– sujet de thèse –

Robustesse et Fragilité des Infrastructures en Réseaux

Matthieu Latapy

LIP6 – CNRS et Sorbonne Université

Matthieu.Latapy@lip6.fr

1. Titre ou intitulé de la thèse

Robustesse et Fragilité des Infrastructures en Réseaux

2. Objet de la thèse

Des infrastructures clés, comme l'internet ou les réseaux de transports, subissent divers types de pannes et d'attaques. Estimer leur résistance à ces événements nécessite le développement de méthodes et outils dédiés, aptes à modéliser la structure du réseau aussi bien que sa charge, et leurs dynamiques couplées. Cette thèse propose une approche basée sur les données de terrain et la science des réseaux pour répondre à ce besoin.

3. Descriptif de la thèse

De nombreuses infrastructures d'importance stratégique sont structurées en réseaux ; par exemple les réseaux de distribution (d'eau, de gaz, d'électricité), de transport (terrestres, maritimes, aériens), ou de communication (téléphone, internet, pair-à-pair). Ces réseaux évoluent au cours du temps : de nouveaux éléments (nœuds ou liens) sont ajoutés, d'autres sont supprimés, des pannes peuvent advenir. De plus, ces infrastructures acheminent des ressources (énergie, marchandises, individus, informations). A tout instant, **le réseau est donc occupé par une certaine charge** pour chacun de ses éléments, et cette charge est en constante évolution. Il peut s'agir de la quantité de ressources ou d'individus présents sur chaque élément du réseau, par exemple.

La structure du réseau et sa charge sont couplées. Par exemple, les trajectoires des ressources ou individus dépendent fortement de la structure du réseau : ils suivent des chemins de coût minimal (les plus courts, les plus rapides, ou les moins chers, typiquement). Réciproquement, une forte charge ou une surcharge de certains éléments du réseau vont les rendre temporairement inopérants, et les trajectoires devront les éviter. Elles seront alors susceptibles de créer des surcharges à d'autres endroits du réseau, pendant que la charge des éléments qu'elles auront évité, elle, diminuera. L'exemple le plus intuitif est probablement le trafic routier et les bouchons, mais des phénomènes similaires se retrouvent dans la plupart des infrastructures en réseaux.

Dans cette optique, **l'efficacité et la robustesse sont deux caractéristiques essentielles d'un réseau** : connecte-t-il bien l'ensemble des éléments du réseau ? permet-il des trajectoires majoritairement ou toujours efficaces ? souffre-t-il de congestions, de ruptures de connectivité ? peut-il facilement et rapidement s'adapter à des changements de sa structure et/ou de sa charge ? son efficacité est-elle facilement dégradée lorsque le réseau ou sa charge sont modifiés ? Par exemple, si un réseau est très efficace mais que toutes ses trajectoires passent par un même élément, alors une panne, une surcharge, ou une attaque de cet élément peuvent être dévastatrices ; le réseau est extrêmement fragile, et c'est certainement une information aussi importante que son efficacité.

On sait par exemple que, si on modélise les pannes par des retraits aléatoires de nœuds ou de liens, et si on mesure l'efficacité du réseau par le nombre de nœuds qu'il interconnecte, alors les réseaux d'infrastructures sont généralement très robustes : l'efficacité reste grande jusqu'à un nombre de pannes énormes. Si par contre on suppose que l'infrastructure subit une attaque visant les nœuds de forts degrés (c'est-à-dire ayant beaucoup de liens), alors on sait que ces réseaux s'avèrent très fragiles. Cette propriété a même été appelée le « talon d'Achille » des réseaux de ce type. Notre référence 4 ci-dessous présente une synthèse approfondie des connaissances sur ce sujet.

L'état-de-l'art reste toutefois largement insuffisant : les réseaux sont souvent représentés par de simples graphes, les suppressions visent essentiellement les forts degrés, l'efficacité est mesurée par le nombre de nœuds interconnectés, et la charge du réseau n'est pas prise en compte. Ces modélisations apportent des éclairages fondamentaux importants (par exemple le rôle crucial de la distribution de degrés). Mais ils manquent de réalisme lorsqu'on veut aller plus loin que les grands principes, comme par exemple identifier les vulnérabilités d'une infrastructure réelle.

L'objectif de cette thèse est de combler ce manque. Nous voulons concevoir des **modélisations enrichies** des réseaux d'infrastructures, capables de coder des informations clés comme la capacité des nœuds et/ou des liens. Nous voulons tirer parti des données très riches aujourd'hui disponible pour modéliser ainsi des **infrastructures réelles**, notamment les réseaux routiers et l'internet. Nous voulons ensuite définir des métriques avancées pour **caractériser l'efficacité et la robustesse** des réseaux d'infrastructures. L'objectif sera de mettre en évidence leurs points faibles, sous forme d'ensembles d'éléments dont la surcharge ou la destruction aurait un impact majeur. Ces travaux devront **prendre en compte l'activité sur le réseau**, en modélisant de façon dynamique la charge de chaque nœud et chaque lien. Nous voulons en particulier que ces modèles capturent des rétro-actions et des phénomènes en cascades dans lesquels, par exemple, une surcharge induit une redistribution de la charge, qui engendre de nouvelles surcharges, et ainsi de suite. Les avancées récentes en science des réseaux, en particulier nos références 2 et 3 ci-dessous, fournissent les formalismes nécessaires.

4. Programme de la thèse

Les objectifs explicités ci-dessus soulèvent des **problématiques relevant de science des données, science des réseaux, et algorithmique**. En effet, nous reposons sur des données massives, hétérogènes, peu structurées, et imparfaites. Ces données sont modélisées comme des réseaux complexes, dont il s'agit d'analyser la structure et la dynamique, y compris en termes de charge. La taille des graphes sous-jacents, et les métriques envisagées, rendent les calculs souvent coûteux ; leur complexité doit alors être réduite, notamment par des approximations maîtrisées. Enfin, les résultats obtenus sont complexes car ils fournissent des éclairages multi-critères (par exemple, efficacité en termes de rapidité ou de coût), faisant à nouveau appel aux sciences des données pour leur interprétation.

Afin de répondre à ces besoins, nous identifions une **séquence de travaux à réaliser**, que nous présentons maintenant.

1. Collecte et prétraitement des données.

Afin d'assurer un réalisme maximal à nos travaux, nous utiliserons intensivement des données de **cartographies d'infrastructures à grandes échelles**. Nous utiliserons notamment OpenStreetMaps pour les cartographies routières à l'échelle de villes ou de pays. Nous utiliserons également les cartographies de l'internet au niveau AS (c'est-à-dire les systèmes autonomes qui le constituent, observables via les tables de routage BGP), et au niveau IP (c'est-à-dire au niveau des adresses des routeurs et des sauts réseau visibles typiquement par l'outil *traceroute* et ses variantes plus fiables). Nous explorerons également la disponibilité de données concernant d'autres types d'infrastructures, comme les aéroports par exemple, ou les réseaux de distribution d'eau ou d'électricité.
Les cartographies de l'internet et les cartographies OpenStreetMaps sont immédiatement disponibles, mais elles nécessitent un **prétraitement approfondi** pour pouvoir être utilisées et modélisées sous forme de graphes valués et étiquetés. Nous devons en particulier les renseigner avec la capacité et le coût de chaque nœud et lien (par exemple, la largeur et la longueur des routes, la taille des carrefours, la capacité des routeurs ou la bande passante des liens). Ces informations sont partiellement disponibles dans les cartographies, mais dans certains cas nous devrons les approximer par des mesures ou des heuristiques, comme les temps d'aller retour des paquets sur internet, par exemple.
Afin d'avoir une modélisation complète, nous avons besoin également d'information sur la **charge de l'infrastructure et ses variations** au cours du temps. Ce type de données est beaucoup plus rare et difficile à obtenir, mais de plus en plus de sources permettent de les approximer. Par exemple, des traces GPS de déplacements individuels sont disponibles via OpenStreetMaps, et plusieurs grandes villes dans le monde fournissent des mesures de trafic sur leurs voies par exemple via les déplacements en taxi.
Soulignons que la cartographie détaillée et précise d'infrastructures, avec leur charge réelle, est un défi en soi. Il n'est pas au cœur de cette thèse, qui se limitera donc aux données disponibles et à des approximations simples. Celles-ci pourront être améliorées par ailleurs.

2. Métriques de graphes pour la robustesse.

La science des réseaux a défini une **multitude de métriques** permettant de décrire un graphe et ses caractéristiques principales, comme par exemple la distribution de degrés, la densité locale (coefficient de *clustering*), les structures de communautés hiérarchiques, diverses notions de centralité basées sur les plus courts chemins (*closeness*/eccentricité, *betweenness*, diamètre, etc), des mesures basées sur les marches aléatoires (*pagerank*, centralité de *katz*, etc), des mesures spectrales basées sur une vision matricielles, etc. La plupart de ces mesures caractérisent les nœuds, mais certaines s'appliquent aux liens (*betweenness*, et coefficient de Jaccard par exemple) ; on a également recours aux *line graphs*, qui permettent d'interchanger nœuds et liens.

La plupart de ces métriques, sinon toutes, peuvent en principe donner des éclairages sur la robustesse du réseau. On sait par exemple que les nœuds de fort degré jouent un rôle important, et au contraire qu'une forte densité locale indique des liens redondants. D'autres métriques ont également été utilisées dans ce contexte, en particulier les métriques de centralité. Nous devrons ici dresser un **panorama complet de ces métriques et de leur intérêt pour la robustesse**. Nous les calculerons sur les graphes modélisant les infrastructures et observerons leur pertinence en termes de description de ces graphes, d'une part, et en termes de stratégies pour identifier des points faibles, d'autre part. Outre leur pertinence descriptive, nous prendrons en compte leur coût de calcul, qui peut être prohibitif pour certains (par exemple la *betweenness*). Nous étudierons alors les possibilités connues d'approximer ces métriques.

3. Métriques de graphes pour l'efficacité.

En parallèle des métriques décrivant le graphe lui-même et ses composantes (nœuds et liens), nous avons besoin de métriques permettant d'évaluer l'efficacité d'un graphe, c'est-à-dire à quel point il répond bien au besoin. **La méthode classique consiste à compter le nombre de nœuds reliés par le réseau** : s'il y a possibilité d'acheminer des ressources de la plupart des nœuds à la plupart des nœuds, alors on considère que le réseau est opérant. Algorithmiquement, ceci revient à calculer la plus grande composante connexe, ce qui se fait très efficacement (temps et espace linéaires). Cette métrique a aussi l'avantage de la simplicité, mais il est bien clair qu'elle mesure l'efficacité du réseau de façon très insatisfaisante. Par exemple, elle ne tient aucun compte de la **distance** entre les nœuds dans le réseau ; il serait pourtant naturel de mesurer l'efficacité par la distance moyenne, ou par le diamètre (la plus grande distance), par exemple. Toutefois, calculer ces métriques a un coût prohibitif (temps quadratique) sur de grands graphes, et elles doivent être approximées. Nous fournissons par exemple un algorithme donnant un encadrement du diamètre dans la référence 6 ci-dessous.
Dans le cadre de ce projet, nous avons besoin de **quantifier l'efficacité des réseaux de façon bien plus fine que par la connexité**. Nous ferons l'état-de-l'art des métriques considérées dans la littérature, et le compléterons de métriques plus avancées, basées sur les distances en tenant compte des capacités ou des coûts des nœuds et liens. Nous prendrons également en compte des caractéristiques plus fines mais cruciales, comme la bonne répartition des trajectoires sur le réseau (afin d'évaluer les risques de congestion), et les corrélations entre les différentes métriques. Nous veillerons à la pertinence pratique de ces métriques, en focalisant sur les calculs en temps et espace linéaires ou quasi-linéaires, ou approximables avec ces complexités de façon maîtrisée. Nous obtiendrons finalement une librairie de concepts et d'algorithmes implémentés, permettant une description complète de l'efficacité d'un réseau donné sous plusieurs angles complémentaires.

4. Couplage entre la structure et la charge.

Les travaux ci-dessus focalisent sur l'analyse de réseaux sans prise en compte d'une charge explicite. Or cette charge elle-même a un impact sur la structure et l'efficacité de l'infrastructure : elle peut ralentir ou rendre ineffectives des trajectoires surchargées, par exemple. **Etudier la robustesse et la fragilité des infrastructures en réseaux doit prendre en compte ce couplage**.
Nous proposons en un premier temps de modéliser les trajectoires de façon simple, comme des plus courts chemins (ou chemins de moindre coût) ou des marches aléatoires biaisées, par exemple. Mais **nous intégrerons dans les calculs de ces trajectoires les capacités et la charge** des différents éléments du réseau. En particulier, un nœud ou un lien ne pourra être employé que par un nombre maximal de trajectoires, correspondant à sa capacité. En augmentant progressivement le nombre de trajectoires supposées, nous serons donc en mesure de créer une dynamique dans les choix de trajectoires, et d'identifier des points de saturation. Limiter ainsi le nombre de trajectoires empruntant les mêmes éléments du réseau forcera une répartition des trajectoires dans le réseau. Nous évaluerons ainsi la capacité du réseau à assurer cette répartition, en lien avec les métriques d'efficacité ci-dessus.

Dans cette optique, nous définirons de **nouvelles métriques pour les réseaux d'infrastructures *avec charge*** : nous étendrons les concepts de centralité (notamment la *betweenness*) et ceux basés sur les marches aléatoires (notamment le *pagerank*) pour tenir compte de cette charge. Ceci renouvellera les métriques de graphes pour la robustesse et pour l'efficacité définies ci-dessus, et nous obtiendrons un *framework* complet pour l'étude des réseaux avec charge.

5. Modélisation complètement dynamique.

La prise en compte de la charge et de sa dynamique est **naturellement modélisée par des *stream graphs* valués**, modèle de réseaux temporels introduits par l'équipe, voir références 2 et 3 ci-dessous. Ce modèle permet de capturer complètement la dynamique de la charge sur les nœuds aussi bien que sur les liens, ainsi que les ajouts et/ou suppressions de nœuds et/ou de liens. Dans ce formalisme, les trajectoires partent d'un nœud *à un certain moment* et le déroulé temporel de la trajectoire est entièrement intégré. La charge de chaque nœud et chaque lien est alors connue à chaque instant. Nous intégrerons donc tous les concepts ci-dessus dans cette modélisation afin de capturer complètement la dynamique de l'infrastructure.
Nous définirons et calculerons alors des **métriques sur les *stream graphs* aptes à mieux capturer les caractéristiques de l'infrastructure**, notamment en termes de robustesse et d'efficacité. Par exemple, nous quantifierons la connectivité offerte par l'infrastructure en termes de nombre de nœuds accessibles à un instant d'arrivée donné à partir d'un nœud pour un temps de départ donné. Nous étudierons les chemins les plus rapides, les chemins arrivant au plus tôt, et les chemins de coût minimal. Toutes ces notions de chemins optimaux sont définies dans les *stream graphs*, mais n'ont jamais été exploitées pour l'étude de la robustesse. Elles permettront ici de capturer finement la dynamique de la charge et les potentiels effets en cascades de la dynamique.

6. Modélisation avancée de la charge.

Forts de la modélisation complètement dynamique et des métriques de *stream graphs* ci-dessus, nous serons outillés pour une modélisation réaliste de la charge. Dans la littérature et dans les travaux ci-dessus, elle est généralement estimée par des modèles de trajectoires simples, à base de chemins optimaux et/ou de marches aléatoires. **La réalité est toutefois bien plus complexe**, surtout quand on prend en compte la dimension temporelle. Il est par exemple bien évident que les infrastructures sont souvent soumises à des périodicités dans leur charge, ou à des pics d'activité comme des heures de pointe.
Nous nous appuierons sur les **données ouvertes pour une meilleure modélisation** de tels effets. Ces données sont de plusieurs types. D'abord, comme explicité ci-dessus, certaines sources fournissent directement une estimation de la charge au cours du temps ; une telle information pourra être intégrée sous forme d'un poids sur les nœuds et liens temporels dans notre modélisation en stream graphs. Une autre approche repose sur la modélisation plus réaliste des trajectoires : les travaux en science des transports, par exemple, ou les travaux en réseaux informatique, fournissent des modèles des déplacements et du routage ; nous les utiliserons pour estimer nos métriques de robustesse et d'efficacité avec des trajectoires plus réalistes. Enfin, nous disposons également de données de mobilité, typiquement sous la forme de trace GPS, de mesures de matrices origine-destination, des captures de trafic réseau, ou de suivi de paquets sur internet. Ces données fournissent des indications précieuses sur la charge des infrastructures au cours du temps, que nous intégrerons à nos modèles.

7. Scénarios de pannes ou d'attaques.

Afin d'évaluer plus précisément la robustesse et la fragilité des infrastructures en réseau considérées, nous voulons finalement les soumettre à différents scénarios de pannes ou d'attaques. Nous nous baserons tout d'abord sur la **recherche de pires cas** : quel est le nombre minimal de nœuds et/ou de liens à supprimer (c'est-à-dire à surcharger ou à détruire) pour obtenir un effet escompté (par exemple,

la saturation globale de l'infrastructure) ? Ce type de problématique est déjà étudié dans la littérature de façon théorique, et trouver la réponse optimale est généralement hors de portée. Des heuristiques permettent toutefois de trouver des solutions approchant l'optimal, et nous les appliquerons à nos réseaux d'infrastructures.
Nous nous inspirerons ensuite d'attaques réelles pour définir des scénarios plus réalistes. En particulier, plusieurs attaques visent à **saturer certains éléments du réseau y faisant converger un grand nombre de trajectoires.** C'est pas exemple le cas des attaques de déni de service distribués, dans les réseaux informatiques : un ensemble de machines attaquantes envoient conjointement une masse de trafic vers la cible. C'est aussi le cas des *convois de la liberté*, mouvement social récent qui a consisté à bloquer les transports en faisant converger des camions vers de grandes villes cibles. De telles attaques soumettent l'infrastructure à une charge anormale (en quantité mais surtout en structure et en temporalité) ; nous évaluerons leur potentiel de perturbation avec les méthodes développées ci-dessus.
D'autres attaques reposent sur des **blocages temporaires d'éléments du réseau**, comme par exemple des manifestations (les actions de type *rebellion of one* lancées par des activistes écologistes consistent à s'asseoir au milieu de certaines rues, typiquement), ou la destruction physique ou virtuelle de matériels (câbles sectionnés, par exemple, ou exploitation de failles de sécurité permettant de désactiver des matériels). Ceci correspond à la suppression temporaire de nœuds et/ou liens dans notre modélisation en *stream graphs*, avec un impact sur les trajectoires et donc la charge du réseau. La réparation des matériels ou le rétablissement de la circulation par les forces de l'ordre peuvent être modélisés par un retour des nœuds et/ou des liens. Toutes ces dynamiques seront prises en compte grâce aux modèles et métriques de la thèse.

La **première année** de thèse sera consacrée essentiellement aux travaux des tâches 1 à 3, l'objectif étant d'obtenir une vision consolidée de l'ensemble de l'état-de-l'art ainsi qu'une librairie de concepts et d'outils sur la robustesse dans les infrastructures en réseaux, modélisées par des graphes.
La **seconde année** sera dédiée aux tâches 4 à 6, afin d'obtenir un cadre d'analyse réaliste et complètement dynamique, reposant sur les *stream graphs*, de l'infrastructure et de sa charge au cours du temps.

Outre la rédaction du mémoire et la valorisation des résultats, la **troisième année** sera consacrée à l'étude approfondie de scénarios de pannes et d'attaques de la tâche 7, à un niveau de réalisme sans précédent.

Au final, nous comptons obtenir **un ensemble d'avancées méthodologiques et d'outils permettant d'étudier en profondeur la robustesse et la fragilité de multiples infrastructures en réseaux**. Ces avancées permettront de mieux comprendre ces aspects, et notamment d'évaluer la capacité de divers événements, dont diverses stratégies d'attaques, à perturber le système.

5. Références

Les six références sélectionnées ci-dessous (ordre chronologique inverse) illustrent l'activité de l'équipe en lien avec le sujet.

La **référence 4**, dont la version complète disponible en ligne fait 74 pages, est ici la plus importante : elle traite de la robustesse des réseaux en cas de pannes ou d'attaques (ici, les réseaux sont modélisés par des graphes, et les pannes ou attaques par des suppressions aléatoires ou ciblées de nœuds et/ou liens). On y propose une synthèse exhaustive de l'état-de-l'art sur le sujet, qu'on unifie dans un formalisme cohérent. Ceci permet d'en délimiter les contours et de le compléter par la compréhension de plusieurs subtilités essentielles, notamment le rôle des liens par rapport aux nœuds.

Les **références 2 et 3 sont** importantes également car elles présentent nos travaux sur la modélisation de graphes à forte dynamique, y compris valués, bipartis, ou orientés, que nous comptons mobiliser dans ce projet. Nous sommes pionniers dans ce domaine, avec le formalisme des flots de liens ou *stream graphs*, que nous avons développé. Il permet une prise en compte complète de la dynamique des liens et des nœuds, dans un cadre unifiant graphes et séries temporelles.

La **référence 5** montre notre expertise en mesure et analyse de réseaux d'infrastructures, avec ici des mesures égo-centrées à haute fréquence de l'internet, afin de détecter des anomalies dans sa dynamique. Nous avons conçu et implémenté l'outil et le protocole de mesure, mené la collecte à une échelle sans précédent à partir de plusieurs centaines de machines, fourni le code et les données publiquement, et analysé les dynamiques observées.

Enfin, les **références 1 et 6** concernent le calcul (éventuellement approximé) efficace de métriques essentielles pour modéliser l'efficacité et la robustesse des réseaux. La première concerne la *betweenness* dans les réseaux à forte dynamique, qui est un critère clé pour identifier des points faibles. L'autre concerne le calcul très rapide d'encadrements resserrés du diamètre d'un graphe, une métrique cruciale pour l'efficacité.

1. *Computing Betweenness Centrality in Link Streams.* Frédéric Simard, Clémence Magnien, Matthieu Latapy. Submitted. 2021.

2. *Weighted, Bipartite, or Directed Stream Graphs for the Modeling of Temporal Networks.* Matthieu Latapy, Clémence Magnien, Tiphaine Viard. Springer, 2019.

3. *Stream graphs and link streams for the modeling of interactions over time.* Matthieu Latapy, Tiphaine Viard, Clémence Magnien. Social Network Analysis and Mining (SNAM), 2018.

4. *Impact of random failures and attacks on poisson and power-law random networks.* Clémence Magnien, Matthieu Latapy, Jean-Loup Guillaume. ACM Computing Surveys, 2011.

5. *A radar for the internet. Matthieu Latapy, Clémence Magnien, Frédéric Ouédraogo.* Complex Systems, 2011.

6. *Fast computation of empirically tight bounds for the diameter of massive graphs.* Clémence Magnien, Matthieu Latapy, Michel Habib. ACM Journal of Experimental Algorithmics (JEA) 13, 2009.