Concept Drift and Anomaly Detection in Graph Streams

Daniele Zambon, *Student Member, IEEE*, Cesare Alippi, *Fellow, IEEE*, and Lorenzo Livi, *Member, IEEE*

***Abstract*— Graph representations offer powerful and intuitive ways to describe data in a multitude of application domains. Here, we consider stochastic processes generating graphs and propose a methodology for detecting changes in stationarity of such processes. The methodology is general and considers a process generating attributed graphs with a variable number of vertices/edges, without the need to assume a one-to-one correspondence between vertices at different time steps. The methodology acts by embedding every graph of the stream into a vector domain, where a conventional multivariate change detection procedure can be easily applied. We ground the soundness of our proposal by proving several theoretical results. In addition, we provide a specific implementation of the methodology and evaluate its effectiveness on several detection problems involving attributed graphs representing biological molecules and drawings. Experimental results are contrasted with respect to suitable baseline methods, demonstrating the effectiveness of our approach.**

***Index Terms*— Anomaly detection, attributed graph, change detection, concept drift, dynamic/evolving graph, embedding, graph matching, stationarity.**

Nomenclature

$\mathcal{G}$	Graph domain.
$d(\cdot,\cdot)$	Graph distance $\mathcal{G}\times\mathcal{G}\to\mathbb{R}_+$.
g_t	Generic graph in $\mathcal{G}$ generated at time t.
$\boldsymbol{g}$	Set $\{g_{t_1},\dots,g_{t_n}\}$ of graphs.
$\mathcal{D}$	Dissimilarity domain $\mathbb{R}^M$.
$d'(\cdot,\cdot)$	Distance $\mathcal{D}\times\mathcal{D}\to\mathbb{R}_+$.
y_t	Dissimilarity representation of generic graph g_t.
$\boldsymbol{y}$	Set $\{y_{t_1},\dots,y_{t_n}\}$ of dissimilarity representations of $\boldsymbol{g}$.
$\zeta(\cdot)$	Dissimilarity representation $(\mathcal{G},d)\to(\mathcal{D},d')$.
R	Set of prototypes $\{r_1,\dots,r_M\}\subset\mathcal{G}$.
$\mathcal{P}$	Stochastic process generating graphs.
$\tau,\widehat{\tau}$	Change time and its estimate.
$(\mathcal{G},\mathcal{S},Q)$	Probability space for $\mathcal{G}$ in nominal regime.
$(\mathcal{D},\mathcal{B},F)$	Probability space for $\mathcal{D}$ in nominal regime.
H_0, H_1	Null and alternative hypotheses of a statistical test.
S_t	Statistic used by the change detection test.
h_t	Threshold for the change detection test.
α_t	Significance level of the change detection test.
$\overline{\boldsymbol{y}}, \mathbb{E}[F]$	Sample mean and expected value with respect to F.
$\mu[\boldsymbol{g}], \mu[Q]$	Fréchet sample and population means with respect to Q.

Manuscript received June 22, 2017; revised November 7, 2017 and February 4, 2018; accepted February 5, 2018. This work was supported by the Swiss National Science Foundation Project under Grant 200021_172671: "ALPSFORT: A Learning graPh-baSed framework FOr cybeR-physical sysTems". *(Corresponding author: Daniele Zambon.)*

D. Zambon is with the Faculty of Informatics, Università della Svizzera italiana, 6900 Lugano, Switzerland (e-mail: daniele.zambon@usi.ch).

C. Alippi is with the Department of Electronics, Information, and Bioengineering, Politecnico di Milan, 20133 Milan, Italy, and also with the Faculty of Informatics, Università della Svizzera italiana, 6900 Lugano, Switzerland (e-mail: cesare.alippi@polimi.it; cesare.alippi@usi.ch).

L. Livi is with the Department of Computer Science, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter EX4 4QF, U.K. (e-mail: l.livi@exeter.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2018.2804443

I. Introduction

LEARNING in nonstationary environments is becoming a hot research topic, as proven by the increasing body of literature on the subject, e.g., [1], [2] for a survey. Within this learning framework, it is of particular relevance the detection of changes in stationarity of the data generating process. This can be achieved by means of either passive approaches [3], which follow a pure online adaptation strategy, or active ones [4], [5], enabling learning only as a proactive reaction to a detected change in stationarity. In this paper, we follow this last learning strategy, though many results are general and can be suitably integrated in passive learning approaches as well.

Most change detection mechanisms have been proposed for numeric independent and identically distributed (i.i.d.) sequences and either rely on change point methods or change detection tests. Both change point methods and change detection tests are statistical tests; the former works offline over a finite number of samples [6] while the latter employs a sequential analysis of incoming observations [7] to detect possible changes. These techniques were originally designed for univariate normal distributed variables, and only later developments extended the methodology to non-Gaussian distributions [8], [9] and multivariate data streams [10], [11].

A somehow related field to change detection tests is one-class classification (e.g., [12] and references therein). There, the idea is to model only the nominal state of a given system and detect nonnominal conditions (e.g., outliers, anomalies, or faults) by means of inference mechanisms. However, one-class classifiers typically process data batches with no specific presentation order, while change detection problems are sequential in nature.

The important role played, nowadays, by graphs as description of dynamic systems is boosting, also thanks to recent discoveries of theoretical frameworks for performing signal processing on graphs [13], [14] and for analyzing

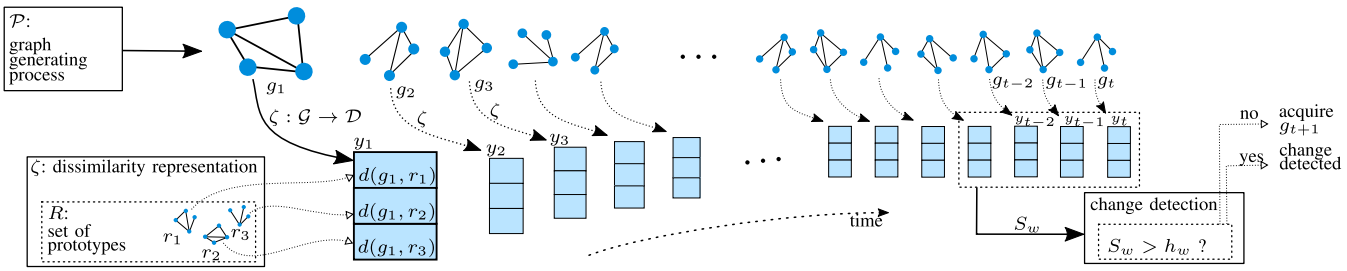

Fig. 1. Fundamental steps of the methodology. At the top of the figure, the stochastic process $\mathcal{P}$ generates over time a stream of graphs $g_1, g_2, g_3, \ldots$. The embedding procedure $\zeta(\cdot)$ is described in the bottom-left corner. The embedding of graph g_t is computed by considering the dissimilarity $d(g_t, r_m)$ with respect to each prototype graph $r_m \in R$, and returns the embedded vector y_t (here 3-dimensional) lying in the dissimilarity space $\mathcal{D}$. The embedding procedure proceeds over time and generates the multivariate vector stream $y_1, y_2, y_3, \ldots$ A change detection method (bottom-right) is applied to the wth window extracted from the y-stream to evaluate whether a change was detected or not, hence, iterating the procedure with the acquisition of a new graph.

temporal (complex) networks [15]–[17]. However, very few works address the problem of detecting changes in stationarity in streams (i.e., sequences) of graphs [18]–[20], and, to the best of our knowledge, none of them tackles the problem by considering a generic family of graphs (e.g., graphs with a varying number of vertices/edges, and arbitrary attributes on them). In our opinion, the reason behind such a lack of research in this direction lies in the difficulty of defining a sound statistical framework for generic graphs that, once solved, would permit to also detect changes in time variance in time-dependent graphs. In fact, statistics grounds on concepts like average and expectation, which are not standard concepts in graph domains. Fortunately, recent studies [21], [22] have provided some basic mathematical tools that allow us to move forward in this direction, hence, addressing the problem of detecting changes in stationarity in sequences of graphs.

A key problem in analyzing generic graphs refers to assessing their dissimilarity, which is a well-known hard problem [23]. The literature proposes two main approaches for designing such a measure of dissimilarity [24]–[26]. In the first case, graphs are analyzed in their original domain $\mathcal{G}$, whereas the second approach consists of mapping (either explicitly or implicitly) graphs to numeric vectors. A well-known family of algorithms used to assess dissimilarity between graphs relies on the graph edit distance (GED) approach [27]. More specifically, GED algorithms count and weight the edit operations that are needed in order to make two input graphs equal. Differently, other techniques take advantage of kernel functions [28], spectral graph theory [29], [30], or assess graph similarity by looking for recurring motifs [31]. The computational complexity associated with the graph matching problem inspired researchers to develop heuristics and approximations (e.g., [32]–[34] and references therein.)

A. Problem Formulation

In this paper, we consider sequences of *attributed graphs*, i.e., directed or undirected labeled graphs characterized by a variable number of vertices and edges [21]. Attributed graphs associate vertices and edges with generic labels, e.g., scalars, vectors, categorical, and user-defined data structures. In addition, multiple attributes can be associated with the same vertex/edge, whenever requested by the application. By considering attributed graphs, we position ourselves on a very general framework covering most of application scenarios. However, generality requires a new operational framework, since all assumptions made in the literature to make the mathematics amenable, e.g., graphs with a fixed number of vertices and/or scalar attributes, cannot be accepted anymore. In order to cover all applications modellable through attributed graphs, we propose the following general problem formulation for change detection.

Given a generic premetric distance $d(\cdot, \cdot)$ on $\mathcal{G}$, we construct a σ-algebra $\mathcal{S}$ containing at least all open balls of $(\mathcal{G}, d)$ and associate a generic probability measure Q to $(\mathcal{G}, \mathcal{S})$. The generated probability space $(\mathcal{G}, \mathcal{S}, Q)$ allows us to consider graphs as a realization of a structured random variable g on $(\mathcal{G}, \mathcal{S}, Q)$. Define $\mathcal{P}$ to be the process generating a generic graph $g_t \in \mathcal{G}$ at time t according to a stationary probability distribution Q (nominal distribution). We say that a change in stationarity occurs at (unknown) time τ when, from time τ on, $\mathcal{P}$ starts generating graphs according to a *nonnominal* distribution $\widetilde{Q} \neq Q$, i.e.,

$$g_t \sim \begin{cases} Q & t < \tau \\ \widetilde{Q} & t \geq \tau. \end{cases}$$

In this paper, we focus on persistent (abrupt) changes in stationarity affecting the population mean. However, our methodology is general and potentially can detect other types of change, including drifts and transient anomalies lasting for a reasonable lapse of time.

B. Contribution and Paper Organization

A schematic description of the proposed methodology to design change detection tests for attributed graphs is shown in Fig. 1, and consists of two steps: 1) mapping each graph g_t to a numeric vector y_t through a prototype-based embedding and 2) using a multivariate change detection test operating on the y-stream for detecting changes in stationarity.

The novelty content of this paper can be summarized as follows.

1) A methodology to detect changes in stationarity in streams of *attributed* graphs. To the best of our knowledge, this is the first research contribution tackling change detection problems in streams of varying-size graphs with nonidentified vertices and user-defined vertex/edge attributes.

2) A method derived from the methodology to detect changes in stationarity in attributed graphs. We stress that the user can design his/her own change detection method by taking advantage of the proposed methodology.
3) A set of theoretic results grounding the proposed methodology on a firm basis.

The proposed methodology is general and advances the few existing approaches for change detection in graph sequences mostly relying on the extraction and processing of topological features of fixed-size graphs (e.g., [35]).

It is worth emphasizing that the proposed approach assumes neither one-to-one nor partial correspondence between vertices/edges across time steps (i.e., vertices do not need to be uniquely identified). This fact has important practical implications in several applications. As a very relevant example, we refer to the identification problem of neurons in extracellular brain recordings based on their activity [36]. In fact, each electrode usually records the activity of a neuron cluster, and single neurons need to be disentangled by a procedure called spike sorting. Hence, a precise identification of neurons (vertices) is virtually impossible in such an experimental setting, stressing the importance of methods that do not require one-to-one correspondence between vertices over time.

The remainder of this paper is structured as follows. Section II contextualizes our contribution and discusses related works. Section III-A presents the proposed methodology for change detection in generic streams of graphs. Theoretical results are sketched in Section III-B; related proofs are given in Appendix C. A specific implementation of the methodology is presented in Section IV and related proofs in Appendix D. Section V shows experimental results conducted on data sets of attributed graphs. Finally, Section VI draws conclusions and offers future directions. Appendixes A and B provide further technical details regarding the problem formulation.

II. Related Works

The relatively new field of temporal networks deals with graph-like structures that undergo events across time [16], [17]. Such events mostly realize in instantaneous or persistent *contacts* between pairs of vertices. With such structures one can study dynamics taking place on a network, like epidemic and information spreading, and/or dynamics of the network itself, i.e., structural changes affecting vertices and edges over time. Further relevant directions in temporal networks include understanding the (hidden) driving mechanisms and generative models [15].

The literature in statistical inference on time-varying graphs (or networks) is rather limited [16], [22], especially when dealing with attributed graphs and nonidentified vertices. Among the many, anomaly detection in graphs emerged as a problem of particular relevance, as a consequence of the ever growing possibility to monitor and collect data coming from natural and man-made systems of various size. An overview of proposed approaches for anomaly and change detection on time-variant graphs is reported in [35] and [37], where the authors distinguish the level of influence of a change. They identify changes affecting vertices and edges, or involving entire subnetworks of different size; this type of change usually concerns static networks, where the topology is often fixed. Other changes have a global influence, or might not be ascribed to specific vertices or edges.

We report that there are several applications in which the vertices are labeled in such a way that, from a time step to another, we are always able to create a partial one-to-one correspondence (identified vertices). This case arises, e.g., when the identity of vertices plays a crucial role and must be preserved over time. Here, we put ourself in the more general scenario where vertices are not necessarily one-to-one identifiable through time.

Within the anomaly detection context, only few works tackle the problem in a classical change detection framework. Among the already published works in detecting changes in stationarity, we mention Barnett and Onnela [19], whose paper deals with the problem of monitoring correlation networks by means of a change point method. In particular, at every time step t, the authors construct the covariance matrix computed from the signals up to time t and the covariance matrix of the remaining data. As statistic for the change point model, they adopt the Frobenius norm between the covariance matrices. The authors evaluate their method on functional magnetic resonance imaging and stock returns. A different way to approach the problem consists in modeling the network-generating process within a probabilistic framework. Graphs with N vertices and disjoint communities can be described by the degree corrected stochastic block model, where some parameters represent the tendency of single vertices to be connected and communities to interact. This model has been adopted by Wilson *et al.* [18] for monitoring the U.S. Senate co-voting network. As monitoring strategy, they consider the standard deviation of each community, and then apply exponential weighted moving average control chart. A further example of change point method for fixed-size graphs combines a generative hierarchical random graph model with a Bayesian hypothesis test [38].

III. Change Detection in a Stream of Graphs

The structure of the section is as follows. Section III-A describes the proposed methodology at a high level to ease the understanding. In Section III-B, we present theoretical results grounding our proposal; their proofs are given in the appendixes.

A. Proposed Methodology

The methodology operates on an input sequence of attributed graphs $g_1, g_2, \ldots, g_t, \ldots \in \mathcal{G}$ and, as sketched in Fig. 1, it performs two steps.

1) Map (embed) input graphs to a vector domain $\mathcal{D} = \mathbb{R}^M$. Embedding is carried out by means of the dissimilarity representation $\zeta : \mathcal{G} \to \mathcal{D}$, which embeds a generic graph $g_t \in \mathcal{G}$ to a vector $y_t \in \mathcal{D}$.
2) Once a multivariate i.i.d. vector stream $y_1, y_2, \ldots, y_t, \ldots$ is formed, change detection is carried out by inspecting such a numerical sequence

with the designer favorite method. The two phases are detailed in the sequel.

1) Dissimilarity Representation: The embedding of a generic graph $g \in \mathcal{G}$ is achieved by computing the dissimilarity between g and the prototype graphs in $R = \{r_1, \ldots, r_M\} \subset \mathcal{G}$

$$y = \zeta(g) := [d(g, r_1), \ldots, d(g, r_M)]^\top. \tag{1}$$

The vector y is referred to as the dissimilarity representation of g. Set R has to be suitably chosen to induce informative embedding vectors. For a detailed discussion about dissimilarity representations and possible ways to select prototypes, we suggest [39].

In order to make the mathematics more amenable, here we assume $d(\cdot, \cdot)$ to be a metric distance; nevertheless, in practical applications, one can choose more general dissimilarity measures.

2) Multivariate Vector Stream: At time step t the process $\mathcal{P}$ generates graph g_t, and the map $\zeta(\cdot)$ embeds g_t onto vector $y_t = \zeta(g_t) \in \mathcal{D}$, inducing a multivariate stream $y_1, y_2, \ldots, y_t, \ldots$ whose elements lie in $\mathcal{D}$. Fig. 1 depicts the continuous embedding of graph process.

Under the nominal condition for process $\mathcal{P}$, graphs $\{g_t\}_{t<\tau}$ are i.i.d. and drawn from probability space $(\mathcal{G}, \mathcal{S}, Q)$. Consequently, also vectors $y_t \in \mathcal{D}$ are i.i.d.. We now define a second probability space $(\mathcal{D}, \mathcal{B}, F)$ associated with embedded vectors y_t; in particular, here we propose to consider for $\mathcal{B}$ the Borel's σ-algebra generated by all open sets in $\mathcal{D}$. F is the push forward probability function of Q by means of $\zeta(\cdot)$, namely,

$$F(B) = Q(\zeta^{-1}(B)) \quad \forall\ B \in \mathcal{B}. \tag{2}$$

With such a choice of F, we demonstrate in Appendix A that F is a probability measure on $(\mathcal{D}, \mathcal{B})$.

3) Change Detection Test: By observing the i.i.d. vector stream $y_1, y_2, \ldots, y_t, \ldots$ over time we propose a multivariate change detection procedure to infer whether a change has occurred in the vector stream and, in turn, in the graph stream.

The change detection test is the statistical hypothesis test

$$\begin{aligned} H_0 &: \quad \mathbb{E}[S_t] = 0 \\ H_1 &: \quad \mathbb{E}[S_t] > 0 \end{aligned}$$

where 0 is the expected value during the nominal—stationary—regime and $\mathbb{E}[\cdot]$ is the expectation operator. Statistic S_t, which is applied to windows of the vector stream, is user-defined and is requested to increase when the process $\mathcal{P}$ becomes nonstationary.

Often, the test comes with a threshold h_t so that if

$$S_t > h_t \quad \Rightarrow \quad \text{a change is detected} \tag{3}$$

and the estimated change time is

$$\widehat{\tau} = \inf\{t\ :\ S_t > h_t\}.$$

Whenever the distribution of S_t under hypothesis H_0 is available—or can be estimated—the threshold can be related to an user-defined significance level α_t so that $\alpha_t = \mathbb{P}(S_t > h_t | H_0)$.

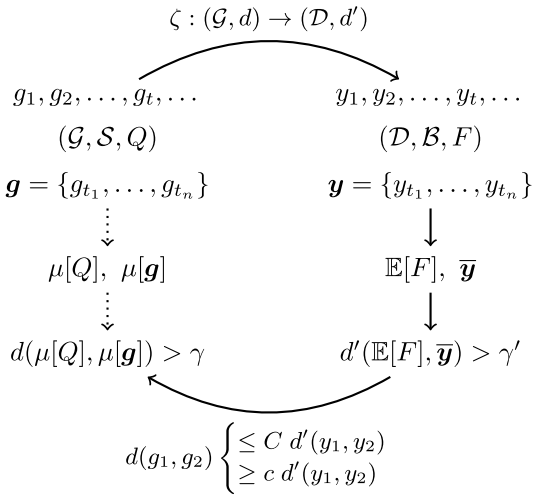

Fig. 2. Conceptual scheme of the proposed methodology that highlights the theoretical results. The key point is to show that objects close in the graph domain $\mathcal{G}$, onto which metric $d(\cdot, \cdot)$ is defined, under certain conditions, are close also in the embedded domain $\mathcal{D}$ controlled by metric $d'(\cdot, \cdot)$. It comes up that if objects in the embedding domain $\mathcal{D}$ are distant in probability, then also the related graphs in $\mathcal{G}$ are distant, hence indicating a change in stationarity according to a given false positive rate.

B. Theoretical Results

In this section, we show some theoretical results related to the methodology presented in Section III-A. In particular, we prove the following claims.

(C1) Once a change is detected in the dissimilarity space according to a significance level α', then a change occurs in probability with significance level α also in the graph domain;

(C2) If a change occurs in the graph domain having set a significance level α, then, with a significance level α', a change occurs also in the dissimilarity space.

Fig. 2 depicts the central idea behind the methodology. Through transformation $\zeta(\cdot)$, we map graphs to vectors. In the transformed space, we consider the expectation $\mathbb{E}[F]$ and the sample mean $\overline{\boldsymbol{y}}$ associated with set $\boldsymbol{y} = \{y_{t_1}, \ldots, y_{t_n}\}$ obtained by embedding the graph set $\boldsymbol{g} = \{g_{t_1}, \ldots, g_{t_n}\}$, and design a hypothesis test of the form

$$\begin{aligned} H_0 &: d'(\mathbb{E}[F], \overline{\boldsymbol{y}}) \leq \delta \\ H_1 &: d'(\mathbb{E}[F], \overline{\boldsymbol{y}}) > \delta \end{aligned} \tag{4}$$

where δ is a positive threshold and hypothesis H_0 is associated with a nominal, change-free condition. In this paper, we relate the y-test of (4) to a correspondent test $d(\mu[Q], \mu[\boldsymbol{g}]) > \gamma$ operating in the graph domain, where $\mu[Q]$, $\mu[\boldsymbol{g}]$ are, respectively, the population and sample mean defined according to Fréchet [40]; for further details about Fréchet statistics refer to Appendix B.

Define α and α' to be two significance levels, such that

$$\begin{aligned} \alpha &= \mathbb{P}(d(\mu[Q], \mu[\boldsymbol{g}]) > \gamma\,|H_0) \\ \alpha' &= \mathbb{P}(d'(\mathbb{E}[F], \overline{\boldsymbol{y}}) > \gamma'|H_0). \end{aligned} \tag{5}$$

In the sequel, we relate the threshold γ to γ', so that also the significance levels α and α' are in turn related to each other.

In order to address our problem, we introduce mild assumptions to obtain closed-form expressions. Such assumptions are satisfied in most applications. (A3)

(A1) We assume that the attributed graph space $(\mathcal{G}, d)$ and the dissimilarity space $(\mathcal{D}, d')$ are metric spaces; in particular, $(\mathcal{G}, d)$ is chosen as a graph alignment space [22]—i.e., a general metric space of attributed graphs—and $d'(\cdot,\cdot)$ has to be induced by a norm.

(A2) We put ourselves in the conditions of [22] in order to take advantage of results therein; specifically, we assume that the Fréchet function $\mathcal{F}_Q(g) = \int_{\mathcal{G}} d^2(g, f)\, dQ(f)$ is finite for any $g \in \mathcal{G}$, and there exists a sufficiently asymmetric graph f such that the support of Q is contained in a cone around f. In this way, we are under the hypotheses of Theorems 4.1 and 4.2 of [22], which ensure the existence and uniqueness of the Fréchet population and sample mean in $\mathcal{G}$.

(A3) The embedding function $\zeta : (\mathcal{G}, d) \to (\mathcal{D}, d')$ is bilipschitz, i.e., there exist two constants $c, C > 0$, such that for any pair $g, f \in \mathcal{G}$

$$d(g, f) \geq c\, d'(\zeta(g), \zeta(f)) \tag{6}$$

$$d(g, f) \leq C\, d'(\zeta(g), \zeta(f)). \tag{7}$$

Let $\Psi(\cdot)$ and $\Upsilon(\cdot)$ be the cumulative density functions (CDFs) of $d(\mu[\boldsymbol{g}], \mu[Q])$ and $d'(\overline{\boldsymbol{y}}, \mathbb{E}[F])$, respectively. Proposition 1 bounds the distribution $\Psi(\cdot)$ in terms of $\Upsilon(\cdot)$. This fact yields the possibility to derive significance levels α, α' and thresholds γ, γ' (5) that are related.

In order to prove Proposition 1, we make use of two auxiliary results, Lemmas 1 and 2. At first, we need to comment that, although in general $\overline{\boldsymbol{y}} \neq \zeta(\mu[\boldsymbol{g}])$ and $\mathbb{E}[F] \neq \zeta(\mu[Q])$, differences are bounded in practice, as shown by Lemma 1.

Lemma 1: Considering a set $\boldsymbol{g}$ of i.i.d. random graphs and the associated embedded set $\boldsymbol{y}$, there exists a constant v_2 such that

$$\|\mathbb{E}[F] - \zeta(\mu[Q])\|_2^2 \leq v_2$$
$$\mathbb{P}(\|\overline{\boldsymbol{y}} - \zeta(\mu[\boldsymbol{g}])\|_2^2 \geq \delta) \leq \frac{v_2}{\delta} \quad \forall \delta > 0.$$

Lemma 2 is used to derive bounds on the marginal distributions from bounds on the joint distributions that are useful in Proposition 1 to relate the threshold and the significance level in the graph space with the ones in the dissimilarity space.

Lemma 2: Consider a random variable $x \in \mathcal{X}$ and two statistics $d_1(\cdot), d_2(\cdot) : \mathcal{X} \to \mathbb{R}_+$ with associated CDFs $\Phi_1(\cdot)$, $\Phi_2(\cdot)$, respectively. If function $u : \mathbb{R}_+ \to \mathbb{R}_+$ is increasing and bijective and p is a constant in $[0, 1]$, then

$$\mathbb{P}(d_1(x) \leq u(d_2(x))) \geq p \Rightarrow \Phi_1(\cdot) \geq p \cdot \Phi_2(u^{-1}(\cdot)).$$

Claims (C1) and (C2) now follow from the relation between the CDFs $\Psi(\cdot)$ and $\Upsilon(\cdot)$ proven by the subsequent Proposition 1. In particular, regarding Claim (C1) Proposition 1 provides a criterion for setting a specific threshold γ' for the test (4) for which we can state that $d(\mu[Q], \mu[\boldsymbol{g}])$ is unexpectedly large with a significance level at most α; similarly, we obtain Claim (C2).

Proposition 1: Consider a sample $\boldsymbol{g}$ of i.i.d. graphs. Under assumptions (A1)–(A3), if $\Psi(\cdot)$ and $\Upsilon(\cdot)$ are the CDFs of statistics $d(\mu[Q], \mu[\boldsymbol{g}])$ and $d'(\mathbb{E}[F], \overline{\boldsymbol{y}})$, respectively, then for every $\delta > 0$ there exist two values b_δ and p_δ, depending on δ but independent of $\boldsymbol{g}$, such that

$$p_\delta \cdot \Upsilon\left(\frac{\gamma}{C} - b_\delta\right) \leq \Psi(\gamma) \leq \frac{1}{p_\delta} \cdot \Upsilon\left(\frac{\gamma}{c} + b_\delta\right).$$

The proofs are given in Appendix C. Results of Proposition 1 allows us to state the major claim (C1): given a sample $\boldsymbol{g}$, for any significance level α and threshold γ, as in (5), we can set up a γ' such that the confidence level of detecting a change in $\mathcal{D}$ is at least $p_\delta(1-\alpha)$. Specifically, for $\gamma' = (\gamma/c) + b_\delta$ we have that

$$\mathbb{P}(d'(\mathbb{E}[F], \overline{\boldsymbol{y}}) \leq \gamma' \mid H_0) = \Upsilon\left(\frac{\gamma}{c} + b_\delta\right) \geq p_\delta(1-\alpha). \tag{8}$$

Equation (8) states that if no change occurs in $\mathcal{G}$ with confidence level $1-\alpha$, then a change will not be detected in $\mathcal{D}$ according to threshold γ' at least with confidence level $p_\delta(1-\alpha)$. Indeed, this proves (C1) by contraposition. Similarly, Proposition 1 allows to prove Claim (C2). In fact, for any α' and γ' as in (5), we can set γ so that $\gamma' = (\gamma/C) - b_\delta$ and obtain

$$\mathbb{P}(d(\mu[Q], \mu[\boldsymbol{g}]) \leq \gamma \mid H_0) = \Psi(\gamma) \geq p_\delta(1-\alpha').$$

IV. Implementations of the Methodology

This section describes two examples showing how to implement the proposed methodology and derive actual change detection tests. In Section IV-A, we present a specific method for generic families of graphs, whereas in Section IV-B, we further specialize the methodological results to a special case considering graphs with identified vertices.

A. Change Detection Based on Central Limit Theorem

Here, we consider specific techniques for prototype selection and change detection test. For the sake of clarity, we keep the subsection structure of Section III-A. Both the prototype selection and change detection test require a training phase. For this reason, the first observed graphs of the stream will serve as training set T, that we assume to be all drawn under nominal conditions.

1) Dissimilarity Representation: Since the change detection method operates in the dissimilarity space, we need to define the embedding $\zeta(\cdot)$ that, at each time step, maps a generic graph g_t to a vector y_t.

We comment that the embedding $\zeta(\cdot)$ is completely defined once the graph distance metric $d(\cdot,\cdot)$ and prototype set R are chosen. Here, we adopt a metric GED as graph distance, since it meets Assumption (A1) of $(\mathcal{G}, d)$ being a graph alignment space.

Many approaches have been proposed to deal with the prototype selection problem (e.g., [41] and references therein). While the proposed methodology is general and one can choose any solution to this problem, here we adopt the k-Centres method [42]. The method selects prototypes so as

to cover training data with balls of equal radius. Specifically, the algorithm operates as follows:

1) select M random prototypes $R := \{r_1, \ldots, r_M\} \subseteq T$;
2) for each $r_m \in R$, consider the set C_m of all $g \in T$ such that $d(r_m, g) = \min_{r \in R} d(r, g)$;
3) for $m = 1, \ldots, M$ update the prototypes r_m with a graph $c \in T$ minimizing $\max_{g \in C_m} d(c, g)$;
4) if the prototype set R did not change from the previous iteration step then exit, otherwise go to step 2.

In order to improve the robustness of the k-Centres algorithm, we repeat steps 1–4 by randomizing initial conditions and select the final prototype set R to be

$$R = \underset{R \in \{R^{(i)}\}}{\arg\min} \left\{ \max_{r_m \in R} \left[\max_{c \in C_m} d(r_m, c) \right] \right\}$$

where $\{R^{(i)}\}$ is the collection of prototype sets found at each repetition.

2) Multivariate Vector Stream: Every time the process $\mathcal{P}$ generates a graph g_t, we embed it as $y_t = \zeta(g_t)$ by using the prototype set R identified with the k-Centres approach. This operation results in a multivariate vector stream $y_1, y_2, \ldots, y_t, \ldots$ on which we apply the change detection test.

3) Change Detection Test: We consider here a variation of the cumulative sum (CUSUM) test [8] to design the change detection test. CUSUM is based on the cumulative sum chart [43], it has been proven to be asymptotically optimal [7] and allows for a simple graphical interpretation of results [44]. Here, the CUSUM is adapted to the multivariate case. However, we remind that, in principle, any change detection test can be used on the embedded sequence.

We batch the observed embedding vectors into nonoverlapping samples $\boldsymbol{y}_w := \{y_{(w-1)n+1}, \ldots, y_{wn}\}$ of length n, where index w represents the wth data window. For each window, we compare the sample mean $\overline{\boldsymbol{y}}_0$ estimated in the training set T with that estimated in the wth window, i.e., $\overline{\boldsymbol{y}}_w$ and compute the discrepancy

$$s_w := d'(\overline{\boldsymbol{y}}_0, \overline{\boldsymbol{y}}_w).$$

By assuming that $y_1, y_2, \ldots, y_t, \ldots$ are i.i.d., and given sufficiently large $|T|$ and n, the central limit theorem grants that $\overline{\boldsymbol{y}}_0$ and $\overline{\boldsymbol{y}}_w$ are normally distributed. In particular, $\overline{\boldsymbol{y}}_0$ and $\overline{\boldsymbol{y}}_w$ share the same expectation $\mathbb{E}[F]$, while covariance matrices are $(1/|T|)\,\mathrm{Var}[F]$ and $(1/n)\,\mathrm{Var}[F]$, respectively.

As a specific choice of $d'(\cdot, \cdot)$, we adopt the Mahalanobis' distance, i.e., $d'(\overline{\boldsymbol{y}}_0, \overline{\boldsymbol{y}}_w) := d_\Sigma(\overline{\boldsymbol{y}}_0, \overline{\boldsymbol{y}}_w)$ where

$$d_\Sigma(\overline{\boldsymbol{y}}_0, \overline{\boldsymbol{y}}_w) = \sqrt{(\overline{\boldsymbol{y}}_0 - \overline{\boldsymbol{y}}_w)^\top \Sigma^{-1} (\overline{\boldsymbol{y}}_0 - \overline{\boldsymbol{y}}_w)} \tag{9}$$

with matrix $\Sigma = ((1/|T|) + (1/n))\,\mathrm{Var}[F]$, i.e., the covariance matrix of $\overline{\boldsymbol{y}}_0 - \overline{\boldsymbol{y}}_w$. In our implementation, we consider as covariance matrix $\mathrm{Var}[F]$ the unbiased estimator $(1/(|T| - 1)) \sum_{y \in T} (y - \overline{\boldsymbol{y}}_0) \cdot (y - \overline{\boldsymbol{y}}_0)^\top$.

For each stationary window w, the squared Mahalanobis' distance s_w^2 is distributed as a χ^2_M.

The final statistic S_w inspired by the CUSUM test is defined as

$$\begin{cases} S_w = \max\{0,\ S_{w-1} + (s_w - q)\} \\ S_0 = 0. \end{cases} \tag{10}$$

The difference $s_w - q$ increases with the increase of the discrepancy in (9) and positive values support the hypothesis that a change occurred, whereas negative values suggest that the system is still operating under nominal conditions. This behavior resembles that of the original CUSUM increment associated with the log-likelihood ratio. In particular, the parameter q can be used to tune the sensitivity of the change detection; in fact, if q^2 is the β-quantile $\chi^2_M(\beta)$, then $s_w - q$ produces a negative increment with probability β and a positive one with probability $1 - \beta$.

The last parameter to be defined is the threshold h_w as requested in (3). Since the nonnominal distribution is unknown, a common criterion suggests to control the rate of false alarms α. In sequential detection tests, a related criterion requires a specific *average run length* under the null hypothesis (ARL0) [45]. ARL0 is connected to the false positive rate in the sense that setting a false alarm rate to α yields an ARL0 of α^{-1}. Since we propose a sequential test, statistic S_w depends on statistics at preceding times. As a consequence, since we wish to keep α fixed, we end up with a time-dependent threshold h_w. As done in [10], [11], we numerically determine thresholds through Monte Carlo sampling by simulating a large number of processes of repeated independent χ^2_M realisations. Threshold h_w is then the $1 - \alpha$ quantile of the estimated distribution of S_w.

We point out that, when setting a significance level for the random variable S_w, we are implicitly conditioning to the event $S_i \leq h_i,\ \forall i < w$; in fact, when S_w exceeds h_w, we raise an alarm and reconfigure the detection procedure.

4) Theoretical Results: The choice of Mahalanobis' distance ensures that almost all assumptions in Section III-B are met. In particular, the Mahalanobis' distance meets the requirements of Assumption (A1). Then, the following Lemma 3 provides a lower bound of the form (6) in Assumption (A3); specifically, the lemma shows that, up to a positive factor, the distance between two graphs is larger than the one between the associated embedding vectors.

From these properties, we can apply Proposition 1 to state that Claim (C1) holds. Hence, for any α, we can set a specific threshold γ' yielding a confidence level at least $p_\delta(1 - \alpha)$.

Lemma 3: For any two graphs $g, f \in \mathcal{G}$, we have that

$$d(g, f) \geq \sqrt{\frac{\lambda_M}{M}}\, d_\Sigma(\zeta(g), \zeta(f))$$

where λ_M is the smallest eigenvalue of Σ.

The proof is reported in Appendix D. Distance $d_\Sigma(\cdot, \cdot)$ is well defined only when Σ is positive definite, a condition implying that selected prototypes are not redundant.

B. Special Case: Graphs With Identified Vertices

Here, we take into account the particular scenario where the attribute function $a(\cdot)$ of $g = (V, E, a) \in \mathcal{G}$ assigns

numerical attributes in $[0, 1]$ to vertices and edges of g; the vertex set V is a subset of a predefined finite set $\mathcal{V}$, with $|\mathcal{V}| = N$. The peculiarity of this space $\mathcal{G}$ resides in the fact that any vertex permutation of a graph leads to a different graph. Many real-world applications fall in this setting, for instance correlation graphs obtained by signals coming from sensors or graphs generated by transportation networks.

We show an example of method for this setting, which complies with the methodology and satisfies Assumption (A3). This fact follows from the existence of an injective map ω from $\mathcal{G}$ to the $[0, 1]^{N\times N}$ matrix set. Indeed, we represent each graph with its weighted adjacency matrix whose row/column indices univocally correspond to the vertices in $\mathcal{V}$. By endowing $\mathcal{G}$ with the Frobenius distance[1] $d_F(g_1, g_2) := \|\omega(g_1) - \omega(g_2)\|_F$, map $\omega : (\mathcal{G}, d_F) \to ([0, 1]^{N\times N}, \|\cdot\|_F)$ is an isometry.

Being the co-domain of ω an Euclidean space, we compute a matrix $X \in \mathbb{R}^{k\times M}$ whose columns x_i constitute a k-dimensional vector configuration related to the prototype graphs; this is done via Classical Scaling [39], that is, $\|x_i - x_j\|_2 = d(r_i, r_j)$ for all pairs of prototypes $r_i, r_j \in R$. As usual, we consider the smallest possible k that preserves the data structure as much as possible. Successively, for any dissimilarity vector $y \in \mathcal{D}$, we define $u := XJ\,y^{*2}$ to be a linear transformation of y^{*2}, obtained by squaring all the components of y; the matrix J is the centering matrix $I - (1/M)\mathbf{1}\mathbf{1}^\top$. We apply the same procedure of Section IV, considering u instead of y: matrix Σ is derived from the nonsingular covariance matrix,[2] and the statistic s_w is the Mahalanobis distance $d_\Sigma(\overline{\boldsymbol{u}}_0, \overline{\boldsymbol{u}}_w)$.

Considering the space $(\mathcal{G}, d_F)$ and the above transform, we claim that the following lemma holds. Lemma 4 proves the fulfillment of Assumption (A3).

Lemma 4: For any positive definite matrix $\Sigma \in \mathbb{R}^{k\times k}$

$$c\; d_\Sigma(u_1, u_2) \le d_F(g_1, g_2) \le C\; d_\Sigma(u_1, u_2)$$

where $c = ((\lambda_k(\Sigma))/(4\lambda_1(XX^\top)))^{1/2}$, $C = ((\lambda_1(\Sigma))/(4\lambda_k(XX^\top)))^{1/2}$. $\lambda_i(\cdot)$ is the ith eigenvalue in descending order of magnitude.

The proof is reported in Appendix D.

V. Experiments

The proposed methodology can operate on very general families of graphs. Besides the theoretical foundations around which the paper is built, we provide some experimental results showing the effectiveness of what proposed in real change detection problems in streams of graphs. In particular, we consider the method introduced in Section IV-A as an instance of our methodology. Source code for replicating the experiments is available in [46].

A. Experimental Description

1) Data: The experimental evaluation is performed on the well-known IAM benchmarking databases [47]. The IAM data sets contain attributed graphs representing drawings and biochemical compounds. Here, we consider the *Letters*, *Mutagenicity*, and *AIDS* data sets.

TABLE I

Experimental Settings. The First Column Contains an Identifier for Each Experiment; in Particular, in the Letter Data Set, "D," "O," and "S," Stand for Disjoint, Overlapping, and Subset, Respectively. The Second Column Reports the Data Set Involved, and the Third and Fourth Columns Show the Set of Classes From Which Nominal/Nonnominal Graphs Are Extracted. The Collections of Letters Are Selected in Alphabetical Order

Experiment ID	Data Set	Collection	
		Nominal	Nonnominal
L-D2	Letter	A,E	F,H
L-D5	Letter	A,E,F,H,I	K,L,M,N,T
L-O	Letter	A,E,F,H	F,H,I,K
L-S	Letter	A,E,F,H,I	F,H,I
MUT	Mutagenicity	nonmutagenic	mutagenic
AIDS	AIDS	inactive	active

The Letters data set contains 2-dimensional geometric graphs. As such, each vertex of the graphs is characterized by a real vector representing its location on a plane. The edges define lines such that the graphical planar representation of the graphs resembles a Latin-script letter. The data set is composed of 15 classes (one for each letter[3]) containing 150 instances each.

Conversely, the Mutagenicity and AIDS data sets contain biological molecules. Molecules are represented as graphs by considering each atom as a vertex and each chemical link as an edge. Each vertex is attributed with a chemical symbol, whereas the edges are labeled according to the valence of the link. Both data sets contain two classes of graphs: mutagenic/nonmutagenic for Mutagenicity and active/inactive for AIDS. The two data sets are imbalanced in terms of size of each classes, in particular Mutagenicity has 2401 mutagenic and 1963 nonmutagenic molecules; AIDS contains 400 active and 1600 inactive molecules.

We considered these data sets because they contain different types of graphs with variegated attributes (numerical and categorical). We refer the reader to [47] and references therein for a more in-depth discussion about the data sets.

2) Simulating the Generating Process $\mathcal{P}$: For each experiment in Table I, we consider two collections of graphs containing all possible observations in the nominal and nonnominal regimes, respectively. Each collection is composed by graphs present in one or more predefined classes of the data set under investigation. The collections have to be different, but they do not need to be disjoint; as such, some graphs can belong to both collections.

Next, we simulate the process $\mathcal{P}$ by bootstrapping graphs from the first collection up to the predefined change time τ; this is the nominal regime. After τ, we bootstrap objects from the second collection hence modeling a change.

[1] In $\mathcal{G}$, $d_F(\cdot, \cdot)$ is a graph alignment distance, as formally shown in Appendixes D–B.

[2] Σ is nonsingular, since there are no isolated points in $\mathcal{G}$ and as a consequence of the selection of k.

[3] The IAM letters database [47] considers only noncurved letters; hence, e.g., letters A and E are considered, whereas B and C are excluded.

TABLE II

RESULTS ATTAINED BY OUR METHOD VARYING NUMBER OF PROTOTYPES (M) AND WINDOW SIZE (n). SYMBOL $*$ INDICATES DCR STATISTICALLY LARGER THAN 0.99, WITH A CONFIDENCE OF 95%. SYMBOLS $\circ$ AND $\bullet$ HIGHLIGHT SIGNIFICANT TRENDS IN THE DCR INCREASING M AND n, RESPECTIVELY

Experiment			DCR		ARL0		DoD		FA1000	
Data Set	M	n	mean	95CI	mean	95CI	mean	95CI	mean	std
L-D2	4	5	$*$ 1.000	[1.000, 1.000]	189	[103, 333]	26	[8, 66]	1.135	0.368
L-D5	4	5	0.380	[0.290, 0.480]	175	[91, 271]	411	[9, 2149]	1.207	0.390
L-D5	4	25	0.970	[0.930, 1.000]	183	[88, 306]	41	[1, 184]	0.233	0.086
L-D5	4	125	$*$ 1.000	[1.000, 1.000]	305	[61, 1148]	2	[1, 5]	0.045	0.038
L-D5	8	25	0.990	[0.970, 1.000]	175	[93, 372]	10	[1, 54]	0.245	0.078
L-D5	8	125	$*$ 1.000	[1.000, 1.000]	255	[50, 928]	1	[1, 1]	0.054	0.047
L-O	4	5	0.230	[0.150, 0.310]	191	[120, 344]	410	[21, 1136]	1.096	0.318
L-O	4	25	$\circ$ 0.790	[0.710, 0.870]	205	[101, 355]	123	[3, 723]	0.205	0.070
L-O	4	125	0.940	[0.890, 0.980]	291	[55, 725]	37	[1, 405]	0.045	0.035
L-O	8	25	$\circ$ 0.980	[0.950, 1.000]	163	[92, 306]	24	[2, 132]	0.260	0.080
L-O	8	125	$*$ 1.000	[1.000, 1.000]	238	[52, 620]	2	[1, 5]	0.052	0.053
L-S	4	5	0.720	[0.630, 0.810]	164	[90, 359]	132	[33, 349]	1.321	0.418
L-S	4	25	$*$ 1.000	[1.000, 1.000]	193	[103, 372]	28	[4, 123]	0.223	0.075
AIDS	4	5	$*$ 1.000	[1.000, 1.000]	175	[100, 337]	1	[1, 1]	1.221	0.364
MUT	4	5	$\bullet$ 0.050	[0.010, 0.100]	37	[18, 89]	70	[39, 124]	6.294	2.159
MUT	4	25	$\bullet$ 0.800	[0.720, 0.880]	52	[19, 145]	53	[4, 348]	1.075	0.521
MUT	4	125	$\bullet$ 0.980	[0.950, 1.000]	153	[8, 820]	5	[1, 22]	0.276	0.284
MUT	8	25	0.920	[0.860, 0.970]	42	[15, 91]	15	[2, 75]	1.242	0.680
MUT	8	125	$*$ 1.000	[1.000, 1.000]	112	[9, 540]	2	[1, 4]	0.292	0.267

Regarding molecular data sets, we considered two distinct experiments. For the Mutagenicity (AIDS) experiment, we set the nonmutagenic (inactive) class as nominal collection and mutagenic (active) class as nonnominal one. On the other side, for the Letter data set we design four different experiments depending on which classes will populate the collections. Table I reports the settings of all the experiments and Section V-A.3 describes the relevant parameters.

3) Parameters Setting: For all experiments, the offset q is set to the third quartile of the $\chi^2(M)$ distribution. The time-dependent threshold h_w has been numerically estimated by Monte Carlo sampling. We drew one million processes of i.i.d. random variables s_w by taking the square root of i.i.d. random variables distributed as $\chi^2(M)$. For each obtained process (stream), we computed the sequence of cumulative sums S_w like in (10) and estimated the threshold h_w as the quantile of order α_w =1/ARL0, with ARL0 = 200 (windows).

We divided the training set T in two disjoint subsets, T_c *and* T_p, used during the prototype selection and change detection learning phases, respectively. We set $|T_p| = 300$ and $|T_c| = 1000$, afterward we generated a stream of graphs containing $20 \cdot n \cdot$ ARL0 observations associated with the operational phase. The change is forced at time $\tau = 12 \cdot n \cdot$ ARL0. As for the distance $d(\cdot, \cdot)$, we considered the bipartite GED implemented in [48], where we selected the Volgenant and Jonker assignment algorithm. The other GED parameters are set according to the type of graphs under analysis, i.e., for geometric graphs we consider the Euclidean distance between numerical attributes, and a binary (0-1 distance) for categorical attributes. The k-Centres procedure is repeated 20 times.

We believe that the selected parameter settings are reasonable. Nevertheless, a proper investigation of their impact with respect to performance metrics is performed in a companion paper [49], hence it is outside the focus of the present one.

4) Figures of Merit: We assess the performance of the proposed methodology by means of the figures of merit here described. Such measurements are obtained by replicating each experiment one hundred times; we report the average of the observed measures with their estimated 95% confidence interval (95CI) or standard deviation (std).

First of all, we consider the observed ARL0 introduced in Section III-A3 and the delay of detection (DoD). Both of them are computed as the average time lapses between consecutive alarms, but limiting to those alarms raised before and after time τ, respectively. From them, we estimate the rate of detected changes [detection rate (DCR)], by assessing the rate of simulations in which the DoD is less than the observed ARL0. Finally, we consider also the estimated rate of false anomalies within 1000 samples (FA1000). This is computed as the ratio between the count of raised false alarms and the total number of thousands of time steps under the nominal condition.

We point out that the measures ARL0, DoD, and FA1000 are computed with the window as unitary time step.

5) Baseline Methods: As previously mentioned, state-of-the-art change detection methods for graph streams usually assume a given topology with a fixed number of vertices and/or simple numeric features for vertices/edges. As reported in [35], considering a variable topology, a common methodology for anomaly detection on graphs consists of extracting some topological features at each time step and then applying a more general anomaly detector on the resulting numeric sequence. Accordingly, in addition to the method proposed in Section IV, we consider two baseline methods for comparison. More precisely, we considered two topological features: the density of edges $\phi_1(g) = |E|/(|V|(|V| - 1))$ and the spectral gap $\phi_2(g) = |\lambda_1(L(g))| - |\lambda_2(L(g))|$ of the Laplacian matrix $L(g)$ [50]. The particular choices of ϕ_1 and ϕ_2 can be justified by considering that both features are suitable for describing

TABLE III

RESULTS ATTAINED BY BASELINE METHODS (INDEX COLUMN) BASED ON THE GRAPH DENSITY (DEN), THE SPECTRAL GAP OF THE LAPLACIAN (SG), AND THE DEGENERATE IMPLEMENTATION OF THE METHODOLOGY WITH $M = 1$ PROTOTYPE (M1). SYMBOL † INDICATES SIGNIFICANTLY BETTER (95% CONFIDENCE) RESULTS

Experiment			DCR		ARL0		DoD		FA1000	
Data Set	index	n	mean	95CI	mean	95CI	mean	95CI	mean	std
L-D2	M1	25	1.000	[1.000, 1.000]	238	[106, 594]	15	[2, 87]	0.194	0.082
L-D2	DEN	1	1.000	[1.000, 1.000]	194	[68, 556]	41	[22, 104]	6.475	3.337
L-D2	SG	1	0.850	[0.780, 0.920]	210	[75, 394]	133	[55, 274]	6.460	3.342
L-D5	M1	25	0.780	[0.700, 0.860]	222	[96, 407]	85	[2, 542]	0.200	0.098
L-D5	DEN	1	0.180	[0.110, 0.260]	219	[70, 485]	369	[91, 1147]	5.775	3.346
L-D5	SG	1	† 1.000	[1.000, 1.000]	209	[74, 674]	33	[18, 53]	6.058	2.971
L-O	M1	25	0.500	[0.400, 0.600]	218	[100, 459]	304	[7, 1199]	0.203	0.082
L-O	DEN	1	† 1.000	[1.000, 1.000]	194	[70, 549]	5	[4, 5]	6.853	3.428
L-O	SG	1	0.130	[0.070, 0.200]	221	[89, 613]	352	[115, 973]	5.558	2.555
L-S	M1	25	0.880	[0.810, 0.940]	230	[116, 469]	93	[4, 636]	0.189	0.066
L-S	DEN	1	† 1.000	[1.000, 1.000]	188	[79, 385]	32	[20, 48]	6.237	2.777
L-S	SG	1	0.080	[0.030, 0.140]	189	[91, 455]	308	[124, 771]	6.100	2.247
AIDS	M1	25	1.000	[1.000, 1.000]	229	[95, 489]	1	[1, 1]	0.198	0.081
AIDS	DEN	1	1.000	[1.000, 1.000]	207	[67, 505]	4	[2, 6]	6.312	3.245
AIDS	SG	1	1.000	[1.000, 1.000]	197	[79, 443]	93	[48, 152]	6.077	2.744
MUT	M1	25	0.260	[0.180, 0.350]	154	[24, 575]	484	[13, 2208]	0.506	0.441
MUT	DEN	1	0.230	[0.150, 0.310]	194	[69, 381]	267	[96, 741]	6.225	3.175
MUT	SG	1	0.030	[0.000, 0.070]	208	[72, 437]	688	[134, 2342]	5.940	2.997

graphs with a variable number of vertices and edges. We implemented two CUSUM-like change detection tests as in Section IV where, for $i = 1, 2$, the statistic s_t is now given by $|\phi_i(g_t) - \mathbb{E}[\phi_i(g_t)]|$, and $\mathbb{E}[\phi_i(g_t)]$ is numerically estimated in the training phase.

In addition, we consider a further baseline implemented as a degenerate case of our method by selecting only $M = 1$ prototype and window size $n = 25$. This last baseline is introduced to show that the strength of our methodology resides also in the embedding procedure, and not only in the graph distance $d(\cdot, \cdot)$.

B. Results on IAM Graph Database

For the sake of readability, we show results for our method and baselines in two different tables, that is, Tables II and III, respectively.

In all experiments shown in Table I, there is a parameter setting achieving a detection rate (DCR) statistically larger than 0.99. Indeed, in Table II, the 95% confidence interval (95CI) of the DCR is above 0.99 (see symbol $*$ in Table II). Looking at Table II more in detail, we notice that both the window size and the number of prototypes yield higher DCR. In particular, this can be seen in L-O experiments, where all DCR estimates have disjoint 95CIs (i.e., differences are statistically significant). The same phenomenon appears also for L-D5, L-S, and mutagenicity (MUT) (e.g., symbols $\circ$ and $\bullet$ in Table II). As far as other figures of merit are concerned, we do not observe statistical evidence of any trend. Still, with the exception of a few cases in MUT, all 95CIs related to ARL0 contain the target value ARL0 $= 200$; hence, we may say that the threshold estimation described in Section V-A3 completed as expected.

Here, we limit the analysis to the proposed parameter settings for n or M, since we already reach the highest possible DCR, achieving one hundred detections out of one hundred. We believe that, by increasing the window size n the false alarms will decrease, as our method relies on the central limit theorem. Concerning the number M of prototypes, we point out that, in the current implementation of the methodology, the number of parameters to be estimated scales as M^2; accordingly, we need to increase the number of samples.

The second experimental analysis addresses the performance assessment of the three baselines of Section V-A5 on the experiments of Table I. The results reported in Table III show that, in some cases, the considered baselines achieve sound performance, which is comparable to the one shown in Table II. Comparing Table III with Table II, by intersecting the 95% confidence intervals, we notice that there is always one of the proposed methods which attains DCR that is statistically equivalent or better than the baselines (see symbol $*$ in Table II). In particular, the proposed method performs significantly better than the baselines on the MUT data set.

Finally, Table III shows that the method based on edge density ϕ_1 is significantly more accurate in terms of DCR than the one based on spectral gap ϕ_2 in almost all experiments; confidence intervals at level 95% do not intersect. The first method performed better than the degenerate one (M1) with only one prototype: see L-O and L-S. Conversely, M1 outperforms ϕ_1 on L-D5.

VI. CONCLUSION

In this paper, we proposed a methodology for detecting changes in stationarity in streams of graphs. Our approach allows to handle general families of attributed graphs and is not limited to graphs with fixed number of vertices or graphs with (partial) one-to-one correspondence between vertices at different time steps (uniquely identified vertices). The proposed methodology consists of an embedding step, which maps input graphs onto numeric vectors, bringing the change detection problem back to a more manageable setting.

We designed and tested a specific method as an instance of the proposed methodology. The embedding has been imple-

mented here as a dissimilarity space representation, hence relying on a suitable set of prototype graphs, which in our case, provide also a characterization of the nominal condition, and a dissimilarity measure between graphs, here implemented as a GED. The method then computes the Mahalanobis' distance between the mean values in two windows of embedded graphs and adopts a CUSUM-like procedure to identify changes in stationarity.

We provided theoretical results proving that, under suitable assumptions, the methodology is able to relate the significance level with which we detect changes in the dissimilarity space with a significance level that changes also occurred in the graph domain; also the vice versa has been proven. We also showed that our methodology can handle more basic, yet relevant scenarios with uniquely identified vertices.

Finally, we performed experiments on IAM graph data sets showing that the methodology can be applied both to geometric graphs (2-dimentional drawings) and graphs with categorical attributes (molecules), as instances of possible data encountered in real-world applications. Results show that the proposed method attains at least comparable (and often better) results with respect to other change detectors for graph streams.

In conclusion, we believe that the proposed methodology opens the way to designing sound change detection methods for sequences of attributed graphs with possibly time-varying topology and nonidentified vertices. In the future studies, we plan to work on real-world applications and focus on the automatic optimization of relevant parameters affecting the performance.

Appendix A
Correspondence Between Probability Spaces

Consider the measurable space $(\mathcal{D}, \mathcal{B})$ introduced in Section III-A2, and the probability space $(\mathcal{G}, \mathcal{S}, Q)$ of Section I-A.

Let us define the preimage function $\zeta^{-1}(B) := \{g \in \mathcal{G} : \zeta(g) \in B\}$, with $\zeta^{-1}(\varnothing) = \varnothing$ and consider the smallest σ-algebra $\mathcal{S}$ containing all open balls[4] $O(\rho, g)$ and preimage sets with respect to any $B \in \mathcal{B}$

$$\{O(\rho, g) | \rho > 0, g \in \mathcal{G}\} \cup \{\zeta^{-1}(B) | B \in \mathcal{B})\}$$

and a generic probability density function $Q : \mathcal{S} \to [0, 1]$ on $\mathcal{S}$. Then, we can define the function $F : \mathcal{B} \to [0, 1]$ as in Equation (2).

The triple $(\mathcal{D}, \mathcal{B}, F)$ is a probability space. The following three properties provide a proof that F is a probability measure on $(\mathcal{D}, \mathcal{B})$:

1) $F(\mathcal{D}) = Q(\zeta^{-1}(\mathcal{D})) = Q(\mathcal{G}) = 1$.
2) $F(\varnothing) = Q(\varnothing) = 0$.
3) For any countable collection $\{B_i\} \in \mathcal{B}$ of pairwise disjoint sets, $\zeta^{-1}(\cup B_i) = \cup \zeta^{-1}(B_i)$, hence the sets $\zeta^{-1}(B_i)$ are pairwise disjoint and $F(\cup B_i) = \cup Q(\zeta^{-1}(B_i)) = \cup F(B_i)$.

[4]A ball $O(\rho, g)$ is defined as a set $\{f \in \mathcal{G} : d(g, f) < \rho\}$ of all graphs $f \in \mathcal{G}$ having distance $d(f, g)$ with respect to a reference graph $g \in \mathcal{G}$ smaller than the radius $\rho > 0$.

Notice also that, indicating with $\mathrm{Im}(\zeta)$ the image set $\{\zeta(g) | g \in \mathcal{G}\} \subseteq \mathcal{D}$ of $\zeta(\cdot)$, we have $F(B) = 0$ for any $B \in \mathcal{B}$ such that $B \cap \mathrm{Im}(\zeta) = \varnothing$.

Appendix B
Fréchet Mean

Given a probability space $(\mathcal{X}, \mathcal{S}, P)$ defined on a metric space $(\mathcal{X}, d)$, we consider a random sample $\boldsymbol{x} = \{x_{t_1}, \ldots, x_{t_n}\}$.

A. Definition of Fréchet Mean and Variation

For any object $x \in \mathcal{X}$, let us define the functions $\mathcal{F}_{\boldsymbol{x}}(x) := (1/n) \sum_{t=t_1}^{t_n} d(x, x_t)^2$ and $\mathcal{F}_P(x) := \int_{\mathcal{G}} d(x, x')^2 \, dP(x')$. A Fréchet sample (population) mean is any object x attaining the minimum of the function $\mathcal{F}_{\boldsymbol{x}}(\cdot)$ ($\mathcal{F}_P(\cdot)$). We point out that the minimum might not exist in $\mathcal{X}$ and, if it does, it can be attained at multiple objects. Whenever the minimum exists and is unique, we refer to it as $\mu[\boldsymbol{x}]$ ($\mu[P]$). In addition, we define Fréchet sample (population) variation as the infimum $V_f[\boldsymbol{x}] := \inf_x \mathcal{F}_{\boldsymbol{x}}(x)$ ($V_f[P] := \inf_x \mathcal{F}_P(x)$).

B. Fréchet Mean in Euclidean Spaces

In the case of a set $\mathcal{X} \subseteq \mathbb{R}^d$ and distance $d(\cdot, \cdot) = \|\cdot - \cdot\|_2$, the space is Euclidean. First, we show that $\mu[P] = \mathbb{E}[P]$ and $\mu[\boldsymbol{x}] = \overline{\boldsymbol{x}}$, then we show that

$$\mathbb{E}[V_f[\boldsymbol{x}]] = \left(1 - \frac{1}{n}\right) V_f[P]. \tag{11}$$

1. The following equality holds:

$$\mathcal{F}_P(z) = \mathcal{F}_P(\mathbb{E}[P]) + \|\mathbb{E}[P] - z\|_2^2 \tag{12}$$

as we will show below, in Part **3**. This result proves that the minimum is attained in the expectation $z = \mathbb{E}[P]$.

2. Similarly, in the "sample" case

$$\mathcal{F}_{\boldsymbol{x}}(z) = \mathcal{F}_{\boldsymbol{x}}(\overline{\boldsymbol{x}}) + \|\overline{\boldsymbol{x}} - z\|_2^2$$

proving that $\mu[\boldsymbol{x}] = \overline{\boldsymbol{x}}$.

3. The results of previous Parts **1** and **2** are derived from the following three equalities: for any $z \in \mathbb{R}^d$

$$\|a + b\|_2^2 = \|a\|_2^2 + 2\, a^\top b + \|b\|_2^2, \quad a, b \in \mathcal{X}$$

$$\int_{\mathcal{X}} (x - \mathbb{E}[P])^\top (\mathbb{E}[P] - z) \, dP(x)$$
$$= \int_{\mathcal{X}} x^\top (\mathbb{E}[P] - z) \, dP(x) - \mathbb{E}[P]^\top (\mathbb{E}[P] - z)$$
$$= 0$$

$$\sum_t (x_t - \overline{\boldsymbol{x}})^\top (\overline{\boldsymbol{x}} - z) = \sum_t x_t^\top (\overline{\boldsymbol{x}} - z) - n\, \overline{\boldsymbol{x}}^\top (\overline{\boldsymbol{x}} - z) = 0.$$

4. Let us move on proving the second result. Notice that

$$\begin{aligned} V_f[P] &= \mathbb{E}\big[\, \|x - \mu[P]\|_2^2 \,\big] \\ &= \mathbb{E}\big[\, \|x\|_2^2 - 2y^\top \mathbb{E}[P] + \|\mu[P]\|_2^2 \,\big] \\ &= \mathbb{E}\big[\, \|x\|_2^2 \,\big] - \|\mu[P]\|_2^2 . \end{aligned}$$

5. Then

$$\begin{aligned} V_f[\boldsymbol{x}] &= \frac{1}{n}\sum_t \|x_t - \mu[\boldsymbol{x}]\|_2^2 = \frac{1}{n}\sum_t x_t^\top x_t - \mu[\boldsymbol{x}]^\top \mu[\boldsymbol{x}] \\ &= \frac{1}{n}\sum_t x_t^\top x_t - \frac{1}{n^2}\sum_{i,j} x_i^\top x_j \\ &= \left(\frac{1}{n} - \frac{1}{n^2}\right)\sum_t x_t^\top x_t - \frac{1}{n^2}\sum_{i\neq j} x_i^\top x_j. \end{aligned}$$

Now, thanks to the independence of the observations

$$\begin{aligned} \mathbb{E}[V_f[\boldsymbol{x}]] &= \left(1-\frac{1}{n}\right)\mathbb{E}[x^\top x] - \left(1-\frac{1}{n}\right)\mu[P]^\top \mu[P] \\ &= \left(1-\frac{1}{n}\right)\left(\mathbb{E}\big[\,\|x\|_2^2\,\big] - \|\mu[P]\|_2^2\right) \\ &= \left(1-\frac{1}{n}\right) V_f[P] \end{aligned}$$

which proves the thesis.

6. Finally, in this part, we prove a result that holds in general. Notice that $(1/n)\sum_t d(x_t, \mu[P])^2 \geq V_f[\boldsymbol{x}]$. Then, since $\boldsymbol{x}$ are i.i.d. so also are the $\{d(x_t, \mu[P])^2\}$. Thanks to the monotonicity of the expectation, we have

$$\mathbb{E}[V_f[\boldsymbol{x}]] \leq \frac{1}{n}\sum_t \mathbb{E}[d(x_t, \mu[P])^2] = V_f[P]. \tag{13}$$

Appendix C
Proofs of Section III-A

A. Proof of Lemma 1

Notice that the means $\mu[\boldsymbol{y}] = \overline{\boldsymbol{y}}$ and $\mu[F] = \mathbb{E}[F]$ are computed with respect to the Euclidean metric, and $d'(\cdot,\cdot)$ is deployed as statistic for the change detection test.

1. From (12), for any $z \in \mathcal{D}$, we have

$$\mathcal{F}_F(z) = \|z - \mathbb{E}[F]\|_2^2 + V_f[F].$$

2. We provide a second inequality that will be useful later. Given three graphs $g, f, r \in \mathcal{G}$, we have $d(g,f) \geq d(g,r) - d(r,f)$ from the triangular inequality (Assumption (A1)); since this holds for any prototype $r \in R$, it proves that

$$d(g,f) \geq \|\zeta(g) - \zeta(f)\|_\infty \geq M^{-\frac{1}{2}}\,\|\zeta(g) - \zeta(f)\|_2\,. \tag{14}$$

3. Exploiting the inequality (14) in Part **2**, and taking $z = \zeta(\mu[Q])$, we obtain

$$\begin{aligned} V_f[Q] &= \int_{\mathcal{G}} d(g,\mu[Q])^2\, dQ(g) = \int_{\mathcal{G}} d(g,\mu[Q])^2\, dF(\zeta(g)) \\ &\geq \int_{\mathcal{G}} M^{-1}\,\|\zeta(g) - \zeta(\mu[Q])\|_2^2\, dF(\zeta(g)) \\ &= M^{-1}\int_{\mathrm{Im}(\zeta)} \|y - \zeta(\mu[Q])\|_2^2\, dF(y). \end{aligned}$$

In Appendix A, we observed that $F(\mathcal{D}\setminus \mathrm{Im}(\zeta)) = 0$, then

$$\begin{aligned} V_f[Q] &\geq M^{-1}\int_{\mathcal{D}} \|y - \zeta(\mu[Q])\|_2^2\, dF(y) \\ &= M^{-1}\mathcal{F}_F(\zeta(\mu[Q])). \end{aligned}$$

Eventually, combining with Part **1**, we obtain the first part of the thesis for $\upsilon_0 := M\ V_f[Q] - V_f[F] \geq 0$, in fact

$$M\ V_f[Q] \geq \mathcal{F}_F(\zeta(\mu[Q])) = \|\zeta(\mu[Q]) - \mathbb{E}[F]\|_2^2 + V_f[F].$$

4. Similarly, we have

$$\mathcal{F}_{\boldsymbol{g}}(z) = V_f[\boldsymbol{y}] + \|z - \overline{\boldsymbol{y}}\|_2^2$$

and, for $\upsilon_{\boldsymbol{g}} := M\ V_f[\boldsymbol{g}] - V_f[\boldsymbol{y}]$,

$$\|\zeta(\mu[\boldsymbol{g}]) - \overline{\boldsymbol{y}}\|_2^2 \leq \upsilon_{\boldsymbol{g}}.$$

By exploiting (11) and (13), and the monotonicity of the expected value

$$\mathbb{E}[\upsilon_{\boldsymbol{g}}] \leq M\ V_f[Q] - (1-1/n)\,V_f[F] =: \upsilon_n.$$

As final remarks, we point out that here we left the dependence from n on purpose, but an independent bound can be easily found, e.g., $\upsilon_2 = M\ V_f[Q] - (1/2)\,V_f[F]$.

5. The random variable $\|\overline{\boldsymbol{y}} - \zeta(\mu[\boldsymbol{g}])\|_2^2$ is nonnegative. As such we can apply Theorem 1, Section V.4 in [51] (sometimes called *Markov's inequality*) and obtain, for each $\delta > 0$,

$$\mathbb{P}\big(\,\|\overline{\boldsymbol{y}} - \zeta(\mu[\boldsymbol{g}])\|_2^2 \geq \delta\big) \leq \frac{\mathbb{E}\big[\,\|\overline{\boldsymbol{y}} - \zeta(\mu[\boldsymbol{g}])\|_2^2\,\big]}{\delta}.$$

Now, from the previous Part **4** we conclude that

$$\mathbb{P}\big(\,\|\overline{\boldsymbol{y}} - \zeta(\mu[\boldsymbol{g}])\|_2^2 \geq \delta\big) \leq \frac{\upsilon_2}{\delta}.$$

B. Proof of Lemma 2

For convenience, let us define the quantities

$$\begin{aligned} A &:= \{x_0 \in \mathcal{X} : d_1(x_0) \leq u(d_2(x_0))\} \\ \rho(-|-) &:= \mathbb{P}(d_1 \leq u(\gamma)\ |\ d_2 \leq \gamma, x \in A) \\ \rho(-|+) &:= \mathbb{P}(d_1 \leq u(\gamma)\ |\ d_2 > \gamma, x \in A). \end{aligned}$$

1. By the law of total probability, and for any $\gamma \geq 0$,

$$\begin{aligned} \mathbb{P}(d_1(x) \leq u(\gamma)) &= \mathbb{P}(d_1(x) \leq u(\gamma)|x \in A)\mathbb{P}(x \in A) \\ &\quad + \mathbb{P}(d_1(x) \leq u(\gamma)|x \notin A)\mathbb{P}(x \notin A). \end{aligned}$$

Lower bounding the second addendum with zero and by hypothesis,

$$\mathbb{P}(d_1(x) \leq u(\gamma)) \geq \mathbb{P}(d_1(x) \leq u(\gamma)|x \in A)\cdot p.$$

2. Notice that $\mathbb{P}(d_1(x) \leq u(\gamma)|d_2(x) = \gamma, x \in A) = 1$ for all $\gamma \geq 0$, thanks to the fact that $x \in A$; hence, we have $\rho(-|-) = 1$. Applying again the law of total probabilities,

$$\begin{aligned} &\mathbb{P}(d_1(x) \leq u(\gamma)|x \in A) \\ &\quad = \rho(-|-)\Phi_2(\gamma) + \rho(-|+)(1-\Phi_2(\gamma)) \geq 1\cdot\Phi_2(\gamma). \end{aligned}$$

Combining with the above Part **1**, we have

$$\Phi_1(u(\gamma)) = \mathbb{P}(d_1(x) \leq u(\gamma)) \geq p\cdot\Phi_2(\gamma).$$

3. A final remark is that Lemma 2 proves also that, if $\ell : \mathbb{R}_+ \to \mathbb{R}_+$ is bijective and increasing providing a lower bound, then for any $\gamma > 0$ and $p > 0$

$$\mathbb{P}(d_1(x) \geq \ell(d_2(x))) \geq p \Rightarrow \Phi_1(\gamma) \leq \frac{1}{p}\cdot\Phi_2(\ell^{-1}(\gamma)) \tag{15}$$

In fact $d_1 \geq \ell(d_2)$ if and only if $d_2 \leq \ell^{-1}(d_1)$, therefore, we obtain the result by applying the lemma to $u(\cdot) = \ell^{-1}(\cdot)$ and inverting the roles of d_1 and d_2.

C. Proof of Proposition 1

By Assumption (A2), we can consider the following function of $\boldsymbol{g}$ and their respective CDFs:

$$\begin{aligned} d(\boldsymbol{g}) &:= d(\mu[\boldsymbol{g}], \mu[Q]), \quad \Psi(\cdot) \\ d'(\boldsymbol{g}) &:= d'(\zeta(\mu[\boldsymbol{g}]), \zeta(\mu[Q])), \quad \Phi(\cdot) \\ d'_0(\boldsymbol{g}) &:= d'(\overline{\boldsymbol{y}}, \mathbb{E}[F]), \quad \Upsilon(\cdot) \end{aligned}$$

recalling that $\boldsymbol{y} = (\dots, \zeta(g_i), \dots)^\top$.

1. From triangular inequality [Assumption (A1)], we have

$$\left|d'_0(\boldsymbol{g}) - d'(\boldsymbol{g})\right| \leq d'(\zeta(\mu[\boldsymbol{g}]), \overline{\boldsymbol{y}}) + d'(\mathbb{E}[F], \zeta(\mu[Q])).$$

By the equivalence of the 1- and 2-norm in $\mathbb{R}^2$,

$$\begin{aligned} &\left(d'_0(\boldsymbol{g}) - d'(\boldsymbol{g})\right)^2 \\ &\qquad \leq \frac{1}{2} \cdot d'(\zeta(\mu[\boldsymbol{g}]), \overline{\boldsymbol{y}})^2 + \frac{1}{2} \cdot d'(\mathbb{E}[F], \zeta(\mu[Q]))^2. \end{aligned}$$

Now, since $d'(\cdot, \cdot)$ is induced by a norm [Assumption (A1)], we can deploy the equivalence of any pair of norms in $\mathbb{R}^M$. Letting m be a constant for relating the distance $d'(\cdot, \cdot)$ with the Euclidean one, and applying Lemma 1, we obtain

$$\begin{aligned} &\left(d'_0(\boldsymbol{g}) - d'(\boldsymbol{g})\right)^2 \\ &\quad \leq \frac{m^2}{2}\left(\|\zeta(\mu[\boldsymbol{g}]) - \overline{\boldsymbol{y}}\|_2^2 + \|\mathbb{E}[F] - \zeta(\mu[Q])\|_2^2\right) \\ &\quad \leq \frac{m^2}{2} \cdot \left(\upsilon_2 + \|\zeta(\mu[\boldsymbol{g}]) - \overline{\boldsymbol{y}}\|_2^2\right). \end{aligned}$$

By exploiting again Lemma 1, for any $\delta > 0$, we have

$$\begin{aligned} \mathbb{P}\left(\left(d'_0(\boldsymbol{g}) - d'(\boldsymbol{g})\right)^2 \leq \frac{m^2}{2} \cdot (\upsilon_2 + \delta)\right) \\ \geq \mathbb{P}\left(\|\zeta(\mu[\boldsymbol{g}]) - \overline{\boldsymbol{y}}\|_2^2 \leq \delta\right) \geq 1 - \frac{\upsilon_2}{\delta}. \end{aligned}$$

So, with $b(\delta) = ((m^2/2)(\upsilon_2 + \delta))^{1/2}$ and $p(\delta) = 1 - (\upsilon_2/\delta)$, we have the following estimates:

$$\begin{aligned} \mathbb{P}(d'_0(\boldsymbol{g}) \leq d'(\boldsymbol{g}) + b(\delta)) &\geq p(\delta) \\ \mathbb{P}(d'(\boldsymbol{g}) \leq d'_0(\boldsymbol{g}) + b(\delta)) &\geq p(\delta). \end{aligned} \tag{16}$$

2. With Assumption (A3), we can exploit Lemma 2 by setting $u(x) = C\,x$, $d_1(\boldsymbol{g}) = d(\boldsymbol{g})$ and $d_2(\boldsymbol{g}) = d'(\boldsymbol{g})$. Here $\mathbb{P}(d_1 \leq u(d_2)) = 1$, hence we obtain that $\Psi(\gamma) \geq \Phi(\gamma/C)$. Accordingly, employing also (15), we have

$$\Phi\left(\frac{\gamma}{C}\right) \leq \Psi(\gamma) \leq \Phi\left(\frac{\gamma}{c}\right).$$

3. Consider this time (16) with $d_1(\boldsymbol{g}) = d'_0(\boldsymbol{g})$ and $d_2(\boldsymbol{g}) = d'(\boldsymbol{g})$. By applying Lemma 2 twice, with $u(x) = x + b(\delta)$ and $\ell(x) = x - b(\delta)$, we obtain

$$\begin{aligned} \Phi(\gamma) &\geq p_\delta \cdot \Upsilon(\gamma - b_\delta) \\ \Phi(\gamma) &\leq \frac{1}{p_\delta} \cdot \Upsilon(\gamma + b_\delta). \end{aligned}$$

4. Combining previous parts, Parts **2** and **3**, we obtain

$$p_\delta \Upsilon\left(\frac{\gamma}{C} - b_\delta\right) \leq \Psi(\gamma) \leq \frac{1}{p_\delta} \Upsilon\left(\frac{\gamma}{c} + b_\delta\right).$$

APPENDIX D
PROOFS OF SECTIONS IV AND IV-B

A. Proof of Lemma 3

Recall that $\|x\|_2^2 \geq \lambda_M x^\top \Sigma^{-1} x$, where λ_M is the smallest eigenvalue of the covariance matrix Σ. By the triangular inequality we have (14), which leads to $c = (\lambda_M/M)^{1/2}$.

B. Instance of Graph Alignment Space

A graph alignment space is a pair $(\mathcal{G}, d)$. According to what defined in Section IV-B, we explicit an attribute kernel determining a graph alignment distance d acting like d_F.

We consider $\mathcal{A} = \mathcal{V} \times [0, 1]$ as attribute set. Following [21, Ex.3.4], for any pair $(x, \upsilon) \in \mathcal{A}$, we define the feature map $\boldsymbol{\Phi} : (x, \upsilon) \mapsto (x, e_\upsilon)$, where $e_\upsilon \in \mathbb{R}^N$ is zero everywhere except for the position υ where it assumes the fixed value ν. The attribute kernel $k : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ is then defined accordingly. We consider a vertex disabled if there are no incoming or outgoing edges, but it has always a label of the form $(0, \upsilon)$.

It might happen that the "cost" of matching vertices is overcome by the improvement of a better topology alignment. This is something which is not contemplated in the distance d_F. For avoiding this, it suffices to set the constant ν sufficiently large, so that the trivial identity alignment is always the optimal alignment (i.e., vertices are uniquely identified). Setting $\nu = N^2$ ensures this behavior and yields an alignment distance $d(\cdot, \cdot)$ acting like $d_F(\cdot, \cdot)$.

C. Proof of Lemma 4

1. For every prototye set R, the dissimilarity matrix $D(R, R) := [d_F(r_i, r_j)]_{i,j}$ is Euclidean, since it has been generated by using the Euclidean distance. In general, the dimension of $\mathbb{R}^k$ is $k = N^2$, although it can be lower. We remind that N is the number of vertices assumed for the input graphs. We use the notation $x_1, \dots, x_M \in \mathbb{R}^k$ to denote the prototypes $r_1, \dots, r_M \in R$ with respect to the classical scaling process and X is obtained by stacking $[x_1|\dots|x_M]$ as columns (see [39, Sec. 3.5.1]). The minimal number of prototypes for representing $\mathbb{R}^k$ is $M = k + 1$, and this holds true only when the matrix X is full rank. For obtaining the embedding z of a generic graph $g \in \mathcal{G} \setminus R$, we compute $y^{*2} := [d_F(g, r_1)^2, \dots, d_F(g, r_M)^2]^\top$ and we can solve the following linear system with respect to z

$$2\,X^\top z = -(J\,y^{*2} - J\,D^{*2}\,\mathbf{1})$$

where $J = I - (1/n)\mathbf{1}\mathbf{1}^\top$ is the centering matrix, I is the identity matrix, and D^{*2} is the componentwise square of $D(R, R)$ (see [39, Cor. 3.7]). The solution is unique, provided that the rank of X is k.

2. The differences $\delta z = z_1 - z_2$ and $\delta y^{*2} = y_1^{*2} - y_2^{*2}$ are related to the equation $2\,XX^\top \delta z = -XJ\,\delta y^{*2}$.

Being $\|a\|_A^2 = a^\top A\,a$ we have that $4\|\delta z\|_{(XX^\top)^2} = 4\left\|XX^\top z\right\|_I = \left\|XJ\,\delta y^{*2}\right\|_I$ and

$$\frac{\|\delta z\|_I}{\left\|XJ\,\delta y^{*2}\right\|_{\Sigma^{-1}}} = \frac{\|\delta z\|_I}{4\,\|\delta z\|_{(XX^\top)^2}} \cdot \frac{\left\|XJ\,\delta y^{*2}\right\|_I}{\left\|XJ\,\delta y^{*2}\right\|_{\Sigma^{-1}}}.$$

We want to show that $c \leq \|\delta z\|_I / \left\|XJ\,\delta y^{*2}\right\|_{\Sigma^{-1}} \leq C$, in fact $\|\delta z\|_I = \|\delta z\|_2 = d_F(g_1, g_2)$, and $\left\|XJ\,\delta y^{*2}\right\|_{\Sigma^{-1}} = d_\Sigma(XJ\,y_1^{*2}, XJ\,y_2^{*2})$.

3. Applying [52, Th. 1], we can bound $\|a\|_A / \|a\|_B$ in terms of the values β for which there exists nonnull $q \in \mathbb{R}^M$ such that $(A - \beta B)\,q = \mathbf{0}$. For the pair $(I, (XX^\top)^2)$, the values β are provided by the square of the inverse eigenvalues $\lambda_1(XX^\top), \ldots, \lambda_k(XX^\top)$ of $XX^\top$. The eigenvalues are reported in descending order. For what concerns (I, Σ^{-1}), instead, we see that the values β corresponds to the eigenvalues $\lambda_1(\Sigma), \ldots, \lambda_k(\Sigma)$ of Σ. In the end, the positive constants c, C are

$$c^2 = \frac{\lambda_k(\Sigma)}{4\,\lambda_1(XX^\top)}, \quad C^2 = \frac{\lambda_1(\Sigma)}{4\,\lambda_k(XX^\top)}.$$

The nonsingularity of $XX^\top$ makes the above fractions feasible.

A final comment on how to exploit Lemma 4 in Proposition 1 is reported in Appendix D–D.

D. Proposition 1 for the Special Case of Section IV-B

When considering identified vertices, while Lemma 2 is still valid, Lemma 1 has to be adapted. Despite the new embedding $\zeta_0(g) = XJ\zeta(g)^{*2}$ is slightly different from the original one in (1), $\zeta_0(\cdot)$ and $\mathcal{D}_0$ can be treated in a similar way. In particular, a result equivalent to Lemma 1 can be proved for $\zeta_0(\cdot)$ by solely adapting Part **2** of the proof shown in Appendixes C–A

$$\begin{aligned} d_F(g_1, g_2) = \|\delta z\|_2 \geq \frac{\left\|XX^\top \delta z\right\|_2}{\lambda_1(XX^\top)} &= \frac{\left\|XJ\,\delta y^{*2}\right\|_2}{\lambda_1(XX^\top)} \\ &= (\lambda_1(XX^\top))^{-1}\,\|u_1 - u_2\|_2\,. \end{aligned}$$

Paying attention in substituting the map $\zeta(\cdot)$ with $\zeta_0(\cdot)$, the rest of the proof holds with $M = \lambda_1(XX^\top)^2$ thus obtaining the Claims (C1) and (C2), with a final result similar to the one of Proposition 1.

REFERENCES

[1] G. Ditzler, M. Roveri, C. Alippi, and R. Polikar, "Learning in nonstationary environments: A survey," *IEEE Comput. Intell. Mag.*, vol. 10, no. 4, pp. 12–25, Apr. 2015.

[2] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Comput. Surv.*, vol. 46, no. 4, pp. 44:1–44:37, Apr. 2014.

[3] R. Elwell and R. Polikar, "Incremental learning of concept drift in nonstationary environments," *IEEE Trans. Neural Netw.*, vol. 22, no. 10, pp. 1517–1531, Oct. 2011.

[4] C. Alippi and M. Roveri, "Just-in-time adaptive classifiers—Part I: Detecting nonstationary changes," *IEEE Trans. Neural Netw.*, vol. 19, no. 7, pp. 1145–1153, Jul. 2008.

[5] C. Alippi, G. Boracchi, and M. Roveri, "Just in time classifiers: Managing the slow drift case," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jun. 2009, pp. 114–120.

[6] D. M. Hawkins, P. Qiu, and C. W. Kang, "The changepoint model for statistical process control," *J. Quality Technol.*, vol. 35, no. 4, pp. 355–366, 2003.

[7] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory Application*, vol. 104. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[8] C. Alippi and M. Roveri, "An adaptive cusum-based test for signal change detection," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2006, pp. 1–4.

[9] G. J. Ross and N. M. Adams, "Two nonparametric control charts for detecting arbitrary distribution changes," *J. Quality Technol.*, vol. 44, no. 2, pp. 102–116, 2012.

[10] K. D. Zamba and D. M. Hawkins, "A multivariate change-point model for statistical process control," *Technometrics*, vol. 48, no. 4, pp. 539–549, 2006.

[11] V. Golosnoy, S. Ragulin, and W. Schmid, "Multivariate CUSUM chart: Properties and enhancements," *AStA Adv. Stat. Anal.*, vol. 93, no. 3, pp. 263–279, 2009.

[12] L. Livi and C. Alippi, "One-class classifiers based on entropic spanning graphs," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 12, pp. 2846–2858, Dec. 2017.

[13] R. Hamon, P. Borgnat, P. Flandrin, and C. Robardet, "Extraction of temporal network structures from graph-based signals," *IEEE Trans. Signal Inf. Process. Over Netw.*, vol. 2, no. 2, pp. 215–226, Jun. 2016.

[14] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.

[15] P. Holme and J. Saramäki, "Temporal networks," *Phys. Rep.*, vol. 519, no. 3, pp. 97–125, 2012.

[16] P. Holme, "Modern temporal network theory: A colloquium," *Eur. Phys. J. B*, vol. 88, no. 9, p. 234, 2015.

[17] N. Masuda and R. Lambiotte, *A Guide to Temporal Networks* (Series on Complexity Science). Singapore: World Scientific, 2016.

[18] J. D. Wilson, N. T. Stevens, and W. H. Woodall. (2016). "Modeling and detecting change in temporal networks via a dynamic degree corrected stochastic block model." [Online]. Available: https://arxiv.org/abs/1605.04049

[19] I. Barnett and J.-P. Onnela, "Change point detection in correlation networks," *Sci. Rep.*, vol. 6, Jan. 2016, Art. no. 18893.

[20] C. Masoller *et al.*, "Quantifying sudden changes in dynamical systems using symbolic networks," *New J. Phys.*, vol. 17, no. 2, p. 023068, 2015.

[21] B. J. Jain, "On the geometry of graph spaces," *Discrete Appl. Math.*, vol. 214, pp. 126–144, Dec. 2016.

[22] B. J. Jain, "Statistical graph space analysis," *Pattern Recognit.*, vol. 60, pp. 802–812, Dec. 2016.

[23] L. Livi and A. Rizzi, "The graph matching problem," *Pattern Anal. Appl.*, vol. 16, no. 3, pp. 253–283, 2013.

[24] P. Foggia, G. Percannella, and M. Vento, "Graph matching and learning in pattern recognition in the last 10 years," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 28, no. 1, p. 1450001, 2014.

[25] F. Emmert-Streib, M. Dehmer, and Y. Shi, "Fifty years of graph matching, network alignment and network comparison," *Inf. Sci.*, vol. 346, pp. 180–197, Jun. 2016.

[26] M. Roy, S. Schmid, and G. Tredan, "Modeling and measuring graph similarity: The case for centrality distance," in *Proc. 10th ACM Int. Workshop Found. Mobile Comput.*, 2014, pp. 47–52.

[27] H. Bunke and G. Allermann, "Inexact graph matching for structural pattern recognition," *Pattern Recognit. Lett.*, vol. 1, no. 4, pp. 245–253, 1983.

[28] G. Da San Martino, N. Navarin, and A. Sperduti, "Tree-based kernel for graphs with continuous attributes," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2017.2705694.

[29] H. J. Qiu and E. R. Hancock, "Graph matching and clustering using spectral partitions," *Pattern Recognit.*, vol. 39, no. 1, pp. 22–34, Jan. 2006.

[30] L. Bai, L. Rossi, A. Torsello, and E. R. Hancock, "A quantum Jensen–Shannon graph kernel for unattributed graphs," *Pattern Recognit.*, vol. 48, no. 2, pp. 344–355, 2015.

[31] F. Costa and K. De Grave, "Fast neighborhood subgraph pairwise distance kernel," in *Proc. 26th Int. Conf. Mach. Learn.*, Jul. 2010, pp. 255–262.

[32] S. Fankhauser, K. Riesen, and H. Bunke, "Speeding up graph edit distance computation through fast bipartite matching," in *Proc. Int. Workshop Graph-Based Representations Pattern Recognit.*, 2011, pp. 102–111.

[33] A. Fischer, C. Y. Suen, V. Frinken, K. Riesen, and H. Bunke, "Approximation of graph edit distance based on Hausdorff matching," *Pattern Recognit.*, vol. 48, no. 2, pp. 331–343, 2015.

[34] S. Bougleux, L. Brun, V. Carletti, P. Foggia, B. Gaüzère, and M. Vento, "Graph edit distance as a quadratic assignment problem," *Pattern Recognit. Lett.*, vol. 87, pp. 38–46, Feb. 2017.

[35] S. Ranshous, S. Shen, D. Koutra, S. Harenberg, C. Faloutsos, and N. F. Samatova, "Anomaly detection in dynamic networks: A survey," *Wiley Interdiscipl. Rev., Comput. Stat.*, vol. 7, no. 3, pp. 223–247, 2015.

[36] C. Rossant *et al.*, "Spike sorting for large, dense electrode arrays," *Nature Neurosci.*, vol. 19, no. 4, pp. 634–641, Apr. 2016.
[37] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: A Survey," *Data Mining Knowl. Discovery*, vol. 29, no. 3, pp. 626–688, 2015.
[38] L. Peel and A. Clauset, "Detecting change points in the large-scale structure of evolving networks," in *Proc. 29th Conf. Artif. Intell.*, 2015, pp. 2914–2920.
[39] E. Pękalska and R. P. W. Duin, *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*. Singapore: World Scientific, 2005.
[40] M. Fréchet, "Les éléments aléatoires de nature quelconque dans un espace distancié," *Ann. Inst. Henri Poincaré*, vol. 10, no. 4, pp. 215–310, 1948.
[41] K. Riesen and H. Bunke, *Graph Classification and Clustering Based on Vector Space Embedding*. Singapore: World Scientific, 2010.
[42] K. Riesen, M. Neuhaus, and H. Bunke, "Graph embedding in vector spaces by means of prototype selection," in *Proc. Int. Workshop Graph-Based Representations Pattern Recognit.*, 2007, pp. 383–393.
[43] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, nos. 1–2, pp. 100–115, 1954.
[44] B. F. J. Manly and D. Mackenzie, "A cumulative sum type of method for environmental monitoring," *Environmetrics*, vol. 11, no. 2, pp. 151–166, 2000.
[45] M. Frisén, "Statistical surveillance. Optimality and methods," *Int. Stat. Rev.*, vol. 71, no. 2, pp. 403–434, 2003.
[46] D. Zambon, L. Livi, and C. Alippi. (2017). *CDG: Change Detection on Graph Streams*. Accessed: Feb. 2, 2018. [Online]. Available: http://www.inf.usi.ch/phd/zambon/#cdg
[47] K. Riesen and H. Bunke, "IAM graph database repository for graph based pattern recognition and machine learning," in *Structural, Syntactic, and Statistical Pattern Recognition*, N. da Vitoria Lobo *et al.*, Eds. Berlin, Germany: Springer, 2008, pp. 287–297.
[48] K. Riesen, S. Emmenegger, and H. Bunke, "A novel software toolkit for graph edit distance computation," in *Proc. Int. Workshop Graph-Based Representations Pattern Recognit.*, 2013, pp. 142–151.
[49] D. Zambon, L. Livi, and C. Alippi, "Detecting changes in sequences of attributed graphs," in *Proc. IEEE Symp. Series Comput. Intell.*, Honolulu, HI, USA, Nov. 2017, pp. 1835–1841.
[50] F. R. K. Chung, *Spectral Graph Theory*, vol. 92. Providence, RI, USA: AMS, 1994.
[51] G. G. Roussas, *A Course in Mathematical Statistics*. Orlando, FL, USA: Academic, 1997.
[52] D. R. Jensen, "Bounds on Mahalanobis norms and their applications," *Linear Algebra Appl.*, vol. 264, pp. 127–139, Oct. 1997.

Daniele Zambon (S'17) received the M.Sc. degree in mathematics from the Università degli Studi di Milano, Milan, Italy, in 2016.

He is currently a Ph.D. Student with the Faculty of Informatics, Università della Svizzera italiana, Lugano, Switzerland. His current research interests include graph representation, statistical processing of graph streams, change and anomaly detection.

Cesare Alippi (F'06) received the M.Sc. degree *(cum laude)* in electronic engineering and Ph.D. degree from the Politecnico di Milano, Milan, Italy, in 1990 and 1995, respectively.

He was a Visiting Researcher at UCL (U.K.), MIS (USA), ESPCI (F), CASIA (RC), and A*STAR (SIN). He is currently a Full Professor of information processing systems with the Politecnico di Milano, and also a Full Professor of cyber-phsical and embedded systems with the Università della Svizzera italiana, Lugano, Switzerland. He has authored a monograph with Springer on "Intelligence for embedded systems" in 2014, and co-authored more than 200 papers in international journals and conference proceedings, and he holds five patents. His current research interests include adaptation and learning in nonstationary environments and intelligence for embedded systems.

Dr. Alippi is a Distinguished Lecturer of the IEEE CIS, and a member of the Board of Governors of INNS. He was a recipient of the IEEE Instrumentation and Measurement Society Young Engineer Award in 2004, the IBM Faculty Award in 2013, the 2016 IEEE TNNLS Outstanding Paper Award, and the 2016 INNS Gabor Award. He is the Vice President Education of IEEE CIS, and an Associate Editor of the IEEE Computational Intelligence Magazine. He was an Associate Editor of the IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENTS, IEEE TRANSACTIONS ON NEURAL NETWORKS, and a member and the Chair of other IEEE committees. Among the others, he was the General Chair of the International Joint Conference on Neural Networks in 2012, the Program Chair in 2014, the Co-Chair in 2011. He was the General Chair of the IEEE Symposium Series on Computational Intelligence 2014.

Lorenzo Livi (M'14) received the B.Sc. degree and M.Sc. degree from the Department of Computer Science, Sapienza University of Rome, Rome, Italy, in 2007 and 2010, respectively, and the Ph.D. degree from the Department of Information Engineering, Electronics, and Telecommunications, Sapienza University of Rome, in 2014. He has been with the ICT industry during his studies. From 2014 to 2016, he was a Post-Doctoral Fellow at Ryerson University, Toronto, ON, Canada. In 2016, he was a Post-Doctoral Fellow at the Politecnico di Milano, Milan, Italy, and Università della Svizzera italiana, Lugano, Switzerland. He is currently a Lecturer (Assistant Professor) in data analytics with the Department of Computer Science at the University of Exeter, Exeter, U.K. His current research interests include computational intelligence methods, time-series analysis, and complex dynamical systems, with focused applications in systems biology and neuroscience. Dr. Livi is a member of the Editorial Board of Applied Soft Computing (Elsevier) and a regular reviewer for several international journals, including the IEEE TRANSACTIONS, *Information Sciences, and Neural Networks (Elsevier).*