2016

Réseaux spatiaux: modèles, classification et flux

Guillaume Guex

UNIL | Université de Lausanne
Faculté des géosciences
et de l'environnement

FACULTÉ DES GÉOSCIENCES ET DE L'ENVIRONNEMENT

INSTITUT DE GÉOGRAPHIE ET DURABILITÉ

Réseaux spatiaux : modèles, classification et flux

THÈSE DE DOCTORAT

présentée à la Faculté des géosciences et de l'environnement
de l'Université de Lausanne
pour l'obtention du grade de Docteur ès géographie

par

Guillaume Guex

Titulaire d'un
Master en statistiques
de l'Université de Neuchâtel

Jury

Directeur de thèse : Prof. François Bavaud, Université de Lausanne
Président : Prof. Suren Erkman, Université de Lausanne
Expert interne : Prof. Mikhail Kanevski, Université de Lausanne
Expert externe : Prof. Marco Saerens, Université Catholique de Louvain

LAUSANNE, 2016

UNIL | Université de Lausanne
Faculté des géosciences
et de l'environnement

FACULTÉ DES GÉOSCIENCES ET DE L'ENVIRONNEMENT

INSTITUT DE GÉOGRAPHIE ET DURABILITÉ

Réseaux spatiaux : modèles, classification et flux

THÈSE DE DOCTORAT

présentée à la Faculté des géosciences et de l'environnement
de l'Université de Lausanne
pour l'obtention du grade de Docteur ès géographie

par

Guillaume Guex

Titulaire d'un
Master en statistiques
de l'Université de Neuchâtel

Jury

Directeur de thèse : Prof. François Bavaud, Université de Lausanne
Président : Prof. Suren Erkman, Université de Lausanne
Expert interne : Prof. Mikhail Kanevski, Université de Lausanne
Expert externe : Prof. Marco Saerens, Université Catholique de Louvain

LAUSANNE, 2016

UNIL | Université de Lausanne
Décanat Géosciences et de l'Environnement
bâtiment Géopolis
CH-1015 Lausanne

IMPRIMATUR

Vu le rapport présenté par le jury d'examen, composé de

Président de la séance publique :	M. le Professeur Suren Erkman
Président du colloque :	M. le Professeur Suren Erkman
Directeur de thèse :	M. le Professeur François Bavaud
Experte interne:	M. le Professeur Mikhail Kanevski
Expert externe :	M. le Professeur Marco Saerens

Le Doyen de la Faculté des géosciences et de l'environnement autorise l'impression de la thèse de

Monsieur Guillaume GUEX

Titulaire d'un
Master en statistiques
de l'Université de Neuchâtel

intitulée

RÉSEAUX SPATIAUX: MODÈLES, CLASSIFICATION ET FLUX

Lausanne, le 7 janvier 2016

Pour le Doyen de la Faculté des géosciences et de l'environnement

Professeur Suren Erkman, Vice-doyen

Résumé

Un *réseau spatial* est une structure constituée d'objets appelés *noeuds*, localisés dans un espace métrique, et possédant des connexions de différentes nature appelées *arêtes*. Cette définition comprend un large panel de structures. Les réseaux routiers, aériens, ferroviaires, de lignes électriques, de rivières, etc. ne sont que quelques exemples de réseaux spatiaux que nous côtoyons tous les jours. Contrairement aux réseaux usuels, qui ne possèdent pas de composante spatiale, maintenir une connexion entre deux noeuds d'un réseau spatial requiert un coût qui va dépendre de la distance séparant ces noeuds, et la structure du réseau tout entière est alors intimement liée à l'espace. L'analyse du rapport existant entre l'espace et la structure du réseau n'est pas chose aisée et les approches permettant d'y arriver sont nombreuses. Ce travail se concentre sur la création de modèles mathématiques facilitant cet objectif, en particulier à l'aide de modèles de *flux*.

Un flux peut être interprété comme une quantité de matière créée par un certain nombre de noeuds appelés *sources*, naviguant sur les arêtes du réseaux pour finalement être absorbée par d'autres noeuds, appelés *cibles*. Le comportement adopté par le flux lors de sa navigation est généralement dicté par la minimisation d'une fonctionnelle, et en choisissant judicieusement cette dernière, il est possible d'obtenir des flux permettant une exploration particulière de la structure du réseau. L'ingrédient principal de cette thèse sera un modèle de flux inédit, appelé le *flux de transport randomisé*. Ce flux, dont le comportement peut être ajusté par un paramètre $T > 0$, est à la croisée entre le flux de *transport optimal* et le *flux aléatoire*. Lorsque $T \to 0$, ce flux permettra de transporter la matière dont il est constitué des sources jusqu'aux cibles en minimisant les coûts. A l'opposé, lorsque $T \to \infty$, ce flux suivra une marche aléatoire, explorant la totalité du réseau avant d'être absorbé. Ces deux extrêmes contiennent des flux qui ont été étudiés dans la littérature, mais nous verrons ici que lorsque la température est intermédiaire, le flux de transport randomisé adoptera alors un comportement inédit, permettant d'appréhender la structure du réseau de manière plus complète. Ce flux nous permettra de construire de nouveaux *indices de centralité*, de nouvelles *dissimilarités* et offrira une solution numérique efficiente à la résolution du problème de transport optimal dans un réseau. Ce nouvel objet soulèvera également quelques questions fondamentales, dépassant largement le cadre des réseaux spatiaux. Un chapitre sur la création de modèles de réseaux spatiaux, similaires aux réseaux existants, viendra compléter l'analyse obtenue par les flux. Si le flux permet, entre autres, d'obtenir de l'information sur la position des noeuds dans l'espace grâce à la structure du réseau, ce dernier chapitre bouclera la boucle en faisant le travail en sens inverse : déduire l'information sur la structure d'un graphe spatial à partir de la position des noeuds dans l'espace.

Cette thèse, constituée d'articles "peer-reviewed" insérés au milieu d'un manuscrit rédigé à la suite de ceux-ci, présente deux niveaux de lecture. En lisant le manuscrit, un lecteur pourra comprendre les enjeux, le contexte et la problématique de notre approche sur l'analyse des réseaux spatiaux, tout en se familiarisant avec le formalisme. Cette lecture s'adresse ainsi à un large public, cependant familier avec les notations mathématiques. La lecture des articles, moins accessible, présente la réalité du travail scientifique accompli durant cinq années. Ces articles ne seront pas toujours parfaitement cohérents les uns par rapport aux autres, et les idées exposées évoluent au fil du temps. Néanmoins, leur lecture permet d'avoir les démonstrations et les résultats les plus significatifs, complétant ainsi le contexte donné par le manuscrit sous la forme d'avancées faites dans le domaine.

Abstract

A *spatial network* is a structure composed of objects lying in a metric space (*nodes*) and connexions between these objects (*edges*). This definition includes a large array of real structures. Roads, airlines, railroads, power grids, etc. are only a few examples of spatial networks encountered in everyday life. Unlike classical networks, which do not contain a spatial component, every connexion in a spatial network has a cost that depends on the distance between nodes, and the entire network structure is closely related to the underlying space. The analysis of the connexion between space and the network structure is not a trivial task, and different approaches can be adopted. This work focuses on creating mathematical models leading to that goal, with an emphasis on *flow* models.
A flow can be interpreted as a quantity of matter, created by a set of nodes (*sources*), transported on graph edges before being eventually absorbed by another set of nodes (*targets*). The behavior of a flow in the network is dictated by the minimization of a functional, and when the latter is chosen properly, the flow will explore the network in a particular way. In this thesis, an emphasis will be placed on previously unseen flow, which we have called the *randomized transportation flow*. This flow possesses a behavior that can be tuned with a free parameter $T > 0$, and is at the crossroads between the *optimal transportation flow* and the *random flow*. When $T \to 0$, this flow carries the matter from sources to targets while minimizing the cost of transportation. By contrast, when $T \to \infty$, it follows a random walk, exploring the entire network before being absorbed. While these two opposite poles are already known in the literature, the flow will adopt an entirely new behavior when the temperature is in-between, and allows us to understand the network structure more accurately. This flow will be used to build new *centrality indices* and new *dissimilarities*, and will offer an efficient way of solving the transportation problem in a network. Moreover, fundamental questions will arise from this new object, greatly outpacing the setting of spatial networks. A chapter about the creation of spatial network models will complete the analysis carried out with the flow. While the flow permits, among other things, the extraction of information about the underlying space through the network structure, the last chapter will work backwards, attempting to extract information on the network structure from node location in space.
This thesis is composed of peer-reviewed articles inserted into a broader text, and offers two different levels of reading. By focusing on the text, readers will familiarize themselves with the issues and the context of our approach to spatial networks, while simultaneously coming into contact with mathematical formalism. This topic concerns a wide audience, however accustomed to mathematical notations. While more difficult to read, the articles show the scientific work accomplished during five years. The notation and formalism used in them are not always identical, and exposed ideas evolved over time. Nevertheless, they contain the most significant demonstrations and results, and are good examples of what can be undertaken on the subject.

Remerciements

L'expérience d'un doctorat à l'Université de Lausanne a été pour moi une fantastique aventure, et ce manuscrit ne peut malheureusement que faiblement refléter ce que j'ai pu vivre durant ces cinq années. Il me tient donc à coeur de remercier toutes les personnes qui y ont participé, de près comme de loin.

Pour commencer, je tiens à remercier François Bavaud, sans qui tout cela n'aurait pas été possible. Merci à lui d'avoir cru en moi, de m'avoir inspiré, guidé et conseillé tout au long de ce travail. Nos discussions de recherche ont grandement participé à l'aboutissement de cette thèse et j'ai été ravi d'avoir pu échanger avec quelqu'un de si inspiré et de si brillant. Je le remercie également d'être la personne fabuleuse qu'il est hors du travail. J'ai eu énormément de chance d'être tombé sur un directeur sympathique, humain, humble et j'ai eu beaucoup de plaisir à partager des moments moins formels, lors de conférences ou à l'université.

Je remercie de même Marco Saerens. En apprenant que nous avions publié un article avec un formalisme similaire au sien, Marco nous a immédiatement contacté et a exprimé son désir de collaborer, plutôt que de se lancer dans une lutte pour la paternité du modèle qui se produit parfois dans le monde académique. Je salue cet état d'esprit grandement favorable à la recherche, sa curiosité et sa sympathie. Merci également à lui d'avoir accepté de faire partie de mon jury de thèse. J'espère que nos collaborations futures, si elles peuvent se produire, seront fructueuses.

Merci également à Aris Xanthos, qui a grandement participé au climat très agréable qui règne dans la section. Je le remercie aussi pour son idée d'article en commun, bien que ce dernier n'apparaisse pas dans cette thèse. Cet article fut l'occasion pour moi de sortir un peu de mon cadre et de découvrir le monde de la linguistique quantitative et les confins de la République Tchèque. Merci à lui d'avoir rendu cela possible.

En restant dans le cadre de l'Université, je tiens à remercier Christelle Cocco, Théophile Emmanouilidis et Sébastien Pabst, de m'avoir si bien accueilli au sein des assistants de la section et de m'avoir montré les ficelles du métier. Ils ont fait partie de la première équipe avec laquelle j'ai pu travailler et les discussions que nous avons partagées m'ont beaucoup aidé à débuter dans ma recherche. Je remercie aussi les deux "nouveaux" assistants : Mattia Egloff et Raphaël Ceré, avec qui j'ai déjà passé de moments très plaisants malgré leur arrivée relativement récente. Je suis certain qu'ils seront dignes de défendre le climat si sympathique qui règne dans le bureau 3132 de l'Anthropôle et je suis ravi de quitter la section en sachant qu'ils font partie de la relève. Je tiens à remercier également les personnes qui sont passées par les postes d'assistants-étudiants durant mon parcours : Aline Corpataux, Vincent Humphrey, Anne-Laure Aeby et Lucas Meylan. Merci à eux d'avoir partagé notre dur labeur. Les surveillances et les corrections des tests de méthodes quantitatives n'étaient sans doute pas la partie la plus passionnante du travail, mais leur présence a participé à rendre cette tâche plus plaisante. Je remercie de même tous les autres membres de la section des sciences du langage et de l'information, pour avoir participé à l'ambiance générale qui a été remarquablement agréable. Merci donc à Anne-Claude Berthoud, Jean-Baptiste Blanc, Xavier Gradoux, Jérôme Jacquin, Johanne Jordan, Marianne Kilani Schoch, Gilles Merminod, Isaac Pante, Florence Parmelin, Davide Picca, Pascal Singy, Nadia Spang Bovey, Antoine Viredaz et Rudolf Wachter, avec qui j'ai passé de bons moments

lors des apéros de section ou toutes autres occasions nous réunissant. Un remerciement spécial revient à Jean-Baptiste Blanc, pour avoir si souvent partagé avec moi le refus de mettre un terme aux choses, des moments inoubliables en ont résulté.

Hors du cadre professionnel, je tiens spécialement à remercier Gaëlle Kovaliv, qui est la personne à avoir partagé la plus grande partie de mon quotidien durant cette thèse. Merci à elle d'être si rayonnante, aimante, sympathique et joyeuse. Grâce elle, mon existence est bien plus agréable et j'espère vraiment qu'elle sera présente dans ma vie le plus longtemps possible. Je veux également la remercier pour ses corrections, ses discussions et ses cris d'animaux qui m'ont beaucoup aidé lors d'un tel travail intellectuel. Merci vraiment pour tout, Gaëlle.

J'aimerais aussi beaucoup remercier ma famille, sans qui je ne serais pas ce que je suis aujourd'hui. Merci à ma mère, Margot, d'être une personne si indépendante et décidée. Je l'ai toujours admirée pour ses choix de vie et j'en suis très fier, même si je ne le lui dit sans doute pas assez souvent. Merci à mon frère Joachim pour ses soirées, son franc parler et ses débats incessants. Bien que son entêtement ait souvent provoqué des discussions interminables entre nous, il a grandement participé à développer mon esprit critique et mes réflexions. Merci aussi à mon frère Thibaud, qui est sans doute l'être qui me ressemble le plus en ce monde, pour les moments emplis de complicité que nous partageons si souvent. J'espère de tout mon coeur qu'il trouvera sa voie comme j'ai eu la chance de trouver la mienne. Je voulais aussi remercier mon père, Pierre, bien qu'il ne lira jamais ces lignes. Il a été un père fabuleux et j'aurais aimé lui dire que sa bonté, sa curiosité et ses réflexions m'habiteront tout au long de ma vie. Finalement, merci à mes oncles et tantes préférés : Sébastien, Janick, Ana et Paul ; j'ai toujours énormément de plaisir à les voir et les discussions que nous partageons sont toujours pour moi des grands moments de bonheur.

Pour terminer, j'aimerai remercier plusieurs de mes amis. Un merci spécial revient à Laura Ces, pour ses conseils juridiques judicieux et absolument sans rapport avec cette thèse, et à Marek Gersbach, pour ses corrections d'anglais pertinentes. Merci également à : Mélina Andronicos, Antoine Berthoud, Stéphane Broillet, Daniel Brulhart, Natacha et Daniel Clerc, Eric Corradin, Lucy Despont, Christophe Diserens, Johanna Frantz, Annie Karlsson, Caroline et David Krzywicki, Aurélie Lasserre, Kim-Anne Le, Sylvain Mayor, Thien Nguyen, Sophie Pedgrift, Thomas Schild, Mina Todorova, Helena Vrtek, Sébastien Wong et Denis Zweifel. C'est grâce à toutes ces amitiés de longue date que j'ai pu trouver de quoi me ressourcer entre la rédaction de ces lignes, et sans les bons moments passés avec eux, je n'aurais sans doute jamais trouvé la force de terminer ce travail.

Et merci encore à toutes les personnes que j'oublierai. Comme de nombreux sociologues, je suis convaincu que les actions d'une personne dépendent en grande partie de son réseau social proche. Vous qui lisez ces lignes, il y a de forte chance pour que vous en fassiez partie, alors, cher lecteur, à vous aussi je vous dis merci. J'espère que ce travail vous plaira...

Table des matières

1 Introduction

1.1 Histoire et contexte

Dans sa définition la plus générale, un *réseau* est une structure d'objets réels, distincts, interconnectés les uns avec les autres. Cette définition est depuis longtemps passée dans le langage courant et s'applique ainsi à de nombreux objets de la vie quotidienne. Les réseaux sociaux, les réseaux aériens, les réseaux de neurones, ne sont que quelques exemples de locutions que nous employons fréquemment et qui désignent des structures constituées d'éléments interconnectés. La particularité des réseaux est que la structure des liens entre les différents objets, aussi appelée *topologie* du réseau, possède une importance primordiale. En effet, peu importe le réseau étudié, son fonctionnement dépend principalement de sa topologie et l'étude de celle-ci est donc fondamentale pour le comprendre. Notons cependant que les liens existant au sein des différents réseaux ne sont pas forcément de même nature : il existe des liens à sens unique (comme dans un réseau d'auteurs, ou un auteur peut en citer un autre sans pour autant se faire citer en retour), des liens plus ou moins forts (par exemple, le nombre de voitures qui fréquentent un tronçon autoroutier peut être plus ou moins élevé) et même parfois, ces liens sont difficilement définissables, bien que l'on soit persuadé de leur existence (par exemple, les "connaissances" dans un réseau social peuvent être de natures très variées). La représentation abstraite d'un réseau se nomme un *graphe*, qui se définit comme un ensemble constitué de points, appelés *noeuds* ou *sommets*, représentant les objets qui constituent le réseau, et de liens entre ces points, nommés *arêtes*, ou dans un cas orienté, *arcs*, représentant les relations entre les objets. La distinction entre réseaux et graphes est parfois ambigüe et bien que ces deux mots soient souvent utilisés comme des synonymes, le terme réseau est généralement utilisé pour une structure réelle, alors que qu'un graphe est un modèle abstrait de ce dernier.

L'étude des graphes est une branche des mathématique discrètes, la *théorie des graphes*. L'origine de cette discipline revient à Leonhard Euler, qui utilisa, en 1736 (?, ?), un modèle mathématique pour répondre à une énigme populaire à son époque, le *problème des sept ponts de Königsberg* (Kaliningrad aujourd'hui). Ce problème consistait à trouver s'il existait un chemin permettant de passer une seule fois par chacun des sept ponts de la ville et de revenir à son point de départ. Euler montra que ce chemin n'existait pas et énonça qu'une telle promenade n'était possible que sur un graphe constitué de noeuds possédant un nombre pair d'arêtes (ce résultat est maintenant connu sous le nom du *Théorème d'Euler*, dont la preuve fut en réalité apportée par Carl Hierholzer en 1873 (?, ?)). Ce résultat, certes anecdotique, est le premier cas retenu par l'histoire, dans lequel une personne, inspirée par un problème réel, a formalisé la structure abstraite d'un réseau sous la forme d'un graphe et en a déduit certaines propriétés. Cette nouvelle façon de raisonner inspira d'autres mathématiciens, du XVIII^{e} siècle jusqu'à aujourd'hui, qui, de plus en plus, se penchèrent sur ce formalisme de graphe pour résoudre certains problèmes théoriques aux allures de casses-têtes. Nous pouvons par exemple citer les problèmes de chemins hamiltoniens (où l'on cherche un chemin couvrant tout le graphe et ne passant qu'une fois par chaque noeud, comme le fameux problème du cavalier qui doit visiter chaque case d'un jeu d'échec), le problème de coloration de graphe (qui consiste à déterminer le nombre de couleurs nécessaires pour colorer entièrement les noeuds d'un graphe sans que deux noeuds de même couleur soient reliés) et l'étude des *facteurs* d'un graphe (cherchant les sous-graphes recouvrants d'un graphe donné, c'est-a-dire les graphes contenant seulement une partie des arêtes et tous les noeuds du graphe initial). L'étude de ces problèmes et les résultats

obtenus permirent ainsi de faire avancer cette nouvelle branche des mathématiques discrètes qu'est l'étude des propriétés fondamentales d'un graphe. Ce domaine est encore très étudié de nos jours et de nombreuses questions restent encore ouvertes.

Nous ne pouvons parler de théorie des graphes sans évoquer son voisin le plus direct : l'*étude des réseaux*, aussi appelée *diktyologie*. Ce domaine consiste en l'observation et la caractérisation des structures réelles formant un réseau et a toujours été intiment lié à la théorie des graphes. A l'image du problème des sept ponts de Königsberg, et comme souvent dans le monde scientifique, les problèmes et théories sur des objets abstraits se nourrissent d'interrogations et d'observations issues du quotidien, et, en retour, ces abstractions nous aident à comprendre le monde réel. Ces deux domaines n'y font pas exception et l'étude des réseaux a toujours été présente en parallèle à la théorie des graphes. La diktyologie en tant que science n'est bien sûr pas uniforme et il existe de multiples théories, modèles et études empiriques qui diffèrent selon le type de réseaux étudié. Cependant, malgré cette grande variété, les même éléments-clés réapparaissent fréquemment. Ainsi, les *flux*, les *plus courts chemins*, la *distribution de degré* et certains indices, comme ceux de *centralité*, jouent un rôle unificateur pour ce domaine aux multiples facettes et il arrive fréquemment qu'un modèle construit pour résoudre un problème posé sur un type de réseaux puisse avoir des applications pour des réseaux de nature différente.

Les premières études du domaine pourraient être attribuées à Gustav Kirchhoff, pour ses travaux sur les réseaux électriques en 1845 (?, ?). Ce dernier modélisa les circuits électriques grâce à un graphe et représenta le courant électrique comme un *flux* sur ce graphe. Il montra que ce flux devait suivre plusieurs lois, appelées *lois de Kirchhoff*, afin que celui-ci se comporte de manière analogue au courant électrique. Cette approche de modélisation par flux est particulièrement intéressante pour nous, et sera l'ingrédient principal de notre étude sur les réseaux spatiaux, ce qui montre encore une fois qu'un modèle créé pour un type de réseaux peut facilement dépasser le cadre pour lequel il a été initialement conçu. Jusqu'au milieu du XX^{e} siècle, les réseaux électriques seront, à quelques exceptions près, les seuls a être étudiés. C'est à cette époque qu'un autre type de réseaux, très différent, commence à être également analysé et apporte son lot de nouvelles interrogations : les réseaux sociaux. En effet, en sociologie, les relations entre les individus a très rapidement pris une dimension fondamentale et il est bien normal que l'étude des réseaux sociaux se développa, en essayant de dégager quelles sont les structures et les interactions clés permettant d'expliquer le fonctionnement du monde social. Malheureusement pour la sociologie, les données sur les réseaux sociaux étaient rares et difficilement disponibles, il faudra attendre la venue d'Internet pour que les chercheurs du domaine puissent se doter de nombreux jeux de données afin de d'étayer leurs théories.

C'est vers la toute fin du XX^{e} siècle, et ce jusqu'à nos jours, grâce à l'accumulation des données grandissantes et l'essor de nombreuses technologies, telles que les ordinateurs, les téléphones mobiles, l'imagerie médicale, Internet, etc. que l'étude des réseaux explosa véritablement. Des applications concrètes, telles que l'algorithme de recherche "PageRank" créé par Google, montrèrent leur incroyable utilité pratique et des chercheurs dans presque tous les domaines commencèrent à s'intéresser aux réseaux. Aujourd'hui, en effectuant une recherche du mot-clé "network" sur Google Scholar, ce dernier nous renvoie le nombre vertigineux de 6.5 millions d'articles (en comparaison, le mot-clé "vector" ne donne "que" 4.6 millions d'articles). Cette abondance peut avoir de mauvais côtés : il devient difficile de suivre l'évolution des progrès faits quant à la compréhension des réseaux lorsque ceux-ci viennent simultanément de nombreux domaines différents, et il n'est pas rare que des résultats deviennent obsolètes rapidement ou ne soient que de simples "redécouvertes". Fort heureusement, en science, c'est souvent un bouillonnement de ce type qui a participé à l'essor de grandes théories unificatrices et sans doute qu'avec le temps, l'essentiel de ce qui est à retenir aujourd'hui apparaîtra plus clairement.

1.2 Objectifs et structure de la thèse

Cette thèse suit la tendance actuelle dans l'étude des réseaux : en utilisant des connaissances, observations et théories issues de domaines variés, elle développe différents modèles mathématiques utiles à l'étude des réseaux *spatiaux*. Cependant, comme nous l'avons mentionné, et nous le verrons par la suite, ces modèles peuvent aisément sortir du cadre pour lequel ils ont été créés.

Les réseaux spatiaux sont des réseaux dans lesquels les noeuds sont plongés dans un espace métrique, et il existe donc une *distance* entre les différents noeuds. Cette distance agit de la façon suivante : un lien entre deux noeuds éloignés sera généralement plus coûteux à maintenir qu'un lien entre deux sommets proches. En d'autres termes, les arêtes possèdent un *coût* qui dépend de leur longueur. La topologie d'un réseau spatial est donc intimement liée à l'espace dans lequel sont plongés les sommets, et les relations qui existent entre les deux sont sujettes à beaucoup d'interrogations. Nous pourrions nous demander, par exemple, s'il est possible de retrouver la position des noeuds dans l'espace uniquement grâce à la topologie du réseau, ou, à l'opposé, s'il est possible de créer un graphe spatial similaire à un réseau existant uniquement avec la position de ses noeuds. Les exemples de réseaux spatiaux sont nombreux surtout, spatialité oblige, en géographie. Les réseaux de transports, les routes dans les villes, les rivières, les canalisations et les réseaux de lignes électriques sont des éléments essentiels de l'urbanisation et leur étude apporte beaucoup dans le développement d'une région ou d'un pays. Face à l'explosion démographique, à l'accroissement du nombre de mégapoles dans le monde et aux défis toujours plus grands que représente le développement durable, les géographes, urbanistes et décideurs en tous genres se sont de plus en plus tournés vers les modèles mathématiques pour leur venir en aide dans la résolution de nombreux problèmes, et il n'est pas rare aujourd'hui de trouver parmi eux des physiciens ou des mathématiciens de formation. Cette thèse va tenter de participer à cet élan en proposant différents modèles permettant une meilleure compréhension des réseaux spatiaux.

Un des concepts-clés de cette thèse sera certainement la notion de *flux*, qui sera formalisé dès le premier chapitre. Un flux peut être vu comme une quantité de matière créée par un groupe de noeuds du réseau nommés les *sources*, se déplaçant sur les arêtes du réseau, pour être finalement absorbée par un autre groupe de noeuds du réseau, appelés les *cibles*. Un flux peut avoir différentes lois régissant son déplacement, et nous étudierons un flux au comportement particulier, nommé le *flux de transport randomisé*. Ce flux de transport randomisé, introduit ici pour le première fois, nous servira à construire des *indices de centralité*, des *dissimilarités* et des *méthodes de classifications des noeuds* sur des réseaux spatiaux. Nous verrons également, cette fois sans l'aide du concept de flux, comment construire différents graphes modélisant des réseaux spatiaux.

Cette thèse est un peu particulière, dans le sens qu'une partie de son contenu est constitué d'*articles scientifiques*, publiés ou en cours de publication. Ces articles créent malheureusement, de par leur langue (l'anglais) et leurs notations utilisées à l'époque, une rupture partielle avec le reste du texte, rédigé à la suite de ceux-ci. Cependant, ces articles contiennent le noyau dur des résultats obtenus tout au long de ce travail, et leur lecture est nécessaire pour la compréhension complète du manuscrit. Aussi, fait qui peut sembler curieux pour les lecteurs issus du domaine des mathématiques, la partie française du manuscrit ne contient aucune démonstration. Les raisons qui nous ont poussé à faire ce choix sont pédagogiques. En lisant ces parties, le lecteur est ainsi invité à se concentrer sur le contexte, les enjeux, le formalisme et le questionnement lié à la matière, et il nous a semblé superflu d'alourdir la lecture par les preuves des énoncés. Mais que nos amis mathématiciens se rassurent, toutes les démonstrations peuvent aisément

être retrouvées, soit dans les articles cités, soit dans les articles contenus dans la thèse. Cette façon de procéder permet donc de choisir son degré de lecture : le compréhension rapide et intuitive de la problématique est donnée par le manuscrit en français, et les articles offrent un travail scientifique plus complet, avec démonstrations et résultats.

Cette thèse est constituée de quatre parties distinctes. La première partie (chapitre 2) sera consacrée au *formalisme*. Nous y introduirons la notation utilisée dans la partie française du manuscrit, ainsi que les différents concepts utilisés tout au long de la thèse, comme par exemple ce nouvel objet qu'est le flux de transport randomisé.

La deuxième partie (chapitre 3) portera sur l'utilisation du flux de transport randomisé lorsque celui-ci ne comporte qu'une seule source et qu'une seule cible. Dans ce contexte, il nous sera possible de créer des *indices de centralité* et des *dissimilarités* sur des réseaux. Nous passerons en revue une partie de la littérature concernant ces sujets et l'innovation qu'apporte le flux de transport randomisé, contenu dans deux articles.

La troisième partie (chapitre 4) s'intéressera à l'utilisation du flux de transport randomisé dans le cadre général, avec des sources et des cibles multiples. Nous verrons que ce cadre correspond au *problème du transport optimal dans un réseau*. A la suite d'un article traitant de la résolution du problème de transport optimal sans flux, nous verrons dans un second article la résolution de ce problème grâce au flux de transport randomisé.

La quatrième et dernière partie (chapitre 5) est un peu particulière, dans le sens où le flux de transport n'apparaîtra plus. Elle portera sur les modèles de réseaux spatiaux, et plusieurs méthodes permettant de créer un graphe ayant une topologie similaire à certains réseaux spatiaux y sont étudiées. Ces modèles de réseaux ont parfois été utilisés dans les autres chapitres afin de disposer de données artificielles similaires à la réalité.

2 Formalisme

Bien que cette thèse comporte plusieurs chapitres aux approches sensiblement différentes, le formalisme et les notations utilisés dans les parties rédigées en français ont été construits pour être cohérents tout au long du manuscrit. Nous ne pouvons malheureusement pas en dire autant des articles, qui sont insérés ici sous la forme qu'ils avaient lors de leur publication. Cependant, bien que le formalisme des articles diverge un peu du point de vue des notations, il reste relativement proche, et un lecteur attentif n'aura aucune difficulté à trouver la correspondance. Dans ce chapitre, nous allons introduire les définitions et les concepts qui réapparaissent de manière récurrentes pour permettre au lecteur de se familiariser avec ceux-ci. Les lecteurs déjà familiers avec les concepts exposés ci-dessous (les graphes, les chaînes de Markov et les flux) pourront passer directement à la dernière section de ce chapitre (section 2.3.2), où est introduit l'objet le plus essentiel à cette thèse : le flux de transport randomisé.

2.1 Les graphes

2.1.1 Généralités

Définition

Un *graphe non orienté* est un couple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, composé de deux ensembles. L'ensemble $\mathcal{V}$ contient un nombre dénombrable d'éléments, $v \in \mathcal{V}$, appelé *noeuds* ou *sommets* et l'ensemble $\mathcal{E}$ contient un nombre dénombrable de paires de noeuds, $e = \{i, j\},\ i, j \in \mathcal{V}$ appelées *arêtes*. Nous nous restreindrons ici au graphe de taille finie, et les cardinalités de ces ensembles sont notées $|\mathcal{V}| = n$ et $|\mathcal{E}| = m$. On appelle le couple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ *graphe orienté* lorsque l'ensemble $\mathcal{E}$ est constitué de paires *ordrées* $e = (i, j),\ i, j \in \mathcal{V}$, alors nommées *arcs*. En réalité, les graphes non-orientés peuvent être interprétés comme un cas particulier d'un graphe orienté, où il existe, pour tout arc donné, un arc dans l'autre direction, c-à-d $(i, j) \in \mathcal{E} \Leftrightarrow (j, i) \in \mathcal{E}$.

Un graphe est dit *simple* lorsqu'il ne contient aucune *boucle*, c-à-d $i \neq j,\ \forall (i, j) \in \mathcal{E}$, et aucune *arête multiple* (ce qui est toujours le cas si $\mathcal{E}$ est un ensemble, c'est-à-dire une collection d'éléments distincts).

Dans cette thèse, notre formalisme se restreint aux cas des graphes simples, et, très fréquemment, aux graphes orientés. C'est pourquoi dans la suite du formalisme, sauf précision du contraire, l'appellation graphe désignera un graphe orienté simple.

Les graphes triviaux

Il existe deux graphes triviaux sur n noeuds donnés : Le *graphe nul*, noté $\mathcal{N}_n$, où $\mathcal{E} = \emptyset$, et le *graphe complet* ou *clique*, noté $\mathcal{K}_n$, où $(i, j) \in \mathcal{E},\ \forall i, j \in \mathcal{V}$.

La matrice d'adjacence

La structure des connections d'un graphe, appelée *topologie* du graphe, est donnée d'une manière "brute" grâce à sa *matrice d'adjacence* $A = (a_{ij})$ de dimensions $(n \times n)$:

$$a_{ij} := \begin{cases} 1 & \text{si } (i,j) \in \mathcal{E} \\ 0 & \text{sinon} \end{cases}$$

Notons que dans le cas d'un graphe non-orienté, cette matrice est symétrique.

Bien que cette matrice contienne toute l'information sur la topologie du graphe, elle n'est généralement pas idéale pour percevoir les particularités du graphe en question. En général, elle est utilisée comme ingrédient dans le calcul d'indices et de quantités qui, de leur côté, fournissent de meilleurs appréciations sur la topologie du graphe.

Les chemins ou chaînes

On définit un *chemin*, noté ξ, comme une séquence de noeuds $(i_0, \dots, i_\tau)$, reliés par des arcs, c-à-d $(i_t, i_{t+1}) \in \mathcal{E}, \forall t \in \{0, 1, \dots, \tau - 1\}$. Par abus de notation, on note $(i,j) \in \xi$ pour dire que le chemin ξ passe par l'arc (i,j). L'*ensemble des chemins* sur un graphe $\mathcal{G}$ se note $\mathcal{P}_\mathcal{G}$ et l'*ensemble des chemins entre i et j*, c-à-d $i_0 = i$ et $i_\tau = j$ se note $\mathcal{P}_{ij}$. On note parfois ξ_{ij} un chemin reliant i à j. Dans un graphe non-orienté, on parle de *chaînes*.

La *longueur d'un chemin* ξ, noté $l(\xi)$, est le nombre de sauts nécessaires pour aller du noeud initial au noeud final, c-à-d $l(\xi) = l((i_0, \dots, i_\tau)) = \tau$.

On nomme *circuit* un chemin ξ_{ii} qui commence et se termine par le même noeud i. Dans le cas des graphes non-orientés, on parle de *cycle*. Un circuit (ou cycle) est *simple* lorsqu'il passe au maximum une fois par chaque arc (arête).

Les plus courts chemins

Un *plus court chemin* entre deux noeuds i et j, noté ξ_{ij}^{sp}, est un chemin de longueur minimale, c'est-à-dire vérifiant :

$$\xi_{ij}^{sp} \in \arg\min_{\xi \in \mathcal{P}_{ij}} l(\xi)$$

Bien qu'il puisse exister plusieurs plus courts chemins entre deux sommets, la *distance du plus court chemin*, notée $d^{sp}(i,j)$, est unique :

$$d^{sp}(i,j) := \min_{\xi \in \mathcal{P}_{ij}} l(\xi)$$

Par convention, cette distance est infinie s'il n'existe pas de chemin entre i et j. On vérifie facilement que cette définition correspond à celle d'une distance (donnée à la section 2.1.2). Nous étudierons plus en détail cette distance dans la section 3.2.

La connectivité

Deux noeuds $i, j \in \mathcal{V}$ sont *faiblement connectés* s'il existe un chemin allant de i à j **ou** un chemin allant de j à i. Ces noeuds sont *fortement connectés* s'il existe un chemin allant de i à j **et** un chemin allant de j à i. Un ensemble constitué de tous les noeuds faiblement (respectivement fortement) connectés entre eux est appelée une *composante faiblement* (resp. *fortement*) *connexe* du graphe. Le graphe est appelé *faiblement* (resp. *fortement*) *connexe* s'il est composé

d'une seule composante faiblement (resp. fortement) connexe. Notons que la distinction entre faiblement et fortement connexe n'est pas nécessaire pour les graphes non-orientés.

Un graphe non-orienté connexe ne contenant aucun circuit simple s'appelle un *arbre*. Dans le cas d'un graphe orienté, on nomme ce dernier *arbre* si le graphe non-orienté correspondant (en supprimant l'orientation des arcs) est un arbre. Un graphe dont les composantes connexes sont des arbres s'appelle une *forêt*.

Dans la plupart des articles, on suppose qu'il existe toujours un chemin entre deux noeuds distincts, nous travaillerons donc généralement avec des graphes fortement connexes.

2.1.2 Les graphes spatiaux

Définition

Un graphe est dit *spatial* lorsque l'ensemble des noeuds $\mathcal{V}$ est muni d'une distance d. Le couple $(\mathcal{V}, d)$ forme ainsi un *espace métrique*. On définit la *distance* ou *métrique* d comme une application $d : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ vérifiant les propriétés suivantes :

1. Positivité : $d(i,j) \geq 0 \quad \forall i,j \in \mathcal{V}$
2. Séparation : $d(i,j) = 0 \Leftrightarrow i = j \quad \forall i,j \in \mathcal{V}$
3. Symétrie : $d(i,j) = d(j,i) \quad \forall i,j \in \mathcal{V}$
4. Inégalité triangulaire : $d(i,j) \leq d(i,k) + d(k,j) \quad \forall i,j,k \in \mathcal{V}$

Dans plusieurs applications concrètes, les noeuds sont des points du plan $\mathbb{R}^2$ muni de la distance Euclidienne classique.

La longueur des arcs

Bien que les arcs ne soient pas, à proprement parler, dans l'espace métrique, le fait que les noeuds possèdent une distance nous permet de définir la *longueur des arcs*, $l^d((i,j))$ de la façon suivante :

$$l^d((i,j)) := d(i,j)$$

C'est à travers cette longueur que la topologie d'un graphe spatial est affectée. En effet, il est généralement plus coûteux de créer un arc d'une longueur élevée ou de faire transiter une quantité à travers celui-ci. C'est à travers cette longueur que les graphes spatiaux prennent toute leur particularité.

La matrice de résistances

Les longueurs des arêtes sont généralement données grâce à la *matrice de résistances* (nommée de la sorte par analogie avec le formalisme électrique) du graphe, $R = (r_{ij})$, de dimensions $(n \times n)$:

$$r_{ij} := \begin{cases} l^d((i,j)) & \text{si } (i,j) \in \mathcal{E} \\ \infty & \text{sinon} \end{cases}$$

Notons que les résistances ne forment pas une métrique, car l'inégalité triangulaire n'est généralement pas vérifiée. Par exemple, on a $r_{ij} > r_{ik} + r_{kj}$, si $(i,j) \notin \mathcal{E}$ et $(i,k),(k,j) \in \mathcal{E}$.

Cette matrice de résistance est utilisée dans tous les articles. Cependant, dans ces articles, la définition de la matrice de résistance est plus large, car elle peut être définie d'une manière asymétrique, c-à-d $r_{ij} \neq r_{ji}$ même si (i,j) et (j,i) sont dans $\mathcal{E}$. Néanmoins, cette situation non symétrique ne correspond pas, à proprement parler, au cas des graphes spatiaux.

La longueur valuée d'un chemin

Sur un graphe spatial, il existe une deuxième version de la longueur d'un chemin, la *longueur valuée* (par la distance) $l^d(\xi)$, qui est en fait la somme de la longueur des arcs le composant :

$$l^d(\xi) = l^d((i_0, \dots, i_\tau)) := \sum_{t=0}^{\tau-1} l^d(i_t, i_{t+1}) = \sum_{t=0}^{\tau-1} r_{i_t i_{t+1}}$$

La longueur totale d'un graphe

On appelle *la longueur valuée totale d'un graphe*, aussi appelé son *coût*, la somme de toutes les longueurs de ses arcs :

$$R(\mathcal{G}) := \sum_{(i,j)\in\mathcal{E}} l^d((i,j)) = \sum_{(i,j)\in\mathcal{E}} r_{ij}$$

Les plus courts chemins valués

De manière similaire aux plus courts chemins, un *plus court chemin valué* ξ_{ij}^{dsp} vérifie :

$$\xi_{ij}^{dsp} \in \arg\min_{\xi\in\mathcal{P}_{ij}} l^d(\xi)$$

et la *distance du plus court chemin valuée*, notée $d^{dsp}(i,j)$, sera définie par :

$$d^{dsp}(i,j); = \min_{\xi\in\mathcal{P}_{ij}} l^d(\xi)$$

avec $d^{dsp}(i,j) = \infty$ s'il n'existe pas de chemin entre i et j. Ces distances seront également étudiées plus en détails dans le chapitre 3.2.

2.1.3 Les graphes pondérés

Définition

Un graphe *pondéré* est un graphe $\mathcal{G}$ muni d'une fonction de poids $c : \mathcal{E} \to \mathbb{R}^+$, qui associe un scalaire positif à tout arc. Ce poids pondère la force du lien : plus celui-ci est élevé, plus le lien entre les deux noeuds est fort.

La matrice des conductances

Encore par analogie avec le formalisme électrique, les poids sont contenus dans la *matrice des conductances* du graphe, $C = (c_{ij})$, de dimensions $(n \times n)$:

$$c_{ij} := \begin{cases} c((i,j)) & \text{si } (i,j) \in \mathcal{E} \\ 0 & \text{sinon} \end{cases}$$

Cette matrice est symétrique dans le cas d'un graphe non-orienté.

Généralement, dans la littérature, les poids peuvent être définis de deux manières différentes. Il existe les poids d'*attraction*, qui correspondent à la définition des poids que nous donnons ici, et les poids de *répulsion*, qui correspondent à la définition de la longueur des arcs. En associant l'attraction aux poids et la répulsion à la spatialité, nous prenons ici la décision de les définir d'une manière indépendante. Un graphe spatial et pondéré aura donc généralement une matrice des conductances autonome face à sa matrice des résistances. Cependant, il est également possible de définir, pour suivre complètement le formalisme électrique, les conductances comme l'inverse des résistances, c-à-d $c_{ij} := 1/r_{ij}$.

2.2 Les chaînes de Markov

Nous allons mettre de côté un moment les graphes pour définir un autre concept fortement utilisé tout au long de cette thèse : les *chaînes de Markov*. Bien que les graphes pondérés et les chaînes de Markov possèdent des liens forts les uns avec les autres, nous allons, en premier lieu, définir ces dernières d'une manière autonome. Les liens et les correspondances possibles entre ces deux concepts seront exposés à la fin de cette section. Il existe un grand nombre de résultats sur les chaînes de Markov et le but de cette section n'est pas d'en dresser une liste exhaustive. Nous ne verrons ici que les notions pertinentes dans le cadre de cette thèse. Un lecteur désireux d'en savoir plus peut néanmoins se reporter aux nombreux ouvrages sur le sujet, notamment les travaux de Kemeny et Snell (?, ?) et d'Aldous et Fill (?, ?), d'où sont tirés la plupart des résultats exposés ici. Il est également possible d'y trouver les preuves des résultats énoncés dans cette section.

2.2.1 Généralités

Définition

Une *chaîne de Markov* sur un espace d'états discrets $\mathcal{N}$ est un processus S_t à temps discret $t \in \mathbb{N}$, possédant la *propriété de Markov faible*, c'est-à-dire que la probabilité d'avoir le processus dans l'état $j \in \mathcal{N}$ au temps $t+1$ ne dépend que de l'état du processus au temps t :

$$\mathbb{P}(S_{t+1} = j|\ S_0 = i_0, S_1 = i_i, \ldots, S_{t-1} = i_{t-1}, S_t = i) = \mathbb{P}(S_{t+1} = j|S_t = i)$$

Tout au long de cette thèse, nous allons nous restreindre aux chaînes de Markov finies, c'est-à-dire définies sur un ensemble $\mathcal{N}$ de cardinalité n. Le processus est donc entièrement décrit par le vecteur de sa *distribution initiale*, $\pi^0 = (\pi^0_i)$, de taille n, et par sa *matrice de transition*, $W = (w_{ij})$ de taille $(n \times n)$:

$$\pi^0_i := \mathbb{P}(S_0 = i)$$
$$w_{ij} := \mathbb{P}(S_{t+1} = j|S_t = i)$$

Toute matrice carrée à coefficients positifs, qui vérifie $w_{i\bullet} = 1, \forall i \in \mathcal{N}$, définit une matrice de transition d'une chaîne de Markov.

On peut trouver la distribution de la chaîne au temps t, $\pi^t = (\pi^t_i) = (\mathbb{P}(S_t = i))$, grâce à la formule suivante :

$$\pi^t = \pi^0 W^t$$

La composante i, j de la matrice W^t se note $w^{(t)}_{ij}$.

Classification des états

Un état j est dit *accessible* à partir de i, noté $i \to j$, si :

$$\exists t \in \mathbb{N} \text{ tel que } w^{(t)}_{ij} > 0$$

Deux états i et j *communiquent* si et seulement si j est accessible de i **et** i est accessible de j. On le note $i \leftrightarrow j$.

On vérifie facilement que la relation $\leftrightarrow$ est réflexive, transitive et symétrique, et forme donc une *relation d'équivalence* entre les états. Cette relation d'équivalence permet de définir des *classes d'équivalences* $\mathcal{C}_i$ sur les états de la chaîne :

$$i_1 \leftrightarrow i_2 \qquad \forall i_1, i_2 \in \mathcal{C}_i$$

On peut définir l'*accessibilité entre deux classes d'équivalences* $\mathcal{C}_i$ et $\mathcal{C}_j$ par :

$$\mathcal{C}_i \to \mathcal{C}_j \iff \exists i \in \mathcal{C}_i \text{ et } \exists j \in \mathcal{C}_j \text{ tel que } i \to j$$

La relation $\to$ entre les classes d'équivalences est réflexive, transitive et antisymétrique, et forme donc une *relation d'ordre partielle* entre les classes d'équivalences de la chaîne.

Les *classes minimales*, par rapport à la relation d'ordre partielle $\to$, s'appellent les *classes récurrentes* de la chaîne. Les classes d'équivalence qui ne sont pas récurrentes se nomment les *classes transientes*. Les termes récurrents et transients s'appliquent également aux états appartenant aux classes correspondantes. Un état i est dit *absorbant* s'il est l'unique élément d'une classe récurrente, on a alors $w_{ij} = \delta_{ij}$.

Période et temps d'accès

La *période* d'un état i, notée θ_i, est définie comme le plus grand commun diviseur des temps nécessaires pour partir de cet état et y revenir. Formellement :

$$\theta_i = \text{PGCD}(\{t \geq 1 |\ w_{ii}^{(t)} > 0\})$$

La période est identique pour tous les états appartenant à la même classe d'équivalence. On nomme *apériodiques* les états ou classes ayant une période égale à 1.

Le *premier temps d'accès* de l'état j à partir de i, noté T_{ij}, est une variable aléatoire qui décrit le nombre minimal de transitions pour atteindre j en partant de i :

$$T_{ij} := \begin{cases} \min(\{t \geq 1 |\ w_{ij}^{(t)} > 0\}) & \text{si } \{t \geq 1 |\ w_{ij}^{(t)} > 0\} \neq \emptyset \\ \infty & \text{sinon} \end{cases}$$

Un état i est *récurrent positif* ssi l'espérance du temps de retour à i en partant de i est finie :

$$\mathbb{E}(T_{ii}) < \infty$$

Dans les chaîne de Markov finies, tous les états récurrents sont également récurrents positifs.

Classification des chaînes

Une chaîne de Markov est *irréductible* ssi elle est constituée d'une seule classe d'équivalence, c-à-d que tous les états communiquent entre eux.

Une chaîne de Markov est *apériodique* ssi la période de tous ses états est égale à 1.

Une chaîne de Markov est *régulière* ssi elle est irréductible et apériodique.

Une chaîne de Markov est *absorbante* ssi toutes ses classes récurrentes sont constituées d'un seul état absorbant.

La distribution stationnaire

On nomme *distribution stationnaire* de la chaîne définie par la matrice W, toute distribution π vérifiant :

$$\pi = \pi W$$

On sait qu'il existe au moins une distribution stationnaire π ssi la chaîne possède au moins un état récurrent positif. Pour cet état récurrent positif i, on aura $\pi_i > 0$.

Si une chaîne ne possède qu'une seule classe récurrente, on a alors une *unique* distribution stationnaire si et seulement si la chaîne possède un état récurrent positif. Dans ce cas, tous les états de cette classe récurrente sont récurrents positifs.

Dans une chaîne régulière définie par W, il existe une *unique* distribution stationnaire π et celle-ci peut être obtenue par :

$$\pi_j = \lim_{t\to\infty} w_{ij}^{(t)} \quad \forall i \in \mathcal{N}$$

On appelle une chaîne de Markov régulière *réversible* ssi sa distribution stationnaire π et sa matrice de transition W vérifient :

$$\pi_i w_{ij} = \pi_j w_{ji}$$

Le terme réversible vient du fait que si l'on observe le processus en remontant le temps, celui-ci reste identique.

La matrice fondamentale des chaînes absorbantes

On se place ici dans le cadre d'une chaîne de Markov *absorbante*. L'ensemble des états $\mathcal{N}$ de cette chaîne peut alors se séparer en deux parties disjointes : $\mathcal{Q}$, l'ensemble des états transients de la chaîne, et $\mathcal{A}$, l'ensemble des états absorbants. La matrice de transition W de cette chaîne peut alors s'écrire comme :

$$W = \left(\begin{array}{c||c|c} & \forall \mathtt{j} \in \mathcal{Q} & \forall \mathtt{j} \in \mathcal{A} \\ \hline\hline \forall \mathtt{i} \in \mathcal{Q} & Q & \Psi \\ \hline \forall \mathtt{i} \in \mathcal{A} & O & I \end{array}\right)$$

où I est la matrice identité de taille $|\mathcal{A}| \times |\mathcal{A}|$, O n'est constitué que de zéros, Q est la matrice du processus dans les états transients et Ψ est la matrice de transition entre les états transients et absorbants. Si le but est de trouver combien de temps le processus reste dans les états transients avant d'être absorbé, la distinction entre les différents états absorbants n'est plus importante. On peut donc considérer tous les états absorbants comme un seul et même état, et la matrice $\Psi = (\psi_{ij})$ devient alors un vecteur, noté $\alpha = (\alpha_i)$, avec $\alpha_i = \sum_{j\in\mathcal{A}} \psi_{ij}$.

La matrice M, définie par :

$$M := I + Q + Q^2 + Q^3 + \ldots = (I - Q)^{-1}$$

s'appelle la *matrice fondamentale* de la chaîne définie par W, dont les composantes, m_{ij}, peuvent être interprétés de la façon suivante :

$$m_{ij} = \mathbb{E}(\text{nombre de passages par } j \text{ en partant de } i \text{ avant d'être absorbé})$$

Il existe également une version de cette matrice fondamentale pour les chaînes réversibles, qui permet alors d'obtenir l'espérance des temps d'accès $\mathbb{E}(T_{ij})$, mais ce cas n'est pas nécessaire pour la suite de cette thèse et n'est donc pas exposé ici. On peut néanmoins trouver sa définition dans (?, ?, ?).

2.2.2 Liens entre chaînes de Markov et graphes

Les chaîne de Markov et les graphes pondérés sont deux concepts très proches : une chaîne peut facilement être représentée par un graphe, l'espace des états $\mathcal{N}$ devenant alors l'ensemble des noeuds $\mathcal{V}$. Il est également aisé de définir une chaîne de Markov entre les noeuds d'un graphe.

Passage d'une chaîne de Markov à un graphe

Une chaîne de Markov peut être représentée par un graphe orienté, pondéré, mais pas nécessairement simple (des boucles existent). L'espace des états $\mathcal{N}$ devient alors l'ensemble des noeuds $\mathcal{V}$, la distribution π^0 reste identique et les transitions sont représentées par des arcs pondérés, $c_{ij} = w_{ij}$.

Lorsque l'on a une chaîne de Markov irréductible, qui possède donc une distribution stationnaire π unique, on peut également définir les conductances C du graphe avec $c_{ij} = \pi_i w_{ij}$. Si la chaîne est en outre réversible, on aura alors :

$$c_{ij} = c_{ji} \geq 0 \quad c_{i\bullet} = \pi_i \quad c_{\bullet j} = \pi_j \quad c_{\bullet\bullet} = 1$$

et C peut alors s'interpréter comme une *matrice d'échange* (?, ?, ?), représentant un flux symétrique et normalisé qui définit l'interaction entre les états de la chaîne. Le graphe résultant est dans ce cas non-orienté.

Passage d'un graphe à une chaîne de Markov

Il également possible d'associer une chaîne de Markov à un graphe orienté, qui sera alors nommée *marche aléatoire*. On pose π^0 comme on le désire et $W = (w_{ij})$ doit vérifier, $w_{ij} = 0$ si $(i, j) \notin \mathcal{E}$.

Lorsque l'on a un graphe pondéré, celui-ci peut induire une matrice de transition $W = (w_{ij})$ entre les noeuds de la manière suivante :

$$w_{ij} := \begin{cases} \frac{c_{ij}}{c_{i\bullet}} & \text{si } c_{i\bullet} > 0 \\ \delta_{ij} & \text{sinon} \end{cases}$$

Définir W de cette manière nous garantit que $w_{ij} \geq 0$ et $w_{i\bullet} = 1$, et donc que W est bien une matrice de transition d'une chaîne de Markov. On appellera cette matrice de transition *matrice induite par les conductances*.

On appelle une marche aléatoire *simple*, une marche aléatoire qui, en chaque noeud, donne la même probabilité de choisir un des arcs (ou arêtes) sortant de ce noeud. Cela correspond à une marche aléatoire définie par une matrice de transition induite par des conductances constantes.

Lorsque le graphe est non-orienté, c-à-d lorsque la matrice des conductances est symétrique, on obtient alors une chaîne de Markov réversible. La distribution stationnaire $\pi = (\pi_i)$ de cette chaîne peut être obtenue par :

$$\pi_i = \frac{c_{i\bullet}}{c_{\bullet\bullet}}$$

2.3 Les flux

Il est temps à présent de définir l'ingrédient principal de cette thèse, c'est-à-dire les *flux*. Dans la vie quotidienne, ce mot est parfois utilisé pour désigner des gens, de la matière ou des marchandises se déplaçant d'un endroit à un autre. Nous verrons que la définition mathématique du flux dans les graphes correspond à ce concept. Le formalisme du flux utilisé ici est spécifique à cette thèse, et les résultats obtenus sont contenus dans les articles présentés. Notons cependant que ce formalisme s'inspire de nombreuses sources, en particulier des travaux de Ravindra Ahuja, Thomas Magnanti et James Orlin (?, ?).

2.3.1 Généralités

Le flux

Dans un graphe $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, un flux est donné par une matrice $X = (x_{ij})$, où $x_{ij} \geq 0$ désigne une quantité de matière se déplaçant par unité de temps du noeud i au noeud j.

Les sources et les cibles

Les noeuds créant du flux sont nommés les *sources* ou les *origines*, et leur ensemble se note $\mathcal{S}$. De leur côté, les noeuds absorbant du flux s'appellent les *cibles* ou les *destinations*, et leur ensemble se note $\mathcal{T}$. On a $\mathcal{S}, \mathcal{T} \in \mathcal{V}$ et on suppose que $\mathcal{S} \cap \mathcal{T} = \emptyset$, c'est-à-dire qu'un noeud ne peut pas être à la fois source et cible. Si, dans un exemple concret, une quantité de matière est à la fois créée et absorbée en un noeud, il faudra alors faire le bilan et, suivant son signe, on désignera ce noeud comme source ou comme cible. Par convention, nous posons que le flux total émis par les sources est égal au flux total absorbé par les cibles. Nous pouvons donc poser, sans perte de généralité, le flux total entrant et sortant comme égal à 1, ce qui revient à travailler avec des proportions. La proportion de flux créée par chaque noeud est décrite par un vecteur de taille n, $f = (f_i)$, avec $f_i \geq 0$ et $f_\bullet = 1$. De manière similaire, la proportion de flux absorbée par les noeuds est définie avec un autre vecteur de taille n, $\rho = (\rho_i)$, avec $\rho_i \geq 0$ et $\rho_\bullet = 1$. Le fait qu'aucun noeud ne puisse à la fois être source et cible s'exprime par $f_i \rho_i = 0,\ \forall i \in \mathcal{V}$.

Dans le chapitre 3, nous allons nous intéresser aux flux créés par une seule source et absorbés par une seule cible. Bien que ce cas de figure soit en réalité un cas particulier du formalisme contenant de multiples sources et cibles, les dérivations mathématiques permettant le calcul du flux s'en trouvent grandement simplifiées.

Les flux admissibles

Pour un graphe simple orienté donné, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, avec des sources et des cibles définies par f et ρ, un flux $X = (x_{ij})$ est *admissible* s'il vérifie les conditions suivantes :

1. Positivité : $x_{ij} \geq 0 \quad \forall i, j \in \mathcal{V}$
2. Déplacement sur les arcs : $x_{ij} = 0 \quad \forall (i,j) \notin \mathcal{E}$
3. Conservation : $x_{i\bullet} - x_{\bullet i} = f_i - \rho_i \quad \forall i \in \mathcal{V}$

Lorsque la source et la cible sont uniques (chapitre 3), la condition de conservation peut s'écrire comme : $x_{i\bullet} - x_{\bullet i} = \delta_{is} - \delta_{it},\ \forall i \in \mathcal{V}$.

L'*ensemble des flux admissibles* est noté $\mathcal{X}$. On peut facilement vérifier que si l'on a deux flux admissibles $X, Y \in \mathcal{X}$, alors leur mélange convexe est également admissible, c-à-d $\alpha X + (1-\alpha)Y \in \mathcal{X},\ \forall \alpha \in [0,1]$. L'ensemble des flux admissibles est donc un *ensemble convexe*.

La marche aléatoire définie par un flux admissible

Lorsque l'on a un flux admissible sur un graphe, celui-ci définit une chaîne de Markov sur ce graphe, il suffit pour cela de poser comme distribution initiale $\pi^0 = f$, et de définir la matrice de transition, $W^X = (w_{ij}^X)$, comme :

$$w_{ij}^X := \begin{cases} \frac{x_{ij}}{x_{i\bullet}} & \text{si } x_{i\bullet} > 0 \\ \delta_{ij} & \text{sinon} \end{cases}$$

On dira que le flux *suit* la chaîne de Markov définie par W^X.

2.3.2 Les flux minimisant une fonctionnelle

Généralement, sur un graphe donné, nous recherchons des flux admissibles *optimaux*, c'est-à-dire des flux admissibles minimisant une certaine fonctionnelle. Le choix de cette fonctionnelle est fondamental pour obtenir un flux avec des caractéristiques désirées. Nous allons voir ici quelques exemples de fonctionnelles à minimiser.

L'énergie

Lorsque le graphe est spatial, donc muni d'une matrice de résistances $R = (r_{ij})$, on appelle *énergie* ou *coût* du flux la fonctionnelle $U(X)$, définie par :

$$U(X) = U(X||R) := \sum_{i,j \in \mathcal{V}} r_{ij} x_{ij}$$

Comme cette fonctionnelle est linéaire, il existe généralement plusieurs flux admissibles la minimisant. Nous verrons que les flux minimisant l'énergie suivent les chemins du *transport optimal* (?, ?, ?, ?, ?, ?) entre les sources et les cibles, c'est-à-dire les chemins de moindre coût, relativement à R, permettant de respecter les contraintes de création et de destruction fixées aux sources et aux cibles. Nous reviendrons sur le problème du transport optimal dans le chapitre 4.

L'énergie électrique

Dans le cas d'un graphe non-orienté pondéré, c'est-à-dire muni de $C = (c_{ij})$ symétrique, on appelle l'*énergie électrique* du flux la fonctionnelle $U^2(X)$, définie par :

$$U^2(X) = U^2(X||C) := \sum_{i,j \in \mathcal{V}} \frac{(x_{ij} - x_{ji})^2}{c_{ij}}$$

Les flux admissibles minimisant cette fonctionnelle ne sont également pas uniques. Pour s'en convaincre, il suffit de voir que deux flux admissibles, x_{ij} et $x_{ij} + c$, avec $c > 0$ une constante, auront la même énergie électrique. Par contre, le *flux net*, $y_{ij} := x_{ij} - x_{ji}$ sera unique, car $\sum_{i,j \in \mathcal{V}} y_{ij}^2 / c_{ij}$ est évidemment convexe par rapport à y_{ij}. Ce flux $Y = (y_{ij})$ décrit le *flux électrique orienté*, en supposant que le graphe est un circuit électrique possédant des conductances C, ainsi que des entrées et sorties de courant définies respectivement par f et ρ (?, ?). Notons cependant que Y n'est pas un flux admissible. Dans le cadre d'un graphe non-orienté spatial, les conductances peuvent être définies grâce aux résistances $c_{ij} = 1/r_{ij}$.

L'entropie

Si l'on a un graphe muni d'une matrice de chaîne de Markov $W = (w_{ij})$, on appelle alors *l'entropie* du flux par rapport à W la quantité $G(X)$, définie comme :

$$G(X) = G(X||W) := \sum_{i,j\in\mathcal{V}} x_{ij} \ln \frac{x_{ij}}{x_{i\bullet} w_{ij}} = x_{\bullet\bullet} \sum_{i\in\mathcal{V}} \frac{x_{i\bullet}}{x_{\bullet\bullet}} K_i(X||W)$$

où $K_i(X||W) := \sum_{j\in\mathcal{V}} \frac{x_{ij}}{x_{i\bullet}} \ln \frac{x_{ij}}{x_{i\bullet} w_{ij}} \geq 0$ est la *divergence de Kullback-Leibler*. Cette divergence mesure, en chaque noeud i du réseau où le flux sortant est non-nul, la divergence entre la chaîne de Markov définie par le flux, $x_{ij}/x_{i\bullet}$, et la distribution de référence w_{ij} (on aura $K_i(X||W) = 0$ ssi $x_{ij}/x_{i\bullet} = w_{ij}, \forall j \in \mathcal{V}$). On fait ensuite la moyenne de cette divergence, pondérée par le nombre de passages en chaque noeud, c-à-d $\sum_{i\in\mathcal{V}} \frac{x_{i\bullet}}{x_{\bullet\bullet}} K_i(X||W)$. Finalement, cette moyenne est multipliée par $x_{\bullet\bullet}$ pour que cette fonctionnelle devienne *homogène*, c'est-à-dire que $G(\nu X) = \nu G(X), \forall \nu > 0$. Si un flux existe sur (i, j) alors que $w_{ij} = 0$, cette divergence est infinie.

On observe facilement que la matrice hessienne de cette fonctionnelle est définie positive, ce qui signifie que cette fonctionnelle est convexe sur $\mathcal{X}$, c'est-à-dire qu'elle vérifie $G(\alpha X + (1-\alpha)Y) \leq \alpha G(X) + (1-\alpha)G(Y),\ \forall \alpha \in [0, 1]$ et $\forall X, Y \in \mathcal{X}$. Comme $\mathcal{X}$ est également convexe, il n'existe qu'un flux admissible minimisant l'entropie. Ce flux doit suivre au possible la chaîne de Markov W, c'est-à-dire être proche de vérifier $x_{ij}/x_{i\bullet} = w_{ij},\ \forall i, j \in \mathcal{V}$, tout en respectant les contraintes. Notons qu'il est possible de définir, comme vu à la section 2.2.2, w_{ij} grâce aux conductances C d'un graphe pondéré, c-à-d $w_{ij} = c_{ij}/c_{i\bullet}$ si $c_{i\bullet} > 0$, et $w_{ij} = \delta_{ij}$ sinon.

L'énergie libre

Dans un graphe spatial pondéré, l'*énergie libre* se définit comme un mélange entre l'énergie et l'entropie. Cette fonctionnelle possède un paramètre ajustable, la *température*, notée $T > 0$, qui donne plus ou moins d'importance à l'entropie relativement à l'énergie (parfois, on utilisera la *température inverse*, $\beta := 1/T$). Formellement, l'*énergie libre* est la fonctionnelle $F(X) = F(X||R, W)$, définie par :

$$\begin{aligned} F(X) = F(X||R, W) &:= U(X||R) + TG(X||W) \\ &= \sum_{i,j\in\mathcal{V}} r_{ij} x_{ij} + T \sum_{i,j\in\mathcal{V}} x_{ij} \ln \frac{x_{ij}}{x_{i\bullet} w_{ij}} \end{aligned}$$

Nous avons vu que sur $\mathcal{X}$, $U(X)$ est linéaire et $G(X)$ est convexe. $F(X)$ est donc également convexe sur $\mathcal{X}$ et le flux admissible minimisant cette fonctionnelle est unique. Ce dernier suit les chemins de transport optimal dans le cas $T \to 0$, et la marche aléatoire définie par W dans le cas $T \to \infty$. Pour des températures intermédiaires, ce flux aura un comportement mixte. Nous appelons ce flux : le *flux de transport randomisé*. Il sera l'ingrédient principal tout au long de cette thèse.

L'énergie libre peut également être interprétée comme une fonctionnelle *régularisant* l'énergie. Cette dernière étant linéaire, elle nécessite des algorithmes peu efficients pour trouver un flux la minimisant, tel que l'*algorithme du simplexe* ou l'*algorithme des points intérieurs*. En introduisant un terme d'entropie, la fonctionnelle devient dérivable et on peut trouver le flux optimal en réduisant drastiquement le temps de calcul (voir section 4.3). En posant une température suffisamment petite, la solution du problème régularisé donnera une bonne approximation d'une des solutions optimales du problème linéaire initial.

L'idée d'ajouter une part d'aléatoire dans le comportement d'un agent se déplaçant sur un réseau n'est pas nouvelle. Au début du XX$^{\text{e}}$ siècle, les modèles visant à simuler le trafic sur un réseau étaient complètement déterministes et consistaient en une superposition de plusieurs plus courts chemins. Cependant, on remarqua rapidement que les comportements des agents sur un

réseau n'étaient généralement pas déterministes et qu'une part d'aléatoire était nécessaire. C'est en 1971 que Robert Dial (?, ?) proposa un premier modèle probabiliste d'attribution de chemins semi-aléatoires entre une source et une cible uniques. Ce modèle se base sur la *vraisemblance* de prendre un arc, qui est plus élevée si l'on s'éloigne au plus de la source et qu'on se rapproche au mieux de la cible. Bien que posant les bases de ce qui va devenir par la suite les modèles probabilistes de trafic actuels, ce premier modèle n'est pas très efficient, car il nécessite de calculer les distances du plus court chemin de tous les noeuds du graphe à la source et à la cible. Ce n'est qu'en 1995 que Michael Bell (?, ?) proposa un modèle très similaire permettant d'éviter ce calcul des plus courts chemins et il fut suivi de près par une série d'articles de Takashi Akamatsu (?, ?, ?, ?) qui finalisèrent de poser les bases des modèles probabilistes d'affectation du trafic. Aujourd'hui, ces modèles semi-aléatoires de déplacement sont très bien représentés par le modèle des *plus courts chemins randomisés* (en anglais, *randomized shortest-paths* ou *RSP*), construit dans une série d'articles publiés par Marco Saerens et son équipe (?, ?, ?, ?, ?, ?). Nous comparerons le modèle des plus courts chemins randomisés avec notre modèle de flux de transport randomisé à la fin du chapitre suivant (voir section 3.3).

3 Source et cible uniques

Dans ce chapitre, nous allons voir comment construire des indices de centralité et des dissimilarités sur des graphes à partir du flux de transport randomisé possédant une seule source et une seule cible, respectivement notées s et t. Lorsque des noeuds uniques sont définis comme source et cible, le flux de transport randomisé explore la structure du graphe séparant ces deux noeuds, avec une trajectoire qui dépend de la valeur de la température. Si la température est basse, la flux de transport randomisé passe par les plus courts chemins existant entre s et t, et n'explore pas le graphe dans sa totalité. A l'opposé, lorsque la température est élevée, le flux est aléatoire et passe par tous les arcs du graphe d'une manière aléatoire. Ces deux cas de figures sont connus dans la littérature, et chacun d'eux contient une différente part de l'information sur la topologie du graphe. En prenant une température intermédiaire, le flux résultant se comporte de manière hybride et le mélange des informations contenues dans ce flux donne alors une appréciation plus complète de la topologie du graphe qui sépare s et t.

Dans la première section de ce chapitre, nous allons montrer comment construire des indices de centralité à partir du flux de transport randomisé. Pour ce faire, il suffit d'observer quel est le flux moyen passant par chaque noeud, ou chaque arc, avec toutes les paires possibles de source et cible. On obtient alors la fréquentation moyenne des éléments constituant le graphe, ce qui peut être interprété comme un indice de centralité. Lorsque la température est aux extrémités de son spectre, nous verrons que les indices de centralité construits de cette façon correspondent à des indices connus dans la littérature. En revanche, dans les cas où la température est intermédiaire, ces nouveaux indices mettent en valeur des noeuds ou des arêtes jusque-là négligés.

La deuxième section couvre la construction de dissimilarités à partir du flux de transport randomisé. Ces dissimilarités s'obtiennent d'une manière relativement aisées, en prenant, par exemple, le distance moyenne parcourue par le flux pour aller de s à t. Encore une fois, nous verrons que ces nouvelles dissimilarités correspondent à des dissimilarités déjà étudiées lorsque des températures extrêmes sont utilisées. Cependant, lorsque la température est intermédiaire, elles ont de nouvelles propriétés et donnent de meilleurs résultats lors de leur utilisation dans des algorithmes de classification.

La dernière section de ce chapitre porte sur un formalisme légèrement différent développé dans une série d'article (?, ?, ?, ?, ?, ?) permettant d'obtenir des résultats similaires. Ce formalisme n'est pas basé sur le flux, mais sur des *probabilités de chemin*, et nous nous efforcerons d'étudier les analogies qui existent entre ces deux approches.

Bien qu'il soit possible d'utiliser le formalisme de flux avec de multiples sources et cibles, celui-ci se simplifie grandement lorsque la cible est unique. En effet, dans ce cas de figure, la cible devient un noeud *absorbant* et, nous le verrons dans les articles, il est nécessaire d'avoir un noeud de ce type pour effectuer les dérivations nous permettant d'obtenir le flux minimisant l'énergie libre.

3.1 Construction d'indices de centralité

Un des principaux objectifs de l'étude des réseaux est de déterminer l'importance des différents noeuds, ou arêtes, par rapport à la structure globale du réseau. L'importance d'un noeud ou d'une arête peut prendre plusieurs facettes : un noeud fortement connecté au reste du réseau est important en raison de son rôle de carrefour, tel les importantes plateformes aéroportuaires, où transitent la plupart des vols internationaux ; une arête qui est le seul lien entre deux parties du réseau est importante, car sans elle, le réseau se diviserait en deux compostantes connexes ; ou encore, une arête faisant office de raccourci est importante, car elle diminue la distance du plus court chemin entre un grand nombre de paires de noeuds. Dans un graphe, cette importance prend le nom de *centralité*. Cette centralité est évaluée grâce aux *indices de centralité*, dont le but est de classer les noeuds ou les arêtes en fonction de certains critères définissant leur importance.

Il existe une grande quantité d'indices de centralité, et en faire une liste exhaustive n'est pas le but de ce chapitre. Néanmoins, dans la section suivante, nous allons passer en revue les indices les plus connus, afin d'observer les similarités existantes entre ces derniers et ceux créés grâce au flux de transport randomisé. Notons que les indices de centralité seront présentés ici dans leur version *valuée*, c'est-à-dire lorsqu'ils sont appliqués à un graphe spatial. La version plus traditionnelle de ces indices est en réalité non-valuée, mais celle-ci n'est qu'un cas particulier des indices valués, en posant la longueur de tous les arcs comme étant égale à un.

3.1.1 Quelques indices de centralité

Le degré

Une des premières mesures de centralité sur les noeuds, et certainement la plus aisée à obtenir, est le *degré*. Celui-ci se définit comme le nombre d'arêtes incidentes à un noeud (ou d'arcs, mais dans ce cas il existe deux degrés différents : le degré entrant et le degré sortant). Cet indice ne prend donc en compte que le voisinage direct de chaque noeud et nous permet d'évaluer la capacité du noeud à répandre ou à intercepter des objets circulant sur le réseau. Malgré sa simplicité, c'est un ingrédient essentiel dans l'étude des réseaux et il est toujours fortement utilisé de nos jours. Dans le cas non-orienté, on note le *degré* : $\deg(i)$. Dans le cas orienté, le *degré entrant* et le *degré sortant* se notent respectivement : $\deg^+(i)$ et $\deg^-(i)$. Ils sont définis par :

$$\begin{aligned} \deg(i) &:= a_{i\bullet} = a_{\bullet i} && \text{(graphe non-orienté)} \\ \deg^+(i) := a_{\bullet i} \quad \deg^-(i) &:= a_{i\bullet} && \text{(graphe orienté)} \end{aligned}$$

La centralité d'intermédiarité

Les indices de *centralité d'intermédiarité* (en anglais : *betweeness centrality*) sont une famille d'indices fortement utilisée, en particulier en sciences sociales. Le plus emblématique d'entre eux est sans doute l'indice de *centralité d'intermédiarité des plus court chemins* (en anglais : *shortest-path betweeness*), développé par Linton Freeman en 1977 (?, ?). Cet indice mesure la proportion de fois qu'un noeud (ou une arête) est fréquenté, si l'on suppose que toutes les paires de noeuds communiquent entre elles via les plus courts chemins valués. Formellement, on note

cet indice pour un noeud : $B^{dsp}(i)$; et pour un arc : $B^{dsp}((i,j))$; définis par :

$$B^{dsp}(i) := \frac{1}{n(n-1)} \sum_{s,t|s\neq t} \frac{|\{\xi_{st}^{dsp} | i \in \xi_{st}^{dsp}\}|}{|\{\xi_{st}^{dsp}\}|}$$

$$B^{dsp}((i,j)) := \frac{1}{n(n-1)} \sum_{s,t|s\neq t} \frac{|\{\xi_{st}^{dsp} | (i,j) \in \xi_{st}^{dsp}\}|}{|\{\xi_{st}^{dsp}\}|}$$

où $|\{\xi_{st}^{dsp} | i \in \xi_{st}^{dsp}\}|$ et $|\{\xi_{st}^{dsp} | (i,j) \in \xi_{st}^{dsp}\}|$ sont le nombre de plus courts chemins valués entre s et t, passant respectivement par i et par (i,j). $|\{\xi_{st}^{dsp}\}|$ désigne le nombre de plus courts chemins valués total entre s et t. Il existe en réalité plusieurs versions de cet indice, suivant si le noeud observé, ou les noeuds composant l'arc observé, sont contenus ou non dans la sommation. En général, ce choix n'affecte pas les rangs de centralité des éléments du graphe (?, ?), et nous présentons ici uniquement la version où la somme inclut les éléments observés.

Une autre mesure de centralité d'intermédiarité est celle de l'*intermédiarité des chemins aléatoires*, appelée aussi l'*intermédiarité du courant électrique* (en anglais : *random-walk betweeness* ou *current-flow betweeness*), proposée par Newman en 2005 (?, ?). Cet indice de centralité est similaire à celui de Freeman, sauf qu'il comptabilise la proportion de passages par un noeud, ou une arête, lorsque les noeuds communiquent entre eux en suivant une marche aléatoire. Cette marche aléatoire est connue pour être similaire, comme nous le verrons plus tard, au courant électrique, d'où sa seconde appellation. Cet indice est noté pour un noeud : $B^{rw}(i)$; et pour un arc : $B^{rw}((i,j))$; définis par :

$$B^{rw}(i) := \frac{1}{n(n-1)} \sum_{s,t|s\neq t} \left(\sum_j |x_{ij}^{st} - x_{ji}^{st}| \right)$$

$$B^{rw}((i,j)) := \frac{1}{n(n-1)} \sum_{s,t|s\neq t} |x_{ij}^{st} - x_{ji}^{st}|$$

où x_{ij}^{st} est un flux suivant une marche aléatoire définie par W commençant en s et se terminant en t, c-à-d $x_{ij}^{st}/x_{i\bullet}^{st} = w_{ij}$ et $x_{i\bullet}^{st} - x_{\bullet i}^{st} = \delta_{is} - \delta_{it}$. Comme vu précédemment, le flux X vérifiant ces conditions n'est pas unique, mais $y_{ij}^{st} = x_{ij}^{st} - x_{ji}^{st}$ l'est. En général, cet indice est défini par rapport à la marche aléatoire simple, c-à-d avec $w_{ij} := 1/\deg^+(i),\ \forall (i,j) \in \mathcal{E}$.

La centralité de proximité

Une autre mesure de centralité pour les noeuds, proposée par Alex Bavelas en 1950 (?, ?), est la *centralité de proximité* (en anglais : *closeness centrality*), qui peut être définie comme celle de la proximité d'un noeud par rapport à tous les autres. Cette proximité est calculée comme l'inverse de l'*excentricité* $E(i)$ d'un noeud, c'est-à-dire la somme de tous les plus courts chemins du noeud i aux autres. On note cet indice $P(i)$, défini par :

$$P(i) := \frac{1}{E(i)} = \frac{1}{\sum_{t\in\mathcal{V}} d^{dsp}(i,t)}$$

où $d^{dsp}(i,t)$ est la distance du plus court chemin valuée entre i et t. Cette centralité pose un problème lorsqu'il n'existe pas de chemin entre i et t, car cette distance est alors infinie et cet indice de centralité sera nul. Pour résoudre ce problème, Yannick Rochat proposa en 2008 (?, ?) un indice de *centralité de proximité harmonique* (en anglais : *harmonic closeness centrality*), noté $P^h(i)$ et défini par :

$$P^h(i) := \sum_{t\in\mathcal{V}, t\neq i} \frac{1}{d^{dsp}(i,t)}$$

3.1.2 Article 1 : "Interpolating between Random Walks and Shortest Paths : a Path Functional Approach"

Ce premier article est paru dans *Social Informatics* (pp. 68-81), Springer Berlin Heidelberg en 2012. Il fait suite à une présentation donnée à la conférence SocInfo en 2012 à l'Ecole Polytechnique Fédérale de Lausanne.

Cet article propose plusieurs indices de centralité basés sur le flux de transport randomisé. Ces indices sont en réalité des généralisations de certains indices déjà étudiés et on peut voir qu'ils révèlent des noeud précédemment négligés. Cet article est important pour la suite de cette thèse, car il a été publié en premier et pose les bases du formalisme pour le flux de transport randomisé avec une source et une cible uniques. Ce formalisme sera réutilisé, puis généralisé dans les articles qui vont suivre.

Points-clés

- Solution du flux de transport randomisé $X^{st} = (x_{ij}^{st})$ pour une source s et une cible t (pp. 24-25).
- Interprétation probabiliste du flux (pp.25-26).
- Etude du flux aux limites $T \to \infty$ et $T \to 0$ (p.26).
- Définition de la *centralité du flux moyen* pour les arcs $\langle x_{ij} \rangle$ et pour les noeuds $\langle x_{i\bullet} \rangle$ (p.29).
- Définition de la *centralité du flux moyen normalisée* pour les arcs c_{ij} et pour les noeuds c_i (p.30).
- Définition de la *centralité du flux net moyen* pour les arcs $\langle \nu_{ij} \rangle$ et pour les noeuds $\langle \nu_{i\bullet} \rangle$ (p.30).
- Études de ces indices sur divers graphes (pp.30-32).
- Suggestion de deux indices généralisant l'indice de centralité de proximité, T_s^{out} et T_t^{in} (p.33).

Remarque

Cet article définit l'énergie comme $U(X) := \sum_{ij} r_{ij}\varphi(x_{ij})$, avec $\varphi(x)$ une fonction dérivable non-décroissante, et propose un algorithme pour trouver un flux de transport randomisé dans ce cas. En réalité, il n'est pas toujours possible de trouver la solution, en particulier, pour $\varphi(x) = x^2$, car l'algorithme proposé ne converge pas. C'est pourquoi, au cours de cette thèse, nous ne considérons que le cas $\varphi(x) = x$. Ce problème pourrait être résolu en appliquant un autre algorithme permettant de trouver le minimum d'une fonctionnelle, par exemple avec un algorithme de type *quasi-Newton* (?, ?). Cette préoccupation fait partie des pistes de recherche à explorer.

Bavaud, F., Guex, G.: *Interpolating between Random Walks and Shortest Paths: a Path Functional Approach*. In: Aberer, K. et al. (Eds.) SocInfo 2012, LNCS 7710, pp. 68–81. Springer, Heidelberg (2012)

Interpolating between Random Walks and Shortest Paths: a Path Functional Approach

François Bavaud and Guillaume Guex *

Department of Computer Science and Mathematical Methods
Department of Geography
University of Lausanne, Switzerland

Abstract. General models of network navigation must contain a deterministic or drift component, encouraging the agent to follow routes of least cost, as well as a random or diffusive component, enabling free wandering. This paper proposes a thermodynamic formalism involving two path functionals, namely an energy functional governing the drift and an entropy functional governing the diffusion. A freely adjustable parameter, the temperature, arbitrates between the conflicting objectives of minimising travel costs and maximising spatial exploration. The theory is illustrated on various graphs and various temperatures. The resulting optimal paths, together with presumably new associated edges and nodes centrality indices, are analytically and numerically investigated.

1 Introduction

Consider a network together with an agent wishing to move (or wishing to move goods, money, information, etc.) from source node s to target node t. The agent seeks to minimise the total cost or duration of the move, but the ideal path may be difficult to realise exactly, in absence of perfect information about the network.

The above context is common to many behavorial and decision contexts, among which "small-word" social communications (Travers and Milgram 1969), spatial navigation (e.g. Farnsworth and Beecham 1999), routing strategy on internet networks (e.g. Zhou 2008, Dubois-Ferrière et al. 2011), and several others (e.g. Borgatti 2005; Newman 2005).

Trajectories can be coded, generally non-univocally, by $X = (x_{ij})$ where x_{ij} = "number of direct transitions from node i to node j". The use of the flow matrix X is central in Operational Research (e.g. Ahuja et al. 1993) and Markov Chains theory (e.g. Kemeny and Snell 1976); four optimal st-paths have in particular been extensively analysed *separately* in the litterature, namely the shortest-path, the random walk, the maximum flow (Freeman et al. 1991) and the electrical current (Kirchhoff 1850; Newman 2005; Brandes and Fleischer 2005).

* The specific remarks of two anonymous reviewers are gratefully acknowledged

This paper investigates the properties of *st*-paths resulting from the minimisation of a *free energy functional* $F(X)$, over the set $X \in \mathcal{X}$ of admissible solutions. $F(X)$ contains a resistance component privileging shortest paths, and an entropy component favouring random walks. The conflict is arbitrered by a continuous parameter $T \geq 0$, the *temperature* (or its *inverse* $\beta := 1/T$), and results in an analytically solvable unique optimum *continuously interpolating* between shortest-paths and random walks. See Yen et al. (2008) and Saerens and al. (2009) for a close proposal, yet distinct in its implementation.

Section 2 introduces the formalism, in particular the *energy functional* (based upon an edge resistance matrix R, symmetrical or not) and the *entropy functional* (based upon a Markov transition matrix W, reversible or not, related to R *or not*). Section 2.5 provides the analytic form of the unique solution minimising the free energy. Section 4 proposes the definition of edge and vertex betweenness centrality indices directly based upon the flow X. They are illustrated in sections 3 and 5 for various network geometries at various temperatures.

2 Definitions and solutions

2.1 Admissible paths

Consider a connected graph $G = (V, E)$ involving $n = |V|$ nodes together with two distinguished and distinct nodes, the source s and target t. The *st*-path or flow matrix, noted $X^{st} = (x_{ij}^{st})$ or simply $X = (x_{ij})$, counts the number of transitions from i to j along conserved unit paths starting at s, possibly visiting s again, and absorbed at t. Hence

$$x_{ij} \geq 0 \qquad \text{positivity} \tag{1}$$

$$x_{i\bullet} - x_{\bullet i} = \delta_{is} - \delta_{it} \qquad \text{unit flow conservation} \tag{2}$$

where δ_{ij} is the Kronecker delta, the components of the identity matrix. Here and in the sequel, $\bullet$ denotes the summation over the values of the replaced index, as in $x_{i\bullet} = \sum_{j=1}^{n} x_{ij}$. In particular, $x_{s\bullet} = x_{\bullet s} + 1$. Also,

$$x_{t\bullet} = 0 \qquad \text{absorbtion at } t \tag{3}$$

entailing $x_{tj} = 0$ for all j, and $x_{\bullet t} = 1$. Normalisation (2) can be extended to *valued flows*

$$x_{i\bullet} - x_{\bullet i} = v(\delta_{is} - \delta_{it}) \qquad \text{conservation for valued flow} \tag{4}$$

where $v \geq 0$, the amount sent through the network, is the *value* of the flow. Further familiar constraints consist of

$$x_{ij} \leq c_{ij} \qquad \text{capacity, where } c_{ij} \geq 0 \tag{5}$$

$$x_{ij} \geq b_{ij} \qquad \text{minimum flow requirement, } b_{ij} \geq 0 \tag{6}$$

$$x_{\bullet j_0} = 0 \qquad \text{forbidden node } j_0 \tag{7}$$

$$x_{i_0 j_0} = 0 \qquad \text{forbidden arc } (i_0 j_0)\ . \tag{8}$$

2.2 Mixtures and convexity

Any of the above constraints (1) to (8) or combinations thereof defines a *convex* set $\mathcal{X}$ of admissible *st*-paths: if X and Y are admissible, so is their *mixture* $\alpha X + (1-\alpha)Y$ for $\alpha \in [0,1]$. Mixture of paths are generally non-integer, and can be given a probabilistic interpretation, as in

- $x_{\bullet\bullet}$ = "average time (number of transitions) for transportation from s to t"
- $x_{ij}/x_{i\bullet}$ = "conditional probability to jump to j from i".

From now on, one considers by default unit flows X, generally non-integer, obeying (1), (2) and (3).

2.3 Path entropy and energy

Let $W = (w_{ij})$ denote the $(n \times n)$ transition matrix of some irreducible Markov chain. A *st*-path constitutes a random walk (as defined by W) iff $x_{ij}/x_{i\bullet} = w_{ij}$ for all visited node i, i.e. such that $x_{i\bullet} > 0$. Random walk *st*-paths X minimise the *entropy* functional

$$G(X) := \sum_{ij} x_{ij} \ln \frac{x_{ij}}{x_{i\bullet} w_{ij}} = \sum_i x_{i\bullet} K_i(X||W) = x_{\bullet\bullet} \sum_i \frac{x_{i\bullet}}{x_{\bullet\bullet}} K_i(X||W)$$

where $K_i(X||W) := \sum_j \frac{x_{ij}}{x_{i\bullet}} \ln \frac{x_{ij}}{x_{i\bullet} w_{ij}} \geq 0$ is the Kullback-Leibler divergence between the transition distributions X and W from i, taking on its minimum value zero iff $\frac{x_{ij}}{x_{i\bullet}} = w_{ij}$. Note $G(X)$ to be *homogeneous*, that is $G(vX) = v\,G(X)$ for $v > 0$, reflecting the *extensivity* of $G(X)$ in the thermodynamic sense.

By contrast, shortest-paths and other alternative optimal paths minimize *resistance* or *energy* functionals of the general form

$$U(X) := \sum_{ij} r_{ij} \varphi(x_{ij})$$

where $r_{ij} > 0$ represent a cost or resistance associated to the directed arc ij, and $\varphi(x)$ is a smooth non-decreasing function with $\varphi(0) = 0$. In particular, minimizing $U(X)$ yields

- *st*-shortest paths for the choice $\varphi(x) = x$, where r_{ij} is the length of the arc ij
- *st*-electric currents from s to t for the choice $\varphi(x) = x^2/2$, where r_{ij} is the resistance of the conductor ij (see section 2.8).

As in Statistical Mechanics, we consider in this paper the class of admissible paths minimizing the *free energy*

$$F(X) := U(X) + T\,G(X) \ . \tag{9}$$

Here $T > 0$ is a free parameter, the *temperature*, controlling for the importance of the fluctuation around the trajectory of least resistance or energy (ground sate), realised in the low temperature limit $T \to 0$. In the high temperature limit $T \to \infty$ (or $\beta \to 0$, where $\beta := 1/T$ is the *inverse temperature*), the path consists of a random walk from s to t governed by W. Hence, minimising the free energy (9) generates for $T > 0$ *"heated extensions"* of classical minimum-cost problems $\min_X U(X)$, with the production of random fluctuations around the classical, "ground state" solution.

Derivating the free energy with respect to x_{ij}, and expressing the conservation constraints (2) through Lagrange multipliers $\{\lambda_i\}$ yields the optimality condition

$$T \ln \frac{x_{ij}}{x_{i\bullet} w_{ij}} + r_{ij}\varphi'(x_{ij}) = \lambda_j - \lambda_i \tag{10}$$

that is

$$x_{ij} = x_{i\bullet} w_{ij} \exp(-\beta[r_{ij}\varphi'(x_{ij}) + \lambda_i - \lambda_j]) \ . \tag{11}$$

The multipliers are defined up to an additive constant (see 15). In any case, $x_{ij} = 0$ when $w_{ij} = 0$ or $i = t$.

2.4 Minimum free energy and uniqueness

Multiplying (10) by x_{ij} and summing over all arcs yields an identity involving the entropy $G(X)$ of the optimal path X. Substitution in the free energy together with (2) demonstrates in turn the identity

$$\min_X F(X) = \sum_{ij} r_{ij}[\varphi(x_{ij}) - \varphi'(x_{ij})x_{ij}] + \lambda_t - \lambda_s \ . \tag{12}$$

The first term is negative for $\varphi(x)$ convex, positive for $\varphi(x)$ concave, and zero for the heated shortest-path problem $\varphi(x) = x$, for which $\min_X F(X) = \lambda_t - \lambda_s$.

Also, the entropy functional is convex, that is $G(\alpha X + (1-\alpha)Y) \leq \alpha G(X) + (1-\alpha)G(Y)$ for two admissible paths X and Y and $0 \leq \alpha \leq 1$. The energy $U(X)$ is convex (resp. concave) iff $\varphi(x)$ is convex (resp. concave).

When a strictly convex functional $F(X)$ possesses a local minimum on a convex domain $\mathcal{X}$, the minimum is unique. In particular, we expect the optimal flows for $\varphi(x) = x^p$ to be unique for $p > 1$, but not anymore for $0 < p < 1$, where local minima may exist; see Alamgir and von Luxburg (2011) on "p-resistances".

In the shortest-path problem $p = 1$, the solution is unique if $T > 0$ (Section 2.5); when $T = 0$, local minima of $U(X)$ may coexist, yet all yielding the same value of $U(X)$.

2.5 Algebraic solution

Solving (11) is best done by considering separately the target node t. Define $v_{ij} := w_{ij} \exp(-\beta r_{ij}\varphi'(x_{ij}))$ as well as the $(n-1) \times (n-1)$ matrix $V = (v_{ij})_{i,j \neq t}$.

Also, define the $(n-1)$ dimensional vectors

$$a_i := x_{i\bullet} \exp(-\beta\lambda_i)|_{i\neq t} \qquad b_j := \exp(\beta\lambda_j)|_{j\neq t} \qquad (13)$$
$$q_i := v_{it}|_{i\neq t} \qquad e_j := \delta_{js}|_{j\neq t}$$

Summing (11) over all i (for $j \neq t$, resp. $j = t$), then over all j for $i \neq t$ yields, using (2) and (3)

$$V'a = a - \exp(-\beta\lambda_s)\, e \qquad a'q = \exp(-\beta\lambda_t) \qquad Vb + \exp(\beta\lambda_t)\, q = b$$

Define the $(n-1) \times (n-1)$ matrix $M = (m_{ij})$ and the $(n-1)$ vector z as

$$M := (I - V)^{-1} = I + V + V^2 \ldots \qquad z := Mq \qquad (14)$$

Then a and b express as

$$a_i = \exp(-\beta\lambda_s)\, m_{si} \qquad b_j = \exp(\beta\lambda_s)\, \frac{z_j}{z_s} = \exp(\beta\lambda_j)$$

implying incidentally

$$\lambda_j = T \ln z_j + C \overset{\text{(Section 2.6)}}{=} T \ln z_j + \lambda_t \ . \qquad (15)$$

Finally

$$x_{i\bullet} = m_{si}\, \frac{z_i}{z_s} \qquad x_{ij} = m_{si}\, v_{ij}\, \frac{z_j}{z_s} \quad (i \neq t) \qquad (16)$$
$$x_{it} = m_{si}\, \frac{q_i}{z_s} \qquad x_{\bullet\bullet} = \frac{(Mz)_s}{z_s} = \frac{(M^2 q)_s}{(Mq)_s} \ . \qquad (17)$$

In general, V, M, q and d depend upon X. Hence (16) and (17) define a recursive system, whose fixed points may be multiple if $U(X)$ is not convex (Section 2.4), but converging to a unique solution for $p > 1$.

In the heated shortest-path case $p = 1$, the above quantities are independent of X. Hence the solution is unique, and particularly easy to compute in one single $O(n^3)$ step, involving matrix inversion, as illustrated in Sections 3 and 5.

2.6 Probabilistic interpretation

In addition to the absorbing target node t, let us introduce another "cemetery" or absorbing state 0, and define an extended Markov chain P on $n+1$ states with transition matrix

$$P = \begin{array}{c|c||c|c|} & \mathtt{i \neq t, 0} & \mathtt{t} & \mathtt{0} \\ \hline\hline \mathtt{i \neq t, 0} & V & q & \rho \\ \hline \mathtt{t} & 0 & 1 & 0 \\ \hline \mathtt{0} & 0 & 0 & 1 \\ \hline \end{array}$$

where $\rho_i = 1 - \sum_{k=1}^{n} v_{ik}$ is the probability of being absorbed at 0 from i in one step.

$M = (m_{ij})$ is the so-called *fundamental matrix* (see (14) and Kemeny and Snell 1976 p.46), whose components m_{ij} give the *expected number of visits from i to j*, before being eventually absorbed at 0 or t. Also, z_i (with $i \neq t, 0$) is the *survival probability*, that is to be, directly or indirectly, eventually absorbed at t rather than killed at 0, when starting from i. The higher the node survival probability, the higher the value of its Lagrange multiplier in view of (15).

Extending the latter to $j = t$ entails the consistency condition $z_t = 1$, making $\lambda_t \geq \lambda_i$ for all i. In particular, the free energy of the heated shortest-path case is, in view of (12),

$$F(X^{st}) = -T \ln z_s(T) \ ,$$

increasing (super-linearly in T) with the risk of being absorbed at 0 from s.

2.7 High-temperature limit

The energy term in (9) plays no role anymore in the limit $T \to \infty$ (that is $\beta \to 0$), and so does the absorbing state 0 above in view of $\rho_i = 0$. In particular, $z_i \equiv 1$ and $x_{ij}^{st} = m_{si} w_{ij}$ for $i \neq t$.

Also, $x_{\bullet\bullet}^{st}$ is the expected number of transitions needed to reach t from s. The *commute time distance* or *resistance distance* $x_{\bullet\bullet}^{st} + x_{\bullet\bullet}^{ts}$ is known to represent a *squared Euclidean distance* between states s and t: see e.g. Fouss et al. 2007, and references therein; see also Yen et al. (2008) and Chebotarev (2010) for further studies on resistance and shortest-path *distances*.

2.8 Low-temperature limit

Equations (11), (16) and (17) show the positivity condition $x_{ij} \geq 0$ to be automatically satisfied, thanks to the entropy term $G(X)$. However, the latter disappears in the limit $T \to 0$, where one faces the difficulty that the optimality condition (10) $r_{ij}\varphi'(x_{ij}) = \lambda_j - \lambda_i$ is still justified only if x_{ij} is freely adjustable, that is if $x_{ij} > 0$.

For the st-shortest path problem $\varphi(x) = x$, one gets, assuming the solution to be unique, the well-known characterisation (see e.g. Ahuja et al. (1993) p.107):

$$\begin{cases} r_{ij} = \lambda_j - \lambda_i & \text{if } x_{ij} > 0 \\ r_{ij} > \lambda_j - \lambda_i & \text{if } x_{ij} = 0 \end{cases}$$

occurring in the dual formulation of the st-shortest path problem, namely *"maximize $\lambda_t - \lambda_s$ subject to $\lambda_j - \lambda_i \leq r_{ij}$ for all i, j"*. Here λ_i is the shortest-path distance from s to i.

For the st-electrical circuit problem $\varphi(x) = x^2/2$, one gets $r_{ij}x_{ij} = \lambda_j - \lambda_i$ if $x_{ij} > 0$, in which case $x_{ji} > 0$ cannot hold in view of the positivity of the resistances, thus forcing $x_{ji} = 0$. Hence

$$\begin{cases} x_{ij} = \frac{\lambda_j - \lambda_i}{r_{ij}} > 0 & \text{if } \lambda_j > \lambda_i \\ x_{ij} = 0 & \text{otherwise} \end{cases}$$

expressing *Ohm's law* for the currrent intensity x_{ij} (Kirchhoff 1850), where λ_i is the electric potential at node i.

3 Illustrations and case studies: simple flow and net flow

Let us restrict on *st*-shortest path problems, i.e. $\varphi(x) = x$, whose free energy is homogeneous in the sense $F(vX) = vF(X)$ where $v > 0$ is the value of the flow in (4).

Graphs are defined by a $n \times n$ Markov transition matrix W together with a $n \times n$ positive resistance matrix R. Fixing in addition s, t and β, yields an unique *simple flow* x_{ij}^{st}, computable for any W (reversible or not) and any R (symmetric or not) - a fairly large set of tractable weighted networks.

An obvious class of networks consists of binary graphs, defined by a symmetric, off-diagonal adjacency matrix, with unit resistances and uniform transitions on existing edges (i.e. a simple random walk in the sense of Bollobás 1998).

Such are the graphs A (Figure 1) and B (Figure 2) below. Graph C (Figure 3) penalises in addition two edges forming short-cut from the point of view of W, but with increased values of their resistance.

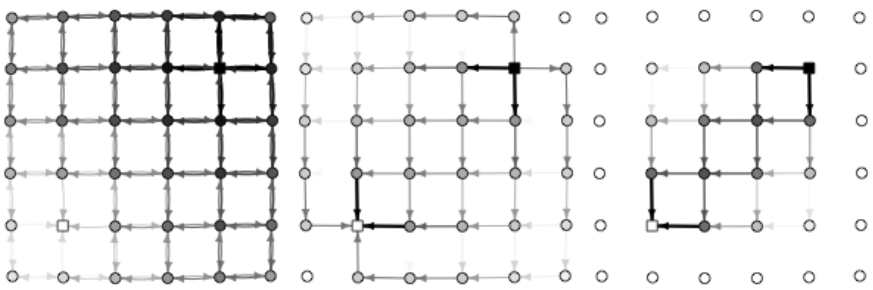

Fig. 1. Graph A is a square grid with uniform transitions and resistances. The resulting (high values in black, low values in light grey) simple flow x_{ij}^{st} and net flow ν_{ij}^{st} from s (black square) to t (white square) are depicted respectively on the left and middle picture with $\beta = 0$ (random walk) and on the right with $\beta = 50$ (shortest-path dominance). Note the simple flow and net flow to be identical at low temperatures.

Among the wide variety of graphs defined by a (W, R) pair, the plain graphs A, B and C primarily aim at illustrating the basic fact that, at high temperature, reverberation among neighbours of the source may dramatically lengthen the shortest path - an expected phenomenon (Figure 4).

Another quantity of interest is the *net flow*

$$\nu_{ij}^{st} := |x_{ij}^{st} - x_{ji}^{st}| \tag{18}$$

discounting "back and forth walks" inside the same edge, as discussed by Newman (2005): as a matter of fact, the presence of such alternate moves mechanically increases the simple flow inside an edge or node, especially near the source

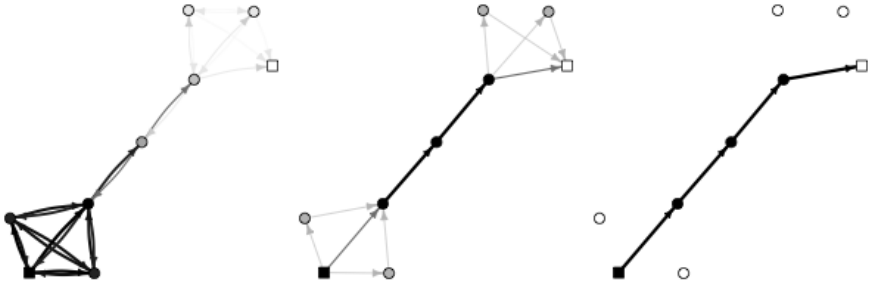

Fig. 2. Graph B consists of two cliques K_4 joined by two edges, with uniform transitions and resistances. Again, the resulting (high values in black, low values in light grey) simple flow x_{ij}^{st} and net flow ν_{ij}^{st} from s (black square) to t (white square) are depicted respectively on the left and middle picture with $\beta = 0$ (random walk) and on the right with $\beta = 50$.

Fig. 3. Graph C consists of two cliques K_5 joined by two paths: the upper one consists of five edges, each with unit resistance, while the upper one contains two edges, each with resistance tenfold larger. The resulting (high values in black, low values in light grey) simple flow x_{ij}^{st} and net flow ν_{ij}^{st} from s (black square) to t (white square) are depicted respectively on the left and middle picture with $\beta = 0$ (random walk) and on the right with $\beta = 50$.

at high temperature (Figures 1, 2 and 3, left), giving the false impression the behaviour is more entropic (that is, random-walk dominated) around the source, which is erroneous.

The net flow "filters out" reverberations and hence captures the resulting "trend" of the agents within their random movements, who rarely go back along the edge from where they came if there is another way; cf. the circulation of "used goods" as defined in Borgatti (2005) along trails exempt of edges repetition. At low temperatures, the simple flow is directed in one way and hence converges to the simple flow (Figures 1, 2 and 3, right).

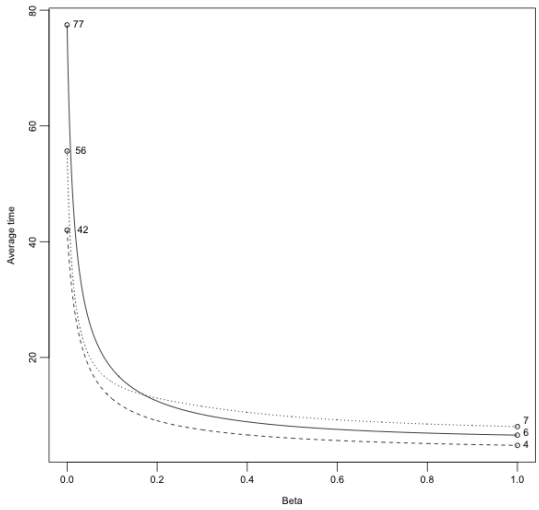

Fig. 4. The average time $x^{st}_{\bullet\bullet}$ to reach t from s is minimum for $T = 0$, and decreases with the inverse temperature β. Solid line: graph A; Dashed line: graph B; Dotted line: graph C.

4 Edge and vertex centrality betweenness

Several flow-based indices of betweenness centrality have been proposed ever since the shortest-path centrality pioneering proposal of Freeman (1977). In particular, random-walk centrality indices have been discussed by Noh and Rieger (2004) and Newman (2005). In this paper, we study the (unweighted) *mean flow betwenness*, defined for edges and vertices respectively (with complexity $O(n^5)$) as

$$\langle x_{ij} \rangle := \frac{1}{n(n-1)} \sum_{s,t|s\neq t} x^{st}_{ij} \qquad \langle x_{i\bullet} \rangle := \sum_j \langle x_{ij} \rangle = \langle x_{\bullet i} \rangle \tag{19}$$

where the latter identity results from the conservation condition (2). Definition (19) is intuitive enough: an edge is central if it carries a large amount of flow *on average*, that is by considering *all pairs of distinct source-targets couples*, thus extending the formalism to flows without specific source or target, such as monetary flows.

A more formal motivation arises from sensitivity analysis, with the result

$$\frac{\partial F(X(R))}{\partial r_{ij}} = \sum_{kl} \frac{\partial F(X(R))}{\partial x_{kl}(R)} \frac{\partial x_{kl}(R)}{\partial r_{ij}} + x_{ij}(R) = x_{ij}$$

where $F(X(R)) = \sum_{ij} r_{ij}\, x_{ij}(R) + TG(X(R))$ is the minimum free energy (9) under the constraints of Section 2.1 and r_{ij} the resistance of the edge ij.

Note that $\langle x_{\bullet\bullet}\rangle := \sum_j \langle x_{\bullet j}\rangle$ represents the average time to go from a vertex s to another vertex t and to return to s, averaged over all distinct pairs st. One can also define the *relative mean flow betweenness* as

$$c_{ij} := \frac{\langle x_{ij}\rangle}{\langle x_{\bullet\bullet}\rangle} \qquad\qquad c_i := \frac{\langle x_{i\bullet}\rangle}{\langle x_{\bullet\bullet}\rangle}$$

with the property $c_{ij} \geq 0$, $\sum_{ij} c_{ij} = 1$ and $c_i = c_{i\bullet} = c_{\bullet i}$.

Another candidate for a flow-based betweenness index is the *mean net flow*, again defined for edges and vertices as

$$\langle \nu_{ij}\rangle := \frac{1}{n(n-1)} \sum_{s,t|s\neq t} \nu_{ij}^{st} \qquad\qquad \langle \nu_{i\bullet}\rangle := \sum_j \langle \nu_{ij}\rangle = \langle \nu_{\bullet i}\rangle \tag{20}$$

Middle pictures in Figures 5, 6 and 7 below demonstrate how the mean net flow "substracts" the mechanical contribution arising from back and forth walks inside the same edge, in better accordance to a common sense notion of centrality.

Also, the sensitivity of the trip duration with respect to the edge resistance

$$\sigma_{ij} := \frac{\partial \langle x_{\bullet\bullet}(R)\rangle}{\partial r_{ij}}$$

constitutes yet another candidate, amenable to analytic treatment, whose study is beyond the size of the paper.

5 Case studies (continued): mean flow and mean net flow

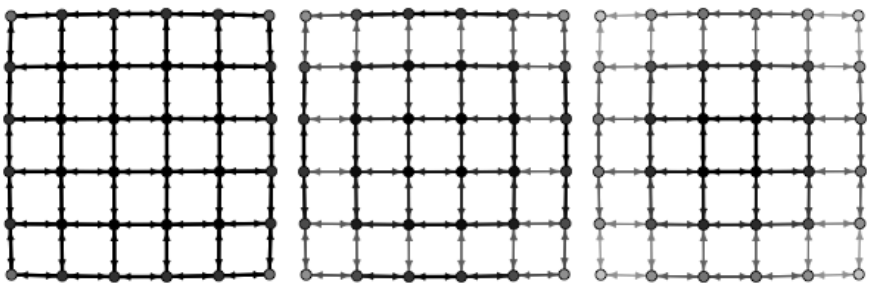

Fig. 5. Graph A: mean flow $\langle x_{ij}\rangle$ and mean net flow $\langle \nu_{ij}\rangle$, with $\beta = 0$ (left and middle) and $\beta = 50$ (right); high values in black, low values in light grey.

Figures 5, 6 and 7 depict the mean flow betweenness and the mean net flow betweenness (19) for the three graphs of Section 3, at high temperatures (left and

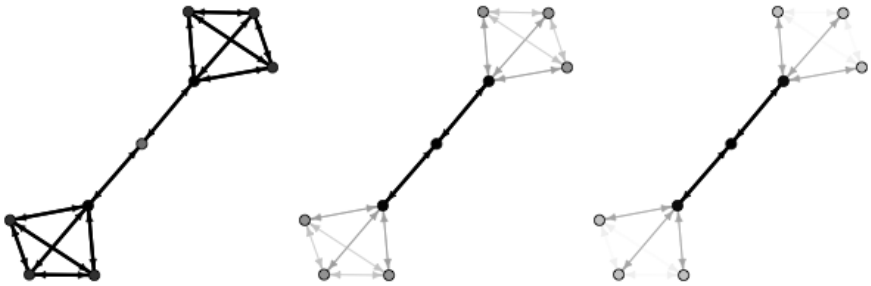

Fig. 6. Graph B: mean flow $\langle x_{ij} \rangle$ and mean net flow $\langle \nu_{ij} \rangle$, with $\beta = 0$ (left and middle) and $\beta = 50$ (right); high values in black, low values in light grey.

Fig. 7. Graph C: mean flow $\langle x_{ij} \rangle$ and mean net flow $\langle \nu_{ij} \rangle$, with $\beta = 0$ (left and middle) and $\beta = 50$ (right); high values in black, low values in light grey.

middle) and low temperatures (right). Here $\langle x_{ij} \rangle = \langle x_{ji} \rangle$ due to the symmetry of R and the reversibility of W. Visual inspection confirms the role of the mean flow as a betweenness index, approaching the shortest-path betweenness at low temperatures.

At high temperatures, the mean flow $\langle x_{ij} \rangle$ turns out to be *constant* for all edges ij, a consistent observation for all "random-walk type" networks we have examined so far. As a consequence, the mean flow centrality of a node $\langle x_{i\bullet} \rangle$ *is proportional to its degree* for $\beta \to 0$, and identical to the shortest-path betweenness for $\beta \to \infty$. The former simply measures the local connectivity of the node, while the latter also takes into account the contributions of the remote parts of the network, in particular penalising high-resistance edges in comparison to low-resistance ones (Figure 7).

At low temperature, the net mean flow converges (together with the simple flow) to the shortest-path betweenness (Figures 5, 6 and 7, right). At high temperatures, the net mean flow betweenness is large for edges connecting clusters, but, as expected, small for edges inside clusters. Hence an original kind of centrality, the "net random walk betweenness", differing from shortest-path and

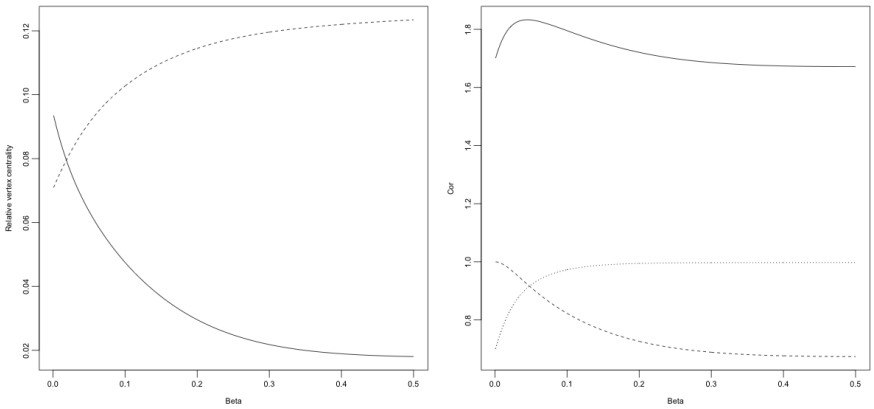

Fig. 8. Left: mean net flow centrality for the vertex in the "high-resistance path" (solid line) of network C, and for one of the nodes in the "low-resistance path" (dashed line) of network C. Right: inter-nodes correlation between the mean net flow centrality with itself at $\beta = 0$ (net random walk centrality; dashed line) and at $\beta = \infty$ (shortest-path node centrality; dotted line), in function of the inverse temperature β, for graph C. The sum of the two lines (solid line) is maximum for $\beta = 0.04$, arguably indicating a transition between an high- and a low-temperature regime.

degree betweeness, can be identified (Figures 5, 6 and 7, middle). As suggested in Figure 8 (right), contributions of both origins manifest themseves in the mean flow node centrality, for intermediate values of the temperature.

6 Conclusion

The paper proposes a coherent mechanism, easy to implement, interpolating between shortest paths and random walks. The construction is controlled by a temperature T and applies to any network endowed with a Markov transition matrix W and a resistance matrix R. The two matrices can be related, typically as (componentwise) inverses of each other (e.g. Yen et al. 2008) *or not*, in which case continuity at $T = 0$ and $T = \infty$ however requires $w_{ij} > 0$ whenever $r_{ij} < \infty$.

Modelling empirical *st*-paths necessitates to define W and R. The "simple symmetric model", namely unit resistances and uniform transitions on existing edges (Section 3) is, arguably, already meaningful in social phenomena and otherwise. For more elaborated applications, one can consider a possible model of tourist paths exploring Kobe (Iryio et al. 2012), consisting in choosing street directions as W with a bias towards "pleasant" street segments identified by low entries in R. Or the situation where a person at s wishes to be introduced to another person at t, by moving over an existing social network (defined by W) of friends, friends of friends, etc., where the resistance r_{ij} can express the

difficulty that actor i introduces the person to actor j. One can also consider general situations where W expresses an average motion, a mass circulation, and R captures an individual specific shift, biased towards preferentially reaching a peculiar outcome t, such as a specific location, or an a-spatial goal such as fortune, power, marriage, safety, etc.

By contrast, the construction seems little adapted to the simulation of replicant agents (such as viruses, gossip or e-mails) violating in general the flow conservation condition (2).

The paper has defined and investigated a variety of centrality indices for edges and nodes. In particular, the mean flow betweenness interpolates between degree centrality and shortest-path centrality for nodes. Regarding edges, the mean net flow embodies various measures ranging from simple random-walk betweenness (as defined in Newman 2005) to shortest-path betweenness, again. The average time needed to attain another node, respectively being attained from another node

$$T_s^{\text{out}} := \frac{1}{n-1} \sum_{t \mid t \neq s} x_{\bullet\bullet}^{st} \qquad\qquad T_t^{\text{in}} := \frac{1}{n-1} \sum_{s \mid s \neq t} x_{\bullet\bullet}^{st}$$

constitute alternative centrality indices, generalising Freeman's *closeness centrality* (Freeman 1979), incorporating a drift component when $T > 0$.

Maximum-likelihood type arguments, necessitating a probabilistic framework not exposed here, suggest for W and R fixed the estimation rule for T

$$U(X^{st}) = U(X^{st}(T))$$

where U is the energy functional in Section 2.3. Here X^{st} is the observed, empirical path, and $X^{st}(T)$ is the optimal path (16, 17) at temperature T. Alternatively, T could be calibrated from the observed total time, using Figure 4 as an abacus.

References

Ahuja, R.K., Magnanti, T.L., Orlin, J.B.: Network Flows. Theory, algorithms and applications. Prentice Hall (1993)

Alamgir, M., von Luxburg, U.: Phase transition in the family of p-resistances. In: Neural Information Processing Systems (NIPS 2011), pp. 379–387 (2011)

Bollobás, B.: Modern Graph Theory. Springer (1998)

Borgatti, S.P.: Centrality and network flow. Social Networks 27, pp. 55–71 (2005)

Brandes, U., Fleischer, D.: Centrality Measures Based on Current Flow. In: Diekert, V., Durand, B. (eds.) STACS 2005. LNCS, vol. 3404, pp. 533–544. Springer (2005)

Chebotarev, P.: A class of graph-geodetic distances generalizing the shortest-path and the resistance distances. Discrete Applied Mathematics 159, pp. 295–302 (2010)

Dubois-Ferrière, H., Grossglauser, M., Vetterli, M.: Valuable Detours: Least-Cost Anypath Routing. IEEE/ACM Transactions on Networking 19, pp. 333–346 (2011)

Iryo, T., Shintaku, H., Senoo, S.: Experimental Study of Spatial Searching Behaviour of Travellers in Pedestrian Networks. In: 1st European Symposium on Quantitative Methods in Transportation Systems, EPFL Lausanne (2012) (contributed talk)

Kemeny, J.G., Snell, J.L.: Finite Markov Chains. Springer (1976)
Farnsworth, K.D., Beecham, J.A.: How Do Grazers Achieve Their Distribution? A Continuum of Models from Random Diffusion to the Ideal Free Distribution Using Biased Random Walks. The American Naturalist 153, pp. 509–526 (1999)
Fouss, F., Pirotte, A., Renders, J.-M., Saerens, M.: Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation. IEEE Transactions on Knowledge and Data Engineering 19, pp. 355–369 (2007)
Freeman, L.C.: Centrality in networks: I. Conceptual claricatión. Social Networks 1, pp. 215–239 (1979)
Freeman, L.C., Borgatti, S.P., White, D.R.: Centrality in valued graphs: A measure of betweenness based on network flow. Social Networks 13, pp. 141–154 (1991)
Kirchhoff, G.: On a deduction of Ohms laws, in connexion with the theory of electrostatics. Philosophical Magazine 37, p. 463 (1850)
Newman, M.E.J.: A measure of betweenness centrality based on random walks. Social Networks 27, pp. 39–54 (2005)
Noh, J.-D., Rieger, H.: Random walks on complex networks. Phys. Rev. Lett. 92, p. 118701 (2004)
Saerens, M., Achbany, Y., Fouss, F., Yen, L.: Randomized Shortest-Path Problems: Two Related Models. Neural Computation 21, pp. 2363–2404 (2009)
Travers, J., Milgram, S.: An experimental study of the small world problem. Sociometry 32, pp. 425–443 (1969)
Yen, L., Saerens, M., Mantrach, A., Shimbo, M.: A Family of Dissimilarity Measures between Nodes Generalizing both the Shortest-Path and the Commute-time Distances. In: Proceedings of the 14th SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–793 (2008)
Zhou, T.: Mixing navigation on networks. Physica A 387, pp. 3025–3032 (2008)

3.2 Construction de dissimilarités

Une grande partie des méthodes d'apprentissage automatique, de classification et de statistique demandent d'avoir des *dissimilarités* entre les objets étudiés. Généralement, les dissimilarités à disposition sont multiples et le choix de la dissimilarité la plus pertinente pour l'étude du problème en question est d'une importance fondamentale. Lorsque les objets étudiés sont les noeuds d'un graphe, la dissimilarité devra représenter au mieux la topologie de celui-ci. Cependant, dans l'état actuel des connaissances, il n'est pas possible de capturer l'entièreté de la topologie du graphe avec une seule dissimilarité. Les différentes dissimilarités existantes ne mettent en valeur qu'un seul aspect structurel du graphe et sont parfois inadéquates pour obtenir le résultat désiré. Développer de nouvelles dissimilarités sur des graphes prend ainsi une valeur particulière, et l'utilisation du flux de transport randomisé permet de le faire aisément.

Dans la première partie de cette section, nous présenterons différentes catégories de distances et quelques exemples de dissimilarités existantes sur un graphe. La seconde partie contiendra un article expliquant comment construire des dissimilarités à partir du flux de transport randomisé, étudiant leurs propriétés et évaluant leurs performances lors de leur utilisation dans différentes méthodes appliquées à un jeu de données. Un lecteur désireux d'en savoir plus sur les dissimilarités pourra consulter la très complète *Encyclopédie des distances* de Michel et Elena Deza (?, ?) ou le chapitre "The partial order by inclusion of the principal classes of dissimilarity on a finite set, and some of their basic properties" de la revue *Classification and dissimilarity analysis* écrit par Frank Critchley et Bernard Fichet (?, ?).

3.2.1 Propriétés générales des dissimilarités

Définitions

Une *dissimilarité* d sur un graphe $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ est une application de $d : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ vérifiant les propriétés suivantes :

1. Positivité : $d(i,j) \geq 0 \quad \forall i,j \in \mathcal{V}$
2. Séparation : $d(i,j) = 0 \Leftrightarrow i = j \quad \forall i,j \in \mathcal{V}$
3. Symétrie : $d(i,j) = d(j,i) \quad \forall i,j \in \mathcal{V}$

Une dissimilarité peut être définie par sa *matrice de dissimilarité* $D = (d_{ij})$, de dimension $(n \times n)$, avec $d_{ij} = d(i,j)$.

Catégories

Une dissimilarité munie de l'*inégalité triangulaire*, $d(i,j) \leq d(i,k) + d(k,j),\ \forall i,j,k \in \mathcal{V}$, s'appelle une *métrique* ou une *distance*. On note l'ensemble des distances $\mathcal{M}$.

Une dissimilarité est *ultramétrique* ssi $d(i,j) \leq \max(d(i,k), d(j,j)),\ \forall i,j,k \in \mathcal{V}$. L'ensemble de ces dissimilarités se note $\mathcal{U}$.

Une dissimilarité est *Minkowski(q)* ssi $\forall i,j \in \mathcal{V}$, il existe deux vecteurs $a_i = (a_{i1}, \ldots, a_{ip})$ et $a_j = (a_{j1}, \ldots, a_{jp})$ dans $\mathbb{R}^p$ tel que $d(i,j) = (\sum_{k=1}^{p} |a_{ik} - a_{jk}|^q)^{1/q}$. L'ensemble de ces dissimilarités se note $\mathcal{L}_q$. $\mathcal{L}_1$ est appelé l'ensemble des dissimilarités *de bloc* ou *de Manhattan*, et $\mathcal{L}_2$ l'ensemble des dissimilarités *Euclidiennes*.

Une dissimilarité est *Euclidienne carrée* ssi $\forall i,j \in \mathcal{V}$, il existe deux vecteurs $a_i, a_j \in \mathbb{R}^p$ tel que $d(i,j) = \sum_{k=1}^{p} (a_{ik} - a_{jk})^2$. L'ensemble de ces dissimilarités se note $\mathcal{L}_2^2$.

Une dissimilarité est *Chebychev* ssi $\forall i, j \in \mathcal{V}$, il existe deux vecteurs $a_i, a_j \in \mathbb{R}^p$ tel que $d(i,j) = \max_{k=1}^{p} |a_{ik} - a_{jk}|$. L'ensemble de ces dissimilarités se note $\mathcal{L}_\infty$ et est confondu avec l'ensemble des distances, c-à-d $\mathcal{L}_\infty = \mathcal{M}$.

Ces différents ensembles de dissimilarités sont inclus les uns dans les autres de la façon suivante (?, ?) :

$$\mathcal{U} \subset \mathcal{L}_2 \subset \mathcal{L}_1 \subset (\mathcal{M} \cap \mathcal{L}_2^2).$$

Critères pour déterminer si une dissimilarité est Euclidienne carrée

De nombreuses méthodes d'analyse de données nécessitent d'avoir des dissimilarités Euclidiennes carrées, c'est pourquoi cette catégorie de dissimilarités prend une dimension particulière. En ayant une dissimilarité entre des noeuds grâce à sa matrice de distance D, on peut vérifier si elle est Euclidienne carrée grâce au critère suivant (?, ?) :

$$D \in \mathcal{L}_2^2 \iff K := -\frac{1}{2} H D H' \text{ est semi définie-positive}$$

où $H = (h_{ij}) := (\delta_{ij} - 1/n)$ est la *matrice de centration* et K la *matrice des produits scalaires* (souvent appelé *noyau* dans la littérature). Il suffit donc de voir si K ne possède que des valeurs propres non-négatives pour vérifier qu'une dissimilarité est Euclidienne carrée.

Dans (?, ?), il a été montré que si l'on élève les composantes d'une matrice de n'importe quelle dissimilarité avec un puissance $p > 0$ suffisamment petite, cette matrice devient une matrice de dissimilarité Euclidienne carrée. On peut donc définir la *puissance* d'une dissimilarité, noté $\text{pow}(d)$, comme la plus grande puissance avec laquelle les composantes d'une matrice de dissimilarité peuvent être élevées avant de ne plus être Euclidienne carrée. On a :

$$\begin{aligned} \text{pow}(d) \geq 1 &\Leftrightarrow d \in \mathcal{L}_2^2 \\ \text{pow}(d) \geq 2 &\Leftrightarrow d \in \mathcal{L}_2 \\ \text{pow}(d) = \infty &\Leftrightarrow d \in \mathcal{U} \end{aligned}$$

La puissance d'une dissimilarité permet ainsi de savoir si elle est Euclidienne carrée, Euclidienne ou ultramétrique.

3.2.2 Quelques dissimilarités sur les graphes

La distance du plus court chemin valuée

Nous avons déjà défini la distance du plus court chemin valuée auparavant. Pour rappel, $d^{dsp}(s,t) = \min_{\xi \in \mathcal{P}_{st}} l^d(\xi)$. C'est la distance la plus "classique" et la plus intuitive sur les graphes. On l'obtient généralement grâce à un algorithme, le plus connu étant celui de développé par Edsger Dijkstra en 1959 (?, ?). Comme son nom l'indique, c'est une dissimilarité métrique, mais généralement pas Euclidienne carrée, sauf si le graphe observé est un arbre (?, ?). Elle n'est donc pas idéale pour utiliser de méthode nécessitant une dissimilarité Euclidienne carrée, car l'information correspondant au valeurs propres négatives de K doit être supprimée. De plus, cette distance ne va prendre en compte que les plus courts chemins entre deux noeuds donnés, et ne contiendra pas d'information, par exemple, sur le nombre de chemins les séparant.

La distance de commutation

La distance de commutation (?, ?) est associée à une marche aléatoire, définie par W, sur un graphe spatial. Généralement, celle-ci correspond à la marche aléatoire simple, où $w_{ij} = 1/\deg^+(i)$, $\forall i, j \in \mathcal{E}$. Cette distance entre s et t correspond à la distance moyenne nécessaire à

la marche aléatoire pour aller de s à t, et revenir en s. On note cette $d^{ct}(s,t)$; et elle est définie par :

$$d^{ct}(s,t) = \frac{1}{2} \sum_{i,j \in \mathcal{V}} r_{ij}(x_{ij}^{st} + x_{ij}^{ts})$$

où x_{ij}^{st} est un flux admissible entre s et t suivant la marche aléatoire, c-à-d vérifiant $x_{ij}^{st}/x_{i\bullet}^{st} = w_{ij}$, $\forall i,j \in \mathcal{V}$. Dans le cas le plus connu, on pose toutes les résistances du graphe égales à 1.

Cette dissimilarité est connue pour être métrique et Euclidienne carrée (?, ?), ses propriétés en font donc une bonne mesure pour la plupart des méthodes. Seulement, si le nombre de noeuds dans le graphe augmente, et donc que le nombre de chemins entre s et t devient grand, cette dissimilarité ne prendra en compte que le voisinage direct des noeuds de départ et d'arrivée. Dans (?, ?), il est même montré que $d^{ct}(s,t) \tilde{\to} 1/\deg^+(s) + 1/\deg^+(t)$ lorsque $n \to \infty$ et cette distance devient alors impuissante à décrire la structure globale du graphe. Notons que sur un arbre, la distance de commutation est équivalente à la distance du plus court chemin (?, ?).

La distance du flux maximal

Cette distance, introduite dans l'article qui suit, est particulière, car elle nécessite des *capacités* associées aux arêtes. Les capacités $\kappa_{ij} \geq 0$ sont des quantités associées aux arcs qui déterminent quelle pourra être la quantité maximale de flux pouvant circuler sur ces derniers. Le flux *maximal* entre s et t, $X^{\max(s,t)} = (x_{ij}^{\max(s,t)})$ doit résoudre le problème d'optimisation suivant :

$$\begin{aligned}
\text{Maximiser} \quad & x_{s\bullet} - x_{\bullet s} \\
\text{sous contraintes :} \quad & x_{s\bullet} - x_{\bullet s} = x_{\bullet t} - x_{t\bullet} \\
& x_{i\bullet} - x_{\bullet i} = 0 \qquad \forall i,j \in \mathcal{V} \setminus \{s,t\} \\
& 0 \leq x_{ij} \leq \kappa_{ij} \qquad \forall i,j \in \mathcal{V}
\end{aligned}$$

La distance du flux maximal sera alors :

$$d^{maxf}(s,t) = \frac{1}{x_{s\bullet}^{\max(s,t)} - x_{\bullet s}^{\max(s,t)}}$$

c'est-à-dire l'inverse du flux maximal sortant de s. Nous n'allons malheureusement que peu nous y intéresser au cours de cette thèse, car notre formalisme ne prend pas en compte les capacités des arcs. Cependant, il est bien de la mentionner car c'est une dissimilarité *ultramétrique* (voir article 2), une propriété relativement rare pour une dissimilarité de graphe.

3.2.3 Article 2 : "Flow-based dissimilarities : shortest path, commute time, max-flow and free energy"

Cet article est paru dans *Data Science, Learning by Latent Structures, and Knowledge Discovery* (pp. 101-111), Springer Berlin Heidelberg, en 2015. Il fait suite à une présentation donnée à l'*European Conference on Data Analysis* donnée en 2013 à l'Université du Luxembourg.

Dans cet article, deux nouvelles dissimilarités sont construites sur les graphes à partir du flux de transport randomisé. La première de ces dissimilarités est égale à la distance du plus court chemin lorsque $T \to 0$, et à la distance de commutation lorsque $T \to \infty$. Etonnamment, cette dissimilarité n'est pas métrique pour des températures intermédiaires. La deuxième distance est construite pour palier à ce problème. Ces distances sont étudiées sur différents graphes, et il est montré qu'elles donnent de meilleurs résultats, en comparaison avec les dissimilarités classiques, lors de leur utilisation dans un algorithme des k-moyennes ("k-means") effectué sur un réseau de documents.

Points-clés

- Définition de la *distance du flux maximal* $D^{\mathrm{mf}} = (d_{ij}^{\mathrm{mf}})$ (p.41).
- Définition de la *dissimilarité d'énergie* $D^U = (d_{st}^U)$ (p.43).
- Définition de la *dissimilarité d'énergie libre* $D^F = d_{st}^F$ (p.43).
- Étude de la dernière valeur propre du noyau des dissimilarités et de leur puissance pow(d) en fonction de T (pp.44-45).
- Utilisation de ces dissimilarités dans une méthode de *positionnement multidimensionnel* (en anglais : "multidimensional scaling" ou MDS) (pp.44-45).
- Utilisation de ces dissimilarités dans l'algorithme du k-moyenne appliqué à un réseau de documents (pp.45-47).

Remarque

Cet article suit le formalisme de l'article précédant et est donc soumis à la même remarque concernant l'incertitude de trouver une solution au delà du cas $\varphi(x) = x$.

Guex, G. , Bavaud, F. : *Flow-Based Dissimilarities: Shortest Path, Commute Time, Max-Flow and Free Energy*. In: Lausen, B. et al. (Eds.) Data Science, Learning by Latent Structures, and Knowledge Discovery, pp. 101–111. (Series: Studies in Classification, Data Analysis, and Knowledge Organization). Springer, Heidelberg (2015)

Flow-based dissimilarities: shortest path, commute time, max-flow and free energy

Guillaume Guex[1] and François Bavaud[2]

[1] Department of Geography, University of Lausanne guillaume.guex@unil.ch
[2] Department of Geography, University of Lausanne francois.bavaud@unil.ch

Abstract. Random-walk based dissimilarities on weighted networks have demonstrated their efficiency in clustering algorithms. This contribution considers a few alternative network dissimilarities, among which a new max-flow dissimilarity, and more general *flow-based dissimilarities*, freely mixing shortest paths and random walks in function of a free parameter - the temperature. Their geometrical properties, and in particular their squared Euclidean nature are investigated through their power indices and multidimensional scaling properties. In particular, formal and numerical studies demonstrate the existence of critical temperatures, where flow-based dissimilarities cease to be squared Euclidean. The clustering potential of medium range temperatures is emphasised.

1 Introduction

The last decade has witnessed an increasing interest in centrality indices, community detection algorithms or network clustering assisted by random walks. Most approaches imply network dissimilarities, among which the shortest path and the commute time, closely linked to the minimisation of path functionals, namely a resistance or energy functional, respectively a relative entropy functional. The maximum flow and p-resistances (Alamgir and Von Luxburg (2011)) constitute alternative path functionals. Optimal flows minimising mixtures of path functionals characterise the global properties of the network, beyond the limited local view provided by binary or weighted adjacency matrices. Optimal flows also generate network dissimilarities, such as the presumably original maximum-flow dissimilarity (Section 2.2).

In particular, the free energy path functional (Saerens et al. (2009)) generates optimal flows interpolating between shortest paths and random walks (Section 2.4), where the edge resistances and transition matrix can be fixed independently (Bavaud and Guex (2012)). After reviewing the main definitions involved in the taxonomy of dissimilarities (Section 2.1), the geometric properties of the energy and free energy path functional dissimilarities (Section 2.5)

are investigated. In particular, the question of their squared Euclidean character is studied through the well-known Torgeson criterium, as well as through the less-known power index criterium of Joly and Le Calvé (1986). Numerical examples and applications demonstrate the existence of phase transitions, where path functional dissimilarities become non-Euclidean below some critical temperatures (Section 3.1). Section 2.5 addresses the issue of multidimensional scaling network reconstruction by path functional dissimilarities. Their better efficiency in clustering and classification of categorical data, as compared to chi-square dissimilarities, is illustrated in Section 3.3 for intermediate temperature ranges.

2 Dissimilarities

2.1 A few definitions and properties

Let us recall a few standard definitions (e.g. Joly and Le Calvé (1986); Citchley and Fichet (1994)): a dissimilarity on a set S of n objects is a $n \times n$ symmetric non-negative matrix $D = (d_{ij})$ with a null diagonal. The dissimilarity is *separable* if $d_{ij} = 0$ iff $i = j$, and *metric* if $d_{ij} \leq d_{ik} + d_{kj}$ (for all triples of S). A *distance* is a metric dissimilarity. One further distinguish between

- **ultrametric** distances ($D \in \mathcal{D}_U$) for which $d_{ij} \leq \max(d_{ik}, d_{jk})$.
- **Minkowski(q)** distances ($D \in \mathcal{D}_q$, where $q \geq 1$) if one can find n vectors $x_{ik} \in \mathbb{R}^p$ such that $d_{ij} = (\sum_{l=1}^{p} |x_{il} - x_{jl}|^q)^{\frac{1}{q}}$. $\mathcal{D}_2$ corresponds to the **Euclidean distance**.
- **squared Euclidean** dissimilarities ($D \in \mathcal{D}_2^2$) if there exists an embedding of the form $d_{ij} = \sum_{l=1}^{p} (x_{il} - x_{jl})^2$.
- **Chebychev** or **Frechet** distances ($D \in \mathcal{D}_\infty$) if there exists an embedding of the form $d_{ij} = \max_{l=1}^{p} |x_{il} - x_{jl}|$.

$\mathcal{D}_\infty$ is the set of all distances, and $\mathcal{D}_U \subset \mathcal{D}_2 \subset \mathcal{D}_1 \subset (\mathcal{D}_\infty \cap \mathcal{D}_2^2)$ holds.

Our study of the squared Euclidean character of the graph dissimilarities (Sections 2.5 and 3.2) mainly relies upon the following results, the first often attributed to Torgeson (1958) (with many precursors e.g. mentioned by Lew (1978)), and the second due to Joly and Le Calvé (1986). Here $f_i > 0$ is the relative weight of object i, normalized to unity, and δ_{ij} is Kronecker's delta:

Proposition 1. *A dissimilarity D on finite set S is $\mathcal{D}_2^2$ iff the* matrix of scalar products $B := -\frac{1}{2} HDH'$ *is positive semi-definite. Here, $H = (h_{ij})$ is the* centering matrix *where $h_{ij} = \delta_{ij} - f_j$, for any fixed normalised distribution f.*

Proposition 2. *For any dissimilarity D, there is a number $a \geq 0$ such that the elementwise power D^a is a squared Euclidean dissimilarity. Also, D^b is $\mathcal{D}_2^2$ as well for $0 \leq b \leq a$.*

Define $\mathrm{pow}(D)$*, the* ***power*** *of D, as the maximum value of a making D^a squared Euclidean. Then* $\mathrm{pow}(D) \geq 1$ *iff* $D \in \mathcal{D}_2^2$, $\mathrm{pow}(D) \geq 2$ *iff* $d \in \mathcal{D}_2$ *and* $\mathrm{pow}(D) = \infty$ *iff* $d \in \mathcal{D}_U$.

2.2 Commute-time, max-flow graph and chi-square distances

A binary graph $G = (V, E)$ on $|V| = n$ nodes is specified by a $n \times n$ symmetric adjacency matrix $A = (a_{ij})$ taking on values 0 or 1. The associated random walk is defined by the Markov transition matrix $W = (w_{ij})$ with $w_{ij} = a_{ij}/a_{i\bullet}$ (here "$\bullet$" denotes the summation over the replaced index).

Conversely, any regular Markov chain $W = (w_{ij})$ with stationary distribution f defines a *weighted graph* with associated node weights f_i, and *edge weights* or *exchange matrix* (Berger and Snell 1957) $e_{ij} := f_i w_{ij}$, giving the probability to select the pair of nodes ij. For unoriented graphs, $e_{ij} = e_{ji}$, that is W is *reversible*. By construction, $e_{i\bullet} = e_{\bullet i} = f_i$ and $e_{\bullet\bullet} = 1$.

Let $\mathcal{X}_{st}$ denote the set of *unit st-flows* from the source node $s \in V$ to the target node $t \in V$, as specified by the edge transitions counts $X = (x_{ij})$ of the trajectories or paths. By construction

$$x_{ij} \geq 0 \qquad , \qquad x_{i\bullet} - x_{\bullet i} = \delta_{is} - \delta_{it} \qquad \text{and} \qquad x_{t\bullet} = 0 \ . \tag{1}$$

The *commute-time distance* d^{ct}_{st} associated with W is the average time to go from s to t and back to i. That is $d^{\mathrm{ct}}_{st} = x^{st}_{\bullet\bullet} + x^{ts}_{\bullet\bullet}$, where x^{st}_{ij} is the random walk flow, obeying (1) and $x^{st}_{ij} = x^{st}_{i\bullet} w_{ij}$. One knows that $D^{\mathrm{ct}} \in (\mathcal{D}_\infty \cap \mathcal{D}_2^2)$, even in the oriented case (e.g. Boley et al. (2011)). Also, D^{ct} is *graph-geodetic* (Klein and Zhu (1998); Chebotarev (2010)), that is obeys $d^{\mathrm{ct}}_{ik} + d^{\mathrm{ct}}_{kj} = d^{\mathrm{ct}}_{jk}$ whenever all ij-paths and ji-paths moving over edges with non-zero weights pass through k.

Let us introduce a presumably new distance on unoriented weighted graphs, the ultrametric *max-flow distance* D^{mf}. In this setup, e_{ij} represents the *edge capacity*, controlling for the flow of maximum value v_{st} between s and t, solution of the problem

$$v_{st} := \max v \quad \text{such that} \quad 0 \leq x_{ij} \leq e_{ij}, \quad x_{i\bullet} - x_{\bullet i} = v(\delta_{is} - \delta_{it}), \quad x_{t\bullet} = 0.$$

By construction, $v_{ij} \geq 0$, $v_{ij} = v_{ji}$, $v_{ii} = \infty$ and $v_{ij} \geq \min(v_{ik}, v_{kj})$ for all triples in V^3. For $e_{ij} = e_{ji}$, define the max-flow distance as $d^{\mathrm{mf}}_{ij} := 1/v_{ij}$. Then d^{mf}_{ij} is a dissimilarity obeying $d^{\mathrm{mf}}_{ij} \leq \max(d^{\mathrm{mf}}_{ik}, d^{\mathrm{mf}}_{jk})$, that is $D^{\mathrm{mf}} \in \mathcal{D}_U$.

Categorial data analysis can also be cast in the above setup: let $N = (n_{il})$ be a $n \times m$ contingency table. Define a pair selection scheme by first choosing a row i, then a category l present in i, then another row j containing l. The resulting edge weight, node weight and transition matrix read (e.g. Bavaud and Xanthos 2005)

$$e_{ij} = \sum_{l=1}^{m} \frac{n_{il} n_{jl}}{n_{\bullet\bullet} n_{\bullet l}} \qquad f_i = \frac{n_{i\bullet}}{n_{\bullet\bullet}} \qquad w_{ij} = \sum_{l=1}^{m} \frac{n_{il} n_{jl}}{n_{i\bullet} n_{\bullet l}} \tag{2}$$

On the other hand, the chi-square dissimilarity $D^\chi = (d^\chi_{ij})$ between rows reads

$$d^\chi_{ij} := n_{\bullet\bullet} \sum_l \frac{1}{n_{\bullet l}} \left(\frac{n_{il}}{n_{i\bullet}} - \frac{n_{jl}}{n_{j\bullet}}\right)^2 \tag{3}$$

$D^\chi \in \mathcal{D}_2^2$, but $D^\chi \notin \mathcal{D}_\infty$. Neither D^χ nor D^{mf} are graph-geodetic.

2.3 Shortest-path distances

Let $r_{ij} \geq 0$ denote the length, cost, travel time or *resistance* of edge ij. The *shortest-path* length from s to t is

$$d^{\mathrm{sp}}_{st} := \min_{X \in \mathcal{X}_{st}} U(X) \qquad \text{where} \quad U(X) := \sum_{ij} r_{ij}\, x_{ij}$$

Then $d^{\mathrm{sp}}_{ii} = 0$, $d^{\mathrm{sp}}_{ik} + d^{\mathrm{sp}}_{jk} \geq d^{\mathrm{sp}}_{ij}$ and $d^{\mathrm{sp}}_{ij} = d^{\mathrm{sp}}_{ji}$ for symmetric $R = (R_{ij})$. In general, $D^{\mathrm{sp}} \notin \mathcal{D}_2^2$ (e.g. Deza and Laurent (1997) or Bavaud (2010)).

An important body of literature considers the *plain setup* $r_{ij} = c/a_{ij}$, where $c > 0$ is a normalisation constant, as e.g. in Yen (2008) and references therein, or as in the celebrated Doyle and Snell monograph (1984); see also the illustrations of Section 3. For $c = 1$, the seemingly counter-intuitive inequality $d^{\mathrm{ct}}_{ij} \leq d^{\mathrm{sp}}_{ij}$ holds, with equality iff the graph is a tree (e.g. Chandra et al. (1989); Deza and Deza (2009); Bavaud (2010)). Also, D^{sp} is graph-geodetic.

2.4 Interpolating random walks and shortest paths

Measuring the navigation effort within a weighted network depends on the nature of the moving agents (people, goods, money, information, etc), either knowledgeable of all networks characteristics, or only aware of their immediate neighborhood: d^{sp} models agents moving directly to their target, while d^{ct} models agents just wandering randomly until the target is reached. The former is more sensitive to short-cuts and to the length of paths in the network, while the latter is more sensitive to the degree and the number of paths between two nodes. Both capture information on the network structure, although of different kind.

This section presents a flow formalism aimed at continuously interpolating between the shortest path and the random walk, already detailed in Bavaud and Guex (2012)); see also Yen et al. (2008) and Saerens and al. (2009) for a close yet independent proposal, distinct in its implementation.

First, consider the general *twofold setup* endowed with two distinct edges valuations, namely the transition matrix $W = (w_{ij})$ of Section 2.2, and the resistances $R = (r_{ij})$ of Section 2.3. Both can be chosen independently, except for the consistency condition $r_{ij} = \infty$ iff $w_{ij} = 0$.

Secondly, define for each path $X = (x_{ij})$ the **flow energy** as

$$U(X) := \sum_{ij} r_{ij}\, \varphi(x_{ij})$$

where $\varphi(x)$ is a smooth non-decreasing function with $\varphi(0) = 0$. Flows of $\mathcal{X}_{st}$ minimizing $U(X)$ yield st-shortest paths for the choice $\varphi(x) = x$ and st-electic currents for the choice $\varphi(x) = x^2/2$ (Alamgir and Von Luxburg (2011); Li et al. (2011)). Define also the **flow entropy**:

$$G(X) := \sum_{ij} x_{ij} \ln \frac{x_{ij}}{x_{i\bullet} w_{ij}} = \sum_i x_{i\bullet} K_i(X||W) = x_{\bullet\bullet} \sum_i \frac{x_{i\bullet}}{x_{\bullet\bullet}} K_i(X||W)$$

where $K_i(X||W) := \sum_j \frac{x_{ij}}{x_{i\bullet}} \ln \frac{x_{ij}}{x_{i\bullet} w_{ij}} \geq 0$ is the Kullback-Leibler divergence between the empirical transitions $x_{ij}/x_{i\bullet}$ and the theoretical transitions w_{ij}. The entropy $G(X)$ takes on its minimum value zero iff $x_{ij}/x_{i\bullet} = w_{ij}$. Note that the multiplicative factor $x_{\bullet\bullet}$ aims at making $G(X)$ *homogeneous*, that is $G(vX) = vG(X)$ for $v > 0$.

Third, the aforementioned interpolation is implemented by considering the minimizing solution, noted $\tilde{X}^{st}$ or simply $\tilde{X}$, of the **free energy**

$$\tilde{X}^{st} := \arg\min_{X \in \mathcal{X}_{st}} F(X) \qquad\qquad F(X) := U(X) + T\, G(X)$$

where $T > 0$ is a free parameter, the **temperature**, which arbiters between the conflicting objectives: st-flows $\tilde{X}^{\mathrm{sp}}$ minimizing $F(X)$ realise shortest paths when $T \to 0$ (and $\varphi(x) = x$), and random walks $\tilde{X}^{\mathrm{rw}}$ when $T \to \infty$. We will also use $\beta = 1/T$, the **inverse temperature**.

On one hand, the feasible set $\mathcal{X}_{st}$ defined by (1) is convex; on the other hand, $F(X)$ is convex iff $\varphi(x)$ is convex, in which case the solution $\tilde{X}^{st}$ is unique and given by (Bavaud and Guex 2012)

$$\tilde{x}_{ij} = \tilde{x}_{i\bullet} w_{ij} \exp(-\beta[r_{ij}\varphi'(\tilde{x}_{ij}) + \lambda_i - \lambda_j]) \tag{4}$$

where λ_i are the Lagrange multipliers associated to (1). Equivalently, defining $v_{ij} := w_{ij} \exp(-\beta r_{ij} \varphi'(\tilde{x}_{ij}))$ as well as $V := (v_{ij})$ (where $i \neq t$ and $j \neq t$), $M := (I - V)^{-1}$, $q := (v_{it})_{i \neq t}$ and $z := Mq$, the solution reads:

$$\tilde{x}_{ij} = m_{si}\, v_{ij}\, \frac{z_j}{z_s} \quad (j \neq t) \qquad\qquad \tilde{x}_{it} = m_{si}\, \frac{q_i}{z_s} \ .$$

The optimal flow $\tilde{X} = (\tilde{x}_{ij})$ can be interpreted as the expected number of passages on edge ij when starting from s until eventually reaching t. The minimum free energy simply express as $F(\tilde{X}) = -T \ln z_s$. In what follows, we assume $\varphi(x) = x$. Then (4) is solved in a single step, instead of iteratively.

2.5 Energy and Free Energy Dissimilarities

Define, for any pair st of nodes, the **energy dissimilarity** D^{U} and the **free energy dissimilarity** D^{F} as

$$d^{\mathrm{U}}_{st} := \frac{1}{2}(U(\tilde{X}^{st}) + U(\tilde{X}^{ts})) \qquad\qquad d^{\mathrm{F}}_{st} := \frac{1}{2}(F(\tilde{X}^{st}) + F(\tilde{X}^{ts})) \ . \tag{5}$$

d^{U}_{st} is the expected resistance for going from s to t and coming back. In general, $D^{\mathrm{U}} \notin \mathcal{D}_\infty$; however, $D^{\mathrm{F}} \in \mathcal{D}_\infty$, and is graph-geodetic as well (see Kivimäki et al. (2012) and references therein). Furthemore, one can prove that

$$\lim_{\beta\to\infty} d^{\mathrm{F}}_{st} = \lim_{\beta\to\infty} d^{\mathrm{U}}_{st} = \frac{1}{2}(d^{\mathrm{sp}}_{st} + d^{\mathrm{sp}}_{ts})$$

$$\lim_{\beta\to 0} d^{\mathrm{F}}_{st} = \lim_{\beta\to 0} d^{\mathrm{U}}_{st} = \frac{1}{2}\sum_{ij} r_{ij}\,(\tilde{x}^{st\ \mathrm{rw}}_{ij} + \tilde{x}^{ts\ \mathrm{rw}}_{ij}) := d^{\mathrm{wct}}_{st}$$

where d^{wct}_{st} is the commute cost or *weighted commute time*, proportional to d^{ct}_{st} (Françoisse et al. (2013); Kivimäki et al. (2012)). The above suggests the possibility of an *Euclidean phase transition*, with dissimilarities in $\mathcal{D}_2^2$ for $T \geq T_c$, but not anymore for $T < T_c$ whenever $D^{\mathrm{sp}} \notin \mathcal{D}_2^2$.

3 Numerical examples and applications

3.1 Experiments with small graphs

Both D^{U} and D^{F} capture network information related to D^{sp} as well as to D^{ct}. Let us investigate their squared Euclidean nature by using the two criteria of Section (2.1). Figure 1 depicts the behaviour of the smallest eigenvalue of B and the power index, in the plain setup $r_{ij} := 1/a_{ij}$ and $w_{ij} := a_{ij}/a_{i\bullet}$.

The first example, K_{23}, demonstrates the existence of *critical temperatures* T_c in the sense of Section 2.5. The second example, C_{15}, shows D^{F} to be $\mathcal{D}_2^2$ in the whole temperature range, contrarily to D^{U} which is not $\mathcal{D}_2^2$ for intermediate values of β. In the third example, both D^{F} and D^{U} are $\mathcal{D}_2^2$ over the whole temperature range, with a contrasted behaviour between the last eigenvalue of B, monotonously decreasing, and the power, maximum around $\beta = 0.5$. On all examples, D^F stays in $\mathcal{D}_2^2$ longer than D^U when β is raised, a potentially interesting property since squared Euclidean dissimilarities are often needed for statistical applications. Lacunary as they are, those results underline the wide behavioral range of simple graphs, as expected from statistical mechanical entities.

3.2 Multidimensional scaling

A planar graph with $n = 50$ nodes, aimed at imitating a realistic road network, is generated following a variant of an algorithm due to Gastner and Newman (2006) (Figure 2). We define r_{ij} as the Euclidean distances between the pairs of nodes, and apply the simple setup $e_{ij} = c/r_{ij}$. After computing D^{U} or D^{F} for various β, we extract the MDS eigen-coordinates in the two first dimensions (regardless of the possible negativity of the last eigenvalue of B^{U}).

In addition, we evaluate the similarity between the original dissimilarities and the path functional dissimilarities by means of a presumably original *configuration similarity* index $\mathtt{CS}_{ab} := \frac{\mathrm{Tr}(B^a B^b)}{\sqrt{\mathrm{Tr}((B^a)^2)\mathrm{Tr}((B^b)^2)}} \in [0,1]$, where B^a and B^b are the scalar products corresponding to configurations D^a and D^b. The maximum similarity of $\mathtt{CS}_{\mathrm{planar,\ U}} = 0.86$ for which D^{U} is still squared Euclidean obtains for $\beta = 0.3$ (Figure 2).

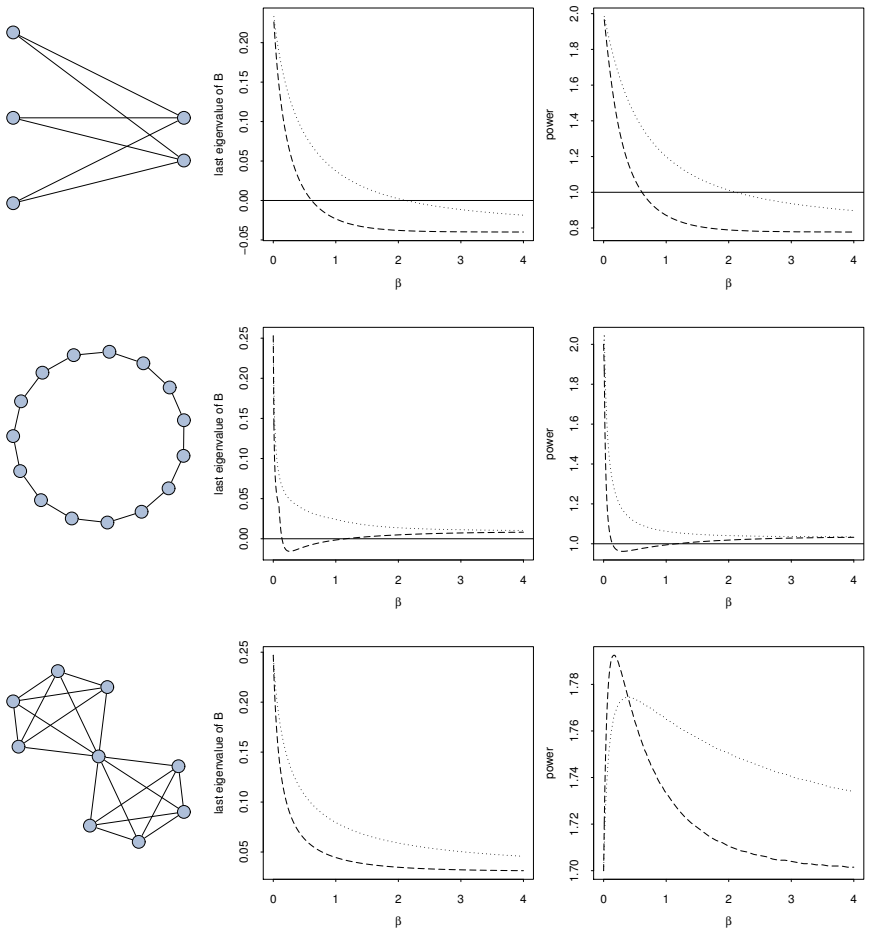

Fig. 1. Three toy graph examples, scanning the squared Euclidean character of D^{U} (dashed line) and D^{F} (dotted line). The first plot depicts the evolution of the last eigenvalue of B versus β. The second plot exhibits the power index versus β.

3.3 Clustering

A standard approach to clustering categorical data consists in computing chi-square dissimilarities D^{χ} (3) between objects, and then applying a k-means procedure. Alternatives based upon D^{U} and D^{F} might reveal more efficient, as demonstrated here in a supervised context with groups known a priori.

Specifically, one considers the document-terms contingency table $N = (n_{il})$ of the $n = 160$ documents of the `Reuters21578` corpus, belonging to $m = 8$ different groups (20 documents in each group). Exchanges e_{ij} obtain as in

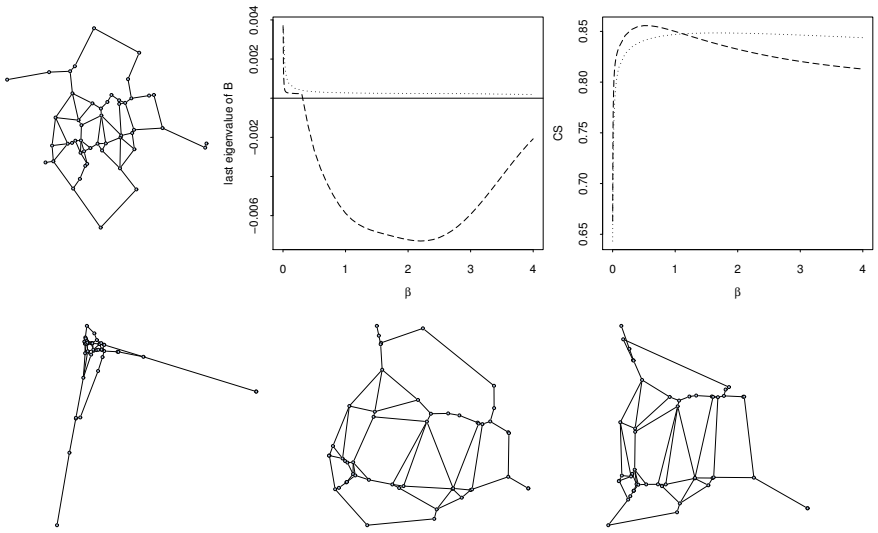

Fig. 2. Top: the original planar graph (left) and the behaviour, regarding β, of the last eigenvalue of B (middle) and the configuration similarity `CS` (right) between the original configuration and the graph dissimilarities D^U (dashed line) and D^{F} (dotted line). Bottom: MDS graph reconstruction from D^{U}, whose first two dimensions respectively explain 42.6% of the inertia (left, for $\beta = 0.001$), 53.8% (middle, for $\beta = 0.3$) and 48.1% (right, for $\beta = 4$). In the last case, the dissimilarity is not squared Euclidean and negative eigenvalues have been removed from the inertia.

(2), and resistances as $r_{ij} = 1/e_{ij}$ (plain setup). Document eigen-coordinates are obtained from *weighted MDS* on D^{U} and D^{F}, that is by considering the spectral decomposition of $K = (k_{ij})$ with $k_{ij} = \sqrt{f_i f_j} b_{ij}$ instead of $B = (b_{ij})$ (e.g. Bavaud (2010)); also, negative eigenvalues of K^{U} or K^{F} are set to zero. A k-means procedure with 8 clusters is then applied on the eigen-coordinates, and the resulting partition C is compared to the true partition C^{true} by means of the *variation of information dissimilarity* $d_{\mathrm{VI}}(C, C^{\mathrm{true}}) := H(C) + H(C^{\mathrm{true}}) - 2I(C, C^{\mathrm{true}})$, where H is the entropy of the partition and I the mutual information (Meila (2003)). Figure 3 shows that both D^{U} and D^{F} yield noticeably better results than D^{χ} for intermediate value of β, but quickly cease to be squared Euclidean. D^{U} gives for $\beta = 5 \cdot 10^{-6}$ the clustering most similar to the true classification, with a rate of correct classification (under optimal group permutation) of 65.63%. See Kivimäki et al. (2102) for further clustering experiments involving D^{U} and D^{F}, and e.g. Liu et al. (2013) for random walk clustering.

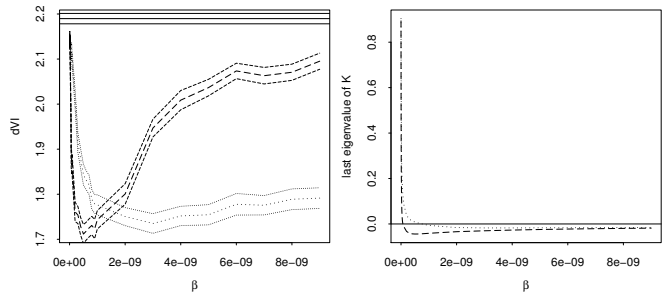

Fig. 3. Left: average information dissimilarity and 95% CI, comparing the true classification and the clustering obtained from D^{U} (dashed line), D^{F} (dotted line) and D^{χ} (baseline, top). Right: last eigenvalue of K^{U} (dashed line) and K^{F} (dotted line) as a function of β.

4 Conclusion

Transforming a graph into a dissimilarity matrix most facilitates the discussion of clustering and visualisation issues. It permits to open graph problems to a large body of statistical and mathematical methods: classical data analysis, machine learning, spectral graph theory and operations research. To that extent, enriching the family of flow-based dissimilarities by considering further sensible yet tractable functionals appears as a research priority. In particular, squared Euclidean dissimilarities immediately allow for MDS visualisation and Ward hierarchical clustering.

Various dissimilarities capture (and hide) various aspects of the graph. Results show that D^U generally gives a better representation of the graph structure than D^F for a precise value of the temperature T^{opt}, while the latter is more stable over temperature changes. This suggest that D^U should be used when T^{opt} is known, whereas D^F is preferable otherwise.

Despite the above results, the questions of knowing how to determine, even approximatively, T^{opt}, which aspect of the graph structure is best enlightened by which dissimilarity, and which dissimilarity is most efficient for a specific clustering or visualisation task, remain largely open.

References

ALAMGIR, M. and VON LUXBURG, U. (2011): Phase transition in the family of p-resistances. *Neural Information Processing Systems (NIPS 2011), 379–38.*

BAVAUD, F. (2010): Euclidean Distances, Soft and Spectral Clustering on Weighted Graphs. In: *Proceedings of ECML-PKDD 2010. LNCS 6321, 103–118.*

BAVAUD, F. and GUEX, G. (2012): Interpolating between random walks and shortest paths: a path functional approach. In: *Proceedings of SocInfo 2012. LNCS 7710, 68–81.*

BAVAUD, F. and XANTHOS, A. (2005): Markov Associativities. *Journal of Quantitative Linguistics 12, 123–137.*

BERGER, J. and SNELL, J.L. (1957): On the concept of equal exchange. *Behavioral Science 2, 111–118.*

BOLEY, D., RANJAN, G. and ZHANG, Z.-L. (2011): Commute Times for a Directed Graph using an Asymmetric Laplacian. *Linear Algebra and its Applications 435, 224–242.*

CHANDRA, A.K., RAGHAVAN, P., RUZZO, W.L., SMOLENSKY, R. and TIWARI, P. (1989): The Electrical Resistance Of A Graph Captures Its Commute And Cover Times. *Proceedings of the twenty-first annual ACM symposium on Theory of computing (STOC '89), 574–586.*

CHEBOTAREV, P. (2010): A class of graph-geodetic distances generalizing the shortest-path and the resistance distances. *Discr. Appl. Math. 159, 295–302.*

CRITCHLEY, F. and FICHET (1994) The partial order by inclusion of the principal classes of dissimilarity on a finite set, and some of their basic properties. In: B. van Cutsem (Ed.): *Classification and dissimilarity analysis. LNS 93, 5–65.*

DEZA, M. and DEZA, E. (2009): *Encyclopedia of Distances.* Springer.

DEZA, M. and LAURENT, M. (1997): *Geometry of cuts and metrics.* Springer.

DOYLE, P. and SNELL, J. (1984): *Random walks and electric networks.* Mathematical Association of America.

FRANÇOISSE, K., KIVIMÄKI, I., MANTRACH, A. ROSSI, F. and SAERENS, M. (2013): A bag-of-paths framework for network data analysis. *arXiv:1302.6766*

GASTNER, M.T. and NEWMAN, M.E.J. (2006): The spatial structure of networks. *Eur. Phys. J. B 49, 247–252.*

JOLY, S. and LE CALVÉ, G. (1986): Etude des puissances d'une distance. *Statistique et analyse des donn?es, 11, 30–50.*

KIVIMÄKI, I., SHIMBO, M. and SAERENS, M. (2012): Developments in the theory of randomized shortest paths with a comparison of graph node distances. *arXiv:1212.1666*

KLEIN, D.J. and ZHU, H.Y. (1998): Distances and volumina for graphs. *J. Math. Chem. 23, 179–195.*

LEW, J.S. (1978): Some counterexamples in multidimensional scaling. *Journal of Mathematical Psychology 17, 247–254.*

LI, Y., ZHANG; Z-L. and BOLEY, D. (2011): The Routing Continuum from Shortest-Path to All-Path: A Unifying Theory. *31st International Conference on Distributed Computing Systems (ICDCS), 847–856.*

LIU, S., MATZAVINOS, A. and SETHURAMAN, S. (2013): Random walk distances in data clustering and applications. *Adv. Data Analysis and Classif. 7, 83–108.*

MEILA, M. (2003): Comparing clusterings by the variation of information. *Proceedings of the Sixteenth Annual Conference of Computational Learning Theory (COLT).* Springer.

SAERENS, M., ACHBANY, Y., FOUSS, F. and YEN, L. (2009): Randomized Shortest-Path Problems: Two Related Models. *Neural Computation 21, pp. 2363–2404.*

TORGESON, W.S. (1958) : *Theory and methods of scaling.* Wiley, New York.

YEN, L., SAERENS, M., MANTRACH, A. and SHIMBO, M. (2008): A Family of Dissimilarity Measures between Nodes Generalizing both the Shortest-Path and the Commute-time Distances. *Proceedings of the 14th SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–793.*

3.3 Cadre probabiliste

Dans une série d'articles écrit par Marco Saerens et ses collaborateurs (?, ?, ?, ?, ?, ?), des résultats similaires à ce qui a été présenté dans ce chapitre ont été obtenus avec un formalisme *probabiliste*. Dans cette partie, nous allons faire un bref exposé de ce formalisme et essayer d'établir de manière intuitive certaines correspondances avec le formalisme de flux. Les preuves des résultats exposés ici se trouvent dans les différents articles cités ci-dessus.

3.3.1 Formalisme

Le cadre

Dans ce formalisme (?, ?), on se place également dans le cadre d'un graphe spatial orienté et pondéré, c'est-à-dire d'un graphe orienté $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ muni d'une matrice de résistances $R = (r_{ij})$ et d'une matrice de conductances $C = (c_{ij})$. On pose l'ensemble $\tilde{\mathcal{P}}_{st}$ comme l'ensemble des chemins *absorbants*, c'est-à-dire tous les chemins ξ_{st} tels que le noeud t n'apparaisse qu'une seule fois (à la fin du chemin). Grâce aux résistances et aux conductances, les chemins possèdent une longueur valuée $l^d(\xi) = l^d((i_0, \dots, i_\tau)) := \sum_{k=0}^{\tau-1} r_{i_k i_{k+1}}$ et il existe une matrice de transition de référence $W = (w_{ij}) = (c_{ij}/c_{i\bullet})$.

Le "sac de chemins" (bag-of-paths)

On peut considérer l'ensemble $\tilde{\mathcal{P}}_{st}$ comme un "*sac de chemins*" (en anglais : "bag-of-paths"), c'est-à-dire qu'on peut définir des mesures de probabilité sur cet ensemble, se notant alors $\mathbb{P}_{st}$.

Une *probabilité de référence* peut être posée sur cet ensemble de la manière suivante :

$$\mathbb{P}^{\text{ref}}_{st}(\xi_{st}) = \mathbb{P}^{\text{ref}}((s = i_0, \dots, i_\tau = t)) := \prod_{k=0}^{\tau-1} w_{i_k i_{k+1}} = \prod_{k=0}^{\tau-1} \frac{c_{i_k i_{k+1}}}{c_{i_k \bullet}}$$

Il s'agit de la probabilité que le chemin se réalise en suivant la chaîne de Markov définie par W.

Le but va être de rechercher une autre mesure de probabilité, la *probabilité des plus courts chemins randomisés* (en anglais : "randomized shortest-paths", abrégé RSP), notée $\mathbb{P}^{\text{RSP}}_{st}(\xi_{st})$. Cette mesure doit vérifier :

$$\begin{aligned} \mathbb{P}^{\text{RSP}}_{st} = \quad & \underset{\mathbb{P}_{st}}{\arg\min} \sum_{\xi \in \tilde{\mathcal{P}}_{st}} \mathbb{P}_{st}(\xi) l^d(\xi) \\ \text{sous contraintes :} \quad & \textstyle\sum_{\xi \in \tilde{\mathcal{P}}_{st}} \mathbb{P}_{st}(\xi) \log\left(\frac{\mathbb{P}_{st}(\xi)}{\mathbb{P}^{\text{ref}}_{st}(\xi)}\right) = J_0 \\ & \textstyle\sum_{\xi \in \tilde{\mathcal{P}}_{st}} \mathbb{P}_{st}(\xi) = 1 \end{aligned}$$

où $J_0 > 0$ est un paramètre libre. Cette mesure doit donc attribuer une probabilité élevée aux chemins courts, et une probabilité faible au chemin longs, sous contrainte que l'entropie relative entre cette mesure et celle de référence soit égale à J_0. Cela revient, dans notre formalisme, à minimiser l'énergie avec une valeur d'entropie donnée.

La solution

La probabilité RSP suit une *distribution de Gibbs-Boltzmann* (?, ?). Donnée par :

$$\mathbb{P}^{\text{RSP}}_{st}(\xi) = \frac{\mathbb{P}^{\text{ref}}_{st}(\xi) \exp(-\beta l^d(\xi))}{\sum_{\xi' \in \tilde{\mathcal{P}}_{st}} \mathbb{P}^{\text{ref}}_{st}(\xi') \exp(-\beta l^d(\xi'))} = \frac{\mathbb{P}^{\text{ref}}_{st}(\xi) \exp(-\beta l^d(\xi))}{\mathcal{Z}}$$

où $\beta > 0$ est un paramètre libre, qui correspond à la *température inverse*. On appelle $\mathcal{Z} := \sum_{\xi' \in \tilde{\mathcal{P}}_{st}} \mathbb{P}^{\text{ref}}_{st}(\xi') \exp(-\beta l^d(\xi'))$ la *fonction de partition*.

En définissant la matrice $\tilde{M} := (\tilde{I} - \tilde{V})^{-1}$, issue de la chaîne de Markov "incomplète" définie par $\tilde{V} = (\tilde{v}_{ij}) := (w^{\text{ref}}_{ij} \exp(-\beta r_{ij}))$, on peut trouver la fonction de partition $\mathcal{Z}$ grâce à :

$$\mathcal{Z} = \frac{\tilde{m}_{st}}{\tilde{m}_{tt}}$$

3.3.2 Les indices de centralité

Notons $\eta_{ij}(s,t)$ l'*espérance du nombre de passages par* (i,j) *lorsque la probabilité est* $\mathbb{P}^{RSP}_{st}$. On peut prouver que (?, ?) :

$$\eta_{ij}(s,t) = \left(\frac{\tilde{m}_{si}}{\tilde{m}_{st}} - \frac{\tilde{m}_{ti}}{\tilde{m}_{tt}} \right) v_{ij}\ \tilde{m}_{jt}$$

Cette quantité correspond exactement à notre définition du flux, comme nous le verrons dans l'article 4. Cette quantité est exprimée d'une manière différente ici, car la matrice $\tilde{M}$ ne correspond pas à la matrice fondamentale M utilisée dans le formalisme de flux. En effet, dans le formalisme du flux, on a $M = (I - V)^{-1}$ avec $V = (v_{ij})_{i,j \neq t}$ et I de taille $((n-1) \times (n-1))$ alors qu'ici, $\tilde{M} = (\tilde{I} - \tilde{V})^{-1}$ avec $\tilde{V} = (v_{ij})_{i,j \in \mathcal{V}}$ et $\tilde{I}$ de taille $(n \times n)$. Trouver la correspondance exacte entre ces deux quantités fait partie des pistes de recherche futures.

La centralité RSP

On peut définir la *centralité RSP du noeud* i, notée $B^{\text{RSP}}(i)$, comme (?, ?) :

$$B^{\text{RSP}}(i) := \sum_{s \in \mathcal{V}} \sum_{t \in \mathcal{V}} \sum_{j \in \mathcal{V}} \eta_{ij}(s,t)$$

c'est-à-dire la somme sur tous les s et t de l'espérance du nombre de passages par i. Cette centralité est similaire à celle de la centralité du flux moyen sur les noeuds $\langle x_{i\bullet} \rangle$.

La centralité RSP nette

La *centralité RSP nette du noeud* i, notée $B^{\text{RSPnet}}(i)$, se définit comme (?, ?) :

$$B^{\text{RSPnet}}(i) := \sum_{s \in \mathcal{V}} \sum_{t \in \mathcal{V}} \sum_{j \in \mathcal{V}} |\eta_{ij}(s,t) - \eta_{ji}(s,t)|$$

c'est-à-dire en sommant sur tous les s et t l'espérance du nombre "net" de passages. Cette centralité est similaire à celle de la centralité du flux net moyen sur les noeuds $\langle \nu_{i\bullet} \rangle$.

3.3.3 Les dissimilarités

La dissimilarité du RSP

Posons l'*espérance de la longueur valuée des chemins entre* s *et* t *en fonction d'une probabilité* $\mathbb{P}_{st}$ comme :

$$\bar{l}^d(\mathbb{P}_{st}) := \sum_{\xi \in \tilde{\mathcal{P}}_{st}} \mathbb{P}_{st}(\xi) l^d(\xi)$$

Cette quantité devrait correspondre, à une constante multiplicative près, au terme d'énergie du flux, $U(X) = \sum_{ij} r_{ij} x_{ij}$.

La *dissimilarité du RSP* entre s et t, notée d_{st}^{RSP}, est définie comme (?, ?) :

$$d^{\mathrm{RSP}}(s,t) = \frac{\bar{l}^d(\mathbb{P}_{st}^{\mathrm{RSP}}) + \bar{l}^d(\mathbb{P}_{ts}^{\mathrm{RSP}})}{2}$$

cette dernière n'est généralement pas une distance (l'inégalité triangulaire n'est pas toujours vérifiée). Cette dissimilarité devrait coïncider, à une constante multiplicative près, à la définition de la dissimilarité d'énergie D^U.

On peut aisément calculer cette dissimilarité grâce à la matrice $\tilde{M}$:

$$\bar{l}^d(\mathbb{P}_{st}^{\mathrm{RSP}}) = -\frac{1}{\beta}\frac{\partial \log \mathcal{Z}}{\partial \beta} = \frac{[\tilde{M}(R \circ \tilde{V})\tilde{M}]_{st}}{\tilde{m}_{st}} - \frac{[\tilde{M}(R \circ \tilde{V})\tilde{M}]_{tt}}{\tilde{m}_{tt}}$$

où "$\circ$" désigne la multiplication composante par composante.

La dissimilarité de l'énergie

On peut définir l'*énergie libre* de la distribution $\mathbb{P}_{st}$ comme (?, ?) :

$$F(\mathbb{P}_{st}) = \bar{l}^d(\mathbb{P}_{st}) + T \sum_{\xi \in \tilde{\mathcal{P}}_{st}} \mathbb{P}_{st}(\xi) \log\left(\frac{\mathbb{P}_{st}(\xi)}{\mathbb{P}_{st}^{\mathrm{ref}}(\xi)}\right)$$

où $T := \frac{1}{\beta}$ est la température. On peut facilement montrer que $\mathbb{P}_{st}^{\mathrm{RSP}}$ est la mesure sur $\tilde{\mathcal{P}}_{st}$ qui minimise l'énergie libre.

La *distance d'énergie libre* $d^{\mathrm{FE}}(s,t)$ entre s et t se définit comme (?, ?, ?) :

$$d^{\mathrm{FE}}(s,t) = \frac{F(\mathbb{P}_{st}^{\mathrm{RSP}}) + F(\mathbb{P}_{st}^{\mathrm{RSP}})}{2}$$

cette dernière est toujours métrique (?, ?). On constate immédiatement que cette définition devrait correspondre à la définition de dissimilarité d'énergie libre D^F calculée à partir du flux.

Pour le moment, les correspondances entre ce formalisme et celui qui est développé dans cette thèse restent intuitives, et les preuves ne sont pas encore trouvées. Un des objectifs majeurs qui suivront cette thèse pourrait être d'unifier ces deux formalismes, afin d'avoir un formalisme probabiliste rigoureux, accompagné de quantités intuitives comme les flux. Certaines étapes amenant à ce but ont déjà pu être franchies dans l'article 4, dans lequel il est montré que le flux de transport randomisé peut avoir une interprétation probabiliste.

4 Sources et cibles multiples

Ce chapitre couvre l'étude du flux de transport randomisé dans le cadre le plus général, c'est-à-dire avec de *multiples* sources et cibles. Rappelons que les sources et les cibles, dans un graphe contenant n noeuds, sont définies par deux vecteurs de taille n : f et ρ. Ces vecteurs contiennent respectivement la proportion de flux entrant et sortant en chaque noeud du graphe. Cette façon de définir les entrées et les sorties est très proche de ce qui est fait dans le contexte du *problème de transport optimal de Monge–Kantorovich* sur un graphe (?, ?, ?, ?, ?, ?). Dans ce problème, on suppose qu'un certain nombre de noeuds du graphe ont une *offre* concernant un type de marchandise, alors que plusieurs autres noeuds possèdent une *demande* de la même marchandise. On admet généralement que le total de l'offre est égal à celui de la demande. Le but du problème du transport optimal est de trouver comment acheminer la marchandise depuis les points d'offre à la demande, tout en minimisant les coûts de transports. L'analogie avec les flux se fait facilement : f représente l'offre, ρ la demande, et les flux admissibles tous les chemins possibles qu'empruntent les marchandises afin de d'attribuer l'offre à la demande. La fonctionnelle d'énergie, vue au sein de la section 2.3.2, correspond aux *coûts de transport*, et les flux la minimisant sont alors les chemins que peuvent prendre les marchandises pour satisfaire une solution optimale du problème de transport. Il est néanmoins important de faire la distinction entre une solution du problème de transport, qui est une *attribution* de l'offre à la demande, et un flux minimisant l'énergie, qui contient le nombre de passages des marchandises sur les différents arcs. En réalité, nous verrons dans le deuxième article de ce chapitre qu'il est possible de trouver une *attribution optimale* entre sources et cibles, c'est-à-dire trouver une solution du transport optimal, grâce au flux de transport randomisé.

La première partie de ce chapitre consiste en un petit historique concernant le problème du transport optimal et contient également la définition de ce problème dans le cadre d'un graphe. Vient ensuite un article, soumis au même moment que la rédaction de ces lignes, qui s'intéresse au problème de transport *régularisé*. Ce problème régularisé, en plus d'être capable de trouver une solution au problème en un temps de calcul fortement réduit, propose d'intéressantes connections avec d'autres modèles connus. Ensuite, comme annoncé préalablement, ce chapitre présente un article proposant de trouver une solution grâce au flux de transport randomisé. Cet article montre également les résultats les plus aboutis concernant le flux de transport randomisé, et est, sans l'ombre d'un doute, le plus important de cette thèse. Finalement, ce chapitre est conclu par une analyse des différentes performances des algorithmes, en comparant leur temps de calcul avec celui de l'algorithme du simplexe.

4.1 Le problème du transport optimal

4.1.1 Historique

La brouette de Monge

C'est dans son mémoire, nommé "Mémoire sur la théorie des déblais et des remblais" (?, ?), que le mathématicien Gaspard Monge fit une première modélisation du problème du *transport optimal*, nommé également *brouette de Monge*. Ce mémoire était issu d'un problème pratique, que Monge décrivit de la sorte :

"Lorsqu'on doit transporter des terres d'un lieu dans un autre, on a coutume de donner le nom de Déblai au volume des terres que l'on doit transporter, et le nom de Remblai à l'espace qu'elles doivent occuper après le transport. Le prix du transport d'une molécule étant, toutes choses d'ailleurs égales, proportionnel à son poids et à l'espace qu'on lui fait parcourir, et par conséquent le produit du transport total devant être proportionnel à la somme des produits des molécules multipliées par l'espace parcouru, il s'ensuit que le déblai et le remblai étant donnés de figure et de position, il n'est pas indifférent que telle molécule du déblai soit transportée dans tel ou tel autre endroit du remblai, mais qu'il y a une certaine distribution à faire des molécules du premier dans le second, d'après laquelle la somme de ces produits sera la moindre possible, et le prix du transport total sera un minimum." (Mémoire sur la théorie des déblais et des remblais, p.1)

Si l'on considère l'ensemble des molécules constituant les déblais comme $\mathcal{S}$, et l'ensemble des places sur le remblais comme $\mathcal{T}$, le problème de Monge revient à trouver une fonction de *transport* $T : \mathcal{S} \to \mathcal{T}$ bijective, tel que, le *coût total du transport*, définit par :

$$c(T) := \sum_{s \in \mathcal{S}} c(s, T(s))$$

où $c(s, T(s))$ est le coût de transport d'une molécule de s à $T(s)$, soit minimal. Bien que Monge n'ait pas trouvé la solution à ce problème, il aura tout de même eu le mérite d'en avoir apporté sa première formalisation et engendra, de par ses réflexions, tout un panel de problèmes riches et féconds, tant au niveau des mathématiques appliquées qu'au niveau des mathématiques "pures". Remarquons tout de même que dans cette première formulation, il s'agit d'un ensemble dénombrable de molécules, et que celle-ci sont indivisibles, il s'agit donc d'un problème d'*optimisation combinatoire*.

Le problème sous sa forme générale

Ce problème fut reformulé dans une version continue par Leonid Kantorovitch au début du XX^{e} siècle (?, ?). Le problème continu peut s'exprimer, en utilisant des notions de *théorie de la mesure* (?, ?), sous la forme suivante : soit $\mathcal{X}$ et $\mathcal{Y}$ deux espaces munis respectivement de deux mesures de probabilité μ et ν. On cherche une mesure π sur $\mathcal{X} \times \mathcal{Y}$, possédant comme marges μ et ν, minimisant le *coût total du transport*, définit par :

$$c(\pi) := \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y)$$

où $c(x, y)$ est le coût de transport pour aller de x à y. Au terme de ses recherches, il trouva également un résultat très intéressant, appelé *dualité de Kantorovitch* :

$$\inf_{\pi} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y) = \sup_{(\phi, \psi)} \left(\int_{\mathcal{X}} \phi(x) d\mu(x) + \int_{\mathcal{Y}} \psi(y) d\nu(y) \right)$$

si π a comme marges μ et ν, et ϕ et ψ sont des fonctions réelles, $\phi : \mathcal{X} \to \mathbb{R}$ et $\psi : \mathcal{Y} \to \mathbb{R}$, vérifiant $\phi(x) + \psi(y) \leq c(x, y)$. On peut donner une interprétation intuitive à ce résultat, mais celle-ci a plus de sens dans le cas discret. Nous y reviendrons à la fin de cette section. C'est sous cette forme générale que nous connaissons aujourd'hui le problème du transport optimal. Bien qu'une solution analytique générale ne soit toujours pas connue aujourd'hui, il est possible de le résoudre pour des cas particuliers, et Leonid Kantorovitch en a trouvé des applications concrètes à économie, en particulier en le formulant sous la forme d'*un problème linéaire*.

Aujourd'hui

Ce problème du transport optimal prit un peu la poussière durant la fin du XX^{e} siècle, mais au début de notre ère, se produisit un formidable regain d'intérêt pour celui-ci. On découvrit que celui-ci possédait des liens avec la mécanique des fluides (?, ?), les équations aux dérivées partielles (?, ?) et de nombreux autres domaines des mathématiques (?, ?). On notera que, malgré la littérature abondante sur le sujet, personne n'a encore trouvé de solution analytique à ce problème dans un cadre général. Dans ce qui suit, nous ne proposons qu'une solution obtenue *numériquement* dans un cadre restreint, le *problème de transport optimal sur un graphe*.

4.1.2 Le problème appliqué à un graphe

Voyons maintenant comment ce problème s'exprime sur un graphe. Notons que sur un graphe, l'aspect discret du problème, comme l'avait formulé Monge, réapparaît de manière naturelle, dans le sens où l'ensemble des éléments à transporter $\mathcal{S}$ (les déblais) et l'ensemble des points d'arrivées $\mathcal{T}$ (les remblais), se situant sur les noeuds d'un graphe, sont obligatoirement *dénombrables*. Cependant, contrairement à Monge, ces éléments pourront être *divisés*. Ainsi, il s'agit plutôt d'une application du problème formulé par Kantorovich, mais possédant des *mesures discrètes*. Sans plus attendre, passons à la définition de ce problème.

Le couplage

Supposons un graphe spatial fortement connexe avec n noeuds. La *proportion de l'offre* en chaque noeud est définie par un vecteur $f = (f_i)$, avec $f_i \geq 0$, $\forall i \in \mathcal{V}$ et $f_\bullet = 1$. La *proportion de la demande* sera définie par un autre vecteur de taille n : $\rho = (\rho_i)$ avec $\rho_i \geq 0$ $\forall i \in \mathcal{V}$ et $\rho_\bullet = 1$. Un noeud i ne peut être à la fois offreur et demandeur, c-à-d $f_i \rho_i = 0$, $\forall i \in \mathcal{V}$. Pour faciliter l'analogie entre ce problème et le formalisme des flux, on appellera l'ensemble des noeuds avec une offre, les *sources*, et les noeuds avec une demande les *cibles*. L'ensemble des sources et des cibles sont notés respectivement $\mathcal{S}$ et $\mathcal{T}$.

On appelle *couplage* la matrice $P = (p_{ij})$, de taille $(n \times n)$ vérifiant :

$$p_{i\bullet} = f_i \quad p_{\bullet i} = \rho_i \quad p_{ij} \geq 0 \qquad \forall i, j \in \mathcal{V}$$

On a donc $p_{\bullet\bullet} = 1$. Une composante p_{ij} de cette matrice peut être vue comme la probabilité jointe de choisir le couple (i, j), c'est-à-dire la probabilité qu'une marchandise prise au hasard soit produite en i et acheminée vers j. C'est le couplage qui va définir l'*attribution* des centres de production à la demande.

Le problème primal

Le *problème du transport optimal* revient à trouver un *couplage optimal*, c'est-à-dire un couplage minimisant les chemins parcourus par toutes les marchandises. Ce couplage optimal

est une solution du problème linéaire suivant :

$$\begin{array}{rll} \text{Touver } P \text{ minimisant} & \sum_{ij\in\mathcal{V}} p_{ij} d^{dsp}(i,j) & \\ \text{sous contraintes} & p_{i\bullet} = f_i & \forall i \in \mathcal{V} \\ & p_{\bullet i} = \rho_i & \forall i \in \mathcal{V} \\ & p_{ij} \geq 0 & \forall i,j \in \mathcal{V} \end{array}$$

où, rappelons-le, $d^{dsp}(i,j)$ est la distance du plus court chemin valuée entre i et j. Ce problème a généralement de multiples solutions.

Ce problème est un problème linéaire tout ce qu'il y a de plus classique et il peut être résolu grâce à un algorithme, comme celui de simplexe ou des points intérieurs. Cependant, ces algorithmes sont relativement lents, c'est pourquoi nous allons présenter une autre méthode de résolution dans l'article qui va suivre.

Le problème dual

Il est intéressant de considérer le problème *dual* du problème de transport :

$$\begin{array}{rll} \text{Touver } u \text{ et } v \text{ maximisant} & \sum_{i\in\mathcal{V}} f_i u_i + \sum_{j\in\mathcal{V}} \rho_j v_j & \\ \text{sous contraintes} & u_i + v_j \leq d^{dsp}(i,j) & \forall i,j \in \mathcal{V} \end{array}$$

Dans cette formulation, le vecteur u est associé aux sources et le vecteur v aux cibles. La *dualité de Kantorovich* s'exprime comme :

$$\min_P \left(\sum_{ij\in\mathcal{V}} p_{ij} d^{dsp}(i,j) \right) = \max_{u,v} \left(\sum_{i\in\mathcal{V}} f_i u_i + \sum_{j\in\mathcal{V}} \rho_j v_j \right)$$

sous réserve que P, u et v respectent les contraintes énoncées ci-dessus.

Cette dualité peut prendre l'interprétation suivante (?, ?) : imaginons qu'une personne propose de prendre en charge le transport des marchandises et donne un prix pour l'embarquement et le débarquement des marchandises (des compensations, c'est-à-dire des prix négatifs, peuvent exister), respectivement donnés par u_i et v_j. Cette personne garantit que pour chaque marchandise, le prix d'embarquement plus le prix d'embarquement seront forcement moins cher que le prix de transport de la marchandise, on aura donc $u_i + v_j \leq d^{dsp}(i,j)$. Avec de telles garanties, la personne est bien entendu engagée. Ce nouveau sous-traitant pour le transport va donc essayer de maximiser son gain, exprimé par $\sum_{i\in\mathcal{V}} f_i u_i + \sum_{j\in\mathcal{V}} \rho_j v_j$, tout en respectant ses promesses. Seulement, la dualité de Kantorovich montre finalement qu'il ne pourra gagner, au mieux, que le coût du transport optimal, c'est-à-dire qu'il ne fera aucun bénéfice.

4.1.3 Article 3 : "Transportation clustering : a regularized version of the optimal transportation problem"

Cet article, dont les résultats ont été présentés à la conférence de l'*International Federation of Classification Societies* à Bologne en juillet 2015, a été soumis en décembre 2015 comme publication dans la série *Studies in Classification, Data Analysis and Knowledge Organization* de Springer.

Il présente une version *régularisée* du transport optimal et apporte un algorithme permettant de trouver une solution à celui-ci. Cet algorithme est nommé les *k-médoïdes contraints*, au vu de sa similarité avec celui des k-moyennes, et s'avère bien plus rapide que l'algorithme du simplexe pour trouver une solution au problème du transport optimal. La solution donnée par cet algorithme est similaire au *modèle gravitaire doublement contraint*, connu en géographie.

Points-clés

- Solution primale P et duale u, v du problème de transport régularisé (pp.63-64).
- Analogie avec le modèle gravitaire doublement contraint (p.63).
- Analogie avec l'algorithme des k-moyennes (p.63).
- Étude de la solution sur une grille 24×24 (p.65-67).
- Application de la solution au problème d'attribution des élèves aux écoles dans la ville de Lausanne (pp.67-69).

Remarque

Contrairement au formalisme vu ci-dessus, cet article possède un formalisme *rectangulaire*. On suppose ici que le graphe contient n sources et m cibles, et le couplage s'exprimera alors par une matrice P de taille $(n \times m)$. En réalité, ce formalisme rectangulaire peut facilement être adapté au formalisme carré étudié dans l'introduction de ce chapitre, et cette façon de procéder ne fait que réduire légèrement le temps de calcul.

Transportation clustering: a regularized version of the optimal transportation problem

Guillaume Guex*, Théophile Emmanouilidis*, François Bavaud*

Abstract The present study is motivated by the search of *journey-to-schools assignment plan* for the schoolchildren in the city of Lausanne, where the pedestrian network, children locations and school capacities are entirely known. This problem can be formalized as an *optimal transportation problem* on a weighted network, where a clustering of origins nodes (children locations) into destination nodes (schools) is searched for. This clustering has to minimize a linear *energy* functional, defined as the total distance travelled by children on the network. Usual linear problem solvers, such as the simplex algorithm, are demanding in presence of many origins and destinations, and yield one particular solution among the generally multiple optima.
In this article, we investigate the properties and behavior of a *regularized* version of the problem, where a convex *entropy* functional is added to the *energy* functional, making the solution easier to compute. It also allows to highlight the spatial boundaries between school attendance areas (clusters), as well as to determine the embarkation and disembarkation costs of the dual formulation of the original linear problem. Moreover, the solution of this regularized problem matches the *doubly constrained gravity flow* of quantitative geography. The algorithm developed here is named *constrained k-medoids*, as it is similar to the *soft k-means* algorithm but with metric dissimilarities instead of squared Euclidean ones, and with additional group size constraints.

1 Introduction

Planning the journey-to-school attribution scheme minimizing the total distance travelled by the children is a classical instance of the *optimal transportation problem*, paradigmatic in operations research, yet various in its expression, formalism and applications, as witnessed by e.g. [1, 7, 9, 12] and many others. This prob-

*Departement of Geography, University of Lausanne, Switzerland
gguex@unil.ch, temmanouilidis@unil.ch, fbavaud@unil.ch

lem aims at finding an *assignment* or *coupling*, $P = (p_{ig})$, i.e. a joint probability of children location i and school g with fixed margins, minimizing a linear energy functional $U(P)$, defined as the total distance travelled by children, assuming they follow shortest-paths between their home and schools. The efficiency of the classical simplex algorithm is questionable when dealing with a large number of origins and destinations, and returns only one optimal solution among many equivalent ones in general.

This paper proposes another approach, inspired by statistical mechanics ([3, 8, 11] and many others): the minimization of a *regularized* functional, the *free energy*, $F(P)$, defined by:

$$F(P) := U(P) + TK(P||P^{\infty}) \tag{1}$$

where $K(P||P^{\infty})$ is the *Kullback-Leibler divergence* [8], or *relative entropy*, measuring the dependence between children locations and schools, P^{∞} is the independent assignment (see section 2.3.1), and $T > 0$ is a free parameter, the *temperature*, controlling the importance of this entropy term relatively to the energy term. Adding a convex functional to the original linear functional has already been shown, in an independent contribution [4], to reduce the computing time needed to solve the problem and to make the solution unique. In the low-temperature limit, $T \to 0$, the resulting soft journey-to-school assignment approaches an optimal solution of the original primal problem, and yields embarkation and disembarkation costs of the dual problem. Moreover, this formulation yields interesting correspondences with other problems, in particular with the *doubly constrained gravity flow of quantitative geography* or the *soft k-means algorithm* [2].

In the first part of this article, the formalism needed to express and solve this regularized problem is exposed, as well as results obtained from mathematical developments. In the second part, this formalism is applied to two case studies: a toy example consisting in a 24×24 lattice, as well as the original journey-to-school attribution problem.

2 Formalism

2.1 Context

Let $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ be a simple undirected connected graph, where two disjoint subsets of nodes are defined, the *set of origins*, noted $\mathscr{S}$ with $|\mathscr{S}| = n$, and the *set of destinations*, noted $\mathscr{T}$ with $|\mathscr{T}| = m$. Suppose that every origin i supplies a countable quantity of good, n_i and every destination g demands a countable quantity of the same good, m_g. We suppose the total quantity of supplies is equal to the total quantity of demands, i.e. $N := \sum_i n_i = \sum_g m_g$. Hence, supply and demand can be ex-

pressed by two *relative weights vectors*, namely $f_i := n_i / \sum_j n_j$ and $\rho_g := m_g / \sum_h m_h$.

An origin-destination *assignment* or *coupling* is defined as the matrix $P = (p_{ig})$ obeying the constraints:

$$p_{ig} \geq 0, \quad p_{\bullet\bullet} = 1, \quad p_{i\bullet} := \sum_g p_{ig} = f_i \quad \text{and} \quad p_{\bullet g} := \sum_i p_{ig} = \rho_g \ . \tag{2}$$

This assignment represents, in the traditional formulation of the transportation problem, the probability of picking a good going from i to g among all goods. The set of all assignments is written $\mathscr{P}$ and is clearly a convex set, i.e. $\alpha P + (1-\alpha)Q \in \mathscr{P}$, $\forall P, Q \in \mathscr{P}$ and $\forall \alpha \in [0,1]$.

Note that an origin-destination *membership matrix* Z can be derived from this assignment, defined by $z_{ig} := p_{ig}/f_i$. Components of this membership matrix z_{ig} yield the proportion of goods at i sent to g.

2.2 Transportation problem

2.2.1 Primal problem

The *energy* or the *cost* of an assignment P, noted $U(P)$, is the sum of all path distances taken by goods:

$$U(P) := \sum_{ig} p_{ig}\, d_{ig} \tag{3}$$

where d_{ig} is the *shortest-path distance* between node i and g. The *optimal transportation problem* in this context consists in finding an assignment minimizing the energy functional. It can be expressed through the following linear problem:

$$\begin{aligned}
\text{Find } P\text{, minimizing} \quad & U(P) = \textstyle\sum_{ig} p_{ig}\, d_{ig} && && (4)\\
\text{subject to} \quad & p_{i\bullet} = f_i && \forall i && (5)\\
& p_{\bullet g} = \rho_g && \forall g && (6)\\
& p_{ig} \geq 0 && \forall i, g && (7)
\end{aligned}$$

2.2.2 Dual problem

The dual version of this problem can be written as:

$$\begin{aligned}
\text{Find } u, v\text{, maximizing} \quad & \textstyle\sum_i f_i\, u_i + \sum_g \rho_g\, v_g && && (8)\\
\text{subject to} \quad & u_i + v_g \leq d_{ig} && \forall i, g && (9)
\end{aligned}$$

where u is a vector of size n and v a vector of size m. These two vectors can be interpreted as follows [7, 12]: imagine an external contractor wishes to take over the delivery of the goods, and s/he has to fix a price for their embarkement and disembarkement (with possibly negative prices). This contractor would like to maximize her/his profit, which is expressed by $\sum_i f_i u_i + \sum_g \rho_g v_g$. However, for every good, s/he has to make the combined embarkment and disembarkment price cheaper or equal than the cost of the displacement, otherwise s/he will not be hired. The latter is expressed through $u_i + v_g \leq d_{ig}$, $\forall ig$. When the latter constraints are respected, the *Kantorovich duality* [7] asserts that:

$$\min_P \left(\sum_{ig} p_{ig}\, d_{ig} \right) = \max_{u,v} \left(\sum_i f_i\, u_i + \sum_g \rho_g\, v_g \right) \tag{10}$$

2.3 Regularized problem

2.3.1 Definition

The regularized problem consists in finding an assignment P minimizing the *free energy* $F(P)$, defined by:

$$F(P) := U(P) + TK(P||P^\infty) \tag{11}$$

where $T > 0$ is a free parameter, the *temperature*, and $K(P||P^\infty)$ is the *Kullback-Leibler divergence* or *relative entropy*, defined as:

$$K(P||P^\infty) := \sum_{ig} p_{ig} \log \frac{p_{ig}}{f_i\, \rho_g} \tag{12}$$

which measures the divergence between the assignment P and the *independent assignment*, $P^\infty = (p^\infty_{ig}) := (f_i \rho_g)$. When $T \to 0$, the free energy converge to the energy $U(P)$ and P minimizing it converge to an optimal assignment of the transportation problem. By contrast, when $T \to \infty$, we have $P \to P^\infty$. Note that this Kullback-Leibler divergence can also be interpreted as the *mutual information* between origins and destinations [8, 11]:

$$K(P||P^\infty) = H(f) + H(\rho) - H(P) \tag{13}$$

where $H(.)$ denotes the *entropy* of distributions defined by f, ρ and P.

Altogether, the regularized problem can be written as:

$$\text{Find } P, \text{ minimizing} \quad F(P) = \sum_{ig} p_{ig}\, d_{ig} + T \sum_{ig}\, p_{ig} \log \frac{p_{ig}}{f_i \rho_g} \tag{14}$$

$$\text{subject to} \quad p_{i\bullet} = f_i \qquad \forall i \tag{15}$$

$$p_{\bullet g} = \rho_g \qquad \forall g \tag{16}$$

$$p_{ig} \geq 0 \qquad \forall ig \tag{17}$$

We can check that $K(P||P^\infty)$ is convex for $P \in \mathscr{P}$, as shown by the positive definiteness of all its Hessian. As the admissible domain $\mathscr{P}$ is also convex, the uniqueness of the solution is ensured.

2.3.2 Solution

By adding Lagrangian multipliers λ_i, μ_g, corresponding respectively to constraints (15) and (16) (constraint (17) turns out to be inactive because of the entropy term), to functional (14), the *Lagrangian* of the regularized problem reads:

$$L(P, \lambda, \mu) := F(P) - \sum_i \lambda_i \left(\sum_g p_{ig} - f_i \right) - \sum_g \mu_g \left(\sum_i p_{ig} - \rho_g \right) \tag{18}$$

By setting the Lagrangian derivative to zero, we get:

$$p_{ig} = f_i \rho_g \exp(\beta \lambda_i) \exp(\beta \mu_g) \exp(-\beta d_{ig}) \tag{19}$$

where $\beta = 1/T$ is the *inverse temperature*. This expression is referred to as the *doubly-constrained gravity model* of Quantitative Geography (see e.g. [5, 6, 13] and references therein).

Substituting (19) in (15), respectively (16), yields:

$$\exp(\beta \lambda_i) = \frac{1}{\sum_g \rho_g \exp(\beta \mu_g) \exp(-\beta d_{ig})} \tag{20}$$

$$\exp(\beta \mu_g) = \frac{1}{\sum_i f_i \exp(\beta \lambda_i) \exp(-\beta d_{ig})} \tag{21}$$

Equations (20) and (21) define an iterative process. Starting with some non-null μ^0, λ^i can be computed with (20) and μ^{i+1} with (21). When convergence occurs, i.e. $\mu^i = \mu^{i+1}$ (within a defined margin of error), the assignment is given by (19). We refer to this algorithm as the *constrained k-medoids*, as the procedure is analogous to the *soft k-means* algorithm (see e.g. [2]), where d_{ig} is metric instead of squared Euclidean, and the groups size are constrained to ρ_g.

Replacing (19) into $F(P)$ yields:

$$\min_P F(P) = \sum_i f_i \lambda_i + \sum_g \rho_g \mu_g \tag{22}$$

Furthermore, the condition $p_{ig} \leq 1$ in (19) entails:

$$\lambda_i + \mu_g \leq d_{ig} + T \log\left(\frac{1}{f_i \rho_g}\right) \tag{23}$$

When $T \to 0$, $\min_P F(P) = \sum_i f_i \lambda_i + \sum_g \rho_g \mu_g$ converges to $\min_P U(P)$, and (23) reads $\lambda_i + \mu_g \leq d_{ig}$. That is (cf. (9)) Lagrangian multipliers λ and μ converge to a solution of the dual problem, and can be interpreted as embarkment and disembarkment costs for low temperatures.

3 Case studies

3.1 Overview

In this section, the constrained k-medoids, defined by (19), (20) and (21), is tested on two graphs. The first one is an artificial graph, a 24×24 lattice, on which the properties and performances of this algorithm are examined. The second case study is the problem which inspired this article, the optimal journey-to-school assignment plan for the city of Lausanne. Note that, while in the formalism section we mainly worked with the assignment matrix P, here we use more frequently the membership matrix, $Z = (z_{ig}) := (p_{ig}/f_i)$, obeying $z_{ig} \geq 0$ and $z_{i\bullet} = 1$.

Besides the optimal assignment P and Lagrangian multipliers λ, μ, different quantities will be studied on these case studies:

1. The *entropy of origins attribution* $H(i)$, defined by:

$$H(i) := -\sum_g z_{ig} \ln z_{ig} \tag{24}$$

   measuring the uncertainty of destination for each origin i. Mapping this entropy reveals destination pools and their boundaries.

2. The *total variation placement error*. This algorithm always results in a soft partition $Z = (z_{ig}) = (p_{ig}/f_i)$ respecting the destinations capacities, i.e. $\sum_i f_i z_{ig} = \sum_i p_{ig} = \rho_g$. However, in real life applications, a hard partition, $Y = (y_{ig})$ with $y_{ig} = 0$ or $y_{ig} = 1$, is sometimes needed (when origins can not be split), and the result z_{ig} needs to be *hardened*, by $y_{ig} = 1 \Leftrightarrow z_{ig} = \max_h z_{ih}$ (solving ties at random). However, the resulting hard partition does not respect the destinations capacities in general, i.e. $\sum_i f_i y_{ig} \neq \rho_g$. The placement error can be measured by the *total variation placement error* $\delta(Y,Z)$, defined by:

$$\delta(Y,Z) := \sum_g |\sum_i f_i (y_{ig} - z_{ig})| = \sum_g |\sum_i f_i y_{ig} - \rho_g| \tag{25}$$

This quantity is bound above by Pinsker inequality [10]: $\delta(Y,Z) \leq \sqrt{2K(Y||Z)}$.

3. The *calculation time* of the algorithm, depending on the computer, language and code used. This calculation time is compared with the one of the simplex under same circumstances, showing how the two algorithms perform against each other. In this article, calculation times have been obtained with R running on an Intel Xeon E5 3.7Ghz with 64 Go RAM DDR3. The simplex algorithm is performed by the package "boot" of R.

3.2 The 24×24 *lattice*

The case study consists of a 24×24 regular lattice, with a distance of 1 between adjacent nodes. The shortest-path distance here correspond to the *Manhattan distance*. $m = 5$ destinations locations are randomly chosen among nodes with an uniform probability and the rest of the nodes are set to be origins, i.e $n = 571$. We set uniform weights for origins and destinations, that is $f_i = \frac{1}{571}$ and $\rho_g = \frac{1}{5}$.

3.2.1 Solution, Lagrangian multipliers and Entropy

Fig. 1 depicts a hardened membership matrix Y for a particular random placement of destination and $\beta = 1$. The solution has a total variation placement error $\delta(Y,Z)$ of 0.097, which is quite good considering that further increasing of β would still lower the placement error. Fig. 2 depicts Lagrangian multipliers λ_i, μ_g, and the entropy of origins $H(i)$ under the same conditions. Lagrangian multipliers are defined up to a common additive constant, and, in this realization, λ_i are negative and μ_g positive. High λ_i multipliers denote origins i placed far away from their assigned destinations, and high μ_g a expensive locations g relatively to origins. The parameter β was not set too large in order to highlight the boundaries between groups/catchment areas, as shown by the entropy $H(i) > 0$.

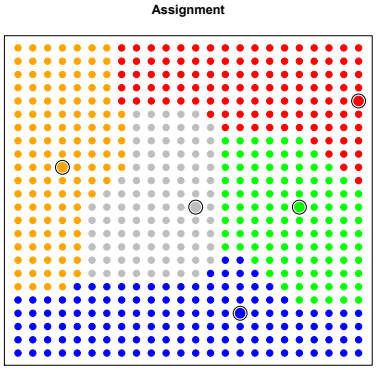

Fig. 1 Example 3.2. A hard membership matrix Y obtained by the constrained k-medoids with $\beta = 1$. This solution has a total variation placement error of $\delta(Y,Z) = 0.097$.

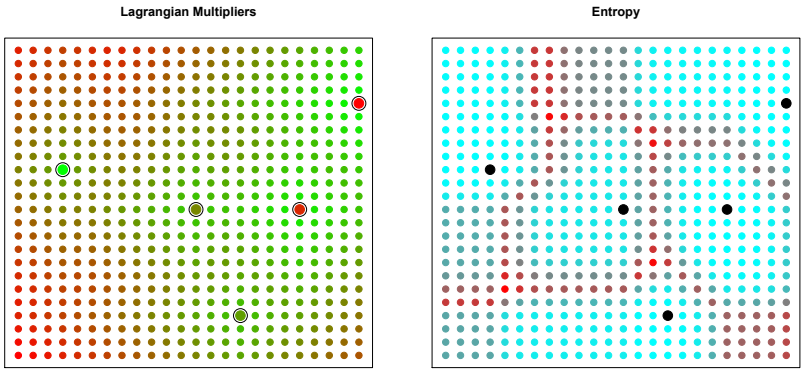

Fig. 2 Example 3.2, continued. Lagrangian multipliers and Entropy of origins for $\beta = 1$. Left: Lagrangian multipliers λ and μ. Red denotes highest values. λ ranges within $[-24.1, -1.1]$ and μ ranges within $[18.4, 26.8]$. Right: Entropy of origins $H(i)$. Red denotes highest values. $H(i)$ ranges within $[10^{-5}, 1.1]$.

3.2.2 Computation time and Error

In Fig. 3, 50 samples of destination locations were generated randomly for each value of β, and the mean computation time is computed for the constrained k-

medoids and the simplex algorithm. The computing time for the simplex algorithm has a mean of 139.6 seconds with a standard deviation of 7.8 seconds. We observe that the constrained k-medoids converges, within a margin of error defined by $\max_g |\mu_g^i - \mu_g^{i+1}| < 10^{-5}$, in considerably less computing time, even for high β. The total variation placement error $\delta(Y,Z)$ is also computed for each β and, as expected, tends to decrease for increasing β.

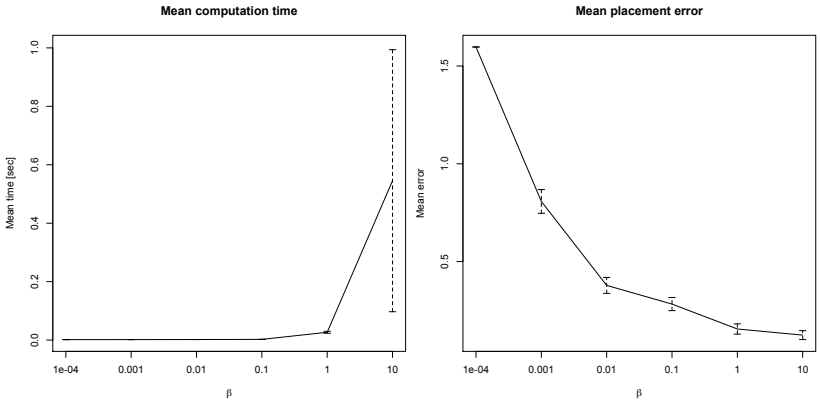

Fig. 3 Example 3.2, continued. Left: Evolution of constrained k-medoids computing time versus β. Right: Evolution of constrained k-medoids placement error vs β. (with 95% confidence intervals)

3.3 The journey-to-school assignment plan for the city of Lausanne

In this real life example, we consider the pedestrian network of the city of Lausanne (total length: 580km). Origins consist in $n = 2887$ children locations, with weights f_i obtained by dividing the number of children in every location by the total number of children (7055). Destinations are made of $m = 44$ schools, with weights ρ_g standing for their relative capacity. Home school distances d_{ig} are computed with the Djiksktra algorithm through the pedestrian network. Note that all school degrees were merged together for this experiment. Hard membership matrix Y and a spatial interpolation [1] of origins entropy are depicted in Fig. 3.3. Hard membership reveals the attendance "area" of each school that can be merged to create new school districts. Entropy depicts parts of the city where the schoolchildren are more likely to be assigned to multiple schools. With $\beta = 0.2$, the placement error is small: $\delta(Y,Z) = 0.057$. Lagrangian multipliers spatial interpolation under same circumstances[1] are shown in Fig. 5. Students in blue areas have a closer access to

[1] Ordinary kriging in ArcGIS. Semi-variogram. Type: spherical, search_radius 100m

schools than the ones in yellow or red areas. Lowest values of μ_g occur when the students of a school come from far away, or when the school is subject to a strong demand, i.e a school with a small capacity located in a high-density neighborhood.

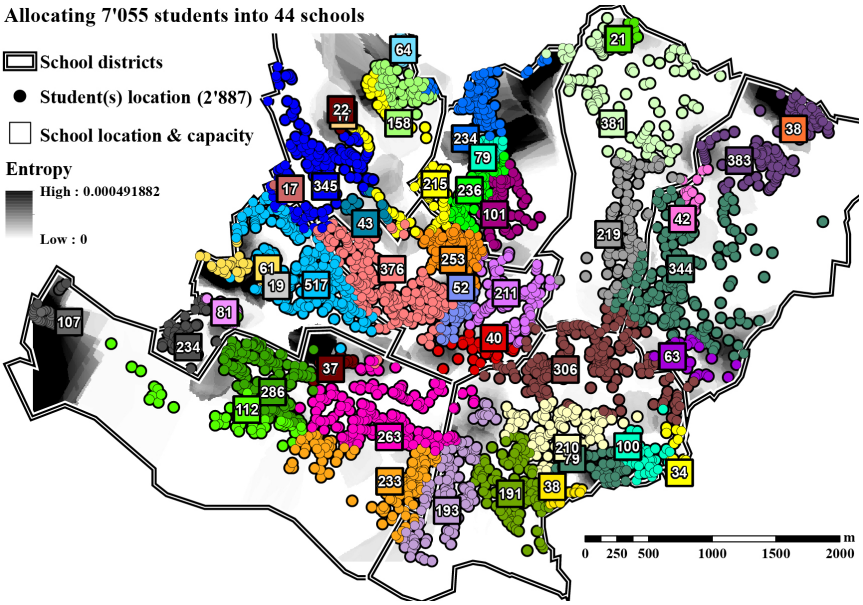

Fig. 4 Example 3.3. The hard children membership matrix Y obtained by the constrained k-medoids algorithm with $\beta = 0.2$, and spatial interpolation of origins entropy. This solution has a total variation placement error of $\delta(Y,Z) = 0.057$.

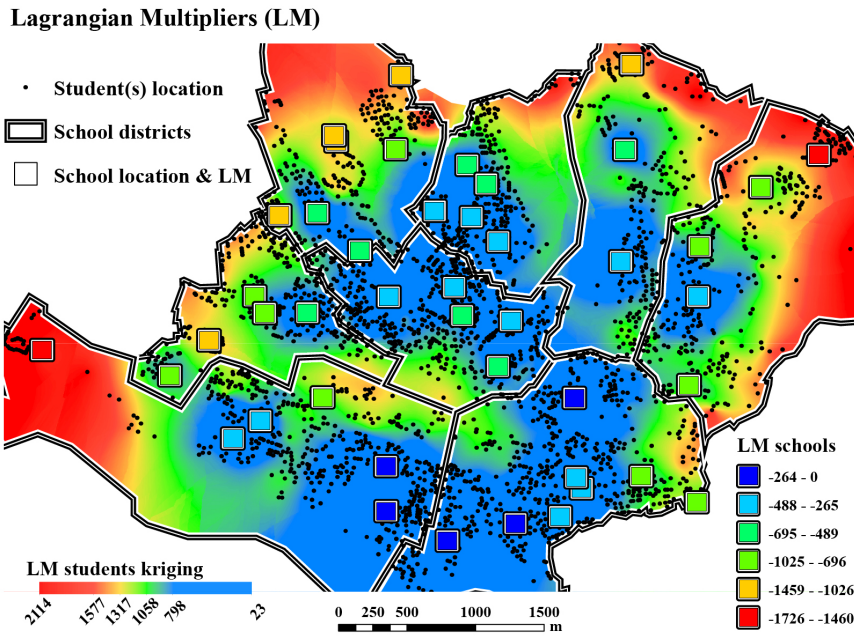

Fig. 5 Example 3.3, continued. Spatial interpolation of Lagrangian multipliers λ, along with multipliers μ, obtained by the constrained k-medoids under same circumstances.

4 Conclusion

Regularized optimal transportation, as developed in this paper, presents quite a few noticeable aspects:

First, the computation time was reduced drastically, as shown in Fig. 3. For solving the children assignment plan for the city of Lausanne, the simplex algorithm was not even attempted, as it would have needed several days to find a solution. Note that the constrained k-medoids computation time, as well as the precision of the optimal solution, strongly depends on the value of β. When the latter algorithm is used, β should be set to a value offering a good balance between speed and precision.

The second aspect is the uniqueness of the solution. Possessing a unique solution is desirable for consistency, and its softness allows the spatial mapping of origins entropy, but it also comes with a price: while the linear problem can be expressed with integer numbers, such as the journey-to-school assignment problem, the simplex algorithm generates a solution respecting the integrity of units on every origins [1]. However, this is not the case anymore with the constrained k-medoids algo-

rithm, as the unique solution is in fact a mixture of all optimal solutions. A way to construct a solution respecting the integrity of units from this mixture of solutions is an open question.

The third aspect is the attractive meaning of quantities revealed by this algorithm, especially the Lagrangian multipliers λ and μ. Their interpretation as embarkment and disembarkment costs when $T \to 0$, as shown by (22) and (23), demonstrates that they both contain information about location and constraints. For example, in the city of Lausanne, large values of the multipliers denote zones where improvements have to be made, either by placing more schools in these spots or by increasing the capacities of the schools in the vicinity. The multipliers appear again in equation (19), allowing the computation of P, and displaying the connection to the doubly-constrained gravity model of the geographers. Also, note the multipliers μ, insuring the group capacity constraints, are absent in the usual soft k-means algorithms.

References

1. Ahuja, R.K., Magnanti, T.L., Orlin, J.B.: Network flows: theory, algorithms, and applications. Prentice hall (1993)
2. Bavaud, F.: Aggregation invariance in general clustering approaches. Advances in Data Analysis and Classification **3**(3) (2009) 205–225
3. Boltzmann, L.: Lectures on gas theory. Univ of California Press (1964)
4. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: Advances in Neural Information Processing Systems. (2013) 2292–2300
5. Erlander, S., Stewart, N.F.: The gravity model in transportation analysis: theory and extensions. Volume 3. Vsp (1990)
6. Fotheringham, A., O'Kelly, M.E.: Spatial interaction models: formulations and applications. Volume 5. Kluwer Academic Pub (1989)
7. Kantorovich, L.V.: On the translocation of masses. In: Dokl. Akad. Nauk SSSR. Volume 37. (1942) 199–201
8. Kullback, S.: Information theory and statistics. Courier Corporation (1968)
9. Monge, G.: Mémoire sur la théorie des déblais et des remblais. De l'Imprimerie Royale (1781)
10. Pinsker, M.S.: On estimation of information via variation. Problems of Information Transmission **41**(2) (2005) 71–75
11. Shannon, C.E., Weaver, W.: The mathematical theory of information. (1949)
12. Villani, C.: Optimal transport: old and new. Volume 338. Springer (2008)
13. Wilson, A.G.: The use of entropy maximising models, in the theory of trip distribution, mode split and route split. Journal of Transport Economics and Policy (1969) 108–126

4.2 Résolution avec les flux

Comme annoncé préalablement, nous allons à présent étudier comment résoudre le problème du transport optimal grâce au flux de transport randomisé. Rappelons que les vecteurs f et ρ, désignant les entrées et les sorties du flux dans le réseau, peuvent également être vus comme l'offre et la demande d'une marchandise sur les noeuds du graphe. Un flux admissible, c'est-à-dire respectant les contraintes sur les sources et les cibles, est alors une représentation des flux de transport permettant de joindre l'offre à la demande. Nous avons vu que le flux de transport randomisé minimise les coûts de déplacements dans le cas où le température approche zéro, et est donc une représentation des flux d'une des solutions du transport optimal. Mais ce n'est pas tout. Il nous est également possible, sans trop de difficultés, de trouver un couplage optimal à partir du flux de transport optimal. Dans la section précédente, nous avons vu comment obtenir un couplage optimal de manière rapide. Seulement, pour obtenir un flux à partir de ce couplage, il faudrait utiliser en outre un algorithme de plus court chemin, tel celui de Dijkstra, entre toutes les paires de noeuds dont le couplage est positif, ce qui est loin d'être efficient. Grâce au flux de transport randomisé, ces deux étapes sont effectuées en même temps, réduisant considérablement le temps de calcul.

Ce n'est pas le seul intérêt du flux de transport randomisé avec sources et cibles multiples. En effet, et cela a été évoqué dans les articles du chapitre 3, le bilan de flux de transport randomisé lorsque la température tend vers l'infini décrit un flux électrique (une preuve formelle est fourni dans l'article qui suit). Lorsque les sources et les cibles sont multiples, elles désignent alors des entrées et des sorties du courant dans un réseau électrique et les capacités C du graphe deviennent des capacités électriques. Ce résultat n'a peut-être pas autant d'utilité pratique que le résolution du transport optimal par le flux, mais cette analogie entre marche aléatoire, flux, couplage et électricité soulève d'importantes questions fondamentales.

L'article contenu dans cette section est particulièrement long, et aborde tous les aspects touchants à cette problématique. C'est pourquoi il n'y aura pas, contrairement aux autres sections, d'explications préalables.

4.2.1 Article 4 : "Interpolating between Random Walks and Optimal transportation routes : flow with multiple sources and targets"

Cet article a été accepté comme publication pour la revue *Physica A* d'Elsevier en décembre 2015 et devrait être disponible durant le cours de l'année 2016.

Il présente un algorithme permettant de calculer le flux de transport régularisé dans le cadre le plus général, avec de multiples sources et cibles. Contrairement au cas où les sources et les cibles sont uniques, où la cible fait office de noeud absorbant, il faut ici ajouter un noeud *cimetière* supplémentaire, relié à toutes les cibles. Le calcul de la probabilité de passage des cibles au cimetière demande de résoludre en premier lien la solution pour $T \to \infty$, et fait apparaître, grâce aux dérivations, un paramètre supplémentaire. La solution finale est étudiée sous toutes ses coutures et nous recommandons au lecteur de lire très attentivement cet article, car il contient le formalisme et les résultats les plus aboutis de cette thèse.

Points-clés

- Extension du graphe avec un noeud absorbant, et résolution du problème pour $T \to \infty$ (pp.77-79).
- Interprétation du flux aléatoire net comme d'un flux électrique (p.79).
- Solution du flux de transport régularisé dans le cadre le plus général (pp.80-81).
- Interprétation probabiliste du flux de transport régularisé (p.81).
- Obtention du couplage à partir du flux de transport régularisé (p.82).
- Étude de la solution sur une grille 10×10 (pp.83-86).
- Application de la solution au problème d'attribution des élèves aux écoles dans un quartier de la ville de Lausanne (pp.87-89).

Interpolating between Random Walks and Optimal transportation routes: flow with multiple sources and targets

Guillaume Guex

Department of Geography,
University of Lausanne
guillaume.guex@unil.ch

Abstract

In recent articles about graphs, different models proposed a formalism to find a type of path between two nodes, the source and the target, at crossroads between the shortest-path and the random-walk path. These models include a freely adjustable parameter, allowing to tune the behavior of the path towards randomized movements or direct routes. This article presents a natural generalization of these models, namely a model with *multiple* sources and targets. In this context, source nodes can be viewed as locations with a supply of a certain good (e.g. people, money, information) and target nodes as locations with a demand of the same good. An algorithm is constructed to display the flow of goods in the network between sources and targets. With again a freely adjustable parameter, this flow can be tuned to follow routes of minimum cost, thus displaying the flow in the context of the *optimal transportation problem* or, by contrast, a random flow, known to be similar to the *electrical current flow* if the random-walk is reversible. Moreover, a *source-target coupling* can be retrieved from this flow, offering an optimal assignment to the transportation problem. This algorithm is described in the first part of this article and then illustrated with case studies.

Keywords: Functional minimization, Random walks, Optimal transportation problem, Multiple sources and targets

1. Introduction

Over the years, the old *Monge-Kantorovich optimal transportation problem* has continued to show its usefulness and richness for numerous applications [1, 2, 3, 4, 5] (and many others). When presented in a graph setting [4], this problem consists in finding the optimal assignment of a resource supplied by a countable set of nodes, the *sources*, to another countable set of nodes, the *targets*, while minimizing the *cost* $U(X)$ of the transportation *flow* X. Different algorithms are capable of finding optimal solutions while respecting constraints of source supplies and target demands [6, 7, 8]. However, none are convenient to display routes of transportation, as it requires to solve the allocation problem first and, subsequently, run a shortest-path algorithm for every source-target pair. In this article, a new algorithm allowing the visualization of optimal transportation routes is presented. This algorithm works with the well-known principle of *regularization*. By adding a suitable nonlinear functional to the linear functional $U(X)$, the resulting functional becomes derivable and an approximation of the optimal solution can be found with considerably less computing effort [9, 10]. Here, this functional is named the *entropy* $G(X)$, as flows X minimizing it will display a random behavior, and the new functional $F(X) := U(X)+TG(X)$ is named the *free energy*, with a freely adjustable parameter $T > 0$, the *temperature*. Defining this new objective functional $F(X)$ does not only serve the purpose of reducing computation time, but also enables modeling uncertainty in transportation, which can be more realistic than using a deterministic model in real-life situations. When $T \to 0$, optimal transportation routes are displayed, but when $T \to \infty$, the cost minimization is negligible and the flow follows a random-walk pattern. If this random-walk is reversible, it has been shown to be similar to the *electrical current flow* with sources and targets being respectively current inputs and outputs [11, 12]. A *source-target coupling*, i.e. an assignment of resources between sources and targets, can be retrieved from this flow. When $T \to 0$, this coupling gives an optimal solution to the transportation problem. To summarize, this algorithm enables to model an array of flows, ranging from optimal transportation routes to random-walk routes, when supply and demand on nodes are fixed, and to find a source-target coupling corresponding to these flows.
This algorithm is, in fact, a generalization of the one exposed in [13, 14] and the notation is similar, even though not entirely compatible. In the latter, and in other articles, e.g. [15, 16, 17, 18], the *randomized shortest-path* computation is limited to two nodes, whereas in the present case, multiples sources and targets are allowed.

The present article is divided in two parts. The formalism needed to construct the algorithm is exposed first, followed by illustrations of the algorithm running on a toy graph and a real graph.

2. Formalism

2.1. Admissible flows

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a simple oriented connected graph with $|\mathcal{V}| = n$, and $\mathcal{S}, \mathcal{T} \subset \mathcal{V}$, $\mathcal{S} \cap \mathcal{T} = \emptyset$, respectively, the set of *sources* and the set of *targets*. Each source generates a defined flow, which is transported on the graph before being absorbed by targets. Let us define $f = (f_i)$ with $f_i > 0 \; \forall i \in \mathcal{S}$, $f_i = 0 \; \forall i \notin \mathcal{S}$ and $\sum_i f_i = 1$, the *in-flow vector*, representing the proportion of the flow created by source nodes. Similarly, let $\rho = (\rho_i)$ with $\rho_i > 0 \; \forall i \in \mathcal{T}$, $\rho_i = 0 \; \forall i \notin \mathcal{T}$ and $\sum_i \rho_i = 1$ be the *out-flow vector*, representing the proportion of the flow absorbed by target nodes. All these quantities are provided initially.

The unknown *flow matrix*, noted $X = (x_{ij})$, represents the quantity of flow on arcs (i, j), $\forall i, j \in \mathcal{V}$. Components x_{ij} must verify:

$$x_{ij} \geq 0 \qquad \text{positivity} \tag{1}$$

$$x_{i\bullet} - x_{\bullet i} = f_i - \rho_i \qquad \text{unit flow conservation} \tag{2}$$

An alternative way to consider the out-flow vector is as a set of constraints on flows coming from each node to a "virtual" node ω, the *ground node*. Similarly, the in-flow vector can be viewed as constraints set on flows coming from another "virtual" node ϕ, the *generator node*, to $\mathcal{V}$. Formally, $\forall i \in \mathcal{V}$:

$$x_{\phi i} = f_i \qquad x_{i\phi} = 0 \tag{3}$$

$$x_{i\omega} = \rho_i \qquad x_{\omega i} = 0 \tag{4}$$

Further in this article, the ground node must be added to the original graph in order to solve the problem.

Note that the *set of admissible flows*, $\mathcal{X}$, that is the set of flows verifying constraints (1) and (2), is a *convex set*, i.e. if X and Y are in $\mathcal{X}$, so is their *mixture* $\alpha X + (1 - \alpha)Y$, $\forall \alpha \in [0, 1]$.

2.2. Flow entropy and energy

Let $W = (w_{ij})$ be the $(n \times n)$ transition matrix of some irreducible Markov chain defined on $\mathcal{G}$. A flow matrix X will follow the random-walk defined by W iff $x_{ij}/x_{i\bullet} = w_{ij}$ for all nodes i with $x_{i\bullet} > 0$. Therefore a "random-walk" flow will minimize the *entropy functional*:

$$G(X) = G(X||W) := \sum_{i,j\in\mathcal{V}} x_{ij} \ln \frac{x_{ij}}{x_{i\bullet} w_{ij}} = x_{\bullet\bullet} \sum_{i\in\mathcal{V}} \frac{x_{i\bullet}}{x_{\bullet\bullet}} K_i(X||W) \qquad (5)$$

where $K_i(X||W) := \sum_{j\in\mathcal{V}} \frac{x_{ij}}{x_{i\bullet}} \ln \frac{x_{ij}}{x_{i\bullet} w_{ij}} \geq 0$ is the Kullback-Leibler divergence between the transition distributions X and W. This divergence is weighted by $x_{i\bullet}/x_{\bullet\bullet}$ to take into account the visit frequency of nodes, and the sum of flows on all arcs, $x_{\bullet\bullet}$, is included in $G(X)$ to transform the entropy in a *homogeneous* functional, that is $G(\nu X) = \nu G(X)$ for $\nu > 0$, reflecting the *extensivity* of $G(X)$ in the thermodynamic sense.

By contrast, flows minimizing the following *energy functional*:

$$U(X) = U(X||R) := \sum_{i,j\in\mathcal{V}} r_{ij} x_{ij} \qquad (6)$$

where $r_{ij} > 0$ represents a cost or a resistance associated to the arc (i, j), will yield the flow in the context of the *optimal transportation* [4], where f represents the proportion of supply and ρ the proportion of demand on every node. This quantity is actually the *total cost* of the flow on the network.

By analogy with Statistical Mechanics, a *free energy functional* can be defined:

$$F(X) = U(X) + TG(X) \qquad (7)$$

where $T > 0$ is a free parameter, the *temperature*, controlling the importance of the entropy functional with respect to the energy functional. When $T \to 0$, flows minimizing free energy follow optimal transportation routes, while $T \to \infty$, the minimization results in flows following the random-walk defined by W.

Combining (1), (2), and (7) results in the following optimization problem:

$$\text{minimize} \quad \sum_{ij} r_{ij} x_{ij} + T \sum_{ij} x_{ij} \ln \frac{x_{ij}}{x_{i\bullet} w_{ij}} \qquad (8)$$

$$\text{subject to} \quad x_{i\bullet} - x_{\bullet i} = f_i - \rho_i \qquad \forall i \in \mathcal{V} \qquad (9)$$

$$x_{ij} \geq 0 \qquad \forall i, j \in \mathcal{V} \qquad (10)$$

The entropy functional is convex, i.e. for two admissible flows $X, Y \in \mathcal{X}$, we have $G(\alpha X + (1-\alpha)Y) \leq \alpha G(X) + (1-\alpha)G(Y)\ \forall \alpha \in [0,1]$. With $U(X)$ being linear, the resulting free energy $F(X)$ is also convex. Thanks to the convexity of the domain $\mathcal{X}$, both the existence and the uniqueness of the solution are ensured.

2.3. Graph extension

To solve this problem, it is convenient to add the terminal ground node ω to the graph. As a matter of fact, solving this problem requires to introduce an additional *absorbing* node in the graph, as in [19]. Let the set $\mathcal{V}^+ = \mathcal{V} \cup \{\omega\}$ denote the set of nodes including ω, and the index "+" a summation over $\mathcal{V}^+$ (e.g. $x_{i+} = \sum_{j \in \mathcal{V}^+} x_{ij}$ while $x_{i\bullet} = \sum_{j \in \mathcal{V}} x_{ij}$). With this new set of nodes, constraints (9) can be rewritten as:

$$x_{i+} - x_{+i} = f_i \qquad \forall i \in \mathcal{V} \tag{11}$$
$$x_{i\omega} = \rho_i \qquad \forall i \in \mathcal{V} \tag{12}$$
$$x_{\omega+} = 0 \tag{13}$$

expressing flow creation on nodes $i \in \mathcal{S}$ and conservation on nodes $i \notin (\mathcal{S} \cup \{\omega\})$ for (11), constrained flow on $x_{i\omega}\ \forall i \in \mathcal{V}$ for (12), and absorption in ω for (13).

By extending the graph, we should use extended functionals $U^+(X) = \sum_{i,j \in \mathcal{V}^+} r_{ij} x_{ij}$ and $G^+(X) = \sum_{i,j \in \mathcal{V}^+} x_{ij} \ln \frac{x_{ij}}{x_{i+} w_{ij}}$ and therefore, R and W should also be extended to R^+ and W^+. However, we can see that extending $U(X)$ and R is not necessary, as it only adds a constant term $\sum_{i \in \mathcal{V}} r^+_{i\omega} \rho_i$ to the original energy, whatever the values of $r^+_{i\omega}$ are.

By contrast, the extension of $G(X)$ and W into $G^+(X)$ and W^+ is more delicate: setting $w^+_{i\omega}$ affects all w^+_{ij}, $\forall i, j \in \mathcal{V}$, as $w^+_{i+} = 1$ for a Markov chain matrix. This extended Markov chain matrix W^+ has the form:

$$\begin{aligned} w^+_{ij} &= (1-\alpha_i) w_{ij} \quad \forall i, j \in \mathcal{V} \\ w^+_{i\omega} &= \alpha_i \quad \forall i \in \mathcal{V} \\ w^+_{\omega i} &= 0 \quad \forall i \in \mathcal{V} \\ w^+_{\omega\omega} &= 1 \end{aligned} \tag{14}$$

where the probabilities $\alpha_i \in [0,1]$ to jump from targets to the absorbing state ω have to be determined (clearly, $\alpha_i = 0$ if $i \notin \mathcal{T}$). The condition we

impose on these α_i is that when $T \to \infty$, the solution $X^\infty = (x_{ij}^\infty)$ needs to verify $x_{ij}^\infty / x_{i+}^\infty = w_{ij}^+$. The next section will detail the solution for this specific problem.

2.4. High-temperature limit solution

To obtain α_i and the high-temperature limit solution X^∞, the following proposition can be used, whose demonstration lies in appendix A.

Proposition 1. *In order to set W^+, given by (14), to have a admissible flow X^∞ verifying $x_{ij}^\infty / x_{i+}^\infty = w_{ij}^+$, the probability to jump to ω from i, i.e. α_i, must be set to:*

$$\alpha_i = \frac{\rho_i}{x_{i+}^\infty} \quad \forall i \in \mathcal{V}$$

where the vector $x_+^\infty := (x_{i+}^\infty)_{\forall i \in \mathcal{V}}$ can be found with

$$\begin{aligned} x_+^\infty &= \widetilde{x}_+^\infty + \gamma q \\ \widetilde{x}_+^\infty &= Q^*(f - W'\rho) \\ q_i &= [I - Q^*Q]_{i1} \\ Q &= I - W' \end{aligned}$$

where Q^* denotes the Moore-Penrose pseudoinverse of Q and γ a free parameter obeying:

$$\gamma \geq \gamma_0 := \max_i \left(\frac{\rho_i - \widetilde{x}_{i+}^\infty}{q_i} \right)$$

Moreover, solutions $(x_{ij}^\infty)_{\forall i,j \in \mathcal{V}}$ read:

$$x_{ij}^\infty = w_{ij}^+ x_{i+}^\infty$$

It is clear that x_{i+}^∞ should be equal or higher that ρ_i for all $i \in \mathcal{T}$, otherwise there is not enough flow escaping from i to respect $x_{i\omega}^\infty = \rho_i$. The vector $\widetilde{x}_+^\infty$ is the first solution found by the Moore-Penrose pseudoinverse and chances are that this case will not respect $\widetilde{x}_{i+}^\infty \geq \rho_i, \forall i \in \mathcal{V}$. By raising γ, we increase all x_{i+}^∞ and setting γ to $\gamma_0 := \max_i \left((\rho_i - \widetilde{x}_{i+}^\infty)/q_i \right)$ results in having one target with $x_{i+}^\infty = \rho_i$ and the rest with $x_{i+}^\infty > \rho_i$ and is the first admissible solution.

The free parameter γ, appearing naturally with calculations, controls the probabilities of being absorbed by ω from targets and therefore the number of jumps in the network before absorption. We will call it the *persistence*

of the flow. Setting the persistence to its minimum value minimizes flow intensities in the network and allows a better visualization of flows. In what follows, we will set $\gamma = \gamma_0$.

Interestingly, we can see that if $\forall i, j \in \mathcal{V}:\ x_{ij}^{\infty}/x_{i+}^{\infty} = w_{ij}^{+} = (1-\alpha_i)w_{ij}$ then:

$$w_{ij} = \frac{x_{ij}^{\infty}}{x_{i+}^{\infty} - \alpha_i x_{i+}^{\infty}} = \frac{x_{ij}^{\infty}}{x_{i+}^{\infty} - w_{i\omega}^{+} x_{i+}^{\infty}} = \frac{x_{ij}^{\infty}}{x_{i+}^{\infty} - \rho_i} = \frac{x_{ij}^{\infty}}{x_{i\bullet}^{\infty}}$$

which means that if the flow follows the Markov chain W^{+} on $\mathcal{V}^{+}$, the restrained flow on $\mathcal{V}$ will follow the original Markov chain W.

An interesting case is when the original Markov chain W is reversible [11, 12], i.e. there is a $(n \times n)$ symmetric matrix $C = (c_{ij}),\ c_{ij} \geq 0$, such that:

$$w_{ij} = \frac{c_{ij}}{c_{i\bullet}}$$

By defining the *net flow* $y_{ij}^{\infty} := x_{ij}^{\infty} - x_{ji}^{\infty}$ on the graph, we have $\forall i, j \in \mathcal{V}$:

$$\begin{aligned} y_{ij}^{\infty} &= x_{ij}^{\infty} - x_{ji}^{\infty} = x_{i+}^{\infty} w_{ij}^{+} - x_{+i}^{\infty} w_{ji}^{+} \\ &= x_{i+}^{\infty}(1-\alpha_i)\tfrac{c_{ij}}{c_{i\bullet}} - x_{j+}^{\infty}(1-\alpha_j)\tfrac{c_{ji}}{c_{j\bullet}} \\ &= c_{ij}\left(\tfrac{x_{i\bullet}^{\infty}}{c_{i\bullet}} - \tfrac{x_{j\bullet}^{\infty}}{c_{i\bullet}}\right) \end{aligned}$$

which describes *Ohm's law* with oriented *current intensities* y_{ij}^{∞}, *electrical potentials* $v_i := x_{i\bullet}^{\infty}/c_{i\bullet}$ and *capacities* c_{ij}. By defining potentials this way, we can see that $v_\omega = 0$, thus ω represents the ground, in the electrical sense. Fixing γ can be seen as defining the potential difference between the lowest potential node of the graph and ground. When $\gamma = \gamma_0$, the lowest potential node has a potential of 0, i.e. it is equivalent to the ground node. Raising γ will increase the whole graph potential, though potential differences between nodes $i, j \in \mathcal{V}$ remain constant. To sum up, Y^{∞} is similar to current intensities in an electrical circuit with resistances equal to $1/c_{ij}$ and with current inputs and outputs on nodes $i \in \mathcal{V}$ equal to $f_i - \rho_i$. Furthermore, it is interesting to see that minimizing an entropy functional is similar to minimizing a *quadratic energy functional*, $\sum_{i,j\in\mathcal{V}} (y_{ij}^{\infty})^2/c_{ij}$ [11, 12].

Note that the matrix W^{+} has to be used in the entropy functional for any $T > 0$, in order to make the general solution $X(T)$ converge to the desired limit when $T \to \infty$, i.e. $\lim_{T\to\infty} X(T) = X^{\infty}$.

2.5. General solution

By introducing Lagrange multipliers λ_i and μ_i to $F^+(X) = U(X||R) + TG^+(X||W^+)$, associated respectively to (11) and (12), the following optimality condition can be obtained by setting the derivative to zero:

$$T \ln \frac{x_{ij}}{x_{i+} w^+_{ij}} + r_{ij} = \lambda_j - \lambda_i - \mu_i \delta_{j\omega} \tag{15}$$

where $\delta_{j\omega}$ is the *Kronecker delta*. It gives $\forall i \in \mathcal{V}, \forall j \in \mathcal{V}^+$:

$$x_{ij} = x_{i+} w^+_{ij} \exp(-\beta r_{ij}) \exp(-\beta \lambda_i) \exp(\beta \lambda_j) \exp(-\beta \mu_i \delta_{j\omega}) \tag{16}$$

where $\beta = \frac{1}{T}$ is the *inverse temperature*. Note that multipliers μ_i are irrelevant, as flows $x_{i\omega}$ are replaced by ρ_i (12) to find the solution.

Define $v_{ij} := w^+_{ij} \exp(-\beta r_{ij}), \forall i, j \in \mathcal{V}^+$, as well as the new matrix and vectors:

$$V := (v_{ij})_{i,j \in \mathcal{V}} \qquad a := (\exp(\beta \lambda_i))_{i \in \mathcal{V}} \qquad b := (x_{i+} \exp(-\beta \lambda_i))_{i \in \mathcal{V}}$$

Summing x_{ij} over all $j \in \mathcal{V}^+$ and using (12) and (16), gives (see appendix B for details):

$$b_i \left(a_i - \sum_{j \in \mathcal{V}} v_{ij} a_j \right) = \rho_i \quad \forall i \in \mathcal{V} \tag{17}$$

and replacing (13) and (16) in (11) yields (see appendix B for details):

$$a_i \left(b_i - \sum_{j \in \mathcal{V}} v_{ji} b_j \right) = f_i \quad \forall i \in \mathcal{V} \tag{18}$$

Hence:

$$\mathrm{Diag}(b)(I - V)a = \rho$$
$$\mathrm{Diag}(a)(I - V')b = f$$

where $\mathrm{Diag}(x)$ is the diagonal matrix with diagonal equal to x. With the matrix $M := (I - V)^{-1}$, we finally have:

$$\boxed{a = M \mathrm{Diag}(b)^{-1} \rho} \tag{19}$$

$$\boxed{b = M' \mathrm{Diag}(a)^{-1} f} \tag{20}$$

which can be solved iteratively starting with any non-zero vector for a or b. The solution $X = (x_{ij})_{i,j\in\mathcal{V}}$ reads:

$$\boxed{x_{ij} = v_{ij} b_i a_j} \tag{21}$$

Notice that a (and, by extension, b) is defined up to a multiplier constant, as multipliers λ_i are defined up to an additive constant, but the solution X is unique due to the convexity of the functional and the domain.

2.6. Probabilistic interpretation

The flow X resulting from this algorithm has a probabilistic interpretation. We will show here that this flow can actually be viewed as the mean number of passages on each edge by a particular Markov process $S_t(T)$. This interpretation will help us in the next section to find a *source-target coupling* corresponding to the flow. For a fixed $T > 0$, let the Markov chain matrix $\widehat{W}(T) = \widehat{W} := (\widehat{w}_{ij})$ on $\mathcal{V}^+$ be defined by:

$$\begin{aligned}
\widehat{w}_{ij} = &\quad w_{ij}^+ \exp(\beta(\lambda_j - \lambda_i - r_{ij})) = v_{ij}\tfrac{a_j}{a_i} && \forall i,j \in \mathcal{V} \\
\widehat{w}_{i\omega} = &\quad 1 - \widehat{w}_{i\bullet} && \forall i \in \mathcal{V} \\
\widehat{w}_{\omega i} = &\quad 0 && \forall i \in \mathcal{V} \\
\widehat{w}_{\omega\omega} = &\quad 1
\end{aligned} \tag{22}$$

As $v_{ij}a_j/a_i = x_{ij}/x_{i+}$, we have $\widehat{w}_{i\bullet} \leq 1$ and this chain is always well defined. Let $(S_t : t \in \mathbb{N})$ be the associated Markov process on $\mathcal{V}^+$ starting with $\mathbb{P}(S_0 = i) = f_i$. In appendix C, we prove that:

$$x_{ij} = \mathbb{E}(\text{number of passages through } (i,j) \text{ by } S_t \text{ before reaching } \omega) \tag{23}$$

$$x_{i+} = \mathbb{E}(\text{number of visits of } i \text{ by } S_t \text{ before reaching } \omega) \tag{24}$$

and from that we have:

$$x_{++} = \mathbb{E}(\text{number of jumps made in } \mathcal{G} \text{ by } S_t \text{ before reaching } \omega)$$

The fact that the flow follows the Markov chain defined by $\widehat{W}$ means that the minimization of $F^+(X) = U(X) + TG^+(X||W^+)$ can be viewed as the a minimization of a unique entropy term, $G^+(X||\widehat{W}(T))$, depending on $T > 0$.

2.7. Source-target coupling

With the help of the previous section, it is possible to find a *source-target coupling* defined as $P = (p_{ij}) :=$ "proportion of units created in i and delivered to j" [4, 5]. This coupling can be expressed in relation to the process S_t:

$$p_{ij} := \mathbb{P}(\text{the process } S_t \text{ starts in } i \text{ and is absorbed through } (j, \omega))$$

The absorbing Markov chain matrix $\widehat{W}$ reads:

$$\widehat{W} = \left(\begin{array}{c||c|c} & \forall \mathtt{i}, \mathtt{j} \in \mathcal{V} & \omega \\ \hline\hline \forall \mathtt{i}, \mathtt{j} \in \mathcal{V} & \widehat{Q} & \widehat{\alpha} \\ \hline \omega & 0 & 1 \end{array}\right)$$

where $\widehat{q}_{ij} := v_{ij}\frac{a_j}{a_i}$ and $\widehat{\alpha}_i := 1 - \widehat{w}_{i\bullet} = 1 - \sum_{j\in\mathcal{V}} v_{ij}\frac{a_j}{a_i}$. As seen in [19], the *fundamental matrix* $\widehat{M} := (I - \widehat{Q})^{-1}$ of the chain defined by $\widehat{W}$ contains the mean number of visits of j when starting from i. The probability to escape through (j, ω) when starting from i is then $\widehat{q}_{ij}\widehat{\alpha}_i$ and therefore:

$$\begin{aligned} p_{ij} = \mathbb{P}(S_0 = i)\widehat{q}_{ij}\widehat{\alpha}_i = f_i\widehat{m}_{ij}(1 - \widehat{w}_{i\bullet}) \\ P = \mathrm{Diag}(f)\mathrm{Diag}(\widehat{\alpha})\widehat{M} \end{aligned} \tag{25}$$

When $T \to 0$, the source-target coupling p_{ij} is an optimal assignment of the transportation problem. There are generally multiple optimal assignments to a transportation problem, and even multiple assignments corresponding to the same unique flow given by X. The solution given by (25) is a convex mixture of all optimal assignments.

2.8. Low temperature limit

While the algorithm can be computed with $\beta = 0$ which correspond to $T \to \infty$, and which will give X^∞, it is impossible to get the exact flow of transportation problem X^0. When $T \to 0$, $\widehat{w}_{ij} = w_{ij}^+ \exp(\beta(\lambda_j - \lambda_i - r_{ij}))$, $\forall i, j \in \mathcal{V}$, becomes:

$$\widehat{w}_{ij} = \begin{cases} 0 & \text{if } (\lambda_j - \lambda_i - r_{ij}) < 0 \\ \text{somewhere in } (0, 1] & \text{if } (\lambda_j - \lambda_i - r_{ij}) = 0 \\ \infty & \text{if } (\lambda_j - \lambda_i - r_{ij}) > 0 \end{cases} \tag{26}$$

The last case, $(\lambda_j - \lambda_i - r_{ij}) > 0$ is impossible as the Markov chain $\widehat{W}$ is always well defined. The case $(\lambda_j - \lambda_i - r_{ij}) = 0$ occurs if and only if $x_{ij}^0 > 0$

and thus will yield the optimal routes. This characterization is similar to the dual optimal routes problem [3], namely:

$$\begin{aligned} \underset{\lambda}{\text{maximize}} \quad & \sum_{i \in \mathcal{S}} f_i \lambda_i - \sum_{i \in \mathcal{T}} \rho_i \lambda_i \\ \text{subject to} \quad & \lambda_j - \lambda_i \leq r_{ij} \quad \forall i, j \in \mathcal{V} \end{aligned}$$

The solution lies on an extreme point of the simplex, i.e. $\lambda_j - \lambda_i = r_{ij}$ for some λ_i, λ_j , implying a passage through (i, j) by the optimal flow.

3. Case studies

To illustrate the behavior of this algorithm, we will run it on two different graphs. The first one is a 10×10 lattice, where the behavior of the algorithm is observed with different values of parameter β. The second case study is made on a real graph, the street network of a neighborhood of the city of Lausanne, Switzerland (source: T. Emmanouilidis [20]). In this setup, sources and targets are, respectively, the residences of pupils and their schools. By setting a high β, x_{ij} will predict the number of pupils crossing each street segment and their optimal assignment to schools. These case studies are computed with the software *R* with the help of the *igraph* [21] library.

3.1. Toy example: 10×10 *lattice*

On this 10×10 lattice, we set 40 sources with $f_i = 1/40$ and 5 targets with $\rho_i = 1/5$. Their placement is randomly chosen and shown in Fig.1. The matrix R is set to be one for every arc and the Markov chain matrix W is the simple random-walk matrix.

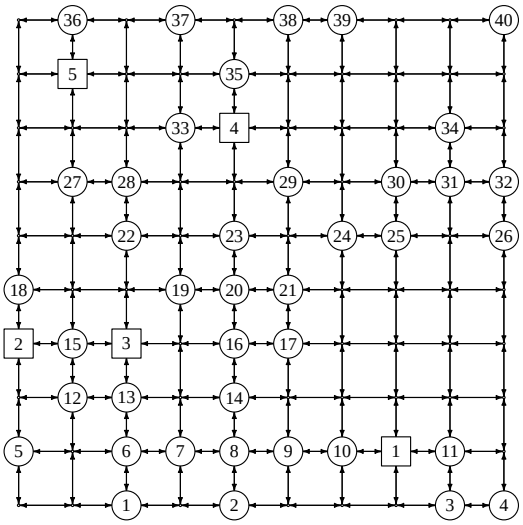

Figure 1: The 10×10 lattice. Sources are drawn with circles and targets with squares.

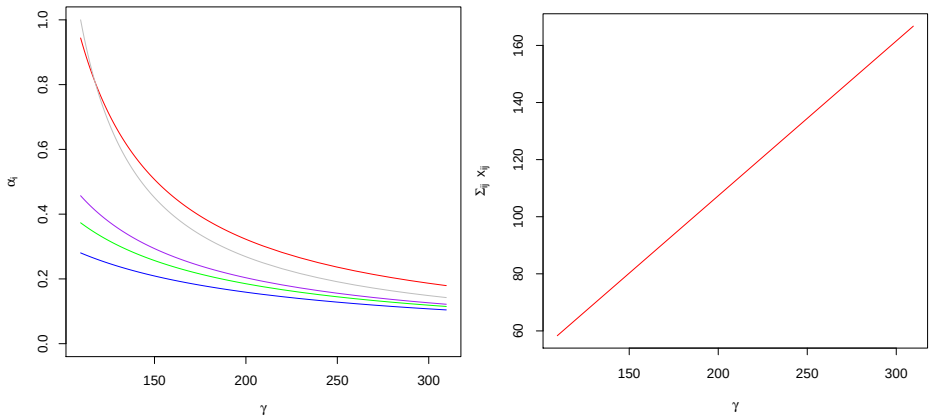

Figure 2: Left: Evolution of α_{t_1} (blue), α_{t_2} (red), α_{t_3} (purple), α_{t_4}(green) and α_{t_5}(gray) versus γ. Right: Evolution of x_{++} when $\beta \to 0$ vs γ.

The modified Markov chain matrix W^+ is computed with (14). Evolutions of α_i for targets versus persistence γ are displayed in the right side of

Fig.2. We notice, as expected, that α_i decreases when γ increases, with the limit $\alpha_i = 0$ when $\gamma \to \infty$. For the sake of visibility, the flow should be as small as possible when β is small. Later, we will set the persistence to its minimum value, $\gamma = \gamma_0 = 109.6$, i.e. where the highest escape probability, i.e α_{t_5}, is equal to one. Setting $\gamma = 109.6$ gives $\alpha_{t_1} = 0.28$, $\alpha_{t_2} = 0.94$, $\alpha_{t_3} = 0.46$ $\alpha_{t_4} = 0.37$ and $\alpha_{t_5} = 1$. It is interesting to see that these α_i only depend on the target position, as they all have the same $\rho_i = 1/5$. The right part of Fig.2 shows a linear increase of time spent in the network when γ is raised, as all α_i, the probabilities of being absorbed, decrease.

Resulting flows and "hard" coupling (i is assigned to the target j maximizing p_{ij}) are displayed in Fig.3. The hard coupling is used here for convenience in the drawing, and is only an approximation of the fuzzy coupling p_{ij}. The bottom right graph shows that x_{++} decreases when β increases, reflecting that the flow "wanders" less and is more geared toward targets. In the top left graph, flows are higher near sources than near targets and every edge is used both ways. We can also see that flows tend to be higher on the left part of the lattice, near t_1, and very low on the upper right side, near t_5. This reflects the low absorbing rate (high potential) of t_1 ($\alpha_{t_1} = 0.28$) and the high absorbing rate (low potential) of t_5 ($\alpha_{t_5} = 1$).

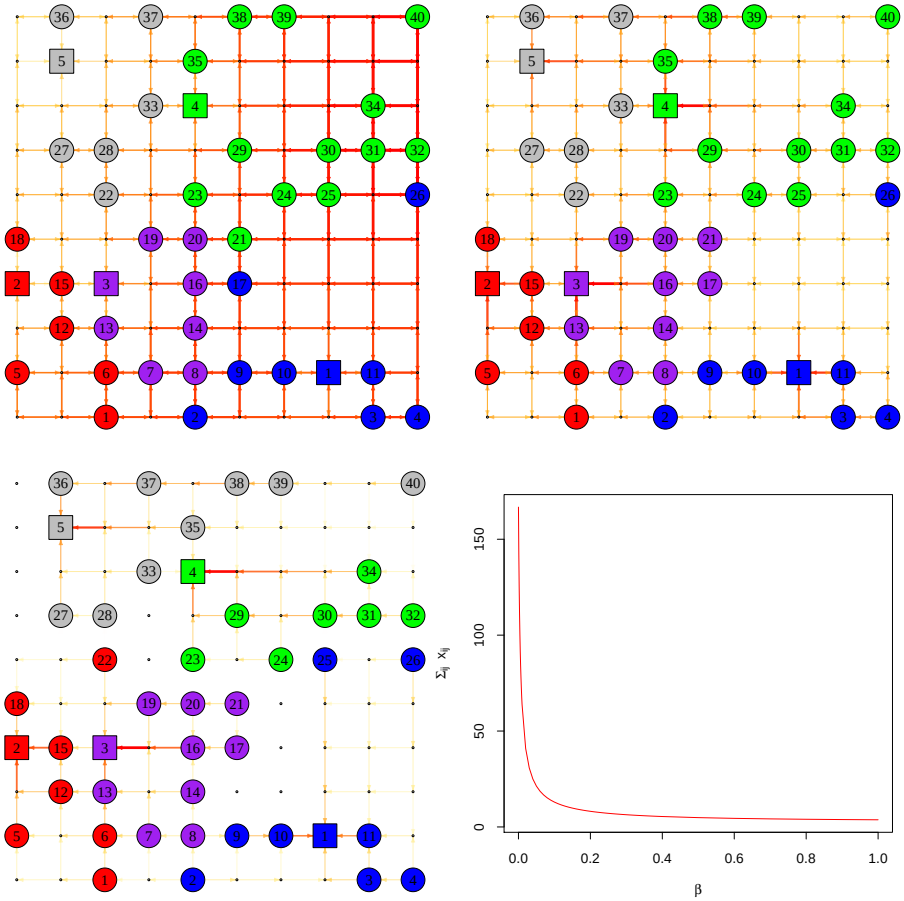

Figure 3: Resulting flow X with $\beta = 0$ (top left), $\beta = 0.1$ (top right) and $\beta = 10$ (bottom left). Source colors indicate to which target the largest part of the flow was send to. Bottom right shows the expected number of transitions, x_{++}, versus β.

In the bottom left graph of in Fig.3, flows are unidirectional on edges and show a combination of all possible constrained shortest-paths. The coupling is in this case optimal, in the transportation sense, and almost deterministic, i.e. almost the entirety of each source is sent to an unique target, therefore the hard coupling showed here gives a good idea of P. We can see that this hard coupling change slightly regarding β. However, most of the change is hidden, as the coupling P is more entropic for low β and not well represented when hardened.

3.2. Real network: A neighborhood of Lausanne

This case study is inspired by a real-life problem. In a neighborhood of the city of Lausanne, authorities are interested in assigning pupils to schools and knowing which road segments will be frequently used by pupils, in order to control traffic in the critical road segments. For confidentiality reasons, f_i and ρ_i have been simulated with distributions close to reality. Suppose the number of pupils in this neighborhood to be 2000 living in 500 different locations. Five schools are present in the neighborhood, with a total capacity of 2000. The map of the neighborhood is shown below.

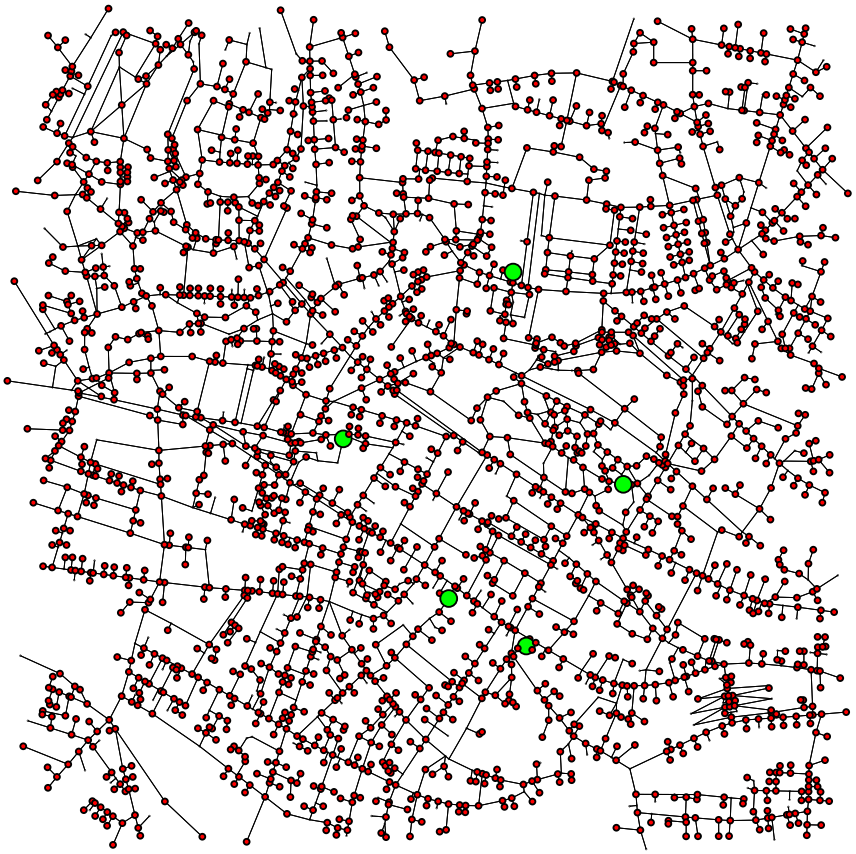

Figure 4: A neighborhood in Lausanne. Schools are shown in green and pupil residences in red.

The resistance matrix R is derived from road lengths in meters and W is taken to be the usual random-walk matrix. f and ρ vectors give, respectively, the proportion of pupils by node and school capacities. The result displayed in Fig.5 is obtained with $\beta = 0.2$, which is sufficiently large in comparison to intensities of resistance and displays an optimal solution (although hardened) of the transportation problem.

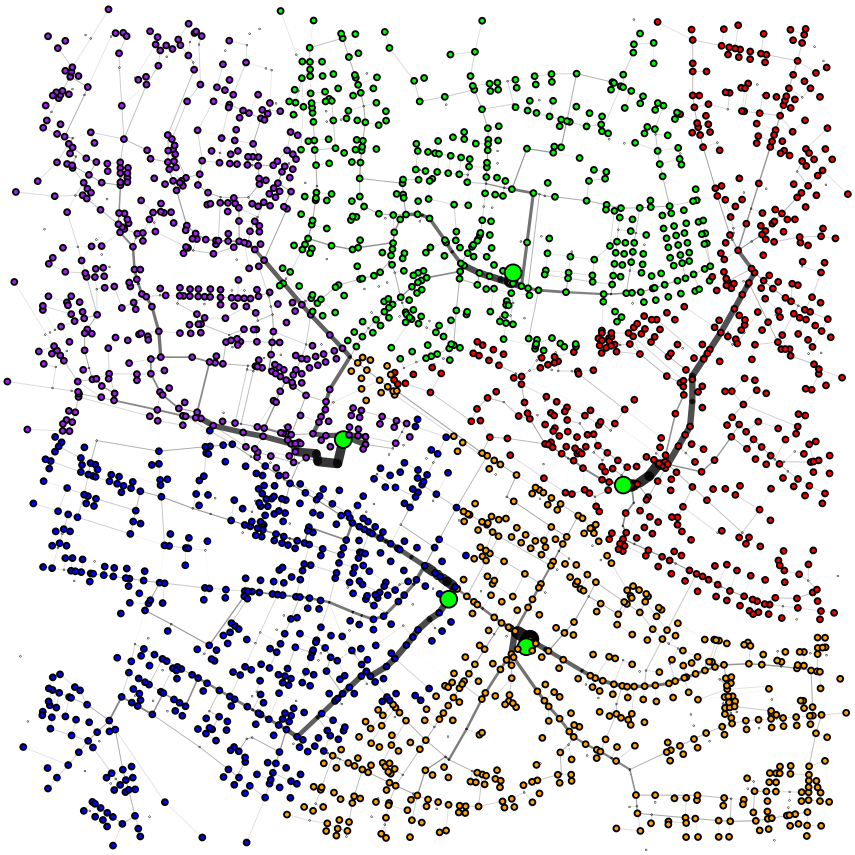

Figure 5: Resulting flow and hard coupling with $\beta = 0.2$. Larger edges denote a larger flow. To give an order of magnitude, the model predicts a flow of 473 pupils on the most frequently used route segment, located next to the yellow school.

In Fig.6, the solution is computed with $\beta = 0$ and with the minimum γ. Here, the flows are much higher on every edges because of the random movements of pupils, especially in the north-west corner. The hard coupling shown here does not represent accurately p_{ij} as it is less deterministic than for $\beta = 0.2$.

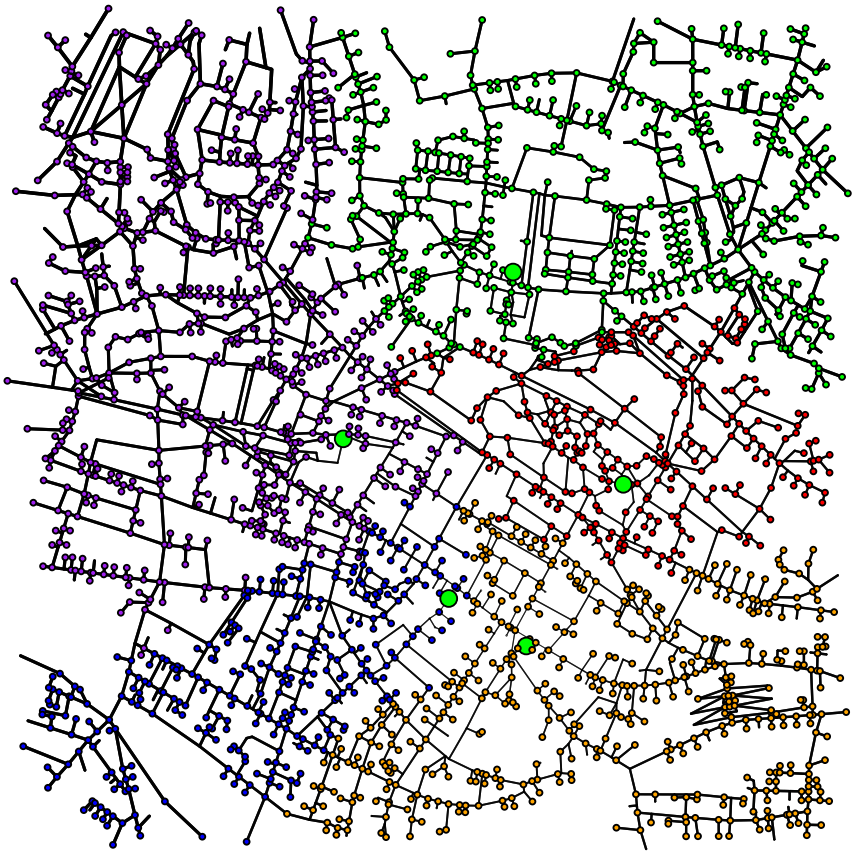

Figure 6: Resulting flow and crisp coupling with $\beta = 0$. Larger edges denote a larger flow. To give an order of magnitude, the model predicts a flow of 2900 pupils on the most frequently used route segment, located somewhere in the north-west.

4. Conclusion

This new algorithm, constructed to generalize the randomized shortest-path [13, 14, 15, 16, 17, 18] with multiple sources and targets, shows some unexpected connections to other problems. To begin with, when the temperature is low, the algorithm is able to display routes taken by goods in the context of the optimal transportation. While other algorithms need to find sources-targets allocations first to subsequently compute paths, this algorithm does it altogether. The source-target coupling found with this algorithm is not the only compatible solution with the resulting flow. As a matter of fact, flows are "anonymous", in the sense that, if two flows from two different sources meet on one node before diverging later, it is impossible to distinguish them when they split. Because of the regularization, the solution of this problem is unique while there are generally multiple solutions for source-target coupling in the transportation problem. In fact, the specific coupling found here is a convex mixture of all optimal couplings.
When the temperature is high, the algorithm gives a solution which follows the random-walk defined by W, or with another surprising analogy, the electrical current. In other words, this algorithm interpolates between a regime where the energy to minimize is linear and a regime where the energy is quadratic [22].
The flow is entirely computed with the help of v_{ij}, x_{i+} and a_i, as seen in (21), $x_{ij} = v_{ij}x_{i+}a_j/a_i$. Let us take a closer look at these quantities: v_{ij} are associated to edges, while x_{i+} and a_i to nodes. When $T \to \infty$, then $a_i = 1$, $x_{ij} = v_{ij}x_{i+}, \forall i, j \in \mathcal{V}$ and the formalism becomes identical to the electrical formalism, constituted of "potentials" x_{i+}, and "capacities", proportional to $v_{ij} = w_{ij}^+$. However, as T decreases, another non-trivial node quantity appears, a_i, weighting the flow by a factor a_j/a_i, and favoring routes of least cost. At the same time, "capacities" v_{ij} transform from w_{ij}^+ to $w_{ij}^+ \exp(-\beta r_{ij})$. When T is close to 0, some a_i turn out to be extremely large, and computers can not handle them anymore. This reflects the unfeasibility for this algorithm to compute solutions with $T = 0$.
The running time of the algorithm is hard to characterize, as the number of iterations needed to converge is unknown. However, analysis of the algorithm demonstrates that inverting $(I - V)$ has a complexity of $\mathcal{O}(n^3)$, while each algorithm iteration is $\mathcal{O}(n^2)$ (the multiplication between M and $\mathrm{Diag}(b)^{-1}$ or between M' and $\mathrm{Diag}(a)^{-1}$). If the number of iterations is low, the total complexity is $\mathcal{O}(n^3)$. But while the algorithm converges in a single step for $\beta = 0$, the number of iterations increases drastically when β is raised. This phenomenon could be the reflection of the difficulty to solve

an optimal transport problem compared to an electrical circuit problem. This algorithm is very flexible due to the large variety of parameters, R, W, f, ρ, β and γ, with little conditions on them. However, this flexibility comes with a price, as estimation of the parameters in order to mimic the behavior of an existing flow is not an easy task. Nevertheless, as shown with the second case study, one can obtain meaningful predictions if origins and destinations distributions are known.

Appendix A. Proof of Proposition 1

If W^+ defined by (14) yields $x^\infty_{ij}/x^\infty_{i+} = w^+_{ij}$, $\forall i, j \in \mathcal{V}$, it means that:

$$x^\infty_{ij} = w^+_{ij} x^\infty_{i+} \qquad \forall i, j \in \mathcal{V}$$

by summing over $j \in \mathcal{V}^+$ and using (12):

$$x^\infty_{i+} = \sum_{j \in \mathcal{V}} w^+_{ij} x^\infty_{i+} + \rho_i \qquad \forall i \in \mathcal{V}$$

which gives:

$$1 - \sum_{j \in \mathcal{V}} w^+_{ij} = \frac{\rho_i}{x^\infty_{i+}} \qquad \forall i \in \mathcal{V}$$

and with $\alpha_i = (1 - \sum_{j \in \mathcal{V}} w^+_{ij})$:

$$x^\infty_{i+} = \frac{\rho_i}{\alpha_i} \qquad \forall i \in \mathcal{V} \tag{A.1}$$

By developing x^∞_{+i} in (11) we have:

$$x^\infty_{i+} - \sum_{j \in \mathcal{V}} w^+_{ji} x^\infty_{j+} - x^\infty_{\omega i} = f_i \qquad \forall i \in \mathcal{V}$$

with $x^\infty_{\omega i} = 0$, $w^+_{ji} = (1 - \alpha_i) w_{ji}$ and (A.1):

$$x^\infty_{i+} - \sum_{j \in \mathcal{V}} w_{ji} x^\infty_{j+} + \sum_{j \in \mathcal{V}} w_{ji} \rho_j = f_i \qquad \forall i \in \mathcal{V}$$

by defining the vector $x^\infty_+ = (x^\infty_{i+})_{i \in \mathcal{V}}$, this last equation reads:

$$(I - W') x^\infty_+ = f - W' \rho \tag{A.2}$$

As W defines an irreducible Markov chain, the matrix $Q := (I - W')$ has a rank of $(n-1)$ [19], and is not invertible. This is why we need Q^*, the Moore-Penrose pseudoinverse of Q. If we write the singular decomposition of $Q = U \Sigma V'$, the Moore-Penrose pseudoinverse is $Q^* = V \Sigma^* U'$, where Σ^* contains $1/\sigma_i$ on its diagonal for all non-zero singular values σ_i of Q, and zero otherwise. One of the solutions for x^∞_+ is then given by:

$$\widetilde{x}^\infty_+ = Q^* (f - W' \rho) \tag{A.3}$$

and all the solutions by [23]:

$$x_+^\infty = \widetilde{x}_+^\infty + (I - Q^*Q)\zeta$$

with ζ an arbitrary vector. It is easy to verify that the rank of $(I - Q^*Q)$ is one, therefore with $q = (q_i) := ([I - Q^*Q]_{i1})$, the first column of $(I - Q^*Q)$, this last equation can be written:

$$x_+^\infty = \widetilde{x}_+^\infty + \gamma q \tag{A.4}$$

where $\gamma \in \mathbb{R}$. It gives for α_i:

$$\alpha_i = \frac{\rho_i}{\widetilde{x}_{i+}^\infty + \gamma q_i} \tag{A.5}$$

Note that not all solutions of $x_+^\infty = \widetilde{x}_+^\infty + \gamma q$ are acceptable for our problem, since we need to have $x_{i+}^\infty \geq \rho_i \; \forall i \in \mathcal{V}$. Thus $\widetilde{x}_+^\infty + \gamma q - \rho_i \geq 0 \; \forall i \in \mathcal{V}$, giving:

$$\gamma \geq \gamma_0 := \max_i \left(\frac{\rho_i - \widetilde{x}_{i+}^\infty}{q_i} \right) \tag{A.6}$$

This free parameter γ controls the probability of being absorbed by ω and therefore the number of jumps in the network before being absorbed in ω. The threshold exists because $\forall i \in \mathcal{V}$, x_{i+}^∞ must be larger than ρ_i.

Appendix B. Derivation for (17) and (18)

With $v_{ij} := w_{ij}^+ \exp(-\beta r_{ij})$, $a_i := \exp(\beta \lambda_i)$ and $b_i := x_{i+} \exp(-\beta \lambda_i)$ $\forall i, j \in \mathcal{V}$, from (16) we have $\forall i, j \in \mathcal{V}$:

$$x_{ij} = v_{ij} b_i a_j$$

If j is summed over $\mathcal{V}^+$, $\forall i \in \mathcal{V}$:

$$x_{i+} = \sum_{j \in \mathcal{V}} v_{ij} b_i a_j + x_{i\omega}$$

$$\Rightarrow x_{i+} = b_i \left(\sum_{j \in \mathcal{V}} v_{ij} a_i + \frac{x_{i\omega}}{b_i} \right) \tag{B.1}$$

which gives with $b_i = x_{i+}/a_i$:

$$1 = \frac{1}{a_i}\left(\sum_{j\in\mathcal{V}} v_{ij}a_j + \frac{x_{i\omega}}{b_i}\right) \tag{B.2}$$

$$\Rightarrow b_i\left(a_i - \sum_{j\in\mathcal{V}} v_{ij}a_j\right) = x_{i\omega} = \rho_i$$

giving (17).

From (B.2) we have:

$$a_i = \left(\sum_{j\in\mathcal{V}} v_{ij}a_j + \frac{x_{i\omega}}{b_i}\right)$$

which gives in (B.1):

$$x_{i+} = a_i b_i$$

Recall (11), $\forall i \in \mathcal{V}$:

$$x_{i+} - x_{+i} = f_i$$

$$\Rightarrow a_i b_i - \left(\sum_{j\in\mathcal{V}} v_{ji}b_j a_i + x_{\omega i}\right) = f_i$$

$$x_{\omega i} = 0 \text{ (13)} \Rightarrow a_i\left(b_i - \sum_{j\in\mathcal{V}} v_{ji}b_j\right) = f_i$$

giving (18).

Appendix C. Proof of (23) and (24)

Let $\widehat{W}$ defined by (22) and $(S_t : t \in \mathbb{N})$ the associated Markov process on $\mathcal{V}^+$ starting with $\mathbb{P}(S_0 = i) = f_i$. Define expectations $u_i := \mathbb{E}$(number of visits of i by S_t before reaching ω) and $u_{ij} := \mathbb{E}$(number of passages through (i, j) by S_t before reaching ω). Similarly to [11] with starting probabilities added, we have:

$$\begin{aligned} u_i &= \textstyle\sum_j u_j\widehat{w}_{ji} + f_i \quad && \forall i \in \mathcal{V} \\ u_{ij} &= u_i\widehat{w}_{ij} && \forall i, j \in \mathcal{V}^+ \\ u_\omega &= 0 \end{aligned} \tag{C.1}$$

By summing on all j in the second equation, we get that $u_i = u_{i+}$, which leads to:

$$\frac{u_{ij}}{u_{i+}} = \widehat{w}_{ij} \Rightarrow u_{ij} = v_{ij}\frac{u_{i+}a_j}{a_i} \tag{C.2}$$

and by developing the first equation with $\widehat{w}_{ij} = v_{ij}a_j/a_i$, we have:

$$u_i = a_i \sum_{j\in\mathcal{V}} \frac{u_j}{a_j} v_{ji} + f_i$$

$$\Rightarrow a_i \left(\frac{u_i}{a_i} - \sum_{j\in\mathcal{V}} \frac{u_j}{a_j} v_{ji} \right) = f_i$$

comparing this last result to (18), we get:

$$\frac{u_i}{a_i} = b_i \quad \forall i \in \mathcal{V} \tag{C.3}$$

For $u_{i\omega}$, $\forall i \in \mathcal{V}$ we have:

$$u_{i\omega} = u_i\widehat{w}_{i\omega} = u_i - \sum_{j\in\mathcal{V}} u_i v_{ij}\frac{a_j}{a_i} = \frac{u_i}{a_i}\left(a_i - \sum_{j\in\mathcal{V}} v_{ij}a_j \right)$$

and this last result with (C.3) and (17) gives:

$$u_{i\omega} = \rho_i \tag{C.4}$$

Finally, if we examine $u_{i+} - u_{+i}$ $\forall i \in \mathcal{V}$:

$$u_{i+} - u_{+i} = u_i - \sum_j u_j\widehat{w}_{ji} + f_i - f_i = f_i \tag{C.5}$$

If we compare the last equation in (C.1), as well as (C.2), (C.4) and (C.5) to, respectively, (13), (21), (12) and (11), and by uniqueness of the solution, we see that x_{ij} and u_{ij} are the same object, thus:

$x_{ij} = u_{ij} = \mathbb{E}$(number of passages through (i,j) by S_t before reaching ω)
$x_{i+} = u_i = \mathbb{E}$(number of visits of i by S_t before reaching ω)

Bibliography

[1] G. Monge, Mémoire sur la théorie des déblais et des remblais, De l'Imprimerie Royale, 1781.

[2] L. V. Kantorovich, On the translocation of masses, in: Dokl. Akad. Nauk SSSR, Vol. 37, 1942, pp. 199–201.

[3] R. K. Ahuja, T. L. Magnanti, J. B. Orlin, Network flows: theory, algorithms, and applications, Prentice hall, 1993.

[4] C. Villani, Topics in optimal transportation, no. 58, American Mathematical Soc., 2003.

[5] C. Villani, Optimal transport: old and new, Vol. 338, Springer, 2008.

[6] G. B. Dantzig, Linear programming and extensions, Princeton university press, 1998.

[7] H. W. Kuhn, The hungarian method for the assignment problem, Naval research logistics quarterly 2 (1-2) (1955) 83–97.

[8] D. P. Bertsekas, The auction algorithm: A distributed relaxation method for the assignment problem, Annals of operations research 14 (1) (1988) 105–123.

[9] S. Ferradans, N. Papadakis, J. Rabin, G. Peyré, J.-F. Aujol, Regularized discrete optimal transport, Springer Berlin Heidelberg, 2013.

[10] M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport, in: Advances in Neural Information Processing Systems, 2013, pp. 2292–2300.

[11] P. G. Doyle, J. L. Snell, Random walks and electric networks, AMC 10 (1984) 12.

[12] D. Aldous, J. A. Fill, Reversible markov chains and random walks on graphs, unfinished monograph, recompiled 2014, available at `http://www.stat.berkeley.edu/$\sim$aldous/RWG/book.html` (2002).

[13] F. Bavaud, G. Guex, Interpolating between random walks and shortest paths: a path functional approach, in: Social Informatics, Springer, 2012, pp. 68–81.

[14] G. Guex, F. Bavaud, Flow-based dissimilarities: Shortest path, commute time, max-flow and free energy, in: Data Science, Learning by Latent Structures, and Knowledge Discovery, Springer, 2015, pp. 101–111.

[15] L. Yen, M. Saerens, A. Mantrach, M. Shimbo, A family of dissimilarity measures between nodes generalizing both the shortest-path and the commute-time distances, in: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2008, pp. 785–793.

[16] M. Saerens, Y. Achbany, F. Fouss, L. Yen, Randomized shortest-path problems: Two related models, Neural Computation 21 (8) (2009) 2363–2404.

[17] P. Chebotarev, A class of graph-geodetic distances generalizing the shortest-path and the resistance distances, Discrete Applied Mathematics 159 (5) (2011) 295–302.

[18] I. Kivimäki, M. Shimbo, M. Saerens, Developments in the theory of randomized shortest paths with a comparison of graph node distances, Physica A: Statistical Mechanics and its Applications 393 (2014) 600–616.

[19] J. G. Kemeny, J. L. Snell, Finite markov chains, Van Nostrand, 1967.

[20] T. Emmanouilidis, Urban network analysis. centrality, sinuosity and shortcut detection, Revue Internationale de Géomatique 23 (3-4) (2013) 431–443. doi:10.3166/rig.23.431-443.
URL `http://dx.doi.org/10.3166/rig.23.431-443`

[21] G. Csardi, T. Nepusz, The igraph software package for complex network research, InterJournal Complex Systems (2006) 1695.
URL `http://igraph.org`

[22] M. Alamgir, U. V. Luxburg, Phase transition in the family of p-resistances, in: Advances in Neural Information Processing Systems, 2011, pp. 379–387.

[23] M. James, The generalised inverse, The Mathematical Gazette 62 (420) (1978) pp. 109–114.
URL `http://www.jstor.org/stable/3617665`

4.3 Performances des algorithmes

Dans cette section, nous allons analyser les performances des deux algorithmes présentés dans ce chapitre : l'algorithme des k-médoïdes contraints et l'algorithme permettant de trouver le flux de transport randomisé avec de multiples sources et cibles. Les deux peuvent être utilisés pour trouver un couplage, c'est-à-dire une solution au problème du transport optimal sur un graphe, mais le flux de transport randomisé permet en plus de trouver les flux de transport. Leurs performances vont être comparées à celles de l'algorithme du simplexe, utilisé dans la résolution de deux problèmes linéaire différents : le problème de couplage optimal et le problème de flux optimal. Dans la première partie de cette section, nous allons tester les différents algorithmes sur une grille dan laquelle le nombre de noeuds sera variable (en modifiant la largeur et la hauteur de la grille), mais la proportion de sources et de cibles restera fixe. La deuxième partie s'intéressera à la variation du nombre de sources et de cibles sur une grille de dimension fixe.

4.3.1 Conditions et paramètres des différents algorithmes

- Tous les résultats sont obtenus grâce à R, tournant sur un MacPro muni d'un processeur Intel Xeon E5 3.7Ghz avec 64 Go de RAM DDR3. L'algorithme du simplexe est obtenu grâce à la librairie "boot" de R.

- Chaque grille aura un nombre de noeuds n, un nombre de sources n_s et un nombre de cibles n_t. Les sources et les cibles sont placées aléatoirement et sont toutes de taille identique. Le graphe est non-orienté, toutes les résistances et conductances sont unitaires.

- L'algorithme du simplexe pour le couplage et l'algorithme des k-médoïdes contraints résolvent un problème comportant $n_s n_t$ variables.

- L'algorithme du simplexe pour les flux et l'algorithme du flux de transport randomisé résolvent un problème comportant n^2 variables.

- Les algorithmes régularisés, c-à-d l'algorithme des k-médoïdes contraints et du flux de transport randomisé, sont effectués avec une température inverse fixée à $\beta = 20$. On estime qu'ils ont convergé si leur fonctionnelle d'énergie libre reste identique, à la cinquième décimale près, après une itération.

4.3.2 Temps de calcul en fonction du nombre de noeuds

Couplage : simplexe, k-médoïdes contraints et flux de transport randomisé

Dans cette partie, les différents tests vont être effectués sur une grille dont la largeur l et la hauteur h vont varier entre 3 et 20, permettant d'obtenir des résultats avec un nombre de noeuds entre 9 et 400. Pour chaque taille, 50 essais sont effectués, avec 40% des noeuds étant des sources, 5% de noeuds des cibles (arrondis vers le haut).

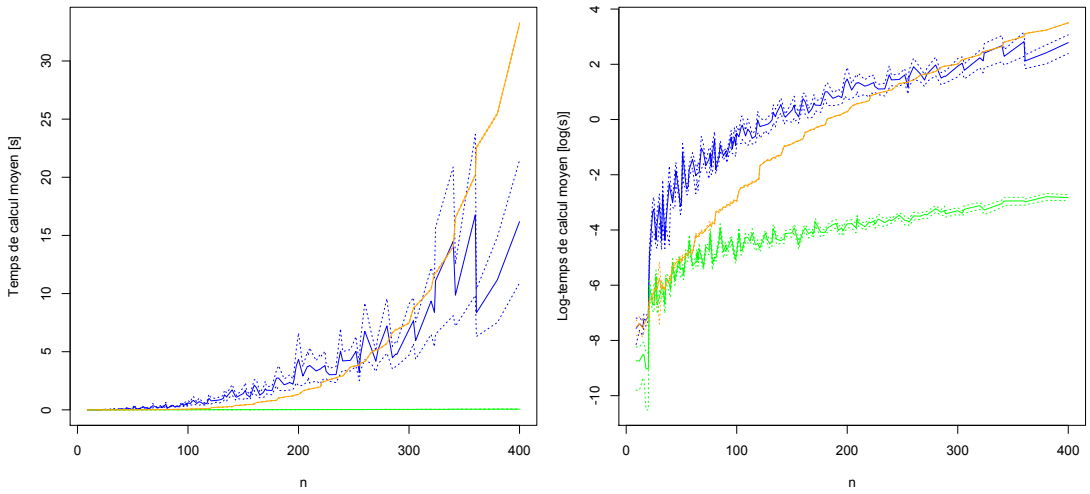

Figure 1 : Gauche : temps de calcul moyen et intervalle de confiance 95% (en secondes) pour obtenir le couplage vs n. Droite : log-temps de calcul moyen et intervalle de confiance 95% (en log-secondes). Jaune : algorithme du simplexe. Vert : k-médoïdes contraints. Bleu : flux de transport randomisé.

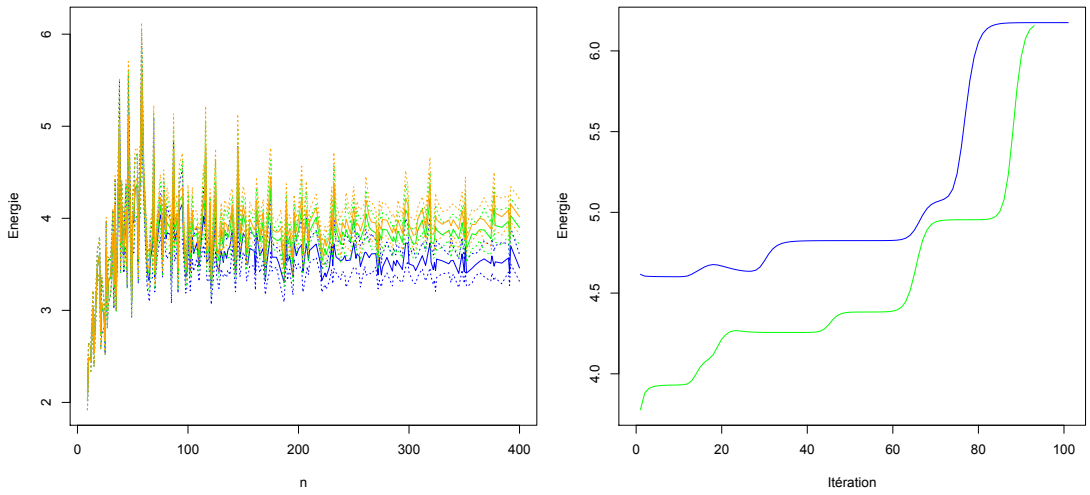

Figure 2 : Gauche : fonctionnelle d'énergie de la solution proposée par les différents algorithmes vs n. Droite : energie de la solution vs itération sur une grille 10×10. Jaune : algorithme du simplexe. Vert : k-médoïdes contraints. Bleu : flux de transport randomisé.

Dans la Figure 1, on voit que le nombre de noeuds agit de manière différente sur les trois algorithmes. Lorsque le graphe est très petit ($n \leq 20$), les trois algorithmes sont sont tous les trois très performants. Avec $20 \leq n \leq 50$, l'algorithme qui trouve le flux de transport randomisé est plus lent que les deux autres, qui affichent des performances relativement équivalentes. On pouvait s'y attendre, compte tenu du fait que les deux algorithmes conçus spécialement pour le couplage ne nécessitent que de trouver $n_s n_t$ variables, contrairement au flux de transport randomisé qui en possède n^2. En revanche, on voit que plus le nombre de noeuds devient élevé, plus l'algorithme du simplexe devient inefficient, montrant même des performances plus faibles que l'algorithme du flux de transport randomisé pour $n \geq 300$, bien que ce dernier soit sujet à de plus grandes variations dans son temps de calcul. Remarquons que l'algorithme du simplexe et celui du k-médoïdes contraints nécessitent le calcul des distances du plus court chemin. Ici, comme nos tests sont effectués sur une grille, ces distances correspondent aux distances de Manhattan et sont obtenues très rapidement. Il est important de garder à l'esprit qu'ils n'afficheront pas d'aussi bonnes performances lorsque ces distances sont plus difficiles à obtenir.

La partie gauche de le Figure 2 nous permet d'observer la différence entre les fonctionnelles d'énergie des différentes solutions. D'une manière surprenante, les algorithmes régularisés ont une fonctionnelle d'énergie plus basse que l'algorithme du simplexe. La raison derrière ce résultat est que les algorithmes régularisés travaillent dans l'espace dual, c'est-à-dire que la fonctionnelle augmente avec le nombre d'itérations (comme le montre la partie droite de la Figure 2) et les contraintes ne sont respectées qu'au point de convergence. En forçant l'arrêt des algorithmes après avoir obtenu une énergie libre ne variant plus selon un certain degré de précision, on observe de légères erreurs dans les contraintes. Cependant, avec les conditions de convergence que nous avons posées sur une grille 10×10, ces dernières sont minimes : l'erreur relative moyenne des distributions marginales du couplage est de 0.5% pour les k-médoïdes contraints et $< 10^{-10}$% pour le flux de transport randomisé.

Flux : simplexe et flux de transport randomisé

L'algorithme du simplexe étant très peu performant lorsqu'il est utilisé pour trouver les flux, on effectuera moins de tests et sur des grilles plus petites. La largeur l de la grille va varier entre 7 et 15 et la hauteur h entre l et 15, permettant d'obtenir des résultats avec un nombre de noeuds entre 49 et 225. Pour chaque taille, 50 essais sont effectués, avec 40% des noeuds étant des sources, 5% de noeuds des cibles (arrondis vers le haut).

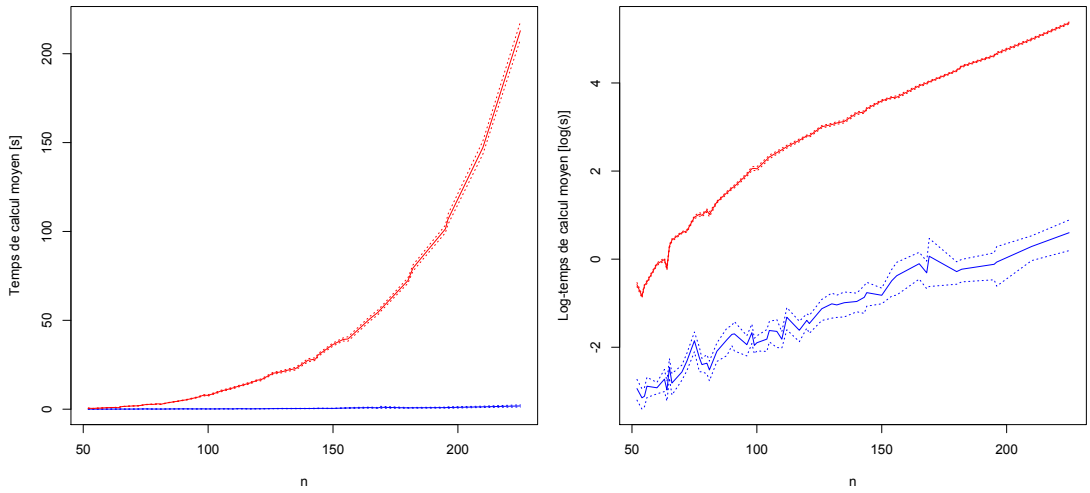

Figure 3 : Gauche : temps de calcul moyen et intervalle de confiance 95% (en secondes) pour obtenir le flux en fonction de n. Droite : log-temps de calcul moyen et intervalle de confiance 95% (en log-secondes). Rouge : algorithme du simplexe. Bleu : flux de transport randomisé.

Comme attendu, la Figure 3 nous montre que l'algorithme utilisé pour obtenir les flux de transport randomisé est bien plus performant que celui du simplexe. Les deux algorithmes s'effectuent sur un système avec n^2 variables, et on observe une croissance exponentielle dans les deux cas.

4.3.3 Temps de calcul en fonction du nombre de sources et de cibles

Ici, la taille de la grille va rester fixe et on posera $l = h = 10$. On aura donc $n = 100$ noeuds dans le graphe. En revanche, nous allons faire varier le nombre de sources n_s et le nombre de cibles n_t entre 1 et 50. Pour chaque nombre de sources et de cibles, on effectuera 50 tirages. Nous observerons les temps de calcul de tous les algorithmes sur les même graphiques.

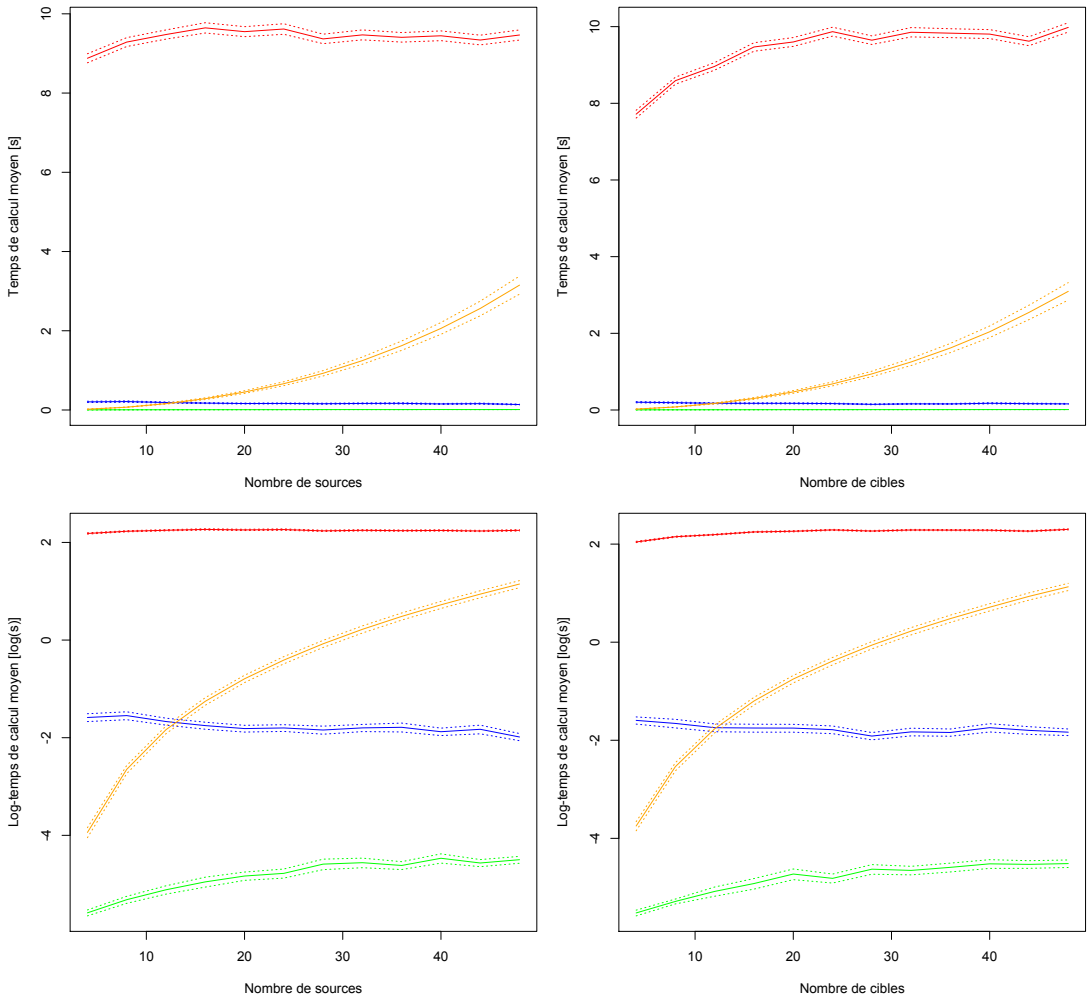

Figure 4 : Haut : temps de calcul moyen et intervalle de confiance 95% (en secondes) pour obtenir le flux en fonction du nombre de sources (gauche) et du nombre de cibles (droite). Bas : idem avec le log-temps de calcul (en log-secondes). Jaune : algorithme du simplexe pour le couplage. Vert : k-médoïdes contraints. Rouge : algorithme du simplexe pour le flux. Bleu : flux de transport randomisé.

Dans la Figure 4, on constate que l'augmentation des sources ou des cibles a le même effet sur le temps de calcul, et l'algorithme le plus affecté semble être celui du simplexe pour le couplage. Ce résultat était attendu, car le système linéaire que doit résoudre cet algorithme possède $n_s n_t$ variables. Dans une moindre mesure, on observe le même effet sur l'algorithme des k-médoïdes contraints. En revanche, l'algorithme du simplexe pour le flux et l'algorithme du flux de transport randomisé ne sont pas du tout affectés par ces changements, le nombre de variables recherchées restant n^2 dans tous les cas.

Ces différents tests nous permettent d'affirmer que dans presque toutes les situations, les algorithmes régularisés sont généralement bien plus efficaces que l'algorithme du simplexe. Il reste à savoir si un utilisateur est intéressé à n'obtenir que le couplage, auquel cas il aura recours à l'algorithme des k-médoïdes contraints, ou s'il désire obtenir un couplage et les flux correspondants, qui lui sera alors fourni plus efficacement grâce l'algorithme du flux de transport randomisé.

5 Les modèles de réseaux spatiaux

Ce chapitre se distingue considérablement des précédents, car il ne présente pas un formalisme de flux dans un graphe, mais s'intéresse aux *modèles de réseaux spatiaux*, c'est-à-dire à la création de graphes spatiaux non-orientés présentant des caractéristiques similaires à celles des réseaux spatiaux existants.

5.1 Historique

Les premiers modèles de réseaux sont apparus en 1959, avec la publication de deux modèles, très proches, permettant d'obtenir des *graphes aléatoires* : le *modèle d'Erdős-Rényi* (?, ?) et le *modèle de Gilbert* (?, ?). Ces modèles permettent de créer des graphes simples, non-orientés et non-pondérés, à partir d'un nombre donné de noeuds. Dans le premier de ces modèles, on fixe en outre le nombre d'arêtes, puis on effectue un tirage sur tous les graphes possédant le nombre d'arêtes et de noeuds donnés (avec une probabilité égale pour tous ces graphes). Dans le second, on pose une probabilité de création des arêtes, puis le graphe est obtenu en effectuant des tirages indépendants entre toutes les paires de noeuds. Ces modèles n'avaient pas comme vocation de modéliser des réseaux existants, mais ont été construits dans le but d'obtenir des résultats sur l'existence de certaines propriétés relatives aux graphes. Des résultats théoriques sur la distribution des degrés, la probabilité de percolation, le nombre de composantes connexes, le nombre de cycles, ont ainsi pu être obtenus à partir de ces modèles, en particulier certaines bornes lorsque le nombre de noeuds tend vers l'infini.

Ces modèles, et leurs variantes, sont longtemps restés les seuls à être proposés. Cependant, l'étude des réseaux, en se développant fortement au cours de la seconde moitié du XX^{e} siècle, a rapidement montré que la structure des réseaux réels différait fortement de celles obtenues par ces premiers modèles. A titre d'exemple, l'expérience empirique menée par le sociologue Stanley Milgram en 1967 (?, ?), a montré que les réseaux sociaux possèdent la caractéristique de *petit monde* ("small-world" en anglais), c'est-à-dire que chaque individu est relié à n'importe quel autre par une courte chaîne de liens de connaissances, ce qui n'était pas le cas dans les modèles d'Erdős-Rényi et de Gilbert. La nécessité de trouver de nouveaux modèles plus en adéquation avec la réalité s'imposa. En 1998, Watts et Strogatz (?, ?) proposèrent un nouveau modèle de réseau, présentant cette fameuse caractéristique de petit monde, suivi de près, en 1999, par Barabasi et Albert (?, ?) qui créèrent un modèle affichant une autre caractéristique présente dans de nombreux réseaux réels : une distribution de degrés suivant une loi de puissance (on appelle ces réseaux des *réseaux sans échelle*). Ces modèles contribuèrent à relancer l'engouement de la recherche pour les modèles de réseaux et, dans les années qui suivirent, ceux-ci se multiplièrent rapidement, chacun modélisant à sa manière certaines caractéristiques présentes dans différents types de réseaux.

5.2 Les modèles spatiaux

Mis à part quelques exceptions (?, ?), la modélisation des réseaux spatiaux commença tardivement, vers le début du XXI^{e} siècle. Le plus souvent, ces modèles furent des variantes de modèles existants, en intégrant la spatialité de différentes manières (en général, en pénalisant la probabilité de créer une arête entre deux noeuds éloignés). Le début de ce chapitre sera

justement consacré à la revue de la littérature sur ces modèles spatiaux, bien qu'elle reprenne en grande partie le travail de synthèse effectué par Marc Barthélemy en 2011 (?, ?). Un lecteur désireux d'en savoir plus pourra d'ailleurs s'y référer. Une emphase particulière sera mise sur un type de modèle, les *graphes optimaux*, c-à-d les graphes minimisant une fonctionnelle donnée. Cette manière de créer un réseau est intéressante, non seulement parce qu'elle reprend la minimisation d'une fonctionnelle, déjà abordée dans les chapitres précédents, mais également parce qu'elle permet l'*exploration de l'espace des graphes* pour un ensemble donné de noeuds dans l'espace. Cette exploration sera détaillée dans l'article clôturant le chapitre.

Bien que ce chapitre puisse surprendre le lecteur par sa différence marquée avec le reste de la thèse, il a en réalité une utilité particulière. En effet, lors de l'élaboration des autres chapitres, il a été très pratique de pouvoir créer, à tout moment, des graphes spatiaux possédant des caractéristiques similaires à celles des réseaux réels, et nous fournissons ici la possibilité au lecteur de reproduire ces résultats.

5.2.1 Les catégories de modèles

Les modèles de réseaux spatiaux que nous allons passer en revue ici sont construits à partir de n noeuds plongés dans un espace muni d'une métrique d (parfois donnée par une matrice $D = (d_{ij})$, de type $(n \times n)$). Dans certains modèles, il n'existe aucune arête entre les noeuds, dans d'autres, un graphe existe déjà entre les noeuds donnés. Les différentes méthodes que nous allons aborder ici, qui pourront être déterministes ou aléatoires, permettent de construire ou de détruire des arêtes entre ces noeuds suivant certaines règles. Parfois, les méthodes utilisées ajouteront également des noeuds supplémentaires aux graphes. La nomenclature des modèles de réseaux proposée Marc Barthélemy (?, ?) est la suivante.

Les graphes géométriques

La première catégorie de modèles sont les *graphes géométriques*. Ces modèles suivent des règles simples utilisant la géométrie présente entre les noeuds. Le plus élémentaire de ces modèles est certainement celui du *graphe géométrique simple* (?, ?), qui illustre parfaitement la logique de cette catégorie de modèles. Dans ce dernier, on pose un seuil r et on relie tous les noeuds i et j tel que $d_{ij} \leq r$. Cette modélisation très simple permet, entre autres, d'obtenir des résultats théoriques sur le seuil de *percolation* du graphe résultant, c'est-à-dire sur une valeur $\tilde{r}$ tel que $\forall r \geq \tilde{r}$, le graphe résultant sera connexe. Ces résultats trouvent des applications concrètes pour les réseaux de télécommunications "ad-hoc", qui utilisent chaque usager comme un relai pour le signal. A l'image de ce modèle, cette catégorie comprend tous les modèles issus de lois se rapportant à la géométrie existant entre les noeuds.

Les généralisations du modèle d'Erdős-Rényi

Cette catégorie comporte tous les modèles qui suivent une logique de création similaire à celle des graphes aléatoires d'Erdős-Rényi, en y intégrant une composante spatiale. On peut donner comme exemple le modèle de Waxman (?, ?) qui altère la probabilité d'apparition des différentes arêtes en fonction de la distance séparant les différents noeuds, c-à-d $\mathbb{P}(\text{"création d'une arête entre } i \text{ et } j\text{"}) = \max(1, \beta \exp(-d_{ij}/d_0))$, où β contrôle la densité des arêtes et d_0 est une distance de référence. En comparant les résultats obtenus par ces modèles avec ceux obtenus par le modèle d'Erdős-Rényi, on arrive à comprendre l'effet que possède l'espace sur la création des liens. Ces modèles furent notamment utilisés pour estimer les risques de congestion dans des réseaux de communications.

Les généralisations du modèle de Watts et Strogatz

Ces modèles, comme leur nom l'indique, généralisent le modèle de Watts et Strogatz, et permettent d'obtenir des graphes spatiaux possédant la caractéristique "petit-monde". Dans le modèle original, on commence à partir d'un graphe dont la structure est similaire à une grille régulière (bien que la position des noeuds dans l'espace n'existe pas, ce graphe est en bijection avec une grille de dimension k, c'est-à-dire que chaque noeud est connecté à $2k$ voisins). On passe ensuite en revue les différentes arêtes, et, avec une probabilité p, on redirige aléatoirement l'une des extrémités de ces arêtes sur un autre noeud, choisi uniformément parmi l'ensemble de ceux-ci. Une des variantes spatiale de ce modèle, proposée dans (?, ?), est la suivante : on prend comme condition initiale une véritable grille dans $\mathbb{R}^k$, et les noeuds sur lesquels les arêtes sont redirigées ne sont plus choisis uniformément, mais selon une probabilité décroissante par rapport à la distance, c-à-d $\mathbb{P}(\text{"\{i,j\} devient \{i,k\}"}) \sim d_{ik}^{-\alpha}$, où $\alpha \geq 0$ est un paramètre libre. Ces variantes du modèles de Watts et Strogatz furent étudiées pour leurs transitions de phases s'opérant sur la longueur espérée des plus courts chemins en fonction de α.

Les modèles de croissance spatiale

Ces modèles sont particuliers, dans le sens où la positon des noeuds n'est pas donnée à l'avance, mais l'on fait "croître" un graphe donné initialement en y ajoutant des noeuds et des arêtes. Ces modèles contiennent, par exemple, les généralisations spatiales du modèle d'attachement préférentiel de Barabasi et Albert, donnant une distribution des degrés suivant une loi de puissance. Mais on trouve également beaucoup de modèles permettant de modéliser des réseaux spatiaux de manière adéquate. Ainsi, il existe des modèles de croissance spatiale permettant d'obtenir des structures similaires à Internet (?, ?), aux réseaux de canalisations de gaz (?, ?) et aux réseaux routiers urbains (?, ?).

Les modèles de graphes optimaux

Cette dernière catégorie contient les réseaux dont nous allons nous intéresser plus en détail, et va être exposée dans la section suivante.

5.2.2 Les modèles de graphes optimaux

Ces modèles sont construits à partir d'un ensemble $\mathcal{V}$ de n noeuds donnés, possédant des distances entre eux, données par $D = (d_{ij})$. Le principe est de trouver des *graphes optimaux* $\mathcal{G}^{\text{opt}}$ sur cet ensemble de noeuds, c'est-à-dire les graphes $\mathcal{G}^{\text{opt}} = (\mathcal{V}, \mathcal{E}^{\text{opt}})$ qui minimisent une certaine fonctionnelle $F(\mathcal{G})$:

$$\mathcal{G}^{\text{opt}} \in \underset{\mathcal{G}=(\mathcal{V},\mathcal{E}) \mid \mathcal{E} \subset (\mathcal{V}\times\mathcal{V})}{\arg\min} F(\mathcal{G})$$

Dans certains cas, cette fonctionnelle est accompagnée de contraintes sur le graphe, afin de réduire le champ des graphes optimaux. Ces contraintes peuvent néanmoins être intégrées dans la fonctionnelle en utilisant le Lagrangien. Nous ne donnerons ici que des fonctionnelles sans contraintes.

Nous pourrions nous questionner quant à la validité de créer des modèles de réseaux à partir d'une minimisation d'une quantité globale. En effet, dans un grand nombre de réseaux réels, tous les acteurs participant à la construction du réseau agissent de leur côté, indépendamment les uns des autres, en essayant de maximiser l'utilité locale du graphe sans qu'il existe de planificateur global. Cependant, à l'image de la "main invisible" d'Adam Smith, ces forces locales génèrent parfois, avec le temps, des réseaux qui présenteront une caractéristique optimale au niveau global. Et même dans les cas où cela ne se produirait pas, le fait de pouvoir modéliser le graphe

optimal par rapport à certains critères peut s'avérer utile, dans le sens où le modèle nous donne un graphe de référence, permettant d'ajuster le réseau existant.
Nous allons passer en revue quelques exemples de modèles dignes d'intérêt obtenus à partir d'une minimisation de fonctionnelle.

L'arbre recouvrant de coût minimal

En posant $F(\mathcal{G})$ comme le coût du graphe, et en ajoutant une contrainte pour n'obtenir qu'une composante connexe, le graphe optimal résultant sera l'*arbre recouvrant de coût minimal*. Formellement, $F(\mathcal{G})$ se définit comme :

$$F(\mathcal{G}) := R(\mathcal{G}) + \lambda \sum_{i \in \mathcal{V}} E(i) = \sum_{\{i,j\} \in \mathcal{E}} l^d(\{i,j\}) + \lambda \sum_{i,j \in \mathcal{V}} d^{dsp}(i,j)$$

où, rappelons-le, $R(\mathcal{G})$ est le coût du graphe, et $\sum_{i \in \mathcal{V}} E(i)$ la somme des excentricités de ses noeuds (voir la section 3.1.1). Le paramètre $\lambda > 0$ doit être suffisamment petit, de manière à ce que les quantités finies dans la somme des excentricités n'aient pas d'impact sur la minimisation du coût. Ainsi, les graphes minimisant cette fonctionnelle ont un coût minimum, mais ne peuvent avoir des excentricités infinies, c'est-à-dire plusieurs composantes connexes.

L'arbre recouvrant de coût minimal est un excellent modèle de référence pour certains réseaux, permettant d'avoir une structure connexe de coût minimal. Cette façon de définir la fonctionnelle comme une balance entre deux quantités antagonistes, comme ici la capacité à connecter tous les noeuds qui s'oppose à son coût, est un principe fréquemment utilisé.

Les graphes modélisant les réseaux de communication

L'article de Markus Brede (?, ?) propose des graphes optimaux pouvant modéliser des réseaux de communication. Ces graphes minimisent la fonctionnelle suivante :

$$F(\mathcal{G}) := \sum_{\{i,j\} \in \mathcal{E}} l^d(\{i,j\}) + \lambda \sum_{i,j \in \mathcal{V}} \frac{2\, d^{dsp}(i,j)}{n(n-1)}$$

avec $\lambda > 0$ un paramètre libre. Lorsque $\lambda \to 0$, la fonctionnelle se réduit au coût du graphe, et le graphe la minimisant est le graphe vide. A l'inverse, lorsque $\lambda \to \infty$, la fonctionnelle devient la moyenne de la longueur des plus court chemins. On peut voir ce paramètre comme l'*investissement* mis dans le graphe afin de diminuer le plus court chemin moyen entre les noeuds. Pour des valeurs intermédiaires, ce graphe présente une structure analogue à celle que l'on trouve dans des réseaux de communication, dans lesquels le temps de communication moyen doit être réduit au maximum tout en minimisant le coût des connexions.

Les graphes interpolant entre des réseaux routiers et aériens

Dans (?, ?), les auteurs proposent un modèle de réseaux se basant sur la minimisation de la fonctionnelle suivante :

$$F(\mathcal{G}) := \sum_{\{i,j\} \in \mathcal{E}} l^d(\{i,j\}) - \lambda \sum_{i,j \in \mathcal{V} \,|\, i \neq j} \frac{1}{\alpha\, d^{sp}(i,j) + (1-\alpha) d^{dsp}(i,j)}$$

avec deux paramètres libres : $\lambda > 0$ et $\alpha \in [0,1]$. Le premier paramètre, $\lambda > 0$, permet d'interpoler entre la minimisation du coût du graphe et la somme d'une variante des proximités harmoniques de chaque noeud (vus à la section 3.1.1). Ainsi, le graphe minimisant cette fonctionnelle est vide pour $\lambda \to 0$, et au fur et à mesure que ce paramètre augmente, des arêtes permettant d'augmenter au plus fort les proximités entre les noeuds vont être créées. D'une manière

similaire à l'exemple précédent, ce paramètre peut être vu comme l'investissement mis dans le graphe afin d'augmenter la proximité entre les noeuds. Le deuxième paramètre, $\alpha \in [0, 1]$, agit sur le sens que nous donnons à la proximité. Lorsque $\alpha = 0$, cette proximité est calculée grâce à la distance du plus court chemin *valuée*, $d^{dsp}(i, j)$, c'est-à-dire que l'on cherche à réduire la distance effectivement parcourue sur le graphe. A l'opposé, lorsque $\alpha = 1$, la proximité ne prend en compte que la distance du plus court chemin *non-valuée*, $d^{sp}(i, j)$, c-à-d le nombre de sauts nécessaires pour aller d'un noeud à l'autre. Cette flexibilité dans la définition de la proximité permet la modélisation de différents types de transports. Prenons un exemple : contrairement aux dépenses énergétiques des voitures, qui dépendent principalement des distances parcourues sur les routes, les avions dépensent une grande quantité d'énergie lors de leurs décollages et atterrissages, et le nombre d'escales prend de l'importance par rapport à la longueur effective du vol. Les réseaux aériens prennent ainsi une structure contenant des noeuds à degré élevé, les fameux "hubs" aéroportuaires. Les graphes minimisant cette fonctionnelle sont donc similaires aux réseaux routier lorsque $\alpha = 0$ et similaire à des réseaux aériens lorsque α est élevé.

Ces fonctionnelles ne sont que quelques exemples qui peuvent être choisis afin de trouver un résultat désiré, car la liste est encore longue (?, ?). Remarquons que ces dernières sont choisies en fonction de l'objet étudié, en réduisant les principes de son fonctionnement à quelques quantités judicieusement choisies. On obtient ainsi des graphes qui, même s'ils ne modélisent pas exactement les réseaux étudiés, donnent un "idéal" dans l'univers simplifié ne contenant que les facteurs retenus.

5.2.3 Obtenir les graphes optimaux et l'exploration de l'espace

Lorsque nous nous plaçons dans le cadre posé dans la section 5.2.2, c'est-à-dire avec un ensemble $\mathcal{V}$ de n noeuds donnés, on peut définir l'*espace de tous les graphes possibles* sur ces noeuds, notés $\Gamma_{\mathcal{V}}$, de taille $2^{n(n-1)/2}$. Cet espace contient les graphes optimaux par rapport à la fonctionnelle et il est aisé de les trouver avec une méthode de type *recuit simulé* (en anglais : *simulated annealing*). Cependant, l'unicité des graphes optimaux n'est absolument pas certaine et quand bien même ce serait le cas, le fait de pouvoir trouver des minima locaux, proches de l'optimum, permet d'offrir une certaine flexibilité par rapport à une solution unique. Dans l'article qui suit, nous allons présenter un moyen d'explorer l'espace des graphes possibles sur les noeuds, $\Gamma_{\mathcal{V}}$, via une chaîne de Markov définie entre les éléments de cet espace, et de trouver les différentes alternatives possibles proches de l'optimalité.

5.2.4 Article 5 : "Spatial graphs cost and efficiency : exploring edges competition by MCMC"

Cet article est paru dans *Geographic Information Science* (pp. 97–108) de Springer, en 2014, faisant suite à une présentation donnée à la conférence du même nom se déroulant à Vienne en 2014.

Ce dernier reprend la fonctionnelle de (?, ?) (voir ci-dessus), avec $\alpha = 0$, et propose une manière d'explorer l'espace des graphes possibles $\Gamma_{\mathcal{V}}$ sur un ensemble de noeuds donnés $\mathcal{V}$. Cette exploration utilise le principe de l'échantillonnage de Gibbs et des chaînes de Markov Monte-Carlo (MCMC). Cette exploration de l'espace résulte en un historique des différents graphes obtenus, permettant de calculer la fréquences d'apparition des arêtes, les corrélations entre les paires d'arêtes, et donnant une représentation graphique de cette exploration.

Points-clés

- Definition de la fonctionnelle à minimiser et illustrations des graphes optimaux (p.107).
- Explication de la procédure d'exploration de l'espace (pp.108-109).
- Etude de l'exploration des graphes possibles à partir de 30 noeuds placés aléatoirement dans l'espace (pp.110-114).
- Etude de l'exploration des graphes possibles à partir des villes américaines de plus de 500'000 habitants (pp.111-116).

Spatial graphs cost and efficiency: exploring edges competition by MCMC

Guillaume Guex

Department of Geography,
University of Lausanne
guillaume.guex@unil.ch

Abstract. Recent models for spatial networks have been built by determining graphs minimizing some functional F composed by two antagonist quantities. Although these quantities might differ from a model to another, methods used to solve these problems generally make use of simulated annealing or operations research methods, limiting themselves to the study of a single minimum and ignoring other close-to-optimal alternatives. This contribution considers the arguably promising framework where the functional F is composed by a graph *cost* and a graph *efficiency*, and the space of all possible graphs on n spatially fixed nodes is explored by MCMC. Covariance between edges occupancy can be derived from this exploration, revealing the presence of cooperative and competition regimes, further enlightening the nature of the alternatives to the locally optimal solution.

Keywords: Spatial graph models, Efficiency, Cost, MCMC, Principal component analysis

1 Introduction

Spatial networks constitute a particular case in networks studies, where nodes and edges are embedded in a metric space. The study of these networks received a special attention in the recent years, as they model a large quantity of complex geographic systems, such as transportation networks (road, railroad and airlines networks), power grids networks and internet ([1, 2, 4–6, 8, 9, 11, 12, 16, 19, 21]). The particularity of these networks is that the underlying space directly controls the cost of edges, thus impacting their topology. Previous empirical studies have examined different spatial network structures and demonstrated that the effect of space greatly differs, depending on the nature of networks (reviewed extensively in [4]). Nevertheless, their designs typically attempt to maximize some utility function while minimizing some kind of cost function, making abstraction of other geographical or economical constraints encountered in real-world situations.

This article attempts to study a particular class of models of optimal networks defined as networks minimizing some functional F specified below. These models exhibit a great variety of interesting results, depending on the ingredients

entering the composition of F, and are aimed at modelling numerous different geographic systems of interest. Here, we will consider the case where $F = C - I \cdot E$, where C is the *cost* of the network (the sum of all edges length) and E the *efficiency* (the mean length of shortest-paths between all pairs of nodes), while the parameter I, the *investment*, acts as a balance between those quantities. This simple and intuitive model, already studied in [1, 2, 4, 11], gives results similar to our railroads, highways or power grid networks. Previous researches concentrated on finding a single graph minimizing F, discarding the study of the nature of the space of all possible graphs on n fixed nodes, controlled by F. By contrast, we attempt here to explore this space with a *Monte Carlo Markov Chain* (MCMC) algorithm ([3, 7, 10, 14, 17, 18]) or more precisely, a variant of simulated annealing model, implying heating as well as cooling schedules (see section 2.4). By examining the history of the algorithm, edge competition and synergies can be revealed, enabling the design of close-to-optimal graphs.
This article is divided in two parts. The first one sets the formalism and the mathematical tools needed to perform the algorithm and the second one examines a few case studies in more detail.

2 Formalism

2.1 Generalities and notations

A *graph* is a couple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ are the vertices (or nodes) set of size n and $\mathcal{E}$ the edges set of size m. A graph is said to be *spatial* when all vertices are embedded in a Euclidean space. A spatial graph is entirely defined by two matrices: X the matrix of vertex coordinates in space and the $n \times n$ symmetric adjacency matrix $A = (a_{ij})$, where $a_{ij} = 1$ if $\{i, j\} \in \mathcal{E}$, $a_{ij} = 0$ otherwise.

This article considers simple unoriented spatial graphs in $\mathbb{R}^2$ equipped with the Euclidean distance d^E. In this context, every edge $e = \{i, j\}$ possesses a *length* l corresponding to the Euclidean distance between nodes composing it, i.e. $l(\{i, j\}) = d^E(i, j)$. Edge lengths permit to define an alternative version of the well-known shortest-path distance, referred to, in the literature, as the *weighted shortest-paths distance* d^{wsp} (or route distance in [1, 2, 4, 11]):

$$d^{wsp}(i, j) = \min_{\xi \in \mathcal{P}(i,j)} \sum_{e \in \xi} l(e)$$

where $\mathcal{P}(i, j)$ is the set of all paths between i and j.

2.2 Functional minimization

Some other quantities can be defined on spatial graphs. Define the *cost* C of a graph $\mathcal{G}$ as the sum of all edge lengths:

$$C(\mathcal{G}) = \sum_{e \in \mathcal{E}} l(e)$$

Futhermore, define the *efficiency* E of graph $\mathcal{G}$ as the mean, along all pairs of vertices, of the inverse of the weighted shortest-path distance:

$$E(\mathcal{G}) = \frac{1}{n(n-1)} \sum_{i \neq j \in \mathcal{V}} \frac{1}{d^{wsp}(i,j)}$$

Obviously, for any set of vertices, the empty graph yields a null cost and efficiency, while the complete graph gives their maximum. From a concrete point of view, the efficiency represents the ability of the network to effectively transport agents from any node to another, while the cost is self speaking. Therefore, an optimal network planning may seek to maximize the efficiency while minimizing the cost, leading to the minimization of the function F defined by:

$$F(\mathcal{G}) = C(\mathcal{G}) - I \cdot E(\mathcal{G})$$

where the parameter $I \geq 0$ is the *investment*, acting as an arbiter between the conflicting objectives. When $I \to 0$ the graph minimizing F is the empty graph, while $I \to \infty$ generates the complete graph. For carefully chosen intermediate values, depending in turn on several parameters such as the real cost of the edges and the insistence on the efficiency of the network, the solutions are similar to some real spatial networks, like railroad, highways or power grid networks ([2, 4, 11]). Note that, unless $I \to \infty$, the resulting graph may not be connected (see e.g. left plot in Fig. 1). If we replace the weighted shortest-path distance by the standard shortest-path distance in the formula for efficiency, optimal graphs will possess a structure of "Hub-and-spoke", similar to an airline network ([4, 11]).

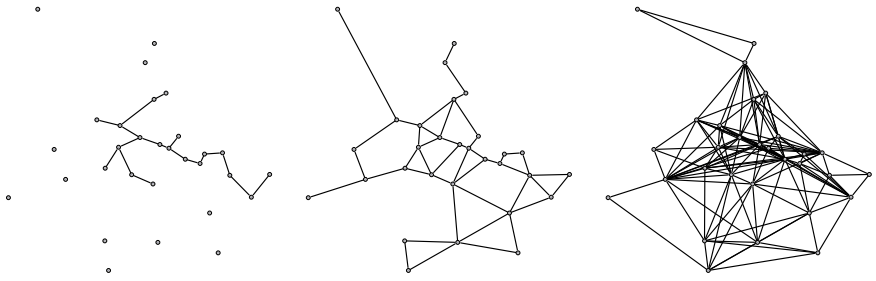

Fig. 1. Local minima for different investment values on 30 fixed points in $\mathbb{R}^2$ with abscissas and ordinates generated as $\mathcal{N}(0, 1)$. On the left $I = 1$, in the middle $I = 100$ and on the right $I = 10^6$.

2.3 MCMC exploration of the space of all graphs

For a fixed set $\mathcal{V}$ of n vertices in space, the *space of all graphs* on these vertices is noted $\Gamma_{\mathcal{V}}$. This space is similar to the atomic spins space in the Ising model, where every possible edges can be in two states { presence, absence } ([10, 15]). Thus the size of $\Gamma_{\mathcal{V}}$ is $2^{n(n-1)/2}$. Let $W = (w_{kl}) = (w_{\mathcal{G}_k\mathcal{G}_l})$ be the transition matrix of a *Markov chain* on $\Gamma_{\mathcal{V}}$ defined by:

$$w_{kl} = w_{lk} = \begin{cases} \frac{2}{n(n-1)} & \text{if } \mathcal{G}_k \text{ and } \mathcal{G}_l \text{ differ exactly by one edge} \\ 0 & \text{otherwise} \end{cases}$$

With this simple transition matrix, the Markov chain will jump from the graph $\mathcal{G}_k$ to any graph $\mathcal{G}_l$ having exactly one more or one less edge with equal probability. It is obvious that this Markov chain can reach any nodes from any starting point, and therefore the chain is *irreducible*. The MCMC *Metropolis-Hasting* (MH) algorithm ([3, 14, 17]) , is designed to create a new Markov chain having a desirable stationary distribution p_k on the states from any irreducible Markov chain:

1. From the state k, generate a new state l with probability w_{kl}
2. Jump to l with probability $\tilde{w}_{kl}$ defined by:

$$\tilde{w}_{kl} = \min\left(1, \frac{p_l w_{lk}}{p_k w_{kl}}\right)$$

   otherwise stay in k
3. Iterate

Since $w_{kl} = w_{lk}$, one has that $\tilde{w}_{kl} = \min(1, \frac{p_l}{p_k})$.

For any initial configuration, the algorithm will converge to the invariant distribution p_k, itself determined so as to favor near-optimal graphs, as in the *Gibbs sampling* of p_k ([3, 7]):

$$p_k = \frac{1}{Z} \exp(-\beta F(\mathcal{G}_k))$$

where $Z = \sum_k \exp(-\beta F(\mathcal{G}_k))$ is the standardization constant and $\beta = \frac{1}{T}$ is the *inverse temperature* parameter ($T \geq 0$ is the *temperature*). In fact, the value of β controls the randomness of the MH algorithm jumps, as seen by:

$$\tilde{w}_{kl} = \min\left(1, \frac{\exp(-\beta F(\mathcal{G}_l))}{\exp(-\beta F(\mathcal{G}_k))}\right) = \min\left(1, e^{\beta(F(\mathcal{G}_k) - F(\mathcal{G}_l))}\right)$$

If $\beta \to 0$, then $\tilde{w}_{kl} = 1$, i.e. the MH algorithm will jump to any candidate state l, while $\beta \to \infty$ implies that $\tilde{w}_{kl} = 1$ iff $F(\mathcal{G}_l) \leq F(\mathcal{G}_k)$ and $\tilde{w}_{kl} = 0$ otherwise.

2.4 Cooling schedule and exploration history

While a local minimum can easily be obtained with the MH algorithm and a simulated annealing cooling schedule (left plot in Fig. 2, as seen in [13, 20]), we are more interested here by the history of the exploration of space $\Gamma_{\mathcal{V}}$. Indeed, local minima are arguably often not really compatible with some real life constraints and we would be interested in finding alternative, but still efficient, ways to build the network. That is why we need our cooling schedule to be reheated periodically (right plot of Fig. 2) in order to avoid to be stuck in the same local minimum and to explore different parts of the space.

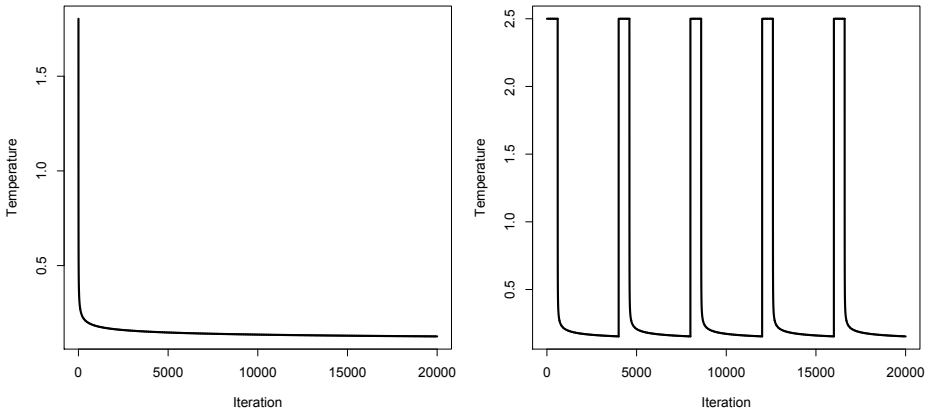

Fig. 2. Two types of cooling schedule: the first case represents a classical simulated annealing cooling schedule designed to find only one local minima: $T(t) = c/\log(1+t)$ with $c = 1.25$. In the second case, we periodically set a high temperature during 400 iterations followed by a similar cool-down, in the hope of finding another minimum.

Recording the graph history $\{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_t\}$ of a MH run by keeping track of the states of every edges modified at least once $\{e_1, e_2, \ldots, e_p\}$, permits to obtain statistics on the behavior of the MH algorithm. Let the *history* matrix $H = (h_{rs})$ defined as followed:

$$h_{rs} = \begin{cases} 1 & \text{if the edge } e_s \text{ was present in the graph } G_r \\ 0 & \text{otherwise} \end{cases}$$

This matrix can be viewed as an usual "individuals × variables" matrix, enabling the computation of various indices. For instance, we can calculate the probability of the appearance of an edge as $p(e_s) = h_{\bullet s}/t$, its variance $\text{var}(e_s) = p(e_s)(1 - p(e_s))$ and the variance-covariance matrix between edges as $\Sigma = \frac{1}{t} H^{c\prime} H^c$ (where H^c is the matrix H after column centration). This variance-covariance matrix permits in turn to apply a *principal component analysis*, where the factor scores

of all observed graphs in the history will underline recurrent configurations in $\Gamma_{\mathcal{V}}$ and the saturations between edges will highlight the competition or the cooperation existing between them.

3 Case studies

3.1 Randomly located nodes

Let us first analyse the behavior of the MH algorithm on small sized graphs. 30 nodes in $\mathbb{R}^2$ with coordinates following a $\mathcal{N}(0,1)$ are drawn, I is set to 50 and the temperature follows the cooling schedule exhibited on the right in Fig. 2 during $t = 20'000$ iterations.

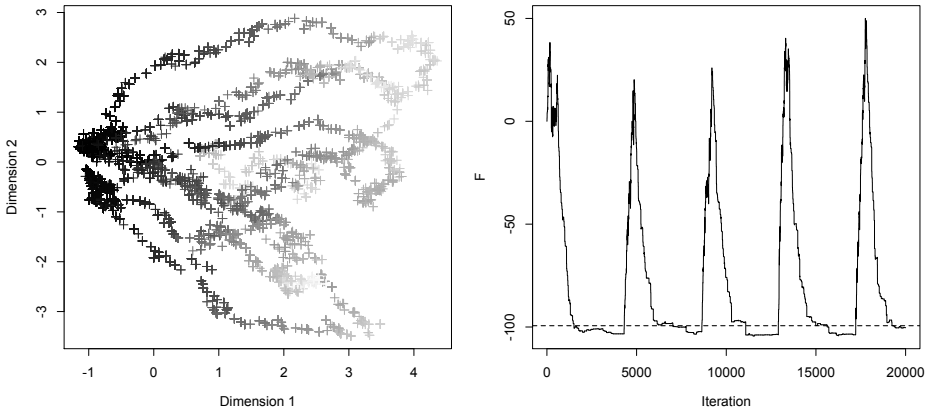

Fig. 3. Results for the graph history in the MH run. Left: first two dimensions of the factor scores, each point represents an iteration (proportion of variance explained: 13%, respectively 9%). The lower the value of F is for each iteration, the darker the point. Right: value of the functional F versus iteration.

As apparent on Fig. 3, the algorithm does not explore very efficiently the graph space. Each time we reheated the system it escaped from a local minimum before converging again on each cool down. The different minima seem to be close to each other, at least according to what appears in the first two factor scores (explaining only 13% and 9% of the variance). Fig. 4 confirms the closeness among the different minima, since some critical edges appear more frequently than others. The saturation plot shows that edges appearing frequently are correlated positively between themselves.

These results, while interesting, are a bit tarnished by the presence of high temperature states. While the presence of these states is essential to escape local

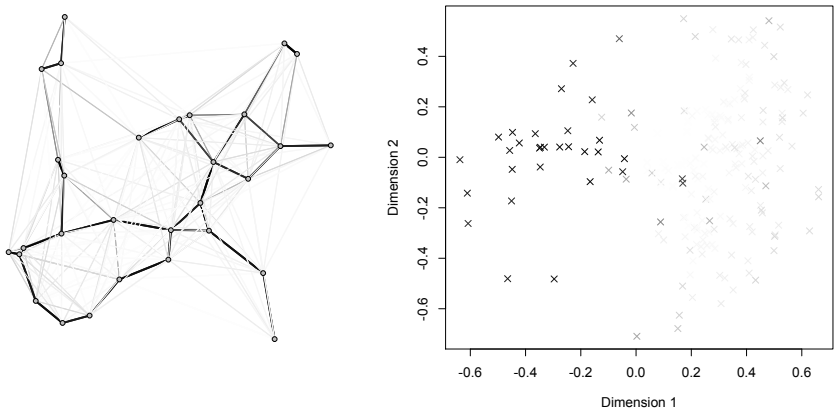

Fig. 4. Results for the edges in the MH run. Left: all edges created at least once during the process. Right: the saturation plot, where each points representing an edge and proximity capture correlation, i.e. two edges appearing frequently together will be close to each other. On both graphics, the darkness of an element is proportional to its apparition frequency during the run.

minima, they bear very little information on optimal and alternative solutions to the efficient network building. Therefore, a second analysis is performed after removing all states having a functional value higher than -100 (corresponding to the dotted line on the Fig. 3, functional plot). By construction, the selected states constitute near-optimal solutions.

The graphic in Fig. 5 illustrates the emergence of five different "cold" temperature regimes during the MH run differ more than what it appear at first glance, showing that they indeed correspond to different local minima of F. Points in the middle yield the lowest value of the functional and correspond to the third cool-down. Graphics in Fig. 6 emphasize the edges created during the process. Here, we can observe some competition between edges. For example, edges numbered 7 and 36, 42 and 49, 20 and 39, 22 and 46, are placed on the opposite side one to another in the saturation plot. In the graph, we can see that both pairs represent building alternative to a close-to-optimal graph. On the other hand, edges 11, 10 and 26 are very centered, meaning that they have a very low variance and represent a kind of "backbone" appearing in any close-to-optimal graph. Iterations 8'643 and 12'903 in Fig. 7 exhibit some built variations. Note that state 12'903 to have a lower functional value than state 8'643 (-103.8 versus -100.4).

3.2 US cities

To study a real life case, the algorithm will be run on nodes representing US cities with more than 500'000 inhabitants (Fig. 8). Latitudes θ_i and longitudes

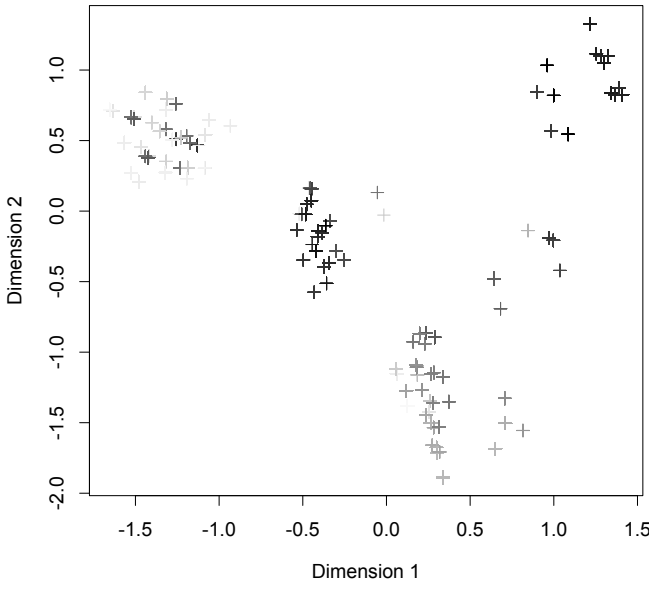

Fig. 5. The first two factorial coordinates of the states of the MH runs where states with high values of F have been removed. The five different cold temperature phases appear clearly, illustrating five different local minima.

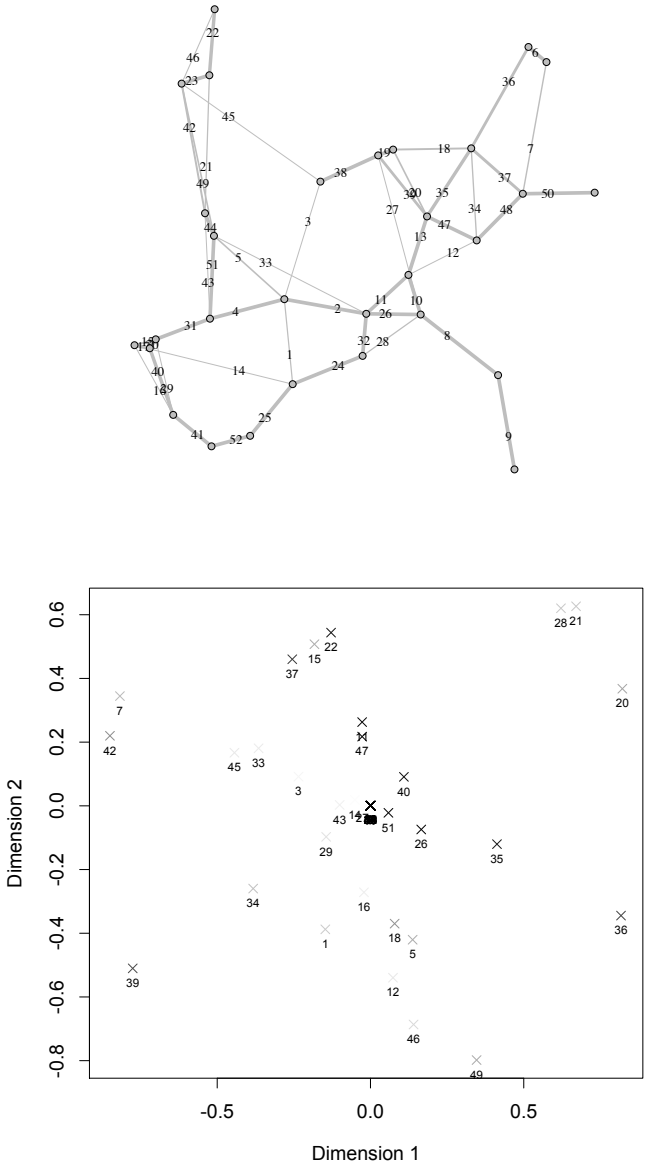

Fig. 6. Top: edges occupation frequencies. Bottom: saturations, with the same labeling.

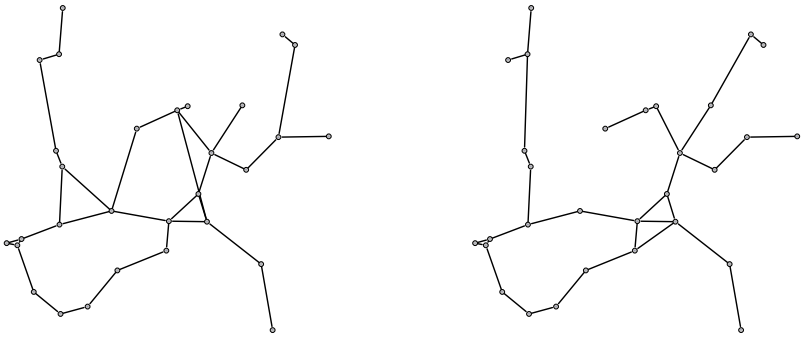

Fig. 7. Graph states at iteration number 8'643 (on the left) and 12'903 (on the right). Their F value are respectively -100.4 and -103.8.

α_i have been extracted from the R data `world.cities{maps}` and we consider the *geodesic dissimilarity* between those cities: $D_{ij} = \arccos^2(\kappa_{ij})$, where $\kappa_{ij} = \sin\theta_i \sin\theta_j + \cos\theta_i \cos\theta_j \cos(\alpha_i - \alpha_j)$. Again, 20'000 iterations of MH are run with an investment of $I = 50$ (distances have been multiplied by 30 to match distances of the previous example) and higher temperature states have been removed from analysis.

Here again, edges frequencies and saturations in Fig. 9 reveal the occurrence of competing edges in the construction of the network together with some robust edges. Note the possibility of weighting each node relatively to its population resulting in a weighted efficiency functional, currently under investigation. Nevertheless, the present result can constitute a good start to explore ways of building real networks, where particular edges can be discarded in a second time, due to some morphological or economical constraints.

4 Conclusion

Exploring the possible graphs states on n nodes by MCMC not only reveals alternatives to the optimal network, but also gives insights on the structure of this space as controlled by the functional F. In the present case, the functional makes the shortest-paths requirement conflicts with the length of the edges, and permits to preliminary explore how the shortest-paths distance is linked to the Euclidean distance in this context. The investment, the cooling schedule, the starting state and the number of iterations are shown to greatly affect this exploration, and a careful design should be made depending on what is searched. The question of how to precisely set the parameters according to spatial configuration in hand

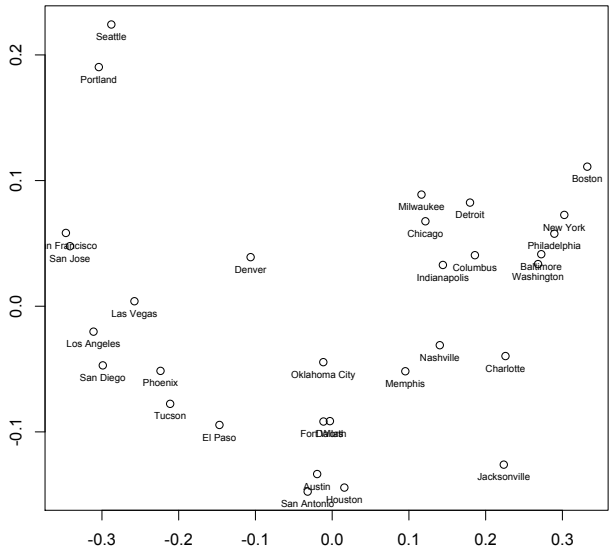

Fig. 8. Representation of the US cities with more than 500'000 inhabitants created by multidimensional scaling from their geodesic dissimilarities.

remains largely open, and a deeper study should be performed before implementing this algorithm in real life applications. The numerical complexity and computational demands of the algorithm are also quite heavy, requiring a way of optimizing the parameters before applying the algorithm to a larger set of nodes. Nevertheless, case studies already show promising results and, provided the procedure can be efficiently refined, its flexibility should permit numerous applications to a large variety of situations.

References

1. Aldous, D.J.: Optimal spatial transportation networks where link costs are sublinear in link capacity. Journal of Statistical Mechanics: Theory and Experiment **2008**(03) (2008) P03006
2. Aldous, D.J., Shun, J., et al.: Connected spatial networks over random points and a route-length statistic. Statistical Science **25**(3) (2010) 275–288
3. Andrieu, C., De Freitas, N., Doucet, A., Jordan, M.I.: An introduction to mcmc for machine learning. Machine learning **50**(1-2) (2003) 5–43
4. Barthélemy, M.: Spatial networks. Physics Reports **499**(1) (2011) 1–101
5. Berg, J., Lässig, M.: Correlated random networks. Physical review letters **89**(22) (2002) 228701
6. Brede, M.: Coordinated and uncoordinated optimization of networks. Physical Review E **81**(6) (2010) 066104

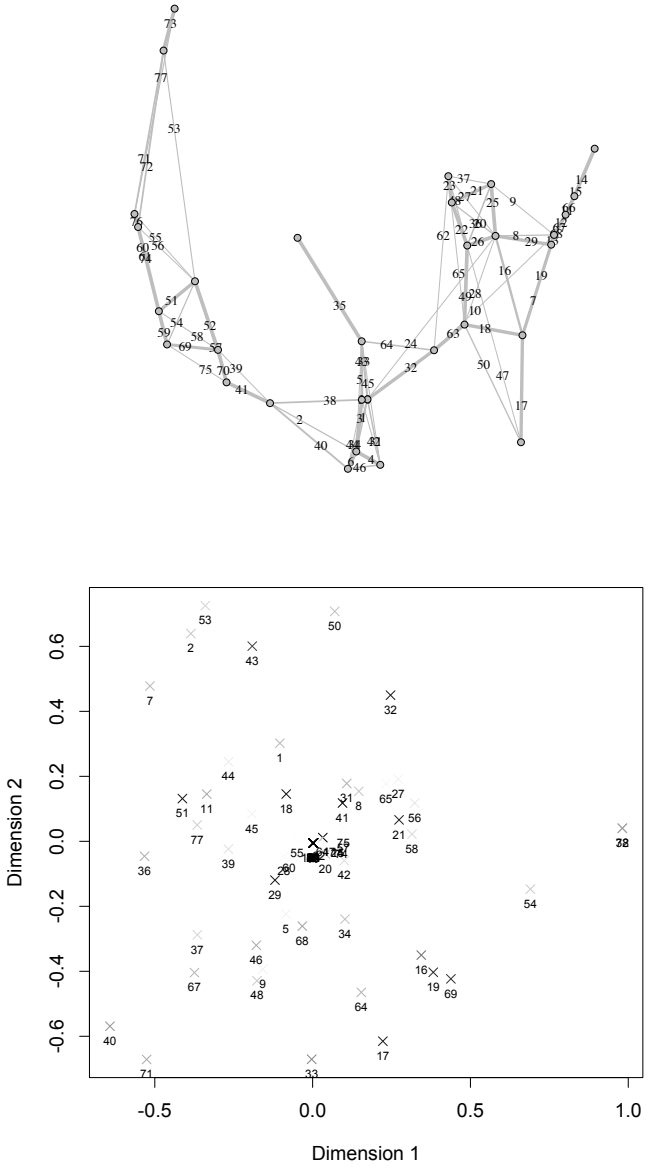

Fig. 9. Top: edges occupation frequencies. Bottom: saturations resulting from a complete MH run with 20'000 iterations.

7. Carter, C.K., Kohn, R.: On gibbs sampling for state space models. Biometrika **81**(3) (1994) 541–553
8. Courtat, T., Gloaguen, C., Douady, S.: Mathematics and morphogenesis of cities: A geometrical approach. Physical Review E **83**(3) (2011) 036106
9. Crucitti, P., Latora, V., Marchiori, M.: A topological analysis of the italian electric power grid. Physica A: Statistical Mechanics and its Applications **338**(1) (2004) 92–97
10. Ferrenberg, A.M., Swendsen, R.H.: New monte carlo technique for studying phase transitions. Physical review letters **61**(23) (1988) 2635
11. Gastner, M.T., Newman, M.E.: The spatial structure of networks. The European Physical Journal B-Condensed Matter and Complex Systems **49**(2) (2006) 247–252
12. Gendron, B., Crainic, T.G., Frangioni, A.: Multicommodity capacitated network design. Springer (1999)
13. Hajek, B.: Cooling schedules for optimal annealing. Mathematics of operations research **13**(2) (1988) 311–329
14. Hastings, W.K.: Monte carlo sampling methods using markov chains and their applications. Biometrika **57**(1) (1970) 97–109
15. Ising, E.: Beitrag zur theorie des ferromagnetismus. Zeitschrift für Physik A Hadrons and Nuclei **31**(1) (1925) 253–258
16. Mathias, N., Gopal, V.: Small worlds: How and why. Physical Review E **63**(2) (2001) 021117
17. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of state calculations by fast computing machines. The journal of chemical physics **21**(6) (1953) 1087–1092
18. Newman, M., Barkema, G.: Monte carlo methods in statistical physics. Oxford University Press (1999)
19. Valverde, S., Cancho, R.F., Sole, R.V.: Scale-free networks from optimal design. EPL (Europhysics Letters) **60**(4) (2002) 512
20. Van Laarhoven, P.J., Aarts, E.H.: Simulated annealing. Springer (1987)
21. Yang, H., H. Bell, M.G.: Models and algorithms for road network design: a review and some new developments. Transport Reviews **18**(3) (1998) 257–278

6 Conclusion

Créer un manuscrit de thèse cohérent à partir d'articles publiés à différentes périodes est un exercice délicat. Bien que ces articles aient été publiés dans le même but, étudier les réseaux spatiaux à l'aide de modélisations mathématiques, les pensées et les buts nous animant lors de leur publication n'ont cessé d'évoluer au fil du temps, et le résultat final a sans doute l'allure d'une grande mosaïque où le lecteur ne sait plus très bien quoi penser sur le sujet. Dans ces circonstances, la conclusion finale prend une importance particulière, dans le sens où elle se doit de résumer au mieux ce qu'on peut relever *globalement* de tout ce travail, les conclusions plus spécifiques sur chaque partie des résultats étant déjà contenues dans les différents articles.

Pour commencer, nous allons voir ce que nous avons appris ici sur les réseaux spatiaux. Comme nous l'avons fréquemment fait remarquer, la spatialité affecte le graphe à travers les *longueurs des arêtes* (ou des arcs), contenues dans la matrice des résistances R. Ces longueurs affectent la structure du graphe, ou simplement les acteurs naviguant sur le graphe, au travers de la minimisation d'une fonction de *coût*, qui pénalise la création, ou l'utilisation, des longues arêtes au détriment des courtes. Ce coût apparaît dans le cadre des flux sous la forme de la fonctionnelle d'énergie et comme la somme de la longueur de toutes les arêtes dans de nombreux modèles de création de réseaux. Seulement, un coût ne peut apparaître seul. Minimiser un coût seul revient à ne rien faire, et c'est là que l'inventivité et la créativité entre en jeu : on se doit de définir un *but*, un *objectif* à notre réseau, qui, par son existence, justifie le fait de dépenser un certain coût.

Dans le chapitre sur la création de modèles, nous avons vu plusieurs manières d'interpréter ce but : avoir un maximum d'accessibilité, obtenir un plus court chemin moyen minimum, ou encore obtenir une unique composante connexe. Définir cet objectif n'est pas une chose aisée : cela va dépendre de la nature du réseau en question et de ce qu'on attend de lui, les facteurs peuvent être multiples et difficiles à interpréter. Malgré cela, les articles cités dans ce chapitre montrent que même si l'on prend une fonctionnelle relativement simple exprimant un objectif, les graphes obtenus présentent de fortes analogies avec des réseaux spatiaux réels, montrant que cet idée de *balance entre coût et but* est une approche qui présente un fort intérêt pour la spatialité. L'apport de cette thèse dans ce domaine a été l'exploration de l'espace des graphes possibles, permettant de nuancer quelque peu le "diktat" d'un minimum global, pas forcément adéquat avec les contraintes réelles, grâce aux alternatives que présentent d'autres minima locaux. Cet exploration de l'espace n'en est, pour le moment, qu'à son balbutiement, mais nous espérons que cette façon de procéder permettra l'émergence d'une méthode permettant de capturer et de jauger efficacement les alternatives, offrant ainsi plus de souplesse aux modèles de réseaux.

En ce qui concerne le flux, l'objectif de celui-ci est facilement défini : transporter le flux depuis les sources jusqu'aux cibles. Ainsi, c'est par des contraintes que nous posons le but qui contrebalance le coût, et ce cadre revient à celui du problème de transport optimal. Un fait intéressant concernant ce problème est qu'un des premiers résultats obtenus par Monge fut d'établir que, si l'on doit transporter un nuage de points $\mathcal{S}$ vers un autre nuage de points $\mathcal{T}$ situés dans un plan, la solution optimale attribue des chemins ne se croisant pas. La chemin de la solution dans le plan est donc un exemple de graphe *planaire*, c'est-à-dire un graphe dans le

plan dont les arêtes ne se croisent pas, un sujet fortement étudié dans le domaine des graphes spatiaux, mais qui, malheureusement, n'entre pas dans la cadre de cette thèse. Sans avoir réellement étudié les conséquences de ce résultat, celui-ci nous montre que des considérations spatiales apparaissent naturellement dans le problème du transport optimal.

Une des grandes nouveautés de cette thèse, qui se base en réalité sur un concept très classique, fut d'ajouter un terme d'entropie au terme d'énergie dans la fonctionnelle que le flux cherche à minimiser, tout en permettant à l'utilisateur de donner de l'importance à l'une des quantités par rapport à l'autre grâce à un paramètre libre. La première idée derrière ce terme d'entropie était d'ajouter une part d'aléatoire dans les déplacements, afin de nuancer les chemins directs des sources au cibles, mais, finalement, cette façon de faire s'est montrée extrêmement riche et fructueuse. Le premier avantage de cet ajout a été de rendre la fonctionnelle convexe. Ainsi, à la place d'utiliser un algorithme résolvant un problème linéaire, on a pu trouver la solution grâce à une dérivation, et le temps de calcul des solutions en a été fortement réduit. Cet aspect a certainement moins d'attraits que d'autres en ce qui concerne la recherche fondamentale, mais est toujours fortement le bienvenu lorsqu'il s'agit d'applications pratiques. Un autre atout de cet ajout a été de généraliser un certain nombre d'indices de centralité et de dissimilarités. En posant la source et la cible comme uniques, le flux de transport randomisé a permis l'interpolation, via la température, entre plusieurs quantités connues sur les réseaux. Un fait intéressant, qui est particulièrement visible avec les dissimilarités créées grâce au flux, est que bien que ces dernières interpolent entre deux dissimilarités connues, elles sont en réalité bien plus qu'une simple mixture des extrémités. En effet, la dissimilarité de l'énergie, vue dans le deuxième article, interpole entre la distance du plus court chemin et la distance de commutation, mais n'est elle-même *pas nécessairement métrique* pour des températures intermédiaires. Ajoutons à cela qu'elle donne de meilleurs résultats lors de son application dans des méthodes de positionnement multidimensionnel et de classification que les deux autres distances classiques, ce qui confirme que cet objet possède réellement de nouvelles caractéristiques. Finalement, la minimisation de cette fonctionnelle composée d'une énergie et d'une entropie soulève plusieurs questions et pistes de recherche.

Une des principales questions fondamentales est la curieuse analogie entre la marche aléatoire et le courant électrique. En effet, il est très curieux de constater que la minimisation de l'entropie, $G(X||W) = \sum_{i,j\in\mathcal{V}} x_{ij} \ln(x_{ij}/(x_{i\bullet}w_{ij}))$, avec $w_{ij} = c_{ij}/c_{i\bullet}$, revient en fait à minimiser une énergie quadratique sur le flux net $U^2(X||C) = \sum_{i,j\in\mathcal{V}}(x_{ij} - x_{ji})^2/c_{ij}$, comme le montre l'analogie entre la marche aléatoire et le courant électrique. On ne peut s'empêcher de penser à l'article d'Alamgir et de von Luxburg (?, ?), qui posent un formalisme de flux orientés $Y = (y_{ij})$ minimisant une "p-énergie", définie par $U^p(Y) = \sum_{i,j\in\mathcal{V}} r_{ij}|y_{ij}|^p$, avec $p \geq 1$. Notre formalisme interpole-il donc entre $U^1(Y)$ et $U^2(Y)$? Rien n'est moins sûr, car il faudrait reformer le problème en terme de flux net, mais cette façon de faire pourrait nous apporter beaucoup d'informations sur cette étonnante analogie entre entropie et énergie quadratique. A l'image de cette question, les liens avec d'autres modèles ne manquent pas. Dans leur article (?, ?), Golnari et Boley posent un formalisme *Bayesian* permettant une interpolation similaire, dans (?, ?), les auteurs montrent la connexion entre le formalisme du RSP (similaire, comme nous l'avons vu, au flux de transport randomisé) et un algorithme de recherche de plus court chemin *local*, l'algorithme de Bellman-Ford (?, ?). La liste des articles proposant un formalisme similaire pourrait encore être allongée (p.ex. (?, ?, ?, ?)). Nous évoquions dans l'introduction le bouillonnement qui précède en science l'éclosion d'une théorie ou d'un formalisme unificateur, et le grand nombre (à l'échelle du domaine, relativisons) de personnes intéressées par le sujet est peut-être le signe, espérons-le, que "quelque chose" est en train de se produire.

Nous sommes donc arrivés au terme de cette thèse plongeant dans le monde des modèles

mathématiques appliqués aux réseaux spatiaux. Les mathématiques ont l'avantage indéniable d'être rigoureusement exactes *dans leur cadre de pensée*. Mais un modèle se doit également d'être une simplification de la réalité, afin d'être suffisamment universel. En effet, en prenant en compte trop de considérations, de paramètres et d'hypothèses, un modèle gagnera certainement en réalisme, mais se contentera d'expliquer des phénomènes dans un cadre extrêmement restreint. Ce phénomène est d'ailleurs bien connu en statistiques, sous le nom de *surparamétrisation*. Un modèle relativement simple, modélisant la réalité dans une certaine mesure, aura l'avantage, de part sa construction, de souligner l'essentiel. Cependant, pour éviter certaines dérives, il est de notre devoir de modélisateurs de ne jamais oublier la citation du statisticien George Box : "Tous les modèles sont faux, mais certains sont utiles". Après la lecture de cette thèse qui met en lumière les nombreux avantages qu'offre l'ajout d'un terme d'entropie, nous sommes sûrs que certaines personnes seraient tentées d'ajouter le petit complément suivant : "Ajoutez de l'incertitude à un modèle, il gagnera en réalisme".

Liste des notations

$\mathcal{T}$	L'ensemble des cibles	13
$f = (f_i)$	Le vecteur des entrées	13
$\rho = (\rho_i)$	Le vecteur des sorties	13
$\mathcal{X}$	L'ensemble des flux admissibles	13
$W^X = (w_{ij}^X)$	La matrice de transition définie par le flux X	14
$U(X) = U(X\|\|R)$	L'énergie du flux X	14
$U^2(X) = U^2(X\|\|C)$	L'énergie quadratique du flux X	14
$Y = (y_{ij})$	Une matrice de flux net	14
$G(X) = G(X\|\|W)$	L'entropie du flux X	15
T	La température	15
β	La température inverse	15
$F(X) = G(X\|\|R, W)$	L'énergie libre du flux X	15
s, t	Le noeud source et le noeud cible	17
$\deg(i)$	Le degré du noeud i	18
$\deg^+(i), \deg^-(i)$	Le degré entrant et le degré sortant du noeud i	18
$B^{dsp}(i), B^{dsp}((i,j))$	La centralité d'intermédiarité des plus courts chemins du noeud i et de l'arc (i,j)	19
$B^{rw}(i), B^{rw}((i,j))$	La centralité d'intermédiarité des chemins aléatoires du noeud i et de l'arc (i,j)	19
$E(i)$	L'excentricité du noeud i	19
$P(i)$	La centralité de proximité du noeud i	19
$P^h(i)$	La centralité de proximité harmonique du noeud i	19
$D = (d_{ij})$	Une matrice de dissimilarité	35
$\mathcal{M}$	L'ensemble des distances	35
$\mathcal{U}$	L'ensemble des dissimilarités ultramétriques	35
$\mathcal{L}_q$	L'ensemble des dissimilarités Minkowski(q)	35
$\mathcal{L}_2$	L'ensemble des dissimilarités Euclidiennes	35
$\mathcal{L}_2^2$	L'ensemble des dissimilarités Euclidiennes carrées	35
$\mathcal{L}^\infty$	L'ensemble des dissimilarités de Chebychev	36
$H = (h_{ij})$	La matrice de centration	36
$K = (k_{ij})$	La matrice des produits scalaires	36
$\mathrm{pow}(d)$	La puissance de la dissimilarité d	36
$d^{ct}(s,t)$	La distance de commutation entre s et t	37
κ_{ij}	La capacité de l'arc (i,j)	37
$d^{maxf}(s,t)$	La distance du flux maximal entre s et t	37
$\tilde{\mathcal{P}}_{st}$	L'ensemble des chemins absorbants de s à t	49
$\mathbb{P}_{st}$	Une mesure de probabilité sur $\tilde{\mathcal{P}}_{st}$	49
$\mathbb{P}_{st}^{\mathrm{ref}}$	La mesure de référence sur $\tilde{\mathcal{P}}_{st}$	49
$\mathbb{P}_{st}^{\mathrm{RSP}}$	La mesure des chemins aléatoires randomisés sur $\tilde{\mathcal{P}}_{st}$	49
$\mathcal{Z}$	La fonction de partition	50
$\eta_{ij}(s,t)$	L'espérance RSP du nombre de passages par l'arc (i,j)	50
$B^{\mathrm{RSP}}(i)$	La centralité RSP du noeud i	50
$B^{\mathrm{RSPnet}}(i)$	La centralité RSP nette du noeud i	50
$\bar{l}^d(\mathbb{P}_{st})$	L'espérance de la longueur valuée des chemins entre s et t par rapport à $\mathbb{P}_{st}$	51
$d^{\mathrm{RSP}}(s,t)$	La dissimilarité du RSP entre s et t	51
$F(\mathbb{P}_{st})$	L'énergie libre de la mesure $\mathbb{P}_{st}$	51
$d^{\mathrm{FE}}(s,t)$	La distance d'énergie libre entre s et t	51
$P = (p_{ij})$	Un couplage	55
$u = (u_i), v = (v_i)$	Les vecteurs du problème dual	56
$\mathcal{G}^{\mathrm{opt}}$	Un graphe optimal	101
$\Gamma_{\mathcal{V}}$	L'espace des graphes sur $\mathcal{V}$	103

Annexe : Pseudo-codes des algorithmes

Cette annexe contient les pseudo-codes des différents algorithmes utilisés. Pour des question de lisibilité, la notation de ces algorithmes est légèrement différente que celle qui se trouve dans le manuscrit.

Notations

$\mathbf{M}$: une matrice M.
m_{ij} : la composante (i, j) de la matrice M
$\mathbf{v}$: un vecteur v.
v_i : la composante i du vecteur v.
s : un scalaire s.
$\top$: l'opération transposée
$\mathbf{I}^n$: la matrice identité de dimensions $(n \times n)$.
$\mathbf{1}^n$: le vecteur de taille n composé uniquement de 1.
$\mathbf{e}_i^n$: un vecteur de taille n dont toute les composantes sont égales à 0, sauf la $i$ème qui est égale à 1.
$\circ$: la multiplication composantes par composantes.
$\div$: la division composantes par composantes.
$\mathbf{M}^{-1}$: la matrice inverse de M.
$\mathbf{M}^*$: la matrice pseudo-inverse de M.
rand() : fonction effectuant le tirage d'un nombre selon une loi uniforme entre $[0, 1]$.
ceiling$(float)$: fonction arrondissant le nombre à l'entier supérieur.

Algorithm 1: L'algorithme permettant d'obtenir le flux de transport randomisé avec une source et une cible (articles 1 et 2).

Input :
- Un graphe orienté fortement connexe $\mathcal{G}$ avec n noeuds.
- Une matrice de résistance $\mathbf{R}$, de taille $(n \times n)$, définie sur $\mathcal{G}$.
- Une matrice de transition de chaîne de Markov $\mathbf{W}$, $(n \times n)$, définie sur $\mathcal{G}$.
- Deux indices, s and t, qui définissent respectivement la source et la cible.
- La température inverse $\beta > 0$.

Output:
- La matrice du flux de transport randomisé $\mathbf{X}$, de taille $(n \times n)$.

$\tilde{\mathbf{V}} \leftarrow \mathbf{W} \circ \exp[-\beta \mathbf{R}]$
$\mathbf{q} \leftarrow (\tilde{\mathbf{V}} \mathbf{e}_t^n) \circ (\mathbf{1}^n - \mathbf{e}_t^n)$
$\mathbf{V} \leftarrow \tilde{\mathbf{V}} \circ ((\mathbf{1}^n - \mathbf{e}_t^n)(\mathbf{1}^n - \mathbf{e}_t^n)^\top)$
$\mathbf{M} \leftarrow (\mathbf{I}^n - \mathbf{V})^{-1}$
$\mathbf{z} \leftarrow \mathbf{M}\mathbf{q}$
$\tilde{\mathbf{X}} \leftarrow \frac{1}{z_s} (\mathbf{V} \circ (\mathbf{1}^n \mathbf{z}^\top) \circ (\mathbf{M}^\top \mathbf{e}_s^n \mathbf{1}^{n\top}))$
$\tilde{\mathbf{x}} \leftarrow \frac{1}{z_s} (\mathbf{M}^\top \mathbf{e}_s^n \circ \mathbf{q})$
$\mathbf{X} \leftarrow \tilde{\mathbf{X}} + \tilde{\mathbf{x}} \mathbf{e}_t^{n\top}$

return X

Algorithm 2: L'algorithme permettant d'obtenir la solution régularisée du problème de transport sur un graphe (article 3).

Input :
- Un graphe orienté fortement connexe $\mathcal{G}$ avec n sources et m cibles.
- Un vecteur $\mathbf{f}$ de taille n, contenant la proportion d'offre aux sources.
- Un vecteur $\boldsymbol{\rho}$ de taille m, contenant la proportion de demande aux cibles.
- Une matrice de dissimilarité entre sources et cibles $\mathbf{D}$, de taille $(n \times m)$.
- La température inverse $\beta > 0$.
- Le paramètre $\sigma > 0$, contrôlant le seuil de tolérance dans la convergence.

Output:
- La matrice de couplage $\mathbf{P}$, de taille $(n \times m)$.

$\tilde{\mathbf{D}} \leftarrow \exp[-\beta \mathbf{D}]$
$\mathbf{v} \leftarrow \mathbf{1}^m$
$\mathbf{v}^{\mathbf{prev}} \leftarrow (1 + 2\sigma)\mathbf{v}$
while $\max_i |v_i^{\mathbf{prev}} - v_i| > \sigma$ **do**
  $\mathbf{v}^{\mathbf{prev}} \leftarrow \mathbf{v}$
  $\mathbf{u} \leftarrow \mathbf{1}^n \div (\tilde{\mathbf{D}} \circ \mathbf{1}^n (\mathbf{v} \circ \boldsymbol{\rho})^\top) \mathbf{1}^m$
  $\mathbf{v} \leftarrow \mathbf{1}^m \div (\tilde{\mathbf{D}} \circ (\mathbf{u} \circ \mathbf{f}) \mathbf{1}^{m\top})^\top \mathbf{1}^n$
end
$\mathbf{P} \leftarrow \tilde{\mathbf{D}} \circ \mathbf{1}^n (\mathbf{v} \circ \boldsymbol{\rho})^\top \circ (\mathbf{u} \circ \mathbf{f}) \mathbf{1}^{m\top}$

return P

Algorithm 3: L'algorithme permettant d'obtenir le flux de transport randomisé avec de multiples sources et cibles et le couplage correspondant (article 4).

Input :

- Un graphe orienté fortement connexe $\mathcal{G}$ avec n noeuds.
- Une matrice de résistance $\mathbf{R}$, de taille $(n \times n)$, définie sur $\mathcal{G}$.
- Une matrice de transition de chaîne de Markov $\mathbf{W}$, $(n \times n)$, définie sur $\mathcal{G}$.
- Un vecteur $\mathbf{f}$, de taille n, définissant les entrées du flux.
- Un vecteur $\boldsymbol{\rho}$, de taille n, définissant les sorties du flux.
- Le paramètre $\tilde{\gamma} > 0$, contrôlant la probabilité d'absorption aux cibles.
- La température inverse $\beta > 0$.
- Le paramètre $\sigma > 0$, contrôlant le seuil de tolérance dans la convergence.

Output:

- La matrice du flux de transport randomisé $\mathbf{X}$, de taille $(n \times n)$.
- La matrice de couplage $\mathbf{P}$ correspondant, de taille $(n \times n)$.

$\mathbf{Q} \leftarrow \mathbf{I}^n - \mathbf{W}^\top$
$\tilde{\mathbf{x}}^\infty \leftarrow \mathbf{Q}^*(\mathbf{f} - \mathbf{W}^\top \boldsymbol{\rho})$
$\mathbf{q} \leftarrow (\mathbf{I}^n - \mathbf{Q}^*\mathbf{Q})\mathbf{e}_1^n$
$\gamma \leftarrow \tilde{\gamma} \ + \max_i \left(\frac{\rho_i - \tilde{x}_i^\infty}{q_i} \right)$
$\mathbf{x}^\infty \leftarrow \tilde{\mathbf{x}}^\infty + \gamma \mathbf{q}$
$\boldsymbol{\alpha} \ \leftarrow \boldsymbol{\rho} \div \mathbf{x}^\infty$
$\mathbf{W}^+ \leftarrow \mathbf{W} \circ (\mathbf{1}^n - \boldsymbol{\alpha})\mathbf{1}^{n\top}$

$\mathbf{V} \leftarrow \mathbf{W}^+ \circ \exp[-\beta \mathbf{R}]$
$\mathbf{M} \leftarrow (\mathbf{I}^n - \mathbf{V})^{-1}$
$\mathbf{a} \leftarrow \mathbf{1}^n$
$\mathbf{a}^{\mathbf{prev}} \leftarrow (1 + 2\sigma)\mathbf{a}$
while $\max_i |a_i^{\mathbf{prev}} - a_i| \ > \sigma$ **do**
    $\mathbf{a}^{\mathbf{prev}} \leftarrow \mathbf{a}$
    $\mathbf{b} \leftarrow \mathbf{M}^\top(\mathbf{f} \div \mathbf{a})$
    $\mathbf{a} \leftarrow \mathbf{M}(\boldsymbol{\rho} \div \mathbf{b})$
end
$\mathbf{X} \leftarrow \mathbf{V} \circ \mathbf{1}^n\mathbf{a}^\top \circ \mathbf{b}\mathbf{1}^{n\top}$

$\widehat{\mathbf{W}} \leftarrow \mathbf{V} \circ (\mathbf{1}^n\mathbf{a}^\top) \div (\mathbf{a}\mathbf{1}^{n\top})$
$\widehat{\mathbf{M}} \leftarrow (\mathbf{I}^n - \widehat{\mathbf{W}})^{-1}$
$\mathbf{P} \leftarrow \widehat{\mathbf{M}} \circ \mathbf{1}^n(\mathbf{1}^n - \widehat{\mathbf{W}}\mathbf{1}^n)^\top \circ \mathbf{f}\mathbf{1}^{n\top}$

return $\mathbf{X}, \mathbf{P}$

Algorithm 4: L'algorithme permettant d'obtenir la matrice de l'historique de création des graphes en suivant la méthode d'exploration de l'espace (article 5).

Input :
- n noeuds dans un espace métrique.
- Un vecteur d'adjacence initial $\mathbf{a}^0$, de taille $n(n-1)$.
(une matrice d'adjacence "déroulée", sans la diagonale).
- Une fonctionnelle $F(\mathbf{a})$, calculant l'utilité du graphe en fonction de $\mathbf{a}$.
- Un vecteur de températures inverses $\boldsymbol{\beta} = (\beta_i)$, de taille t.

Output:
- La matrice de l'historique de création des graphes $\mathbf{H}$, de taille $(t \times n(n-1))$.

$\mathbf{a} = \mathbf{a}^0$
$\mathbf{H} \leftarrow \mathbf{1}^t \mathbf{1}^{n(n-1)^\top}$
$i \leftarrow 1$
while $i < t + 1$ **do**
  $\mathbf{a}^{\mathbf{new}} \leftarrow \mathbf{a}$
  $c = \mathbf{ceiling}(n(n-1) \cdot \mathbf{rand}())$
  $a_c^{\mathbf{new}} \leftarrow 1 - a_c^{\mathbf{new}}$
  if $\min(1, \exp(\beta_i(F(\mathbf{a}) - F(\mathbf{a}^{\mathbf{new}})))) > \mathbf{rand}()$ **then**
    $\mathbf{a} \leftarrow \mathbf{a}^{\mathbf{new}}$
  end
  $\mathbf{H} \leftarrow \mathbf{H} \circ (\mathbf{1}^t - \mathbf{e}_i^t)\mathbf{1}^{n(n-1)}$
  $\mathbf{H} \leftarrow \mathbf{H} + \mathbf{e}_i^t \mathbf{a}^\top$
  $i \leftarrow i + 1$
end

return H