HAL open science

# PageRank computation for Higher-Order Networks

Célestin Coquidé, Julie Queiros, François Queyroi

## To cite this version:

Célestin Coquidé, Julie Queiros, François Queyroi. PageRank computation for Higher-Order Networks. Complex Networks 2021, Nov 2021, Madrid, Spain. hal-03369197v2

**HAL Id: hal-03369197**

**https://hal.science/hal-03369197v2**

Submitted on 3 Nov 2021

**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.

# PageRank computation for Higher-Order Networks

Célestin Coquidé, Julie Queiros, and François Queyroi

Université de Nantes, LS2N, UMR CNRS 6004
44306 Nantes, France

**Abstract.** Higher-order networks are efficient representations of sequential data. Unlike the classic first-order network approach, they capture indirect dependencies between items composing the input sequences by the use of *memory-nodes*. We focus in this study on the variable-order network model introduced in [12,10]. Authors suggested that random-walk-based mining tools can be directly applied to these networks. We discuss the case of the PageRank measure. We show the existence of a bias due to the distribution of the number of representations of the items. We propose an adaptation of the PageRank model in order to correct it. Application on real-world data shows important differences in the achieved rankings.

**Keywords:** Higher-order Networks, Sequential data, Random walks, PageRank

## 1 Introduction

Network representation of real-world sequential data is an effective way to model complexity of interactions between items constituting them (flow of vessels between ports, city traffic, transfers between airports, *etc.*). A classic network model to represent sequential data is the pairwise interactions aggregation, extracted from the input data. This leads to a first-order Markov model of the sequences which can be represented by a first-order network (denoted $1^{\text{st}}$ON). However, the input sequences might reveal higher-order dependencies between items (see Fig. 1). Recent works [12,11] suggest that the $1^{\text{st}}$ON representation is not sufficient since it doesn't capture indirect dependencies in the underlying system. Indeed, if such dependencies exist, a random walk performed on $1^{\text{st}}$ON may result in poor approximations of the flow of movements observed in the system.

Higher-order network (HON) models are an alternative to the first-order Markov model approach. In HON, a node (or *memory-node*) encodes a subsequence of varying length of items rather than a single item. Most of the studied HON are fixed-order networks ($\text{FON}_k$) where the probability to reach the next item depends on the $k$ last visited ones [11,9]. Other studies consider variable-order models leading to variable-order network (VON) models [12,10]. Random walks performed on HON lead to better simulations of input sequences. One can

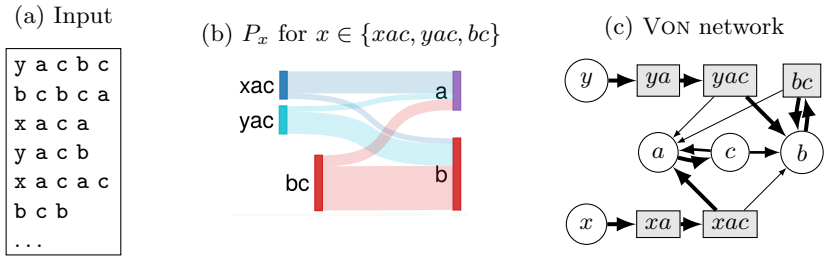



Fig. 1: Example of inputs and variable-order networks representing dependencies between *items* $\Omega = \{x, y, a, b, c\}$. We can identify 2nd or 3rd order dependencies in **(b)**. For instance, when visiting $x$ then $a$ then $c$, the flow tends to return to $a$. The Von($\mathcal{S}$) network in **(c)** only includes relevant dependencies. Subsequence $ac$ is not a relevant extension of the sequence $c$ since knowing we visited $a$ before $c$ does not impact the prediction of the next visited item. Memory-nodes are displayed as grey rectangles. The set $\mathcal{V}(c) = \{xac, yac, bc, c\}$ are *representations* of item $c$.

therefore expect the results of random-walk-based network analysis tools such as PageRank (PR) [3] to be more relevant. In the context of Von, [12] argue that such algorithms could be directly applied on variable-order network as they are still defined as weighted graphs. Nevertheless, authors have not investigated a possible bias in the resulting algorithm output due to the presence of memory-nodes.

In this study, we investigate the existence and nature of such bias when using the PR centrality measure. We start by listing related works and discussing the ambiguity of the term *higher-order* which is used for different purposes in the literature (Section 2). We then describe Von's construction [10] (Section 3). In Section 4, we introduce the standard PR model and its direct application to Von. We show that the teleportation mechanism used for computing PR values introduced a bias when applying it to Von networks. We introduce a correction as a bias-free PR model adapted for Von. In order to assess the effect of the bias (and of our correction) in practice, we compare the various PR models on real-world sequential datasets (described in Section 5). In section 6, we show that a direct application of PageRank on Von networks leads to an overestimation of the centrality of highly represented items. We also discuss the effect on ranking and its stability with a change of the damping factor parameter $\alpha$. Finally, future works perspective are given in Section 7.

## 2 Related Works

In the network science and data mining literature, the terms *higher-order* or *high-order* may relate to different concepts. For instance, in network clustering analysis, the term *higher-order* used in [13] refers to network motifs such as triangles and cycles. Authors designed a clustering algorithm preserving such

motifs. In our study the term *higher-order* refers as a network representation of a sequential data aiming to infer the indirect dependencies occurring in the sequences. In [14], such dependencies were also considered, however, they are not been inferred from sequential data.

Rosvall *et al.* [9] introduced a fixed-order network of order 2 ($\text{Fon}_2$) and generalizations of PR and clustering algorithm to this new model. As an order of 2 is limited for many real world applications, Scholtes [11] introduces a method to infer the order leading to the most accurate fixed-order network. As PR computation may prove cumbersome on fixed-order networks, a Hon PR model called Multilinear PageRank was introduced in [7]. This model is based on a *spacey surfer* whose next step doesn't depend on the previous one but on the most frequently visited ones. We focus in this study on variable-order networks (Von) [12,10]. In this context, the PR computation of the items differs from the fixed-order model.

## 3 Variable-order network representation

We detail in this section the method to build variable-order networks (Von) from input sequential data, as well as some important definitions and notations. We note $\mathcal{S} = \{s_1, s_2, \ldots, s_l\}$ the set of $l$ sequences representing the input data. Each $s_i = \sigma_i^1 \sigma_i^2 \sigma_i^3 \ldots$ consists of a sequence of items (ports, airports, or any locations). The set of all items is denoted $\Omega$. The *order* of a sequence $s$ denoted $|s|$ corresponds to its length. For two sequences $s_1$ and $s_2$, the sequence $s = s_1 s_2$ is a concatenation of $(s_1, s_2)$ and we say that $s_1$ is a *prefix* of $s$ and that $s_2$ is a *suffix* of $s$. For a sequence $s$, we call $c(s)$ the number of occurrences of $s$ in the dataset $\mathcal{S}$. The probability of finding item $\sigma$ next to sequence $s$ is estimated using relative frequencies

$$p(\sigma|s) = \frac{c(s\sigma)}{\sum_{\sigma' \in \Omega} c(s\sigma')} \tag{1}$$

where $p(\sigma|s)$ is read "probability to find $\sigma$ having as context $s$". We define $p_s = \{p(\sigma|s), \sigma \in \Omega\}$ as the distribution of items following $s$.

*Extraction of relevant extensions.* Memory-nodes are added according to *relevant extensions* found in $\mathcal{S}$. The method we used was introduced in [10]. We say $s'$ is an *extension* of $s$ if $|s'| > |s|$ and if $s$ is a suffix of $s'$. The extension $s'$ of $s$ is said to be relevant [10] if

$$D_{KL}(p_{s'}||p_s) > \frac{|s'|}{\log_2(1 + c(s'))} \tag{2}$$

where $D_{KL}$ denotes the *Kullback-Leibler divergence*. Figure 1b shows the distributions $p$ for the relevant extensions found in a toy example. The threshold used (right side of Eq. 2) makes it harder for longer and sparsely observed extensions to be found relevant. The process used for relevant extensions extraction starts

from the first-order sub-sequences. The condition in Eq. 2 is recursively applied to extension of already detected extensions. An upper-bound of $D_{KL}$ is used to stop the recursion. The construction of VON is therefore parameter-free.

*Network construction.* VON is constructed in a way that a memory-less random walk performed on it is a good approximation of the input sequential data $\mathcal{S}$. This network is noted VON $= (\mathcal{V}, \mathcal{E}, w)$, where $\mathcal{V}$ is the set of nodes. These nodes represent all relevant extensions $s$ and all their prefixes (see Fig. 1c). This ensures that any node $v$ representing a relevant extension of an item $\sigma$ is reachable during a random walk. We note $\mathcal{V}(\sigma)$ the set of nodes representing item $\sigma \in \Omega$ and $N_{\text{rep}}(\sigma) = |\mathcal{V}(\sigma)|$ the number of such representations.

For each pair $(s, \sigma)$, where $s$ is a relevant extension and $\sigma$ an item such that $p(\sigma|s) > 0$, a directed link $s \rightarrow s^*\sigma$ is added to the network. The node $s^*\sigma$ represents the longest suffix of $s\sigma$ such that $s^*\sigma \in \mathcal{V}$. The weight of this link is $p(\sigma|s)$.

## 4 Application of PageRank to variable-order networks

In this section, we first introduce the PageRank (PR) measure and its direct application on VON. We discuss the effect of the distribution of $N_{rep}$ on PR probabilities distribution. In order to isolate this effect, we introduced a biased $1^{\text{st}}$ONPR model. A bias-free model called Unbiased VON PR model is then introduced.

*Standard PageRank model ($1^{\text{st}}$ON$PR$).* The PR measure is an efficient eigenvector centrality measure in the context of directed networks. It was implemented in Google's search engine by its inventors Brin and Page [3]. PR definition of node's importance can be interpreted as follows: the more a node is pointed by important nodes, the more it is important. PR is equivalent to the steady state of a random surfer (RS) following a memory-less Markov process. The RS can follow links of the network with probability $\alpha$ or teleport uniformly towards a node of the network with probability $1 - \alpha$ (it will also teleport from any sink node). These teleportations ensure that RS cannot be stuck in a sub-region of the network and that the steady state probability distribution is unique. The PR probability associated to the node $i$ is denoted $P(i)$. As item $i \in \Omega$ is represented by a single node $i$ in $1^{\text{st}}$ON, the PR probability associated to *item* $i$ ($\Pi_1(i)$) is equal to $P(i)$. One can sort items by the decreasing order of their PR probabilities. We note $K_1$ their ranks associated to $\Pi_1$ values.

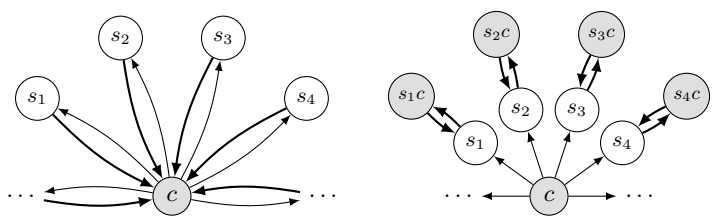
*Variable-order network PageRank (*VON$PR$*).* In the case of VONs, the memory-less Markov process actually simulates the variable-order model as memory is indeed encoded into the nodes. Therefore, [12] suggests that standard PR directly applied to VON will better reflect dependencies between items in the system than $\Pi_1$. Since more than one node represent items in VON, [12] defined the PageRank

of an item as the probability for the RS to reach at least one of its representations (see Eq. 3).

$$\Pi(i) = \sum_{v \in \mathcal{V}(i)} P(v) \tag{3}$$

We denote by $\Pi_{\text{Von}}$ and $K_{\text{Von}}$ the PR values and ranking issued from VonPR model.

Since we use a random surfer, the more representations item $i$ has, the higher is the probability to teleport to one of them. As Eq. 3 sums over representations of item $i$, this translates to a bias that is solely due to the teleportation mechanism. We can illustrate this effect with a simple example (see Fig. 2). The value of $\Pi_{\text{Von}}(c)$ is always greater than or equal to 0.5 in the situation illustrated in Fig. 2b while it is always lower than or equal to 0.5 in Fig. 2a. Equality is achieved when $\alpha = 1$ (i.e. when there is no teleportation). Although order 2 dependencies exist in 2b, it is hard to justify why item $c$ should be "more central" in this case.




(a) Von without round trips (b) Von with round trips

Fig. 2: Example of Von models of trajectories where all flows go through an item $c$. In (a), when leaving $c$, a traveler goes uniformly to any of the satellites $s_i$. In (b), a traveler coming from $s_i$ always goes back to $s_i$.

*($N_{\text{rep}}$)-biased PageRank model (Biased $1^{\text{st}}$onPR)* In order to isolate the bias due to teleportations, we assume the transition probabilities associated to representations of item $i$ are all equal to $p_i$ *i.e.* the representations of $i$ do not encode any different behaviour. This is equivalent to computing PR on $1^{\text{st}}$on using a preferential teleportation vector $\mathbf{v}_{\text{B}}$ depending on $N_{\text{rep}}$ as expressed in Eq. 4.

$$v_{\text{B}}(j) = \frac{N_{\text{rep}}(j)}{\sum_{k \in \Omega} N_{\text{rep}}(k)} \tag{4}$$

The item PR values associated to this model and its resulting ranking are denoted $\Pi_1^{\text{B}}$ and $K_1^{\text{B}}$ respectively. In the example above, $\Pi_1^{\text{B}}$ computed on Fig. 2a is equal to $\Pi_{\text{Von}}$ computed on Fig. 2b since the order-2 dependencies do not affect the centrality of $c$ in this example.

*Unbiased* Von *PageRank model.* In order to remove the bias discussed above, a modification of the teleportation vector is also used. Although several corrections are possible, the one chosen corresponds to the following random surfing process: teleportation is assumed to be the beginning of a new journey. It is therefore only possible to teleport uniformly to first-order nodes (see Eq. 5).

$$v_{\mathrm{U}}(i) = \begin{cases} 1/|\Omega| & \text{if } |i| = 1 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

It is easy to show that each node is reachable during the Markov process and therefore that the RS steady state is still unique using this teleportation vector. The item PR values associated to this model and its resulting ranking are noted $\Pi^{\mathrm{U}}_{\mathrm{Von}}$ and $K^{\mathrm{U}}_{\mathrm{Von}}$ respectively.

## 5 Datasets and Experimental settings

*Datasets.* The three datasets used correspond to spatial trajectories. They differ however in terms of length, number of sequences, number of items, *etc.*. For each dataset and each sequence, we removed any repetition of items. The code used and the datasets are available at https://github.com/ccoquide/unbiased-von-pr/.

- **Maritime :** Sequences of ports visited by shipping vessels, from April the 1st to July the 31st 2009. Data are extracted from the Lloyd's Maritime Intelligence Unit. A variable-order network (Von) analysis of maritime is presented in [12].
- **Airports :** US flight itineraries of the RITA TransStat 2014 database [2], during the 1st quarter of 2011. Each sequence is related to a passenger, it describes passenger's trip in terms of airport stops. In [11] and [9], fixed-order network (Fon) representations of the data set are presented.
- **Taxis :** Taxis rides into Porto City from July the 1st of 2013 to June the 30th of 2014. A sequence reports the succession of positions (recorded every 15 seconds) during a ride. The original data set [1] was part of the ECML/PKDD challenge of 2015. Each GPS location composing the sequences is reported onto the nearest police station as it is suggested in [10].

Sequences and networks statistics are reported in Table 1. We can observe that a large proportion of items have a large number of representations $N_{\mathrm{rep}}$. The $N_{\mathrm{rep}}$ values are far from being uniformly distributed.

*Experimental settings.* For a given dataset, we compute PR values according to the different models with $\alpha = 0.85$ along with the corresponding rankings (see Section 4). Items having the same PR probabilities are ranked using the same order. In addition to the four models described in the previous section, we also report the following measures.

Table 1: Datasets and networks information.

| Dataset | $\lvert\mathcal{S}\rvert$ | $\lvert\Omega\rvert$ | $\lvert\mathcal{V}\rvert$ | $\lvert\mathcal{E}\rvert$ | $max$(order) | Avg. $N_{\text{rep}}$ | $Q_9(N_{\text{rep}})$ | max($N_{\text{rep}}$) |
|---|---|---|---|---|---|---|---|---|
| Maritime | 4K | 909 | 18K | 47K | 8 | 20 | 50 | 674 |
| Airports | 2751K | 446 | 443K | 1292K | 6 | 995 | 1K | 34K |
| Taxis | 1514K | 41 | 4K | 15K | 14 | 99 | 250 | 382 |

- $N_{\text{rep}}$ **ranking** ($K_{\text{rep}}$) is the ranking of items by decreasing order of $N_{\text{rep}}$. We quantify how $N_{\text{rep}}$-biased are other PR models by comparing them with this benchmark.
- **Visit rank** ($K_{\text{V}}$) is the ranking based on the probability of each item to occur in the input sequences.

The visit rank is used as "ground truth" in [11, Eq. 9] for validation of the author's selection of fix-order model. However, we argue that this characterization is limited. For example, in the extreme situation where sequences are composed of only two items and can be viewed as a list of directed arcs, $K_{\text{V}}$ would correspond to the ranking made from node degrees. More generally, using the item count as a centrality measure assumed an underlying symmetry in the system *i.e.* every place is as much a destination as it is a departure.

## 6 Results

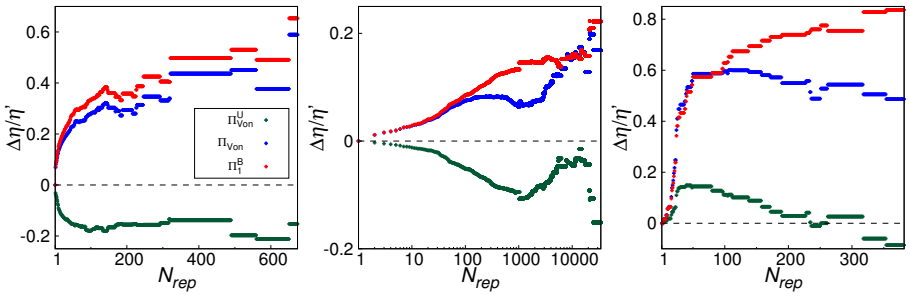
We show here that the bias effect is indeed important when looking at $\Pi$ values or the resulting rankings. Moreover, this is still true when using alternative damping factor values.

*Evolution of $\Pi$ values with $N_{\text{rep}}$.* We note $\eta(N_{\text{rep}})$ the probability that a random surfer (RS) visits any item having at least $N_{\text{rep}}$ representations such as

$$\eta(N_{\text{rep}}) = \sum_{j \in \Omega} \Pi(j) \text{ with } N_{\text{rep}}(j) \geq N_{\text{rep}} \tag{6}$$

The impact of $N_{\text{rep}}$ on PR probabilities is quantified by the relative PR boost $\Delta\eta/\eta^{'} = (\eta - \eta^{'})/\eta^{'}$ where $\eta^{'}$ is related to the 1$^{\text{st}}$onPR model. We show the evolution of $\Delta\eta/\eta$ with $N_{\text{rep}}$ in Fig. 3 for each dataset. Both VonPR ($\Pi_{\text{Von}}$) and Biased 1$^{\text{st}}$onPR ($\Pi_1^{\text{B}}$) models are the ones with the highest relative PR boosts. For example, in case of Maritime dataset (see Fig. 3a), the relative PR boosts, at $N_{\text{rep}} = N_{\text{rep}}^{\text{max}}$, equal to 60% and 65% respectively for these models (compared to 1$^{\text{st}}$onPR probabilities $\Pi_1$). Moreover, we see that the PR boosts relative to $\Pi_{\text{Von}}$ fit well with $\Pi_1^{\text{B}}$ ones. The Unbiased VonPR probabilities ($\Pi_{\text{Von}}^{\text{U}}$) are impacted very differently. For the Airports and Taxis datasets, the distributions shape of boosts is similar to VonPR's but with lower boost values. Although relative PR boosts are the lowest for our model, the higher-order dependencies it encodes still lead to differences with 1$^{\text{st}}$onPR. However, the fact that PR

probabilities are boosted for highly represented items doesn't necessarily lead to resulting biased PR rankings. Therefore, we investigate the changes in rankings when using the different models.




(a) Maritime (b) Airports (c) Taxis

Fig. 3: Relative PageRank boost $\Delta\eta/\eta'$ versus $N_{\text{rep}}$ for three PR models, with $\alpha = 0.85$. For a given number of representations $N_{\text{rep}}$, $\eta(N_{\text{rep}})$ is the probability to reach any item having at least $N_{\text{rep}}$ representations. $\eta'$ is related to 1$^{\text{st}}$ONPR model.

*Rankings comparison.* We quantified similarities between pairs of PR rankings by using both Spearman and Kendall correlation coefficients. Table 2 displays similarities for each dataset. We observe high similarity between VonPR ($K_{\text{Von}}$) and Biased 1$^{\text{st}}$onPR ($K_1^{\text{B}}$) rankings. On the other hand, correlation coefficients between Unbiased VonPR rankings ($K_{\text{Von}}^{\text{U}}$) and $K_1^{\text{B}}$ are lower. We also observe overall lower correlation coefficients with Taxis dataset which are probably due to the lower number of items. Note that the visit rank ($K_{\text{V}}$) is closer to $K_1^{\text{B}}$ or $K_{\text{Von}}$ (for Taxis). If we were to use $K_{\text{V}}$ as a PR selection method, we would likely select our biased 1$^{\text{st}}$onPR model which does not include any higher order dependencies. This highlights the fact that $K_{\text{V}}$ is not an efficient benchmark in the context of Von.

The $N_{\text{rep}}$-bias also affects the Top 10s ranking which is a popular usage of PR ranking. The Top 10s related to Maritime are displayed in Table 3. Ports with bold name are new entries when compared to the previous ranking. Although 80% of entries are common to all Top 10s, the differences come from reordering. Both $K_{\text{Von}}$ and $K_1^{\text{B}}$ fit almost perfectly with the ten most represented ports ($K_{\text{rep}}$). On the other hand, $K_{\text{Von}}^{\text{U}}$ may capture items with bad $K_{\text{rep}}$ *e.g.* the port of Surabaya ($K_{\text{rep}} = 45$ and $K_{\text{Von}}^{\text{U}} = 9$).

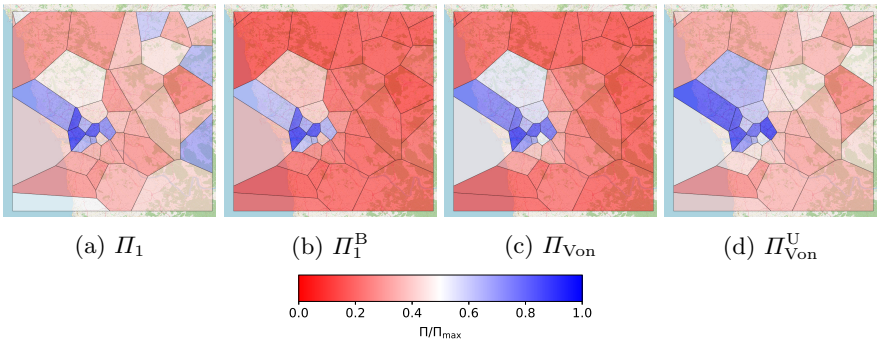
Since the number of items composing Taxis dataset (corresponding to sub-areas of Porto) is small enough, the PR scores of all items are given in Fig. 4. Both $\Pi_{\text{Von}}$ and $\Pi_1^{\text{B}}$ give bad ranks to peripherals. Only 1$^{\text{st}}$onPR and Unbiased VonPR models give importance to peripheral neighborhoods. Finally, central regions have similar rankings whatever the model used.

Table 2: Spearman ($r_s$) and Kendall ($r_\tau$) coefficients between PR rankings.

| | $K_1$ | | $K_1^{\mathrm{B}}$ | | $K_{\mathrm{Von}}$ | | $K_{\mathrm{Von}}^{\mathrm{U}}$ | |
|---|---|---|---|---|---|---|---|---|
| | $r_s$ | $r_\tau$ | $r_s$ | $r_\tau$ | $r_s$ | $r_\tau$ | $r_s$ | $r_\tau$ |
| | a) Maritime | | | | | | | |
| $K_1$ | - | - | 0.96 | 0.85 | 0.95 | 0.81 | 0.95 | 0.81 |
| $K_1^{\mathrm{B}}$ | - | - | - | - | 0.98 | 0.89 | 0.90 | 0.74 |
| $K_{\mathrm{V}}$ | 0.94 | 0.81 | 0.99 | 0.92 | 0.97 | 0.85 | 0.89 | 0.71 |
| | b) Airports | | | | | | | |
| $K_1$ | - | - | 0.98 | 0.91 | 0.96 | 0.86 | 0.95 | 0.83 |
| $K_1^{\mathrm{B}}$ | - | - | - | - | 0.99 | 0.91 | 0.91 | 0.79 |
| $K_{\mathrm{V}}$ | 0.98 | 0.91 | 0.998 | 0.96 | 0.99 | 0.92 | 0.92 | 0.80 |
| | c) Taxis | | | | | | | |
| $K_1$ | - | - | 0.62 | 0.48 | 0.44 | 0.34 | 0.77 | 0.61 |
| $K_1^{\mathrm{B}}$ | - | - | - | - | 0.94 | 0.82 | 0.88 | 0.70 |
| $K_{\mathrm{V}}$ | 0.42 | 0.30 | 0.92 | 0.79 | 0.98 | 0.91 | 0.76 | 0.58 |

Table 3: Maritime's Top 10s PageRank.

| | $K_1$ | | | $K_1^{\mathrm{B}}$ | | | $K_{\mathrm{Von}}$ | | | $K_{\mathrm{Von}}^{\mathrm{U}}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rank | Port | $K_{\mathrm{rep}}$ | $K_{\mathrm{V}}$ | Port | $K_{\mathrm{rep}}$ | $K_{\mathrm{V}}$ | Port | $K_{\mathrm{rep}}$ | $K_{\mathrm{V}}$ | Port | $K_{\mathrm{rep}}$ | $K_{\mathrm{V}}$ |
| 1 | Singapore | 2 | 2 | Singapore | 2 | 2 | Hong Kong | 1 | 1 | Singapore | 2 | 2 |
| 2 | Hong Kong | 1 | 1 | Hong Kong | 1 | 1 | Singapore | 2 | 2 | Busan | 4 | 4 |
| 3 | Rotterdam | 5 | 7 | Shanghai | 3 | 3 | Shanghai | 3 | 3 | Hong Kong | 1 | 1 |
| 4 | Busan | 4 | 4 | Busan | 4 | 4 | Busan | 4 | 4 | Rotterdam | 5 | 7 |
| 5 | Shanghai | 3 | 3 | Rotterdam | 5 | 7 | Rotterdam | 5 | 7 | Shanghai | 3 | 3 |
| 6 | Hamburg | 8 | 10 | Port Klang | 6 | 6 | Port Klang | 6 | 6 | Hamburg | 8 | 10 |
| 7 | Port Klang | 6 | 6 | Kaohsiung | 7 | 5 | Kaohsiung | 7 | 5 | Antwerp | 10 | 12 |
| 8 | Antwerp | 10 | 12 | Hamburg | 8 | 10 | Hamburg | 8 | 10 | **Bremerhaven** | 12 | 19 |
| 9 | Bremerhaven | 12 | 19 | Antwerp | 10 | 12 | Antwerp | 10 | 12 | **Surabaya** | 45 | 36 |
| 10 | Kaohsiung | 7 | 5 | **Jebel Ali** | 9 | 11 | Jebel Ali | 9 | 11 | Port Klang | 6 | 6 |

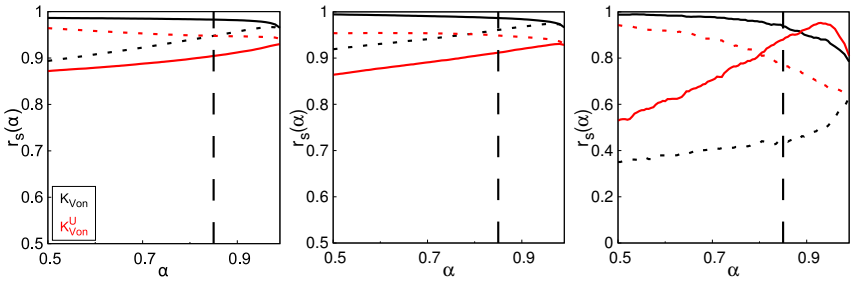



(a) $\Pi_1$ (b) $\Pi_1^{\mathrm{B}}$ (c) $\Pi_{\mathrm{Von}}$ (d) $\Pi_{\mathrm{Von}}^{\mathrm{U}}$

Fig. 4: Distribution of PageRanks values for Porto's neighborhood.

*Dependence of the $N_{\mathrm{rep}}$-bias with the damping factor $\alpha$.* Since in the literature an alternative value of the damping factor $\alpha$ could be used (as in [4]), we investigated

similarities between rankings regarding its choice. The evolution of Spearman correlation coefficient with respect to changes in $\alpha$ is present in Fig. 5 for $\alpha \in [0.5, 0.99]$. Results related to the Kendall correlation coefficient evolution are not reported since they are similar. For $\alpha \leq 0.85$, the observations made earlier are still valid. When teleportations are less frequent, different changes occur. Indeed, for Maritime and Airports, $K_{\text{Von}}$ becomes closer to $K_1$ than $K^{\text{U}}_{\text{Von}}$ is to $K_1$ (dashed lines). Overall the pairs $K_{\text{Von}}$-$K^{\text{U}}_{\text{Von}}$ and $K_1$-$K^{\text{B}}_1$ get closer as $\alpha$ tends to 1 due to the poor contribution of teleportations. For the taxis, we notice a switch at $\alpha \approx 0.9$ for $K^{\text{B}}_1$ (solid lines). We think this is due to the low amount of items related to Taxis dataset. In order to understand this behaviour, we need to further investigate other similar datasets.




(a) Maritime (b) Airports (c) Taxis

Fig. 5: Evolution of the Spearman correlation coefficient $r_s(\alpha)$ with $\alpha$, for couples of rankings. Solid lines (dashed lines) are related to correlations with Biased $1^{\text{st}}$ONPR ($1^{\text{st}}$ONPR). Vertical black dashed line represents $\alpha = 0.85$.

## 7 Future works

This study shows that the application of network measures to the new objects that are VONs are not trivial. We believe the adaptation of other analysis tools is an important challenge for the network science community. We are currently investigating the application of clustering algorithms such as Infomap [8] to VONs. This algorithm indeed uses PR in order to compare clustering qualities. However, [12] also suggests that such algorithm can be directly applied to VONs with no modifications. The PR centrality measure has other applications. A recent method based on the Google matrix (the stochastic matrix which models the random surfer), called *reduced Google matrix*, has shown its efficiency in inferring hidden links between a set of nodes of interests [6] for example with studying Wikipedia networks [5].Using user traces on website rather than usual hypertext click statistics, we will also study the generalization of this tool to VONs.

## Acknowledgement

This work was supported by the *Pays-de-la-Loire* RFI Atlanstic2020[1] research program.

## References

1. Porto taxis Trajectories Data, https://kaggle.com/crailtap/taxi-trajectory
2. Rita tansstat database, https://www.transtats.bts.gov/
3. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30(1), 107–117 (Apr 1998), https://www.sciencedirect.com/science/article/pii/S016975529800110X
4. Coquidé, C., Ermann, L., Lages, J., Shepelyansky, D.L.: Influence of petroleum and gas trade on EU economies from the reduced Google matrix analysis of UN COMTRADE data. The European Physical Journal B 92(8), 171 (Aug 2019), https://doi.org/10.1140/epjb/e2019-100132-6
5. Coquidé, C., Lages, J., Shepelyansky, D.L.: World influence and interactions of universities from Wikipedia networks. The European Physical Journal B 92(1), 3 (Jan 2019), https://doi.org/10.1140/epjb/e2018-90532-7
6. Frahm, K.M., Jaffrès-Runser, K., Shepelyansky, D.L.: Wikipedia mining of hidden links between political leaders. The European Physical Journal B 89(12), 269 (Dec 2016), https://doi.org/10.1140/epjb/e2016-70526-3
7. Gleich, D.F., Lim, L.H., Yu, Y.: Multilinear PageRank. SIAM Journal on Matrix Analysis and Applications 36(4), 1507–1541 (Jan 2015), https://epubs.siam.org/doi/10.1137/140985160, publisher: Society for Industrial and Applied Mathematics
8. Rosvall, M., Axelsson, D., Bergstrom, C.T.: The map equation. The European Physical Journal Special Topics 178(1), 13–23 (Nov 2009), https://doi.org/10.1140/epjst/e2010-01179-1
9. Rosvall, M., Esquivel, A.V., Lancichinetti, A., West, J.D., Lambiotte, R.: Memory in network flows and its effects on spreading dynamics and community detection. Nature Communications 5(1), 4630 (Aug 2014), https://www.nature.com/articles/ncomms5630, number: 1 Publisher: Nature Publishing Group
10. Saebi, M., Xu, J., Kaplan, L.M., Ribeiro, B., Chawla, N.V.: Efficient modeling of higher-order dependencies in networks: from algorithm to application for anomaly detection. EPJ Data Science 9(1), 1–22 (Dec 2020), https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-020-00233-y, number: 1 Publisher: SpringerOpen
11. Scholtes, I.: 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. pp. 1037–1046. KDD '17, Association for Computing Machinery, New York, NY, USA (Aug 2017), https://doi.org/10.1145/3097983.3098145
12. Xu, J., Wickramarathne, T.L., Chawla, N.V.: Representing higher-order dependencies in networks. Science Advances 2(5), e1600028 (May 2016), https://advances.sciencemag.org/content/2/5/e1600028, publisher: American Association for the Advancement of Science Section: Research Article

[1] https://atlanstic2020.fr/en

13. Yin, H., Benson, A.R., Leskovec, J., Gleich, D.F.: Local Higher-Order Graph Clustering. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 555–564. KDD '17, Association for Computing Machinery, New York, NY, USA (Aug 2017), https://doi.org/10.1145/3097983.3098069
14. Zhang, Z., Xu, W., Zhang, Z., Chen, G.: Opinion Dynamics Incorporating Higher-Order Interactions. In: 2020 IEEE International Conference on Data Mining (ICDM). pp. 1430–1435 (Nov 2020), https://doi.org/10.1109/ICDM50108.2020.00189, iSSN: 2374-8486