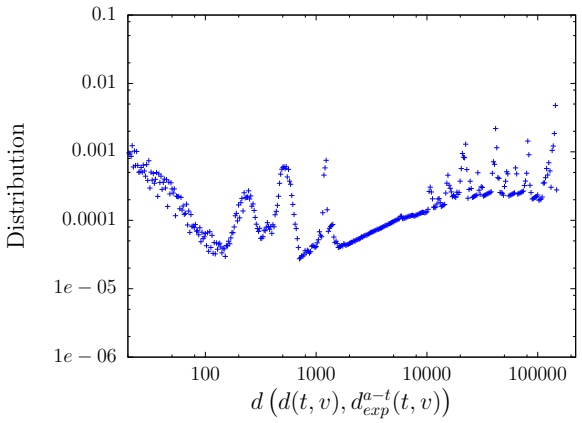
FIGURE 3.22 – **Zoom sur la partie normale de la distribution des valeurs de déviation dans le contexte agrégatif sur le temps.**

Dans les deux jeux de données, on détecte ainsi de nouveaux nœuds-temps dont le degré est moins élevé que celui des nœuds-temps précédemment identifiés. Sur Twitter, ces nœuds correspondent la plupart du temps à des auteurs anonymisés qui reçoivent une grande quantité d'attention sur une durée très limitée. Cependant, les valeurs de déviation extrêmes correspondent toujours aux nœuds de la Figure 3.20 : non seulement leur degré

moyen est faible mais, en plus, la différence entre leur degré moyen et le degré observé au moment des pics d'activité est très importante.

Ainsi, le contexte agrégatif sur le temps permet d'homogénéiser les comportements normaux des nœuds-temps (t, v) dont la différence entre le degré et le degré de v, moyenné sur le temps, est non-significative. L'identification du comportement normal permet ensuite l'identification des comportements anormaux. Comme on le voit sur la Figure 3.23, la multiplicité des comportements anormaux, liée à la multiplicité des valeurs attendues et à la diversité de l'ensemble des nœuds-temps, mène à des distributions à queues lourdes. Or, dans de telles distributions, l'utilisation de notre critère de détection, selon lequel une valeur est anormale si elle dévie de plus de trois écarts-types de la moyenne, mène à 10% d'anomalies dans Twitter et 7.5% dans le trafic IP, ce qui amène à s'interroger quant à la validité de notre modèle de comportement gaussien.

a) Trafic IP



b) Twitter

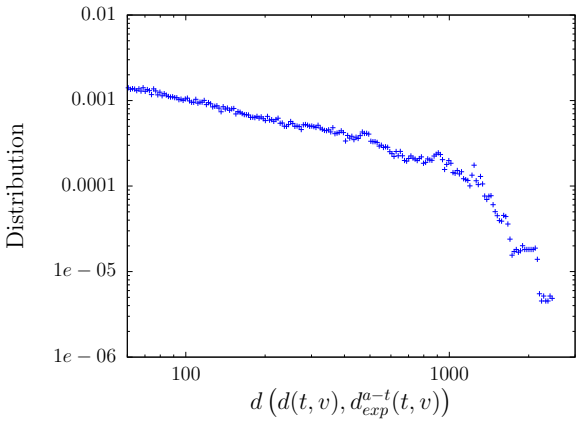


FIGURE 3.23 – **Zoom sur la partie anormale de la distribution des valeurs de déviation dans le contexte agrégatif sur le temps (échelle logarithmique).**

3.7 Conclusion

Dans ce chapitre, nous avons montré que l'utilisation de différents contextes mène à la détection de différentes anomalies. Nous avons également montré qu'à partir d'anomalies détectées dans un contexte basique global, l'utilisation de contextes plus complexes et plus locaux nous permet de nous centrer sur ces anomalies et d'en préciser la nature. L'identification d'anomalies subtiles et pertinentes par l'exploration méthodique des interactions selon leurs dimensions d'intérêt nous a ainsi permis d'améliorer notre compréhension des données, et de dresser une image plus complète de la façon dont des millions d'interactions s'organisent dans le temps. Notre méthode s'applique aux interactions temporelles en général. Ainsi, comme nous l'avons illustré dans la Section 3.5.1, elle pourrait être utilisée dans le cadre de nombreuses autres applications, notamment la caractérisation de

l'utilisation du second écran au travers des médias sociaux et la prédiction de liens.

Nous avons appliqué notre méthode d'une part sur des liens ponctuels, en comptant leur nombre d'occurrence dans une fenêtre de temps, et d'autre part en calculant le degré instantané des nœuds sur un flot de liens avec durée. Le degré d'un nœud étant mis à jour dès lors qu'il change de voisinage, nous détectons, grâce à ce dernier, les anomalies selon la dynamique qui leur est propre, et non selon une fenêtre de temps fixe pour l'ensemble des nœuds.

Cependant, nous avons également observé que tenir compte avec exactitude du comportement de chaque nœud à chaque instant en utilisant une propriété des flots de liens, étant donné la diversité des comportements, mène à des distributions hétérogènes. Considérer les anomalies comme étant les valeurs extrêmes dans ces distributions amène à constamment détecter le même ensemble d'anomalies évidentes. D'autre part, considérer que le comportement normal est caractérisé par une moyenne et un écart type conduit à détecter plus de 7% d'observations anormales. Ces considérations nous amènent donc à nous interroger sur la façon dont caractériser le comportement normal dans ce type de distribution. Ce sera l'objet du chapitre suivant.

Chapitre 4

Anomalies dans une séquence de distributions hétérogènes

Sommaire

Résumé : Nous montrons que, bien que le degré instantané des nœuds suive une distribution très hétérogène, difficile à modéliser, cette distribution est stable dans le temps. Nous concevons une méthode qui exploite la stabilité de cette hétérogénéité pour la détection des anomalies. Premièrement, nous divisons le flot de liens en fenêtres de temps et calculons la distribution des degrés dans chaque fenêtre. En comparant ces distributions, nous détectons des classes de degrés et des fenêtres de temps telles que le fait d'avoir un degré de cette classe au cours de cet intervalle est suspect. En utilisant ces informations, nous identifions finalement des nœuds-temps anormaux. En supprimant ces anomalies des données d'origine, nous validons notre

détection en remarquant que l'on revient à un flot de liens normal vis-à-vis du degré au cours du temps.

Notre objectif est de trouver des nœuds-temps dont le degré est anormal. Cependant, nous avons remarqué dans le chapitre précédent que la diversité des comportements des nœuds-temps vis-à-vis du degré conduisait à des distributions hétérogènes dans lesquelles l'utilisation du modèle gaussien est à proscrire. L'hétérogénéité de la distribution ne se rencontre pas uniquement avec le degré des nœuds-temps. Par exemple, la Figure 4.1 montre que c'est le cas également de la distribution de la variation du degré des nœuds-temps. Une solution serait de rechercher des contextes plus complexes et locaux de façon à normaliser les comportements. Par exemple, comparer uniquement les entités ayant des activités similaires dans un contexte restreint ou construire des valeurs attendues basées sur l'agrégation de communautés. On pourrait également continuer à rechercher des propriétés par rapport auxquelles les nœuds-temps ont un comportement homogène pour pouvoir y détecter des anomalies.

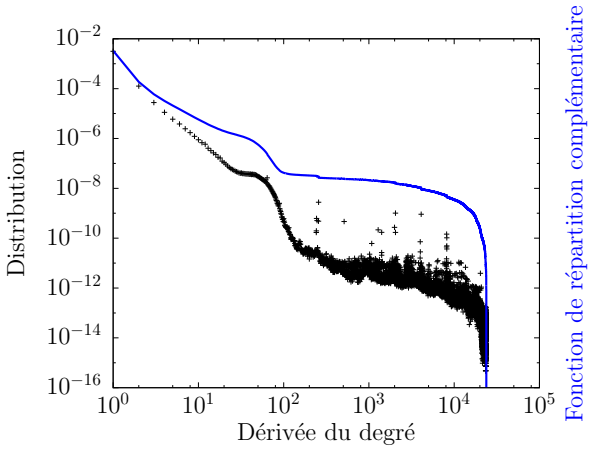


FIGURE 4.1 – **Distribution hétérogène de la dérivée du degré des nœuds-temps dans le flot de liens $\mathcal{L}_1$** – On trace la distribution de la dérivée discrète du degré $\frac{|d(t,v)-d(t-\Delta t,v)|}{\Delta t}$ sur l'ensemble de nœuds-temps avec $\Delta t = 1s$.

Une autre solution, que nous considérons dans ce chapitre, est de caractériser le comportement normal directement à partir de la distribution hétérogène de la propriété. L'état de l'art dans ce domaine est abondant mais également sujet à controverses (Section 4.1). Ainsi, partant de l'observation que la distribution de la propriété est stable dans le temps, nous proposons une nouvelle méthode permettant de détecter des entités anormales en comparant les distributions de la propriété sur différents intervalles de temps (Sections 4.2 et 4.3). Nous introduisons ensuite une technique de validation des résultats basée sur l'observation de la propriété agrégée le long des dimensions structurelles (Section 4.4). Comme dans le chapitre précédent, nous réalisons une série de choix visant à faciliter l'explication de notre méthode. Notamment, dans ce chapitre, nous nous concentrons sur la propriété du degré des nœuds-temps et nous nous plaçons dans un contexte basique. Dans ce contexte, les valeurs attendues sont une constante, ce qui revient à chercher des valeurs anormales directement dans la distribution des valeurs observées (voir Chapitre 3). On s'abstrait ainsi des valeurs attendues et des valeurs de déviations. On fournit des exemples basés sur la trace de trafic IP $\mathcal{L}_1$ de 1 heure, moins massive que les deux autres traces de trafic IP, et plus hétérogène que les interactions sur Twitter.

Cette phase d'explication achevée, on applique notre méthode aux trois autres jeux de données $\mathcal{L}_2$, $\mathcal{L}_3$ et $\mathcal{L}_4$ (Section 4.5). Dans le cas des flots de liens $\mathcal{L}_2$ et $\mathcal{L}_4$, dont l'étendue est plus grande, nous serons amenés à utiliser un contexte agrégatif. Finalement, nous étudions l'influence des paramètres utilisés (Section 4.6) puis concluons et présentons les perspectives de ces travaux (Section 4.7).

4.1 Détection d'anomalies dans des distributions hétérogènes

Dans cette section, nous définissons formellement les distributions hétérogènes. On s'intéresse ensuite aux façons de détecter des anomalies dans de telles distributions.

4.1.1 Distributions hétérogènes

Si nous avons défini les distributions homogènes avec anomalies comme étant des distributions correctement modélisées par une loi normale après suppression des anomalies, nous n'avons pas encore défini formellement ce qu'était une distribution hétérogène.

Soit $f(k)$ la probabilité qu'un nœud-temps pris au hasard ait un degré k dans le flot de liens L. Pour tous $k \in \mathbb{N}$, $f(k)$ est la fraction des couples $(t, v) \in T \times V$ pour lesquels $d(t, v) = k$:

$$f(k) = \frac{|\{(t, v) \in T \times V \ : \ d(t, v) = k\}|}{|T \times V|}.$$

Dans les Chapitres 2 et 3, nous avons qualifié cette distribution d'hétérogène dans les flots de liens $\mathcal{L}_1$ et $\mathcal{L}_4$, étant donné le fait qu'elle est représentative d'une multitude de comportement très différents les uns des autres. Notamment, k s'étend sur plusieurs ordres de grandeur, et plus k est élevé, plus $f(k)$ est faible (voir Figures 2.6.b et 2.7.b). On dit aussi que le flot de liens est *invariant d'échelle*, faisant référence au fait qu'aucune échelle ne le caractérise, quelques nœuds ayant beaucoup plus de connexions que d'autres.

Formellement, nous considérons qu'une distribution hétérogène est une distribution à queue lourde. La queue d'une distribution P traduit son comportement en des valeurs éloignées de sa moyenne. De ce fait, elle peut être décrite par la fonction de répartition complémentaire $\bar{F}(x) = P(X > x)$. Il est possible de comparer les queues de deux distributions P_1 et P_2 en comparant le comportement de leurs fonctions de répartitions complémentaires $\bar{F}_1$ et $\bar{F}_2$ quand x tend vers l'infini [6]. Si

$$\lim_{x \to +\infty} \frac{\bar{F}_1(x)}{\bar{F}_2(x)} = \infty \, ,$$

on dit que P_1 a une queue plus lourde que celle de P_2, ce qui signifie que la probabilité d'obtenir de très grandes valeurs dans P_1 est plus élevée que dans P_2. Nous considérons ici qu'une distribution à queue lourde est une distribution ayant une queue plus lourde que la loi normale [1] (voir Figure 4.2).

1. Traditionnellement, les distributions à queues lourdes sont définies comme n'étant pas exponentiellement bornées, ce qui revient à dire qu'elles ont une queue plus lourde que la loi exponentielle. Nous choisissons ici un critère moins restrictif.

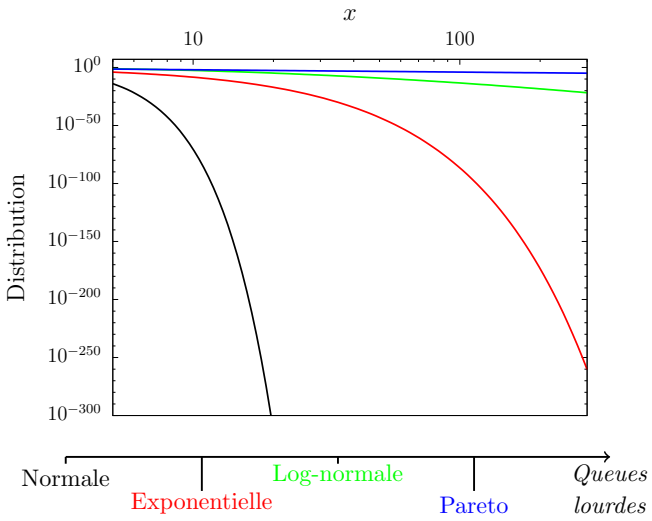


FIGURE 4.2 – **Distributions à queues lourdes** – On trace les distributions des lois normale ($m = 1, \sigma = 0.5$), exponentielle ($\lambda = 2.0$), log-normale ($m = 1, \sigma = 0.5$), et Pareto ($k = 1$, $x_{min} = 1$) sur l'intervalle $x \in [5, 300]$. Plus la distribution décroît lentement, plus sa queue est lourde.

De cette définition, on comprend aisément que l'utilisation du critère $m + 3\sigma$, adapté aux distributions homogènes, conduit à un nombre élevé d'anomalies dans les distributions hétérogènes comme nous l'avons remarqué lors du chapitre précédent.

4.1.2 Valeurs extrêmes de la distribution

Une première façon de détecter des anomalies dans les distributions hétérogènes consiste alors à déterminer un nouveau seuil à partir duquel les observations sont anormales.

Hubert et al. [77] proposent une nouvelle règle d'identification des anomalies en introduisant une boîte à moustaches ajustée en fonction du degré d'asymétrie de la distribution. Comme c'est le cas dans une distribution homogène avec le seuil $m+3\sigma$, après suppression des anomalies, environ 0.7% des observations se situent en dehors de l'intervalle délimité par les deux moustaches.

Verardi et al. [143] proposent une alternative à cette méthode. Ils mesurent le caractère extrême de chacune des observations en utilisant la boîte à moustache ajustée de Hubert et al. [77]. Ils normalisent les valeurs obtenues, puis les ajustent par une distribution de Tukey g-et-h, qui est une transformation de la loi normale permettant d'introduire une asymétrie et une queue lourde via les paramètres g et h. Finalement, ils calculent les quantiles de cette distribution et, par transformation inverse de ces derniers, obtiennent le seuil d'anormalité désiré.

Sur le même principe, Klebanov et al. [88] appliquent à la distribution P la transformation suivante,

$$P'(x) = (1 - p)P(x) + pH(x) \, ,$$

où $p \in (0, 1)$ et où $H(x)$ est la fonction de Heaviside. De cette façon, ils diminuent la variance et « abaissent la queue de la distribution », leur permettant ainsi de détecter des anomalies.

Cependant, les valeurs extrêmes ne sont qu'une sous-catégorie d'anomalies. Comme nous l'avons remarqué lors du chapitre précédent avec le degré, elles mettent toujours en évidence les mêmes nœuds-temps anormaux qui sont ceux ayant les degrés les plus élevés. Dans ce chapitre, nous voulons avoir accès à des anomalies plus subtiles sans changer de propriété. De ce fait, les méthodes ci-dessus ne sont pas adaptées.

4.1.3 Ajustement de la distribution par une loi de puissance

Une autre méthode consiste à ajuster la distribution par un modèle puis à calculer pour chaque point de la distribution empirique sa probabilité d'être généré par le modèle ; une anomalie étant un point dont la probabilité est très faible. Dans la littérature, la loi la plus communément adoptée pour modéliser les distributions hétérogènes est la loi de puissance.

État de l'art

L'équipe de A. Barabási [16] remarque dès 1999 avec le graphe du *World Wide Web* que la distribution du degré des nœuds tracée graphiquement en échelle log-log est proche d'une droite. Ils en déduisent que la queue de la distribution est distribuée selon une loi de puissance telle que $P(k) \approx k^{-\alpha}$ et déterminent l'exposant α en calculant la pente de la droite. Ils publient ensuite un article selon lequel les lois de puissance décrivent également le graphe de collaboration des acteurs, les réseaux électriques et le graphe de citations des publications scientifiques [21]. Dans la longue série d'articles qui en découle, on trouve deux formulations différentes : soit il est dit que le degré des nœuds dans le graphe suit une loi de puissance, soit que la distribution du degré est telle que $P(k) \sim k^{-\alpha}$, ce qui signifie que les deux fonctions ont le même comportement quand k tend vers l'infini.

Dans son livre, M.E.J. Newman [111] caractérise les distributions hétérogènes du degré par la loi de puissance

$$P(k) = ck^{-\alpha},$$

où c est une constante positive. Or, en adoptant ce modèle, une analyse statistique de près de 1 000 graphes réels réalisée par l'équipe de A. Clauset [39, 146] a montré que pour environ 65% des graphes, aucune loi de puissance ne convenait à expliquer la distribution des degrés. Selon leurs tests, cela ne signifie pas que les 35% restants suivent une loi de puissance, mais que cette dernière n'a pas été exclue. Notamment, ils montrent que dans 45% des cas, la loi log-normale est plus adapté que la loi de puissance.

Le contraste entre ces deux affirmations divergentes peut être expliqué de deux façons différentes. La première est que le manque de rigueur menant à l'identification de lois de puissance par la simple observation d'une droite en échelle log-log conduit à des résultats erronés. La deuxième est le manque de définition rigoureuse communément acceptée du comportement hétérogène. Notamment, Voitalov et al. [148] adoptent une définition moins restrictive. Ils considèrent la classe de distributions variant comme une loi de puissance,

$$P(k) = l(k)\, k^{-\alpha},$$

où $l(k)$ est une fonction variant lentement, de façon à ce que toute distribution de cette classe ait la même queue, mais puisse avoir des formes différentes pour des faibles degrés. Ils montrent ainsi que les distributions du degré entrant dans cette catégorie sont plus nombreuses que ce qu'envisage l'équipe de A. Clauset. R. Perline [117] quant à lui, déclare que modéliser ce type de distributions en considérant uniquement le comportement de leurs queues est une simplification excessive. Au contraire, il propose les modéliser par un modèle de mélange de lois de puissance (*mixture model* en anglais), en partant du principe que l'hétérogénéité provient de plusieurs sources différentes (par exemple, dans une population, elle peut être la conséquence à la fois de l'âge, du sexe, ou encore des revenus des personnes considérées).

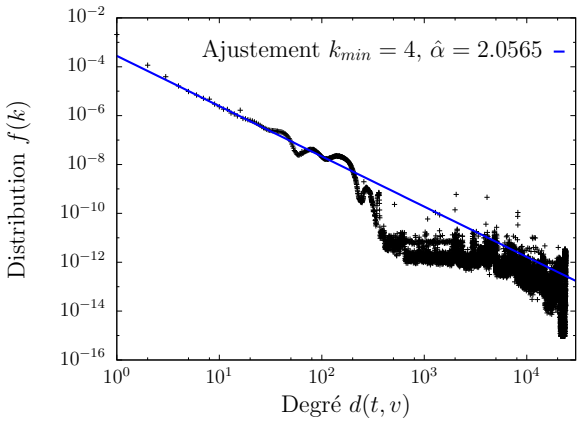
Application

Dans cette sous-section, on réalise un ajustement des distributions du degré des nœuds-temps dans les flots de liens $\mathcal{L}_1$ et $\mathcal{L}_4$ par une loi de puissance de la forme $P(k) = ck^{-\alpha}$, avec $\alpha > 1$ et $k_{\max} \geqslant k \geqslant k_{\min} > 0$.

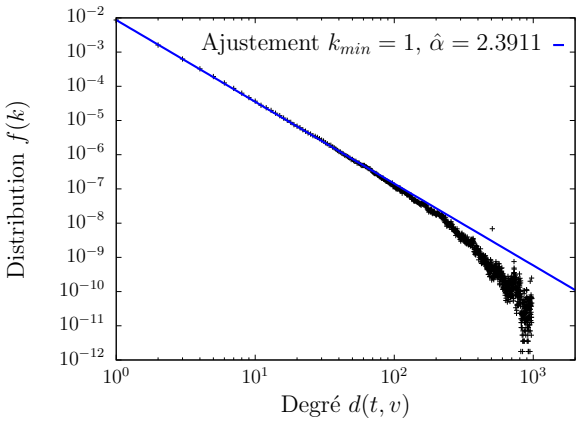
En utilisant l'expression analytique du maximum de vraisemblance de α fournie par Virkar et al. [146], on trouve $(\hat{\alpha}, k_{min}) = (2.0565, 4)$ dans $\mathcal{L}_1$ et $(\hat{\alpha}, k_{min}) = (2.3911, 1)$ dans $\mathcal{L}_4$ (voir Figure 4.3.a et 4.3.b). On teste la qualité de l'ajustement en suivant la procédure proposée par Clauset et al. [39]. On commence par calculer la distance de Kolmogorov-Smirnov (KS) D^* entre la fonction de répartition du modèle estimé et celle de la distribution empirique. Ensuite, on génère 2 500 ensembles de données synthétiques selon une loi de puissance de paramètres $(\hat{\alpha}, k_{min})$. Les résultats des Figures 4.3.c et 4.3.d montrent, aussi bien pour le trafic IP que pour Twitter, que l'ensemble des distances KS entre les données générées et leurs modèles estimés sont plus petites que celles entre la distribution empirique et son modèle estimé D^*. Par conséquent, dans les deux cas, les différences entre la distribution empirique et le modèle estimé ne peuvent pas être attribuées à des fluctuations statistiques, ce qui nous conduit à rejeter l'hypothèse selon laquelle les degrés dans $\mathcal{L}_1$ et $\mathcal{L}_4$ sont distribués selon une loi de puissance.

La distribution du degré dans un flot de liens présente une richesse structurelle qui nécessite l'utilisation de nouveaux outils. Modéliser la distribution $f(k)$ par une loi de la forme $l(k)\, k^{-\alpha}$ comme le font Voitalov et al. [148] comporte des risques, notamment ceux de sur- ou sous-ajuster la distribution. D'une part, en correspondant trop étroitement à l'ensemble de données, le modèle ne permettra pas l'identification d'anomalies (faux-négatifs). D'autre part, en étant trop général, il aura tendance à détecter des nœuds-temps normaux comme étant anormaux (faux-positifs). Nous adoptons donc ici une approche différente qui consiste à déduire le modèle de comportement normal par l'observation des distributions du degré sur des fenêtres de temps successives.

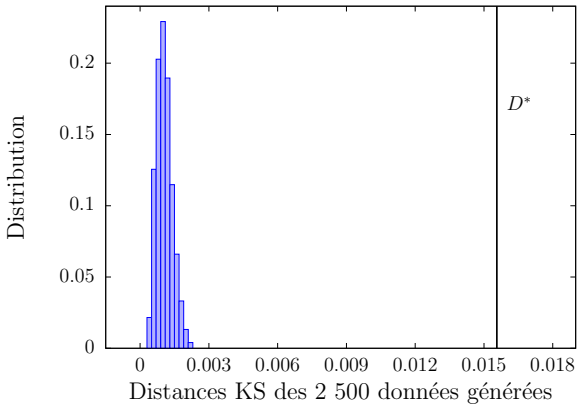
a) Ajustement Trafic IP



b) Ajustement Twitter



c) Qualité de l'ajustement Trafic IP



d) Qualité de l'ajustement Twitter

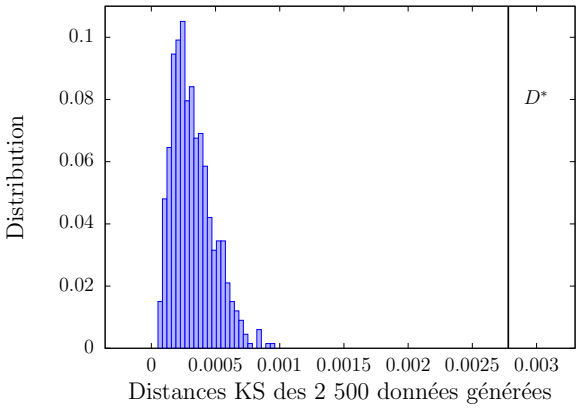


FIGURE 4.3 – **Ajustement des distributions par la méthode de Clauset et al. [39, 146]** – On trouve $D^* = 0.015562$ dans le trafic IP et $D^* = 0.002780$ dans les retweets. Ces distances sont plus élevées que l'ensemble des distances KS obtenues avec des données synthétiques suivant une loi de puissance $(\hat{\alpha}, k_{min})$.

4.2 Hétérogénéité structurelle, homogénéité temporelle

À partir de cette section, nous illustrons notre méthode à l'aide du flot de liens $\mathcal{L}_1$. Lors du Chapitre 2, nous avons fait l'observation que, bien qu'étant hétérogène sur les nœuds-temps, le degré était homogène avec anomalies sur le temps (voir Section 2.3.2). On émet alors l'hypothèse suivante :

Quels que soient les nœuds actifs au cours du temps,
la distribution hétérogène du degré est stable dans le temps.

Afin de vérifier cette hypothèse, nous observons les distributions du degré sur les sous-flots correspondant au trafic IP dans des fenêtres de temps de durée $\tau = 2.0$s. Formellement, nous appelons $T_i = [2i, 2i + 2[$ la $i^{\text{ème}}$ fenêtre temporelle, pour tout $i \in \{0, \ldots, 1799\}$, et nous définissons la distribution de son degré telle que,

$$f(T_i, k) = \frac{|\{(t, v) \in T \times V_i \ : \ d(t, v) = k\}|}{|T \times V_i|}.$$

La Figure 4.4 montre les distributions du degré des quatre premières fenêtres temporelles. Ces dernières sont toujours hétérogènes, mais également très semblables, ce qui conforte notre affirmation.

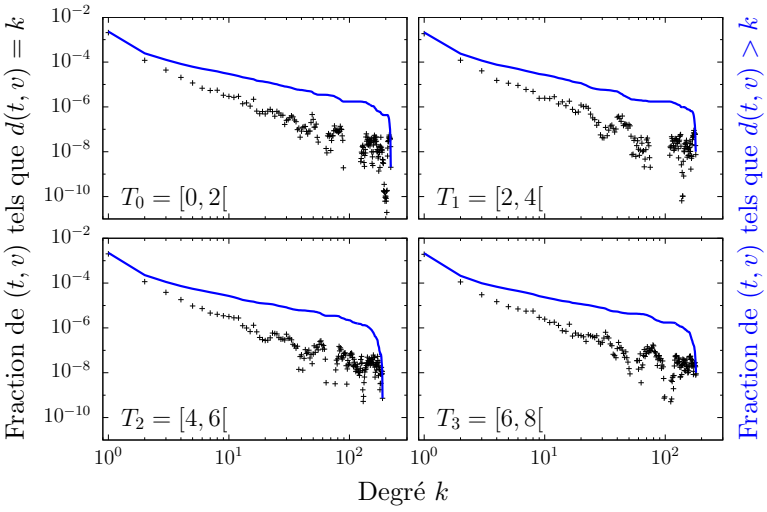


FIGURE 4.4 – **Distributions du degré sur des fenêtres temporelles de 2 secondes** – Pour $T_0 = [0, 2[$, $T_1 = [2, 4[$, $T_2 = [4, 6[$ et $T_3 = [6, 8[$, nous calculons le degré $d(t, v)$ pour tout (t, v) dans le sous-flot correspondant et traçons la distribution de l'ensemble de valeurs $\mathcal{O}_i = \{d(t, v) : (t, v) \in T_i \times V\}$, pour $i = \{0, 1, 2, 3\}$. La fraction exprime la probabilité de tirer un instant $t \in T_i$ et un nœud $v \in V$ tel que $d(t, v) = k$.

Pour quantifier cette similarité sur l'ensemble de la trace, nous effectuons des KS-tests à deux échantillons sur toutes les paires de distributions $(f_i, f_j)_{i \neq j}$ [121]. En fonction de la position relative entre la distance KS entre f_i et f_j, notée $D_{i,j}$, et une valeur critique c, ce test détermine si les deux échantillons, $\mathcal{O}_i = \{d(t, v) : (t, v) \in T_i \times V\}$ et $\mathcal{O}_j = \{d(t, v) : (t, v) \in T_j \times V\}$ sont susceptibles de provenir de la même distribution. Soit n_i la taille de l'échantillon $\mathcal{O}_i$, avec un intervalle de confiance à 90%,

$$c = 1.073\sqrt{\frac{n_i + n_j}{n_i n_j}} \quad [7].$$

La Figure 4.5 montre la distribution du ratio entre $D_{i,j}$ et c pour toutes les paires $(f_i, f_j)_{i \neq j}$. On voit que la plupart des valeurs sont inférieures à 1, ce qui signifie que la plupart des distances KS sont inférieures à c. Sur la base du KS-test à deux échantillons, on en déduit que les degrés sont distribuées de manière similaire sur la plupart des fenêtres temporelles. Par ailleurs, 1% des ratios sont supérieurs à 1. Ils sont le résultat de la comparaison de quelques distributions divergentes avec le reste des distributions et mettent en évidence la présence d'anomalies dans les sous-flots correspondants.

Ainsi, nous avons montré que les distributions du degré sont hétérogènes de la même manière sur la plupart des fenêtres temporelles. De cette façon, nous avons réussi à caractériser un comportement normal : il est normal que la fraction des couples (t, v) ayant un degré k dans la fenêtre de temps T_i soit similaire à cette fraction dans les autres fenêtres de temps, et inversement, il est anormal d'observer un écart significatif par rapport à cette fraction habituelle.

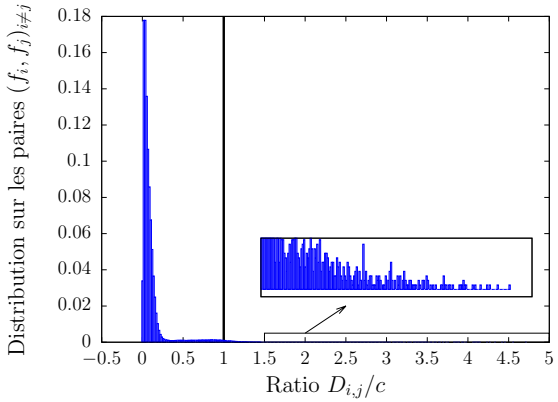


FIGURE 4.5 – **Similarité des distributions du degré sur des fenêtres temporelles de 2 secondes.**

4.3 Détection de fenêtres temporelles et de classes de degrés anormales

Partant de ces observations, nous concevons une méthode permettant d'isoler des anomalies en comparant structurellement les distributions du degré sur l'ensemble des fenêtres temporelles.

Afin d'atténuer les fluctuations observées dans la queue de la distribution, nous ne considérons pas exactement les distributions mais les histogrammes dans lesquels les valeurs du degré sont regroupées dans des classes logarithmiques. Dans des classes linéaires, les bornes de la classe $C_j = \{k_j, \cdots, k_{j+1}-1\}$ sont espacées régulièrement telles que $k_{j+1} = k_j + r$. Les bornes des classes logarithmiques, quant à elles, sont espacées régulièrement en échelle logarithmique telles que $\log(k_{j+1}) = \log(k_j) + r$. Étant donné que C_j est un ensemble d'entier, on notera dans ce cas $C_j = \{\lceil k_j \rceil, \ldots, \lfloor k_{j+1} \rfloor - 1\}$. Dans $\mathcal{L}_1$, le degré minimal est $k_1 = 1$, ainsi, en choisissant $r = 0.1$, on aboutit à 41 classes de degrés telles que $C_1 = \{1\}$, $C_2 = \{2\}$, $C_3 = \{3\}$, $C_4 = \{4, 5\}$, etc., jusqu'à $C_{41} = \{19953, \ldots, 25117\}$.

Soit $f(T_i, C_j)$ la fraction de nœuds-temps (t, v) ayant des degrés dans la classe C_j durant la fenêtre temporelle T_i :

$$f(T_i, C_j) = \frac{|\{(t, v) \in T_i \times V : \ d(t, v) \in C_j\}|}{|T_i \times V|}.$$

Pour comparer les distributions du degré, nous étudions comment la fraction des nœuds-temps qui ont un degré compris dans une classe donnée C_j durant T_i est répartie sur l'ensemble des fenêtres temporelles.

Ainsi, dans la classe de degrés C_j, les entités sont les fenêtres temporelles $\{T_i \ : i \in \{0, \ldots, 1799\}\}$ et la propriété est la fraction $f(T_i, C_j)$, de telle sorte que $\mathcal{O} = \{f(T_i, C_j) \ : i \in \{0, \ldots, 1799\}\}$ est l'ensemble des valeurs observées (voir Chapitre 3). Dans ce contexte, une fenêtre de temps T_i est anormale si sa fraction de nœuds-temps ayant un degré dans la classe C_j dévie significativement de celles observées pour les autres fenêtres de temps. [2]

2. Comme précisé dans l'introduction, le contexte basique ne requiert pas l'introduction de valeurs attendues et de valeurs de déviation.

Il y a autant de contextes différents qu'il y a de classes de degrés.

La Figure 4.6 montre les distributions de la fraction dans les classes C_1, C_2, C_{19}, C_{22}, C_{31} et C_{41}. Comme attendu étant donné l'homogénéité temporelle, les distributions sont homogènes avec anomalies : la plupart des fractions sont réparties autour d'une moyenne et quelques-unes seulement en sont éloignées [3]. De même, comme attendu étant donné l'hétérogénéité des degrés, plus la classe de degrés est élevée, plus la fraction de nœuds-temps au sein de cette classe est faible. Nous voyons dans C_1 que la fraction moyenne sur tous les intervalles de temps est $2.1 \cdot 10^{-3}$. Dans C_2, elle chute à $1.15 \cdot 10^{-4}$ et diminue progressivement pour atteindre 0 dans les classes de degrés supérieurs à 252. Dans ces classes de degrés élevés, le pic sur la fraction 0 indique que dans la plupart des fenêtres de temps, aucun nœud-temps n'atteint de tels degrés. Il s'agit là d'une particularité du flot de liens $\mathcal{L}_1$: un ou plusieurs nœuds peuvent avoir un degré élevé constant et conduire à une fraction moyenne non nulle (voir Section 4.5).

Parmi les classes de la Figure 4.6, nous détectons 151 fenêtres de temps anormales dans la première classe contenant uniquement le degré 1, 5 fenêtres de temps anormales dans C_2 et 12 dans C_{19}. Dans les classes C_{22}, C_{31} et C_{41}, les fenêtres de temps anormales sont celles pour lesquelles la fraction est supérieure à 0.

Par commodité, nous notons $\mathcal{A} = \{(T_i, C_j)\}$ l'ensemble des fenêtres anormales détectées dans les 41 classes de degrés, où (T_i, C_j) indique que la fenêtre de temps T_i est anormale dans la classe C_j. Ainsi, toutes classes de degrés confondues, nous détectons 1 358 couples (T_i, C_j) anormaux.

Après cette première étape, nous n'avons pas encore rempli notre objectif qui est d'identifier des nœuds-temps anormaux. Néanmoins, nous sommes maintenant en possession de deux informations cruciales qui permettent, pour chaque anomalie $(T_i, C_j) \in \mathcal{A}$, de cibler la recherche des nœuds-temps anormaux aux nœuds-temps de l'ensemble $\{(t, v) \in T_i \times V \ : d(t, v) \in C_j\}$.

4.4 Détection de nœuds-temps anormaux

En nous basant sur l'assertion suivante, « *le retrait d'une anomalie ne perturbe pas le comportement normal* », nous introduisons dans cette section une méthode de suppression itérative permettant d'identifier quels sont les nœuds-temps anormaux. Dans la suite, nous notons $\mathcal{I}_{(T_i,C_j)}$ l'ensemble de nœuds-temps anormaux correspondant à l'anomalie (T_i, C_j).

4.4.1 Méthode

Si une anomalie (T_i, C_j), nous permet de cibler la recherche des nœuds-temps anormaux sur l'ensemble $\{(t, v) \in T_i \times V \ : d(t, v) \in C_j\}$, elle ne suffit pas à affirmer qu'avoir un degré

3. Les ajustements sont effectués à l'aide de la procédure proposée par F. Grubb [61] énoncée dans la Section 2.5.3. Dans ces distributions, une valeur f est anormale si $f > m + 3\sigma$.

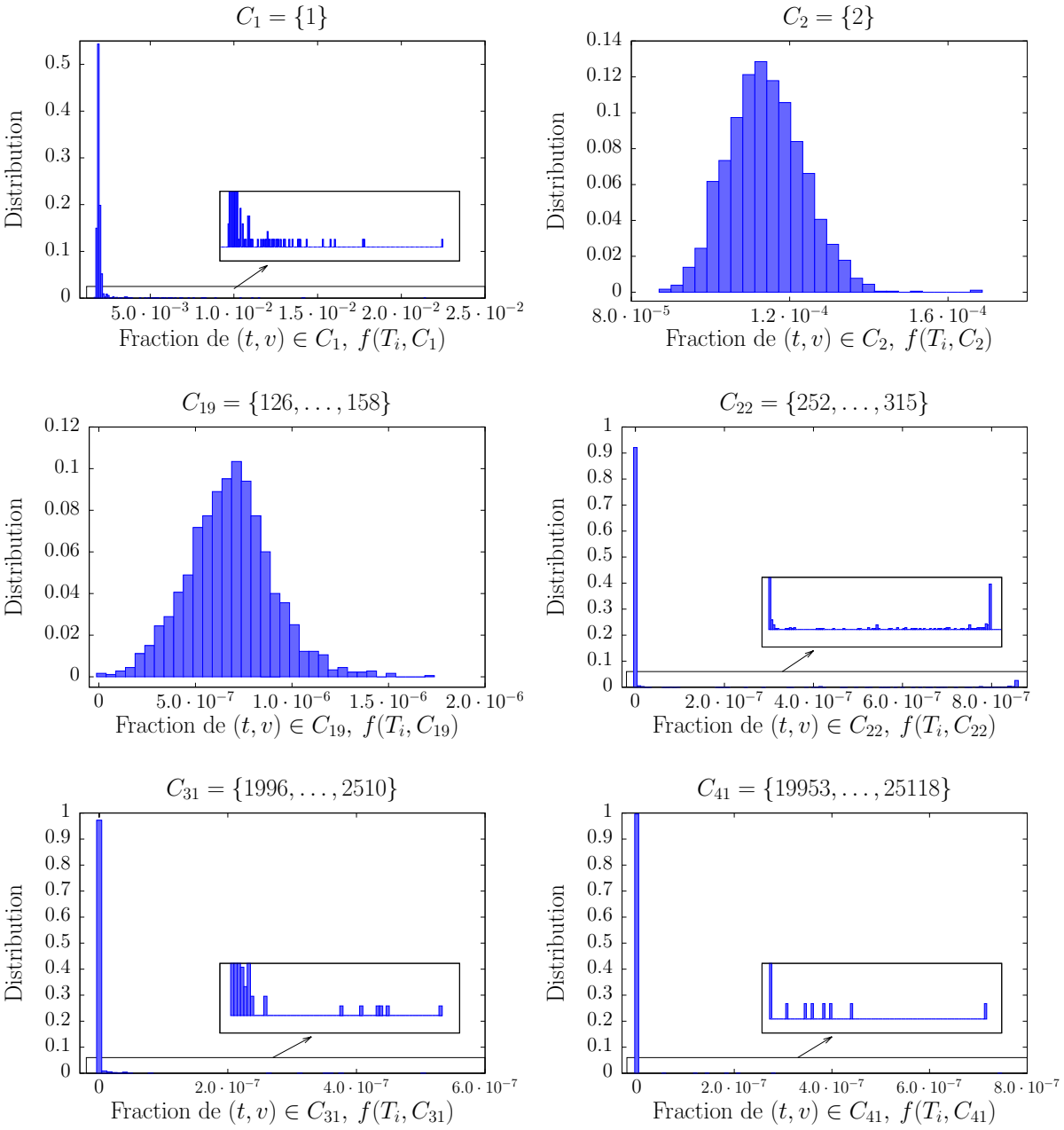


FIGURE 4.6 – **Distributions des fractions $f(T_i, C)$ sur l'ensemble des fenêtres temporelles T_i pour les classes de degrés C_j dans $\{C_1, C_2, C_{19}, C_{22}, C_{31}, C_{41}\}$** – Les distributions sur C_1, C_2 et C_{19} sont homogènes avec anomalies. Les distributions sur C_{22}, C_{31} et C_{41} ont un pic en zéro car, dans la plupart des fenêtres temporelles, il n'y a pas de nœuds-temps dans la classe correspondante.

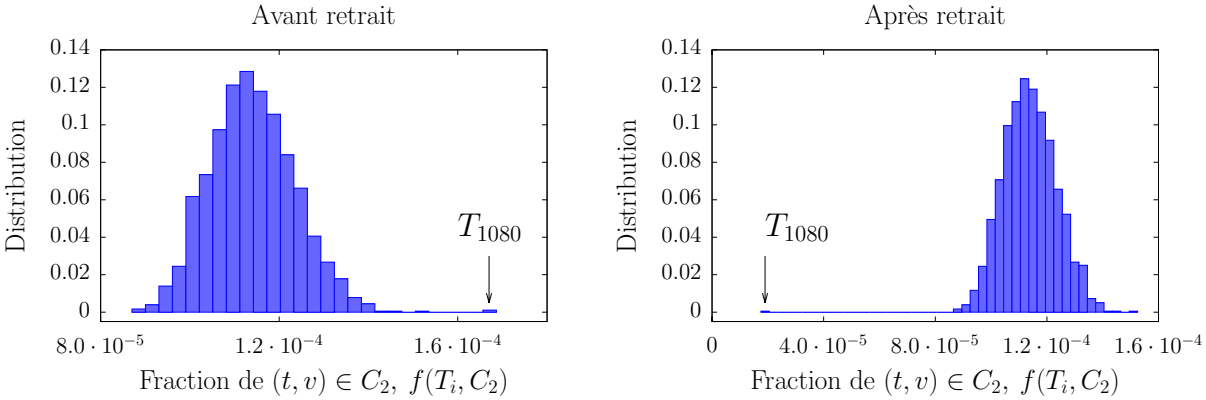


FIGURE 4.7 – **Identification erronée dans C_2** – La suppression de toutes les interactions (t, u, v) telles que $d(t, v)$ soit dans C_2 durant la fenêtre de temps T_{1080} provoque l'apparition d'une anomalie négative. La fraction résultante sur T_{1080} n'est pas nulle car la suppression de certaines interactions réduit le degré de nœuds-temps de classes supérieures qui finissent par avoir un degré dans C_2.

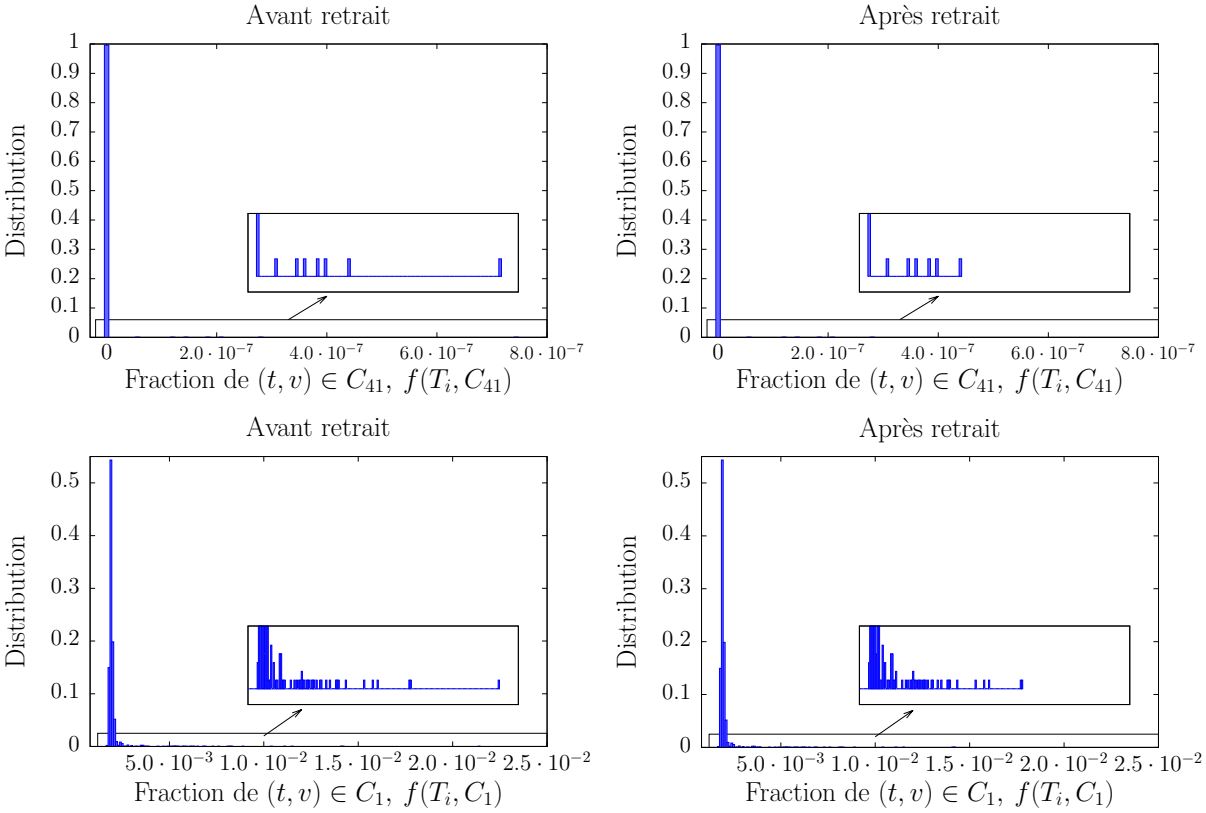


FIGURE 4.8 – **Identification des nœuds-temps anormaux dans les classes C_{41} et C_1** – Le retrait des nœuds-temps responsables de l'anomalie extrême de la classe C_{41} permet l'identification des nœuds-temps responsables de l'anomalie extrême de la classe C_1.

dans C_j au cours de T_i est anormal. Prenons par exemple l'anomalie (T_{1080}, C_2) et le flot de liens L' dans lequel nous avons supprimé les interactions impliquant les nœuds-temps ayant un degré dans C_2 durant T_{1080} :

$$L' = (T, V, E') \quad \text{avec} \quad E' = E \setminus \{(t, u, v) : t \in T_{1080} \text{ et } d(t, u) \in C_2 \text{ ou } d(t, v) \in C_2\}.$$

Nous voyons sur la Figure 4.7 que cette suppression provoque l'apparition d'une anomalie négative dans la distribution des fractions sur C_2.[4] Ainsi, en supprimant ces nœuds-temps, nous supprimons du trafic anormal, mais également du trafic normal. Ce retrait perturbe donc le comportement normal et, par conséquent, selon l'assertion énoncée en début de section, l'identification des nœuds-temps anormaux par $\mathcal{I}_{(T_i,C_j)} = \{(t, v) \in T_i \times V : d(t, v) \in C_j\}$ n'est pas validée.

Cette observation suggère qu'il est impossible d'identifier directement des nœuds-temps anormaux dans les classes dont la moyenne est non nulle. En effet, une fraction anormale dans ces classes est constituée de nœuds-temps anormaux, mais également de nœuds-temps normaux, ce qui nous empêche d'identifier ceux responsables de l'anomalie sans perturber le trafic normal.

Au contraire, dans les classes pour lesquelles la fraction moyenne est zéro, les nœuds-temps contribuant à des fractions non nulles sont explicitement anormaux. Dans ce cas, ils sont correctement identifiés par l'ensemble $\mathcal{I}_{(T_i,C_j)} = \{(t, v) \in T_i \times V : d(t, v) \in C_j\}$. Pour l'illustrer, considérons la classe C_{41}. Sa fraction anormale extrême correspond à la fenêtre temporelle T_{315}. La Figure 4.8 montre le résultat de la suppression de l'ensemble $\mathcal{I}_{(T_{315},C_{41})} = \{(t, v) \in T_{315} \times V : d(t, v) \in C_{41}\}$. Comme attendu, la fraction anormale dans C_{41} disparaît sans créer d'anomalie négative. Qui plus est, nous notons la disparition de l'anomalie (T_{315}, C_1). Cette répercussion s'explique par le fait que les nœuds v étaient, avant la suppression, liés à un nombre significatif de nœuds de degré 1. De ce fait, la suppression de l'ensemble $\mathcal{I}_{(T_{315},C_{41})}$ conduit également à l'identification des nœuds-temps responsables de l'anomalie (T_{315}, C_1) tels que

$$\mathcal{I}_{(T_{315},C_1)} = \{(t, u) \in N(t, v) \times T_{315} : d(t, v) \in C_{41} \text{ et } d(t, u) \in C_1\},$$

où $N(t, v)$ est l'ensemble des voisins de v au temps t.

> Par commodité, nous étiquetons les classes en fonction de leur type de distribution :
> – celles ayant une fraction moyenne nulle sont appelées *classes A*, car elles contiennent uniquement du trafic anormal,
> – celles ayant une fraction moyenne supérieure à 0, qui contiennent un trafic anormal et un trafic normal, sont appelées *classes AN*,
> – celles qui n'entrent dans aucune des deux catégories, à savoir celles qui ne sont pas homogènes avec anomalies, sont appelées *classes R*, pour classes rejetées.

Finalement, la 2ème étape de notre méthode de détection de nœuds-temps anormaux est la suivante. Pour chaque anomalie $(T_i, C_j) \in \mathcal{A}$ où C_j est une classe A, nous identifions les

4. Nous appelons anomalie négative une anomalie ayant une valeur inférieure à la moyenne : $f < m - 3\sigma$.

nœuds-temps anormaux par l'ensemble $\mathcal{I}_{(T_i,C_j)} = \{(t,v) \in T_i \times V \ : d(t,v) \in C_j\}$. Ensuite, nous supprimons les interactions les impliquant tel que, lors de la $n^{\text{ème}}$ suppression, le flot de liens résultant est $L_n(T,V,E_n)$ avec

$$E_n = E_{n-1} \setminus \{(t,u,v) : t \in T_i \text{ et } d(t,u) \in C_j \text{ ou } d(t,v) \in C_j\}, \quad \text{et} \quad E_0 = E.$$

En plus de supprimer le trafic anormal des classes A, cette suppression itérative permet également d'identifier les nœuds-temps anormaux des classes AN. De plus, si le retrait de $\mathcal{I}_{(T_i,C_j)}$ crée une anomalie négative dans une classe de degrés alors il est annulé et les nœuds-temps anormaux correspondant à l'anomalie (T_i, C_j) ne sont pas identifiés.

Dans $\mathcal{L}_1$, aucun retrait n'a engendré d'anomalies négatives. Au total, nous supprimons 205 anomalies dans les classes A. Ces retraits nous permettent d'identifier les nœuds-temps anormaux responsables de 1 163 anomalies sur les 1 358 détectées précédemment, soit plus de 85%. Pour ce faire, nous avons supprimé 7.4% du trafic (c'est-à-dire 7.4% des liens). La Figure 4.9 montre l'allure finale, après suppressions, des classes C_1 et C_2. On remarque la disparition de la quasi totalité des anomalies. Dans la Figure 4.10, on montre un exemple de 4 nœuds supprimés durant les périodes au cours desquelles ils ont un degré anormal. Selon Mazel et al. [105], les profils de degré de ces nœuds suggèrent qu'ils constituent une activité malveillante, notamment, le nœud v_3 atteint des degrés égaux à des puissances de deux, indiquant qu'il effectue des scans de réseau. Nous observons un comportement similaire autour du degré 256 pour le nœud v_1.

4.4.2 Validation

En plus des validations effectuées après chaque retrait en termes d'anomalies négatives, nous validons l'exactitude de notre méthode en examinant la conséquence des retraits sur le degré $d(t)$ à l'instant t.

La Figure 4.11.a montre qu'après la suppression des nœuds-temps anormaux identifiés par notre méthode, les pics et les changements soudains de tendance disparaissent, tandis que la moyenne du degré au cours du temps reste inchangée. De même, dans la Figure 4.11.b, on voit que toutes les anomalies disparaissent sans modifier la partie homogène de la distribution. Quantitativement, avant les retraits, on détecte une durée cumulée anormale de 158.1 secondes, contre seulement 2.4 secondes après l'application de notre méthode. D'autre part, la moyenne temporelle du degré à l'instant t passe de $1.706 \cdot 10^{-3}$ à $1.695 \cdot 10^{-3}$ après les retraits, soit une déviation de $1.1 \cdot 10^{-5}$%. Ainsi, si l'on s'en tient à l'assertion de départ selon laquelle les retraits des anomalies ne doivent pas perturber le comportement normal, ces résultats prouvent la validité des nœuds-temps anormaux détectés par notre méthode.

> Après l'étape 1, on sait que les nœuds-temps anormaux (t,v) sont liés aux anomalies $(T_i, C_j) \in \mathcal{A}$ tels que
>
> $$(t,v) \in \{(t,v) \in T_i \times V : d(t,v) \in C_j\}.$$

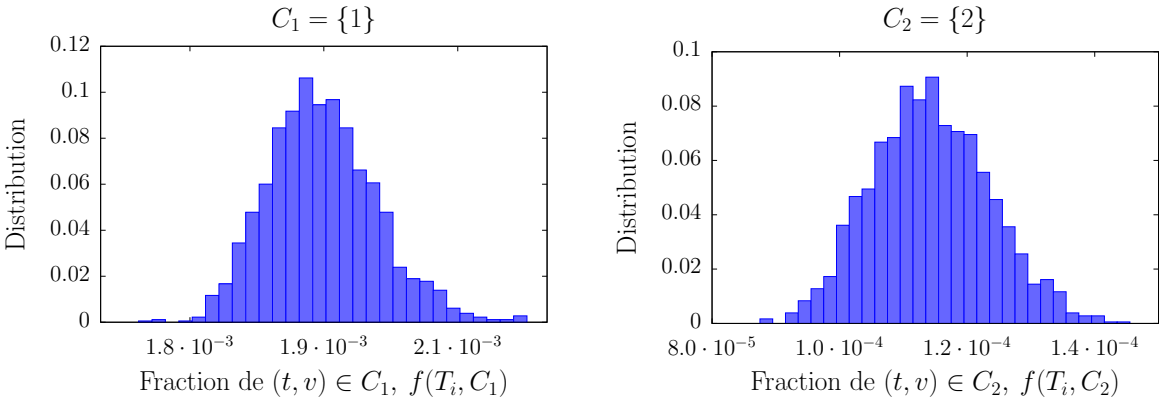


FIGURE 4.9 – **Distributions des fractions sur l'ensemble des fenêtres temporelles dans les classes C_1 et C_2 après suppression des événements** – Avant les suppressions, C_1 comporte 151 anomalies et C_2 5. Après les suppressions, il reste 10 anomalies non identifiées dans C_1 et 2 dans C_2.

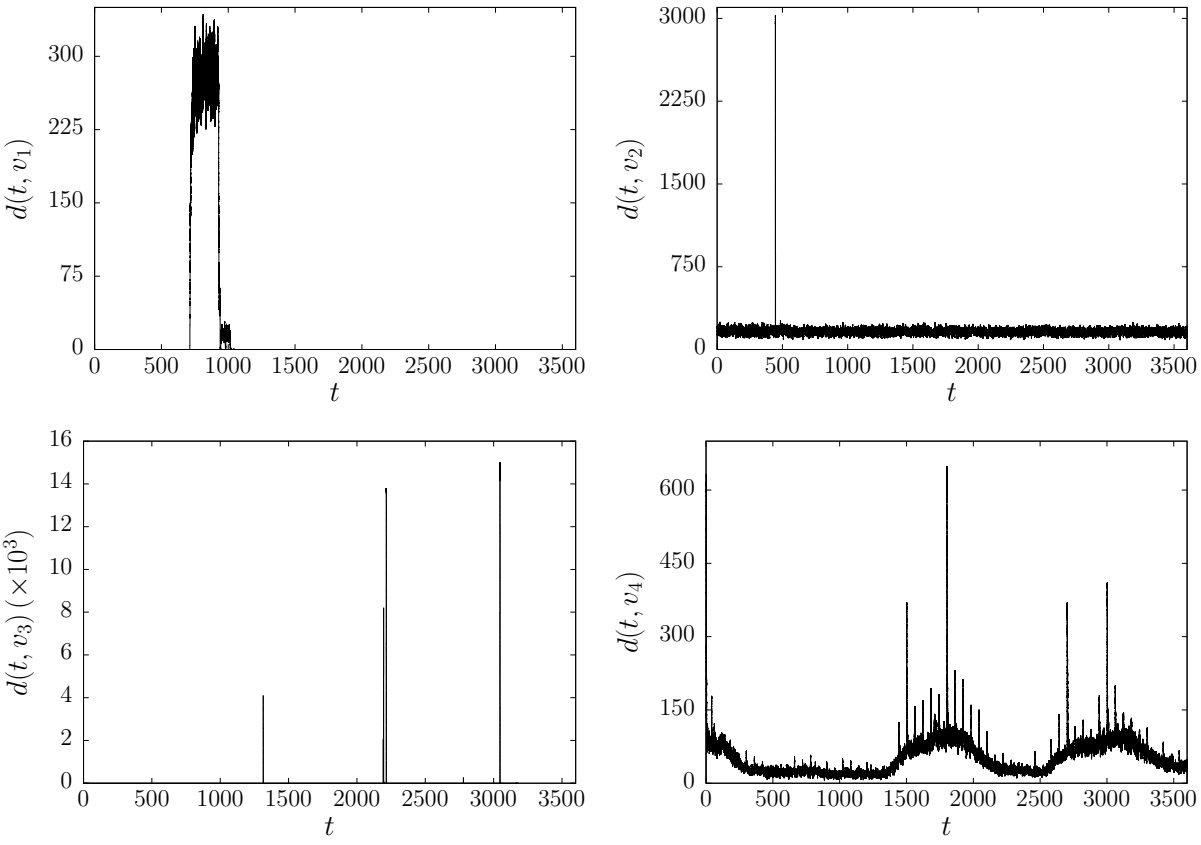
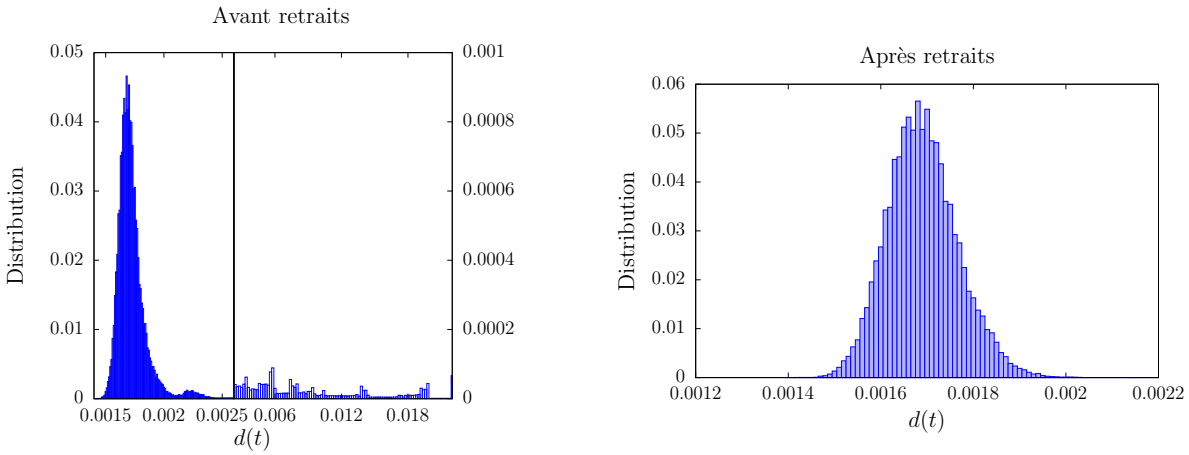


FIGURE 4.10 – **Profils de degré de quatre nœuds identifiés** – Les ensembles $\{(t, v_1) \in [712, 940[\times V\ :\ d(t, v) \in C_{22}\}$, $\{(t, v_2) \in [446, 448[\times V\ :\ d(t, v) \in C_{32}\}$ ont été identifiés et supprimés. De même, pour les ensembles $\{(t, v_3)\}$ dans lesquels v_3 est actif ainsi que les quatre pics pour lesquels le degré de v_4 est supérieur à 300.

a) Distribution du degré $d(t)$ à l'instant t avant et après les retraits dans $\mathcal{L}_1$.



b) Évolution du degré $d(t)$ à l'instant t avant et après les retraits dans $\mathcal{L}_1$.

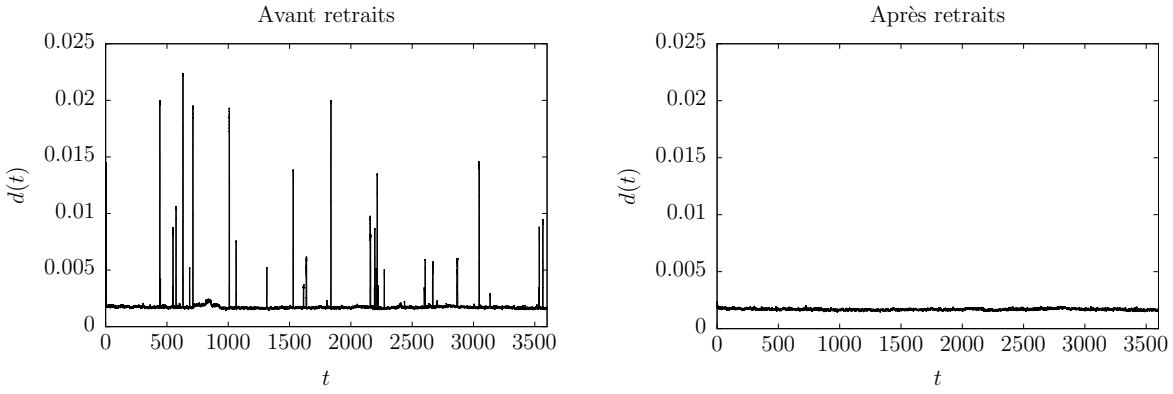


FIGURE 4.11 – **Conséquences des suppressions d'anomalies sur le degré au cours du temps** – Notre méthode permet de supprimer les anomalies identifiées sans impact significatif sur le trafic normal sous-jacent.

De ce fait, on restreint le contexte aux nœuds-temps de cet ensemble. On a donc un contexte par anomalie $(T_i, C_j) \in \mathcal{A}$. Les valeurs observées consistent en un score d'anormalité binaire $s(t, v) \in \{0, 1\}$, où (t, v) est anormal si $s(t, v) = 1$, normal sinon :
– Si C_j est une classe A et que le retrait de l'ensemble $\mathcal{I} = \{(t, v) \in T_i \times V : d(t, v) \in C_j\}$ entraîne la disparition de l'anomalie (T_i, C_j) sans créer d'anomalies négatives alors

$$(t, v) \in \mathcal{I} \Rightarrow s(t, v) = 1.$$

– Si C_j est une classe AN et que le retrait de l'ensemble $\mathcal{I} = \{(t', v') \in T_i \times V : d(t', v') \in C_{j'}\}$, entraîne la disparition de l'anomalie (T_i, C_j) sans créer d'anomalies négatives alors

$$(t, v) \in \{(t, v) \in T_i \times V : d(t, v) \in C_j \text{ et } \exists (t', v') \in \mathcal{I} \text{ t.q. } v \in N(t', v')\} \Rightarrow s(t, v) = 1.$$

– $s(t, v) = 0$ sinon.
Plus concrètement, un nœud-temps est anormal s'il est responsable de l'augmentation de la fraction dans une classe de degrés et dans une fenêtre de temps détectées lors de l'étape 1. À la fin de cette seconde étape, nous avons donc atteint notre objectif : nous avons détecté des nœuds-temps anormaux vis-à-vis du degré. Nous les avons de plus validés en vérifiant que leurs retraits n'altéraient pas le comportement normal.

4.5 Application aux autres jeux de données

Pour montrer la généralité et l'applicabilité de notre méthode, nous la testons sur les deux autres traces de trafic IP ainsi que sur les retweets de Twitter. Dans cette section, nous présentons les principaux résultats et différences observées avec ces jeux de données.

4.5.1 Trace de trafic IP d'une journée

Nous appliquons notre méthode au flot de liens $\mathcal{L}_2$ qui consiste en la trace de trafic IP longue d'une journée collectée le 25 juin 2013. Nous conservons une taille de fenêtre temporelle et une taille de classe de degrés identiques à l'expérience précédente.

La Figure 4.12.a montre la distribution de la fraction de nœuds-temps dans la classe C_2. Nous identifions trois comportements distincts liés en partie au rythme circadien. Afin de normaliser cette distribution, nous considérons un contexte agrégatif sur les nœuds (voir Section 3.3.1). Dans ce contexte, la valeur attendue du nœud-temps (t, v) est le degré $d(t)$ à l'instant t et les valeurs de déviation, basées sur le ratio et notées $\overline{d(t, v)}$, sont telles que

$$\overline{d(t, v)} = \frac{d(t, v)}{d(t)} .$$

Par commodité, nous nous y référons en tant que degré normalisé des nœuds-temps. Nous voyons sur la Figure 4.12.b que les distributions locales sur les fenêtres de temps sont similaires et sur Figure 4.12.c que la distribution globale du degré normalisé est hétérogène.

Ainsi, les deux contraintes requises pour appliquer notre méthode sont satisfaites.

Nous obtenons 34 classes de degrés allant de $C_1 = [1, 2[$ à $C_{34} = [3982, 5011[$. [5] [6] Parmi celles-ci, 3 classes sont rejetées car elles ne correspondent pas à une distribution homogène avec anomalies. Les 11 classes allant du degré $k = 1$ au degré $k = 26$ sont des classes AN. Les 20 restantes, allant du degré $k = 51$ au degré $k =$5 011 sont des classes A. Nous détectons 22 669 couples (T_i, C_j) anormaux et réussissons à identifier les nœuds-temps responsables de 63% d'entre eux. Pour ce faire, nous avons supprimé 8.7% du trafic.

Une fois encore, nous voyons sur la Figure 4.13 que le retrait des nœuds-temps anormaux entraîne la suppression des anomalies dans le degré au cours du temps. On passe d'une durée cumulée anormale de 1 681.74 secondes avant les retraits, à durée cumulée anormale de 679.71 secondes après les retraits. On remarque également que le trafic normal reste inchangé : la moyenne du degré à l'instant t est égale à $1.761 \cdot 10^{-3}$ avant les suppressions et à $1.760 \cdot 10^{-3}$ après, soit uniquement 10^{-6}% de déviation. Nous discutons des possibilités envisageables afin d'identifier l'origine des 679.71 secondes restantes dans le Chapitre 5.

4.5.2 Trace de trafic IP de 15 minutes : comparaison avec MAWILab

Nous appliquons maintenant notre méthode au flot de liens $\mathcal{L}_3$ qui consiste en la trace de trafic IP de 15 minutes du 3 novembre 2018. Cette trace est associée à une liste d'anomalies indexées par MAWILab [53] auxquelles nous pouvons comparer nos résultats. Étant donné l'étendue temporelle plus courte, nous prenons des fenêtres de taille $\tau = 1.0$s au lieu de $\tau = 2.0$s, afin d'en conserver un nombre important. La taille des classes de degrés reste inchangée.

Nous observons une distribution de degré globale hétérogène et des distributions de degrés locales similaires (voir Figures 4.14.a et 4.14.b). Nous obtenons 43 classes de degrés allant de $C_1 = \{1\}$ à $C_{43} = \{31623, \ldots, 39810\}$. [7] Parmi celles-ci, il y a 23 classes AN, 17 classes A et 3 classes R. Contrairement aux flots de liens précédents, trois classes AN sont des classes de degrés élevés : $C_{24} = \{399, \ldots, 502\}$, $C_{27} = \{795, \ldots, 1001\}$ et $C_{40} = \{15849, \ldots, 19953\}$. On voit sur la Figure 4.15 que cela est dû à trois nœuds qui ont un degré constant fluctuant dans chacune de ces classes et qui, de ce fait, forment le trafic normal observé dans chacune d'entre elles.

Dans ce flot de liens, plusieurs suppressions génèrent des anomalies négatives. Par exemple, le retrait des nœuds-temps $\mathcal{I}_{(T_{752}, C_{40})} = \{(t, v) \in T_{752} \times V \ : \ d(t, v) \in C_{40}\}$ crée une anomalie négative dans la classe C_1. Ces derniers correspondent au nœud v_3 durant son pic d'activité de $t_1 = 755, 3$ à $t_2 = 756, 5$ (voir Figure 4.15). Le comportement normal de v_3 étant d'être lié en permanence à 18 178 nœuds de degré 1 en moyenne, la suppression de son activité durant cette période génère une anomalie négative dans C_1. En effet, en plus de supprimer les interactions anormales de v_3, elle supprime également ses interactions

5. Nous redimensionnons le degré normalisé par une constante afin d'obtenir une plage de valeurs plus étendue. Le degré normalisé étant un nombre décimal, les classes sont des intervalles et non des ensembles d'entiers.

6. Le nombre de classes est moins (*resp.* plus) élevé que lors de l'expérience précédente étant donné la plus faible (*resp.* grande) étendue de valeurs.

7. Voir note 6.

a) Contexte basique

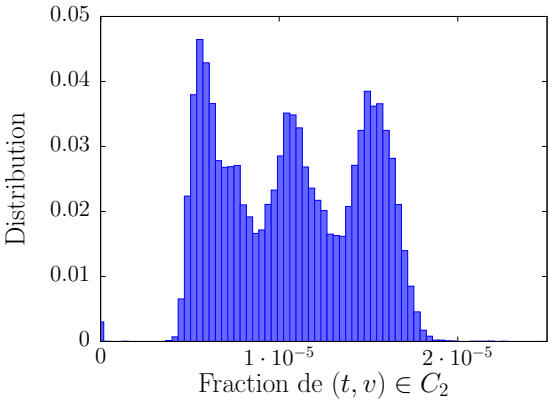
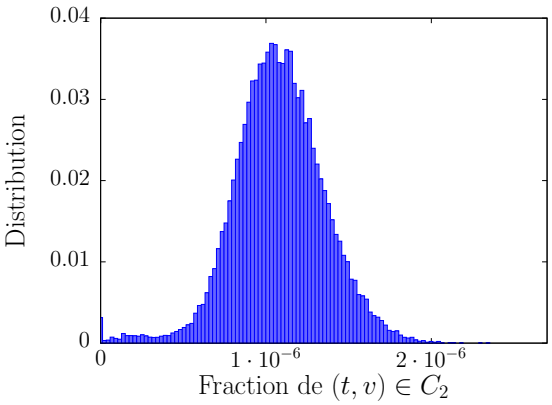
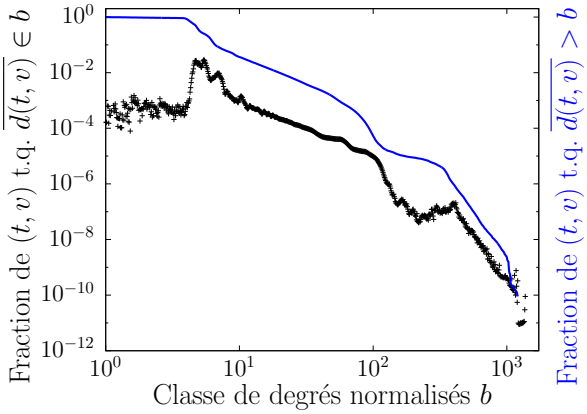


FIGURE 4.12 – **Homogénéité temporelle et hétérogénéité structurelle du degré normalisé dans $\mathcal{L}_2$** – Note : dans c., les valeurs du degré ont été regroupées par classes afin de lisser la distribution.

b) Contexte agrégatif sur les nœuds



c) Distribution du degré normalisé



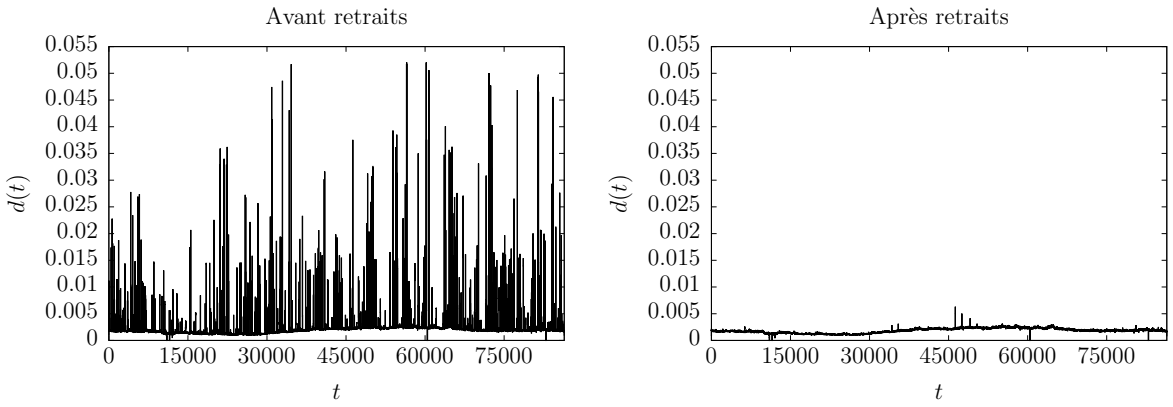


FIGURE 4.13 – **Évolution du degré $d(t)$ à l'instant t avant et après les retraits dans $\mathcal{L}_2$.**

a) Distribution du degré

b) KS-tests sur les distributions locales

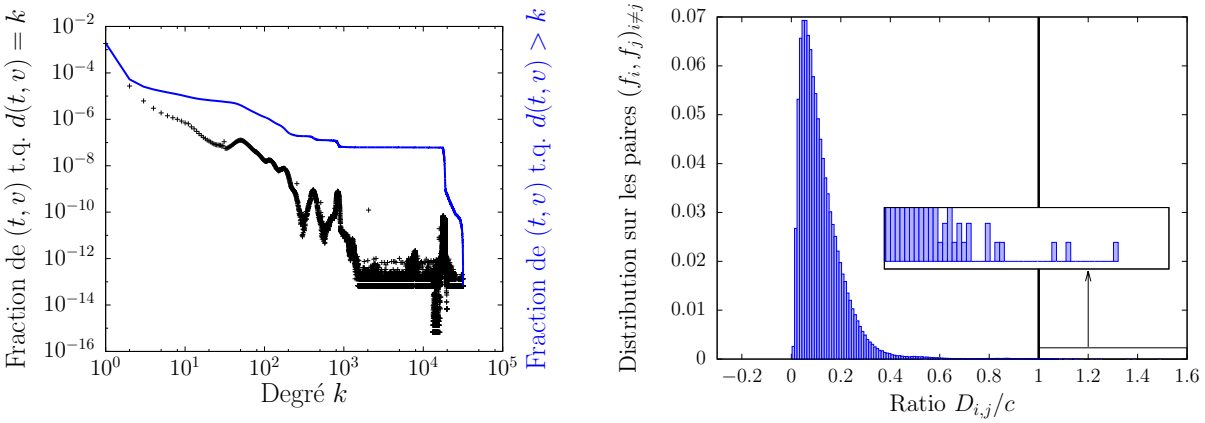


FIGURE 4.14 – **Hétérogénéité structurelle et homogénéité temporelle du degré dans $\mathcal{L}_3$.**

normales.

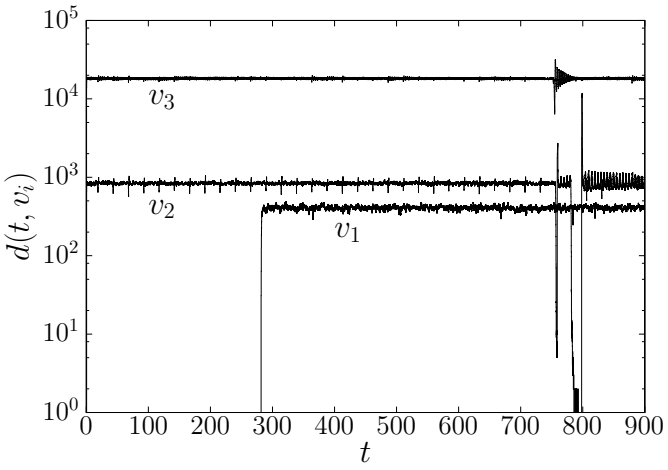


FIGURE 4.15 – **nœuds ayant un degré constant élevé** – Des classes AN sont observées dans les classes de degrés élevés : v_1 est responsable du trafic normal observé dans $C_{24} = \{399, \ldots, 502\}$; v_2 de celui observé dans $C_{27} = \{795, \ldots, 1001\}$ et v_3 de celui observé dans $C_{40} = \{15849, \ldots, 19953\}$.

Au final, notre méthode nous permet de détecter 827 anomalies et d'en identifier 796 (96%). Pour ce faire, nous avons supprimé 1.2% du trafic. Comme c'est le cas dans les deux autres flot de liens, les suppressions conduisent à un trafic exempt de la plupart des anomalies liées au degré comme on peut le voir dans la Figure 4.16. La durée cumulée anormale est de 24.43 secondes avant les retraits et de 5.78 secondes après les retraits. De même, la moyenne du degré à l'instant t passe de $1.809 \cdot 10^{-3}$ à $1.801 \cdot 10^{-3}$ après les retraits, soit une déviation de $8 \cdot 10^{-6}$%.

Dans la suite, nous comparons nos résultats à la base de données MAWILab [53]. Le 3 novembre 2018, de 14h00 à 14h15, elle indique un total de 287 adresses IP auxquelles sont associées les périodes de temps durant lesquelles elles sont anormales. Étant donné que nous n'avons pas accès aux numéros de ports dans nos données et que le degré ne tient pas compte du nombre de paquets échangés, les anomalies dans les catégories déni de service point à point, scans de port et flot alpha ne peuvent pas être détectées par notre méthode.

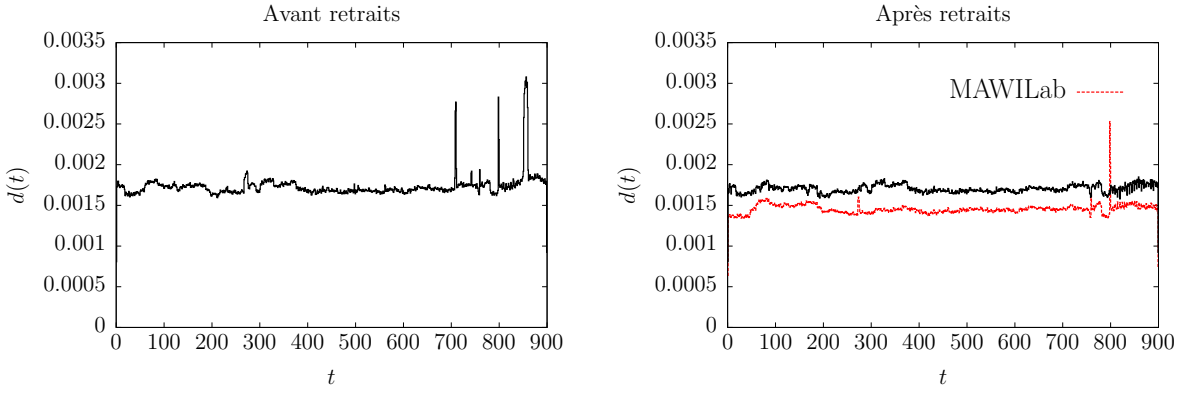


FIGURE 4.16 – **Évolution du degré $d(t)$ à l'instant t avant et après les retraits dans $\mathcal{L}_3$** – Le nœud v_3 a été supprimé des calculs pour plus de clarté. Contrairement à notre méthode, les événements identifiés par MAWILab entraînent une diminution de la moyenne lors de leur suppression.

a) nœuds non détectés par MAWILab

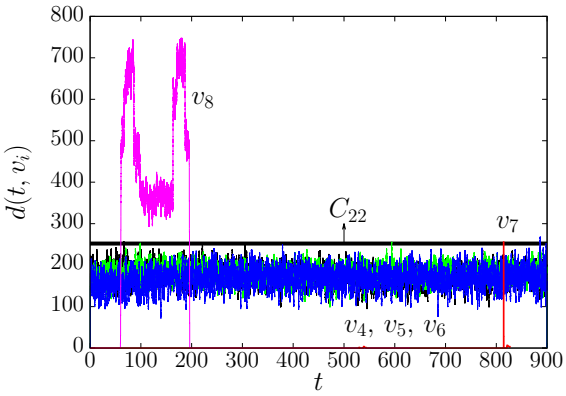
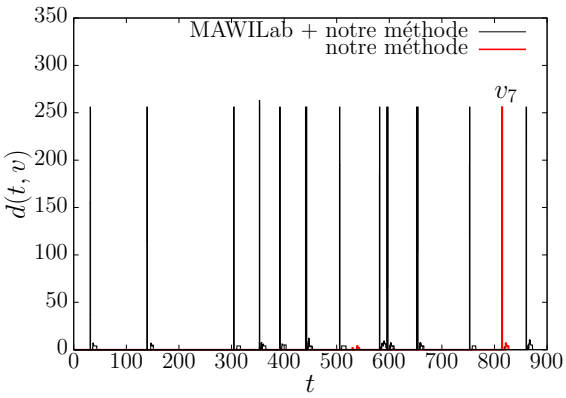
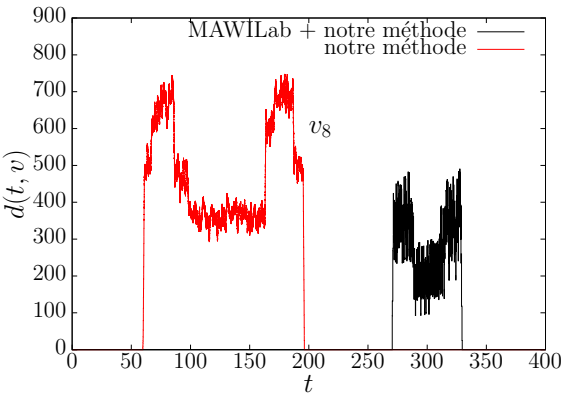


FIGURE 4.17 – **Comparaison à la base de données MAWILab** – Les nœuds v_2, v_4, v_5, v_6, v_7 et v_8 ne sont pas détectés par MAWILab. Cependant, la suppression de l'activité anormale de v_2 est responsable de la disparition du pic autour de $t = 800$ dans le degré au cours du temps en Figure 4.16 et les nœuds v_7 et v_8 ont des activités suspectes habituellement détectées par MAWILab.

b) Activité anormale de v_7



c) Activité anormale de v_8



Si, de plus, nous nous restreignons aux anomalies correspondant à du trafic non-légitime, cela réduit le nombre d'adresses IP listées à 77.

Avec notre méthode, nous trouvons 33 adresses IP anormales. Six d'entre elles ne sont pas répertoriés par MAWILab. Elles correspondent au nœud v_2 de la Figure 4.15 et aux nœuds v_4, v_5, v_6, v_7 et v_8 de la Figure 4.17.a. Le nœud v_2 a été supprimé lors de son pic d'activité de 798.18 à 799.87. De même, le nœud v_7 a été supprimé de 814.06 à 815.00 et le nœud v_8 sur l'intégralité des périodes durant lesquelles il est actif. Comme on peut le voir sur les Figures 4.17.b et 4.17.c, les activités des nœuds v_7 et v_8 sont typiques de celles des nœuds effectuant des scans de réseau, généralement détectées par MAWILab. De ce fait, elles auraient dû être identifiées par leurs détecteurs. Les trois nœuds restants v_4, v_5 et v_6 ont été supprimés respectivement sur des intervalles de 0.0768s, 0.0677s et 0.181s, en raison de leur activité éphémère au sein de la classe $A\ C_{22} = \{252, \ldots, 317\}$. Étant donné leur profil de degré stable, ces nœuds-temps n'auraient pas dû être identifiés. On montre comment ces suppressions pourraient être évitées en utilisant des classes de taille plus élevée en Section 4.6.2.

Dans la catégorie d'anomalies correspondant aux scans de réseau, nous avons identifié 24 adresses IP parmi les 76 ($32,6\%$) répertoriées par MAWILab. Tous les scans impliquant plus de 250 destinations différentes ont été identifiés avec notre méthode. Comme mentionné ci-dessus, nous avons également identifié deux adresses IP omises par les détecteurs de MAWILab (voir Figures 4.17.b et 4.17.c). De plus, la précision temporelle fournie par notre méthode est meilleure. Cependant, nous ne parvenons pas à identifier les adresses IP liées en permanence au réseau car elles ont des profils de degré constants et conduisent donc à des classes AN. Plus généralement, nous ne détectons aucune adresse IP scannant des réseaux comportant moins de 250 de destinations : d'une part, toutes les classes inférieures à $C_{22} = \{252, \ldots, 317\}$ sont des classes AN, d'autre part leurs activités ne sont pas liées à celles de nœuds-temps anormaux supprimés. Néanmoins, les fenêtres de temps au cours desquelles se produisent la plupart de ces scans sont détectées dans leurs classes de degrés respectives.

Dans la catégorie de déni de service distribué, une seule anomalie est identifiée par MAWILab. Le nœud correspondant a un degré maximum de 53. Par conséquent, nous ne le détectons pas pour les mêmes raisons.

Parmi les 33 adresses IP détectées, les trois restantes entrent dans la catégorie point à multipoint que nous ne considérons pas ici étant donné qu'elle constitue un trafic normal de routeur.

Enfin, la Figure 4.16 montre le degré au cours du temps après la suppression des nœuds-temps anormaux identifiés par MAWILab. Contrairement à notre méthode, nous constatons que ces suppressions affectent la moyenne de $d(t)$, ce qui est principalement dû à la faible précision temporelle avec laquelle MAWILab décrit les anomalies (63% des adresses IP sont identifiées comme anormales sur l'ensemble de la trace).

Bien que ne constituant pas une vérité-terrain, la comparaison avec la base de données MAWILab nous permet de mettre en évidence les avantages et les inconvénients de notre méthode. Notamment, on montre, grâce aux suppressions, que la modélisation en flots de liens permet une meilleure précision des résultats. Nous montrons également que notre méthode permet de détecter tous les scans de réseaux ayant plus de 250 destinations. Cependant, elle met également en lumière l'incapacité de notre méthode à identifier la totalité des nœuds-temps responsables d'anomalies dans une fenêtre de temps et dans une classe de degrés, lorsque cette dernière est une classe AN. Pour pallier à cela, nous pourrions tenter de trouver une propriété pour laquelle leurs activités dévie de manière plus significative. Nous développons cette perspective dans la conclusion (voir Chapitre 5).

4.5.3 Twitter

Finalement, nous appliquons notre méthode au flot de liens $\mathcal{L}_4$ qui regroupe l'ensemble des retweets liés à la politique française au cours du mois d'août 2016.

L'activité temporelle de ce flot de liens n'est pas homogène étant donnés le rythme circadien et la tendance globale du mois. Comme dans le cas de la trace $\mathcal{L}_1$ de trafic IP d'une journée, on considère le degré des nœuds-temps dans un contexte agrégatif tel que :

$$\overline{d(t,v)} = \frac{d(t,v)}{d(t)} \,.$$

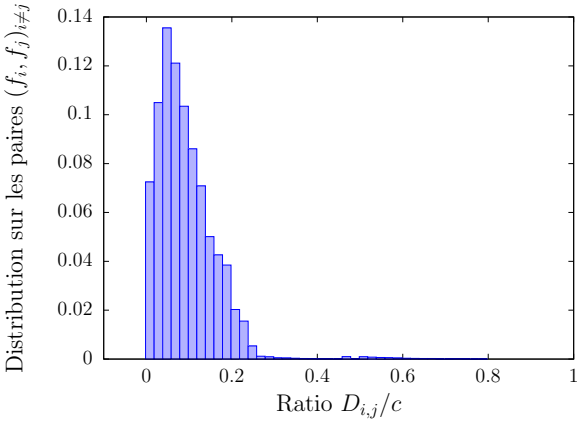
En considérant des fenêtres temporelles d'une heure, nous voyons sur la Figure 4.18.a que les distributions locales du degré normalisé sont similaires. On remarque également sur la Figure 4.18.b que sa distribution globale est hétérogène.

En conservant une taille de classes identique, nous obtenons 28 classes de degrés allant de $C_1 = [1, 2[$ à $C_{28} = [1001, 1258[$.[8] Parmi celles-ci, il y a 16 classes AN, 10 classes A et 2 classes R. Comme le laissent supposer les distributions des KS-tests et du degré normalisé au cours du temps (voir Figures 4.18.a et 4.18.c), il y a beaucoup moins d'anomalies dans ce flot de liens que dans les autres : on détecte uniquement 208 couples (T_i, C_j) anormaux. On identifie les nœuds-temps correspondant à 151 d'entre eux. Parmi ceux-ci, on trouve des auteurs retweetés par un grand nombre de diffuseurs distincts, comme marseille et fhollande les 11 et 12 août aux alentours de minuit, ou encore, deux diffuseurs qui retweetent un grand nombre d'auteurs durant la nuit, mais aussi MLP_officiel, et Marion_M_Le_Pen le 3 août vers 11h, et Nicolas Sarkozy les 22 et 24 août.

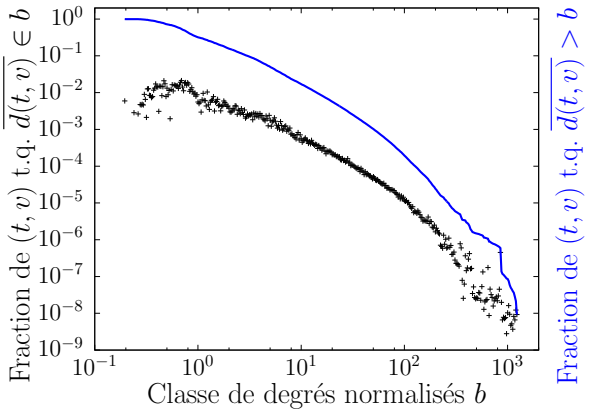
L'homogénéité du degré normalisé moyen implique un très faible taux d'anomalies (voir Figure 4.18.c). De ce fait, nous ne traçons pas son évolution au cours du temps avant et après les retraits étant donné le peu de différences observées.

8. Voir notes 5 et 6.

a) KS-tests sur les distributions locales



b) Distribution du degré



c) Distribution du degré normalisé au cours du temps avant les retraits

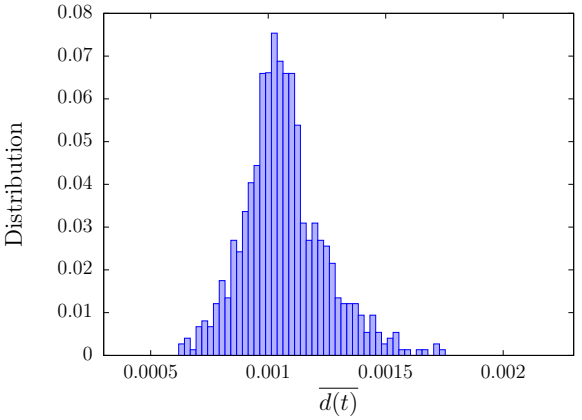


FIGURE 4.18 – **Anomalies dans le flot de liens des rewteets $\mathcal{L}_4$.**

Dans cette section, nous appliquons notre méthode à plusieurs flots de liens différents aussi bien au niveau de leurs origines, des traces de trafic IP et des retweets de Twitter, que de leurs étendues temporelles, de quinze minutes, une heure, une journée et un mois. L'ensemble de ces expériences, menant à l'identification de nœuds-temps anormaux d'intérêt dans chacun des cas, prouve la généralité de notre méthode. Notre méthode est applicable à tous flots de liens et propriété p tels que la distribution de p sur le flot de liens est (1) hétérogène et (2) stable dans le temps. Cette deuxième restriction, moyennant l'utilisation de contextes adaptés, peut être réduite à la présence de régularités temporelles (voir Chapitre 3).

4.6 Influence des paramètres

Nous avons montré l'efficacité de notre méthode sur plusieurs jeux de données. Dans cette section, nous effectuons une série d'expériences pour étudier l'influence des paramètres τ, la durée des fenêtres temporelles, et r, la taille des classes de degrés. Nous reprenons à cet effet le flot de liens $\mathcal{L}_1$.

4.6.1 Variation de la taille des fenêtres de temps

Nous divisons le flot de liens en fenêtres de temps de taille τ afin de comparer les distributions de degré locales. Le succès de notre méthode repose sur le fait qu'elles soient similaires d'une fenêtre de temps à l'autre. En raison de l'agrégation sur une période plus longue et donc d'une diminution des variations statistiques, plus la durée τ est élevée, plus la similarité entre les fenêtres de temps est élevée, et inversement, lorsque la taille diminue.

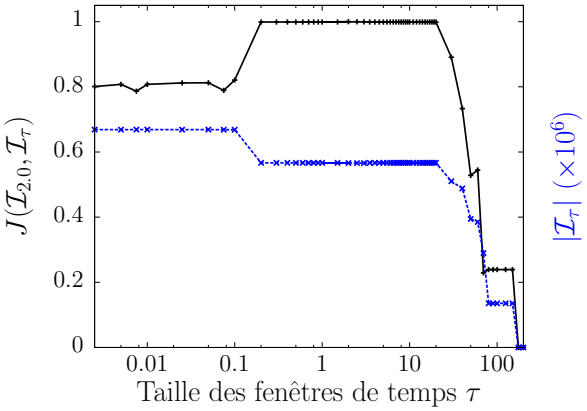
Soit $\mathcal{I}_\tau = \cup_{i,j} \mathcal{I}_{(T_i,C_j)}$ l'ensemble de nœuds-temps anormaux détectés en utilisant des fenêtres temporelles de taille τ. Afin d'évaluer l'impact de τ, nous mesurons le coefficient de similarité de Jaccard entre $\mathcal{I}_{2.0}$, obtenu lors de la première expérience, et d'autres ensembles obtenus en faisant varier τ :

$$J(\mathcal{I}_{2.0}, \mathcal{I}_\tau) = \frac{|\mathcal{I}_{2.0} \cap \mathcal{I}_\tau|}{|\mathcal{I}_{2.0} \cup \mathcal{I}_\tau|} \ .$$

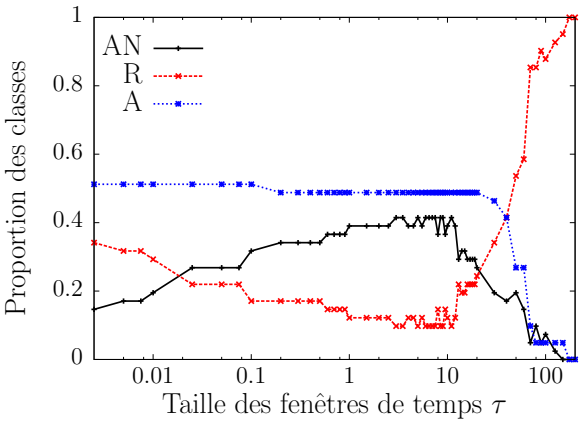
Les résultats sont illustrés dans la Figure 4.19.a. Nous voyons que les ensembles identifiés $\mathcal{I}_\tau$ sont identiques de $\tau = 0.2$ à $\tau = 20.0$ ($J(\mathcal{I}_{2.0}, \mathcal{I}_\tau) = 1$), ce qui montre que notre méthode est stable par rapport à ce paramètre. Lorsque $\tau < 0.2$, nous identifions légèrement plus d'anomalies. Au contraire, lorsque la taille augmente, on en identifie de moins en moins jusqu'à ne plus en détecter pour $\tau \geqslant 175.0$. Ces observations s'expliquent par les nombres de classes AN, A et R selon τ, visibles sur la Figure 4.19.b : plus τ augmente, plus le nombre de classes R est élevé et plus le nombre de classes de A, dans lesquelles nous identifions les anomalies, est faible. Lorsque l'on atteint $\tau = 175.0$, toutes les classes sont rejetées et, par conséquent, aucune anomalie n'est détectée. Cette augmentation du nombre de classes rejetées est provoquée par le très petit nombre de fenêtres lorsque τ augmente. En effet, ces dernières sont insuffisamment nombreuses pour établir un comportement normal et les ajustements entre les fractions $f(T_i, C)$ et une distribution homogène sont davantage rejetés.

Nous avons remarqué que nous identifions plus de nœuds-temps lorsque τ est petit. Cependant, ce résultat doit être pris avec prudence. Comme on peut le voir sur la Figure 4.19.b, lorsque τ diminue, le nombre de classes rejetées augmente et le nombre de classes AN diminue. En effet, plus la fenêtre de temps est petite, moins le comportement entre les fenêtres est similaire étant donné les variations statistiques liées à des échantillons plus petits. Cela entraîne alors un rejet du comportement normal. Les classes A ne sont pas affectées : dans la plupart des fenêtres de temps, aucun nœud n'atteint un degré au sein de la classe, quel que soit τ. De plus, leur nombre augmente. Ces observations sont expliquées par l'exemple de la Figure 4.20.a. On voit que pour $\tau = 0.25$, il y a 83% de T_i pour lesquels la fraction $f(T_i, C_{34})$ vaut zéro, contre 67% pour $\tau = 2.0$. Ainsi, lorsque τ diminue, la proportion de fenêtres de temps sans trafic par rapport à celles contenant du trafic est beaucoup plus élevée que dans les expériences avec un τ plus grand. Or, si l'augmentation des classes A permet d'identifier davantage d'anomalies, la diminution des classes AN nous empêche de déterminer si un retrait est erroné ou non par l'apparition d'anomalies négatives. Par conséquent, notre critère de validation par suppression ne peut être appliqué ce qui pourrait avoir des conséquences sur le trafic normal résultant.

a) Similarité des nœuds-temps identifiés



b) Conséquences sur les classes AN, R et A



c) Temps de calcul

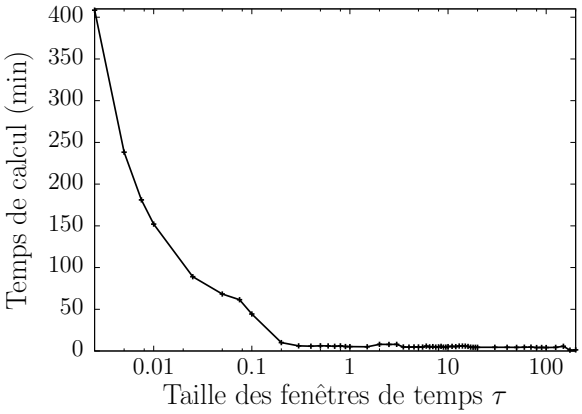
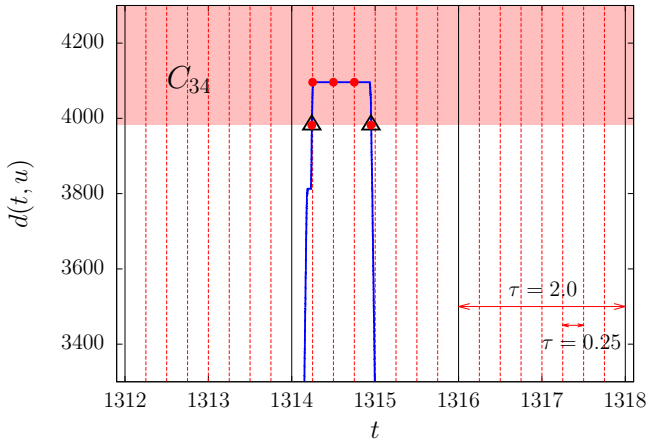


FIGURE 4.19 – **Influence de la taille des fenêtres de temps** – a) L'indice de Jaccard entre les ensembles de nœuds-temps identifiés montre que notre méthode est stable de $\tau = 0.2$ à $\tau = 20.0$. b) Les petites tailles de fenêtres entraînent une augmentation du nombre de classes A et R, tandis que des les grandes tailles entraînent un nombre élevé de classes R mais un petit nombre de classes AN et A. c) Le temps de calcul augmente de manière significative à mesure que τ diminue.

a) Exemple de nœud anormal dans la classe C_{34}



b) Précision temporelle des retraits

	τ	
	0.25	2.0
Classes et fenêtres détectées	$(T_{5256}, C_{34}), (T_{5257}, C_{34})$ $(T_{5258}, C_{34}), (T_{5259}, C_{34})$	(T_{657}, C_{34})
nœuds-temps détectés	$\{(t, v) \in [1314, 1315[\times V$ $: d(t, v) \in C_{34}\}$	$\{(t, v) \in [1314, 1316[\times V$ $: d(t, v) \in C_{34}\}$
Retrait	u de 1314.24 à 1314.95	u de 1314.24 à 1314.95

FIGURE 4.20 – **Proportion de classes A et retraits selon τ** – a) Le nœud u a un trafic anormal dans la classe C_{34}. Pour $\tau = 0.25$, il y a 4 fenêtres sur 24 ayant un trafic anormal, pour $\tau = 2.0$, il y en a 1 sur 3. b) La proportion de classes A dépend de τ mais pas la précision temporelle : avec $\tau = 0.25$ (points rouges), nous supprimons u de 1314.24 à 1314.25 (T_{5256}), puis de 1314.25 à 1314.75 (T_{5257}, T_{5258}), et finalement de 1314.75 à 1314.95 (T_{5259}). Avec $\tau = 0.25$ (triangles noirs), nous supprimons u de 1314.24 à 1314.95 (T_{657}). Ainsi, la taille des fenêtres temporelles n'affecte pas la précision des résultats.

Nous voyons également sur la Figure 4.19.c que l'utilisation de petites fenêtres temporelles augmente considérablement le temps de calcul.

Enfin, il important de souligner que, grâce à la modélisation du trafic en flot de liens, la taille des fenêtres temporelles n'affecte pas la précision avec laquelle nous identifions les anomalies étant donné que les interactions ne sont pas agrégées à l'intérieur des fenêtres. Nous détaillons cette affirmation à l'aide de l'exemple de la Figure 4.20.b.

4.6.2 Variation de la taille des classes de degrés

Nous divisons les distributions de degré locales en classes de degrés de taille logarithmique r. Le succès de notre méthode repose sur l'homogénéité des distributions des fractions $f(T_i, C_j)$ sur l'ensemble des fenêtres temporelles T_i pour chaque classe C_j.

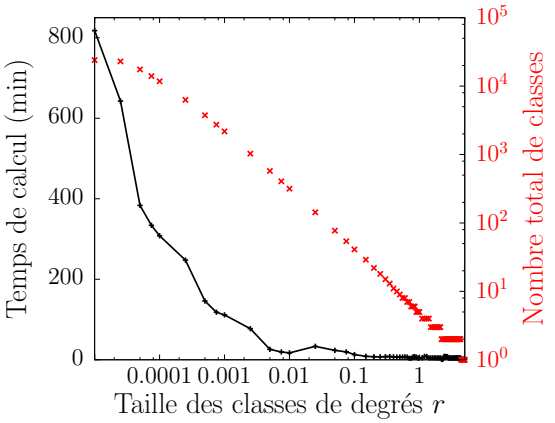
Tout d'abord, étant donné leur construction logarithmique, le nombre total de classes est très élevé lorsque r est petit et diminue rapidement lorsque r augmente (voir Figure 4.21.a). Pour $r = 10^{-5}$, on observe 23 983 classes ; pour $r > 1.6$, ce nombre est inférieur à 4 et atteint 1 dès lors que $r \geqslant 4.4$. En découle que plus r est petit, plus le temps de calcul est long.

On fait plusieurs observations concernant la taille et la similarité des ensembles de nœuds-temps détectés (voir Figure 4.21.b) :
1) le nombre de nœuds-temps identifiés augmente pour $r < 0,2$;
2) nous n'identifions pas de nœuds-temps anormaux pour $r \in [2.2, 2.3]$ et $r > 4.4$;
3) l'indice de Jaccard est supérieur à 0.8 pour $r \in [10^{-5}, 1.6]$ et $r \in [2.4, 3.3]$;
4) l'indice Jaccard fluctue entre 0.8 et 1 pour $r \in [0.2, 1.6]$;
5) le nombre de nœuds-temps identifiés diminue à partir de $r = 1.7$, augmente à partir de $r = 2.4$, puis diminue à nouveau à partir de $r = 3.4$.
Une fois de plus, ces observations sont liées aux proportions des trois types de classes.

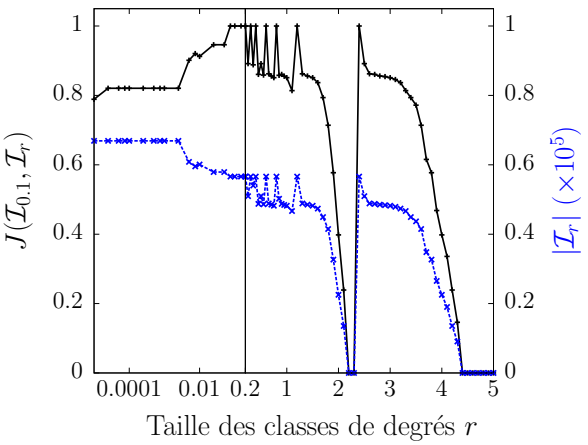
Lorsque la taille des classes de degrés est trop petite, les classes n'incorporent plus les fluctuations temporelles du degré des nœuds comme on le voit avec le nœud v et une taille de classe $r = 0.01$ dans la Figure 4.22. Comme on peut le voir sur la Figure 4.21.c, cela entraîne une augmentation du nombre de classes A et R au détriment des classes AN qui diminuent. Comme dans le cas des tailles de fenêtres temporelles, alors que l'augmentation des classes A permet d'identifier davantage d'anomalies (observation 1), la diminution des classes AN nous empêche d'utiliser notre critère de validation.

Au contraire, lorsque la taille des classes de degrés est trop élevée, ces dernières intègrent trop de trafic et le nombre de classes A diminue (voir Figure 4.21.c). Par conséquent, plus r augmente, moins nous sommes capables d'identifier de nœuds-temps. C'est ainsi que nous expliquons l'observation 2 : pour $r = 2.2$, il y a deux classes AN ; pour $r = 2.3$; il y a une classe AN et une classe R ; et finalement, pour $r \geqslant 4.4$, il n'y a plus qu'une seule classe AN. De plus, nous voyons sur la Figure 4.23 que lorsque la classe C_1 contient plusieurs valeurs du degré, la distribution résultante est identique à celle pour laquelle elle ne contient que le degré 1, étant donné le nombre écrasant de

a) Temps de calcul et nombre de classes



b) Similarité des nœuds-temps identifiés



c) Conséquences sur les classes AN, R et A

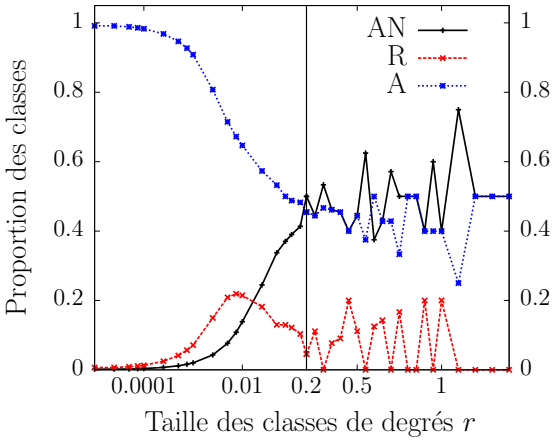


FIGURE 4.21 – **Influence de la taille des classes de degrés** – a) En raison du grand nombre de classes, le temps de calcul augmente de manière significative lorsque r diminue. b) Notre méthode est stable de $r = 10^{-5}$ à $r = 1,6$. c) Les petites classes conduisent à une augmentation du nombre de classes A et à une diminution du nombre de classes AN. Pour $r < 0,02$, la proportion des classes R est supérieure à celle des classes AN. Pour $r > 0,2$, les proportions des classes A et AN sont similaires et la proportion des classes R est faible. Note : nous ne traçons pas la proportion des classes pour $r > 1,4$ en raison des fluctuations dues au petit nombre de classes.

nœuds-temps ayant ce degré. Par conséquent, nous détectons moins d'anomalies et des retraits erronés impliquant des nœuds-temps de degré supérieur à 1 pourraient être acceptés.

Enfin, les observations 3, 4 et 5 s'expliquent par des effets de discrétisation. Dans la trace de trafic d'une heure, les classes sont disposées de telle sorte que les classes de faibles degrés sont des classes AN et les classes de degrés élevés sont des classes A. Soit k_{id} le plus petit degré à partir duquel nous pouvons identifier des nœuds-temps anormaux, soit, dans ce cas, la borne inférieure de la première classe A. Comme on le voit sur la Figure 4.24, k_{id} dépend de r. Plus k_{id} est petit, plus le nombre d'anomalies détectées est élevé. Jusqu'à $r = 0.2$, k_{id} est inférieur à 250 et le nombre d'anomalies identifiées est maximal. Ensuite, k_{id} fluctue entre 250 et 2 512, ce qui explique l'observation 4. Beaucoup d'anomalies sont situées dans cette plage de degré, ainsi, si $k_{id} \in [250, 2512]$, ces dernières sont identifiées, sinon elles ne le sont pas, ce qui explique l'observation 3. Pour $r \in [2.2, 4.3]$, on observe un total de 2 classes. Le nombre d'anomalies détectées dépend des proportions de chacune des classes (observation 5) : il y a soit 2 classes AN ($r = 2.2$), soit une classe AN et une classe R ($r = 2.3$), soit une classe AN et une classe A ($r \in [2.4, 4.3]$). Nous observons le même processus pour $r \in [2.4, 4.3]$, k_{id} augmente avec r, cela entraîne la baisse du nombre de nœuds-temps identifiés jusqu'à ce qu'il ne reste plus qu'une classe pour $r > 4.3$.

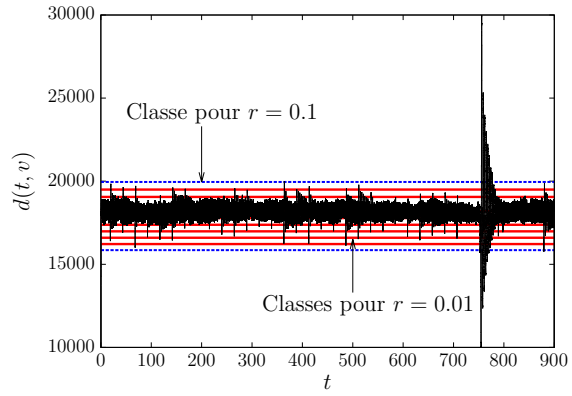


FIGURE 4.22 – **Conséquences de l'utilisation de petites classes de degrés** – Lorsque $r = 0,01$, les classes sont trop petites et ne contiennent pas les fluctuations du degré. Par conséquent, pour $r = 0.01$, nous observons 6 classes AN et 4 classes R au lieu d'une seule classe AN pour $r = 0.1$. Note : ce nœud provient du flot de liens $\mathcal{L}_3$.

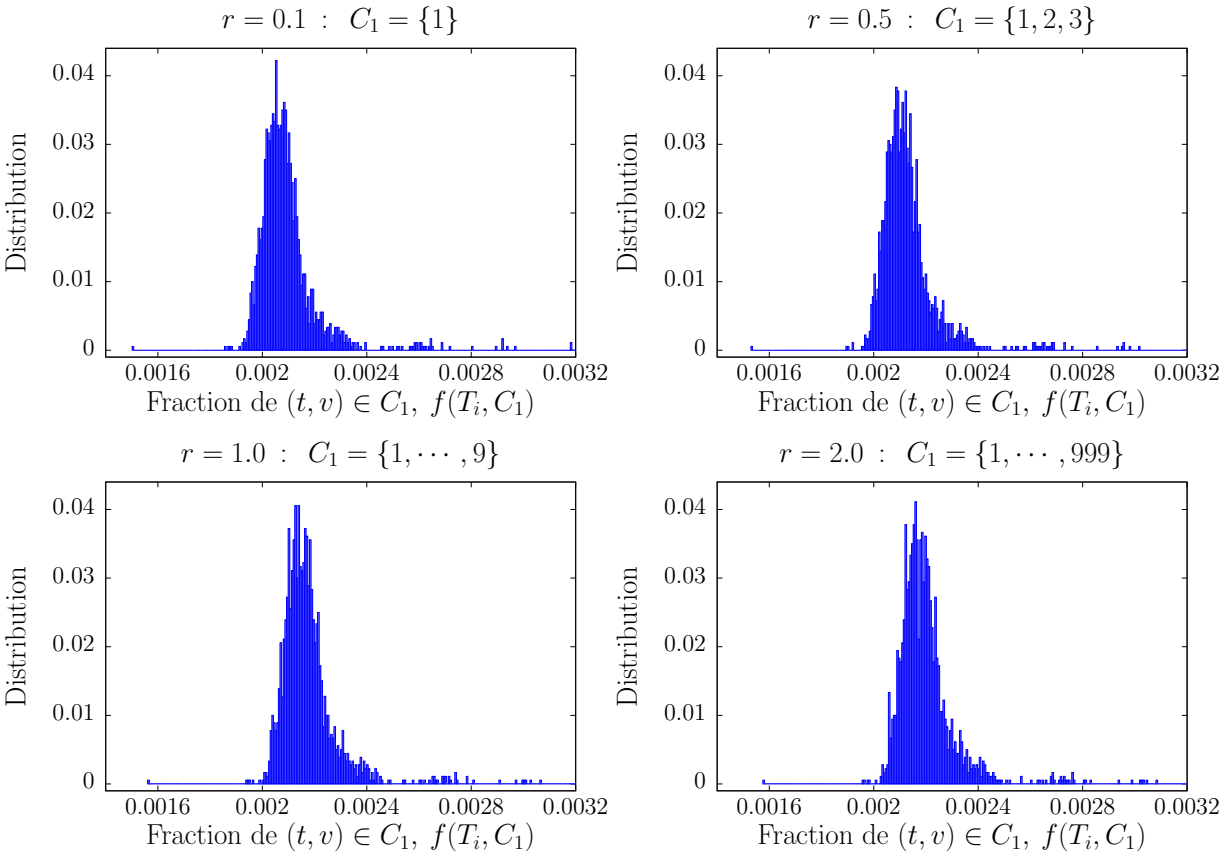


FIGURE 4.23 – **Classe C_1 selon r** – Pour $r > 0,3$, nous détectons les mêmes anomalies que pour $r = 0,1$, car les anomalies impliquant les nœuds-temps tels que $d(t, v) > 1$ sont incluses dans la gaussienne. Note : nous faisons un zoom sur la partie homogène de la distribution pour plus de clarté.

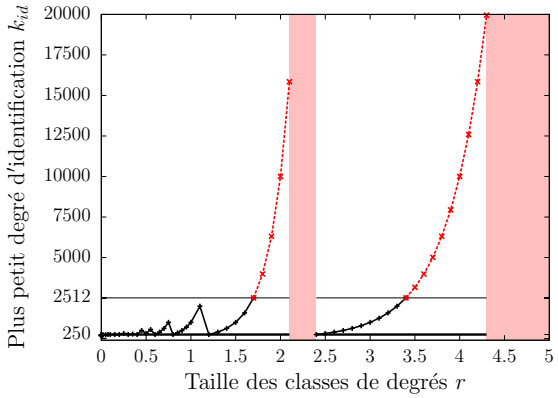


FIGURE 4.24 – **Plus petit degré d'identification selon** r – Lorsque $k_{id} \in [250, 2512]$, l'indice de Jaccard est supérieur à 0.8. Lorsque $k_{id} > 2512$, le nombre d'anomalies identifiées diminue (ligne rouge pointillée) et atteint 0 dans les zones rouges.

Notre méthode est stable par rapport aux paramètres τ et r. La taille des fenêtres temporelles τ peut varier de deux ordres de grandeur de $0.2s$ (18 000 fenêtres) à $20s$ (180 fenêtres) sans que les résultats soient affectés. De même, malgré l'instabilité apparente de la méthode vis-à-vis de la taille des classes de degrés r, pour un nombre de classes allant de 24 000 ($r = 10^{-5}$) à 4 ($r = 1.6$), la méthode est stable et présente des résultats très similaires.

4.7 Conclusion

Nous proposons une approche permettant de différencier les comportements normaux des comportements anormaux dans une distribution hétérogène. Nous introduisons à cet effet une méthode qui s'appuie sur l'homogénéité temporelle du caractère hétérogène des entités sous étude. Dans ce chapitre, nous l'appliquons au cas du degré des nœuds-temps dans un flot de liens. L'originalité de notre méthode est d'admettre l'hétérogénéité comme propriété intrinsèque des nœuds-temps : un nœud-temps anormal ne signifie pas nécessairement qu'il a un degré extrême, mais plutôt qu'il agit anormalement vis-à-vis du contexte temporel et structurel établi par les autres nœuds-temps.

Nous avons illustré la pertinence de notre approche sur 15 minutes, 1 heure et 1 jour de trafic IP ainsi qu'un mois de retweets de Twitter. Ces données satisfont les deux conditions nécessaires à l'application de notre méthode : leur degré se distribue (1) hétérogènement sur les nœuds-temps, et (2) homogènement sur le temps, une fois agrégé sur les nœuds. De nombreux travaux ont montré que les nœuds dans les graphes réels avaient souvent des connectivités très différentes menant à des distributions du degré hétérogènes. Ainsi, l'hétérogénéité structurelle ne semble pas être une restriction limitant l'applicabilité de notre méthode. Le caractère universel de l'homogénéité temporelle, par contre, semble moins évident. Notamment, Karsai et al. [86, 85] parlent de « *processus temporels inhomogènes* » pour caractériser les comportements temporels par salves d'activités (*bursty behavior*). Néanmoins, s'ils ne sont pas homogènes, ils ont tendances à être périodiques, en particulier lorsqu'ils sont corrélés aux comportements humains. Dans le cas des appels téléphoniques par exemple, on observe toujours moins d'activité pendant la nuit que le

jour, à 15h qu'à 19h, ou un 7 juillet qu'un 1er janvier. Dans ce cas, il est toujours possible, notamment avec les outils proposés dans le Chapitre 3, de normaliser l'activité comme nous l'avons fait dans les cas des flots de liens $\mathcal{L}_2$ et $\mathcal{L}_4$. L'homogénéité temporelle n'apparaît ainsi pas non plus comme étant une condition limitant drastiquement les domaines d'application de notre méthode. Il serait toutefois intéressant d'effectuer une étude statistique sur un ensemble conséquent de jeux de données afin d'en cerner les limites.

Une première perspective de ces travaux serait de les appliquer à d'autres propriétés. Cela pourrait nous permettre de détecter d'autres types d'anomalies et également de détecter les entités responsables de l'anormalité des couples (T_i, C_j) pour lesquels nous n'avons pas réussi à identifier la cause avec le degré uniquement. Nous discutons de cette perspective en détail dans le chapitre suivant.

Une autre perspective serait d'étudier l'influence du paramètre Δ, utilisé lors de la modélisation des interactions ponctuelles en flot de liens avec durée. Notamment, analyser en quoi les résultats sont affectés par ce paramètre, ou si l'utilisation de Δ différents mène à l'identification de nouvelles anomalies.

Jusqu'à présent, nous avons appliqué notre méthode à des données d'interactions temporelles stockées sur disque. Nous pourrions facilement l'adapter au traitement d'interactions temporelles arrivant en continu (*streaming*) : le degré des nœuds-temps pouvant facilement être mis à jour, il suffirait de comparer la distribution du degré de la fenêtre temporelle au temps présent à un échantillon des distributions passées pour détecter des anomalies.

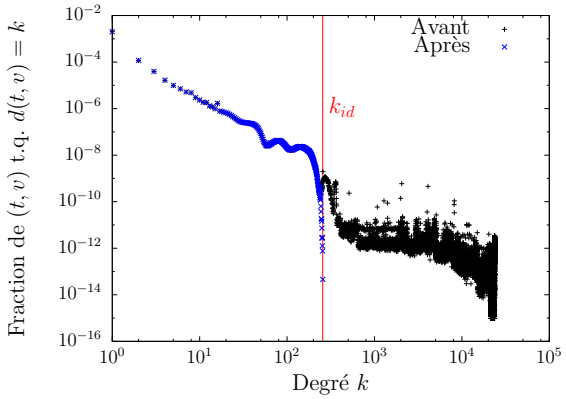


FIGURE 4.25 – **Distribution du degré avant et après les retraits** – $k_{id} = 256$ est le seuil des valeurs extrêmes anormales de la distribution.

Enfin, nous avons constaté dans les flots de liens $\mathcal{L}_1$, $\mathcal{L}_2$ et $\mathcal{L}_4$, que les classes A correspondaient uniquement à des classes de degrés élevés. Ainsi, nous détectons les mêmes nœuds-temps anormaux qu'en prenant les valeurs extrêmes de la distribution. Or, contrairement à l'approche adoptée dans le Chapitre 3, le seuil n'est pas fixé arbitrairement. De ce fait, dans les cas où les valeurs extrêmes de la distribution sont effectivement des nœuds-temps anormaux dans notre contexte [9], notre méthode apporte une contribution au domaine de la recherche de valeurs extrêmes dans une distribution hétérogène : le plus petit degré d'identification k_{id} représente le seuil à partir duquel l'ensemble des degrés, et

9. C'est-à-dire lorsque les degrés de la queue de la distribution ne sont pas ceux de nœuds ayant un degré constant ou, de façon équivalente, qu'il n'y a pas de classes AN parmi les classes de degrés élevés.

des nœuds-temps qui y sont associés, sont anormaux (voir Figure 4.25).

Chapitre 5

Conclusion et Perspectives

Sommaire

5.1 Résumé des contributions

Dans un flot de liens, les liens arrivent au cours du temps. Pour y détecter des sous-ensembles d'interactions anormales, il est nécessaire de déterminer dans quelle mesure cette dynamique est régulière et dans quels cas des irrégularités se produisent. Dans cette thèse, nous nous sommes fixés pour objectif d'apporter des éléments de réponse à ce problème.

Notre première contribution est la caractérisation de la notion d'anomalie dans un flot de liens dans le Chapitre 3. Dans ce chapitre, nous avons d'abord montré que considérer à la fois la structure et la dynamique des interactions menait à de nombreuses façons de définir leur anormalité. Afin de faciliter la recherche d'anomalies, nous avons proposé une procédure permettant de construire des contextes variés de façon systématique et de considérer ainsi l'anormalité des interactions selon différentes perspectives. Notamment, nous avons défini une entité anormale dans un flot de liens selon quatre critères différents : (1) la propriété utilisée pour décrire son comportement, (2) le sous-ensemble d'entités du même type et (3) le comportement attendu auxquels on la compare, ainsi que (4) un critère de déviation qui lui attribue un score d'anormalité par rapport à l'ensemble de ces éléments. Au travers de nombreux exemples et études de cas, nous avons montré que l'utilisation de différents contextes menait à la détection d'anomalies sémantiquement différentes et que l'utilisation de contextes plus complexes et plus locaux permettait de se centrer sur une anomalie et d'en préciser la nature et l'origine. La flexibilité de notre méthode permet de s'adapter à de multiples situations : dans le cas de l'étude de la communication politique sur Twitter et de la détection d'attaques dans le trafic IP, nous nous sommes concentrés sur l'émergence de certains évènements ; dans le cas de l'étude du

second écran, nous nous sommes concentrés sur un intervalle de temps restreint autour de la diffusion d'une émission télévisée et nous avons analysé l'évolution des hashtags anormaux dans ce contexte ; enfin, dans le cas de la prédiction des liens utilisateurs/sujets, nous nous sommes intéressés à la similarité des hashtags anormaux retweetés par une même communauté politique.

Cependant, en appliquant cette procédure au degré instantané dans les flots de liens, nous n'avons pas été capables de caractériser correctement le comportement normal des degrés des nœuds-temps étant donnée leur distribution hétérogène. Nous avons alors envisagé plusieurs solutions. Concevoir des valeurs attendues plus adaptées nécessite d'avoir plus d'informations sur le comportement normal propre à chaque entité, ce que nous ne sommes pas en mesure de faire sans introduire de nouvelles propriétés. Considérer des contextes plus locaux regroupant des entités aux comportements similaires nécessite de partitionner les flots de liens, ce qui constitue un domaine de recherche à part entière. Dans le Chapitre 4, nous avons alors envisagé d'ajuster les distributions du degré de façon à détecter des anomalies dans la distribution des écarts entre les probabilités empiriques et un modèle de normalité (voir Section 4.1.3). Or, en les ajustant par une loi de puissance, nous avons remarqué que leurs subtilités structurelles n'étaient pas prises en compte et que la recherche d'un modèle plus adapté pouvait mener, à l'inverse, à un sur-ajustement des données.

De ce fait, notre deuxième contribution dans le Chapitre 4 est la conception d'une méthode permettant de trouver des anomalies dans une séquence de distributions individuellement hétérogènes, mais homogènes entre elles dans le temps. Nous avons remarqué, dans chacun des flots de liens, que les distributions du degré étaient hétérogènes mais similaires sur des fenêtres de temps successives. Nous avons à la fois exploité ce comportement normal, en comparant dans chaque fenêtre de temps la fraction de nœuds-temps dans une même classe de degrés, et respecté le comportement hétérogène en considérant des classes de degrés logarithmiques. D'une certaine façon, nous avons ainsi considéré des contextes plus locaux, en comparant les nœuds-temps ayant des degrés similaires, et conçu des valeurs attendues plus adaptées en comparant les fractions de nœuds-temps dans chaque fenêtre à une fraction moyenne caractérisant correctement le comportement normal.

Notre troisième contribution est la conception d'une méthode de validation des nœuds-temps anormaux. Nous sommes partis de l'assertion selon laquelle le retrait d'une anomalie ne devait pas perturber le comportement normal. De là, nous avons supprimé les nœuds-temps anormaux détectés dans le flot de liens et avons vérifié que ces retraits n'affectaient pas les distributions des fractions de nœuds-temps. En plus de permettre de valider nos détections, ces suppressions nous ont permis d'identifier l'origine d'anomalies détectées dans d'autres distributions et avec d'autres propriétés, ainsi que d'obtenir un flot de liens « nettoyé » dans lequel les nœuds-temps sont normaux vis-à-vis de leur degré.

Indépendamment des outils méthodologiques proposés, et bien que nous l'ayons exploité au travers du degré uniquement, nous avons montré que l'utilisation du formalisme des flots de liens pour l'analyse des interactions temporelles était particulièrement adapté.

Les degrés des nœuds-temps sont mis à jour dès lors que leur voisinage change, ce qui nous permet d'identifier exactement sur quelles périodes de temps les nœuds sont anormaux, puis de supprimer l'ensemble des interactions correspondantes de façon chirurgicale, sans perturber le comportement normal. De plus, l'ensemble des contextes par rapport auxquels les interactions peuvent être considérées anormales, bien que complexifiant leur recherche, constitue également une richesse de ce formalisme en nous permettant d'accéder à des anomalies plus subtiles que ce que permettent les autres représentations.

Pour conclure, l'objectif de notre recherche dans cette thèse n'était pas l'amélioration d'une technique de détection d'anomalies. Nous nous sommes plutôt concentrés sur la recherche d'anomalies (1) subtiles, (2) pertinentes et (3) interprétables dans un flot de liens. Pour cela, nous avons (1) caractérisé la structure et la dynamique des liens à l'aide de propriétés du formalisme des flots de liens et constaté que la suppression des anomalies n'altérait pas le comportement normal, (2) construit des contextes décrivant exactement les circonstances dans lesquelles les observations étaient anormales, et (3) déterminé la présence d'anomalies dans des distributions homogènes uniquement, en parvenant à caractériser un comportement normal par la recherche de contextes et de propriétés adaptés. Cette thèse, de ce fait, contribue à une des premières approches de la détection d'anomalies dans des flots de liens.

5.2 Perspectives

Cette thèse ouvre plusieurs perspectives. Nous en avons déjà évoquées plusieurs résultant directement de chaque chapitre. Dans la suite, nous présentons deux pistes de recherche plus générales : une extension de nos méthodes et la génération de flots de liens normaux.

5.2.1 Généralisation

La suite logique concernant l'amélioration de notre méthode serait de l'étendre à d'autres propriétés que le degré afin de chercher des anomalies plus subtiles.

Dans un premier temps, nous pourrions par exemple tenter d'identifier les anomalies dans les classes de faibles degrés en utilisant une propriété pour laquelle l'activité des nœuds-temps anormaux dévie de manière plus significative. Par exemple, dans le Chapitre 4, nous avons remarqué que les scans de réseau impliquant moins de 250 destinations n'étaient pas identifiés par notre méthode (voir Section 4.5.2). Lorsque l'on s'intéresse aux profils de degré de ces nœuds, on remarque que beaucoup présentent une variation de degré très brusque. Ainsi, on s'attend à ce que la dérivée de leur degré au cours du temps

$$\frac{|d(t,v) - d(t-\Delta t, v)|}{\Delta t},$$

où Δt est l'échelle de temps sur laquelle est quantifiée la variation, dévie significativement de celles des autres nœuds-temps. C'est ce que nous observons en pratique : en appliquant cette propriété sur le flot de liens $\mathcal{L}_3$ correspondant au trafic IP de 15 minutes du 3 novembre 2018, nous parvenons à identifier les scans à 64 et 128 destinations, ainsi que

a) Distribution des fractions de nœuds-temps dans C_1 selon la dérivée du degré

b) Profils de degré de nœuds anormaux selon la dérivée du degré

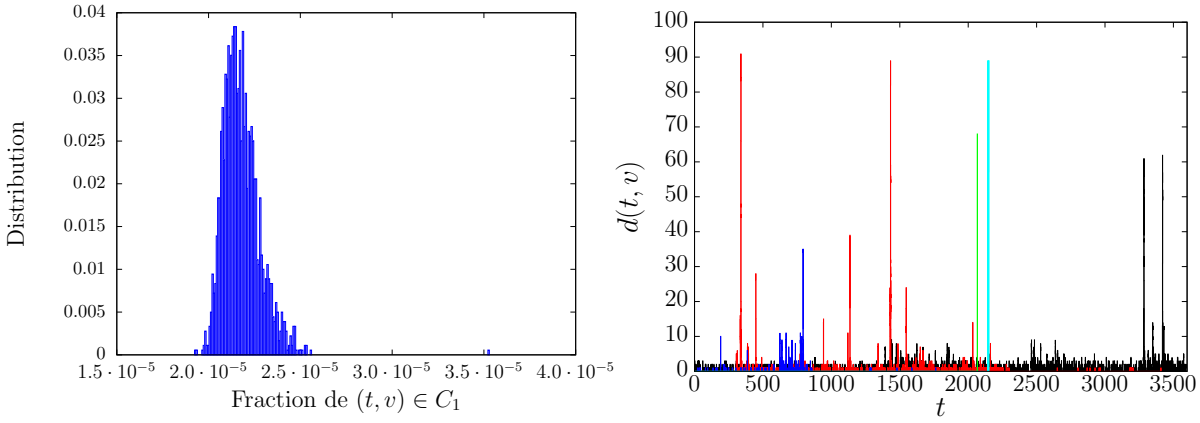


FIGURE 5.1 – **Détection d'anomalies avec la dérivée du degré** – Les résultats affichés sont obtenus par application de notre méthode à la dérivée du degré dans le flot de liens $\mathcal{L}_1$ après suppression des anomalies liées au degré instantané des nœuds réalisée au Chapitre 4.

d'autres nœuds aux profils de degré suspects (voir Figure 5.1).

Utiliser d'autres propriétés permet donc de détecter de nouvelles anomalies. D'autre part, l'utilisation de propriétés plus complexes pourrait nous permettre d'identifier des anomalies détectées, mais pour lesquelles l'identification des entités qui en sont à l'origine n'a pas été possible (voir Section 4.5).
Indirectement, c'est déjà ce que nous avons fait en passant du degré moyen au cours du temps $d(t)$ au degré instantané des nœuds $d(t, v)$. Dans la Section 2.3.2, nous avons détecté des instants anormaux dans la distribution de $d(t)$. Or, le degré au cours du temps est une propriété trop agrégée pour permettre l'identification des interactions anormales. En effet, la suppression d'une anomalie dans ce contexte implique la suppression de l'intégralité des interactions ayant lieu à l'instant correspondant. Cependant, nous avons remarqué que le retrait des noeuds-temps détectés avec le degré instantané des nœuds $d(t, v)$ entraîne la disparition des instants anormaux du degré moyen au cours du temps $d(t)$, sans perturber son comportement normal.
Ce résultat montre qu'utiliser des propriétés plus complexes et moins agrégées, permet d'identifier l'origine d'anomalies détectées avec une propriété plus agrégée. Ainsi, de façon similaire, nous pourrions tenter d'identifier les anomalies pour lesquelles l'identification avec le degré des nœuds-temps est erronée en caractérisant le comportement des liens $(t, u, v) \in E$ (comme dans le cas du nœud de la Figure 5.2 par exemple).

Ces deux points laissent envisager une généralisation de notre méthode. Notamment, on pourrait réitérer le processus de retrait des anomalies avec un ensemble de propriétés hiérarchisées selon deux niveaux. On considérerait, d'une part, un même type de propriété puis des contextes de plus en plus locaux, comme ce que nous avons fait avec la quantité d'interaction et le degré. La suppression des anomalies locales permettrait de valider ces dernières ainsi que déterminer l'origine des anomalies globales. On utiliserait ensuite

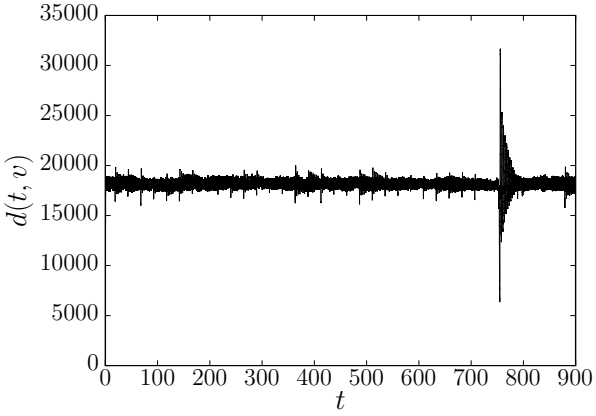


FIGURE 5.2 – **Anomalie non-identifiée avec le degré instantané des nœuds** – La suppression de l'ensemble des interactions impliquant le nœud v de $t_1 = 753$ à $t_2 = 775$ perturbe le comportement normal du flot de liens. Pour identifier cette anomalie, il est nécessaire de déterminer plus précisemment quels sont les liens $(t, u, v) \in T \times V \times V$ anormaux.

des propriétés plus complexes, impliquant une complexité de calculs plus élevée. Leur utilisation serait simplifiée par le fait que les anomalies les plus invasives aient déjà été supprimées du flot de liens à l'aide des propriétés de plus bas niveau. Pour permettre la détection d'anomalies, l'ensemble des propriétés considérées devront, soit être distribuées homogènement, soit être hétérogènes localement au niveau de leur structure mais homogènes sur le temps. À terme, notre méthode permettrait alors, pour un ensemble de propriétés donné, de récupérer à la fois un flot de liens normal et un ensemble d'anomalies correspondant à ces propriétés. Isoler les anomalies permettrait d'étudier leurs caractéristiques en profondeur, séparément du reste du flot de liens. Générer un flot de liens normal pourrait d'autre part être utile à d'autres applications. Nous en discutons dans la section suivante.

5.2.2 Génération de flots de liens normaux réalistes

La génération de flots de liens normaux vis-à-vis d'une propriété et d'un contexte donnés peut être mise à profit pour la validation des résultats d'un détecteur d'anomalies et pour la prédiction de liens.

Validation des résultats

Les jeux de données étiquetés représentant une vérité-terrain sont très rares, en particulier concernant les jeux de données en libres accès. Ainsi, il est souvent très difficile pour les chercheurs d'évaluer la qualité de leur méthode de détection.

Une première façon de faire, la plus courante, est l'évaluation interne [14]. Elle consiste à caractériser un comportement normal puis à quantifier l'anormalité des observations. Par exemple, déterminer la distribution empirique des observations puis calculer leurs valeurs-p, ce qui est équivalent à ce que nous avons fait dans cette thèse, en considérant le seuil d'anormalité $m + 3\sigma$ dans une distribution homogène [8]. Cependant, en utilisant une évaluation interne, la comparaison entre les différents travaux est difficilement envisageable étant donné que la représentation des interactions temporelles et les techniques utilisées influent beaucoup sur la nature des anomalies détectées, y compris lorsque les travaux portent sur les mêmes jeux de données.

Une autre façon de faire est alors d'évaluer les outils méthodologiques en les appliquant à des graphes synthétiques dans lesquels des anomalies ont été injectées [15, 11]. Cette évaluation permet ainsi de tester leurs comportements en faisant varier plusieurs paramètres tels que la taille du graphe ou la distribution du degré. Cependant, dans de tels graphes, le comportement normal est connu ce qui amène à s'interroger sur le succès de la méthode une fois appliquée à des graphes réels. Pour pallier à cela, il est également possible d'injecter des anomalies synthétiques dans un graphe réel. Cependant, l'évaluation de la méthode proposée est plus complexe étant donné que le graphe d'origine peut contenir des anomalies du même type que celles injectées.

Une fois notre méthode mise au point, nous serions capables de séparer les interactions normales des interactions anormales vis-à-vis de propriétés et de contextes donnés dans un flot de liens. Nous posséderions donc, d'une part, de flots de liens normaux selon différents contextes, et d'autre part, d'une base d'anomalies dans laquelle les anomalies sont classées selon la propriété et le contexte avec lesquels elles ont été identifiées. De ce fait, un chercheur pourrait choisir un flot de lien normal particulier, puis y injecter le nombre et les types d'anomalies qu'il souhaite pour évaluer sa méthode. De cette façon, nous améliorerions les techniques de validation des résultats de deux manières. Premièrement, les anomalies injectées correspondraient exactement à celles que la méthode est censée être capable de détecter. Deuxièmement, bien que le comportement normal du flot de liens choisi serait connu selon certaines propriétés, il n'en resterait pas moins un flot de liens réaliste comportant des structures complexes et dont le processus de formation n'est pas connu. Ainsi, la validation des résultats évaluerait d'une part le nombre d'anomalies injectées détectées (vrais positifs) et d'autre part sa capacité à être transparente aux autres anomalies (faux positifs).

Prédiction de liens

La prédiction de liens à partir d'un flot de liens $L(T, V, E)$, $T = [\alpha, \omega]$, consiste à prédire tout ou un sous-ensemble des liens du flot $L_p(T_p, V_p, E_p)$ où $T_p = [\alpha', \omega']$, $\omega < \alpha' < \omega'$ [18]. Comme nous l'avons mentionné dans la Section 3.5.2, la prédiction de liens est complémentaire à la détection d'anomalies. Notamment, elle permet d'identifier des liens anormaux parmi ceux s'écartant de la prédiction. Tout comme pour la détection d'anomalies, elle nécessite de comprendre et de caractériser les mécanismes responsables de l'apparition et de la disparition des liens au cours du temps. Or, les anomalies perturbent le comportement normal du flot de liens et, de ce fait, représentent un obstacle à l'apprentissage de leur dynamique. La génération d'un flot de liens normal à partir d'un ensemble de propriétés à l'aide de notre méthode pourrait donc s'avérer utile dans ce cas également. Par exemple, Arnoux et al. [19] cherchent à prédire le nombre d'interactions entre chaque paire de nœuds au cours d'une fenêtre de temps donnée. Pour cela, ils caractérisent le flot de liens à l'aide de plusieurs propriétés structurelles et temporelles, puis les combinent au sein d'un algorithme d'apprentissage supervisé. On peut alors supposer que leur méthode appliquée à un flot de liens dans lequel les anomalies correspondant à cet ensemble de propriétés ont été supprimées soit plus robuste et fournisse de meilleurs résultats.

Bibliographie

[1] MAWI Working Group Traffic Archive. http ://mawi.wide.ad.jp/mawi/. Consulté le 14/05/2019.

[2] National Centers for Environmental Information. https :www.ncdc.noaa.gov. Consulté le 28/03/2019.

[3] Netsimile : A scalable approach to size-independent network similarity, author=Berlingerio, Michele and Koutra, Danai and Eliassi-Rad, Tina and Faloutsos, Christos, journal=arXiv preprint arXiv :1209.2684, year=2012.

[4] Projet Politoscope, CNRS Institut des Systèmes Complexes Paris Ile-de-France (ISC-PIF). http ://politoscope.org. Consulté le 06/08/2018.

[5] Virus et malwares : les chercheurs contre-attaquent, C. Zeitoun, CNRS Le Journal. https ://lejournal.cnrs.fr/articles/virus-et-malwares-les-chercheurs-contre-attaquent. Consulté le 10/05/2019.

[6] R. Adler, R. Feldman, and M. Taqqu. *A practical guide to heavy tails : statistical techniques and applications*. Springer Science & Business Media, 1998.

[7] B. L. Agarwal. *Basic statistics*. New Age International, 2006.

[8] C. Aggarwal. *Outlier Analysis*. Springer International Publishing, 2016.

[9] C. C. Aggarwal and P. S. Yu. Online analysis of community evolution in data streams. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 56–67. SIAM, 2005.

[10] W. Ahmed, P. A. Bath, and G. Demartini. Using Twitter as a Data Source : An Overview of Ethical, Legal, and Methodological Challenges. In *The Ethics of Online Research*, pages 79–107. Emerald Publishing Limited, 2017.

[11] L. Akoglu and C. Faloutsos. RTG : a recursive realistic graph generator using random typing. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages=13–28, year=2009, organization=Springer.*

[12] L. Akoglu and C. Faloutsos. Event detection in time series of mobile communication graphs. In *Army science conference*, pages 77–79, 2010.

[13] L. Akoglu, M. McGlohon, and C. Faloutsos. Oddball : Spotting anomalies in weighted graphs. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 410–421. Springer, 2010.

[14] L. Akoglu, H. Tong, and D. Koutra. Graph based anomaly detection and description : a survey. *Data mining and knowledge discovery*, 29(3) :626–688, 2015.

[15] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1) :47, 2002.

[16] R. Albert, H. Jeong, and A.-L. Barabási. Internet : Diameter of the world-wide web. *nature*, 401(6749) :130, 1999.

[17] M. Araujo, S. Papadimitriou, S. Günnemann, C. Faloutsos, P. Basu, A. Swami, E. E. Papalexakis, and D. Koutra. Com2 : fast automatic discovery of temporal ('comet') communities. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 271–283. Springer, 2014.

[18] T. Arnoux. *Prédiction d'interactions dans les flots de liens, Combiner les caractéristiques structurelles et temporelles*. PhD thesis, Sorbonne Université, 2018.

[19] T. Arnoux, L. Tabourier, and M. Latapy. Combining structural and dynamic information to predict activity in link streams. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 935–942. ACM, 2017.

[20] H. Asai, K. Fukuda, P. Abry, P. Borgnat, and H. Esaki. Network application profiling with traffic causality graphs. *International Journal of Network Management*, 24(4) :289–303, 2014.

[21] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439) :509–512, 1999.

[22] P. Barford, J. Kline, D. Plonka, and A. Ron. A signal analysis of network traffic anomalies. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*, pages 71–82. ACM, 2002.

[23] V. Batagelj and S. Praprotnik. An algebraic approach to temporal network analysis based on temporal quantities. *Social Network Analysis and Mining*, 6(1) :28, 2016.

[24] D. R. Bild, Y. Liu, R. P. Dick, Z. M. Mao, and D. S. Wallach. Aggregate characterization of user behavior in Twitter and analysis of the retweet graph. *ACM Transactions on Internet Technology (TOIT)*, 15(1) :4, 2015.

[25] A. W. Bohannon, B. M. Sadler, and R. V. Balan. A Filtering Framework for Time-Varying Graph Signals. In *Vertex-Frequency Analysis of Graph Signals*, pages 341–376. Springer, 2019.

[26] P. Borgnat, G. Dewaele, K. Fukuda, P. Abry, and K. Cho. Seven years and one day : Sketching the evolution of internet traffic. In *INFOCOM 2009, IEEE*, pages 711–719. IEEE, 2009.

[27] R. Bro. PARAFAC. Tutorial and applications. *Chemometrics and intelligent laboratory systems*, 38(2) :149–171, 1997.

[28] A. Bruns, J. E Burgess, K. Crawford, and F. Shaw. # qldfloods and@ QPSMedia : Crisis communication on Twitter in the 2011 south east Queensland floods. *ARC Centre of Excellence for Creative Industries and Innovation*, 01 2012.

[29] F. M. Cardoso, S. Meloni, A. Santanche, and Y. Moreno. Topical homophily in online social systems. *arXiv preprint arXiv :1707.06525*, 2017.

[30] P. Casas, J. Mazel, and P. Owezarski. Unada : Unsupervised network anomaly detection using sub-space outliers ranking. In *International Conference on Research in Networking*, pages 40–51. Springer, 2011.

[31] A. Casteigts, P. Flocchini, W. Quattrociocchi, and N. Santoro. Time-varying graphs and dynamic networks. *International Journal of Parallel, Emergent and Distributed Systems*, 27(5) :387–408, 2012.

[32] D. Chakrabarti. Autopart : Parameter-free graph partitioning and outlier detection. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 112–124. Springer, 2004.

[33] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection : A survey. *ACM computing surveys (CSUR)*, 41(3) :15, 2009.

[34] N. Chavoshi, H. Hamooni, and A. Mueen. Temporal patterns in bot activities. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1601–1606. International World Wide Web Conferences Steering Committee, 2017.

[35] Z. Chen, W. Hendrix, and N. F. Samatova. Community-based anomaly detection in evolutionary networks. *Journal of Intelligent Information Systems*, 39(1) :59–85, 2012.

[36] F. Chierichetti, J. M. Kleinberg, R. Kumar, M. Mahdian, and S. Pandey. Event Detection via Communication Pattern Analysis. In *ICWSM, year=2014*.

[37] K. Cho. Recursive lattice search : hierarchical heavy hitters revisited. In *Proceedings of the 2017 Internet Measurement Conference*, pages 283–289. ACM, 2017.

[38] B. Claise. Cisco systems netflow services export version 9. Technical report, 2004.

[39] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4) :661–703, 2009.

[40] M. Coletto, K. Garimella, A. Gionis, and C. Lucchese. Automatic controversy detection in social media : A content-independent motif-based approach. *Online Social Networks and Media*, 3 :22–31, 2017.

[41] E. Colleoni, A. Rozza, and A. Arvidsson. Echo chamber or public sphere ? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of communication*, 64(2) :317–332, 2014.

[42] G. Cormode and S. Muthukrishnan. An improved data stream summary : the count-min sketch and its applications. *Journal of Algorithms*, 55(1) :58–75, 2005.

[43] D. Corney, C. Martin, and A. Göker. Spot the ball : Detecting sports events on Twitter. In *European Conference on Information Retrieval*, pages 449–454. Springer, 2014.

[44] E. C. Costa, A. B. Vieira, K. Wehmuth, A. Ziviani, and A. P. C. Da Silva. Time centrality in dynamic complex networks. *Advances in Complex Systems*, 18(07n08) :1550023, 2015.

[45] L. Cruickshank, E. Tsekleves, R. Whitham, A. Hill, and K. Kondo. Making interactive TV easier to use : Interface design for a second screen approach. *The Design Journal*, 10(3) :41–53, 2007.

[46] X. Dong, D. Mavroeidis, F. Calabrese, and P. Frossard. Multiscale event detection in social media.

[47] D. Duan, Y. Li, Y. Jin, and Z. Lu. Community mining on dynamic weighted directed graphs. In *Proceedings of the 1st ACM international workshop on Complex networks meet information & knowledge management*, pages 11–18. ACM, 2009.

[48] W. Eberle and L. Holder. Discovering structural anomalies in graph-based data. In *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*, pages 393–398. IEEE, 2007.

[49] D. Eswaran and C. Faloutsos. Sedanspot : Detecting anomalies in edge streams. *ICDM. IEEE*, 2018.

[50] D. Eswaran, C. Faloutsos, S. Guha, and N. Mishra. Spotlight : Detecting anomalies in streaming graphs. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1378–1386. ACM, 2018.

[51] G. Fernandes and P. Owezarski. Automated classification of network traffic anomalies. In *International Conference on Security and Privacy in Communication Systems*, pages 91–100. Springer, 2009.

[52] R. Fontugne, P. Abry, K. Fukuda, D. Veitch, K. Cho, P. Borgnat, and H. Wendt. Scaling in internet traffic : a 14 year and 3 day longitudinal study, with multiscale analyses and random projections. *IEEE/ACM Transactions on Networking (TON)*, 25(4) :2152–2165, 2017.

[53] R. Fontugne, P. Borgnat, P. Abry, and K. Fukuda. Mawilab : combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking. In *Proceedings of the 6th International COnference*, page 8. ACM, 2010.

[54] D. Freelon and D. Karpf. Of big birds and bayonets : Hybrid Twitter interactivity in the 2012 presidential debates. *Information, Communication & Society*, 18(4) :390–406, 2015.

[55] F. Gama, E. Isufi, G. Leus, and A. Ribeiro. Control of Graph Signals Over Random Time-Varying Graphs. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4169–4173. IEEE, 2018.

[56] J. Gao, F. Liang, W. Fan, C. Wang, Y. Sun, and J. Han. On community outliers and their efficient detection in information networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 813–822. ACM, 2010.

[57] N. Gaumont, M. Panahi, and D. Chavalarias. Reconstruction of the socio-semantic dynamics of political activist Twitter networks-Method and application to the 2017 French presidential election. *PLoS ONE*, 13(9), 2018.

[58] F. Giglietto and D. Selva. Second screen and participation : A content analysis on a full season dataset of tweets. *Journal of Communication*, 64(2) :260–277, 2014.

[59] H. Gil de Zúñiga, V. Garcia-Perdomo, and S. C. McGregor. What is second screening ? Exploring motivations of second screen use and its effect on online political participation. *Journal of Communication*, 65(5) :793–815, 2015.

[60] C. Grasland, R. Lamarche-Perrin, B. Loveluck, and H. Pecout. International agenda-setting, the media and geography : A multi-dimensional analysis of news flows. *L'Espace géographique*, 45(1) :25–43, 2016.

[61] F. E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1) :1–21, 1969.

[62] M. Gupta, C. C. Aggarwal, and J. Han. Finding top-k shortest path distance changes in an evolutionary network. In *International Symposium on Spatial and Temporal Databases*, pages 130–148. Springer, 2011.

[63] M. Gupta, C. C. Aggarwal, J. Han, and Y. Sun. Evolutionary clustering and analysis of bibliographic networks. In *2011 International Conference on Advances in Social Networks Analysis and Mining*, pages 63–70. IEEE, 2011.

[64] M. Gupta, J. Gao, Y. Sun, and J. Han. Community trend outlier detection using soft temporal pattern mining. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 692–708. Springer, 2012.

[65] M. Gupta, J. Gao, Y. Sun, and J. Han. Integrating community matching and outlier detection for mining evolutionary community outliers. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 859–867. ACM, 2012.

[66] S. Gurukar and B. Ravindran. Temporal analysis of telecom call graphs. In *2014 Sixth International Conference on Communication Systems and Networks (COMSNETS)*, pages 1–6. IEEE, 2014.

[67] R. Hamon, P. Borgnat, P. Flandrin, and C. Robardet. Networks as signals, with an application to a bike sharing system. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 611–614. IEEE, 2013.

[68] R. Hamon, P. Borgnat, P. Flandrin, and C. Robardet. Transformation de graphes dynamiques en signaux non stationnaires. In *Colloque GRETSI 2013*, page 251, 2013.

[69] R. Hamon, P. Borgnat, P. Flandrin, and C. Robardet. From graphs to signals and back : Identification of network structures using spectral analysis. *arXiv preprint arXiv :1502.04697*, 2015.

[70] R. Hamon, P. Borgnat, P. Flandrin, and C. Robardet. Extraction of temporal network structures from graph-based signals. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2) :215–226, 2016.

[71] R. Hamon, P. Borgnat, P. Flandrin, and C. Robardet. Transformation from Graphs to Signals and Back. In *Vertex-Frequency Analysis of Graph Signals*, pages 111–139. Springer, 2019.

[72] J. Han, J. Pei, and M. Kamber. *Data mining : concepts and techniques*. Elsevier, 2011.

[73] A. Hanbanchong and K. Piromsopa. SARIMA based network bandwidth anomaly detection. In *2012 Ninth International Conference on Computer Science and Software Engineering (JCSSE)*, pages 104–108. IEEE, 2012.

[74] D. M. Hawkins. *Identification of outliers*, volume 11. Springer, 1980.

[75] S. Hirose, K. Yamanishi, T. Nakata, and R. Fujimaki. Network anomaly detection based on eigen equation compression. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1185–1194. ACM, 2009.

[76] P. Holme and J. Saramäki. Temporal networks. *Physics reports*, 519(3) :97–125, 2012.

[77] M. Hubert and E. Vandervieren. An adjusted boxplot for skewed distributions. *Computational statistics & data analysis*, 52(12) :5186–5201, 2008.

[78] T. Idé and H. Kashima. Eigenspace-based anomaly detection in computer systems. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 440–449. ACM, 2004.

[79] M. Iliofotou, M. Faloutsos, and M. Mitzenmacher. Exploiting dynamicity in graph-based traffic analysis : techniques and applications. In *Proceedings of the 5th international conference on Emerging networking experiments and technologies*, pages 241–252. ACM, 2009.

[80] M. Iliofotou, P. Pappu, M. Faloutsos, M. Mitzenmacher, S. Singh, and G. Varghese. Network monitoring using traffic dispersion graphs (TDGs). In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 315–320. ACM, 2007.

[81] K. Ishibashi, T. Kondoh, S. Harada, T. Mori, R. Kawahara, and S. Asano. Detecting anomalous traffic using communication graphs. In *Telecommunications : The Infrastructure for the 21st Century (WTC), 2010*, pages 1–6. VDE, 2010.

[82] D. Jiang, Z. Xu, P. Zhang, and T. Zhu. A transform domain-based anomaly detection approach to network-wide traffic. *Journal of Network and Computer Applications*, 40 :292–306, 2014.

[83] M. Jiang, A. Beutel, P. Cui, B. Hooi, S. Yang, and C. Faloutsos. A general suspiciousness metric for dense blocks in multimodal data. In *2015 IEEE International Conference on Data Mining*, pages 781–786. IEEE, 2015.

[84] Y. Kanda, K. Fukuda, and T. Sugawara. Evaluation of anomaly detection based on sketch and PCA. In *2010 IEEE Global Telecommunications Conference GLOBECOM 2010*, pages 1–5. IEEE, 2010.

[85] M. Karsai, H.-H. Jo, and K. Kaski. *Bursty human dynamics*. Springer, 2018.

[86] M. Karsai, K. Kaski, A.-L. Barabási, and J. Kertész. Universal features of correlated bursty behaviour. *Scientific reports*, 2 :397, 2012.

[87] A. Kato, J. Murai, S. Katsuno, and T. Asami. An Internet traffic data repository : The architecture and the design policy. In *INET'99 Proceedings*, 1999.

[88] L. B. Klebanov. Big Outliers Versus Heavy Tails : what to use ? *arXiv preprint arXiv :1611.05410*, 2016.

[89] E. M. Knorr and R. T. Ng. Finding intensional knowledge of distance-based outliers. In *VLDB*, volume 99, pages 211–222, 1999.

[90] V. Kostakos. Temporal graphs. *Physica A : Statistical Mechanics and its Applications*, 388(6) :1007–1023, 2009.

[91] A. Lakhina, M. Crovella, and C. Diot. Diagnosing network-wide traffic anomalies. In *ACM SIGCOMM Computer Communication Review*, volume 34, pages 219–230. ACM, 2004.

[92] M. Latapy, A. Hamzaoui, and C. Magnien. Detecting events in the dynamics of ego-centred measurements of the Internet topology. *Journal of Complex Networks*, 2(1) :38–59, 2013.

[93] M. Latapy, T. Viard, and C. Magnien. Stream graphs and link streams for the modeling of interactions over time. *Social Network Analysis and Mining*, 8(1) :61, 2018.

[94] Y. Léo, C. Crespelle, and E. Fleury. Non-altering time scales for aggregation of dynamic networks into series of graphs. *Computer Networks*, 148 :108–119, 2019.

[95] R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang. Tedas : A twitter-based event detection and analysis system. In *Data engineering (icde), 2012 ieee 28th international conference on*, pages 1273–1276. IEEE, 2012.

[96] K.-W. Lim, S. Secci, L. Tabourier, and B. Tebbani. Characterizing and predicting mobile application usage. *Computer Communications*, 95 :82–94, 2016.

[97] N. Liu, D. Shin, and X. Hu. Contextual outlier interpretation. *arXiv preprint arXiv :1711.10589*, 2017.

[98] A. Loukas and D. Foucard. Frequency analysis of time-varying graph signals. In *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 346–350. IEEE, 2016.

[99] W. Lu and A. A. Ghorbani. Network anomaly detection based on wavelet analysis. *EURASIP Journal on Advances in Signal Processing*, 2009 :4, 2009.

[100] O. Macindoe and W. Richards. Graph comparison using fine structure analysis. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 193–200. IEEE, 2010.

[101] J. Maddock, K. Starbird, and R. M. Mason. Using historical Twitter data for research : Ethical challenges of tweet deletions. In *CSCW 2015 Workshop on Ethics for Studying Sociotechnical Systems in a Big Data World. ACM*, 2015.

[102] E. Manzoor, S. M. Milajerdi, and L. Akoglu. Fast memory-efficient anomaly detection in streaming heterogeneous graphs. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1035–1044. ACM, 2016.

[103] N. Masuda and P. Holme. Detecting sequences of system states in temporal networks. *Scientific reports*, 9(1) :795, 2019.

[104] J. Mazel, P. Casas, R. Fontugne, K. Fukuda, and P. Owezarski. Hunting attacks in the dark : clustering and correlation analysis for unsupervised anomaly detection. *International Journal of Network Management*, 25(5) :283–305, 2015.

[105] J. Mazel, R. Fontugne, and K. Fukuda. A taxonomy of anomalies in backbone network traffic. In *Wireless Communications and Mobile Computing Conference (IWCMC), 2014 International*, pages 30–36. IEEE, 2014.

[106] H. Z. Moayedi and M. Masnadi-Shirazi. Arima model for network traffic prediction and anomaly detection. In *2008 International Symposium on Information Technology*, volume 4, pages 1–6. IEEE, 2008.

[107] M. Mongiovi, P. Bogdanov, R. Ranca, E. E. Papalexakis, C. Faloutsos, and A. K. Singh. Netspot : Spotting significant anomalous regions on dynamic networks. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 28–36. SIAM, 2013.

[108] J. Moody. The importance of relationship timing for diffusion. *Social Forces*, 81(1) :25–56, 2002.

[109] H. Moonesinghe and P.-N. Tan. Outrank : a graph-based outlier detection framework using random walk. *International Journal on Artificial Intelligence Tools*, 17(01) :19–36, 2008.

[110] H. J. Motulsky and R. E. Brown. Detecting outliers when fitting data with nonlinear regression–a new method based on robust nonlinear regression and the false discovery rate. *BMC bioinformatics*, 7(1) :123, 2006.

[111] M. Newman. *Networks : An Introduction*. Oxford University Press, Inc., New York, NY, USA, 2010.

[112] S. M. T. Nezhad, M. Nazari, and E. A. Gharavol. A novel DoS and DDoS attacks detection algorithm using ARIMA time series model and chaotic system in computer networks. *IEEE Communications Letters*, 20(4) :700–703, 2016.

[113] C. C. Noble and D. J. Cook. Graph-based anomaly detection. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636. ACM, 2003.

[114] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls : Linking text sentiment to public opinion time series. In *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[115] P. Papadimitriou, A. Dasdan, and H. Garcia-Molina. Web graph similarity for anomaly detection. *Journal of Internet Services and Applications*, 1(1) :19–30, 2010.

[116] L. Pépin, J. Blanchard, F. Guillet, P. Kuntz, and P. Suignard. Visual analysis of topics in Twitter based on co-evolution of terms. In *Data Science, Learning by Latent Structures, and Knowledge Discovery*, pages 169–178. Springer, 2015.

[117] R. Perline. Strong, weak and false inverse power laws. *Statistical Science*, pages 68–88, 2005.

[118] N. Perraudin, A. Loukas, F. Grassi, and P. Vandergheynst. Towards stationary time-vertex signal processing. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3914–3918. Ieee, 2017.

[119] B. Pincombe. Anomaly detection in time series of graphs using arma processes. *Asor Bulletin*, 24(4) :2, 2005.

[120] S. Praprotnik and V. Batagelj. Spectral centrality measures in temporal networks. *Ars Mathematica Contemporanea*, 11(1) :11–33, 2015.

[121] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. Numerical Recipes in C : The Art of Scientific Computing. Second Edition, 1992.

[122] S. Ranshous, S. Harenberg, K. Sharma, and N. F. Samatova. A scalable approach for outlier detection in edge streams using sketch-based approximations. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 189–197. SIAM, 2016.

[123] S. Ranshous, S. Shen, D. Koutra, S. Harenberg, C. Faloutsos, and N. F. Samatova. Anomaly detection in dynamic networks : a survey. *Wiley Interdisciplinary Reviews : Computational Statistics*, 7(3) :223–247, 2015.

[124] F. Ren and Y. Wu. Predicting user-topic opinions in twitter with social and topical context. *IEEE Transactions on Affective Computing*, 4(4) :412–424, 2013.

[125] A. Reyes-Menendez, J. Saura, and C. Alvarez-Alonso. Understanding# WorldEnvironmentDay user opinions in Twitter : A topic-based sentiment analysis approach. *International journal of environmental research and public health*, 15(11) :2537, 2018.

[126] M. H. Ribeiro, P. H. Calais, Y. A. Santos, V. A. Almeida, and W. Meira Jr. Characterizing and Detecting Hateful Users on Twitter. *arXiv preprint arXiv :1803.08977*, 2018.

[127] H. Ringberg, A. Soule, J. Rexford, and C. Diot. Sensitivity of PCA for traffic anomaly detection. *ACM SIGMETRICS Performance Evaluation Review*, 35(1) :109–120, 2007.

[128] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users : real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[129] R. Saxena, S. Kaur, D. Dash, and V. Bhatnagar. Leveraging Structural Hierarchy for Scalable Network Comparison. In *International Conference on Database and Expert Systems Applications*, pages 287–302. Springer, 2016.

[130] Y. Shimada, T. Ikeguchi, and T. Shigehara. From networks to time series. *Physical review letters*, 109(15) :158701, 2012.

[131] K. Shin, B. Hooi, and C. Faloutsos. M-zoom : Fast dense-block detection in tensors with quality guarantees. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 264–280. Springer, 2016.

[132] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst. The emerging field of signal processing on graphs : Extending high-dimensional data analysis to networks and other irregular domains. *arXiv preprint arXiv :1211.0053*, 2012.

[133] J. Si, A. Mukherjee, B. Liu, Q. Li, H. Li, and X. Deng. Exploiting topic based twitter sentiment for stock prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, volume 2, pages 24–29, 2013.

[134] J. Song, S. Lee, and J. Kim. Spam filtering in twitter using sender-receiver relationship. In *International workshop on recent advances in intrusion detection*, pages 301–317. Springer, 2011.

[135] S. Stieglitz and L. Dang-Xuan. Political communication and influence through microblogging–An empirical analysis of sentiment in Twitter messages and retweet behavior. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 3500–3509. IEEE, 2012.

[136] J. Sun, C. Faloutsos, S. Papadimitriou, and P. S. Yu. Graphscope : parameter-free mining of large time-evolving graphs. In *Proceedings of the 13th ACM SIGKDD*

international conference on Knowledge discovery and data mining, pages 687–696. ACM, 2007.

[137] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 8–pp. IEEE, 2005.

[138] J. Sun, Y. Xie, H. Zhang, and C. Faloutsos. Less is more : Sparse graph mining with compact matrix decomposition. *Statistical Analysis and Data Mining : The ASA Data Science Journal*, 1(1) :6–22, 2008.

[139] T. Takaguchi, Y. Yano, and Y. Yoshida. Coverage centralities for temporal networks. *The European Physical Journal B*, 89(2) :35, 2016.

[140] M. Ten Thij, T. Ouboter, D. Worm, N. Litvak, H. van den Berg, and S. Bhulai. Modelling of trends in twitter using retweet graph dynamics. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 132–147. Springer, 2014.

[141] E. Tsekleves, L. Cruickshank, A. Hill, K. Kondo, and R. Whitham. Interacting with digital media at home via a second screen. In *Ninth IEEE International Symposium on Multimedia Workshops (ISMW 2007)*, pages 201–206. IEEE, 2007.

[142] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini. Online human-bot interactions : Detection, estimation, and characterization. *arXiv preprint arXiv :1703.03107*, 2017.

[143] V. Verardi and C. Vermandele. Univariate and multivariate outlier identification for skewed or heavy-tailed distributions. *The Stata Journal*, 18(3) :517–532, 2018.

[144] T. Viard. *Flots de liens pour la modélisation d'interactions temporelles et application à l'analyse de trafic IP*. PhD thesis, Paris 6, 2016.

[145] T. Viard, R. Fournier-S'niehotta, C. Magnien, and M. Latapy. Discovering patterns of interest in IP traffic using cliques in bipartite link streams. *In Proceedings of the International Conference on Complex Networks (CompleNet), 2018*, 2018.

[146] Y. Virkar and A. Clauset. Power-law distributions in binned empirical data. *The Annals of Applied Statistics*, pages 89–119, 2014.

[147] M. Vlachos, P. Yu, and V. Castelli. On periodicity detection and structural periodic similarity. In *Proceedings of the 2005 SIAM international conference on data mining*, pages 449–460. SIAM, 2005.

[148] I. Voitalov, P. van der Hoorn, R. van der Hofstad, and D. V. Krioukov. Scale-free Networks Well Done. *CoRR*, abs/1811.02071, 2018.

[149] M. Walther and M. Kaisser. Geo-spatial event detection in the twitter stream. In *European conference on information retrieval*, pages 356–367. Springer, 2013.

[150] H. Webb, M. Jirotka, B. C. Stahl, W. Housley, A. Edwards, M. Williams, R. Procter, O. Rana, and P. Burnap. The ethical challenges of publishing Twitter data for research dissemination. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 339–348. ACM, 2017.

[151] K. Wehmuth, A. Ziviani, and E. Fleury. A unifying model for representing time-varying graphs. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pages 1–10. IEEE, 2015.

[152] J. Whitbeck, M. Dias de Amorim, V. Conan, and J.-L. Guillaume. Temporal reachability graphs. In *Proceedings of the 18th annual international conference on Mobile computing and networking*, pages 377–388. ACM, 2012.

[153] N. Williams, S. Zander, and G. Armitage. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. *ACM SIGCOMM Computer Communication Review*, 36(5) :5–16, 2006.

[154] A. Wilmet and R. Lamarche-Perrin. Multidimensional Outlier Detection in Interaction Data : Application to Political Communication on Twitter. In *International Workshop on Complex Networks*, pages 147–155. Springer, 2019.

[155] A. Wilmet and R. Lamarche-Perrin. Multidimensional Outlier Detection in Temporal Interaction Networks : An Application to Political Communication on Twitter. *En cours de soumission à SNAM*, 2019.

[156] A. Wilmet, T. Viard, M. Latapy, and R. Lamarche-Perrin. Degree-based Outliers Detection within IP Traffic Modelled as a Link Stream. In *2018 Network Traffic Measurement and Analysis Conference (TMA)*, pages 1–8. IEEE, 2018.

[157] A. Wilmet, T. Viard, M. Latapy, and R. Lamarche-Perrin. Degree-based Outlier Detection within IP Traffic Modelled as a Link Stream. *Accepté à Computer Networks, en cours de révision*, 2019.

[158] F. M. F. Wong, C. W. Tan, S. Sen, and M. Chiang. Quantifying political leaning from tweets, retweets, and retweeters. *IEEE transactions on knowledge and data engineering*, 28(8) :2158–2172, 2016.

[159] W. Xie, F. Zhu, J. Jiang, E.-P. Lim, and K. Wang. Topicsketch : Real-time bursty topic detection from twitter. *IEEE Transactions on Knowledge and Data Engineering*, 28(8) :2216–2229, 2016.

[160] K. Xu, F. Wang, and L. Gu. Behavior Analysis of Internet Traffic via Bipartite Graphs and One-Mode Projections. *IEEE/ACM Transactions on Networking*, 22 :931–942, 2014.

[161] W. Yu, C. C. Aggarwal, S. Ma, and H. Wang. On anomalous hotspot discovery in graph streams. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 1271–1276. IEEE, 2013.