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

Julie Queiros, François Queyroi

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 *

14 mars 2024

Type de Soumission

Software paper

Titre Anglais

Methods and Tools for the Construction of Higher-Order Networks from Sequences.

Résumé

Les réseaux d'ordre séquentiel supérieur sont une classe de réseaux qui intègre des « nœuds-mémoires » afin de prendre en compte les interactions pouvant exister dans des données séquentielles (traces), 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 (*graph mining*) qui peuvent être employées. Nous présenterons le paquet *Python* `honyx` [1], qui rassemble des algorithmes présents dans la littérature. Nous proposons un didacticiel sur son utilisation à travers un cas d'étude portant sur les itinéraires de vols commerciaux aux États-Unis. Nous abordons également certains des défis et des orientations futures dans ce domaine.

Mots-clés

Séquences, Traces, Réseaux d'Ordre Supérieur, PageRank, Python

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

1. https://pypi.org/project/honyx/

Abstract

Sequential Higher-order networks are a class of graphs that incorporate "memory nodes" in order to take into account the indirect interactions that can exist in sequential data. They differ from so-called "order 1" networks, which only take direct relationships into account. In this article, we provide an overview of this concept, detailing their construction and the mining techniques that can be employed. We present the Python package `honyx`, which contains algorithms already available in the literature. We propose a tutorial on its use through a case study of commercial flight itineraries in the United States. We also discuss some of the challenges and future directions in the field.

Keywords

Sequences, Trajectories, Higher-order Networks, PageRank, Python

1 Introduction

Les graphes et la fouille de graphes (*graph mining*) sont des outils fondamentaux pour la compréhension des réseaux, systèmes complexes composés d'interactions entre un grand nombre d'entités. Rencontrés dans divers domaines, tels que la physique, la biologie et les sciences sociales, les graphes modélisant ces réseaux sont caractérisés par leurs propriétés topologiques non triviales, telles que la présence de nœuds hautement connectés, la présence d'une structure communautaire ou l'existence de corrélations entre les nœuds. La compréhension de la structure et de l'évolution des réseaux est cruciale pour de nombreuses applications, comme la propagation des maladies ou la diffusion de l'information. L'étude des réseaux complexes est un domaine interdisciplinaire qui s'appuie sur des idées et des méthodes issues de nombreux autres domaines tels que la physique, les mathématiques, l'informatique et les sciences sociales. La quantité de données disponibles ne cessant d'augmenter, la capacité à comprendre et analyser les réseaux devient de plus en plus importante pour un large éventail d'applications.

Dans certains cas, les données sont séquentielles, c'est-à-dire qu'elles représentent une suite de changements d'états (on parlera aussi de «séquences» ou de «traces»). L'analyse de réseau peut être utilisée pour étudier les relations entre ces états à partir de ces transitions. Des exemples courants sont les trajets maritimes (séquences de ports d'escale d'un navire) ou bien les traces d'utilisateurs sur Internet (séquences des pages Web visitées). Comme illustration, nous prendrons les séquences artificielles données dans la figure 1a.

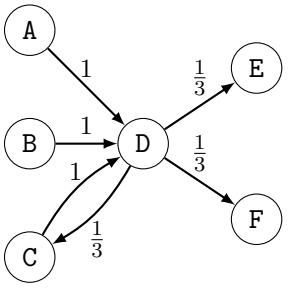
Ces données séquentielles peuvent être représentées par des réseaux[2] dits de *premier ordre*, ou bien *sans-mémoire*, qui seront notés FON_1 (pour *fixed-order network*), modélisant les transitions entre les états $A, B, \ldots, F$. Les réseaux FON_1 sont construits en agrégeant les occurrences entre paires d'états dans le jeu de données. 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 la sous-séquence D (les définitions précises sont données dans la section 3.1). Il y a 6 occurrences de DE pour un total de 18 séquences avec D. Dans ce réseau, on remarque que, partant de D, la probabilité de passer au nœud C, E ou F est la même. Cependant, dans les séquences, il existe des relations indirectes. Par exemple, en suivant uniquement le réseau FON_1, il serait

2. Bien que la distinction entre le système modélisé (réseau) et l'objet mathématique (graphe) est importante, nous utiliserons ici «réseau» de façon interchangeable.

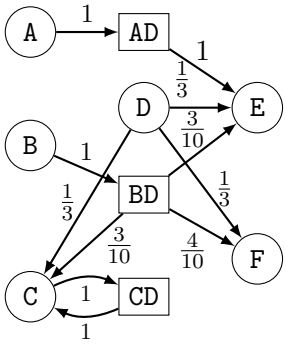
possible d'avoir une séquence CDE, en suivant les relations directes entre C et D puis D et E. Or, cette séquence n'apparaît pas dans le jeu de données de base. En ne prenant en compte que les relations d'ordre 1, ou directes, les relations indirectes sont perdues et les réseaux sont moins fidèles aux comportements des données séquentielles.

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

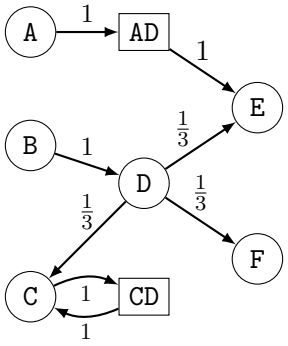
(a) Tableau de comptage des séquences $\mathcal{S}$



(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 dénombrées en 1a. Pour un ensemble d'états possibles $\mathcal{A} = \{A, B, C, D, E, F\}$, nous observons, par exemple, la séquence BDF quatre fois. Le réseau d'ordre 1 (fig. 1b) est obtenu en prenant les probabilités de transitions directes. Dans l'état D, il y a une chance sur trois d'aboutir à l'état E. Les réseaux d'ordre séquentiel supérieur FON$_2$ et VON (fig. 1c et 1d) incluent des dépendances indirectes sous la forme de nœuds-mémoire (représentés sous la forme de rectangle). L'état D a ainsi 4 et 3 représentations respectivement.

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 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 les réseaux cités plus haut, les interactions entre les entités 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 (algorithmes de centralité ou de *clustering*) (Xu, Wickramarathne, & Chawla, 2016 ; Rosvall et al., 2014). Nous illustrerons cette différence par un cas d'étude basé sur des données réelles.

Pour représenter ces interactions d'ordre supérieur, les chercheurs ont développé le concept de réseau d'ordre séquentiel supérieur (Torres, Blevins, Bassett, & Eliassi-Rad, 2021). Ces réseaux d'ordre séquentiel supérieur ont pour particularité d'inclure les relations indirectes. Les figures 1c et 1d sont respectivement des exemples de réseau d'ordre fixe et d'ordre variable. Des «nœuds-mémoires» encodant les dépendances indirectes sont inclus au réseau. 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». Le réseau en figure 1c inclut toutes les sous-séquences (suivies d'un autre symbole) d'ordre inférieur ou égale à 2 présentes dans les séquences d'entrée. Le réseau de la figure 1d permet de faire un choix entre les sous-séquences. Cette différence sera expliquée en détail dans la section 3. 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.

Malgré les nœuds-mémoire ajoutés, les réseaux d'ordre séquentiel supérieur restent des graphes de même nature que les réseaux d'ordre 1 *i.e.* en omettant les labels, les graphes des figures 1c et 1d sont des graphes dirigés et pondérés «classiques». Des algorithmes de fouille (*graph mining algorithms*) pourraient ainsi y être directement appliqués (Xu et al., 2016), même si, comme nous allons le voir, certaines adaptations sont nécessaires.

Il existe de multiples façons de construire un réseau d'ordre séquentiel 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 (cette notion sera discutée dans la section 3.1). 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 ou simplement permettre l'utilisation finale de ce genre de réseau.

Nous ferons dans cet article un état de l'art sur les méthodes existantes pour construire et analyser les réseaux d'ordre séquentiel supérieur. Dans un premier temps, nous discuterons de l'utilisation de la terminologie « ordre supérieur » et de la place qu'occupent les réseaux d'ordre supérieur dans la littérature. Nous proposerons ensuite des définitions formelles des réseaux à ordre fixe, FON$_k$ et des réseaux à ordre variable, VON. Nous présenterons le paquet Python `honyx` permettant de comparer différents modèles. Afin d'illustrer l'intérêt de notre outil, nous utilisons un cas d'étude sur les itinéraires de vol de passagers aux États-Unis. Une tâche est notamment de savoir si la prise en compte des dépendances indirectes affectent le résultat de l'algorithme de centralité *PageRank*. Nous finirons par une discussion sur les limitations et les enjeux de l'étude de ce genre de réseaux.

2 Contexte et utilisations

Dans cette section, nous préciserons la terminologie d'ordre supérieur. Celle-ci peut recouvrir plusieurs concepts existant en analyse des réseaux. Nous discutons ensuite des domaines d'applications de ces objets issus de la littérature.

La notion d'«ordre supérieur» est employée pour désigner des concepts différents. Eliassi-Rad, Latora, Rosvall, et Scholtes (2021) définissent les «réseaux d'ordre supérieur» comme tous les réseaux conçus pour capturer davantage que les relations *dyadiques* (entre deux entités). Les relations pouvant inclure plus de deux entités que sont les relations de co-autorat (Battiston et al., 2020) sont un exemple notable. Dans ce cadre, la transformation de ces relations en graphe simple représente une perte d'information. Par exemple, un article de trois auteur·rices ou trois articles entre chaque paire d'auteur·rices ajouteront dans les deux cas une 3-clique au graphe. Un *hypergraphe* peut être utilisé pour encoder ces relations sans perte d'information. Les dépendances séquentielles sont un autre exemple de relations dépassant les relations dyadiques. Afin d'éviter les confusions entre ces concepts, nous utilisons le terme «réseaux d'ordre séquentiel supérieur» pour parler des réseaux construits à partir de séquences. Dans le cas de modèles plus spécifiques tels que les réseaux d'ordre fixe ou d'ordre variable définis dans la section 3, nous n'emploierons pas l'adjectif «séquentiel».

Les champs d'application des réseaux d'ordre séquentiel supérieur sont larges. 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 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 non considérées. 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 conversations sur le même sujet.

Enfin, l'utilisation de données séquentielles dans les réseaux d'ordre supérieurs doit être distinguée de l'« analyse de séquences » (*sequence analysis*). Dans cette dernière, les individus sont définis par une séquence d'états. Une application possible consiste à utiliser des mesures de distances entre ces séquences (Studer & Ritschard, 2015) 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 encore des «parcours de vie» (Robette, 2011).

L'analyse de séquences incorpore des dimensions temporelles, à savoir la durée passée par l'individu dans un état donné ou l'instant du passage dans cet état (avec une ligne temporelle souvent discrétisée). Dans un ouvrage sur le présent et le futur de l'analyse de séquences, Cornwell (2018) présente la construction et l'analyse de *Sequence-Networks*. Dans cette construction, les nœuds du réseau correspondent à un couple (t, s) où t est une date (sur une trame temporelle discrète) et s est un état. Les liens correspondent à des paires $(t, s) \rightarrow (t + 1, s')$ comptant le nombre d'individus étant passés de l'état s à l'état s' entre t et $t + 1$. L'auteur suggère que des propriétés de ce graphe temporel (puisque les nœuds sont associés à un temps ou une période donnée) peuvent être pertinentes «en complément» de l'analyse de séquence. Cette construction doit donc être distinguée des réseaux d'ordre supérieur étudiés ici car, premièrement, elle inclut la dimension temporelle et, deuxièmement, seules les transitions directes sont

comptabilisées. Cependant, Cornwell suggère qu'inclure des «transitions d'ordre supérieur» (*i.e.* entre t et $t+2$) pourrait être pertinent. Ainsi, malgré la différence des dimensions prises en compte dans les séquences, nous pensons qu'un rapprochement entre l'analyse de réseaux d'ordre séquentiel supérieur et l'analyse de séquences est une perspective prometteuse.

3 Modèles de réseaux d'ordre séquentiel supérieur

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

3.1 Définitions

Nous donnons ici des définitions préliminaires qui permettront une meilleure compréhension du sujet.

Définition 1 (Séquences)**.** *Pour un ensemble d'états donné $\mathcal{A}$, une séquence est une suite finie d'états de $\mathcal{A}$ i.e. $s = \sigma_1\sigma_2\dots\sigma_m$.*
Un jeu de données *est un multi-ensemble de séquences $\mathcal{S}$ sur les états de $\mathcal{A}$ (une même séquence peut être observée plusieurs fois).*
*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 la sous-séquence s dans $\mathcal{S}$.*

Dans l'exemple donné en figure 1a, on a $\mathcal{A} = \{A, B, C, D, E, F\}$ et $\mathcal{S} = \{ADE, ADE, DE, BDE, BDE, BDE, DF, DF, BDF, BDF, BDF, BDF, BDC, BDC, BDC, CDC, CDC, CDC\}$ pour la séquence $s = BD$ d'ordre 2, on a ainsi $c(BD) = 10$ car la sous-séquence BD apparaît bien 10 fois (notons qu'elle ne correspond pas forcément à une séquence complète).

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'.*

Par exemple, la séquence BDF admet BDF, BD ou B comme préfixes et BDF, 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. Les modèles étudiés vont exclusivement utiliser l'information séquentielle et aucun autre prédicat. Ces modèles séquentiels 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 de probabilités des états possibles après la séquence s. Le point '.' correspond à un état quelconque.*

La probabilité de transition donnée dans l'équation 1 correspond aux nombre d'occurrences de $s^*\sigma$ sur le nombre d'occurrences de s^* observées dans $\mathcal{S}$ suivie par d'autres éléments (*i.e.* on ignore les fois où s^* apparaît à la fin d'une séquence).
Si le modèle noté M_1 n'inclut comme contextes que les séquences d'ordre 1 (*i.e.* les états de $\mathcal{A}$) alors la probabilité $P_{M_1}(C|CD)$ d'aller en C sachant qu'on a visité C puis D est donnée par $P_{M_1}(C|CD) = P_{M_1}(C|D) = \frac{6}{18} = \frac{1}{3}$. Ce modèle ne va pas utiliser d'informations précédentes hormis le dernier état visité. Notons les distributions de probabilité obtenues

$$P_{M_1}(.|CD) = P_{M_1}(.|D) = \left[0, 0, \frac{1}{3}, 0, \frac{1}{3}, \frac{1}{3}\right]$$

i.e. si $\mathcal{A} = \{A, B, C, D, E, F\}$ alors les seuls états possibles après D sont C, E et F avec la même probabilité.

Comme pour n'importe quel modèle statistique, on peut évaluer la *précision* de M *i.e.* sa capacité à correctement prédire des séquences. Il est également possible de parler de fidélité aux données. Ainsi, le modèle M_1 aura 2 chances sur 3 de se tromper s'il lui est demandé de prédire l'état suivant CD. La précision peut être calculée en utilisant différentes méthodes d'échantillonnage courantes en statistiques et en apprentissage (*machine learning*) *e.g.* séparer le jeu de données en deux parties ; une partie utilisée pour construire M et une autre pour calculer la précision de M. Cette précision sera équivalente à la capacité d'une marche aléatoire sur un réseau construit à partir de M (voir définition ci-dessous) à bien simuler le jeu de données séquentielles modélisé.

Définition 4 (Réseau d'ordre séquentiel supérieur)**.** *Étant donné un modèle séquentiel* M*, le réseau d'ordre séquentiel 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*. Par exemple, si $s = abc$ et $s^*\sigma = bc\sigma$ sont des extensions pertinentes de la séquence c et σ respectivement alors le graphe contiendra l'arc $s \to s^*\sigma$ si $abc\sigma \notin M$ et $P^M(\sigma|s) > 0$.*

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 séquentiel 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

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 transition 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-mémoire d'ordres différents. 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[3]. 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 contextes 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 quel que soit le k choisi. En pratique, cependant, les probabilités de second ordre ne seront pas exactement semblables du fait de l'échantillonnage. La question est de savoir si ces différences sont significatives ou non. Il peut également être intéressant de savoir quel ordre est le plus adapté pour une application donnée.

Afin de répondre à ces enjeux, Scholtes (2017) propose une méthode pour déterminer l'ordre le plus adapté par rapport au jeu de données utilisé. Cette méthode repose sur un test statistique de rapport de vraisemblance. Ce rapport correspond au gain de précision (tel que discuté dans la section 3.1) obtenu en considérant FON_{k+1} au lieu de FON_k. En faisant l'hypothèse que ce changement n'augmente pas la précision, ce rapport suit la loi du χ^2 ayant pour degré de liberté le nombre de probabilités de transitions supplémentaires à définir. La p-valeur (*i.e.* la probabilité, sous cette hypothèse, d'observer un gain de précision au moins aussi grand) est calculée et comparée à un seuil de significativité γ (généralement fixé à 10^{-3}). En partant de $k = 1$, la méthode consiste donc à tester chaque accroissement de l'ordre. L'ordre optimal k^* est soit 1 soit le dernier ordre pour lequel la p-valeur est inférieure à γ (*i.e.* pour lequel l'hypothèse que la précision n'augmente pas est rejetée).

3. Les sous-séquences qui ne sont pas suivies d'un état (*i.e.* qui apparaissent systématiquement en fin d'une ou plusieurs séquences) seront ignorées car elles n'apportent rien au modèle et correspondraient à des puits (nœuds sans arcs sortants) dans le graphe.

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. Les modèles dits à ordre variable et notés VON (pour *Variable-Order Network*), qui ne reposent pas sur cette hypothèse, 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), on a $P(.|AD) = [0, 0, 0, 0, 0, 1, 0]$. 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 car 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. Ne conserver que les contextes pertinents conduit à obtenir le réseau d'ordre variable illustré en figure 1d.

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

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

pour deux distributions P, Q sur l'ensemble $\mathcal{A}$. Cette mesure, très courante en statistiques et en théorie de l'information, indique quelle quantité supplémentaire d'information (en *bits*) est obtenue lorsque P est utilisé alors que Q est déjà connu. Elle est ainsi proche de 0 quand les deux distributions sont similaires. Nous avons, 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 proposé de comparer cette divergence à un seuil α fixé 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, et Chawla (2020) ont proposé une fonction de seuil dépendant du nombre d'occurrences d'une séquence et de son ordre. Un 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 l'ordre de grandeur du nombre d'occurrences de s' (l'ajout de 1 permet d'éviter un dénominateur nul dans les cas où $c(s') = 1$). Ainsi, une séquence sera plus facilement jugée pertinente si elle est souvent observée et qu'elle n'est pas trop longue. Dans notre exemple, la séquence 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, il serait également possible de considérer, par exemple, le double ou la moitié de cette fonction afin d'obtenir des réseaux respectivement plus petits ou plus grands. Les auteurs ont remplacé le paramètre α par une règle arbitraire bien que validée par des expériences comparant leur modèle à des modèles d'ordre fixe.

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 efficace des réseaux d'ordre variable. En effet, cette construction se fait de façon récursive en rallongeant les contextes (*i.e.* en rajoutant des états précédents) aux 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. Les réseaux VON doivent également inclure les préfixes des contextes jugés comme pertinents. En effet, si nous souhaitons ajouter le contexte ABC au modèle alors le réseau final doit nécessairement intégrer le nœud AB. Dans le cas contraire, il n'est pas possible d'arriver à l'événement «passer par A puis B pour arriver à C» sans disposer de «passer par A pour arriver à B». Dans ce cas, le contexte AB peut être pertinent ou non. Remarquons que cette situation n'apparaît pas dans les réseaux FON$_k$.

3.4 Analyse des réseaux d'ordre séquentiel supérieur

Nous avons présenté différents modèles de réseaux d'ordre séquentiel supérieur. Nous détaillons ici leur utilisation. 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 du réseau.

Une première observation est que des mesures structurelles se basant sur le degré entrant/sortant, les plus-courts-chemins ou la densité du graphe (ou d'une de ses parties) ont une pertinence limitée. En effet, les réseaux présentés ici sont stochastiques ; le poids pour l'arc $A \to B$ correspond à la

probabilité observée d'aller de A à B. Une mesure telle que le degré indique à quel point chaque transition est observée au moins une fois. Les probabilités de transition sont donc totalement ignorées, un arc de poids proche de 0 comptant autant qu'un avec un poids proche de 1. Cette observation vaut déjà pour FON_1, qui n'est pas un réseau d'ordre supérieur. Dans le cadre des réseaux FON_k ou VON, 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 séquentiel supérieur apparaît principalement lorsque des mesures reposant sur les relations indirectes (en terme de probabilités de transitions) entre entités sont utilisées, 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)[4]. Notons que des mesures comme la centralité d'intermédiarité (*betweenness centrality*) (Barthélémy, 2004) n'est pas applicable car elle se base sur l'énumération de plus-courts-chemins qui n'ont pas de sens dans un réseau stochastique. L'intermédiarité de M. J. Newman (2005), basée sur des marches aléatoires, pourrait toutefois présenter un intérêt. Cette alternative ne sera pas traitée ici.

Nous nous focaliserons ici sur le *PageRank* qui a été davantage étudié dans le contexte des réseaux d'ordre séquentiel supérieur. 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 séquentiel 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 nous voulons ramener la mesure effectuée sur les réseaux aux états, une transformation s'impose. Cette transformation est assez intuitive dans le cas de *PageRank* : le *PageRank* d'un état est défini comme la somme des valeurs de *PageRank* des représentations de cet état (Xu et al., 2016).

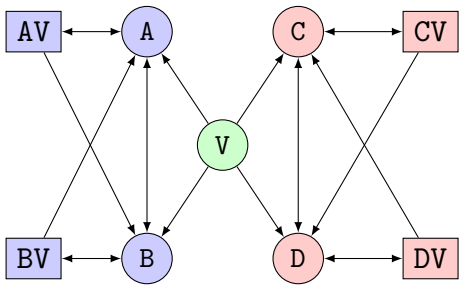
Des analyses récentes (Coquidé, Queiros, & Queyroi, 2021) ont toutefois démontré que l'algorithme de calcul de *PageRank* doit être adapté pour éliminer un biais lié à l'existence de plusieurs nœuds représentant un même

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

état. En effet, la valeur de *PageRank* d'un nœud est équivalente à la part du temps qu'un «*surfeur aléatoire*» passe sur ce nœud. Un surfeur aléatoire diffère d'un marcheur aléatoire car il peut à n'importe quel moment se téléporter sur n'importe quel nœud 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 séquentiel supérieur, il existe 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 nœuds-mémoire, 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. Ainsi, lorsque le surfeur se téléporte, il est considéré comme commençant 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) Tableau de comptage de séquences $\mathcal{S}'$



(b) Réseau VON

FIGURE 2 – Exemple d'utilisation des réseaux d'ordre séquentiel supérieur pour le *clustering*. Dans les séquences $\mathcal{S}'$ (Tableau 2a), en passant par `V` en venant de $\{\texttt{A}, \texttt{B}\}$ ou $\{\texttt{C}, \texttt{D}\}$, il est uniquement possible de retourner dans le même ensemble. Le réseau VON (Figure 2b) contient deux composantes fortement connexes pouvant être retrouvées avec un algorithme de partitionnement de graphe (couleurs des noeuds). L'état `V` est ainsi présent dans les deux groupes.

Les réseaux d'ordre séquentiel supérieur peuvent également être utilisés pour la détection de communautés. Comme expliqué au début de cette section, une détection de communautés basée sur des mesures de densité n'est pas adaptée pour les réseaux d'ordre séquentiel supérieur *i.e.* des mesures se basant sur la présence ou l'absence de liens telles que la modularité (M. E. J. Newman, 2006). 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 aura tendance à évoluer dans une même communauté. Or, encore une fois, les marches

aléatoires réalisées sur les réseaux FON$_k$ et VON représentent mieux les séquences analysées. Il est possible de 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 séquentiel 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 Cas d'études avec le paquet `honyx`

Nous allons ici illustrer la construction et l'analyse de réseaux d'ordre séquentiel supérieur à travers un cas d'étude portant sur des itinéraires de vols commerciaux aux États-Unis. Les échanges aériens constituent un aspect important des flux entre villes (Derudder, Witlox, Faulconbridge, & Beaverstock, 2008). Une problématique est, dans ce contexte, de connaître la «hiérarchie» des villes émergeant des réseaux formés par ces échanges (Cattan, 2004). Nous pouvons suspecter, dans le cas des itinéraires de vols, l'existence de dépendances séquentielles telles que discutées dans cet article (notamment des aller-retours). Nous allons tenter de les détecter avec les modèles décrits précédemment et d'évaluer l'effet du choix du modèle sur la mesure *PageRank*. Les deux questions importantes auxquelles nous allons tenter de répondre sont donc :

— L'hypothèse de Markov (*i.e.* il n'y a pas de dépendances indirectes au-delà de l'ordre 1) est-elle réaliste ?
— Si oui, l'abandon de cette hypothèse conduit-elle à des résultats de centralité différents ?

Ces deux questions peuvent se poser dans d'autres contextes où les réseaux sont créés à partir des séquences d'événements.

Précisons toutefois que ce cas d'étude, limité aux seuls États-Unis et à une unique dimension des flux entre villes (les itinéraires), a un but essentiellement démonstratif. Des analyses et une réflexion plus poussées sont nécessaires pour tirer des conclusions sur ce sujet.

Les outils logiciels permettant la création et l'exploitation de réseaux d'ordre séquentiel supérieur sont limités et correspondent principalement

aux programmes mis à disposition par les auteurs des articles précédemment discutés. Ainsi, la librairie `pathpy`[5] 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)[6]. 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 librairie Python nommé `honyx`[7] permettant de générer les différents modèles décrits dans la section précédente à partir d'un jeu de séquences. Les graphes ainsi générés suivent le format `NetworkX`, qui est une librairie très utilisée en analyse et fouille de réseaux. Le but de `honyx` est aussi de fournir des outils d'analyse de ces objets. Le jeu de données et un notebook Jupyter permettant de reproduire l'étude suivante sont disponibles en ligne (Queiros & Queyroi, 2024).

4.1 Données

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'autres détails sur les passagers transportés. Les séquences que nous utilisons sont construites à partir de ces données (Scholtes, 2017).

Chacune des $286\,810$ 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-retours. Les séquences étudiées peuvent inclure de 2 à 14 aéroports qui correspondent donc aux «états» étudiés. Il y a en tout 175 aéroports présents sur le territoire des États-Unis.

4.2 Construction des réseaux

Pour construire les réseaux grâce à `honyx`, 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 de réduire l'espace de recherche. Dans notre cas, il n'est pas utilisé et est 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. Le modèle FON_k est obtenu en fixant le paramètre à k. Pour tenter de déterminer l'ordre optimal tel que décrit dans la Section 3.2, l'algorithme recherche celui-ci entre l'ordre 1 et k. Notons

5. `www.pathpy.net`
6. `www.higherordernetwork.com`
7. `https://pypi.org/project/honyx/`

TABLEAU 1 – Statistiques sur les réseaux issus des différents modèles

Statistique	FON$_1$	FON$_2$	VON
Nombre de noeuds	175	1 716	58 092
Nombre d'arcs	1 598	32 204	149 078
Ordre maximum	1	2	6
Nombre de représentations moyen	1	9, 80	332, 19
Nombre de représentations minimum	1	2	1
Nombre de représentations maximum	1	120	13 554
Probabilité d'aller-retour	10, 90%	30, 93%	30, 91%
Précision moyenne	19, 47%	27, 41%	39, 36%

qu'une valeur trop grande de ce paramètre mènera à des temps de calculs extrêmement longs dans le cas de jeux de données de taille importante. Pour la recherche d'ordre optimal, nous utiliserons 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.

Des informations sur la topologie des réseaux sont disponibles dans le Tableau 1. 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 nœuds. Cette différence est liée à l'ajout de dépendances indirectes sous forme de nœuds dans le réseau. Un aéroport a en moyenne presque 10 représentations dans FON$_2$. De même, le réseau VON est 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.

Ce jeu de données représentant des mobilités aériennes, il comporte un nombre important d'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. Il faudra vérifier si, en ajoutant les dépendances indirectes, il est possibe de retrouver ce comportement. Pour cela, nous calculons la probabilité d'un marcheur aléatoire de faire deux sauts dans le réseau et de revenir à son état de départ. En comparant les résultats sur les aller-retours de du tableau 1, on note un impact non négligeable des dépendances indirectes sur les résultats des probabilités d'aller-retour. Il y a par ailleurs peu de différences entre les probabilités des deux réseaux d'ordre séquentiel supérieur. Cela s'explique par le fait que VON inclut presque tous les nœuds d'ordre 2 possibles. Il serait aussi intéressant

de 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 voyageurs avec plus d'escales.

Comme indiqué en section 3.1, il est également possible d'évaluer la précision du modèle. La dernière ligne du tableau 1 indique la probabilité moyenne, avec chaque réseau, de prédire l'aéroport suivant un contexte aléatoire donné. Un gain notable est constaté en utilisant les réseaux d'ordre séquentiel supérieur. Avec VON, il est ainsi possible de prédire environ deux fois mieux l'aéroport suivant pour une trajectoire donnée que le réseau FON_1. Il faut toutefois relativiser ce gain car ce modèle contient environ cent fois plus de probabilité de transitions (c'est-à-dire d'arcs dans le graphe).

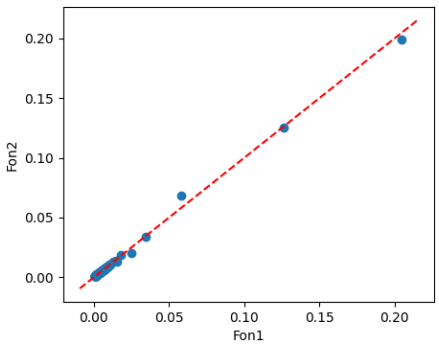
Pour répondre à la première question sur la pertinence de l'hypothèse markovienne, il est possible de dire que des dépendances séquentielles indirectes existent dans ces données. L'ordre optimal tel que défini par Scholtes (2017) est de deux et le modèle d'ordre variable (Saebi et al., 2020) détecte également des dépendances d'ordre supérieur à deux. Nous allons maintenant voir si la prise en compte de ces dépendances peut avoir un effort sur les mesures d'importance des aéroports.

4.3 Comparaison des *PageRank* selon le modèle

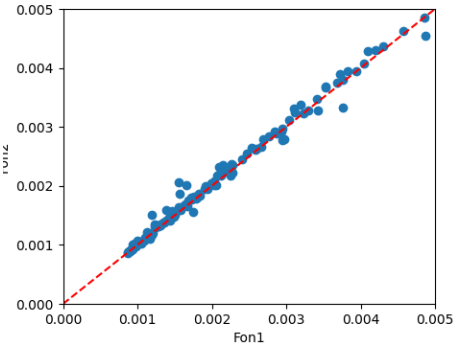
Dans cette partie nous allons calculer l'importance des aéroports dans les trois types de réseaux créés en utilisant la fonction de calcul *PageRank* du paquet `honyx`. Elle permet d'interdire la téléportation vers des nœuds d'ordre supérieur à un afin d'annuler le biais de représentations. Notons que, malgré les différences topologiques entre les réseaux, il est théoriquement possible d'obtenir des résultats similaires.

La figure 3 représente sous forme de nuage de points les corrélations entre les valeurs de *PageRank* pour chaque aéroport et ceux pour les trois différents modèles FON_1, FON_2, VON. Dans les trois sous-figures 3a-3c-3e, les points se rapprochent de la droite indiquant que les valeurs de *PageRank* sont plutôt proches, malgré les différences entre les modèles, qui n'ont pas du tout les mêmes dimensions et topologies. Pour tous les réseaux, il y a une hiérarchie similaire avec des écarts du même ordre. Cependant, les valeurs plus petites des figures 3b-3d-3e révèlent qu'il existe tout de même des différences, notamment entre VON et les deux autres réseaux.

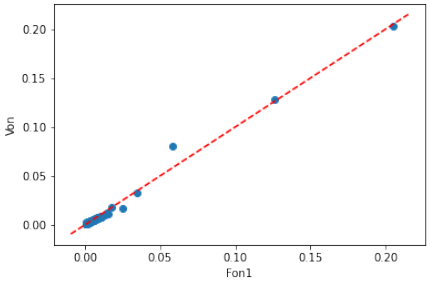
Afin de mieux se rendre compte des différences et similitudes entre les résultats de *PageRank*, nous allons regarder leur impact en terme de classement des aéroports. Les classements de rang de *PageRank* pour les 20



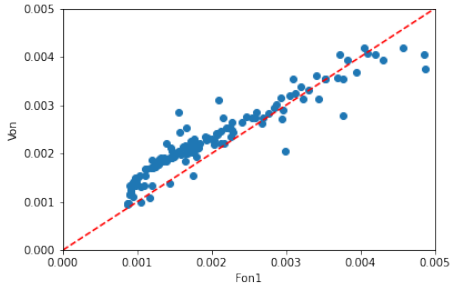
(a) Comparaison entre FON_1 et FON_2 ...



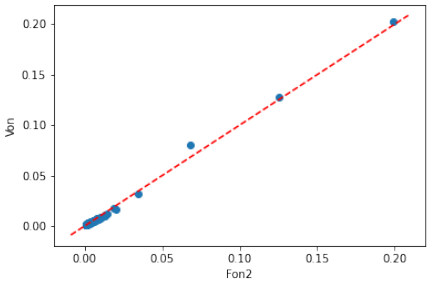
(b) ... pour les petites valeurs.



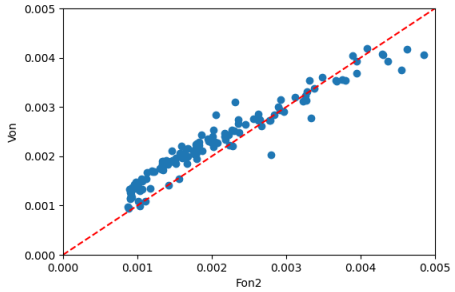
(c) Comparaison entre FON_1 et VON ...



(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. La ligne rouge en pointillé correspond à une égalité entre abscisses et ordonnées.

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 séquentielles d'ordre supérieur, celles-ci n'ont finalement que très peu de conséquences sur les rangs *PageRank* des

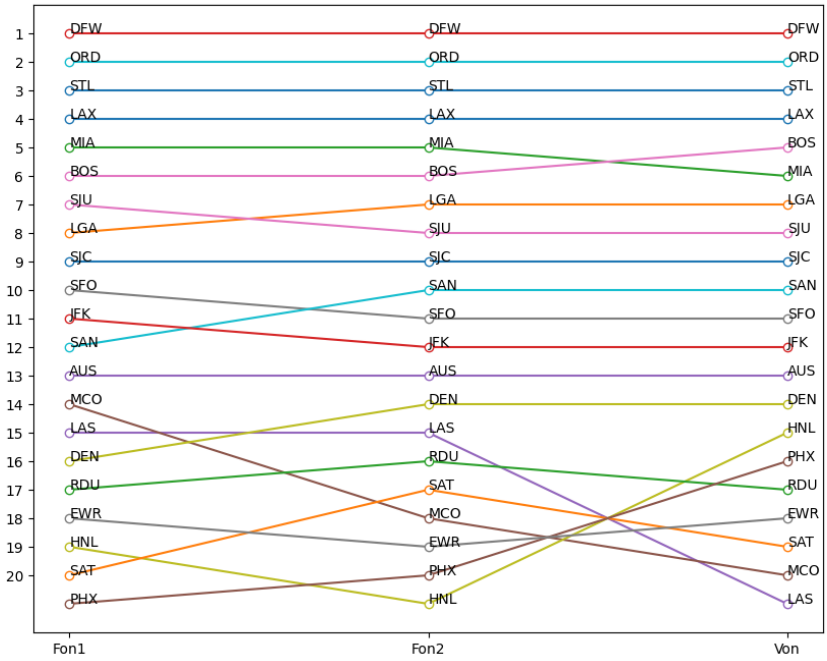


FIGURE 4 – Évolution des rangs pour les aéroports classés dans le top 20 d'au moins un réseau. *Lecture : l'aéroport de Phoenix (PHX) a le 20e plus grand PageRank dans le réseau* FON_2 *et le 16e dans le réseau* VON*, il est toutefois hors du top 20 avec le réseau d'ordre* 1, FON_1.

aéroports. Les réseaux FON_1 et FON_2 restent les plus similaires ; cela suggère que ce sont principalement les dépendances au-delà de l'ordre 2 qui ont des conséquences sur les résultats. Les variations n'impactent toutefois que de façon limitée la hiérarchie : quel que soit le modèle, les hubs de Dallas, Chicago et Saint-Louis ont des valeurs de PageRank très grandes par rapport au reste.

5 Conclusion

Nous avons présenté plusieurs façons de construire des réseaux d'ordre séquentiel supérieur. Peut-on toutefois 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, le plus parcimo-

nieux sera privilégié. Par ailleurs, les modèles possibles pour les réseaux d'ordre séquentiel 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. Il est également possible de réduire le nombre de nœuds-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 séquentiel supérieur pour un même jeu de données. Nous suggérons, à l'instar de la méthodologie présentée en 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.

Les réseaux d'ordre séquentiel supérieur peuvent être fouillés comme n'importe quel réseau «classique» d'ordre 1. Cependant, des travaux récents ont soulevé des problèmes liés à l'ajout de représentations d'un même état (Coquidé et al., 2021 ; Queiros et al., 2022). 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, le paquet Python `honyx` [8] 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

Barthélémy, M. (2004). Betweenness centrality in large complex networks. *The European Physical Journal B*, *38*(2), 163–168. doi: 10.1140/epjb/e2004-00111-4

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). doi: 10.1016/j.physrep.2020.05.004

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *30*(1), 107–117. doi: 10.1016/S0169-7552(98)00110-X

Cattan, N. (2004). Le monde au prisme des réseaux aériens. *Flux - Cahiers scientifiques internationaux Réseaux et territoires*, *58*, 32-43. doi: 10.3917/flux.058.0032

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. doi: 10.1145/2187836.2187919

8. https://pypi.org/project/honyx/

Coquidé, C., Queiros, J., & Queyroi, F. (2021). Pagerank computation for higher-order networks. In *International workshop on complex networks & their applications*. doi: 10.48550/arXiv.2109.03065

Cornwell, B. (2018). Network analysis of sequence structures. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches : Innovative methods and applications* (pp. 103–120). Cham : Springer International Publishing. doi: 10.1007/978-3-319-95420-2_7

Derudder, B., Witlox, F., Faulconbridge, J., & Beaverstock, J. (2008). Airline data for global city network research : reviewing and refining existing approaches. *GeoJournal*, *71*, 5–18. doi: 10.1007/s10708-008-9148-6

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. doi: 10.1098/rspb.2003.2629

Ducruet, C., & Notteboom, T. (2012). The worldwide maritime network of container shipping : spatial structure and regional dynamics. *Global networks*, *12*(3), 395–423. doi: 10.1111/j.1471-0374.2011.00355.x

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. 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. doi: 10.1098/rsif.2009.0495

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. doi: 10.1016/j.comcom.2010.06.007

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577-8582. doi: 10.1073/pnas.0601602103

Newman, M. J. (2005). A measure of betweenness centrality based on random walks. *Social Networks*, *27*(1), 39-54. doi: https://doi.org/10.1016/j.socnet.2004.11.009

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 doi: 10.1017/nws.2022.36

Queiros, J., & Queyroi, F. (2024). *Tutoriel sur l'utilisation du paquet python honyx*. Consulté sur https://doi.org/10.5281/zenodo.10797902 doi: 10.5281/zenodo.10797902

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. doi: 10.4000/cybergeo.36572

Robette, N. (2011). *Explorer et décrire les parcours de vie : les typologies de trajectoires*. Collections Du CEPED. Consulté sur https://shs.hal.science/halshs-01016125

Rosvall, M., Axelsson, D., & Bergstrom, C. T. (2009, 11). The map equation. *The European Physical Journal Special Topics*, *178*(1), 13–23. 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. *European Physical Journal (EPJ)*, *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

Studer, M., & Ritschard, G. (2015, 07). What Matters in Differences Between Life Trajectories : A Comparative Review of Sequence Dissimilarity Measures. *Journal of the Royal Statistical Society Series A : Statistics in Society*, *179*(2), 481-511. doi: 10.1111/rssa.12125

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. doi: 10.1137/20M1355896

TransStat, R. (2001). *Origin and destination survey database DB1B*. Consulté sur 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