AZ-whiteness test: a test for uncorrelated noise on spatio-temporal graphs

Daniele Zambon[*1] and Cesare Alippi[1,2]

[1]The Swiss AI Lab IDSIA & Università della Svizzera italiana, Switzerland.
[2]Politecnico di Milano, Italy.

April 26, 2022

Abstract

We present the first whiteness test for graphs, i.e., a whiteness test for multivariate time series associated with the nodes of a dynamic graph. The statistical test aims at finding serial dependencies among close-in-time observations, as well as spatial dependencies among neighboring observations given the underlying graph. The proposed test is a spatio-temporal extension of traditional tests from the system identification literature and finds applications in similar, yet more general, application scenarios involving graph signals. The AZ-test is versatile, allowing the underlying graph to be dynamic, changing in topology and set of nodes, and weighted, thus accounting for connections of different strength, as is the case in many application scenarios like transportation networks and sensor grids. The asymptotic distribution — as the number of graph edges or temporal observations increases — is known, and does not assume identically distributed data. We validate the practical value of the test on both synthetic and real-world problems, and show how the test can be employed to assess the quality of spatio-temporal forecasting models by analyzing the prediction residuals appended to the graphs stream.

1 Introduction

In recent years, machine learning methods based on graph-structured data have made significant advances in the field of multivariate time-series analysis and resulted in major achievements, *e.g.*, in forecasting performance and missing values imputation [7, 22]. In this paper, we address the problem of testing whether given time series can be considered white noise or not and, by that, we aim at assessing the optimality of machine learning models trained to solve associated forecasting tasks.

We focus on spatio-temporal time series where the spatial domain is defined by a graph $G = (V, E)$ and stochastic time series (or node signals) $\mathbf{x}_v[t] \in \mathbb{R}^F$, with $t = 1, 2, \ldots$, are associated with the nodes $v \in V$ of G. That said, the framework can be extended to deal with graphs seen as a realization of a random variable. Graph G can be static, meaning that is constant over time, or dynamic, hence modeling frameworks where topology and number of nodes can change. Dynamic graphs appear frequently in cyber-physical systems where sensors can be added or removed, or node data is missing for certain lapses of time following communication problems or faults. Another example is provided by social networks where users expand their set of friends or change preferences and interests. It is also extremely important to consider graphs with weighted edges encoding, for instance, the capacity or the strength of the link. Figure 1 provides a visualization of a dynamic graph G with associated graph signal $\mathbf{X}$, that is, the collection of all node signals $\mathbf{x}_v[t]$, for nodes v and time steps t.

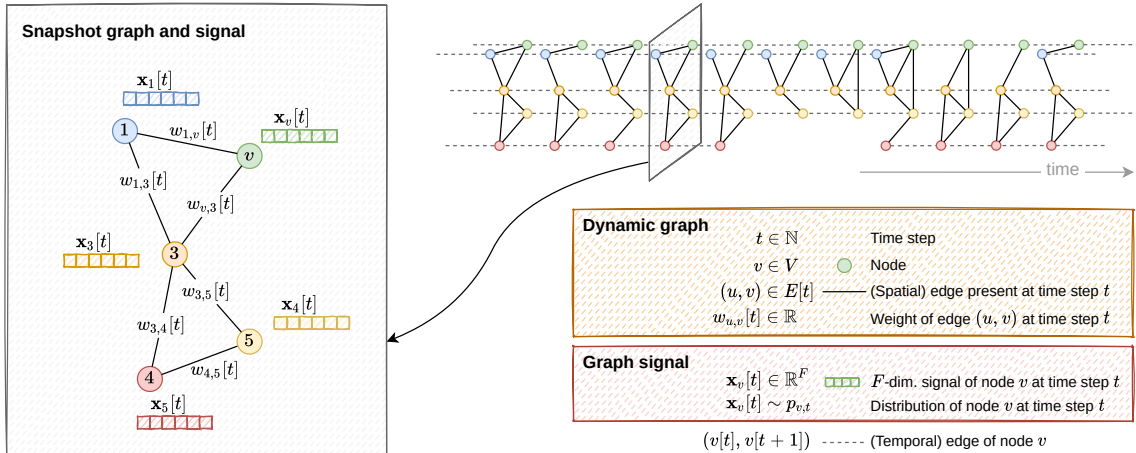[*]Corresponding author, `daniele.zambon@usi.ch` .

Figure 1: A dynamic weighted graph and a graph signal. The dynamic graph is defined over a set V of nodes; without requesting all nodes to be always available over time. The topology of the graph is represented by solid lines. Dashed lines represent temporal edges connecting the same node at consecutive time steps. The graph signal at generic time step t and node v is multivariate. On the left-hand side, a single snapshot of the temporal graph and graph signal are extracted.

In the presented setting, we address the problem of assessing the optimality of a predictive model f_θ trained to solve a forecasting task associated with stochastic graph signal $\mathbf{X}$. We tackle this problem by inspecting the graph signal $\mathbf{R}$ composed of the residuals $\mathbf{r}_v[t] = \hat{\mathbf{x}}_v[t] - \mathbf{x}_v[t]$ between the values estimated by model f_θ and the measured ones in $\mathbf{X}$, respectively. Indeed, finding data dependency within $\mathbf{R}$ entails that there is further structural information that can be learned from the data and, therefore, the predictive model f_θ can be improved by relying on a better training procedure or considering a more appropriate family of models.

We propose a novel whiteness test to decide whether a signal on a graph G can be considered white noise or there is prominent evidence of either serial or spatial correlations among data observations. In general, testing the assumption of independent signals is ill-posed as it requires studying a number of unknown distributions $p_{v,t}$ (associated with each $\mathbf{x}_v[t] \in \mathbf{X}$) proportional to the number of nodes and time steps, and for each of which we have only a single observation. The test we propose extends traditional methods for serial correlation [13, 17, 18] that study the autocorrelation of the signals at different time lags. In this paper, we provide methods allowing for testing the whiteness of the graph signal by exploiting known relational inductive biases existing among the data observations represented by the given graph topology and the temporal coherence. In doing so, we expand the analysis of serial correlation to the more general case of spatio-temporal correlation. To the best of our knowledge, this is the first test proposing such a general whiteness test. The proposed test statistic is based on counting the positive and negative signs of the product between spatially and temporally adjacent observations, whose disproportion indicates either direct or inverse correlation among the variables. Indeed, authors can consider different statistics. In summary, the contributions of the paper are the following:

- We propose the first statistical test in the literature to check the whiteness hypothesis for a graph signal $\mathbf{X}$ defined over a, possibly weighted and dynamic, graph G [Sections 3 and 4].

- We derive the limit distribution of the test statistics under the null hypothesis [Theorem 1].

- We present a statistical procedure based on the proposed test to assess whether a given forecasting model can be considered optimal with respect to the given data or not [Section 5.2].

The proposed whiteness test is computationally efficient for sparse graphs as the number of operations scales linearly with the number of edges and time steps. Moreover, the test is very

Table 1: List of possible configurations to which the proposed whiteness test is applicable.

Aspect:	Temporal dimension	Edge weights	Node signals	Neighborhood
Options:	Absent	Absent	Scalar ($F = 1$)	1-hop
	$T > 0$, G is static	Present	$F > 1$	K-hop ($K > 1$)
	$T > 0$, G is dynamic			

general and allows to incorporate very relevant application contexts and operational scenarios as listed in Table 1. In particular, the test is applicable when $\mathbf{x}_v[t]$ is univariate or multivariate, when G is static or dynamic, when the edges of G are weighted, and when the graph signal is composed of a set of time series or it is static, namely, $\mathbf{X} = \{\mathbf{x}_v \in \mathbb{R}^F \mid v \in V\}$, and there is no temporal dimension involved.

The remainder of the paper is structured as follows. Section 2 reviews the related work. Section 3 presents the test in the simplified case of static graph and signal. Section 4 shows how the test from Section 3 can be applied to spatio-temporal signals on dynamic graphs. Section 5 reports empirical evidence of the statistical power of the test and shows how to assess the optimality of forecasting models by applying the test to the prediction residuals. Finally, Section 6 draws some conclusions and provides pointers to future research.

2 Related Work

Among the most renowned whiteness tests, there are the Durbin-Watson test [9] and the Ljung-Box test [18]. Both have been introduced to test serial dependency in a univariate time series. In particular, the Ljung-Box test [18] is able to inspect the autocorrelation of the signal at multiple time lags at the same time; these types of tests are known as "portmanteau" tests. More recently, Drouiche [8] proposed a test that, differently from the previous ones, operates on the spectral density of the signal which, under the null hypothesis of white noise, has to be constant.

Several whiteness tests for multivariate data have been developed as extensions of the Box-Pierce or Ljung-Box tests [5, 13, 15]. These tests operate in the asymptotic regime where the data dimensionality, say D, is negligible with respect to the number T of available temporal observations. More recently, the literature has proposed tests overcoming the lack of power and poor approximation of the critical region. Examples are the test by Bose and Hachem [2], Li et al. [17] operating under the condition where the ratio D/T approaches a constant, as T approaches infinity.

To the best of our knowledge, no work has presented a whiteness test for spatio-temporal signals, and what proposed here is pioneering in this direction.

For other purposes, few tests consider a graph structure to represent relationships among the given observations. For example, Friedman-Rafsky test [10] compares the distribution of two given sets of data by constructing a minimum spanning tree connecting the given data. The test statistic is based on counting the number of connected components (trees) left after removing the edges linking data points coming from different samples. Friedman-Rafsky test generalizes the run test by Wald and Wolfowitz [21]. Other statistical tests based on topological information computed from the data have been proposed by Rosenbaum [19] and Chen and Zhang [4].

In this paper, we take inspiration from Geary test [11] for uncorrelated residuals in univariate data that counts the number of sign changes between consecutive observations. Geary [11] shows that, empirically, the test performance was on par with more sophisticated tests, such as the Durbin-Watson test [9]. Geary's comparison considered another test, the Wald-Wolfowitz run test [21] for scalar sequences, which counts the number of runs (subsequences of values all of the same sign) and, hence, closely related to the number sign changes. It turned out that the Walf-Wolfowitz test performed slightly worse than the Geary test, even though the run test does not assume that the observations have an equal probability of being positive and negative.

The statistical test we propose shares the fundamental idea of counting signs of the Geary test, but introduces several advancements that make it a substantial and original contribution. Apart

from the fact we are considering graphs, which is a major contribution per se, the most important result is the applicability of the test to spatio-temporal signals defined over graphs. Secondly, the designed test is general enough to operate on weighted and dynamic graphs with multivariate node signals.

3 Whiteness test for graph signals

For the sake of clarity, we first present the test in a simplified — yet relevant — setting characterized by a static weighted graph and a static signal with a single observation (scalar or vector) associated with each node; an example is given in Figure 2. The most general case where time is involved and the graph is dynamic, as in the scenario depicted in Figure 1, is provided in next Section 4 as an extension of what we are presenting here. In particular, we will see that we can apply the test here designed by considering a suitably constructed (multiplex) graph involving temporal edges alongside the spatial edges present in the original graph.

Consider a weighted graph $G = (V, E, \mathbf{W})$ defined over node set V, edge set $E \subseteq V \times V$ and scalar weights

$$\mathbf{W} = \{w_e \in \mathbb{R}_+ \mid e \in E\}.$$

The graph topology and weights define the spatial structure underlying a graph signal

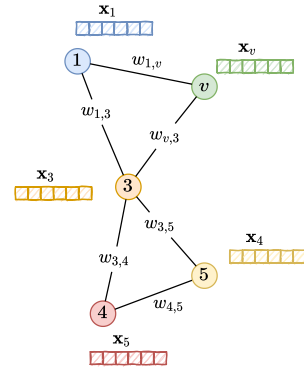$$\mathbf{X} = \{\mathbf{x}_v \in \mathbb{R}^F \mid v \in V\}$$

Figure 2: A static graph signal is defined over a weighted graph $G = (V, E, \mathbf{W})$. Multivariate node signals $\mathbf{x}_v$, $v \in V$, are generic random vectors and, here, not associated with time series.

defined over the nodes of G. We assume that the edge weights are positive values and encode the strength or the capacity of the links; without loss of generality, we consider absent those edges characterized by null weight. Node signals $\mathbf{x}_v \in \mathbb{R}^F$ can be scalars ($F = 1$) or vectors ($F > 1$), but they are static, meaning that no temporal information is associated with them, as shown in Figure 2. Essentially, we consider a single snapshot of the dynamic graph depicted in Figure 1.

The ultimate goal is to test if *graph* signal $\mathbf{X}$ is white noise, implying that there is no further dependency left among node signals in $\mathbf{X}$. The statistical hypotheses of the test are

$$\begin{cases} H_0: & \mathbf{x}_u, \mathbf{x}_v \text{ are independent for all } u \neq v \in V; \\ H_1: & \mathbf{x}_u, \mathbf{x}_v \text{ are dependent for some } u \neq v \in V. \end{cases} \tag{1}$$

We propose the following whiteness test

$$\text{If } |C_G(\mathbf{X})| > \gamma \implies \text{ Reject null hypothesis } H_0, \tag{2}$$

based on statistic

$$C_G(\mathbf{X}) = \frac{\widetilde{C}_G(\mathbf{X})}{W_2^{1/2}}. \tag{3}$$

Statistic $C_G(\mathbf{X})$ is asymptotically distributed a standard Gaussian $\mathcal{N}(0, 1)$ as $|E| \to \infty$ (we prove it in following Theorem 1), and is defined over quantities

$$\widetilde{C}_G(\mathbf{X}) = \sum_{(u,v) \in E} w_{u,v} \, \text{sgn}(\mathbf{x}_u^\top \mathbf{x}_v), \tag{4}$$

$$W_2 = \sum_{(u,v) \in E \,:\, u < v} (w_{u,v} + w_{v,u})^2, \tag{5}$$

4

with $w_{u,v} = 0$ if $(u,v) \notin E$, and sign function $\mathrm{sgn}(x) = 1$ if $x > 0$, $\mathrm{sgn}(x) = -1$ if $x < 0$, and $\mathrm{sgn}(x) = 0$ otherwise; note that for undirected graphs W_2 reduces to $\sum_{(u,v) \in E} w_{u,v}^2$.

The intuition behind (2) to test null hypothesis H_0 against H_1 in (1) is that, when the node signals in $\mathbf{X}$ are mutually independent and centered around zero, then random variables $\mathrm{sgn}(\mathbf{x}_u^\top \mathbf{x}_v)$ for all $u \neq v \in V$ should also be centered on zero, as we prove in Lemma 1, Appendix A. In particular, in the case of scalar signals, i.e., $F = 1$, we can factorize $\mathrm{sgn}(\mathbf{x}_u \mathbf{x}_v) = \mathrm{sgn}(\mathbf{x}_u)\mathrm{sgn}(\mathbf{x}_v)$ so that it represents a sign change between the signals of nodes u and v. We observe that large values of $C_G(\mathbf{X}) \gg 0$ indicate the presence of few sign changes between observations of adjacent nodes, which in turn suggests correlation among variables. Similarly, inverse correlation is revealed by $C_G(\mathbf{X}) \ll 0$. Therefore, $|C_G(\mathbf{X})| \gg 0$ is symptomatic of dependent data, and justifies our test in (2). When $F > 1$, we can interpret the scalar product $\mathbf{x}_u^\top \mathbf{x}_v$ as a correlation between the node signals so that, under the null hypothesis, the expected value of the signs is null and the observed values of $C_G(\mathbf{X})$ are around zero. Another interpretation is more geometrical and treats $\mathbf{x}_u, \mathbf{x}_v$ as points in a Euclidean space. Here, the sign of the scalar (inner) products $\mathbf{x}_u^\top \mathbf{x}_v$ indicates whether the two vectors point in similar or opposing directions, concluding once again that $|C_G(\mathbf{X})|$ is expected to be small when the node signals are independent.

Next Theorem 1 supports the soundness of test (2) and provides a distribution-free criterion to select threshold γ granting a user-defined significance level $\alpha = \mathbb{P}(\text{Reject } H_0 | H_0)$, e.g., $\alpha = 0.05$.

Theorem 1. *Consider a weighted graph $G = (V, E, \mathbf{W})$ without self-loops and a stochastic graph signal $\mathbf{X} = \{\mathbf{x}_v \in \mathbb{R}^F \mid v \in V\}$ on it, with $\mathbf{x}_v \neq 0$ almost surely. Under assumptions*

   (**Ass1**) *$\{\mathbf{x}_v \mid v \in V\}$ are all mutually independent (hypothesis H_0 in (1)),*

   (**Ass2**) *$\mathbb{E}_{\mathbf{x}_v}\left[\mathrm{sgn}(\bar{\mathbf{x}}^\top \mathbf{x}_v)\right] = 0$ for all $\bar{\mathbf{x}} \in \mathbb{R}^F \setminus \{\mathbf{0}\}$ and $v \in V$,*

   (**Ass3**) *$w_{u,v} \in (0, w_+]$ for all $(u,v) \in E$, and $W_2 \to \infty$ as $|E| \to \infty$,*

the distribution of $C_G(\mathbf{X})$ in (3) converges weakly to a standard Gaussian distribution $\mathcal{N}(0,1)$ as the number $|E|$ of edges goes to infinity.

In light of above Theorem 1, threshold γ is selected to be the quantile $1 - \alpha/2$ of the standard Gaussian distribution so that

$$\mathbb{P}(|C_G(\mathbf{X})| > \gamma \mid H_0) = \alpha,$$

thus ensuring to meet the user-defined significance level α. Before sketching the proof, we comment on the assumptions under which the theorem is derived.

Remark 1 (Assumption of no self-loops and $\mathbb{P}(\mathbf{x}_v = 0) = 0$). *Quantities $\mathrm{sgn}(\mathbf{x}_v^\top \mathbf{x}_v)$ arising from self-loops do not carry relevant information to test the independence of node signals. Moreover, we could have expressed (4) as a sum over all edges that are not self-loop. Therefore, the assumption of no self-loop is requested to simplify only the notation. Secondly, requesting null probability of null node signals $\mathbf{x}_v$ is another simplifying assumption that eliminates the probability that $\mathrm{sgn}(\mathbf{x}_v^\top \bar{\mathbf{x}}) = 0$.*

Remark 2 (Assumption (**Ass2**)). *Assumption (**Ass2**) requests that all node signals have a distribution yielding the same probability of being in either of the two half-spaces of $\mathbb{R}^F$ defined by the sign of the scalar product with $\bar{\mathbf{x}}$. Note that in the scalar case with $F = 1$, (**Ass2**) implies that $\mathbb{P}(\mathbf{x}_v > 0) = \mathbb{P}(\mathbf{x}_v < 0)$, which is equivalent to asking that the median of all $\mathbf{x}_v$ is zero. Therefore it does not impose any real constraint on the data distribution because: (i) if the median is not zero, then the model is already not optimal in the sense of the mean absolute deviation and (ii) we can apply the test to the graph signal translated to have zero median. Conversely, for $F > 1$, it is not always possible to center the data so that (**Ass2**) holds.[1] Nevertheless, in this case we can perform separate tests, one for each of the F components, and then either study them individually, employ some multiple hypothesis test correction (e.g., [12]), or sum them together if the components are independent. In any case, this assumption allows node signals to have different distributions.*

---

[1]A counterexample is given by considering all components of all node signals to be i.i.d. with distribution $1/2\, U[-4, 0] + 1/2\, U[0, 1)$, i.e. a mixture of two uniform distributions. For $F = 2$, we have that if (**Ass2**) holds for $\bar{\mathbf{x}} = [0, 1]^\top$ and $[1, 0]^\top$, then it cannot hold also for $\bar{\mathbf{x}} = [1, 1]^\top$.
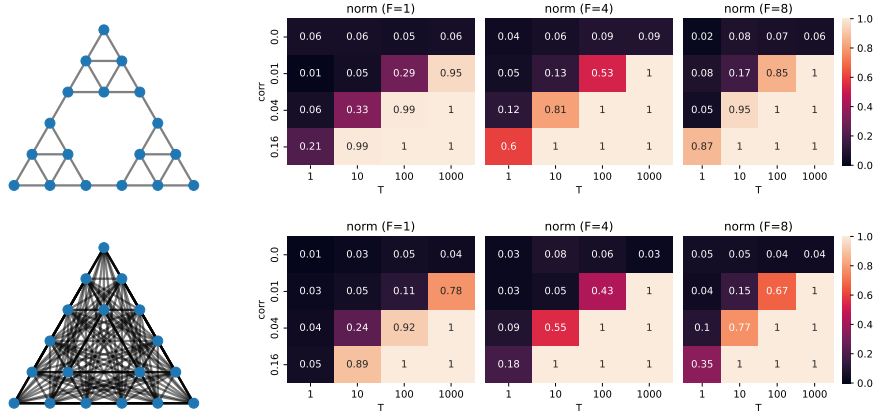
Figure 3: Rate of rejected null hypotheses for different node signal dimensions F, number of time steps T, and correlation parameters $c_{\mathrm{sp}} = c$ and $c_{\mathrm{tm}} = 0$; a detailed description of the data generating process is provided in Section 5.1. The first row corresponds to the summation in (4) only over the edges of the original graph, while the second one to a summation over all pairs of nodes (fully connected graph).

Remark 3 (Assumption (**Ass3**)). *Finally, in Assumption (**Ass3**) we request the weights to be positive. We stress that, as already mentioned, the edge weights are assumed to encode the strength of the relation between the corresponding node so that higher weights imply a stronger impact on statistics (3). An example is when weights come from the absolute value of Pearson's correlation between signals. Not rarely, however, the graph we are given does not come with edge weights or the provided edge attributes do not reflect the criterion assumed above. In all such situations, we can still apply test (2) straightforwardly considering that all weights are equal to 1; accordingly, constant W_2 in (5) becomes equal to the number $|E|$ of edges in the graph. Other criteria to re-weight the graph can be designed for specific cases. The assumption of bounded weights, instead, is technical and only takes part in the limit case of $|E| \to \infty$. Intuitively, (**Ass3**) ensures that all edges bring a tangible contribution to the final statistics $C_G(\mathbf{X})$. The same theorem can be proven under milder assumptions; see Appendix A.*

Sketch of the proof. The proof of Theorem 1 is based on the fact that, under (**Ass1**) and (**Ass2**), random variables $\mathrm{sgn}(\mathbf{x}_v^\top \mathbf{x}_u)$, for all $(u, v) \in E$, are mutually independent, including edges that share one of the two ending nodes. It follows that statistic $\widetilde{C}_G(\mathbf{X})$ in (4) is a weighted sum of independent Bernoulli random variables. Finally, we prove that, with bounded weights, the Lindeberg condition [Equation 14, Appendix A] holds for $\widetilde{C}_G(\mathbf{X})$ under (**Ass3**), and we conclude the thesis by applying the central limit theorem [1] for independent, but not equally distributed, summands of $\widetilde{C}_G(\mathbf{X})$. Detailed development of the proof is given in Appendix A. □

Despite the apparently little amount of information required to perform the test (only the sign of the scalar products $\mathbf{x}_v^\top \mathbf{x}_u$ of all edges $(u, v) \in E$), Geary [11] have shown that for scalar time-series (equivalent to a line graph) the test performed on par of more sophisticated tests. We give empirical evidence of the effectiveness of our whiteness test in Section 5.

3.1 Sparse versus fully connected graphs

We observe that, if we were to consider all pairs of nodes $(u, v) \in V \times V$ in (4), instead of only the edges present in graph G, then several additional terms that, usually, carry negligible functional dependency would be considered in the sum (4). As a result, the impacts of more sensible correlations between nodes close in the graph can wash out. We empirically demonstrate this effect in Figure 3 where, for the same configurations of the simulated graph signal, considering only the edges in E

instead of all possible pairs of nodes, results in a higher statistical power, namely, a higher rate of correctly rejecting the null hypothesis.

Moreover, from formula (4), we immediately see that the computational complexity of the test scales linearly in the number of edges $|E|$. Therefore, the test is more scalable for sparse graphs.

3.2 Whiteness test under different problem settings

Referring to Table 1, we see that test (2) has been presented for (i) static graph signals (absent temporal dimension), (ii) weighted and unweighted graphs, (iii) scalar and multivariate node signals, and (iv) 1-hop neighbors. In this section, we conclude all scenarios where the temporal dimension is absent and leave the case of temporal signals and dynamic graphs to Section 4.

Typically, the correlation between node signals decays as the get farther apart in graph. Therefore, the proposed test considers dependencies only among adjacent nodes, that is, 1-hop neighboring nodes, which should be those displaying the strongest dependency. However, when we have reason to think that the dependency among signals covers more than 1-hop neighbors, we can consider K-hop relationships as well for $K > 1$. To do so, it is enough to extend the edge set E with $\cup_{k=1}^{K} E_k$ where

$$E_k = \left\{ (v_1, v_k) \mid \exists (v_1, \ldots, v_k) \text{ with distinct } v_1 \ldots v_k \in V \text{ and } (v_i, v_{i+1}) \in E, \forall i < k \right\},$$

that is, all pairs of nodes connected by paths of length $k \leq K$. The presented strategy can be seen as an extension to graphs of the portmanteau lack-of-fit test presented in [3]. When appropriate, edges can be weighted differently, for instance, according to the associated path length.

From the current section and Remark 3, we see that, before applying test (2), graph G can be suitably modified, or augmented, to best fit the application at hand. In the next section, we demonstrate how to apply test (2) to temporal signals and dynamic graphs, hence covering the remaining configurations in Table 1.

4 Whiteness test for spatio-temporal graph signals

In this section, we show how to apply the test from previous Section 3 to a generic spatio-temporal signal associated with a possibly dynamic and weighted graph.

We consider a time frame $[1, T]$, and a discrete-time dynamic graph

$$G = \left(V[t], E[t], \mathbf{W}[t] \mid t = 1, \ldots, T \right). \tag{6}$$

At each time step t, $(V[t], E[t], \mathbf{W}[t])$ is a (static) weighted graph like the one depicted in Figure 2. Since we are dealing with node-level time series, we have a correspondence between the nodes of G at different time steps and, in general, $V[t_1] \cap V[t_2] \neq \emptyset$ if $t_1 \neq t_2$. Note that the possibility of having a variable node set is important to model scenarios like the integration of new intersections in a street map, the removal of a broken device in a cyber-physical system, or missing data from smart meters in a power grid. The edge weights can be dynamic, as well, and we denote with $w_{u,v}[t] \in \mathbf{W}[t]$ the weight associated with edge (u, v) present in the edge set $E[t]$ at time t.

On top of graph G, we consider a stochastic graph signal

$$\mathbf{X} = \left\{ \mathbf{x}_v[t] \in \mathbb{R}^F \mid v \in V[t],\ t = 1, 2, \ldots, T \right\}. \tag{7}$$

In this new setting, each node v of the graph is associated with a (multivariate) time series $\mathbf{x}_v[t]$ available for all time steps $t = 1, 2, \ldots, T$, if $v \in V[t]$ for all t, or only for a subset of $\{1, 2, \ldots, T\}$. This setting is depicted in Figure 1 and expresses the full potential of the whiteness test we propose.

Here, we are interested in identifying not only spatial dependencies among different nodes, but also temporal ones appearing among observations from the same node at different time steps. Accordingly, we adapt the null and alternative hypotheses in (1) to account for the temporal dimension, and define

$$\begin{cases} H_0: & \text{All pairs } \mathbf{x}_u[t_u], \mathbf{x}_v[t_v] \in \mathbf{X},\ \text{with } (u, t_u) \neq (v, t_v),\ \text{are independent;} \\ H_1: & \text{Some pairs } \mathbf{x}_u[t_u], \mathbf{x}_v[t_v] \in \mathbf{X},\ \text{with } (u, t_u) \neq (v, t_v),\ \text{are dependent.} \end{cases} \tag{8}$$

By suitably constructing a static graph $G_T = (V_T, E_T, \mathbf{W}_T)$ from G as a stack of snapshot graphs $(V[t], E[t], \mathbf{W}[t])$ and a graph signal $\mathbf{X}_T$ with the same observations of $\mathbf{X}$, but located at the nodes of graph G_T, we are able to test hypotheses H_0, H_1 with the static test (2) presented in previous Section 3. The proposed whiteness test for spatio-temporal graph signals becomes

$$\text{If } |C_{G_T}(\mathbf{X}_T)| > \gamma \implies \text{ Reject null hypothesis } H_0, \tag{9}$$

where $C_{G_T}(\mathbf{X}_T)$ is statistic (3) applied to G_T and $\mathbf{X}_T$, and threshold γ is derived from Theorem 1 according to a user-defined significance level α.

As we detail in next Section 4.1, the construction of G_T and $\mathbf{X}_T$ results in a single larger graph preserving all spatial and temporal relations in G and $\mathbf{X}$, for which there is no constraint on the node and edge sets to be constant throughout the entire time frame $[1, T]$. Whiteness test (9) completes the coverage of the configurations reported in Table 1 and, to emphasize its generality and versatility, we name it AZ-test.

4.1   Multiplex graph and associated signal

The idea is to construct graph $G_T = (V_T, E_T, \mathbf{W}_T)$ as a multiplex graph by stacking graphs $(V[t], E[t], \mathbf{W}[t])$ of dynamic graph (6) according to their temporal index t. A representation of graph G_T is given in Figure 1, where the nodes of G are replicated for all their occurrences across time preserving their spatial connectivity (solid lines), and temporal edges (dashed lines) are added by connecting corresponding nodes at subsequent time steps. When graph $G = (V, E, \mathbf{W})$ is static, we simply replicate it for all considered time steps. Formally, G_T and $\mathbf{X}_T$ are constructed as follows.

Nodes We define node set V_T as the disjoint union of $V[t]$, for all t, that is,

$$V_T = \left\{ v[t] := (v, t) \mid v \in V[t], t = 1, 2, \ldots, T \right\}.$$

If node v is not present in node set $V[t']$, for some t', then $(v, t') \notin V_T$. Therefore, the number of nodes $|V_T|$ is equal to $\sum_t |V[t]|$ and, for a static graph G, we would have $|V_T| = |V| \cdot T$.

Edges The edge set E_T is composed of two types of edges: spatial and temporal edges. The set of spatial edges is the disjoint union of sets $E[t]$, for all t, that is

$$E_{\text{sp}} = \left\{ (u[t], v[t]) \mid (u, v) \in E[t], \ t = 1, 2, \ldots, T \right\}.$$

Temporal edges, instead, connect consecutive occurrences of the same nodes:

$$E_{\text{tm}} = \left\{ (v[t], v[t+1]) \mid v \in V[t] \cap V[t+1], \ t = 1, 2, \ldots, T-1 \right\}.$$

The edge set E_T is then given by

$$E_T = E_{\text{tm}} \cup E_{\text{sp}}$$

Referring to Figure 1, set E_{sp} collects all edges represented by solid lines, while edges in E_{tm} are represented as dashed lines.

Weights The edge weights in $\mathbf{W}_T$ are constructed similarly to E_T: edges $(u[t], v[t]) \in E_{\text{sp}}$ are associated with weight $w_{u,v}[t] \in \mathbf{W}[t]$ corresponding to edge $(u, v) \in E[t]$, while the weight of the temporal edges can be set arbitrarily. In Section 4.2, we propose a reasonable way to define a single weight w_{tm} for all temporal edges in E_{tm}.

Signals Defined G_T, the graph signal $\mathbf{X}$ in (7) can be arranged over the new node set V_T:

$$\mathbf{X}_T = \left\{ \mathbf{x}_{v[t]} := \mathbf{x}_v[t] \in \mathbb{R}^F \mid v[t] \in V_T, \text{ with } \mathbf{x}_v[t] \in \mathbf{X} \right\}.$$

To conclude, we observe that when $T = 1$ the described procedure degenerates to that presented in Section 3. Secondly, we note that the construction of G_T and $\mathbf{X}_T$ is convenient for the theoretical development because it allows applying all the results from Section 3, however, a practical implementation does not need the full G_T graph to be explicitly constructed, as will be more evident from the next section.

8

4.2 Weights for temporal edges

Commenting further about weighting the temporal edges, we note that by definition of G_T there is typically a larger number of spatial edges, than temporal edges; in the case of a static unweighted graph G, $|E_{\text{tm}}| = (T-1) \cdot |V|$, while

$$|E_{\text{sp}}| = T \cdot |V| \cdot d \approx d \cdot |E_{\text{tm}}|,$$

where d is the average degree of the nodes in G. Consequently, an imbalance of positive and negative signs encountered along the temporal edges has, overall, a lower impact on the final test statistic $C_{G_T}(\mathbf{X}_T)$ than that of the spatial edges. Accordingly, we may find it appropriate to weight temporal edges so that spatial and temporal relations are of comparable importance in $C_{G_T}(\mathbf{X}_T)$. In the remainder of this section, we derive a weight w_{tm} that meets the above criterion, when associated with all temporal edges.

Statistics $C_{G_T}(\mathbf{X}_T)$ can be decomposed into a sum of contributions of the spatial and temporal edges, and can be rewritten as

$$C_{G_T}(\mathbf{X}_T) = \frac{\widetilde{C}_{\text{sp}} + \widetilde{C}_{\text{tm}}}{(W_{2,\text{sp}} + W_{2,\text{tm}})^{\frac{1}{2}}}$$

where

$$\widetilde{C}_{\text{sp}} = \sum_{(u[t], v[t]) \in E_{\text{sp}}} w_{u,v}[t] \cdot \text{sgn}\left(\mathbf{x}_u[t]^\top \mathbf{x}_v[t]\right),$$

$$\widetilde{C}_{\text{tm}} = \sum_{(v[t], v[t+1]) \in E_{\text{tm}}} w_{\text{tm}} \cdot \text{sgn}\left(\mathbf{x}_v[t]^\top \mathbf{x}_v[t+1]\right),$$

$$W_{2,\text{sp}} = \sum_{(u[t], v[t]) \in E_{\text{sp}} \,:\, u < v} (w_{u,v}[t] + w_{v,u}[t])^2,$$

$$W_{2,\text{tm}} = |E_{\text{tm}}| \cdot w_{\text{tm}}^2;$$

as in the definition of W_2 in (5), if $(u,v) \notin E[t]$, then we assume $w_{u,v}[t] = 0$.

Under the assumptions of Theorem 1, we have that $\mathbb{E}[\widetilde{C}_{\text{sp}}] = 0 = \mathbb{E}[\widetilde{C}_{\text{tm}}]$, while $\text{Var}[\widetilde{C}_{\text{sp}}] = W_{2,\text{sp}}$ and $\text{Var}[\widetilde{C}_{\text{tm}}] = W_{2,\text{tm}}$. We conclude that balancing the spatial and temporal contributions amounts to select w_{tm} such that the variance is the same, *i.e.*, $W_{2,\text{sp}} = W_{2,\text{tm}}$. We result in

$$w_{\text{tm}} = \left(\frac{W_{2,\text{sp}}}{|E_{\text{tm}}|}\right)^{\frac{1}{2}}.$$

When appropriate, a convex combination of spatial and temporal components can be defined in terms of a scalar value $\lambda \in [0,1]$ and obtain statistics

$$C_{G_T}(\mathbf{X}_T; \lambda) := \frac{\lambda \, \widetilde{C}_{\text{sp}} + (1 - \lambda) \, \widetilde{C}_{\text{tm}}}{(\lambda^2 \, W_{2,\text{sp}} + (1 - \lambda)^2 \, W_{2,\text{tm}})^{\frac{1}{2}}}. \tag{10}$$

Note that the new statistic defined in (10) enjoys all the properties presented in the previous section; in fact, it is equivalent to scaling all the weights of G_T of either λ or $(1 - \lambda)$, and $C_{G_T}(\mathbf{X}_T; 1/2) = C_{G_T}(\mathbf{X}_T)$. In particular, as a corollary of Theorem 1, $C_{G_T}(\mathbf{X}_T; \lambda)$ is approximately distributed as a standard Gaussian under the null hypothesis, and it can be directly employed in AZ-test 9.

5 Experiments

We consider two experimental setups. In the first one, Section 5.1, we generate correlated graph signals and we study the ability of the AZ-test to detect data dependency of various amplitude. In the second one, Section 5.2, we test the optimality of machine-learning models trained to solve forecasting problems on synthetic and real-world data.

5.1 Detection of correlated residuals

We consider the undirected unweighted graph $G = (V, E)$ depicted in Figure 4 and denote with $\mathbf{A}$ its adjacency matrix, that is $\mathbf{A}_{u,v} = 1$ if $(u, v) \in E$ and $\mathbf{A}_{u,v} = 0$ otherwise. In this set of experiments, we generate white-noise graph signals and correlated graph signals with the following procedure.

5.1.1 Signal generation and testing procedure

We consider a scalar probability distribution P with null median, and we sample $|V| \cdot T$ independent values $\eta_v[t]$ from P, one for all $t = 1 \ldots T$ and $v \in V$. We obtain that $\mathbf{X} = \{\mathbf{x}_v[t] = \eta_v[t] \mid \forall t, v\}$ is a graph signal of independent observation, meeting our null hypothesis H_0 in (8).

We generate correlated signals by propagating an independent signal $\{\eta_v[t] \mid \forall t, v\}$ across graph G and the time dimension. Specifically, we consider the following model

$$\mathbf{x}_v[t] = \eta_v[t] + c_{\text{tm}}\, \eta_v[t-1] + c_{\text{sp}} \left(\sum_{(u,v) \in E} w_{u,v}[t]\, \eta_u[t] \right) - m, \qquad (11)$$

for all $t = 2, 3, \ldots, T$ and $v \in V$. The correlation between observations is controlled by positive parameters c_{sp} and c_{tm}. Finally, a scalar offset m is subtracted to have a process $\mathbf{X}$ of null median, thus removing side effects that could make the AZ-test fail.

We consider different distributions P, including unimodal symmetric distributions, as well as asymmetric and bimodal ones. In particular, we select the standard Gaussian distribution $\mathcal{N}(0, 1)$, the chi-squared distributions[2] $\chi_2(d)$ with $d = 1, 5$ degrees of freedom, a mixture of $\mathcal{N}(-3, 1)$ and $\mathcal{N}(+3, 1)$, a mixture of $\chi_2(1)$ and $-\chi_2(5)$, and a mixture of uniform distributions $U[-4, 0)$ and $U[0, 1)$. We generate graph signals for variable number T of time steps, dimension F of the node features, and parameters c_{sp} and c_{tm}. We also include the case with $c_{\text{sp}} = c_{\text{tm}} = 0$ to test the rate of false positives, and to verify that we are able to replicate the user-defined significance level α.

Figure 4 displays sample signals generated from different values of c. We observe that the higher the value of c, the more correlated the signals appear on both the temporal and spatial axes. Temporal correlation is visible as horizontal strikes of similar color in the heatmaps. Instead, having numbered the nodes so that neighboring nodes have similar indices,[3] the spatial correlation is suggested by vertical strikes.

5.1.2 Results

All experiments in the current section are repeated 100 times, each time generating a different sequence according to (11) and applying test (9) with threshold γ set to meet a user-defined significance level $\alpha = 0.05$. When not specified, we consider test (9), *i.e.*, statistic $C_{G_T}(\mathbf{X}_T; 0.5)$ in (10). The considered figure of merit is the rate of rejected null hypotheses, that is, the number of times the test statistic is greater than γ over the total number of repetitions of the same experiment. This setting also applies to Figure 3 discussed in Section 3.1.

Power of the test. Figure 5 shows the ability of the test to identify deviations from the null hypothesis for different values of parameter $c_{\text{tm}} = c_{\text{sp}} = c$, distribution P and number of time steps T. We observe that (i) the higher the value of c, the easier it is for the AZ-whiteness test to identify the correlation, and (ii) the test becomes more powerful as the number T of time steps increases. Similar results are displayed in Figure 3, where we varied the node feature dimension F.

---

[2]Distributions $\chi_2(d)$ are shifted to have null median.

[3]We numbered the nodes according to the magnitude of components of the eigenvector associated with the smallest non-null eigenvalue of the Laplacian matrix $\mathbf{L}$ of G. The Laplacian of an undirected graph with adjacency matrix $\mathbf{A}$ is defined by $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}_{ij} = 0$ if $i \neq j$, and $D_{ii} = \sum_j \mathbf{A}_{ij}$.
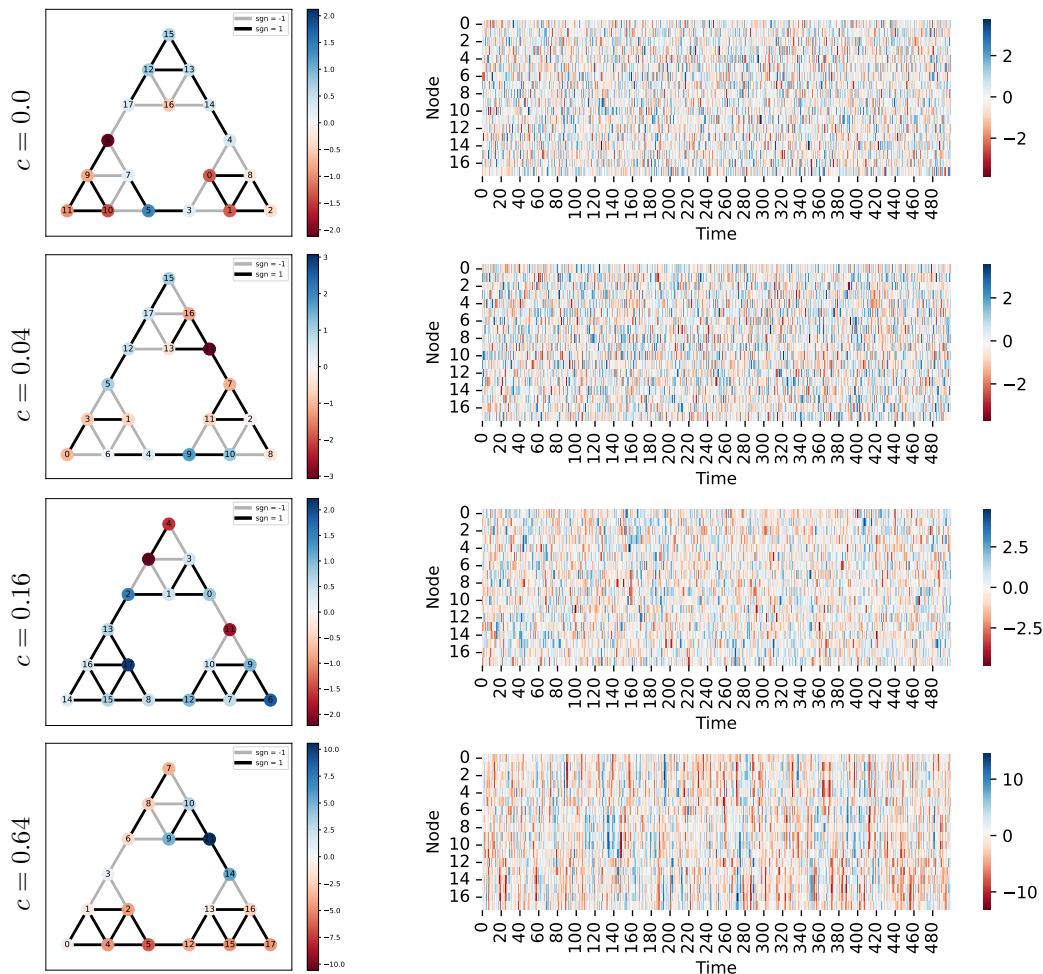
Figure 4: Graph signals displaying different levels of correlation. Each row corresponds to a different value of $c_{\mathrm{tm}} = c_{\mathrm{sp}} = c$ in (11). The number of time steps is $T = 500$, the node feature dimension is $F = 1$, and the distribution of all $\eta_v[t]$ is a standard Gaussian. On the right-hand side, we draw a heatmap of graph signals $\mathbf{X}$. On the left-hand side, we draw the underlying graph with nodes numbered according to the ordering used in the heatmaps. The node color encodes the value of the node signals at time $T/2$, and the edge color encodes the sign of the product of the corresponding node signals, $i.e.$, $\mathrm{sgn}(\mathbf{x}_u[T/2]\mathbf{x}_v[T/2])$ for all edges (u, v).
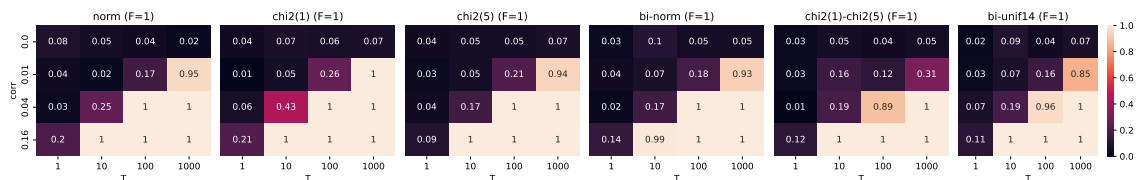

Figure 5: Rate of rejected null hypotheses for significance level $\alpha = 0.05$. Each block corresponds to a different distribution P. Every block has different correlation parameters c on the rows and number of time steps T on the columns. The node feature dimension is $F = 1$. The underlying graph is that of Figure 4.
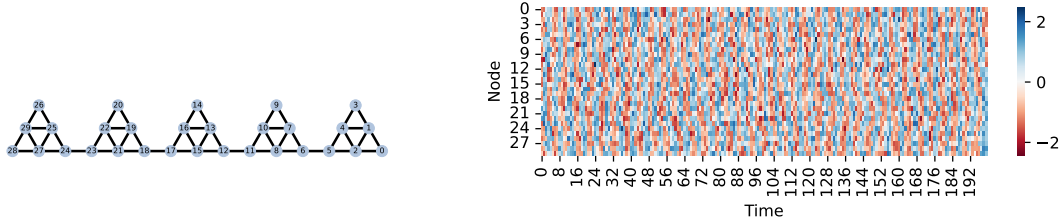
11

Figure 6: Graph and graph signal from the GPVAR dataset of Section 5.2. The displayed graph signal consists of the last $T = 200$ consecutive time steps of the entire dataset.

Calibration of the test. From Figure 5, we can also analyze the correct calibration of the test in terms of desired significance level α. We can see that, regardless of the data distribution P (symmetric/asymmetric, unimodal/bimodal), when the signal observations are independent — *i.e.*, $c = 0$ — the rate of rejected null hypotheses is around the predefined significance level $\alpha = 0.05$, hence suggesting that the calibration is appropriate. For node feature dimensions $F > 1$, we refer to Figure 3 where, again, the observed rejection rates are close to the desired value α.

5.2   Optimality of forecasting models

As a second set of experiments, we train state-of-the-art forecasting models on different synthetic and real-world graph signals, and we assess if they produce correlated residuals, hence implying that there is further information in the data that the models were not able to learn.

5.2.1   Experimental setting

We consider three datasets: GPVAR, PemsBay, and MetrLA.

GPVAR GPVAR is a synthetic dataset we generated from a graph polynomial vector autoregressive system model [14]. The model generates each observation $\mathbf{x}_v[t] \in \mathbb{R}$ at time t and node $v \in V$ from $Q \in \mathbb{N}$ observations in the past and L-hop neighboring nodes, with $L \in \mathbb{N}$, as follows:

$$\mathbf{x}[t] = \tanh \left(\sum_{l=0}^{L} \sum_{q=1}^{Q} \Theta_{l,q} \mathbf{S}^l \mathbf{x}[t - p] \right) + \eta[t] \tag{12}$$

where $\mathbf{x}[t] \in \mathbb{R}^N$ concatenates all scalar node signals at time step t, $\mathbf{S}$ is a graph shift operator, $\Theta \in \mathbb{R}^{(L+1) \times Q}$ collects the model parameters, and $\eta[t] \in \mathbb{R}^N$ is white noise generated form the standard Gaussian distribution [Section 5.1.1]. In this paper, we consider the undirected and unweighted graph shown in Figure 6 with the following shift operator $\mathbf{S} = \mathbf{D}^{-1/2}(\mathbf{I} + \mathbf{A})\mathbf{D}^{-1/2}$ where $\mathbf{A}$ is the adjacency matrix (with now self-loops), $\mathbf{D}$ is the diagonal degree matrix, and $\mathbf{I}$ the identity matrix. Model parameters Θ are set to $[[5, 2], [-4, 6], [-1, 0]]^\top$ and result in $L = Q = 2$. We generate $T = 30000$ time steps. Figure 6 displays a graph signal generated according to (12).

PemsBay PemsBay is a traffic dataset collected by the California Transportation Agencies Performance Measurement System (PeMS) [16]. It presents $T = 52128$ scalar observations from $N = 325$ sensors in the Bay Area.

MetrLA MetrLA is a dataset containing traffic information from $N = 207$ detectors along the Los Angeles County highway [16]. Data observations are collect for 4 months and amount to $T = 34272$ time steps.

The graphs for PemsBay and MetrLA are constructed from the distances $\Delta_{u,v}$ between nodes $u, v \in V$; not all pairwise distances are available. When $\Delta_{u,v} \in (0, \kappa)$, then the weighted adjacency

Table 2: Analysis of the observed residuals on GPVAR data. MAE is in the unit of measure of the graph signal, whereas p-values are reported as results of the statistical tests. AZ-tests with p-values larger than 0.01 are represented in bold.

	MAE	Median=0	AZ-test($\lambda = 0$)	AZ-test($\lambda = 1/2$)	AZ-test($\lambda = 1$)
Optimal Predictor	0.319	0.082	**0.354**	**0.416**	**0.823**
FCRNN	0.384	0.009	<0.001	<0.001	**0.066**
FCRNM-m	0.384	1.000	<0.001	<0.001	**0.056**
GWNet	0.323	<0.001	**0.705**	**0.709**	**0.881**
GWNet-m	0.323	1.000	**0.730**	**0.629**	**0.736**
GatedGN	0.321	<0.001	0.005	**0.171**	**0.414**
GatedGN-m	0.321	1.000	**0.015**	**0.257**	**0.411**
DCRNN	0.327	<0.001	**0.533**	**0.955**	**0.587**
DCRNN-m	0.327	1.000	**0.428**	**0.776**	**0.69**

matrix $\mathbf{A}_{u,v} = 1$ and $w_{u,v} = e^{-\Delta_{u,v}^2/\sigma}$, otherwise is null; σ is chosen to be the standard deviation of all distances in $(0, \kappa)$ [16].

The task to solve is a 1-step-ahead forecasting problem where, for every t, we want to predict graph signal $\mathbf{x}[t]$ from window $\mathbf{X}_{t-\tau:t-1} = \{\mathbf{x}[t-j] \in \mathbb{R}^N \mid j = 1, \ldots, \tau\}$; τ is the size of the considered temporal window. In addition, for PemsBay and MetrLA, a positional encoding of the day of observation is added as exogenous variable. On the above forecasting tasks, we trained a Graph WaveNet [22] (GWNet), a Gated Graph Network [20] (GatedGN), and a Diffusion Convolutional RNN [16] (DCRNN). We also consider a baseline model (FCRNN) composed of a 1-layer fully-connected encoder $\mathbf{z}[t] = f(\mathbf{x}[t]; \mathbf{u}[t])$ processing graph signal $\mathbf{x}[t]$ at each time step t and the associated exogenous variables $\mathbf{u}[t]$, and a 2-layer GRU decoder applied to the resulting window of representations $[\mathbf{z}[t - \tau], \ldots, \mathbf{z}[t - 1]]$.

Because our test assumes that the median of the residuals is null, we report the results when the predictions of the trained models are subtracted by their empirical median, and denote the corrected models with suffix "-m", *i.e.*, FCRNN-m, GWaveNet-m, etc.

The size τ of the temporal window is set to 12 for all experiments. Train, validation and test sets contain 0.7, 0.1, and 0.2 of the original data. The models are trained until convergence with patience of 50 epochs. All models and datasets are available in the TorchSpatiotemporal library [6], except for GPVAR dataset.

In the following section, we report the mean absolute deviation (MAE) achieved by the above methods on the different datasets and the results of a statistical test on the median of the residuals. Metric MAE is, in fact, related to the median of the residuals in that the minimum of the MAE has to produce residuals with null median; see also Remark 2. The test on the median is implemented as a test on the parameter of Bernoulli random variable sgn($\mathbf{x}_v[t]$). Then, we report the results produced by the proposed whiteness test. In addition to the AZ-test run with parameter $\lambda = 1/2$ [Equation 10] accounting for both spatial and temporal correlations, we also considered $\lambda = 0$ and 1 to assess the impact of the temporal and spatial components alone, respectively.

5.2.2 Results

We start by considering the synthetic dataset GPVAR. The results are reported in Table 2. In this experiment, we know the data generating process (12) and we are able to contrast the considered methods against the optimal predictor, *i.e.*, the graph polynomial VAR filter with the same parameter Θ that generated the data. From the first row of Table 2, we observe that the whiteness test (AZ-test) for all values of λ produced p-values significantly different from zero, suggesting uncorrelated residuals. Baseline models FCRNN and FCRNN-m produce MAE substantially higher than the optimal MAE, regardless of the subtraction of the median. Consequently, we observe that AZ-test($\lambda = 1/2$) suggests rejecting the null hypothesis. From the same test with $\lambda = 0$ (only temporal

Table 3: Analysis of the observed residuals on PemsBay data. MAE is in the unit of measure of the graph signal, whereas p-values are reported as results of the statistical tests. AZ-tests with p-values larger than 0.01 are represented in bold.

	MAE	Median=0	AZ-test($\lambda = 0$)	AZ-test($\lambda = 1/2$)	AZ-test($\lambda = 1$)
FCRNN	2.015	<0.001	<0.001	<0.001	<0.001
FCRNN-m	2.015	0.993	<0.001	<0.001	<0.001
GWNet	0.840	<0.001	<0.001	<0.001	<0.001
GWNet-m	0.840	0.990	<0.001	<0.001	<0.001
GatedGN	0.838	<0.001	<0.001	<0.001	<0.001
GatedGN-m	0.838	0.990	<0.001	<0.001	<0.001
DCRNN	0.845	<0.001	<0.001	<0.001	<0.001
DCRNN-m	0.845	0.988	<0.001	<0.001	<0.001

Table 4: Analysis of the residuals observed on MetrLA data. MAE is in the unit of measure of the graph signal, whereas p-values are reported as results of the statistical tests. AZ-tests with p-values larger than 0.01 are represented in bold.

	MAE	Median=0	AZ-test($\lambda = 0$)	AZ-test($\lambda = 1/2$)	AZ-test($\lambda = 1$)
FCRNN	2.841	<0.001	<0.001	<0.001	<0.001
FCRNN-m	2.841	0.999	<0.001	<0.001	<0.001
GWNet	2.114	<0.001	<0.001	<0.001	<0.001
GWNet-m	2.114	1.000	<0.001	<0.001	<0.001
GatedGN	2.150	<0.001	**0.021**	<0.001	<0.001
GatedGN-m	2.150	0.996	**0.016**	<0.001	<0.001
DCRNN	2.141	<0.001	<0.001	<0.001	<0.001
DCRNN-m	2.140	1.000	<0.001	<0.001	<0.001

edges are considered) and $\lambda = 1$ (only spatial edges), we see that the temporal correlation appears more prominent than the spatial one. In contrast, we can consider the training of GWNet, GatedGN, and DCRNN successful because they produce MAE close to the target value and relatively high p-values in the AZ-tests.

Regarding PemsBay and MetrLA datasets, we do not have a reference MAE, as the optimal predictor is unknown. From Tables 3 and 4, we observe that none of DCRNN, GatedGN, and GWNet (and their variants) produces uncorrelated residuals. However, in Table 4 we see that the GatedGN produces residuals with a temporal correlation higher than the spatial one because when $\lambda = 0$ the p-values are significantly larger than when λ is 1.

Finally, we point out that the null hypothesis is not rejected because Assumption (**Ass2**) is not valid, in fact, in all cases, almost identical results are observed with the original methods (FCRNN, GWNet, GatedGN, DCRNN) and those with the median offset removed (FCRNN-m, GWNet-m, GatedGN-m, DCRNN-m).

6   Conclusions

In this work, we propose the first whiteness test for spatio-temporal time series defined over the nodes of a graph. By exploiting the known structural and functional relations defined by the underlying graph, we are able to identify both temporal and spatial correlations that exist among data observations. We show that the test is asymptotically distribution-free with respect to the number of time steps and graph edges and, therefore, we can apply to data coming from arbitrary distributions and, even, nonidentically distributed observations.

The designed test is general, from which the definitive alphabetic encompassing name, and based

on sign changes between observations that are close in time or with respect to the graph connectivity. The test is very versatile, too, as it can exploit edge weights encoding the strength of the link, allows the graph topology to vary over time, and can operate when nodes are inserted or removed from the graph; these scenarios are frequent when dealing with real-world graph signals, like those coming from sensor networks, *e.g.*, transportation networks, and smart grids. Finally, the test is computationally scalable for sparse graphs with complexity linear in the number of edges and time steps.

We empirically show that the proposed test is indeed capable of identifying dependencies among graph signals; when applied to analyze prediction residuals, the AZ-test can assess whether given forecasting models can be assumed to be optimal or not.

The proposed whiteness test is pioneering in its nature, and allows for future extensions involving stochastic graphs and more sophisticated edge statistics.

Acknowledgements

References

[1] P. Billingsley. Probability and measure 3rd edition new york, 1995.

[2] A. Bose and W. Hachem. A whiteness test based on the spectral measure of large non-hermitian random matrices. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8768–8771. IEEE, 2020.

[3] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time series analysis: forecasting and control.* John Wiley & Sons, 2015.

[4] H. Chen and N. R. Zhang. Graph-based tests for two-sample comparisons of categorical data. *Statistica Sinica*, pages 1479–1503, 2013.

[5] R. V. Chitturi. Distribution of residual autocorrelations in multiple autoregressive schemes. *Journal of the American Statistical Association*, 69(348):928–934, 1974.

[6] A. Cini and I. Marisca. Torch Spatiotemporal, 3 2022. URL `https://github.com/TorchSpatiotemporal/tsl`.

[7] A. Cini, I. Marisca, and C. Alippi. Filling the g_ap_s: Multivariate time series imputation by graph neural networks. In *International Conference on Learning Representations*, 2021.

[8] K. Drouiche. A new test for whiteness. *IEEE transactions on signal processing*, 48(7):1864–1871, 2000.

[9] J. Durbin and G. S. Watson. Testing for serial correlation in least squares regression: I. *Biometrika*, 37(3/4):409–428, 1950.

[10] J. H. Friedman and L. C. Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, pages 697–717, 1979.

[11] R. C. Geary. Relative efficiency of count of sign changes for assessing residual autoregression in least squares regression. 1970.

[12] Y. Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75 (4):800–802, 1988.

[13] J. R. Hosking. The multivariate portmanteau statistic. *Journal of the American Statistical Association*, 75(371):602–608, 1980.

[14] E. Isufi, A. Loukas, N. Perraudin, and G. Leus. Forecasting time series with varma recursions on graphs. *IEEE Transactions on Signal Processing*, 67(18):4870–4885, 2019.

[15] W. Li and A. McLeod. Distribution of the residual autocorrelations in multivariate arma time series models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(2):231–239, 1981.

[16] Y. Li, R. Yu, C. Shahabi, and Y. Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SJiHXGWAZ.

[17] Z. Li, C. Lam, J. Yao, and Q. Yao. On testing for high-dimensional white noise. *The Annals of Statistics*, 47(6):3382–3412, 2019.

[18] G. M. Ljung and G. E. Box. On a measure of lack of fit in time series models. *Biometrika*, 65 (2):297–303, 1978.

[19] P. R. Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):515–530, 2005.

[20] V. G. Satorras, S. S. Rangapuram, and T. Januschowski. Multivariate time series forecasting with latent graph inference. *arXiv preprint arXiv:2203.03423*, 2022.

[21] A. Wald and J. Wolfowitz. On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*, 11(2):147–162, 1940.

[22] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *The 28th International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence Organization, 2019.

A   Proof of Theorem 1

We start by proving two auxiliary lemmas, Lemma 1 and Lemma 2, which will be used to prove Theorem 1.

For brevity, in the following we use the short forms

$$s(e) := s((u, v)) := \operatorname{sgn}(\mathbf{x}_u^\top \mathbf{x}_v),$$
$$\pi_v(\bar{\mathbf{x}}) := \mathbb{P}_{\mathbf{x}_v}(\bar{\mathbf{x}}^\top \mathbf{x}_v),$$
$$\pi_e := \pi_{(u,v)} := \mathbb{E}_{\mathbf{x}_v}\left[\pi_u(\mathbf{x}_v)\right].$$

Lemma 1. *Under assumption* (**Ass1**) *of independent* $\mathbf{x}_v$ *for* $v \in V$, *we have that*

$$\mathbb{P}(s(e) = a) = \mathbb{E}_{\mathbf{x}_v}[\pi_u(a\mathbf{x}_v)] = \mathbb{E}_{\mathbf{x}_u}[\pi_v(a\mathbf{x}_u)]$$
$$= \begin{cases} 0 & a = 0, \\ \pi_e & a > 0, \\ 1 - \pi_e & a < 0. \end{cases}$$

Furthermore, if (**Ass2**) *holds, then* $\mathbb{E}[s(e)] = 0$.

Proof. **1)** For any value $a \in \{-1, 0, 1\}$, edge $e = (u, v) \in E$, and distributions P_v of all node signal $\mathbf{x}_v$ for $v \in V$, we have:

$$\mathbb{P}(s(e) = a) = \mathbb{P}(a\,\mathbf{x}_v^\top \mathbf{x}_u > 0)$$

$$= \int \mathbb{P}(a\,\bar{\mathbf{x}}^\top \mathbf{x}_u > 0 | \mathbf{x}_v = \bar{\mathbf{x}})\,\mathrm{d}P_v(\bar{\mathbf{x}})$$

$$(\mathbf{Ass1}) \quad = \int_{\mathbb{R}^F \setminus \{0\}} \mathbb{P}(a\,\bar{\mathbf{x}}^\top \mathbf{x}_u > 0)\,\mathrm{d}P_v(\bar{\mathbf{x}})$$

$$= \int \pi_u(a\bar{\mathbf{x}})\,\mathrm{d}P_v(\bar{\mathbf{x}}) = \mathbb{E}_{\bar{\mathbf{x}} \sim P_v}[\pi_u(a\bar{\mathbf{x}})] = \pi_e$$

With analogous developments, we obtain $\mathbb{P}(s(e) = a) = \mathbb{E}_{\bar{\mathbf{x}} \sim P_u}[\pi_v(a\bar{\mathbf{x}})]$, so $\pi_e = \pi_{(u,v)} = \pi_{(v,u)}$.

2) Because for all v

$$\pi_v(\bar{\mathbf{x}}) = \mathbb{P}_{\mathbf{x}_v}(\mathbf{x}_v^\top \bar{\mathbf{x}} > 0) = 1 - \mathbb{P}_{\mathbf{x}_v}(\mathbf{x}_v^\top \bar{\mathbf{x}} \leq 0) = 1 - \pi_v(-\bar{\mathbf{x}}),$$

we conclude that

$$\mathbb{P}(s(e) = a) = \begin{cases} 0 & a = 0, \\ \pi_e & a > 0, \\ 1 - \pi_e & a < 0, \end{cases}$$

hence proving the first part of the thesis.

3) We observe that for all $u \in V, \bar{\mathbf{x}} \neq \mathbf{0}$:

$$\mathbb{E}_{\bar{\mathbf{x}} \sim P_v}[\mathrm{sgn}(\bar{\mathbf{x}}^\top \mathbf{x}_u)] = 1 \cdot \mathbb{P}(\mathrm{sgn}(\bar{\mathbf{x}}^\top \mathbf{x}_u) = 1) + 0 + (-1)\mathbb{P}(\mathrm{sgn}(\bar{\mathbf{x}}^\top \mathbf{x}_u) = -1)$$

$$= \pi_u(\bar{\mathbf{x}}) - (1 - \pi_u(\bar{\mathbf{x}})) = 2\pi_u(\bar{\mathbf{x}}) - 1.$$

Therefore, we have the following equivalence

$$\text{Assumption } (\mathbf{Ass2}) \iff \pi_v(\bar{\mathbf{x}}) = \frac{1}{2} \text{ for all } v \in V, \ \bar{\mathbf{x}} \neq \mathbf{0}. \tag{13}$$

We conclude that, under $(\mathbf{Ass2})$,

$$\mathbb{E}[s(e)] = 1 \cdot \pi_e + 0 + (-1)(1 - \pi_e) = 2\pi_e - 1$$

$$= 2\mathbb{E}_{\mathbf{x}_v}[\pi_u(\mathbf{x}_v)] - 1 = 2\frac{1}{2} - 1 = 0.$$

$\square$

Lemma 2. *Consider two distinct edges $e, f \in E$ that are not self-loop.*

- *If $(\mathbf{Ass1})$ holds and e, f are not adjacent, then random variables $s(e)$ and $s(f)$ are independent.*

- *Otherwise, if $e = (u, v), f = (u, w)$ for $v \neq w$, then $s(e)$ and $s(f)$ are independent if both $(\mathbf{Ass1})$ and $(\mathbf{Ass2})$ hold.*

Proof. To prove the thesis, we need to show that, for every $a, b \in \{-1, 0, 1\}$,

$$(*) := \mathbb{P}(s(e) = a)\mathbb{P}(s(f) = b) = \mathbb{P}(s(e) = a, s(f) = b) =: (**)$$

1) If e and f are not adjacent, then $s(e), s(f)$ are independent because computed from different node pairs, which are independent by $(\mathbf{Ass1})$. This proves the first part.

2) Consider now $e = (u, v)$ and $f = (u, w)$ for some distinct $u, v, w \in V$. We start from $(*)$ and, in light of Lemma 1, we have that

$$(*) = \mathbb{E}_{\mathbf{x}_v}[\pi_u(a\mathbf{x}_v)] \cdot \mathbb{E}_{\mathbf{x}_v}[\pi_u(b\mathbf{x}_w)].$$

3) On the other side

$$(**) = \mathbb{P}(s(e) = a, s(f) = b)$$

$$= \int \mathbb{P}(a\,\bar{\mathbf{x}}^\top \mathbf{x}_v > 0, b\,\bar{\mathbf{x}}^\top \mathbf{x}_w > 0 | \mathbf{x}_u = \bar{\mathbf{x}})\,\mathrm{d}P_u(\bar{\mathbf{x}})$$

$$(\mathbf{Ass1}) = \int \mathbb{P}(a\,\bar{\mathbf{x}}^\top \mathbf{x}_v > 0)\mathbb{P}(b\,\bar{\mathbf{x}}^\top \mathbf{x}_w > 0)\,\mathrm{d}P_u(\bar{\mathbf{x}})$$

$$= \mathbb{E}_{\mathbf{x}_u}\left[\pi_v(a\,\mathbf{x}_u)\pi_w(b\,\mathbf{x}_u)\right].$$

4) For example, in the case of P_v is the same for all $v \in V$, we reduce to Jensen's inequality, yielding $(*) = (**)$ only in particular cases, like when $\pi_v(\mathbf{x})$ is affine on the support of $\mathbf{x}$. However, under the additional assumption of $(\mathbf{Ass2})$, we have that $\pi_v(\mathbf{x}) = 1/2$ for all v and $\mathbf{x}$, so

$$(*) = \begin{cases} 0 & ab = 0, \\ 1/4 & ab \neq 0 \end{cases} = (**).$$

This proves the independency of $s(e)$ and $s(f)$.

$\square$

Proven Lemmas 1 and 2, we return to Theorem 1. Without loss of generality, we can assume that the given graph $G = (V, E, \mathbf{W})$ is undirected. Otherwise, we can always construct an undirected graph $\widetilde{G} = (\widetilde{V}, \widetilde{E}, \widetilde{\mathbf{W}})$ from directed graph G that produces the same test statistics $C_G(\mathbf{X}) = C_{\widetilde{G}}(\mathbf{X})$ and for which W_2 in (5) simplifies to $\sum_{e \in E} w_e^2$; to do so, it is enough to consider $\widetilde{V} = V$, construct $\widetilde{E}_u$ by removing the orientation of the edges in E, and considering $\widetilde{w}_{u,v} \in \widetilde{\mathbf{W}}$ as $w_{u,v} + w_{v,u}$ if both (u, v) and $(v, u) \in E$, otherwise $\widetilde{w}_{u,v}$ equals either $w_{u,v}$ or $w_{v,u}$, depending which edge is present in E. Moreover, we note that the concept of correlation is not directional.

Under the assumption of undirected graph with no self-loops, we see that $\widetilde{C}_G(\mathbf{X})$ in (4) is a weighted sum of i.i.d. Bernoulli random variables:

$$\widetilde{C}_G(\mathbf{X}) = \sum_{(e) \in E} w_e s(e)$$

or, equivalently, a sum of independent [Lemma 2], but not identically distributed terms. Nevertheless, a central limit theorem [1, Th. 27.3] can still be applied to show the thesis if the following Lindeberg condition on random variables $\{w_e\, s(e) \mid e \in E\}$ holds.

Consider sequence of independent processes

$$\{\{X_{n,k} \mid k \leq n\} \mid n \in \mathbb{N}\}$$

with $\mathbb{E}[X_{n,k}] = 0$ and $\mathrm{Var}[X_{n,k}] = \sigma_{n,k}^2$ for all $k \leq n, n \in \mathbb{N}$, and

$$s_n = \left(\sum_k \sigma_{n,k}^2\right)^{1/2}.$$

The Lindeberg condition [1, Eq. 27.16] is the following

$$\lim_{n \to \infty} \frac{1}{s_n^2} \sum_{k=1}^n \int_{|X_{|E|,e}| \geq \varepsilon s_n} |X_{n,k}|^2\,\mathrm{d}P_{n,k} = 0, \quad \text{for all } \varepsilon > 0, \tag{14}$$

where $P_{n,k}$ is the distribution of $X_{n,k}$. We now show that (14) holds for $X_{|E|,e} = w_e s(e)$.

1) Under assumptions (**Ass1**) and (**Ass2**), by Lemma 2 $\{X_{|E|,e} \mid e \in E\}$ are independent and, by Lemma 1, $\mathbb{E}[X_{|E|,e}] = 0$ with

$$\mathbb{E}[X_{|E|,e}^2] = w_e^2 \mathbb{E}[s(e)^2] = w_e^2 \mathbb{E}[1] = w_e^2.$$

So $s_{|E|}^2 = \sum_{e \in E} w_e^2 \cdot 1 = W_2$. Substituting the quantities we have just calculated in (14), we can express the Lindeberg condition as follows

$$\frac{1}{s_{|E|}^2} \sum_{e \in E} \int_{|X_{|E|,e}| \geq \varepsilon s_{|E|}} |w_e\, s(e)|^2 \, \mathrm{d}P_{s(e)} = \frac{1}{s_{|E|}^2} \sum_{e \in E} \int_{|w_e||s(e)| \geq \varepsilon s_{|E|}} w_e^2 \, \mathrm{d}P_{s(e)}$$

$$= \frac{1}{s_{|E|}^2} \sum_{e \in E} w_e^2 I(|w_e| \geq \varepsilon s_{|E|})$$

2) From (**Ass3**), $s_{|E|}^2 = W_2 \to \infty$ as $|E| \to \infty$. Therefore, for every $\varepsilon > 0$ there is a number $N(\varepsilon, w_+)$ of edges such that $w_+ < \varepsilon s_{N(\varepsilon, w_+)}$, and for which $I(|w_e| \geq \varepsilon s_{N(\varepsilon, w_+)}) = 0$ for all e. We conclude that (14) holds true and, in turn, $C_G(\mathbf{X})$ weakly converges to $\mathcal{N}(\mathbb{E}[C_G(\mathbf{X})], \mathrm{Var}[C_G(\mathbf{X})])$ by Theorem 27.3 in [1].

3) Finally, observe that the expected value of $C_G(\mathbf{X})$ is

$$\mathbb{E}[C_G(\mathbf{X})] = \sum_{e \in E} w_e \mathbb{E}[s(e)] W_2^{-1/2} = 0$$

thanks to Lemma 1, and the variance is

$$\mathrm{Var}[C_G(\mathbf{X})] = \frac{1}{W_2} \mathrm{Var}\left[\sum_e w_e s(e)\right]$$

$$(\text{Lem. 2}) = \frac{1}{W_2} \sum_{e \in E} w_e^2 (\mathbb{E}[s(e)^2] - \mathbb{E}[s(e)]^2)$$

$$= \frac{1}{W_2} W_2 (1 - 0) = 1.$$

We concluded the thesis of Theorem 1.