**THÈSE DE DOCTORAT DE L'UNIVERSITÉ PIERRE ET MARIE CURIE**

Spécialité

**Informatique**

Présentée par

Assia HAMZAOUI

épse. GHOUTI

Pour obtenir le grade de

**Docteur de l'Université Pierre et Marie Curie**

# Détection d'événements dans la dynamique des grands graphes de terrain : une approche statistique et son application au radar de l'internet

Thèse dirigée par

Matthieu LATAPY

Soutenue le 28 Juin 2011

**Jury :**

| | | | |
|---|---|---|---|
| *Rapporteurs :* | Hugues FAUCONNIER | - | MdC HDR, Paris-Diderot |
| | Damien MAGONI | - | Professeur, Université Bordeaux 1 |
| *Examinateurs :* | Pierre BORGNAT | - | CR1 CNRS, ENS de Lyon |
| | Christophe CRESPELLE | - | MdC, Université Lyon 1 |
| | Guy MÉLANÇON | - | Professeur, Université Bordeaux 1 |
| | Michèle SORIA | - | Professeur, UPMC Sorbonne |
| *Directeur :* | Matthieu LATAPY | - | DR CNRS, UPMC Sorbonne |

**THÈSE DE DOCTORAT DE L'UNIVERSITÉ PIERRE ET MARIE CURIE**

Spécialité

**Informatique**

Présentée par

Assia HAMZAOUI

épse. GHOUTI

Pour obtenir le grade de

**Docteur de l'Université Pierre et Marie Curie**

# Détection d'événements dans la dynamique des grands graphes de terrain : une approche statistique et son application au radar de l'internet

Thèse dirigée par

Matthieu LATAPY

Soutenue le 28 Juin 2011

**Jury :**

| | | | |
|---|---|---|---|
| *Rapporteurs :* | Hugues FAUCONNIER | - | MdC HDR, Paris-Diderot |
| | Damien MAGONI | - | Professeur, Université Bordeaux 1 |
| *Examinateurs :* | Pierre BORGNAT | - | CR1 CNRS, ENS de Lyon |
| | Christophe CRESPELLE | - | MdC, Université Lyon 1 |
| | Guy MÉLANÇON | - | Professeur, Université Bordeaux 1 |
| | Michèle SORIA | - | Professeur, UPMC Sorbonne |
| *Directeur :* | Matthieu LATAPY | - | DR CNRS, UPMC Sorbonne |

# Remerciements

Mes premiers remerciements reviennent à Hugues Fauconnier et Damien Magoni pour avoir accepté de rapporter cette thèse. Je remercie tout autant Pierre Borgnat, Guy Mélançon, Michèle Soria et Christophe Crespelle d'avoir bien voulu faire partie de mon jury.

Je tiens à exprimer toute ma gratitude envers mon directeur et mon encadrant de thèse, Matthieu Latapy, pour m'avoir accepté dans son équipe, de m'avoir permis de travailler pour cette thèse dans les meilleures conditions possibles, de m'avoir encadré avec beaucoup de rigueur, ses conseils et ses encouragements sont derrière chacun des résultats obtenus.

Je souhaite également remercier les membres de l'équipe "ComplexNetworks" du Laboratoire d'Informatique de Paris 6, avec qui j'ai partagé cette expérience enrichissante sur tous les plans et qui me laisseront tous de très agréables souvenirs. Je remercie particulièrement Clémence Magnien pour ses relectures, remarques et discussions pertinentes. Merci à Jean Loup Guillaume pour "ses courses à pied" et à Bénédicte Le Grand pour son soutien. Ces années passées au LIP6 m'ont donné l'opportunité de côtoyer de nombreux collègues dans l'équipe. En particulier, je tiens à mentionner dans cette page mes collègues du premier (ou deuxième) bureau Hamid et Oussama, pour l'ambiance au bureau et pour leur complicité, sans oublier Thomas, Lamia, Lionel, Massoud et Raphaël. Merci à tous les autres membres de l'équipe.

Un grand Merci à Marguerite Sos et Véronique Varenne, pour leur gentillesse et pour avoir répondu avec tant de professionnalisme et de bonne humeur à toutes mes sollicitations.

Je ne peux m'empêcher de penser à toutes mes copines de la fac de Constantine, Lamy, Wafy et Emira.

C'est plus que des remerciements que j'adresse à mes parents Ali et Alia, à qui je dédie cette thèse, leur affection, leur soutien et leurs encouragements sont depuis plus de vingt ans mes principales sources de motivation. C'est avec plaisir que je mentionne aussi ma soeur Nawel et sa petite famille, son mari Fares et ses petits choux, et mes frères : Salim, Fethi, Houssem et Seddik, pour les remercier d'être toujours là quand j'en ai besoin.

Enfin, merci à mon mari et ma moitié Adel, pour sa compréhension, ses encouragements et toute son aide afin de me laisser consacrer un maximum de temps à cette thèse, merci d'être là, à mes côtés... à notre fée Lynn.

# Table des matières

Chapitre 1

# Introduction

On peut modéliser de nombreux objets issus du monde réel par des graphes. Citons par exemple la topologie de l'internet (au niveau des routeurs ou des systèmes autonomes), les graphes du web (ensembles de pages web et liens hypertextes entre elles), les réseaux métaboliques (réactions entre protéines au sein d'une cellule), les connexions dans le cerveau, les réseaux sociaux comme les réseaux d'amitié, de communication ou de collaboration, et les réseaux de transport.

Depuis une dizaine d'années, un grand nombre de travaux s'intéresse à ces objets, suite à la découverte que, bien qu'ils soient issus de contextes différents, ils se ressemblent au sens de certaines propriétés statistiques [Watts 1998, Barabási 1999]. Ces travaux ont étudié une grande variété de graphes et fait des observations importantes dans ces divers contextes. D'autre part, le fait qu'ils se ressemblent a fait émerger plusieurs questions transversales, c'est-à-dire s'appliquant à l'ensemble de ces graphes. On peut citer en particulier l'étude de leur structure en communautés, c'est-à-dire en groupes de nœuds fortement reliés les uns aux autres et faiblement liés à des nœuds d'autres groupes [Wasserman 1994, Girvan 2002, Fortunato 2004, Blondel 2008], l'introduction de diverses propriétés statistiques pour leur description, ou encore leur robustesse face à des pannes ou des attaques [Guillaume 2006, Callaway 2000]. Ceci a montré qu'il est effectivement pertinent de considérer ces objets comme un ensemble cohérent. C'est pourquoi on les désigne sous le terme général de graphes de terrain (*complex networks* en anglais).

Il est possible de regrouper les différentes questions liées à l'étude des graphes de terrain en grandes familles de problématiques [Latapy 2007]. En particulier, les graphes de terrain ne sont pas donnés a priori et la connaissance des nœuds et liens qui les composent passe par une opération de *mesure*, dont la nature dépend du graphe étudié. Remarquons que dans l'immense majorité des cas, la mesure ne peut espérer capturer l'intégralité du graphe, en particulier en raison de sa taille et de différentes autres contraintes. Il a été montré que l'échantillon collecté peut avoir des propriétés statistiques différentes de celles du graphe de départ, et que donc la mesure induit un biais dans les observations. La métrologie vise à étudier ce biais, le corriger et/ou proposer des méthodes capables de capturer certaines propriétés de manière fiable. La taille par ailleurs, des graphes de terrain est en général très grande (on parle dans beaucoup de cas de centaines de milliers, voire de millions ou de milliards de nœuds

et de liens), ce qui rend impossible, une fois les données collectées, de comprendre leur structure par une observation directe. L'*analyse* vise à décrire cette structure, par l'introduction de propriétés statistiques (distribution des degrés, coefficient de *clustering*, etc.) ou structurelles (hiérarchie de communautés par exemple) qui résument l'information et en soulignent les principales caractéristiques. On peut citer également les problématiques liées à la *modélisation*, l'*algorithmique*, et les *phénomènes* ayant lieu sur des graphes de terrain (par exemple des phénomènes de diffusion).

La plupart des graphes de terrain sont de plus dynamiques, c'est-à-dire que leur structure évolue au fil du temps par l'ajout et/ou le retrait de nœuds et/ou de liens, à des fréquences plus ou moins grandes. La grande majorité des travaux qui les ont étudiés les ont cependant considérés comme statiques, c'est-à-dire qu'ils ont considéré un instantané d'un graphe, capturé à un instant donné. Ceci est naturel, car entamer d'emblée l'étude de la dynamique sans connaissances préalables de la structure serait une tâche extrêmement ardue, et l'étude des graphes statiques a produit un grand nombre de résultats importants. La dynamique est cependant une composante fondamentale des graphes de terrain, et le domaine a aujourd'hui atteint une maturité suffisante pour aborder son étude. Ceci s'est traduit par l'apparition, depuis quelques années, de travaux sur la dynamique des graphes de terrain. Certains se sont intéressés à des cas particuliers, comme les contacts entre personnes mesurés à l'aide de capteurs ou de téléphones bluetooth par exemple [Calegari 2007, Chaintreau 2005, Clauset 2007, Scherrer 2008, Tournoux 2009], les échanges pair-à-pair [Le-Blond 2005, Guillaume 2004, Blond 2009], la topologie de l'internet [Latapy 2008, Oliveira 2007, Magnien 2009, Park 2004, Pansiot 2007, Lad 2006], des réseaux biologiques [Babu 2004, Tarissan 2009, Steuer 2007], ou des réseaux de citations d'articles [Leicht 2007]. et plusieurs types de réseaux sociaux en ligne [Cointet 2009, Stoica 2009, Schneider 2009].

Malgré ces efforts dirigés vers l'étude du comportement dynamique des graphes de terrain, il subsiste aujourd'hui un manque crucial de formalismes et de notions adaptés au contexte dynamique. Exactement de la même façon qu'une intense activité d'analyse sur ces dix dernières années a mené à un ensemble aujourd'hui relativement large et cohérent de notions pour le cas statique, il s'agit maintenant de développer l'équivalent dans le cas dynamique. Bien sûr, la tendance naturelle est de se contenter d'étendre les notions statiques, mais ceci n'est pas pleinement satisfaisant : des nouvelles notions, qui ne prennent sens que dans le contexte dynamique, doivent être développées.

Une problématique type mettant en évidence ces besoins de formalises et de notions adaptés au contexte dynamique est la détection d'événements. Cette problématique est motivée par deux objectifs principaux. Tout d'abord, il s'agit de surveiller l'objet et localiser (dans le temps et l'espace) des problèmes comme par exemple des attaques, pannes, congestions, dans le cas de la topologie de l'internet ou des échanges au niveau

IP. Dans d'autres cas, comme par exemple les réseaux sociaux, la motivation se rapproche plus de la fouille de données, puisqu'il s'agit d'identifier des moments ou des endroits où quelque chose d'inhabituel (événement) a lieu ; on étudiera plus en profondeur cet événement détecté. Cette problématique suppose toutefois qu'il y a une notion de normalité concernant la dynamique de ces réseaux, et que nous sommes capables de faire la distinction entre ce régime et des événements qui seraient anormaux. Dans de nombreux cas, ceci est loin d'être une évidence : les objets se révèlent beaucoup plus dynamiques qu'on le pensait, cette dynamique est extrêmement hétérogène, elle dépend de la localisation dans le réseau, etc. Il est possible aussi que la question n'ait pas de sens si elle est posée ainsi : il peut être normal qu'il y ait des anomalies... ou encore l'anomalie peut être si omniprésente qu'elle est la norme. Dans cette direction, et malgré de premiers travaux prometteurs, l'essentiel reste à faire.

## Structure du mémoire

Nous présentons dans le chapitre 2 le contexte de nos travaux : les graphes de terrain dans les différents domaines concernés, leurs caractéristiques, leurs problématiques communes ainsi que la problématique qui nous intéresse particulièrement qui est leur dynamique. Nous introduisons ensuite la problématique concernée par notre travail, à savoir la détection d'événements dans la dynamique des graphes de terrain, et un état de l'art des travaux sur ce sujet.

Dans le chapitre 3 nous proposons et décrivons une approche générique pour automatiquement et rigoureusement détecter des événements dans la dynamique des graphes de terrain. Nous fournissons également des approches pour interpréter les événements détectés afin de mieux les comprendre, ainsi que leur impact sur les graphes de terrain.

Nous détaillons dans le chapitre 4 l'application de notre approche générique de détection d'événements dans les graphes de terrain au radar de l'internet, c'est-à-dire l'observation égo-centrée et periodique de la topologie de l'internet. Le but de ce chapitre est double : illustrer l'usage de notre méthodologie et en démontrer sa pertinence en pratique.

Enfin, le chapitre5 résume les contributions de cette thèse, analyse nos résultats et discute des directions futures possibles pour l'amélioration de notre travail.

Chapitre 2

# Contexte et état de l'art

## Sommaire

## 2.1 Introduction

De récentes avancées dans le domaine des graphes de terrain ont montré leur importance pour l'étude de nombreux phénomènes. Ces grands graphes permettent de modéliser les interactions entre les différents acteurs de phénomènes complexes, qui interviennent dans de très nombreux domaines : informatique, physique, sociologie, épidémiologie, biologie, linguistique, etc. Ces graphes possèdent des propriétés structurelles communes non triviales qui ont fait l'objet de nombreuses études [Wasserman 1994, Strogatz 2001, Albert 2002, Newman 2003, Dorogovtsev 2003]. Ces propriétés communes permettent d'envisager des problématiques nouvelles et transversales qui en découlent et ont des applications potentielles dans les nombreux domaines concernés.

Les travaux de cette thèse s'inscrivent dans ce contexte interdisciplinaire, en se concentrant sur la problématique de la dynamique des graphes de terrain, c'est-à-dire leur évolution au fil du temps par l'ajout et/ou le retrait de nœuds et/ou de liens. Nous abordons cette étude par une problématique transversale clé : la détection d'événement.

Dans ce chapitre nous présentons le contexte de nos travaux qui sont les graphes de terrain dans les différents domaines concernés (section 2.2) et leurs caractéristiques et problématiques communes. Ensuite nous passons en revue une des problématiques qui nous intéresse particulièrement qui est leur dynamique (section 2.3). Nous introduisons ensuite la problématique concernée par notre travail, à savoir la détection d'événements dans la dynamique des graphes de terrain, et un état de l'art des travaux sur ce sujet à la fin de ce chapitre (section 2.4).

## 2.2 Les graphes de terrain

Nous allons maintenant présenter quelques exemples clés qui illustrent la diversité des domaines d'applications possibles. Pour chaque cas, nous identifierons les acteurs du phénomène, modélisés par les nœuds du graphe, et les interactions entre eux, modélisées par des liens entre les nœuds. L'ensemble de ces domaines fait l'objet d'une littérature très abondante, nous nous contenterons de donner des références vers certaines publications centrales. Une bibliographie plus complète peut être obtenue dans [Wasserman 1994, Strogatz 2001, Albert 2002, Newman 2003, Dorogovtsev 2003, Barrat 2008, Cohen 2010].

– Les réseaux sociaux [Wasserman 1994] constituent un champ d'application ancien et important dans lequel les nœuds sont des individus ou entités sociales (associations, entreprises, pays). Les liens entre eux peuvent être de différentes natures, conduisant ainsi à observer plusieurs types de réseaux : les réseaux de connaissance (deux individus sont reliés s'ils se connaissent), les réseaux de

contact physique (deux individus sont reliés s'ils ont été physiquement en contact), les réseaux de collaboration (deux individus sont reliés s'ils ont travaillé ensemble, en particulier de nombreux travaux ont étudié les collaborations scientifiques [Palla 2005]), les réseaux d'appels téléphoniques [Resende 2000] (deux individus ou numéros de téléphone sont reliés s'il y a eu un appel entre eux), les réseaux d'échanges (deux entités sont reliées si elles ont échangé un fichier [Guillaume 2004] ou un courrier électronique [Ebel 2002] par exemple), etc.

– Les réseaux biologiques sont également de nature différente. On peut citer par exemple les réseaux métaboliques [Jeong 2001] (les nœuds sont des gènes ou des protéines qui sont liées par leurs interactions chimiques), les réseaux de neurones (chaque neurone est connecté à plusieurs autres neurones) ou les réseaux trophiques [Martinez 2003] (les espèces d'un écosystème sont reliées pour représenter les chaînes alimentaires).

– Les réseaux d'infrastructure représentent des connexions matérielles entre objets distribués dans un espace géographique. Nous pouvons citer les réseaux de transport (routes entre villes ou liaisons aériennes entre aéroports), les réseaux de distribution électrique (câbles entre les lieux de production et de consommation) ou encore le réseau physique de l'internet (câbles entre routeurs).

– Les réseaux d'information représentent des liens abstraits de référencement ou de similarité entre des supports d'information. Parmi eux, les exemples typiques sont les réseaux de citation d'articles ou les graphes du Web [Salah-Ibrahim 2010] (les nœuds sont des pages Web liées par des liens hypertextes).

– Les réseaux linguistiques relient les mots d'un langage donné et regroupent entre autres les réseaux des synonymies (deux mots sont reliés s'ils sont synonymes), les réseaux de cooccurrences (deux mots sont reliés s'ils apparaissent dans une même phrase d'un ouvrage) ou encore les réseaux de dictionnaires (deux mots sont liés si l'un est utilisé dans la définition de l'autre).

Les problématiques d'étude peuvent être très variées selon les disciplines. Par exemple, on étudie la propagation des épidémies [Satorras 2001] grâce à un réseau social de contact physique modélisant les possibilités de contamination. Un opérateur de télécommunications peut vouloir établir les profils de ses clients en analysant le réseau des appels téléphoniques [Tam 2009]. Les réseaux métaboliques sont utilisés pour comprendre le fonctionnement de la cellule [Jeong 2001]. L'analyse des réseaux d'infrastructure permet de détecter leurs points faibles susceptibles d'être la cible d'attaques ou de pannes ayant des conséquences majeures [Albert 2000, Callaway 2000].

### 2.2.1 Des caractéristiques communes

Les grands graphes de terrain que l'on peut rencontrer dans les différentes disciplines que nous venons de citer n'ont, à première vue, pas de raison de se ressembler. Cependant, plusieurs études ont révélé l'existence de caractéristiques structurelles communes et non triviales [Watts 1998, Strogatz 2001, Albert 2002, Newman 2003, Dorogovtsev 2003, Barrat 2008, Cohen 2010]. Nous présentons dans cette section un résumé de ces caractéristiques communes.

Tout d'abord du point de vue de la taille, le terme *grand* graphe de terrain fait référence au nombre de nœuds qui peut aller de quelques centaines à plusieurs milliards (par exemple pour les graphes du Web ou la structure neuronale du cerveau). Cependant chaque nœud est lié à relativement peu d'autres nœuds : une des caractéristiques des grands graphes de terrain est de posséder un faible degré moyen $d_{moy}$ (nombre moyen de voisins d'un nœud dans le graphe) : $d_{moy} = 2m/n$, où $m$ et $n$ sont le nombre de liens et de nœuds respectivement. Cette faible moyenne peut correspondre à des limites individuelles de chaque acteur du réseau. De nombreuses études et modèles visent à considérer que les grands graphes de terrain possèdent un degré moyen borné (indépendamment de la taille du graphe). Ceci implique en particulier que le nombre de liens d'un grand graphe de terrain peut être considéré comme linéaire par rapport au nombre de nœuds du graphe ($m = O(n)$) ; on parle alors de graphes creux ou peu denses.

Bien que le degré moyen soit relativement faible par rapport à la taille du graphe, la distribution des degrés est en général très hétérogène : il existe dans les graphes de terrain un nombre non négligeable de nœuds possédant un très fort degré par rapport à une majorité de nœuds possédant un faible degré. Cette distribution est souvent bien approximée par une loi de puissance pour laquelle le nombre $P_k$ de nœuds de degré $k$ est proportionnel à $k^{-\alpha}$ , avec des puissances $\alpha$ estimées entre 2 et 3 pour la plupart des graphes de terrain. Ceci a engendré l'appellation *réseau sans-échelle* (ou *scale-free networks*), issue de la physique, pour certains grands graphes de terrain. Cette caractéristique implique qu'il existe des rares (mais non exceptionnels) nœuds de très fort degré qui jouent nécessairement des rôles particuliers par rapport aux nœuds de faibles degrés. Cette propriété est l'une des premières propriétés surprenantes qui rapprochent les différents des grands graphes de terrain.

La propriété de degré moyen petit implique que la densité des grands graphes de terrain est faible. Cette densité $\sigma$ est la proportion de liens existantes par rapport au nombre possible $\theta = 2m$ ($\sigma = n(n-1)$). Si, comme nous l'avons vu, le nombre de liens croît linéairement par rapport au nombre de nœuds, comme le nombre de liens possibles croît de manière quadratique, ceci implique que la densité tend à se rapprocher de 0 lorsque la taille du graphe croît.

Cette forte densité traduit le fait que les liens entre les nœuds qui sont proches (possédant des voisins communs par exemple) sont beaucoup plus probables que les liens entre nœuds éloignés. Cela s'explique par la tendance des nœuds à se regrouper en modules ou communautés. Cette densité est souvent capturée par le coefficient de *clustering* qui compte la probabilité que deux voisins d'un même nœud soient eux-mêmes liés par un lien. La différence entre forte densité locale et faible densité globale fonde la problématique de détection de communautés.

Une dernière caractéristique attribuée aux grands graphes de terrain est celle de petits-mondes (en référence à l'effet *small-word* de l'expérience de Milgram [Travers 1969]). Celle-ci traduit le fait qu'il existe des très courts chemins pour relier chaque paire de nœuds (un chemin étant une succession de nœuds reliés par des liens).

Cette dernière propriété, bien qu'intuitivement surprenante, peut en réalité s'expliquer de manière probabiliste. En effet elle est vérifiée par le modèle de graphes aléatoires d'Erdős-Rényi [Erdös 1959], qui construit un graphe en plaçant aléatoirement des liens entre les paires de nœuds, la probabilité (identique pour chaque paire de nœuds) fixant la densité du graphe aléatoire construit. Au-delà, tout graphe auquel est ajouté une faible quantité de liens aléatoires voit sa distance moyenne devenir très faible [Travers 2000, Travers 1969]. La propriété *petits-mondes* peut donc être considérée comme une propriété mécanique découlant de la présence d'aléatoire plutôt qu'une propriété caractéristique des graphes de terrain. [1]

Les trois premières propriétés (faible densité globale, distribution des degrés en loi de puissance, forte densité locale) sont au contraire des propriétés caractéristiques que l'on ne retrouve pas dans les graphes aléatoires. C'est en ce sens que nous les considérons comme non-triviales.

### 2.2.2 Des problématiques communes

Outre le fait que la plupart des grands graphes de terrain ont des propriétés non triviales en commun, il est apparu dans cette dernière décennie, ce qui est peut-être plus important encore, que de très nombreuses questions qui se posent sur ces divers graphes sont en fait très générales et transversales. Ceci est bien sûr un point clé pour le domaine, puisqu'il permet de travailler sur des problématiques indépendantes d'un cas d'étude particulier, et de produire des résultats applicables à de très nombreux cas. Il permet également de faire ressortir des questions fondamentales, d'une portée théorique et/ou scientifique forte, qu'il aurait été difficile d'identifier sans passer par

1. Une autre vision du phénomène *petit-monde* plus algorithmique, dans laquelle ce n'est pas tant l'existence de chemins courts qui est étudiée que notre capacité à les trouver avec une information locale seulement a amené à de nombreux éclairages importants [Kleinberg 2000, Kleinberg 2006, Fraigniaud 2007].

une phase d'étude de cas concrets divers.

Aujourd'hui, ces problématiques se répartissent naturellement en quatre grandes familles qui semblent capturer de façon pertinente les principales interrogations du domaine et résister au passage du temps [Latapy 2007].

**Mesure et métrologie**. La plupart des grands graphes de terrain n'étant pas directement disponibles, la connaissance qu'on en a passe par une opération de mesure. Celle-ci peut s'avérer extrêmement complexe, et la mettre en œuvre est alors un défi en soi. De plus, elle fournit généralement une vision partielle et biaisée de l'objet réel ; il est alors nécessaire d'étudier ce biais, de tenter de le corriger, et de mener une réflexion plus approfondie sur les conclusions que l'on peut effectivement tirer de nos observations.

**Analyse**. Étant donné un grand graphe de terrain, une première étape naturelle est de tenter d'en décrire la forme, c'est-à-dire les propriétés principales, les caractéristiques. Ceci se fait par le biais de notions statistiques et/ou structurelles, visant à synthétiser de façon pertinente les principales caractéristiques du graphe. La définition de telles propriétés est toutefois loin d'être triviale, ainsi que l'évaluation de leur pertinence. De même, l'interprétation des descriptions obtenues peut s'avérer délicate.

**Modélisation**. Afin d'expliquer la nature des observations, de pouvoir développer des résultats mathématiquement rigoureux et de mener les simulations adéquates, il est important de capturer les propriétés observées en pratique dans des modèles de graphes de terrain. Ceci se fait généralement par le tirage aléatoire de graphes dans certaines classes ou par un processus explicite de construction de graphes. On obtient ainsi des graphes artificiels, représentatifs des propriétés choisies.

**Algorithmique**. Enfin, l'étude de très grands graphes interpelle naturellement l'algorithmique, et ceci à deux titres. Tout d'abord, le contexte des graphes de terrain soulève des questions algorithmiques (comme la détection de communautés) originales, c'est-à-dire qui ne se posaient pas précédemment. De plus, les solutions usuelles à des problèmes algorithmiques classiques (comme le calcul du diamètre) ne sont plus applicables du fait de la taille des graphes considérés. Par contre, les propriétés rencontrées en pratique peuvent être mises à profit pour concevoir des algorithmes efficaces sur les graphes de terrain ; ceci ouvre la voie à tout un nouveau pan de l'algorithmique de graphes.

Ces quatre grandes familles de problématiques ont un complément naturel : un axe transversal centré sur l'étude des phénomènes ayant lieu sur ces réseaux, comme la diffusion d'information ou de virus, la résistance aux pannes ou aux attaques, les

comportements sociaux ou biologiques, le routage et les congestions, etc. Cet axe soulève lui-même des questions de mesure et métrologie, d'analyse, de modélisation et d'algorithmique, et joue naturellement un rôle important dans le domaine (notamment parce qu'il motive l'étude des topologies et montre que les propriétés observées ont effectivement un impact fort sur les phénomènes qui nous intéressent).

## 2.3 La dynamique des graphes de terrain

Une caractéristique importante de nombreux graphes de terrain est leur dynamique : des nœuds et/ou des liens arrivent et partent au cours du temps ; les gens changent leurs relations, de nouvelles interactions moléculaires se créent, de nouvelles machines sont ajoutées aux réseaux de l'internet, des liens de communication échouent, et ainsi de suite. Bien que des contributions au cas par cas aient développées (notamment dans le contexte des réseaux mobiles [Fleury 2007], et du pair-à-pair [Stutzbach 2005]), il existe aujourd'hui peu de résultats généraux sur la dynamique des graphes de terrain. De plus, beaucoup d'études se limitent à des suites de visions instantanées du réseau, avec un grain de temps bien supérieur à la dynamique de ces objets.

L'une des approches développées pour étudier les graphes dynamiques est la prédiction des liens [Liben-Nowell 2003, Ceyhan 2011], qui consiste à prédire le lien qui va probablement apparaître à l'avenir, étant donné un *snapshot* de graphe considéré à un instant donné. Plusieurs travaux l'étudient en se basant sur les mesures de similarité entre les nœuds. Nous pouvons citer par exemple le travail [Liben-Nowell 2003] les auteurs examinent plusieurs mesures topologiques (comme le coefficient de Jaccard, le coefficient adamique/Adar et le *SimRank*) basées sur les nœuds de voisinage et l'ensemble de tous les chemins entre eux. Dans [Huang 2005] les auteurs proposent d'utiliser une autre mesure topologique qui est le coefficient de *clustering*. Les auteurs dans [Hasan 2006, O'Madadhain 2005] ajoutent plusieurs mesures non topologiques basées sur des attributs de nœuds (comme la correspondance des mots clés et la proximité géographique) et ils utilisent un algorithme d'apprentissage supervisé pour effectuer la prévision de lien. Les auteurs dans [Clauset 2008] utilisent une décomposition hiérarchique d'un réseau social pour prédire les liens manquants. Enfin, dans [Allali 2011] les auteurs définissent la notion de liens internes dans les graphes bipartis et proposent une méthode de prédiction basée sur ces liens.

L'étude de l'évolution des communautés est une autre problématique qui s'intéresse aux graphes dynamiques. Elle consiste à trouver un partition optimale des nœuds du graphe qui peut être évaluée par une fonction de qualité telle que la modularité. Plusieurs études ont été réalisées pour intégrer la dynamique dans la détection

de communautés. On peut par exemple chercher à détecter une décomposition différente à plusieurs instants et essayer de suivre les communautés entre ces multiples décompositions [Hopcroft 2004, Palla 2007]. L'instabilité des algorithmes de détection de communautés, qui ont tendance à donner des résultats très différents entre deux instants proches, a conduit à proposer d'autres fonctions de qualité que la modularité. Ainsi, dans [Kumar 2006] les auteurs proposent d'ajouter à la fonction de qualité un terme de stabilité, représentant la proximité entre la partition à $t$ et celle à $t-1$ et dans [Song 2007] il est proposé d'ajouter un terme imposant que la partition à $t$ soit également bonne à $t-1$. Enfin les auteurs dans [Aynaud 2010] proposent une méthode pour trouver une partition unique de qualité dans les graphes dynamiques couvrant une longue période, cette décomposition peut être trouvée efficacement via une adaptation de la méthode de Louvain [Blondel 2008].

Il existe deux obstacles majeurs pour aller plus loin dans l'étude des graphes de terrain dynamiques. Tout d'abord, et malgré certains efforts, il y a aujourd'hui un manque crucial de données de qualité (c'est-à-dire de grande taille, recouvrant des périodes de temps significatives, avec une précision temporelle fine, et non biaisées). Bien sûr, mener de telles mesures, en analyser le biais et le corriger est encore plus complexe que dans le cas statique. L'autre point bloquant est le manque de formalismes et de notions adaptés au contexte dynamique. Exactement de la même façon qu'une intense activité d'analyse sur ces dix dernières années a mené à un ensemble aujourd'hui relativement large et cohérent de notions pour le cas statique, il s'agit maintenant de développer l'équivalent dans le cas dynamique. Bien sûr, la tendance naturelle est de se contenter d'étendre les notions statiques, mais ceci n'est pas pleinement satisfaisant : des nouvelles notions, qui ne prennent sens que dans le contexte dynamique, doivent être développées.

## 2.4 La détection d'événements

La problématique de la détection d'événements est loin d'être nouvelle. Il s'agit d'une problématique classique présente dans une grande variété de contextes tels que : la détection d'anomalie dans le trafic internet [Brutlag 2000, Mark 2004, Roughan 2004, Lakhina 2005], la détection d'intrusion dans la cybersécurité [Hofmeyr 1998, Lane 1998, Phoha 2002], la détection de dysfonctionnements industriels [Keogh 2002, Basu 2007], la détection d'anomalie dans les images [Torr 1995, Diehl 2002, Pokrajac 2007], la détection de défauts dans les systèmes critiques pour la sécurité et la surveillance militaire [Brotherton 2001, Mackey 2001], etc.

La détection d'événements est la capacité à mesurer et/ou à isoler des modifications particulières dans les systèmes qui ne sont pas conformes au *comportement attendu*. Cette fonction est très importante dans la surveillance des performances, car elle peut

fournir l'avertissement de conditions de défaut possible, ou du moins aider à identifier les causes et les lieux des problèmes connus. Les modifications particulières qui ne sont pas conformes au *comportement attendu* sont souvent appelées des anomalies, des événements, des *outliers* (valeurs aberrantes), des exceptions, des observations discordantes, etc. Cela dépend de la perception qu'on donne à ces modifications particulières, aussi et plus fortement, et dans la plupart des cas, du contexte visé. Nous allons dans toute la suite de ce mémoire utiliser le terme *événement* pour désigner ces modifications particulières dans les systèmes qui ne sont pas conformes au *comportement attendu*.

L'importance de la détection d'événements est due au fait que les événements traduisent d'importantes (et souvent critiques) informations exploitables dans un large éventail de domaines d'études. Par exemple, une anomalie dans les données de transaction de carte de crédit pourrait indiquer un vol de carte de crédit ou une usurpation d'identité [Kumar 2005]. Un délai de réception anormal de paquets sur internet pourrait signifier des problèmes de congestion ou de routage [Lakhina 2005]. Des lectures anormales d'un capteur de vaisseau spatial pourraient signifier un défaut dans certains composants de l'engin spatial [Fujimaki 2005]. Une image IRM anormale peut indiquer la présence de tumeurs malignes [Lucas 2001], etc.

La détection d'anomalie a fait l'objet de nombreuses études. Une recherche importante sur la détection d'*outliers* dans les statistiques a été traitée dans plusieurs études [Hawkins 1980, Rousseeuw 1987, Barnett 1994, Bakar 2006]. Les auteurs de [Hodge 2004] fournissent une vaste études sur les approches de détection d'anomalie utilisant des techniques d'apprentissage et des outils statistiques. L'étude [Hofmeyr 1998] présente les approches de détection d'événements utilisées spécifiquement dans le domaine de la cyber-intrusion. Enfin un état de l'art complet de techniques et d'approches de détection d'événements a été effectué et présenté dans [Chandola 2009].

D'une manière générale, les approches de détection d'événements peuvent être classées en deux grandes familles, l'approche basée anomalie et l'approche basée signature. Dans cette section nous discutons les deux familles, afin de situer le contexte nécessaire pour réaliser la détection des événements dans les graphes de terrain.

### 2.4.1 Approche basée anomalie

Une approche directe de détection d'événements appelée approche basée anomalie [Salem 2010] consiste à définir un ensemble de comportements qui représente les comportements *normaux* du système et déclarer toute observation qui ne ressemble pas au comportement normal comme un événement.

Plusieurs facteurs rendent cette approche d'apparence simple très difficile. Le premier facteur est celui lié à la caractérisation ou la définition du comportement normal. Cette tâche consiste en la délimitation d'une région qui énumère tous les comportements *normaux* possibles, chose qui est très difficile [Jensen 2006] en pratique. En outre, la limite entre un comportement normal et anormal n'est souvent pas précise [Schölkopf 1999]. Ainsi, une observation qui à l'allure d'un événement et qui se trouve à proximité (dans le temps ou dans l'espace) de la limite d'un comportement normal peut effectivement être normal, et vice-versa. Un deuxième facteur concerne l'imitation de l'événement, en d'autres termes, lorsque des événements sont le résultat d'actions malveillantes, les adversaires malveillants adaptent souvent leurs actions pour qu'elles aient une apparence normale, ce qui rend la tâche de définir un comportement normal plus difficile. En conclusion, cette approche est très attrayante, car elle est capable de détecter tout type d'événement, y compris les types qui n'ont jamais été observés. Elle s'appuie cependant sur une connaissance précise de la dynamique de l'objet considéré rarement possible en pratique.

### 2.4.2 Approche basée signature

L'approche basée signature [Li 2005, Paxson 1999] se base, quant à elle, sur la connaissance des caractéristiques des événements à détecter, qui peut être déduite d'un ensemble d'événements connus dont on dérive une *signature* [Hopcroft 2006], obtenue généralement avec des techniques d'apprentissage statistique [Liao 2005]. Si la dynamique observée correspond à ces signatures à un instant donné de la dynamique, alors on considère cette dernière comme étant un événement.

Cette approche est très efficace dans les cas où les événements dans leur globalité peuvent être décrits, comme certains virus informatiques par exemple [Webster 2008]. Mais, se baser sur une définition figée des mêmes événements peut ne pas s'avérer efficace, car la notion exacte d'un même événement peut être différente dans plusieurs contextes. Par exemple, dans le domaine médical un petit écart à la normale (par exemple, les fluctuations de la température corporelle) pourrait être une anomalie, alors que des écarts similaires dans le domaine des marchés financiers (par exemple, les fluctuations de la valeur d'une action) peuvent être considérées comme normaux. Pour remédier à ce genre de problèmes, une mise à jour intense des bases de signatures est fortement nécessaire [Shafi 2009] afin d'atteindre une efficacité suffisante. De plus, cette méthode ne permet pas de détecter des événements de forme nouvelle et/ou inconnue. Elle suppose donc encore une relativement fine connaissance du système observé.

## 2.5 Conclusion

Dans ce chapitre, nous avons présenté le contexte de nos travaux qui sont les graphes de terrain, ainsi que la problématique traitée, à savoir la détection des événements dans leur dynamique. De nombreuses études sur la détection d'événements ciblent la dynamique de différents systèmes. Deux principales approches sont utilisées, l'approche basée anomalie et l'approche basée signature.

Nous allons voir dans le chapitre suivant les différents facteurs qui ont motivé la proposition d'une nouvelle approche de détection d'événements dans la dynamique des graphes de terrain. Nous allons aussi détailler le principe de notre approche ainsi que la méthodologie et les outils utilisés pour l'appliquer.

Chapitre 3

# Notre approche

## Sommaire

## 3.1 Introduction

Le but de nos travaux est de proposer une approche générique pour la détection d'événements dans la dynamique des graphes de terrain. Nous avons souligné dans le chapitre 1 que les graphes de terrain sont dynamiques et que les efforts dédiés à l'étude de ces dynamiques n'ont pas fourni jusqu'alors un formalisme complet adapté à l'étude des graphes dynamiques. Ce manque rend notre tache de détection d'événements plus ardue.

Aussi, l'état de l'art actuel des approches de détection d'événements, comme nous l'avons vu dans la section 2.4, reste très limité en ce qui concerne spécifiquement la dynamique des graphes de terrain. Rappelons que le principe sous-jacent des approches basées anomalies est que l'on connaît le comportement normal du système. Puis, toute observation qui diffère de ce comportement normal est considérée comme un événement. Cette approche s'appuie donc sur une connaissance précise de la dynamique de l'objet considéré, et l'évolution du comportement normal rend la méthode inapplicable. Dans le cas des graphes de terrain, ces deux limitations rendent cette approche inadéquate. L'approche basée signature, quant à elle, repose sur la connaissance des caractéristiques des événements à détecter, qui peut être déduite d'un ensemble d'événements connus. Si la dynamique observée correspond à ces caractéristiques à un moment donné, alors on considère qu'il s'agit d'un événement. Cette approche est très efficace dans les cas où les événements peuvent être décrits. Cependant, dans le cas des graphes de terrain, la connaissance préalable d'événements à détecter dans la dynamique est en général inexistante. De plus, il n'y a pas de connaissance de leur impact sur la structure des graphes, ni même de notions pour décrire cet impact. Par conséquent, ne nous pouvons nous appuyer sur des définitions d'événements connus pour rechercher leur signature dans la dynamique observée.

Pour ces raisons, nous proposons une approche qui ne nécessite aucune connaissance préalable, ni de la dynamique du système considéré, ni des signatures des événements à détecter. Ceci donne une nature profondément générique et robuste à notre approche. En contrepartie, son principal inconvénient est que les événements détectés peuvent être difficiles à interpréter. Nous expliquons dans ce chapitre le principe de notre approche ainsi que la méthodologie permettant son application en pratique aux graphes de terrain.

## 3.2 Événement statistiquement significatif

Notre approche repose sur l'intuition que les graphes de terrain peuvent être soumis à une dynamique *de fond*, qui peut être vue comme régulière et permanente. Cependant, cette dynamique subit de façon exceptionnelle des changements profonds et

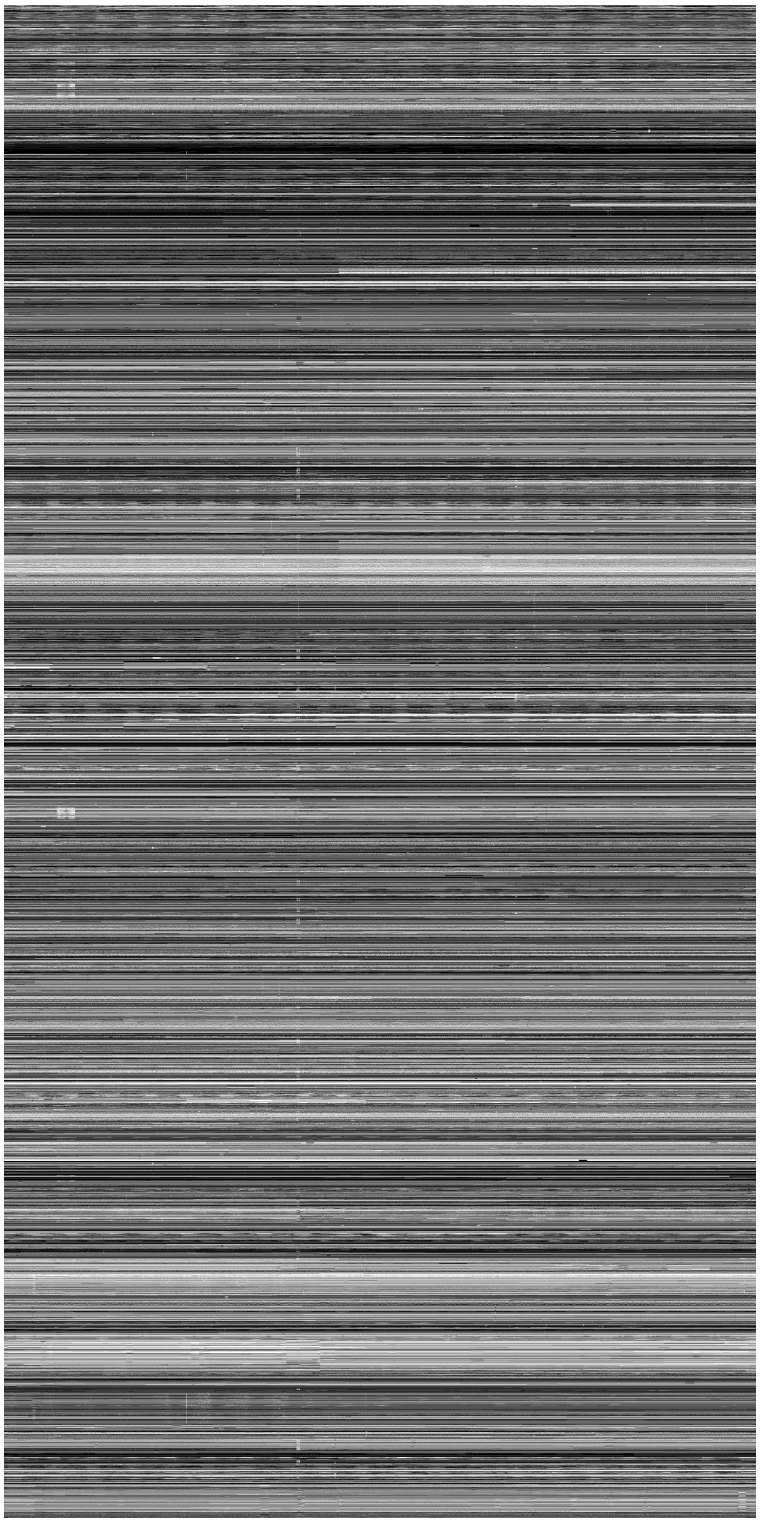


FIGURE 3.1 – Visualisation en *code barres* de 2000 visions instantanées d'un graphe dynamique de 2000 nœuds : chaque ligne représente la dynamique d'un nœud du graphe, et chaque colonne représente une vision instantanée. La présence d'un point blanc aux coordonnées $(i, j)$ signifie que le nœud $i$ est présent à la vision instantanée $j$ ; sinon le point est noir.

inhabituels, que nous appelons des événements. Par exemple, pour le routage dans l'internet, l'équilibrage de charge [1] peut être considéré comme une dynamique de fond, c'est-à-dire une dynamique normale. Seulement, certaines pannes ou congestions à certains endroits du réseau, créant un nombre important de changements de routage, peuvent être considérées comme étant des événements. Ainsi, nous formulons le principe de cette intuition comme suit : certaines propriétés de la dynamique de graphes de terrain ont une évolution *normale* la plupart du temps, néanmoins, elles s'écartent parfois *significativement* de cette évolution, révélant ainsi des événements.

Cette intuition est illustrée dans la Figure 3.1 sur le cas que nous allons détaillé dans le prochain chapitre. Cette visualisation permet d'observer des apparitions et/ou des disparitions de nœuds (ou idéalement de blocs de nœuds, qui sont plus facilement repérables que des nœuds isolés). Comme on peut le remarquer un effet *code barres* c'est-à-dire des alternances de lignes blanches et noires, indiquant des nœuds (presque toujours présents (lignes blanches) et d'autres (presque) toujours absents (lignes noires). On distingue également des lignes en pointillés, indiquant des nœuds alternativement présents et absents, mais régulièrement, traduisant ce que nous appelons la dynamique de fond. Cependant on distingue aussi des dynamiques plus anormales, typiques de ce que nous appelons des *événements*.

Pour caractériser ces événements, notre approche s'appuie sur l'étude des distributions statistiques (c'est-à-dire le nombre d'occurrences de chaque valeur possible) des propriétés de graphes dynamiques (comme le nombre de nœuds dans le graphe à chaque instant), et l'identification d'écarts *statistiquement significatifs* dans l'évolution de ces propriétés.

Quand on considère une distribution statistique d'un ensemble de valeurs numériques associées à une propriété dynamique, trois situations typiques peuvent se produire :

- les valeurs observées peuvent être homogènes, ce qui signifie qu'elles sont toutes semblables, ont une valeur moyenne bien identifiée, et que l'on n'observe jamais d'écarts importants à cette moyenne ;
- les valeurs observées peuvent être de nature hétérogène, ce qui signifie qu'il n'y a pas de notion de valeur *normale*, et ce qui caractérise cette distribution est plutôt cette hétérogénéité ;
- les valeurs observées peuvent être homogènes avec quelques *outliers* (valeurs statistiquement aberrantes), c'est-à-dire que la plupart des valeurs ont un caractère homogène mais certaines s'écartent considérablement d'elles.

Dans les deux premières situations, la propriété considérée n'a pas d'utilité pour la détection des événements : soit toutes les valeurs de sa distribution sont normales

1. L'équilibrage de charge améliore les performances du réseau en distribuant le trafic de manière (en principe) équilibrée sur les différentes routes disponibles.

et il n'y a pas de notion d'événements (distributions homogènes), soit il n'y a pas de notion de comportement normal et, par conséquent il n'y a pas d'événement (distributions hétérogènes). Contrairement aux deux cas précédents, dans le troisième cas, la propriété peut être utilisée pour la détection d'événements *statistiquement significatifs*. Les *outliers* détectés indiquent des événements, alors que la plupart des autres valeurs de la propriété sont proches d'une valeur *normale* et correspondent donc à un comportement normal.

Ainsi, caractériser un événement dans la dynamique des graphes de terrain, reviendra pour nous à déterminer des propriétés de dynamiques de graphes dont la distribution est normale avec *outliers*. Les valeurs *suffisamment* loin des valeurs *normales* de la distribution seront celles qui identifient les événements.

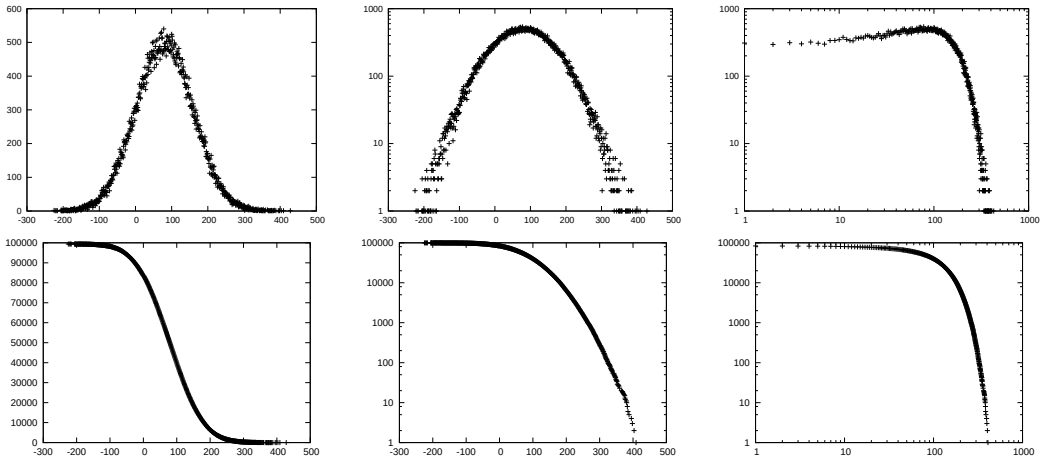


FIGURE 3.2 – Cas typique d'une distribution homogène. Première ligne (de gauche à droite) : la distribution dans les échelles lin-lin, lin-log et log-log. Deuxième ligne : la cumulative inverse de la distribution dans les mêmes échelles.

Les Figures 3.2 à 3.4 illustrent les différentes cas typiques et leur inspection visuelle. Nous considérons trois distributions (une figure pour chaque cas) et nous les traçons dans les trois échelles lin-lin, lin-log et log-log (première ligne, de gauche à droite). Nous avons aussi tracé la distribution cumulative inverse dans les trois échelles (deuxième ligne, de gauche à droite), conduisant à six courbes pour chaque distribution. La figure 3.2 illustre le cas homogène : il existe une valeur *normale*, et aucune autre valeur observée ne s'écarte beaucoup de cette dernière. Cette distribution est capturée par la pente exponentielle de la distribution, révélée par sa forme droite dans les échelles lin-log. La figure 3.3 illustre le cas hétérogène : il n'existe pas de notion de valeur *normale* ; plutôt une distribution caractérisée par le fait que les valeurs observées couvrent une large plage, avec une décroissance polynomiale, révélée par sa forme droite en échelles log-log. Enfin, la figure 3.4 illustre le cas homogène avec des *outliers* :

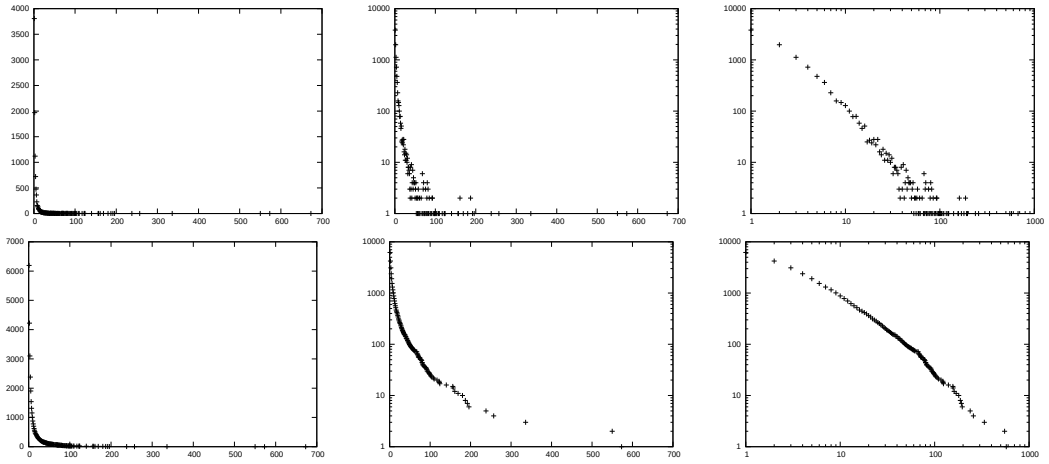


FIGURE 3.3 – Cas typique d'une distribution hétérogène. Première ligne (de gauche à droite) : la distribution dans les échelles lin-lin, lin-log et log-log. Deuxième ligne : la cumulative inverse de la distribution dans les mêmes échelles.

il existe une valeur *normale*, mais certaines valeurs s'écartent de manière significative de cette dernière. Ces valeurs indiquent des événements statistiquement significatifs. Les distributions correspondantes présentent donc deux régimes : une décroissance exponentielle (révélée par une forme droite à l'échelle lin-log) et certaines valeurs qui s'écartent de façon significative.

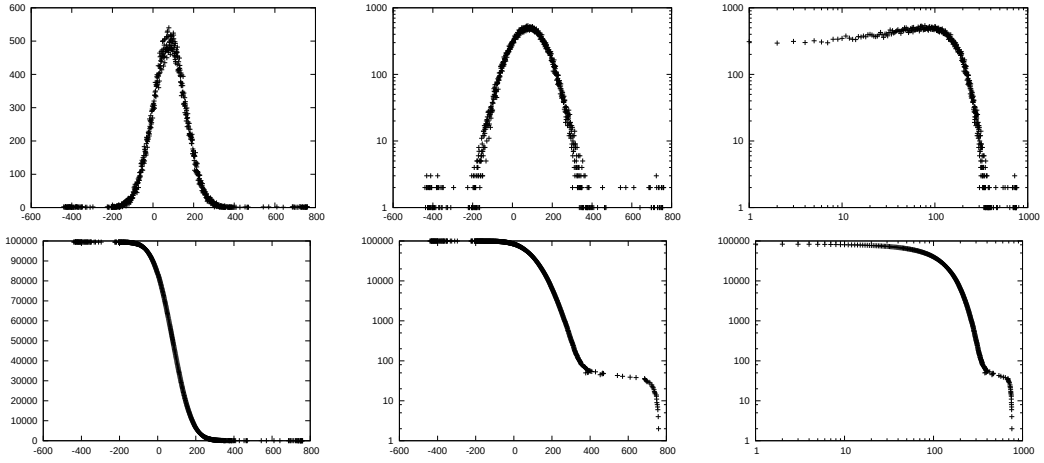


FIGURE 3.4 – Cas typique d'une distribution homogène avec *outliers*. Première ligne (de gauche à droite) : la distribution dans les échelles lin-lin, lin-log et log-log. Deuxième ligne : la cumulative inverse de la distribution dans les mêmes échelles.

## 3.3 Méthodologie

Après avoir expliqué le principe utilisé pour caractériser les événements, nous pouvons détailler la méthodologie par laquelle nous procédons et son application à la détection d'événements dans les graphes de terrain. Elle se décompose en quatre grandes étapes :

1. définir les propriétés décrivant la dynamique du graphe de terrain considéré. Un exemple de propriété est le nombre de nœuds du graphe à chaque instant ;
2. calculer les valeurs des propriétés définies dans l'étape précédente à partir du graphe de terrain considéré ;
3. étudier les distributions des valeurs de ces propriétés, et décider de leur nature : homogène, hétérogène, ou homogène avec *outliers* ;
4. sélectionner les propriétés homogènes avec *outliers* et considérer ces valeurs comme étant des événements dans la dynamique du graphe.

Pour pouvoir appliquer notre méthodologie, nous devons donc répondre à un ensemble de questions clés. La première concerne la définition de propriétés pertinentes, générales et susceptibles de capturer la dynamique des graphes considérés, et répondre à notre principe de caractérisation d'événements (c'est-à-dire ayant des valeurs homogènes avec *outliers*). La deuxième question concerne le calcul des valeurs des propriétés dynamiques. À ce sujet, des questions délicates se posent : quelles sont les propriétés qui peuvent être mesurées (donc observées) en pratique ? quel(s) niveau(x) de granularité ou quelles résolutions de l'observation sont pertinents, que ce soit dans le temps ou l'espace ? Nous expliquons nos réponses à ces questions dans la section 3.4.

L'autre question difficile se rapporte à l'étude des distributions empiriques et la décision quant à leur nature. En principe, cela devrait être possible en utilisant des méthodes classiques d'ajustement (*fit*) et des tests d'hypothèses appropriés. En pratique, cependant, les distributions empiriques sont rarement très proches des modèles, et les méthodes automatiques peuvent être trompeuses [Wasserman 2005, Chen 2002]. Nous répondons à cette problématique dans la section 3.5.

Avant de répondre aux différentes questions posées par notre méthodologie, nous introduisons quelques notations utiles dans la suite. Nous considérons une série de graphes indexée par un entier $i$. Intuitivement, cette série, que nous notons $G_i$, correspond à une suite de visions instantanée du graphe dynamique considéré, un peu comme un film est une suite d'images fixes. On note $G_i = (V_i, E_i)$ où $V_i$ est un ensemble de nœuds et $E_i \subseteq V_i \times V_i$ est un ensemble des liens. Nous notons par $N_i = |V_i|$ le nombre de nœuds du graphe $G_i$. Dans la Figure 3.5, par exemple, $N_{t-2} = 6$, $N_{t-1} = 7$..., et $N_{t+2} = 8$.

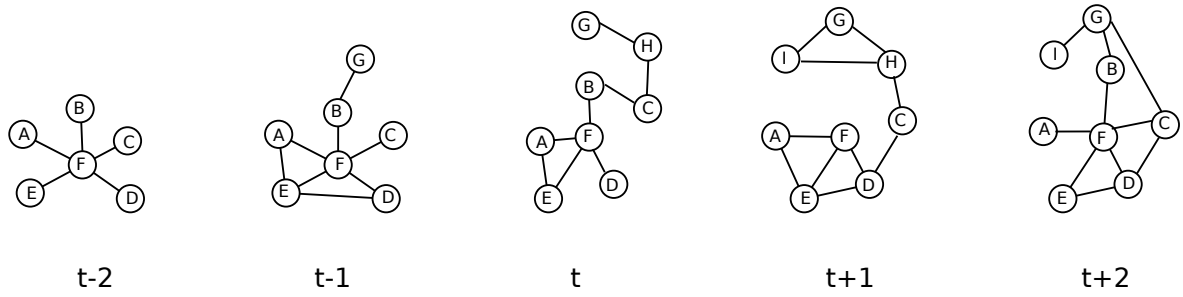



FIGURE 3.5 – Exemple d'une série de visions instantanées (à cinq instants successifs) d'un graphe dynamique

Étant donné deux entiers $i$ et $j$, nous notons $G_i^j = (V_i^j, E_i^j)$ le graphe obtenu par l'union des graphes à partir du $i$-ème jusqu'au $i+j-1$-ème (c'est-à-dire $j$ graphes à partir du $i$-ème) : $V_i^j = \cup_{k=i}^{k=i+j-1} V_k$ et $E_i^j = \cup_{k=i}^{k=i+j-1} E_k$.

Pour tout entier $i$, étant donnés deux entiers $p$ et $c$, nous appelons $G_i^c = (V_i^c, E_i^c)$ le graphe *courant* et $G_{i-p}^p = (V_i^{i-p}, E_i^{i-p})$ le graphe *précédent*. Le graphe courant est l'union de $c$ visions instantanées à partir de la $i$-ème, et le graphe précédent est l'union des $p$ visions instantanées précédant la $i$-ème. Nous utilisons ces notations dans les sections suivantes pour définir des propriétés de graphes dynamiques.

Dans l'exemple de la Figure 3.5, en considérant $i = t$ et en prenant $p = 2$ et $c = 3$, le graphe précédent et celui constitué de l'union des visions instantanées de l'instant $t-1$ à $t-2$ et le graphe courant est l'union des visions de l'instant $t$ à $t+2$. Voir la Figure 3.6.

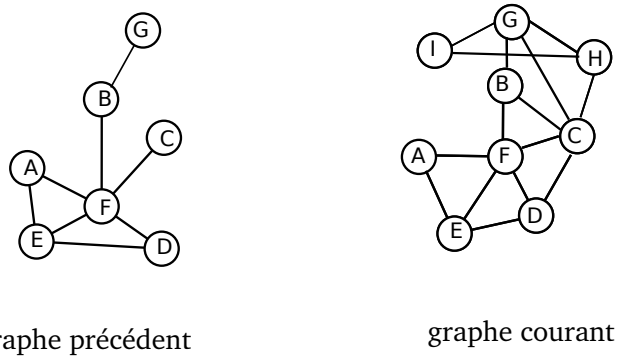



FIGURE 3.6 – Les graphes précédent et courant de la série de visions instantanées du graphe dynamique de la Figure 3.5, c'est-à-dire ceux constitué de l'union des visions instantanées de l'instant $t-1$ à $t-2$ et de l'instant $t$ à $t+2$ respectivement.

## 3.4 Propriétés dynamiques

Dans cette section nous proposons un ensemble de propriétés pour la description de la dynamique de graphes, conçues pour être susceptibles de révéler des événements tout en restant relativement simples. Nous classons ces propriétés en deux grandes familles, celles basées sur des nombres de nœuds et celles sur des distances.

### 3.4.1 Nombres de nœuds

La propriété la plus naturelle à observer est certainement le nombre de nœuds $N_i$ présents à chaque instant $i$, dans l'exemple de la Figure 3.5, comme déjà dit, on a $N_{t-2} = 6$, $N_{t-1} = 7$..., et $N_{t+2} = 8$. Cependant, la granularité et la résolution temporelle de l'observation du graphe impactent directement cette propriété. Pour explorer ceci, nous considérons également le nombre de nœuds distincts observés pendant un intervalle de temps continu, c'est-à-dire pas uniquement à un instant donnée. Dans l'exemple de la Figure 3.5 le nombre de nœuds dans le graphe courant (de l'instant $t$ à $t+2$) est de 9 par exemple. Cependant, la détection automatique en utilisant le nombre de nœuds dans un intervalle de temps continu n'est pas triviale : la valeur moyenne observée peut changer au cours du temps, et les pics vers le haut peuvent être inférieurs à ces variations de la moyenne. Faire la différence entre un événement statistiquement *significatif* et la dynamique normale peut être difficile. Afin de résoudre ce problème, nous introduisons la notion de nœuds qui apparaissent définie comme les nœuds observés à certain intervalle de temps mais pas dans l'intervalle de temps précédent. Dans l'exemple de la Figure 3.5 le nombre de nœuds qui apparaissent dans le graphe courant et qui n'étaient pas présent dans le graphe précédent est de 2 (les nœuds $I$ et $H$). Notons que l'observation du nombre de nœuds qui disparaissent est aussi naturelle que les liens qui apparaissent. Nous avons observé des résultats similaires pour toutes ces notions, et nous nous concentrons dans ce mémoire sur les apparitions des nœuds.

Ces propriétés sont bien des propriétés de graphes, mais elles ne capturent aucune information leur sa structure. Elles constituent donc une base de référence mais nous devons les compléter par des propriétés plus subtiles.

Une direction naturelle est de se demander si les nouveaux nœuds apparaissant sont dispersés dans le graphe, s'ils sont regroupés ou, si au contraire ils appartiennent à plusieurs petits groupes. Intuitivement, par exemple, un changement important dans le graphe peut mener à la découverte d'une nouvelle partie de celui-ci, ce qui serait révélé par l'apparition de nœuds formant une composante connexe. Afin d'approfondir ces questions nous étudions les composantes connexes de nœuds qui apparaissent. Plus précisément, pour tout instant $i$, nous sélectionnons les nœuds apparaissant comme définis ci-dessus, et nous considérons les liens entre ces nœuds. Nous calculons alors les composantes connexes de ces derniers, que nous appelons les composantes connexes

de nouveaux nœuds. Les nœuds $G$ et $I$ et le lien entre eux consituent une composante connexe des nœuds qui apparaissent dans l'exemple de la Figure 3.5.

Notons qu'on pourrait aller plus loin en calculant les diverses propriétés des composantes connexes (leur densité, degré moyen, ou coefficient de *clustering* par exemple), puis en observant leurs distributions. Cela pourrait conduire à l'identification d'événements statistiquement significatifs. Cependant explorer ceci est hors de la portée du présent travail et constitue une de nos perspectives.

### 3.4.2 Propriétés basées distance

Les propriétés considérés précédemment restent basiques, et ne capturent que pauvrement les caractéristiques structurelles de graphes (taille et connexité). Dans cette section, nous considérons des propriétés basées sur les distances [2] afin de tenter de capturer des caractéristiques plus subtiles des dynamiques.

De même que pour les nœuds qui apparaissent, considérés ci-dessus, nous définissons pour tout entier $i$ les *liens qui apparaissent* comme étant les liens observés dans l'intervalle de temps courant, c'est-à-dire l'union de $c$ visions instantanées à partir de la $i$-ème de $i$ jusqu'à $i + c - 1$, mais pas dans l'intervalle de temps précédent, c'est-à-dire l'union des $p$ visions instantanées de $i - p$ jusqu'à $i - 1$, ($c$ et $p$ étant deux entiers donnés). Ce sont donc des liens dans $E_i^c \setminus E_{i-p}^p$. Notons que tous les liens d'un nœud qui apparait sont nécessairement des liens qui apparaissent. Nous les appellerons des liens qui apparaissent *triviaux*, et nous nous concentrons sur les liens qui apparaissent *non triviaux*. Les liens qui apparaissent *non triviaux* sont donc les liens qui apparaissent au cours de l'intervalle de temps courant entre les nœuds qui étaient déjà présents, mais pas reliés entre eux dans l'intervalle de temps précédent : les liens qui apparaissent *non triviaux* sont ceux dans $(E_i^c \setminus E_{i-p}^p) \cap (V_{i-p}^p \times V_{i-p}^p)$ (un exemple de lien qui apparait *non trivial* dans la Figure 3.5 est le lien $(G, C)$, la distance entre ses deux extrémités dans le graphe précédent est 3).

On peut s'attendre à ce que les liens qui apparaissent *non triviaux* aient tendance à apparaître entre des nœuds qui étaient déjà très proches précédemment, c'est-à-dire ceux dont la distance entre eux dans $G_{i-p}^p$ est petite. Une grande distance pourrait révéler un événement. Notons qu'il existe une valeur de distance pour chaque lien de ce type, et nous obtenons donc plusieurs valeurs à chaque instant (une par lien non trivial qui apparait). Ceci signifie que nos statistiques seront plus fiables (car sur des échantillons plus grands) et surtout qu'on peut explorer et détecter *plusieurs* événements au même instant.

Un problème majeur avec ces statistiques est que toutes les distances dans les graphes considérés sont petites, et sont très similaires entre elles (la distribution de

2. La *distance* entre deux nœuds est la longueur d'un plus court chemin entre eux, c'est-à-dire le nombre de sauts nécessaires pour aller de l'un à l'autre dans le graphe.

toutes les distances dans le graphe est homogène). En conséquence, il y a peu d'espoir que les distributions des distances entre les nœuds, même pour des paires sélectionnées de nœuds, puissent présenter des distributions assez riches pour la détection d'événements. Afin d'obtenir de telles distributions, il faut considérer des propriétés avec des valeurs couvrant un large intervalle.

Afin de trouver une telle propriété, tout en s'appuyant sur les distances, nous définissons pour tout lien $(u, v)$ non trivial qui apparait, le nombre de nœuds dans $V^p_{i-p}$ tels que leur distance à $u$ ou $v$ n'est pas la même dans les graphes précédents et courants, $G^p_{i-p}$ et $G^c_i$ respectivement (le nombre de tels nœuds dans l'exemple de la Figure 3.5 pour le lien qui apparait *non trivial* $(G, C)$ correspondant au nœud $D$ : sa distance à $G$ dans le graphe précédent était de 3 et elle est de 2 dans le graphe courant). Nous notons ce nombre $\delta(u, v)$ et soulignons qu'il peut prendre des valeurs de 0 au nombre total de nœuds. Il est donc plus susceptible de permettre de détecter des événements que les distances elles-mêmes, à l'espace de valeurs très resérré. Comme pour la propriété précédente, pour chaque passe $i$, nous définissons $\Delta_i = \{\{\delta(u, v),\ (u, v)\ \textit{est un lien qui apparait}\}\}$ comme le multi-ensemble[3] de toutes les valeurs $\delta(u, v)$.

### 3.4.3 Autres propriétés

Remarquons que de très nombreuses autres propriétés peuvent être définies, en se basant sur le nombre de liens par exemple. Cependant, notre but n'est pas d'être exhaustif, ce qui serait impossible, mais de fournir un premier ensemble de propriétés permettant d'explorer l'approche proposée.

## 3.5 Analyse statistique

Une problématique clé pour notre méthode est l'étude des distributions empiriques et en particulier la décision de leur nature (homogène, hétérogène ou homogène avec *outliers*). Nous procédons comme suit : nous combinons une inspection visuelle des distributions dans différentes échelles et des techniques d'ajustement automatiques, les deux ayant leurs propres forces et limites, et étant complémentaires l'une de l'autre.

### 3.5.1 Inspection visuelle

Afin d'effectuer une inspection visuelle des distributions, nous les représentons graphiquement avec des courbes dans les échelles lin-lin, lin-log et log-log, comme illustré dans les Figures 3.2 à 3.4. Ceci permet d'observer les distributions directement

3. Un multi-ensemble, que nous notons entre double accolades, est un ensemble avec multiplicité (les éléments peuvent être présents plusieurs fois).

(en échelles lin-lin), et de mettre en évidence les décroissances exponentielles (lignes droites dans les échelles lin-log) et les décroissances polynomiales (lignes droites dans les échelles log-log). La décroissance exponentielle dans les distributions sont une caractéristique d'homogénéité (les valeurs sont exponentiellement plus rares quand elles croissent) et les décroissances polynomiales sont un indicateur de l'hétérogénéité (les valeurs sont *seulement* polynomialement plus rares quand elles croissent). Notons que nous parlons ici de distributions décroissantes, mais le même raisonnement est applicable pour leurs parties croissantes.

Nous étudions également, et de façon similaire, les distributions cumulatives inverses (c'est-à-dire pour chaque valeur possible, le nombre d'occurrences de valeurs plus grandes que celle-ci), qui sont souvent plus faciles à lire. L'observation de ces distributions dans les échelles lin-lin, lin-log et log-log aboutit à des interprétations similaires à ce que nous avons décrit ci-dessus.

### 3.5.2 Ajustement (*fit*) automatique

L'inspection visuelle des distributions est l'approche la plus fiable car elle permet d'identifier plusieurs régimes dans les distributions et de confronter plusieurs visualisations. Elle souffre toutefois d'une limitation majeure : elle n'est pas automatique. Nous expliquons dans cette section comment il est possible d'automatiser la prise de décision concernant la nature des distributions homogènes, hétérogènes et homogènes avec *outliers* des distributions que nous rencontrons. Il faut toutefois garder à l'esprit que de telles méthodes automatiques peuvent induire en erreur et ne doivent être utilisées qu'avec précautions.

Ces techniques reposent sur l'utilisation de modèles de distributions qui peuvent ajuster des distributions empiriques. Il y a une grande variété de modèles de distributions possibles, mais notre objectif ici n'est pas de trouver une adéquation parfaite : nous voulons décider si la distribution empirique est plutôt homogène, plutôt hétérogène ou plutôt homogène avec *outliers*, comme décrit dans la section 3.3. Nous considérons donc que deux modèles de distribution [4] pour capturer la nature homogène ou hétérogène d'une distribution empirique : la distribution normale $P(x) = \frac{1}{\nu\sqrt{2\pi}} e^{\frac{-1}{2}(\frac{x-\mu}{\nu})^2}$, qui est une distribution homogène typique, de moyenne bien définie $\mu$, d'écart type $\nu$ et une décroissance exponentielle ; et la distribution de loi de puissance $P(x) \sim x^{-\alpha}$, qui est une distribution hétérogène typique, caractérisée par son exposant $\alpha$ et qui n'a pas de valeur *normale*.

Dans ce cadre, décider de la nature homogène ou hétérogène d'une distribution empirique donnée consiste à : (1) la comparer aux différents modèles de distribution

4. Nous avons effectué les mêmes calculs avec des modèles de poisson et gamma, conduisant à des résultats très similaires. La loi normale à l'avantage d'avoir deux paramètres, une moyenne bien définie $\mu$ et un écart type $\nu$, ce qui permet des ajustements de qualité tout en restant très simple

choisis (c'est-à-dire déterminer les paramètres de chaque modèle qui lui correspond le mieux), (2) déterminer, parmi les modèles envisagés et leur ajustement, le modèle qui correspond le mieux à la distribution empirique. Nous détaillons ces deux étapes ci-dessous.

Afin de traiter le cas crucial où la distribution empirique est homogène avec *outliers*, nous ajoutons une étape supplémentaire : d'abord nous identifions et supprimons les *outliers* possibles, puis nous comparons la distribution obtenue à la distribution modèle normale, comme la distribution d'origine, ce que nous détaillons ci-dessous.

Identifier les *outliers* possibles dans une distribution est une tâche difficile et un problème en soi [Hawkins 1980] ; nous allons utiliser ici une méthode simple basée sur le test de Grubb's [Thode 2002]. Elle consiste à comparer chaque valeur à la moyenne et l'écart-type de la distribution : si la valeur est supérieure ou inférieure à la moyenne multipliée par un nombre donné de fois l'écart-type, elle est alors considérée comme un *outlier*.

Les méthodes les plus classiques pour effectuer l'ajustement de distributions empiriques aux modèles de distribution sont probablement celles qui ont pour principe de minimiser l'erreur (*Minimum Error* ME) à chaque point de la distribution empirique, et celles qui ont pour principe de maximiser la vraisemblance (*Maximum Likelihood* ML) entre la distribution empirique et le modèle. Ici, nous utilisons l'estimation classique de maximum de vraisemblance (*Maximum Likelihood Estimation* MLE) [Eliason 1993, Crowder 2007]. La raison est que l'estimation par ME utilise seulement les moments de la distribution empirique plutôt que toutes les valeurs et est très sensible aux valeurs d'initialisation ; puisque nous n'avons pas une idée à priori des valeurs de l'initialisation, cette méthode est impossible à utiliser dans notre cas.

Notons que, dans notre contexte, l'ajustement n'est pas un but en soi : nous sommes intéressés plutôt à estimer combien chaque ajustement est pertinent, afin de pouvoir comparer les différents ajustements produits par les différents modèles. D'une manière conforme à la façon dont l'ajustement MLE est calculé, nous pourrions utiliser la vraisemblance obtenue pour estimer cette pertinence. Toutefois, la notion de vraisemblance dépend du modèle considéré, et donc comparer les vraisemblances obtenues avec différents modèles n'a guère de sens.

Nous avons donc jugé de la pertinence des ajustements des distributions empiriques en les comparant directement. Une méthode classique pour comparer plusieurs ajustements est l'utilisation du test Kolmogorov-Smirnov (KS) [Press 1992]. Ce test permet de fournir une distance $D$, entre la distribution cumulative d'un modèle noté $F_{model}$ et la distribution cumulative empirique notée $F_{empirical}$. $D = max|F_{model} - F_{empirical}|$. Le modèle qui produit la valeur la plus basse de D est celui qui correspond le mieux à la distribution empirique. Cela donne une comparaison selon le *pire des cas*, puisque le moment où l'ajustement est le plus faible détermine entièrement la valeur de la distance $KS$. Si l'ajustement est médiocre sur ce point, mais excellent partout ailleurs,

la distance KS sera élevée. Pour obtenir plus d'efficacité, nous calculons la distance de Monge-Kantorovich (MK) [Georgiou 2009], définie comme la distance $D'$ calculée pour toutes les $N$ valeurs de la distribution comme : $D' = (\Sigma_{1..N}|F_{model} - F_{empirical})|/N$. Cela nous donne une distance moyenne entre la distribution cumulative empirique et son ajustement.

Nous obtenons finalement une méthode complète pour décider automatiquement si une distribution empirique est homogène, hétérogène ou homogène avec *outliers* : nous ajustons les modèles de distributions normale et en loi de puissance à la distribution empirique en utilisant le $MLE$ ; nous supprimons les *outliers* potentiels avec le Grubb's test et nous ajustons la distribution modèle empérique à la distribution obtenue ; nous calculons les distances KS et MK entre les distributions empiriques et chacun des modèle ajustés, et concluons que la distribution empirique est plutôt homogène si la distance est la plus petite pour le modèle normal, hétérogène si elle l'est pour le modèle en loi de puissance, ou homogènes avec *outliers* si après avoir enlevé ces dernières la distance avec le modèle normal est la plus petite. Nous supposons à priori que les distances KS et MK sont en accord ; dans le cas contraire, on obtient un bon indicateur du fait que la méthode automatique ne devrait pas être appliquée à ce cas.

## 3.6 Corrélation entre les événements détectés

Dans les sections précédentes, nous avons défini diverses propriétés visant à détecter des événements dans des graphes dynamiques. Nous pouvons nous s'attendre ou pas à ce que ces propriétés conduisent à des distributions homogènes avec des *outliers*. Nous pouvons, toutefois, nous demander si toutes les propriétés permettant la détection d'événements révélent les mêmes événements, dans ce cas, les propriétés subtiles et plus coûteuses ne seraient pas utiles, ou si elles détectent différents événements, dans ce cas, elles sont toutes utiles et complémentaires. Afin d'explorer cela, nous étudions les corrélations entre les événements détectés par chaque propriété. La corrélation des événements détectés entre eux, n'est pas une question triviale, pour l'aborder, nous allons juste affronter des propriétés, notamment les séries temporelles correspondantes afin d'avoir une idée sur la relation entre les événements détectés par ces dernières.

## 3.7 Interprétation

Dans tout ce chapitre, nous avons présenté notre méthodologie pour détecter des événements statistiquement significatifs. Plus précisément, nous sommes en mesure d'identifier des moments dans le temps où des événements se produisent, et d'identifier les nœuds et les liens impliqués dans ces événements. Le but ultime est toutefois d'être capable d'*interpréter* les événements détectés, c'est-à-dire les décrire en terme

de ce qui s'est passé dans le réseau. Ceci est crucial pour une véritable compréhension de la dynamique du graphe de terrain considéré. L'interprétation d'événements est cependant difficile, car les connaissances sur la dynamique des graphes de terrain sont limitées. Pour aider dans cette tâche, nous proposons deux approches d'interprétation : la corrélation avec des événements connus, et la visualisation.

Idéalement, l'interprétation par corrélation avec des événements connus, pourrait se réaliser en faisant la correspondance entre une base de données d'événements survenus dans le graphe considéré, et ceux que nous détectons (et inversement). Cependant, ce n'est pas possible en général, car, il n'existe pas de base de données fiable pour le faire.

L'interprétation d'événements par visualisation, consiste en le dessin des graphes correspondants, afin de voir l'impact qu'a cet événement sur le graphe de terrain considéré. Les méthodes actuelles de visualisation étant incapables de traiter de façon satisfaisante de grands et/ou de produire des dessins faciles à interpréter, nous utilisons une technique de réduction de données qui est d'une grande aide. Cela consiste à se concentrer sur la partie du graphe de terrain impliquée dans l'événement. Pour ce faire nous identifions d'abord l'ensemble $S$ de nœuds et/ou liens impliqués dans l'événement (cela dépend de la propriété considérée). Nous sélectionnons la zone englobant cette partition du graphe, par exemple le sous-graphe induit par $S$ et les voisins des nœuds de $S$, ou l'ensemble des plus courts chemins passant par $S$. Enfin, nous visualisons cette partie du graphe, en identifiant par des couleurs les nœuds et/ou liens impliqués dans l'événement.

## 3.8 Conclusion

Dans ce chapitre nous avons proposé et décrit une approche générique pour automatiquement et rigoureusement détecter des événements dans la dynamique des graphes de terrain. Elle repose sur une notion d'événements statistiquement significatifs. Nous fournissons également des approches pour interpréter les événements détectés afin de mieux les comprendre, ainsi que leur impact sur les graphes de terrain.

Ce travail est à l'intersection de deux sujets : la détection d'événements et la dynamique des graphes de terrain. En ce qui concerne la détection d'événements, c'est à notre connaissance, le premier travail qui considère les graphes dynamiques. Cela nous a conduit à introduire des propriétés de telles dynamiques, des plus simples comme le nombre de nœuds à chaque passe de mesure à d'autres plus subtiles comme celles basées sur la distance, avec plusieurs valeurs par unité de temps (et un nombre variable de valeurs). Soulignons que notre méthode est générique et peut être appliquée directement à de nombreux graphes dynamiques (mesures de l'internet, réseaux sociaux, etc), ce qui est une contribution importante en soi.

La problématique de la détection d'événements est cependant loin d'être nouvelle. Il s'agit d'une problématique classique, qui nécessite toutefois en général une connaissance préalable du fonctionnement du système et/ou des caractéristiques des événements, ce qui rend la plupart des méthodes antérieures inapplicables dans notre contexte. Notre approche vise à pallier ce manque.

Dans le chapitre suivant, nous illustrons notre approche en l'appliquant à la détection d'événements dans les mesures de l'internet. Ce faisant, nous démontrons sa pertinence en pratique, qui reste à établir à l'issue du présent chapitre.

Chapitre 4

# Cas d'étude : radar de l'internet

## Sommaire

## 4.1 Introduction

L'internet joue aujourd'hui un rôle clé dans notre société, notre économie et notre vie quotidienne. Cependant, notre connaissance des problèmes liés à son infrastructure (défaillances, attaques, congestions, bogues) reste très limitée. En conséquence, lorsque l'on est confronté à des pertes importantes, à des dégradations de la connectivité ou d'un service, on se trouve face à une compréhension limitée des phénomènes sous-jacents, de leur impact sur l'internet et des solutions pour les prévenir.

Comme nous l'avons souligné dans le chapitre 3, la détection d'événements dans la dynamique des graphe de terrain en général, et en particulier dans la dynamique de l'internet, est confronté à la problématique liée à l'état de l'art des approches de détection d'événements, mais aussi à la nature des objets concernés.

L'étude de la topologie de l'internet, obtenue à partir de mesures, attire beaucoup d'attention, voir par exemple [Magoni 2001, Li 2004, Guillaume 2006]. Toutefois, l'obtention d'informations à son sujet est extrêmement difficile, car nos capacités de mesure sont limitées [Lun 2005, Crespelle 2011] et la dynamique de l'objet est très complexe [Albert 1999, Augustin 2006, Oliveira 2007, Lad 2007, Magnien 2009].

En effet, le *monitoring* du réseau complet à l'échelle mondiale est hors de portée. Le premier défi à résoudre, comme mentionné ci-dessus, serait d'effectuer des mesures qui capturent l'intégralité du réseau. Ces mesures reposent généralement sur des outils type `traceroute` et/ou les tables BGP. Ces cartes sont cependant très partielles et biaisées, et leur collecte est trop coûteuse (en temps et en charge du réseau) pour pouvoir les répéter à une fréquence suffisamment élevée pour une surveillance du réseau.

Pour contourner ces problèmes, il a été proposé de se concentrer sur une partie de la topologie appelée vue *ego-centrée*. Elle consiste en ce qu'une seule machine, appelée *moniteur*, voit de la topologie de l'internet. Elle est essentiellement capturée en exécutant des mesures `traceroute` à partir du moniteur vers un ensemble donné de destinations choisies au hasard, et en itérant ce processus toutes les quelques minutes. Cela peut être fait de façon très efficace, et des mesures de ce type ont été réalisées à grande échelle. Voir ci-après (section 4.2) et [Latapy 2009] pour plus de détails. L'ensemble des données obtenues et les principaux résultats de ces travaux sont disponibles (voir [Latapy 2009]). L'étude de ces vues ego-centrées a notamment montré que leur dynamique est beaucoup plus forte qu'on le pensait [Magnien 2009].

Il doit être clair que ces mesures donnent une image très partielle de la topologie globale. En outre, la nature ego-centrée de ces mesures a un impact fort sur les observations. En particulier, en cas de perte de connectivité proche du moniteur, la vue devient vide (ou presque), ce qui peut donner l'impression qu'un problème majeur s'est produit sur l'internet. Cela n'est pas vrai, car les pertes de connectivité peuvent être d'une portée très limitée (peut-être au niveau du moniteur lui même). Ces mesures

ont toutefois l'avantage d'avoir un sens clair et intuitif (ce sont essentiellement des arbres de routage à partir d'un moniteur), et elles peuvent être répétées à une fréquence relativement élevée (toutes les quelques minutes en général).

Nous proposons dans ce chapitre, l'application de notre approche empirique et générique pour la détection d'événements, à la dynamique des vues égo-centrées de l'internet, et nous démontrons son efficacité. Nous commençons par une description des données utilisées (section 4.2), telles qu'elles sont décrites dans [Latapy 2009]. Ensuite, nous introduisons les différentes observations (section 4.3 et section 4.5) des propriétés dynamiques introduites au chapitre 3. Enfin, nous évaluons leur pertinence pour la détection d'événements, avec la corrélation entre les événements détectés (section 4.6) ainsi qu'avec nos méthodes d'interprétation des événements détectés (section 4.7).

## 4.2 Données radar

Nous présentons ici les données que nous utilisons dans l'ensemble du chapitre. Bien qu'il soit parfois inexact, l'outil `traceroute` [Jacobson 1989] donne essentiellement les chemins dans l'internet au niveau IP suivi par les paquets envoyés par un moniteur vers une destination. Chaque nœud sur ce chemin est une adresse IP et chaque lien représente un saut au niveau IP. Cet outil est à la base de la plupart des mesures de niveau IP de la topologie de l'internet, et des cartes sont construites par l'union de ces routes mesurées à partir de plusieurs moniteurs et vers de nombreuses destinations, voir par exemple [Claffy 2001, Huffak 2002, Pineda 2002, Viger 2008, Madhyastha 2009].

L'outil `tracetree` [Latapy 2009] fonctionne d'une manière très similaire, mais donne un arbre de routage à partir d'un moniteur vers un ensemble de destinations, qui est appelé une mesure ego-centrée car il donne une vue de l'internet à partir de ce moniteur spécifique. Il est pratiquement équivalent à l'exécution de `traceroute`à partir du moniteur vers chaque destination de l'ensemble. Toutefois, `tracetree` induit une charge plus faible et plus équilibrée sur le réseau, et est reste plus rapide. En conséquence, il peut être répété à une fréquence relativement élevée, conduisant à ce qu'on appelle des mesures *radar*. Ces mesures consistent en une série de mesures égo-centrées périodiques, chacune étant appelée une *passe* de mesure.

Les mesures radar sont présentées dans [Latapy 2009] et les auteurs fournissent des données accessibles librement. Ils ont conduit ces mesures depuis plus d'une centaine de nœuds vers 3000 cibles aléatoires chacune, pendant plusieurs semaines en continu, avec environ une passe toutes les 15 minutes. Chaque moniteur possède donc sa propre vue ego-centrée de la dynamique de la topologie de l'internet au niveau IP. Nous avons effectué nos calculs sur plusieurs de moniteurs et obtenu des observations similaires. Comme notre but n'est pas de discuter des différences subtiles entre les points de vue

des moniteurs, nous présentons dans la suite les résultats d'un seul moniteur représentatif, qui a effectué 5000 passes de mesure (pendant plus de 7 semaines). Comparer les observations depuis plusieurs moniteurs pourrait conduire à une meilleure détection des événements, avec plus de précision ; c'est une de nos principales perspectives.

Notons enfin que, comme pour celles de `traceroute`, les mesures `tracetree` ne reçoivent pas nécessairement une réponse. Cela conduit à nœuds non identifiés sur les chemins (traditionnellement représentés par des étoiles '*'). Faire la correspondance entre les différents nœuds non identifiés d'une passe à une autre est un problème difficile, et peut interférer avec la détection d'événements. Nous avons donc décidé de simplement les ignorer (c'est-à-dire les retirer de la mesure) et de nous concentrer sur les nœuds avec une adresse IP réelle. En outre, nous avons supprimé les nœuds isolés (nœuds sans lien), qui ne fournissent aucune information topologique. Ceci est classique dans l'analyse des mesures de l'internet.

Afin d'étudier l'impact de la résolution temporelle de l'observation évoquée dans le chapitre 3 section 3.4, nous utilisons des intervalles de passes de tailles $c$ et $p$ variables. Nous avons effectué les calculs pour des gammes de valeurs allant de 1 à 100 pour les entiers $p$ et $c$. Nous utilisons, ci-après les valeurs $p = 10$ et $c = 2$, qui donnent des résultats représentatifs de ce que nous avons observé sur de larges plages de valeurs.

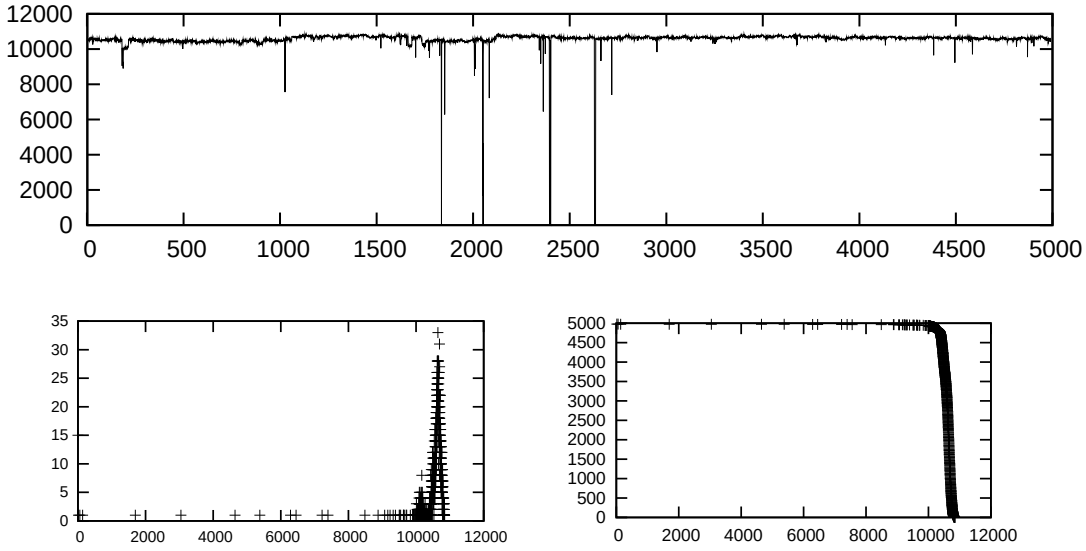
## 4.3 Nombres de nœuds

Comme expliqué dans la section 3.4 du chapitre 3, nous commençons par tenter de détecter des événements avec des propriétés basées sur des nombres de nœuds. Nous présentons dans cette section les observations des propriétés suivantes : le nombre de nœuds observé par passe, le nombre de nœuds distincts dans plusieurs passes consécutives, et enfin le nombre de nouveaux nœuds qui apparaissent.

### 4.3.1 Nœuds par passe

Nous présentons dans la Figure 4.1 le nombre $N_i$ de nœuds observés par passe de mesure $i$ effectuée. Cette courbe montre que le nombre de nœuds à chaque passe est très stable, à quelques exceptions près. La plupart du temps, il oscille à proximité d'une valeur moyenne légèrement supérieure à 10600. Toutefois, un examen plus attentif montre que cette valeur change à proximité des valeurs 1100 et 2100 : pendant un certain temps après ces passes, le nombre de nœuds oscille à proximité d'une valeur différente. En plus de ces changements dans la valeur moyenne, la courbe présente également de grands pics vers le bas. Néanmoins, aucun pic vers le haut n'est visible.

Ces observations sont confirmées par la distribution des valeurs de $N_i$ et les distances KS et MK, voir Figure 4.1. La distribution met en évidence deux régimes




| | KS | MK |
|---|---|---|
| normale | 0.3138 | 0.2394 |
| avec outliers | 0.1003 | 0.0993 |
| loi de puissance | 0.6215 | 0.5798 |

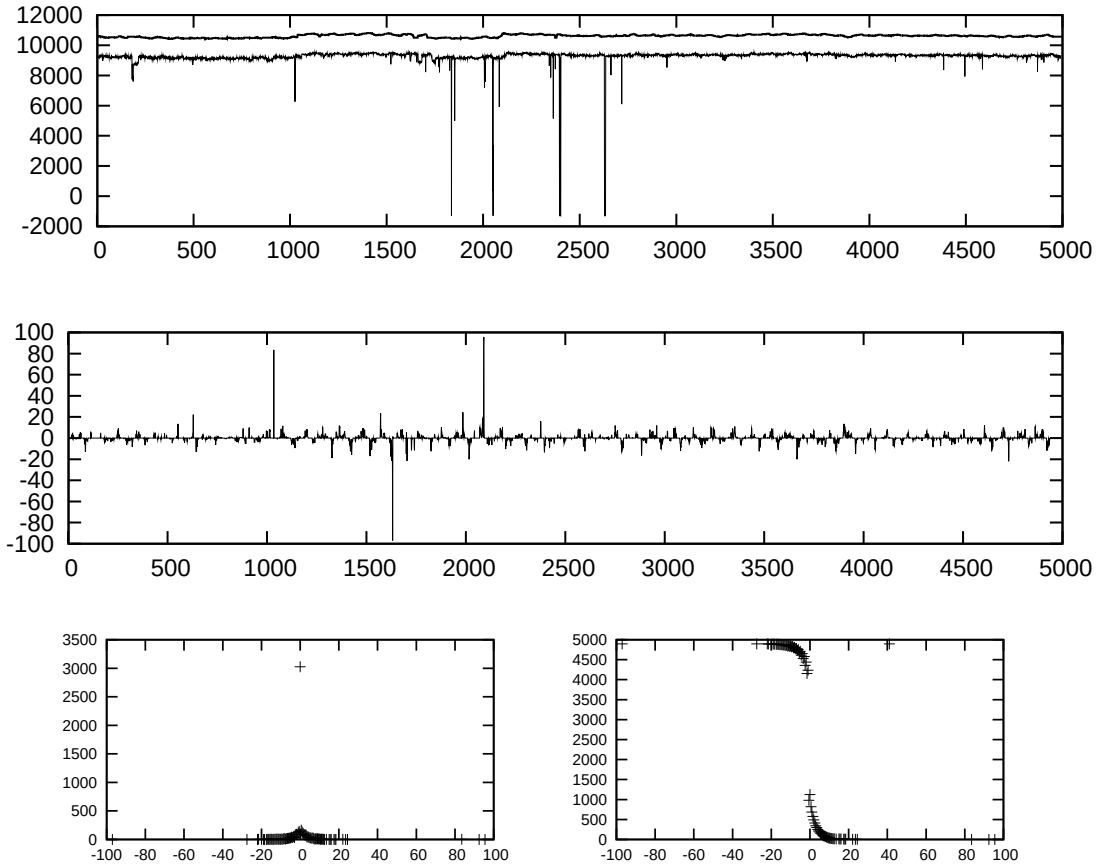
FIGURE 4.1 – De haut en bas : nombre $N_i$ de nœuds observés, en fonction de la passe de mesure $i$ effectuée ; à gauche, la distribution de valeur $N_i$ ; à droite, la distribution cumulative inverse de $N_i$ ; les distances KS et MK de la distribution avec les trois modèles étudiés.

distincts, avec de nombreuses valeurs autour de 10050 et 10060. Autrement, la distribution est clairement homogène avec des *outliers*. La présence de valeurs anormalement petites (les points sur la gauche de la distribution) mais aucune valeur anormalement grande, cela correspondant à la présence de pics vers le bas mais pas vers le haut. Il existe également un régime moins stable entre 1600 et 1800.

Il doit être clair que les pics vers le bas, même s'ils sont très clairement des *outliers*, donnent et dans ce cas précis peu d'informations : ils peuvent être causés par des défaillances de connectivité locale (chute de la connexion internet du moniteur, typiquement), qui ont pour effet de produire des vues ego-centrées vides ou presque pendant une ou quelques passes. Ce genre d'événement est trivial.

Au contraire, un pic vers le haut indiquerait un événement intéressant : cela signifierait que nous avons soudainement observé significativement plus de nœuds à une passe donnée. Cependant, il n'y a pas de pic vers le haut sur cette courbe, ce qui est un fait non trivial : on peut facilement imaginer un scénario où de tels pics seraient

visibles. La Figure 4.1 montre que de tel scénario ne se produit pas en pratique. En conséquence, on ne peut pas détecter des événements en observant des valeurs anormalement grandes de $N_i$.




| | KS | MK |
|---|---|---|
| normale | 0.1365 | 0.1321 |
| avec outliers | 0.0581 | 0.0547 |
| loi de puissance | 0.4725 | 0.4390 |

FIGURE 4.2 – De haut en bas : nombres $N_i$ et $M_i$ en fonction du nombre de passe $i$ de mesures effectuées ($N_i$ et $M_i$ se chevauchant, nous les avons décalés pour des raisons de lisibilité) ; le filtre médian $D_i$ en fonction du nombre de passe $i$ de mesures effectuées ; à gauche la distribution de $D_i$ ; à droite la distribution cumulative inverse de de ces valeurs ; les distances KS et MK de la distribution de $D_i$ avec les trois modèles étudiés.

Enfin, la dynamique la plus remarquable dans le nombre $N_i$ de nœuds observés par passe sont les changements dans la valeur moyenne autour de laquelle il oscille. Nous détectons ces changements en utilisant le filtre médian [1] comme suit : nous associons

1. Le filtre médian est une technique de filtrage numérique permettant de réduire le bruit sur un

à chaque passe $i$ la médiane des valeurs $N_i$ à $N_{i+100}$, que nous désignons par $M_i$, puis nous considérons les variations de la médiane, c'est-à-dire $D_i = M_i - M_{i-1}$ pour toute passe $i$. Nous avons tracé ces valeurs dans la Figure 4.2. Il est clair que ces valeurs réussissent (à la fois visuellement et avec notre méthode automatique) à détecter des événements, c'est-à-dire des *outliers* dans la distribution. Ceci est notre premier moyen de détecter des événements statistiquement significatifs.

### 4.3.2 Nœuds distincts dans des passes consécutives

Nous présentons dans la Figure 4.3 le nombre $N_i^5$ de nœuds distincts dans $c = 5$ passes consécutives en fonction de $i$. La courbe montre que, comme pour $N_i$, $N_i^5$ est plutôt stable et oscille autour d'une valeur moyenne[2]. Comme prévu, cette valeur est plus grande que celle de $N_i$, mais elle est loin d'être 5 fois plus grande. Cela montre que de nombreux nœuds apparaissent dans plusieurs passes consécutives. En outre, des pics vers le haut apparaissent sur cette courbe, ce qui la différencie fortement de celle de $N_i$ dans la Figure 4.1. La méthode automatique confirme la présence d'une valeur moyenne, et souligne clairement aussi des *outliers*, à la fois anormalement bas (comme pour $N_i$) et anormalement haut (ce qui est nouveau).

Cette observation est importante pour la détection d'événements : il y a des moments spécifiques (indiqués par les pics vers le haut de la Figure 4.3) auxquels un nombre anormal de nouveaux nœuds apparait dans une série de passes consécutives. Cela donne une nouvelle façon de détecter des événements statistiquement significatifs.

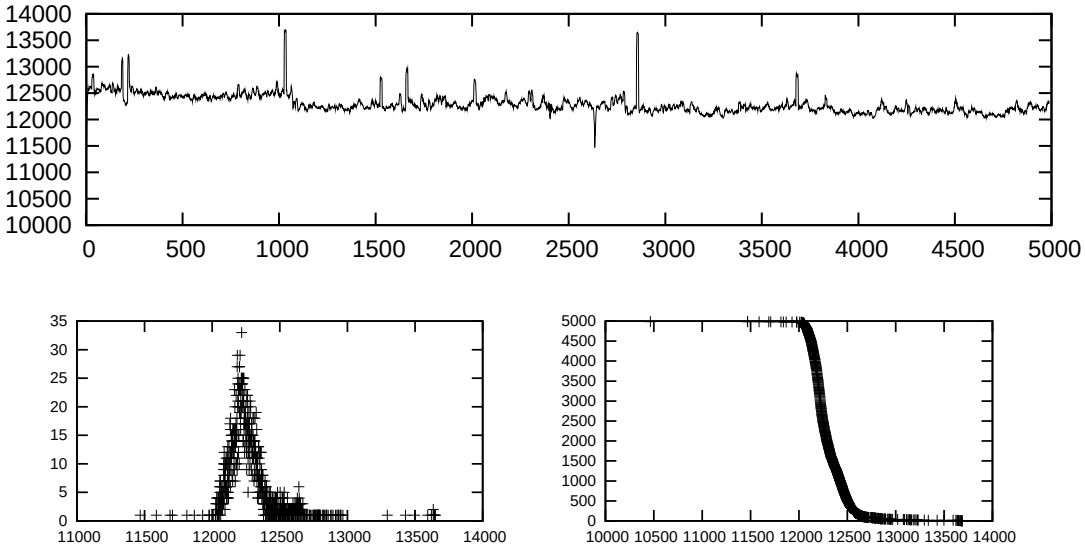
### 4.3.3 Nouveaux nœuds qui apparaissent

Nous définissons les nœuds qui apparaissent par ceux observés dans une série de $c$ passes consécutives, mais dans aucune des $p$ passes précédentes, $p$ et $c$ étant deux entiers. Avec nos notations, les nœuds qui apparaissent à une passe $i$ sont les nœuds dans $V_i^c \setminus V_{i-p}^p$.

Pour étudier cette propriété nous la traçons dans la Figure 4.4. La courbe obtenue présente clairement des pics vers le haut, indépendants de la valeur actuelle moyenne de $N_i^c$, ce qui est confirmé par les méthodes d'ajustements et les distances de la distribution des nœuds qui apparaissent. Nous obtenons ainsi une méthode de détection automatique d'événements statistiquement significatifs, définis comme des *outliers* dans le nombre de nœuds qui apparaissent.

signal [Pearson 2005]

2. Elles connaissent également des changements de régime, comme pour $N_i$, par exemple autour de la passe 1100 (voir Figure 4.3). Notons que, dans ce cas, la nouvelle valeur moyenne de $N_i^5$ est plus grande qu'avant, alors qu'elle était inférieure pour $N_i$. Cela signifie que, bien que nous voyons moins de nœuds à chaque passe, les nœuds que nous voyons varient plus d'une passe à l'autre.



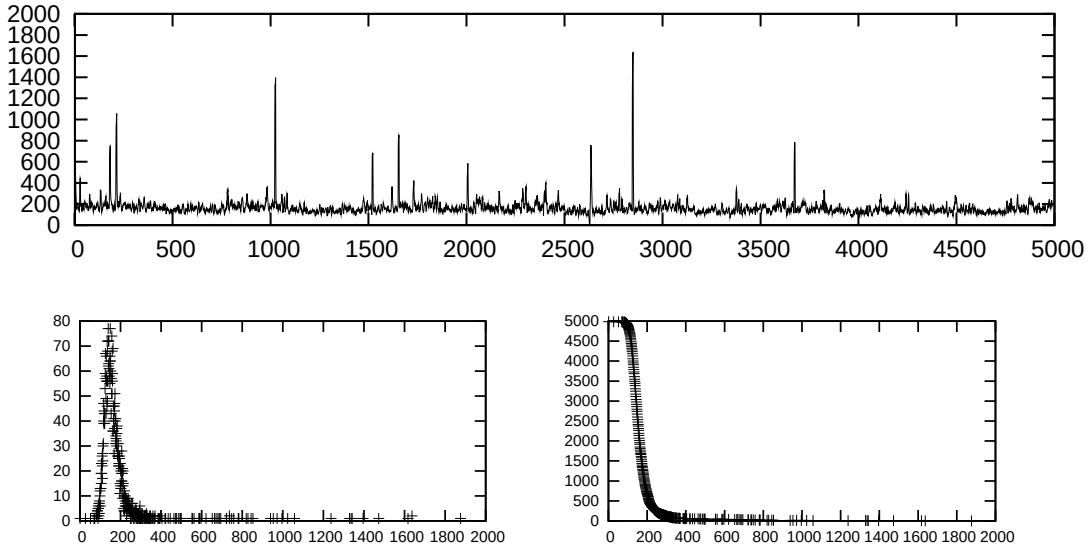

| | KS | MK |
|---|---|---|
| normale | 0.0795 | 0.0712 |
| avec outliers | 0.0495 | 0.0369 |
| loi de puissance | 0.6374 | 0.6118 |

FIGURE 4.3 – De haut en bas : nombre $N_i^5$ de nœuds distincts observés pendant cinq passes consécutives, en fonction du nombre de passe $i$ de mesure effectué ; à gauche, la distribution de $N_i^5$ ; à droite, la distribution cumulative inverse de $N_i^5$ ; les distances KS et MK de la distribution avec les trois modèles étudiés.

## 4.4 Composantes connexes

Nous avons vu dans la section précédente que, à certains moments particuliers, un nombre anormal des nouveaux nœuds qui apparaissent dans nos vues ego-centrées de la topologie de l'internet. Cependant, nous n'avons rien dit sur leur *structure* : sont-ils dispersés dans la topologie ? sont-ils regroupés ? ou appartiennent-ils à plusieurs petits groupes ? Intuitivement, par exemple, un changement important de routage peut mener à la découverte d'une nouvelle partie du réseau, ce qui serait révélé par l'apparition de nœuds formant une composante connexe dans nos vues ego-centrées.

Afin d'approfondir cette question, comme expliqué au chapitre 3 nous étudions les composantes connexes des nœuds qui apparaissent. Nous montrons le nombre de composantes connexes dans la Figure 4.5, et la taille de la plus grande composante connexe dans la Figure 4.6, pour chaque numéro de passe $i$ et avec les mêmes valeurs des paramètres $p = 5$ et $c = 2$ que dans la section précédente. Les événements statistiquement significatifs détectés avec le nombre de composantes connexes des nœuds qui apparaissent sont les mêmes que ceux détectés dans la section précédente. Cela




| | KS | MK |
|---|---|---|
| normale | 0.2381 | 0.2269 |
| avec outliers | 0.2000 | 0.1460 |
| loi de puissance | 0.9606 | 0.8995 |

FIGURE 4.4 – De haut en bas : le nombre $a_i$ de nœuds qui apparaissent, en fonction du nombre de passes de mesure $i$ effectué ; à gauche, la distribution de $a_i$ ; à droite, la distribution cumulative inverse de $a_i$ ; les distances KS et MK de la distribution avec les trois modèles étudiés. Ici nous considérons $p = 10$ et $c = 2$.

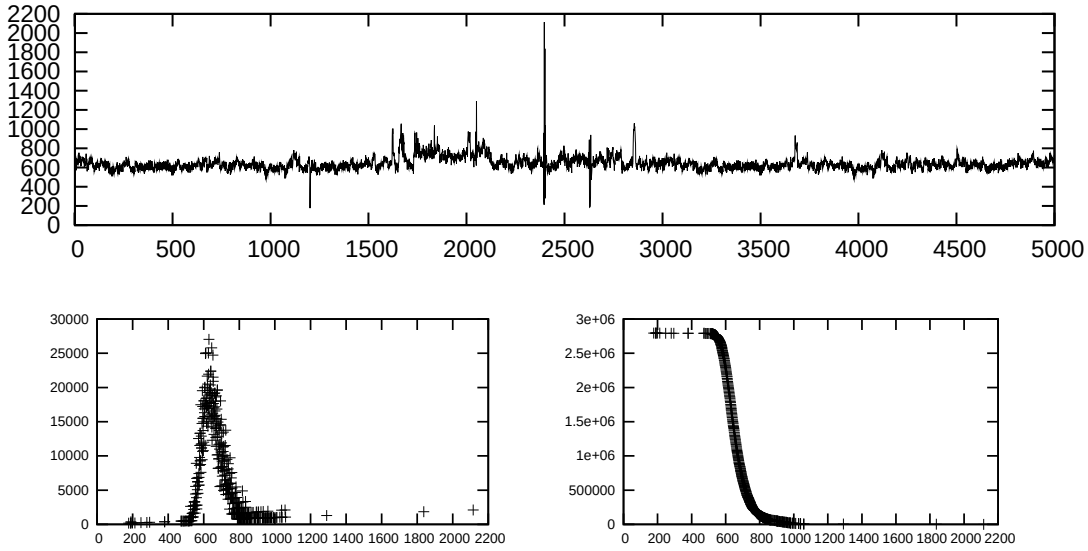
indique que le nombre de composantes connexes a peu d'intérêt de ce point de vue.

L'observation de la taille des composantes connexes nous permet d'aller plus loin dans nos réflexions sur les propriétés pertinentes pour la détection d'événements. En effet, elle a l'avantage qu'à chaque passe plusieurs valeurs sont obtenues ; une par composante connexe observée. Cela conduit à la distribution de la taille de *toutes* les composantes connexes qui apparaissant, quelle que soit leur passe, voir la Figure 4.7.

Cette distribution ne présente toutefois pas une différence claire entre des valeurs *normales* et d'autres *anormales* : la distribution est proche d'une loi de puissance, comme le montre les distances KS et MK dans la Figure 4.6. En conséquence, nous ne pouvons pas l'utiliser pour détecter des événements qui seraient révélés par l'apparition d'une composante connexe anormalement grande.

## 4.5 Distances

Dans cette section, nous considérons les propriétés basées sur la notion de distance. Comme expliqué dans le chapitre 3, nous observons les liens présents dans le graphe



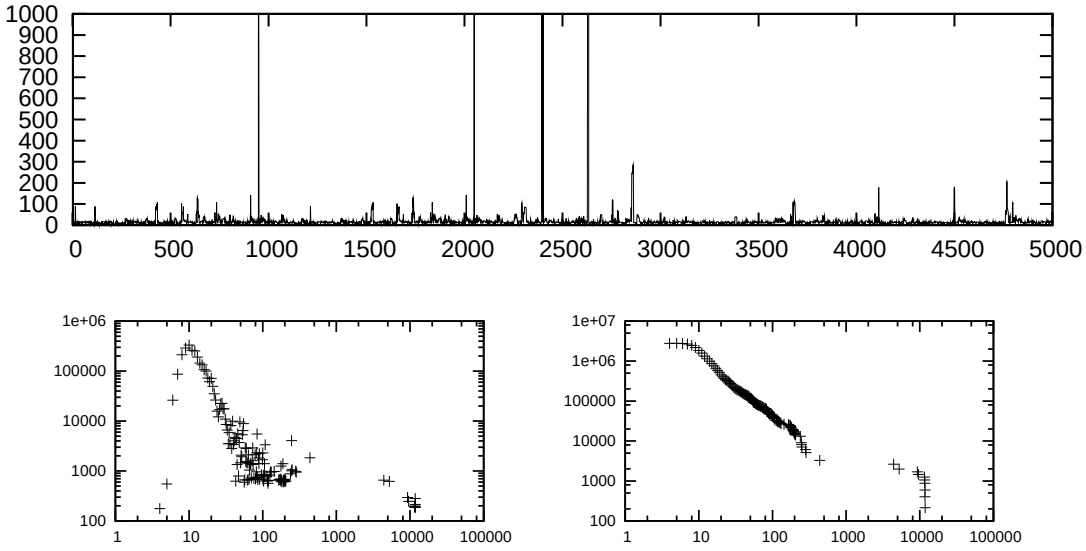

| | KS | MK |
|---|---|---|
| normale | 0.0807 | 0.0773 |
| avec outliers | 0.0519 | 0.0451 |
| loi de puissance | 0.4338 | 0.4192 |

FIGURE 4.5 – De haut en bas : nombre de composantes connexes des nœuds qui apparaissent, en fonction du nombre de passe de mesures $i$ effectuées ; à droite, la distribution du nombre de composantes connexes des nœuds qui apparaissent, à droite, la distribution cumulative inverse de ces valeurs ; les distances KS et MK de la distribution avec les trois modèles étudiés. Ici nous considérons $p = 10$ et $c = 2$.

courant entre des nœuds qui n'étaient pas liés dans le graphe précédent (mais y étaient présents) et calculons le multi-ensemble $D_i$ des distances entres ces paires de nœuds dans le graphe précédent (avant l'apparition du lien qui amène cette distance à 1).

Nous présentons dans la Figure 4.8 la valeur de $\max(D_i)$ en fonction de $i$, ainsi que la distribution de ces valeurs. Nous présentons aussi la distribution de toutes les distances observées, c'est-à-dire le multi-ensemble $\Gamma_i = \biguplus D_i$ dans la Figure 4.9. Pour ces valeurs, il n'y a pas de série temporelle significative. Cependant, comme pour les composantes connexes, le fait qu'il existe plusieurs valeurs par passe $i$ a l'avantage de rendre les résultats statistiquement plus fiables, et nous pouvons espérer détecter plusieurs événements qui se produisent en même temps.

Comme on s'y attendait, les distances observées sont assez petites, même si des valeurs relativement importantes (jusqu'à 35) apparaissent. Cela peut être considéré comme important, surtout que les distances dans les graphes considérés sont connues pour être petites (pas significativement supérieure à cette valeur extrémale). Ceci étant dit, la distribution de la distance maximale (Figure 4.8) est homogène. Les distances



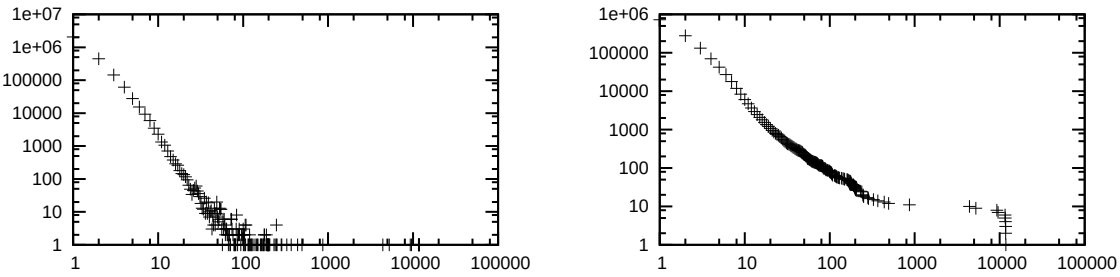

| | KS | MK |
|---|---|---|
| normale | 0.3852 | 0.3374 |
| avec outliers | 0.3986 | 0.3749 |
| loi de puissance | 0.3129 | 0.2740 |

FIGURE 4.6 – De haut en bas : la taille de la plus grande composante connexe des nœuds qui apparaissent, en fonction du nombre de passe de mesures $i$ effectuées ; à gauche, la distribution des ces valeur ; à droite, la distribution cumulative inverse de ces valeurs ; les distances KS et MK de la distribution avec les trois modèles de distribution étudiés. Ici nous considérons $p = 10$ et $c = 2$.

montrent que les valeurs extrêmes peuvent éventuellement être considérées comme des *outliers*, mais la différence avec une distribution purement homogène est faible. Si nous considérons *toutes* les distances, voir Figure 4.9, la distribution devient hétérogène, bien que le maximum soit par définition identique. Le meilleur ajustement est la loi de puissance, ce qui signifie que cette propriété ne peut pas être directement utilisée pour détecter des événements.

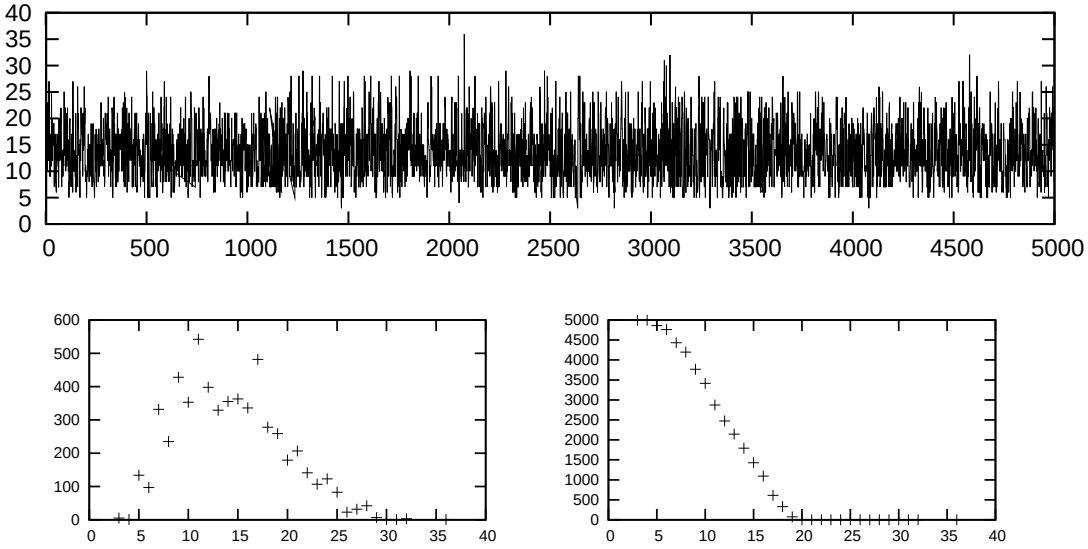
Nous nous intéressons maintenant au multi-ensemble $\Delta_i = \{\{\delta(u, v)\}\}$ dont le $\delta(u, v)$ (voir chapitre 3 section 3.4). Nous présentons dans la Figure 4.10 la valeur de $\max(\Delta_i)$ en fonction de $i$, ainsi que la distribution de ces valeurs. Dans la Figure 4.11 nous présentons aussi la distribution de toutes les valeurs obtenues, c'est-à-dire les valeurs du $\Delta_i$ (on se restreint à $i > 1100$ pour éviter le changement de régime visible dans la Figure 4.10).

Comme espéré, les valeurs de $\max(\Delta_i)$ couvrent un large intervalle, beaucoup plus large que de simples distances. Plus important encore, il réussit à identifier deux types d'événements : la valeur maximale pour chaque $i$ oscille à proximité d'une valeur



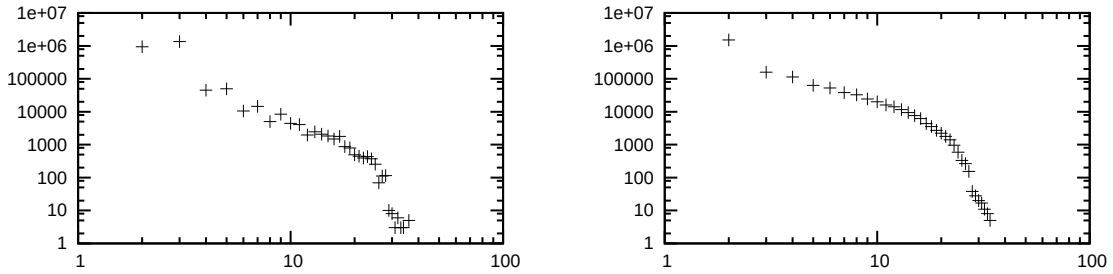

| | KS | MK |
|---|---|---|
| normale | 0.6842 | 0.5268 |
| avec outliers | 0.7742 | 0.6260 |
| loi de puissance | 0.2542 | 0.2007 |

FIGURE 4.7 – De haut en bas : à gauche, la distribution de la taille de *toutes* les composantes connexes qui apparaissant ; à droite, la distribution inverse de ces valeurs ; les distances KS et MK de la distribution avec les trois modèles de distribution étudiés. Ici nous considérons $p = 10$ et $c = 2$.




| | KS | MK |
|---|---|---|
| normale | 0.1100 | 0.0952 |
| avec outliers | 0.1142 | 0.0898 |
| loi de puissance | 0.5329 | 0.4921 |

FIGURE 4.8 – De haut en bas : la valeur $\max(D_i)$ de la plus grande distance entre les extrémités des liens qui apparaissent, en fonction du nombre de passes $i$ effectuées ; à gauche, la distribution de $\max(D_i)$ ; à droite, la distribution cumulative inverse de $\max(D_i)$ ; les distances KS et MK de la distribution avec les trois modèles de distribution étudiés. Ici nous considérons $p = 10$ et $c = 2$.




| | KS | MK |
|---|---|---|
| normale | 0.6842 | 0.5268 |
| avec outliers | 0.7742 | 0.6260 |
| loi de puissance | 0.2542 | 0.2007 |

FIGURE 4.9 – De haut en bas : à gauche, la distribution de $\Gamma_i$ le nombre de toutes les distances (plusieurs valeurs par passe) entre les extrémités des liens qui apparaissent, en fonction du nombre de passes $i$ effectuée ; à droite, la distribution cumulative inverse de ces valeurs ; les distances KS and MK de la distribution avec les trois modèles de distribution étudiés. Ici nous considérons $p = 10$ et $c = 2$.

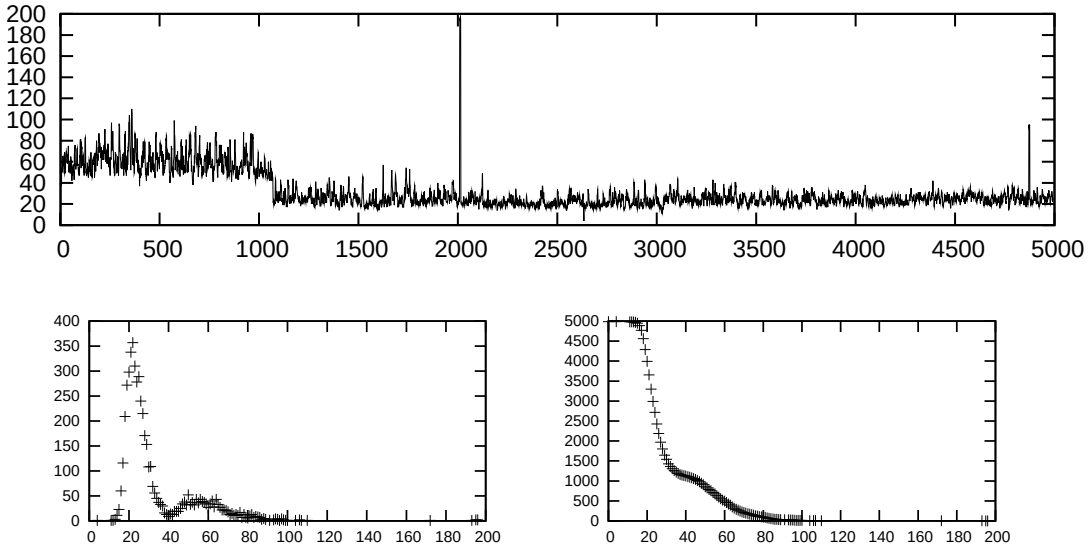
moyenne, mais présente des valeurs anormalement élevées, ainsi que des changements dans la valeur moyenne. Les deux phénomènes sont clairement visibles dans la Figure 4.10, et sont confirmé par les distributions. Les deux fournissent un moyen de détecter automatiquement des événements avec une propriété liée à la distance.

La situation n'est pas aussi claire lorsque nous considérons *toutes* les valeurs. La Figure 4.11 montre que la distribution est assez hétérogène, même si l'inspection visuelle peut indiquer que les valeurs de plus de 100 constituent des événements. Pourtant, notre méthode automatique considère la distribution comme hétérogène et manque ces événements. Des techniques statistiques plus subtiles pourraient être utilisées pour améliorer cette situation, mais cela sort du champ de ce travail.

## 4.6 Corrélations entre les événements détectés

Dans les sections précédentes, nous avons étudié diverses propriétés visant à détecter des événements dans la dynamique des vues égo-centrées de l'internet. Plusieurs ont conduit à des distributions homogènes avec *outliers*, et sont donc efficaces pour ce faire. Nous pouvons toutefois nous demander si toutes les propriétés détectent les mêmes événements, auquel cas les propriétés subtiles et plus coûteuses ne seraient pas utiles, ou si elles détectent des événements différents, auquel cas, elles seraient toutes utiles et complémentaires. Afin d'explorer cela, nous étudions dans cette section les corrélations entre les événements détectés par chaque propriété.

Pour ce faire, nous présentons ensemble dans la Figure 4.12 les trois propriétés

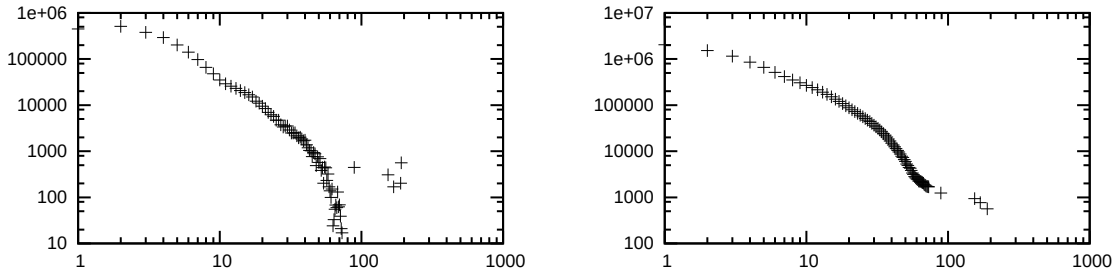



| | KS | MK |
|---|---|---|
| normale | 0.2875 | 0.2116 |
| avec outliers | 0.1651 | 0.0993 |
| loi de puissance | 0.3025 | 0.3002 |

FIGURE 4.10 – De haut en bas : le nombre maximal de nœuds qui changement leur distance de la première extrémité de nouveaux liens, en fonction du nombre de passes $i$ des mesures effectuée ; à gauche, la distribution de ces valeurs, la distribution cumulative de ces valeurs ; les distances KS et MK de la distribution avec les trois modèles de distribution étudiés. Ici nous considérons $p = 10$ et $c = 2$.

principales qui se sont révélées pertinentes pour la détection d'événements : le nombre de nœuds qui apparaissent ; les variations de la médiane du nombre de nœuds à chaque passe (section 4.3) ; le nombre maximal de nœuds qui changent de distance à l'extrémité d'un nouveau lien (section 4.5).

Plusieurs observations importantes peuvent être tirées de ces trois courbes. Tout d'abord, pour chaque propriété, il existe un événement (un pic dans les courbes) qui est détecté par cette propriété uniquement. Cela montre qu'elles ont toutes un intérêt et doivent être considérées comme complémentaires. Certains événements ne sont par contre détectés que par deux propriétés, comme celui légèrement après $i = 1000$. Cela montre que certains événements ont un impact sur plusieurs propriétés et d'autres pas, et donc notre approche identifie différentes classes d'événements.




| | KS | MK |
|---|---|---|
| normale | 0.3943 | 0.2862 |
| avec outliers | 0.4956 | 0.2190 |
| loi de puissance | 0.1307 | 0.1192 |

FIGURE 4.11 – De haut en bas : à gauche, la distribution de $\Delta_i(u, v)$ pour tous les liens apparaissant pour tout $i > 1100$ (plusieurs valeurs par passe) ; à droite, la distribution cumulative de ces valeurs ; les distances KS et MK de la distribution avec les trois modèles de distribution étudiés. Ici nous considérons $p = 10$ et $c = 2$.

## 4.7 Interprétation

Nous présentons dans cette section l'application de nos méthodes d'interprétation d'événements, telles qu'elles sont expliquées dans le chapitre 3 section 4.7, sur notre cas d'étude, qui est la topologie de l'internet.

### 4.7.1 Corrélation avec des événements connus

Afin d'aider à leur maintenance et offrir de meilleurs services, certains FAI enregistrent les *événements* survenant sur leur réseau et les documentent dans des bases de données. Ces bases offrent des informations partielles, mal structurées, et nécessitent une d'inspection manuelle minutieuse [Kuatse 2007, Huang 2008]. Néanmoins elles sont d'un grand intérêt dans notre contexte, car elles permettent d'essayer de correspondre des événements statistiquement significatifs, que nous détectons, aux événements réseaux connus signalés dans ces bases de données.

Abilene [Houweling 2005] est l'un des principaux FAI à fournir des informations riches sur des événements survenus dans leur réseau. Une base de données de tickets d'incidents décrivant de tels événements est disponible gratuitement en ligne ; nous présentons des exemples typiques dans la Figure 4.13.

Pour faire correspondre un événement statistiquement significatif à un ticket d'incident, nous procédons comme suit. Nous sélectionnons d'abord un événement statistiquement significatif avec notre approche, puis nous extrayons les *timestamps* auxquels il arrive (ils correspondent aux pics de la courbe permettant la détection ainsi que

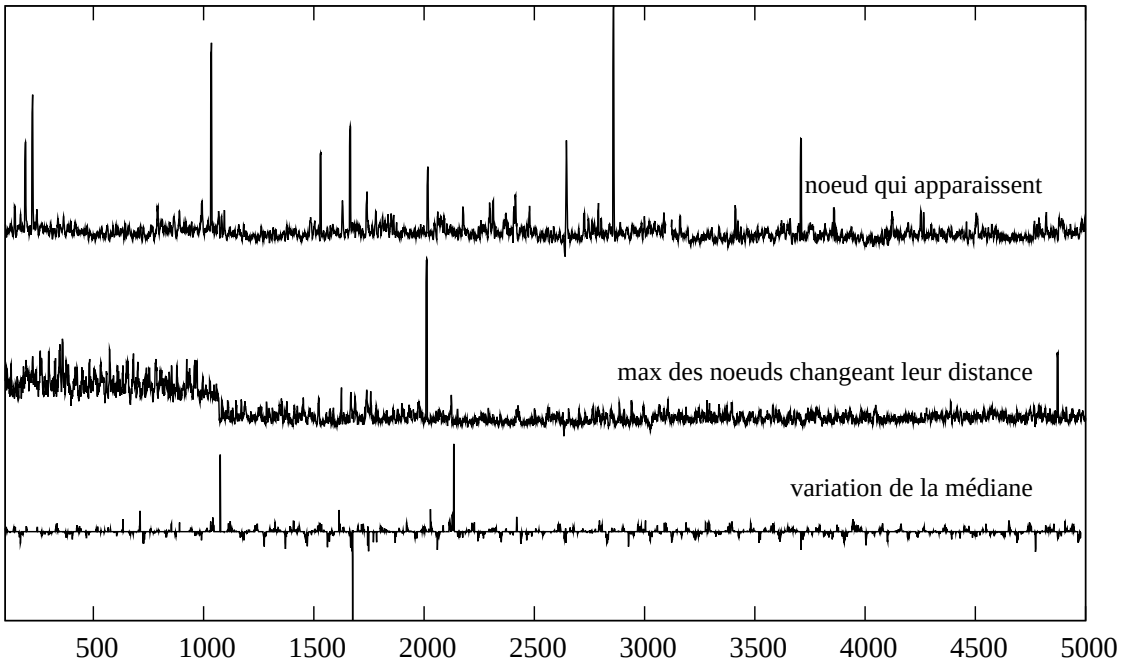



FIGURE 4.12 – Les trois propriétés principales qui se sont révélées pertinentes pour la détection d'événements. De bas en haut : les variations de la médiane du nombre de nœuds à chaque passe (section 4.3) ; le nombre maximal de nœuds qui changent de distance à l'extrémité d'un nouveau lien (section 4.5) ; et le nombre de nœuds qui apparaissent (section 4.3). Comme précédemment, nous avons utilisé $p = 10$ et $c = 2$. Nous avons réduit des courbes pour une meilleure lisibilité (les valeurs sur l'axe vertical n'auraient aucun sens, donc nous ne les affichons pas).

l'ensemble des nœuds impliqués dans l'événement). Nous cherchons ensuite dans la base de données Abilene un ensemble de tickets, tel que, les *timestamps* de ces tickets chevauchent les *timestamps* de notre événement, et les champs *AFFECTED* du ticket, citant les éléments réseaux affectés par l'événement, aient des adresses appartenant à l'ensemble $S$ de nœuds concernés. Nous avons donc à collecter par ailleurs les adresses IP des éléments cités dans le ticket afin de vérifier leur présence dans $S$.

Un exemple de résultat est affiché dans la Figure 4.14. Deux événements détectés sont en corrélation avec les tickets d'incident. Le premier événement statistiquement significatif et est caractérisé par une diminution significative du nombre de nœuds qui apparaissent. Cet événement est en corrélation avec le ticket d'incident Abilene illustré à la Figure 4.13 (à gauche). Un deuxième événement se produit plus tard, caractérisé par une augmentation équivalente du nombre de nœuds qui apparaissent. L'inspection du deuxième événement conduit à sa corrélation avec l'autre ticket dans la Figure 4.13 (à droite), qui s'avère être le ticket déclarant que le problème cité dans le premier ticket est réglé. Dans ce cas, donc, il y a une adéquation parfaite entre les deux événements

```
SUBJECT:        Internet2 IP Network Peer SINET (CHIC) Outage
AFFECTED:       Peer SINET (CHIC)
STATUS:         Unavailable
START TIME:     Thursday, May 17, 2007, 11:47 AM (1147) UTC
END TIME:       Pending
DESCRIPTION:    Peer SINET's connection the Internet2 IP
                Community is unavailable.  SINET Engineers
                have been contacted, however, no cause of
                outage has been provided yet. SINET is multi-homed.
TICKET NO.:     10201:45
TIMESTAMP:      07-05-17 00:40:43 UTC


SUBJECT:        Internet2 IP Network Peer SINET (CHIC) Resolved
AFFECTED:       Peer SINET (CHIC)
STATUS:         Available
START TIME:     Friday, May 25, 2007, 3:30 AM (0330) UTC
END TIME:       Friday, May 25, 2007, 10:00 AM (1000) UTC
DESCRIPTION:    Peer SINET was unavailable to the Internet2 IP
                Network Community.   SINET Engineers reported the
                reason for outage was due to a fiber cut in New York.
                SINET is multi-homed.
TICKET NO.:     10211:45
TIMESTAMP:      07-05-25 07:39:16 UTC
```

FIGURE 4.13 – Deux exemples de tickets d'incidents Abilene qui correspondent à des événements détectés dans la Figure 4.14 ; en haut, celui correspondant au premier événement, qui décrit une intervention technique sous le numéro 10201 : 45, les éléments de réseau concernés étant cités dans le champ *AFFECTED* ; les *timestamps* de début et de fin sont donnés, et les détails sont fournis dans le champs *DESCRIPTION* ; en bas, le ticket correspondant au deuxième événement.

statistiquement significatifs en question, et ceux présentés dans la Figure 4.13.

### 4.7.2 Visualisation

On peut aussi examiner un événement détecté par la manipulation du graphe sous-jacent. Bien que de nombreuses méthodes de dessin de graphe existent [Herman 2000], avec des avantages et des limitations, dans la plupart des cas la taille de nos données est prohibitive. À cet égard, être capable d'identifier un moment dans le temps où un événement se produit et de se concentrer sur les nœuds impliqués dans l'événement, tel que décrit dans le chapitre 3 section 4.7, est crucial. En effet la réduction des données

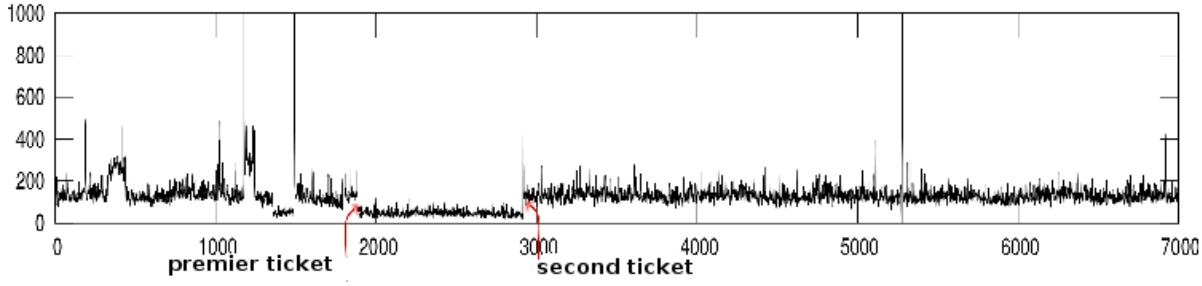



FIGURE 4.14 – Le nombre $a_i$ des noeuds qui apparaissent, en fonction du nombre de passes $i$ de mesures effectuées, Les deux flèches indiquent deux événements statistiquement significatifs qui ont été corrélés avec deux événements connus suivis par les tickets d'incident Abilene dans la Figure 4.13.

conduit à des graphes de quelques milliers de nœuds, que plusieurs logiciels sont en mesure de manipuler.

On peut alors dessiner en différentes couleurs les nœuds et/ou des liens qui apparaissent, qui disparaissent et les stables. La Figure 4.15 affiche un exemple typique. Nous observons que, alors que les nouveaux nœuds et liens sont en général dispersés dans tout le réseau, cet événement correspond à un changement significatif dans une partie spécifique de la topologie.

Cet examen manuel des événements, en utilisant des logiciels de dessins et/ou de manipulation de graphes, ouvre la voie à une compréhension plus détaillée des événements détectés, et à leur interprétation en termes d'événements réseaux.

## 4.8 Conclusion

Dans ce chapitre nous avons appliqué notre approche générique de détection d'événement dans la dynamique de graphes de terrain sur les mesures égo-centrées de la topologie de l'internet, que nous avons détaillé dans le chapitre 3.

Pour ce faire nous avons observé le comportement de la dynamique de ces vues, à travers les différentes propriétés définies. Toutes les sortes de distributions de ces propriétés sont apparues : homogènes et hétérogènes, qui ne conduisent pas à la détection d'événements, et homogènes avec *outliers*, qui eux le font. En outre, nous montrons que les différentes propriétés conduisent à la découverte de différents événements, et que par conséquent il est utile de définir plusieurs propriétés de graphes dynamiques afin de mieux la cerner.

Pour aller plus loin dans la compréhension des événements, on pourrait observer les événements détectés à partir de différents moniteurs. Certains événements peuvent être invisibles d'un moniteur particulier, ou paraître importants alors qu'ils ne le sont

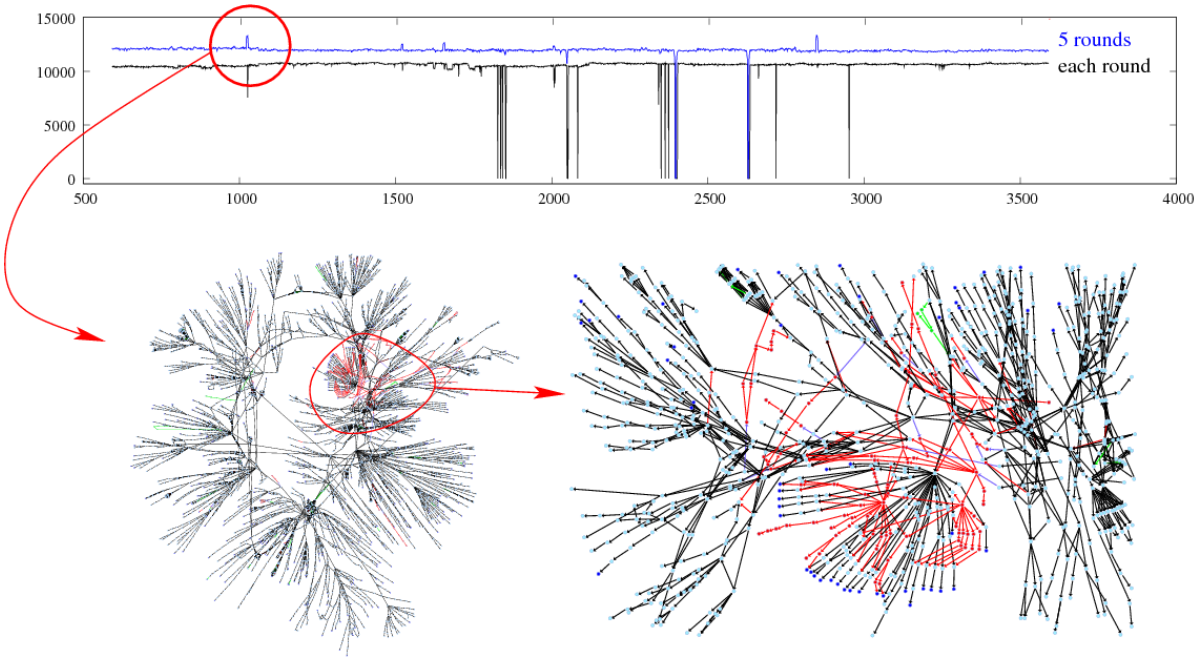



FIGURE 4.15 – De haut en bas : le nombre $N_i^5$ de nœuds distincts observés pendant cinq passes consécutives et le nombre $N_i$ de nœuds observés à chaque passe de mesure, en fonction du nombre de passes de mesure $i$ effectuées ; le graphe des changements topologiques observés lors de la détection d'un événement avec un zoom sur la zone correspondante sur la vue ego-centré. Les nœuds et les liens qui apparaissent sont en rouge.

pas. Cette approche donnerait des vues complémentaires et permettrait une détection plus fine et plus sûre.

Nous avons également exploré notre approche de corrélation des événements détectés ; les observations sont non-triviales : toutes les situations possibles se produisent, ce qui montre que l'ensemble des événements détectés est probablement très complexe et riche (bien que limité en taille ce qui est un élément crucial). Explorer cette direction entièrement demeure une de nos principales perspectives.

Nous avons ensuite interprété nos événements détectés en les corrélant avec des événements des tickets d'incident d'Abile. Ceci a permis de *traduire* nos événement statistiquement significatifs en des événements réseaux, ce qui permet de vérrifier la validité de notre approche générique de détection d'événements.

Enfin nous avons visualisé les événements détectés en dessinant leurs graphes sous-jacents. Cette piste a montré son efficacité dans une compréhension plus profonde et avancée des impacts que peuvent avoir les événements détectés sur la topologie des graphes de terrain.

Chapitre 5

# Conclusion et perspectives

Nous nous sommes focalisés dans cette thèse sur la problématique de la détection d'événements, définie comme la capacité à pointer des modifications particulières dans les systèmes qui ne sont pas conformes au *comportement attendu*. La contribution principale de cette thèse réside dans la proposition et la mise en œuvre d'une approche générique pour détecter, automatiquement et rigoureusement, des événements dans les dynamiques de graphes de terrain.

Nous avons détaillé notre approche et son principe de base dans le Chapitre 3. Ce dernier s'appuie sur une notion d'événements statistiquement significatifs. Selon ce principe, caractériser un événement dans la dynamique d'un graphe de terrain nécessite d'identifier des propriétés de la dynamique du graphe dont la distribution est normale avec *outliers*. Afin d'appliquer cette notion d'événements *statistiquement significatifs* à la détection des événements dans les dynamiques des graphes de terrain, nous avons proposé un ensemble de propriétés de graphes dynamiques.

Afin d'explorer l'apport spécifique de chaque propriété, nous avons étudié les corrélations entre les événements détectés par chacune d'elle. Dans la même optique, et afin de mieux les cerner et comprendre l'impact des événements détectés sur les graphes de terrain concernés, nous avons complémenté notre méthode de détection d'événements avec deux approches d'interprétation : la corrélation avec des événements connus, et la visualisation.

Pour démontrer l'efficacité de notre approche, nous l'avons appliquée aux mesures égo-centrées de l'internet dans le Chapitre 4. Nous avons rencontré les trois types de distributions des modèles considérés. Pour certaines propriétés, les distributions sont homogènes ou hétérogènes, et donc elles ne conduisent pas à la détection d'événements. Nous avons aussi observé des distributions homogènes avec *outliers*, qui permettent elles de détecter des événements statistiquement significatifs dans la dynamique des vues égo-centrées. Nous avons montré qu'elles conduisent à la découverte de différents événements, et que par conséquent il est pertinent de définir plus de propriétés de graphes dynamiques afin de mieux les cerner.

Après avoir localisé des événements dans les vues égo-centrées de l'internet, nous avons étudié leurs corrélations et leur interprétation. Les corrélations entre les événements détectés entre eux ne sont pas triviales. Toutes les situations possibles se pro-

duisent : des événements détectés uniquement par une propriété donnée ; des événements observés par la plupart des propriétés ; et enfin des événements détectés par un sous-ensemble de propriétés. Cela montre que l'ensemble des événements détectés est probablement très complexe et riche (bien que limité en taille, ce qui est un élément crucial). Explorer cette direction demeure une de nos principales perspectives. L'interprétation par corrélation avec des événements connus et par visualisation quant à elle, nous a permis d'avoir des éléments de réponse pour une compréhension plus détaillée des événements statistiquement significatifs, à travers leur correspondance avec des événements connus et leur l'impact sur la topologie observée.

Notons que, comme on ne sait pas ce qui caractérise la dynamique normale des graphes de terrain et des événements, notre approche peut être considérée comme un moyen d'explorer les données de mesure. Elle donne un aperçu sur ce qui peut être considéré comme dynamique normale et ce qui indique des événements. Elle localise des moments précis dans le temps où quelque chose d'inhabituel se produit et identifie les ensembles de nœuds et les liens en cause. Ceci permet d'approfondir l'étude de la dynamique, en fournissant ainsi l'un des outils les plus efficaces actuellement disponibles pour l'étude empirique de la dynamique de graphes.

Dans ce qui suit, nous décrivons trois directions de recherches pour les travaux futurs qui nous semblent les plus pertinents pour étendre les résultat de cette thèse : travailler à améliorer notre méthode de détection d'événements ; travailler sur l'interprétation des événements qu'on peut voir comme une approche *top-down* (on part de la mesure) ; développer une approche *bottom-up* pour mieux comprendre les dynamiques observées, à partir de dynamiques simulées.

On peut espérer améliorer notre approche en se concentrant successivement sur les différentes phases de sa méthodologie. Dans cette direction, une première piste serait d'explorer des propriétés plus subtiles, notamment celles qui permettent d'avoir plusieurs valeurs par unité de mesure. C'est d'autant plus pertinent que les propriétés étudiées montrent que l'ensemble des événements détectés est complexe et riche. Un exemple de propriété pertinente pourrait être les changements *inhabituels* dans la structure des communautés selon l'évolution de leur modularité [1]. Une deuxième piste concerne l'étude des distributions empiriques. Nous proposons pour garantir de meilleures résultats de compléter les techniques d'ajustement automatique avec des méthodes d'inspection visuelle des distributions. Afin d'aboutir à une automatisation plus complète de notre approche, des techniques statistiques d'ajustement automatiques plus raffinées sont à envisager, pour garantir un niveau de confiance plus élevé dans les décisions sur la nature des distributions des propriétés étudiées. On pourrait aussi

1. La modularité est une fonction de qualité utilisée dans la détection de communautés dans les graphes. Elle quantifie la qualité d'un partitionnement d'un graphe en communautés.

relaxer les critères sur la présence d'*outliers* (par exemple en cas de distribution en loi de puissance, mais avec des valeurs anormales).

En ce qui concerne l'amélioration des approches d'interprétation, une première piste serait d'approfondir l'étude des corrélations avec des événements connus, en explorant d'autres événements en provenance de plusieurs bases de données, détectés par plusieurs FAI par exemple. Une deuxième piste concerne naturellement l'interprétation par visualisation. Il serait avantageux d'intégrer notre méthode à des outils de visualisation de graphe dynamique et ainsi examiner beaucoup plus finement les événements détectés et leur impact sur la topologie.

La troisième direction de recherche qui nous semble prometteuse pour mieux comprendre la dynamique et les événements à détecter, serait de partir de dynamiques connues (simulées) et d'étudier la vision que notre méthode en donne dans le contexte de notre cas d'études qui est les mesures égo-centrées de l'internet. Une piste serait de faire des simulation de mesures égo-centrées sur des graphes avec des dynamiques simulées (en enlevant et/ou en ajoutant des liens et/ou des nœuds aléatoires par exemple). Ceci éclairerait les relations entre ce que nous observons et les mesures et les événements réels de la topologie, ce qui est crucial dans notre contexte.

Enfin, maintenant que nous avons introduit une méthode générique de détection d'événement dans les dynamique de graphes de terrain, une perspective naturelle est de l'appliquer à d'autres cas d'études (comme certains réseaux sociaux, par exemple). Dans de tels cas, l'interprétation d'événements peut être plus facile et donc permettre une validation plus forte de notre approche. Dans tous les cas, nous espérons avoir ouvert ici la voie à l'étude des événements dans la dynamique de nombreux graphes de terrain.

# Bibliographie


[Albert 1999] Réka Albert et Albert L. Barabási. *Topology of Evolving Networks : Local Events and Universality*. Physical Review Letters, vol. 85, no. 24, pages 5234–5237, 1999. (Cité en page 34.)

[Albert 2000] Réka Albert, Hawoong Jeong et Albert-László Barabási. *Error and attack tolerance of complex networks*. Nature, vol. 406, no. 6794, pages 378–382, 2000. (Cité en page 7.)

[Albert 2002] Réka Albert et Barabási. *Statistical mechanics of complex networks*. Rev. Mod. Phys., vol. 74, pages 47–97, Juin 2002. (Cité en pages 6 et 8.)

[Allali 2011] Oussama Allali, Clémence Magnien et Matthieu Latapy. *Link prediction in bipartite graphs using internal links and weighted projection*. In Proceedings of the third International Workshop on Network Science for Communication Networks (Netscicom 2011), In Conjuction with IEEE Infocom, 2011. (Cité en page 11.)

[Augustin 2006] Augustin, Brice, Cuvellier, Xavier, Orgogozo, Benjamin, Vigerand Fabien, Friedman, Timur, Latapy Matthieu, Magnien Clémence et Teixeira Renata. *Avoiding traceroute anomalies with Paris traceroute*. In IMC '06 : Proceedings of the 6th ACM SIGCOMM conference on Internet measurement, pages 153–158, New York, NY, USA, 2006. ACM. (Cité en page 34.)

[Aynaud 2010] Thomas Aynaud et Jean-Loup Guillaume. *Détection de communautés à long terme dans les graphes dynamiques*. In Journée thématique Fouille de grands graphes, en conjonction avec la première conférence sur les Modèles et l'-Analyse des Réseaux : Approches Mathématiques et Informatique (MARAMI), Toulouse, France, 2010. (Cité en page 12.)

[Babu 2004] M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein et S. A. Teichmann. Curr Opin Struct Biol, vol. 14, no. 3, pages 283–291, Juin 2004. (Cité en page 2.)

[Bakar 2006] Z.A. Bakar, R. Mohemad, A. Ahmad et M.M. Deris. *A Comparative Study for Outlier Detection Techniques in Data Mining*. In Cybernetics and Intelligent Systems, 2006 IEEE Conference on, pages 1 –6, 2006. (Cité en page 13.)

[Barabási 1999] A. L. Barabási et R. Albert. *Emergence of Scaling in Random Networks*. Science, vol. 286, no. 5439, pages 509–512, 1999. (Cité en page 1.)

[Barnett 1994] Vic Barnett et Toby Lewis. Outliers in statistical data. Wiley Series in Probability & Statistics. Wiley, Avril 1994. (Cité en page 13.)

[Barrat 2008] Alain Barrat, Marc Barthelemy et Alessandro Vespignani. Dynamical processes on complex networks. Cambridge University Press, 2008. (Cité en pages 6 et 8.)

[Basu 2007] Sabyasachi Basu. *Automatic outlier detection for time series : an application to sensor data*. Knowledge and Information Systems, vol. 11, pages 137–154(18), February 2007. (Cité en page 12.)

[Blond 2009] Stevens Blond, Fabrice Fessant et Erwan Merrer. *Finding Good Partners in Availability-Aware P2P Networks*. In Proceedings of the 11th International Symposium on Stabilization, Safety, and Security of Distributed Systems, SSS '09, pages 472–484, Berlin, Heidelberg, 2009. Springer-Verlag. (Cité en page 2.)

[Blondel 2008] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte et Etienne Lefebvre. *Fast unfolding of communities in large networks*. Journal of Statistical Mechanics : Theory and Experiment, vol. 2008, no. 10, pages P10008+, 2008. (Cité en pages 1 et 12.)

[Brotherton 2001] Tom Brotherton. *Anomaly detector fusion processing for advanced military aircraft*. In IEEE Aerospace Conference, Big Sky, pages 3125–3137, 2001. (Cité en page 12.)

[Brutlag 2000] Jake D. Brutlag. *Aberrant Behavior Detection in Time Series for Network Monitoring*. In Proceedings of the 14th USENIX conference on System administration, pages 139–146, Berkeley, CA, USA, 2000. USENIX Association. (Cité en page 12.)

[Calegari 2007] Roberta Calegari, Mirco Musolesi, Franco Raimondi et Cecilia Mascolo. *CTG : A Connectivity Trace Generator for Testing the Performance of Opportunistic Mobile Systems*. In Proceedings of the European Software Engineering Conference and the International ACM SIGSOFT Symposium on the Foundations of Software Engineering (ESEC/FSE07), Dubrovnik, Croatia, September 2007. ACM Press. (Cité en page 2.)

[Callaway 2000] D. S. Callaway, M. E. J. Newman, S. H. Strogatz et D. J. Watts. *Network Robustness and Fragility : Percolation on Random Graphs*. Physical Review Letters, vol. 85, no. 25, pages 5468–5471, 2000. (Cité en pages 1 et 7.)

[Ceyhan 2011] Simla Ceyhan, Xiaolin Shi et Jure Leskovec. *Dynamics of bidding in a P2P lending service : effects of herding and predicting loan success*. In Proceedings of the 20th international conference on World wide web, WWW '11, pages 547–556, New York, NY, USA, 2011. ACM. (Cité en page 11.)

[Chaintreau 2005] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, et J. Scott. *Pocket Switched Networks : Real-Word Mobility and its Consequences for Opportunistic Forwarding*. In WDTN, page 244–251, 2005. (Cité en page 2.)

[Chandola 2009] Varun Chandola, Arindam Banerjee et Vipin Kumar. *Anomaly detection : A survey*. ACM Comput. Surv., vol. 41, no. 3, pages 1–58, 2009. (Cité en page 13.)

[Chen 2002] Qian Chen, Hyunseok Chang, Ramesh Govindan, Sugih Jamin, Scott J. Shenker et Walter Willinger. *The Origin of Power Laws in Internet Topologies Revisited*, 2002. (Cité en page 23.)

[Claffy 2001] K. C. Claffy. *CAIDA : Visualizing the Internet*. IEEE Internet Computing, vol. 5, January 2001. (Cité en page 35.)

[Clauset 2007] A. Clauset et N. Eagle. *Persistence and periodicity in a dynamic proximity network*. In DIMACS Workshop, 2007. (Cité en page 2.)

[Clauset 2008] Aaron Clauset, Cristopher Moore et M. E. J. Newman. *Hierarchical structure and the prediction of missing links in networks*. Nature, 2008. (Cité en page 11.)

[Cohen 2010] Reuven Cohen et Shlomo Havlin. Complex networks : Structure, robustness and function. Cambridge University Press, Aot 2010. (Cité en pages 6 et 8.)

[Cointet 2009] Jean-Philippe Cointet et Camille Roth. *Socio-semantic dynamics in a blog network*. In Proceedings of the 2009 IEEE International Conference on Social Computing (SocialCom-09), Aug 2009, Vancouver, Canada IEEE International Conference on Social Computing (SocialCom-09), pages 114–121, Vancouver Canada, 2009. (Cité en page 2.)

[Crespelle 2011] Christophe Crespelle et Fabien Tarissan. *Evaluation of a new method for measuring the internet degree distribution : Simulation results*. Comput. Commun., vol. 34, pages 635–648, April 2011. (Cité en page 34.)

[Crowder 2007] M.J. Crowder. *Parameter Estimation for Scientists and Engineers by Adriaan van den Bos*. International Statistical Review, vol. 75, no. 3, pages 436–437, December 2007. (Cité en page 29.)

[Diehl 2002] Christopher P. Diehl et John B. Hampshire Ii. *Real-time Object Classification and Novelty Detection for Collaborative Video Surveillance*. In In Proceedings of the International Joint Conference on Neural Networks, pages 2620–2625, 2002. (Cité en page 12.)

[Dorogovtsev 2003] S. N. Dorogovtsev et J. F. F. Mendes. Evolution of networks : From biological nets to the Internet and WWW. Oxford University Press, 2003. (Cité en pages 6 et 8.)

[Ebel 2002] Holger Ebel, Lutz-Ingo Mielsch et Stefan Bornholdt. *Scale-free topology of e-mail networks*. Phys. Rev. E, vol. 66, no. 3, page 035103, Sep 2002. (Cité en page 7.)

[Eliason 1993] Scott R Eliason et Michael S Lewis Beck. Maximum likelihood estimation : Logic and practice. Sage Publications (CA), 1993. (Cité en page 29.)

[Erdös 1959] P. Erdös et A. Rényi. *On random graphs, I.* Publicationes Mathematicae (Debrecen), vol. 6, pages 290–297, 1959. (Cité en page 9.)

[Fleury 2007] Céline Robardet Antoine Scherrer Fleury Jean-Loup Guillaume. *Tools for the analysis of evolving sensor networks.* In In IEEE Conference on Communication System Software and Middleware, 2007. (Cité en page 11.)

[Fortunato 2004] Fortunato, Vito Latora et Massimo Marchiori. *Method to find community structures based on information centrality.* Physical Review, no. 056104, 2004. (Cité en page 1.)

[Fraigniaud 2007] Pierre Fraigniaud. *Small worlds as navigable augmented networks : model, analysis, and validation.* In Proceedings of the 15th annual European conference on Algorithms, ESA'07, pages 2–11, Berlin, Heidelberg, 2007. Springer-Verlag. (Cité en page 9.)

[Fujimaki 2005] Ryohei Fujimaki. *An approach to spacecraft anomaly detection problem using kernel feature space.* In in Proc. PAKDD-2005 : Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining. ACM Press, 2005. (Cité en page 13.)

[Georgiou 2009] Tryphon T. Georgiou, Johan Karlsson et Mir Shahrouz Takyar. *Metrics for Power Spectra : An Axiomatic Approach.* IEEE Transactions on Signal Processing, vol. 57, no. 3, pages 859–867, 2009. (Cité en page 30.)

[Girvan 2002] Michelle Girvan et Mark E. J Newman. *Community structure in social and biological networks.* 2002. (Cité en page 1.)

[Guillaume 2004] Jean-Loup Guillaume, Matthieu Latapy et Le-Blond Stevens. *Statistical analysis of a P2P query graph based on degrees and their timeevolution.* In 6th International Workshop on Distributed Computing (IWDC 04), Kolkata Inde, 2004. (Cité en pages 2 et 7.)

[Guillaume 2006] Jean-Loup Guillaume, Matthieu Latapy et Magoni Damien. *Relevance of massively distributed explorations of the internet topology : qualitative results.* Comput. Netw., vol. 50, pages 3197–3224, November 2006. (Cité en pages 1 et 34.)

[Hasan 2006] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem et Mohammed Zaki. *Link prediction using supervised learning.* In Proceedings of SDM 06 workshop on Link Analysis, Counterterrorism and Security, 2006. (Cité en page 11.)

[Hawkins 1980] D.M. Hawkins. dentifications of outliers, monograph on applied probability and statistic. Reading. London Chapman and Hall, 1980. (Cité en pages 13 et 29.)

[Herman 2000] Herman, G. Mélançon et M. S. Marshall. *Graph Visualization and Navigation in Information Visualization : A Survey*. IEEE Transactions on Visualization and Computer Graphics, vol. 6, no. 1, pages 24–43, slash 2000. (Cité en page 49.)

[Hodge 2004] Victoria J. Hodge et Jim Austin. *A Survey of Outlier Detection Methodologies*. Artificial Intelligence Review, vol. 22, pages 85–126, 2004. 10.1007/s10462-004-4304-y. (Cité en page 13.)

[Hofmeyr 1998] Steven A. Hofmeyr, Stephanie Forrest et Anil Somayaji. *Intrusion Detection using Sequences of System Calls*. Journal of Computer Security, vol. 6, pages 151–180, 1998. (Cité en pages 12 et 13.)

[Hopcroft 2004] J. Hopcroft, O. Khan, B. Kulis et B. Selman. *Tracking evolving communities in large linked networks*. In PNAS, volume 101, pages 5249–5253. National Acad Sciences, 2004. (Cité en page 12.)

[Hopcroft 2006] John E. Hopcroft, Rajeev Motwani et Jeffrey D. Ullman. Introduction to automata theory, languages, and computation (3rd edition). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2006. (Cité en page 14.)

[Houweling 2005] Van Houweling, Douglas et Ted Hanss. Internet2 : The promise of truly advanced broadband," in the broadband explosion. R. Austin and S. Bradley, Editors, Harvard Business School Press, 2005. (Cité en page 47.)

[Huang 2005] Zan Huang, Xin Li et Hsinchun Chen. *Link prediction approach to collaborative filtering*. In Proceedings of the Joint Conference on Digital Libraries (JCDL05). ACM, 2005. (Cité en page 11.)

[Huang 2008] Yiyi Huang, Nick Feamster et Renata Teixeira. *Practical issues with using network tomography for fault diagnosis*. Computer Communication Review 38(5), pages 53–58, 2008. (Cité en page 47.)

[Huffak 2002] Bradley Huffak, Daniel Plummer, David Moore, Claffy et k. *Topology Discovery by Active Probing*. In Proceedings of the 2002 Symposium on Applications and the Internet (SAINT) Workshops, SAINT-W '02, page 90, Washington, DC, USA, 2002. IEEE Computer Society. (Cité en page 35.)

[Jacobson 1989] V. Jacobson. *traceroute*, 1989. The most recent version is available at : `ftp://ftp.ee.lbl.gov/traceroute.tar.gz`. (Cité en page 35.)

[Jensen 2006] April Jensen et Marina Gavrilova. *Normal vs. Abnormal Behavior*. Rapport technique, 2006. (Cité en page 14.)

[Jeong 2001] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai et A.-L. Barabási. *The Large-Scale Organization of Metabolic Networks*. Nature, no. 411, pages 41–42, 2001. (Cité en page 7.)

[Keogh 2002] Eamonn Keogh, Stefano Lonardi et Bill 'Yuan chi' Chiu. *Finding Surprising Patterns in a Time Series Database in Linear Time and Space*. In In In

proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 550–556. ACM Press, 2002. (Cité en page 12.)

[Kleinberg 2000] Jon Kleinberg. *The Small-World Phenomenon : An Algorithmic Perspective*. In in Proceedings of the 32nd ACM Symposium on Theory of Computing, pages 163–170, 2000. (Cité en page 9.)

[Kleinberg 2006] Jon Kleinberg. *Complex Networks and Decentralized Search Algorithms*. In In Proceedings of the International Congress of Mathematicians (ICM, 2006. (Cité en page 9.)

[Kuatse 2007] Amelie Medem Kuatse, Renata Teixeira et Mickael Meulle. *Characterizing network events and their impact on routing*. CoNEXT, page 59, 2007. (Cité en page 47.)

[Kumar 2005] V. Kumar. *Parallel and distributed computing for cybersecurity*. Distributed Systems Online, IEEE, 2005. (Cité en page 13.)

[Kumar 2006] Ravi Kumar, Andrew Tomkins et D Chakrabarti. *Evolutionary clustering*. In In Proc. of the 12th ACM SIGKDD, pages 554–560. ACM Press, 2006. (Cité en page 12.)

[Lad 2006] Mohit Lad, Dan Massey et Lixia Zhang. *Visualizing Internet Routing Changes*. IEEE Transactions on Visualization and Computer Graphics, vol. 12, pages 1450–1460, 2006. (Cité en page 2.)

[Lad 2007] Mohit Lad, Ricardo V. Oliveira, Daniel Massey et Lixia Zhang. *Inferring the Origin of Routing Changes using Link Weights*. ICNP, pages 93–102, 2007. (Cité en page 34.)

[Lakhina 2005] Anukool Lakhina, Mark Crovella et Christophe Diot. *Mining anomalies using traffic feature distributions*. In In ACM SIGCOMM, pages 217–228, 2005. (Cité en pages 12 et 13.)

[Lane 1998] Terran Lane et Carla E. Brodley. *Temporal Sequence Learning and Data Reduction for Anomaly Detection*. ACM Transactions on Information and System Security, vol. 2, pages 150–158, 1998. (Cité en page 12.)

[Latapy 2007] Matthieu Latapy. *Grands graphes de terrain – mesure et métrologie, analyse, modélisation, algorithmique*. UPMC, 2007. (Cité en pages 1 et 10.)

[Latapy 2008] Matthieu Latapy, Clémence Magnien et Frédéric Ouédraogo. *A Radar for the Internet*. In Proceedings of the 2008 IEEE International Conference on Data Mining Workshops, pages 901–908, Washington, DC, USA, 2008. IEEE Computer Society. (Cité en page 2.)

[Latapy 2009] Matthieu Latapy, Clémence Magnien et Frédéric Ouédraogo. *A Radar for the Internet*. In ADN 08 : 1st International Workshop on Analysis of Dynamic Networks — in conjunction with IEEE ICDM 2008, pages 901–908, 2009. Data available at `http://data.complexnetworks.fr/Radar/`. (Cité en pages 34 et 35.)

[Le-Blond 2005] Stevens Le-Blond, Jean-Loup Guillaume et Matthieu Latapy. *Clustering in P2P exchanges and consequences on performances*. In 4th International Workshop on Peer-To-Peer Systems (IPTPS'05), Ithaca, NY États-Unis, 2005. (Cité en page 2.)

[Leicht 2007] E. A. Leicht, G. Clarkson, K. Shedden et M. E.J. Newman. *Large-scale structure of time evolving citation networks*. The European Physical Journal B - Condensed Matter and Complex Systems, vol. 59, pages 99 75–83, 2007. 10.1140/epjb/e2007-00271-7. (Cité en page 2.)

[Li 2004] Lun Li, David Alderson, Walter Willinger et John Doyle. *A first-principles approach to understanding the internet's router-level topology*. In SIGCOMM '04 : Proceedings of the 2004 conference on Applications, technologies, architectures, and protocols for computer communications, pages 3–14, New York, NY, USA, 2004. ACM. (Cité en page 34.)

[Li 2005] Ninghui Li, Wenliang Du et Dan Boneh. *Oblivious signature-based envelope*. Distrib. Comput., vol. 17, pages 293–302, May 2005. (Cité en page 14.)

[Liao 2005] Yihua Liao. *Machine learning in intrusion detection*. PhD thesis, Davis, CA, USA, 2005. AAI3191152. (Cité en page 14.)

[Liben-Nowell 2003] David Liben-Nowell et Jon Kleinberg. *The link prediction problem for social networks*. In Proceedings of the twelfth international conference on Information and knowledge management(CIKM '03), 2003. (Cité en page 11.)

[Lucas 2001] Clay Spence Lucas, Lucas Parra et Paul Sajda. *Detection, Synthesis and Compression in Mammographic Image Analysis with a Hierarchical Image Probability Model*. In In L. Staib (editor), IEEE Workshop on Mathematical Methods in Biomedical Image Analysis. 2001, pages 3–10. IEEE Press, 2001. (Cité en page 13.)

[Lun 2005] Li Lun, David Alderson, John Doyle et Walter Willinger. *Towards a Theory of Scale-Free Graphs : Definition, Properties,and Implications*. Internet Mathematics, vol. 2, no. 4, 2005. (Cité en page 34.)

[Mackey 2001] R. Mackey. *Generalized cross-signal anomaly detection on aircraft hydraulic system*. In Aerospace Conference, 2001, IEEE Proceedings., volume 2, pages 2/657 –2/668 vol.2, 2001. (Cité en page 12.)

[Madhyastha 2009] Harsha V. Madhyastha, Ethan Katz-Bassett, Thomas Anderson, Arvind Krishnamurthy et Arun Venkataramani. *iPlane Nano : path prediction for peer-to-peer applications*. In Proceedings of the 6th USENIX symposium on Networked systems design and implementation, pages 137–152, Berkeley, CA, USA, 2009. USENIX Association. (Cité en page 35.)

[Magnien 2009] Clémence Magnien, Frédéric Ouédraogo, Guillaume Valadon et Matthieu Latapy. *Fast dynamics in Internet topology : preliminary observa-*

*tions and explanations.* CoRR, vol. abs/0904.2716, 2009. (Cité en pages 2 et 34.)

[Magoni 2001] Damien Magoni et Jean Jacques Pansiot. *Analysis of the autonomous system network topology.* SIGCOMM Comput. Commun. Rev., vol. 31, no. 3, pages 26–37, 2001. (Cité en page 34.)

[Mark 2004] Anukool Lakhina Mark, Mark Crovella et Christophe Diot. *Characterization of Network-Wide Anomalies in Traffic Flows.* In In ACM/SIGCOMM IMC, pages 201–206, 2004. (Cité en page 12.)

[Martinez 2003] N. Martinez. *Simple Rules Yield Complex Food Webs.* APS Meeting Abstracts, 2003. (Cité en page 7.)

[Newman 2003] M. E. J. Newman. *The Structure and Function of Complex Networks.* SIAM Review, vol. 45, no. 2, pages 167–256, 2003. (Cité en pages 6 et 8.)

[Oliveira 2007] Ricardo V. Oliveira, Beichuan Zhang et Lixia Zhang. *Observing the evolution of internet as topology.* In Proceedings of the 2007 conference on Applications, technologies, architectures, and protocols for computer communications, SIGCOMM '07, pages 313–324, New York, NY, USA, 2007. ACM. (Cité en pages 2 et 34.)

[O'Madadhain 2005] Joshua O'Madadhain, Jon Hutchins et Padhraic Smyth. *Prediction and ranking algorithms for event-based network data.* ACM SIGKDD Explorations Newsletter, 2005. (Cité en page 11.)

[Palla 2005] Gergely Palla, Imre Derényi, Illés Farkas et Tamás Vicsek. *Uncovering the overlapping community structure of complex networks in nature and society.* Nature, vol. 435, no. 7043, pages 814–818, Juin 2005. (Cité en page 7.)

[Palla 2007] Gergely Palla, Albert-Laszlo Barabasi et Tamas Vicsek. *Quantifying social group evolution.* Nature, vol. 446, pages 664–667, 2007. (Cité en page 12.)

[Pansiot 2007] Pansiot. *Local and dynamic analysis of Internet multicast router topology Analyse locale et dynamique de la topologie des routeurs multicast d'Internet.* In Annales des télécommunications, pages 62 :408–425,, 2007. (Cité en page 2.)

[Park 2004] Park, David M. Pennock et C. Lee Giles. *Comparing static and dynamic measurements and models of the internet's topology.* In In IEEE INFOCOM, 2004. (Cité en page 2.)

[Paxson 1999] Vern Paxson. *Bro : a System for Detecting Network Intruders in Real-time.* Computer Networks (Amsterdam, Netherlands : 1999), vol. 31, no. 23–24, pages 2435–2463, 1999. (Cité en page 14.)

[Pearson 2005] Ronald K. Pearson. Mining imperfect data : Dealing with contamination and incomplete records. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2005. (Cité en page 39.)

[Phoha 2002] Vir V. Phoha. Internet security dictionary. The Springer, 2002. (Cité en page 12.)

[Pineda 2002] Luis Alberto Pineda, Antonio Massé Márquez, Ivan Meza, Miguel Salas Zúñiga, Eric Schwarz, Esmeralda Uraga et Luis Villaseñor Pineda. *The DIME Project*. In Proceedings of the Second Mexican International Conference on Artificial Intelligence : Advances in Artificial Intelligence, MICAI '02, pages 166–175, London, UK, UK, 2002. Springer-Verlag. (Cité en page 35.)

[Pokrajac 2007] Dragoljub Pokrajac. *Incremental local outlier detection for data streams*. In In Proceedings of IEEE Symposium on Computational Intelligence and Data Mining, pages 504–515, 2007. (Cité en page 12.)

[Press 1992] William H. Press, Saul A. Teukolsky, William T. Vetterling et B. P. Flannery. Numerical recipes : The art of scientific computing. Cambridge University Press, Cambridge, England, 1992. (Cité en page 29.)

[Resende 2000] Mauricio G. C. Resende. *Detecting dense subgraphs in massive graphs*. In 17th international Symposium on Mathematical Programming, 2000. (Cité en page 7.)

[Roughan 2004] Matthew Roughan, Tim Griffin, Morley Mao, Albert Greenberg et Brian Freeman. *Combining Routing and Traffic Data for Detection of IP Forwarding Anomalies*. In In ACM SIGCOMM NeTs Workshop, 2004. (Cité en page 12.)

[Rousseeuw 1987] P. J. Rousseeuw et A. M. Leroy. Robust regression and outlier detection. John Wiley & Sons, Inc., New York, NY, USA, 1987. (Cité en page 13.)

[Salah-Ibrahim 2010] A. Salah-Ibrahim, B. Le Grand et M. Latapy. *Some Insight on Dynamics of Posts and Citations in Different Blog Communities*. In Communications Workshops (ICC), 2010 IEEE International Conference on, pages 1 –6, Mai 2010. (Cité en page 7.)

[Salem 2010] Osman Salem, Sandrine Vaton et Annie Gravey. *A scalable, efficient and informative approach for anomaly-based intrusion detection systems : theory and practice*. Int. J. Netw. Manag., vol. 20, pages 271–293, 2010. (Cité en page 13.)

[Satorras 2001] Romualdo Satorras et Alessandro Vespignani. *Epidemic Spreading in Scale-Free Networks*. PHYSICAL REVIEW LETTERS, vol. 86, no. 14, 2001. (Cité en page 7.)

[Scherrer 2008] Fleury Eric Guillaume Jean-Loup Robardet Céline Scherrer Borgnat Pierre. *Description and simulation of dynamic mobility networks*. Computer Networks, page 2842–2858, 2008. (Cité en page 2.)

[Schneider 2009] Fabian Schneider, Anja Feldmann, Balachander Krishnamurthy et Walter Willinger. *Understanding online social network usage from a network*

*perspective*. In Anja Feldmann et Laurent Mathy, editeurs, Internet Measurement Conference, pages 35–48. ACM, 2009. (Cité en page 2.)

[Schölkopf 1999] Bernhard Schölkopf, John C. Platt, John S. Taylor, Alexander J. Smola et Robert C. Williamson. *Estimating the Support of a High-Dimensional Distribution*. Rapport technique MSR-TR-99-87, 1999. (Cité en page 14.)

[Shafi 2009] Kamran Shafi. Online and adaptive signature learning for intrusion detection. VDM Verlag, Saarbrücken, Germany, Germany, 2009. (Cité en page 14.)

[Song 2007] Xiaodan Song, Yun Chi, Belle L. Tseng, D. Zhou et K. Hino. *Evolutionary spectral clustering by incorporating temporal smoothness*. In Proc. of the 13th ACM SIGKDD, pages 153–162. ACM, 2007. (Cité en page 12.)

[Steuer 2007] Ralf Steuer, Adriano Nunes Nesi, Alisdair R. Fernie, Thilo Gross, Bernd Blasius et Joachim Selbig. *From structure to dynamics of metabolic pathways : application to the plant mitochondrial TCA cycle*. Bioinformatics, vol. 23, no. 11, pages 1378–1385, 2007. (Cité en page 2.)

[Stoica 2009] Alina Stoica et Christophe Prieur. *Structure of Neighborhoods in a Large Social Network*. In Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04, pages 26–33, Washington, DC, USA, 2009. IEEE Computer Society. (Cité en page 2.)

[Strogatz 2001] Steven H. Strogatz. *Exploring complex networks*. Nature, vol. 410, no. 6825, pages 268–276, Mars 2001. (Cité en pages 6 et 8.)

[Stutzbach 2005] Daniel Stutzbach et Reza Rejaie. *Characterizing unstructured overlay topologies in modern p2p file-sharing systems*. In In Internet Measurement Conference, pages 49–62, 2005. (Cité en page 11.)

[Tam 2009] Wai M. Tam, Francis C. M. Lau et Chi K. Tse. *Complex-network modeling of a call network*. Trans. Cir. Sys. Part I, vol. 56, pages 416–429, February 2009. (Cité en page 7.)

[Tarissan 2009] Fabien Tarissan, Matthieu Latapy et Christophe Prieur. *Efficient Measurement of Complex Networks Using Link Queries*. CoRR, vol. abs/0904.3222, 2009. (Cité en page 2.)

[Thode 2002] Henry C. Jr. Thode. Identification of outliers. CRC Press, 2002. (Cité en page 29.)

[Torr 1995] P. H. S. Torr et D. W. Murray. *Outlier Detection and Motion Segmentation*. pages 432–443, 1995. (Cité en page 12.)

[Tournoux 2009] P. Tournoux, Jérémie Leguay, Farid Benbadis, Vania Conan, Marcelo Dias De Amorim et John Whitbeck. *The Accordion Phenomenon : Analysis, Characterization, and Impact on DTN Routing*. page 1116–1124. In Proceedings of the 28rd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM), 2009. (Cité en page 2.)

[Travers 1969] Jeffrey Travers et Stanley Milgram. *An Experimental Study of the Small World Problem*. Sociometry, vol. 32, no. 4, pages 425–443, 1969. (Cité en page 9.)

[Travers 2000] Jeffrey Travers et Stanley Milgram. *Navigation in a small world*. Nature, vol. 406, no. 6798, page 845, Aot 2000. (Cité en page 9.)

[Viger 2008] Fabien Viger, Brice Augustin, Xavier Cuvellier, Clémence Magnien, Matthieu Latapy, Timur Friedman et Renata Teixeira. *Detection, understanding, and prevention of traceroute measurement artifacts*. Comput. Netw., vol. 52, pages 998–1018, April 2008. (Cité en page 35.)

[Wasserman 1994] S. Wasserman et K. Faust. Social network analysis : Methods and applications. Cambridge Univ Pr, 1994. (Cité en pages 1 et 6.)

[Wasserman 2005] Larry Wasserman. All of statistics : A concise course in statistical inference. Springer, 2005. (Cité en page 23.)

[Watts 1998] D. J. Watts et S. H. Strogatz. *Collective Dynamics of 'Small-World' Networks*. Nature, vol. 393, no. 6684, pages 440–442, 1998. (Cité en pages 1 et 8.)

[Webster 2008] Matt Webster et Grant Malcolm. *Formal affordance-based models of computer virus reproduction*. Journal in Computer Virology, vol. 4, no. 4, pages 289–306, 2008. (Cité en page 14.)

## Event detection in the dynamics of complex networks : a statistical approach and its application to the internet radar

This work addresses the problem of event detection in the dynamics of complex networks, defined as the ability to point out specific changes in systems that do not conform to the *expected behavior*. The main contribution of this thesis is the proposal and implementation of a generic approach to automatically and rigorously detect events in the dynamics of complex networks.

According to the principle of our approach, characterizing an event needs to identify characteristics of the dynamic graph whose distribution is normal with *outliers*. To apply this notion of *statistically significant* event for event detection in the dynamics of real networks, we propose a set of dynamic properties of graphs.

To explore the specific contribution of each property, we study the correlations between events detected by each one. In a similar way, and to better identify and understand the impact of events detected on real networks, we have complemented our event detection method with two interpretation approaches : the correlation with known events, and visualization.

To demonstrate the effectiveness of our empirical and generic approach of event detection in complex networks, we apply it to the internet radar, composed of ego-centered and periodic measurements of the internet topology.

**Keywords : event detection, complex network, statistical approach, internet, measurement.**

## Détection d'événements dans la dynamique des grands graphes de terrain : une approche statistique et son application au radar de l'internet

Ce travail traite de la problématique de la détection d'événements dans la dynamique des graphes de terrain, définie comme la capacité à pointer des modifications particulières systèmes qui ne sont pas conformes au *comportement attendu*. La contribution principale de cette thèse réside dans la proposition et la mise en œuvre d'une approche générique pour détecter, automatiquement et rigoureusement, des événements dans les dynamiques de graphes de terrain.

Selon le principe de notre approche, caractériser un événement dans la dynamique d'un graphe de terrain nécessite d'identifier des propriétés de la dynamique du graphe dont la distribution est normale avec *outliers*. Afin d'appliquer cette notion d'événements *statistiquement significatifs* à la détection des événements dans les dynamiques des graphes de terrain, nous avons proposé un ensemble de propriétés de graphes dynamiques.

Afin d'explorer l'apport spécifique de chaque propriété, nous avons étudié les corrélations entre les événements détectés par chacune d'elle. Dans la même optique, et afin de mieux les cerner et comprendre l'impact des événements détectés sur les graphes de terrain concernés, nous avons complémenté notre méthode de détection d'événements avec deux approches d'interprétation : la corrélation avec des événements connus, et la visualisation.

Pour démontrer l'efficacité de notre approche empirique et générique de détection d'événements dans les graphes de terrain, nous l'avons appliqué au radar de l'internet, c'est-à-dire l'observation égo-centrée et periodique de la topologie de l'internet.

**Mots clés : détection d'événements, graphe de terrain, approche statistique, internet, mesure.**