Graphe de contacts et ondelettes : étude d'une diffusion bactérienne

Benjamin GIRAULT[1], Paulo GONÇALVES[1], Éric FLEURY[1]

[1]École Normale Supérieure de Lyon – Inria – CNRS : UMR5668 – Université de Lyon – UCBL Lyon I
Laboratoire de l'Informatique et du Parallélisme – Équipe DANTE
{benjamin.girault, paulo.goncalves, eric.fleury}@ens-lyon.fr

Résumé – Cet article présente la série temporelle des contacts entre personnes d'un établissement hospitalier ainsi que la propagation épidémiologique de différentes souches bactériennes. Ce jeu de données expérimentales collecté sur un système à grande échelle et sur une longue durée offre un cadre exceptionnel pour l'analyse des fonctions sur graphe, avec pour objectif à terme, de mettre en lumière des liens de causalité entre dynamiques structurelles des réseaux et modes de diffusions bactériennes. À titre d'illustration, nous comparons l'évolution temporelle d'une souche à partir d'un foyer localisé, avec les réponses impulsionnelles d'ondelettes sur le graphe des contacts, à différentes échelles spatiales.

Abstract – This article presents the time series of contacts between individuals of an hospital together with measures of the epidemiological spread of different bacterial strains. This experimental dataset collected on a large scale system, over a long period of time provides an outstanding framework for the analysis of functions defined on graphs. Long term objective of this work is to emphasise relations of causality between structural network dynamics and diffusion modes of bacteria. As a mere illustration, we compare the time spreading of a strain from a localised source with the impulse response of a wavelet on the graph of contacts, at different spatial scales.

1 Motivations

Les paramètres de la diffusion et des mutations des souches bactériennes nosocomiales sont aujourd'hui encore mal compris. Les mécanismes macroscopiques en jeu lors de la diffusion sont à opposer à des mécanismes microscopiques qui sont eux bien connus et compris. Le passage à l'échelle d'un hôpital conduit alors à l'étude d'un système complexe qui doit être simplifié, modélisé, avant d'en faire une étude épidémiologique.

Ce travail entre dans le cadre d'un travail plus large qui entend donner un début de réponse en étudiant une corrélation entre le réseau (dynamique) de contacts et la diffusion microbiologique. Pour cela, dans le cadre du projet MOSAR (*Mastering hOSpital Antimicrobial Resistance*) au sein du groupe i-Bird (*Individual Based Investigation of Resistance Dissemination*), un réseau de capteurs sans fil fut déployé pendant plusieurs mois sur le site de l'Hôpital Maritime de Berck-sur-Mer conjointement à des analyses microbiologiques.

Des études en cours montrent que le réseau des contacts comporte plusieurs échelles spatio-temporelles. Par exemple, le réseau présente des pseudo-périodicités journalière et hebdomadaire. Spatialement, les contacts sont organisés par service, avec une affinité plus forte pour les contacts intra-service, intra-batiment, ou intra-étage, mais également par catégorie socio-professionnelle pour les personnels. Ce graphe des contacts est un graphe de terrain présentant de multiples dynamiques à des échelles différentes. Par exemple, l'hôpital support de l'étude est un hôpital qui peut être divisé en deux ailes (bâtiments), eux-mêmes divisés en deux et trois étages, pour un total de cinq étages dans l'hôpital. Chaque étage possède alors ses propres spécificités quant aux types de pathologies traitées (avec quelques recouvrements entre les étages). Parallèlement à cela, alors que la majeure partie des personnels est affectée à un service donné, certains personnels sont beaucoup plus mobiles et entrent en contact avec d'autres personnes de manière plus homogène (par exemple, les personnels de nuit ou les ergothérapeutes). On voit donc que si le graphe des contacts expose les mêmes caractéristiques que l'organisation de l'hôpital, alors nous avons affaire à un problème ayant une dynamique structurelle multi-échelle. En outre, les analyses microbiologiques présentent également plusieurs dynamiques. En effet, les souches bactériennes étudiées peuvent se propager entre individus, mais elles peuvent également muter et développer des résistances aux traitements. Ceci nous force à considérer les interactions inter-souches pour un modèle épidémique précis.

Ce jeu de données est donc remarquable par ses multiples dynamiques structurelles, temporelles et de transformation, et donne donc un cadre idéal pour mettre en application les méthodes d'ondelettes sur graphe récemment introduites. Il est également idéal pour mieux maîtriser et comprendre cet outil d'ondelettes sur graphe. On pourra voir par exemple [2] pour une méthode d'ondelettes sur graphe, et [4] pour une application de cette méthode à la détection de communautés.

2 Données i-Bird

Le présent travail a pour but d'analyser le jeu de données issu du projet i-Bird, projet mené de Mai 2009 à Octobre 2009 à l'Hôpital Maritime de Berck-sur-Mer. Cet hôpital de moyenne à longue durée (de séjour) a été choisi en raison du caractère

de « réservoir » à infection de ce type d'établissement de soins. Lors de la période de l'étude, c'est l'ensemble des personnels et des patients ayant donné leur accord qui a été mis à contribution, à savoir plus de 450 patients et près de 350 personnels. La question posée était alors de quantifier l'influence des contacts sur une diffusion microbienne.

Dans toute l'étude i-Bird, il est supposé qu'un contact physique entre deux personnes peut être approximé par une courte distance entre ces personnes (distance inférieure à un mètre et demi). Partant de cette supposition, les participants à l'étude ont été équipés de capteurs sans fil enregistrant toutes les trente secondes les capteurs détectés, afin d'obtenir une image du réseau de contacts au cours du temps.

L'objectif de cette étude de terrain est en outre d'étudier la propagation et la résistance de souches bactériennes responsables d'infections nosocomiales. Deux grandes familles de souches ont fait l'objet de prélèvements (écouvillons) hebdomadaires : les staphylocoques dorés (*Staphylococcus Aureus*, responsables de 13% des infections nosocomiales) et les entérobactéries (responsables de 5% de ces infections). Ces prélèvements ont ensuite été mis en culture et analysés à la recherche des souches présentes et de leurs résistances aux différents traitements antibiotiques. Nous nous intéressons dans ce travail à la propagation et à l'évolution des souches de staphylocoques dorés au sein de l'hôpital, ces bactéries se propageant par contact physique.

À l'issue de l'étude, l'ensemble des données disponibles comprend près de 70000 heures de contacts bidirectionnels répartis sur 1.6 millions de contacts, et près de 7000 résultats de prélèvements. Cet ensemble de données peut alors être vu comme un grand graphe dynamique (des contacts) pour lequel les nœuds sont le support de plusieurs (une par souche microbienne) fonctions de portage (de la souche). Notre but est alors d'étudier ces fonctions.

Dans la suite de ce document, nous considérons un sous-ensemble représentatif de ces données. Ce sous-ensemble s'étend sur cinq semaines et comprend 166 individus.

3 Diffusion Microbienne

Afin de mieux appréhender le problème de la modélisation d'une diffusion microbienne, nous avons isolé dans le jeu de données i-Bird une souche de staphylocoques dorés remarquable. Les souches de staphylocoques sont usuellement classées en deux catégories : les souches résistantes (*SARM* pour *staphylococcus aureus* résitant à la méticiline) et les souches non résistantes à la *méticiline* (*SASM* pour *staphylococcus aureus* susceptible à la méticiline), un antibiotique. Cette caractérisation permet de différencier les souches communes facilement traitables des souches multi-résistantes responsables d'infections nosocomiales. La souche isolée a donc été sélectionnée parmi les souches de SARM, un des sujets d'étude du projet i-Bird. Deux critères ont ensuite été retenus : une diffusion nette depuis un nombre aussi restreint que possible d'individus puis une diffusion au sein de la population.

La Figure 1 (a-d) montre le portage de cette souche à différentes semaines. On peut observer que la diffusion se fait à partir d'un patient qui reste à la fin le seul porteur de la souche. De plus, il est à noter que cette souche est principalement transmise à des personnels.

4 Ondelettes sur graphe

La transformée en ondelettes d'une fonction définie sur les nœuds d'un graphe permet d'analyser les variations spatiales de cette fonction en prenant en compte la topologie du graphe telle que définie par les arêtes (pondérées). Ce document reprend les grandes lignes de l'approche développée dans [2].

Transformée de Fourier d'une fonction sur graphe Dans un premier temps, il est nécessaire d'obtenir une notion de fréquence dans l'espace du graphe considéré. Pour cela, nous utilisons l'analyse spectrale du Laplacien du graphe $\mathcal{L} = D - A$ où D est la matrice diagonale des degrés des nœuds et A la matrice d'adjacence du graphe (où A_{ij} est le poids de l'arête ij). Le graphe est supposé non orienté si bien que A est symétrique semi-définie positive [3]. La décomposition spectrale du Laplacien donne $\mathcal{L}\chi_l = \lambda_l \chi_l$. De plus nous supposerons que le graphe est connexe donc que la multiplicité de la valeur propre 0 est 1. Par analogie avec les fonctions temporelles, on associe les valeurs propres aux fréquences et les vecteurs propres aux modes de Fourier. En notant N le nombre de nœuds du graphe et par analogie avec le cas classique, on définit la transformée de Fourier :

$$\hat{f}(l) = \langle \chi_l, f \rangle = \sum_{n=1}^{N} \chi_l^*(n) f(n). \tag{1}$$

Transformée en ondelettes d'une fonction sur graphe Cette décomposition spectrale permet de définir une transformée en ondelettes. Pour cela, nous définissons un opérateur d'ondelette par multiplication dans l'espace de Fourier du spectre de f avec le spectre d'un noyau noté $g : \mathbb{R}_+ \to \mathbb{R}_+$, gabarit spectral d'un filtre passe-bande, dilaté à l'échelle s :

$$\widehat{T_g^s f}(l) = g(s\lambda_l)\hat{f}(l). \tag{2}$$

À partir de cette définition, l'ondelette $\psi_{s,n}$ à l'échelle s et centrée sur le nœud n est obtenue par application de l'équation (2) avec $\hat{f}$ la transformée de Fourier de la fonction Dirac localisée sur le nœud n. Exprimée dans l'espace du graphe, celle-ci s'écrit :

$$\psi_{s,n} = T_g^s \delta_n \tag{3}$$

Les coefficients d'ondelettes d'une fonction f sont alors obtenus par projection sur les ondelettes :

$$W_f(s,n) = \langle \psi_{s,n}, f \rangle \tag{4}$$

Comme pour les ondelettes définies sur les séries temporelles, ψ est très souvent associée à une fonction d'échelle ϕ duale d'un filtre passe-bas h (filtre miroir de g dans le cas d'une analyse multirésolutions [1]). Sans le rappeler explicitement dans la suite, c'est cette fonction d'échelle ϕ plutôt que l'ondelette associée que nous utilisons dans cette étude, avec comme choix particulier $h(x) = exp(-x^4)$ [2].

5 Ondelettes sur graphe des contacts

Ce travail a pour but de présenter un cadre d'étude expérimental particulièrement bien adapté à l'application de certaines approches du traitement du signal orienté graphes. Si le but à terme est l'étude simultanée de toutes les dynamiques du jeu de données, comme énoncé en Section 2, nous nous intéressons dans un premier temps à une structure de données réduite. Nous faisons ici l'hypothèse qu'un contact entre deux individus est une variable aléatoire et que cette variable aléatoire vue comme une série temporelle est stationnaire. Sous cette hypothèse, il est raisonnable de caractériser statistiquement la variable contact par une mesure globale, comme par exemple la durée moyenne de contact, la durée moyenne d'inter-contact, le nombre de contacts (par jour, par semaine), ou plus simplement la durée cumulée de contact. Le choix de cette mesure est déterminant en termes de signification épidémiologique pour permettre de mieux comprendre et d'expliquer la diffusion macroscopique d'une souche bactérienne.

Cette grandeur statistique doit en outre être calculée sur une échelle de temps suffisamment grande pour ne pas être trop dépendante de phénomènes individuels ou des périodes liées à l'activité des services. Étant donné que les prélèvements n'ont eu lieu qu'au plus une fois par semaine et que l'on observe une périodicité des contacts à l'échelle de la semaine, nous calculons la valeur moyenne choisie par agrégation des données de contact sur plusieurs semaines consécutives. La question est alors de comprendre comment la structure du graphe agrégé à cette échelle influence la diffusion, et en particulier quelle mesure sur les contacts explique le mieux la diffusion observée.

Comme on l'a vu précédemment, l'organisation de l'hôpital est hiérarchisée. Dès lors, on s'attend à ce que le graphe des contacts révèle les mêmes caractéristiques que cette organisation, et en particulier, ait une dynamique structurelle multi-échelle. Ainsi, les outils d'analyse multirésolutions tels que les ondelettes sur graphes (*cf.* Section 4), semblent un choix pertinent pour caractériser la dynamique (spatiale) de ces objets, tout en tenant compte de la structure du support. Plus précisément nous nous focalisons sur la diffusion microbienne en étudiant les mesures de portage, comme fonctions à analyser. Les questions posées sont donc comment se diffusent les souches microbiennes et quels nœuds du graphe des contacts sont des vecteurs de diffusion ou au contraire en sont des freins.

Nous nous intéressons alors à l'évolution de l'ondelette (de la fonction d'échelle) sur le graphe avec l'échelle d'analyse, telle que donnée par la relation (3). Nous ne nous attendons pas à ce que l'ondelette se diffuse sur les nœuds du graphes selon les mêmes modes que la propagation microbienne. Notre intention est simplement de voir dans quelle mesure ces deux schémas présentent des similitudes, et plus précisément si en révélant certaines caractéristiques du graphe, les ondelettes offrent des éléments d'explication à la dynamique épidémiologique. La Figure 1 montre quelques exemples d'évolution d'ondelettes (fonction d'échelle), correspondant à différentes positions nodales de la fonction Dirac, et à différents choix de mesure des contacts.

Avant de commenter les résultats, il est nécessaire de préciser que les nœuds du graphe ont été spatialisés par un algorithme d'attraction-répulsion se basant sur le poids des arêtes du graphe, à savoir le temps cumulé de contact dans le reste de ce document. Cette spatialisation est en accord avec les différentes échelles spatiales de l'hôpital. Nous remarquons également que le portage au cours du temps de la souche considérée ne reste pas confinée au service d'origine, mais se délocalise aux autres services.

Il nous est maintenant possible d'étudier la dilatation du support des fonctions d'échelle dans différents cas, et notamment lorsque la fonction est centrée sur différents Dirac, ou pour des pondérations différentes des arêtes. Afin de conserver un lien avec la souche isolée, nous nous proposons de choisir deux individus porteurs très tôt de la souche : un patient et un personnel. Les Figures 1 (e-h) et (i-l) montrent la dilatation du support des fonctions d'échelle pour ces individus. On peut observer que ces deux dilatations sont globalement très différentes. En particulier, les figures 1(h) et 1(l) montrent que le transfert de masse entre les nœuds « patient » et « personnel » n'est pas identique dans les deux sens. On note également que les amplitudes maximales restent localisées dans le cluster de la source dans le cas où le Dirac est placé sur le nœud « personnel », mais qu'elles se délocalisent sur un autre service que celui du Dirac lorsque celui-ci coïncide avec le nœud « patient ». Ces observations confirment qu'un modèle de diffusion (type équation de la chaleur) basé sur la seule durée des contacts entre individus n'explique pas la dynamique de contagion.

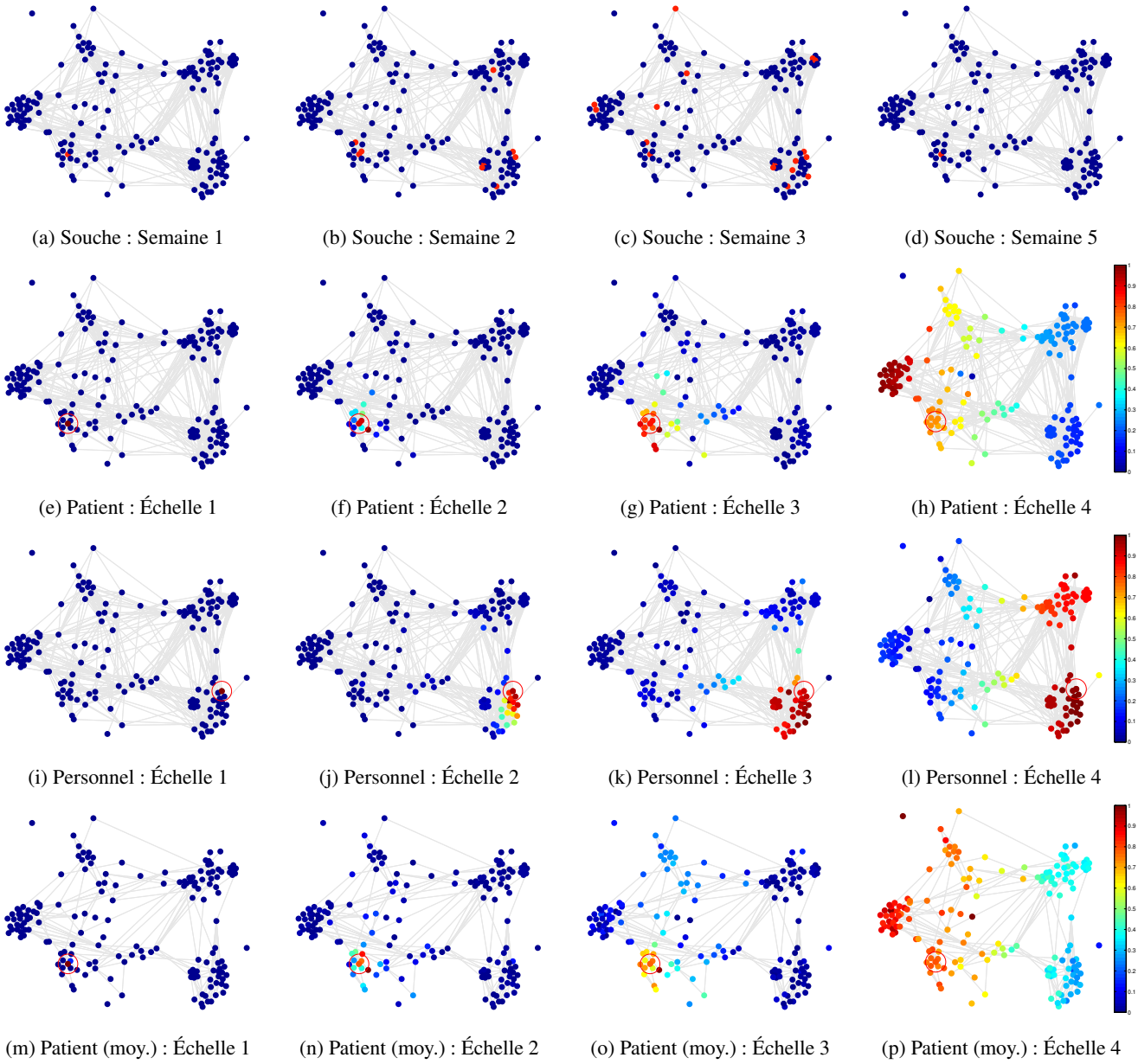
Enfin, les figures 1 (m-p) montrent la dilatation du support des fonctions d'échelle avec comme pondération des arêtes, le temps moyen de contact. Comme pour la pondération précédente, ce modèle de diffusion ne permet pas *per se*, d'expliquer le mécanisme de diffusion de la souche. Néanmoins ce cas de figure semble accélérer (au sens des échelles) l'étalement de la fonction d'échelle aux services voisins de celui du foyer.

6 Perspectives

Cette analyse descriptive est un travail prépatoire à une étude approfondie des données i-Bird par les outils de traitement du signal. En particulier, la question de la bonne grandeur représentative d'un contact qui expliquerait au mieux les différentes diffusions bactériennes est posée et nous invite à poursuivre des comparaisons systématiques entre ondelettes sur graphe et dynamique des mesures de portage.

D'un point de vue épidémiologique, comprendre quelle est le facteur le plus explicatif d'une diffusion bactériologique pourrait avoir un impact sur les protocoles de soins et sur l'organisation structurelle des services.

Enfin, les différentes souches bactériennes peuvent s'échanger de l'information, et à ce titre développer des résistances aux différents traitements. Il s'agit aussi pour le projet i-Bird de comprendre comment les différents individus au sein d'un hôpital favorisent un échange d'information génétique des bac-



(a) Souche : Semaine 1 (b) Souche : Semaine 2 (c) Souche : Semaine 3 (d) Souche : Semaine 5

(e) Patient : Échelle 1 (f) Patient : Échelle 2 (g) Patient : Échelle 3 (h) Patient : Échelle 4

(i) Personnel : Échelle 1 (j) Personnel : Échelle 2 (k) Personnel : Échelle 3 (l) Personnel : Échelle 4

(m) Patient (moy.) : Échelle 1 (n) Patient (moy.) : Échelle 2 (o) Patient (moy.) : Échelle 3 (p) Patient (moy.) : Échelle 4

FIGURE 1 – (a-d) : portage au cours des semaines d'une souche. (e-p) : coefficients de la fonction d'échelle de la réponse impulsionnelle d'un patient ou d'un personnel sur le graphe des contacts pondéré par la durée cumulée des contacts (figures (e-l)) ou par la durée moyenne de contact (figures (m-p)) normalisés par le coefficient maximal de l'échelle. Quatre échelles sont ici représentées, de la plus fine à gauche à la plus grossière à droite. Seules les arêtes de poids suffisamment grand sont représentées (temps cumulé supérieur à 5 minutes ou temps moyen supérieur à 2 minutes).

téries propre à développer des multi-résistances. Dans ce cadre, identifier les mécanismes macroscopiques en jeu dans le développement des multi-résistances est un pré-requis pour comprendre comment les limiter.

Références

[1] I. DAUBECHIES : *Ten lectures on wavelets*. SIAM, 1992.

[2] D.K. HAMMOND, P. VANDERGHEYNST et R. GRIBONVAL : Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011.

[3] D. SPIELMAN : Spectral graph theory. *Lecture Notes, Yale University*, 2009.

[4] N. TREMBLAY et P. BORGNAT : Multiscale Community Mining in Networks Using Spectral Graph Wavelets. Preprint, décembre 2012.