Supplemental Material
AZ-whiteness test: a test for signal uncorrelation on spatio-temporal graphs

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

$$\begin{aligned} s(e) \triangleq s((u,v)) &\triangleq \operatorname{sgn}(\mathbf{x}_u^\top \mathbf{x}_v), \\ \pi_v(\bar{\mathbf{x}}) &\triangleq \mathbb{P}_{\mathbf{x}_v}(\bar{\mathbf{x}}^\top \mathbf{x}_v > 0), \\ \pi_e \triangleq \pi_{(u,v)} &\triangleq \mathbb{E}_{\mathbf{x}_v}\left[\pi_u(\mathbf{x}_v)\right]. \end{aligned}$$

Lemma 1. *Under assumption* (**A1**) *of independent* $\mathbf{x}_v$ *for* $v \in V$*, we have that*

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

Furthermore, if (**A2**) *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:

$$\begin{aligned} \mathbb{P}(s(e)=a) &= \mathbb{P}(a\,\mathbf{x}_v^\top \mathbf{x}_u > 0) \\ &= \int \mathbb{P}(a\,\bar{\mathbf{x}}^\top \mathbf{x}_u > 0 | \mathbf{x}_v = \bar{\mathbf{x}})\, \mathrm{d}P_v(\bar{\mathbf{x}}) \\ \textbf{(A1)} \;&= \int_{\mathbb{R}^F \setminus \{0\}} \mathbb{P}(a\,\bar{\mathbf{x}}^\top \mathbf{x}_u > 0)\, \mathrm{d}P_v(\bar{\mathbf{x}}) \\ &= \int_{\mathbb{R}^F \setminus \{0\}} \pi_u(a\,\bar{\mathbf{x}})\, \mathrm{d}P_v(\bar{\mathbf{x}}) = \mathbb{E}_{\bar{\mathbf{x}} \sim P_v}[\pi_u(a\bar{\mathbf{x}})] \end{aligned}$$

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 $\pi_{(u,v)} = \pi_{(v,u)}$. We conclude that

$$\mathbb{P}(s(e)=a) = \begin{cases} 0 & a=0, \\ \pi_e & a=1, \\ 1-\pi_e & a=-1, \end{cases} \tag{11}$$

hence proving the first part of the thesis.

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

$$\begin{aligned} \mathbb{E}_{\mathbf{x}_u \sim P_u}\left[\operatorname{sgn}\left(\bar{\mathbf{x}}^\top \mathbf{x}_u\right)\right] &= 1 \cdot \mathbb{P}\left(\operatorname{sgn}\left(\bar{\mathbf{x}}^\top \mathbf{x}_u\right) = 1\right) + 0 + (-1)\mathbb{P}\left(\operatorname{sgn}\left(\bar{\mathbf{x}}^\top \mathbf{x}_u\right) = -1\right) \\ &= \pi_u(\bar{\mathbf{x}}) - (1 - \pi_u(\bar{\mathbf{x}})) = 2\pi_u(\bar{\mathbf{x}}) - 1. \end{aligned}$$

Therefore, we have the following equivalence

$$\text{Assumption (\textbf{A2})} \iff \pi_u(\bar{\mathbf{x}}) = \frac{1}{2} \text{ for all } u \in V,\ \bar{\mathbf{x}} \neq \mathbf{0}. \tag{12}$$

From (11) and (**A2**) we conclude that

$$\begin{aligned} \mathbb{E}[s(e)] &= 1 \cdot \pi_e + 0 + (-1)(1-\pi_e) = 2\pi_e - 1 \\ &= 2\mathbb{E}_{\mathbf{x}_v}[\pi_u(\mathbf{x}_v)] - 1 = 2\frac{1}{2} - 1 = 0. \end{aligned}$$

□

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

- *If* (**A1**) *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* (**A1**) *and* (**A2**) *hold.*

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

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

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 (**A1**). 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

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

3) On the other side

$$\begin{aligned}
(**) &= \mathbb{P}(s(e) = a, s(f) = b) \\
&= \int \mathbb{P}(a\,\bar{\mathbf{x}}^\top \mathbf{x}_v > 0, b\,\bar{\mathbf{x}}^\top \mathbf{x}_w > 0 | \mathbf{x}_u = \bar{\mathbf{x}})\, \mathrm{d}P_u(\bar{\mathbf{x}}) \\
\textbf{(A1)} \;&= \int \mathbb{P}(a\,\bar{\mathbf{x}}^\top \mathbf{x}_v > 0)\mathbb{P}(b\,\bar{\mathbf{x}}^\top \mathbf{x}_w > 0)\, \mathrm{d}P_u(\bar{\mathbf{x}}) \\
&= \mathbb{E}_{\mathbf{x}_u}\left[\pi_v(a\,\mathbf{x}_u)\pi_w(b\,\mathbf{x}_u)\right].
\end{aligned}$$

4) Under the additional assumption of (**A2**), we have $\pi_v(\mathbf{x}) = 1/2$ 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 $s(f)$.

□

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 $C(G) = C(\widetilde{G})$ and for which W in (5) simplifies to $\sum_{e \in E} w_e^2$; to do so, it is enough to consider $\widetilde{V} = V$, construct $\widetilde{E}$ 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 on which edge is present in E.

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

$$\widetilde{C}(G) = \sum_{(e) \in E} w_e s(e)$$

or, equivalently, a sum of independent [Lemma 2], but *not* identically distributed terms. However, a central limit theorem [1, Th. 27.3] can still be applied to prove the thesis if the following Lindeberg condition (13) on random variables $\{w_e\, s(e) \mid e \in E\}$ holds.

Consider sequence of independent processes

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

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

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{k=1}^{n} \int_{|X_{|E|,e}|\geq \varepsilon s_n} |X_{n,k}|^2 \, \mathrm{d}P_{n,k} = 0, \quad \text{for all } \varepsilon > 0, \tag{13}$$

where $P_{n,k}$ is the distribution of $X_{n,k}$. Now we show that (13) holds for $X_{|E|,e} = w_e s(e)$.

1) Under assumptions (**A1**) and (**A2**), 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 $s_{|E|}^2 = \sum_{e\in E} w_e^2 \cdot 1 = W$. Substituting the quantities we have just calculated in (13), we can express the Lindeberg condition as follows

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

2) From (**A3**), $s_{|E|}^2 = W \to \infty$ 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 (13) holds true and, in turn, $C(G)$ weakly converges to $\mathcal{N}(\mathbb{E}[C(G)], \mathrm{Var}[C(G)])$ by Theorem 27.3 in [1].

3) Finally, observe that the expected value of $C(G)$ is

$$\mathbb{E}[C(G)] = \sum_{e\in E} w_e \mathbb{E}[s(e)] W^{-1/2} = 0$$

thanks to Lemma 1, and the variance is

$$\begin{aligned}\mathrm{Var}[C(G)] &= \frac{1}{W} \mathrm{Var}\left[\sum_e w_e s(e)\right] \\ \text{(Lem. 2)} \ &= \frac{1}{W} \sum_{e\in E} w_e^2 (\mathbb{E}[s(e)^2] - \mathbb{E}[s(e)]^2) \\ &= \frac{1}{W} W(1-0) = 1.\end{aligned}$$

We concluded the thesis of Theorem 1.

B Comments on the assumptions of Theorem 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-loops. Therefore, the assumption of no self-loops simplifies the notation only. Also, requesting null probability for null node signals $\mathbf{x}_v$ is another simplifying assumption that eliminates the probability that $\mathrm{sgn}(\mathbf{x}_v^\top \bar{\mathbf{x}}) = 0$.

Assumption (A2) Assumption (**A2**) requests 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 $\bar{\mathbf{x}}$, and allows node signals to have different distributions. Note that in the scalar case with $F = 1$, (**A2**) implies that $\mathbb{P}(\mathbf{x}_v > 0) = \mathbb{P}(\mathbf{x}_v < 0)$, which is equivalent to requesting that the median of all $\mathbf{x}_v$ is zero. Therefore it does not impose any real constraint on the data distribution as we can always shift the data to have null median. Conversely, for $F > 1$, it is not always possible to center the data by shifting all signals by vector $\mathbf{x}_0$ so that (**A2**) holds for signals $\mathbf{x}_v - \mathbf{x}_0$. A counterexample is given by considering all components of all node signals to be i.i.d. from the mixture distribution $U[-4, 0) + U[0, 1)$ where, for $F = 2$, $\mathbf{x}_0$ must be $\mathbf{0}$ for (**A2**) to hold for $\bar{\mathbf{x}} = [0, 1]^\top$ and $[1, 0]^\top$,

however, it results that for $\bar{\mathbf{x}} = [1, 1]^\top$ (**A2**) does not hold. That said, for $F > 1$ with invalid assumption (**A2**), we can perform F individual tests with statistic $C(G^{(f)})$, one for each component $\mathbf{X}^{(f)}$, $f = 1, \dots, F$, of the graph signal and then combine the results with a multiple hypothesis test correction, like [13]. Alternatively, when $C(G^{(f)}), f = 1, \dots, T$, are statistically independent, their sum $C^F(G) \triangleq \sum_f C(G^{(f)})$ is again distributed as a Gaussian [Theorem 1] and we can design a test on it.

Assumption (A3) Assumption (**A3**) requests the weights to be positive. Here, 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). As an example, consider the case where weights come from the absolute value of Pearson's correlation between signals. Not rarely, however, the given graph 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 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, (**A3**) ensures that all edges bring a tangible contribution to the final statistics $C(G)$.

C Weights for temporal edges

Commenting further about weighting the temporal edges, we note that by definition of G^* 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 $C(G^*)$ 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 $C(G^*)$. In the remainder of this section, we derive a weight w_{tm} that meets the above criterion when associated with all temporal edges.

Statistics $C(G^*)$ can be decomposed into a sum of contributions of the spatial and temporal edges, and can be rewritten as

$$C(G^*) = \frac{\widetilde{C}_{\text{sp}} + \widetilde{C}_{\text{tm}}}{(W_{\text{sp}} + W_{\text{tm}})^{\frac{1}{2}}}$$

where $\widetilde{C}_{\text{tm}}$ and $\widetilde{C}_{\text{sp}}$ collect the temporal and spatial terms in $\widetilde{C}(G^*)$, respectively, *i.e.*

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

Instead, $W_{\text{tm}} = |E_{\text{tm}}| \cdot w_{\text{tm}}^2$ and W_{sp} is defined as

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

that is similar to (5), but limiting to the weights of the spatial edges, and with $E_{\text{sp}}^{\leftrightarrow}$ containing all self-loops in E_{sp}.

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

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

When appropriate, the user can consider trading off the spatial and temporal components in terms of a scalar value $\lambda \in [0, 1]$ defining a convex combination of $\widetilde{C}_{\mathrm{sp}}$ and $\widetilde{C}_{\mathrm{tm}}$. Accordingly, we define statistic

$$C(G^*; \lambda) \triangleq \frac{\lambda\, \widetilde{C}_{\mathrm{sp}} + (1-\lambda)\, \widetilde{C}_{\mathrm{tm}}}{\left(\lambda^2\, W_{\mathrm{sp}} + (1-\lambda)^2\, W_{\mathrm{tm}}\right)^{\frac{1}{2}}}, \tag{14}$$

which turns out to be equivalent to $C(G^*)$, where the weights of the spatial and temporal edges are scaled by λ and $1-\lambda$, respectively, and $C(G^*; 1/2) = C(G^*)$. Therefore, statistic (14) enjoys all the properties of statistic $C(G^*)$. In particular, as a corollary of Theorem 1, $C(G^*; \lambda)$ is approximately distributed as a standard Gaussian under the null hypothesis, and can be directly employed in AZ-test (8).

D Whiteness test for K-hop and K-lag correlation

When the dependency among signals covers more than 1-hop neighbors, or more than 1 time lag, we can consider K-hop relationships, with $K > 1$. It is enough to extend the edge set E to $\cup_{k=1}^{K} E_k$ where

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

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.

E Synthetic graph signals

In Figure 5, we report graph signals generated according to the procedure of Section 5.1 and displaying different levels of correlation. 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,[6] the spatial correlation is suggested by vertical strikes.

Figure 6 shows the graph considered in the GPVAR dataset and a portion of the graph signal generated according to (10).

F Hardware and software

Experiments are implemented in Python 3.8 and run on an NVIDIA Titan V GPU with 12 GB of memory. The code is based on PyTorch-Geometric [10] and TorchSpatiotemporal [6].

The run time of the experiments in Section 5.1 is in the order of milliseconds for signals with $T = 10^4$ time steps and $F = 1$ node features on an 18-node graph, since only the computation of the statistical test is involved. Instead, the run time of the experiments of Section 5.2, which includes the training time, is about 3 hours for each problem setting, except for the training of DCRNN and GWNet on MetrLA, which took about 20 hours. Apart from minor adjustments (like increased number of epochs), the hyperparameters are the default ones provided by TorchSpatiotemporal (and are detailed in the config files).

G Extended experiments on multivariate node signals

In Figure 7, we extend the analysis of the of Table 3 to multivariate node signals, *i.e.*, when $\mathbf{x}_v[t] \in \mathbb{R}^F$, with $F > 1$. We considered the data generating process in (9), with the components of $\mathbf{Z} \in \mathbb{R}^{|V| \times T \times F}$ sampled i.i.d. from standard Gaussian and different values of c. We observe a slight performance improvement when F increases.

[6] 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 $D_{ii} = \sum_j \mathbf{A}_{ij}$.

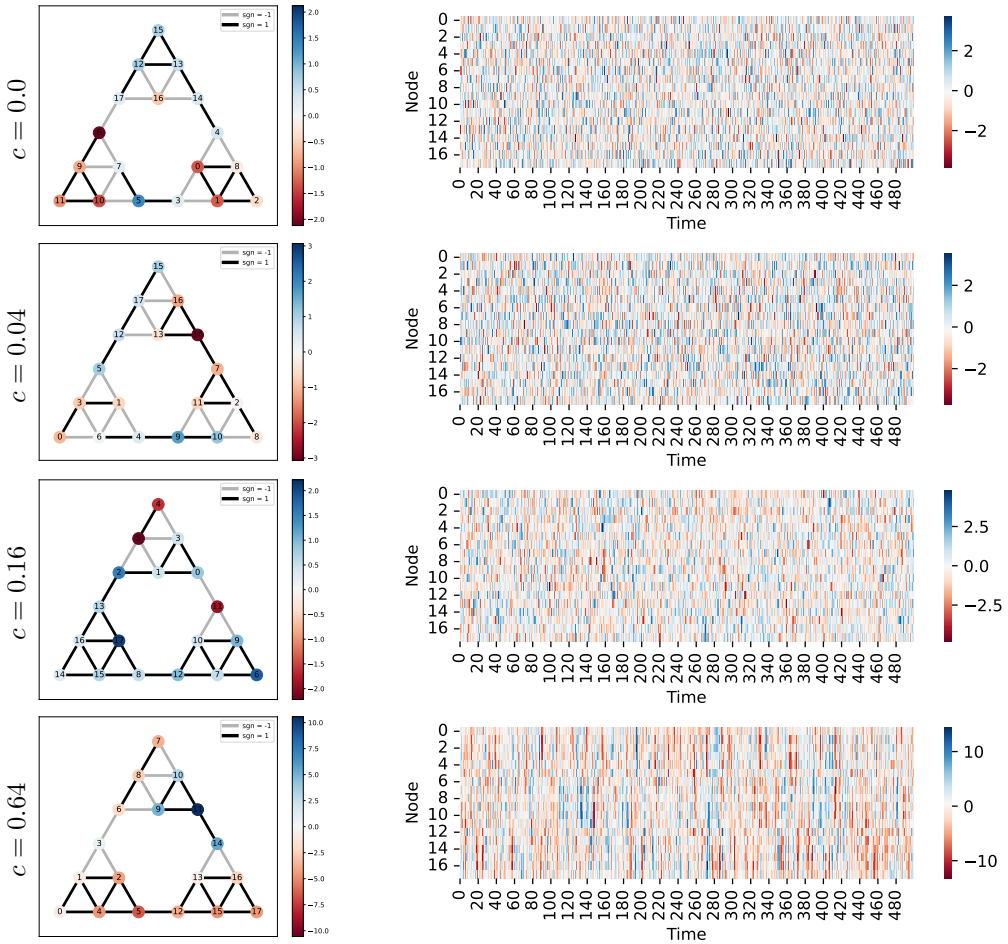

Figure 5: Graph signals displaying different levels of correlation. Each row corresponds to a different value of c, with $c_{\text{sp}} = c$ and $c_{\text{tm}} = c\,d$ in (9), and where d is the average node degree. The number of time steps is $T = 500$, the node feature dimension is $F = 1$, and the distribution of all $\mathbf{z}_v[t]$ is the 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.*, $\text{sgn}(\mathbf{x}_u[T/2]\mathbf{x}_v[T/2])$ for all edges (u, v).

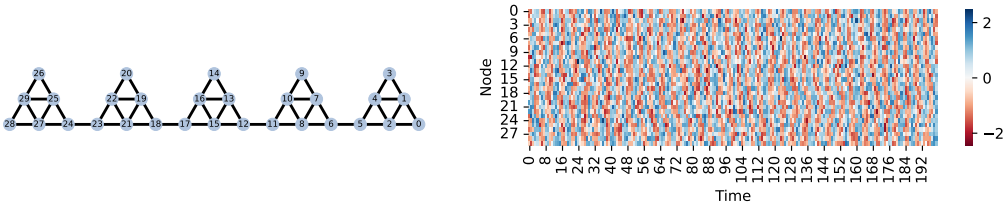

Figure 6: Graph and graph signal from the GPVAR dataset of Section 5.2. The displayed graph signal consists of the last $T = 200$ consecutive time steps of the entire dataset.

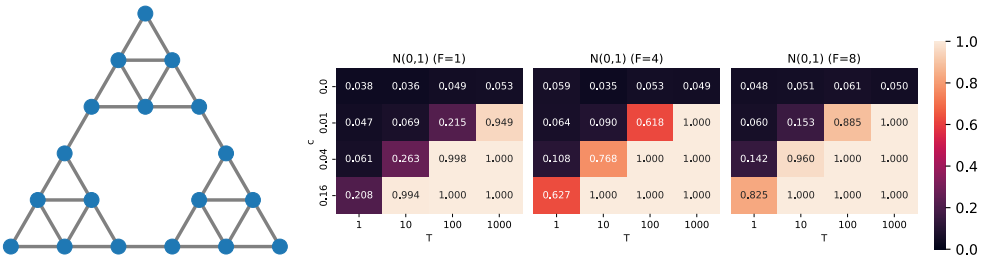

Figure 7: Rate of rejected null hypotheses for different node signal dimensions F, number of time steps T, and correlation parameters $c_{\text{sp}} = c$ and $c_{\text{tm}} = 0$. The considered graph is displayed on the left-hand side of the figure.

H Extended experiments on the optimality of forecasting models

We report an extended version of the experiments in Section 5.2 that includes the analysis of the prediction residuals shifted to force null median. In particular, for each model f_θ among FCRNN, GWNet, GatedGN, and DCRNN, we compute the associated residuals $\mathbf{R} \in \mathbb{R}^{T \times N}$ and reported the outcomes of the analysis as described in Section 5.2. In addition, we carry out the same analysis on shifted signal $\mathbf{R}' = \mathbf{R} - \text{median}(\mathbf{R})$, where $\text{median}(\mathbf{R})$ is the empirical median of residuals $\mathbf{R}$, so that $\text{median}(\mathbf{R}') = 0$. The results on the adjusted residuals $\mathbf{R}'$ are denoted in Table 3 with "-m" as suffix to the original model name, *e.g.*, FCRNN-m, GWaveNet-m, etc.

In all settings, we observe almost identical results before and after subtracting the median from the residuals.

Table 3: Analysis of the observed residuals on the considered datasets. The tests on the null-median report the estimated median with the associated p-value subscripted. The AZ-tests report the statistic $C(G^*)$ and the associated p-value subscripted. Results with p-value larger than 0.01 are highlighted in bold. Values reported as 0.000 are intended as < 0.001. Extended version of Table 2

Dataset	Model	MAE	Test Median=0	AZ-test Spatio-temporal	AZ-test Temporal	AZ-test Spatial
GPVAR	Optimal Pred.	0.319	**0.001** $_{\mathbf{0.083}}$	**-0.8** $_{\mathbf{0.416}}$	**-0.9** $_{\mathbf{0.355}}$	**-0.2** $_{\mathbf{0.823}}$
GPVAR	FCRNN	0.385	0.003 $_{0.010}$	5.0 $_{0.000}$	8.9 $_{0.000}$	**-1.8** $_{\mathbf{0.067}}$
GPVAR	FCRNN-m	0.385	**0.000** $_{\mathbf{1.000}}$	4.8 $_{0.000}$	8.7 $_{0.000}$	**-1.9** $_{\mathbf{0.057}}$
GPVAR	GWNET	0.324	0.004 $_{0.000}$	**0.3** $_{\mathbf{0.709}}$	**0.3** $_{\mathbf{0.706}}$	**0.1** $_{\mathbf{0.881}}$
GPVAR	GWNET-m	0.324	**0.000** $_{\mathbf{1.000}}$	**0.4** $_{\mathbf{0.630}}$	**0.3** $_{\mathbf{0.731}}$	**0.3** $_{\mathbf{0.736}}$
GPVAR	GATEDGN	0.321	0.008 $_{0.000}$	**1.3** $_{\mathbf{0.172}}$	2.7 $_{0.006}$	**-0.8** $_{\mathbf{0.414}}$
GPVAR	GATEDGN-m	0.321	**0.000** $_{\mathbf{1.000}}$	**1.1** $_{\mathbf{0.257}}$	**2.4** $_{\mathbf{0.015}}$	**-0.8** $_{\mathbf{0.411}}$
GPVAR	DCRNN	0.328	0.013 $_{0.000}$	**-0.0** $_{\mathbf{0.955}}$	**-0.6** $_{\mathbf{0.534}}$	**0.5** $_{\mathbf{0.587}}$
GPVAR	DCRNN-m	0.328	**0.000** $_{\mathbf{1.000}}$	**-0.2** $_{\mathbf{0.777}}$	**-0.7** $_{\mathbf{0.428}}$	**0.3** $_{\mathbf{0.696}}$
PemsBay	FCRNN	2.016	0.032 $_{0.000}$	1107.4 $_{0.000}$	1035.1 $_{0.000}$	531.0 $_{0.000}$
PemsBay	FCRNN-m	2.015	**0.000** $_{\mathbf{0.994}}$	1108.1 $_{0.000}$	1036.9 $_{0.000}$	530.2 $_{0.000}$
PemsBay	GWNET	0.841	-0.003 $_{0.000}$	422.7 $_{0.000}$	7.1 $_{0.000}$	590.7 $_{0.000}$
PemsBay	GWNET-m	0.841	**0.000** $_{\mathbf{0.992}}$	422.4 $_{0.000}$	6.9 $_{0.000}$	590.4 $_{0.000}$
PemsBay	GATEDGN	0.838	0.018 $_{0.000}$	454.6 $_{0.000}$	25.2 $_{0.000}$	617.7 $_{0.000}$
PemsBay	GATEDGN-m	0.838	**0.000** $_{\mathbf{0.995}}$	454.2 $_{0.000}$	24.7 $_{0.000}$	617.6 $_{0.000}$
PemsBay	DCRNN	0.845	-0.004 $_{0.000}$	433.0 $_{0.000}$	14.2 $_{0.000}$	598.1 $_{0.000}$
PemsBay	DCRNN-m	0.845	**0.000** $_{\mathbf{0.989}}$	432.3 $_{0.000}$	13.6 $_{0.000}$	597.8 $_{0.000}$
MetrLA	FCRNN	2.842	-0.016 $_{0.000}$	415.3 $_{0.000}$	238.5 $_{0.000}$	348.8 $_{0.000}$
MetrLA	FCRNN-m	2.842	**0.000** $_{\mathbf{0.999}}$	414.7 $_{0.000}$	238.2 $_{0.000}$	348.3 $_{0.000}$
MetrLA	GWNET	2.115	0.014 $_{0.000}$	162.6 $_{0.000}$	-6.5 $_{0.000}$	236.5 $_{0.000}$
MetrLA	GWNET-m	2.115	**0.000** $_{\mathbf{1.000}}$	162.7 $_{0.000}$	-6.2 $_{0.000}$	236.4 $_{0.000}$
MetrLA	GATEDGN	2.151	0.010 $_{0.000}$	200.1 $_{0.000}$	**2.2** $_{\mathbf{0.022}}$	280.7 $_{0.000}$
MetrLA	GATEDGN-m	2.151	**0.000** $_{\mathbf{0.996}}$	200.4 $_{0.000}$	**2.4** $_{\mathbf{0.016}}$	281.0 $_{0.000}$
MetrLA	DCRNN	2.141	-0.018 $_{0.000}$	177.4 $_{0.000}$	6.7 $_{0.000}$	244.1 $_{0.000}$
MetrLA	DCRNN-m	2.141	**0.000** $_{\mathbf{1.000}}$	176.5 $_{0.000}$	6.4 $_{0.000}$	243.2 $_{0.000}$