Knowledge and Information Systems (2023) 65:2347–2374
https://doi.org/10.1007/s10115-022-01800-7

REGULAR PAPER

CASPITA: mining statistically significant paths in time series data from an unknown network

Andrea Tonon[1] · **Fabio Vandin**[1]

Received: 28 January 2022 / Revised: 7 November 2022 / Accepted: 13 November 2022 /
Published online: 2 February 2023

Abstract
The mining of time series data has applications in several domains, and in many cases the data are generated by networks, with time series representing paths on such networks. In this work, we consider the scenario in which the dataset, i.e., a collection of time series, is generated by an *unknown* underlying network, and we study the problem of *mining statistically significant paths*, which are paths whose number of observed occurrences in the dataset is unexpected given the distribution defined by some features of the underlying network. A major challenge in such a problem is that the underlying network is unknown, and, thus, one cannot directly identify such paths. We then propose CASPITA, an algorithm to mine statistically significant paths in time series data generated by an unknown and underlying network that considers a generative null model based on meaningful characteristics of the observed dataset, while providing guarantees in terms of false discoveries. Our extensive evaluation on pseudo-artificial and real data shows that CASPITA is able to efficiently mine large sets of significant paths, while providing guarantees on the false positives.

Keywords Pattern mining · Statistically sound pattern mining · Time series · Graph mining

1 Introduction

Time series data mining [2–4] is a fundamental data mining task that covers a wide range of real-life problems in various fields of research such as economic forecasting, telecommunications, intrusion detection, and gene expression analysis. Even if the common purpose is to extract meaningful knowledge from the data, many different problems and approaches have

A preliminary version of this work appeared in the proceedings of IEEE ICDM'21 [1].
Caspita: Italian exclamation indicating surprise,
e.g., "Caspita! Such significant paths are really surprising.".

✉ Fabio Vandin
fabio.vandin@unipd.it

Andrea Tonon
andrea.tonon@dei.unipd.it

[1] Department of Information Engineering, University of Padova, 35131 Padua, Italy

been proposed over the years, ranging from anomaly detection [5, 6] to motif discovery [7], from clustering [8] to classification [9]. However, in many real-life scenarios, time series data are generated by networks, and thus represent paths constrained by the structures and the distributions that define such networks. Very often, one has access to a collection of time series but does not know the distribution on the network that generated them, or the structure of such network. As an example, consider a survey on the paths traveled by people with the underground service of a given city. In such a scenario, one has a dataset that represents a limited number of paths from a network, defined by the underground structure, but does not know the distribution defined by the entire population that uses such service.

In this work, we study the problem of mining statistically significant paths from an unknown network. We assume to have a time series dataset, defined as a collection of time series, and that such time series are paths generated from an unknown network. In such a scenario, we are interested in mining unexpected paths from the dataset, i.e., paths that appear in the dataset more or less than expected given the distribution of the underlying network. To find interesting paths, standard techniques usually directly use the frequency or the number of occurrences as extraction criteria but, in many real applications, such metrics are not enough to find paths that provide useful knowledge. For example, paths that appear only few times in a dataset may be over-represented if we consider the distribution of the network underlying the data, or vice versa, paths that appear a lot of times may be under-represented. Thus, techniques based on such metrics may led to several spurious discoveries. In addition, since we do not know the network underlying the data, we cannot directly find over- or under-represented paths.

We then introduce CASPITA, an algorithm to find statistically significant paths over- (or under-)represented from time series data considering a generative null model based on meaningful characteristics of the observed dataset, while providing guarantees in terms of the false positives employing the Westfall–Young (WY) method. Our generative null model is based on the observed number of occurrences of paths of a given length, with the idea that such paths represent well-known substructures of the underlying network. In the simplest case, they are paths of length one, that are edges of the underlying network, and thus the generative null model is exactly the underlying network that we are able to reconstruct from the dataset (i.e., containing only the edges appearing in the dataset). Then, such null model is used to test the significance of paths of a given, higher, length, which are the paths of interest mined from the observed dataset. The intuition is to create a generative null model that is able to explain the number of occurrences of shorter paths, and to check whether such generative null model is able to also explain the observed number of occurrences of the paths of interest. Otherwise, such paths can be considered significant, in the sense that they appear more (or less) times than expected under such generative null model.

Let us consider, as an example, a network composed by the web pages of a website, and let us suppose that the dataset $\mathcal{D}$ shown in Fig. 1 (left) contains paths that are sequences of web pages visited by some users in such a website. From $\mathcal{D}$, we may be interested in finding sequences of web pages of a given length visited more (or less) than expected with respect to the distribution of the underlying network, defined by the navigation of the whole population of users which visit the website. For example, we may be interested in mining significant over- or under-represented paths of length 3. Let us assume that the only paths of length 3 in $\mathcal{D}$ are $CBAD$ and $CBAC$, which occur, respectively, 10 and 90 times in $\mathcal{D}$. Since we do not know the underlying network, we cannot directly assess their significance and thus we first need to reconstruct the underlying network from $\mathcal{D}$. The 1-st order generative null model constructed from $\mathcal{D}$ is shown in Fig. 1 (right). Let us note that such generative null model is based on the number of occurrences of the paths of length 1 observed in $\mathcal{D}$, i.e., CB, BA, AD, and AC.

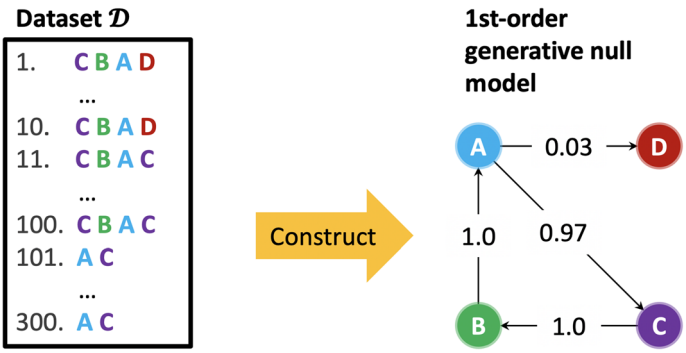

Fig. 1 Example of application of CASPITA. It shows a dataset $\mathcal{D}$ containing the following transactions: 10 $CBAD$, 90 $CBAC$, and 200 AC ($|\mathcal{D}| = 300$) (left) and the 1-st order generative null model constructed from $\mathcal{D}$ (right)

Since after C we always observed B, after B we always observed A, while after A we observed 10 times D and 290 times C, then the corresponding edges have weight, respectively, 1, 1, $10/300 \approx 0.03$, and $290/300 \approx 0.97$. As explained in Sect. 3, the generative null model is then used to check whether the observed number of occurrences of the paths of interest, i.e., 10 for $CBAD$ and 90 for $CBAC$, can be explained by the distribution described by the null model or if they are, instead, significant since they appear in $\mathcal{D}$ more (or less) times than expected. In particular, in this example, the path $CBAD$ results over-represented while the path $CBAC$ results under-represented. (Let us note that generating 100 paths of length 3 in accordance with the distribution described by the generative null model and starting from vertex C, the expected number of occurrences of $CBAD$ is 3 while the one of $CBAC$ is 97.) Let us notice that the path that appears more times is under-represented, while the path that appears only few times is over-represented, highlighting that the number of occurrences alone is not a useful extraction criterion in this scenario. Finally, while in this example we consider the 1-st order generative null model (and thus based on the number of occurrences of paths of length 1), given the application or the structure of the network, there may be some well-known substructures, defined as short sequences of web pages, that are traversed by the users that visit such website with a particular distribution. Thus, in such a scenario, one may be interested in finding whether such substructures also explain the number of observed occurrences of longer paths, by constructing from $\mathcal{D}$ a more general h-th order generative null model, with $h > 1$, or if such longer paths are significant due to some external factors causing their number of occurrences.

1.1 Our contributions

In this work, we introduce the problem of mining *statistically significant paths* in time series data from an unknown network. In this regard, our contributions are:

- We introduce the problem of mining statistically significant paths in time series data from an unknown network, defining a generative null model based on meaningful characteristics of the observed dataset.
- We introduce CASPITA, an algorithm to mine statistically significant paths (over- or under-represented) from a time series dataset, while providing guarantees on the proba-

bility of reporting at least one false positive, i.e., the FWER. We also discuss an extension of CASPITA to mine both types of paths at the same time.
– We introduce *g*-CASPITA, a variant of CASPITA to mine statistically significant paths (over- or under-represented) while providing guarantees on the *generalized* FWER, which allows to increase the statistical power of the algorithm, by tolerating the presence of a few false positives.
– We introduce an alternative interesting scenario in which CASPITA can be applied, which consists in mining paths that are significant with respect to a null model based on data from a different dataset.
– We perform an extensive suite of experiments that demonstrates that CASPITA is able to efficiently mine statistically significant paths in real datasets while providing guarantees on the false positives.

Let us note that throughout the paper, unless otherwise noted, we only describe the scenario in which one is interested in mining statistically significant paths that occur more times than expected under the null hypothesis (*over-represented paths*), for clarity of presentation. However, all the reasoning are still valid to mine paths that occur less times than expected (*under-represented paths*). In particular, we discuss the mining of under-represented paths in Sect. 3.5 and our open-source implementation of CASPITA mines over- and under-represented paths. In addition, results for both scenarios are shown in the experimental evaluation.

1.2 Related works

We now discuss the relation of our work to prior art on significant pattern mining, anomaly detection in sequential data, and temporal anomaly detection in graphs, which are the areas most related to our work. Since the nature of our work, we only consider unsupervised approaches.

In significant pattern mining, the dataset is seen as a collection of samples from an unknown distribution, and one is interested in finding patterns significantly deviating from an assumed *null hypothesis*, i.e., *distribution*. Many variants and algorithms have been proposed for the problem. We point interested reader to the survey [10], and recent works that employ permutation testing [11–13]. Even if our work falls within the framework of significant pattern mining, such approaches are orthogonal to our work, which focuses on finding significant paths, i.e., patterns, from time series that are constrained by a network structure.

Many works have been proposed to detect anomalies in sequential data [5, 14], employing several definition of anomalies, and considering different types of patterns. For example, [6] defines a pattern as *surprising* if its frequency differs substantially from that expected by chance, given some previously seen data. Lemmerich et al. [15], instead, consider the mining of subgroups, defined by subsets of attributes, that exhibit exceptional transition behavior, i.e., induce different transition models compared to the ones of the attributes that describe the entire dataset. Although our approach adopts a definition of significant pattern based on how the number of its occurrences differs from the one expected under an appropriate model, similarly to other works, we consider the setting in which the data represent paths from a weighted and directed graph, which results in a different problem. In fact, this aspect makes our work closer to the task of detecting anomalies in temporal graph [16, 17], i.e., graphs that evolve over time. However, even if our work considers data generated by a network and aims to find paths whose number of occurrences is significant with respect to the network's distribution, we consider the scenario in which we do not know the network, and we have only access to a sample.

The only work that considers the problem of finding anomalous paths in time series data from an unknown network is [18]. In this work, the authors propose an algorithm, HYPA, to find anomalous length k paths using a null model based on length $k-1$ paths. In particular, they aim to find length k paths whose number of occurrences in a dataset is anomalous with respect to a null model based on the number of occurrences of length $k-1$ paths in the same dataset. Reducing the difficult problem of detecting anomalous length k paths to the easier problem of detecting anomalous edges in a k-th order De Bruijn graph, they describe a strategy based on the hypergeometric distribution to compute a score for each length k path, where the score describes the level of anomaly of such a path. Even if our approach is inspired by [18], our work differs from it in many key aspects. First of all, we aim to find length k paths whose number of occurrences in a dataset is significant with respect to a null model based on the number of occurrences of length h paths, with $h \in \{1, \ldots, k-1\}$ provided in input by the user, and not only with $h = k-1$ as in [18]. In such a direction, it is not clear if HYPA can be modified to consider a more general length $h \in \{1, \ldots, k-1\}$. Finally, while our approach employs the WY method to correct for multiple hypothesis testing providing guarantees in terms of false positives, [18] uses fixed thresholds to define interesting patterns, which does not provide any guarantee.

To the best of our knowledge, our work is the first approach that employs the statistical hypothesis testing framework to mine paths, i.e., patterns, from time series constrained by the structure and the distribution of an unknown network, while providing rigorous guarantees on the probability of reporting at least one false positive, i.e., the FWER.

This version of our work differs in many ways from the preliminary one that appeared in the proceedings of IEEE ICDM'21 [1]. The major changes are the following:

- We provide a detailed explanation on how to mine under-represented paths. We also present a strategy to mine over- and under-represented paths at the same time, including also an approach to obtain false positives guarantees for both types of paths simultaneously (Sect. 3.5).
- We introduce g-CASPITA, a variant of CASPITA to mine statistically significant paths (over- or under-represented) while providing guarantees on the *generalized* FWER, which allows to increase the statistical power of the algorithm by tolerating the presence of a few false positives (Sect. 3.6).
- We extend our experimental evaluation, executing g-CASPITA on real and pseudo-artificial datasets, proving that it is able to increase the statistical power of CASPITA while controlling the generalized FWER (Sect. 5).
- We include additional explanations and experimental results for the extraction of over- and under-represented paths.

1.3 Organization of the paper

The rest of the paper is structured as follows. Section 2 contains the definitions and concepts used throughout this work. Section 3 describes our algorithm CASPITA, beside all related concepts, and the discussion of possible extensions of our approach. Section 4 describes an alternative scenario in which CASPITA can be applied, which considers two datasets. Section 5 reports the results of an extensive suite of experiments performed to evaluate the effectiveness of CASPITA on real and pseudo-artificial datasets. Section 6 concludes the paper with some final remarks.

2 Preliminaries

We now provide the definitions and concepts used in the paper. First, in Sect. 2.1, we describe the task of mining paths in time series data from a network. Then, in Sect. 2.2, we define, similarly to [19], the concept of k-th order De Bruijn graph used in the paper to define our generative null model. Finally, in Sect. 2.3, we describe concepts of hypothesis and multiple hypothesis testing for significant path mining.

2.1 Mining paths in time series data from a network

Let us define a *network* $N = (G, \omega)$ as a *directed graph* $G = (V, E)$ and a *weight function* $\omega : E \to [0, 1]$. $V = \{v_1, v_2, \ldots, v_{|V|}\}$ is the *vertices set*, where each $v \in V$ is called *vertex*, and $E = \{(u, v) : u, v \in V\}$ is the *edges set*, where each (u, v) is an *ordered* pair of vertices, called *edge*. An edge (u, v) is an *incoming edge* of the vertex v and an *outgoing edge* of the vertex u. Denoting with $(u, :)$ an outgoing edge of u, for each vertex $u \in V$, we have

$$\sum_{(u,:)\in E} \omega((u, :)) = 1,$$

that is, the weights of the edges from u represent a probability distribution. Figure 2 (left) shows an example of network.

A *path* $w = \{v_{i_0}, v_{i_1}, \ldots, v_{i_{|w|}}\}$ of length $|w|$ on the network N is an ordered sequence of $|w| + 1$ vertices such that $(v_{i_j}, v_{i_{j+1}}) \in E \ \forall j \in \{0, \ldots, |w| - 1\}$. Let us note that a vertex $v \in V$ is a path $w = \{v\}$ of length $|w| = 0$. A path $w = \{w_0, w_1, \ldots, w_{|w|}\}$ *occurs* in a path $q = \{q_0, q_1, \ldots, q_{|q|}\}$ starting from position $s \in \{0, \ldots, |q| - |w|\}$, denoted by $w \subset q^{(s)}$, if and only if $w_0 = q_s, w_1 = q_{s+1}, \ldots, w_{|w|} = q_{s+|w|}$. We say that the path w is a *sub-path* of the path q. The number of *occurrences* $Occ_q(w)$ of w in q is the number of times that w occurs in q, that is,

$$Occ_q(w) = |\{s \in \{0, \ldots, |q| - |w|\} : w \subset q^{(s)}\}|.$$

A *time series dataset* $\mathcal{D} = \{\tau_1, \tau_2, \ldots, \tau_{|\mathcal{D}|}\}$ from a network N is a bag of $|\mathcal{D}|$ *transactions*, which are paths on N. Given a path w on N, the number of occurrences $Occ_{\mathcal{D}}(w)$ of w in $\mathcal{D}$ is the sum of the number of occurrences $Occ_\tau(w)$ of w in $\tau, \forall\ \tau \in \mathcal{D}$, that is,

$$Occ_{\mathcal{D}}(w) = \sum_{\tau\in\mathcal{D}} Occ_\tau(w).$$

Given a positive integer ℓ, the task of *mining paths* of length ℓ from a time series dataset $\mathcal{D}$ from a network N is the task of mining the set $\mathcal{W}_{\mathcal{D}}(\ell)$ of all paths of length ℓ that occur at least once in $\mathcal{D}$ and the number of their occurrences, that is,

$$\mathcal{W}_{\mathcal{D}}(\ell) = \{(w, Occ_{\mathcal{D}}(w)) : |w| = \ell \wedge Occ_{\mathcal{D}}(w) > 0\}.$$

With an abuse of notation, in the following we use $w \in \mathcal{W}_{\mathcal{D}}(\ell)$ to indicate that $\exists (w, Occ_{\mathcal{D}}(w)) \in \mathcal{W}_{\mathcal{D}}(\ell)$.

2.2 *k*-th order De Bruijn graph

Given a directed graph $G = (V, E)$ and an integer $k > 0$, the *k-th order De Bruijn graph* $G^k = (V^k, E^k)$ of G is a directed graph where each vertex $v^k \in V^k$ is a path of length $k-1$ on

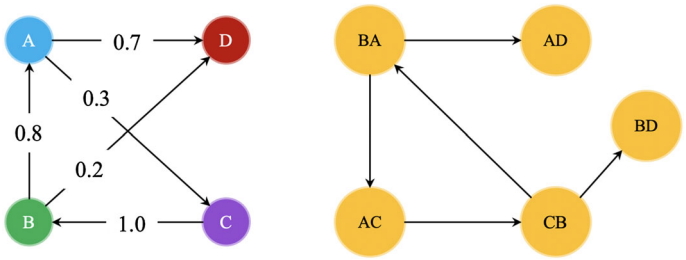

Fig. 2 Example of network and of De Bruijn graph. It shows the network $N = (G, \omega)$ (left) and the 2-nd order De Bruijn graph $G^2 = (V^2, E^2)$ of G (right)

G, i.e., $v^k = \{v_{i_0}, v_{i_1}, \ldots, v_{i_{k-1}}\}$, and an ordered pair (v^k, u^k), with $v^k = \{v_{i_0}, v_{i_1}, \ldots, v_{i_{k-1}}\}$, $u^k = \{u_{j_0}, u_{j_1}, \ldots, u_{j_{k-1}}\} \in V^k$, it is an edge of G^k if and only if $v_{i_t} = u_{j_{t-1}}$ $\forall t \in \{1, \ldots, k-1\}$. Thus, each edge $(v^k, u^k) \in E^k$ is a path of length k on G, since $(v^k, u^k) = \{v_{i_0}, v_{i_1} = u_{j_0}, v_{i_2} = u_{j_1}, \ldots, v_{i_{k-1}} = u_{j_{k-2}}, u_{j_{k-1}}\}$. Let us note that G itself is a 1-st order De Bruijn graph of G. Figure 2 (right) shows an example of k-th order De Bruijn graph.

Example 1 Let us consider as an example the network $N = (G, \omega)$ (left) and the 2-nd order De Bruijn graph $G^2 = (V^2, E^2)$ of G (right), both shown in Fig. 2. The network $N = (G, \omega)$ is composed by the directed graph $G = (V, E)$, with $V = \{A, B, C, D\}$ and $E = \{(A, C), (A, D), (B, A), (B, D), (C, B)\}$, and by the weight function ω, such that $\omega((A, C)) = 0.3, \omega((A, D)) = 0.7, \omega((B, A)) = 0.8, \omega((B, D)) = 0.2$, and $\omega((C, B)) = 1.0$. The paths $w = CBAC$ and $q = BAD$ are examples of paths on N, respectively, of length $|w| = 3$ and $|q| = 2$. The 2-nd order De Bruijn graph $G^2 = (V^2, E^2)$ of G is composed by $V^2 = \{AC, AD, BA, BD, CB\}$, where each vertex $v^2 \in V^2$ represents a path of length 1 on G, and by $E^2 = \{(AC, CB), (BA, AC), (BA, AD), (CB, BA), (CB, BD)\}$, where each edge $(v^2, u^2) \in E^2$ represents a path of length 2 on G, i.e., ACB, BAC, BAD, CBA, and CBD.

2.3 Multiple hypothesis testing for paths

The task of mining statistically significant paths requires to identify paths whose number of occurrences in a dataset $\mathcal{D}$ is *significant*, or unexpected, with respect to the distribution of the weight function of the network that generated such data. To assess the significance of a path, a common solution is to employ the framework of statistical hypothesis testing. For each path w, let H_w be the null hypothesis that the number of occurrences $Occ_{\mathcal{D}}(w)$ of w on $\mathcal{D}$ well conforms to the number of its occurrences in *random* time series data generated from the network $N = (G, \omega)$. In particular, we define a *random dataset* $\tilde{\mathcal{D}}$ as a collection of random transactions which contains the same number of paths of interest of the original dataset $\mathcal{D}$ and that is generated from the graph G in accordance with the weight function ω. That is, a path from a vertex $u \in V$ continues with a vertex $v \in V$ with probability $\omega(u, v)$. In addition, the starting vertices (of full transactions or of single paths) observed in the original dataset $\mathcal{D}$ are preserved in the random dataset $\tilde{\mathcal{D}}$.

Under the null hypothesis, the number of occurrences of w is described by a random variable X_w, and in order to assess the significance of w, a p-value p_w is commonly computed. The p-value p_w of w is the probability of observing a number of occurrences, under the null

hypothesis, at least as large as the number of occurrences $Occ_{\mathcal{D}}(w)$ of w in $\mathcal{D}$, that is,

$$p_w = \Pr\left[X_w \geq Occ_{\mathcal{D}}(w) | H_w\right].$$

For complex null hypotheses, the p-values cannot be computed analytically, since there is not a closed form for X_w. However, when one can generate random data from the distribution described by the null hypothesis, the p-values can be estimated by a simple Monte Carlo (MC) procedure as follows: to generate M random time series datasets $\tilde{\mathcal{D}}_i$, with $i \in \{1, \ldots, M\}$, from the distribution described by the null hypothesis. Then, the p-value p_w is estimated as

$$p_w = \frac{1}{M+1}\left(1 + \sum_{i=1}^{M} \mathbb{1}\left[Occ_{\tilde{\mathcal{D}}_i}(w) \geq Occ_{\mathcal{D}}(w)\right]\right), \tag{1}$$

where $\mathbb{1}[\cdot]$ is the indicator function of value 1 if the argument is true, and 0 otherwise.

The statistical hypothesis testing framework is commonly used to provide guarantees on the false discoveries, i.e., paths flagged as significant while they are not. When a single path w is tested for significance, flagging w as significant, i.e., rejecting the null hypothesis, when $p_w \leq \alpha$, where $\alpha \in (0, 1)$ is the *significance threshold* fixed by the user, guarantees that the probability that w corresponds to a false discovery $\leq \alpha$.

The situation is completely different when several paths are tested simultaneously, as in the case of path mining. If d paths are tested with the approach used for a single path, i.e., each path is flagged as significant if its p-value is $\leq \alpha$, then the expected number of false discoveries can be as large as αd. To solve this issue, one identifies a *corrected significance threshold* $\delta \in (0, 1)$ such that all paths with p-value $\leq \delta$ can be reported as significant while providing some guarantees on the number of false discoveries. A common approach is to identify δ that provides guarantees on the family-wise error rate (FWER), defined as the probability of reporting at least one false positive, that is, if FP is the number of false positives, then FWER $= \Pr[\text{FP} > 0]$. For a given value δ, let FWER(δ) be the FWER obtained when δ is used as corrected significance threshold, that is, by reporting as significant all paths with p-value $\leq \delta$. Often FWER(δ) cannot be evaluated in closed form, and thus approaches, as the Bonferroni correction or based on permutation testing, described below, must be employed.

A simple approach to correct for multiple hypothesis testing is to use the *Bonferroni correction* [20], setting $\delta = \alpha/d$. Using the union bound, it is easy to show that the resulting FWER satisfies FWER(δ) $\leq d\delta = \alpha$. However, to properly correct for multiple hypothesis testing, one must consider the number of all paths that can be generated from the distribution described by the null hypothesis, and when such number d is very large, as in the case of path mining, δ is very close to 0, resulting in low *statistical power* and many *false negatives*, i.e., significant paths that are not correctly reported in output.

The Westfall–Young (WY) method [21] is a multiple hypothesis testing procedure based on permutation testing that results in high statistical power and that has been successfully applied in other pattern mining scenarios [11–13]. The WY method directly estimates the joint distribution of null hypotheses using permuted datasets, i.e., datasets obtained from the distribution described by the null hypothesis. In detail, the WY method considers P random datasets $\tilde{\mathcal{D}}_i$, with $i \in \{1, \ldots, P\}$, generated from the distribution described by the null hypothesis. Then, for every dataset $\tilde{\mathcal{D}}_i$, with $i \in \{1, \ldots, P\}$, it computes the minimum p-value $p_{min}^{(i)}$ over all paths of interest in $\tilde{\mathcal{D}}_i$. The FWER FWER(δ) obtained using δ as corrected significance threshold can then be estimated as

$$\text{FWER}(\delta) = \frac{1}{P}\sum_{i=1}^{P} \mathbb{1}\left[p_{min}^{(i)} \leq \delta\right]. \tag{2}$$

Thus, given a *FWER threshold* $\alpha \in (0, 1)$, the corrected significance threshold δ^* is obtained as

$$\delta^* = \max\{\delta : \text{FWER}(\delta) \leq \alpha\}. \tag{3}$$

3 CASPITA: mining statistically significant paths

In this section, we describe our method CASPITA, mining statistiCAlly Significant Paths In Time series dAta, to mine *statistically significant paths* in time series data generated by a network, while controlling the probability of having at least one false discovery, i.e., the FWER. Given a time series dataset $\mathcal{D}$ from an *unknown* network N, we aim to mine statistically significant paths, which are paths that have a number of occurrences on $\mathcal{D}$ that is surprising, i.e., higher than the expected number of their occurrences under the null hypothesis. In particular, given two natural values $k, h \in \mathbb{N}^+$, with $k > h$, we aim to mine length k paths from $\mathcal{D}$ whose number of occurrences are not due to the number of occurrences of length h paths observed in $\mathcal{D}$, with the idea that such paths of length h represent some well-known substructures in the underlying and unknown network.

The idea behind CASPITA is the following. First, we mine all the paths $\mathcal{W}_{\mathcal{D}}(k)$ of length k from the time series dataset $\mathcal{D}$. Since we do not know the network N from which $\mathcal{D}$ has been generated, we cannot directly infer the statistical significance of such paths with respect to N, and thus we need to construct a new network, i.e., a generative null model, from the dataset $\mathcal{D}$. Such generative null model is then used to generate random time series datasets in order to estimate the p-values and to compute the corrected significance threshold δ^* using the WY method. We now describe the generative null model employed by CASPITA.

3.1 Generative null model

In this work, we aim to find length k paths whose number of occurrences in $\mathcal{D}$ are not due to the number of occurrences of shorter length h paths in $\mathcal{D}$. Thus, we construct a generative null model in accordance with the number of occurrences of the paths of length h in $\mathcal{D}$, and then we test the significance of the paths of length k using such model. Given a time series dataset $\mathcal{D}$, generated by an unknown network $N = (G, \omega)$, and $h \in \mathbb{N}^+$, we define the *h-th order generative model* $N^h(\mathcal{D})$ of the time series dataset $\mathcal{D}$ as a network $N^h(\mathcal{D}) = (G^h, \omega^h)$, where G^h is the h-th order De Bruijn graph of G (based on $\mathcal{D}$, since the entire structure of G is unknown), and ω^h is a weight function. The h-th order De Bruijn graph $G^h = (V^h, E^h)$ is composed as follows: $V^h = \{w \in \mathcal{W}_{\mathcal{D}}(h-1)\}$, while E^h is constructed as defined in Definition 2.2. Thus, each vertex $v^h \in V^h$ is a path of length $h-1$ in $\mathcal{D}$ (and thus on G), while each edge $(u^h, v^h) \in E^h$ represents a path of length h in $\mathcal{D}$ (and thus on G). With an abuse of notation, in the following we use (u^h, v^h) to indicate both the edge in G^h and the corresponding path on G. Finally, the weight function ω^h is defined as follows: $\forall (u^h, v^h) \in E^h$,

$$\omega^h\left((u^h, v^h)\right) = \frac{Occ_{\mathcal{D}}\left((u^h, v^h)\right)}{\sum_{(u^h,:)\in E^h} Occ_{\mathcal{D}}\left((u^h, :)\right)}.$$

Let us note that the weight function ω^h of the h-th order generative null model $N^h(\mathcal{D})$ is defined using the observed number of occurrences of length h paths, and that each edge of $N^h(\mathcal{D})$ represents a path of length h. Thus, $N^h(\mathcal{D})$ correctly represents the distribution of the

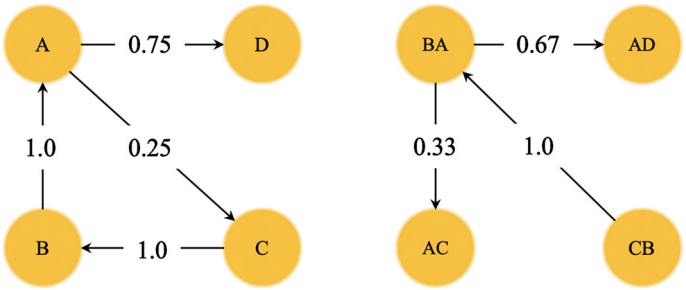

Fig. 3 Example of generative null models. It shows $N^1(\mathcal{D})$ (left) and $N^2(\mathcal{D})$ (right), respectively, the 1-st and the 2-nd order generative null model of $\mathcal{D} = \{\tau_1 = BAD, \tau_2 = CBAC, \tau_3 = CBAD, \tau_4 = AD\}$

number of occurrences of the paths of length h in the dataset $\mathcal{D}$. An example of generative null models is shown in Fig. 3.

Example 2 Let us consider the following dataset $\mathcal{D} = \{\tau_1, \tau_2, \tau_3, \tau_4\}$, which is a possible time series dataset from the network N shown in Fig. 2 (left), as an example:

$$\begin{aligned}\tau_1 &= BAD\\ \tau_2 &= CBAC\\ \tau_3 &= CBAD\\ \tau_4 &= AD.\end{aligned}$$

Fig. 4 shows $N^1(\mathcal{D})$ (left) and $N^2(\mathcal{D})$ (right), respectively, the 1-st and the 2-nd order generative null model of $\mathcal{D}$. Let us note that the two generative models represent different probability distributions for the paths of length $> h$, e.g., the path $w = BAD$ has a probability of 0.75 of being generated in $N^1(\mathcal{D})$ while it has a probability of 0.67 in $N^2(\mathcal{D})$. In addition, let us note that they are based on the dataset $\mathcal{D}$ and not on the network N that generated $\mathcal{D}$. Indeed, they have some missing edges with respect to the network N (that is a 1-st order De Bruijn graph of itself) and its 2-nd order De Bruijn graph G^2, respectively, both shown in Fig. 2.

As explained in Sect. 2.3, to compute the p-values p_w of the paths $w \in \mathcal{W}_{\mathcal{D}}(k)$ and to estimate the corrected significance threshold δ^*, we require random data generated from the generative null model. In the following two sections, we introduce two different strategies to generate random datasets $\tilde{\mathcal{D}}$ that contain the same total number T of paths of length k of the original dataset $\mathcal{D}$, that is,

$$T = \sum_{w \in \mathcal{W}_{\mathcal{D}}(k)} Occ_{\mathcal{D}}(w) = \sum_{w \in \mathcal{W}_{\tilde{\mathcal{D}}(k)}} Occ_{\tilde{\mathcal{D}}}(w). \tag{4}$$

First, we describe the *transactions oriented generation* (TOG) strategy, a natural way to generate random datasets performing a series of random walks that generate random transactions with characteristics similar to the ones of the transactions in $\mathcal{D}$. To overcome some issues of this strategy when it is applied to large generative null model, we then introduce the *paths oriented generation* (POG) strategy, an alternative approach that directly generates random length k paths. For this second strategy, we also introduce an approximation based on the binomial distribution that allows to estimate the p-values avoiding expensive MC procedures.

3.2 Transactions oriented generation (TOG) strategy

In this section, we explain how to generate random datasets $\tilde{\mathcal{D}}$ from the generative null model $N^h(\mathcal{D})$ defined above using the TOG strategy. The idea is to perform a series of random walks that generate random transactions $\tilde{\tau}$ with characteristics similar to the ones of the transactions $\tau \in \mathcal{D}$. Characteristics that are natural to consider and that we want to preserve are:

- the dataset $\tilde{\mathcal{D}}$ has the same number of transactions of $\mathcal{D}$;
- each transaction $\tilde{\tau} \in \tilde{\mathcal{D}}$ has the same length of the corresponding transaction $\tau \in \mathcal{D}$;
- each transaction $\tilde{\tau} \in \tilde{\mathcal{D}}$ starts from the same vertex (of the generative null model) of the corresponding transaction $\tau \in \mathcal{D}$.

Let us note that to preserve such characteristics guarantees to preserve also the total number T of length k paths in the dataset. As a motivation to consider such characteristics, let us consider the case in which the dataset $\mathcal{D}$ contains transactions that represent visits of some users in a website. In such a scenario, we are interested in preserving the web pages from which the visits start, since they probably are homepages (or web pages from which the users typically start their navigation). In addition, by preserving the length of the transactions, we preserve the number of web pages that the users visit in a single navigation on the website.

Let s_τ be a path of length $|s_\tau| = h-1$ such that $s_\tau \subset \tau^{(0)}$, that is, s_τ is the vertex of $N^h(\mathcal{D})$ from which the transaction τ starts. The TOG strategy is the following. For each $\tau_i \in \mathcal{D}$, we perform a random walk on $N^h(\mathcal{D})$ of $|\tau_i| - (h-1)$ steps, starting from the vertex s_{τ_i}. At each step, the random walk moves from a vertex v^h to a vertex u^h with probability $\omega^h\left((v^h, u^h)\right)$. The path generated from such random walk is then the transaction $\tilde{\tau}_i \in \tilde{\mathcal{D}}$ that corresponds to the transaction $\tau_i \in \mathcal{D}$, with $|\tau_i| = |\tilde{\tau}_i|$. Performing all the $|\mathcal{D}|$ random walks, we generate the random dataset $\tilde{\mathcal{D}}$. Let us note that a random walk may reach a vertex without outgoing edges before performing the desired number of steps, generating a shorter transaction, and thus not preserving the second characteristic (and neither T). In such a case, we discard the transaction and repeat the random walk until we generate a transaction $\tilde{\tau}_i$ with $|\tilde{\tau}_i| = |\tau_i|$.

The TOG strategy is the most natural way to generate random data from the generative null model. However, when the generative null model is large, as happens in many real applications, the number of paths contained in a dataset is only a small fraction of the gargantuan number of paths that can be generated as sub-paths of such long transactions. Thus, the corrected significance threshold δ^* obtained with the WY method could be very small, resulting in few or even zero reported significant paths. In addition, depending on the structure of the generative null model, to generate such long transactions may be computationally expensive for the high number of transactions that we need to generate and discard before reaching the desired lengths.

3.3 Paths oriented generation (POG) strategy

To overcome the issue of the TOG strategy, we now describe an alternative approach to generate random data. In the POG strategy, instead of generating long transactions, we generate single random paths w of length $|w| = k$. In particular, from the generative null model $N^h(\mathcal{D})$ defined above, we generate random datasets $\tilde{\mathcal{D}}$, which are bags of paths of length k, where the number of paths w of length $|w| = k$ that start in each vertex (of the generative null model) is the same in the two datasets $\mathcal{D}$ and $\tilde{\mathcal{D}}$. Let us note that to preserve such characteristic guarantees to also preserve the total number T of length k paths in the dataset.

Let us remember that s_w is a path of length $|s_w| = h-1$ such that $s_w \subset w^{(0)}$, that is, s_w is the vertex of $N^h(\mathcal{D})$ from which the path w starts, and let $\mathcal{S} = \{s_w : w \in \mathcal{W}_\mathcal{D}(k)\}$ be the set

of vertices of $N^h(\mathcal{D})$ from which starts at least one path $w \in \mathcal{W}_{\mathcal{D}}(k)$. To generate random paths, for each vertex $s \in \mathcal{S}$, we perform a series of random walks of $k - (h-1)$ steps on $N^h(\mathcal{D})$, until we generate

$$n_s = \sum_{w \in \mathcal{W}_{\mathcal{D}}(k): s_w = s} Occ_{\mathcal{D}}(w) \tag{5}$$

random paths w of length $|w| = k$ that start from such vertex s. Then, the bag of all the paths of length k generated from all the vertices $s \in \mathcal{S}$ is the random dataset $\tilde{\mathcal{D}}$. Let us note that $|\tilde{\mathcal{D}}| = \sum_{s \in \mathcal{S}} n_s = T$. As explained above, let us remember that at each step the random walk moves from a vertex v^h to a vertex u^h with probability $\omega^h((v^h, u^h))$. Thus, each random walk generates a path of length k or a path of length $< k$ that ends in a vertex without outgoing edges. Since we are interested in paths of length k, we discard all generated paths of length shorter than k.

While the POG strategy overcomes the issue of the TOG strategy explained above reducing the space of paths that can be generated from the generative null model, it still requires expensive MC procedures to estimate the p-values, and such procedures could be computationally prohibitive for large datasets. In the following section, we introduce a method to approximate the p-values for the POG strategy avoiding the MC procedure.

3.3.1 Binomial approximation for the *p* values

In this section, we illustrate an approach to approximate the p-values p_w of paths w of length $|w| = k$ when the POG strategy is used to generate random data. First, we compute with which probabilities such paths are generated under the POG strategy. Let us consider a random walk that starts from a vertex $s \in \mathcal{S}$ and that performs $k - (h-1)$ steps on $N^h(\mathcal{D})$, and let $\mathcal{W}_s$ be the set of all paths that can be generated by such random walk. As explained above, the set $\mathcal{W}_s$ contains paths of length $|w| = k$ and, eventually, paths of length $|w| < k$ that end in a vertex without outgoing edges. Let $RW(w)$ be the set of edges of $N^h(\mathcal{D})$ that the random walk traverses to generate the path $w \in \mathcal{W}_s$. From the definition of random walk, the probability $\Pr(w)$ that the random walk generates $w \in \mathcal{W}_s$ starting from s is

$$\Pr(w) = \prod_{(u^h, v^h) \in RW(w)} \omega^h\left((u^h, v^h)\right).$$

Let us note that $\sum_{w \in \mathcal{W}_s} \Pr(w) = 1$. Let E_k be the event that the random walk starting from s generates a path of length exactly k and let $\mathcal{W}_s^k \subseteq \mathcal{W}_s$ be the set of paths $w \in \mathcal{W}_s$ with $|w| = k$. Since in the POG strategy we discard paths shorter than k which could be generated by the series of random walks, then the probability of generating the path $w \in \mathcal{W}_s^k$ is

$$\Pr(w \mid E_k) = \frac{\Pr(w \cap E_k)}{\Pr(E_k)}, \tag{6}$$

where $\Pr(w \cap E_k) = \Pr(w)$ for all $w \in \mathcal{W}_s^k$ and 0 otherwise, and $\Pr(E_k) = \sum_{w \in \mathcal{W}_s^k} \Pr(w)$. Again, let us note that $\sum_{w \in \mathcal{W}_s^k} \Pr(w \mid E_k) = 1$, and that if $\mathcal{W}_s \backslash \mathcal{W}_s^k = \emptyset$, then $\Pr(w \mid E_k) = \Pr(w)$.

Example 3 Let us consider the 2-nd order generative null model $N^2(\mathcal{D})$ shown in Fig. 4 (left), as an example. For $k = 3$, starting from the vertex BA and performing $k - (h-1) = 2$ steps, a random walk can generate the paths of length 3 $BAAE$, $BAAB$, and $BACE$, or can

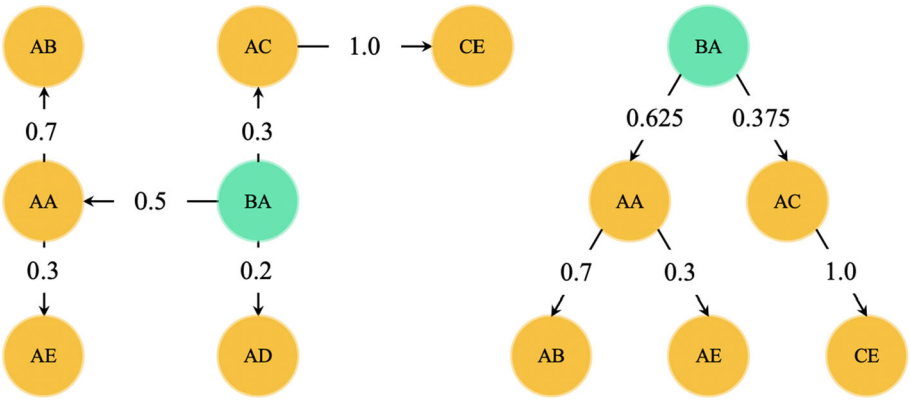

Fig. 4 Example of paths oriented generation. It shows the 2-nd order generative null model $N^2(\mathcal{D})$ and the starting vertex BA (left), and the probabilities of all paths of length 3 under the POG strategy (right)

reach the vertex AD just after one step, generating the path of length 2 BAD. The probabilities of all these paths are: $\Pr(BAAE) = 0.15$, $\Pr(BAAB) = 0.35$, $\Pr(BACE) = 0.3$, and $\Pr(BAD) = 0.2$, and then the probability of generating a path of length 3 starting from the vertex BA is $\Pr(E_3) = 0.8$. Thus, the probabilities of the length 3 paths under the POG strategy are: $\Pr(BAAE \mid E_3) = 0.1875$, $\Pr(BAAB \mid E_3) = 0.4375$, and $\Pr(BACE \mid E_3) = 0.375$, shown in Fig. 4 (right).

Since from a given vertex $s \in \mathcal{S}$, we generate exactly n_s (see Eq. 5) length k paths, then the number of occurrences $Occ_{\tilde{\mathcal{D}}}(w)$ of a path $w \in \mathcal{W}_s^k$ in the random dataset $\tilde{\mathcal{D}}$ follows a binomial distribution, that is, $Occ_{\tilde{\mathcal{D}}}(w) \sim \text{Bin}(n_s, \Pr(w \mid E_k))$. For a fixed vertex $s \in \mathcal{S}$, this is true for all the paths $w \in \mathcal{W}_s^k$, but the binomial distributions corresponding to these paths are not independent, and thus, the computation of the p-values p_w as

$$p_w = \Pr\left[\text{Bin}(n_s, \Pr(w \mid E_k)) \geq Occ_{\mathcal{D}}(w)\right] \tag{7}$$

considers a number of paths that is in *expectation* n_s, and not exactly n_s as for the original POG strategy. (Let us note that, as a consequence, the *total* number of considered paths of length k is T *in expectation*.) However, in our experimental evaluation, we empirically show that the p-values for the binomial approximation are within one order of magnitude of the corresponding MC p-values, and, thus, that the binomial approximation is a valid approach to approximate the p-values for the POG strategy, avoiding expensive MC procedures.

Let us note that while this approximation does not require the generation of M random datasets $\tilde{\mathcal{D}}$ to estimate the p-values, CASPITA still requires the generation of P random datasets $\tilde{\mathcal{D}}$ for the WY method. However, the binomial approximation can also be used to approximate the minimum p-value in the P random datasets. Thus, given the observed dataset $\mathcal{D}$, we compute the p-values for all paths $w \in \mathcal{W}_{\mathcal{D}}(k)$ using Eq. 7. Then, we generate a series of P random datasets $\tilde{\mathcal{D}}$ required by the WY method using the POG strategy. For all the P random datasets $\tilde{\mathcal{D}}$, we compute the minimum p-value over all the paths $w \in \mathcal{W}_{\tilde{\mathcal{D}}}(k)$, where the p-value p_w of w is computed with Eq. 7 replacing $Occ_{\mathcal{D}}(w)$ with $Occ_{\tilde{\mathcal{D}}}(w)$.

3.4 Analysis

In this section, we describe in detail our algorithm CASPITA and formally prove its false positives guarantees. Algorithm 1 shows the pseudo-code of CASPITA. Its inputs are the time series dataset $\mathcal{D}$, the FWER threshold $\alpha \in (0, 1)$, the order $h > 0$ of the generative null model, and the paths length $k > h$. For a given generation strategy, (i.e., TOG or POG), CASPITA first mines the set $\mathcal{W}_{\mathcal{D}}(k)$, of all paths w of length $|w| = k$ that occur at least once in $\mathcal{D}$ (line 1). Then, it constructs the generative null model $N^h(\mathcal{D})$ (line 2) as explained in Sect. 3.1, and it uses $N^h(\mathcal{D})$ to compute the p-values of the paths $w \in \mathcal{W}_{\mathcal{D}}(k)$ (lines 3-4). The p-values can be computed with a MC procedure using Eq. 1 (for both generation strategies), and thus generating M random datasets, where M is a parameter set by the user, or with the binomial approximation using Eq. 7 (for the POG strategy). To compute the corrected significance threshold δ^*, it then employs the WY method, which requires the generation of P random datasets (lines 5-6), where P is a parameter set by the user. For each random dataset $\tilde{\mathcal{D}}_i$, with $i \in \{1, \ldots, P\}$, it mines the set $\mathcal{W}_{\tilde{\mathcal{D}}_i}(k)$ (line 7) and then it computes the minimum p-value $p^{(i)}_{min}$ over all paths $\tilde{w} \in \mathcal{W}_{\tilde{\mathcal{D}}_i}(k)$ (lines 8-10). For the computation of such p-values, the considerations made above are still valid. The corrected significance threshold δ^* is then computed using Eq. 3 (line 11). If $\delta^* > \alpha$, then we set $\delta^* = \alpha$, corresponding to an uncorrected threshold. Finally, the output is the set of paths $w \in \mathcal{W}_{\mathcal{D}}(k)$ such that $p_w < \delta^*$ (line 12). Theorem 1 proves that the output of CASPITA has FWER $\leq \alpha$.

Algorithm 1: CASPITA

Data: Time Series Dataset $\mathcal{D}$, FWER Threshold $\alpha \in (0, 1)$, Order $h > 0$ of the Generative Null Model, Paths Length $k > h$.
Result: Set $\mathcal{SW}$ with FWER $\leq \alpha$.

1 $\mathcal{W} \leftarrow$ `MinePaths`$(\mathcal{D}, k)$;
2 $N^h \leftarrow$ `GenerativeNullModel`$(\mathcal{D}, h)$;
3 **foreach** $w \in \mathcal{W}$ **do**
4 $\quad p_w \leftarrow$ `PValue`$\left(N^h, w, Occ_{\mathcal{D}}(w)\right)$;
5 **for** $i \leftarrow 1$ **to** P **do**
6 $\quad \tilde{\mathcal{D}}_i \leftarrow$ `RandomDataset`(N^h, k, h);
7 $\quad \mathcal{W}_i \leftarrow$ `MinePaths`$(\tilde{\mathcal{D}}_i, k)$;
8 $\quad$ **foreach** $\tilde{w} \in \mathcal{W}_i$ **do**
9 $\quad\quad p_{\tilde{w}} \leftarrow$ `PValue`$\left(N^h, \tilde{w}, Occ_{\tilde{\mathcal{D}}_i}(\tilde{w})\right)$;
10 $\quad p^{(i)}_{min} \leftarrow \min\{p_{\tilde{w}} : \tilde{w} \in \mathcal{W}_i\}$;
11 $\delta^* \leftarrow \max\left\{\delta : \sum_{i=1}^{P}\left(\mathbb{1}[p^{(i)}_{\min} \leq \delta]\right) \leq \alpha P\right\}$;
12 $\mathcal{SW} \leftarrow \{\left(w, Occ_{\mathcal{D}}(w), p_w\right) : w \in \mathcal{W} \wedge p_w < \delta^*\}$;
13 **return** $\mathcal{SW}$;

Theorem 1 *The output of* CASPITA *has FWER* $\leq \alpha$.

Proof Let us consider the P random datasets $\tilde{\mathcal{D}}_i$, with $i \in \{1, \ldots, P\}$, generated by CASPITA for the WY method. Let us note that they do not contain any significant paths of length k, since they are generated from the generative null model N^h, and thus from the distribution described by the null hypothesis. Given $\delta \in (0, 1)$, the FWER FWER(δ) obtained using δ as significance threshold can be estimated using Eq. 2. That is, estimated as the fraction, over P,

of the number of datasets $\tilde{\mathcal{D}}_i$ that contain at least one path with p-value $\leq \delta$, and thus a path that would be reported as significant while it is not when δ is used as significance threshold. Since CASPITA uses the corrected significance threshold $\delta^* = \max\{\delta : \text{FWER}(\delta) \leq \alpha\}$, then its output has FWER $\leq \alpha$, which concludes our proof. □

We now provide a brief analysis of the time complexity of CASPITA. The time complexity t_W to mine $\mathcal{W}_{\mathcal{D}}(k)$ and t_N to construct $N^h(\mathcal{D})$ are $t_W = t_N = \mathcal{O}(D)$, with $D = \sum_{\tau \in \mathcal{D}} |\tau|$, since they can be done with a single scan of the entire dataset. The time complexity t_P^{MC} to estimate the p-values of all paths $w \in \mathcal{W}_{\mathcal{D}}(k)$ using MC procedures is $t_P^{MC} = \mathcal{O}(M \cdot t_{\tilde{\mathcal{D}}} + |\mathcal{W}_{\mathcal{D}}(k)|)$, where $t_{\tilde{\mathcal{D}}}$ is the time complexity to generate a random dataset $\tilde{\mathcal{D}}$ and M is the number of random datasets to create. Let us note that $t_{\tilde{\mathcal{D}}}$ depends on the generation strategy, and that it is not trivial to bound it since we do not know in advance the number of random walks that we need to generate $\tilde{\mathcal{D}}$. However, our experimental evaluation empirically proves that such random datasets can be generated with feasible computational time. In addition, the MC procedure is well-suited to parallelization: when C cores are used to compute the p-values considering M random datasets, each core computes the p-values on M/C random datasets, and the results are then aggregated at the end. The time complexity t_P^B to compute the p-values of all paths $w \in \mathcal{W}_{\mathcal{D}}(k)$ using the binomial approximation is instead $t_P^B = \mathcal{O}(|\mathcal{W}_{\mathcal{D}}(k)|)$, but it first requires the computation of the probabilities of Eq. 6. Such computation has time complexity $\mathcal{O}(|\mathcal{S}| \cdot R^{MAX})$, where $|\mathcal{S}|$ is the number of starting nodes and R^{MAX} is the maximum, over all $s \in \mathcal{S}$, number of vertices that can be reached in $k - h$ steps on $N^h(\mathcal{D})$, starting from each vertex s. Finally, the time complexity of the WY method is $t_{WY} = \mathcal{O}(P \cdot (t_{\tilde{\mathcal{D}}} + t_P))$, with P the number of random datasets to generate and t_P one of the two time complexity described above to compute the p-values.

As previously stated, while here we consider the mining of over-represented paths, all our reasoning can be easily adapted to the mining of under-represented paths: the details are provided in the next section.

3.5 Mining under-represented paths

In this section, we describe how CASPITA can be modified to mine under-represented statistically significant paths. Let us remember that in order to assess the significance of a path w, we require to compute a p-value p_w. To mine under-represented paths, the p-value p_w of w is the probability of observing a number of occurrences, under the null hypothesis, at least as small as the number of occurrences $Occ_{\mathcal{D}}(w)$ of w in $\mathcal{D}$, that is,

$$p_w = \Pr\left[X_w \leq Occ_{\mathcal{D}}(w) | H_w\right],$$

where X_w is the random variable which describes the the number of occurrences of w under the null hypothesis. As already said for the over-represented paths, since there is not a closed formula for X_w under our null hypotheses, the p-value p_w must be estimated with a MC procedure (for both generation strategies) or with the binomial approximation (for the POG strategy). Using a MC procedure, p_w can be estimated as

$$p_w = \frac{1}{M+1}\left(1 + \sum_{i=1}^{M} \mathbb{1}\left[Occ_{\tilde{\mathcal{D}}_i}(w) \leq Occ_{\mathcal{D}}(w)\right]\right), \tag{8}$$

where $\tilde{\mathcal{D}}_i$, with $i \in \{1, \ldots, M\}$, are M random time series datasets generated from the distribution described by the null hypothesis. Instead, using the binomial approximation, p_w

can be estimated as

$$p_w = \Pr\left[\text{Bin}(n_s, \Pr(w \mid E_k)) \leq Occ_{\mathcal{D}}(w)\right], \tag{9}$$

where $\Pr(w \mid E_k)$ is the probability of generating w under the POG strategy and n_s is the number of paths to generate from the same starting vertex of w. Thus, Algorithm 1 can be simply modified to mine under-represented paths computing the p-values p_w of the paths $w \in \mathcal{W}$ (line 4) with a MC procedure using Eq. 8 (for both generation strategies) or with the binomial approximation using Eq. 9 (for the POG strategy). The same approach must be employed to compute the p values $p_{\tilde{w}}$ of the paths $\tilde{w} \in \mathcal{W}_i$ (line 9) in the P random datasets for the WY method. Let us note that the complexity analysis above is valid for the mining of under-represented paths as well.

In the case one is interested in mining both types of paths, over- and under-represented, it is possible to speed up the execution of CASPITA mining both types of paths at the same time, avoiding the execution of CASPITA twice. Indeed, since the only difference in the two versions of CASPITA is the computation of the p-values, it is possible to perform all the other operations only once and to compute both types of p-values (both for the paths of the original and of the P random datasets). Thus, we obtain two different significance thresholds, one for the over- and one for the under-represented paths, to test the respective p-values. Let us note that this approach is a valid strategy to speed up the execution avoiding unnecessary re-computations but the false positives guarantees are still valid for the two types of paths separately, i.e., the set of returned over-represented paths has FWER $\leq \alpha$ and the set of returned under-represented paths has FWER $\leq \alpha$, separately. Instead, if one is interested in mining over- and under-represented paths obtaining false positives guarantees for both types of paths simultaneously, i.e., the set of returned over- and under-represented paths has FWER $\leq \alpha$, the following strategy is a possible solution. Given the FWER threshold $\alpha \in (0, 1)$, to compute the two corrected significance thresholds considering $\alpha/2$, i.e.,

$$\delta^* \leftarrow \max\left\{\delta : \sum_{i=1}^{P}\left(\mathbb{1}[p_{\min}^{(i)} \leq \delta]\right) \leq \frac{\alpha P}{2}\right\},$$

each with the respective p-values. Using the union bound, it is easy to prove that the resulting output, consisting of both over- and under-represented paths, has FWER $\leq \alpha$. Let us note that, for both scenarios, the computational complexity is asymptotically the same of the one of the original CASPITA that mines over-represented paths. Indeed, once that the p-values for the over-represented paths are computed (and, thus, the quantities required for such a computation have been obtained), the computation of the p-values for the under-represented paths has constant complexity.

3.6 Controlling the generalized FWER

In this section, we illustrate how CASPITA can be modified to mine statistically significant paths while controlling the *generalized* FWER [22]. In several real applications, one may be interested in tolerating a small amount of false discoveries in order to increase the power of detecting significant paths, still obtaining guarantees on the false positives. In such cases, methods to discover significant paths while controlling the generalized FWER are preferred to methods controlling the FWER. Given a positive integer g, the generalized FWER g-FWER is defined as the probability of reporting at least g false positives, that is, if FP is the number of false positives, then g-FWER $= \Pr[\text{FP} \geq g]$. For a given value δ, let g-FWER(δ) be the

g-FWER obtained when δ is used as corrected significance threshold, that is, by reporting as significant all paths with p-value $\leq \delta$. The WY method can be used to estimate the FWER FWER(δ) obtained using δ as corrected significance threshold as

$$g\text{-FWER}(\delta) = \frac{1}{P}\sum_{i=1}^{P} \mathbb{1}\left[p_g^{(i)} \leq \delta\right],$$

where $p_g^{(i)}$ is the g-th smallest p-values over all paths of interest in the random dataset $\tilde{\mathcal{D}}_i$. Thus, given a g-FWER threshold $\alpha \in (0, 1)$, the corrected significance threshold δ^* is obtained as

$$\delta^* = \max\{\delta : g\text{-FWER}(\delta) \leq \alpha\}.$$

Algorithm 1 can be simply modified to mine statistically significant paths with g-FWER $\leq \alpha$. To obtain such guarantees, it is sufficient to substitute lines 10 and 11, respectively, with

$$p_g^{(i)} \leftarrow g\text{-th } \min\{p_{\tilde{w}} : \tilde{w} \in \mathcal{W}_i\}$$

and

$$\delta^* \leftarrow \max\left\{\delta : \sum_{i=1}^{P}\left(\mathbb{1}[p_g^{(i)} \leq \delta]\right) \leq \alpha P\right\}.$$

Let g-CASPITA be such modified version of our algorithm. (Let us note that 1-CASPITA corresponds to the original CASPITA.) Theorem 2 proves that the output of g-CASPITA has g-FWER $\leq \alpha$.

Theorem 2 *The output of* g-CASPITA *has* g-*FWER* $\leq \alpha$.

The proof is analogous to the proof of Theorem 1.

Let us note that the computational complexity of g-CASPITA is the same of CASPITA, since the computation of the P g-th smallest p-values still requires the computation of the p-values of all the paths of interest in each of the P random dataset $\tilde{\mathcal{D}}_i$.

4 Mining statistically significant paths from different datasets

In this section, we illustrate another interesting scenario in which our algorithm CASPITA can be applied. Let us suppose to have two datasets, $\mathcal{D}_1$ and $\mathcal{D}_2$, and that such two datasets are taken from the same network N, but in different circumstances, e.g., in different temporal points, or maybe that they represent data generated from two different populations, e.g., men and women. In such a scenario, one may be interested in finding paths from one of the two datasets that are statistically significant considering the distribution represented by the other dataset. Thus, it is possible to use a slightly modified version of CASPITA, considering the dataset $\mathcal{D}_1$ to generate the h-th order generative null model $N^h(\mathcal{D}_1)$ and then, to consider the paths mined from the other dataset $\mathcal{W}_{\mathcal{D}_2}(k)$, and to compute their significance using $N^h(\mathcal{D}_1)$. Differently from the scenario described above, in this setting it is also possible to mine statistically significant paths of length k considering the h-th order generative null model with $k = h$. Let us note that, even in this alternative scenario, the complexity analysis provided above is still valid for this modified version of CASPITA. The only differences are in the time complexity t_N required to construct the generative null model $N^h(\mathcal{D}_1)$ that

now depends on the dataset $\mathcal{D}_1$, i.e., $t_N = \mathcal{O}(D_1)$, with $D_1 = \sum_{\tau \in \mathcal{D}_1} |\tau|$, and in the time complexity t_W to mine $\mathcal{W}_{\mathcal{D}_2}(k)$ that now depends on the dataset $\mathcal{D}_2$, i.e., $t_W = \mathcal{O}(D_2)$, with $D_2 = \sum_{\tau \in \mathcal{D}_2} |\tau|$.

5 Experimental evaluation

In this section, we report the results of our experimental evaluation on multiple pseudo-artificial and real datasets to assess the performance of CASPITA for mining statistically significant paths from an unknown network.

The goals of the evaluation are the following:

- To prove that for small datasets, CASPITA is able to find statistically significant paths with both generation strategies, i.e., TOG and POG, using the MC procedure, while for larger datasets the binomial approximation is necessary to provide useful results.
- To prove that the binomial approximation is a valid approach to approximate the p-values for the POG strategy.
- Focusing on the POG strategy with the binomial approximation, to prove that CASPITA and its modified version g-CASPITA are able to find large sets of statistically significant paths in pseudo-artificial and real large datasets, while avoiding false positives, and compare CASPITA with HYPA [18].
- To prove that CASPITA is able to find statistically significant paths in the scenario in which the generative null model is constructed considering data from another dataset (see Sect. 4).

5.1 Environment and datasets

We implemented CASPITA in Java. We performed all the experiments on the same machine with 512 GB of RAM and 2 Intel(R) Xeon(R) CPU E5-2698 v3 @ 2.3GHz, using Java 1.8.0_201. To parallelize the MC procedures, we used Apache Spark Java API version 3.1.1. Our open-source implementation of CASPITA and the code developed for the tests and to generate the datasets are available at https://github.com/VandinLab/CASPITA. In all the experiments, we fixed the FWER threshold to the commonly used value $\alpha = 0.05$, and we mined over- and under-represented paths at the same time, considering the FWER guarantees separately for the two types of paths. To compare with HYPA [18], we used their implementation available online,[1] considering the "rpy2" version.

In the following, we describe the datasets used in the evaluation, and how we generated them. Their characteristics are shown in Table 1:

- BIKE: data on the bike sharing service of Los Angeles. Each vertex is a bike station, while each transaction represents the sequence of bike stations that a given bike visits. We considered the 2019 data of the "Los Angeles Metro Bike Share trip data,"[2] containing single trips in the format of starting station, ending station, and an unique numerical identifier of the bike, among other information. We collected the temporal ordered sequence of bike stations that each bike visited. Such sequence is a transaction in our dataset. In the case of data anomalies, i.e., an ending station of a trip does not correspond to the starting

[1] https://github.com/tlarock/hypa.

[2] https://bikeshare.metro.net/about/data/.

Table 1 Datasets characteristics. $|\mathcal{D}|$: number of transactions; Avg $|\tau|$: average transaction length; Max $|\tau|$: maximum transaction length; for the 1-st generative null model $N^1(\mathcal{D})$, $|V^1|$: number of vertices; $|E^1|$: number of edges

Dataset $\mathcal{D}$	$\lvert\mathcal{D}\rvert$	Avg $\lvert\tau\rvert$	Max $\lvert\tau\rvert$	$N^1(\mathcal{D})$	
				$\lvert V^1\rvert$	$\lvert E^1\rvert$
BIKE10	3025	1.54	11	10	76
BIKE20	5080	1.90	21	20	279
BIKE	38,651	7.51	232	237	10,269
FLIGHT	17,447,803	1.63	15	455	69,234
WIKI	51,307	5.76	434	4169	59,530

station of the following trip, we split the sequence where the gap happens, creating two transactions.

- BIKE10 and BIKE20: smaller versions of the BIKE dataset. From BIKE, we only considered the 10 or 20 vertices, respectively, that occur most frequently times, and collect all the transactions that only contain such vertices.
- FLIGHT: data of the commercial flights in the USA. Each vertex is an airport, while each transaction represents the sequence of airports visited in a single itinerary by a passenger. We considered the 2019 data of the "Origin and Destination Survey: DB1BCoupon".[3] Such data contains single flights in the format of origin and destination airports, an unique numerical identifier of the itinerary that contains the flight, and the sequence number of the flight inside the itinerary. We collected the temporal ordered sequence of airports that each passenger visited in a single itinerary, sorting the airports using the sequence numbers. Such sequence is a transaction in our dataset.
- WIKI: it contains human navigation paths on Wikipedia, collected through the human-computation game Wikispeedia [23]. Each vertex is a Wikipedia web page, while each transaction is a sequence of web pages visited by an user during a game. We considered the data "paths finished,"[4] that represent finished games.

5.2 Generation strategies comparison

In this section, we compare the results obtained by CASPITA with the TOG or POG strategies that employ MC procedures, and the POG strategy that uses the binomial approximation, on BIKE10 and BIKE20.

The experiments have been performed with $P = 1000$, $M = 10^5$, $k \in \{2, \ldots, 5\}$, and $h \in \{1, \ldots, k-1\}$. The results are reported in Table 2. For BIKE10, the smallest dataset, the number of significant paths obtained with the TOG and POG strategies with MC procedures differs from at most 1, for all combinations of parameters. The same is true when the POG strategy with the binomial approximation is used. In all the cases, CASPITA reported at most 3 statistically significant paths, which is not surprising since BIKE10 only contains few distinct paths. For BIKE20, the situation is different. For some combinations of parameters (shown in bold in Table 2), CASPITA with the MC procedures did not report any significant (over-represented) paths, while it reported some paths (from 1 to 8) when the binomial approximation is used. In all such cases, the MC estimates resulted in a corrected threshold $\delta^* = 1/(M+1)$, corresponding to the minimum achievable p-value considering M random datasets. Thus, to be able to mine paths, one has to consider a larger value of M, which is

[3] https://www.transtats.bts.gov/Fields.asp?gnoyr_VQ=FLM.

[4] https://snap.stanford.edu/data/wikispeedia.html.

Table 2 CASPITA results with BIKE10 and BIKE20. k: paths length; h: order of the null model; for each dataset, BIKE10 and BIKE20, $|\mathcal{W}|$: number of distinct paths of length k; T: number of total paths of length k; for each generation strategy, TOG (T), POG (P), and POG with binomial approximation (B), $|\mathcal{SW}|$: number of significant paths reported, over- (+) and under- (−) represented

k	h	BIKE10								BIKE20							
		$\lvert\mathcal{W}\rvert$	T	$\lvert\mathcal{SW}^T\rvert$		$\lvert\mathcal{SW}^P\rvert$		$\lvert\mathcal{SW}^B\rvert$		$\lvert\mathcal{W}\rvert$	T	$\lvert\mathcal{SW}^T\rvert$		$\lvert\mathcal{SW}^P\rvert$		$\lvert\mathcal{SW}^B\rvert$	
				+	−	+	−	+	−			+	−	+	−	+	−
2	1	164	1630	3	3	3	3	3	3	978	4553	11	12	9	8	11	8
3	1	163	575	1	2	2	2	2	2	997	2320	0	9	0	7	**8**	7
	2			1	2	1	2	1	2			2	8	1	6	1	5
4	1	104	21	0	2	0	2	1	2	713	1220	0	4	0	5	**4**	5
	2			0	1	1	1	1	1			0	7	0	4	0	4
	3			0	1	0	0	0	0			0	4	0	2	0	2
5	1	57	83	0	2	0	2	0	2	450	661	0	2	0	2	**2**	2
	2			0	1	1	0	1	0			0	3	0	2	**1**	2
	3			0	0	0	0	0	0			0	1	0	1	0	1
	4			0	0	0	0	0	0			0	0	0	1	0	1

In bold the values for which CASPITA did not report any significant path with the MC procedures, while it reported some paths with the binomial approximation

infeasible with larger datasets, or to resort to the binomial approximation. This phenomenon appeared with $k > h - 1$, that is, when a large number of distinct paths can be generated, even for a small dataset such as BIKE20. This emphasizes the issue of the TOG strategy described above, that is, the gargantuan number of paths that must be considered with the generation of long transactions.

We then compared the p-values from the POG strategy obtained with the MC procedure and the binomial approximation. Let us note that while in the MC procedure the total number of length k paths starting from a vertex is fixed to the value observed in the data, using the binomial approximation such property holds only in expectation. Thus, the p-values from the two approaches will be different. However, by comparing the p-values[5] for all paths (over- and under-represented) of BIKE10 and BIKE20 with $k \in \{2, \ldots, 5\}$ and $h \in \{1, \ldots, k-1\}$, and considering $M \in \{10^4, 10^5, 10^6\}$ random datasets for the MC estimates, we observed that the p-values for the binomial approximation are within one order of magnitude of the corresponding MC p-values, and that the difference between binomial p-values and MC p-values is lower than the standard deviation of the MC estimates (obtained from 5 estimates of the MC p-values). Furthermore, the binomial approximation is several orders of magnitude faster than the MC procedure (few milliseconds against over 40 s, considering the maximum execution time for both strategies and using 8 cores to parallelize the MC estimates).

5.3 Results for POG strategy with binomial approximation

Since the results of the previous section demonstrated that the POG strategy with the binomial approximation is necessary to mine statistically significant paths from large datasets, and that the p-values for the binomial approximation are within one order of magnitude of the corresponding MC p-values, in this section we focus on such version of CASPITA. First,

[5] We only considered p-values $\geq 1/(M+1)$, since lower p-values require larger M to be correctly estimated with the MC procedure.

Table 3 g-FWER estimates for g-CASPITA with pseudo-artificial datasets obtained from BIKE10 and BIKE20. P: number of random datasets for the WY method; g: number of false positives considered in the g-FWER; for each real dataset, BIKE10 and BIKE20, g-FWER (%): estimate of the g-FWER in percentage

P\g	BIKE10				BIKE20			
	g-FWER (%)				g-FWER (%)			
	1	2	5	10	1	2	5	10
100	0.75	0.50	0.75	0.25	2.25	1.25	0.25	0.50
1000	1.25	0.50	0.75	0.25	1.75	0.25	1.25	0.75
10,000	0.50	1.25	0.10	0.10	3.25	1.25	0.25	0.50

we investigated the false positives guarantees of CASPITA on pseudo-artificial datasets, also performing a comparison with HYPA, and then we executed it on real datasets.

5.3.1 False positives guarantees

In this section, we report the results of our experimental evaluation to assess the false positives guarantees of CASPITA using pseudo-artificial datasets. Starting from a real dataset, we created its h-th order generative null model, which we used to generate random datasets using the POG strategy. Each random dataset is then a bag of paths of a given length $k > h$ that does not contain any significant path of length k (since they have been generated in accordance with the generative null model). We then executed CASPITA on each random dataset, with parameters h and k corresponding to the ones used to generate the random dataset, and we checked whether CASPITA reported some paths, which would be false positives by construction. We considered BIKE10 and BIKE20 as starting real datasets, $k \in \{2, \ldots, 5\}$, and $h \in \{1, \ldots, k-1\}$, mining over- and under-represented paths. Given a real dataset, we generated 20 random datasets for each combinations of h and k, obtaining a total number of 400 runs for each real dataset. We then estimated the FWER as the fraction of runs with at least one false positive. Table 3 ($g = 1$) shows the obtained results. For BIKE10, CASPITA obtained an estimated FWER of 0.75% with $P = 100$, 1.25% with $P = 1000$, and 0.5% with $P = 10{,}000$. Instead, for BIKE20, it obtained an estimated FWER of 2.25% with $P = 100$, 1.75% with $P = 1000$, and 3.25% with $P = 10{,}000$. These results show that the false positives guarantees of CASPITA are even better than the theoretical ones, which are $\leq 5\%$ using $\alpha = 0.05$, and that $P = 100$ is enough to obtain such guarantees. Using these random datasets, we also made a comparison with HYPA. Let us remember that HYPA employs a fixed threshold β to flag as anomalous a path, without any theoretical guarantees. We used $\beta \in \{0.00001, 0.001, 0.05\}$, which are, respectively, the minimum, the most commonly used, and the maximum value used in [18], and $k \in \{2, \ldots, 5\}$. (For h, HYPA always considers $h = k - 1$, thus we only used this value.) For BIKE10, it obtained an estimated FWER of 40.63% with $\beta = 0.05$, 15.63% with $\beta = 0.001$, and 0.0% with $\beta = 0.00001$. Instead, for BIKE20, it obtained an estimated FWER of 60.63% with $\beta = 0.05$, 32.50% with $\beta = 0.001$, and 1.25% with $\beta = 0.00001$. These results show that HYPA is able to return anomalous paths achieving low FWER with the correct threshold, but also emphasize the importance of having a strategy, as the one that we employ, to compute such threshold in an automatic way, since the usage of fixed thresholds may lead to many spurious discoveries, or to a low statistical power.

Finally, considering the same pseudo-artificial datasets, we assessed the false positives guarantees of g-CASPITA with $g \in \{2, 5, 10\}$. We executed g-CASPITA in each random dataset and then we estimated the g-FWER as the fraction of runs with at least g false

Table 4 g-CASPITA results with BIKE. k: paths length; h: order of the null model; $|\mathcal{W}|$: number of distinct paths of length k; T: number of total paths of length k; for each g, number of reported significant paths: over- (+) and under- (−) represented; Time (s): execution time in seconds with $g = 1$

k	h	$\lvert\mathcal{W}\rvert$	T	$g=1$		$g=2$		$g=5$		$g=10$		Time (s)
				+	−	+	−	+	−	+	−	
2	1	90.5K	252K	197	28	256	47	374	84	453	98	140
3	1	172K	221K	118	40	159	60	226	85	311	117	279
	2			1	8	4	17	7	29	10	38	262
4	1	179K	195K	80	15	127	22	153	40	184	65	315
	2			10	7	14	15	19	25	21	33	333
	3			0	3	0	4	0	12	0	15	246
5	1	166K	174K	71	6	98	6	115	9	130	19	437
	2			17	3	19	5	31	10	39	16	360
	3			0	1	1	3	1	3	2	9	270
	4			0	1	0	2	0	4	1	7	246

Table 5 g-CASPITA results with WIKI. k: paths length; h: order of the null model; $|\mathcal{W}|$: number of distinct paths of length k; T: number of total paths of length k; for each g, number of reported significant paths: over- (+) and under- (−) represented; Time (s): execution time in seconds with $g = 1$

k	h	$\lvert\mathcal{W}\rvert$	T	$g=1$		$g=2$		$g=5$		$g=10$		Time (s)
				+	−	+	−	+	−	+	−	
2	1	155K	244K	160	53	222	82	346	134	424	175	254
3	1	169K	194K	219	6	334	16	443	27	549	34	303
	2			8	12	11	13	22	25	27	34	250
4	1	139K	147K	193	1	319	2	425	3	535	3	294
	2			16	6	20	8	39	10	51	16	247
	3			2	2	4	6	6	12	7	12	196
5	1	106K	108K	113	0	154	0	243	0	327	0	646
	2			7	0	9	0	18	1	39	1	190
	3			0	1	0	2	0	5	4	8	160
	4			0	0	0	0	1	1	1	1	150

positives. Table 3 shows all the obtained results. Similarly to what has been observed with 1-CASPITA, i.e., the original version of CASPITA, the results show that the false positives guarantees of g-CASPITA are even better than the theoretical ones since all the g-FWER estimates are (far) below 5%, and $P = 100$ is enough to obtain them.

5.3.2 Results with real datasets

We then executed (g-)CASPITA on the real datasets BIKE, WIKI, and FLIGHT. Tables 4, 5, and 6 report the results obtained with $P = 100$, $k \in \{2, \ldots, 5\}$, $h \in \{1, \ldots, k-1\}$, and $g \in \{1, 2, 5, 10\}$. For the computational time, we only reported the values for $g = 1$, since the others are analogous. Considering $g = 1$, for all the datasets, and for almost all combinations of parameters, CASPITA reported some significant paths. It is interesting to

Table 6 g-CASPITA results with FLIGHT. k: paths length; h: order of the null model; $|\mathcal{W}|$: number of distinct paths of length k; T: number of total paths of length k; for each g, number of reported significant paths: over- (+) and under- (−) represented; Time (s): execution time in seconds with $g = 1$

k	h	$\lvert\mathcal{W}\rvert$	T	$g = 1$		$g = 2$		$g = 5$		$g = 10$		Time (s)
				+	−	+	−	+	−	+	−	
2	1	574K	11.1 M	96.5K	16.4K	102K	19.2K	111K	23.2K	115K	26.1K	3.95K
3	1	849K	5.16 M	132K	36	140K	50	144K	75	149K	92	9.18K
	2			128K	1.63K	134K	2.13K	140K	2.81K	144K	3.25K	6.40K
4	1	406K	530K	19.2K	0	22.0K	0	25.1K	0	26.6K	0	1.07K
	2			10.4K	30	14.8K	34	19.2K	43	22.0K	52	864
	3			3.11K	19	3.84K	24	5.00K	50	5.95K	66	721
5	1	127K	155K	6.66K	0	7.92K	0	9.39K	0	10.1K	0	29.2K
	2			4.28K	0	5.16K	0	6.35K	0	7.07K	0	426
	3			1.80K	5	2.32K	7	3.30K	8	4.01K	8	306
	4			884	2	1.30K	2	1.87K	3	2.23K	4	259

notice that the number of over-represented paths is almost always (some orders of magnitude) greater than the number of under-represented paths. In addition, for a fixed value of k, the number of over-represented paths always decreases considering higher values of h. The number of under-represented paths, instead, always decreases for BIKE, while increases and then decreases for FLIGHT and WIKI, highlighting different substructures in the three underlying networks. The computational time ranges from under 3 minutes (BIKE with $k = 2, h = 1$) to over 8 hours (FLIGHT with $k = 5, h = 1$). Let us note that FLIGHT has over $17M$ transactions, but we are still able to analyze it with a reasonable running time. The combination $k = 5$ and $h = 1$ is always the most expensive from a computational point of view, since it requires to generate longer paths and also to analyze a large number of vertices to compute the probabilities of Eq. 7. For the same reasons, for a fixed path length k, the computational time decreases as h increases. Overall, these results show that CASPITA is able to efficiently mine significant paths from real datasets, with feasible computational time even in huge datasets. Considering $g > 1$, for all the datasets, and for almost all combinations of parameters, g-CASPITA reported more significant paths, both over- and under-represented, with respect to the paths reported by the original version of CASPITA. In particular, for BIKE, g-CASPITA with $g = 2$ reported on average 77% more paths than CASPITA,[6] 188% more with $g = 5$, and 330% more with $g = 10$. For WIKI, instead, g-CASPITA with $g = 2$ reported on average 70% more paths than CASPITA, 194% more with $g = 5$, and 286% more with $g = 10$. Finally, for FLIGHT, g-CASPITA with $g = 2$ reported on average 22% more paths than CASPITA, 61% more with $g = 5$, and 86% more with $g = 10$. Let us note that even if we were allowing very few false positives, the number of reported paths increased considerably. Overall, these results show that to control the g-FWER is a valid strategy to increase the statistical power for those applications which can tolerate a small number of false positives.

Finally, we provide a brief analysis of some paths returned by CASPITA from FLIGHT. The over-represented path of length 2 with the lowest p-value and the highest number of occurrences (4813) is an itinerary which starts from the "Philadelphia International" airport, goes to the "Charlotte Douglas International" airport, and then comes back to the "Philadelphia International." Instead, among the over-represented paths of length 2 with the lowest p-value, the one with the lowest number of occurrences (5) is an itinerary which starts from the "Trenton Mercer" airport, goes to the "Hartsfield-Jackson Atlanta International" airport, and then comes back to the "Trenton Mercer." Since both these over-represented paths represent returning itineraries, i.e., they start and end in the same airports, we investigated which is the percentage of returning over-represented paths, that resulted almost 5%. Thus, even if it is not very surprising that returning paths are over-represented, 95% of the over-represented paths of length 2 are not returning itineraries. The under-represented path of length 2 with the lowest p-value and the highest number of occurrences (137) is instead an itinerary that starts from the "Daniel K Inouye International" airport (located in the Hawaii), goes to the "Hilo International" airport (again located in the Hawaii), and then finishes in the "Los Angeles International" airport, while the percentage of returning under-represented paths of length 2 is 0.18%, significantly lower than that of the over-represented paths. Considering paths of length 3 and the 1-st order generative null model, the over-represented path with the lowest p-value and the highest number of occurrences (1398) is an itinerary which starts from the "Ted Stevens Anchorage International" airport, goes first to the "McCarran International" airport (located in Las Vegas), then goes to the "Seattle/Tacoma International" airport, and finally comes back to the "Ted Stevens Anchorage International." Let us note that this path

[6] We only considered combinations of parameters for which CASPITA reported at least one significant path.

results the one with the lowest p-value and the highest number of occurrences also considering the 2-nd order null model. Again, this path represents a returning itinerary and the percentages of returning over-represented paths of length 3, considering, respectively, the 1-st and the 2-nd order generative null model, are 94% and 98%. Finally, we observed that for the over-represented paths of length 4, considering the 1-st order generative null model, the percentage of returning itineraries is 64%, but only 3% of these paths are symmetrical, i.e., the outward and return flights are the same, while the others consider different airports even if they start and end in the same one. Instead CASPITA did not report any significant under-represented paths of length 4 considering the 1-st order generative null model.

5.4 Analysis of BIKE

In this section, we provide a brief analysis of some paths returned by CASPITA from BIKE. The over-represented path of length 2 with the lowest p-value and the highest number of occurrences is a path which starts and ends in "Ocean Front Walk & Navy" located in Venice Beach. The fact that this path is over-represented indicates that people tend to leave and then come back to this place, instead of moving to other parts of the city. For example, such pattern may capture the fact that people leave the beach to buy some food and then immediately come back. Instead, the under-represented path of length 2 with the lowest p-value is a path which starts from "Union Station West Portal," goes to "Main & 1st," and then comes back. "Main & 1st" is located near the Los Angeles City Hall, the center of the government of the city, while "Union Station West Portal" is near the Union Station, the main railway station of the city. The fact that this path is under-represented is probably due to the fact that a lot of people move from the station to the city hall, and vice versa, but in particular moments of the day, i.e., in the morning and in the evening. Thus, even if the two direct links are very popular, it is uncommon to see this entire path. These are only two examples of paths mined by CASPITA, but they highlight its capability in detecting real-life trends.

Finally, we investigated the capability of CASPITA in mining significant paths in the scenario in which the generative null model is created considering a different dataset (see Sect. 4). Using the procedure to generate BIKE described above, we generated a new dataset, NEWBIKE, considering the 2020 data from the same website. We then used the original BIKE dataset to create the h-th order generative null model and we tested on it the significance of length k paths mined from NEWBIKE. Given the pandemic situation that involved the world, one may be interested in finding changes in the habits of the people defining a model based on paths traveled in 2019 to test paths traveled in 2020. Table 7 reports the results obtained with $k = h \in \{1, \ldots, 5\}$ and $P = 100$. (For NEWBIKE, we only considered full transactions that can be generated by the generative null model obtained from BIKE.) Again, it is possible to notice that CASPITA returned paths for almost all combinations of parameters, and that the number of over-represented paths is always higher than the number of under-represented paths. Overall, these results demonstrate that CASPITA is able to mine paths also in such a scenario.

6 Conclusions

In this work, we introduced the problem of mining *statistically significant paths* in time series data from an unknown network, which naturally arises in several applications. We described CASPITA, an algorithm to mine statistically significant paths (over- and under-

Table 7 CASPITA results on BIKE considering different datasets. k: paths length; h: order of the null model; $|\mathcal{W}|$: number of distinct paths of length k; T: number of total paths of length k; (+): reported over-represented paths; (−): reported under-represented paths

$k = h$	$\lvert\mathcal{W}\rvert$	T	+	−
1	5.79K	68.0K	256	120
2	5.51K	12.2K	30	9
3	1.19K	2.45K	14	1
4	436	786	4	0
5	124	203	0	0

represented) while controlling the probability of reporting at least one false positive, i.e., the FWER, employing the Westfall–Young method. We also described some extensions of CASPITA, one of which, g-CASPITA, aims to mine statistically significant paths while controlling the generalized FWER, which allows to increase the statistical power of our strategy by tolerating a few false positives. We then introduced an alternative scenario which considers a different dataset, with respect to the one used to mine the paths of interest, to construct the generative null model. Our extensive experimental evaluation shows that (g-)CASPITA is able to efficiently mine large sets of significant paths from real datasets, while correctly controlling the (generalized) FWER. Interesting future directions are the extension of CASPITA to mine statistically significant paths while controlling the *false discovery rate* (FDR) and to identify the k most significant paths.

Acknowledgements Part of this work was supported by the MIUR, the Italian Ministry of Education, University and Research, under PRIN Project No. 20174LF3T8 AHeAD (Efficient Algorithms for HArnessing Networked Data) and the initiative "Departments of Excellence" (Law 232/2016), by the University of Padova under project SEED 2020 RATED-X, and by the project "POR FESR 2014-2020: Piattaforma Integrata di Cruscotti Intelligenti per la Gestione Avanzata della Qualità e del Marketing Innovativo."

Funding Open access funding provided by Università degli Studi di Padova within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Tonon A, Vandin F (2020) caSPiTa: mining statistically significant paths in time series data from an unknown network. In: Proceedings of the 21st IEEE international conference on data mining. IEEE, ICDM'21, pp 639–648
2. Keogh E, Kasetty S (2003) On the need for time series data mining benchmarks: a survey and empirical demonstration. Data Min Knowl Disc 7(4):349–371

3. Roddick JF, Hornsby K, Spiliopoulou M (2000) An updated bibliography of temporal, spatial, and spatio-temporal data mining research. International workshop on temporal, spatial, and spatio-temporal data mining. Springer, Berlin, Heidelberg, pp 147–163
4. Esling P, Agon C (2012) Time-series data mining. ACM Comput Surv 45(1):1–34
5. Weiss GM (2004) Mining with rarity: a unifying framework. ACM SIGKDD Explor Newslett 6(1):7–19
6. Keogh E, Lonardi S, Chiu BYC (2002) Finding surprising patterns in a time series database in linear time and space. In: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining, pp 550–556
7. Lin J, Keogh E, Lonardi S, Lankford JP, Nystrom DM (2004) Visually mining and monitoring massive time series. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, pp 550–556
8. Keogh E, Lin J (2005) Clustering of time-series subsequences is meaningless: implications for previous and future research. Knowl Inf Syst 8(2):154–177
9. Wei L, Keogh E (2006) Semi-supervised time series classification. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, pp 748–753
10. Hämäläinen W, Webb GI (2019) A tutorial on statistically sound pattern discovery. Data Min Knowl Disc 33(2):325–377
11. Llinares-López F, Sugiyama M, Papaxanthos L, Borgwardt K (2015) Fast and memory-efficient significant pattern mining via permutation testing. In: Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining, pp 725–734
12. Pellegrina L, Vandin F (2020) Efficient mining of the most significant patterns with permutation testing. Data Min Knowl Disc 34:1201–1234
13. Tonon A, Vandin F (2019) Permutation strategies for mining significant sequential patterns. In: Proceedings of the 19th IEEE international conference on data mining, IEEE, ICDM'19, pp 1330–1335
14. Gupta M, Gao J, Aggarwal CC, Han J (2013) Outlier detection for temporal data: a survey. IEEE Trans Knowl Data Eng 26(9):2250–2267
15. Lemmerich F, Becker M, Singer P, Helic D, Hotho A, Strohmaier M (2016) Mining subgroups with exceptional transition behavior. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 965–974
16. Noble CC, Cook DJ (2003) Graph-based anomaly detection. In: Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining, pp 631–636
17. Akoglu L, Tong H, Koutra D (2015) Graph based anomaly detection and description: a survey. Data Min Knowl Disc 29(3):626–688
18. LaRock T, Nanumyan V, Scholtes I, Casiraghi G, Eliassi-Rad T, Schweitzer F (2020) Hypa: Efficient detection of path anomalies in time series data on networks. In: Proceedings of the 2020 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, pp 460–468
19. 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, pp 1037–1046
20. Bonferroni C (1936) Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze 8:3–62
21. Westfall PH, Young SS (1993) Resampling-based multiple testing: examples and methods for p-value adjustment. Wiley, New York
22. Lehmann EL, Romano JP (2012) Generalizations of the familywise error rate. Selected works of EL Lehmann. Springer, Boston, pp 719–735
23. West R, Leskovec J (2012) Human wayfinding in information networks. In: Proceedings of the 21st international conference on world wide web, pp 619–628

Andrea Tonon received the Laurea Triennale degree (2015) in Information Engineering from the University of Padova, Italy, and the Laurea Magistrale degree (summa cum laude, 2018) in Computer Engineering from the University of Padova, Italy. He received the Ph.D. (2022) in Information Engineering from the University of Padova, Italy. He has been a Research Intern at the Huawei Ireland Research Center in the AIOps team. His research interests focus on algorithms for data mining and AI. In particular, the development of novel efficient methods for knowledge and pattern discovery from large collections of sequential data, often exploiting distributed/parallel approaches and statistically sound algorithms.

Fabio Vandin received the Laurea Triennale degree (summa cum laude, 2004) and the Laurea Specialistica degree (summa cum laude, 2006) in Computer Engineering from the University of Padova, Italy. He received the Ph.D. (2010) in Information Engineering from the University of Padova, Italy. He has been a postdoctoral researcher at Brown University and an Assistant Professor at the University of Southern Denmark. Since 2020 he is Professor at the Department of Information Engineering of the University of Padova. His research interests are in algorithms for data mining and machine learning and their application for the analysis of large biological datasets, in particular cancer genomic datasets. He has authored over 70 papers in international peer-reviewed venues, and he has used his methods for analyses published in Nature, Nature Genetics, Cell, NEJM. He has been a co-PI for two projects funded by the NSF (USA) and a participant in several European projects.