Learning in nonstationary environments

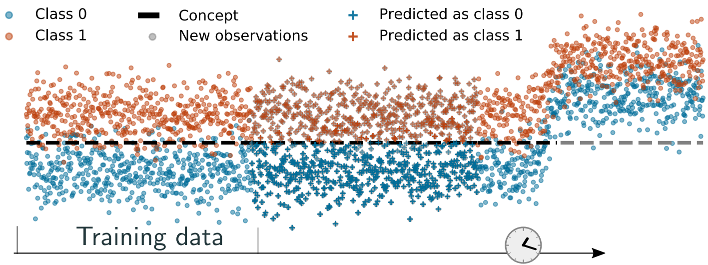
Anomaly and change detection in sequences of graphs

Daniele Zambon

* IDSIA, USI, Switzerland
  daniele.zambon@usi.ch

Graph Deep Learning, USI, Spring 2024

Change in the data distribution



- **Train** a model $f\,(x;\theta)$ on training set $\mathcal{T} = \{(x_i, y_i)\}$
- **Predict** on new observations $\hat{y}_t = f\left(x_t; \hat{\theta}\right) \approx y_t$

Performance as expected until the data distribution changes (**stationarity assumption**)

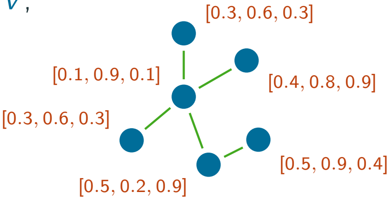
Attributed graphs

$g = (V, E, \alpha) \in \mathcal{G}_{\mathcal{A}}$

finite vertex set V;

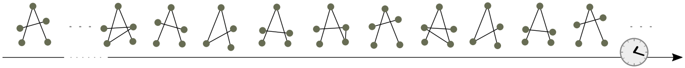
attribute function $\alpha : V \cup E \to \mathcal{A}$, with $\mathcal{A}$ attribute space.

edge set $E \subseteq V \times V$;



Detecting changes in stationarity

Consider a graph space $\mathcal{G}$ and a graph-generating stochastic process $\mathcal{P}$



Assumption: Process $\mathcal{P}$ is **stationary** and generates i.i.d. graphs $g_t \sim P_0$ for all $t \in \mathbb{N}$.
→ Nominal regime

Goal: **Identify** whether a change in stationarity occurred or not

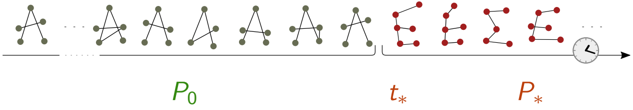
$$\exists\ t_* \in \mathbb{N} \text{ s.t. } g_t \sim \begin{cases} P_0, & t < t_* & \text{(Nominal regime)} \\ P_t, & t \geq t_* \text{ with } P_{t_*} \neq P_0. & \text{(Post-change distribution)} \end{cases}$$

→ all distributions are unknown!

Anomalies and abrupt changes in stationarity

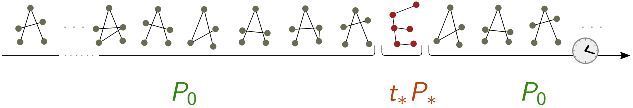
Abrupt change in stationarity:

$$g_t \sim \begin{cases} P_0, & t < t_* \\ P_* \neq P_0, & t \geq t_* \end{cases}$$



Anomalous observation:

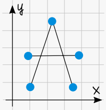
$$g_t \sim \begin{cases} P_0, & t \neq t_* \\ P_* \neq P_0, & t = t_* \end{cases}$$

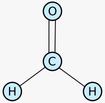


Considered types of graphs

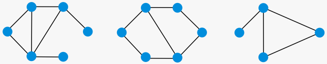
... compatible with
i.i.d. assumption.

Generic attributes

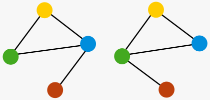




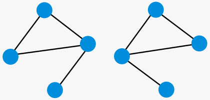
Variable topology and order



Identified vertices

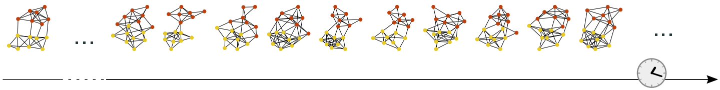


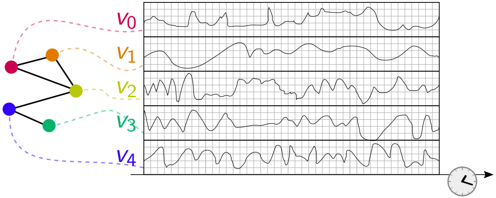
Non-identified vertices

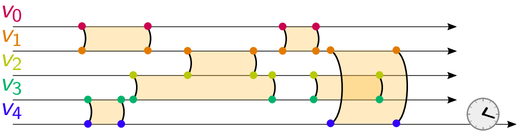


Graphs and time

we focus mainly here



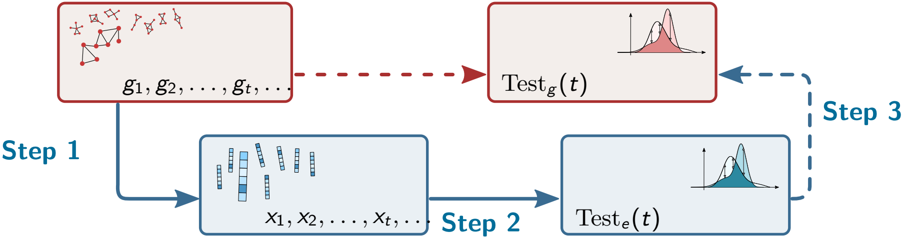




Graph-level embeddings

A 3-step methodology

1. **Embed** each graph $g_t \in \mathcal{G}$ to a point $\phi(g_t) = x_t \in \mathcal{X}$.
2. Carry out **statistical analyses** on the embedding sequence $x_1, x_2, \ldots$.
3. **Relate** drawn conclusions back to the graph domain.



Zambon et al. IEEE TNNLS 2018, Zambon et al. IEEE TSP 2019.

Graph distances, kernels and embeddings

$$\mathrm{d}(g_1, g_2)^2 = \kappa(g_1, g_1) + \kappa(g_2, g_2) - 2\kappa(g_1, g_2)$$

$$\kappa(g_1, g_2) = \frac{1}{2}\left[\mathrm{d}(g_1, g_{ref})^2 + \mathrm{d}(g_{ref}, g_2)^2 - \mathrm{d}(g_1, g_2)^2\right]$$

$$\begin{tikzcd}
\textbf{Distances} \arrow[rr, leftrightarrow] \arrow[dr, leftrightarrow] & & \textbf{Kernels} \arrow[dl, leftrightarrow] \\
& \textbf{Embeddings} &
\end{tikzcd}$$

$$\mathrm{d}(g_1, g_2) = \left|\phi(g_1) - \phi(g_2)\right|_{\mathcal{H}}$$

$$\kappa(g_1, g_2) = \left\langle \phi(g_1),\, \phi(g_2) \right\rangle_{\mathcal{H}}$$

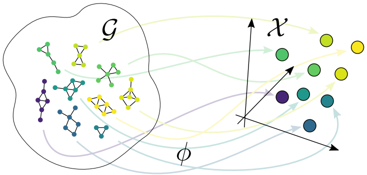
we focus mainly here

See e.g. Zambon 2022, Sejdinovic et al. 2013, Schölkopf 2001, Steinwart and Christmann 2008.

Step 1: Graph-level embeddings

Graph-level embedding: a function $\phi : \mathcal{G} \to \mathcal{X}$.

- amenable mathematics;
- approximate distances and kernels;
- define new distances and kernels;
- well-established statistical tools.



Examples:

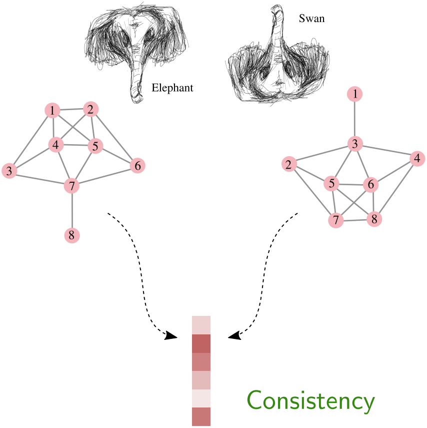
1. Adaptive filtering with kernel $\kappa(g_i, g_j) = \langle \phi(g_i), \phi(g_j) \rangle_{\mathcal{X}}$

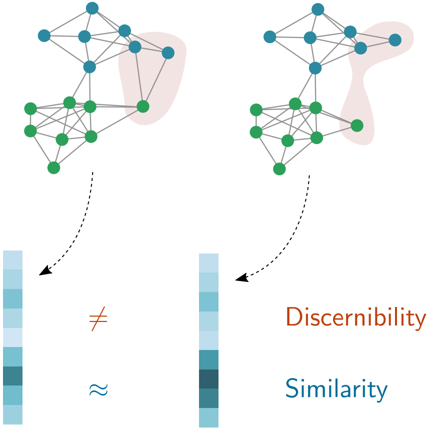
   implicit: several kernel evaluations with graphs in memory,
   $f(g_t) = \sum_{i<t} \alpha_i \kappa(g_i, g_t)$,
   $\vec{\alpha} \leftarrow \mathrm{Update}(\ldots, g_i, \alpha_i, y_i, \ldots, g_t, y_t)$

   explicit: past graphs used only once,
   $\phi_t = \phi(g_t)$, $\quad f(\phi_t) = \langle \vec{w}, \phi_t \rangle_{\mathcal{X}}$, $\quad \vec{w} \leftarrow \mathrm{Update}(\vec{w}, \phi_t, y_t)$

2. Sequential hypothesis testing

Desirable properties

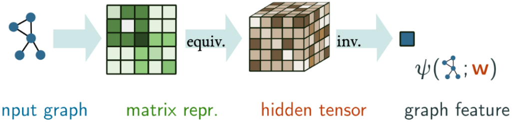




Graph-level embeddings

Graph Random Neural Features

GRNF: Graph random neural features



(E.g., Maron et al. 2019, Keriven and Peyré 2019)

Consider a parametric family

$$\mathcal{F}(\mathcal{W}) = \{\psi(\,\cdot\,;\vec{w}) : \mathcal{G} \to \mathbb{R} \mid \vec{w} \in \mathcal{W}\},$$

of **graph functions** $\psi(\,\cdot\,;\vec{w})$ such that

- ψ is continuous in $\vec{w}$.
- $g_i \neq g_j \implies \exists\ \vec{w}_* \ :\ \psi(g_i;\vec{w}_*) \neq \psi(g_j;\vec{w}_*)$.

Theorem [Zambon et al. ICML 2020]

$$\mathrm{d}_P(g_i, g_j) = \left(\mathbb{E}_{\vec{w}\sim P}\left[\left(\psi(g_i;\vec{w}) - \psi(g_j;\vec{w}) \right)^2 \right] \right)^{\frac{1}{2}}.$$

If distribution P has support $\mathrm{Supp}(P) = \mathcal{W}$, then $\mathrm{d}_P(g_i, g_j)$ is **metric**.

Zambon et al. ICML 2020

GRNF: The embedding

Graph random neural features (GRNF):

Construct the embedding from a set of M

- graph functions $\psi(\,\cdot\,;\vec{w}_1)\dots\psi(\,\cdot\,;\vec{w}_M)$
- weights $\alpha_1,\dots,\alpha_M$.

$$\phi(g) = \begin{bmatrix} \alpha_1 \cdot \psi(g;\vec{w}_1) \\ \vdots \\ \alpha_M \cdot \psi(g;\vec{w}_M) \end{bmatrix}$$

Training-free:

- Select P such that $\mathrm{Supp}(P) = \mathcal{W}$
- Sample $\vec{w}_1,\dots,\vec{w}_M \sim P$
- Set $\alpha_1,\dots,\alpha_M = M^{-1/2}$

Learnable:

- Tune weights $\alpha_1,\dots,\alpha_M$
  - $\rightarrow$ relatively cheap to compute
  - $\rightarrow$ equivalent to training P
- Train all parameters $\vec{w}_1,\dots,\vec{w}_M$

Zambon et al. ICML 2020

GRNF: Trading off quality with computation

Distance-preserving embedding: If $\vec{w}_1, \ldots, \vec{w}_M \sim P$ and $\alpha_1, \ldots, \alpha_M = M^{-1/2}$, then

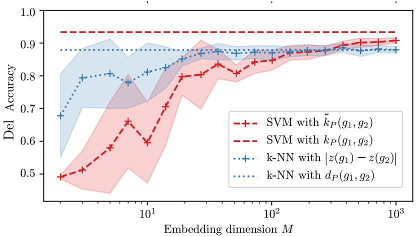
$$\mathbb{E}_{\vec{w}_1 \ldots \vec{w}_M}\left[|\phi(g_i) - \phi(g_j)|_2^2\right] = \mathrm{d}_P(g_i, g_j)^2, \qquad |\phi(g_i) - \phi(g_j)|_2 \xrightarrow[\mathrm{M} \to \infty]{\text{prob.}} \mathrm{d}_P(g_i, g_j).$$

Trade-off: For any value of $\varepsilon > 0$ and $\delta \in (0,1)$, if $M \geq \frac{16}{\delta \varepsilon^2}$, then

$$\left| \; \mathrm{d}_P(g_i, g_j)^2 \; - \; |\phi(g_i) - \phi(g_j)|_2^2 \; \right| \; \geq \; \varepsilon, \qquad \text{with prob. } < \delta.$$

→ allows to select M to satisfy (ε, δ) constraints;

→ M trades off approximation accuracy with memory and computation.



Zambon et al. ICML 2020

Change detection

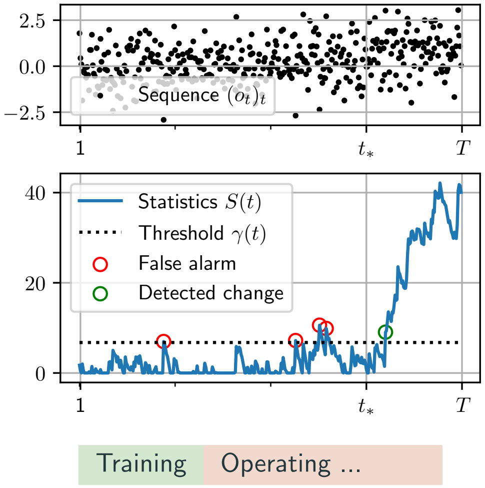
Sequential change detection test

Consider an i.i.d. sequence $x_1, x_2, \ldots, x_t, \ldots$

Cumulative sum

- compute local statistic $s(t) = |x_t - \mu|$
- accumulate information
  $S(t) = \max\{0, S(t-1) + s(t) - q\}$
- change detected if $S(t) > \gamma(t)$

$$\gamma(t) : \ \mathbb{P}(S(t) > \gamma(t) \mid H_0) \leq \alpha \ll 1$$



See also Page 1954.

Offline change detection test

Consider an i.i.d. sequence $[x_1, x_2, \dots, x_T]$

Change point method

- for every t, split sequence:

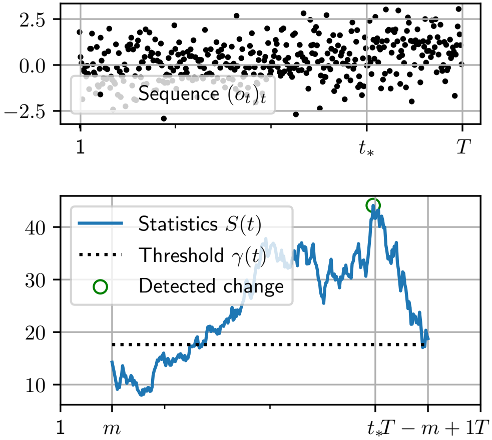
$$x_1, \dots, x_{t-1} \quad \Big| \quad x_t, \dots, x_T$$

- two-sample tests

$$S(t) = Test(\{x_1, \dots, x_{t-1}\}, \{x_t, \dots, x_T\})$$

- change detected if $\max\{S(t)\} > \gamma$

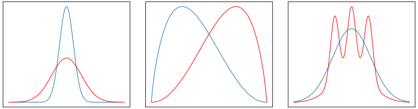
$$\gamma : \; \mathbb{P}(\max S(t) > \gamma \mid H_0) \leq \alpha \ll 1$$



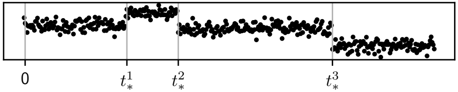
See e.g. Hawkins et al. 2003.

Change detection on graph sequences

- ✓ **Sequential** and **offline** tests,
- ✓ tests for changes in the **Fréchet mean** graph
- ✓ for **arbitrary** distribution changes,



- ✓ **multiple** changes.



Fréchet sample mean:

Given n graphs $\{g_i\}$,

$$\arg\min_{g\in\mathcal{G}} \sum_{i=1}^{n} \mathrm{d}(g, g_i)^2.$$

Fréchet population mean:

Given distribution P

$$\arg\min_{g\in\mathcal{G}} \mathbb{E}_{g'\sim P}\left[\mathrm{d}(g', g)^2\right]$$

Zambon et al. IEEE TNNLS 2018, Zambon et al. IEEE IJCNN 2018, Grattarola, Zambon et al. IEEE TNNLS 2019, Zambon et al. IEEE TSP 2019, Zambon, PhD thesis, 2021.

Change detection

Test on constant-curvature manifolds

Example: Test on manifold data

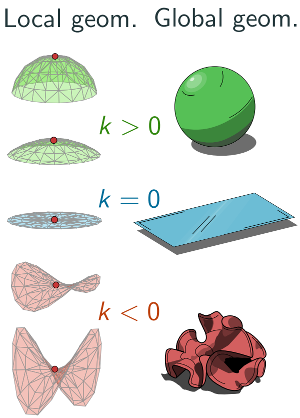
Test for change in the mean

$$\mathbb{E}^{\mathrm{f}}[P_0] := \arg\min_{g\in\mathcal{G}} \mathbb{E}_{g'\sim P_0}\left[\mathrm{d}(g,g')^2\right].$$

- Embedding $\phi : \mathcal{G} \to \mathcal{M}_\kappa$ to Riemannian manifold of constant curvature κ;
- From $g_1, g_2, \dots \in \mathcal{G}$

  $\to$ we get $x_1, x_2, \dots \in \mathcal{M}_\kappa$;
- Implement a test on the mean in the manifold

$$\mathbb{E}^{\mathrm{f}}[P_0^\sharp] := \arg\min_{x\in\mathcal{M}_\kappa} \mathbb{E}_{x'\sim P_0^\sharp}\left[\rho_\kappa(x,x')^2\right].$$



Anomaly detection

Anomaly and outlier detection

Consider some graph data $\{g_1, g_2, \dots, g_N\} \subseteq \mathcal{G}$,

Assume: the majority of the graphs is such that

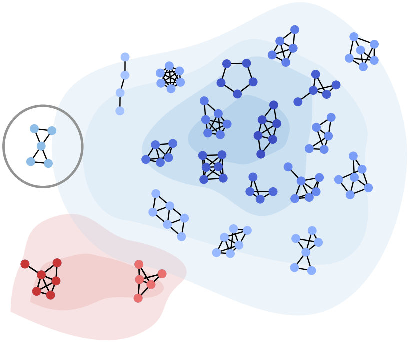
$$g_i \sim p_0.$$

Goal: Find

- anomalies ($g_* \sim p_* \neq p_0$)
- outliers (like, g_*, with $p_0(g_*) \approx 0$)

Define: anomaly/outlier score $s_\theta(g) : \mathcal{G} \to \mathbb{R}_+$

$$\text{If} \begin{cases} s_\theta(g) \leq \gamma & \text{consider } g \text{ nominal} \\ s_\theta(g) > \gamma & \text{consider } g \text{ anomalous/outlier} \end{cases}$$

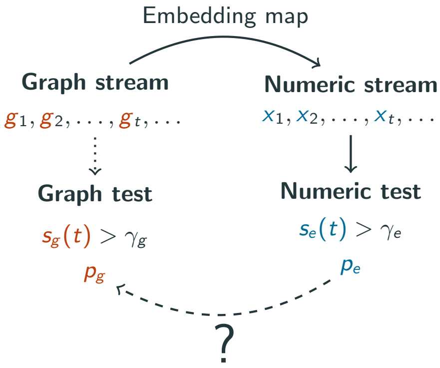


⚠ anomalies/outliers are rare

Zambon and Alippi, IEEE IJCNN 2022.

Back to the graph domain

Step 3: Linking the inference in graph and embedding domains



Questions:

1. Does the change observed in $\mathcal{X}$ correspond to a change in $\mathcal{G}$? ✓

2. Are all changes occurring in $\mathcal{G}$ detectable in $\mathcal{X}$?

Step 3: Linking the inference in graph and embedding domains …cont'd

Deterministic bound [Zambon et al. IEEE TNNLS 2018]

If Test on Fréchet mean and $\exists\, c$:

$$\frac{1}{c} d(g_i, g_j) \leq |x_i - x_j|_2 \leq c\, d(g_i, g_j)$$

Then $\exists q, b > 0$ such that, for any γ,

$$q \cdot \mathrm{cdf}_e \left(\frac{\gamma}{c} - b\right) \leq \mathrm{cdf}_g(\gamma) \leq \frac{1}{q} \cdot \mathrm{cdf}_e(\gamma \cdot c + b)$$

Probabilistic bound [Zambon et al. IEEE TSP 2019]

If $\exists \lambda, q > 0$ such that

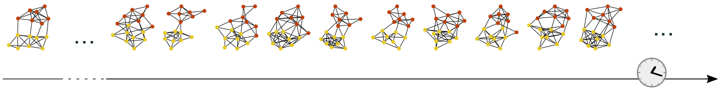
$$\Pr(|s_e(t) - s_g(t)| \leq \lambda) \geq q$$

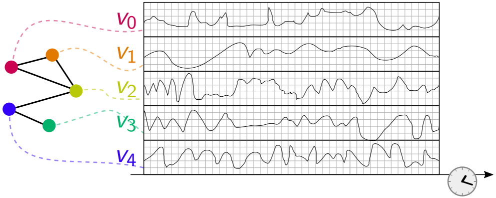
Then, for any γ,

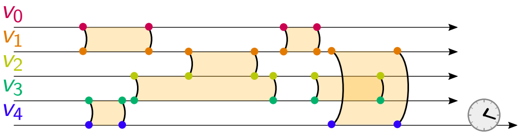
$$q \cdot \mathrm{cdf}_e\,(\gamma - \lambda) \leq \mathrm{cdf}_g(\gamma) \leq \frac{1}{q} \cdot \mathrm{cdf}_e(\gamma + \lambda).$$

Assessing model optimality

Graphs and time





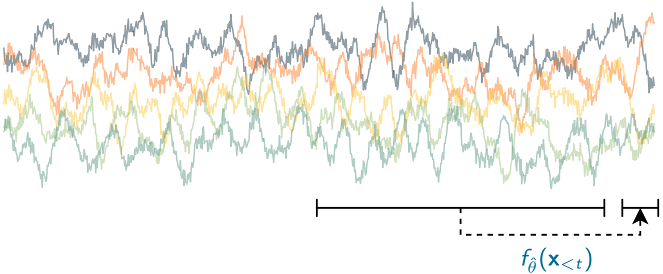


Let us move here now!

Residual analysis and model optimality

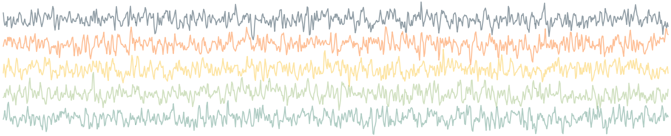
Consider:

- a multivariate time series
- a forecasting problem
- a trained model $f_{\hat{\theta}}(\cdot)$



Inspect the prediction residuals

$$\mathbf{r}_t \doteq \mathbf{x}_t - f_{\hat{\theta}}(\mathbf{x}_{<t})$$



If residuals $\mathbf{r}_t$ are correlated (not white noise) $\Longrightarrow$ there is still information to be learned

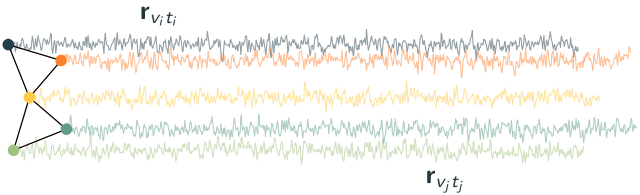
$\Longrightarrow$ $f_{\hat{\theta}}(\cdot)$ not optimal ⚠

Assessing model optimality

AZ-whiteness test

AZ-whiteness test: what is it?

We consider **time series** associated with the nodes of a **graph**



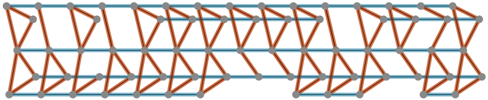
We propose a **test** for both temporal and spatial correlation.

- ✓ Graphs can be (un)directed, weighted, and dynamic;
- ✓ Asymptotically distribution-free;
- ✓ Applicable for evaluating optimality of forecasting **models**.

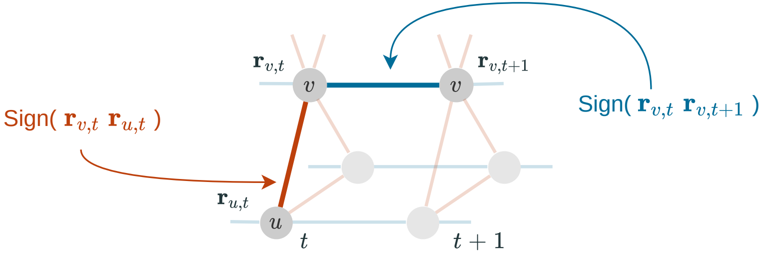
Zambon and Alippi, NeurIPS 2022.

Test statistic based on signs

Construct a **multiplex** graph,
and compute **sign changes** for all edges.



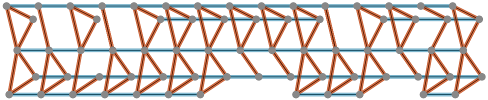
spatial and temporal edges.



$$\mathrm{Sign}(a) = \begin{cases} +1 & a > 0 \\ 0 & a = 0 \\ -1 & a < 0 \end{cases}$$

Zambon and Alippi, NeurIPS 2022.
Simplified version for undirected graphs, and scalar $\mathbf{r}_{v,t}$.

Test statistic



spatial and temporal edges.

The **test statistic** is

$$C(\lambda) \doteq \frac{\lambda \tilde{C}_{\text{sp}} + (1-\lambda)\tilde{C}_{\text{tm}}}{(\lambda^2 W_{\text{sp}} + (1-\lambda)^2 W_{\text{tm}})^{\frac{1}{2}}}$$

- $\tilde{C}_{\text{sp}} \doteq \sum w_{uv}\, \text{Sign}(\mathbf{r}_{u,t}\, \mathbf{r}_{v,t})$, $\quad \tilde{C}_{\text{tm}} \doteq \sum w_{vt}\, \text{Sign}(\mathbf{r}_{v,t}\, \mathbf{r}_{v,t+1})$
- λ trading off spatial and temporal contributions, and
- W_{sp}, W_{tm} normalization terms

Zambon and Alippi, NeurIPS 2022.

Assess model optimality

		$\lambda = 1/2$	$\lambda = 0$	$\lambda = 1$
Model	MAE	AZ-test spatiotemporal	AZ-test temporal	AZ-test spatial
Optimal Predictor	0.319			
Model A	0.385	4.8 $_{0.000}$	8.7 $_{0.000}$	**-1.9** $_{\mathbf{0.057}}$
Model B	0.328	**-0.2** $_{\mathbf{0.777}}$	**-0.7** $_{\mathbf{0.428}}$	**0.3** $_{\mathbf{0.696}}$

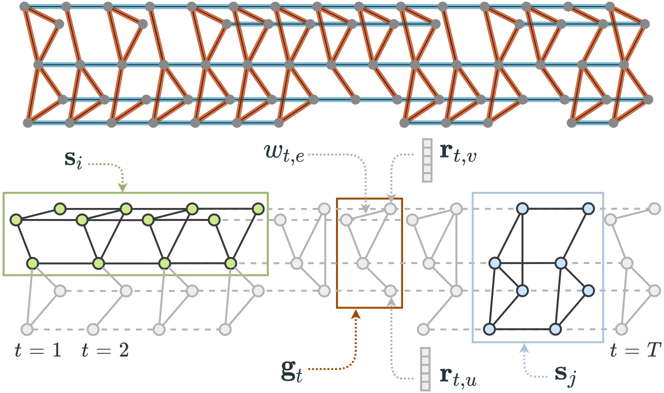
Test statistic and associated p-value: Stat $_{p\text{-val}}$.

Test on synthetic forecasting problem.

Assessing model optimality

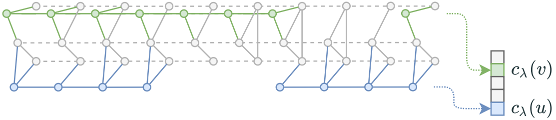
Where and how improving our models

Testing subgraphs

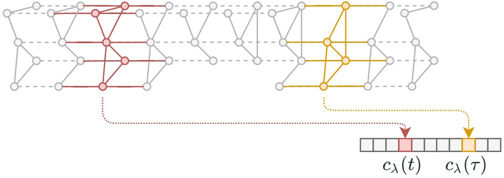


Families of subgraphs

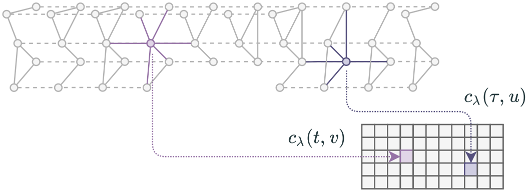
Node-level analysis



Temporal analysis



Spatio-temporal analysis

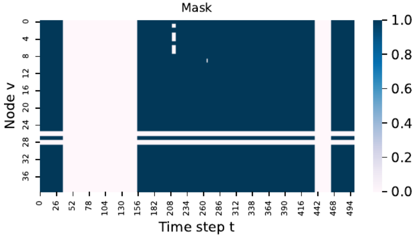


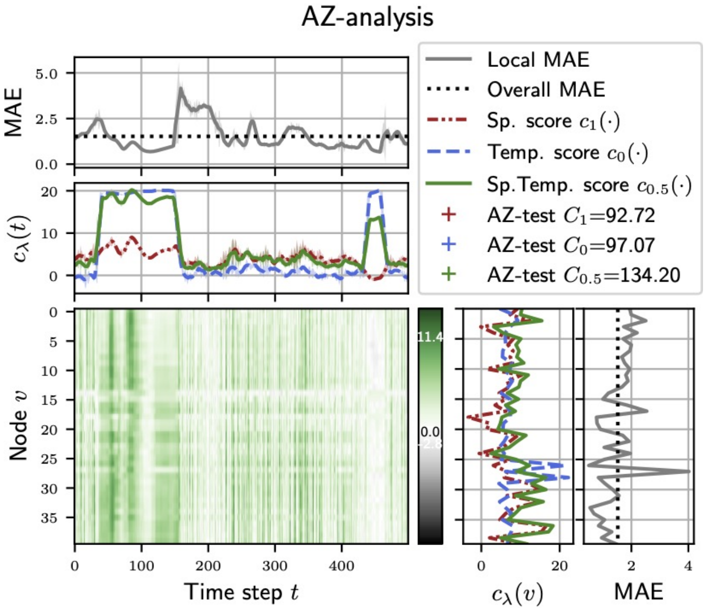
Zambon and Alippi 2023

Analysis of prediction residuals.

Traffic forecasting in MetrLA

Zambon and Alippi 2023





D. Grattarola, D. Zambon, C. Alippi, and L. Livi. Change detection in graph streams by learning graph embeddings on constant-curvature manifolds. **IEEE Transactions on Neural Networks and Learning Systems**, 2019. doi: 10.1109/TNNLS.2019.2927301.

S. Hariri, M. C. Kind, and R. J. Brunner. Extended isolation forest. **IEEE Transactions on Knowledge and Data Engineering**, 2019.

D. M. Hawkins, P. Qiu, and C. W. Kang. The Changepoint Model for Statistical Process Control. **Journal of Quality Technology**, 35(4):355–366, Oct. 2003. ISSN 0022-4065. doi: 10.1080/00224065.2003.11980233.

N. Keriven and G. Peyré. Universal invariant and equivariant graph neural networks. In **Advances in Neural Information Processing Systems**, pages 7090–7099, 2019.

H. Maron, E. Fetaya, N. Segol, and Y. Lipman. On the universality of invariant networks. In **International Conference on Machine Learning**, pages 4363–4371, 2019.

C. Morris, N. M. Kriege, F. Bause, K. Kersting, P. Mutzel, and M. Neumann. TUDataset: A collection of benchmark datasets for learning with graphs. In **ICML 2020 Workshop on Graph Representation Learning and beyond (GRL+ 2020)**, 2020.

E. S. Page. Continuous inspection schemes. **Biometrika**, 41(1/2):100–115, 1954.

B. Schölkopf. The kernel trick for distances. In **Advances in Neural Information Processing Systems**, pages 301–307, 2001.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. **The Annals of Statistics**, pages 2263–2291, 2013.

I. Steinwart and A. Christmann. **Support Vector Machines**. Springer Science & Business Media, 2008.

S. C. Tan, K. M. Ting, and T. F. Liu. Fast anomaly detection for streaming data. In **Twenty-Second International Joint Conference on Artificial Intelligence**, 2011.

D. Zambon. **Anomaly and Change Detection in Sequences of Graphs**. PhD thesis, Università della Svizzera italiana / Università della Svizzera italiana, 2022.

D. Zambon and C. Alippi. AZ-whiteness test: A test for signal uncorrelation on spatio-temporal graphs. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, **Advances in Neural Information Processing Systems**, 2022.

D. Zambon and C. Alippi. Where and How to Improve Graph-based Spatio-temporal Predictors, Feb. 2023.

D. Zambon, C. Alippi, and L. Livi. Concept drift and anomaly detection in graph streams. **IEEE Transactions on Neural Networks and Learning Systems**, pages 1–14, 2018a. doi: 10.1109/TNNLS.2018.2804443.

D. Zambon, L. Livi, and C. Alippi. Anomaly and change detection in graph streams through constant-curvature manifold embeddings. In **IEEE International Joint Conference on Neural Networks**, 2018b. doi: 10.1109/IJCNN.2018.8489762.

D. Zambon, C. Alippi, and L. Livi. Change-point methods on a sequence of graphs. **IEEE Transactions on Signal Processing**, 2019. doi: 10.1109/TSP.2019.2953596.

D. Zambon, C. Alippi, and L. Livi. Graph random neural features for distance-preserving graph representations. In H. D. III and A. Singh, editors, **Proceedings of the 37th International Conference on Machine Learning (ICML)**, volume 119 of **Proceedings of Machine Learning Research**, pages 10968–10977, Virtual, July 2020. PMLR.

D. Zambon, L. Livi, and C. Alippi. Graph iForest: Isolation of anomalous and outlier graphs. In **2022 International Joint Conference on Neural Networks (IJCNN)**, pages 1–8. IEEE, 2022.