Construction de Réseaux d'Ordre Supérieur à partir de Traces : Méthodes et Outils

Julie Queiros, François Queyroi

▶ **To cite this version:**

Julie Queiros, François Queyroi. Construction de Réseaux d'Ordre Supérieur à partir de Traces : Méthodes et Outils. 2023. hal-04085138v1

HAL Id: hal-04085138

https://hal.science/hal-04085138v1

Preprint submitted on 28 Apr 2023 (v1), last revised 25 Mar 2024 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Construction de Réseaux d'Ordre Supérieur à partir de Traces : Méthodes et Outils

Julie Queiros
François Queyroi
Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004
F-44000 Nantes, France *

28 avril 2023

Résumé

Les réseaux d'ordre supérieur sont une classe de réseaux qui intègrent des "nœuds-mémoires" afin de prendre en compte les interactions pouvant exister dans des données séquentielles, par opposition aux réseaux dits d'"ordre 1" qui ne prennent en compte que les relations directes. Dans cet article, nous donnons un aperçu de ce concept en détaillant leur construction et les techniques de fouille qui peuvent être employées. Nous proposons un didacticiel sur un cas d'étude utilisant une implémentation de notre part des algorithmes présents dans la littérature. Nous abordons également certains des défis et des orientations futures dans ce domaine.

1 Introduction

Les réseaux sont un outil fondamental pour la modélisation de systèmes complexes, composés d'un grand nombre d'éléments. Utilisés dans divers domaines, tels que la physique, la biologie et les sciences sociales, ces systèmes sont caractérisés par leurs propriétés topologiques non triviales, telles que la présence de nœuds hautement connectés, la présence d'une structure communautaire et l'existence de corrélations entre les nœuds, etc. La compréhension de la structure et de la dynamique des réseaux complexes est cruciale pour de nombreuses applications, telles que la propagation des maladies ou la diffusion de l'information. Ces réseaux peuvent être modélisés et analysés à l'aide d'outils informatiques et mathématiques ; la théorie des graphes permet d'en étudier les propriétés. L'étude des réseaux complexes est un domaine interdisciplinaire qui s'appuie sur des idées et

*Autrice correspondante : julie.queiros@univ-nantes.fr

des méthodes issues de nombreux autres domaines tels que la physique, les mathématiques, l'informatique et les sciences sociales. La quantité de données disponibles sur les réseaux du monde réel ne cessant d'augmenter, la capacité à comprendre, analyser et contrôler les réseaux complexes devient de plus en plus importante pour un large éventail d'applications.

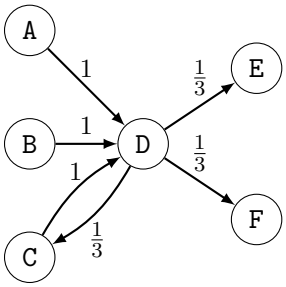
Dans certains cas, les données sont séquentielles. Elles représentent une suite de changements d'*états* et l'analyse de réseau peut être utilisée pour étudier les relations entre ces états à partir de ces transitions. Des exemples courants sont des trajets maritimes (séquences de ports d'escale d'un navire) ou bien des trajets d'utilisateurs sur Internet (séquences des pages Web visitées). Pour illustrer, nous prendrons comme exemple les séquences donnés dans la Figure 1a.

Ces données séquentielles peuvent être représentées par des réseaux dits de *premier ordre* ou bien *sans-mémoire* que l'on notera FON_1, modélisant les interactions par paire entre les états $A, B, \ldots, F$. Les réseaux FON_1 sont construits en agrégeant les occurrences entre paires d'états dans le jeu de données d'entrée. La Figure 1b est le réseau FON_1 construit à partir des données 1a. En prenant pour exemple les états D et E, la *probabilité de transition* entre les deux états est égale au nombre d'occurrences de la sous-séquence DE divisé par le nombre total d'occurrences de sous-séquence D. On a 6 occurrences de DE pour un total de 18 séquences avec D. Dans ce réseau, on remarque que, partant de D, on aura la même probabilité de passer au nœud C,E ou F. Cependant, en regardant les séquences, on se rend compte qu'il existe des relations indirectes. Par exemple, si on suit uniquement le réseau FON_1, il serait possible d'avoir une séquence CDE, en suivant les relations directes entre C et D puis D et E. Or, cette séquence n'apparaît pas dans le jeu de données de base. En ne prenant en compte que les relations d'ordre 1, ou directes, on perd les relations indirectes et les réseaux sont moins fidèles aux comportements des données séquentielles d'entrée.

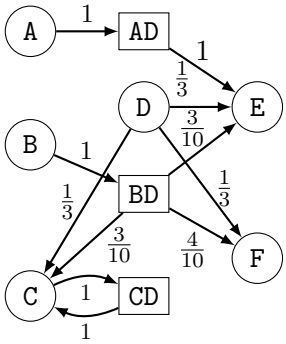
Les transitions d'un réseau de premier ordre se rapprochent d'un *processus de Markov* traditionnel, où l'état futur d'un système dépend uniquement de l'état actuel, et non des états précédents. C'est ce qu'on appelle *la propriété de Markov*. L'hypothèse que les relations entre états sont markoviennes a été étudiée et remise en question (Chierichetti, Kumar, Raghavan, & Sarlos, 2012; Rosvall, Esquivel, Lancichinetti, West, & Lambiotte, 2014). Dans de nombreux systèmes du monde réel, les interactions entre les nœuds seraient ainsi plus complexes et impliqueraient des dépendances séquentielles. La perte d'information liée à l'utilisation de réseaux de premier ordre peut avoir un impact sur les résultats obtenus par les outils de fouille appliqués aux réseaux.

Séquences	#
A D E	2
D E	1
B D E	3
D F	2
B D F	4
B D C	3
C D C	3

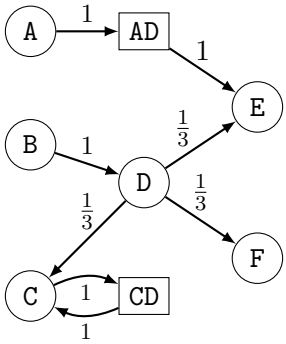
(a) Données séquentielles



(b) Réseau d'ordre 1 (FON$_1$)



(c) Réseau d'ordre 2 (FON$_2$)



(d) Réseau d'ordre variable (VON)

FIGURE 1 – Exemple de construction de réseaux à partir des séquences 1a. Pour un ensemble d'états possibles $\mathcal{A} = \{A, B, C, D, E, F\}$, on observe par exemple la séquence BDF quatre fois. Le réseau d'ordre 1 (Fig. 1b) est obtenu en prenant les probabilité de transitions directes. Dans l'état D, on a une chance sur trois d'aboutir à l'état E. Les réseaux d'ordres supérieurs FON$_2$ et VON (Fig. 1c et 1d) incluent des dépendances indirectes sous la forme de nœuds-mémoire (rectangles). L'état D a ainsi 4 et 3 représentations respectivement.

Pour représenter ces interactions d'ordre supérieur, les chercheurs ont développé le concept de réseaux d'ordre supérieur (Torres, Blevins, Bassett, & Eliassi-Rad, 2021). Ceux-ci viennent ajouter les relations indirectes directement sur le réseau. Les Figures 1c et 1d sont respectivement des exemples de réseau d'ordre fixe et d'ordre variable. On va prendre en compte ces dépendances d'ordre supérieur afin de construire des réseaux plus fidèles aux données d'entrée. Pour cela, sont ajoutés directement sur le réseau des "noeuds-mémoires" qui encodent directement les dépendances indirectes. Ces nœuds sont à lire comme suit : AD est une représentation d'ordre 2 de D, et correspond à l'événement "être dans l'état D après l'état A". La réseau

en Figure 1c inclut toutes les sous-séquences d'ordre 2 présentes dans les séquences d'entrée, à la différence du réseau de la Figure 1d qui permet de faire un choix entre les sous-séquences à ajouter. Cette différence sera expliquée en détail plus bas. Dans notre exemple, l'ordre maximal des réseaux est de 2. Cependant, il serait possible, si le jeu de données le permettait, d'avoir des ordres plus grands. Si remonter k étapes en arrière permet de révéler des interactions non présentes dans les ordres inférieurs, alors il serait possible d'avoir un réseau avec des nœuds d'ordre k.

Une question importante est de savoir si la prise en compte de ces dépendances peut avoir un impact sur l'analyse faite des réseaux (algorithmes de centralité ou de *clustering*). Malgré les nœuds-mémoire ajoutés, les réseaux d'ordre supérieur restent des graphes de même nature que les réseaux d'ordre 1. Des algorithmes de fouille pourraient ainsi y être directement appliqués (Xu, Wickramarathne, & Chawla, 2016), même si, comme nous allons le voir, certaine adaptations sont nécessaires.

Mis à part les différences d'échantillonnage ou de stratégie de filtrage des données, il n'y a qu'un seul réseau d'ordre 1, FON_1, possible pour un ensemble de traces données. Cependant, il existe de multiples façons de construire un réseau d'ordre supérieur. Différents modèles ont été étudiés dans la littérature sur ce sujet relativement récent, le but étant de trouver le meilleur modèle qui représente les données de façon fidèle. Il s'agit d'un compromis entre la fidélité aux données et la taille du réseau, ce dernier ne doit pas être excessivement grand, afin de faciliter l'utilisation finale de ce genre de réseau.

Dans la suite, nous allons faire un état des lieux des méthodes proposées afin de construire des réseaux d'ordre supérieur. Dans un premier temps, nous discuterons de la place de ce sujet dans la littérature. Nous proposerons ensuite des définitions formelles des réseaux à ordre fixe, FON_k et des réseaux à ordre variable, VON. Nous continuerons avec un cas d'étude comparant ces différents modèles et finirons avec une discussion sur les limitations et les enjeux de l'étude de ce genre de réseaux.

2 Contexte et utilisations

Nous allons dans cette section apporter des précisions sur la terminologie d'ordre supérieur qui peut recouvrir plusieurs concepts existant en analyse des réseaux. Nous allons par ailleurs discuter des domaines d'applications de ces objets issus de la littérature.

La notion d'"ordre supérieur" (*"higher-order"*) couvre un champ concep-

tuel assez large. Ainsi (Eliassi-Rad, Latora, Rosvall, & Scholtes, 2021) définissent les "réseaux d'ordres supérieurs" comme tous réseaux conçus pour capturer d'avantage que les relations *dyadiques*. On peut par exemple citer les interactions de dépendances d'ensemble que sont les relations de co-auteurs (Battiston et al., 2020). À l'instar de (Torres et al., 2021), nous utiliserons exclusivement ce terme pour décrire les réseaux permettant la prise en compte des dépendances séquentielles entre des états.

Le champ d'applications des réseaux d'ordres supérieurs est également large. Notons tout d'abord que les données séquentielles se démarquent des données temporelles. En effet, pour ces dernières l'instant où un changement d'état est observé ou la durée entre deux événements sont des dimensions essentielles tandis que c'est principalement l'enchaînement des états qui importe dans le cas des données séquentielles. Beaucoup de données de déplacements ou de flux géographiques sont ainsi concernées. Dans le cas des échanges maritimes (Ducruet & Notteboom, 2012 ; Kaluza, Kölzsch, Gastner, & Blasius, 2010), les séquences correspondent aux ports dans lesquels des navires de type porte-conteneurs font successivement escale. Dans ce cadre, la durée de l'escale ou de navigation entre escales sont des informations secondaires. L'utilisation de réseaux d'ordre supérieur permet notamment, pour cette application, l'étude de la propagation d'algues ou de micro-organismes transportés dans les eaux de ballast entre différentes régions du globe (Drake & Lodge, 2004 ; Xu et al., 2016). Les séquences de "déplacements" peuvent également s'envisager dans un espace numérique. Les réseaux d'ordre supérieur ont ainsi été utilisés afin de mieux prendre en compte les dépendances séquentielles observées dans les pratiques de navigation sur le Web (Chierichetti et al., 2012 ; Rosvall et al., 2014). Notons que la différence entre séquentiel et temporel n'exclut pas une réflexion sur les durées séparant les évènements. Ainsi (Scholtes, 2017) construit des séquences d'échanges entre personnes à partir de données de courriels en utilisant un seuil de durée pour différencier les conservations sur le même sujet.

Enfin, l'utilisation de données séquentielles dans les réseaux d'ordre supérieurs doit être distinguée de l'"analyse de séquences". Dans cette dernière, les individus sont en partie définis par une séquence d'état. Des mesures de distances entre ces séquences sont mobilisées pour en déduire des classes d'individus. Ces techniques sont notamment utilisées dans le cadre de l'étude des évolutions de composition socio-professionnelle de quartiers (Rivière, Madoré, Batardy, Garat, & Raimbault, 2021) ou des "parcours de vie" (Robette, 2011). À l'inverse, l'analyse des réseaux d'ordre supérieur s'intéresse aux relations entre états, les composantes individuelles étant agrégées.

3 Modèles de réseaux d'ordre supérieur

Nous donnons ici une définition formelle des réseaux d'ordre supérieur et des concepts qui y sont liés. Nous discuterons en particulier deux modèles : les réseaux d'ordre fixe (noté FON$_k$) et d'ordre variable (noté VON). Enfin, nous évoquerons différentes utilisations de ces réseaux dans le cadre d'analyses quantitatives.

3.1 Définitions

Dans cette partie, nous allons nous attacher à fournir certaines définitions préliminaires qui permettront une meilleure compréhension du sujet.

Définition 1 (Séquences)**.** *Pour un ensemble d'état donné $\mathcal{A}$, s est appelée une* séquence *d'éléments de A et correspond à une suite finie de $\mathcal{A}$ i.e. $s = \sigma_1\sigma_2 \dots \sigma_m$. Un* jeu de données *est constitué d'un ensemble de séquences $\mathcal{S}$ sur les états de $\mathcal{A}$. L'*ordre *de la séquence s est sa longueur et est noté $|s|$. Nous utilisons la notation $c(s)$ pour désigner le nombre d'occurrences de s dans l'ensemble de données $\mathcal{S}$.*

Dans l'exemple donné en Figure 1a, pour la séquence $s = BD$ d'ordre 2, on a $c(BD) = 10$.

Définition 2 (Suffixe et préfixe)**.** *Pour une séquence donnée $s = \sigma_1\sigma_2 \dots \sigma_m$, la séquence s' est appelée* suffixe *de s si les $|s'|$ derniers états de s forment la sous-séquence s'. De plus, on dira que s' est un préfixe de s si les $|s'|$ premiers états de s forment la sous-séquence s'.*

Toujours dans le même exemple, si on prend la séquence BDF, celle-ci admet BD ou B comme préfixes et DF ou F en suffixes.

Un modèle séquentiel est une estimation construite à partir du jeu de données de la probabilité d'observer un état donné en tenant compte des états précédents. Dans l'exemple de la Figure 1, partant de l'état C pour aller en D, on veut savoir la probabilité de revenir en C. Dans notre cas, les modèles étudiés vont exclusivement utiliser l'information séquentielle et aucun autre prédicat. Pour des raisons détaillées ci-dessous, les modèles séquentiels étudiés ne sont pas exhaustifs : les probabilités de transition ne sont définies que pour certaines sous-séquences appelées *contextes*. Sans perte de généralité, on considère qu'un modèle séquentiel correspond à un ensemble de contextes associés à des probabilités de transition définies comme suit.

Définition 3 (Probabilité de transition)**.** *Pour une séquence s donnée et un modèle séquentiel* M*, la probabilité de transition vers l'état $\sigma \in A$ sachant les états précédents s est :*

$$P_{\mathrm{M}}(\sigma|s) = \frac{c(s^*\sigma)}{\sum_{\sigma' \in A} c(s^*\sigma')} \tag{1}$$

où s^* est le plus grand contexte suffixe de s dans M. *On note également $P_{\mathrm{M}}(.|s) = \{P_{\mathrm{M}}(\sigma|s)\}_{\sigma \in \mathcal{A}}$ la distribution des états possibles après la séquence s.*

Les estimations que nous allons utiliser correspondent à la formule 1 qui est l'estimation du maximum de vraisemblance étant donné l'ensemble de données $\mathcal{S}$. Elle correspond grossièrement aux nombre d'occurrences de $s^*\sigma$ sur le nombre d'occurrences de s^* observées dans $\mathcal{S}$.

Les contextes dans le FON$_1$ sont uniquement les séquences d'ordre 1 *i.e.* l'ensemble des états $\mathcal{A}$. Dans ce cadre, la probabilité $P(C|CD)$ d'aller en C sachant qu'on a visité C puis D est donnée par l'estimation $P(C|D) = \frac{1}{3}$. Ainsi, le modèle d'ordre 1 ne va pas utiliser d'informations précédentes hormis le dernier état visité.

Comme pour n'importe quel modèle statistique, on peut ainsi évaluer la *précision* de M *i.e.* sa capacité à correctement prédire des séquences. Cette précision sera équivalente à la capacité pour une marche aléatoire sur un réseau construit à partir de M (voir définition ci-dessous) à bien simuler le jeu de données séquentiel modélisé.

Définition 4 (Réseau d'ordre supérieur)**.** *Étant donné un modèle séquentiel* M*, le réseau d'ordre supérieur $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ représentant* M *est un graphe orienté pondéré où chaque sommet de $\mathcal{V}$, appelé nœud-mémoire correspond à un contexte de M. Par simplicité, on considère le sommet et le contexte comme un même objet. On parlera ainsi de l'ordre (longueur) d'un nœud-mémoire.*
Soit $\sigma \in \mathcal{A}$ et un contexte s de M *pour lesquels $P_{\mathrm{M}}(\sigma|s) > 0$, $\mathcal{E}$ inclura un arc $(s \to s^*\sigma)$ de poids $P_{\mathrm{M}}(\sigma|s)$ où s^* est le plus grand suffixe de s dans* M.

Le réseau donné en Figure 1b correspond au FON$_1$ pour les séquences 1a et est appelé FON$_1$. Comme dit plus haut, on voit que, pour cet exemple, ce modèle ne permet pas de reproduire les dépendances indirectes existantes dans les données. Les réseaux d'ordre supérieur définis par la suite vont ainsi s'opposer à ce premier modèle dans le sens où ils utiliseront des nœuds avec un ordre *supérieur* à 1. Notons que le modèle FON$_1$ reste un modèle possible. En particulier, s'il n'existe effectivement pas de dépendances indirectes, c'est le modèle le plus simple pour rendre compte des transitions entre états.

Ces réseaux sont dit stochastiques car le poids des arcs sortants définissent une distribution sur les états suivants possibles. Une marche aléatoire sur ce type de graphe consiste à passer d'un nœud à l'autre en utilisant ces poids pour déterminer le prochain saut. On dit ainsi qu'une marche

aléatoire sur le graphe FON_1 est “sans-mémoire” : on considère qu'une fois arrivé en D on "oublie" être venu de C ou que, implicitement, cette information n'est pas pertinente (hypothèse de Markov mentionnée plus haut).

3.2 Réseaux d'ordre fixe

Le réseau d'ordre fixe k (noté FON_k) s'obtient naturellement en considérant comme contextes du modèle les sous-séquences observées de longueur inférieure ou égale à k. La Figure 1c représente un réseau FON_2. Dans cet exemple, le nœud d'ordre 1, A, est connecté au nœud d'ordre 2 AD. Une marche aléatoire débutant en A et allant à D va donc indirectement utiliser les probabilités de transitions correspondant aux arcs sortants de AD pour déterminer le prochain état visité. Notons que, dans cet exemple simplifié, les seuls contextes d'ordre 2 possibles forment tous des représentations de D. De même, FON_3 correspondrait au même réseau car il n'y a pas de séquence de longueur supérieure à 3. En général, les représentations de plusieurs états différents sont possibles.

Dans la littérature, ces réseaux sont qualifiés de “mixtes” ou “d'ordre multiple” (Rosvall et al., 2014 ; Scholtes, 2017) car ils contiennent des nœuds-mémoires d'ordre différent. Toutefois, afin de simplifier les notations, nous appellerons simplement ce type de réseau “Réseaux d'ordre fixe k” ou FON_k car le paramètre k doit être fixé *a priori* et est valable pour tout le système. L'avantage de ce type de réseau est la simplicité avec laquelle il est possible de les construire, il suffit en effet d'énumérer les sous-séquences de longueur inférieure ou égale à k présentes dans le jeu de données en entrée. Ce sont les réseaux d'ordres supérieurs qui furent les premiers étudiés (Rosvall et al., 2014).

Or, deux problèmes principaux se posent avec le modèle FON_k. Le premier est l'augmentation exponentielle de la taille des réseaux avec le paramètre k. En effet, le nombre de sous-séquences de taille k est de l'ordre de $\mathcal{O}(N^k)$ où $N = |\mathcal{A}|$ est le nombre d'états possibles. Ainsi le réseau FON_k peut être impossible à construire et encore moins à analyser même pour des valeurs de k relativement faibles.

Le second problème est le choix de la valeur du paramètre k. Considérons le cas où les séquences analysées ne contiennent pas de dépendances indirectes. Dans cette situation, utiliser un $k > 1$ ne devrait en théorie pas poser de problème d'analyse majeur car les probabilités de transition de la forme $P(.|AD)$ vont correspondre aux transitions $P(.|D)$ (le problème lié à l'explosion combinatoire se pose toujours). Les marches aléatoires ne seront ainsi que peu affectées qu'importe le k choisi. En pratique, cependant, les probabilités de second ordre ne seront pas exactement semblables. La

question est de savoir si ces différences sont significatives ou non. Il peut par ailleurs être en soi intéressant de savoir quel ordre est le plus adapté pour une application donnée.

Afin de répondre à ces enjeux, (Scholtes, 2017) a proposé une méthode pour déterminer l'ordre k le plus adapté. Cette méthode repose sur un test de rapport de vraisemblance comparant le gain de précision obtenu en accroissant l'ordre du modèle à l'accroissement de la taille du réseau. Ce modèle d'ordre fixe remplace donc le paramètre k par un seuil de significativité γ généralement fixé à 10^{-3}.

3.3 Réseaux d'ordre variable

Les deux problèmes inhérents aux modèles FON décrits plus haut sont liés à l'hypothèse qu'il existe un ordre k valable pour n'importe quelle séquence. En dépassant cette hypothèse, les modèles dits à ordre variable, notés VON permettent d'obtenir des modèles plus parcimonieux (*i.e.* nécessitant moins de nœuds-mémoires) sans pour autant sacrifier la fidélité aux données d'entrée.

Dans ces modèles, l'idée principale est de conserver uniquement les nœuds-mémoires qui sont considérés comme statistiquement pertinents. Reprenons l'exemple de la Figure 1, on a $P(.|D) = [0, 0, \frac{1}{3}, 0, \frac{1}{3}, \frac{1}{3}]$. Dans le modèle FON_2 (Fig. 1c), le nœud-mémoire AD a pour distribution d'état suivant $P(.|AD) = [0, 0, 0, 0, 0, 1, 0]$. Autrement dit, le seul état possible après D en venant de A est l'état E. On peut dire que AD est un contexte pertinent par rapport à D, il ajoute de l'information. Au contraire, on a $P(.|BD) = [0, 0, \frac{3}{10}, 0, \frac{3}{10}, \frac{4}{10}]$, ce qui est très proche de $P(.|D)$; la séquence BD ne semble donc pas ajouter beaucoup d'information. En gardant les contextes pertinents, on a abouti sur le réseau d'ordre variable VON illustré en Figure 1d.

Les premiers auteurs à avoir étudié l'application des réseaux d'ordre variable utilisent la divergence de *Kullback-Leibler* D_{KL} afin de quantifier la différence entre distributions :

$$D_{KL}(P||Q) = \sum_{\sigma \in \mathcal{A}} P(x) \log_2 \left(\frac{P(x)}{Q(x)} \right) \tag{2}$$

pour deux distributions P, Q sur l'ensemble $\mathcal{A}$. Cette mesure est proche de 0 quand les deux distributions sont similaires. On a, pour notre exemple, $D_{KL}\left(P(.|AD)||P(.|D)\right) = 1.5849$ et $D_{KL}\left(P(.|BD)||P(.|D)\right) = 0.014$.

(Xu et al., 2016) ont ainsi proposé de comparer cette divergence à un seuil fixé α par l'utilisateur. Ainsi le problème de définition d'un ordre

donné des modèles FON devient maintenant le problème de fixation de ce seuil définissant cette "pertinence statistique".

Par la suite, (Saebi, Xu, Kaplan, Ribeiro, & Chawla, 2020) ont proposé une fonction de seuil dépendant du nombre d'occurrences d'une séquence et de son ordre. Une séquence s' est jugée pertinente par rapport une séquence s suffixe de s' si

$$D_{KL}\left(P(.|s')||P(.|s)\right) > \frac{|s'|}{\log_2(1+c(s'))} \tag{3}$$

Cette fonction de seuil est croissante avec l'ordre mais décroissante avec le nombre d'occurrences de s'. Ainsi, une séquence sera plus facilement jugée pertinente si elle est souvent observée et qu'elle n'est pas trop longue. Pour notre exemple AD est significative car ce seuil vaut $\frac{2}{\log_2(3)} = 1.2618$. En revanche BD ne l'est pas car le seuil vaut $\frac{2}{\log_2(11)} = 0.5781$, ce qui est inférieur à la divergence calculée précédemment. La définition de (Saebi et al., 2020) ne requiert pas de paramètres mais il faut toutefois noter que la définition de la fonction de seuil est *ad hoc*. En effet, on pourrait aussi bien considérer par exemple le double ou la moitié de cette fonction afin d'obtenir des réseaux respectivement plus petit ou plus grand. Les auteurs ont remplacé le paramètre par une règle arbitraire bien que validée par leur expériences.

Quelle que soit la définition de "pertinence statistique" utilisée, l'algorithme de (Xu et al., 2016) (que nous ne détaillerons pas ici) permet une construction des réseaux d'ordre variable efficace. En effet, cette construction se fait de façon récursive en testant les extensions possibles des contextes déjà identifiés comme pertinents, en commençant par les contextes d'ordre 1 (toujours considérés comme pertinents). Il n'est ainsi pas nécessaire de tester toutes les séquences possibles. Au final, les contextes inclus dans VON sont toutes les séquences statistiquement pertinentes par rapport à leurs suffixes également présents dans le modèle.

3.4 Analyse des Réseaux d'ordre supérieur

Nous avons présenté différents modèles de réseaux d'ordre supérieur. Nous détaillons ici leur utilisation dans le cadre d'analyses quantitatives. L'analyse des réseaux utilise en effet régulièrement des mesures ou algorithmes permettant d'extraire des informations sur la structure globale ou sur certaines parties.

Une première observation est que l'ensemble des outils pertinents pour l'analyse de réseaux d'ordre supérieur est limité à ceux qui sont pertinents pour l'analyse du réseau FON_1. Rappelons que ce dernier est un graphe

orienté et pondéré, avec un poids pour l'arc $A \to B$ correspondant soit au nombre de transitions directes observées entre A et B, soit au poids relatif de ces transitions (*i.e.* la probabilité observée d'aller de A à B). Ce graphe est ainsi équivalent au graphe dirigé complet où toutes les transitions non-observées sont ajoutées en tant qu'arc avec un poids de 0. Dans ce cadre, des mesures structurelles comme la densité du graphe ou d'une de ses parties a une portée d'analyse assez limitée car elles indiquent à quel point chaque transition est observée au moins une fois. Ces mesures incluent le degré entrant/sortant ou le *clustering* basé sur des mesures de densité. Dans le cadre des réseaux FON_k ou VON, on peut noter que le nombre d'arcs et la densité peuvent varier de manière importante selon les paramètres de construction. Toutefois, l'intérêt de ces mesures est tout aussi limité que pour le réseau FON_1.

L'intérêt des réseaux d'ordre supérieur apparaît principalement lorsque des mesures reposant sur les relations indirectes (en terme de probabilités de transitions) entre entités sont utilisées. On peut évoquer en premier lieu les mesures de centralité qui permettent de mesurer l'"importance" d'un nœud dans un réseau. Ces mesures incluent notamment le *PageRank* (Brin & Page, 1998) ou la centralité de second ordre (Kermarrec, Le Merrer, Sericola, & Trédan, 2011)[1]. Nous nous focaliserons ici sur le *PageRank*. La notion d'"importance" pour cette mesure est définie informellement de manière récursive : un sommet est important si il est probable d'atteindre ce sommet à partir de sommets importants. Dans le réseau donné en Figure 1b, l'importance de `E` va autant bénéficier de l'importance de `A`, `B` et `C`. L'utilisation de réseaux d'ordre supérieur peut ainsi affecter la centralité. En effet, dans les réseaux FON_2 et VON, il est par exemple impossible d'observer une transition de l'état `C` vers `E`.

Dans les réseaux d'ordre supérieur, un état peut avoir différentes représentations. Si on veut ramener la mesure effectuée sur les réseaux aux états, une transformation s'impose. Cette transformation est assez intuitive dans le cas de *PageRank* : on définit le *PageRank* d'un état comme la somme des valeurs de *PageRank* des représentations de cet état (Xu et al., 2016).

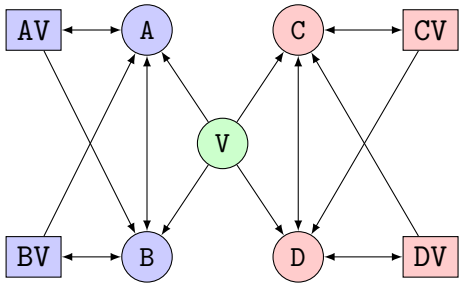
Des analyses récentes (Coquidé, Queiros, & Queyroi, 2022) ont toutefois démontré que l'algorithme de calcul de *PageRank* doit être adapté pour éliminer un biais lié à l'existence de plusieurs nœuds représentant un même état. En effet, la valeur de *PageRank* d'un noeud est équivalente à la part du temps qu'un *"surfeur aléatoire"* passe sur ce noeud. Un surfeur aléatoire diffère d'un marcheur aléatoire car il peut à n'importe quel moment se télé-

1. Notons que le second "ordre" fait ici référence au second moment d'une distribution et non à la longueur d'une séquence.

porter sur n'importe quel noeud avec une probabilité (généralement fixée à 15%). Ce mécanisme de téléportation permet une convergence de l'algorithme vers des valeurs de *PageRank* uniques. Cependant, dans les réseaux d'ordre supérieur, on observe généralement une importante hétérogénéité dans le nombre de représentations par état. Un surfeur aléatoire aura donc plus de chance de se téléporter sur un état représenté par de nombreux nœud-mémoires, ce qui augmentera mécaniquement l'importance de cet état. Ce biais peut être annulé en interdisant la téléportation vers des nœuds d'ordre supérieur à 1 *i.e.* lorsque le surfeur se téléporte, on doit considérer que c'est le début d'une nouvelle séquence.

Séquences	#
A B V A	3
B V A B V B	7
V A V B	5
B A V A V	2
C D V C D	4
V D C V C	3
D V C	5

(a) Données séquentielles



(b) Réseau VON

FIGURE 2 – Exemple d'utilisation des réseaux d'ordre supérieur pour le *clustering*. On remarque dans les séquences 2a qu'en passant par `V` en venant de $\{\texttt{A}, \texttt{B}\}$ ou $\{\texttt{D}, \texttt{F}\}$ on retourne dans le même ensemble. Le réseau VON 2b contient deux composantes fortement connectées pouvant être retrouvées avec un algorithme de partitionnement de graphe (couleurs des noeuds). L'état `V` est ainsi présent dans les deux groupes.

Bien qu'une détection de communauté basée sur des mesures de densité n'est pas adaptée pour les réseaux d'ordre supérieur, d'autres mesures telles que la *Map equation* (Rosvall, Axelsson, & Bergstrom, 2009) peuvent avoir un intérêt. Cette mesure évalue à quel point une marche aléatoire va avoir tendance à rester souvent dans une même communauté. Or, encore une fois, les marches aléatoires réalisées sur les réseaux FON_k et VON représentent mieux les séquences analysées. On peut directement appliquer un algorithme de détection de communauté maximisant la *Map equation* à ces réseaux. Notons que différentes représentations d'un même état peuvent alors se retrouver dans différentes communautés. Ceci est un apport majeur des réseaux d'ordre supérieur. En effet, un partitionnement simple des nœuds-mémoires va correspondre à un *clustering* chevauchant des états (un exemple est donné en Fig. 2). Notons toutefois que, comme pour la mesure

de *PageRank*, l'existence de plusieurs représentations peut biaiser les résultats des algorithmes de détection de communautés. Le développement de méthodes de *clustering* adaptées à ces réseaux est une perspective de recherche future (Queiros, Coquidé, & Queyroi, 2022).

4 Outils logiciels et cas d'étude

Nous allons ici illustrer la construction et l'analyse de réseaux d'ordre supérieur à travers un cas d'étude portant sur des itinéraires de vols commerciaux aux États-Unis.

Notons d'abord que les outils logiciels sont limités et correspondent principalement aux programmes mis à disposition par les auteurs. Ainsi, la librairie `pathpy` [2] permet la génération de réseaux d'ordre optimal (Scholtes, 2017) décrits en Section 3.2 tandis que les réseaux d'ordre variable décrits en Section 3.3 peuvent être construits en utilisant le code des auteurs (Saebi et al., 2020) [3]. Toutefois, il n'y a pas de raisons *a priori* de préférer un modèle aux autres, il faut donc pouvoir comparer les réseaux. Dans ce but, nous proposons une implémentation en Python qui permet de générer les différents modèles décrits dans la section précédente à partir d'un jeu de séquences. Nous analysons ensuite les réseaux en utilisant la bibliothèque `NetworkX` [4] qui est très utilisée en analyse et fouille de réseaux.

Notre code est disponible en ligne [5] et contient un notebook Jupyter utilisé pour réaliser le cas d'étude suivant.

4.1 Données

Pour le cas d'étude, nous allons travailler sur des séquences de voyages domestiques américains de passagers en 2001 (TransStat, 2001). Le jeu de données est un échantillon de 10% des billets d'avion des transporteurs participants. Les variables incluent l'origine, la destination du voyageur mais aussi d'autre détails sur les passagers transportés. Les séquences que nous utilisons sont construites à partir de ces données.

Chacune des 286810 séquences correspond aux aéroports où un passager a fait escale (incluant l'origine et la destination). Notons que ces itinéraires ne sont pas séparés si l'escale correspond en fait à un séjour et incluent ainsi de nombreux aller-retour. Les séquences étudiées peuvent

2. www.pathpy.net
3. www.higherordernetwork.com
4. networkx.org
5. https://gitlab.univ-nantes.fr/queyroi-f/higherordernetworks/-/tree/Tuto

	FON$_1$	FON$_2$	VON
Nombre de noeuds	175	1716	58092
Nombre d'arcs	1598	322024	148975
Ordre maximum	1	2	6
Nombre de représentations moyen	1	9.80	331.95
Nombre de représentations minimum	1	2	1
Nombre de représentations maximum	1	120	13554
Probabilité d'aller-retour	10.90%	30.93%	30.91%

TABLEAU 1 – Statistiques sur les réseaux issus des différents modèles

inclure de 2 à 14 aéroports qui correspondent donc aux "états" étudiés. Il y a en tout 175 aéroports différents présents sur le territoire des États-Unis.

4.2 Construction des modèles et des réseaux

Pour construire les réseaux grâce à notre outil, il faut fixer un certain nombre de paramètres valables pour tous les modèles.
Le premier est le *support minimum* qui écarte les contextes observés moins souvent que la valeur donnée du paramètre. Il permet donc de réduire l'espace de recherche. Dans notre cas, il n'est pas utilisé et est donc fixé à 1.
Le second est l'*ordre maximal* qui fixe une limite à la longueur des contextes (et donc des nœuds-mémoires) présents dans le réseau construit. On obtient ainsi le modèle FON$_k$ en fixant le paramètre à k. Si on cherche l'ordre optimal tel que décrit dans la Section 3.2, alors l'algorithme recherchera entre l'ordre 1 et k. Notons qu'une valeur trop grande de ce paramètre peut mener à des temps de calculs extrêmement longs. Pour la recherche d'ordre optimal on utilisera ainsi un ordre maximal de 4. Dans le cas du modèle VON, nous n'utilisons pas ce paramètre ; l'algorithme va déterminer seul les contextes pertinents sans contraintes sur leur ordre.

Afin de comparer les différents modèles de réseau nous allons commencer par comparer leur topologie, informations présentes dans le Tableau 1. Les réseaux FON$_2$ et VON seront comparés au réseau FON$_1$ qui consiste "simplement" à agréger tous les vols entre paires d'aéroports. L'ordre fixe optimal détecté en utilisant la méthode de (Scholtes, 2017) est de 2. Pour VON, l'ordre maximum observé est 6. Une première différence évidente entre le réseau FON$_2$ et le réseau FON$_1$ est l'augmentation significative du nombre de noeuds. Cette différence est liée à l'ajout de dépendances indirectes sous forme de nœuds dans le réseau. Un nœud de FON$_2$ a en moyenne presque 10 représentations. De même, le réseau VON est aussi bien plus grand que FON$_2$ et FON$_1$. Effectivement, son ordre maximal étant plus élevé, celui-ci inclut un nombre bien plus important de représentations

par nœud, en moyenne 332 par nœud. Toutefois, la plupart des nœuds-mémoires sont d'ordre 3 ou 4.

Un comportement qui est intuitivement présent dans le jeu de données d'entrée sont les aller-retours. Un voyageur qui fait un trajet pour un voyage sans escale fera son trajet depuis son aéroport de départ puis d'arrivée et l'inverse pour rentrer chez lui. On va vérifier si, en ajoutant les dépendances indirectes, on retrouve ce comportement. Pour cela, on va calculer la probabilité d'un marcheur de faire deux sauts dans le réseau et de revenir à son état de départ. En comparant les résultats sur les aller-retours de la Table 1, on remarque un impact non négligeable des ordres supérieurs sur les résultats des probabilités d'aller-retour. Il y a par ailleurs peu de différences entre les probabilités des deux réseaux d'ordre supérieur, cela s'explique par le fait que VON inclut presque tous les nœuds d'ordre 2 possibles. On pourrait aussi prendre en compte des ordres plus élevés pour calculer la probabilité de faire un aller-retour en n étapes, pour prendre en compte les trajets de voyageur avec plus d'escales.

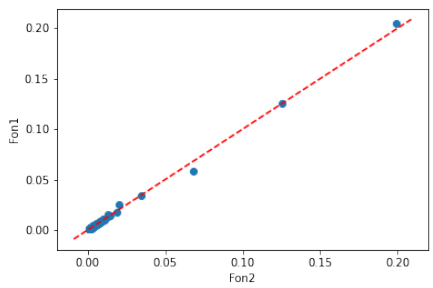
4.3 Comparaison des *PageRank* selon le modèle

Dans cette partie nous allons calculer l'importance des aéroports dans les 3 types de réseaux créés en utilisant l'algorithme *PageRank* de *NetworkX*. Notons que, malgré les différences topologiques entre les réseaux, il est possible d'obtenir des résultats similaires.

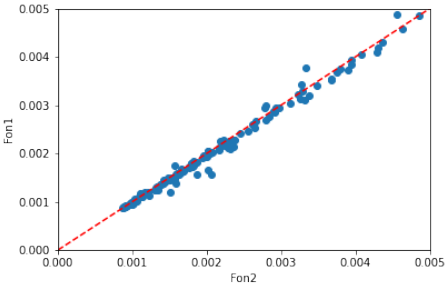
L'algorithme prend deux paramètres principaux ; la valeur α (usuellement 0.85), qui correspond à la probabilité d'un marcheur de s'arrêter et de se téléporter sur n'importe quel nœud du réseau. L'algorithme prend aussi comme paramètre le vecteur *"personalization"*. Il nous permet, comme évoqué précédemment, d'interdire la téléportation vers des nœuds d'ordre supérieur afin d'annuler le biais.

La Figure 3 représente sous forme de nuage de points les corrélations entre les valeurs de *PageRank* pour chacun des aéroports et ceux pour les trois différents modèles FON_1, FON_2, VON. Pour chacun des trois sous-figures 3a-3c-3f, les points se rapprochent de la droite indiquant que les valeurs de *PageRank* sont plutôt corrélées, malgré les différences entre les modèles, qui n'ont pas du tout les mêmes dimensions et topologies. Pour tous les réseaux, on observe une hiérarchie similaire avec des écarts du même ordre. Cependant, en regardant de plus près les valeurs plus petites des Figures 3b-3d-3f, on remarque tout de même des différences, notamment entre FON_1 et VON.

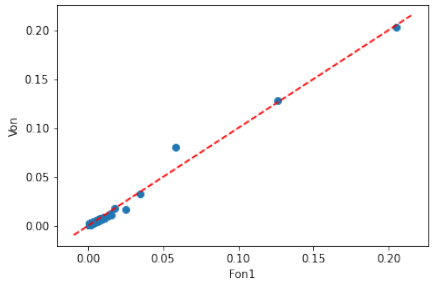
Afin de mieux se rendre compte des différences et similitudes entre les



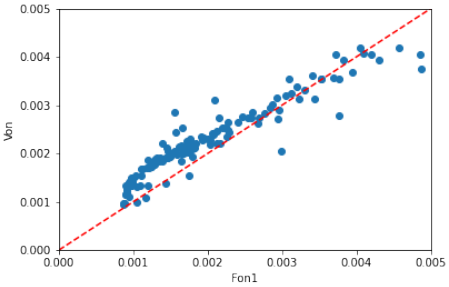
(a) Comparaison entre FON_1 et FON_2 ...



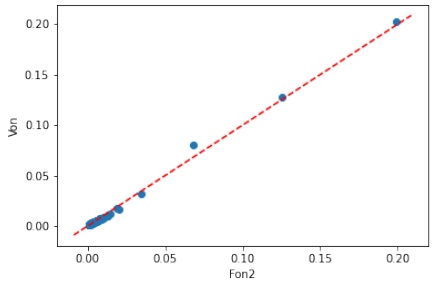
(b) ... pour les petites valeurs.



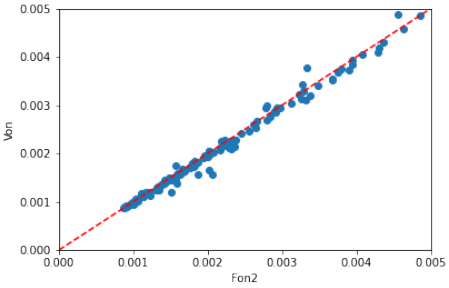
(c) Comparaison entre FON_1 et VON ...



(d) ... pour les petites valeurs.



(e) Comparaison entre FON_2 et VON ...



(f) ... pour les petites valeurs.

FIGURE 3 – Comparaison entre les valeurs de *PageRank* pour les modèles présentés. Chaque point représente un aéroport.

résultats de *PageRank*, on va regarder leur impact en terme de classement des états. Les classements de rang de *PageRank* pour les 20 aéroports les plus importants des trois modèles sont représentés dans la Figure 4.
Les rangs sont sensiblement identiques entre les trois modèles, à quelques exceptions près, notamment *HNL Honolulu* qui fait partie du top 20 dans les modèles VON et FON_1 mais pas dans FON_2. De plus, même en prenant en compte des dépendances d'ordre supérieur, celles-ci n'ont finalement que très peu de conséquences sur les rangs *PageRank* des aéroports. Les réseaux FON_1 et FON_2 restent les plus similaires ; cela suggère que les différences

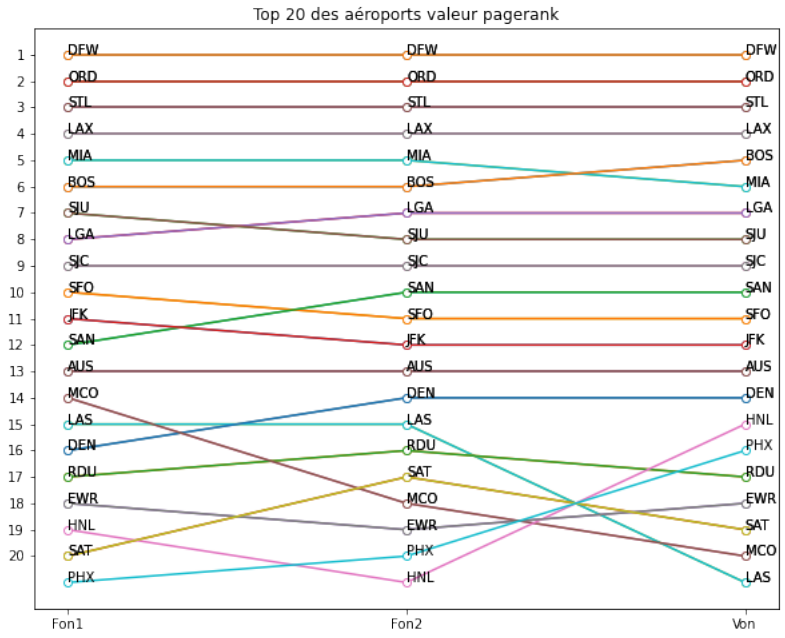


FIGURE 4 – Évolution des rangs pour les aéroports classés dans le TOP 20 d'au moins un réseau. *Lecture : L'aéroport de Phoenix (PHX) a le 20ème plus grand PageRank dans le réseau* FON_2 *et le 16ème dans le réseau* VON*, il est toutefois hors du TOP 20 avec le réseau d'ordre* 1, FON_1.

de construction impactent la structure du réseau et par conséquent les résultats des algorithmes de fouille.

5 Conclusion

Après avoir présenté plusieurs façons de construire des réseaux d'ordre supérieur, il est important de s'interroger sur la question suivante : peut-on obtenir le "meilleur" modèle pour nos données ? Un critère usuel est l'application du *rasoir d'Ockham* : entre deux modèles qui représentent aussi bien les données, on va préférer le plus parcimonieux. Par ailleurs, les modèles possibles pour les réseaux d'ordre supérieur ne se limitent pas à ceux présentés dans cet article. Par exemple, il est tout à fait possible de remettre en question le seuil utilisé par (Saebi et al., 2020) et décrit en Section 3.3. On peut également envisager de réduire le nombre de noeuds-mémoires en agrégeant les noeuds se comportant de façon très similaire (Queiros et

al., 2022). Ainsi, il n'existe pas un unique réseau d'ordre supérieur pour un même jeu de données. Nous suggérons, à l'instar de la méthodologie présentée dans la Section 4, de comparer les résultats obtenus avec différents modèles ou tester des hypothèses *a priori* sur les dépendances séquentielles pouvant être formulées sur le système étudié.

Comme expliqué précédemment, les réseaux d'ordre supérieur peuvent être fouillés comme n'importe quel réseau "classique". Cependant, des travaux récents ont soulevé des problèmes liés à l'ajout de représentations d'un même état. L'hétérogénéité des nombres de représentations peut mécaniquement mener à des biais dans les résultats d'algorithmes de fouille notamment ceux basés sur des marches aléatoires. Des travaux de recherche supplémentaires pour adapter les principaux outils utilisés en analyse de réseaux semblent ainsi nécessaires. Dans ce cadre, une plateforme logicielle comme celle que nous proposons permettra de diffuser et tester ces outils. Toutefois, nous allons, à l'avenir, travailler à l'intégration de nos implémentations dans des bibliothèques populaires existantes.

Références

Battiston, F., Cencetti, G., Iacopini, I., Latora, V., Lucas, M., Patania, A., … Petri, G. (2020). *Networks beyond pairwise interactions : Structure and dynamics* (Vol. 874) (N° 0). Consulté sur https://www.sciencedirect.com/science/article/pii/S0370157320302489 (Networks beyond pairwise interactions : Structure and dynamics) doi: https://doi.org/10.1016/j.physrep.2020.05.004

Brin, S., & Page, L. (1998, avril). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *30*(1), 107–117. Consulté le 2021-03-30, sur https://www.sciencedirect.com/science/article/pii/S016975529800110X doi: 10.1016/S0169-7552(98)00110-X

Chierichetti, F., Kumar, R., Raghavan, P., & Sarlos, T. (2012). Are web users really markovian ? In *Proceedings of the 21st international conference on world wide web* (p. 609–618). New York, NY, USA : Association for Computing Machinery. Consulté sur https://doi.org/10.1145/2187836.2187919 doi: 10.1145/2187836.2187919

Coquidé, C., Queiros, J., & Queyroi, F. (2022). Pagerank computation for higher-order networks. In R. M. Benito, C. Cherifi, H. Cherifi, E. Moro, L. M. Rocha, & M. Sales-Pardo (Eds.), *Complex networks & their applications x* (pp. 183–193). Cham : Springer International Publishing.

Drake, J. M., & Lodge, D. M. (2004). Global hot spots of biological invasions : evaluating options for ballast–water management. *Proceedings*

of the Royal Society of London. Series B : Biological Sciences, *271*(1539), 575–580.

Ducruet, C., & Notteboom, T. (2012). The worldwide maritime network of container shipping : spatial structure and regional dynamics. *Global networks*, *12*(3), 395–423.

Eliassi-Rad, T., Latora, V., Rosvall, M., & Scholtes, I. (2021). *Higher-Order Graph Models : From Theoretical Foundations to Machine Learning (Dagstuhl Seminar 21352)* (Vol. 11) (N° 7). Dagstuhl, Germany : Schloss Dagstuhl – Leibniz-Zentrum für Informatik. Consulté sur https://drops.dagstuhl.de/opus/volltexte/2021/15592 doi: 10.4230/DagRep.11.7.139

Kaluza, P., Kölzsch, A., Gastner, M. T., & Blasius, B. (2010). The complex network of global cargo ship movements. *Journal of the Royal Society Interface*, *7*(48), 1093–1103.

Kermarrec, A.-M., Le Merrer, E., Sericola, B., & Trédan, G. (2011). Second order centrality : Distributed assessment of nodes criticity in complex networks. *Computer Communications*, *34*(5), 619–628.

Queiros, J., Coquidé, C., & Queyroi, F. (2022). Toward Random Walk Based Clustering of Variable-Order Networks. *Network Science*. Consulté sur https://hal.science/hal-03863570

Rivière, J., Madoré, F., Batardy, C., Garat, I., & Raimbault, N. (2021). Les divisions socioprofessionnelles en mouvement d'une métropole attractive. le cas de l'aire urbaine de nantes (1975-2015). *Cybergeo : European Journal of Geography*.

Robette, N. (2011). *Explorer et décrire les parcours de vie : les typologies de trajectoires*. CEPED.

Rosvall, M., Axelsson, D., & Bergstrom, C. T. (2009, novembre). The map equation. *The European Physical Journal Special Topics*, *178*(1), 13–23. Consulté le 2021-03-30, sur https://doi.org/10.1140/epjst/e2010-01179-1 doi: 10.1140/epjst/e2010-01179-1

Rosvall, M., Esquivel, A. V., Lancichinetti, A., West, J. D., & Lambiotte, R. (2014). Memory in network flows and its effects on spreading dynamics and community detection. *Nature communications*, *5*(1), 1–13. doi: 10.1038/ncomms5630

Saebi, M., Xu, J., Kaplan, L. M., Ribeiro, B., & Chawla, N. V. (2020). Efficient modeling of higher-order dependencies in networks : from algorithm to application for anomaly detection. *EPJ Data Sci.*, *9*(1), 15. doi: 10.1140/epjds/s13688-020-00233-y

Scholtes, I. (2017). When is a network a network? multi-order graphical model selection in pathways and temporal networks. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (p. 1037–1046). New York, NY, USA : Association for Computing Machinery. doi: 10.1145/3097983.3098145

Torres, L., Blevins, A. S., Bassett, D., & Eliassi-Rad, T. (2021). The why,

how, and when of representations for complex systems. *SIAM Review*, *63*(3), 435–485.

TransStat, R. (2001). *Origin and destination survey database. db1b.* https://www.transtats.bts.gov/.

Xu, J., Wickramarathne, T. L., & Chawla, N. V. (2016). Representing higher-order dependencies in networks. *Science Advances*, *2*(5), e1600028. doi: 10.1126/sciadv.1600028