Node Metadata Can Produce Predictability Crossovers in Network Inference Problems

Oscar Fajardo-Fontiveros,[1,‡] Roger Guimerà,[2,1,*] and Marta Sales-Pardo[1,†]

[1]*Department of Chemical Engineering, Universitat Rovira i Virgili, 43007 Tarragona, Catalonia*
[2]*ICREA, 08010 Barcelona, Catalonia*

(Received 8 April 2021; accepted 16 November 2021; published 14 January 2022)

Network inference is the process of learning the properties of complex networks from data. Besides using information about known links in the network, node attributes and other forms of network metadata can help solve network inference problems. Indeed, several approaches have been proposed to introduce metadata into probabilistic network models and to use them to make better inferences. However, we know little about the effect of such metadata in the inference process. Here, we investigate this issue. We find that, rather than affecting inference gradually, adding metadata causes a crossover in the inference process and in our ability to make accurate predictions, from a situation in which metadata do not play any role to a situation in which metadata completely dominate the inference process. When network data and metadata are partly correlated, metadata optimally contributes to the inference process at the crossover between data-dominated and metadata-dominated regimes.

DOI: 10.1103/PhysRevX.12.011010 Subject Areas: Complex Systems, Statistical Physics

I. INTRODUCTION

Many systems can be represented as networks, with nodes representing units (for example, people in a social network, or proteins in a protein-protein interaction network), and links representing interactions between the units (for example, friendship relationships between people or physical binding interactions between proteins). Network inference is the process of inferring the properties of those networks from data; typical network inference problems include the identification of groups of nodes with similar connection patterns or the identification of unobserved interactions, that is, link prediction [1–6]. Network inference and, in particular, link prediction are increasingly important in problems with applications ranging from the prediction of interactions between drugs [7–9] to the prediction of human preferences and decisions [10–14].

Typically, network inference starts from observations of some of the links in the network, which are used to predict unobserved links or to infer other network properties. However, other sources of information such as system dynamics [15,16] or node attributes [13,17–25] can also be used to aid in the inference process. Here, we study how node attributes are introduced in the inference process and what the effect of using such metadata is. In this regard, our work is in line with previous work treating metadata as additional data [13,17–25] and warning against the practice of using metadata as "ground truth" against which inference approaches should be evaluated [23]. Indeed, as we show below, metadata are sometimes uninformative and even misleading to the inference process; therefore, in general, they should not be used as ground truth without good theoretical arguments.

We present our work in terms of the problem of link prediction in recommender systems [11,12,26], in which the goal is to predict the association between users and items (for example, books or movies). However, our conclusions apply to model-based, probabilistic approaches to network inference, in general. We introduce a multipartite network model that encompasses and generalizes previous attempts to use node metadata in network inference problems (Fig. 1). Within this framework, the problem of link prediction in general unipartite or bipartite networks is just a particular case. Unlike most previous approaches, our multipartite network model allows us to control the importance of the node metadata and thus to investigate when and how metadata help in the inference.

We find that, contrary to what one may expect, node metadata do not affect the inference problem gradually. Rather, even when the weight of metadata increases smoothly, the inference process undergoes a crossover from a situation in which metadata do not play any role to a situation in which metadata completely dominate the inference process. When network data and metadata are

[*] Corresponding author.
roger.guimera@urv.cat
[†] Corresponding author.
marta.sales@urv.cat
[‡] oscar.fajardo@urv.cat

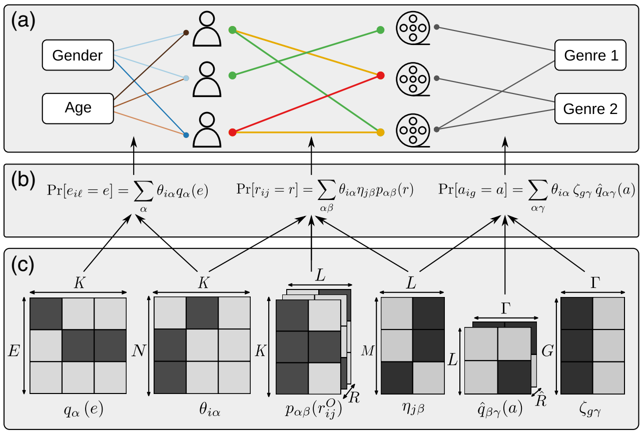
(a)
Gender
Age
Genre 1
Genre 2
(b)
$\Pr[e_{i\ell}=e]=\sum_{\alpha}\theta_{i\alpha}q_{\alpha}(e)$
$\Pr[r_{ij}=r]=\sum_{\alpha\beta}\theta_{i\alpha}\eta_{j\beta}p_{\alpha\beta}(r)$
$\Pr[a_{ig}=a]=\sum_{\alpha\gamma}\theta_{i\alpha}\,\zeta_{g\gamma}\,\hat{q}_{\alpha\gamma}(a)$
(c)
K
E
$q_{\alpha}(e)$
K
N
$\theta_{i\alpha}$
L
K
R
$p_{\alpha\beta}(r^{O}_{ij})$
L
M
$\eta_{j\beta}$
Γ
L
$\hat{R}$
$\hat{q}_{\beta\gamma}(a)$
Γ
G
$\zeta_{g\gamma}$

FIG. 1. Multipartite mixed-membership stochastic block model with labeled links. In panel (a), we cast the recommendation problem (in which one aims to predict how users will rate certain items) into a network inference problem. Here, users rate movies with three possible ratings (green, orange, or red). Additionally, we have excluding attributes for users (two excluding genders and three excluding age groups, represented by different shades of the same color) and nonexcluding attributes for movies (two movie genres; the connection to these attributes is binary, yes or no, but, in general, it does not need to be). Similar to ratings, we represent these attributes as bipartite networks. Although we frame our description of the model in terms of recommendations or link prediction in a bipartite network, the problem of link prediction in regular unipartite networks is just a particular case in which user nodes and item nodes are the same. In panel (b), each bipartite network in the multipartite network is modeled using a mixed-membership stochastic block model (see text). The individual block models are coupled by the user and item membership vectors (**θ** and **η**, respectively), shown in panel (c) along with all other model parameters and their dimensions (see text).

partly correlated, metadata optimally contribute to the inference process at the crossover between data-dominated and metadata-dominated regimes. This crossover is reminiscent of (but distinct from, in that it is induced by the metadata) transitions in the detectability of node groups [27–29] or in semisupervised network inference problems [30].

II. MULTIPARTITE MIXED-MEMBERSHIP STOCHASTIC BLOCK MODELS WITH LABELED LINKS

We introduce a very general network model based on stochastic block models [3,31,32] that allows us to deal with (directed or undirected) unipartite and bipartite networks, whose links are binary or labeled, and with node attributes of different types that can be combined as needed (Fig. 1). As we discuss below, this model extends and generalizes previous models.

In what follows, we use the terminology of recommender systems [11,12,26] although, as previously mentioned, the model is completely general and applicable to any type of relational data with node attributes. Our objective is to model a bipartite network with labeled links connecting N users to M items (for example, movies or books). Links r_{ij} represent ratings of users i to items j and are labeled; that is, r_{ij} can take values in a finite discrete set such as $\{\text{like}, \text{dislike}\}$, $\{\text{green}, \text{yellow}, \text{red}\}$, or $\{0, 1, \ldots, R\}$. To model these ratings, we assume that (i) there are user and item groups, and users and items belong to mixtures of such groups; (ii) the probability that a user i rates item j with r_{ij} depends only on the groups to which they belong.

These assumptions lead to a bipartite [10,11,33] mixed-membership [34] stochastic block model [12] in which the probability that user i gives item j a rating r is

$$\Pr[r_{ij}=r]=\sum_{\alpha\beta}\theta_{i\alpha}\eta_{j\beta}p_{\alpha\beta}(r). \tag{1}$$

Here, $\boldsymbol{\theta}_i$ is the normalized membership vector of user i, and each element $\theta_{i\alpha}$ represents the probability that user i belongs to group α (with $\sum_{\alpha}\theta_{i\alpha}=1$). Similarly, $\boldsymbol{\eta}_j$ is the normalized membership vector of item j; $\eta_{j\beta}$ represents the probability that item j belongs to group β. Finally, $p_{\alpha\beta}(r)$ is the probability that a user in group α and an item in group β are connected with a rating r. The normalization condition here is $\sum_r p_{\alpha\beta}(r)=1$.

We note that the association between nodes (users and items) and their attributes can also be represented as bipartite networks. Therefore, we can model node-attribute associations in a similar manner to ratings. Because we are interested in how node attributes can help in the inference of the model parameters $(\boldsymbol{\theta},\boldsymbol{\eta},\mathbf{p})$ for ratings, we consider that membership vectors for users $(\boldsymbol{\theta})$ and items $(\boldsymbol{\eta})$ in their respective attribute networks are the same as in the model for the ratings.

We consider both excluding and nonexcluding attributes. For excluding attributes, having one attribute precludes from having another; for example, a user's age group cannot be 30–39 years old and 40–49 years old simultaneously. We model each set of excluding attributes as a single attribute node (for example, an age node) that is connected to users or items through labeled links (each label representing a mutually excluding age group in the example). The probability that user i has an excluding attribute e (that is, the probability that the link $e_{i\ell}$ between user i and attribute node ℓ is of type e) is

$$\Pr[e_{i\ell}=e]=\sum_{\alpha}\theta_{i\alpha}q_{\alpha}(e), \tag{2}$$

where $q_{\alpha}(e)$ is the probability that a user of group α has an attribute of type e, and $\sum_e q_{\alpha}(e)=1$. For items, the expression is identical except that we use item membership vectors $\boldsymbol{\eta}$ instead of user membership vectors $\boldsymbol{\theta}$.

We also consider nonexcluding attributes, such as item genre (for example, a movie could be both "action" and "western"). We model each of these nonexcluding attribute types as individual attribute nodes connected to user or item nodes by links that are typically binary (either do or do not have the attribute) but that could, in general, also be labeled. Then, the probability that item i has attribute g of type a is also modeled using a mixed-membership, bipartite stochastic block model

$$\Pr[a_{ig} = a] = \sum_{\alpha\gamma} \theta_{i\alpha}\,\zeta_{g\gamma}\hat{q}_{\alpha\gamma}(a), \tag{3}$$

where $\zeta_{g\gamma}$ is the membership vector of attribute g and $\hat{q}_{\alpha\gamma}(a)$ is the probability that a user in group α has an attribute of type a for an attribute in attribute group γ. As before, the expression for item nonexcluding attributes is identical, just replacing user membership vectors $\boldsymbol{\theta}$ by item membership vectors $\boldsymbol{\eta}$.

III. MODEL POSTERIOR AND INFERENCE

Our objective is to model the observed ratings R^O and to predict the value of some unobserved ratings R. For this, and given Eq. (1), we need to infer the parameters $\boldsymbol{\theta}$, $\boldsymbol{\eta}$, and $\boldsymbol{p}$ from R^O. The posterior distribution over these parameters is given by

$$\begin{aligned} P(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{p}|R^O) &\propto P(R^O|\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{p})P(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{p}) \\ &= L^R(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{p})P(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{p}), \end{aligned} \tag{4}$$

where $L^R(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{p}) \equiv P(R^O|\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{p})$ is the likelihood of the model and $P(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{p})$ is the prior over model parameters. According to Eq. (1), the likelihood is

$$L^R(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{p}) = \prod_{(i,j)\in R^O}\left[\sum_{\alpha\beta}\theta_{i\alpha}\eta_{j\beta}p_{\alpha\beta}(r^O_{ij})\right]. \tag{5}$$

Similarly, if we decide to jointly model the ratings and the metadata encoded in the observed user and item attributes A^O, we also need to infer the values of the parameters $\boldsymbol{\zeta}$, $\boldsymbol{q}$, and $\hat{\boldsymbol{q}}$) using the posterior

$$\begin{aligned} P(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\zeta},\boldsymbol{p},\boldsymbol{q},\hat{\boldsymbol{q}}|R^O,A^O) \propto L^R(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{p})\prod_k L^{A_k}(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\zeta},\boldsymbol{q},\hat{\boldsymbol{q}}) \\ \times P(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\zeta},\boldsymbol{p},\boldsymbol{q},\hat{\boldsymbol{q}}), \end{aligned} \tag{6}$$

where $L^{A_k}(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\zeta},\boldsymbol{q},\hat{\boldsymbol{q}}) \equiv P(A^O_k|\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\zeta},\boldsymbol{q},\hat{\boldsymbol{q}})$ is the likelihood of the kth attribute network (for example, the age attribute network for users or the genre attribute network for items). For the kth excluding attribute, this likelihood reads

$$L^{A_k}(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{q}) = \prod_{(i,\ell_k)\in A^O_k}\left[\sum_{\alpha}\theta_{i\alpha}q^k_{\alpha}((e^O_k)_{i\ell_k})\right], \tag{7}$$

where ℓ_k is the kth excluding attribute and the product is over all nodes i for which we observe attribute ℓ_k.

For the kth nonexcluding attribute, we have

$$L^{A_k}(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\zeta},\hat{\boldsymbol{q}}) = \prod_{(i,g)\in A^O_k}\left[\sum_{\alpha\gamma}\theta_{i\alpha}\zeta^k_{g\gamma}\hat{q}^k_{\alpha\gamma}((a^O_k)_{ig})\right], \tag{8}$$

where the product is over all observed associations between nodes i and attributes g within the kth class of nonexcluding attributes.

Ignoring normalizing constants, and in a spirit similar to Refs. [18,25], we define a parametric log-posterior as

$$\begin{aligned} &\pi(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\zeta},\boldsymbol{p},\boldsymbol{q},\hat{\boldsymbol{q}}|R^O,A^O) \\ &\quad= \mathcal{L}^R(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{p}) + \sum_k \lambda_k\mathcal{L}^{A_k}(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\zeta},\boldsymbol{q},\hat{\boldsymbol{q}}), \end{aligned} \tag{9}$$

where

$$\begin{aligned} \mathcal{L}^R(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{p}) &= \log L^R(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{p}), \\ \mathcal{L}^{A_k}(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\zeta},\boldsymbol{q},\hat{\boldsymbol{q}}) &= \log L^{A_k}(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\zeta},\boldsymbol{q},\hat{\boldsymbol{q}}) \end{aligned}$$

are the log-likelihoods of ratings and attributes, respectively. For $\lambda_k = 0$, we recover Eq. (4) with uniform priors on the parameters, thus completely ignoring all metadata. Conversely, for $\lambda_k = 1$, we are jointly modeling the network of ratings and the network of attributes as in Eq. (6), with uniform priors on the parameters. By tuning the values of λ_k, we can interpolate between these situations and extrapolate to others with $\lambda_k > 1$; in the limit $\lambda_k \to \infty$, we only model the attribute network.

To interpret the parametric log-posterior in Eq. (9), consider the case where $\lambda_k \in \mathbb{N}^+$ is a natural number. In this case, for each observed attribute link, there are λ_k identical terms in the likelihood, exactly as if the corresponding attribute links were observed λ_k times independently. For example, the posterior we obtain with $\lambda_k = 2$ is identical to what we would obtain if we had two independent sources for attribute k, if both sources coincided in all cases, and if we modeled each of them with the $\lambda_k = 1$ generative model. Note that, because of this, each attribute likelihood is automatically normalized as in the $\lambda_k = 1$ case, except that, if there are M_k observed attribute links, the normalization is over the space of networks with $\lambda_k M_k$ links, instead of M_k. The case with other positive values of λ, $\{\lambda > 0; \lambda \notin \mathbb{N}^+\}$, can be interpreted as interpolating between integer values of λ_k.

Given this interpretation of the log-posterior as a whole, the terms corresponding to the attribute models can indistinctly be interpreted as part of the likelihood of a joint model of ratings and attributes, similar to Refs. [18–20,24,25], or as a nonuniform prior over membership vectors as in Refs. [17,21,22]. If interpreted as part of a joint model, then λ_k can be seen as some factors that are needed because attribute data are somehow less (or more) reliable than rating data, perhaps because we have reason to believe that attributes are more (or less) subject to noise, or because each rating corresponds, in fact, to a mean over several observations. Conversely, if interpreted as priors over the partitions, λ_k should be interpreted as hyperparameters defining how certain we are *a priori* about the importance of node attributes.

Either way, this parametrized posterior allows us to investigate how the metadata encoded in the attribute networks enter the inference process for the ratings, and under which conditions this results in better and more predictive models for those ratings. To do this, we maximize the posterior for fixed values of λ_k using an expectation-maximization (EM) algorithm [12,21,24,25] (see the Appendix A), which gives the most plausible parameter values (including group memberships). Because the posterior landscape is, in general, rugged, we perform several runs of the EM algorithm and compute the average probability for each unobserved rating to make predictions (see Ref. [12] and Appendix A).

IV. RELATIONSHIP TO PREVIOUS WORK

The literature on using metadata for link prediction and recommender systems is vast and includes all sorts of approaches ranging from simple heuristics to sophisticated machine learning methods. However, our interest here is more closely related to probabilistic approaches to network inference, even when those approaches are not applied directly to link prediction [13,17,18,20–23]—as shown in Refs. [19,24,25], once model parameters are inferred for, for example, community detection, they can easily be used to predict links as well.

Our focus on approaches based on probabilistic generative models is motivated by three characteristics of such approaches: (i) All assumptions in them are explicit; (ii) principled (as opposed to heuristic) and sometimes even exact inference approaches are possible; and (iii) their results are more readily interpretable. These three characteristics make probabilistic approaches especially appropriate for our ultimate goal of understanding how node attributes enter and help in the inference process.

From this perspective, the multipartite mixed-membership stochastic block model is useful because it extends and generalizes previous models. By introducing excluding and nonexcluding attributes, the model can simultaneously accommodate attributes like those considered in Refs. [21,25] (excluding) and in Refs. [18,20] (nonexcluding). It can also combine an arbitrary number of attributes of different types, unlike approaches that can only deal with single attributes [21,25] or, more often, with a single type of attribute; and it naturally deals with missing attribute data, unlike approaches that require all node attributes to be known [17,22]. Since attributes are modeled with a stochastic block model, our approach also automatically clusters attributes that have similar effects on the data (for example, age groups that show similar behavior) as in Ref. [20]. Unlike most previous approaches for attributed networks, nodes and attributes in our model belong to mixtures of groups, which makes the model more expressive [12]; links between nodes and to attributes can be labeled; and the influence of the attributes can be adjusted (as in Ref. [25]). As stated above, this last feature is precisely the main focus of our work.

V. SYNTHETIC DATA

We first use synthetic data to validate the expectation-maximization inference approach and to investigate the role of introducing node attributes. Here and throughout the validations in the coming sections, we quantify link prediction performance by measuring rating prediction accuracy, that is, the fraction of correctly predicted ratings in cross-validation experiments.

Our synthetic rating networks consist of 200 users and 200 items, partitioned into $K = 2$ groups of users and $L = 4$ groups of items. Users have an excluding attribute labeled "male" or "female," and items have an excluding attribute labeled from 0 to 3, which may represent four different genres. We generate the synthetic ratings $r_{ij} \in \{0, 1, 2\}$ with the model depicted in the central part of Fig. 1, the $\boldsymbol{p}$ matrices in Appendix C, and the membership vectors that we describe next.

Attribute links are generated as follows. In the simplest case, in which ratings and attributes are completely correlated, all female users have membership vectors $\boldsymbol{\theta}_{\mathrm{f}} = (0.8, 0.2)$; conversely, all male users have $\boldsymbol{\theta}_{\mathrm{m}} = (0.2, 0.8)$. Similarly, an item with attribute a has a membership of 0.8 to group a and 0.067 to all other groups. To simulate partial correlation c or even no correlation ($c = 0$) between membership vectors and attributes, with probability $1 - c$, we reassign each node attribute to a value selected uniformly at random among all possibilities (2 for users and 4 for items).

For the experiments reported in Fig. 2, we consider all attribute links but only a number $|R^O| = 400$ of observed ratings (that is, 1% of all generated ratings). Although the synthetic data are created with item genre as an excluding attribute, we carry out the inference process assuming that genre is a nonexcluding attribute. We do this for two reasons. First, in the empirical data set discussed below, most movies only have one assigned genre despite the fact that they could (and sometimes do) have more than one. Therefore, it seems reasonable to assign one genre and evaluate if the algorithm discovers this pattern despite having the freedom to assign several genres to each movie. Second, no empirical data set can be expected to be drawn exactly from a proposed generative model, so this provides us with a way to validate the expressiveness of the model, that is, its ability to fit data sets that were not drawn exactly from the model.

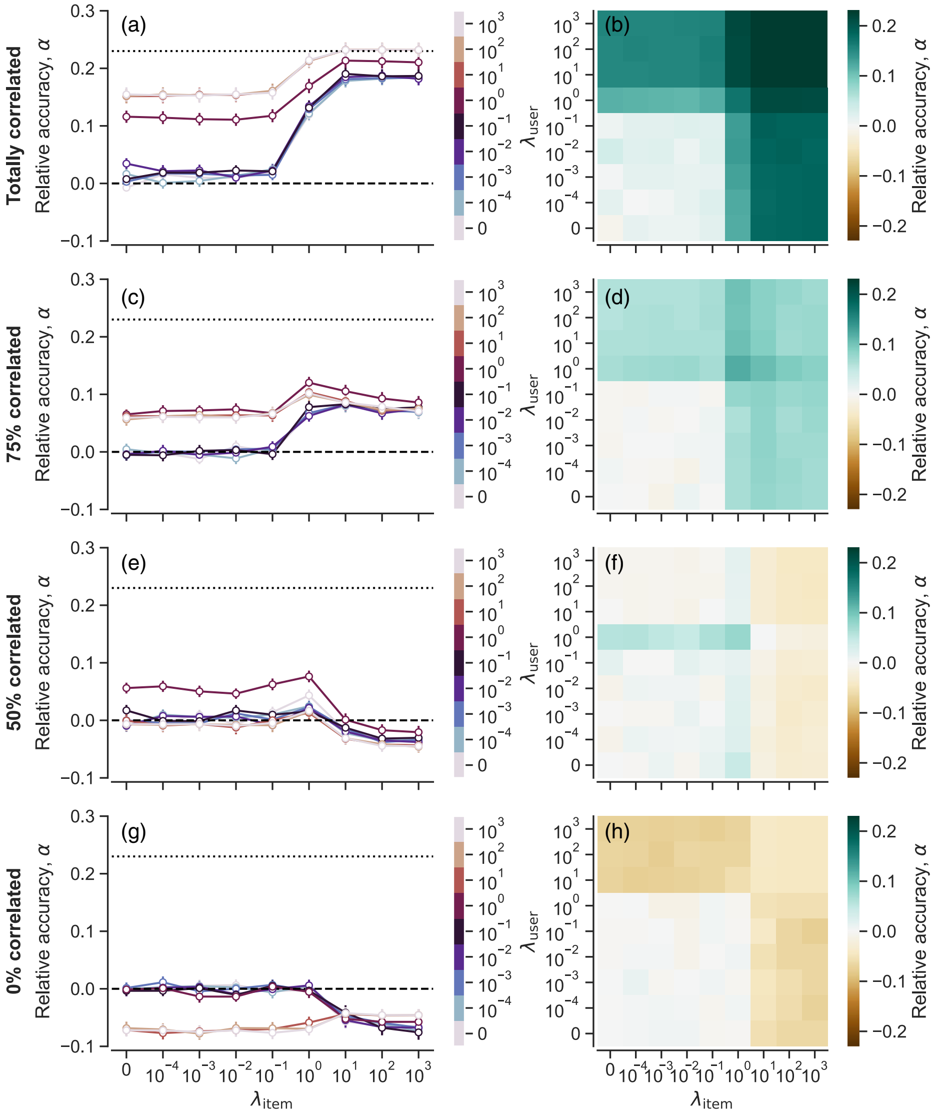

FIG. 2. Predictive performance and effect of metadata on synthetic ratings. We create synthetic ratings from 200 users on 200 items, with different levels of correlation c between ratings and node attributes (see text). We then use fivefold cross-validation to calculate the performance of the expectation-maximization equations at predicting unobserved ratings. We use accuracy a (that is, the fraction of correctly predicted ratings) as our performance metric; we take as a reference the predictive accuracy $a_0 = 0.468$ of the algorithm when all attributes are ignored ($\lambda_{\text{user}} = \lambda_{\text{item}} = 0$) and measure relative accuracy α for a given pair $(\lambda_{\text{user}}, \lambda_{\text{item}})$ as the log-ratio $\alpha(\lambda_{\text{user}}, \lambda_{\text{item}}) = \log[a(\lambda_{\text{user}}, \lambda_{\text{item}})/a_0]$. The value $\alpha(\lambda_{\text{user}}, \lambda_{\text{item}}) = 0$ (dashed line) thus indicates no change with respect to the reference $a_0 = 0.468$, and $\alpha(\lambda_{\text{user}}, \lambda_{\text{item}}) > 0$ [respectively, $\alpha(\lambda_{\text{user}}, \lambda_{\text{item}}) < 0$] indicates predictions that are more (less) accurate than those obtained by ignoring node attributes. The maximum possible relative performance ($a_{\max} = 0.580$; dotted line) is obtained when each rating is assigned the exact probability that was used to generate it. For each value of the correlation [(a,b) full correlation, $c = 1$; (c,d) $c = 0.75$; (e,f) $c = 0.50$; (g,h) no correlation, $c = 0$], we show the variation of $\alpha(\lambda_{\text{user}}, \lambda_{\text{item}})$ with λ_{item} for different values of λ_{user} (left), and the whole dependence of $\alpha(\lambda_{\text{user}}, \lambda_{\text{item}})$ on both λ_{user} and λ_{item} (right).

We infer the values of the model parameters using the expectation-maximization equations, and we use the inferred parameters to predict unobserved ratings in the bipartite ratings network. We do this for different levels of correlation c between the ratings and the attribute networks (Fig. 2), from a situation $c = 1$ in which the attributes are perfectly correlated with user and item membership vectors (all male users belong to one group and have identical parameters, and all females belong to another group with different parameters; items of the same genre belong to the exact same mixture of groups) to a situation $c = 0$ in which user and item memberships and attributes are completely uncorrelated (Fig. 2).

Since we focus on sparse observations in which the number of observed ratings is low (only 1% of all ratings), model parameters cannot be inferred accurately from the ratings alone. Therefore, when we only consider the observed ratings R^O and ignore all attributes A^O by setting $\lambda_{\mathrm{user}} = \lambda_{\mathrm{item}} = 0$ in Eq. (9) (λ_{user} and λ_{item} correspond to the user and item attribute networks, respectively), the prediction of unobserved links is suboptimal; that is, the inferred probabilities of unobserved links differ significantly from the actual probabilities used to build the network. Therefore, in this regime, $\lambda_{\mathrm{user}} = \lambda_{\mathrm{item}} = 0$, the prediction accuracy (that is, the fraction of correctly predicted ratings) is always lower than the theoretical maximum accuracy (Fig. 2).

When there is perfect correlation between node attributes and group memberships, considering the attributes A^O by setting $\lambda_{\mathrm{user}} > 0$ and $\lambda_{\mathrm{item}} > 0$ should, in principle, help in the inference process. In fact, since attributes are perfectly correlated to group memberships, in the limit $\lambda_{\mathrm{user}} \to \infty$ and $\lambda_{\mathrm{item}} \to \infty$, nodes will be forced into the correct groups and predictions should be near optimal. This is what we observe in our numerical experiments [Figs. 2(a) and 2(b)]. Interestingly, as we increase the weight of the attributes in the log-posterior from $\lambda_{\mathrm{user}} = \lambda_{\mathrm{item}} = 0$, the effect on prediction accuracy is not smooth. Rather, below certain threshold values of λ_{user} and λ_{item}, using the attributes does not have any significant effect on prediction accuracy. Then, at those threshold values, a crossover occurs and prediction accuracy increases after a certain point until it reaches its theoretical maximum, as expected.

When attributes and ratings are completely uncorrelated [Figs. 2(g) and 2(h)], the role of attributes is reversed. Predictions are equally suboptimal at $\lambda_{\mathrm{user}} = \lambda_{\mathrm{item}} = 0$, but then, as λ_{user} and λ_{item} cross certain threshold values, predictions suddenly worsen as user and item nodes are forced into groups that are uncorrelated with their real membership vectors and, thus, with the observed ratings. In this situation, clearly, treating metadata as ground truth would be particularly misleading [23]; our approach enables us to show when that would be the case.

Unlike the extreme cases of total correlation or zero correlation, when attributes are partly correlated with the true group memberships of the nodes, the change in performance is not monotonic as we increase the importance of the attributes. As before, when λ_{user} and λ_{item} are small enough, we observe no difference with the situation in which the attributes are ignored entirely. In the other limit, when $\lambda_{\mathrm{user}} \to \infty$ and $\lambda_{\mathrm{item}} \to \infty$, user and item nodes are forced into groups that partly, but not perfectly, match the true group memberships of the nodes, so the performance may increase or decrease with respect to the situation with no attributes, depending on whether the correlation is high [Figs. 2(c) and 2(d)] or low [Figs. 2(e) and 2(f)]. However, we find that the most predictive models in this case are those at intermediate values of λ_{user} and λ_{item}, precisely at the crossover region where both the observed ratings and the observed attributes play a role in determining the most plausible group memberships.

VI. THEORETICAL INTERPRETATION OF THE CROSSOVER

To better understand this crossover, we look at the posterior of the two models corresponding to the maximum *a posteriori* estimates for $\lambda_{\mathrm{user}} = \lambda_{\mathrm{item}} = 0$ and for $\lambda_{\mathrm{user}} = \lambda_{\mathrm{item}} \to \infty$ (Fig. 3). These are the most plausible models when only data (ratings) and only metadata (attributes), respectively, are taken into consideration.

If we draw upon the analogy between Bayesian statistics and statistical mechanics [35,36], we can equate the log-posterior to the (minus) free energy of a physical system and interpret the crossover in terms of a transition in which λ plays the role of the tuning (temperature-like) parameter.

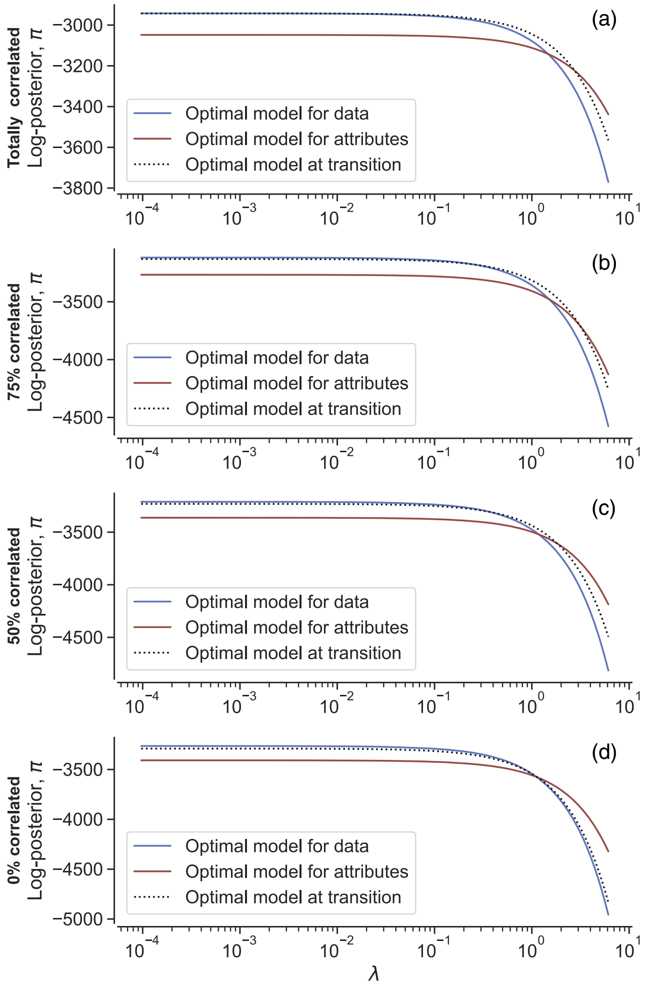

FIG. 3. Crossover between data-dominated and metadata-dominated inference regimes. For the synthetic data in Fig. 2, we plot the log-posterior $\pi(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{p}, \boldsymbol{q}, \hat{\boldsymbol{q}} | R^O, A^O)$ as a function of the hyperparameter $\lambda = \lambda_{\mathrm{item}} = \lambda_{\mathrm{user}}$ for three models: the model that maximizes the data likelihood L^R, the model that maximizes the metadata likelihood L^A, and the model that maximizes the posterior when two previous cases cross (that is, have equal posteriors). The position of the crossing coincides with the crossover and the maxima observed in Fig. 2.

Within this framework, these extreme models are the dominating maxima in the posterior landscape (or, by analogy, the states of the system at the two sides of the transition); therefore, the predictability crossover occurs when the data-dominated and metadata-dominated log-posteriors cross, that is, for a value λ^* such that

$$\mathcal{L}_0^R + \lambda^* \mathcal{L}_0^A = \mathcal{L}_\infty^R + \lambda^* \mathcal{L}_\infty^A \tag{10}$$

or

$$\lambda^* = \frac{\mathcal{L}_0^R - \mathcal{L}_\infty^R}{\mathcal{L}_\infty^A - \mathcal{L}_0^A}. \tag{11}$$

Here, the subindex 0 (or ∞) indicates the quantities corresponding to the model that maximizes the posterior for $\lambda_{\rm user} = \lambda_{\rm item} = 0$ (respectively, $\lambda_{\rm user} = \lambda_{\rm item} \to \infty$), and we group all attributes in a single term $\mathcal{L}^A$. As we show in Fig. 3, we find that, indeed, the crossover in predictability in Fig. 2 coincides (at least in order of magnitude) with the point where the data-dominated and metadata-dominated log-posteriors cross. This crossover is reminiscent of the detectability transition in graph clustering, which is associated to a similar crossover phenomenon between the eigenvalues of certain matrices [28,29].

Importantly, all log-likelihoods are extensive quantities. Therefore, the dependency on the number of observed ratings N_R and attributes N_A can be made explicit by defining intensive (that is, *per-link*) log-likelihoods $\ell^R = \mathcal{L}^R / N_R$ and $\ell^A = \mathcal{L}^A / N_A$. Then,

$$\lambda^* \sim \frac{N_R}{N_A}, \tag{12}$$

and at the crossover λ^*, we have that both $\mathcal{L}^R \sim N_R$ and $\lambda^* \mathcal{L}^A \sim N_R$ are of the same order. By considering Eq. (9), we see that this must be the case. Indeed, for each attribute network, we find three regimes—one dominated by the $\mathcal{L}^R$ term, one dominated by the $\mathcal{L}^A$ term, and one in which both terms are comparable. Unless there is perfect or almost perfect correlation between attributes and node memberships, any improvement in predictive power must come from considering both the observed ratings and the observed attributes, and therefore in the crossover region that we have identified.

VII. REAL DATA

Finally, we analyze two empirical data sets and study whether we observe the same behaviors as in the synthetic data. First, we consider the 100-K MovieLens data set [37], which contains 100 000 ratings of movies by users. Age and gender attributes are available for users, which we model as excluding attributes (Fig. 4). Movies have genre attributes, which we model as nonexcluding attributes. The relative weights of user and movie attributes are given by the parameters $\lambda_{\rm users}$ and $\lambda_{\rm items}$.

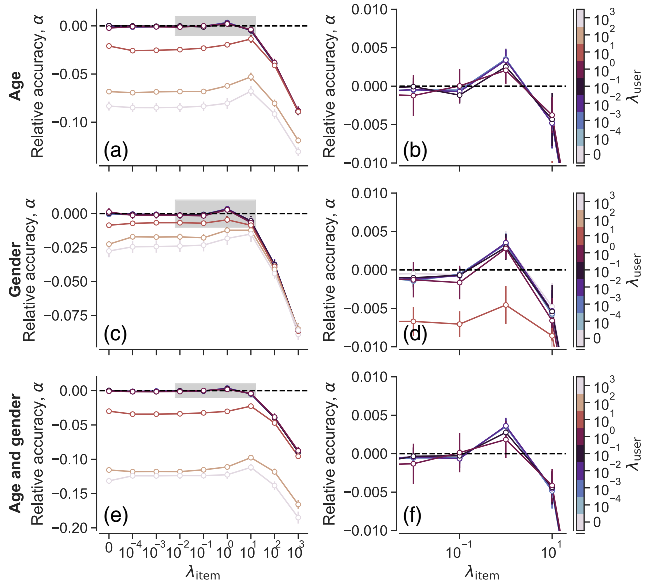

FIG. 4. Predictive performance and effect of metadata on the MovieLens data set. As in Fig. 2, we take as a reference the predictive accuracy a_0 of the algorithm when all attributes are ignored ($\lambda_{\rm user} = \lambda_{\rm item} = 0$), and we measure relative accuracy α for a given pair $(\lambda_{\rm user}, \lambda_{\rm item})$ as the log-ratio $\alpha(\lambda_{\rm user}, \lambda_{\rm item}) = \log[a(\lambda_{\rm user}, \lambda_{\rm item})/a_0]$. Accuracy is the fraction of correctly predicted ratings in cross-validation experiments, and $a_0 = 0.448$. We consider three different attributes for user nodes: (a,b) age; (c,d) gender; and (e,f) age and gender combined as a single attribute. We plot the whole range of $\lambda_{\rm user}$ (left) and zoom into the intermediate (shaded) region of $\lambda_{\rm user}$ in which predictions are significantly more accurate than the reference (right).

Just as in the synthetic networks with small but finite correlation, we observe an intermediate value of $\lambda_{\rm user}$ and $\lambda_{\rm item}$ that provides more accurate rating predictions than either considering the observed ratings alone or considering the node attributes alone. This behavior is similar when we consider age only, gender only, or age and gender simultaneously. As in synthetic networks, the optimal combination of rating data and node metadata occurs for values of λ such that the ratings network and the attributes networks have comparable contributions to the log-posterior.

Second, we consider a data set on the votes of 441 members of the U.S. House of Representatives in the 108th U.S. Congress [38] (Fig. 5). Between January 2003 and January 2005, these representatives voted on 1217 bills, casting one of 9 different types of vote, which, following previous analyses, we simplify to Yes, No, and Other [38]. In this data set, "users" are the representatives and "items" are the bills. The ratings represent the votes of the representatives on the bills. For representatives, we have attribute data indicating their party and state, which we model as excluding attributes. Although all votes of all members are recorded in the data set (in total,

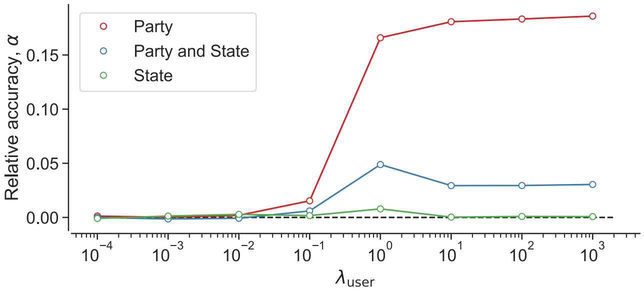

FIG. 5. Predictive performance and effect of metadata on the U.S. Congress data set. As in Fig. 2, we take as a reference the predictive accuracy a_0 of the algorithm when all attributes are ignored ($\lambda_{\text{user}} = 0$) and measure relative accuracy α for a given λ_{user} as the log-ratio $\alpha(\lambda_{\text{user}}) = \log [a(\lambda_{\text{user}})/a_0]$. Accuracy is the fraction of correctly predicted ratings in cross-validation experiments, and $a_0 = 0.677$. We consider three different attributes for user nodes: party, state, and party and state simultaneously.

536 698 votes), for the purpose of our analysis, we infer the parameters of the multipartite mixed-membership stochastic block model using 1% of the data and predict the remaining 99% (and repeat this using each 1% of the data as the training set).

Again, the effects of introducing the attributes in the inference process are very similar to those we encounter in synthetic data (Fig. 5). When using only the state of the representatives, we observe a behavior that is compatible with small but finite correlation between attribute and voting patterns since the optimal predictive performance is observed at intermediate values of λ_{user}. Rather, when we consider party affiliation, we observe a behavior that is compatible with almost perfect correlation between attribute and voting behavior. Indeed, in this case, the predictive performance of the model increases monotonically with λ_{user}, with a crossover at $\lambda_{\text{user}} \approx 1$, just as for perfectly correlated attributes in synthetic data. When state and party are combined into a single excluding attribute (for example, "Democrat from Texas" is a group), we observe a behavior compatible with strong (but imperfect) correlation between attributes and voting behavior. In this case, predictive accuracy does not improve monotonically with λ_{user} because, for very large values, representatives are forced into small groups that are more prone to fluctuations; that is, the model overfits the data, thus worsening the predictive power with respect to considering large groups associated to party affiliation alone.

VIII. CONCLUSION

There is ample evidence that using node metadata can help solve network inference problems. As we have discussed, several approaches have been proposed in recent years to introduce node attributes into probabilistic network models and to use them to make better inferences about, for example, the group structure of networks or the existence of unobserved interactions. In these approaches, node attributes are introduced either as part of a whole-system model (including both the links between nodes and node attributes) or as priors over the parameters of the model for the links (for example, as priors for the node group memberships that, in turn, determine the probability of the existence of links). However, beyond the improvement in performance that they may entail in a given task such as group detection or link prediction, we know little about the effect that node attributes have in the inference process. Here, our goal has been to clarify this issue.

Regardless of whether attributes are introduced as part of a whole model or as a prior for model parameters, they appear in probabilistic models as additional terms in the likelihood or the posterior. As we have shown, our results depend on this simple observation alone—only when all terms in these likelihoods or posteriors are comparable in magnitude, or when attributes are perfectly correlated with ratings, can we expect attributes to improve the inference process. In this sense, our findings here may be expected to be universal.

Our results are also general in that they should apply to all model-based, probabilistic approaches to network inference problems and not only link prediction [39]. Indeed, although, in general, the task of community detection is different from link prediction, our probabilistic approach to rating prediction involves obtaining the most plausible assignment of nodes to (mixtures of) groups—we first obtain the maximum *a posteriori* (MAP) partition of the nodes into groups and then use that partition to make link predictions [40]. Therefore, all our claims with regards to the crossover apply equally to the MAP estimation of node partitions with our model. By extension, our results also apply to all probabilistic approaches to community detection in which data and metadata enter in the likelihood or posterior as separate, competing terms.

From a practical point of view, our work helps us to understand when certain approaches will not work. For example, our results suggest that modeling data and metadata jointly will only improve link predictions (or other network inference problems) if two conditions are fulfilled simultaneously: (i) The metadata are correlated to the data; and (ii) the balance between the amount of data and metadata is such that their likelihoods (L^R and L^A above) are of the same order. If the first condition is not fulfilled, using metadata will, in general, worsen predictions rather than improving them; if the second condition is not fulfilled, one may, in practice, inadvertently ignore either the data or the metadata and thus make, again, suboptimal predictions.

Some works have intuitively addressed this problem by introducing tuning parameters akin to our λ_k [18,25]. However, the impact of those parameters has not been studied in detail; instead, their values are typically chosen among a very limited set by means of cross-validation.

Our work clarifies how the value of those parameters should be chosen and why.

Our work may also suggest new questions to investigate. For example, we have shown that the phenomenology we observe is driven by a crossover of posteriors, similar to the crossover of eigenvalues driving the detectability transition in community detection. There, connectivity correlations can affect the transition in counterintuitive ways [28,29]. So, perhaps in the case of metadata, there are similar effects, and the distribution of attributes among nodes (correlated or anticorrelated with the amount of data available for each node) shifts the optimal metadata weight or makes metadata more or less useful, in general. These are important practical questions.

From a broader perspective, our work opens the door to understanding the role of different terms in probabilistic network models, as well as the crossovers that occur between the regimes in which one term or another dominates. This sets the stage for more systematic approaches to building better probabilistic models of network systems.

ACKNOWLEDGMENTS

This research was funded by Projects No. FIS2016-78904-C3-P-1 and No. PID2019–106811GB-C31 from MCIN/AEI/10.13039/501100011033 and by the Government of Catalonia (2017SGR-896).

APPENDIX A: EXPECTATION-MAXIMIZATION EQUATIONS

We aim to maximize the parametric log-posterior in Eq. (9) as a function of the model parameters $\boldsymbol{\theta}$, $\boldsymbol{\eta}$, $\boldsymbol{p}$, $\boldsymbol{\zeta}$, $\boldsymbol{q}$, and $\hat{\boldsymbol{q}}$. Because logarithms of sums are hard to deal with, we use a variational trick that first introduces an auxiliary distribution $p(x)$ with $\sum_x p(x) = 1$ into a sum of terms as $\sum_x x = \sum_x p(x)(x/p(x))$. Then, because $\sum_x p(x)(x/p(x)) = \langle x/p(x)\rangle$, we can use Jensens' inequality $\log\langle y\rangle \geq \langle \log y\rangle$ to write $\log\left[\sum_x p(x)(x/p(x))\right] \geq \sum_x p(x)\log\left[x/p(x)\right]$.

Because both rating and attribute terms in Eq. (9) contain logarithms of sums, we introduce an auxiliary distribution for each of the terms as follows. For the ratings, we have

$$\begin{aligned}\mathcal{L}^R &= \sum_{(i,j)\in R^O}\log\sum_{\alpha\beta}\theta_{i\alpha}\eta_{j\beta}p_{\alpha\beta}(r_{ij}^O)\\ &= \sum_{(i,j)\in R^O}\log\sum_{\alpha\beta}\omega_{ij}(\alpha,\beta)\frac{\theta_{i\alpha}\eta_{j\beta}p_{\alpha\beta}(r_{ij}^O)}{\omega_{ij}(\alpha,\beta)}\\ &\geq \sum_{(i,j)\in R^O}\sum_{\alpha\beta}\omega_{ij}(\alpha,\beta)\log\frac{\theta_{i\alpha}\eta_{j\beta}p_{\alpha\beta}(r_{ij}^O)}{\omega_{ij}(\alpha,\beta)},\end{aligned} \tag{A1}$$

where $\omega_{ij}(\alpha,\beta)$ is the auxiliary distribution.

For the term corresponding to excluding node attributes, we have

$$\begin{aligned}\mathcal{L}^{A_k} &= \sum_{(i,\ell_k)\in A_k^O}\log\sum_{\alpha}\theta_{i\alpha}q_\alpha^k(i\ell_k)\\ &= \sum_{(i,\ell_k)\in A_k^O}\log\sum_{\alpha}\sigma_{i\ell_k}^k(\alpha)\frac{\theta_{i\alpha}q_\alpha^k(i\ell_k)}{\sigma_{i\ell_k}(\alpha)}\\ &\geq \sum_{(i,\ell_k)\in A_k^O}\sum_{\alpha}\sigma_{i\ell_k}^k(\alpha)\log\frac{\theta_{i\alpha}q_\alpha^k(i\ell_k)}{\sigma_{i\ell_k}^k(\alpha)},\end{aligned} \tag{A2}$$

where $\sigma_{i\ell_k}^k(\alpha)$ is the auxiliary distribution, and to simplify the notation, we have defined $q_\alpha^k(i\ell_k) \equiv q_\alpha^k((e_k^O)_{i\ell_k})$.

Finally, for the term corresponding to nonexcluding node attributes, we have

$$\begin{aligned}\mathcal{L}^{A_k} &= \sum_{(i,g)\in A_k^O}\log\sum_{\alpha\gamma}\theta_{i\alpha}\zeta_{g\gamma}^k\hat{q}_{\alpha\gamma}(ig)\\ &= \sum_{(i,g)\in A_k^O}\log\sum_{\alpha\gamma}\hat{\sigma}_{ig}^k(\alpha,\gamma)\frac{\theta_{i\alpha}\zeta_{g\gamma}^k\hat{q}_{\alpha\gamma}(ig)}{\hat{\sigma}_{ig}^k(\alpha,\gamma)}\\ &\geq \sum_{(i,g)\in A_k^O}\sum_{\alpha\gamma}\hat{\sigma}_{ig}^k(\alpha,\gamma)\log\frac{\theta_{i\alpha}\zeta_{g\gamma}^k\hat{q}_{\alpha\gamma}(ig)}{\hat{\sigma}_{ig}^k(\alpha,\gamma)},\end{aligned} \tag{A3}$$

where $\hat{\sigma}_{ig}^k(\alpha,\gamma)$ is the auxiliary distribution, and to simplify the notation, we have defined $\hat{q}_\alpha^k(ig) \equiv \hat{q}_{\alpha\gamma}^k((a_k^O)_{ig})$.

Note that, in Eqs. (A1)–(A3) above, the equality is satisfied when maximizing with respect to the auxiliary distributions. By solving these optimization problems, we obtain

$$\omega_{ij}(\alpha,\beta) = \frac{\theta_{i\alpha}\eta_{j\beta}p_{\alpha\beta}(r_{ij}^O)}{\sum_{\alpha'\beta'}\theta_{i\alpha'}\eta_{j\beta'}p_{\alpha'\beta'}(r_{ij}^O)}, \tag{A4}$$

$$\sigma_{i\ell_k}^k(\alpha) = \frac{\theta_{i\alpha}q_\alpha^k(i\ell_k)}{\sum_{\alpha'}\theta_{i\alpha'}q_{\alpha'}^k(i\ell_k)}, \tag{A5}$$

$$\hat{\sigma}_{ig}^k(\alpha,\gamma) = \frac{\theta_{i\alpha}\zeta_{g\gamma}^k\hat{q}_{\alpha\gamma}(ig)}{\sum_{\alpha'\gamma'}\theta_{i\alpha'}\zeta_{g\gamma'}\hat{q}_{\alpha'\gamma'}(ig)}. \tag{A6}$$

Therefore, the auxiliary distributions have the following interpretations: $\omega_{ij}(\alpha,\beta)$ is the contribution of user group α and item group β to the probability that user i gives item j a rating r_{ij}^O; $\sigma_{i\ell_k}^k(\alpha)$ is the contribution of user group (or item group) α to the probability that user (item) i has attribute type $(e_k^O)_{i\ell_k}$ in the kth excluding attribute; and, finally, $\hat{\sigma}_{ig}^k(\alpha,\gamma)$ is the contribution of groups α and γ to the probability that, for the kth nonexcluding attribute, the association between node i and attribute g is of type $(a_k^O)_{ig}$.

Using Lagrange multipliers for the normalization constraints and equating to zero the derivatives of the log-posterior with respect to the model parameters yields

$$\theta_{i\alpha} = \frac{\sum_{j\in\partial i}\sum_{\beta}\omega_{ij}(\alpha,\beta) + \sum_k \lambda_k \sigma^k_{i\ell_k}(\alpha) + \sum_l \lambda_l \sum_{g\in\partial_i{}^k}\sum_\gamma \hat{\sigma}^l_{ig}(\alpha,\gamma)}{d_i + \sum_k \lambda_k \delta^k_i + \sum_l \lambda_l \Delta^l_i}, \tag{A7}$$

where ∂^k_i is the set of kth attributes associated with user i, d_i is the degree of user i in the network of ratings, and $\Delta^l_i = |\partial_i{}^l|$. Note that the term $\sigma^k_{i\ell_k}(\alpha)$ is equal to zero if user i does not have attribute ℓ_k, so $\delta^k_i = 1$ if user i has exclusive attribute ℓ_k and zero otherwise:

$$\eta_{j\beta} = \frac{\sum_{i\in\partial j}\sum_{\alpha}\omega_{ij}(\alpha,\beta) + \sum_k \lambda_k \sigma^k_{j\ell_k}(\beta) + \sum_l \lambda_l \sum_{i\in\partial^k_j}\sum_\gamma \hat{\sigma}^l_{ij}(\beta,\gamma)}{d_j + \sum_k \lambda_k \delta^k_j + \sum_l \lambda_l \Delta^l_j} \tag{A8}$$

where ∂^k_j is the set of kth attributes associated with item j, d_j is the degree of item j in the network of ratings, and $\Delta^l_j = |\partial_j{}^l|$. As before, the term $\sigma^k_{j\ell_k}(\beta)$ is equal to zero if item j does not have attribute ℓ_k, so $\delta^k_j = 1$ if item j has exclusive attribute ℓ_k and zero otherwise,

$$\zeta^k_{g\gamma} = \frac{\sum_{i\in\partial^k_g}\sum_\alpha \hat{\sigma}^k_{ig}(\alpha,\gamma)}{\Delta^k_g}, \tag{A9}$$

where ∂^k_g is the set of nodes associated with attribute g, and $\Delta^k_g = |\partial_g{}^k|$. Additionally, we have

$$p_{\alpha\beta}(r) = \frac{\sum_{(i,j)\in R^O|r^0_{ij}=r}\omega_{ij}(\alpha,\beta)}{\sum_{(i,j)\in R^O}\omega_{ij}(\alpha,\beta)}, \tag{A10}$$

$$q^k_\alpha(e) = \frac{\sum_{(i,\ell_k)\in A^O_k|(e^O_k)_{i\ell_k}=e}\sigma^k_{i\ell_k}(\alpha)}{\sum_{(i,\ell_k)}\sigma^k_{i\ell_k}(\alpha)}, \tag{A11}$$

$$\hat{q}^k_{\alpha\gamma}(a) = \frac{\sum_{(i,g)\in A^O_k|(a^O_k)_{ig}=a}\hat{\sigma}^k_{ig}(\alpha,\gamma)}{\sum_{(i,g)\in A^O_k}\hat{\sigma}^k_{ig}(\alpha,\gamma)}. \tag{A12}$$

APPENDIX B: EXPECTATION-MAXIMIZATION ALGORITHM

To obtain a maximum of the posterior, we start by generating random initial conditions for each model parameter $\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{p}, \boldsymbol{\zeta}, \boldsymbol{q}, \hat{\boldsymbol{q}}$.

Then, we iteratively perform the following two steps until model parameters converge:

(1) Expectation step: Compute the auxiliary functions $\omega_{ij}(\alpha,\beta)$, $\sigma^k_{i\ell_k}(\alpha)$, and $\hat{\sigma}^k_{ig}(\alpha,\gamma)$ using current values for $\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{p}, \boldsymbol{\zeta}, \boldsymbol{q}, \hat{\boldsymbol{q}}$ using Eqs. (A.4)–(A.6).

(2) Maximization step: Compute the new values for the model parameters using the values for the auxiliary functions and Eqs. (A.7)–(A.12).

Because the posterior landscape is very rugged, to make predictions, we run the EM algorithm 10 times and consider all of the models to estimate the probability that user i rates item j with rating r (see Ref. [12]) as follows:

$$p(r_{ij} = r|R^O, A^O_k) \approx \frac{1}{N}\sum_{n=1}^{N} p_n(r_{ij} = r|R^O, A^O_k, (\ldots)), \tag{B1}$$

where $(\ldots) = \{\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{p}, \boldsymbol{\zeta}, \boldsymbol{q}, \hat{\boldsymbol{q}}\}$, and $p_n(r_{ij} = r|R^O, A^O_k, (\ldots))$ is the probability that user i rates item j with rating r in run n of the EM algorithm.

APPENDIX C: PARAMETERS FOR GENERATING SYNTHETIC DATA

For the generation of synthetic data, we use the following group-to-group connection probability matrices:

$$\boldsymbol{p}(r=0) = \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.8 \\ 0.8 & 0.1 & 0.1 & 0.1 \end{pmatrix},$$

$$\boldsymbol{p}(r=1) = \begin{pmatrix} 0.1 & 0.1 & 0.8 & 0.1 \\ 0.1 & 0.8 & 0.1 & 0.1 \end{pmatrix},$$

$$\boldsymbol{p}(r=2) = \begin{pmatrix} 0.8 & 0.8 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.8 & 0.8 \end{pmatrix}.$$

[1] D. Liben-Nowell and J. Kleinberg, *The Link-Prediction Problem for Social Networks*, J. Am. Soc. Inf. Sci. Technol. **58**, 1019 (2007).

[2] A. Clauset, C. Moore, and M. E. J. Newman, *Hierarchical Structure and the Prediction of Missing Links in Networks.*, Nature (London) **453**, 98 (2008).

[3] R. Guimerà and M. Sales-Pardo, *Missing and Spurious Interactions and the Reconstruction of Complex Networks.*, Proc. Natl. Acad. Sci. U.S.A. **106**, 22073 (2009).

[4] L. Lü, L. Pan, T. Zhou, Y.-C. Zhang, and H. E. Stanley, *Toward Link Predictability of Complex Networks*, Proc. Natl. Acad. Sci. U.S.A. **112**, 2325 (2015).

[5] A. Ghasemian, H. Hosseinmardi, A. Galstyan, E. M. Airoldi, and A. Clauset, *Stacking Models for Nearly Optimal Link Prediction in Complex Networks*, Proc. Natl. Acad. Sci. U.S.A. **117**, 23393 (2020).

[6] R Guimerà, *One Model to Rule Them All in Network Science?*, Proc. Natl. Acad. Sci. U.S.A. **117**, 25195 (2020).

[7] R. Guimerà and M. Sales-Pardo, *A Network Inference Method for Large-Scale Unsupervised Identification of Novel Drug-Drug Interactions*, PLoS Comput. Biol. **9**, e1003374 (2013).

[8] M. Tarrés-Deulofeu, A. Godoy-Lorite, R. Guimerà, and M. Sales-Pardo, *Tensorial and Bipartite Block Models for Link Prediction in Layered Networks and Temporal Networks*, Phys. Rev. E **99**, 032307 (2019).

[9] M. P. Menden *et al.*, *Community Assessment to Advance Computational Prediction of Cancer Drug Combinations in a Pharmacogenomic Screen*, Nat. Commun. **10**, 2674 (2019).

[10] R. Guimerà and M. Sales-Pardo, *Justice Blocks and Predictability of U.S. Supreme Court Votes*, PLoS One **6**, e27188 (2011).

[11] R. Guimerà, A. Llorente, E. Moro, and M. Sales-Pardo, *Predicting Human Preferences Using the Block Structure of Complex Social Networks*, PLoS One **7**, e44620 (2012).

[12] A. Godoy-Lorite, R. Guimerà, C. Moore, and M. Sales-Pardo, *Accurate and Scalable Social Recommendation Using Mixed-Membership Stochastic Block Models*, Proc. Natl. Acad. Sci. U.S.A. **113**, 14207 (2016).

[13] S. Cobo-López, A. Godoy-Lorite, J. Duch, M. Sales-Pardo, and R. Guimerà, *Optimal Prediction of Decisions and Model Selection in Social Dilemmas Using Block Models*, Eur. Phys. J. Data Sci. **7**, 48 (2018).

[14] G. Poux-Médard, S. Cobo-López, J. Duch, R. Guimerà, and M. Sales-Pardo, *Complex Decision-Making Strategies in a Stock Market Experiment Explained as the Combination of Few Simple Strategies*, Eur. Phys. J. Data Sci **10**, 26 (2021).

[15] M. Timme, *Revealing Network Connectivity from Response Dynamics*, Phys. Rev. Lett. **98**, 224101 (2007).

[16] T. P. Peixoto, *Network Reconstruction and Community Detection from Dynamics*, Phys. Rev. Lett. **123**, 128301 (2019).

[17] C. Tallberg, *A Bayesian Approach to Modeling Stochastic Blockstructures with Covariates*, J. Math. Sociol. **29**, 1 (2004).

[18] J. Yang, J. McAuley, and J. Leskovec, *Community Detection in Networks with Node Attributes*, in *2013 IEEE 13th International Conference on Data Mining* (2013), pp. 1151–1156, 10.1109/ICDM.2013.167.

[19] Y. Zhu, X. Yan, L. Getoor, and C. Moore, *Scalable Text and Link Analysis with Mixed-Topic Link Models*, in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13 (Association for Computing Machinery, New York, 2013), p. 473–481, 10.1145/2487575.2487693.

[20] D. Hric, T. P. Peixoto, and S. Fortunato, *Network Structure, Metadata, and the Prediction of Missing Nodes and Annotations*, Phys. Rev. X **6**, 031038 (2016).

[21] M. E. J. Newman and A. Clauset, *Structure and Inference in Annotated Networks*, Nat. Commun. **7**, 11863 (2016).

[22] A. White and T. B. Murphy, *Mixed-Membership of Experts Stochastic Blockmodel*, Netw. Sci. **4**, 48 (2016).

[23] L. Peel, D. B. Larremore, and A. Clauset, *The Ground Truth About Metadata and Community Detection in Networks*, Sci. Adv. **3**, e1602548 (2017), 10.1126/sciadv.1602548.

[24] N. Stanley, T. Bonacci, and R. Kwitt, *Stochastic Block Models with Multiple Continuous Attributes*, Appl. Netw. Sci. **4**, 54 (2019).

[25] M. Contisciani, E. A. Power, and C. De Bacco, *Community Detection with Node Attributes in Multilayer Networks*, Sci. Rep. **10**, 15736 (2020).

[26] Y. Koren, R. Bell, and C. Volinsky, *Matrix Factorization Techniques for Recommender Systems*, Computer **42**, 30 (2009).

[27] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, *Inference and Phase Transitions in the Detection of Modules in Sparse Networks*, Phys. Rev. Lett. **107**, 065701 (2011).

[28] F. Radicchi, *Detectability of Communities in Heterogeneous Networks*, Phys. Rev. E **88**, 010801(R) (2013).

[29] F. Radicchi, *A Paradox in Community Detection*, Europhys. Lett. **106**, 38001 (2014).

[30] P. Zhang, C. Moore, and L. Zdeborová, *Phase Transitions in Semisupervised Clustering of Sparse Networks*, Phys. Rev. E **90**, 052802 (2014).

[31] P. W. Holland, K. B. Laskey, and S. Leinhardt, *Stochastic Blockmodels: First Steps*, Soc. Networks **5**, 109 (1983).

[32] K. Nowicki and T. A. B. Snijders, *Estimation and Prediction for Stochastic Blockstructures*, J. Am. Stat. Assoc. **96**, 1077 (2001).

[33] T.-C. Yen and D. B. Larremore, *Community Detection in Bipartite Networks with Stochastic Block Models*, Phys. Rev. E **102**, 032309 (2020).

[34] E. M. Airoldi, D. M. Blei, S. E Fienberg, and E. P. Xing, *Mixed Membership Stochastic Blockmodels*, J. Mach. Learn. Res. **9**, 1981 (2008).

[35] E. T. Jaynes, *Information Theory and Statistical Mechanics*, Phys. Rev. **106**, 620 (1957).

[36] E. T. Jaynes, *Information Theory and Statistical Mechanics. II*, Phys. Rev. **108**, 171 (1957).

[37] F. M. Harper and J. A. Konstan, *The Movielens Datasets: History and Context*, ACM Trans. Interact. Intell. Syst. **5**, 19 (2015).

[38] A. S. Waugh, L. Pei, J. Fowler, P. Mucha, and M. A. Porter, *Party Polarization in Congress: A Network Science Approach*, arXiv:0907.3509v3.

[39] The task of predicting ratings considered here is a generalization of the task of link prediction that allows for (i) unobserved links (that is, links whose existence is unknown) and (ii) more than two states for each link (that is, it goes beyond exists or does not exist). Indeed, the standard link-prediction problem can be mapped exactly to our rating-prediction problem by assign "rating 1" to all links and "rating 0" to all nonlinks in a network; with this mapping, all the discussion in the paper remains valid, and we are strictly solving the link prediction problem in a regular network.

[40] Strictly speaking, EM gives a local maximum of the posterior rather than the global MAP.