Social Network Analysis and Mining (2021) 11:99
https://doi.org/10.1007/s13278-021-00823-2

ORIGINAL ARTICLE

X-Mark: a benchmark for node-attributed community discovery algorithms

Salvatore Citraro[1,2] · Giulio Rossetti[2]

Received: 3 February 2021 / Revised: 22 July 2021 / Accepted: 23 July 2021 / Published online: 15 October 2021

Abstract
Grouping well-connected nodes that also result in label-homogeneous clusters is a task often known as attribute-aware community discovery. While approaching node-enriched graph clustering methods, rigorous tools need to be developed for evaluating the quality of the resulting partitions. In this work, we present *X-Mark*, a model that generates synthetic node-attributed graphs with planted communities. Its novelty consists in forming communities and node labels contextually while handling categorical or continuous attributive information. Moreover, we propose a comparison between attribute-aware algorithms, testing them against our benchmark. Accordingly to different classification schema from recent state-of-the-art surveys, our results suggest that *X-Mark* can shed light on the differences between several families of algorithms.

Keywords Network models · Synthetic benchmarks · Labeled community discovery · Node-attributed community discovery

1 Introduction

Networks are the natural way to express phenomena whose unit elements exhibit complex interdependent organization. During the last decades, the availability of data expressing meaningful complex structures has increased significantly; hence, the definition of *network science [as] the study of the collection, management, analysis, interpretation, and presentation of relational data* (Brandes et al. 2013), built on top of the mathematical tools of graph theory. Among the massive number of complex network fields and sub-fields, community discovery (henceforth, CD) is one of the most important and critical tasks, aiming to group the actors of a system according to the relations they form. The lacking of general criteria—from the *ill-posed* definition of *community* to the uncountable number of alternative approaches—leads to the challenging problem of evaluating the quality of the resulting CD partitions. Classically, both internal measures and external methodologies have been provided to test the goodness or the quality of the CD algorithms. An internal evaluation adopts a quality measure to assess the well-defined structural segmentation of the communities; conversely, an external evaluation aims to estimate the agreement between the communities and a possible *ground-truth* partition. In real-world networks, ground-truths are often defined by one specific property/attribute whose values are attached to the nodes. Several epistemological issues behind the practice of evaluating CD outputs against such ground-truths were recently investigated (Peel et al. 2017); although some possible variants to the issue (Rabbany and Zaïane 2015), real-world networks are not recommended for testing purposes. Another option consists of adopting synthetic benchmarks designed explicitly to mimic the meso-scale level of real-world networks by building artificially planted sets of communities and evaluate the CD algorithm performances on various difficulty levels. Moreover, driven by the *homophily* principle (McPherson et al. 2001), node attributes are often used to improve CD—or, at least, redefine it w.r.t. external aspects (Chunaev et al. 2020)—by leveraging both topological and label-homogeneous clustering criteria. The *node-attributed network* encodes information about the node's properties/qualities, in form of attributes, accordingly to the general purposes of *feature-rich networks* (Interdonato

✉ Salvatore Citraro
salvatore.citraro@phd.unipi.it

Giulio Rossetti
giulio.rossetti@isti.cnr.it

[1] Department of Computer Science, University of Pisa, Largo Bruno Pontecorvo, 3, Pisa, Italy

[2] KDD-Lab, ISTI (CNR), G. Moruzzi, 1, Pisa, Italy

et al. 2019), where the goal is to merge the graph topology together with other possibly meaningful external information. In the redefinition of the CD task—known as node-attributed or labeled CD task (henceforth, LCD)—the aim is to find well-connected communities that are also homogeneous w.r.t. the attributes carried by the nodes. It follows that the evaluation environment should be improved at the same time. Thus, for testing LCD algorithm outputs, only connectivity-based benchmarks are not enough. Motivated by all the above-mentioned evaluation issues, often not approached in a systematic manner in the LCD task, we aim to address them in this work (i) by building a synthetic generator with attribute-aware planted communities, *X-Mark*, and (ii) by testing different LCD approaches against it. In detail, our two main contributions are to provide a new benchmark for testing LCD algorithms, then carefully evaluate them being aware of the class they belong according to state-of-the-art taxonomies, by highlighting their ability to perform better/worse on incrementally complex real-world scenarios.

The rest of the paper is organized as follows. In Sect. 2, we will review the state-of-the-art of attribute-aware network models, synthetic benchmarks, and LCD approaches. In Sect. 3, we will introduce *X-Mark* our node-attribute enriched network generator that handles label-homogeneous communities, embedding both categorical and continuous attributes. In Sect. 4, we will test some LCD families of approaches against *X-Mark*, to prove to what extent the algorithms can reconstruct the artificial communities embedded in the benchmark. Finally, Sect. 5 will conclude the work, summarizing the results and possible future lines of research.

2 Related work

An overview of several topics is needed to provide the full context surrounding the present work, i.e., the state-of-art about network models, synthetic generators, and LCD techniques.

Network models Network models aim to capture and replicate some essential properties underlying real-world phenomena, from heavy-tailed degree distributions to high clustering coefficients and short average path lengths [i.e., small-world properties (Watts and Strogatz 1998)], as well as nonzero degree-degree correlation, community structure, and homophily. The well-known *Preferential Attachment* mechanism (henceforth, *PA*) of the Barabási-Albert model (Barabási and Albert 1999) generates scale-free networks with a power-law degree distribution, following the principle that the more connected a node is, the more likely it is to receive new links. Extensions of *PA* include steps for the formation of triads (Holme and Kim 2002), or for allowing the growth of degree-assortative networks (Catanzaro et al. 2004) or communities with power-law distributions (Xie et al. 2007). Alternative approaches—such as the *Community Guidance Attachment* and *Forest Fire Model*s (Leskovec et al. 2005)—can exploit other network properties, e.g., self-similarity and hierarchies, for generating community structure.

Network models that include homophily in the generative process aim to study how such a principle can influence the properties and the evolution of a system. A standard procedure shared by several models is that the probability of forming connections depends both on the degree (i.e., *PA*) and the attributes the nodes encode (Gong et al. 2012; Pasta et al. 2014; Kim and Altmann 2017; Shah et al. 2019). Several analytical experiments suggest that modeling homophily-aware networks produces interesting results. In Kim and Altmann (2017), the authors observe different shapes of the cumulative degree distributions, which transform from concave to convex when homophily is forced to have a substantial role in the generative process; such a convexity is interpreted as the power of homophily to amplify the rich-get-richer effect (more than considering only the *PA*); in Pasta et al. (2014), it is observed that high degree assortativity acts as a negative force to generate homophilic networks. moreover, the mechanism of focal closure (i.e., the formation of links between similar nodes without common neighbors) differs from structural closure (Murase et al. 2019), and their cumulative effects imply the formation of core-periphery structures (Asikainen et al. 2020). In the context of opinion dynamics, several works introduce homophily-aware network generators for exploiting controlled analysis of human dynamics: false uniqueness and false consensus are amplified in heterophilic and homophilic networks, respectively (Lee et al. 2017); higher homophilic networks exhibit meaningful community structure and have a role in the formation and cohesion of groups (Gargiulo and Gandica 2016). In such models, it is worth noticing that communities are not built-in, since they are extracted *a-posteriori* with a CD algorithm. These example leads us to make an important distinction between network modeling and synthetic benchmarks.

Synthetic benchmarks Synthetic benchmarks allow researchers to evaluate their algorithms on data whose characteristics resemble those observed in real-world networks. Contrary to network models, the rationale behind the construction of synthetic benchmarks is to use *ground-truths* to evaluate the fitness of the partitions resulting from CD methods. Among the most famous generators used for classic CD, we find the Girvan–Newman (GN) (Girvan and Newman 2002) and the Lancichinetti–Fortunato–Radicchi (LFR) (Lancichinetti et al. 2008) benchmarks, as well as the family of stochastic blockmodels (SBMs) (Holland et al. 1983; Karrer and Newman 2011). The GN benchmark Girvan and Newman (2002) is a graph of 128 nodes with an expected degree of 16, divided into four communities of equal sizes. Two parameters identify the probabilities of intra- and inter-clusters links, respectively. The LFR benchmark (Lancichinetti et al. 2008) allows for a user-defined number of nodes and distributes both node degrees and communities size according to a power-law. A parameter (i.e., the structure mixing μ) identifies the fraction of links that a node has to share with other nodes in its cluster, while the remaining fraction is shared with random nodes in other parts of the graph. In the SBM (Holland et al. 1983), nodes are assigned to one of k user-defined communities; then, the links are placed independently between nodes with probabilities that are a function of the community membership of the nodes; a degree-corrected version of SBM allows to identifying heterogeneous node degrees (Karrer and Newman 2011).

Such methods are designed to evaluate static graph partitions and do not natively support the generation/analysis of node-attributed graphs. Homophily-aware synthetic benchmarks are developed to cope with the limitation of such classic benchmarks, allowing for a more reliable controlled environment testing for LCD methods. Among the benchmarks specifically designed to generate node-attributed networks with communities, we find LFR-EA (Elhadi and Agam 2013), ANC (Largeron et al. 2015), and acMark (Maekawa et al. 2019). In LFR-EA (Elhadi and Agam 2013), the LFR benchmark is extended with a noise parameter that controls the percentage of homogeneity within communities. The user can define the number of attributes and the number of values for each attribute, as well as the percentage of random sampling with or without replacement (i.e., how the values distribute among the communities). Interesting LCD testing against LFR-EA can be found in Pizzuti and Socievole (2018) and Berahmand et al. (2020). In ANC (Largeron et al. 2015), nodes with only continuous attributes are generated, whose values are sparse out through a user-defined standard deviation parameter; some representative nodes of each community are initialized, then a K-medoids clustering is performed to build communities, and a user-defined number of intra and inter-links is generated. The node community assignment depends only on the labels of representative nodes. An LCD testing against ANC can be found in Falih et al. (2017) and Liu et al. (2020). In acMark (Maekawa et al. 2019), a bayesian approach is used to generate node-attributed graphs with communities. It enables to specify various degree distributions, cluster sizes, and both categorical and continuous attribute types.

Finally, it is also worth mentioning a set of works modifying SBMs to cope with node covariates, as in Tallberg (2004), where this is achieved via a multinomial probit model. Often referred as CSBMs (covariate stochastic blockmodels) (Sweet 2015), they consist of a hybrid between the network models and synthetic benchmarks previously mentioned. Since they can create networks with communities correlated with node attributes, they often purpose to test the ability of algorithms to make use of metadata (i.e., whether they can be helpful to the LCD task). The work in Newman and Clauset (2016) gives a prototypical example of this, where a correlation between structure and attributes is created matching the latter ones with the true community assignments of nodes in an SBM; this approach is found to be effective also for generating multi-layer synthetic networks with ground-truth (Contisciani et al. 2020), and in the network inference problem by systematically studying the influence of the attributes on the correlation between network data and metadata (Fajardo-Fontiveros et al. 2021). Other attributed SBMs can be found in Hric et al. (2016), where a multi-layer-based approach allows developing one layer modeling relational information between attributes and the other one modeling connectivity, then assigning nodes to communities maximizing the likelihood of the observed data in each layer; in Stanley et al. (2019), a similar approach is able to handle multiple continuous attributes. Other than augmented-SBMs, in Emmons and Mucha (2019), instead, the map equation is modified to control the varying importance of metadata with a tuning parameter.

Labeled or node-attributed community discovery LCD focuses on obtaining structurally well-defined partitions that also result in label-homogeneous communities. Several comparative studies and survey have been proposed to classify the large and increasing amount of node-attributed CD algorithms by leveraging taxonomies that allow grouping the algorithms according to the *point-of-view* adopted for the *clustering step*. Figure 1 summarizes them. While Bothorel et al. (2015) proposes a preliminary low-level classification, Falih et al. (2018) aggregates the algorithms into three general families: (i_a) *topological-based*, (ii_a) *attributed-based*, and (iii_a) *hybrid* approaches. Such a taxonomy focuses primarily on *how* the original graph is manipulated for taking attributive information into account, namely (i_a) *attaching* it to the topology, or (ii_a) *merging* them together at the expense of the original links, or (iii_a) using an ensemble method. The important aspect of *time* (e.g., modifying the original structure *before* or *contextually* to the clustering step), leads (Chunaev 2020) to propose a different classification schema: algorithms are grouped according to the moment when structure and attributes are fused, distinguishing between (i_b) *early-fusion*, (ii_b) *simultaneous fusion*, and (iii_b) *late-fusion* approaches.

Just to give an idea of the complexity of defining appropriate taxonomies, an approach like CESNA (Yang et al. 2013), built on top of a probabilistic generative process while treating node attributes as latent variables, can be viewed either as a hybrid or a simultaneous fusion approach, but also as an approach similar to the hybrid network models outlined in the previous paragraph.

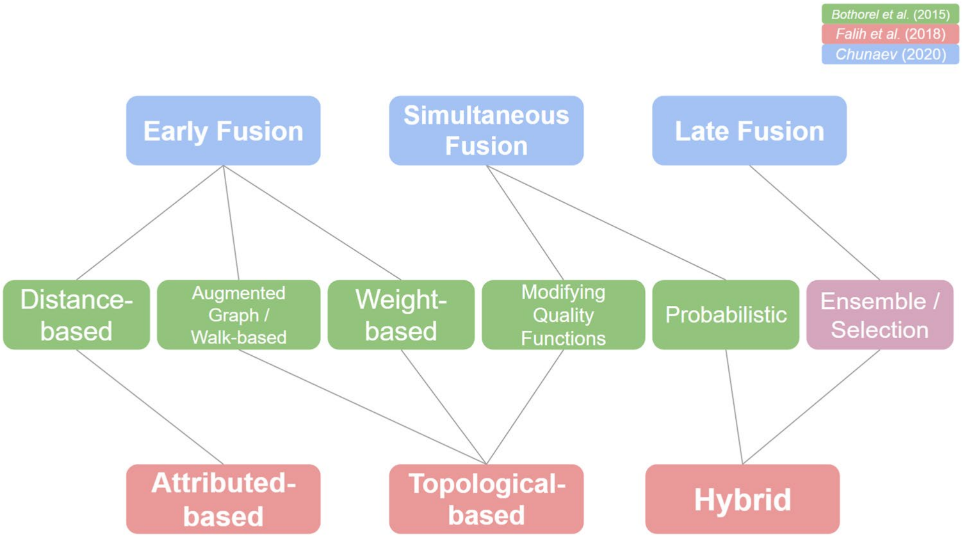

Fig. 1 LCD taxonomies, where the middle level (in green) is a sub-hierarchy of two possible higher-level schema (in blue and red, respectively) (Color figure online)

For a review of some specific LCD algorithms, demanding detailed information is left to the mentioned surveys. Nevertheless, the LCD approaches that we test against *X-Mark* will be described better in the appropriate analytical section.

3 *X-Mark*

Throughout the work, we refer to the following definition of the node-attributed graph:

Definition 1 (**Node-attributed Graph**) $\mathcal{G} = (V, E, A)$ is a node-attributed graph, where V is the set of nodes, E the set of edges, and A a set of categorical or continuous attributes such that $A(v)$, with $v \in V$, identifies the set of categorical or continuous values associated to v.

X-Mark[1] aims to generate an undirected and unweighted node-attributed graph $\mathcal{G}$ along with an attribute-aware planted partition $\mathcal{C}$ while guaranteeing: (i) power-law node degree and (ii) community size distribution; (iii) user-defined noise distribution within homogeneous communities; (iv) user-defined intra/inter-community edge distribution. In detail, *X-Mark* network generation procedure works as reported in Algorithm 1 - subject to the controlling parameters summarized in Table 1. In detail, it articulates into four steps:

Algorithm 1 *X-Mark*

```
Require: Parameters as defined in Table 1
Generate degree sequence                              ▷ subject to $|V|$, $\langle k \rangle$, $\gamma$
Generate community size sequence                      ▷ subject to $s_{min}$, $s_{max}$, $\beta$
for each attribute do
    Generate community label sequence                 ▷ subject to $m_{cat}$ or $m_{cont}$
for each iteration do
    Assign nodes to communities
    for each attribute do
        Assign node attribute value                   ▷ subject to $\theta$ or $\sigma$
Generate intra/inter-community edges                  ▷ subject to $\mu$
```

[1] Python code available at: https://github.com/dsalvaz/XMark

Table 1 Description of tunable parameters

Parameter	Description
$\|V\|$	Number of nodes
$\langle k \rangle$	Average degree of nodes
γ	Power-law exponent for node degree sequence
β	Power-law exponent for community size sequence
m_{cat}	Number of values in the domain of a categorical attribute (at least 2)
$\theta \in [0, 1]$	noise parameter
m_{cont}	Number of peaks (at least a bimodal distribution)
σ	Standard deviation
$\mu \in [0, 1]$	Mixing parameter

Step 1: Nodes generation and degree assignation - subject to the average degree $\langle k \rangle$ and the power-law exponent γ (line 1, Algorithm 1);

Step 2: Community size sequence generation imposing the power-law exponent β (line 2), and identification, for each attribute, of the *representative label* of each community, sampled from m_{cat} or m_{cont} (line 3-4)[2]; in detail: (i) for each categorical attribute, a random assignment from the m_{cat} possible values in the domain of the attribute, where $m_{\text{cat}} \geq 2$; (ii) for each continuous attribute, a random assignment from an ad-hoc multimodal distribution having m_{cont} possible peaks, where $m_{\text{cont}} \geq 2$, and the first peak has mean 0, while the other ones are ϵ values positively distant from the previous peak;

Step 3: Communities and node's attribute generation (lines 5-8) handling different strategies for categorical and continuous attributes, i.e.,: (i) for each categorical attribute, assign to the node the same value of its community with probability $1 - \theta$; (ii) for each continuous attribute, assign to the node a value picked from a normal distribution assuming the community label as the distribution mean and σ as it standard deviation;

Step 4: Edge sampling - subject to the expected ratio among intra/inter-community edges as expressed by the mixing parameter μ (line 9), as previously defined in Lancichinetti et al. (2008).

Among the model hyper-parameters reported in Table 1, the following are peculiar to *X-Mark*: (i) θ: it tunes the level of noise within each community. A low value of θ implies the emergence – within each community – of a majority label, with $\theta = 0$ modelling the extreme scenario where all the nodes within a community share the same categorical attribute value;

(ii) ϵ: it affects the *speed* at which the benchmark starts to produce less well-separated clusters according to the attribute values distribution: in this work, we impose $\epsilon = 10$; (iii) m_{cat} and m_{cont} are integers modeling the domain for categorical and numerical attributes respectively; in the rest of the article, for the sake of simplicity, we will implicitly treat such parameters as lists of integers, meaning that each attribute has its proper m value in the range expressed by the list.

3.1 *X-Mark* characterization

In this subsection, we provide an overview of some *X-Mark* characteristics. For this purpose, we introduce a set of measures for the analysis; then, we split the study according to the differences between the categorical and the continuous attributes modeling.

Evaluation Measures To characterize the behaviour of the model in presence of categorical attributes, we relate the observed and expected label homophily. In detail, we calculate the observed homophily, H, as the probability that two nodes share the same attribute value, and compare it to the expected one, H_{exp}, namely, the probability that a randomly chosen node pair shares the same attribute value. Formally:

$$H = \frac{|(u, v) \in E : A(u) = A(v)|}{|E|} \tag{1}$$

$$H_{exp} = \frac{|(u, v) : A(u) = A(v)|}{|N|(|N| - 1)} \tag{2}$$

Since H and H_{exp} do not take the homophilic contribution of each community/node into account explicitly, we also provide (i) a function capturing noise within communities (i.e., the percentage of the majority attribute value within a cluster), namely Purity (Citraro and Rossetti 2019), and (ii) two measures for explaining the homophilic contribution of each node, namely Peel's assortativity (Peel et al. 2018) and Conformity (Rossetti et al. 2020). Given a community C, its purity P_c is the product of the frequencies of the most frequent categorical attribute values carried by the nodes within C, formally:

$$P_c = \prod_{a \in A} \frac{\max_{a \in A}(\sum_{v \in c} a(v))}{|c|} \tag{3}$$

where A is the attribute value set, $a \in A$ is an attribute value, and $a(v)$ is an indicator function that takes value 1

[2] Note that the same expressiveness can be preserved with a single parameter, m: the distinction aims to show the qualitative difference between the two attribute types, i.e., categorical or continuous attributes.

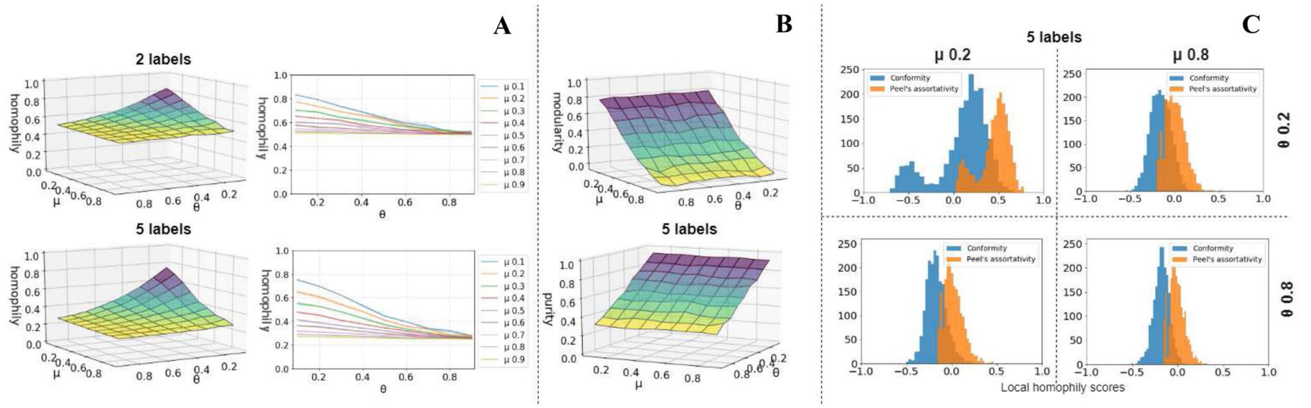

Fig. 2 **a**: 3D plots with relative 2D projections including IQR ranges (15 iterations) in functions of μ and θ (1 attribute, $m_{\text{cat}} = [2]$ or $m_{\text{cat}} = [5]$); **b**: purity and modularity in functions of μ and θ (1 attribute, $m_{\text{cat}} = [5]$); **c**: Peel's assortativity and conformity on four networks with different combinations of low and high μ and θ values (1 attribute, $m_{\text{cat}} = [5]$)

iff $a \in A(v)$. The purity of a complete partition is then the average of the purities of the communities that compose it:

$$P = \frac{1}{|C|} \sum_{c \in C} P_c \tag{4}$$

Since homophily H gives only one global score, we might not identify the contribution of single nodes or observe differences between the intra- and inter-homophilic connections. Peel's assortativity and Conformity compute for each node its homophilic embeddedness within the *neighborhood* it belongs.

We evaluate continuous attributes, using the Within-Cluster Sum of Squares (WCSS):

$$WCSS = \sum_{i=1}^{k} \sum_{v \in C_i} |v - M|^2 \tag{5}$$

where, for each community C from i to k, M is the centroid of nodes within the community.

Moreover, we leverage the concept of *silhouette* for represent graphically how well clusters are tight and separated to each other. Detailed information is left to the reference paper (Rousseeuw 1987).

Finally, to analyze the degree of connectivity of homogeneous clusters, we compute the *modularity* score, e.g., the fraction of the edges that fall within the given community C minus the expected fraction if they were distributed following a null model.

$$Q = \frac{1}{(2m)} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{(2m)} \right] \delta\left(c_v, c_w\right) \tag{6}$$

where m is the number of graph edges, $A_{v,w}$ is the entry of the adjacency matrix for $v, w \in N$, k_v, k_w the degree of v, w and $\sigma(c_v, c_w)$ identifies an indicator function taking value 1 iff v, w belong to the same community, 0 otherwise.

Categorical attributes. In this scenario, homogeneous communities are well-connected sets of nodes within which most of them share the same attribute value. The θ parameter models the percentage of nodes labeled according to a randomly assigned attribute value among the user-defined m_{cat} possible ones; the remaining fraction is labeled according to the *preferred* community value. Thus, imposing $\theta = 0.2$ means that at least 80% of nodes within a community share the same attribute value. The rationale behind the inclusion of the majority value justify the case of a binary categorical attribute (i.e., $m_{\text{cat}} = 2$), where $\theta = 1$ leads to a lower bound of observed homophily of 0.5.

Figure 2 a shows the value of H as function of μ and θ. We focus on two different setups, $m_{\text{cat}} = 2$ and $m_{\text{cat}} = 5$: in the former, the minimum observed homophily is around 0.5 (as H_{exp}, not displayed); in the latter, the minimum observed homophily is around 0.3 (as H_{exp}, not displayed). In general, the plots in Fig. 2 a show us how *X-Mark* can implicitly model homophily by only considering clusters homogeneity. Indeed, H decreases as both randomly rewired connections and attribute noise within communities increase; e.g., for high values of μ and θ (i.e., from 0.6 to 0.9), H and H_{exp} tend to coincide, with the consequence of creating a very hard scenario for all structural-only, attribute-only and attribute-aware CD strategies.

To better understand how homophily emerges from such parameters, we analyzed the network node-centric homophilic behaviour. Peel's assortativity and Conformity give us two different *points of view*. In Fig. 2 c, we show the local homophily scores of the two measures for the outlined setups. In particular, two peaks emerge when well-defined (i.e, well-connected and homogeneous) communities are modelled (i.e., $\mu = 0.2$ and $\theta = 0.2$), telling us that the network

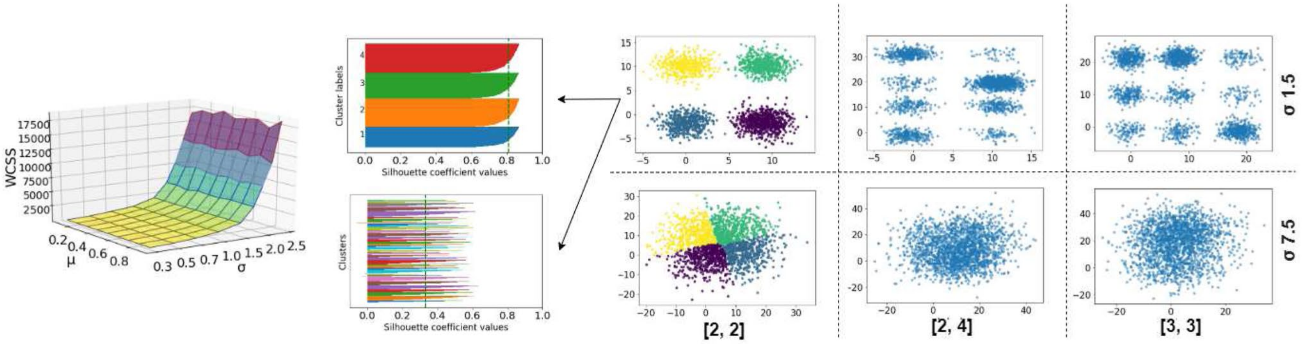

Fig. 3 From left to right, WCSS values in functions of μ and σ; silhouette analysis of K-Means graph segmentation ($m_{\text{cont}} = [2, 2]$ and $\sigma = 1.5$), by using two strategies for determining k, i.e., $k = 4$ (elbow method), and k equal to the community size sequence cardinality; other attribute distributions of graphs generated from several m_{cont} and σ combinations, and a focus on K-Means graph segmentation ($m_{\text{cont}} = [2, 2]$ and $\sigma = 7.5$) by selecting $k = 4$ (elbow method): higher σ values lead to ill-defined clusters

has a large (majority) homophilic behavior, but smaller heterophilic zones emerge mostly from inter-cluster noise. Noisy communities decrease the within-cluster homophilic contribution even if the former ones are well-connected (i.e., $\mu = 0.2$ and $\theta = 0.8$). The distributions observed for both measures describe similar scenarios: nodes tend to concentrate around a mean value neither homophilic nor heterophilic, except for very well-defined and homogeneous communities. To conclude, it follows that clustering modularity only depends on the parameter μ, and clustering purity, only on the parameter θ. Figure 2b summarizes it.

Continuous attributes Considering a continuous attribute scenario, homogeneous communities are clusters with low standard deviations. As outlined in Fig. 3 (the leftmost 3D plot of the figure), the Within-Cluster Sum of Squares (WCSS) increases as σ increases, independently from the structure mixing parameter μ. Modeling continuous attributes by controlling m_{cont} allows deducing the number of dense and well-separated clusters – in particular, when using low σ values. In Fig 3, we show some examples, by using the following m_{cont} configurations on two networks with low ($\sigma = 1.5$) and relatively high ($\sigma = 7.5$) standard deviations, respectively: $m_{\text{cont}} = [2, 2]$, $m_{\text{cont}} = [2, 4]$, and $m_{\text{cont}} = [3, 3]$. Indeed, well-separated clusters are visible when σ is low. We executed K-Means (MacQueen 1967) over the network configured with $m_{\text{cont}} = [2, 2]$ to show that the centroid-based clustering algorithm is able to recognize automatically the number of planted components from the attribute point-of-view. On the other hand, such well-separated clusters do not match with the planted component of communities emerging from the structural point-of-view, i.e., the number of communities subject to the β parameter. We can continue to refer to the first ones as the *attribute-component* of the partition, and to the second ones as its *structural-component*. Indeed, the differences among those two components are relevant since they induce potentially distinct, although meaningful, clustering. In Fig. 3, we show the silhouette scores of each clustering found by K-Means with (i) $k = 4$ (optimal value suggested by the elbow method), and (ii) k equal to the number of planted communities generated by *X-Mark*. The silhouette scores are different, and less qualitatively good clusters are found according to the latter strategy, i.e., while considering the structure-point-of-view to tune an attribute-only clustering approach.

With this last point, we anticipate one of the fundamental problems dissected in the next section: how to combine the attribute-component view and the structural one while performing an attribute-aware graph clustering?

4 Experiments

This section provides an analytical framework of comparison between LCD algorithms against *X-Mark*. We compare the algorithms by considering the several classification schema emerging in LCD literature, as we discussed in Section 2.

Algorithms We compare (i_a) *topological-based*, (ii_a) *attributed-based*, and (iii_a) *hybrid* algorithms, contextually to (i_b) *early-fusion*, (ii_b) *simultaneous-fusion*, and (iii_b) *late fusion* ones.

Ensemble/Selection (iii_a, iii_b): methods falling within this category aim to fuse (or choose between) topological and attribute information after that both CD (for structure) and classic clustering methods (for attributes) are performed. We consider: (i) *CSPA* (Strehl and Ghosh 2002; Elhadi and Agam 2013), a method that uses a graph representation to solve cluster ensemble, by partitioning an induced similarity graph built on top of the binary similarity matrices extracted from the partitions; (ii) *MCLA* (Strehl and Ghosh 2002), another graph-based approach, where each partition is represented as a node, then linked to the other ones by considering their similarity; (iii)

Table 2 List of parameter values used for the analyses

Parameter	Value(s)
$\|V\|$	2000
$\langle k \rangle$	10
γ	3
β	2
m_{cat}	[2,4]
θ	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]
m_{cont}	[2, 2]; [2, 4]; [$\|\mathcal{C}\|$, $\|\mathcal{C}\|$] where $\|\mathcal{C}\|$ is the cardinality of the partition set
σ	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]
μ	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]

Selection (Elhadi and Agam 2013), that chooses a preferable partition between a structural and an attributive one (Louvain (Blondel et al. 2008) and K-Means, respectively, in this work); the choice is made by looking at the estimated mixing parameter μ of the graph: if such a value is less than a certain experimental value μ_{lim} (i.e., 0.55, in the current study), Louvain is selected, K-Means otherwise; (iv) *Late-Fusion* (Liu et al. 2020), that combines two partitions (again, a structural and an attributive one) by integrating their adjacency matrices through a linear combination; then, a CD algorithm segments the final induced graph.

Modifying Quality Functions (i_a, ii_b): methods falling within this category aim to modify the objective functions of classical CD algorithms by integrating attribute-aware criteria for attributes. We consider: (i) *EVA* (Citraro and Rossetti 2019, 2020), a Louvain extension that integrates an attribute-aware function (i.e., Purity) for grouping homogeneous communities through a linear combination. It works with categorical and ordinal attributes; (ii) *I-Louvain* (Combe et al. 2015), a Louvain extension that includes an attribute-aware objective function called Inertia; no parameters are involved, but the algorithm works only with continuous attributes;

Distance-based (ii_a, i_b): methods falling within this category perform the attribute-aware clustering on a distance matrix obtained by fusing structure and attributes distance functions; common metrics for structure distance are the shortest path lengths or Jaccard similarity. We consider: (i) *ANCA* (Falih et al. 2017), that selects a set of seeds toward which each nodes characterize their topological and semantic similarity, then computes a distance matrix factorization and runs K-Means over it; we apply the BiCC criteria for seed selection and the shortest path length to compute topological similarity, as suggested in the original paper; (ii) *StoC* (Baroni et al. 2017), that uses a multi-objective distance to fuse structure and attribute node similarities; the user is assumed to provide a semantic attraction ratio α_s and a topological one α_t, to let the method compute from itself a distance threshold τ extracting τ-close clusters, i.e., nodes which are within a maximum distance τ from a given random seed, and a distance length l to define the l-neighborhood of a node; in this work, several values of α_s and α_t are selected.

CSPA and MCLA were implemented in python[3]; Late-Fusion[4], ANCA[5] and EVA[6] implementations are the ones of the original authors; the latter is also available on the CDLib Python library (Rossetti et al. 2019), together with the I-Louvain one. The code of SToC was gently released by the corresponding authors on our requests.

X-Mark settings and evaluation

We report in Tab. 2 the *X-Mark* parameter values used for the graphs generation. We leverage the widely adopted (Fortunato and Hric 2016) Normalized Mutual Normalized Information (henceforth, NMI) to compare *X-Mark* communities to the ones identified by the selected algorithms. NMI is formally defined as in the following:

$$NMI(X,Y) = \frac{H(X) + H(Y) - H(X,Y)}{(H(X) + H(Y))/2} \tag{7}$$

where $H(X)$ is the entropy of the random variable X associated with an algorithm partition, $H(Y)$, the one related to the ground-truth one, and $H(X, Y)$, the joint entropy. NMI ranges in [0, 1], and it is maximized when the algorithm partition and the ground-truth one are identical.

Evaluation: ensemble/selection As previously introduced while analyzing the continuous attributes generation, the *naıve* number of communities subject to the sequence obtained by tuning the β parameter (i.e., the *structural-component* of the ground-truth partition) might not correspond to the *naıve* number of clusters subject to the attribute value distribution (i.e., the *attribute-component* one), in particular when the benchmark is instantiated to model well-connected communities that also produce well-separated clusters (i.e., imposing low μ and σ values).

To test the ensemble algorithms on *X-Mark*, we define three different case of scenarios, identified as a, b, and c - subject to specific m_{cont} values, namely:

(i) $m_{\text{cont}} = [\,|\mathcal{C}|, |\mathcal{C}|\,]$, where $|\mathcal{C}|$ is the cardinality of the partition set; we aim to generate as much peaks as the number of graph communities, in order to avoid any issue related to the differences between the structural- and the attribute-component, i.e., the fact that similar nodes w.r.t. they attributes actually do not correlate with the connections they

[3] https://github.com/GGiecold/Cluster_Ensembles

[4] https://github.com/changliu94/attributed-community-detection

[5] https://github.com/Issamfalih/ANCL

[6] https://github.com/GiulioRossetti/Eva

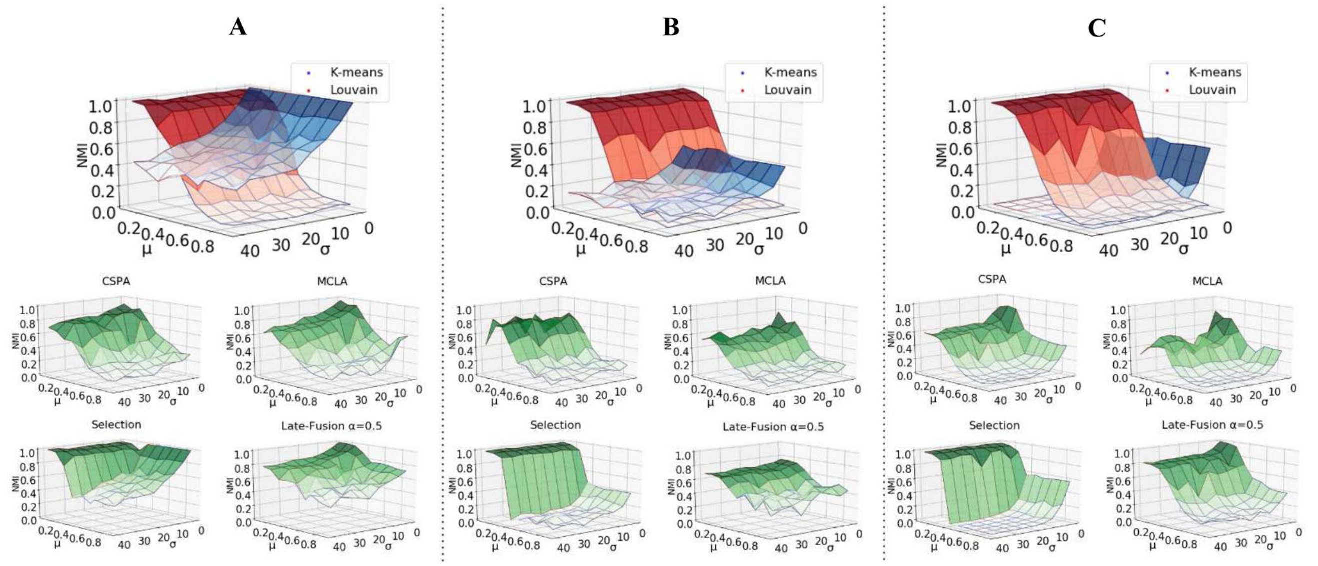

Fig. 4 *Ensemble/selection analysis*: NMI similarities in functions of μ and σ; the letters above each experimental framework correspond to different *X-Mark* benchmark instances against which the algorithms are tested (i.e., **a**: $m_{\text{cont}} = [\,|\mathcal{C}|, |\mathcal{C}|\,]$, $k = elbow(\mathcal{G})$; **b**: $m_{\text{cont}} = [\,|\mathcal{C}|, |\mathcal{C}|\,]$, $k = |\mathcal{C}|$, c: $m_{\text{cont}} = [2, 4]$, $k = elbow(\mathcal{G})$)

establish; to cope with this framework, two solutions are proposed to infer the number of k required by the attribute-component: a k is the one chosen by the elbow method, that picks the elbow of the curve described by the WCSS values as the number of clusters to use; b: $k = |\mathcal{C}|$, i.e., the number of structural-component communities.

(ii) $m_{\text{cont}} = [2, 4]$, where c k is chosen according to the elbow method.

The proposed analysis is designed to increasingly resemble real-world scenarios, since the gap between structural- and attribute-components increases from $m_{\text{cont}} = [\,|\mathcal{C}|, |\mathcal{C}|\,]$ to $m_{\text{cont}} = [2, 4]$, and an only-attribute clustering algorithm can find the cluster cardinality estimation more difficult. In other words, the algorithm performances should decrease when their attribute-component needs to determine the number of clusters k by only looking at attribute information and, contextually, this one does not match with the heavy-tailed topological constraints of the community size sequence. Thus, in the former scenario (i.e., m_{cont} =[$|\mathcal{C}|, |\mathcal{C}|$] with k chosen according to the WCSS elbow curve), such a gap is flattened, because the attribute domains equal the number of topological communities, i.e., we have a different peak for each graph community. Then, on the same benchmark instance, we test an alternative solution for the estimation of k (i.e., $m_{\text{cont}} = [\,|\mathcal{C}|, |\mathcal{C}|\,]$), to observe how the algorithms perform if we use only topological information to determine k. Finally, a more likely real-world scenario generates an attribute-aware planted partition where the attribute domains do not match with the number of communities (i.e., $m_{\text{cont}} = [2, 4]$) and where an elbow method is used to determine k, because, in real-world contexts, we cannot have information about the real number of graph clusters.

Figure 4 shows a selection of the obtained results. The letters above the plots (A, B, C) refer to the three scenarios previously introduced. All the plots report the NMI between the *X-Mark* ground-truth partitions and the ones obtained by the algorithms, as functions of μ and σ parameters. Above each ensemble/selection method (whose results are highlighted in green), we focus on the only-topological and only-attribute algorithmic approaches that each method uses to obtain a consensus partition from their fusion/selection, i.e., Louvain Blondel et al. (2008) (values highlighted in red) and K-means (MacQueen 1967) (in blue). Intuitively, Louvain is only affected by the mixing parameter tuning; conversely, K-means is only affected by the value dispersion due to the standard deviation increase. When the attribute domains equal the number of topological communities (i.e., Fig 4 a), we also observe partition similarities when σ is relatively high, contrary to the other two scenarios. Most importantly, the similarity between the benchmark ground-truths and K-means clustering decreases when k is supposed to match the real number of communities (i.e., Fig. 4 b) or in the most likely real-world network simulation (i.e., Fig. 4 c).

Briefly, consensus and selection methods depend on both the two output types. Among the consensus methods, the Late-Fusion one seems to perform better than CSPA and MCLA, in particular because the α parameter, when is set to 0.5, can tune a better trade-off between the two clustering typologies. The Selection method chooses between a topological-only and an only-attribute algorithm according to the

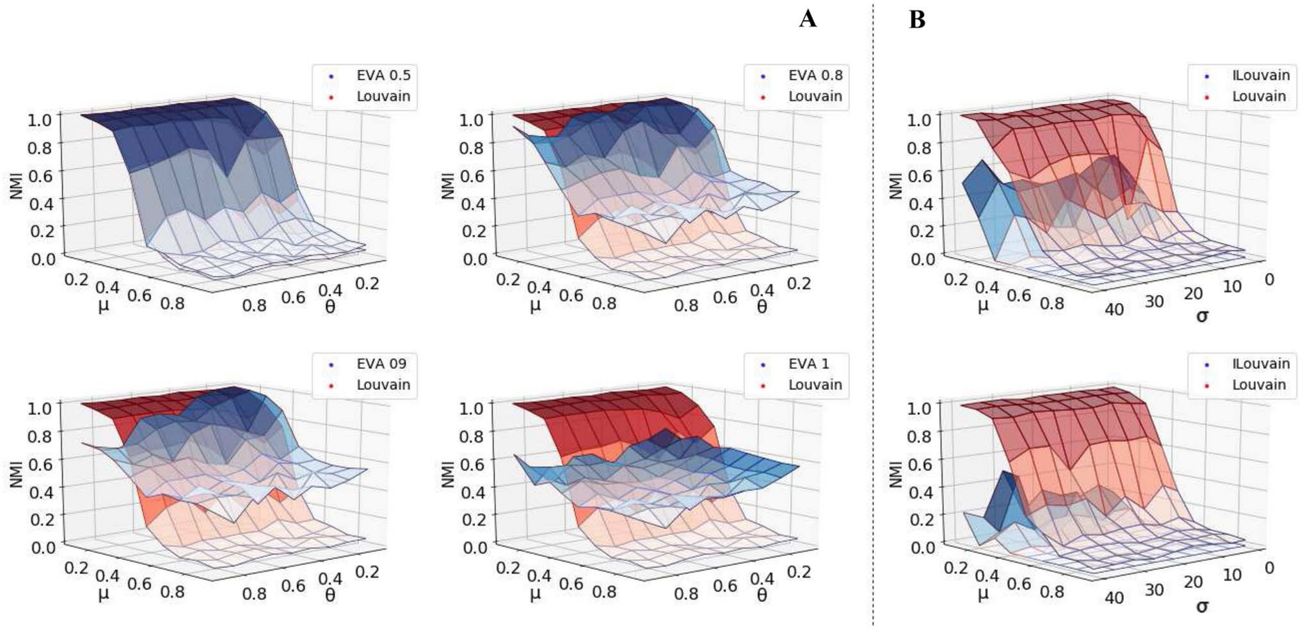

Fig. 5 *Modifying quality function analysis*: NMI similarities in functions of μ and θ (**a**) or σ (**b**); in detail, EVA is run against a *X-Mark* instance generated using $m_{\mathrm{cat}} = [2, 4]$, while ILouvain, using $m_{\mathrm{cont}} = [\,|\mathcal{C}|, |\mathcal{C}|\,]$ (above) and $m_{\mathrm{cont}} = [2, 4]$ (below)

moment when the graph structure is ambiguous. Until a very low μ value, Louvain is maintained as a clustering choice, then KMeans is selected, but its performances depend on the attribute dispersion tuned by σ: if the structure is ambiguous and the attributes are clear, the Selection method performs well (and better than a consensus method, since it only uses K-means and not a combination of clustering); however, such achievement is strongly affected by the involved scenario (a or c).

Within the LCD context, these approaches work well if the two types of outputs *correct* each other. Again, observing the Louvain and KMeans NMI in Fig 4 a, we can see how both the methods can recognize the true *X-Mark* synthetic communities, respectively, when a well-separated structure (low μ) and well-separated attributes (low σ) are generated; thus, switching from Louvain to K-means in the Selection method gives such a method a similarity continuity (w.r.t. the true communities) from an ambiguous structure to clear attributes. In some sense, since communities from a network point of view do not exist, a classic clustering method is performed. However, the switching from an ambiguous structure to clear attributes gives worse results when more likely real-world scenarios are simulated (Fig 4 c), that is when two well-separated and poorly interconnected dense communities sharing the same majority attribute values exist.

Evaluation: modifying quality functions Contrary to ensemble/selection methods, algorithms that modify a topological quality function do not fuse the clustering of two already performed only-topological and only-attributes methods, but they extend an only-topological approach including the attributes into the maximization of a function aiming to find well-connected (and homogeneous) communities. Here, we will focus on EVA and ILouvain, that work, respectively, on categorical and continuous attributes. They do not need to specify a required number of clusters. EVA needs to tune the parameter of the linear combination used to balance the topological and semantic importance when grouping nodes, i.e., the α parameter. ILouvain does not need any parameter tuning since its function is normalized to give the same importance to relational and attribute information.

Figure 5 shows the NMI between the *X-Mark* ground-truth partitions and the ones obtained by EVA (Fig 5 a) and ILouvain (Fig 5 b), as functions of μ and θ (EVA) or σ (ILouvain). We test EVA only against a benchmark instances generated with $m_{\mathrm{cat}} = [2, 4]$ (results, not showed, with $m_{\mathrm{cat}} = [\,|\mathcal{C}|, |\mathcal{C}|\,]$ were similar). When $\alpha = 0$, only the topological function component (i.e., modularity) is optimized, and it is equivalent to run Louvain; when $\alpha = 1$, only the attribute component (i.e., purity) is optimized, equivalent to cluster the set of the biggest connected components whose nodes share the same label profiling. In the figure, we show results for $\alpha = [0.5, 0.8, 0.9, 1]$: we focus only on values towards the homogeneity optimization to see to what extent the attributes influence clustering. EVA matches the *X-Mark* communities outperforming its *natural* baseline, Louvain: when μ increases, EVA can exploit

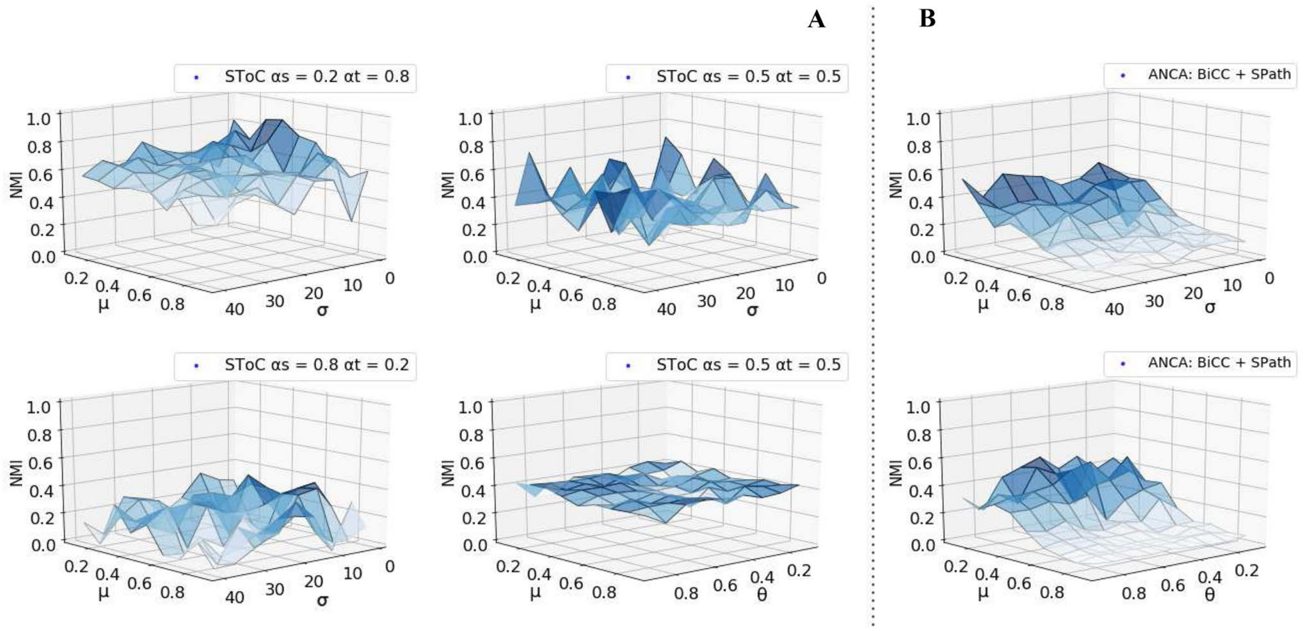

Fig. 6 *Distance-based analysis*: NMI similarities in functions of μ and σ or θ for SToC (**a**) and ANCA (**b**); $m_{\text{cat}}, m_{\text{cont}} = [2, 4]$ are the parameters used

attribute information to find the homogeneous communities that emerge from the random configuration of links between communities. In other words, a flat surface means that an algorithm focuses only on the attribute information: regarding EVA, this is quite evident when $\alpha = 1$. A good trade-off is the one able to maintain high NMI when μ is low, and to not decrease to zero when μ is high contextually to a low level of attribute noise. Conversely, ILouvain performs poorly on *X-Mark*. Similarly to the framework proposed for ensemble/selection methods, we tested ILouvain against a benchmark generated using $m_{\text{cont}} = [\,|\mathcal{C}|, |\mathcal{C}|$ (Fig 5 b, above), and $m_{\text{cont}} = [2, 4]$ (Fig 5 b, below). The obtained results underline that ILouvain is not able to exploit attributes (i.e., NMI equal to 0 for high μ). Even fusing the two components (attributes and modularity) does not allow to recognize structurally well-defined clusters (i.e., NMI low for low μ); by consequence, ILouvain performs *worse* than its baseline, Louvain, being possible the case that the ILouvain objective function cannot tune the relative contribution of structure and attributes.

Evaluation: Distance-based Finally, we focus on two distance-based methods, ANCA and SToC. Such methods can use both categorical and continuous attributes, which can be exploited even together. We focus only on the two attributes types taken individually, generating *X-Mark* networks with $m_{\text{cat}}, m_{\text{cont}} = [2, 4]$ (not showed, results with $m_{\text{cat}}, m_{\text{cont}} = [\,|\mathcal{C}|, |\mathcal{C}|\,]$ were similar). Regarding SToC, the user is allowed to tune some dummy parameters, α_s, that forces towards attribute similarities, and α_t, that forces towards topological similarity: we noticed that similar results are achieved testing SToC against the categorical benchmark instance, thus we show only $\alpha_s = \alpha_t = 0.5$ (Fig. 6 a, right, below), that is also one of the setting parameter solution proposed in the reference paper (Baroni et al. 2017); for the continuous attributes, instead, we tested SToC also both with $\alpha_s = 0.2$, $\alpha_t = 0.8$, performing a topological clustering, and $\alpha_s = 0.8$, $\alpha_t = 0.2$, a more attribute-aware one. As we can observe from Fig. 6 b, ANCA performs relatively worse than the other approaches, particularly if compared with the ensemble/selection methods, or EVA. The trend of the ANCA 3D plots appears reasonable, but (i) the NMI decreases only as function of μ, suggesting that only the topological component is taken into account for the clustering task, and (ii) maximal NMI values are lower than the ensemble/selection methods or EVA. Similarly, the trend of the SToC 3D plots are reasonable, but (i) it resembles a flat surface (particularly, while clustering categorical attributes, Fig. 6 a, below, right), suggesting that only the attribute component is taken into account for the clustering task (as we already saw for EVA when its α parameter is equal to 1), and (ii), again, the maximal NMI values are lower than other methods. SToC performances are better while clustering continuous attributes, when the discovery of communities is forced towards the topological component (Fig. 6 a, above, left), but it decreases for other parameter settings, suggesting that the algorithm is, in some sense, confounded by the attribute component of the graph.

5 Discussion and conclusion

In this work, we proposed a solution for evaluating labeled community discovery (LCD) algorithms. Thus, we modeled *X-Mark*, a synthetic tool for generating node-attributed networks with planted communities. Extending some already existent intuitions for the generation of only topological-based benchmarks (e.g., LFR (Lancichinetti et al. 2008)), *X-Mark* firstly generates both the community size and degree distribution, then use them to associate each node to a partition. Label-homogeneity within communities is controlled by the probability to have within each community a user-defined percentage of similar nodes, encoded in a noise parameter θ for categorical attributes, and the community standard deviation σ for continuous ones. Once inserted each node into its preferable community, the edge rewiring automatically generates assortative patterns within communities, contributing to the homophilic network behavior. We guarantee community homogeneity and network homophily, resembling scenarios for simulating node-attributed real-world network representations.

Indeed, several lines of discussion span from *X-Mark*, among them: (i) how to exploit the *X-Mark* ability to specify different structures and attribute combinations (e.g., clear structure vs. clear attributes or clear structure vs. noisy attributes), and, generally, (ii) how to fairly compare the quality of clustering testing the algorithms against synthetic benchmarks. Firstly, we designed our model to be as general as possible, leaving the analyst to specify how to combine different structure and attribute combinations. Analyzing the algorithm performances as functions of the whole range of structure and attribute parameter values allowed us to have a broad vision of how algorithms perform. Nevertheless, as well remarked by several discussions (Fortunato and Hric 2016; Chunaev 2020; Chunaev et al. 2020), a strong rationale behind many of LCD approaches is often assumed by the researchers: the algorithms can exploit nodes' attributes in the CD task because homophily strongly contributes to community formation. In other words, since node similarities match with the connections they made, it is useful to consider such similarities while grouping closer nodes. Nevertheless, it is intuitive to think that some attributes might match with the node connections, while others are independent from the relational realm of a dataset (see Peel et al. 2017; Newman and Clauset 2016). *X-Mark* can model situations where attributes align/not align to topology.

In the future, we plan to extend our tests to LCD algorithms that explicitly exploit attribute information by looking at the combination of clear/noisy structures and clear/noisy attributes. Moreover, we plan to test LCD algorithms against different attribute-aware benchmarks to see if other external comparison methods can lead to different results. Being based on the same algorithmic schema of LFR, we can also plan to extend *X-Mark* to cope with overlapping communities, as well as weighted and directed networks, as done for the classic LFR extension (Lancichinetti and Fortunato 2009). Dealing with such task variants and different representations is not trivial in the presence of node metadata. Since a benchmark aims to resemble real-world scenarios, we also need more investigations into real-world weighted or directed node-attributed networks. The actual lack of a large corpus of studies in this direction makes it more difficult to find valuable solutions for these extensions.

Attribute-aware CD, which identifies well-connected and label-homogeneous nodes, is a rising theme in complex network analysis. We are far away from reaching standard procedures for handling attribute information embedded in the nodes as well as evaluating different clustering outputs. We aimed to take some first steps towards a more careful evaluation analysis of attribute-aware CD algorithms, as recently provided only in Vieira et al. (2020). Based on the present findings, thanks to *X-Mark*, we can evaluate algorithms performances within a controlled environment, i.e., adopting systematic tuning parameters strategies. Among others, we observed that ensemble clustering methods can suffer the selection of the best k number of communities, while algorithms modifying only-structure quality functions can outperform their only-structure baseline only when the new fitness function is well defined.

Acknowledgements This work is supported by the scheme 'INFRAIA-01-2018-2019: Research and Innovation action', Grant Agreement n. 871042 'SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics'

Funding Open access funding provided by Università di Pisa within the CRUI-CARE Agreement.

References

Asikainen A, Iñiguez G, Ureña-Carrión J, Kaski K, Kivelä M (2020) Cumulative effects of triadic closure and homophily in social networks. Sci Adv 6(19):eaax7310

Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512

Baroni A, Conte A, Patrignani M, Ruggieri S (2017) Efficiently clustering very large attributed graphs. In: 2017 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), pp 369–376. IEEE

Berahmand K, Haghani S, Rostami M, Li Y (2020) A new attributed graph clustering by using label propagation in complex networks. J King Saud Univ-Comput Inf Sci **(in press)**

Blondel VD, Guillaume J-L, Lambiotte R (2008) Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech: Theory Exp 10:P10008

Bothorel C, Cruz JD, Magnani M, Micenková B (2015) Clustering attributed graphs: models, measures and methods. Netw Sci 3(03):408–444

Brandes U, Robins G, McCranie A, Wasserman S (2013) What is network science? Netw Sci 1(1):1–15

Catanzaro M, Caldarelli G, Pietronero L (2004) Social network growth with assortative mixing. Physica A 338(1–2):119–124

Chunaev P (2020) Community detection in node-attributed social networks: a survey. Comput Sci Rev 37:100286

Chunaev P, Gradov T, Bochenina K (2020) Community detection in node-attributed social networks: How structure-attributes correlation affects clustering quality. Procedia Comput Sci 178:355–364

Citraro S, Rossetti G (2019) Eva: attribute-aware network segmentation. In: International conference on complex networks and their applications. Springer, Berlin pp 141–151

Citraro S, Rossetti G (2020) Identifying and exploiting homogeneous communities in labeled networks. Appl Netw Sci 5(1):1–20

Combe D, Largeron C, Géry M, Egyed-Zsigmond E (2015) I-louvain: an attributed graph clustering method. In: International symposium on intelligent data analysis. Springer, pp 181–192

Contisciani M, Power EA, De Bacco C (2020) Community detection with node attributes in multilayer networks. Sci Rep 10(1):1–16

Elhadi H, Agam G (2013) Structure and attributes community detection: comparative analysis of composite, ensemble and selection methods. In: Proceedings of the 7th workshop on social network mining and analysis, pp 1–7

Emmons S, Mucha PJ (2019) Map equation with metadata: varying the role of attributes in community detection. Phys Rev E 100(2):022301

Fajardo-Fontiveros O, Sales-Pardo M, Guimera R (2021) Node metadata can produce predictability transitions in network inference problems. arXiv preprint arXiv:2103.14424

Falih I, Grozavu N, Kanawati R, Bennani Y (2017) Anca: attributed network clustering algorithm. In: International conference on complex networks and their applications, Springer, pp 241–252

Falih I, Grozavu N, Kanawati R, Bennani Y (2018) Community detection in attributed network. Companion Proc Web Conf 2018:1299–1306

Fortunato S, Hric D (2016) Community detection in networks: a user guide. Phys Rep 659:1–44

Gargiulo F, Gandica Y (2016) The role of homophily in the emergence of opinion controversies. arXiv preprint arXiv:1612.05483

Girvan M, Newman ME (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99(12):7821–7826

Gong NZ, Xu W, Huang L, Mittal P, Stefanov E, Sekar V, Song D (2012) Evolution of social-attribute networks: measurements, modeling, and implications using google+. In: Proceedings of the 2012 internet measurement conference, pp 131–144

Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: first steps. Soc Netw 5(2):109–137

Holme P, Kim BJ (2002) Growing scale-free networks with tunable clustering. Phys Rev E 65(2):026107

Hric D, Peixoto TP, Fortunato S (2016) Network structure, metadata, and the prediction of missing nodes and annotations. Phys Rev X 6(3):031038

Interdonato R, Atzmueller M, Gaito S, Kanawati R, Largeron C, Sala A (2019) Feature-rich networks: going beyond complex network topologies. Appl Netw Sci 4(1):1–13

Karrer B, Newman ME (2011) Stochastic blockmodels and community structure in networks. Phys Rev E 83(1):016107

Kim K, Altmann J (2017) Effect of homophily on network formation. Commun Nonlinear Sci Numer Simul 44:482–494

Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms. Phys Rev E 78(4):046110

Lancichinetti A, Fortunato S (2009) Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Phys Rev E 80(1):016118

Largeron C, Mougel P-N, Rabbany R, Zaïane OR (2015) Generating attributed networks with communities. PLoS ONE 10(4):e0122777

Lee E, Karimi F, Wagner C, Jo H-H, Strohmaier M, Galesic M (2017) Homophily and minority size explain perception biases in social networks. arXiv preprint arXiv:1710.08601

Leskovec J, Kleinberg J, Faloutsos C (2005) Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp 177–187

Liu C, Largeron C, Zaïane OR, Gharaghooshi SZ (2020) A late-fusion approach to community detection in attributed networks. In: International symposium on intelligent data analysis. Springer, pp 300–312

MacQueen J (1967) et al Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol 1, pp 281–297, Oakland, CA, USA

Maekawa S, Zhang J, Fletcher G, Onizuka M (2019) General generator for attributed graphs with community structure. In: Proceeding of the ECML/PKDD graph embedding and mining workshop, pp 1–5

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Ann Rev Sociol 27(1):415–444

Murase Y, Jo H-H, Török J, Kertész J, Kaski K (2019) Structural transition in social networks: the role of homophily. Sci Rep 9(1):1–8

Newman ME, Clauset A (2016) Structure and inference in annotated networks. Nat Commun 7(1):1–11

Pasta MQ, Zaidi F, Rozenblat C (2014) Generating online social networks based on socio-demographic attributes. J Complex Netw 2(4):475–494

Peel L, Larremore DB, Clauset A (2017) The ground truth about metadata and community detection in networks. Sci Adv 3(5):e

Peel L, Delvenne J-C, Lambiotte R (2018) Multiscale mixing patterns in networks. Proc Natl Acad Sci 115(16):4057–4062

Pizzuti C, Socievole A (2018) A genetic algorithm for community detection in attributed graphs. In: International conference on the applications of evolutionary computation, pp 159–170, Springer

Rabbany R, Zaïane OR (2015) Evaluation of community mining algorithms in the presence of attributes. In: Trends and applications in knowledge discovery and data mining, pp 152–163, Springer

Rossetti G, Milli L, Cazabet R (2019) Cdlib: a python library to extract, compare and evaluate communities from complex networks. Appl. Netw Sci. 4(1):52

Rossetti G, Citraro S, Milli L (2020) Conformity: a path-aware homophily measure for node-attributed networks.

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65

Shah H, Kuma S, Sundaram H (2019) Growing attributed networks through local processes. The World Wide Web conference, pp 3208–3214

Stanley N, Bonacci T, Kwitt R, Niethammer M, Mucha PJ (2019) Stochastic block models with multiple continuous attributes. Appl Netw Sci 4(1):1–22

Strehl A, Ghosh J (2002) Cluster ensembles–a knowledge reuse framework for combining multiple partitions. J Mach Learn Res 3(Dec):583–617

Sweet TM (2015) Incorporating covariates into stochastic blockmodels. J Edu Behav Stat 40(6):635–664

Tallberg C (2004) A bayesian approach to modeling stochastic blockstructures with covariates. J Math Sociol 29(1):1–23

Vieira AR, Campos P, Brito P (2020) New contributions for the comparison of community detection algorithms in attributed networks. J Complex Netw 8(4):cnaa044

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–442

Xie Z, Li X, Wang X (2007) A new community-based evolving network model. Physica A 384(2):725–732

Yang J, McAuley J, Leskovec J (2013) Community detection in networks with node attributes. In: 2013 IEEE 13th international conference on data mining, pp 1151–1156. IEEE