Flexible inference in heterogeneous and attributed multilayer networks

Martina Contisciani[a,1,2], **Marius Hobbhahn**[b,1], **Eleanor A. Power**[c,d], **Philipp Hennig**[b], **and Caterina De Bacco**[a,2]

This manuscript was compiled on January 7, 2025

Networked datasets can be enriched by different types of information about individual nodes or edges. However, most existing methods for analyzing such datasets struggle to handle the complexity of heterogeneous data, often requiring substantial model-specific analysis. In this paper, we develop a probabilistic generative model to perform inference in multilayer networks with arbitrary types of information. Our approach employs a Bayesian framework combined with the Laplace matching technique to ease interpretation of inferred parameters. Furthermore, the algorithmic implementation relies on automatic differentiation, avoiding the need for explicit derivations. This makes our model scalable and flexible to adapt to any combination of input data. We demonstrate the effectiveness of our method in detecting overlapping community structures and performing various prediction tasks on heterogeneous multilayer data, where nodes and edges have different types of attributes. Additionally, we showcase its ability to unveil a variety of patterns in a social support network among villagers in rural India by effectively utilizing all input information in a meaningful way.

Significance Statement

Network models are a powerful tool for analyzing complex interactions in a variety of domains. These models are widely used to explore data or detect hidden patterns such as communities. However, they often struggle to incorporate additional information encoded in the data, such as nodes or edge attributes of various types. Here, we develop an approach that flexibly integrates extra information without requiring ad-hoc preprocessing steps or derivations. Our method enhances performance in prediction tasks and uncovers hidden patterns by fully exploiting the rich information encoded in complex network datasets.

probabilistic generative models | attributed multilayer networks | heterogeneous information | network inference | overlapping communities | automatic differentiation | Laplace approximation

Networks effectively represent real-world data from various fields, including social, biological, and informational systems. In this framework, nodes within the network correspond to individual components of the system, and their interactions are illustrated through network edges (1). With the advancement of data collection and representation techniques, networks have evolved to become more versatile and informative. Notably, attributed multilayer networks have emerged as a significant development, allowing the inclusion of additional information related to nodes and edges. This enriches the representation of real-world systems, where nodes naturally have specific characteristics and are connected through different types of interactions. For instance, in social networks, individuals can be described by attributes like age, gender, and height, while engaging in various types of relationships like friendship, co-working, and kinship.

It is important to treat these systems as multilayer networks, rather than reducing them to a single layer (e.g., through aggregation), to avoid losing valuable information. This is particularly significant when the properties of the entire system cannot be derived from a simple linear combination of the properties of each layer in isolation, or through other forms of aggregation into a single layer representation (2). Relevant examples of this non-trivial behavior are observed in interbank markets, patent citation networks, time-dependent networks and brain networks (3–6). A similar rationale applies to datasets with multiple types of edges and node attributes, as explored in this work. Approaches to multilayer networks that consider only one type of attribute – or none at all – risk losing critical information.

The analysis of attributed multilayer networks has primarily been tackled using techniques like matrix factorization (7, 8), network embedding (9, 10), and deep learning (11–14). While effective for learning low-dimensional node representations for inference tasks, these methods have significant limitations when applied to network data. First, the resulting representations lack inherent interpretability. Since their main focus is prediction tasks, these models do not impose constraints on the parameters to improve interpetability, such as non-negativity or shared parameters across layers. Instead, they often require arbitrarily post-processing techniques (e.g., k-means clustering) for interpretation. Second, these methods prioritize node-level tasks (e.g., classification or clustering) and treat edge-level tasks, like link prediction, as secondary, relying on ad hoc mappings from nodes to edges. Lastly, they are typically parameter-heavy and require a substantial amount of training data, making them inefficient and overparametrized for real-world networks,

Author affiliations: [a]Max Planck Institute for Intelligent Systems, Tübingen 72076, Germany; [b]Tübingen AI Center, University of Tübingen, Tübingen 72076, Germany; [c]Department of Methodology, London School of Economics and Political Sciences, London WC2A 2AE, United Kingdom; [d]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA

Author contributions: M.C., M.H., P.H., and C.D.B. designed research; M.C., M.H., and C.D.B. performed research; M.C., E.A.P., and C.D.B. analyzed data; and M.C., M.H., E.A.P., P.H., and C.D.B. wrote the paper.

The authors declare no competing interest.

[1]M.C. and M.H. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: martina.contisciani@tuebingen.mpg.de, caterina.debacco@tuebingen.mpg.de

which are typically sparse, label-scarce for supervised settings, and often provide only a single observed sample.

To overcome these limitations, we adopt a different approach based on probabilistic generative models (15). Unlike the aforementioned methods, these models provide a principled and flexible framework that incorporates prior knowledge and specific assumptions, resulting in more interpretable representations. Importantly, they explicitly capture the dependencies between nodes and edges, rather than relying on ad hoc mappings. They also account for the inherent uncertainty present in real-world network data (16), providing a more robust and comprehensive understanding of this type of data. Furthermore, these models can be applied to perform various network tasks, such as edge and attribute prediction, detecting statistically meaningful network structures, and generating synthetic data.

Our goal is to develop a probabilistic generative model that can flexibly adapt to any attributed multilayer network, regardless of the type of information encoded in the data. Acting as a "black box", our method can enable practitioners to automatically analyze various datasets, without the need to deal with mathematical details or new derivations. This approach aligns with some practices in the machine learning community, where black box methodologies have been introduced to simplify the inference of latent variables in arbitrary models (17, 18). In this context, more specific probabilistic methods have been developed to address the challenge of performing inference on heterogeneous data (19, 20). However, these techniques are tailored for tabular data and do not provide a general solution to adapt them to network data.

Probabilistic generative models specifically designed for attributed networks aim to combine node attributes effectively with network interactions. Existing methods (21–28) have highlighted the importance of incorporating extra information to enhance network inference, resulting in improved prediction performance and deeper insights on the interplay between edge structure and node metadata. However, these models mainly focus on single-layer networks, assume the same generative process for all interactions, and consider only one type of attribute – typically categorical. These limitations restrict their capability to represent complex scenarios characterized by heterogeneous information. As a consequence, addressing the challenge of effectively incorporating various sources of information and evaluating their collective impact on downstream network inference tasks remains an open issue.

We address this gap by introducing PIHAM, a generative model explicitly designed to perform Probabilistic Inference in directed and undirected Heterogeneous and Attributed Multilayer networks. Our approach differs from previous studies in that PIHAM flexibly adapts to any combination of input data, while standard probabilistic methods rely on model-specific analytic derivations that highly depend on the data types given in input. This can dramatically hinder the flexibility of a model, as any small change in the data, e.g., adding a new node attribute or a new type of interaction, usually requires new derivations. As a result, the vast majority of these models work only with one type of edge weight for all layers, and one type of attribute. In contrast, PIHAM takes in input any number of layers and attributes, regardless of their data types.

At its core, PIHAM assumes the existence of a mixed-membership community structure that drives the generation of both interactions and node attributes. In addition, the inference of the parameters is performed within a Bayesian framework, where both prior and posterior distributions are modeled with Gaussian distributions. Importantly, PIHAM employs the Laplace matching technique (29) and conveniently maps the posterior distributions to various desired domains, to ease interpretation. For instance, to provide a probabilistic interpretation of the inferred communities, our method properly maps the parameters of a Gaussian distribution into those of a Dirichlet distribution. The latter operates within a positive domain and enforces normalization on a simplex, making it a valuable tool for this purpose. Notably, the inference process is flexible and scalable, relying on automatic differentiation and avoiding the need for explicit derivations. As a result, PIHAM can be considered a "black box" method, as practitioners only need to select the desired probabilistic model and a set of variable transformation functions, while the remaining calculations and inference are performed automatically. This versatility enables our model to be flexibly applied to new modeling scenarios.

We apply our method on a diverse range of synthetic and real-world data, showcasing how PIHAM effectively leverages the heterogeneous information contained in the data to enhance prediction performance and provide richer interpretations of the inferred results.

Methods

We introduce PIHAM, a versatile and scalable probabilistic generative model designed to perform inference in attributed multilayer networks. Our method flexibly adapts to any combination of input data, regardless of their data types. For simplicity, in what follows, we present examples with Bernoulli, Poisson, Gaussian, and categorical distributions, which collectively cover the majority of real-world examples. Nevertheless, our model can be easily extended to include new distributions, as well as applied to single-layer networks with or without attributes.

General framework. Attributed multilayer networks provide an efficient representation of complex systems in which the individual components have diverse attributes (often referred to as covariates or metadata) and are involved in multiple forms of interactions. Mathematically, these interactions are depicted by an adjacency tensor $\boldsymbol{A}$ of dimension $L \times N \times N$, where N is the number of nodes common across all L layers. Each entry A_{ij}^{ℓ} in this tensor denotes the weight of a directed interaction of type ℓ from node i to node j. Notably, different layers can incorporate interactions of diverse data types, depending on the nature of the underlying relationship. For instance, in social systems, one layer might represent binary relationships like friendships, another could describe nonnegative discrete interactions such as call counts, and a third might contain continuous real-valued measurements such as geographical distances between locations. In this scenario, the adjacency tensor would be represented as $\boldsymbol{A} = \{\boldsymbol{A^1} \in \{0,1\}^{N\times N}, \boldsymbol{A^2} \in \mathbb{N}_0^{N\times N}, \boldsymbol{A^3} \in \mathbb{R}_+^{N\times N}\}$. Node metadata describes additional information about the nodes. They are stored in a design matrix $\boldsymbol{X}$ with dimensions $N \times P$, where P is the total number of attributes and the entries X_{ix}

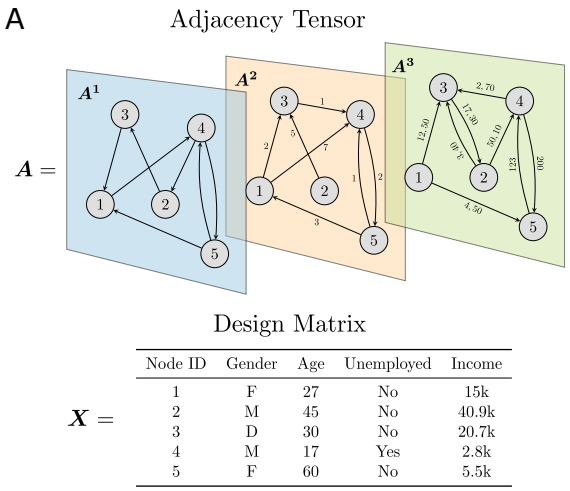

Fig. 1. Input data and graphical model representation. (A) The attributed multilayer network is represented by the interactions A^{ℓ}_{ij} and the node attributes X_{ix}. (B) PIHAM describes the observed data through a set of latent variables $\boldsymbol{\Theta} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{H})$. $\boldsymbol{U_i}$ and $\boldsymbol{V_i}$ respectively depict the communities of node i determined by the out-going and in-coming edges; $\boldsymbol{W^{\ell}}$ is the affinity matrix associated to the layer ℓ and characterizes the edge density between different community pairs in the given layer; $\boldsymbol{H_{\cdot x}}$ is a K-dimensional vector that explains how an attribute x is distributed among the K communities. All latent variables are independent and normally distributed, and $f(\cdot)$ and $g(\cdot)$ are transformation functions to ensure that the expected values λ^{ℓ}_{ij} and π_{ix} belong to the correct parameter space for the various distribution types.

represent the value of an attribute x for a node i. Similar to network interactions, different attributes can have different data types. An example of input data is given in Fig. 1A.

PIHAM describes the structure of attributed multilayer networks, represented by $\boldsymbol{A}$ and $\boldsymbol{X}$, through a set of latent variables $\boldsymbol{\Theta}$. The goal is to infer $\boldsymbol{\Theta}$ from the input data. In particular, we want to estimate posterior distributions, as done in a probabilistic framework. These can be approximated as:

$$\begin{aligned} P(\boldsymbol{\Theta} \,|\, \boldsymbol{A}, \boldsymbol{X}) &\propto P(\boldsymbol{A}, \boldsymbol{X} \,|\, \boldsymbol{\Theta})\, P(\boldsymbol{\Theta}) \\ &= P(\boldsymbol{A} \,|\, \boldsymbol{\Theta})\, P(\boldsymbol{X} \,|\, \boldsymbol{\Theta})\, P(\boldsymbol{\Theta})\,. \end{aligned} \qquad [1]$$

In this general setting, the proportionality is due to the omission of an intractable normalization term that does not depend on the parameters. The term $P(\boldsymbol{A}, \boldsymbol{X} \,|\, \boldsymbol{\Theta}) = P(\boldsymbol{A} \,|\, \boldsymbol{\Theta})\, P(\boldsymbol{X} \,|\, \boldsymbol{\Theta})$ represents the likelihood of the data, where we assume that $\boldsymbol{A}$ and $\boldsymbol{X}$ are conditionally independent given the parameters. This assumption allows to model separately the network structure and the node metadata. The term $P(\boldsymbol{\Theta})$ denotes the prior distributions of the latent variables, which we assume to be independent and Gaussian distributed, resulting in $P(\boldsymbol{\Theta}) = \prod_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu^{\theta}}, \boldsymbol{\Sigma^{\theta}})$. Importantly, we also make the assumption that the posterior distributions of the parameters can be approximated with Gaussian distributions, for which we have to estimate mean and covariance matrices: $P(\boldsymbol{\Theta} \,|\, \boldsymbol{A}, \boldsymbol{X}) \approx \prod_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{N}(\boldsymbol{\theta}; \hat{\boldsymbol{\mu}}^{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}^{\boldsymbol{\theta}})$.

In the following subsections, we provide additional details on the role of the latent variables in shaping both interactions and node attributes, as well as the methods for inferring their posterior distributions.

Modelling the network structure. The interactions encoded in the adjacency tensor $\boldsymbol{A}$ are assumed to be conditionally independent given the latent variables, resulting in a decomposition of the likelihood across individual entries A^{ℓ}_{ij}. This factorization can be further unpacked by explicitly considering the distributions that describe each layer. For instance, in the scenario with binary, count-based, and continuous interactions, we can express the likelihood as follows:

$$\begin{aligned} P(\boldsymbol{A} \,|\, \boldsymbol{\Theta}) &= \prod_{\ell,i,j} P(A^{\ell}_{ij} \,|\, \boldsymbol{\Theta}) \\ &= \prod_{\ell \in L_B, i, j} \mathrm{Bern}(A^{\ell}_{ij}; \lambda^{\ell}_{ij}(\boldsymbol{\Theta})) \\ &\quad \times \prod_{\ell \in L_P, i, j} \mathrm{Pois}(A^{\ell}_{ij}; \lambda^{\ell}_{ij}(\boldsymbol{\Theta})) \\ &\quad \times \prod_{\ell \in L_G, i, j} \mathcal{N}(A^{\ell}_{ij}; \lambda^{\ell}_{ij}(\boldsymbol{\Theta}), \sigma^2)\,, \end{aligned} \qquad [2]$$

where σ^2 is a hyperparameter and L_B, L_P, and L_G are the sets of Bernoulli, Poisson, and Gaussian layers, respectively. We assume that each distribution is fully parametrized through the latent variables $\boldsymbol{\Theta}$ and these explicitly define the expected values λ^{ℓ}_{ij}, regardless of the data type.

Specifically, we adopt a multilayer mixed-membership model (30), and describe the observed interactions through K overlapping communities shared across all layers. Following this approach, the expected value of each interaction of type ℓ from node i to j can be approximated as:

$$\lambda^{\ell}_{ij}(\boldsymbol{\Theta}) \approx \sum_{k,q=1}^{K} U_{ik} W^{\ell}_{kq} V_{jq}\,, \qquad [3]$$

where the latent variables U_{ik} and V_{jq} denote the entries of K-dimensional vectors $\boldsymbol{U_i}$ and $\boldsymbol{V_i}$, which respectively represent the communities of node i determined by the out-going and in-coming edges. In undirected networks, we set $\boldsymbol{U} = \boldsymbol{V}$. Moreover, each layer ℓ is associated with an affinity matrix $\boldsymbol{W^{\ell}}$ of dimension $K \times K$, which characterizes the edge density between different community pairs in the given layer ℓ. This setup allows having diverse structural patterns in each layer, including arbitrarily mixtures of assortative, disassortative and core-periphery structures.

As a final remark, the approximation in Eq. (3) arises from a discrepancy between the parameter space of the latent

variables and that of the expected values of the distributions. In fact, while all variables are normally distributed, λ_{ij}^{ℓ} has to satisfy different constraints according to the distribution type. For instance, $\lambda_{ij}^{\ell} \in [0,1]\,\forall\,\ell \in L_B$ and $\lambda_{ij}^{\ell} \in (0,\infty)\,\forall\,\ell \in L_P$. For further details, we refer to the section Parameter space and transformations.

Modelling the node metadata. Similarly to the network edges, the node metadata are also considered to be conditionally independent given the latent variables. Therefore, when dealing with data that encompass categorical, count-based, and continuous attributes, the likelihood can be formulated as follows:

$$\begin{aligned} P(\boldsymbol{X} \mid \boldsymbol{\Theta}) &= \prod_{i,x} P(X_{ix} \mid \boldsymbol{\Theta}) \\ &= \prod_{i,x \in C_C} \mathrm{Cat}(X_{ix}; \pi_{ix}(\boldsymbol{\Theta})) \\ &\times \prod_{i,x \in C_P} \mathrm{Pois}(X_{ix}; \pi_{ix}(\boldsymbol{\Theta})) \\ &\times \prod_{i,x \in C_G} \mathcal{N}(X_{ix}; \pi_{ix}(\boldsymbol{\Theta}), \sigma^2)\,, \end{aligned} \quad [4]$$

where C_C, C_P, and C_G are the sets of categorical, Poisson, and Gaussian attributes, respectively.

Following previous work (22, 27, 28), we assume that the attributes are also generated from the node community memberships, thereby creating dependencies between node metadata and network interactions. In particular, we approximate the expected value of an attribute x for node i as:

$$\pi_{ix}(\boldsymbol{\Theta}) \approx \frac{1}{2} \sum_{k=1}^{K} (U_{ik} + V_{ik})\, H_{kx}\,, \quad [5]$$

where $\boldsymbol{H}$ is a $K \times P$-dimensional community-covariate matrix, explaining how an attribute x is distributed among the K communities. For instance, if we consider income as node metadata and expect communities to group nodes with similar income values, then the column vector $\boldsymbol{H}_{\cdot\boldsymbol{x}}$ describes how income varies across groups. It is important to observe that when the attribute x is categorical, the expression in Eq. (5) becomes more complex because it must consider the total number of attribute categories Z. We provide additional details in the Supporting Information.

Notice that like λ_{ij}^{ℓ}, π_{ix} also needs to satisfy specific constraints depending on the distribution type. We clarify this in the next subsection.

Parameter space and transformations. A key technical aspect of PIHAM is the use of Gaussian distributions to model priors and posteriors of the latent variables $\boldsymbol{\Theta} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{H})$. This choice simplifies the inference by an additional step that ensures the expected values λ_{ij}^{ℓ} and π_{ix} belong to the correct parameter space for the various distribution types. To achieve this, we apply specific transformation functions to the latent variables, and model the expected values as follows:

$$\lambda_{ij}^{\ell}(\boldsymbol{\Theta}) = f(\boldsymbol{U_i})\, g(\boldsymbol{W^{\ell}})\, f(\boldsymbol{V_j}) \quad [6]$$

$$\pi_{ix}(\boldsymbol{\Theta}) = \frac{1}{2}\big(f(\boldsymbol{U_i}) + f(\boldsymbol{V_i})\big)\, g(\boldsymbol{H}_{\cdot\boldsymbol{x}})\,. \quad [7]$$

Table 1. Functions $g(\cdot)$ used in our implementation to transform the latent variables as defined in Eq. (6) and Eq. (7).

Distribution	Parameter space	Transformation function
Bernoulli	$[0,1]$	Logistic
Poisson	$(0,\infty)$	Exponential
Gaussian	$\mathbb{R}$	Identity
Categorical	$p_z \geq 0\,\forall z, \sum_z p_z = 1$	Softmax

For Gaussian distributions, we model only the mean; for categorical distributions, we apply the softmax by row, i.e., across categories.

The functions $f(\cdot)$ and $g(\cdot)$ can take various forms, as long as they adhere to the required constraints. In our implementation, we select $f(\cdot)$ to be the softmax function, which is applied to every row of the community membership matrices. This allows interpretability of the communities, as they result in quantities that are positive and normalized to one, as discussed in the section Parameter interpretation. Meanwhile, the choice of $g(\cdot)$ varies depending on the distribution type, as illustrated in Table 1.

One might argue that it would be simpler to employ a single link function for λ_{ij}^{ℓ} and π_{ix}, rather than applying individually transformations to the latent variables, as done in standard statistical approaches (31). However, this may not ensure interpretability of the communities, as we do with the softmax $f(\cdot)$. In addition, empirically we discovered that the approach outlined in Eq. (6) and Eq. (7) gives more stable results, and it does not result in over- or under-flow numerical errors. Alternatively, another approach considers treating the transformed parameters as random variables and applies the probability transformation rule to compute their posterior distributions (32). While this method is theoretically well-founded, it comes with constraints regarding the choice of the transformation functions, which directly affects the feasibility of the inference process. Conversely, PIHAM offers the flexibility to use any set of transformation functions that respects the parameter space of the distribution types given by the network and covariates.

In Fig. 1, we illustrate the input data and the graphical model representation of our approach.

Posterior inference. PIHAM aims at estimating the posterior distributions of the latent variables, as outlined in Eq. (1), where the normalization term is omitted as it does not depend on the parameters. More precisely, this equation can be reformulated as:

$$\begin{aligned} P(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{H} \mid \boldsymbol{A}, \boldsymbol{X}) &= P(\boldsymbol{A} \mid \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \\ &\times P(\boldsymbol{X} \mid \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{H}) \\ &\times P(\boldsymbol{U})\, P(\boldsymbol{V})\, P(\boldsymbol{W})\, P(\boldsymbol{H})\,. \end{aligned} \quad [8]$$

In general, this posterior distribution lacks a closed-form analytical solution and requires the use of approximations.

Common methods for inference in attributed networks typically rely on Expectation-Maximization (EM) (33) or Variational Inference (VI) (34) techniques. However, these approaches have limitations, as they require model-specific analytic computations for each new term added to the likelihood. For instance, an EM-based approach involves taking derivatives with respect to a given latent variable and setting them to zero. In a specific class of models where

the likelihood and prior distributions are compatible, solving the resulting equation for the variable of interest can yield closed-form updates. Nonetheless, for generic models, there is no guarantee of a closed-form solution. Even when this does exist, slight variations in the input data may require entirely new derivations and updates. Consequently, most of these models are designed to handle only a single type of edge weight and a single type of attribute.

In contrast, our model takes a different approach and flexibly adapts to any combination of input data, regardless of their data types. We begin by assuming that the latent variables are conditionally independent given the data, allowing us to model each posterior distribution separately:

$$P(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{H}\,|\,\boldsymbol{A},\boldsymbol{X}) = P(\boldsymbol{U}\,|\,\boldsymbol{A},\boldsymbol{X})\,P(\boldsymbol{V}\,|\,\boldsymbol{A},\boldsymbol{X}) \times P(\boldsymbol{W}\,|\,\boldsymbol{A},\boldsymbol{X})\,P(\boldsymbol{H}\,|\,\boldsymbol{A},\boldsymbol{X})\,. \quad [9]$$

Subsequently, we employ a Laplace Approximation (LA) to approximate each posterior with a Gaussian distribution, resulting in:

$$P(\boldsymbol{\theta}\,|\,\boldsymbol{A},\boldsymbol{X}) \approx \mathcal{N}(\boldsymbol{\theta};\hat{\boldsymbol{\mu}}^{\boldsymbol{\theta}},\hat{\boldsymbol{\Sigma}}^{\boldsymbol{\theta}})\,,\ \forall\,\boldsymbol{\theta}\in\boldsymbol{\Theta}\,. \quad [10]$$

LA involves a second-order Taylor expansion around the Maximum A Posteriori estimate (MAP) of the right-hand side of Eq. (8). We compute this estimate using Automatic Differentiation (AD), a gradient-based method that, in our implementation, employs the Adam optimizer to iteratively evaluate derivatives of the log-posterior. One might question the validity of this approximation in a potentially multimodal landscape. However, it is important to recognize that the existence of local minima is a problem independent of the approximation method employed. Nonetheless, in our model class – multilevel networks of log-convex likelihood functions connected by log-convex link functions – this issue is less likely to occur.

The MAP estimate found with AD also constitutes the mean $\hat{\boldsymbol{\mu}}^{\boldsymbol{\theta}}$ of $P(\boldsymbol{\theta}\,|\,\boldsymbol{A},\boldsymbol{X})$. To go beyond point estimates and quantify uncertainty, one can further estimate the covariance matrix $\hat{\boldsymbol{\Sigma}}^{\boldsymbol{\theta}}$, which is given by the inverted Hessian around the MAP:

$$\hat{\boldsymbol{\Sigma}}^{\boldsymbol{\theta}} \approx \Big[-\frac{\partial^2 P(\boldsymbol{\theta}\,|\,\boldsymbol{A},\boldsymbol{X})}{\partial\boldsymbol{\theta}}(\hat{\boldsymbol{\mu}}^{\boldsymbol{\theta}})\Big]^{-1}. \quad [11]$$

Other inference methods can be employed to approximate Gaussian distributions, such as VI. However, in such situations, utilizing AD directly might not be feasible due to the involvement of uncertain expectations in the optimization cost function. On the other hand, LA naturally combines with AD, providing a flexible and efficient inference procedure.

Parameter interpretation. We approximate the posterior distributions of the latent variables using Gaussian distributions, as outlined in Eq. (10). Consequently, all our estimated parameters belong to the real space. Although this approach is advantageous for developing an efficient and automated inference method, practitioners may desire different variable domains to enhance interpretability. In some instances, achieving this transformation is straightforward, involving the application of the probability transformation rule to obtain a distribution for the transformed variable within the desired constrained support. For example, if we are interested in expressing $\bar{\boldsymbol{U}} := \exp(\hat{\boldsymbol{U}}) \in \mathbb{R}_{>0}^{N\times K}$, we can simply employ the Lognormal$(\bar{\boldsymbol{U}};\hat{\boldsymbol{\mu}}^{\boldsymbol{U}},\hat{\boldsymbol{\Sigma}}^{\boldsymbol{U}})$ distribution. Similarly, when seeking $\bar{\boldsymbol{U}} := \text{logistic}(\hat{\boldsymbol{U}}) \in (0,1)^{N\times K}$, we can just compute the Logitnormal$(\bar{\boldsymbol{U}};\hat{\boldsymbol{\mu}}^{\boldsymbol{U}},\hat{\boldsymbol{\Sigma}}^{\boldsymbol{U}})$.

However, certain functions lack closed-form transformations. For instance, obtaining a probabilistic interpretation of the mixed-memberships of nodes requires applying the softmax function to each row of the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, which is not a bijective function. To address this challenge, our framework employs the Laplace Matching (LM) (29) to approximate the distributions of such transformations. This technique yields a bidirectional, closed-form mapping between the parameters of the Gaussian distribution and those of the approximated transformed distribution. In this scenario, we can derive:

$$\bar{\boldsymbol{U}}_{\boldsymbol{i}} := \text{softmax}(\hat{\boldsymbol{U}}_{\boldsymbol{i}})\,,\ \bar{U}_{ik}\in[0,1] \text{ and } \sum_{k=1}^{K}\bar{U}_{ik}=1 \quad \text{with } P(\hat{\boldsymbol{U}}_{\boldsymbol{i}}) = \text{Dir}(\bar{\boldsymbol{U}}_{\boldsymbol{i}};\hat{\boldsymbol{\alpha}}_{\boldsymbol{i}}^{\boldsymbol{U}})\,, \quad [12]$$

where $\hat{\boldsymbol{\alpha}}_{\boldsymbol{i}}^{\boldsymbol{U}}$ is a K-dimensional vector obtained with LM, whose entries are described as:

$$\hat{\alpha}_{ik}^{U} = \frac{1}{\hat{\Sigma}_{ikk}^{U}}\Big(1-\frac{2}{K}+\frac{\exp(\hat{\mu}_{ik}^{U})}{K^2}\sum_{l=1}^{K}\exp(\hat{\mu}_{il}^{U})\Big)\,. \quad [13]$$

This approach is theoretically grounded and enables us to provide closed-form posterior distributions for the latent variables across a diverse range of domains. Consequently, it consistently allows for the estimation of uncertainties and other relevant statistical measures. Nonetheless, PIHAM can also be utilized for the sole purpose of determining point estimates of the latent variables, which are essentially given by the MAP estimates. In such scenarios, it remains feasible to map these point estimates to different supports by applying any desired function, without worrying about the transformation process. Although this approach lacks full posterior distributions, it significantly simplifies the inference process by avoiding the computation of the Hessian. The choice between these two approaches should be guided by the specific application under study.

Results

We demonstrate our method on both synthetic and real-world datasets, presenting a comprehensive analysis through quantitative and qualitative findings. Further explanations about the data generation and pre-processing procedures can be found in the Supporting Information, which also includes additional results. The settings used to run our experiments and the choice of the hyperparameters are also described in the Supporting Information. The code implementation of PIHAM is accessible at: https://github.com/mcontisc/PIHAM.

Simulation study.

Comparison with existing methods in a homogeneous scenario. We first investigate the behavior of our model in a simpler and common scenario, characterized by attributed multilayer networks with nonnegative discrete weights and one categorical node attribute. This represents the most general case addressed by existing methods, which are specifically designed

for homogeneous settings, where there is only one attribute and one data type. For comparison, we use MTCOV (28), a probabilistic model that assumes overlapping communities as the main mechanism governing both interactions and node attributes. In contrast to PIHAM, MTCOV is tailor-made to handle categorical attributes and nonnegative discrete weights. Additionally, it employs an EM algorithm, with closed-form derivations for parameters inference strongly relying on the data type, making MTCOV a bespoke solution compared to the more general framework proposed by PIHAM. The results of this comparison are depicted in Fig. S1 of the Supporting Information, accompanied by additional details about the data generation and experiment settings. In principle, we expect MTCOV to exhibit better performance in this specific scenario due to its tailored development for such data and also its generative process aligning closely with the mechanism underlying the synthetic data. Nonetheless, despite the generality of our approach, we observe that PIHAM achieves comparable performance to MTCOV in link and attribute prediction, as well as community detection, especially in scenarios involving denser networks. These results collectively show that PIHAM is a valid approach even in less heterogeneous scenarios, as it can compete effectively with bespoke existing methods.

Validation on heterogeneous data. Having demonstrated that PIHAM performs comparably well to existing methods for attributed multilayer networks, we now demonstrate its behavior on more complex data containing heterogeneous information. To the best of our knowledge, this is the first probabilistic generative model designed to handle and perform inference on such data, and as a result, a comparative analysis is currently unavailable. Additionally, due to the absence of alternative benchmarks for data generation, we validate the performance of our method on synthetic data generated using the model introduced in this work.

We analyze attributed multilayer networks with $L = 3$ heterogeneous layers: one with binary interactions, one with nonnegative discrete weights, and one with real values. In addition, each node is associated with three covariates: one categorical with $Z = 4$ categories, one representing nonnegative discrete values, and one involving real values. To generate these networks, we initially draw the latent variables $\boldsymbol{\Theta} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{H})$ from Gaussian distributions with specified hyperparameters. Subsequently, we generate $\boldsymbol{A}$ and $\boldsymbol{X}$ according to the data types, following Eq. (2) and Eq. (4). Our analysis spans networks with varying number of nodes $N \in \{100, 200, \dots, 1000\}$ and diverse number of overlapping communities $K \in \{3, 4, 5\}$. Additional details on the generation process can be found in the Supporting Information.

We assess the effectiveness of PIHAM by testing its prediction performance. To this end, we adopt a 5-fold cross-validation procedure, where we estimate the model's parameters on the training set and subsequently evaluate its prediction performance on the test set (see the Supporting Information for details). The presence of heterogeneous information complicates the measurement of goodness of fit, as distinct data types impose different constraints and domains. To address this complexity, we employ different metrics tailored to assess the prediction performance of each type of information. Specifically, we use the Area Under the receiver-operator Curve (AUC) for binary interactions, the Maximum Absolute Error (MAE) for nonnegative discrete values, the Root Mean Squared Error (RMSE) for real values, and the accuracy for categorical attributes. Further exploration to determine a unified metric could be a subject of future research.

The results are illustrated in Fig. 2, where the performance of PIHAM is compared against baselines given by the predictions obtained from either the average or the maximum frequency in the training set. For the categorical attribute, we also include the uniform random probability over Z, and for the AUC, the baseline corresponds to the random choice 0.5. Overall, PIHAM outperforms the baselines significantly for each type of information, with performance slightly decreasing as K increases. This is somewhat expected, considering the increased complexity of the scenarios. On the other hand, the performance remains consistent across varying values of N, indicating the robustness of our method and its suitability for larger networks.

Interpretation of posterior estimates. We have showcased the prediction performance of PIHAM across diverse synthetic datasets, and we now delve into the qualitative insights that can be extracted from the inferred parameters. In particular, we focus on the membership matrix $\boldsymbol{U}$. For this purpose, we examine the results obtained through the analysis of the synthetic data used in the section Comparison with existing methods in a homogeneous scenario, where ground truth mixed-memberships are represented as normalized vectors summing to 1. This scenario is particularly relevant for illustrating an example where the desired parameter space, defined by the simplex, differs from the inferred one existing in real-space.

To ease visualizations, we investigate a randomly selected network and focus on three representative nodes with distinct ground truth memberships: Node A has extreme mixed-membership, Node B slightly less mixed-membership, and Node C exhibits hard-membership. The results are depicted in Fig. 3, with the top row displaying the ground truth membership vectors for these representative nodes. In the middle row, we plot the inferred posterior distributions $\hat{U}_{ik} \sim \mathcal{N}(\hat{U}_{ik}; \hat{\mu}^U_{ik}, (\hat{\sigma}^U_{ik})^2)$, where different colors represent distinct communities (in this case, $K = 3$). Through a comparative analysis of the three distributions for each node, we can gain insights into the nodes' behaviors: Node A exhibits greater overlap among the three distributions, Node B shows a slighter shift toward K_1, while Node C distinctly aligns more with community K_3. This preliminary investigation leads to the conclusion that the inferred communities reflect the ground truth behaviors. However, interpreting such patterns can be challenging, if not unfeasible, especially when dealing with large datasets. To address this issue, we quantitatively compute the area of overlap between every pair of distributions for each node and then calculate the average. For this purpose, we use the implementation proposed in (35) and we name this measure as *Overlap*. This metric ranges from 0 (indicating no overlap) to 1 (representing perfect matching between the distributions). Notably, the overlap decreases as we move from Node A to Node C, in line with the decreasing degree of mixed-membership.

Computing the Overlap for many communities can be computationally expensive due to the need to calculate all

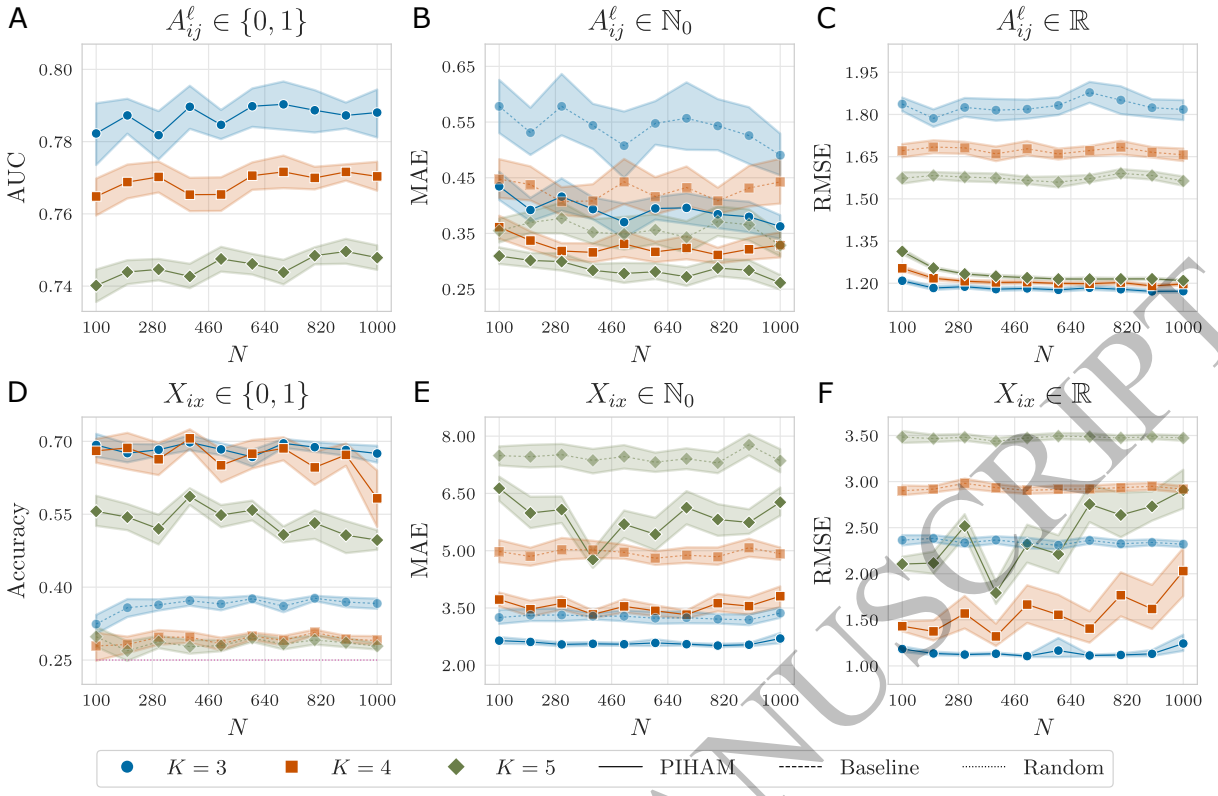

Fig. 2. Prediction performance on synthetic data. We analyze synthetic attributed multilayer networks with $L = 3$ heterogeneous layers (one with binary interactions (A), one with nonnegative discrete weights (B), and one with real values(C)), three node covariates (one categorical with $Z = 4$ categories (D), one representing nonnegative discrete values (E), and one involving real values (F)), varying number of nodes N, and diverse number of overlapping communities K. We employ a 5-fold cross-validation procedure and plot averages and confidence intervals over 20 independent samples. The prediction performances are measured with different metrics according to the data type: Area Under the receiver-operator Curve (AUC) for binary interactions (A), the Maximum Absolute Error (MAE) for nonnegative discrete values (B, E), the Root Mean Squared Error (RMSE) for real values (C, F), and accuracy for categorical attributes (D). The baselines are given by the predictions obtained from either the average or the maximum frequency in the training set. For the categorical attribute, we also include the uniform random probability over Z, and for the AUC, the baseline corresponds to the random choice 0.5. Overall, PIHAM outperforms the baselines significantly for each type of information.

pairwise combinations. As an alternative solution, we suggest utilizing the L_2-barycenter distribution, which essentially represents a weighted average of the node-community distributions (36, 37). We show the barycenter distributions in gray in the second row of Fig. 3. This approach allows focusing on a single distribution per node, instead of K different ones. To quantify this distribution, we calculate its variance (σ^2), where higher values indicate nodes with harder memberships, as the barycenter is more spread due to the individual distributions being more distant from each other. Conversely, lower variance suggests more overlap among the distributions, indicating a more mixed-membership scenario. We observe that σ^2 increases as we decrease the degree of mixed-membership, a trend consistent with that of the Overlap. Further details on the barycenter distribution and the metrics are provided in the Supporting Information.

To facilitate interpretability, a practitioner may desire to work within the simplex space. This also reflects the ground truth parameter space, as opposed to the normal posterior distributions. As discussed in the section Parameter interpretation, PIHAM employs the LM technique. This has the capability to transform in a principled way every membership vector $\hat{\boldsymbol{U}}_{\boldsymbol{i}}$ into the simplex space using Dirichlet distributions. The outcomes of this transformation are depicted in the bottom row of Fig. 3. By investigating these plots, it becomes even more apparent how the inferred memberships closely resemble the ground truth: the Dirichlet distributions gradually concentrate more towards a specific corner (K_1 for Node B and K_3 for Node C), instead of spreading across the entire area (as observed for Node A).

With this example, we presented a range of solutions for interpreting the posterior distributions associated with the inferred node memberships. These options are not exhaustive, and other approaches may also be considered. For instance, a practitioner might focus solely on analyzing the point estimates for the sake of facilitating comparisons with the ground truth. In such cases, as discussed in the section Parameter interpretation, two procedures can be employed: i) applying a transformation to the point estimates, such as softmax, to align them with the ground truth space, or ii) using a sufficient statistic of the posterior distribution, where the mean of the Dirichlet distribution is a suitable option. The choice between these various approaches should be guided by the specific application under study, and the provided example serves as just one illustration.

Analysis of a social support network of a rural Indian village. We now turn our attention to the analysis of a real-world dataset describing a social support network within a village in Tamil Nadu, India, referred to as "Aḷakāpuram" (38, 39). The

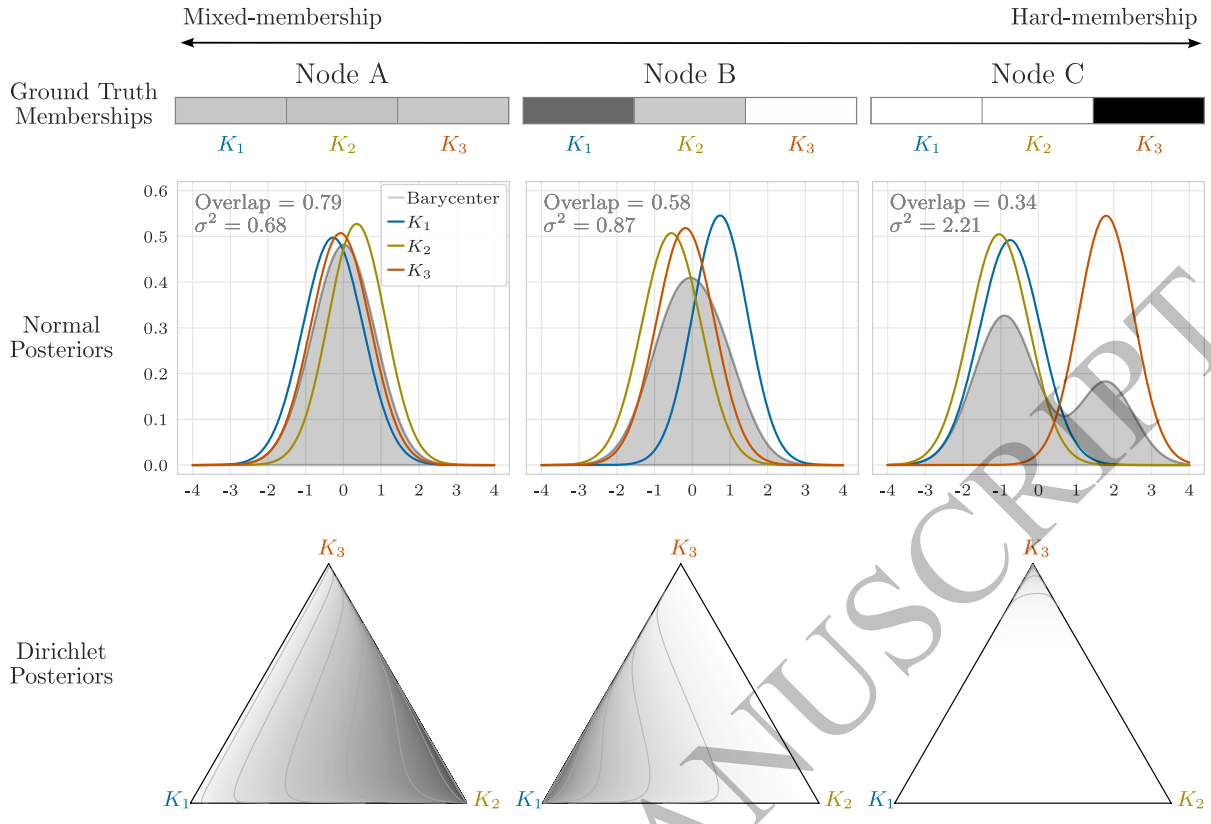

Fig. 3. Interpretation of posterior distributions in comparison with ground truth memberships. We analyze a synthetic attributed multilayer network with ground truth mixed-memberships represented as normalized vectors summing to 1. In this case, $K = 3$. (Top row) Ground truth membership vectors for three representative nodes: Node A displays extreme mixed-membership, Node B shows a slightly lower mixed-membership, and Node C exhibits hard-membership. (Middle row) Inferred posterior distributions $\hat{U}_{ik} \sim \mathcal{N}(\hat{U}_{ik}; \hat{\mu}^U_{ik}, (\hat{\sigma}^U_{ik})^2)$, where different colors represent distinct communities, and the distribution in gray consists of the L_2-barycenter distribution. Overlap is the average of the area of overlap between every pair of distributions, and σ^2 is the variance of the barycenter distribution. (Bottom row) Transformed posterior distributions into the simplex space using the LM technique and employing Dirichlet distributions. The inferred node memberships reflect the ground truth behavior, as evidenced by the trends of Overlap and σ^2, which align with the decreasing degree of true mixed-membership. Additionally, the Dirichlet transformation provides a more straightforward interpretation, further supporting this conclusion.

data were collected in 2013 through surveys, in which adult residents were asked to nominate individuals who provided various types of support, such as running errands, offering advice, and lending cash or household items. Additionally, several attributes were gathered, encompassing information like gender, age, and caste, among others. The pre-processing of the dataset is described in the Supporting Information. The resulting heterogeneous attributed multilayer network comprises $N = 419$ nodes, $L = 7$ layers, and $P = 3$ node attributes. The initial six layers depict directed binary social support interactions among individuals, with average degree ranging from 1.8 to 4.2. The seventh, instead, contains information that is proportional to the geographical distance between individuals' households. The adjacency tensor is then represented as $\boldsymbol{A} = \{\boldsymbol{A^\ell} \in \{0,1\}^{N\times N}\, \forall \ell \in [1,6], \boldsymbol{A^7} \in \mathbb{R}_+^{N\times N}\}$. As node covariates, we consider the caste attribute with $Z_{caste} = 14$ categories, the religion attribute with $Z_{religion} = 3$ categories, and the attribute representing the years of education, that is $\boldsymbol{X_{\cdot 3}} \in \mathbb{N}_0^N$. Ethnographic work and earlier analyses (39, 40) suggest that these attributes play an important role in how villagers relate to one another, with certain relationships being more strongly structured by these identities than others.

Inference, prediction performance, and goodness of fit. We describe the likelihood of the real-world heterogeneous attributed multilayer network according to Eq. (2) and Eq. (4), customized to suit the data types under examination. In particular, we employ Bernoulli distributions for the binary layers $[\boldsymbol{A^\ell}]_{\ell\in[1,6]}$ and Gaussian distributions for the distance layer $\boldsymbol{A^7}$. Moreover, we characterize the attributes caste $\boldsymbol{X_{\cdot 1}}$ and religion $\boldsymbol{X_{\cdot 2}}$ using Categorical distributions, and model the covariate $\boldsymbol{X_{\cdot 3}}$ with a Poisson distribution. The choice of the model hyperparameters and the algorithmic settings used in our experiments are described in the Supplementary Information.

Similarly to many real-world datasets, we lack the information about the true parameters underlying the network, including the node memberships. Hence, to determine the number of communities K, we employ a 5-fold cross-validation procedure for $K \in [1, 10]$ and select the value that exhibits the optimal performance. Detailed results are displayed in Table S2 of the Supporting Information. We set $K = 6$ as it achieves the best performance across the majority of prediction metrics. In fact, selecting a single metric to summarize and evaluate results in a heterogeneous setting is nontrivial, as discussed in the section Validation on heterogeneous data. The results in Table S2 additionally validate PIHAM's performance in inference tasks like edge

and covariate prediction. Overall, our method demonstrates robust outcomes with the chosen fixed value of K and consistently outperforms the baselines, which are omitted for brevity.

We further evaluate our model's goodness of fit through a posterior-predictive assessment (41, 42), comparing the input data to synthetic data generated by the fitted model. A well-fitted model should produce synthetic data that closely resemble the original input. To accurately assess performance, we test whether two samples from the posterior-predictive distribution are generally more, equally, or less distant from each other than a sample from the posterior-predictive distribution compared to the input data (41). We measure the distance using different metrics depending on the data type, and the results are shown in Fig. S3. The discrepancies between synthetic data samples consistently exceed those between the observed data and synthetic samples, indicating that PIHAM provides a good fit for the data.

Qualitative interpretation of the inferred parameters. We now shift our attention to analyze the results qualitatively, specifically focusing on the inferred communities. For easier interpretation, we apply a softmax transformation to the MAP estimates $\hat{\boldsymbol{\mu}}_i^U$, allowing us to treat node memberships as probabilities. Opting for the softmax over the mean of the posterior Dirichlet distributions is primarily for visualization purposes, as it results in slightly less mixed-memberships, thereby improving clarity. The middle and bottom rows of Fig. 4 depict the inferred out-going communities $\hat{\boldsymbol{U}}_i$, where darker values in the grayscale indicate higher values in the membership vector. In addition, the top row of Fig. 4 displays the node attributes included in our analysis. Note also that the nodes' position reflects the geographical distance between individuals' households, and the depicted interactions refer to the first layer (talk about important matters). A full representation of the six binary layers is shown in Fig. S2 of the Supporting Information.

Upon initial examination, we observe a correspondence between various detected communities and the covariate information. For instance, the first and second communities predominantly consist of nodes belonging to the Yātavar and Paṟaiyar castes, respectively. Similarly, K_3 comprises nodes from the Kulālar and Maṟavar castes. This observation is supported by the inferred $K \times Z_{caste}$-dimensional matrix $\hat{\boldsymbol{H}}_{\cdot\mathbf{1}}$ (see Fig. S4 in the Supporting Information), which explains the contributions of each caste category to the formation of the k-th community. Furthermore, the affinity tensor $\hat{\boldsymbol{W}}$ (see Fig. S6 in the Supporting Information) suggests that these communities have an assortative structure, where nodes tend to interact more with individuals belonging to the same community as with those from different communities. This pattern reflects a typical behavior in social networks (43). Additionally, note that these communities contain nodes that are geographically close to each other and, in some cases, very distant from the majority.

In contrast to the first three, communities K_4, K_5, and K_6 are more nuanced. In fact, they are predominantly comprised of nodes from the Paḷḷar caste, which, however, is also the most represented caste in the dataset. Despite that, we observe some differences by examining other parameters. For instance, K_4 exhibits a strong assortative community structure, contrasting with the less structured nature of K_5 and K_6. This suggests that interactions play a more relevant role than attributes in determining the memberships of K_4. On the other hand, the attribute $\boldsymbol{X}_{\cdot\mathbf{3}}$ seems to play a bigger role in determining K_6, which includes nodes with more years of education. This correlation is depicted in Fig. S5 of the Supporting Information, where the posterior distribution $\mathcal{N}(\hat{H}_{63}; \hat{\mu}_{63}^H, (\hat{\sigma}_{63}^H)^2)$ of education years in K_6 significantly differs and is distant from the others.

By looking at the affinity matrices of the seven layers in Fig. S6, we see how layers have predominantly an assortative structure, but show also variations for certain layers. For instance, L_2 (help finding a job) has few non-zero diagonal values, suggesting that this type of support is one for which people must sometimes seek out others in different communities. In particular, L_7, corresponding to the geographical distance between nodes, has several off-diagonal entries, particularly for communities K_4, K_5, and K_6, suggesting a weakened effect for physical proximity for those communities.

Taken together, these findings suggest that the inferred communities do not solely correlate with one type of information, which may be the most dominant. Instead, PIHAM utilizes all the input information to infer partitions that effectively integrate all of them in a meaningful manner. In addition, the inferred affinity matrices illustrate how different layers can exhibit different community structures, a diversity that can be captured by our model.

Discussion

In this work, we have introduced PIHAM, a probabilistic generative model designed to perform inference in heterogeneous and attributed multilayer networks. A significant feature of our approach is its flexibility to accommodate any combination of the input data, made possible through the use of Laplace approximations and automatic differentiation methods, which avoid the need for explicit derivations. However, it is important to note that having a method capable of handling complex network datasets does not automatically ensure the quality of the input data. For example, if only certain attributes are relevant for explaining the networked dataset, or if only specific layers contain valuable information for the task, adding unnecessary information could be detrimental. Practitioners must carefully assess which information is useful based on their specific objectives. Alternatively, model selection tests, such as the cross-validation routine demonstrated in the manuscript, can be used to determine the optimal combination of layers and attributes.

When compared to other methods tailored for scenarios with only one type of attribute and interaction, PIHAM demonstrates comparable performance in prediction and community detection tasks, despite its broader formulation. Moreover, our approach significantly outperforms baseline metrics in more complex settings characterized by various attribute and interaction types, where existing methods for comparison are lacking. Additionally, PIHAM employs a Bayesian framework, enabling the estimation of posterior distributions, rather than only providing point estimates for the parameters. And, through the use of the Laplace matching technique, it maps these posterior distributions to various desired domains in a theoretically sound manner, facilitating interpretation.

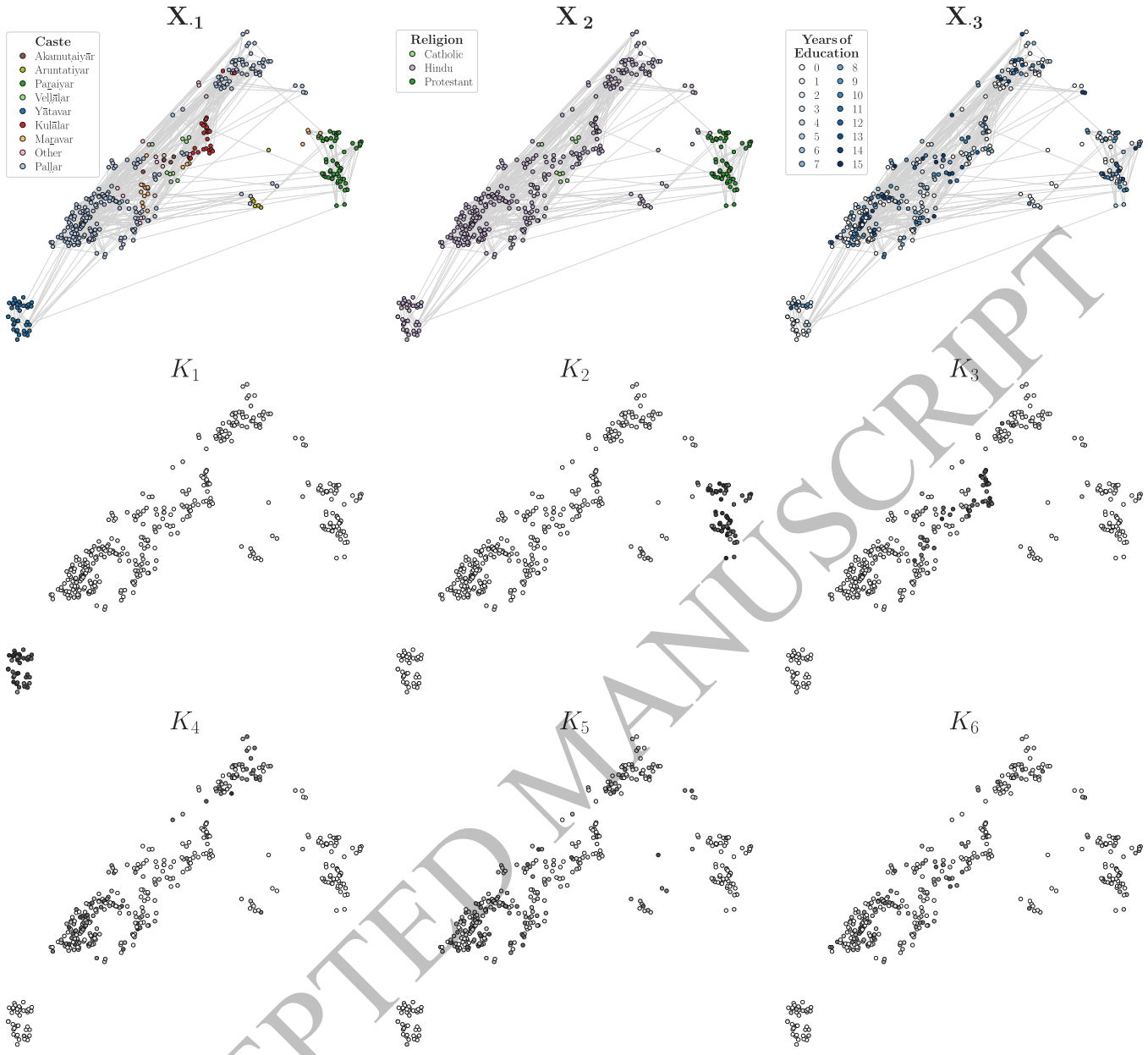

Fig. 4. Inference of overlapping communities in a social support network. We analyze a real-world heterogeneous attributed multilayer network, which was collected in 2013 through surveys in the Indian village. This network comprises six binary layers representing directed social support interactions among individuals, alongside an additional layer reflecting information proportional to the distance between individuals' households. (Top row) As node covariates, we consider caste $\boldsymbol{X}_{\cdot 1}$, religion $\boldsymbol{X}_{\cdot 2}$, and years of education $\boldsymbol{X}_{\cdot 3}$. For privacy reasons, nodes belonging to castes with fewer than five individuals are aggregated into an "Other" category. Moreover, the displayed interactions refer only to the first layer (talk about important matters) to enhance clarity in visualization. (Middle-Bottom rows) We display the MAP estimates of the out-going communities inferred by PIHAM. For easier interpretation, we apply a $\mathrm{softmax}$ transformation to the MAP estimates of the membership vectors, and darker values in the grayscale indicate higher values in the membership vector $\hat{\boldsymbol{U}}_i$. The position of the nodes reflects the geographical distance between individuals' households. In summary, the inferred communities do not exclusively align with a single type of information. Rather, PIHAM incorporates all input information to infer partitions that effectively integrate them in a meaningful way.

While PIHAM constitutes a principled and flexible method to analyze heterogeneous and attributed multilayer networks, several questions remain unanswered. For example, determining the most appropriate metric for summarizing prediction performance in heterogeneous scenarios, where information spans different spaces, is not straightforward. This aspect also influences the selection of the optimal model during cross-validation procedures. While we have provided explanations for our choices, we acknowledge that this remains an open question. Similarly, when dealing with many communities, summarizing posterior distributions becomes challenging due to computational constraints. We addressed this issue by employing L_2-barycenter distributions and proposing their variance to guide interpretation. Nevertheless, we believe there is still considerable room for improvement and exploration in this area. Moreover, we have considered here mixed-memberships, but in certain data-scarce scenarios hard-membership approaches with fewer parameters could be better suited. Future work should consider how to flexibly drive parameters' inference towards mixed or hard memberships,

based on the input data. Our method could be further extended to accommodate distinct community-covariate contributions by integrating two separate $\boldsymbol{H}$ matrices for both in-coming and out-going communities, respectively. This modification will offer clearer insights into how covariates influence the partitions, especially when discrepancies arise between in-coming and out-going communities. Moreover, we treated the number of communities K as given, and for real-world data it was selected via cross-validation. While this procedure is grounded, it can be computationally expensive. An alternative approach would be to treat K as a model parameter and infer it directly from the data. Lastly, it would be interesting to expand this framework to incorporate higher-order interactions, an emerging area that has shown relevance in describing real-world data (44).

In summary, PIHAM offers a flexible and effective approach for modeling heterogeneous and attributed multilayer networks, which arguably better captures the complexity of real-world data, enhancing our capacity to understand and analyze the organization of real-world systems.

Data availability

Synthetic data used in the paper are explained in the Supplementary Material. Anonymized data of the social support network are available from the corresponding author E. A. Power, upon reasonable request.

Code availability

An open-source algorithmic implementation of the model is publicly available and can be found at https://github.com/mcontisc/PIHAM.

ACKNOWLEDGMENTS. M.C. and C.D.B. were supported by the Cyber Valley Research Fund. M.C. acknowledges support from the International Max Planck Research School for Intelligent Systems (IMPRS-IS). M.C., C.D.B., and E.A.P. were supported by a UKRI Economic and Social Research Council Research Methods Development Grant (ES/V006495/1). M.H. and P.H. acknowledge support from the European Research Council through ERC CoG Action 101123955 ANUBIS; the DFG Cluster of Excellence "Machine Learning - New Perspectives for Science", EXC 2064/1, project number 390727645; the German Federal Ministry of Education and Research (BMBF) through the Tübingen AI Center (FKZ: 01IS18039A); as well as funds from the Ministry of Science, Research and Arts of the State of Baden-Württemberg. This manuscript was posted on a preprint: http://arxiv.org/abs/2405.20918.

1. M Newman, *Networks.* (Oxford university press), (2018).
2. M De Domenico, More is different in real-world multilayer networks. *Nat. Phys.* **19**, 1247–1262 (2023).
3. L Bargigli, G Di Iasio, L Infante, F Lillo, F Pierobon, The multiplex structure of interbank networks. *Quant. Finance* **15**, 673–691 (2015).
4. K Higham, M Contisciani, C De Bacco, Multilayer patent citation networks: A comprehensive analytical framework for studying explicit technological relationships. *Technol. forecasting social change* **179**, 121628 (2022).
5. PJ Mucha, T Richardson, K Macon, MA Porter, JP Onnela, Community structure in time-dependent, multiscale, and multiplex networks. *science* **328**, 876–878 (2010).
6. F Battiston, V Nicosia, M Chavez, V Latora, Multilayer motif analysis of brain networks. *Chaos: An Interdiscip. J. Nonlinear Sci.* **27** (2017).
7. J Liu, J Wang, B Liu, Community detection of multi-layer attributed networks via penalized alternating factorization. *Mathematics* **8**, 239 (2020).
8. S Xu, Y Zhen, J Wang, Covariate-assisted community detection in multi-layer networks. *J. Bus. & Econ. Stat.* **41**, 915–926 (2023).
9. Z Pei, X Zhang, F Zhang, B Fang, Attributed multi-layer network embedding in *2018 IEEE International Conference on Big Data (Big Data).* (IEEE), pp. 3701–3710 (2018).
10. Y Cen, et al., Representation learning for attributed multiplex heterogeneous network in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining.* pp. 1358–1368 (2019).
11. J Cao, D Jin, L Yang, J Dang, Incorporating network structure with node contents for community detection on large networks using deep learning. *Neurocomputing* **297**, 71–81 (2018).
12. C Park, D Kim, J Han, H Yu, Unsupervised attributed multiplex network embedding in *Proceedings of the AAAI Conference on Artificial Intelligence.* Vol. 34, pp. 5371–5378 (2020).
13. B Han, Y Wei, L Kang, Q Wang, Y Yang, Node classification in attributed multiplex networks using random walk and graph convolutional networks. *Front. Phys.* **9**, 763904 (2022).
14. L Martirano, L Zangari, A Tagarelli, Co-mlhan: contrastive learning for multilayer heterogeneous attributed networks. *Appl. Netw. Sci.* **7**, 1–44 (2022).
15. A Goldenberg, AX Zheng, SE Fienberg, EM Airoldi, , et al., A survey of statistical network models. *Foundations Trends Mach. Learn.* **2**, 129–233 (2010).
16. L Peel, TP Peixoto, M De Domenico, Statistical inference links data and theory in network science. *Nat. Commun.* **13**, 6794 (2022).
17. R Ranganath, S Gerrish, D Blei, Black box variational inference in *Artificial intelligence and statistics.* (PMLR), pp. 814–822 (2014).
18. D Tran, R Ranganath, DM Blei, Variational gaussian process in *4th International Conference on Learning Representations, ICLR 2016.* (2016).
19. I Valera, MF Pradier, M Lomeli, Z Ghahramani, General latent feature models for heterogeneous datasets. *The J. Mach. Learn. Res.* **21**, 4027–4075 (2020).
20. A Nazabal, PM Olmos, Z Ghahramani, I Valera, Handling incomplete heterogeneous data using vaes. *Pattern Recognit.* **107**, 107501 (2020).
21. C Tallberg, A bayesian approach to modeling stochastic blockstructures with covariates. *J. Math. Sociol.* **29**, 1–23 (2004).
22. J Yang, J McAuley, J Leskovec, Community detection in networks with node attributes in *2013 IEEE 13th international conference on data mining.* (IEEE), pp. 1151–1156 (2013).
23. D Hric, TP Peixoto, S Fortunato, Network structure, metadata, and the prediction of missing nodes and annotations. *Phys. Rev. X* **6**, 031038 (2016).
24. ME Newman, A Clauset, Structure and inference in annotated networks. *Nat. communications* **7**, 11863 (2016).
25. A White, TB Murphy, Mixed-membership of experts stochastic blockmodel. *Netw. Sci.* **4**, 48–80 (2016).
26. N Stanley, T Bonacci, R Kwitt, M Niethammer, PJ Mucha, Stochastic block models with multiple continuous attributes. *Appl. Netw. Sci.* **4**, 1–22 (2019).
27. O Fajardo-Fontiveros, R Guimerà, M Sales-Pardo, Node metadata can produce predictability crossovers in network inference problems. *Phys. Rev. X* **12**, 011010 (2022).
28. M Contisciani, EA Power, C De Bacco, Community detection with node attributes in multilayer networks. *Sci. Reports* **10**, 15736 (2020).
29. M Hobbhahn, P Hennig, Laplace matching for fast approximate inference in latent gaussian models. *arXiv preprint arXiv:2105.03109* (2021).
30. C De Bacco, EA Power, DB Larremore, C Moore, Community detection, link prediction, and layer interdependence in multilayer networks. *Phys. Rev. E* **95**, 042317 (2017).
31. P McCullagh, *Generalized linear models.* (Routledge), (2019).
32. A Kucukelbir, D Tran, R Ranganath, A Gelman, DM Blei, Automatic differentiation variational inference. *J. machine learning research* (2017).
33. AP Dempster, NM Laird, DB Rubin, Maximum likelihood from incomplete data via the em algorithm. *J. royal statistical society: series B (methodological)* **39**, 1–22 (1977).
34. DM Blei, A Kucukelbir, JD McAuliffe, Variational inference: A review for statisticians. *J. Am. statistical Assoc.* **112**, 859–877 (2017).
35. MP Wand, JT Ormerod, SA Padoan, R Frühwirth, Mean Field Variational Bayes for Elaborate Distributions. *Bayesian Analysis* **6**, 847 – 900 (2011).
36. JD Benamou, G Carlier, M Cuturi, L Nenna, G Peyré, Iterative bregman projections for regularized transportation problems. *SIAM J. on Sci. Comput.* **37**, A1111–A1138 (2015).
37. CL Coz, A Tantet, R Flamary, R Plougonven, A barycenter-based approach for the multi-model ensembling of subseasonal forecasts. *arXiv:2310.17933* (2023).
38. EA Power, *Building bigness: Religious practice and social support in rural South India.* (Stanford University), (2015).
39. EA Power, Social support networks and religiosity in rural south india. *Nat. Hum. Behav.* **1**, 0057 (2017).
40. EA Power, E Ready, Building bigness: Reputation, prominence, and social capital in rural south india. *Am. Anthropol.* **120**, 444–459 (2018).
41. A Gelman, XL Meng, H Stern, Posterior predictive assessment of model fitness via realized discrepancies. *Stat. sinica* pp. 733–760 (1996).
42. A Gelman, et al., *Bayesian Data Analysis.* (CreateSpace, United States), 3rd ed edition, (2013).
43. ME Newman, Assortative mixing in networks. *Phys. review letters* **89**, 208701 (2002).
44. A Badalyan, N Ruggeri, C De Bacco, Structure and inference in hypergraphs with node attributes. *Nat. Commun.* **15**, 7073 (2024).