Persistence-based Motifs Discovery in Time Series

T. Germain[1,2],

C. Truong[1,2] and L. Oudre[1,2]

[1]Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, F-91190, Gif-sur-Yvette, France
[2]Université de Paris, CNRS, Centre Borelli, F-75005 Paris, France

MLMDA, July 2024

Motif discovery in time series

Definition

Motif Discovery consists of finding repeated patterns and locating their occurrences in a time series without prior knowledge about their shape or location.

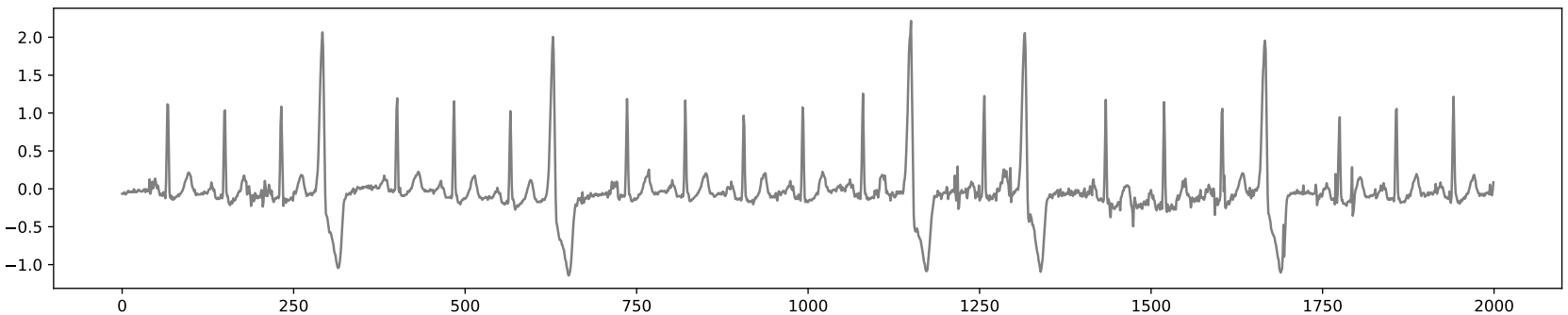
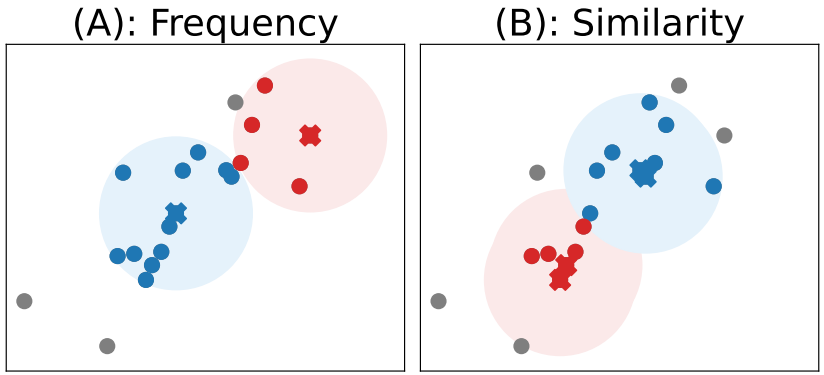


Figure: ECG with premature ventricular contraction (PVC)

Constraints:

- A single long univariate time series
- May contain motifs of different length

State-of-the-art

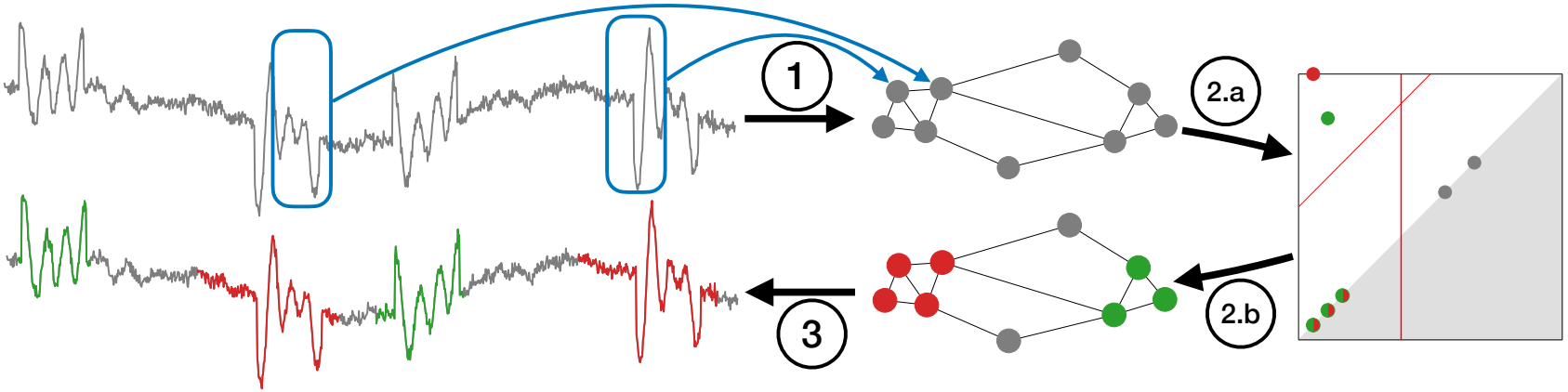


Two main families of algorithms:

- **Frequency-based:** algorithms identify sets of subsequences that represent the most frequently repeated patterns.
- **Similarity-based:** algorithms identify sets of subsequences that represent repeated patterns with minimal variability between occurrences, regardless of frequency.

Observation: Most algorithms rely on three core parameters: the number of motifs, the motif length, and a similarity threshold.
➔ In practice, the setting often results from a trial-and-error strategy.

PEPA algorithm overview



Algorithm steps:

1. From time-series to graph
2. Graph clustering with persistent homology
   a. From graph to persistent diagram
   b. From persistent diagram to clusters
3. From cluster to motifs sets

From time series to graph

Aim: Map a time series $S = (s_1, \ldots, s_n) \in \mathbb{R}^n$ into an undirected weighted graph $\mathcal{G}_S^K = (V, E)$.

Graph specification

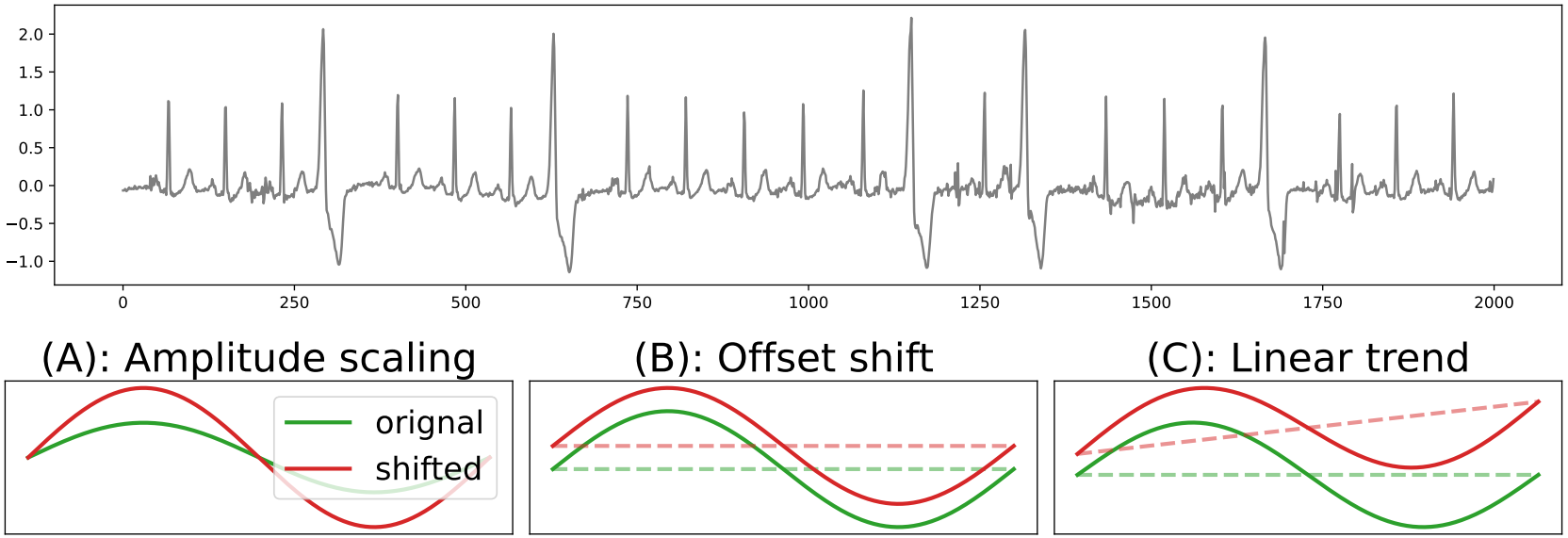
- **Vertices**: $V = (S_i^w)_{i=1\ldots n-w+1}$, subsequences of length l of S
- **Edges**: $E = E_1 \cup E_2$, union of two edges sets:
  - **Similarity set:** each subsequence S_i^w is connected to its K most similar non-overlapping subsequences.

$$E_1 = \bigcup_{i=1}^{n-w+1} \left\{ \left(S_i^w, N_i^k, d_i^k \right) \mid k = 1, \ldots, K \right\}, \tag{1}$$

  - **Time set:** edges connecting time adjacent subsequences.

$$E_2 = \bigcup_{i=1}^{n-w} \{ (S_i^w, S_{i+1}^w, \max \left(d_i^1, d_{i+1}^1 \right) \}. \tag{2}$$

Distance between subsequences



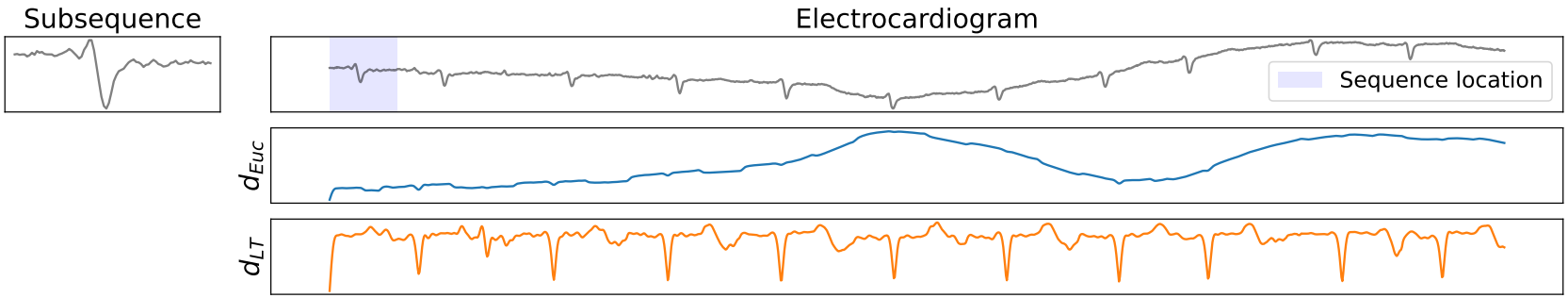
Definition (LT-normalized Euclidean distance)

The LT-normalized (Euclidean) distance between $x \in \mathbb{R}^w$ and $y \in \mathbb{R}^w$ is:

$$d_{LT}(x, y) = \left\| \frac{x - (\alpha_x \mathbf{t} + \beta_x \mathbf{1})}{\|x - (\alpha_x \mathbf{t} + \beta_x \mathbf{1})\|} - \frac{y - (\alpha_y \mathbf{t} + \beta_y \mathbf{1})}{\|y - (\alpha_y \mathbf{t} + \beta_y \mathbf{1})\|} \right\|$$

where $\mathbf{t} = (0, \ldots, w-1)$, $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^w$ and (α_x, β_x) are solutions of the linear regression problem $\underset{(a,b)\in\mathbb{R}^2}{\operatorname{argmin}} \|x - (a\mathbf{t} + b\mathbf{1})\|^2$.

Distance between subsequences



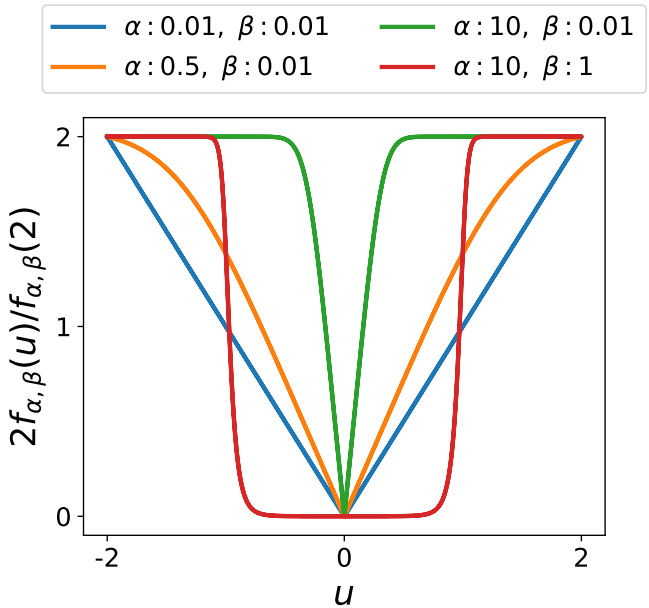
(α, β)-rectified distance

Let $\alpha \in \mathbb{R}_+^*$ and $\beta \in]0, 2[$, the (α, β)-rectified distance between $x \in \mathbb{R}^w$ and $y \in \mathbb{R}^w$ is:

$$d_{\alpha,\beta}(x, y) = 2f_{\alpha,\beta}(d_{LT}(x, y))/f_{\alpha,\beta}(2)$$

with

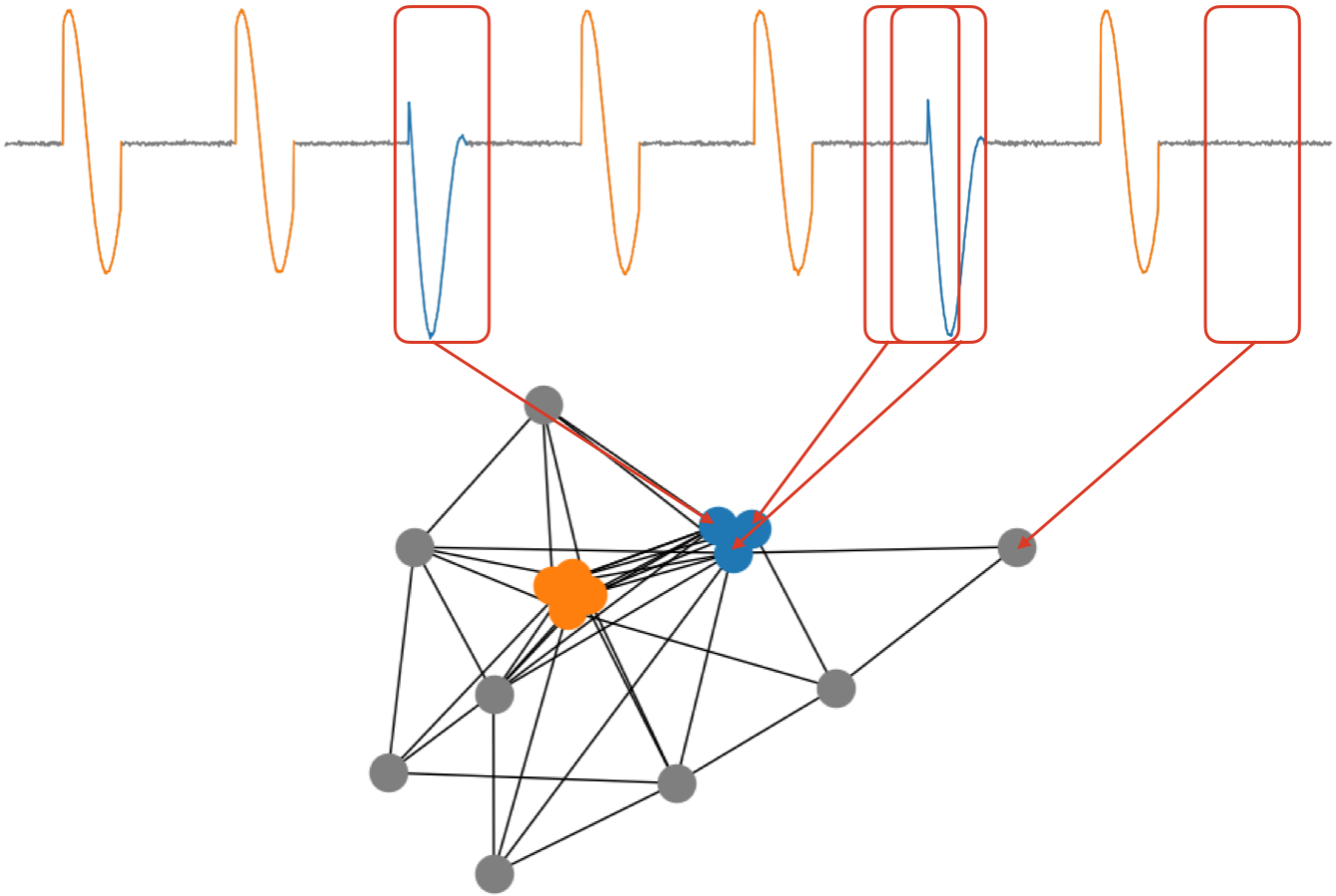
$$f_{\alpha,\beta}(x) = \sqrt{\tanh(\alpha\beta^2) + \tanh(\alpha(x^2 - \beta^2))}$$



Graph computation complexity

For any time series $S \in \mathbb{R}^n$, the graph $\mathcal{G}_S^K$ is computed in $\mathcal{O}(Kn^2)$.

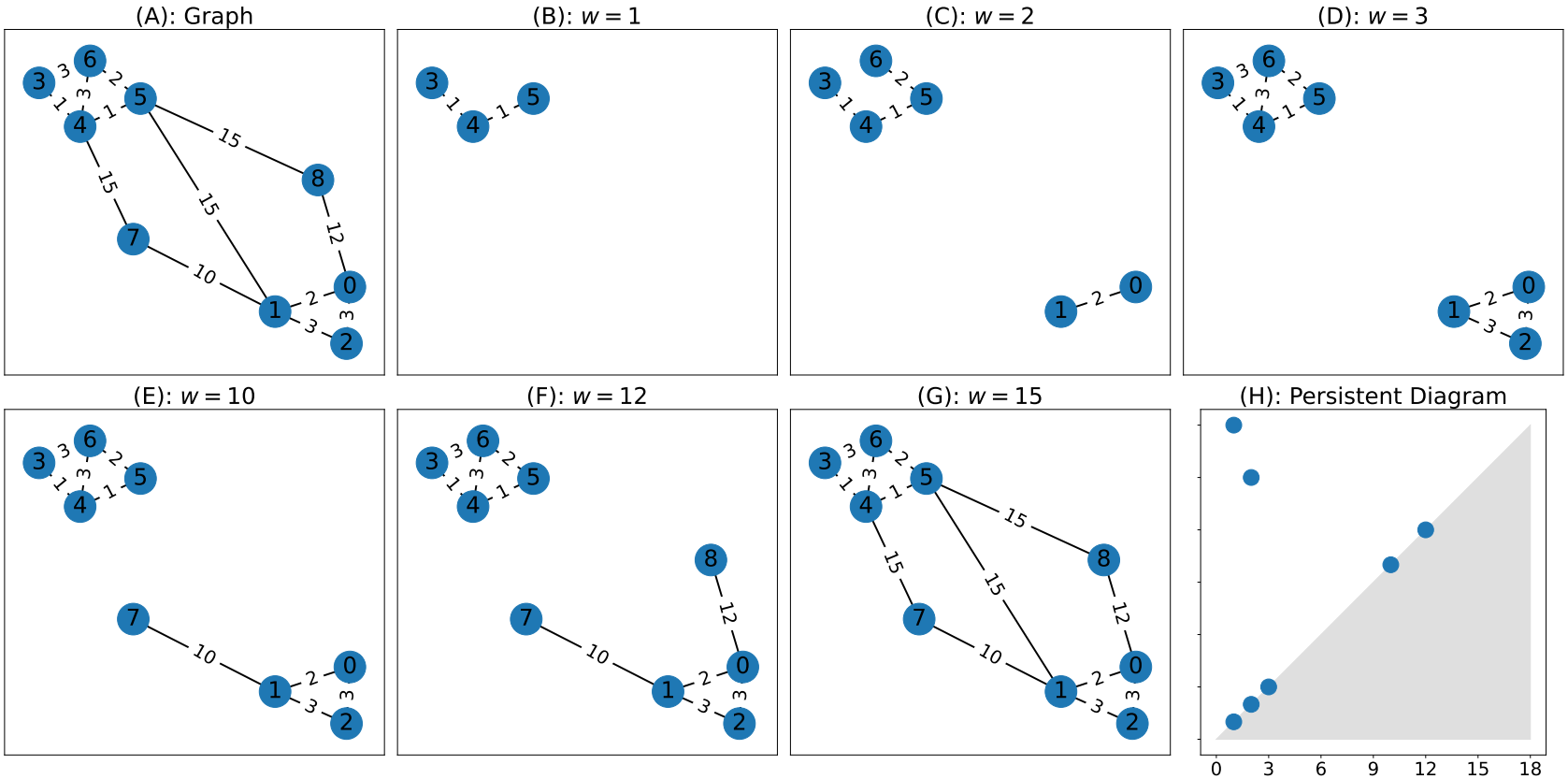
Graph clustering through persistent homology



Main Idea

Subsequences that overlap the same motif are close to each other and far from any other subsequences.

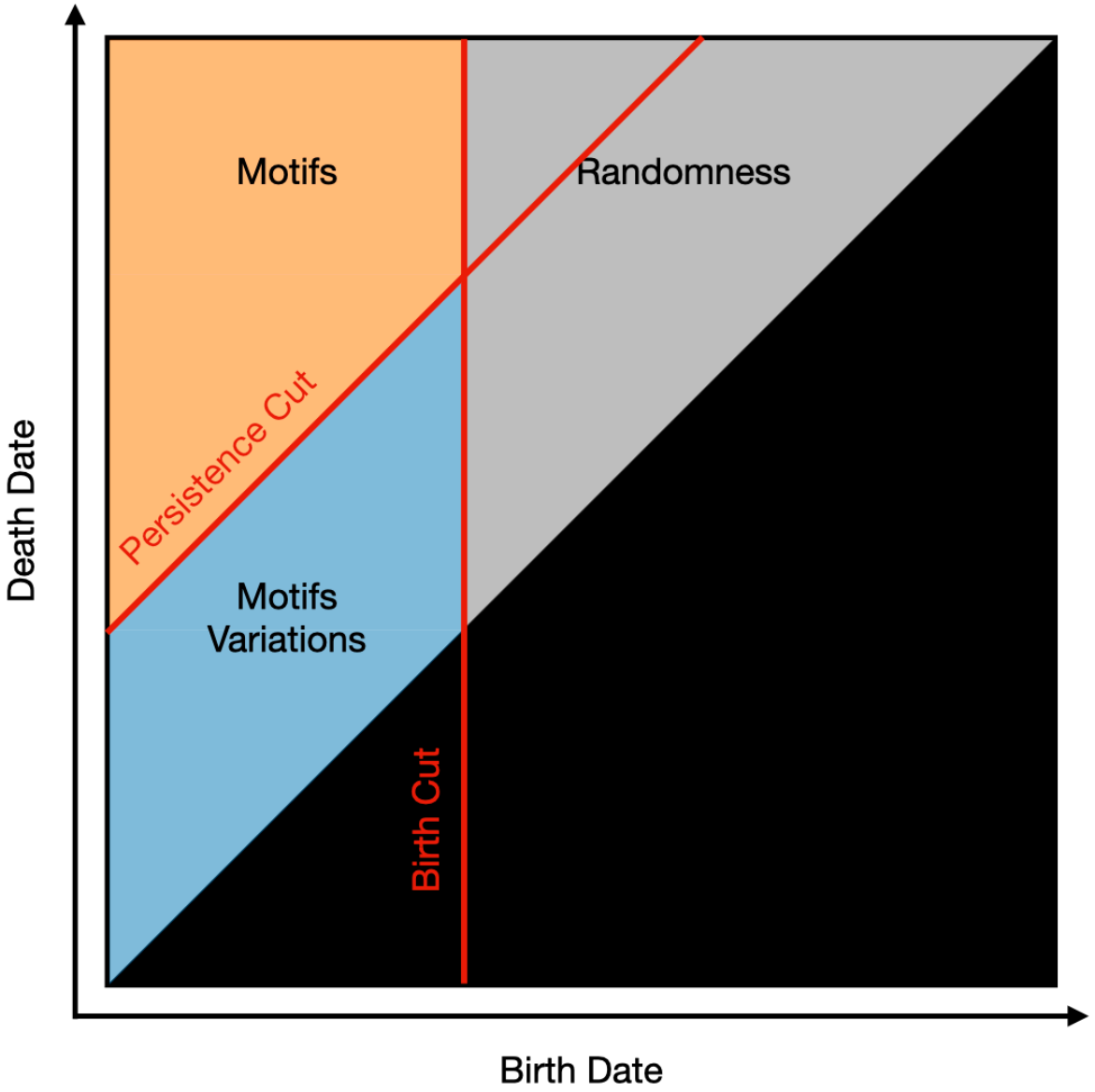
Graph clustering through persistent homology



Connected subgraph tracking rule:

- Birth Date: distance at which its first edge appears.
- Death Date: distance at which the subgraph gets connected an older subgraph.

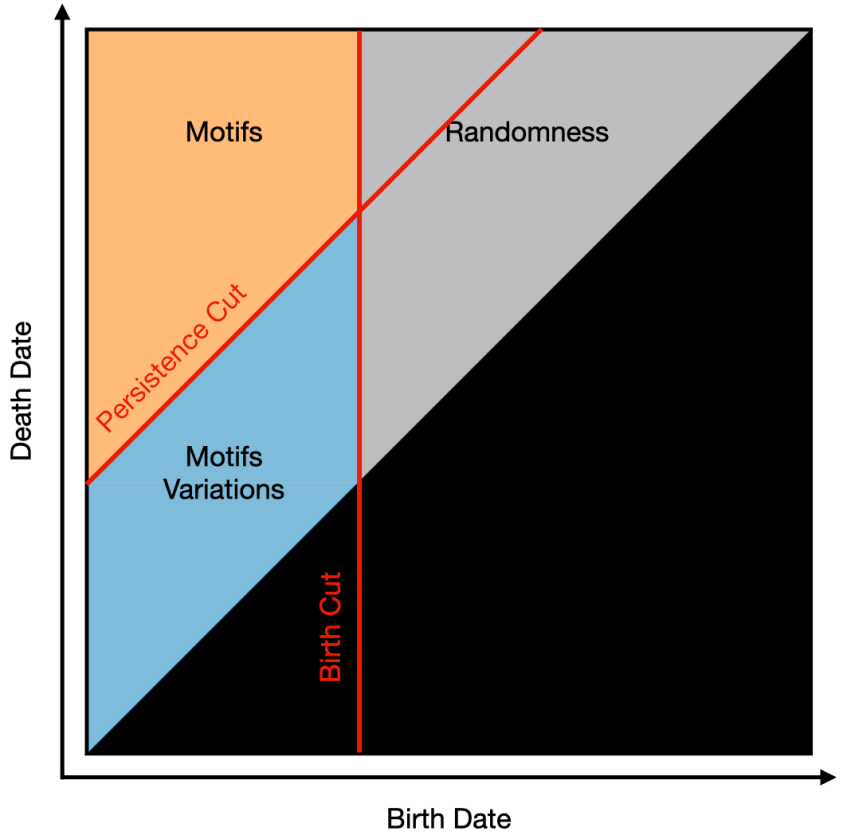
Graph clustering through persistent homology



Graph clustering through persistent homology

Some properties:

- All nodes (subsequences) have birth and death dates.
- Births and deaths are tracked with an algorithm similar to the Kruskall's algorithm for computing the minimum spanning tree (MST).
- The filtration of $\mathcal{G}_S^K$ can be reduced to the filtration of its MST.



From persistence diagram to motif sets

Input: the graph of a time series, $\mathcal{G}_S^K$.
Parameters: persistent cut, birth cut.

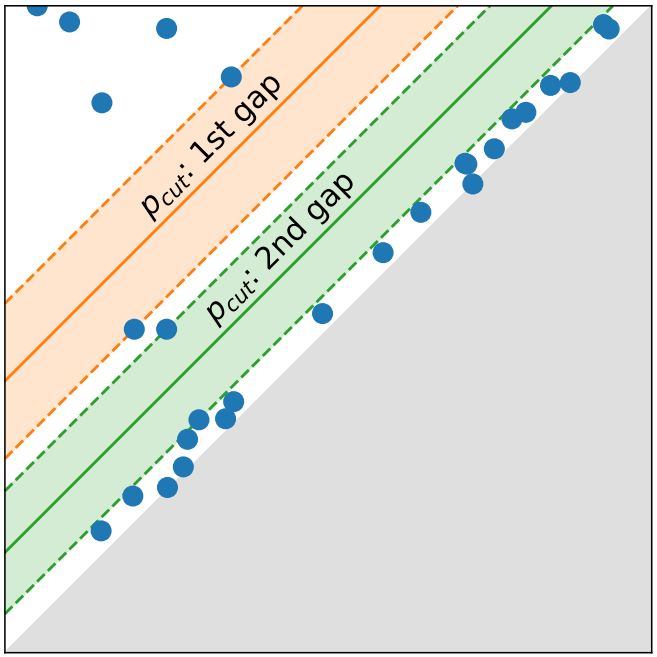
Procedure

1. Perform filtration and prevent subgraphs from merging when their persistence exceeds the persistence cut.
2. Remove all subsequences whose date of birth is greater than the birth cut.
3. For each subgraph (motif), merge subsequences that are time-adjacent (occurrences) while preventing overlapping between occurrences.

Parameter heuristics

Birth cut:
Otsu method: the cut maximizes the inter-class variance.

Persistence cut:
Fixed cut: the number of motifs is given.
Adaptive cut: based on the persistence gap.

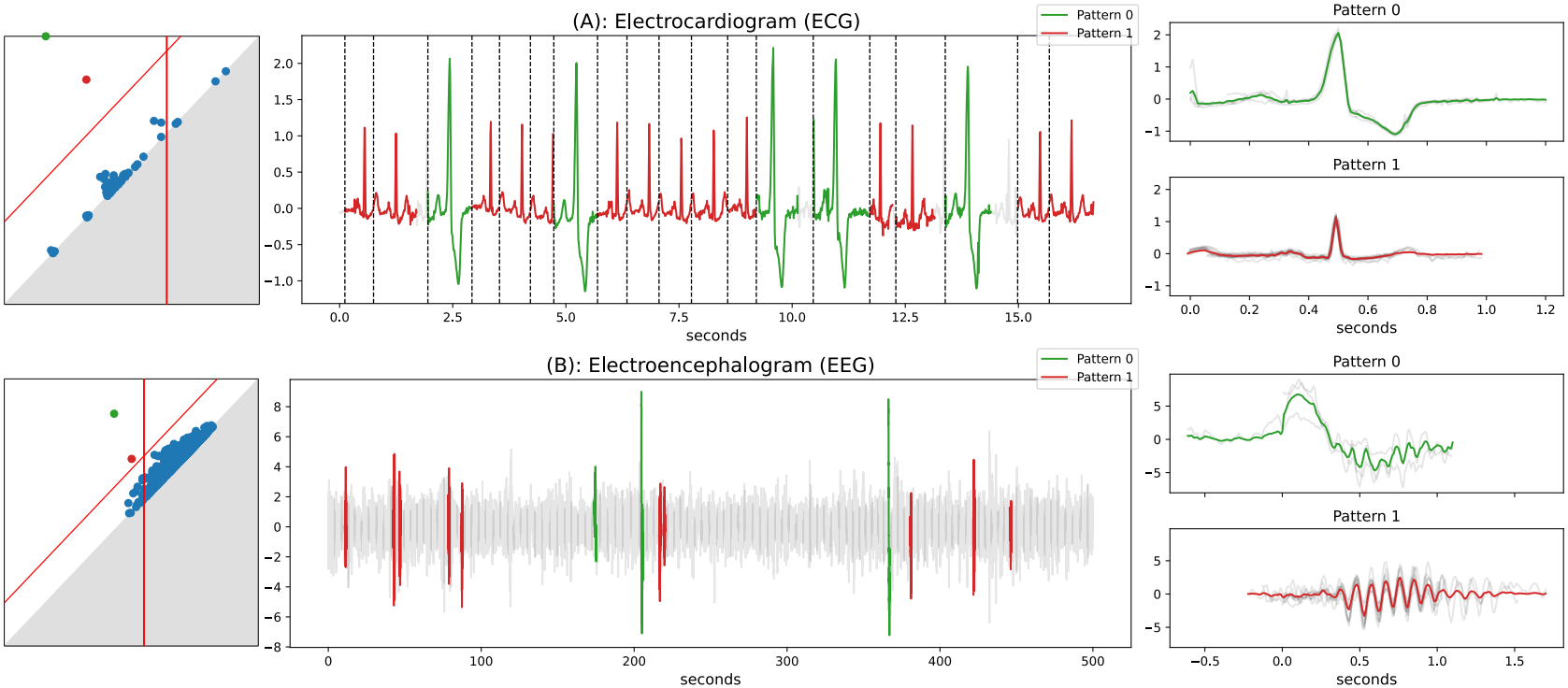


Distance parameter (α, β)**:**
For any $\alpha \in \mathbb{R}_+$ and $\beta \in]0, 2[$, $f_{\alpha,\beta}$ is bijective on $[0, 2]$ and strictly increasing.
Thus, edges' weight order is preserved while their value changes when $f_{\alpha,\beta}$ is modified: the filtration and the MST are preserved.
The optimization criteria is given by:

$$\underset{\alpha,\beta}{\operatorname{argmin}}\, D_{C.S}(B_{\alpha,\beta}, U_{[0,2]})$$

where $D_{C.S}$ is the Cauchy-Schwartz divergence, and $\mathcal{B}_{\alpha,\beta}$ is the kernel density estimation of vertices' birth, and $U_{[0,2]}$ is the density of the uniform distribution on $[0, 2]$.

Examples



Web App

Experiment: comparison with state-of-the-art

Protocol:

- task: motif set discovery
- datasets: 6 real-worlds, 3 synthetics
- metrics: f1-score
- algorithms: PEPA, A-PEPA (adaptative version), and 6 competitors.

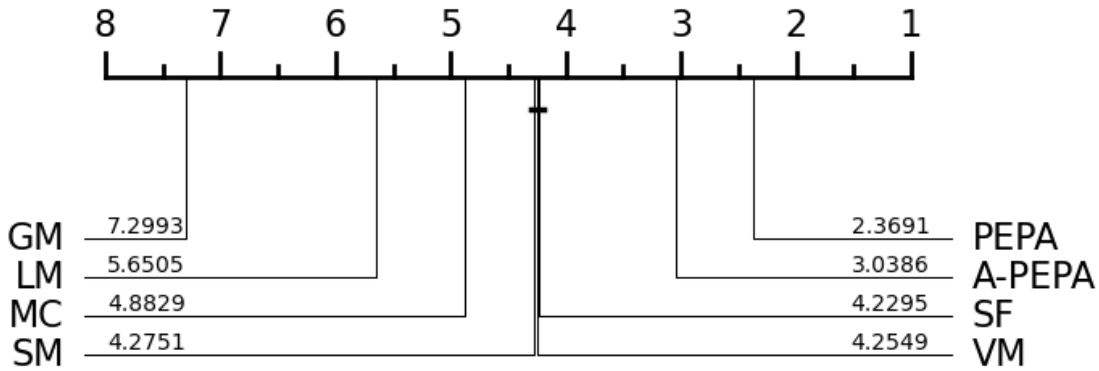
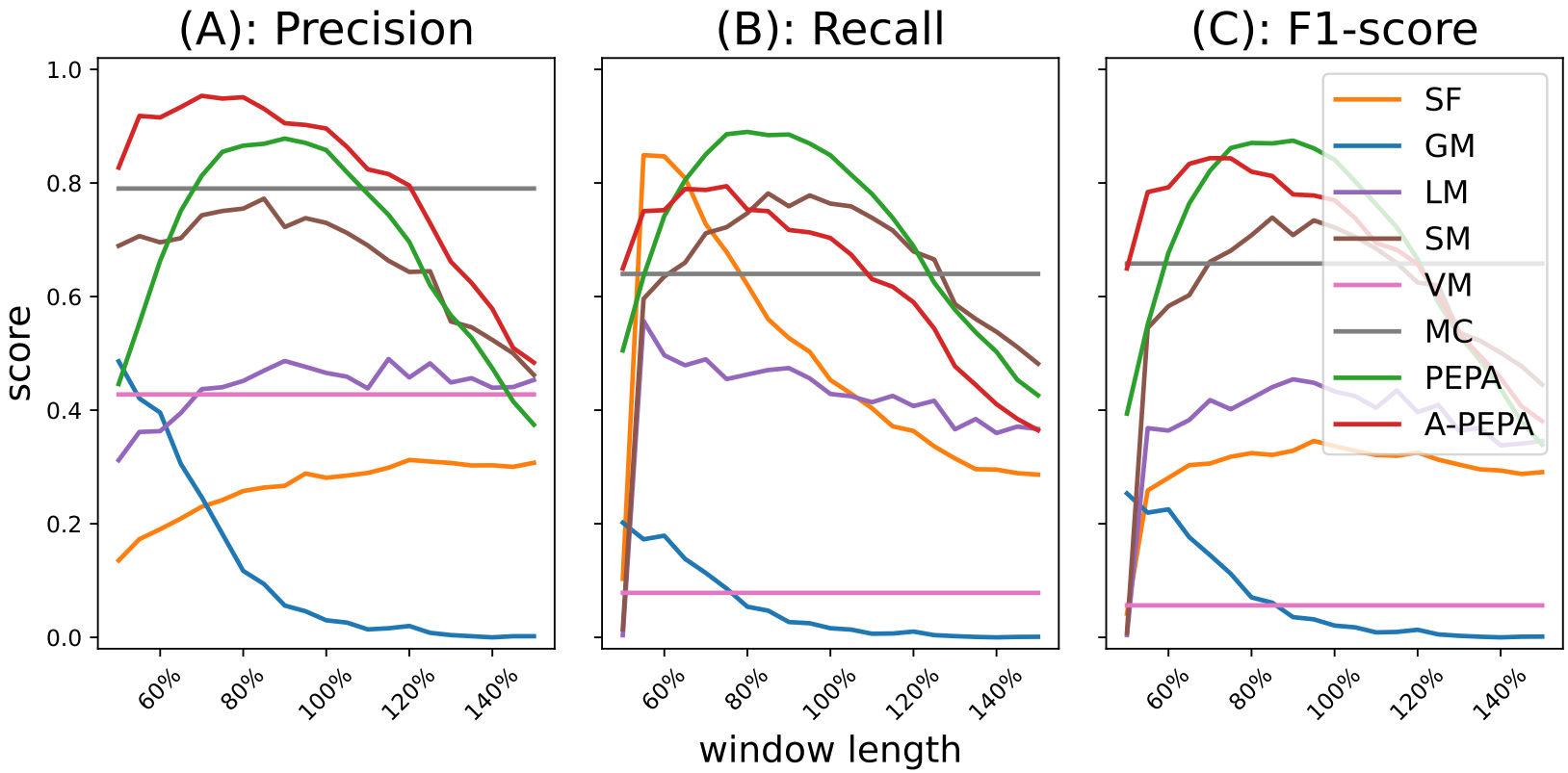


Figure: Critical difference diagram (f1-score based rank, Friedman's test & Nemenyi post-hoc test). PEPA and A-PEPA performs significantly better than other algorithms.

Experiment: Influence of the subsequence length

Question: Does the subsequence length parameter affect the retrieval of variable-length motif sets?



Thank you!

- **PEPA algorithm**: T. Germain, C. Truong, and L. Oudre. "Persistence-Based Motif Discovery in Time Series". In: *IEEE Transactions on Knowledge and Data Engineering* (2024)
- **Graph clustering**: Alexandre Bois, Brian Tervil, and Laurent Oudre. "Persistence-based clustering with outlier-removing filtration". In: *Frontiers in Applied Mathematics and Statistics* 10 (2024), p. 1260828
- **Distance**: Thibaut Germain, Charles Truong, and Laurent Oudre. *Linear-trend normalization for multivariate subsequence similarity search. In Proceedings of the International Conference on Data Engineering Workshops (ICDEW), Utrecht, Netherlands*. 2024
- **Web App:** Thibaut Germain, Charles Truong, and Laurent Oudre. *Interactive motif discovery in time series with persistent homology. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Vilnius, Lithuania*. 2024

Test

$$\mathbb{A}$$