A graph representation based on fluid diffusion model for multimodal data analysis: theoretical aspects and enhanced community detection

Andrea Marinoni, Christian Jutten, Mark Girolami

Abstract—Representing data by means of graph structures identifies one of the most valid approach to extract information in several data analysis applications. This is especially true when multimodal datasets are investigated, as records collected by means of diverse sensing strategies are taken into account and explored. Nevertheless, classic graph signal processing is based on a model for information propagation that is configured according to heat diffusion mechanism. This system provides several constraints and assumptions on the data properties that might be not valid for multimodal data analysis, especially when large scale datasets collected from heterogeneous sources are considered, so that the accuracy and robustness of the outcomes might be severely jeopardized. In this paper, we introduce a novel model for graph definition based on fluid diffusion. The proposed approach improves the ability of graph-based data analysis to take into account several issues of modern data analysis in operational scenarios, so to provide a platform for precise, versatile, and efficient understanding of the phenomena underlying the records under exam, and to fully exploit the potential provided by the diversity of the records in obtaining a thorough characterization of the data and their significance. In this work, we focus our attention to using this fluid diffusion model to drive a community detection scheme, i.e., to divide multimodal datasets into many groups according to similarity among nodes in an unsupervised fashion. Experimental results achieved by testing real multimodal datasets in diverse application scenarios show that our method is able to strongly outperform state-of-the-art schemes for community detection in multimodal data analysis.

Index Terms—Multimodal data analysis, fluid diffusion, information propagation, community detection, clustering.

I. INTRODUCTION

Graph-based data analysis identifies one of the most effective tools for modern data analysis, as it allows the investigation of complex datasets by means of a compact representation of mathematical manifolds, groups and varieties across different domains [1], [2]. In particular, the samples in the considered dataset are typically represented as nodes in the graph structure, whilst the edges are meant to model and quantify the similarity between each pair of nodes, according to a given informativity criterion and/or distance metric [1]–[6]. In this way, it is possible to obtain flexible and efficient investigation of complex data acquired by multiple acquisition strategies (or modalities), i.e., multimodal datasets. These properties are particularly interesting when semisupervised and unsupervised characterization of the records is required, e.g., clustering, outlier detection, ranking, label propagation [2], [7].

In technical literature, several strategies have been proposed to adequately model the graph structure and topology, from geometrical to statistical analysis, to probabilistic description of the similarities among samples (see for instance [7]–[12]). In fact, these frameworks assume that the information would propagate across the graph according to the heat diffusion mechanism [13], [14]. Specifically, assuming that the data distribution would be uniform across the samples and smooth in the feature space, it is possible to consider the connections as transition probabilities in larger Markov models, of which the nodes represent the states. In this analogy, the weight that each edge of the graph can be associated with is then computed in terms of affinity between the nodes according to a dynamical system considered at equilibrium, which results in the ability to use Euclidean distances to quantify the similarities among samples [13], [14].

This idea represents the bedrock on which the most advanced data analytics methods - from nonlinear optimization to eigenanalysis, from inference to regression, from geometrical deep learning to motif networks - rely on [1], [2], [7], [8], [15], [16]. Indeed, this assumption has been proven to be successful in a number of applications, and numerous artificial intelligence schemes take advantage of this to design the main characterization of decision making procedures. Nonetheless, it is also true that this approach might fail in detailing the observations and records acquired for the most modern challenges in data analysis [17], [18]. In fact, when investigating large scale datasets that are collected by heterogeneous modalities, the records can be affected by multiple and diverse sources of noise, leading to incomplete, non-uniform, and/or unbalanced data to be analyzed [17], [19]. Moreover, the data would show different degrees of reliability, so that information ends up being extracted by datasets showing corrupted and/or missing records. Finally, the efficiency and the explainability of the data analysis must be addressed, so that near real time applications and thorough interpretation of the data can be carried out. This situation is typically not addressed by the graph signal processing methods in technical literature, as these architectures might not be able to adequately extract information from multimodal datasets, especially in an unsu-
To give an intuitive understanding of these points, let us consider a practical example. Let us say that we are interested in studying genes within genome-wide association studies in order to determine specific markers that lead to a disease. In some cases, this is simple, with one DNA marker resulting in the expression of the disease. Other diseases may have multiple genes that together result in the expression of a disease. When studying large volumes of genes from sources taken globally (e.g., from nations as globally diverse as Finland, Kenya, and Japan) the common mistake made by learning methods is to classify at a level that may group the results into three classes, corresponding to the three nations. In reality, there are other markers (less ‘obvious’ to the machine) that could correspond to the same disease expressed in genes that are common across all three classes. This population stratification leads to differences in gene frequencies as a result of systematic differences in ancestry rather than association of genes with disease [21].

To address these points, in this work we propose the consideration of a new model for information propagation across graphs. Specifically, we introduce a fluid diffusion model to shape the graph design, with special focus to the topology and connectivity of the data structure [22], [23]. In this way, the global and local interactions among samples and records are taken into account in terms of tensor representation, which can be expressed as permeability, diffusivity and flow velocity across the graph. This representation allows one to take into account a large variety of data characteristics, so to ensure versatility and applicability of this approach in several operational scenarios. By taking advantage of this novel data representation, we provide an efficient method for community detection that can be easily implemented in terms of spectral representation, ensuring versatility and applicability of this approach in several domains. By taking advantage of this novel data representation, we provide an efficient method for community detection that can be easily implemented in terms of spectral representation, ensuring versatility and applicability of this approach in several domains.

The main contributions of this paper can be then summarized as follows:

- a new paradigm to model information propagation - based on fluid diffusion - on graph structures which is able to grasp global and local scale interactions and patterns induced by multimodal datasets;
- the analysis of the proposed fluid diffusion system in terms of eigenvectors of the flow velocity matrix that can be employed to characterize the dependency among samples and the relevance of the features the considered dataset consists of. In this way, an effective understanding of the graph can be carried out in terms of eigenanalysis, enabling valid characterization of the data structures and topologies;
- an efficient method for non-overlapping community detection, taking advantage of the eigenanalysis of the flow velocity matrix used to describe the graph connections.

The paper is organized as follows. Section III introduces the theoretical aspects of classic graph representation of datasets based on the heat diffusion model. It continues with the motivation of the proposed novel graph representation based on fluid diffusion, and its main properties. Section IV reports a thorough overview of the main works introduced in technical literature for the application task used as test case in this paper - community detection - as well as the proposed method for community detection based on the novel fluid graph representation of multimodal datasets. Section V reports the performance results obtained over three multimodal datasets, as well as heuristic confirmation of the motivation and validity of the proposed fluid graph representation. Finally, Section VI delivers our final remarks and some ideas on future research.

II. FLUID GRAPH REPRESENTATION

In this Section, we introduce the main motivations for the novel graph representation for multimodal data analysis we proposed in this work. First, the connection between classic graph representation and heat diffusion is summarized. Then, the main issues for classic graph representation are presented, leading to the motivation and the description of the graph representation based on fluid diffusion that we introduce in this paper.

A. Classic graph representation and heat diffusion

Graph representation of data manifolds is a valuable tool in extracting information from records, understanding their interactions, and providing a thorough interpretation of their semantics. Indeed, graph-based signal processing has enabled exploiting data structure and relational priors, improving data and computational efficiency, and enhancing model interpretability in various domains [1], [2].

The structure and meaning of the edges and nodes, as defined within graph representation, affects the accuracy and reliability of any information derived from it [2]. [24]–[26]. In fact, graph representation identifies a favorable trade-off between simplicity and explainability of the relationships between the samples in the dataset. The similarity and interactions among samples are represented by means of the weights of the edges of the graphs. The edge weight is then typically computed as function of the proximity of the corresponding data points in the feature space induced by the records collected in the considered dataset. Hence, a connection between two samples in the dataset could be considered stronger as the proximity of their corresponding feature vectors increases [13].

Characterizing the complex geometry of the data is therefore crucial to obtain an accurate graph representation and therefore a reliable understanding of the interactions among samples. To this aim, combining the main properties of random walks and spectral analysis is a proven approach in finding relevant structures in complex manifolds, enabling the detection and classification of thematic clusters within the data [8], [14], [27]. Indeed, using the eigenfunctions of a Markov matrix defining a random walk on the data can help in achieving a new description of data sets, as well as in providing a thorough interpretation of the similarity modeled by the edge weights [14], [27]. To embed these samples in a Euclidean space, these quantities can be associated with transition mechanisms described in terms of diffusion processes.
Further, processing the higher order moments of the Markov matrix this strategy aims to connect the spectral properties of the diffusion process to the geometrical characteristics of the dataset.

Specifically, let $X = \{x_i\}_{i=1,...,N}$, $x_i \in \mathbb{R}^n$ be the considered dataset consisting of $N$ samples characterized by $n$ features. In general, it is possible to translate $X$ into a graph structure consisting of nodes and edges connecting them. Specifically, the $i$-th node identifies the sample $x_i$ in the dataset $X$. On the other hand, the weight of the edge connecting node $i$ to node $j$ is computed according to a function (or kernel) $\eta(x_i, x_j)$ of the features associated with the considered samples. In the classic derivation of graph structure, the goal of the metric $\eta$ is to capture the characteristics of the local geometry of the given dataset. It is then possible to construct a Markov matrix associated with $X$ that can describe the local geometry of the dataset by summarizing the node-to-node similarities. In other terms, the $(i, j)$ element of the Markov matrix is defined as probability of transition in one time step from node $i$ to node $j$ in the graph. As such, the $(i, j)$ element of the Markov matrix is also proportional to the edge weight $\eta(x_i, x_j)$. Moreover, it is possible to retrieve the transition probability in $t$ steps by elevating the Markov matrix to the power $t$ [13], [14], [27].

These properties of the Markov matrix are particularly interesting for the characterization of the graph structure and connections. Analyzing the behavior of the Markov matrix for long transitions, i.e., large power of the Markov matrix, can help in detecting and understanding the actual relationships among the samples in the dataset [13], [14]. To this aim, spectral theory plays a crucial role. In particular, the eigenanalysis of the aforesaid Markov matrix can help unveil hidden patterns among the samples, leading to a precise understanding of the interactions among samples. Moreover, a compact description of the random walk processes based on the eigenvectors and eigenvalues of the Markov matrix can be used to identify the information propagation mechanisms that can be drawn within the dataset according to the geometrical properties of the samples in the feature space [27].

The metric $\eta$ is expected to provide a characterization of the local geometry of the dataset [13]. On the other hand, the Markov matrix defines the direction of propagation according to the transition probabilities, which can lead to an exhaustive understanding of the overall properties of the dataset when long random walks induced by the Markov matrix are considered [27]. This scenario can be investigated in terms of a stochastic dynamical system where the transitions summarized in the Markov matrix can be described as results of a differential equation. This can lead to a global characterization of the system when integrated on a long term scale [13], [14]. Hence, the graph is considered as a realization of a dynamical process at equilibrium [13].

This analogy is particularly interesting, since it enables a robust description of the data interactions with respect to noise perturbation, as well as a multiscale analysis of the considered dataset [14]. This investigation relies once again on the transition probability proportional to the weight of the edge connecting the two nodes. In particular, the inference mechanism is based on the transition probability density $p(x(t + \epsilon) = x_j|x(t) = x_i)$ of finding the system at location $x_j$ at time $t + \epsilon$, given an initial location $x_i$ at time $t$, where $x_i$ identifies the point in the $n$-dimensional feature space corresponding to the $i$-th sample, using the notation previously introduced in this Section [13], [27], [28].

In this way, the analysis of the relationships and interactions among samples can be less affected by the density of the data and the local geometry of the dataset [14]. Nevertheless, it is also true that the characterization of the dataset in terms of dynamic system analysis requires that the Markov matrix and the metric used to quantify the edge weight in the graph representation would fulfill a number of conditions. Specifically:

- the transition probabilities must only depend on the current state and not on the past ones (first-order Markov chain). In this way, since the graph is connected, the Markov chain is irreducible, that is, every state can be reached from any other state [13], [27];
- the Markov matrix must be characterized by a unique stationary distribution to allow the existence of the eigenvalues of the Markov matrix [14], [27];
- the Markov chain must be ergodic since the space of the Markov chain associated with the matrix of the node transitions is finite and the corresponding random walk is aperiodic [14];
- the kernel $\eta(x_i, x_j)$ used to quantify the edge weight must capture the relationships between pair of samples in $X$, so it is not surprising that it must be non-negative. Moreover, the function $\eta$ must be a rotation invariant kernel [29], so that is possible to retrieve the manifold structure regardless of the distribution of the samples [14].

When these conditions are satisfied, it is possible to prove that the solution of the aforementioned problem satisfies the forward Fokker-Planck equation associated with the heat diffusion process which can be written for the density $p(x(t + \epsilon) = x_j|x(t) = x_i)$ as follows [13]:

$$\frac{\partial p}{\partial t} = \nabla \cdot (\nabla p + p \nabla U(x)), \quad (1)$$

where $\nabla = \left[\frac{\partial}{\partial x_i}\right]_{i=1,\ldots,n}$, and the state $x(t) \in \mathbb{R}^n$ (i.e., each sample in the dataset) is a realization of the dynamical system that can be written as follows:

$$\dot{x} = -\nabla U(x) + \sqrt{2w}, \quad (2)$$

where $\dot{x}$ and $w$ identify the derivatives with respect to $t$ of $x$ and $w$, respectively. Moreover, $U$ is the free energy at $x$ (which can be also called the potential at $x$), and $w(t)$ is an $n$-dimensional Brownian motion process. From a practical point of view, in this scenario the considered dataset $X = \{x_i\}_{i=1,...,N}$, $x_i \in \mathbb{R}^n$ is assumed to be sampled from the aforesaid dynamical system in equilibrium [13], [14].

In general, the solution of (1) can be written in terms of an eigenfunction expansion of the Fokker-Planck operator [13], [14]. In low dimensions, it is possible to calculate approximations to these eigenfunctions via numerical solutions
of the relevant partial differential equations. In high dimensions, however, this approach is in general infeasible and one typically resorts to simulations of trajectories of (2). In this case, there is a need to employ statistical methods to analyze the simulated trajectories, identify the slow variables, the metastable states, the reaction pathways connecting them and the mean transition times between them [13], [14], [27].

In particular, for the analysis of (2), a key role is played by the normalized graph Laplacian matrix, i.e., the matrix which $(i,j)$ element is set to $\sum_{k=1}^{N} \eta(X_i, X_k)$ if $i = j$, and $-\eta(X_i, X_j)$ otherwise [13], [14], [27], [28]. In fact, it is possible to prove that the eigenvalues and eigenfunctions of the normalized graph Laplacian operator asymptotically correspond to the Fokker-Planck equation with a potential $U(x)$ [13]. The crucial role of the Laplacian operator is further highlighted by considering that under special conditions it can have isolated eigenvalues [27], [32], [35]. These conditions can be quantified by taking into account a few metrics derived from the heat diffusion model can be written as a random matrix generated according to a spiked model with the same eigenvalues [32]–[35]. This allows the analytic investigation of the characteristics of the graph Laplacian, with special focus to its covariance [32].

In fact, it is possible to identify several conditions that the mean and covariance matrices associated with each thematic cluster in the dataset must show so that the Laplacian matrix can have isolated eigenvalues [27], [32], [35]. These conditions can be quantified by taking into account a few metrics derived from the heat diffusion model set-up and the inter-class covariance matrices [32]. Specifically, let us assume that the samples in the dataset $X$ can be associated with $k$ classes. It is therefore possible to compute for each class a length-$n$ vector $m_l = [m_{t,l}]_{t=1,...,n}$ ($l \in \{1, \ldots, k\}$), where $m_{t,l} \in \mathbb{R}$ identifies the average value that the $i$-th record assumes across the samples belonging to class $l$. Each class can be analogously characterized by a non-negative definite covariance matrix $C_l \in \mathbb{R}^{n \times n}$ computed across the samples associated with the $l$-th class. Moreover, let $N_l$ be the population of the $l$-th class within the dataset, i.e., the amount of samples belonging to the $l$-th class: hence, $\sum_{l=1}^{k} N_l = N$. It is thus possible to write $m'_l = m_l - \bar{m} = m_l - \sum_{j=1}^{k} \frac{N_j}{N} m_j$. Analogously, $C'_l = C_l - C = C_l - \sum_{j=1}^{k} \frac{N_j}{N} C_j$. Finally, we can define $T_l = \{T_{ij}\}_{i,j \in \{1, \ldots, k\}} \in \mathbb{R}^{k \times k}$, where $T_{ij} = \frac{1}{n} \text{Tr}(C'_i C'_j)$, and $t_l = [t_l]_{l=1,...,k}$, $t_l \in \mathbb{R}$.

With this in mind, it is possible to prove [32] that the Laplacian matrix would show isolated eigenvalues if the inter-class mean and covariance matrices must show as much energy (modeled by the $m$, $T$ and $t$ factors) as possible when the number of samples and/or features to be considered in the dataset increase, assuming that the first derivatives of the kernel function $\eta$ would not tend to 0 when $N \to +\infty$ and/or $n \to +\infty$ [32], [33], [36]. In other terms, one or more of the following conditions must hold:

$$\begin{align*}
|m_l'| & \stackrel{N \to +\infty, n \to +\infty}{\to} +\infty \\
t_l & \stackrel{N \to +\infty, n \to +\infty}{\to} +\infty \\
T_{ij} & \stackrel{N \to +\infty, n \to +\infty}{\to} +\infty,
\end{align*}$$

for some $l, j \in \{1, \ldots, k\}$ [32].

These conditions are sufficient to guarantee that the graph Laplacian matrix derived under the heat diffusion model can show isolated eigenvalues. In other terms, it is possible to summarize the main properties of the dataset, i.e., to lead to a thorough characterization of the interactions and relationships among the samples [32]. Nonetheless, when the data are sparse, these conditions might not be matched [7]. In this case,
the graph Laplacian matrix should be regularized to ensure that
the energy of the higher order statistics could be concentrated,
thus avoiding the aforementioned vanishing phenomenon that
could jeopardize the presence of isolated eigenvalues [7], [8], [32]. On the other hand, it is possible to show that this process
can be valid only when the number of thematic clusters the
considered records are meant to describe is very low (e.g.,
two) [7].

B. From heat to fluid: a new graph representation

1) Motivations of a new graph representation: As previously mentioned, investigating the graph structures induced by
the datasets by exploiting the heat propagation analogy in
terms of information inference has been proven to be effective and
efficient for a wide range of applications and method-
ological research instances. Nonetheless, these architectures
might fail in addressing several data analysis issues that can
occur when dealing with multimodal records, especially in
operational scenarios [17], [20].

Specifically, we can summarize the major limitations of
the classic heat diffusion model for graph investigation in the
following points [1], [2], [17]:

- adaptivity: The learning system would have to deal with
records showing multiple resolutions (either in time,
spatial, metrical units). Moreover, noise (i.e., any un-
dersired effect) might affect attributes/features/classes in
different ways across the whole dataset, as well as in
intra- and inter-class relationships. Hence, a single data
model (in terms of propagation mechanisms, label as-
signment, similarity computation) might not be adequate
for obtaining accurate and solid characterization of the
records;

- sparsity/missing data: Not all the attributes of each
sample might be relevant (by corruption, or by linear
correlation). Using all the records to compute the simi-
arity among samples might lead to dramatic degrada-
tion and/or bias of the analysis. Further, the complexity of
the data to be investigated might make classic impainting/interpolation techniques inadequate, thus jeopardizing
the validity of the outcomes;

- data mismatch/unbalance: the distribution of the thematic
clusters in the dataset might be strongly unbalanced,
and/or the training set might not contain samples associ-
ated with all the classes actually present in the dataset.
Thus, relying on a uniform statistical distribution as the
source of the samples to be investigated might be an
assumption too hard to match;

- prediction/inference: The dynamics of the phenomena
captured by the data may be too complex to follow on
a large scale, so classic graph-based approaches may
lead to a strong informative outlier removal. Furthermore, the amount of samples to be used for training the
learning/assimilation/inference models is typically very
scarce, either in quantity and quality. Not all estimates
can therefore be drawn with sufficient confidence.

These issues would result in strong limitations of the data
analysis schemes used to characterize the records. They would
in fact limit the full exploitation of the available training
set, either in terms of information extraction or context-
aware inference. Moreover, the aforementioned points would
reverberate in terms of degradation of confidence and precision
of the analysis, as well as restriction of the ability to fully
explain and interpret the records under exam [17], [18], [20],
[37].

With this in mind, the graph representation based on the
heat diffusion model might sound intuitively inadequate to
deal with all these limitations induced by modern data anal-
ysis. Nevertheless, several approximate solutions have been
proposed in technical literature, in order to mitigate the effect
of these conditions whilst maintaining the data analysis steps
compliant to the main assumptions presented in Section II-A
[1], [2], [15], [38], [39]. Thus, it is useful to provide a practical
eXample to show how the properties in the previous Section
that motivate the use of heat diffusion model are not matched
when multimodal datasets are considered. In particular, we
can focus on the conditions in (3), as they must be fulfilled for
the classic graph representation to be adequate for information
extraction from the considered datasets.

To this aim, we report in Appendix A an analysis we
conducted on a multimodal dataset that is considered as a
benchmark in the remote sensing community [40]. Investi-
gating this dataset from a theoretical, methodological, and
experimental perspective supports the need for a novel graph
representation so that the major limitations of the classic heat
diffusion model could be addressed. In particular, we have
shown how the necessary conditions for the heat diffusion
model to reliably characterize the data interactions (i.e., the
properties in (3) might not hold.

For these reasons, we propose using a fluid diffusion model
to derive a new graph representation, that could be more
flexible and versatile to address the modern data analysis needs
and limitations. Our findings are reported in the following
subsection.

2) Proposed approach: We need to define the graph topol-
ygy and the diffusion model to be applied in order to take into
account the data analysis needs mentioned in the previous sub-
section. To this aim, the definition of the process underlying
the diffusion mechanisms across the graph should reflect a
higher flexibility of the system, so to address the relevance
of the features and the modeling of a confidence score for
the propagation structure [41], [42]. Hence, the system in (2)
should be replaced by a more complex stochastic differential
model, such as follows:

\[ \dot{x} = a(x) + B(x)w, \quad (4) \]

where \( w(t) \) is a \( N \)-dimensional Wiener process, \( a(x) \) is a
length-\( n \) vector, whilst \( B(x) \) identifies a \( n \times N \) matrix [41–
[43].

In a fluid diffusion system model, \( a \) regulates the flow rate,
i.e., the velocity by which the diffusion can take place from
one node to another in the system [23], [41]. In general, it
depends on the characteristics of the \( x \) state, as well as on
the local conductivity properties of the fluid diffusion at local
scale, and on the diffusivity properties of the model at global
such, it is possible to write as follows [46]:

\[ g \text{ considering that considerations the expansion in Taylor series of identifies a Wiener process with } d \text{ scalar function} \]

This approach first introduces an arbitrary twice-differentiable different components of \( x \), i.e., a function of the conductivity and the state \( x \), \( g \). We can investigate the properties of this stochastic differential density for the system in (4). To achieve this goal, we are interested in deriving the expression of the transition probability density of a particle starting at \((x_0, t_0)\), i.e., \( p(x, t) = p(x, t|x_0, t_0) \). Specifically, we can write (7) as follows:

\[
\frac{d}{dt} \mathbb{E}[g(x)] = \frac{d}{dt} \int g(x)p(x, t)dx = \int g(x) \frac{\partial p(x, t)}{\partial t} dx \\
= \int a(x) \cdot \nabla g(x)p(x, t)dx + \sum_{i,j=1}^{n} \int \tilde{B}_{ij} \partial x_i \partial x_j g(x)p(x, t)dx.
\]

Integrating by parts, this equation can be rewritten as [46]:

\[
\int \left[ \frac{\partial p(x, t)}{\partial t} + \nabla \cdot [a(x)p(x, t)] - \beta'(x, t) \right] g(x)dx = 0, \quad (9)
\]

where \( \beta'(x, t) = \sum_{i,j=1}^{n} \partial x_i \partial x_j(\tilde{B}_{ij}p(x, t)) \). Then, since the \( g \) function is arbitrary by construction, the aforementioned equation is satisfied when the term inside the square brackets is null. With this in mind and expanding the \( \beta' \) term, it is possible to write as follows (the details of the algebraic steps are detailed in Appendix B):

\[
\frac{\partial p(x, t)}{\partial t} + \nabla \cdot [a(x)p(x, t)] - \left[ \nabla \cdot \left( \nabla \mathbb{B}(x) \right) \right] p(x, t) \\
= \left[ \nabla \mathbb{B}(x) \right] \cdot \nabla p(x, t) - \nabla \cdot \left[ \mathbb{B}(x) \nabla p(x, t) \right]. \quad (10)
\]

At this point, using the representation in (3), and considering the linearity of the divergence operator and its product rule [46], it is now possible to write the diffusion equation for this system as follows:

\[
\frac{\partial p(x, t)}{\partial t} = - \nabla \cdot \left( [a(x) - \nabla \mathbb{B}(x)] p(x, t) \right) \\
+ \nabla \cdot \mathbb{B}(x) \nabla p(x, t). \quad (11)
\]

It is possible to recognize in this expression the Fokker-Planck equation for fluid diffusion in porous media [22]. Analogously to the heat diffusion analysis, the solution of (11) can be retrieved by eigenanalysis of the Fokker-Planck operator. Furthermore, the asymptotic analysis of the trajectories of the system in (3) can help in obtaining a thorough characterization of its solution [22].

It is indeed possible to analyze the geometry of the dataset by investigating the data by means of an approach based on Markov chain scheme [13]. In fact, it is possible to characterize the transition probability density \( p(x(t+\epsilon) = x_j|x(t) = x_i) \) by using a time domain random walk approach [47], [48]. This scheme achieves the characterization of the whole system by exploiting the adjacency of the nodes. Specifically, it aims to solve a Green function problem derived from (3) by imposing initial conditions and absorbing boundary conditions to the diffusion system centered on the node \( i \) [47].

This strategy aims to determine the transition probability density \( p(x(t+\epsilon) = x_j|x(t) = x_i) \) by means of the first arrival time density \( \phi_{ij} \) at the boundary between nodes \( i \) and \( j \), that would denote the joint probability of the transition to...
occur from node $i$ to node $j$ \cite{47}. Specifically, it is possible to write this transition probability density as follows:

$$p(x(t + \epsilon) = x_j | x(t) = x_i) = p_{ij} = \int_0^{+\infty} \phi_{ij} dt$$ \hspace{1cm} (12)$$

The details of the time domain random walk strategy are detailed in Appendix C. By projecting the solution of the Green function problem in a Laplace space, we can write $p_{ij}$ as follows \cite{47, 48}:

$$p_{ij} = \frac{|v_{ij}^+| \exp[v_{ij}^+] \csch[v_{ij}^-]|}{\sum_{u \in \{+, -\}|v_{ui}^+| \exp[u \cdot v_{ui}^-] \csch[v_{ui}^-]|}$$ \hspace{1cm} (13)$$

where $\csch[z] = 1/\sinh[z] = 2/(\exp[z] - \exp[-z])$. Moreover, $\tilde{u}$ is set to 1 when $u = +$, whilst $\tilde{u} = -1$ if $u$ is -. Finally, $v_{ij}^\pm = v_{ij} / 2B_{ij}$, being:

$$v_+ = v_{ij}, \quad v_\mp = \sum_{m \in N(i) \cup j} \frac{|v_{im}|}{|N(i)| - 1}$$

$$B_\mp = B_{ij}, \quad B_\mp = \sum_{m \in N(i) \cup j} \frac{|B_{im}|}{|N(i)| - 1}$$

where $N(i)$ identifies the neighborhood of node $i$, i.e., the set of nodes adjacent to node $i$. Finally, $B_{ij} \in [0, 1]$ is the $(i, j)$-th element of the matrix $B$, whereas $v_{ij}$ is the transport velocity between node $i$ and node $j$ as in \cite{5}. This quantity is typically computed as $v_{ij} = -|\mathcal{K}_{ij}^-| \odot |x_i - x_j|_2 \in [0, 1]$, where $\mathcal{K}_{ij}$ is the length-$n$ row vector collecting the third dimension elements of the conductivity tensor $\mathcal{K}$ on the $(i, j)$ coordinates and $\odot$ is the Hadamard product \cite{22, 23, 41, 42, 47, 48}.

Hence, it is possible to define a Markov matrix $Q = \{Q_{ij}\}_{(i, j) \in \{1, \ldots, N\}^2}$ that summarizes the edge weights of the new graph representation according to the transition probabilities in \cite{13}, i.e., $Q_{ij} = p_{ij} / \sum_i p_{ij}$ \cite{13, 14, 42}. The matrix $Q$ can be explored and used to address several tasks in multimodal data analysis and to improve the information extraction from the considered datasets. In particular, the eigenanalysis of the Laplacian matrix associated with $Q$ (whose $(i, j)$ element can be defined as $\sum_k \{1, \ldots, N\} \setminus i Q_{ik}$ if $i = j$, and as $-Q_{ij}$ otherwise) can be directly connected to the solution of the system in \cite{11}. In fact, in general the solution of the fluid diffusion equation can be written in terms of the eigenfunction expansion, i.e., $p(x, t)$ can be expressed as follows:

$$p(x, t) = \sum_{i=0}^{+\infty} \omega_i \exp[-\lambda_i t] \varphi_i(x)$$ \hspace{1cm} (15)$$

where $\lambda_i$ are the sorted eigenvalues of the fluid Fokker-Planck operator (with $\lambda_0 = 0$), $\varphi_i$ are their corresponding eigenfunctions, and the coefficients $\omega_i$ depend on the initial conditions \cite{13, 14, 42}.

It is worth noting that numerical approximations of these eigenfunctions can be computed when the considered data are characterized by a low amount of records (e.g., three). On the other hand, when high dimensional data are considered - such as the multimodal data we are considering in this work - it is not possible to use numerical solutions to solve this equation. The only valid approach would be to simulate the trajectories of the stochastic differential model in (4), which implies the use of statistical methods to analyze the simulated results and explore the validity of low and high frequency trends, as well as the mean transition times among them \cite{13, 41, 42}.

It has been proven that the solution of (13) can be described by reduced set of $\kappa$ eigenfunctions, which can carry significant information on the density and geometry of the data under exam \cite{13, 22, 42}. To obtain a reliable characterization of the relevant eigenvectors and eigenvalues, it is useful to explore the asymptotic behavior of the diffusion process in the probability space. This analysis shows that the eigenvalues of the matrix $Q$ asymptotically correspond to the relevant $\kappa$ eigenfunctions that can be used to achieve a solid understanding of the solution in (15), as previously mentioned \cite{13, 14, 22, 23, 41, 42}. This result is extremely interesting, and it summarizes the key-role that the matrix $Q$ can have in providing a thorough and reliable understanding of the properties underlying the graph topology induced by the considered datasets, as well as the information propagation mechanisms.

In Appendix D, we provide an example to visualize how the proposed definition of the $Q$ matrix could improve the characterization of data interactions with respect to the classic graph representation based on heat diffusion mechanism.

In this work, we focus our attention on the investigation of $Q$ in order to learn the structure of the data under exam, and to enable an effective functional analysis of the records, with special focus to multimodal data analysis. The next Section summarizes the main steps of the proposed method for fluid community detection.

### III. FLUID COMMUNITY DETECTION

#### A. Background and related works

Several criteria at global and local scale can be used to identify communities in graphs. Since in this work we focus our attention towards the detection of communities that are separated, we report in this Section an overview of non-overlapping community detection methods. Moreover, we summarize the main categories community detection algorithms can be grouped in, according to the strategy they employ \cite{6}.

In particular, it is possible to categorize these methods in seven main groups: 1) graph partitioning; 2) hierarchical clustering; 3) partitional clustering; 4) spectral clustering; 5) dynamic community detection; 6) statistical inference-based community detection; and 7) hybrid methods. We report in \cite{6} a brief overview of these algorithms. The methods are typically designed to address problems that could be described in a monovariate data analysis system \cite{6, 31, 49}. In particular, these architectures are developed at theoretical level.
to address community detection problems when a single source of information is used to generate the data to be analyzed. Nevertheless, multimodal community detection is typically addressed by extending these approaches to records acquired by multiple modalities \( [17], [18], [37], [49] \). In particular, the similarity between samples (either in terms of edge betweenness, modularity, Euclidean distance, geodesic metric) is computed along all the features available \( [1], [2], [50] \). In other terms, the aforesaid methods are applied to datasets where the records acquired by the diverse modalities are stacked and vectorized, so that each sample could be considered as a point in an extended multidimensional feature space.

This approach is very popular within the scientific community, because of its high degree of implementability. However, this does not always reflect in good performance in terms of community detection, especially in operational scenarios \( [2], [15], [17], [20], [49] \). Hence, directly applying these architectures to multimodal data analysis might lead to strong limitations of the multimodal community detection performance. In other terms, successful community detection is achieved only when datasets characterized by low diversity, low sparsity, high reliability, and low variability can be found across the considered records (see for instance \( [51]–[54] \)). As these characteristics are hard to be found in multimodal datasets, especially when addressing operational scenarios \( [17], [49] \). Hence, this makes this research avenue an open field for a successful development of multimodal data analysis methods \( [49], [55] \).

Nevertheless, it is also true in recent years methods designed to address multimodal data analysis have been proposed. One first approach relies on the design of \( n \)-partite graphs, where different kind of nodes are used to represent the diverse modalities under exam \( [56]–[58] \). On these graph structures, partitional clustering techniques are applied to retrieve the community structures hidden within the data. These methods are typically showing pretty high computational complexity, such that it is proven that they might achieve successful performance when the trade-off between diversity and thematic clusters to be identified is good (i.e., when few modalities with numerous communities are considered, or when multiple modalities and a small amount of communities are taken into account). This is a major factor that must be considered when employing these techniques \( [6], [49], [55] \).

Similar results can be registered when the records acquired by diverse modalities are separately processed, to be then fused at a later stage \( [59]–[61] \). In this case, methods retrieved from graph partitioning, partitional clustering, and/or hybrid approaches are used to analyze the different sources of information. Then, the obtained information is integrated according to optimization criteria designed to maximize the alignment between modalities and therefore identify communities are are more homogeneous across the diverse sources of information \( [60] \). Although this approach can be performed with pretty low latency (especially when high performance computing platforms are available), it can hardly be generalized for multimodal datasets where diverse records, characterized by various statistical distributions, high variability and sparsity are considered, limiting the range of applications that could actually benefit of this strategy.

Recently, methods relying on the multiview data analysis approach have been introduced for community detection \( [62], [63] \). In this case, the different \( n \) modalities are assumed to generate \( n \) graphs that are then investigated by means of spectral clustering techniques. Then, an optimization process based on normalized cut approach is performed, in order to identify the most informative clusters across the whole dataset. This approach typically shows low computational complexity. However, it is also true that it requires the modalities to be as less as diverse as possible, so that the joint cut across the graphs can be accurately performed \( [62] \). Moreover, the multiple graphs are expected to show homogeneous characteristics imposed by the communities underneath. This is a pretty strong requirement, since in multimodal data analysis not all the features are typically reliable, significant and/or informative at the same level \( [63] \). As such, using this scheme to integrate heterogeneous sources of information at large scale might be cumbersome.

\section*{B. Proposed approach}

Taking into account the definition of the \( Q \) matrix as a result of the fluid diffusion model (as introduced in Section II-B2), there are several properties that can be particularly interesting to address the community detection task in an accurate and efficient way. In this work, we take advantage of the aforementioned characteristics of the eigenanalysis of the flow velocity matrix to identify the most informative clusters within the considered multimodal dataset. As such, the approach we propose in this paper could fall within the spectral clustering category of community detection mentioned in Section II-A. This choice helps us in achieving accurate and reliable understanding of the data interactions in closed form and with rigorous convergence, while guaranteeing a simple implementation and high efficiency of the unsupervised community detection approach. It is also worth noting that the focus of this paper is on unsupervised analysis of the functional relationships among samples, i.e., no contextual information (either in shape of side information, or \textit{a priori} knowledge, or semantic knowledge) could be taken into account to achieve a fully data drive investigation. As such, the diffusivity term \( B \) in (5) can be then set to the identity matrix throughout the following Sections.

\textbf{1) Definition of the \( Q \) matrix:} In order to provide a thorough investigation of the complex relationships hidden within the records in multimodal datasets according to the fluid graph representation previously introduced, we first need to define the permeability tensor \( \mathcal{H} \) in (5). To this aim, it would be instrumental to investigate the significance of the features associated with each sample in the considered dataset. In this work, we propose to address this task by exploring the relevance of the features at global (i.e., across all modalities) and local (i.e., across samples for each feature) scale. Following the successful approach proposed in [64], we quantify the multiscale significance of the features by using information theory-based metrics. Specifically, we consider to measure the degree of redundancy and intercorrelation between features.
across the whole dataset by employing mutual information [65], [66]. This choice allows us to assess the redundancy and dependence among features we could record across the dataset. In fact, mutual information quantifies the shared information between two random variables [66]. This is especially relevant when complex datasets, which lead to fully connected graph between two random variables [66]. This is especially relevant.

In other terms, let us consider a dataset \( \mathbf{X} \) that consists of \( N \) samples and \( n \) features, i.e., \( \mathbf{X} = \{ \mathbf{x}_i^T \}_{i=1,...,N} \), \( \mathbf{x}_i = \{ x_{ij} \}_{j=1,...,n} \), that induces a graph \( \mathcal{G} = (\mathcal{V}, \mathcal{E}) \) where \( \mathcal{V} \) and \( \mathcal{E} \) identify the node and edge sets, respectively. Moreover, the \( i \)-th node \( v_i \) is associated with the \( i \)-th sample \( \mathbf{x}_i \). Then, we can write the mutual information between two features \( n_1 \) and \( n_2 \) (where \( n_j \in \{1, \ldots, n\} \)) as follows:

\[
W_{n_1n_2}^{\text{MI}} = \sum_{i=1}^{N} \sum_{j=1}^{N} p(x_{i,n_1}, x_{j,n_2}) \log \frac{p(x_{i,n_1}, x_{j,n_2})}{p(x_{i,n_1})p(x_{j,n_2})}, \tag{16}
\]

where \( p(y, z) \) is the joint density function of \( y \) and \( z \), and \( p(z) \) is marginal of \( z \). It is worth to note that, according to this definition, large values of \( W_{n_1n_2}^{\text{MI}} \) imply redundancy in information. On the other hand, low values of \( W_{n_1n_2}^{\text{MI}} \) imply synergy (novelty) [65].

On the other hand, it is important to evaluate the significance of the local properties of the features for each sample. In this way, we can take into account the local characteristics of each feature, so to address the variability of the statistical properties of the data across the complete dataset. In other terms, we should identify a metric for which, if the features \( n_1 \) and \( n_2 \) are very similar across the \( m \)-th sample, the value will be large. In this case, it would be possible to assume that using just one of these features would be enough to obtain a robust and precise understanding of the given sample. Conversely, a small value of this metric would mean that the two features are independent from each other such that they should be both taken into account to characterize the sample [26], [31].

The distance metric based on Gaussian kernel would shows all these properties. Thus, we propose to quantify the difference between two features \( n_1 \) and \( n_2 \) for the \( m \)-th sample as follows:

\[
W_{m_{n_1n_2}}^{\text{GK}} = \exp \left[ -\frac{||\mathbf{z}_{mn_1} - \mathbf{z}_{mn_2}||^2}{2\sigma^2} \right], \tag{17}
\]

where the value of \( \sigma \) controls the width of the Gaussian kernel [31], [32].

At this point, we can build for each sample \( m \) a graph \( \mathcal{G}_m = (\mathcal{V}, \mathcal{E}^{\text{GK}}, \mathcal{E}^{\text{MI}}) \). The \( n \)-th vertex \( v_n \) in \( \mathcal{V} \) identifies the \( n \)-th feature of the \( m \)-th sample. Two vertices \( v_{n_1} \) and \( v_{n_2} \) are connected by two kinds of edges, \( \mathcal{E}^{\text{GK}} \) and \( \mathcal{E}^{\text{MI}} \), whose weights are computed according to (17) and (16), respectively. This structure can be used as a platform to perform adaptive feature selection across the dataset according to the guidelines of spectral clustering approach [7], [8], [24]–[26], [31]. In particular, the aforesaid weights can be arranged in matrix form, so to generate two adjacency matrices associated with \( \mathcal{G}_m \), i.e., \( \mathbf{W}^{\text{GK}} = \{w_{m_{ij}}^{\text{GK}}\}_{(i,j)\in\{1,...,n\}^2} \) and \( \mathbf{W}^{\text{MI}} = \{w_{m_{ij}}^{\text{MI}}\}_{(i,j)\in\{1,...,n\}^2} \). For the first matrix, it is possible to define a degree matrix \( \mathbf{D}^{\text{GK}} = \text{diag}(\mathbf{d}^{\text{GK}}) \),

\[
\mathbf{d}^{\text{GK}}_m = \{d^{\text{GK}}_i = \sum_{l=1}^{n} w_{m_{il}}^{\text{GK}} \}_{j=1,...,n},
\]

where \( d^{\text{GK}}_i \) are the number of edges connected to the \( i \)-th node, and we define \( \mathbf{I}_m = \mathbf{I} - \mathbf{D}^{\text{GK}}^{-1/2} \mathbf{W}^{\text{GK}} \mathbf{D}^{\text{GK}}^{-1/2} \). Analogously, we can define \( \mathbf{D}^{\text{MI}} = \text{diag}(\mathbf{d}^{\text{MI}}) \), \( \mathbf{d}^{\text{MI}}_m = \{d^{\text{MI}}_i = \sum_{l=1}^{n} w_{m_{il}}^{\text{MI}} \}_{j=1,...,n} \), as the degree matrix associated with \( \mathbf{W}^{\text{MI}} \), and \( \mathbf{I}^{\text{MI}} = \mathbf{I} - \mathbf{D}^{\text{MI}}^{-1/2} \mathbf{W}^{\text{MI}} \mathbf{D}^{\text{MI}}^{-1/2} \) as its associated normalized Laplacian matrix.

With this in mind, identifying the relevant features for each sample in the dataset can be described as partitioning the graph \( \mathcal{G}_m \) such that the vertices of the same subgraph have strong connections via both links, while the vertices from different subgraphs have one or two weak connections. In spectral clustering, this problem can be written as follows:

\[
\begin{align*}
\min_{\mathbf{H}} \quad & \text{Tr}(\mathbf{H}^T \mathbf{L}^{\text{GK}} \mathbf{H}) & \\
\text{subject to} & \quad \text{Tr}(\mathbf{H}^T \mathbf{L}^{\text{MI}} \mathbf{H}) = 1
\end{align*}
\]

where \( \mathbf{H} \) represents the matrix of the indicator vectors, and \( \mathbf{HH}^T = \mathbf{I} \) [25], [26], [31]. The solution of (18) is given by the common eigenspace of the two normalized Laplacian matrices. Hence, this problem translates in identifying the set of joint eigenvectors \( \mathbf{v}_m = [v_{m1}, \ldots, v_{mn}] \) that solves the following [68], [69]:

\[
\begin{align*}
\min_{\mathbf{v}_m} \quad & \log \frac{\text{diag}(\mathbf{v}_m^T \mathbf{L}^{\text{GK}} \mathbf{v}_m)}{\text{diag}(\mathbf{v}_m^T \mathbf{L}^{\text{MI}} \mathbf{v}_m)} & \\
\text{subject to} & \quad \mathbf{v}_m^T \mathbf{1} = 1
\end{align*}
\]

At this point, \( \mathbf{H} \) in (18) contains the eigenvectors corresponding to the \( K_m \) lowest and non-null eigenvalues. It is worth recalling that the cardinality \( K_m \) identifies the number of relevant features in the \( m \)-th sample according to the spectral clustering guidelines [7], [8], [24]–[26], [31]. In fact, the number of relevant features equals the number of informative eigenvalues that can be defined as the local minima of the eigenvalues’ difference curve [7], [26], [31]. To this aim, the kneedle algorithm can be employed to select the optimal \( K_m \) [70]. It is therefore crucial that the difference among the eigenvalues is well pronounced, so that the separation between eigenvalues associated with relevant and non-informative features can be easily carried out. Once the \( K_m \) eigenvalues have been identified, it is possible to select the set of relevant features for the \( m \)-th sample as the centroids of the associated clusters.

This information will finally be used to define the elements of the \( \mathcal{K} \) tensor in [5]. Specifically, if the \( l \)-th feature has been selected as relevant for both sample \( m_1 \) and \( m_2 \), then \( \mathcal{K}_{m_1m_2l} = 1 \); otherwise, \( \mathcal{K}_{m_1m_2l} = 0 \). It is worth noting that this simple set-up could be made more sophisticated by allowing the values of \( \mathcal{K} \) to live in \( \mathbb{R} \). Future works will be dedicated to investigate the impact of this choice in the effective use of the proposed fluid diffusion model in multimodal data analysis. Analogously, the definition of the distance operator to be used to determine the \( v \) values in [13] can be subject for deep investigation in the near future. Nevertheless, in this work we can assume without losing generality (and considering the observations on the continuity
of the information propagation drawn in Section II-B that the each \( v_{ij} \) in (15) will be based on the norm-2 between nodes [47], [48].

At this point, we can compute the \( Q \) matrix that has been introduced in Section II-B2. The next steps of the proposed community detection strategy consist in the eigenanalysis of the \( Q \) matrix. The next paragraphs summarize the main steps of the approach we present in this work.

2) Community detection based on fluid Laplacian matrix:
As previously mentioned, the \( Q \) matrix is the core of the community detection algorithm based on fluid diffusion that we introduce in this work. We can indeed build a new matrix \( F = D - Q \), where \( D \) is a \( N \times N \) diagonal matrix such that \( D_{ii} = \sum_{j=1}^{N} Q_{ij} \). As such, \( F \) is a matrix where all the diagonal elements are positive, and the other elements are negative. Therefore, \( F \) is invertible. Let us further analyze the properties of \( F \). Specifically, let us consider a generic vector \( z \), and let us derive the analytical solution of the \( z^T F z \) function. It is possible to prove that the following holds [7], [10]:

\[
 z^T F z = z^T Dz - z^T Qz = \sum_{i=1}^{N} D_{ii} z_i^2 - \sum_{i,j=1}^{N} Q_{ij} z_i z_j 
\]

\[
 = \frac{1}{2} \left( \sum_{i=1}^{N} D_{ii} z_i^2 - 2 \sum_{i,j=1}^{N} Q_{ij} z_i z_j + \sum_{j=1}^{N} D_{jj} z_j^2 \right) 
\]

\[
 = \frac{1}{2} \sum_{i,j=1}^{N} Q_{ij} (z_i - z_j)^2.
\]

Therefore, the matrix \( F \) can be considered as a Laplacian matrix, and will take the name of fluid Laplacian matrix. Moreover, this system can be used to construct a graph that could be partitioned in communities. In particular, in order to find a partition of the graph such that the edges between different communities have lower weight and the edges within the same community have a higher weight, we can apply the Normalized Cut algorithm [71]. In other terms, the graph induced by \( Q \) can be partitioned in \( K_F \) connected components \( C_k, k = 1, \ldots, K_F \) (where \( \bigcup_{k=1}^{K_F} C_k = \mathcal{V} \), and \( C_{k_1} \cap C_{k_2} = \emptyset \) \( \forall k_1, k_2 \in \{1, \ldots, K_F\}, k_1 \neq k_2 \)) by minimizing over \( \{C_k\}_{k=1, \ldots, K_F} \) the NormalizedCut function \( NC_{K_F} \), which can be written as follows:

\[
 NC_{K_F} = \frac{1}{2} \sum_{k=1}^{K_F} \frac{\zeta(C_k, \bar{C}_k)}{C_k},
\]

where \( \bar{C}_k \) represents the complement of the \( k \)-th partition over the vertex set \( \mathcal{V} \), \( C_k \) is a measure of the width of the \( k \)-th partition (typically expressed in volume \( \text{vol}(C_k) = \sum_{v_i \in C_k} \sum_{j=1}^{N} Q_{ij} \)), and \( \zeta(C_k, \bar{C}_k) = \sum_{v_i \in C_k, v_j \in \bar{C}_k} Q_{ij} \).

The minimization of \( NC_{K_F} \) leads to have large weights for the edges connecting the nodes within \( C_k \), while the edges connecting nodes within \( C_k \) with the nodes in its complement \( \bar{C}_k \) will show small weights. Furthermore, this operation can be described in terms of the eigenvectors of the normalized fluid Laplacian matrix \( \bar{F} = D^{-1/2} F D^{-1/2} \). In other terms, the \( NC_{K_F} \) optimization can be written as follows:

\[
 \min_{J} \text{Tr}(J^T F J) \quad \text{subject to} \quad J^T J = I,
\]

where \( J \) is the matrix of the first smallest \( K_F \) eigenvectors of \( \bar{F} \). Hence, in order to solve this problem, it is possible to employ the kneedle algorithm [70] to select the best value of \( K_F \), and then run a traditional \( K_F \)-means algorithm over \( J \) (considering the rows of \( J \) as nodes) in order to identify the \( K_F \) communities [8], [10], [31]. In this way, we can guarantee a high degree of implementability and efficiency of the system, whilst achieving a thorough unsupervised characterization of the multimodal data under exam. The following Section provides tests to validate this set-up with respect to state-of-the-art methods.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

Several multimodal datasets representative of different research fields have been used to validate the novel graph representation based on fluid diffusion model and to test the performance of the proposed community detection method. In this Section, we first summarize the main characteristics of the datasets we have taken into account. Then, we report the performance of the proposed community detection framework based on fluid diffusion model.

A. Datasets

We tested the proposed approach on three very diverse datasets, focusing on three different research fields: remote sensing, brain-computer interface, and photovoltaic energy.

1) Multimodal remote sensing (RS): First, we considered a multimodal dataset consisting of LiDAR and hyperspectral records acquired over the University of Houston campus and the neighboring urban area, and was distributed for the 2013 IEEE GRSS Data Fusion Contest [72]. Specifically:

- the size of the dataset is 1905x349 pixels, with spatial resolution equal to 2.5m;
- the final dataset consists of \( N=151 \) features. In fact, the hyperspectral dataset includes 144 spectral bands ranging from 0.38 to 1.05 \( \mu m \), whilst the LiDAR records includes one band and 6 textural features;
- the available ground truth labels consists of \( K_C = 15 \) classes.

2) Multimodal brain-computer interface (BCI): The second dataset we considered was collected by means of brain-computer interface [73]. Specifically, the records were collected by means of 60-channel electroencephalography (EEG), 7-channel electromyography (EMG) and 4-channel electro-oculography (EOG) on \( K_C = 11 \) intuitive upper extremity movements from 25 participants. A 3-sessions experiment was carried out, and 3300 trials per participant were collected. According to the notation we have used in Section II-B the final dataset consists of \( N = 71(= 60 + 4 + 7) \) multimodal features for a total sum of 82500 samples across all the participants.
3) Multimodal photovoltaic energy (PV): The final dataset was acquired in order to monitor the photovoltaic energy produced between July 21 and Aug. 17, 2018 at the University of Queensland, Australia \[74\]. The records that have been collected by weather ground stations can be listed as follows:
- instantaneous and average wind speed [km/h] and direction [deg];
- temperature [deg];
- relative humidity [%];
- mean surface level pressure [hPa];
- rain intensity [mm/h];
- accumulated rain [mm];
- accumulated hail [hits/cm\(^2\)];
- hail intensity [hits/cm\(^2\)hr];
- solar mean [W/m\(^2\)].

This summed to 1440 samples acquired for each day, summing up to a dataset of 1440 \(\times\) 28 records. For each sample, the photovoltaic energy [W/h] is recorded: \(K = 10\) classes uniformly drawn based on this parameter are considered. The considered data analysis task consists of assigning a class to all the samples by investigating the \(N = 12\) heterogeneous features.

B. Results

In order to provide a thorough investigation of the actual impact of the proposed approach, we conducted several experiments at different levels. Specifically, we tested the proposed architecture in terms of design choices, validity of the proposed model, parameter sensitivity, and community detection performance.

First, we investigated whether the proposed adaptive dimensionality reduction method for the definition of the \(\mathcal{X}\) tensor in \(13\) that is used then to determine the fluid Laplacian matrix in Section III-B2 in particular, we tested the ability of the method in Section III-B1 to reliably identify relevant features across complex datasets, so that the construction of the \(\mathcal{X}\) tensor could be carried out. To this aim, we considered the three aforementioned datasets, and added records generated by noise-like process characterized by a Gaussian distribution and a signal-to-noise ratio (SNR) set to 20dB to each sample. Hence, we obtained for each dataset an additional subset of attributes (approximately 30\% of the original amount, i.e., we added 50, 23, and 4 features for the datasets in Section IV-A1-IV-A3 respectively) that were clearly irrelevant for the characterization of the phenomena under exam and therefore simulating a set of corrupted, noisy, and non-informative records. Then, we applied the adaptive dimensionality reduction approach for the determination of \(\mathcal{X}\), as well as other four state-of-the-art feature selection methods on the new total datasets: the algorithms we considered were based on diverse strategies, i.e., genetic algorithm (GA) \[75\], structure preserving feature selection (SPFS) \[76\], regularization-based feature selection (RegFS) \[77\], and filter-based feature selection (FilterFS) \[78\]. Our goal is to check whether these methods are able to discriminate the real records from the added (irrelevant) ones in the obtained extended datasets. We carried out 100 runs, and summarized the results in Fig. 1.

Taking a look to these histograms, it is possible to appreciate how the approach we proposed in this work is actually able to outperform the state-of-the-art methods by selecting almost exclusively the original (relevant) features for dimensionality reduction. Moreover, the high kurtosis of the histograms achieved by means of the adaptive dimensionality reduction method emphasizes the robustness of this algorithm to select relevant features across all the tests we have performed. On the other hand, the variability of the performance of the other methods appear very high, so that their outcomes do not appear solid. Thes observations are valid for all the datasets.
we presented in Section IV-A hence supporting our choice of using the adaptive dimensionality reduction method presented in Section III-B1 for the construction of the fluid Laplacian matrix.

Let us now focus our attention on the actual procedure for community detection based on the fluid diffusion model that we propose in this work. In particular, it is worth to recall that the main steps for the community detection technique reported in Section III-B2 are fundamentally based on the eigenanalysis of the fluid Laplacian matrix \( \mathbf{F} \). Thus, the method in Section III-B2 could be considered as an instance of the spectral clustering approach for community detection, according to the characteristics summarized in Appendix E4. As such, the ability to discriminate the lowest eigenvalues from the overall eigenvalues set, so that the identification of the communities in the dataset can be accurately carried out [31]. Therefore, it is important to analyze the eigenspectrum of the computed eigenvalues, so to retrieve a solid understanding of the actual characterization ability the considered spectral clustering-based architecture might have. Hence, in order to obtain a first assessment of the actual impact of the proposed community detection method, we assessed the improvement provided by the use of the fluid Laplacian matrix for spectral clustering.

Specifically, we computed the eigenspectrum resulting from the analysis of the datasets in Section IV-A by means of the scheme proposed in Section III-B2 and several spectral clustering methods introduced in technical literature, i.e., using unnormalized and normalized Laplacian matrix [31], graph distance-based spectral clustering [10], covariate-assisted spectral clustering [9], spectral clustering using probability matrix [11], self tuning spectral clustering [12], and regularized Laplacian matrix [7]. It is worth noting that all these methods are relying on the graph representation based on heat diffusion. In particular, the parameter \( \zeta_p \) for the regularized Laplacian matrix was set to \( \sqrt{\epsilon} \), according to the guidelines in [7]. We reported the retrieved eigenspectra associated with these approaches applied to the three datasets. In other terms, we displayed in these figures the histogram showing on the y-axis the percentage of occurrence of all the values each eigenvalue can assume, that are reported on the x-axis. By visual inspection, it is possible to appreciate that the proposed approach is able to provide a better identification of the isolated eigenvalues in the dataset by increasing the spread among them. On the other hand, the eigenvalues obtained by using the Laplacian matrices defined in the aforesaid methods appear to be all compressed close to 0. As such, it is possible to assume that the subsequent eigenanalysis for the detection of communities in the data might be cumbersome and difficult to perform.

In this respect, when considering a dataset composed by \( K_C \) communities, the gap between the \( K_C \)-th and the \( K_C + 1 \)-th eigenvalues plays a crucial role to the achievement of an effective clustering of the datasets, according to the theoretical aspects of spectral clustering (briefly reported in Appendix E4). In particular, it is possible to achieve a more accurate community detection for large eigengaps. Thus, to quantify the difference between the approaches considered in Figs. 4-6, we computed the difference \( |\lambda_{K_C + 1} - \lambda_{K_C}| \) (where \( \lambda_i \) identifies the \( i \)-th eigenvalue) for all the methods in these figures. We reported the eigengaps we obtained for the considered datasets in Table I where \( K_C \) is set to 15, 11, and 10 for the datasets in Section IV-A1, IV-A2, and IV-A3 respectively. At this point, the improvement provided by using the fluid Laplacian matrix as in Section III-B2 with respect to state-of-the-art methods appears dramatic. Indeed, the results in Table I emphasize the ability of the fluid diffusion model in addressing the complex interactions among samples that can occur at global and local scale in multimodal datasets. This impact of this result is further highlighted by Table II, where the gap between the \( K_C + 1 \)-th and \( K_C + 2 \)-th eigenvalue is displayed. In fact, it is possible to appreciate how this difference is sensibly smaller than their corresponding values in Table I.

| TABLE I |
|---|
| \textbf{Eigengap} \( |\lambda_{K_C+1} - \lambda_{K_C}| \) FOR THE METHODS IN FIG. 4 TO 6 |
| \textbf{Method} | \textbf{RS} (multimodal remote sensing) | \textbf{BCI} (multimodal brain-computer interface) | \textbf{PV} (multimodal photovoltaic energy) |
| Fluid Unnormalized | 45 | 48 | 54 |
| Fluid Normalized | \( \times 10^{-14} \) | \( \times 10^{-14} \) | \( \times 10^{-15} \) |
| Fluid Graph Distance | \( \times 10^{-14} \) | \( \times 10^{-14} \) | \( \times 10^{-14} \) |
| Fluid Covariate | \( \times 10^{-14} \) | \( \times 10^{-14} \) | \( \times 10^{-14} \) |
| Fluid Probability | \( \times 10^{-14} \) | \( \times 10^{-14} \) | \( \times 10^{-14} \) |
| Fluid Self tuning | \( \times 10^{-14} \) | \( \times 10^{-14} \) | \( \times 10^{-14} \) |
| Fluid Regularized | \( \times 10^{-14} \) | \( \times 10^{-14} \) | \( \times 10^{-14} \) |

With this in mind, we can further explore the capacity of the proposed method by assessing its actual ability of detecting communities within the considered datasets. In order to obtain a quantitative assessment in this sense, we
compute cluster assignments on the graph and evaluate the partitions that are delivered. We evaluate the performance of the methods by computing three indices: modified purity ($mP$), modified adjusted Rand index ($mARI$), and modified normalized mutual information ($mNMI$) [79, 80]. These metrics are able to quantify the ability of the methods to understand the graph properties, and to provide a proper setup to extract information at semantic and functional level from the considered dataset. Moreover, in order to obtain a more reliable evaluation of the graph learning performance, these metrics take into account graph topology whilst their “traditional” counterparts do not [79, 80].

In order to give a compact definition of the aforesaid metrics, let $\Omega = \{\omega_k\}_{k=1,\ldots,C}$ be the set of detected clusters, whilst $\Psi = \{\psi_l\}_{l=1,\ldots,C}$ is the set of ground-truth classes. Then, let us consider the modified purity index. To define $mP$, it is important to take into account the purity of a node for a partition $\Omega$ relatively to another partition $\Psi$, i.e., a function that identifies if the class of $\Psi$ containing node $u$ is majority in that of $\Omega$ also containing $u$, and otherwise. This function can be written as:

$$P(u, \Omega, \Psi) = \delta(\psi_j \; s.t. \; |\omega_i \cap \psi_j| \; is \; maximum),$$  \hspace{1cm} (23)

where $u \in \omega_i$, $u \in \psi_j$, and $\delta$ is the Kronecker delta function. At this point, assuming that $w_u$ is the weight of node $u$ the modified purity ($mP$) can be defined as follows:

$$mP(\Omega, \Psi) = \sum_{i} \sum_{u \in \omega_i} \frac{w_u}{|\omega_i|} P(u, \Omega, \Psi)$$  \hspace{1cm} (24)

TABLE II

| Method               | RS (multimodal remote sensing) | BCI (multimodal brain-computer interface) | PV (multimodal photovoltaic energy) |
|----------------------|-------------------------------|------------------------------------------|-------------------------------------|
| Fluid                | 0.2                           | 1.2                                      | 0.3                                 |
| Unnormalized         | $3.2 \times 10^{-14}$         | $1.8 \times 10^{-14}$                   | $2.2 \times 10^{-15}$               |
| Normalized           | $6.4 \times 10^{-15}$         | $3.9 \times 10^{-15}$                   | $2.3 \times 10^{-15}$               |
| Graph Distance       | $5.5 \times 10^{-14}$         | $2.8 \times 10^{-14}$                   | $3.8 \times 10^{-14}$               |
| Covariate            | $6.8 \times 10^{-14}$         | $2.9 \times 10^{-14}$                   | $2.77 \times 10^{-14}$              |
| Probability          | $7.7 \times 10^{-15}$         | $2.1 \times 10^{-15}$                   | $2.8 \times 10^{-14}$               |
| Self tuning           | $3.8 \times 10^{-15}$         | $1.5 \times 10^{-16}$                   | $2.4 \times 10^{-14}$               |
| Regularized          | $2.8 \times 10^{-15}$         | $1.26 \times 10^{-15}$                  | $1.03 \times 10^{-14}$              |

Fig. 4. Spectrum of the eigenvalues obtained when analyzing the dataset in Section IV-A1 (multimodal remote sensing) by means of spectral clustering methods based on different set-ups of the Laplacian matrix: fluid Laplacian matrix as in Section II-B2, unnormalized and normalized Laplacian matrix [31], Graph Distance-based Laplacian matrix [10], covariate-assisted Laplacian matrix [9], probability-based Laplacian matrix [11], self tuning Laplacian matrix [12], regularized Laplacian matrix [7].

Fig. 5. Spectrum of the eigenvalues obtained when analyzing the dataset in Section IV-A2 (multimodal brain-computer interface) by means of spectral clustering methods based on different set-ups of the Laplacian matrix: the same notation as in Fig. 4 applies here.

Fig. 6. Spectrum of the eigenvalues obtained when analyzing the dataset in Section IV-A3 (multimodal photovoltaic energy) by means of spectral clustering methods based on different set-ups of the Laplacian matrix: the same notation as in Fig. 4 applies here.
It is worth noting that $mP$ measures whether each detected community is assigned to the ground-truth label which is most frequent in the community without bias (that could affect the classic purity metric [79], [80]). Moreover, the upper bound of this metric is 1, which corresponds to a perfect match between the partitions $\Omega$ and $\Psi$. On the other hand, its lower bound is 0 and corresponds to a complete mismatch between partitions $\Omega$ and $\Psi$.

With the notion of weight of a node in mind, it is possible to modify the classic definition of the adjusted Rand index in order to take into account the topological configuration of the graph. In this respect, it is necessary to define for any subset of nodes $\Phi$ the quantity $\kappa(\Phi) = \sum_{t,u \in \Phi} w_t w_u$. Hence, the modified adjusted Rand index ($mARI$) can be written as follows [79], [80]:

$$mARI(\Omega, \Psi) = \frac{\sum_{i,j} \kappa(\omega_i \cap \psi_j) - \frac{1}{\kappa(\Phi)} \sum_j \kappa(\omega_i) \sum_j \kappa(\psi_j)}{\frac{1}{2} (\sum_i \kappa(\omega_i) + \sum_j \kappa(\psi_j)) - \frac{1}{\kappa(\Phi)} \sum_i \kappa(\omega_i) \sum_j \kappa(\psi_j)}$$

(25)

This metric interprets the evaluation of community detection performance in terms of the decisions that have been taken for the nodes in the graph. In particular, $mARI$ is used to assess the community detection performance against the case for which the detected clusters $\Omega$ would be randomly generated, taking into account the individual effect of each node in the graph and in a chance-corrected manner [79], [80]. Specifically, $mARI$ is upper bounded to 1. I.e., $mARI$ assumes the value 1 when $\Omega$ and $\Psi$ perfectly match. On the other hand, $mARI$ is equal to or less than 0 when the similarity between $\Omega$ and $\Psi$ is equal or less than what is expected from two random partitions.

Finally, to define the modified normalized mutual information index ($mNMI$), we should consider a modified joint probability of cluster $\omega_i$ and class $\psi_j$, which can be written as $\tilde{p}_{ij} = \frac{w_t}{\sum_{u \in \omega_i} w_u}$. Accordingly, we can define $mNMI = mNMI(\Omega, \Psi)$ as follows:

$$mNMI = \frac{-2 \sum_{i,j} \tilde{p}_{ij} \log(\sum_{i,j} \tilde{p}_{ij} \frac{\tilde{p}_{ij}}{\sum_j \tilde{p}_{ij} \sum_i \tilde{p}_{ij}})}{\sum_i \sum_j \tilde{p}_{ij} \log(\sum_j \tilde{p}_{ij}) + \sum_j \sum_i \tilde{p}_{ij} \log(\sum_i \tilde{p}_{ij})}$$

(26)

It is possible to notice that the metrics in (24), (25), and (26) all rely on the definition of weight of a node, that is supposed to help in emphasize the role of the graph topology in computing how well the proposed graph representation can help in extracting functional information from the given dataset. To this aim, several definitions of $w_u$ can be drawn out [79], [80]. In particular, three configurations can be considered in order to cover the main characteristics to be taken into account when assessing the relevance of a node in graph learning frameworks:

1) $w_u = d_u / \max_i d_i$, where $d_u$ identifies the degree of node $u$ - in this case, we can talk about a degree measure;
2) $w_u = \frac{d^{ext}_u}{d_u}$, where $d^{ext}_u$ is the internal degree of node, i.e. the number of connections it has in the cluster it belongs to - in this case, we can talk about an embeddedness measure;
3) $w_u = \frac{\delta^{RT}_u}{\max_i \delta_i}$ - in this case, we can talk about a weighted embeddedness measure.

Thus, when assessing the functional information retrieval performance of the different graph learning frameworks, we considered the metrics as defined according to these set-ups, i.e., three values for each metric, for a total of nine metrics for performance comparison. In particular, we used these metrics to compare the strategy we introduced in this work with the following state-of-the-art methods:

- clustering via hypergraph modularity (CNM) [81];
- hierarchical community detection (HCD) [82];
- community detection based on distance dynamics (Attractor) [83];
- joint criterion for community detection (JCDC) [84];
- a standard k-means algorithm, where $k = K_C$;
- variational Bayes community detection (VB) [85];
- node importance-based label propagation (NI-LPA) [86];
- fluid label propagation (FLP) [87];
- weighted stochastic block model (WSBM) [88];
- multiview spectral clustering (Multiview) [89];
- covariate-assisted spectral clustering (CASC) [90];
- regularized spectral clustering (RegularizedSC) [7];
- deep multimodal clustering (MMClustering) [58].

All these methods rely on graph representation based on heat diffusion. It is worth to recall that the method introduced in [87] does not show any overlap whatsoever with the strategy for community detection we introduce in this work. The authors in [87] do not discuss fluid diffusion indeed, nor introduce any novel graph representation of datasets.

Figures 7 to 9 report the results we achieved by assessing the ability of extracting functional information by applying clustering methods to the outcomes of graph learning frameworks. For each column we report the value of $mP$, $mARI$, and $mNMI$ computed according to the setting mentioned in the previous bullet point list. Independently by the configuration we used to assess the results, the proposed method is able to outperform the other schemes in all datasets and set-ups. Indeed, the approach we introduce appears to be more suitable to characterize multimodal data. In particular, it is worth noting that the proposed diffusion model provides a solid platform that can be used by several graph analysis approaches to effectively explore the data properties (stronger fluctuations of the proposed metrics are registered for heat diffusion-based schemes across different clustering algorithms). This is crucial to extract functional characteristics of the records taken into account, so that an analysis at semantic level can be accurately performed. Hence, the flexibility produced by the fluid diffusion model to analyze the graph representation provides a great advantage with respect to the heat diffusion-based approaches.

Finally, Fig. 10 displays the execution time (in log[sec]) for the aforesaid algorithms to achieve community detection on the multimodal remote sensing, brain-computer interface, and photovoltaic energy datasets: for each scheme, these results are shown in blue, red, and orange bars, respectively. It is possible
to appreciate how the size of the datasets is typically the driving force behind these outcomes. The proposed community detection algorithm based on fluid graph representation delivers a performance that is comparable with the other methods in this respect. Nevertheless, these results show how taking advantage of the modularity property of deep learning-based approach (such as that in [58]) could reduce the computational load of the architecture. Hence, to improve the scalability of the approach we presented in this work, a deep learning analysis relying on the proposed fluid graph representation will be considered in future works.

V. CONCLUSION

In this paper, we introduce a novel approach for graph representation with special focus of multimodal data analysis. The proposed scheme is based on the use of a fluid diffusion model to characterize the interactions among samples, and hence the mechanism for information propagation in graphs. This approach is meant to address several issues in modern multimodal data analysis, when large scale datasets collected by heterogeneous sources of information are investigated. In particular, the proposed framework aims to provide an accurate and versatile automatic characterization of the relationships among samples, so that a robust community detection can be derived for complex datasets where multiple statistical, geometrical, and semantic distributions are collected. In this respect, the main contributions of this work are:

- the introduction of a novel model for graph information propagation based on fluid diffusion;
- the development of a compact description of the interactions among data that takes advantage of the eigenanalysis;
of flow velocity matrix, so to guarantee a data driven set-up for multimodal data characterization;

- the development of an architecture for community detection based on fluid dynamics, which allows to obtain a solid characterization of the connections among samples in complex datasets (e.g., where samples show different levels of reliability and where the relevance of the feature might vary across the data).

We tested our approach on three diverse real multimodal datasets in terms of functional information retrieval. The experimental results we achieved show the solidity of our approach, as the proposed framework is able to outperform the state-of-the-art methods in community detection, which are all based on heat diffusion model. Thus, it is possible to state that the fluid diffusion model could be a valid option to improve the characterization of multimodal data analysis and to enhance the understanding of the functions and phenomena underlying the multimodal records, so to fully exploit the potential provided by the diversity of modalities collected in the datasets under exam. This approach can thus represent the platform on which multimodal data analysis could be based to address the main issues in modern multimodal data analysis. Future works will be then devoted to explore the development of fluid diffusion-based data analysis schemes for specific tasks, from semisupervised learning to explainable data analysis, to prediction and inference in complex operational scenarios. Moreover, the results we achieved might signal the need to invest the geometry of multimodal spaces by means of new computational strategies, which theoretical and methodological basis will be studied in the next steps of this work.

VI. ACKNOWLEDGEMENTS

This work is funded in part by Centre for Integrated Remote Sensing and Forecasting for Arctic Operations (CIRFA) and the Research Council of Norway (RCN Grant no. 237906), the Visual Intelligence Centre for Research-based Innovation funded by the Research Council of Norway (RCN Grant no. 309439), the Automatic Multisensor remote sensing for Sea Ice Characterization (AMUSIC) Fransenteret ”Polhavet” flagship project 2020, and the IMPETUS project funded by the European Union Horizon 2020 research and innovation program under grant agreement nr. 101037084.

APPENDIX

A. Motivations of a new graph representation: a multimodal example

In this Section, we provide an experimental example to support the statements we made in Section II-B1 to support the need for a new graph representation for multimodal datasets.

To this aim, let us consider a multimodal dataset that is considered a benchmark in the remote sensing community [40]. This dataset consists of hyperspectral and Lidar observations acquired over a rural area in the south of the city of Trento, Italy. Hence, by means of the hyperspectral sensor it is possible to obtain a characterization of the physical composition of the scene. As such, this source of information can help in discriminating between different elements in the considered region of interest, e.g., vegetation vs. asphalt. On the other hand, LiDAR observations lead to the generation of digital surface models (DSMs), which provide details on the height of the objects showing up in the considered area. Therefore, combining these records can help in obtaining an accurate characterization of specific elements (e.g., tree species) on the Earth’s surface. The main properties of this dataset are summarized as follows: i) the size of the dataset is 600 × 166 pixels (each pixel is a sample of the dataset); ii) the sensor used to generate the LiDAR DSM and the associated features (summing up to 14 features in total) was Optech ALTM 3100EA sensor, having spatial resolution of 1m; iii) the hyperspectral data acquired by AISA Eagle sensor consist of 63 bands ranging from 402.89 to 989.09nm, where the spectral resolution is 9.2nm, with spatial resolution of 1m; iv) six classes of interests were extracted, including Building, Woods, Apple trees, Roads, Vineyard, and Ground.

Fig. [11] displays the RGB composite of this dataset, as well as its groundtruth. It is worth noting that pixels in which geographical unit more materials associated with the aforementioned six thematic classes in the dataset are grouped in the "Background", which is not used for learning and performance comparison purposes. Furthermore, the geographical extent of the dataset avoids the occurrence of heterogeneous atmospheric effects throughout the data. Hence, it is possible to state that the atmosphere composition could show uniform properties across the whole region, hence avoiding the need for focused correction processing across the records. Moreover, the classes showing up in the dataset are pretty distinct from each other, both in the spectral and spatial domains. However, it is possible to prove that analyzing this dataset by means of classic graph representation might not lead to accurate and reliable understanding of the considered scene. Indeed, although acquired in ideal conditions, the diversity of the records make the classic graph representation based on heat diffusion model not adequate for the characterization of this dataset, as it can be proven by considering the description in Section II-B1 and especially computing the trend of the quantities in [49].

Specifically, let us start to consider the relevance of the features for each class in the dataset. In fact, investigating and selecting the most relevant features in the samples associated with each class we can obtain an initial estimate of how the different records could contribute to the characterization of the dataset. To this aim, considering an algorithm for feature selection that can take into account the nonlineairties and sparsity of the information across the samples can help in understanding the actual interactions among features, hence their statistical relationships. Therefore, without losing generality, we can run over the samples associated with each class the algorithm for adaptive dimensionality reduction proposed in [89] to select the most relevant features for each subregion showing homogeneous properties in the dataset in Fig. [11].

In order to visualize the features (i.e., spectral channels in case of the hyperspectral sensor, properties of the digital surface model for the LiDAR sensor) that are selected for different pixels belonging to the same class, we can use a circular
Fig. 11. Multimodal remote sensing dataset acquired over the region of Trento, Italy [40]: RGB composite (top) and groundtruth map (bottom).

Fig. 12. Circular graph plot showing the selected features for five pixels associated with the class “Apple trees” in the dataset in Fig. 11. The features selected for each pixel by means of the algorithm in [64] are displayed by lines of different colors connecting circles associated with the corresponding indices, according to Table III.

Let us take a closer look to Fig. 12. In particular, it is possible to observe that each considered pixel is characterized by a different subset of features, as the lines of different colors connect distinct circles (i.e., features) in the disk. This is a very interesting result, since one might expect that pixels belonging to one single class would be characterized by one very homogeneous set of features. Fig. 12 clearly contradicts this intuition. Indeed, the graph in Fig. 12 results from the complex set of nonlinearities induced by different conditions of illumination, scattering, morphology that are intrinsically affecting the observations of the multimodal remote sensing dataset, and that are taken into account by the feature selection algorithm in [89]. As such, this method is able to highlight how different subsets of features are actually channeling the relevant information for different pixels. In particular, it is worth noting that the selected features are not only different from pixel to pixel, but also associated with very distinct observations in the dataset. For instance, to characterize the pixel associated with the red line in Fig. 12 the features provided by the LiDAR sensor are apparently playing a key role, whilst the records acquired in the near infrared portion of the electromagnetic spectrum are not so significant. On the other hand, the pixel associated with the blue line in Fig. 12 shows a dual behavior, i.e., the features in the near infrared seem very relevant for it characterization, whilst the LiDAR records could be discarded.

The results in Fig. 12 provide material to discuss the composition of the matrices and vectors in (3). In fact, since the relevant features for each pixel belonging to one class might vary substantially, it is possible to expect that the off-diagonal elements of the covariance matrices $C_i$ might show not negligible values. As a result, the energy will not be concentrated in the diagonal elements of these matrices, which is implicitly a major factor for the conditions in (3) to hold [32]. In fact, when progressively computing the quantities in (3) as the number of samples increases, it is possible to obtain their trends: Fig. 13 reports the results we achieved as the number of pixels used in their computation increases.

| Index | Meaning                                      |
|-------|----------------------------------------------|
| 1 - 14| LiDAR                                        |
| 15 - 19| 402.89 nm - 440.7 nm [Violet]                |
| 20 - 24| 450.16 nm - 487.98 [Blue]                    |
| 25 - 32| 497.43 nm - 563.62 nm [Green]                |
| 33 - 35| 573.07 nm - 591.91 nm [Yellow]               |
| 36 - 39| 601.44 nm - 629.8 nm [Orange]                |
| 40 - 45| 639.2 nm - 686.53 nm [Light red]             |
| 46 - 52| 695.99 nm - 752.71 nm [Dark red]             |
| 53 - 77| 762.17 nm - 989.09 [Near Infrared]          |
Specifically, the values of $|m'|$ and $l_1$ have been computed for the class "Apple trees" (top and middle plots, respectively). On the other hand, the bottom plot in Fig. 13 displays the trend of the maximum value of the matrix $T$.

As these trends are valid also for the other classes in the dataset, it is therefore apparent that the quantities in (3) converge to a finite non-zero value. More importantly, this proves that the values of the parameters in (3) would not diverge to infinity, hence jeopardizing (according to the discussion we previously reported) the validity and appropriateness of a classic graph representation based on heat diffusion model for the considered dataset in Fig. 11. As a result, carrying out a characterization of the dataset by means of classical graph-based data analysis (i.e., processing and investigating the data over a classic graph representation structure) might not be the best choice, and indeed might lead to a strong degradation of the overall analysis.

To further support this statement, we analyzed the dataset in Fig. 11 by means of state-of-the-art supervised and semi-supervised methods based on the classic graph representation introduced in Section II-A [90–93]. Moreover, for comparison, we used a method based on ensemble learning approach [64]. It is worth noting that deeply investigating the capacity and limitations of all these methods is out of the scope of this paper. Nonetheless, taking a look to the results that these algorithms provide when analyzing the dataset in Fig. 11 with comparable hyperparameter set-ups can help us in consolidating the statements and observations we previously made.

In particular, focusing once again on the "apple trees" class, it is possible to compute the average misclassification error between the "apple trees" and "wood" classes obtained over all the experiments we ran: Table IV reports these results. It is therefore possible to appreciate that the algorithms based on the analysis of the graph structure defined according to the guidelines in Section II-A show a substantial increase of the error with respect to the ensemble learning-based architecture in [64]. Hence, the classic graph representation based on heat diffusion model apparently leads to a substantial degradation of the ability of the architectures to characterize and interpret the samples in the dataset. Recalling the considerations we previously made on the quality of the Trento dataset, these error distributions emphasize that the classic graph representation might not be adequate for a lot of situations encompassed by modern data analysis, leading to unacceptable degradation in the characterization performance.

| Method                                      | Error    |
|---------------------------------------------|----------|
| Graph convolutional network (GCN) [91]     | 14 ± 4.3 |
| Simple GCN (S-GCN) [92]                    | 12.3 ± 2.5 |
| Deep neural network                        | 11.8 ± 1.9 |
| Chebyshev polynomial approximation (ChebNet) [90] |          |
| Scalable inception graph neural network (SIGN) [93] |         |
| Ensemble learning with adaptive dimensionality reduction (ADR-EL) [64] | 2.1 ± 1.2 |

### B. Derivation of the Fokker-Planck equation for fluid diffusion in porous media

In this Section, we report the steps that are taken to derive the Fokker-Planck equation for fluid diffusion in porous media starting from the analysis of the diffusion process in the form of (4) by means of the Itô’s lemma [46]. In particular, let us consider (9). Since the function $g(x)$ is arbitrary by construction, this equation is satisfied when the following condition applies:

$$\frac{\partial p(x,t)}{\partial t} + \nabla \cdot [a(x)p(x,t)] = \beta(x,t),$$  \hspace{1cm} (27)

where $\beta(x,t) = \sum_{i,j=1}^n \partial x_i \partial x_j (\tilde{B}_{ij} p(x,t))$. Let us focus on each term of this sum: we can then write as follows:

$$\partial x_i \partial x_j (\tilde{B}_{ij} p(x,t)) = \left[ \partial x_i \partial x_j (\tilde{B}_{ij}) \right] p(x,t) + \left[ \partial x_i \tilde{B}_{ij} \right] \partial x_j p(x,t) + \partial x_i [\tilde{B}_{ij} \partial x_j p(x,t)].$$  \hspace{1cm} (28)

At this point, $\beta(x,t)$ becomes:

$$\sum_{i,j=1}^n \partial x_i \partial x_j (\tilde{B}_{ij} p(x,t)) = \left[ \nabla \cdot (\nabla \tilde{B}(x)) \right] p(x,t) + \left[ \nabla \tilde{B}(x) \right] \cdot \nabla p(x,t) + \nabla \cdot [\tilde{B}(x) \nabla p(x,t)].$$  \hspace{1cm} (29)

where $\nabla = \left[ \frac{\partial}{\partial x_i} \right]_{i=1,\ldots,n}$. Substituting this term in (27), then we obtain:
\[ \frac{\partial p(x, t)}{\partial t} + \nabla \cdot [a(x)p(x, t)] = \big[ \nabla \cdot (\nabla B(x)) \big] p(x, t) \]
\[ = \big[ \nabla B(x) \big] \cdot \nabla p(x, t) - \nabla \cdot \big[ B(x) \nabla p(x, t) \big] , \]

i.e., eq. (10). This hence leads to the formulation of the diffusion process represented by (4) in terms of the Fokker-Planck equation for fluid diffusion in porous media.

**C. Derivation of transition probability density function in the fluid diffusion model**

In this Section, we report the details of the derivation of the transition probability density function \( p(x(t + \epsilon) = x_j | x(t) = x_i) \) characterizing the fluid diffusion model as described in Section II-B2.

The time domain random walk strategy \([47],[48]\) solves the Green function problem derived from (11) by imposing initial conditions and absorbing boundary conditions to the diffusion system centered on the node \( i \). In other terms, this strategy considers boundary conditions at the outer edges of the \( i \)-th node and a pulse initial condition of unit mass at \( x = 0 \), where the \( i \)-th node is located. Furthermore, we assume to locate the \( j \)-th node at the right boundary \( l^+ \), while the remaining nodes in the neighborhood of node \( i \) \( \mathcal{N}(i) \) (i.e., the nodes adjacent to node \( i \)) are located at the left boundary at \(-l^-\), where \( l^+, l^- > 0 \). Fig. 14 displays the aforesaid set-up. Hence, the boundary and initial conditions of the Green function \( G(x, t) \) (solution of the system in (11)) are defined as follows:

\[ G(l^-, t) = G(l^+, t) = 0 \]
\[ G(x, 0) = \delta(x) , \quad (30) \]

where \( \delta(\cdot) \) is the Dirac delta. Moreover, solutions for the Green function problem can be written as follows:

\[ G(x, t) = G_-(x, t)\Theta(-x) + G_+(x, t)\Theta(x), \quad (31) \]

where \( \Theta(x) \) identifies the Heaviside step function \([47]\). At this point, we can impose the continuity condition of \( G(x, t) \) at \( x = 0 \), which implies that \( G_-(0, t) = G_+(0, t) \), according to the aforesaid notation.

With this in mind, it is possible to derive the definition of \( G(x, t) \) by considering the projection in the Laplace space of the fluid diffusion system. In fact, the Laplace transform of (11) over the time variable (denoted by \( G^\mathbb{L} \)) with the condition \( G(x, 0) = \delta(x) \) leads to the following \([47],[48]\):

\[ \xi G^\mathbb{L}(x, \xi) + \nabla \cdot [v(x)G^\mathbb{L}(x, \xi)] - \nabla \cdot \tilde{B}(x)\nabla G^\mathbb{L}(x, \xi) = \delta(x). \quad (32) \]

Moreover, it worth recalling that the \( G(x, t) \) function can be expressed as in (31). As such, it is possible to write as follows:

\[ G^\mathbb{L}_\pm(x, \xi) = A^{(1)}_\pm(\xi) \exp \left[ \frac{v_\pm x(1 - \alpha_\pm(\xi))}{2B_\pm} \right] \]
\[ + A^{(2)}_\pm(\xi) \exp \left[ \frac{v_\pm x(1 + \alpha_\pm(\xi))}{2B_\pm} \right] , \quad (33) \]

where \( G^\mathbb{L}_-(x, \xi) \) and \( G^\mathbb{L}_+(x, \xi) \) represent the Laplace transform of \( G_-(x, t) \) and \( G_+(x, t) \), respectively. Moreover, taking into account the quantities that have been defined in Section II-B2 we can define \( v_\pm \) and \( B_\pm \) as follows:

\[ v_+ = v_{ij}, \quad v_- = \sum_{m \in \mathcal{N}(i) \setminus j} \frac{v_{im}}{|\mathcal{N}(i) - 1|}, \quad (34) \]
\[ \tilde{B}_+ = \tilde{B}_{ij}, \quad \tilde{B}_- = \sum_{m \in \mathcal{N}(i) \setminus j} \frac{\tilde{B}_{im}}{|\mathcal{N}(i) - 1|}, \quad (35) \]

where \( \mathcal{N}(i) \) identifies the neighborhood of node \( i \), i.e., the set of nodes adjacent to node \( i \). Finally, we define \( \alpha_\pm(\xi) = \text{sign}(v_\pm) \sqrt{1 + 4B_\pm \xi / v_\pm^2} \).

Recalling that \( G(i^-, t) = G(i^+, t) = 0 \), this equation can be rewritten as follows:

\[ G^\mathbb{L}_\pm(x, \xi) = A^{(1)}_\pm(\xi) \exp \left[ \frac{v_\pm x}{2B_\pm} \right] \left\{ \exp \left[ -\frac{v_\pm x}{2B_\pm} \alpha_\pm(\xi) \right] \right\} \quad (36) \]

Recalling the condition for which \( G_-(0, t) = G_+(0, t) \), it is possible to write \( G^\mathbb{L}_- \) and \( G^\mathbb{L}_+ \) as follows:

Fig. 14. Diffusion system centered on node \( i \) to analyze the transition probability to node \( j \) as set by the time domain random walk approach \([47]\). The red box identify the set \( \mathcal{N}(i) \setminus j \).

- \( l^- \)
- \( 0 \)
- \( l^+ \)
\[ G_Z^G(x, \xi) = \frac{1}{2} A(\xi) \left\{ \begin{array}{l} 1 - \exp \left[ \frac{v_- - v_+}{B_-} \alpha_-(\xi) \right] \exp \left[ \frac{v_+ - x}{2B_+} \right] \\
- \left( v_+ \alpha_+(\xi) \right) \\
- 1 - \exp \left[ \frac{v_- - v_+}{2B_+} \alpha_-(\xi) \right] \end{array} \right\}. \]  

Thus, it is possible to summarize these equations as follows (where for convenience the variable \( A(\xi) \) is renamed \( A(\xi) \)):

\[ G_Z^G(x, \xi) = A(\xi) \sinh \left[ (l^\pm \pm x) v^\pm_\alpha(\xi) \right] \text{csch} \left[ v^\pm_\alpha(\xi) \right], \] 

where \( v^\pm_\alpha = l^\pm v_+/2B_\pm \) and \( \text{csch}[z] = 1/\sinh[z] = 2/(\exp[z] - \exp[-z]) \). At this point, it is worth to recall that the definition of the Green function \( G(x, t) \) is regulating the diffusion process through the nodes. Furthermore, the functions \( G_0 \) and \( G_1 \) determine the fluxes through the boundaries that the time domain random walk approach imposes. Hence, they can be used to determine the first arrival time densities at the boundaries, which can be written as follows:

\[ \phi^\pm(t) = 2B_\pm \nabla G_Z^G(x, t)|_{x=\pm t^\pm}, \]  

where \( \nabla = \sum_{|i|=1,\ldots,n} \alpha_i = 1, \ldots, n \).

According to the notation that we used so far, \( \phi^\pm(t) dt \) would then denote the joint probability of the transition to occur towards the right boundary (i.e., from node \( i \) to node \( j \)) with an arrival time in \( t, t + \epsilon \) \( [4,7, 48] \). As such, we can then define the transition probability density \( p(x(t + \epsilon) = x_j | x(t) = x_i) \) as follows \( [47] \):

\[ p(x(t + \epsilon) = x_j | x(t) = x_i) = p_+ = \int_0^{+\infty} \phi_+ dt. \]  

Therefore, it is crucial to derive the analytical expression of the \( A(\xi) \) parameter in \( [48] \) so that the transition probability in \( [40] \) can be retrieved. To this aim, let us integrate \( [32] \) over \( x \); this would lead to the following equation:

\[ \xi \int G_Z^G(x, \xi) dx = 1 + A(\xi) \left\{ v_\alpha(\xi) \sinh \left[ \frac{v_- - v_+}{2B_-} \right] - v_\alpha(\xi) \exp \left[ \frac{v_-}{2B_-} (1 - \alpha_-(\xi)) \right] \right\}. \] 

On the other hand, considering the definition of \( G_Z^F(x, \xi) \) and \( G_Z^G(x, \xi) \) in \( [36] \) and \( [38] \), integrating over \( x \) and multiplying by \( \xi \) the Laplace transform of \( [31] \), we can write as follows:

\[ \xi \int G_Z^F(x, \xi) dx = \frac{1}{2} A(\xi) \left\{ v_\alpha(\xi) \exp \left[ \frac{v_- - v_+}{2B_-} \alpha_-(\xi) \right] + \left[ 1 - \exp \left[ \frac{v_-}{2B_-} (1 - \alpha_-(\xi)) \right] \right] \right\}. \] 

By equating \( [42] \) and \( [43] \), it is possible to obtain the definition of \( A(\xi) \) as follows:

\[ A(\xi) = \left\{ \sum_{u \in \pm^1} v_u \left[ \alpha_u(\xi) \text{coth} \left( v^\pm_\alpha(\xi) \right) + \bar{u} \right] \right\}^{-1}, \]  

where \( \bar{u} = -1 \) if \( u \) is \( - \), and \( \bar{u} = +1 \) if \( u \) is \( + \). Consequently, we can derive the Laplace transform of the \( \phi \) functions in \( [39] \).

In particular, \( \phi_+^F(\xi) \) can be written as follows:

\[ \phi_+^F(\xi) = v_\alpha^+ \alpha_+(\xi) A(\xi) \text{csch} \left[ v^\pm_\alpha(\xi) \right]. \]  

Finally, we can derive the value of \( p_+ \) in \( [40] \) that can be written as follows:

\[ p_+ = \frac{\sum_{u \in \pm^1} |v^\pm_u| \exp[|v^\pm_u|] \text{csch}(|v^\pm_u|)}{\sum_{u \in \pm^1} |v^\pm_u| \exp[|u^\prime \cdot v^\prime_u|] \text{csch}(|v^\prime_u|)}. \]
D. Effect of the Q matrix: a visual example

To visualize how the proposed definition of the Q matrix (introduced in Section II-B2) could help in obtaining more accurate and reliable characterization of the data interactions in the considered dataset, let us consider a toy example. In particular, we synthetically generated a small dataset consisting of 15 samples. Each sample was characterized by 77 features. Specifically, the dataset was set up as follows. Initially, 8 samples, 3 samples, and 4 samples were associated with three sets of features (or signatures) corresponding to three pixels labeled as "Vineyard", "Road", and "Wood" in the Trento dataset that has been previously mentioned (see Fig. 11), respectively. Then, we applied noise to the samples in a not uniform fashion across the features: the graph induced by the resulting dataset can be summarized in Fig. 15, where the nodes represent the samples and the width of the edges is proportional to the similarity between the nodes they connect. It is worth noting that in principle the graph would be fully connected. However, for sake of visualization, we avoided to display in Fig. 15 the connections having weights close to 0.

At this point, we computed the Q matrix and the weight matrix W according to the classic graph representation based on heat diffusion mechanism. For the Q matrix, we considered the definition as in (13), we used the method in [64] to determine the values of the K tensor, and assumed that the distribution of the ˜B matrix would be uniform. For the classic definition of the weight matrix W based on heat diffusion mechanism, we used a Gaussian kernel to define the function η mentioned in Section II-A [6], [31]. We displayed the heatmaps associated with the Q and W matrices in Fig. 16 and [7] respectively: for sake of visualization, the values of the elements were quantized in 10 levels. Moreover, we can recall that high values in these matrices (i.e., lighter colors in the heatmaps) mean high degree of similarity between the considered samples.

It is worth noting that by construction the ideal configuration of the matrix summarizing the similarities among samples should show as a symmetric diagonal block matrix, since the samples are originally associated with three distinct classes. The Q matrix clearly shows this property. In fact, the Q matrix is apparently able to retrieve the main characteristics of the informative groups of samples although the noise we added to the samples' signatures we used. On the other hand, it is possible to appreciate from Fig. 17 that the classic approach to derive the similarities among samples would lead to a more confused distribution of the weights for graph characterization, leading to a less accurate understanding of the samples' interactions. In particular, the higher precision of the Q matrix directly translates in a more accurate eigenanalysis of the associated Laplacian matrix, which is crucial to identify the relevant features and the informative patterns within the data, i.e., to retrieve a reliable information extraction from the considered datasets [6], [31], [32]. As such, the Q matrix seems to represent a valid candidate to achieve a more accurate and reliable understanding of the considered datasets.

E. Community detection algorithms: a short review

As mentioned in Section III-A, community detection algorithms can be grouped in categories defined according to the strategy they employ [6]. In particular, it is possible to categorize these methods in seven main groups: 1) graph partitioning; 2) hierarchical clustering; 3) partitional clustering; 4) spectral clustering; 5) dynamic community detection; 6) statistical inference-based community detection; and 7) hybrid methods. It is worth noting that these methods have been originally developed considering heat diffusion to derive the graph representation. Future works could be dedicated to explore the opportunity to develop these algorithms while considering the graph representation based on fluid dynamics proposed in this paper. The following subsections report the main characteristics of these classes.

1) Graph partitioning: Graph partitioning methods aim to minimize the number of edges lying between the groups, so to discriminate communities of predefined size in the dataset [6].
Graph partitioning problem has been historically addressed as an optimization problem where a function (e.g., modularity, connectivity, conductance, ratio cut, normalized cut) of the difference between the number of edges inside the modules and the number of edges lying between them must be maximized \[0, 7]. \[94, 97\]. It is worth to recall that graph partitioning algorithms require to provide as input the number of clusters (and in some cases also some additional information on the clusters, such as size or degree of compactness) \[6, 81, 94, 96\]. This is a major drawback in several application scenarios, as in most of the cases no information on the characteristics of the data to be explored is available: this is particularly true when addressing operational scenarios. Also, these algorithms are typically prone to high sensitivity with respect to the initial state of the optimization process (e.g., bisectioning the graph) \[81, 97, 99\]. Hence, to overcome these issues graph partitioning algorithms are often combined with hierarchical, spectral and/or dynamic community detection principles.

2) Hierarchical clustering: Several times the graph might show a hierarchical structure, where it is possible to identify a tree-like interplay among the communities in the dataset. Hierarchical clustering aim to exploit this property to detect the communities in the dataset \[6\]. Two main categories of hierarchical clustering techniques can be mentioned: i) agglomerative algorithms, i.e., clusters are iteratively merged if their similarity is sufficiently high; ii) divisive algorithms, i.e., clusters are iteratively split by removing edges connecting vertices with low similarity. These approaches are intrinsically opposite strategies, since agglomerative algorithms are bottom-up strategies (i.e., nodes are initially considered to be all separate clusters and are iteratively merged according to their similarity), while divisive algorithms are top-down (i.e., nodes are initially considered to be all part of a single cluster and are iteratively separated according to their similarity). Although these schemes have shown good performance in several application scenarios, it is true that they are very sensitive to design choices (e.g., choice of similarity metric). Moreover, to reduce the latency of the algorithms, nonlinear processing solutions have been considered, although this leads to higher sensitivity of the initial conditions \[6, 82, 84, 100, 102\].

3) Partitional clustering: Another popular approach for community detection is represented by so-called partitional clustering. In this case, the nodes of the graph are mapped into a multidimensional space, and the edges connecting nodes show weights proportional to the distance between the nodes in this space \[6, 103\]. This distance is thus used as a metric to assess the dissimilarity among nodes, so that it can be used to design objective functions to be optimized in order to identify the group of nodes that would maximize a similarity condition. The methods belonging to this category differentiate according to the function to be optimized. The most popular algorithm in this class is k-means scheme, and its fuzzy variant \[6, 103, 104\]. These schemes are typically brilliant in terms of efficiency, as the mapping and the optimization tasks can be performed very rapidly and with low computational complexity. On the other hand, these algorithms require to specify the number of clusters to be detected, which might not be always possible in practical scenarios, as previously mentioned. Additionally, the mapping onto the multidimensional space used for the optimization step must be properly designed according to the given dataset, in order to avoid the introduction of artificial bias for the community detection task \[6, 104\].

4) Spectral clustering: Spectral clustering methods rely on the definition of a similarity matrix \(W\) that aims to summarize the affinity of the samples \[31\]. Typically, the similarity matrix can be defined in terms of distance between samples computed according to a specific criterion (traditionally, Euclidean distance in the multidimensional feature space), although several metrics have been proposed in technical literature to measure the similarity between samples \[7, 12\], \[31\]. Spectral clustering methods investigate the space induced by the eigenvectors of the Laplacian matrix (introduced in Section II-A), associated with \(W\). The most common definition of the Laplacian matrix is \(D - W\), where \(D\) is the diagonal matrix whose element \(D_{ii}\) equals the degree of the \(i\)-th node in the graph, i.e., the sum of the weights in the \(i\)-th row of \(W\).

It is possible to appreciate that the sum of the elements of each row of the Laplacian is zero by construction. This implies that the Laplacian matrix would always show at least one zero eigenvalue, corresponding to the eigenvector with all equal components. Moreover, eigenvectors corresponding to different eigenvalues are all orthogonal to each other \[8, 25, 26, 31, 105\]. It is possible to state that the ideal case of \(K\) non-overlapping communities showing up neatly separated in the graph would lead to \(K\) eigenvalues of the Laplacian matrix equal to zero. On the other hand, in a realistic scenario, \(K\) subgraphs would be weakly linked to each other, so that the spectrum of the unnormalized Laplacian will have one zero eigenvalue, while the others would be positive \[26, 105\]. Nevertheless, it is possible to prove that the lowest \(K - 1\) eigenvalues would still be close to zero. Hence, identifying
the $K + 1$-th eigenvalue that is clearly different from zero would lead to a robust community detection. Furthermore, the detection of significant gaps between eigenvalues is crucial for the effectiveness of spectral clustering algorithms, as well as for their efficiency [7], [8], [25], [26], [51], [105]. To address this problem several spectral clustering methods have been introduced in technical literature [7]–[12], [26]. However, it is also true that these schemes might show some limitations in addressing community detection in large scale and sparse graphs, where the separability of the aforesaid eigenvalues may be cumbersome to guarantee and achieve [7], [8].

5) Dynamic community detection: Another approach for community detection relies on the representation of graphs as results of dynamic processes occurring from node to node, so that the nodes that minimize the energy of the given process would be assumed to belong to the same community [106]–[109]. These methods are typically based on a Markov random field description of the state dynamics, so that higher order moment of informativity can be incorporated in iterative detection schemes in a very straightforward way [106], [107]. These algorithms are typically pretty easy to implement, which represent the main properties for their success in scientific community. On the other hand, it is also true that their complexity might become very high when sparse graphs are taken into account. Moreover, these algorithms are typically prone to high sensitivity to initial conditions and might not be reliable when communities’ distribution is very skewed (i.e., high degree of imbalanced data to be analyzed), such that the convergence of the methods might be jeopardized [6], [107], [109].

6) Statistical inference-based community detection: In this category, the methods aim to fit the graph topology to a (set of) statistical hypotheses that are drawn according to the diverse set-ups that could be considered. These methods employ different strategies in achieving this goal. Nonetheless, it is also true that these schemes rely on a common trait, i.e., the definition of the transition probability (used to define the inference mechanisms) as a function of the similarity between nodes [65]. These functions might be linear or nonlinear, or eventually based on higher order moments of the features to be considered [110]–[112]. The main idea that drive these methods is that the nodes of the same group are linked with a probability $p_x$, while nodes of different groups are linked with a probability $p_y$. Indeed, $p_x$ and $p_y$, as well as the number of the communities, should be input to the system a priori: this might limit the impact of the statistical inference-based architectures, although in recent times methods for the automatic optimized selection of these probabilities have been proposed [85], [110], [113]. The complex characteristics that might show up in the dataset (especially in operational scenario) might provide a major issue of these algorithms. In fact, the probabilities used in these schemes are typically assumed to be independent. As previously mentioned, this might not be matched by the considered dataset: this is particularly true when multimodal community detection is addressed [6], [110].

7) Hybrid methods: Finally, it is worth to recall that several methods for community detection that have been used in diverse research fields rely on the combination of multiple strategies that have been previously summarized in this Section (e.g., hierarchical clustering and graph partitioning, spectral clustering and partitional clustering), with the ultimate goal to combine the major benefits of the considered techniques whilst overcoming their possible drawbacks.

The local optimization based on objective functions retrieved from combining modularity and compactness is used to design community detection methods [86]–[88], [114]–[117]. The local nature of the process enables high efficiency and low computational complexity, although increasing the sensitivity to initial conditions [118]. Leveraging on edge betweenness in the regions close to boundaries might reduce the impact of this issue. This idea has been extended to manifold analysis, where similarities among nodes are computed in terms of geodesic distances. Communities appear as portions of the graph with a large curvature [119]. The manifold approach has been proven to be successfully applied to massive datasets obtained form one single source of information (e.g., social media networks) [88], [114]. However, they also requires the communities to show up uniformly across the dataset, as well as the features to be affected by linear perturbations across the samples [6]. This is typically a major limiting factor for this approach.

Deep learning-based community detection has then triggered the attention of a large number of scientists worldwide, either in terms of autoencoder-based structures or deep investigation of latent features by means of convolutional neural networks [55], [120]–[125]. In fact, embedding communities in deep learning fashion has proven to be an efficient approach, since the network aims to learn node distributions of communities in a low-dimensional space [122], [124]. Nevertheless, there are several open issues that can be identified for this community detection strategy [55]. In fact, in order to perform accurate community detection, deep learning techniques require the number of communities to be identified as an input parameter of the processing, so that the analysis of the latent space they rely on can be carried out accurately [49], [55]. Moreover, the different probability distributions associated with each data type to be processed need to be addressed at the design step of models and algorithms. This is a relevant limiting factor for the use of deep learning-based community detection, especially when complex and/or multimodal datasets are considered [55], [120], [125]. Finally, the community detection performance of these methods can be jeopardized when unbalanced datasets (eventually showing non-uniform statistical properties across the records) are analyzed.

F. Codes for performance comparison

The results in Fig. 79 have been produced by using codes that can be found at these websites: https://github.com/HPAI-BSC/Fluid-Communities; https://gist.github.com/pszufe; https://github.com/michaelschaub/michaelschaub.github.io; https://github.com/chocolates/Community-detection-based-on-distance-dynamics; https://aaronclauset.github.io/wsbm/; https://github.com/norbertbin/SpecClustPack; https://github.com/DTaoo/DMC.
[114] B. Long et al. A probabilistic framework for relational clustering. In Proc. ACM international conference on Knowledge discovery and data mining (SIGKDD), San Jose, CA, 2019.

[115] A. Clauset. Finding local community structure in networks. Phys. Rev. E, 72(026132), 2005.

[116] W. Li et al. Stepping community detection algorithm based on label propagation and similarity. Physica A: Statistical Mechanics and its Applications, 472:145–155, 2017.

[117] U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E, 76(036106), 2007.

[118] J. Liu et al. An approach of community search with minimum spanning tree based on node embedding. Complexity, (6673444), 2021.

[119] W. Leal et al. Forman-ricci curvature for hypergraphs. arXiv e-prints, page arXiv:1811.07825, 2018.

[120] P. Xu et al. Link Prediction with Signed Latent Factors in Signed Social Networks. In Proc. ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Anchorage, AK, 2019.

[121] V. Bhatia and R. Rani. Dfuzzy: A deep learning-based fuzzy clustering model for large graphs. Knowl. Inf. Syst., 57(1):159–181, 2018.

[122] S. Xue, J. Lu, and G. Zhang. Cross-domain network representations. Pattern Recognit., 94:135–148, 2019.

[123] Y. Xie et al. Community discovery in networks with deep sparse filtering. Pattern Recognit., 81:50–59, 2018.

[124] Y. Li et al. Community Detection in Attributed Graphs: An Embedding Approach. In Proc. AAAI Conference on Artificial Intelligence, New Orleans, LA, 2018.

[125] X. Shen and F. Chung. Deep network embedding for graph representation learning in signed networks. IEEE Trans. on Cybernetics, 50(4):1556 – 1568, 2020.