Sampling based sequential dependencies discovery in Higher-Order Network Models

Julie Queiros, François Queyroi, Samuel Maistre

FRCCS

Sampling based sequential dependencies discovery in Higher-Order Network Models

Julie Queiros[1✓], François Queyroi[1], Samuel Maistre[2]

[1] *Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004*
[2] *Université de Strasbourg, CNRS, IRMA, UMR 7501*
[✓] *Presenting author*

Abstract. Higher-order networks are a general form of network models that include memory nodes used to capture indirect dependencies between entities. When built from sequential or pathway data, random walks performed on these networks usually represent input sequences better than traditional first-order models. Unlike the latter, there are various ways to build higher-order networks that already exist in the literature. We introduce a new variable-order network where nodes can encode sequences of varying length. Nodes are selected by looking at sub-sequences that are unlikely extensions of already detected sub-sequences. Using experiments on real-world datasets, we demonstrate that our method produces smaller and as accurate models compared to the main variable-order model in the literature. We also study the differences achieved when ranking items using a higher-order reformulation of the PageRank metric.

Keywords. *Higher-order networks; Sequential data; Monte-Carlo; PageRank*

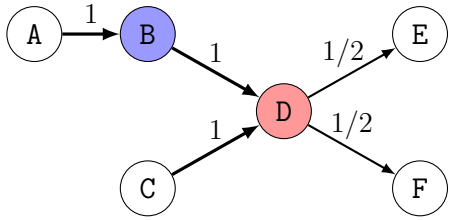
1 Introduction

Networks can be used to represent dependencies found in sequential data. One direct approach is to aggregate pairwise interactions between items in the input sequences. Examples include the number of clicks between two web pages or the number of times a ship navigates from one port to another. Most of the time, network mining algorithms use the indirect dependencies induced by the network topology, *e.g.* the PageRank metric is linked to the behaviour of a random walker on the graph.

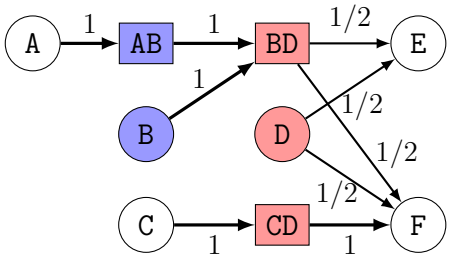
The use of this traditional representation raises an issue in the case of networks built from sequential data as the indirect relations induced by the network topology may not correspond to observed sub-sequences. For example, the graph in Fig. 1b suggests an indirect relation between item `C` and `E` going through `D`. But no such transitions exist in the input dataset the graph was built with (see Fig. 1a). Indeed, using this network representation presupposes that the modeled system respects the Markov property *i.e.* the information useful for predicting the future state is contained only in the current state (the process is also said to be "memory-less").

Sequences
D E
A B D E
B D F
C D F

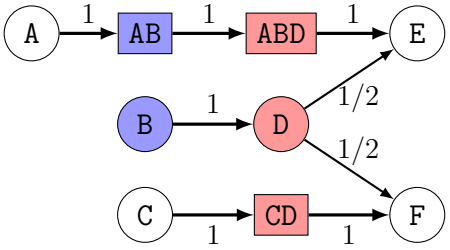
(a) Sequential Data



(b) FON$_1$



(c) FON$_2$



(d) VON

Figure 1: Examples of network representation of sequential data (1a). We assume here that each sequence is observed several times. There is a 50% chance to leave D to go to either E or F. First-order networks can encode these transitions probabilities (arc labels in 1b). However, when looking at the sequences, we can see that coming from C before D then its next destination is always F. Higher-order networks (1c and 1d) can encode these indirect dependencies by using multiple *representations* of each item. FON$_2$ network (1c) include all order-2 dependencies even those that do not help in predicting next item (*i.e.* BD). VON Network (1d) keeps only relevant dependencies (*i.e.* excluding BD but adding ABD) leading to a sparser and better representation of the sequential dependencies. The three red nodes are representations of the item D. Node labeled ABD is then a memory node of order 3.

In order to address this issue, *higher-order networks* can be used instead. In these networks, the memory on the previous steps is encoded into *memory nodes*. An item can then be represented not by one (as in first-order models) but by several nodes. On this new class of networks, a random walk will better simulate the input sequences. In the example, coming for C, the only possible destination for a random walker after D is F. We focus specifically on variable-order networks (see Fig. 1d) where there are no constraint on the length of memory nodes.

The main question is: how do we select the memory nodes to add to the network? In the Fig. 1d, the sequence ABD is included as a memory node while BD is not. Indeed, only knowing that B is observed before D does not add more information. We say that the sequence ABD is a *relevant context* as it adds information about the following items. The issue of *relevant context* definition is the central topic of this paper. The use of variable-order network was introduced by Xu *et al.* [23, 18] who provided a first answer to this question. However, as there are several ways to define "relevance", several variable-order network representations of the same sequential dataset are possible. In this paper, we introduce a new approach called MC-VON to construct variable-order networks where the relevance of a context is assessed using a statistical test. We show on real-world data that this new model can be as efficient as predicting sequences while being sparser that the model of Xu *et al.*. Furthermore, we evaluate the effect of model selection on higher-order PageRanks rankings.

2 Related works

The concept "higher-order networks" (HON) is used to describe graph models that are not limited to *dyadic* relations and design to capture different *system dependencies* (*e.g.* temporal, sequential, subset, spatial *dependencies*) [7, 21]. Like several authors [21], we use here the term "higher-order" to refer exclusively to networks that encode sequential dependencies (transition probabilities for *contexts* of length (or order) longer than 1). Applications of HON include the identification of overlapping modules or the evaluation of items centrality [23, 17, 13].

Most of the existing literature focus on fixed-order networks we call FON$_k$ [17]. These models generalize classic *memory-less* models as they are networks where the probability of a random walker to reach *item* σ depends on the previous k steps rather than only the last one. The parameter k can be set to a given value [17] but the estimation of the optimal order for a given dataset was also investigated [20, 19].

We focus in this paper on a less discussed family of models called variable-order networks (VON). Variable-order Markov models of discrete sequences have been studied in the past [3]. But, to the best of our knowledge, the only applications to network analysis is the seminal work of Xu *et al.* [23, 18]. Their main idea is that orders should be found locally rather than determining a global order for the system *i.e.* memory nodes should only be included if they indeed impact the behaviour of a random walker.

The number of memory nodes in a FON$_k$ grows exponentially with k. Therefore, VON models are useful when higher-order dependencies existing in the system are sparse. This is a safe hypothesis since real-world networks are considered sparse. Still, Xu *et al.* rightly suggest that one should find a balance between the resulting network size and the corresponding model's goodness-of-fit *w.r.t.* input sequences. In the present paper, we compare their network model (denoted D_{KL}-VON) to ours (denoted MC-VON). These models are defined in the following section.

3 Methods

Let $\mathcal{A}$ be the set of items. An input dataset corresponds to a set $\mathcal{S} = (s^1, s^2, \ldots)$ of sequences $s^i = \sigma_1^i \sigma_2^i \sigma_3^i \ldots$ where all $\sigma_j^i \in \mathcal{A}$. For a sequence s of symbols in $\mathcal{A}$, the *order* of s denoted $|s|$ is the length of s and the *support* of s denoted $c(s)$ is the number of occurrences of s in dataset $\mathcal{S}$. We will also use $C_s = (c(s\sigma), \sigma \in \mathcal{A})$ to denote the occurrences of items following the sequence s. Let $s = s_1 s_2$ be a sequence resulting in the concatenation of sequences s_1 and s_2. Here, s_1 is a *prefix* of s, s_2 is a *suffix* of s while we call s an *extension* of s_2. Based on a vector $K = \{K_\sigma, \sigma \in \mathcal{A}\} \in \mathbb{N}^{|\mathcal{A}|}$, where K_σ represents the number of occurences of element σ, we define the probability measure $Q_K(\sigma) = K_\sigma / \sum_{\sigma' \in \mathcal{A}} K_{\sigma'}$, *i.e.* the probability to choose σ when drawing a random element from K.

In discrete sequences prediction, we want to estimate $P(\sigma|\sigma_1 \ldots \sigma_k)$ *i.e.* the probability to encounter item σ after the sequence $\sigma_1 \ldots \sigma_k$. In a *higher-order* Markov model, we assume we have a set $\mathcal{R}$ (the relevant contexts) of sequences of items including $\mathcal{A}$. The probability $P(\sigma|\sigma_1 \ldots \sigma_k)$ will be estimated using $P^{\mathcal{R}}(\sigma \mid \sigma_1 \ldots \sigma_k) \coloneqq Q_{C_{s'}}(\sigma)$ where s' is the longest suffix

of $\sigma_1 \dots \sigma_k \in \mathcal{R}$. In a *memory-less* Markov model (or first-order model), we have $\mathcal{R} = \mathcal{A}$ and only the last visited item is taken into account *i.e.* $P^{\mathcal{A}}(\sigma \mid \sigma_1 \dots \sigma_k) = P^{\mathcal{A}}(\sigma \mid \sigma_k)$.

The HON models studied here are built from the same general procedure described in the Section 3.1. We defined known HON models using this procedure. The difference in the method we are proposing comes down to the identification of the set of relevant contexts $\mathcal{R}$ which is detailed in 3.2.

3.1 Generic Higher-Order Networks construction procedure

Higher-order networks aim at encoding transition probabilities of higher-order Markov models into a regular weighted directed graph (as the example in Fig. 1b). From a set of relevant contexts $\mathcal{R}$, the weighted directed graph $G(\mathcal{R}) = (\mathcal{R}, E, w)$ is built with each item $\sigma \in \mathcal{A}$ represented by multiple nodes corresponding to the contexts having σ as the last entry. We say that these *memory nodes* are the *representations* of item σ. The edge set E and the weights w are defined as follows. Let $s \in \mathcal{R}$ and $\sigma \in \mathcal{A}$ such that $P^{\mathcal{R}}(\sigma|s) > 0$, $G(\mathcal{R})$ contains a link $s \to s^*\sigma$ of weight $w(s \to s^*\sigma) = P^{\mathcal{R}}(\sigma|s)$ where s^* is the longest suffix of $s \in \mathcal{R}$. For example, let $s = abc$ and $s^*\sigma = bc\sigma$ be relevant extensions of c and σ respectively then there will be a link $s \to s^*\sigma$ if $abc\sigma \notin \mathcal{R}$ and $P^{\mathcal{R}}(\sigma|s) > 0$.

Algorithm 1: VON Generic Algorithm

Data: $\mathcal{S}$: set of sequences over itemset $\mathcal{A}$
Input: s_c, s_v: current and last relevant contexts
Result: R: set of relevant contexts
if *existRelevantExts*(s_c, s_v) **then**
  for $\sigma \in \mathcal{A}$ **do**
    if *isRelevant*$(\sigma s_c, s_v)$ **then**
      $R \leftarrow R \cup \text{VON}(\sigma s_c, \sigma s_c)$
    else
      $R \leftarrow R \cup \text{VON}(\sigma s_c, s_v)$
return $R \cup$ *prefixes*(s_v)

As previously said, the difference between HON models mainly comes from the way the set of contexts $\mathcal{R}$ is defined. Algorithm 1 is a general framework used in [18] to extract such a set. Relevant contexts are recursively found as *extensions* of contexts found at lower orders. For a dataset $\mathcal{S}$ and an itemset $\mathcal{A}$, the final set of contexts is defined as $\mathcal{R} \coloneqq \bigcup_{\sigma \in \mathcal{A}} \text{VON}(\sigma, \sigma)$.

The functions `isRelevant` and `existRelevantExts` depend on the model used. The test `isRelevant`(s_c, s_v) is passed when s_c is judged relevant when compared to the last relevant suffix identified s_v. The function `existRelevantExts` is used to identify situations where no relevant extensions of s_v are possible and, therefore, where the recursion must be stopped. As such, this function should not need to count sub-sequences $\sigma_1 s_c \sigma_2$. This operation is to be done only if a relevant extension is possible. If well designed, it should make possible the use of Algorithm 1 on large datasets [18].

Finally, function `prefixes` returns the set of prefixes of s_v (including itself). Indeed, a random

walker on a higher-order network can only reach memory node $s_1 s_2 \dots s_k$ if there is a path $s_1 \to s_1 s_2 \to \dots \to s_1 s_2 \dots s_k$. Therefore, every prefix of s_v are added in the network even if some are not relevant. In the example of Fig. 1d, knowing that we observed `A` before `B` does not provide more information than simply knowing it came from `B`. However, the representation `AB` is here as a prefix of `ABD` which is a relevant context.

Definition 1 *The fixed-order network* FON$_k$ *is obtained by treating a sub-sequence as relevant if its order is lower or equal to* k*.*

With this definition, the network FON$_1$ is the traditional first-order network (Fig. 1b). Fixed-order networks are usually defined as subgraphs of the k order De Bruijn graph over $\mathcal{A}$ in the literature. But this formalism does not allow to keep track of transition probabilities for contexts of order lower than k. This model is also called *multi-order* model [19].

Definition 2 *The variable-order network* D_{KL}-VON(λ) *[18] is obtained by treating the sub-sequence* σs_c *as relevant w.r.t* s_v *iff*

$$D_{KL}\left(P_{\sigma s_c} || P_{s_v}\right) > \frac{\lambda |\sigma s_c|}{\log_2(1 + c(\sigma s_c))} \tag{1}$$

with $\lambda \in \mathbb{R}^+$ *the threshold multiplier,* $P_s \coloneqq Q_{C_s}$ *the distribution of items following* s *and* D_{KL} *the Kullback-Leibler divergence (in bits).*

One main advantage of D_{KL}-VON when compared to FON$_k$ is that the length of contexts is locally defined in order to best fit the data. The right side of Eq. 1 makes longer and sparsely observed contexts harder to be recognized as relevant.
The parameter λ is not actually included in the original definition of [18] (we have an equivalent definition for $\lambda = 1$). The original definition of the authors is indeed parameter-free. However, the interpretation of the right-side threshold function in relation to the D_{KL} divergence is hard to grasp. We argue that the definition of the threshold function actually hides an arbitrary choice of "scale" made by the authors. Therefore, the "parameter-freeness" of D_{KL}-VON is limited in our opinion. We shall use different value of λ in order to compare D_{KL}-VON model to the one defined below.

3.2 MC-Von model definition

Our proposal to construct a variable-order network model is to use to the quantity $D_{KL}\left(P_{\sigma s_c} || P_{s_v}\right)$ as in [18], but as a test statistic in a hypothesis testing paradigm to avoid relying on an *ad hoc* threshold function. Indeed, if σs_c is not a relevant context, then $C_{\sigma s_c}$ should behave like a draw of $c(\sigma s_c)$ elements from C_{s_v} without replacement, i.e. from a multivariate hypergeometric distribution $\mathcal{MH}(C_{s_v}, c(\sigma s_c))$. Therefore, we will decide that σs_c is a relevant context when $C_{\sigma s_c}$ does not behave like a random draw, that is while we can reject the null hypothesis

$$H_0 : C_{\sigma s_c} \sim \mathcal{MH}(C_{s_v}, c(\sigma s_c)) \quad \text{vs.} \quad H_1 : C_{\sigma s_c} \nsim \mathcal{MH}(C_{s_v}, c(\sigma s_c)).$$

The nominal level $\alpha \in (0, 1)$ of the test allows us to choose how surprising we want the draw $C_{\sigma s_c}$ to be in order to consider s_c as a relevant context. It is also an upper bound for the probability of a context being considered relevant when it is not.

Definition 3 *The variable-order network MC-VON is obtained by treating the sub-sequence σs_c as relevant w.r.t s_v iff*

$$D_{KL}\left(P_{\sigma s_c}||P_{s_v}\right) > q_{1-\alpha}\left(c(\sigma s_c), s_v\right) \quad \textit{or equivalently} \quad p < \alpha \tag{2}$$

where $P_s \coloneqq Q_{C_s}$ the distribution of items following s, $q_{1-\alpha}\left(c(\sigma s_c), s_v\right)$ is the $(1-\alpha)$-th quantile of the distribution of $D_{KL}\left(Q_D||P_{s_v}\right)$ where D is a random draw from $\mathcal{MH}(C_{s_v}, c(\sigma s_c))$ and

$$p = \mathbb{P}\left(D_{KL}\left(Q_D||P_{s_v}\right) \geq D_{KL}\left(P_{\sigma s_c}||P_{s_v}\right)\right)$$

is the p-value of the test.

Example 1 *Assume we have an order 1 subsequence s_v with $C_{s_v} = (1, 2, 5, 0, \ldots)$ and we want to assess the relevancy of the extension σs_c with $C_{\sigma s_c} = (1, 0, 0, 0, \ldots)$. We have $D_{KL}\left(P_{\sigma s_c}||P_{s_v}\right) = -\log_2(1/8) = 3$. Since it is the hightest possible D_{KL}, the probability that a draw from $\mathcal{MH}(C_{s_v}, 1)$ has a divergence higher or equal is the probability to draw $C_{\sigma s_c}$ i.e. $p = \frac{1}{8}$. Taking a standard threshold of $\alpha = 10^{-3}$, we would accept H_0 and declare that this extension is not relevant. On the other hand, for D_{KL}-VON(1), the threshold function of Eq. 1 is equal to $\frac{2}{\log_2(2)} = 2$. The extension σs_c would here be considered as relevant.*

However, p (or $q_{1-\alpha}\left(c(\sigma s_c), s_v\right)$) can be difficult to compute, particularly if $c(\sigma s_c)$ is neither small nor close to $c(s_v)$. Therefore, we propose different approximations to estimate it and decide whether (2) holds or not. The first possibility is to use a Monte-Carlo algorithm that draws M independent replications $\{D_i, 1 \leq i \leq M\}$ from $\mathcal{MH}(C_{s_v}, c(\sigma s_c))$ and estimate p by $\widehat{p} = S_M/M$ where

$$S_M = \sum_{i=1}^{M} \mathbb{I}\left\{D_{KL}\left(Q_{D_i}||P_{s_v}\right) \geq D_{KL}\left(P_{\sigma s_c}||P_{s_v}\right)\right\}$$

follows a binomial distribution of size M and probability p, i.e.

$$\mathbb{P}(S_M = k) = \binom{M}{k} p^k (1-p)^{M-k} =: b(n, p, k).$$

The choice of M will affect the precision of our decision, particularly if p is close to α. On the contrary, if the conclusion is more obvious, i.e. $p \ll \alpha$ or $p \gg \alpha$, we might have chosen a smaller value for M to get a reasonable precision. Methods that adapt the number of replications to the distance between p and α were proposed, e.g. in [8] and [6] among others. These two papers both control the resampling risk defined by

$$RR_p(\widehat{p}) = \begin{cases} \mathbb{P}_p\left(\widehat{p} > \alpha\right) & \text{if } p \leq \alpha \\ \mathbb{P}_p\left(\widehat{p} \leq \alpha\right) & \text{if } p > \alpha. \end{cases}$$

This resampling risk measures the probability to take the wrong decision regarding (2). For a given $\epsilon > 0$, [8] and [6] propose procedures that ensures that $RR_p \leq \epsilon$. Nevertheless, there is no bound on the number of replications needed and the procedure might not end if $p = \alpha$, i.e. $D_{KL}\left(P_{\sigma s_c}||P_{s_v}\right) = q_{1-\alpha}\left(c(\sigma s_c), s_v\right)$. A maximum number of iterations must be chosen and the resampling risk is therefore not truly controlled. For this reason, we divide α into a lower value α^- and a higher value α^+ so that the number of iterations is always finite. The cost of this

finite number of iterations is made by accepting slightly less relevant sequences ($\alpha^- < p \leq \alpha^+$) rather than missing sequences that are relevant ($p \leq \alpha^-$) and define

$$\widetilde{RR}_p(\widehat{p}) = \begin{cases} \mathbb{P}_p\left(\widehat{p} > \alpha^*\right) & \text{if } p \leq \alpha^- \\ 0 & \text{if } p \in]\alpha^-, \alpha^+] \\ \mathbb{P}_p\left(\widehat{p} \leq \alpha^*\right) & \text{if } p > \alpha^+ \end{cases}$$

where $\alpha^* \in]\alpha^-, \alpha^+[$ is a critical value for $\widehat{p}$ such that we will reject H_0 iff $\widehat{p} < \alpha^*$. We construct a procedure that ends after a finite number of steps and such that

$$\sup_{p \in [0,1]} \widetilde{RR}_p(\widehat{p}) \leq \epsilon. \tag{3}$$

We define the algorithm 2 that ensures (3), where

$$\alpha^* = 1 - \frac{\log\left(\alpha^+/\alpha^-\right)}{\log\left(\dfrac{\alpha^+/(1-\alpha^+)}{\alpha^-/(1-\alpha^-)}\right)},$$

is such that $b(n, \alpha^-, n\alpha^*) = b(n, \alpha^+, n\alpha^*)$. α^* is also the value of p that will require the highest number of draws on average. The termination of the algorithm comes from the fact that the function $x \mapsto (n+1)b(n, x, n\hat{p})$ is the beta density with parameters $n\hat{p}+1$ and $n(1-\hat{p})$ and tends to a Dirac measure in p as $n \to \infty$. Therefore at least one of the two values $b(n, \alpha^-, S_n)$ and $b(n, \alpha^+, S_n)$ must tend to zero.

Algorithm 2: MC-Von Decision Algorithm

Data: $D_{KL,obs}$: observed KL divergence
Input: α^-, α^+, ϵ: test levels and bound for resampling risk
Result: $\widehat{p}$: estimated p-value
$S = 0$; $n = 0$
while $b(n, \alpha^-, S) > \epsilon/(n+1)$ *and* $b(n, \alpha^+, S) > \epsilon/(n+1)$ **do**
  $D \sim \mathcal{MH}(C_{s_v}, c(\sigma s_c))$
  if $D_{KL}\left(Q_D || P_{s_v}\right) \geq D_{KL,obs}$ **then**
    $S = S + 1$
  $n = n + 1$
return $\widehat{p} = S/n$

For the function `existRelevantExts`(s_c, s_v), we use a simple lower-bound on the p-value. Indeed, there are at most $z = \binom{c(s_v)}{\min\left(\frac{c(s_v)}{2}, c(s_c)\right)}$ draws from $\mathcal{MH}(C_{s_v}, c(\sigma s_c))$ for any $\sigma \in \mathcal{A}$. Therefore if $z^{-1} > \alpha^+$ there is no possible extensions of s_c that can be found relevant.

4 Experiments

We use four different datasets (see Table 1) for the experiments. They offer a variety not only in terms of origin of the data but also in terms of size of the itemsets and sequences. Two of them (AIR and PORTS) correspond to spatial pathway data. AIR contains the itineraries of US passengers for domestic flights extracted from the *RITA TransStat* database. PORTS contains the sequences of ports where shipping vessels stop over between April and October 2009.

Table 1: Summary of datasets used

Dataset	$\lvert\mathcal{A}\rvert$	$\lvert\mathcal{S}\rvert$	**min/max** $\lvert s\rvert$	**Refs.**
Shipping path. (PORTS)	909	4243	2/183	[23, 5]
US Airflight (AIR)	175	286,810	2/14	[19]
Wikipedia clicks (WIKI)	100	29,573	2/22	[20, 19]
MSNBC clicks (MSNBC)	17	388,434	2/1810	[20, 19]

This is an extract from the Lloyd's Maritime Intelligence Unit database. WIKI and MSNBC are clickstream datasets. WIKI is the result of Wikipedia navigation games. Following [19], we only retain the sequences going through the top 100 visited pages. MSNBC gathers click streams of visitors of the website of the channel. The pages are grouped into 17 categories (*e.g.* "business", "local", "sports",...). We also removed all consecutive repetitions from the input sequences.

To construct the networks MC-VON we use a standard value for the confidence threshold $\alpha^- = 10^{-3}$ with $\alpha^+ = \alpha^- + 2.10^{-3}$ to control a risk $\epsilon = 0.05$. This means we make the correct decision for p-values outside $)\alpha^-, \alpha^+($ with a probability at least $1-\epsilon$. We report results for the D_{KL}-VON [18] model as it is the the main other model of variable-order networks existing in the literature. In order to compare the contexts retained by each approach in terms of accuracy, we will also determine λ^* such that D_{KL}-VON(λ^*) contains a number of nodes equivalent to MC-VON. Similarly, we determine the α_*^- such that MC-VON(α_*^-) contains a number of nodes equivalent to D_{KL}-VON(1), the other parameters being equal. We also include the results obtained with the FON$_1$ network and the FON with the optimal order according to [19]. The `honyx` python package[1], developed by the authors, was used to generate the higher-order networks. The datasets and the scripts used for the experiments can be found at [15].

4.1 Networks size and models accuracy

We investigate here the difference between the constructed HON and whether a better or similar accuracy with a smaller model can be achieved using MC-VON. Table 2 reports the results for each constructed network on the four datasets using the whole set $\mathcal{S}$. Networks size is represented by the number of nodes $|V|$ in the networks. The order correspond to the maximum order among the vertices. The last columns reports each model accuracy score *Acc* (Eq. 4) when splitting $\mathcal{S}$ into a 90% training set and a 10% testing set $\mathcal{S}_T$. It corresponds to the average probability to identify the correct next item in $\mathcal{S}_T$:

$$Acc(\mathcal{R}, \mathcal{S}_T) \coloneqq \frac{100}{|\mathcal{S}_T|} \sum_{s \in \mathcal{S}_T} \frac{1}{|s|-1} \sum_{i=1}^{|s|-1} P^{\mathcal{R}}(s_{i+1} | s_1 s_2 \dots s_i) \tag{4}$$

The increase in accuracy between FON$_1$ and the other models justifies the use of higher-order models. For example we can correctly predict almost half of the ports visited by ships in the PORTS dataset using D_{KL}-VON or MC-VON. This score drops to 13% for the regular FON$_1$ network. This difference is less important for WIKI. Accordingly, the optimal order found using Scholtes' method [19] is 1 for this dataset.

[1] `https://pypi.org/project/honyx/`

Table 2: Comparison of the network models used.

Dataset	**Network**	$\|V\|$	**Order**	**Acc** $\pm 2sd$
PORTS	FON$_1$	909	1	13.71 ± 0.73
	FON$_2$	9,437	2	31.73 ± 1.38
	D_{KL}-VON(1.95)	9,559	6	38.56 ± 1.63
	D_{KL}-VON(1)	18K	8	46.48 ± 1.89
	MC-VON(0.001)	9,553	16	42.93 ± 2.22
	MC-VON(0.05)	18K	27	48.17 ± 2.23
AIR	FON$_1$	175	1	19.48 ± 0.09
	FON$_2$	1,716	2	27.44 ± 0.10
	D_{KL}-VON(2.85)	28K	6	36.50 ± 0.15
	D_{KL}-VON(1)	58K	6	39.37 ± 0.19
	MC-VON(0.001)	28K	6	37.11 ± 0.15
	MC-VON(0.29)	58K	6	39.19 ± 0.20
MSNBC	FON$_1$	17	1	13.82 ± 0.07
	FON$_3$	4,061	3	22.18 ± 0.16
	D_{KL}-VON(1.585)	5,774	8	22.04 ± 0.15
	D_{KL}-VON(1)	28K	11	22.29 ± 0.17
	MC-VON(0.001)	5,771	122	22.44 ± 0.17
	MC-VON(0.027)	28K	145	22.43 ± 0.16
WIKI	FON$_1$	100	1	21.48 ± 0.65
	D_{KL}-VON(3.39)	306	4	21.87 ± 0.67
	D_{KL}-VON(1)	2,260	4	23.29 ± 0.64
	MC-VON(0.001)	304	4	22.85 ± 0.65
	MC-VON(0.35)	2,257	12	23.39 ± 0.70

We now compare the networks created using D_{KL}-VON and MC-VON. For a given order, the set of contexts found relevant is different even if the parameters are tuned to have sets of similar size. This suggests that the difference between the two methods is not just a matter of parameterization. The relevant contexts occur less frequently on average in D_{KL}-VON than the contexts found using MC-VON. This effect may be similar to the situation shown in Example 1: the low-frequency contexts may have a large D_{KL} value that easily passes the test of Eq. 1. Networks constructed using MC-VON have a larger maximum order which can be very large. This is expected since the criterion used does not inherently penalize large contexts. The largest discrepancies are obtained with MSNBC and PORTS. Note, however, that such contexts are rare; the vast majority of memory nodes are of order 2 or 3.

When comparing variable-order networks of similar size, MC-VON seems to match D_{KL}-VON in terms of accuracy. For PORTS or WIKI, it clearly outperforms it. For MSNBC and AIR, the results are closer. This supports the idea that the criterion used for MC-VON helps to produce higher-order networks that are more consistent with the data. A final observation is that the computational time required to construct networks using MC-VON is several orders of magnitude higher than the time required using D_{KL}-VON (*e.g.* half an hour versus a few seconds for MSNBC). Although network analysis tasks rarely come with online computational constraints, a future challenge would be to improve the computation of MC-VON.

Table 3: Spearman correlations between higher-order PageRank rankings and Levenshtein edit distance between the Top10s.

Dataset	**Networks**	FON$_*$	D_{KL}-VON(1)	MC-VON
PORTS	FON$_1$	0.96 / 3	0.95 / 5	0.98 / 5
	FON$_*$		0.99 / 4	0.96 / 5
	D_{KL}-VON(1)			0.96 / 4
AIR	FON$_1$	0.99 / 3	0.98 / 4	0.97 / 4
	FON$_*$		0.99 / 2	0.98 / 2
	D_{KL}-VON(1)			0.99 / 0
MSNBC	FON$_1$	0.99 / 4	0.97 / 4	0.98 / 3
	FON$_*$		0.98 / 2	0.99 / 2
	D_{KL}-VON(1)			0.98 / 4
WIKI	FON$_1$	1. / 0	0.90 / 6	0.96 / 5
	D_{KL}-VON(1)			0.94 / 6

4.2 Higher-order PageRank

We now investigate how the choice of model impacts network mining algorithms results. In particular, we look at the rankings achieved using a higher-order version of the PageRank (PR) centrality measure. Since higher-order networks are still weighted graph, we can compute nodes' PRs and then define *items*' PR as the sum of their representations' PR values [23]. To be more precise, we use a corrected version of the PR metric for higher-order networks [5]. This reformulation corrects a bias due to the multiplicity and the non-normal distribution of the representation of each item. With this correction, we can compare PR items for HON of different sizes.

Table 3 reports the Spearman correlation coefficient as well as the edit distance between the top 10 found for each network. We observe strong similarities in rankings between all of the networks including FON$_1$. For all of Spearman correlations, the hypothesis of independence between the samples can be rejected. Therefore, the choice of HON model does not completely reverse the hierarchy of items. Even if sequential dependencies exist in the dataset, PR-based centrality analysis is still relevant without taking them into account. However, there are still differences between the ranking found as suggested by the difference between the top 10 most important items. These rank differences are more common when the PR values are more evenly distributed (*e.g.* for the WIKI dataset). In order to better see these differences, Fig. 2 shows the actual top 10s for the PORTS dataset.

5 Conclusion

We introduced a variable-order network model MC-VON that uses statistical significance for the identification of relevant contexts in a variable-order Markov model. Experiments have shown that we can construct sparser networks in which random walks will represent input sequences almost as well or even better than using known models. We therefore argue that our approach is a good alternative to D_{KL}-VON. On the other hand, the difference between the networks is not as important when looking at the items PageRank rankings. This suggests

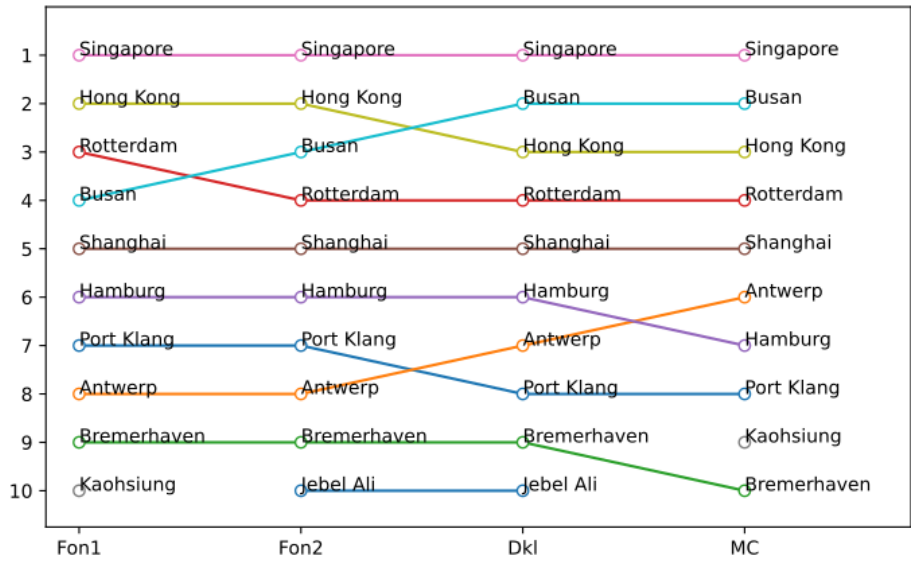


Figure 2: Differences between the top 10 ports in terms of PageRank according to the network model (PORTS dataset).

the effects of the choice of model on the information we extract from those network may be more limited. However, the choice of model may be more important when using other network mining algorithms [13].

A direction for future work is the improvement of the computation of our model p value. The method here is designed to obtain a stable solution that is not affected a lot by Monte-Carlo innate randomness. We believe that faster approximations and still stable procedures are possible, for example using Sequential Monte-Carlo techniques.

References

[1] Ron Begleiter, Ran El-Yaniv, and Golan Yona. On prediction using variable order markov models. *Journal of Artificial Intelligence Research*, 22:385–421, 2004.

[2] Irad Ben-Gal, Gail Morag, and Armin Shmilovici. Context-based statistical process control: A monitoring procedure for state-dependent processes. *Technometrics*, 45(4):293–311, 2003.

[3] Jose Borges and Mark Levene. Evaluating variable-length markov chain models for analysis of user web navigation sessions. *IEEE Transactions on Knowledge and Data Engineering*, 19(4):441–452, 2007.

[4] Brenno Caetano Troca Cabella, Márcio Júnior Sturzbecher, Walfred Tedeschi, Oswaldo Baffa Filho, Dráulio Barros de Araújo, and Ubiraci Pereira da Costa Neves. A numerical study of the kullback-leibler distance in functional magnetic resonance imaging. *Brazilian Journal of Physics*, 38:20–25, 2008.

[5] Célestin Coquidé, Julie Queiros, and François Queyroi. Pagerank computation for higher-order networks. In *International Conference on Complex Networks and Their Applications*, pages 183–193, 2021.

[6] Dong Ding, Axel Gandy, and Georg Hahn. A simple method for implementing monte carlo tests. *Computational Statistics*, 35:1373–1392, 2020.

[7] Tina Eliassi-Rad, Vito Latora, Martin Rosvall, and Ingo Scholtes. Higher-Order Graph

Models: From Theoretical Foundations to Machine Learning (Dagstuhl Seminar 21352), 2021.

[8] Axel Gandy. Sequential implementation of monte carlo tests with uniformly bounded resampling risk. *Journal of the American Statistical Association*, 104(488):1504–1511, 2009. https://www.jstor.org/stable/40592357.

[9] Richard E Korf. A complete anytime algorithm for number partitioning. *Artificial Intelligence*, 106(2):181–203, 1998.

[10] Renaud Lambiotte, Martin Rosvall, and Ingo Scholtes. Understanding complex systems: From networks to optimal higher-order models. *arXiv preprint arXiv:1806.05977*, 2018.

[11] Tiago P Peixoto and Martin Rosvall. Modelling sequences and temporal networks with dynamic community structures. *Nature communications*, 8(1):1–12, 2017.

[12] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. *J. Graph Algorithms Appl*, pages 191–218, 2006.

[13] Julie Queiros, Célestin Coquidé, and François Queyroi. Toward random walk-based clustering of variable-order networks. *Network Science*, 10(4):381–399, 2022.

[14] François Queyroi. Least likely sample in multivariate hypergeometric distributions. Mathematics Stack Exchange. (version: 2022-03-02).

[15] François Queyroi, Julie Queiros, and Samuel Maistre. Code and Datasets "Sampling based sequential dependencies discovery in Higher-Order Network Models", February 2024.

[16] Martin Rosvall, Daniel Axelsson, and Carl T Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23, 2009.

[17] Martin Rosvall, Alcides V Esquivel, Andrea Lancichinetti, Jevin D West, and Renaud Lambiotte. Memory in network flows and its effects on spreading dynamics and community detection. *Nature communications*, 5(1):1–13, 2014.

[18] Mandana Saebi, Jian Xu, Lance M. Kaplan, Bruno Ribeiro, and Nitesh V. Chawla. Efficient modeling of higher-order dependencies in networks: from algorithm to application for anomaly detection. *EPJ Data Sci.*, 9(1):15, 2020.

[19] Ingo Scholtes. 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*, page 1037–1046, New York, USA, 2017. ACM.

[20] Philipp Singer, Denis Helic, Behnam Taraghi, and Markus Strohmaier. Detecting memory and structure in human navigation patterns using markov chain models of varying order. *PLOS ONE*, 9(7):1–21, 07 2014.

[21] Leo Torres, Ann S Blevins, Danielle Bassett, and Tina Eliassi-Rad. The why, how, and when of representations for complex systems. *SIAM Review*, 63(3):435–485, 2021.

[22] Jian Xu, Mandana Saebi, Bruno Ribeiro, Lance M Kaplan, and Nitesh V Chawla. Detecting anomalies in sequential data with higher-order networks. *arXiv preprint arXiv:1712.09658*, 2017.

[23] Jian Xu, Thanuka L. Wickramarathne, and Nitesh V. Chawla. Representing higher-order dependencies in networks. *Science Advances*, 2(5), 2016.

[24] Hao Yin, Austin R. Benson, Jure Leskovec, and David F. Gleich. Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 555–564, New York, NY, USA, 2017. Association for Computing Machinery.