Quantifying metadata relevance to network block structure using description length

Lena Mangold, Camille Roth

Quantifying metadata - block structure relationships in networks using description length

Lena Mangold[1,2,3,*] and Camille Roth[1,2,3]

[1] *Centre national de la recherche scientifique (CNRS), 3 rue Michel-Ange, 75 016 Paris, France*
[2] *Computational Social Science team, Centre Marc Bloch, Friedrichstr. 191, 10117 Berlin, Germany,*
[3] *Centre d'Analyse et de Mathématique Sociales (CAMS), École des hautes études en sciences sociales (EHESS), 54 Bd Raspail, 75006 Paris, France*
[*] *lena.mangold@cmb.hu-berlin.de*

July 23, 2024

Abstract

Network analysis is often enriched by including an examination of node metadata. In the context of understanding the mesoscale of networks it is often assumed that node groups based on metadata and node groups based on connectivity patterns are intrinsically linked. This assumption is increasingly being challenged, whereby metadata might be entirely unrelated to structure or, similarly, multiple sets of metadata might be relevant to the structure of a network in different ways. We propose the *metablox* tool to quantify the relationship between a network's node metadata and its mesoscale structure, measuring the strength of the relationship and the type of structural arrangement exhibited by the metadata. Our tool incorporates a way to distinguish significantly relevant relationships and can be used as part of systematic meta analyses comparing large numbers of networks, which we demonstrate on a number of synthetic and empirical networks.

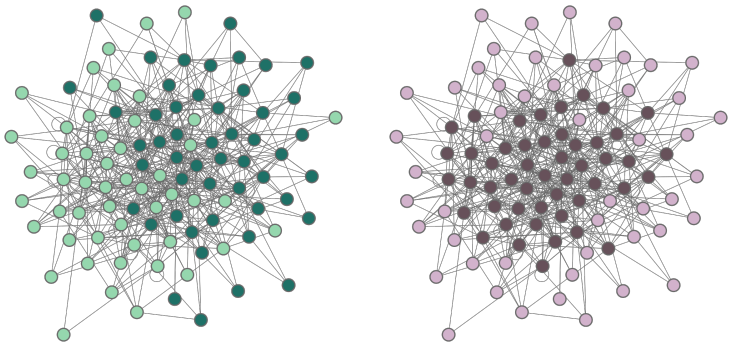
Introduction

Block structure in networks is characterised by the grouping of nodes on the basis of shared connectivity patterns [1, 2]. Such networks can be generated by Stochastic blockmodels (SBMs) [3] which – in turn – can be used as baseline models to infer block structure from observed networks. The latter is becoming increasingly popular due to a number of considerable advances in the development of SBM variants with closer resemblance of real network structure [4], the introduction of flexible, nonparametric inference approaches [5], and increasingly efficient inference algorithms [6, 7]. Blocks may take the shape of commonly studied mesoscale structures, such as assortative communities or internally cohesive clusters. However, other structural arrangements on the mesoscale, such as core-periphery structures,

disassortative (bipartite) structures as well as (nested) combinations of the above, are also possible, owing to the relatively general definition of similarity of the SBM. It is often assumed that blocks – whichever specific structural arrangement they may have – correspond to a latent 'meaning', i.e. some external similarity exhibited by the nodes that has made it more likely for them to connect to other nodes in the network in a similar way, or vice versa. In practice, this 'meaning' is often attributed to additional information on the network nodes, which we call *metadata*. In the literature, node attributes available as part of network analyses have sometimes been assumed to be intrinsically linked to the network's structure, an assumption that – as has been demonstrated on multiple occasions [8–10] – cannot be readily made.

Imagine a set of users who interact with each other on some social media platform and for whom – in an idealised scenario – some metadata is known to us: for each user we know their preference for one of two political parties. We can construct a network that represents the users' conversation around some political topic, by placing an edge between user nodes who interacted within the context of the topic. Assume we observe *homophily* in political leaning (called *assortativity* in network science): users with shared party preferences are more likely to endorse each other's content than that of users who support other parties, something that has been observed in the literature around online interactions repeatedly [11–13]. The static 'snapshot' of an interaction network between these users, which might look something like the toy network in Figure 1a, would show an alignment of a partition of users by party preference and one according to connectivity patterns. We could also say that the *generative process* of this network was, at least to some extent, likely governed by this particular set of metadata under an assortativity modeling assumption.

Such alignment between metadata partitions and block structure has been observed in social networks on many occasions [14–16] which may partly explain the widespread assumption of an intrinsic connection between metadata and node structure. Motivated also by a lack of knowledge on the 'real' generative processes of empirical networks, node metadata has often been viewed as *the* ground truth for the block structure of a network, to evaluate [4, 17, 18] or to improve [19–23] the performance of community detection or other inference algorithms. However, a number of recent works in the complex networks community has challenged the notion of an intrinsic alignment between metadata and community structure [8, 9, 24, 25]; metadata might be entirely unrelated to structure or, similarly, multiple sets of metadata might be relevant to the structure of a network in different ways [10]. We can illustrate this on our example set of social media users for which we imagine a second set of node metadata: whether or not a user is an expert on the political topic that is being discussed in this particular network. Assume that in terms of structural features, the network exhibits some assortativity (of party preference) but the network also has a relatively well connected core of users – in which there are connections even between users of different party preference – and a loosely connected set of peripheral nodes. This structure then corresponds to the node attribute of expert (core) vs non-expert (periphery), an example of this can be seen in 1b. Overall, we thus have two sets of metadata that are both linked to the network structure



(a) Bicommunity partition **(b)** Core-periphery partition

Figure 1: An example graph with nodes coloured according to their block memberships in the planted bicommunity partition (a) and the planted core-periphery partition (b), using the network drawing functionality from the graph-tool library [26] (which is used for all network visualisations in this paper).

while exhibiting different structural arrangements.

Besides emphasising the possible existence of multiple relevant metadata, this example can also be understood in the context of the diversity of likely node partitions exhibited by real networks: when focusing purely on the connectivity patterns in a network we can, in many cases, identify multiple, potentially qualitatively different partitions that divide the nodes in a 'plausible' way [27]. This, in turn, accentuates the notion that multiple sets of metadata can be related to the network structure, even if they divide the network's nodes in very different ways. It has been demonstrated, for example, that we can generate synthetic networks whose mesoscale – like the one in our toy example – is similarly well explained by (a) a division into two assortative communities and (b) a division into a well-connected core and a sparsely connected periphery [28]. By fitting an SBM to such a network and sampling from the posterior distribution of partitions, we are likely to find two locally optimal partitions – a bicommunity and a core-periphery partition – which may be aligned with different sets of metadata. In other words: not only might multiple sets of metadata be relevant to the network structure in general, but they might be relevant in structurally very different ways.

To add to the complication, we can extend our social media thought experiment by considering the 'reversed' situation: instead of one network representing a political discussion with multiple sets of metadata, we imagine multiple networks with node attributes of one particular type. Firstly, we imagine that we are able to construct *multiple* networks for the same set of user nodes, such as networks of different interaction types (e.g. endorsement vs exchange of opinions) or snapshots recorded at different points in time. Secondly, we might have collected data on different topics from a variety of categories, leading to the construction of multiple topic-induced networks with different (but potentially overlapping) user node sets but shared type of metadata. To summarise and in all generality, we can

divide our thought experiment into three scenarios worth thinking about: when studying social networks, we might be dealing with a situation with (I) a single network and multiple sets of metadata, (II) one set of metadata and multiple networks with the *same* user sets, and finally (III) one set of metadata and multiple networks in the same *context*, but with possibly different node sets.

Having outlined three types of scenarios, we come back to the questions we might be asking to investigate metadata-block structure relationships: Is there a way to measure the strength of the relationship between metadata and network block structure, to enable comparisons between sets of metadata for one given network (I) and multiple metadata-network pairs (II and III)? If a set of node attributes *is* relevant to the structure of a particular network, can we quantify the particular structural arrangement at hand (assortativity vs some other structure)? Answering these questions requires rigorous, statistically grounded methods and we argue that generative models lend themselves particularly well in this context: they provide mathematical justification, help us deal with uncertainties connected to model selection and facilitate comparisons of different models and networks.

To our knowledge, we lack a framework that makes it possible, for a single given network and in a comparative setting for a collection of networks, to appraise the relative strength of the connection between various sets of metadata and various types of intermediate-scale structures. Related work that has gone furthest in this direction, and that serves as a motivation for our measure, is that by Peel at al [10]. The authors demonstrated convincingly that multiple ground-truth partitions can be responsible for the generation of a given network. They proposed two separate approaches, one to measure the statistical significance of the metadata-structure connection and one to explore the relationship between specific sets of metadata and the partition landscape of a network. Their first measure (which we outline later on in this work) serves as a p-value and thus answers a simple yes-no question with respect to the significance of metadata-structure relevance; their second measure is an inferential approach based on SBMs, which – upon visual inspection of the results – provides insights into the extent to which different sets of metadata are related to different parts of the partition landscape of a network. While these measures enable an in-depth analysis of individual networks and their metadata, they cannot be easily used for direct comparative purposes, since the strength of metadata-structure relationships cannot be measured and visual analysis is required for comparative studies of multiple sets of metadata; a direct comparison between networks is also not possible. Another relevant method is a label propagation approach to calculate a p-value for the significance of metadata and structure [29] has similar problems and thus does not lend itself well for large-scale comparative meta analyses of multiple networks either.

In this work, we propose the *metadata block structure exploration (metablox)* tool, which utilises microcanonical SBMs and methods from information theory to quantify the connection between node metadata and the block structure of a network. Our measure exploits the feature of the minimum description length (MDL) principle [30], which penalises overly complex models, and enables comparison across multiple sets of network-metadata pairs with

respect to the *strength* of the relevance of the metadata to the block structure and regarding the prominent *type* of block structure exhibited by the metadata. We design the measure to enable such comparative analyses for the scenarios I-III discussed above.

General approach

To derive our measure, we first note that a set of categorical node attributes naturally divides the nodes of a network into a partition. We will refer to this division of the nodes as the *metadata partition* and denote it by the vector $\mathbf{d} = \{d_i\}$, where d_i denotes the metadata label of node i. The perhaps most obvious way to measure the relationship between a metadata partition and the block structure of a network would be to infer the optimal partition $\mathbf{b}_{\text{opt}}$ of the network in some way and then measure the alignment between the metadata partition and the optimal partition through a similarity measure $s = f(\mathbf{b}_{\text{opt}}, \mathbf{d})$, such as the Rand index [31], variation of information [32], or partition overlap [27]. However, this approach may fail as soon as the detection method yields multiple similarly likely partitions that are not well-aligned in terms of their node groups: if the metadata partition is similar to one of the highly likely partitions that is not *the* optimal one, we would not capture the metadata relevance.

Instead of directly measuring partition similarities, we take a Bayesian approach to the problem. Specifically, we note that in the Bayesian SBM framework, one generally considers the entire distribution over the possible partitions that could have been responsible for generating the network under the given SBM. With this in mind, we hypothesise that a metadata partition could – in theory – appear among the partitions in the posterior distribution of the SBM with some non-zero probability. Unfortunately, calculating the posterior distribution of the SBM exactly turns out to be an intractable problem and one therefore needs to resort to approximations, e.g. by sampling from it using an MCMC method. Since we cannot rely on finding an exact copy of our metadata partition among the samples taken from the posterior, we instead take the reverse approach: we determine where in the posterior distribution the metadata partition *would* fall, were it to have been inferred by fitting an SBM to our observed network. The schematic in Figure 2 shows an idealised partition landscape of a synthetic network (upon fitting an SBM) and illustrates the idea of the metadata being positioned more closely – in terms of description length (i.e. model fit) – to a non-optimal partition than to the optimal partition; below the landscape, we show side-by-side the naive approach described above and the approach we will be taking in this work.

Microcanonical SBMs and description length

To lay the groundwork for this approach, we first outline the microcanonical SBM framework and its connection to the concept of description length, which we will use to quantify the metadata relevance. We focus on the so-called degree-corrected SBM [4], which configures an ensemble of networks with block structure, while making it possible to take into account heterogeneous degree distributions (in contrast to the traditional SBM). For a network with

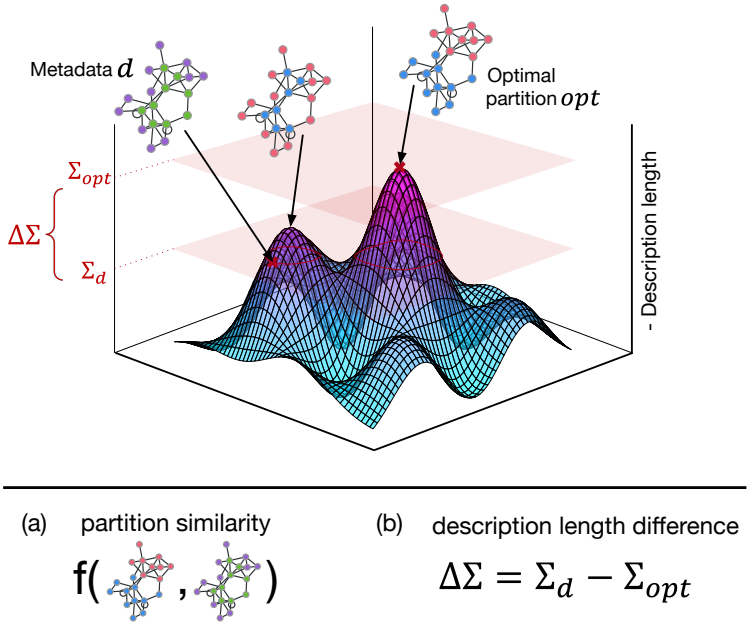


Figure 2: Schematic of a partition landscape[10, 27] for a toy network, for which we have highlighted the positions of the optimal inferred partition, a second partition, and a metadata partition d. Underneath the schematic, we include the naive approach for measuring metadata relevance (a) and the core idea of the metablox approach (b).

N nodes and B blocks, it is parameterised by an N-dimensional block membership vector $\mathbf{b} = \{b_i\}$, a degree sequence $\mathbf{k} = \{k_i\}$, and a $B \times B$ edge count matrix $\mathbf{e} = \{e_{rs}\}$. In its microcanonical form, the degree sequence and edge count matrix are specified *exactly*, making it possible to count the number of possible networks that can be generated with a set of parameters, so that the likelihood $P(\mathbf{A}|\mathbf{k},\mathbf{e},\mathbf{b})$, i.e. the probability of observing a network $\mathbf{A}$ given the SBM parameters, can be calculated as $\Omega(\mathbf{k},\mathbf{e},\mathbf{b})^{-1}$, the inverse of the ensemble cardinality[33]. When using a parametric framework to detect the optimal partition from a network, one would find the partition $\mathbf{b}$ that maximises the log-likelihood, which amounts to minimising the microcanonical entropy $S = \log\Omega(\mathbf{k},\mathbf{e},\mathbf{b})$.[33, 34].

In the *nonparametric* approach – which allows the model parameters to be inferred from the data –, one considers the joint distribution of the generative model, including the notion of priors on the model parameters in addition to the model likelihood. The hard constraints of the microcanonical approach imply that one does not need to sum over the remaining parameters to calculate the marginal likelihood. The joint distribution can therefore be written as $P(\mathbf{A},\mathbf{k},\mathbf{e},\mathbf{b}) = P(\mathbf{A}|\mathbf{k},\mathbf{e},\mathbf{b})P(\mathbf{k}|\mathbf{e},\mathbf{b})P(\mathbf{e}|\mathbf{b})P(\mathbf{b})$. This is where an information theoretical lens lets us write $P(\mathbf{A},\mathbf{k},\mathbf{e},\mathbf{b}) = e^{-\Sigma}$, where $\Sigma = -\ln P(\mathbf{A}|\mathbf{k},\mathbf{e},\mathbf{b}) - \ln P(\mathbf{k},\mathbf{e},\mathbf{b})$ is called the description length of the data: the amount of information necessary to describe

a network given a model plus the information required to describe the model itself, via its parameters. Maximising the posterior distribution is therefore equivalent to minimising the SBM description length, which – in turn – is the same as finding the partition which provides the most compact *compression* of the network and therefore amounts to being the best model. Similarly, one can identify the more likely of two partitions $\mathbf{b}_1$ and $\mathbf{b}_2$ (under SBM variants m_1 and m_2 respectively) by comparing the network's description length for each of them, which amounts to calculating their posterior odds ratio. For $\Sigma_1 < \Sigma_2$, for example, partition $\mathbf{b}_1$ (under variant m_1) is the more likely of the two (see Methods section for a detailed explanation). Note that we use description length as the basis for model selection (i.e. finding the best model by penalising overly complex modeling choices) since popular existing information criteria (such as AIC [35] and BIC [36]) or cross-validation are built on assumptions that do not hold for the SBM or are not suitable for relational data [37].

Measuring metadata relevance

To tie in these concepts with our metablox measure, we realise that one can simply count the nodes in each *metadata block* (i.e. those that share the same metadata category), and the edges within and between them, and arrive at quantities that are equivalent to the parameters of the microcanonical SBM. Given the metadata partition by $\mathbf{d}$, we denote the edge counts within and between metadata blocks by $\mathbf{e}'$. By plugging the quantities derived from the metadata and from the observed network into the SBM description length calculations – which, in the microcanonical case, can be done exactly by using combinatorics [5] – we can straightforwardly calculate Σ_d^m, the description length of the network under the metadata partition $\mathbf{d}$ and a given SBM variant m. In this work, we use the description length calculations that originate in Ref. [5], which we provide in the Methods section. Note that a more consistent notation would be to denote a metadata partition by $\mathbf{b}'$, but we choose $\mathbf{d} = \mathbf{b}'$, to make a clearer distinction from the inferred partitions.

Calculating the metadata description length in this way does not yet tell us anything about the fit of the metadata *relative to the partition landscape* of the network. We therefore also separately fit the SBM variant m to our observed network (ignoring the metadata labels), using the graph-tool library [26], and compute the description length Σ_{opt}^m of the optimal partition. We can then simply calculate the description length difference of the two quantities, $\Delta\Sigma = \Sigma_d^m - \Sigma_{opt}^m$ to understand how well the metadata partition fits the given network, compared to the optimal partition: if $\Delta\Sigma$ is close to zero we conclude that the metadata partition is strongly relevant to the block structure; if $\Delta\Sigma \gg 0$, the metadata partition fits the network much less well than the optimal partition, under the SBM variant m. Note that for better readability, we dropped the superscript m for $\Delta\Sigma = \Delta\Sigma^m$.

Statistical significance and comparability

There are two major caveats of quantifying the metadata relevance by simply using $\Delta\Sigma$: description length increases with growing networks size, which means that simply taking the absolute description length difference does not enable inter-network comparisons and we therefore need to normalise the measure in some way. Furthermore, we need to include a notion of statistical significance, to ensure that the fit of the metadata partition is, in fact, better than a partition we would find at random. We address both caveats at once by turning our measure into a ratio and including a distinction from randomness as follows: additionally to Σ_d^m we calculate the description lengths of the same network under an SBM with multiple sets of *randomised* metadata. Each time we randomise the nodes' metadata labels, we can plug in the resulting quantities $\mathbf{d}_*$ and $\mathbf{e}_*$ in the description length formulas again. This approach is inspired by the blockmodel entropy significance test (BESTest) [10], which produces a metadata p-value, i.e. the probability that a randomised version of the observed metadata describes the network better than (or equally well as) the observed metadata. The BESTest p-value is calculated by generating a large set of randomised metadata partitions (whereby the number of elements in each metadata category is fixed), computing the SBM entropy under each of the randomised partitions, and finally comparing the SBM entropy under the observed metadata partition to that of the randomised partitions. However, SBM entropy decreases for an increasing number of groups which greatly complicates direct comparison of metadata models (i.e. different metadata partitions and/or different SBM variants). Description length enables model selection and can thus be used to compare different partitions, whether induced by metadata categories or inferred based on connectivity patterns.

Instead of calculating a p-value from distributions of description lengths, we identify the *randomised metadata description length* Σ_*^m, which – for a significance level of $\alpha = 0.01$ – is equal to the first percentile of the description length distribution of the randomised metadata partitions under SBM variant m. The choice of α is, to some extent, arbitrary; here, we make the choice based on conventions in measuring statistical significance, according to which $\alpha = 0.01$ denotes strong evidence against the null hypothesis, i.e. that the network is described equally well by randomised metadata. We emphasise that other choices can be made when our measure is used, as the required level of significance may depend on the specific context. In our measure, we take Σ_*^m to be the maximum description length for which we would still consider the description length of the *observed* metadata to be relevant, and we use to normalise $\Delta\Sigma$ by $\Delta\Sigma_* = \Sigma_*^m - \Sigma_{opt}^m$.

Probing structural arrangement

Before we put these components together, we note that we can use the above approach to formulate a measure that not only quantifies metadata relevance but that also helps us probe for specific *types of structural arrangements* of the metadata blocks: Since we can follow the above procedure for any SBM variant for which there is a microcanonical formulation with a straightforward description length calculation, we can directly plug in the structure-specific SBM variant of interest. In this work, we focus on three SBM variants that are specific

to three different types of block structures, which we briefly outline in the following. In the 'standard' SBM (i.e. non-degree-corrected, NDC), the placement of an edge between two nodes depends solely on the block membership of each node and on the probability of two nodes from the two blocks being connected. Similar to the random graph model, one major drawback of this model is that the generative process produces networks that have blocks within which node degrees are Poisson distributed. To overcome this caveat, that makes the model unlike many real networks whose node degrees often have power-law degree distributions, the degree-corrected variant (DC) – which we introduced above to motivate out approach – was proposed [4]. In this variant, edge placement depends on node degrees as well as block membership, thus accounting for heterogeneous degree distributions. A third variant we include in our analysis is the assortative 'planted partition' SBM (PP), which – as part of the generative process – assumes assortativity [38]. The description length formulations for the three variants used in our measure are based on the work in Refs. [5, 33, 38], as detailed in the Methods section.

Note that we are focusing here on undirected simple graphs, so that any network for which our measure is used has to be treated as such. An extension of our measure to directed networks and/or networks with multi-edges is straightforward and should be considered as part of future research.

Measure

We finally put together these components to create the metadata block structure exploration (metablox) tool, which produces the vector $\boldsymbol{\gamma}_d$ for a metadata partition d, where each element represents the metadata relevance under a different SBM variant m. It can consist of as many elements as we include SBM variants to probe for different structural arrangements. Here, we will work with the two variants used in the motivating example, DC and PP as well as the non-degree-corrected SBM (NDC), thus $\boldsymbol{\gamma}_d$ is a vector with three elements. For a metadata partition d, the metablox vector $\boldsymbol{\gamma}_d$ thus consists of elements

$$\gamma_d^m = \frac{\Delta\Sigma^m}{\Delta\Sigma_*^m} = \frac{\Sigma_d^m - \Sigma_{opt}^m}{\Sigma_*^m - \Sigma_{opt}^m} \tag{1}$$

for each SBM variant m. In summary, Σ_d^m is the description length of metadata partition d under model m, Σ_*^m is the randomised metadata description length, and Σ_{opt}^m is the description length of the optimal partition of the network. Recall that the difference in description length between two models is equal to the log of the posterior odds ratio of the two models. The numerator of γ_d^m thus measures how much more likely the optimal partition is compared to the metadata partition under model m. Similarly, the denominator represents how much more likely the optimal partition is to the randomised metadata partitions under all models. The total measure is the former normalised by the latter.

With this definition, each element can then be interpreted as follows: for $\gamma_d^m \geq 1$ we say that this set of metadata is not relevant to the structure under model m, since more than

1% of the randomised metadata partitions compress the network more efficiently than our observed metadata, under the given model. For $\gamma_d^m < 1$, the metadata is relevant to the structure under the given model, and the closer γ_d^m is to 0, the stronger the relevance of the set of metadata to the block structure and the closer the structural arrangement of the metadata to the typical type of structure generated under the given model m. Another way of interpreting γ_d^m is: relative to the best compression we can find, how efficient is the compression of the observed metadata d under m compared to partitioning the network by randomised versions of the metadata.

Limitations and network compressibility

One notable limitation of our approach is that the inequality $\gamma_d^{m_1} < \gamma_d^{m_2}$ for a network with some metadata partition d merely indicates that, *given the observed metadata*, model m_1 provides a better explanation of the network than m_2. It does not necessarily mean that the block structure of our network as a whole is optimally explained by m_1: since our measure is a ratio of description length differences, it is possible that in fact $\Sigma_{\text{opt}}^{m_1} > \Sigma_{\text{opt}}^{m_2}$ or that the model m_2 provides a better explanation for the network overall. This limitation arises from the comparative nature of the measure, which inherently focuses on relative rather than absolute fit to enable comparisons across metadata and networks.

Another limitation is related to the case of inter-network comparison, in particular in the context of different network topology. While metablox is robust with respect to the number of nodes in a network (see Methods section), certain aspects of a network's topology do affect the measured metadata relevance. Let us assume that we are comparing two node attributed networks A and A', with the signal in the community structure of network A being significantly stronger than that of network A' (i.e. the communities of network A are more clearly separated). All other things being equal – including the level of alignment of the metadata partition with the optimal detected partition – we will have $\gamma < \gamma'$. Since Σ_{opt}^m decreases when the network becomes more compressible due to a stronger signal in the block structure, $\Delta\Sigma*$ increases (since Σ_*^m does not decrease with Σ_{opt}^m) and does so more quickly than $\Delta\Sigma$. This apparent conflation of block structure signal strength and metadata block-structure relevance points to an interesting question of the definition of this 'relevance', which in turn relates back to our earlier point of going beyond the direct measurement of partition alignment. We argue that the concept of measuring the relevance of metadata to block structure in fact *should* include the notion of block structure strength: we do not want to retrieve the same γ for the described networks A and A' merely because there is alignment of metadata partition and optimal partition. Instead, the relevance of the metadata should be quantified as stronger (i.e., a better γ) when the network exhibits more significant block structure.

The elements of the measure that lead to these limitations are part of the design that enables comparative work and what enables the measure to capture metadata-structure relationships in the intended way. However, there might be reasons for a researcher using this measure to require insights – alongside the metablox vector $\boldsymbol{\gamma}_d$ – into the absolute fit of the optimal

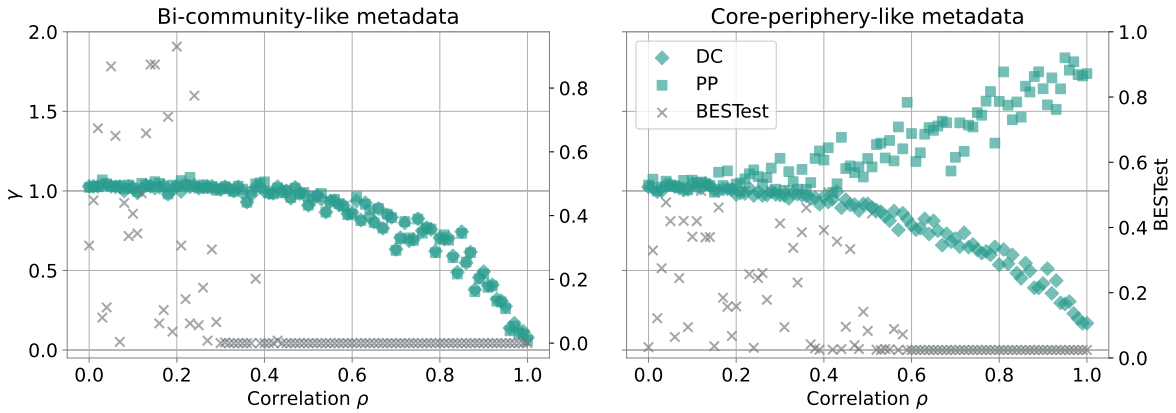


Figure 3: Metablox dimensions γ^{DC} and γ^{PP} for one network with multiple sets of of bi-community-like (left) and core-periphery-like metadata (right), showing increasing correlation ρ between metadata and block structure on the x-axis.

partition, which acts as the main point of comparison in the partition landscape, or to disentangle block structure signal and metadata partition correlation. To account for this, the edge compression $c^m = \Sigma^m_{opt}/E$ can be referred to as a second dimension in our measure, which quantifies the absolute compression of the optimal partition per network edge. It can serve as a point of reference for the structure of the optimal partition and help us understand how much of γ is explained by correlation vs compressibility of the network under a given model. In the Methods section we include an analysis of the sensitivity of our measure to network topology and the role of the second dimension.

Results

We now summarise the metablox pipeline and state the choices we make for any calculation of $\boldsymbol{\gamma}_d$ in the remainder of this work: We calculate Σ^m_d by plugging the required quantities, as obtained by the metadata partition $\mathbf{d}$, into the description length formulas for SBM variant m. For Σ^m_{opt}, we use the graph-tool library to fit variant m to the observed network and find the partition $\mathbf{b}^m_{\text{opt}}$ that minimises the description length, refining the result of the agglomerative algorithm by running 1000 sweeps of the merge-flip MCMC [7]. We calculate Σ^m_* by performing 500 permutations of the metadata labels, calculating the description length of the network under each of them, and finding the description length for which only 1% of permutations have a lower description length.

Synthetic network

To illustrate the metablox measure, we generate a network with multiple, non-aligned, plausible explanations in terms of its block structure, create multiple sets of synthetic metadata and calculate metablox on two dimensions: under a degree-corrected (DC) and a planted

partition (PP) SBM. More specifically, we generate the network with the Stochastic cross block model (SCBM) [28], which facilitates the generation of such networks with 'ambiguous' mesoscales structures, by 'planting' (similar to the planted partition model [39]) multiple coexisting partitions. We use the SCBM to generate a network with $N = 100$ nodes, $B = 2$ blocks and expected degree $k = 10$. We plant two coexisting partitions, so that the network exhibits bicommunity (BC) structure (i.e. two assortative communities) as well as core-periphery (CP) structure (see the Methods section for details on how this network was constructed). Figure 1 shows this network in two separate visualisations, with nodes painted according to their block membership in the BC partition in Figure 1a and according to their block membership in the CP partition in Figure 1b. We also generate multiple sets of synthetic metadata, in a way such that the labels align with the block structure of the network to varying degrees, similar to the related work in Ref. [10]. We correlate half of the synthetic metadata with the BC partition and the other half with the CP partition, with varying level of correlation ρ. The metadata label of each node is equal to the node's block assignment of the planted structure with probability $(1+\rho)/2$. We increase ρ from 0 to 1 at steps of 0.01, yielding a total of 202 sets of metadata, half of which being 'bicommunity-like' and the other half being 'core-periphery-like'.

In Figure 3, each dot represents one set of metadata d for which we are plotting γ_d^{DC} and γ_d^{PP} for increasing correlation ρ, for the bicommunity-like metadata (left) and the core-periphery-like metadata (right). For Figure 3 only, we additionally calculate the BESTest p-value to illustrate the contribution of metablox compared to the most similar existing measure [10].

For the bi-community-like metadata, we can see that both dimensions of γ are a decreasing function of the correlation ρ: the more strongly correlated the metadata, the higher the relevance assigned to it by the metablox measure. As expected, we cannot make a distinction between the general (DC) and assortative (PP) SBM variant, since – by design – the metadata in this case is similar to the block structure that was created under an assortativity assumption. For core-periphery-like metadata, however, γ_d^{DC} decreases similarly to the previous case, while γ_d^{PP} is now an increasing function of ρ, suggesting that – as we know is true – if we were to assume that this metadata was responsible for generating this network, assortativity would be much less likely to explain its structure.

For comparison, we observe that the BESTest values become non-distinguishable once the correlation ρ has crossed a certain threshold. This is unsurprising: when the SBM entropy under a given metadata partition is lower than the SBM entropy under *all* sets of randomised metadata partitions then the p-value is $1/n_p$ (where n_p is the number of randomised metadata partitions) and we cannot compare this metadata to another set of metadata for which this is also true. In our example, BESTest does not allow us to compare a metadata partition that is nearly perfectly correlated with the planted block structure, to one for which $\rho = 0.7$, in both the bicommunity and core-periphery case. For comparison, the metablox values for DC cross the significance line $\gamma = 1$ at a similar point as BESTest becomes significant, for both the bicommunity and core-periphery-like metadata. However, it proceeds in the shape of a continuous decreasing function with a minimum of 0 at $\rho = 1$, which is what we

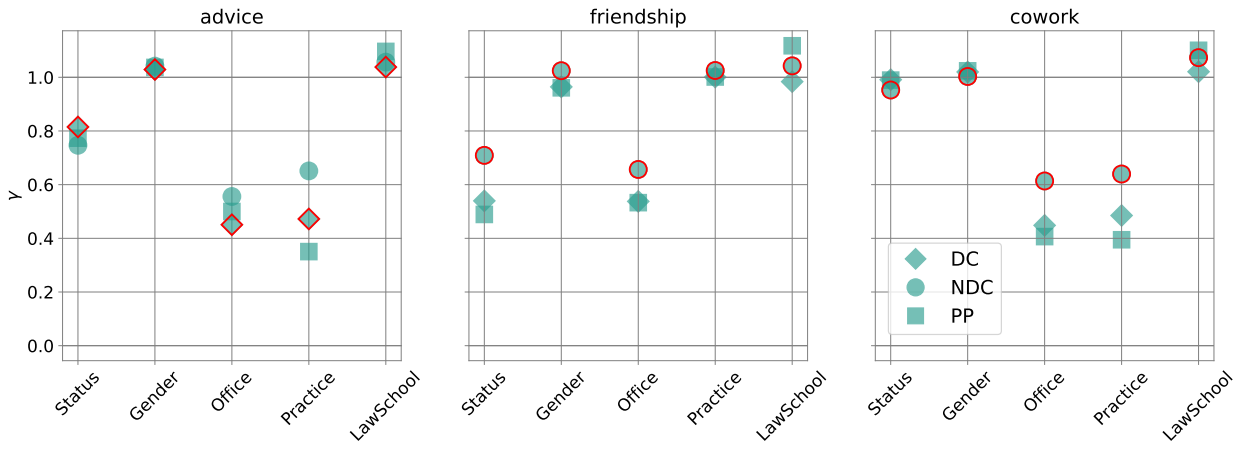


Figure 4: Metablox dimensions $(\gamma^{DC}, \gamma^{PP}, \gamma^{NDC})$ for different law firm networks[40]; on each figure, the SBM variant that gives the lowest edge compression for the network is highlighted in red.

would expect as the metadata partition is equal to the planted block structure at this point. The first contribution of metablox is therefore that, by comparing the metadata partition to the *optimal* partition (while still taking into account statistical significance), we are able to make direct comparisons between metadata sets. The second contribution is illustrated by the fact that BESTest does not tell us anything about likely structural arrangements of the metadata, while the difference between the trajectory of γ_d^{PP} on the left and right figure demonstrate the way in which metablox does allow for this.

Applications

We demonstrate the metablox measure on three real-world applications, in each case referring to one of the three scenarios (I-III) introduced above. As outlined previously, our measure can be extended to include any number of structure-specific SBMs. For our applications to real networks, we focus on the non-degree-corrected (NDC), the degree-corrected (DC), and the assortative 'planted partition' (PP) SBM. This means that we calculate a three-dimensional metablox vector for every network-metadata pair.

Law firm We first demonstrate our method on a number of networks from the Lazega law firm networks collection [40]. In this collection, there are three different networks among the same set of nodes, in which edges represents different types of connections between the employees of a corporate law firm: coworkers, friendship, and advice. For the employees, we have five sets of node attributes - their status in the firm, gender, the office in which they work, which type of law they practice, and which law school they attended. Figure 4 shows the metablox results for the three networks, each on a separate figure. On each figure, we show the results for the respective network and compare the five different metadata partitions. Each figure thus represents a case of scenario I, while we can also compare

network-metadata pairs which share a type of node attribute (scenario II). Note that we have highlighted the SBM variant which gave the best per edge compression with a red marker. We can see some considerable variation in how the sets of metadata are related to the networks' block structure under each of the SBM variants. Firstly, we observe that the strongest metadata-block structure relevance can be observed between the type of law practiced by an employee and the block structure in the advice network, under an assortative SBM assumption. This implies that employees are more likely to seek advice from colleagues who practice the same type of law as themselves. The same set of metadata is also strongly relevant in the cowork network. However, in the friendship network the law practiced by an employee does not seem relevant to the formation of blocks under any of the SBM variants. Employees' status, on the other hand, is the most strongly relevant metadata out of all attributes we have for the friendship network, under PP and DC. This indicates that a division of the network's nodes into the two available categories (partner and associate) in this set of metadata is more strongly aligned with a plausible partition according to shared connectivity patterns in the network, under the modelling assumptions of both PP and DC. In other words, the nodes that share a metadata attribute (partner or associate) are not only likely to share connectivity patterns independent of their node degrees (DC), they are also more likely to connect to employees with whom they share the status than with others (PP). Status matters less in the advice network (albeit still significantly relevant) and is essentially irrelevant in the cowork network. Remarkably, the office in which employees were located is the only metadata that is highly relevant for all three networks, indicating that the physical location of employees' plays a large role in determining connections. Both employees' gender and the law school they attended are not related to the block structure under any SBM variant and therefore appear to be essentially irrelevant in the tie generation process. In this example, we have been able to both compare different sets of metadata for a given network (scenario I) as well as different networks with the same node set and shared metadata (scenario II).

Peel at al. [10] used the same networks to demonstrate both of their methods, which we outlined above. Using the BESTest significance test, they reached similar conclusions for the network-metadata pairs. Using a second method – which admittedly relies on visual interpretation and does not explicitly serve to compare the strength of metadata relevance – the authors concluded that for the friendship network, the law school attended by employees was more strongly structure-relevant than the office in which employees are located. This finding is the opposite of that given by metablox. While the methods proposed by the above authors can provide insights into the significance of metadata relevance and the quality of the relationship for a given network-metadata pair, they do not enable a *direct quantification* of the strength of the relationship and of the likely prominent structure. As a result, they also do not enable a direct comparison of different networks (either made up of different or the same node sets), something our measure is designed to do as shown in the following paragraphs.

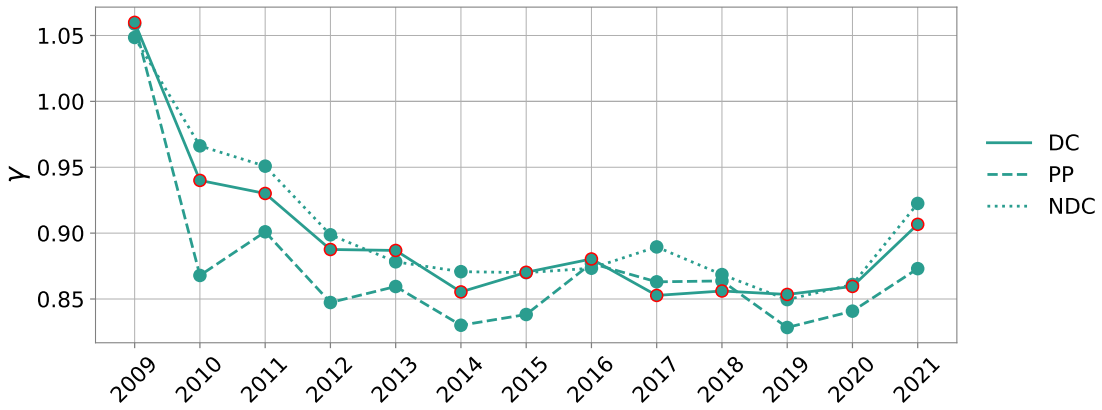


Figure 5: Static snapshots, representing a non-overlapping one year period each, of a Twitter/X retweet network among users discussing the topic of impact investing, with user location (country) as shared metadata (scenario II). The SBM variant that gives the lowest edge compression for each network is highlighted in red.

Impact investing We continue with a second example for scenario II, in which we investigate the metadata-block structure relationship in Twitter/X retweet networks of users discussing the topic of 'impact investing' [41]. In October 2021 we exhaustively collected tweets mentioning any term belonging to a small subset of hashtags directly relevant to this topic. This amounted to 1.89m tweets published by around 300k accounts since 2007. We deemed users who posted less than a tweet a month on average or who never posted more than one year apart to be insufficiently active in that field and filtered them, leading to a core set of around 16k users. We subsequently inferred likely home countries from the geolocation voluntarily shared by users in their profile description. Interestingly, most users appeared to come from an English-speaking country: countries with at least 1,000 active accounts include only the US, Canada, Australia and the United Kingdom, covering 72% of the data. Focusing on these country users, we eventually study 13 static networks representing each a snapshot for each year between 2009 and 2021 (we excluded 2007 and 2008 which yielded trivial networks).

This time, the edges in the different networks represent the same type of interaction at different moments in time. The metadata here is the country which the user declares being located in and which we consider to be fixed over time, being inferred at the moment of data collection in 2021. We calculated the three metablox dimensions for each network snapshot and also highlighted, for each network, the SBM variant that provided the best compression per edge for its optimal partition. In Figure 5, we can see that the country in which a user is located starts being relevant to the network's block structure in 2010, under all SBM variants but most strongly under PP. In the next five years (until 2015) the metadata becomes more strongly relevant in general and assortativity remains the prominent arrangement of the metadata. In 2016, there seems to be a change in the structural organisation of the network in relation to the metadata: the country in which users are located remains relevant to the

block structure but in 2017 and 2018, PP is replaced by DC as the best fit for the metadata, before it swaps again starting in 2019 up until the last year of data collection. It is worth noting that for each year, DC is the SBM variant for which the optimal inferred partition provided the best compression. Overall, the quantification of the relevance of the metadata afforded by metablox enables us to observe some interesting changes over time that should serve as a starting point to 'zoom into' the networks for those particular events to gain a deeper understanding of the relationship between the users' location and their tweeting behaviour: the change from 2009 to 2010 when the country user location becomes significant, the years 2014 and 2019 when the metadata is most strongly relevant under PP; and the years 2016-2018, in which the networks seem to exhibit a change in the metadata-structure relationship.

Twitter/X In our final example, we proceed with another set of networks created from Twitter/X, this time demonstrating metablox for networks in the same context and not necessarily with the same set of nodes (scenario III). In particular, we use data from three debates on political topics in the US, for which we are studying the metadata-block structure relationship for the same type of node attributes. The original set of users on which these networks are based were collected by the authors of Ref [42] and the ones used in this paper were recollected by the authors of Ref [43]. The tweets from which these networks were created were collected between 2015 and 2016, and are based on conversations on abortion, Obamacare, and gun control. In all three networks, the available node metadata reflects two categories of political orientation as either liberal or conservative, based on calculations in Ref [43] to estimate political opinion scores from shared URLs. The liberal-conservative opinion categories are based on a continuous score between -1 and +1 that were calculated [43] based on URLs shared by Twitter accounts and the categorisation of websites behind these URLs on https://mediabiasfactcheck.com – a method originally used by Ref [44]. The two categories used as node metadata are based on users below and above a 'neutral' score of 0. Note that we do not know to what extent there is an overlap between the set of users participating in the debates represented by the three networks, neither do we care; here, the subject of comparison is the topic that is discussed in each network and to what extent the political stance of users is related to the block structure for each topic.

In Figure 6, we plot the metablox vectors for these interaction networks with binary political orientation metadata. We observe that for the topics of gun control and abortion, the metadata partition into liberals and conservatives is relevant under all variants but most strongly under PP, while DC provides – again – the best overall fit. For the obamacare network, the metadata is barely relevant under NDC and not significantly relevant under the other variants. We can conclude that for gun control and abortion, the metadata partitions were likely to be at least somewhat related to the edge generating process and that assortativity is the prominent structural arrangement of the metadata in those cases.Interestingly, our findings seem equivalent to previous findings on polarisation levels in these networks, which classified all three networks to be polarised to some extent with the Obamacare network being the least polarised [43]. Similar to our work, their polarisation measure also considers

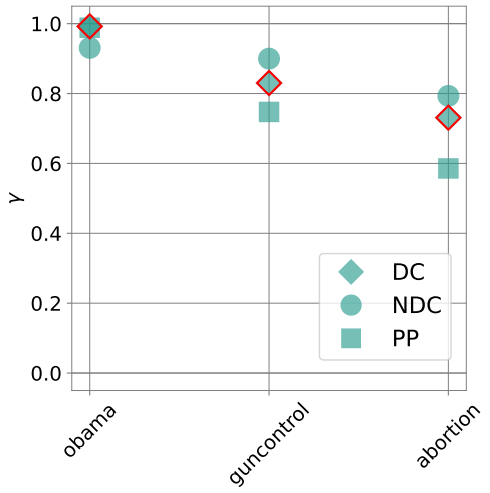


Figure 6: Three Twitter/X interaction networks[42, 43] with shared metadata representing the users' political stance (liberal vs conservative. The SBM variant that gives the lowest edge compression for each network is highlighted in red.

both network structure as well as a metadata dimension. However, their measure considers continuous node attributes – specifically political ideology on a continuous left-right scale – and specialises on measuring polarisation rather than the more general approach we take here. So although our measure has a different purpose, it is interesting to see that it might be able to pick up specific network properties (such as polarisation or fragmentation) while also being general enough to enable broader comparisons.

Discussion

In this paper, we have introduced a novel measure for probing the relationship between network metadata and its structural organisation. Our metablox pipeline, which produces the vector $\boldsymbol{\gamma}$, is designed to provide insights into the relevance of metadata to a network's block structure and the likely structural arrangement in various scenarios: for the comparison of multiple sets of metadata for one given network (scenario I) and to compare the metadata-block structure relationships for entire collections of networks that share the same type of metadata, for networks that have the same node set (scenario II) and for networks within the same context (scenario III).

We have demonstrated the metablox measure on a synthetic network and applied it to a number of real networks, including the Lazega law firm networks and networks representing different social media debates. The results reveal variations in the relevance of metadata partitions to block structure, and the likely structural arrangement, therefore providing insights into the underlying dynamics of these networks. By covering examples from the three scenarios (I-III), we have demonstrated that our measure allows for a comprehensive inter- and

intra-network comparison, enabling researchers to quickly identify networks where specific metadata partitions are closely related to structure or where certain structural arrangements are likely or unlikely under the metadata.

We have discussed a number of limitations of the measure, related to the specific design of metablox as a comparative measure, and proposed ways to address these. Additional limitations are connected with the specific type of node attributes for which relevance can be measured. In its current framework, metablox can only quantify the relevance of non-overlapping categorical metadata, while overlapping, real-valued or multidimensional metadata cannot be evaluated. For the case of overlapping metadata, it is conceivable that an overlapping SBM variant may be used. For real-valued and multidimensional metadata, the options are less clear. We therefore see an interesting research direction in an alternative version of the metablox tool for real-valued metadata. While this clearly requires a number of non-trivial decisions with respect to suitable generative models, providing an equivalent measure for non-discrete metadata seems a valuable extension.

Other potentially interesting research direction evolve around the particular type of structural arrangements of metadata categories. Currently, we focus on degree-corrected (DC), non-degree-corrected (NDC), and planted partition (PP) stochastic block models. However, networks often exhibit more complex structural arrangements beyond these models and so might their metadata. Future work could explore the integration of other SBM variants that are tailored to specific structural motifs, such as core-periphery structures, bipartite structures or nested patterns. The current measure has been implemented for undirected graphs, but extensions to more complex network structures such as directed graphs are straightforward and should also be considered as part of future research.

Our measure has the potential to serve as a tool for conducting large-scale comparisons of collections of network-metadata pairs, for networks coming from a variety of research fields – as long as categorical node metadata is available. Obvious examples are: different types of social networks, for which metadata may include a range of demographics or affiliations; biological networks such as gene regulatory networks, protein-protein interaction networks, and ecological networks with metadata related to genes, proteins, or species; economic and financial networks, such as trade networks, supply chains, and stock market networks, for which metadata may relate to industries, sectors, or companies; or networks from science of science, where networks represent collaborations and knowledge flows, and metadata may include fields of scientific research.

Methods

Data compression & microcanonical SBMs

To justify the use of description length as part of the metablox formulation, we expand on the Results section in the main text and more throughly explain the minimum description length (MDL) principle and its relationship with microcanonical SBMs. MDL is a model

selection criterion according to which one should favour the model that achieves the smallest compression of the data. The idea behind this is that compression is possible when we find regularities in the data which, in turn, means that we 'learn' about patterns in the data [30]. MDL is sometimes described as a formal interpretation of Occam's razor – also known as the principle of parsimony – which is the idea that one should try to find the explanation with the smallest number of assumptions possible. In more formal terms, the best hypothesis H (e.g. a model with its parameters) for a data set D, is the one that minimises the sum $S(H) + S(D|H)$, where $S(D|H)$ is the amount of information required to describe the data D when it has been encoded with the hypothesis H and $S(H)$ is the amount of information necessary to describe the hypothesis itself. This demonstrates the 'automatic' overfitting-prevention property of MDL, which makes it an attractive model selection criterion: with a more complex hypothesis, we need less information to describe the data given the hypothesis, but we need more information to describe the hypothesis itself.

There is a strong relationship between MDL and Bayesian inference in general [30] and the Bayesian interpretation of the SBM more specifically, where it was first used to infer network partitions without knowing the number of blocks in advance [45]. To calculate the description length of a network under a particular model, one needs to derive the entropy of the individual components of the SBM, which is an ensemble of networks that can be generated from a set of parameters (i.e. the partition). In general, *microcanonical* network ensembles – for which structural constraints need to be satisfied exactly rather than on average – can be described by their entropy $S = \ln \Omega(\theta)$, where $\Omega(\theta)$ is the total number of networks that can be generated under the given set of parameters θ [34]. The higher the entropy of a network ensemble, the more 'disordered' (or 'random') is the ensemble. More concretely, let us assume that we have a network $\mathbf{A}$ generated by a model with parameter set θ. $P(\mathbf{A}|\theta)$ is the probability of observing the network A in an ensemble generated by the model with these parameters (i.e. the likelihood) and we assume that all networks occur with the same probability $P(\mathbf{A}|\theta) = \frac{1}{\Omega(\theta)}$. From this assumption, one can straightforwardly make a connection between the microcanonical entropy S and the log-likelihood: $L = \ln P = -\ln \Omega(\theta) = -S$ [33]. By minimising the entropy S one could therefore find the maximum likelihood parameters, such as the most likely partition, given an observed network. While maximum likelihood methods work well in many cases, in the particular case of model selection with SBMS, maximum likelihood estimation can lead to overfitting if the number of model parameters is not fixed. For example, if the number of blocks B is not known, minimising the entropy would lead to the trivial partition of every node being its own block, i.e. $B = N$, where N is the number of nodes. Peixoto proposed the MDL principle as part of a microcanonical *nonparametric* approach to SBM inference, to address the overfitting issue and enable the use of flexible priors and hyperpriors on the model parameters [5, 45]. Specifically, this means considering the full joint distribution of the network and the SBM model parameters as part of the inference process, rather than just the SBM likelihood, which can be interpreted as calculating the description length of the network under the given model, via its parameters.

We recall that for a network A of size N with E edges, we denote the degree of a node i by k_i, and we use the following notation for the parameters of a microcanonical SBM with B blocks. The block assignments of the nodes is denoted by the vector $\mathbf{b} = \{b_i\}$ of length N, the $B \times B$ matrix $\mathbf{e} = \{e_{rs}\}$ represents the edge counts within and between two blocks r and s (with twice the number of edges on the diagonal, as is convention), and the B-dimensional vector $\mathbf{n} = \{n_r\}$ describes the number of nodes in each block r. The full joint distribution of the degree-corrected SBM is, as introduced above, given by $P(\mathbf{A}, \mathbf{k}, \mathbf{e}, \mathbf{b}) = P(\mathbf{A}|\mathbf{k}, \mathbf{e}, \mathbf{b})P(\mathbf{k}|\mathbf{e}, \mathbf{b})P(\mathbf{e}|\mathbf{b})P(\mathbf{b})$. In terms of the generative process of this model, this means that one first samples a partition of the nodes into blocks, then samples the numbers of edges within and between the blocks, then samples the half-edges according to the node degrees, and finally connects half-edges accordingly to create the network (i.e. sampling from the networks that are possible given the partition and number of edges).

When using the nonparametric Bayesian framework for inference purposes, one can then use this formulation to maximise (or sample from) the *posterior* distribution of partitions

$$P(\mathbf{b}|\mathbf{A}) = \frac{P(\mathbf{A}, \mathbf{b})}{P(\mathbf{A})} = \frac{P(\mathbf{A}|\mathbf{b})P(\mathbf{b})}{P(\mathbf{A})}. \tag{2}$$

This is where the microcanonical formulation helps simplify the inference problem, which allows us to write $P(\mathbf{A}|\mathbf{b}) = P(\mathbf{A}|\mathbf{k}, \mathbf{e}, \mathbf{b})P(\mathbf{k}|\mathbf{e}, \mathbf{b})P(\mathbf{e}|\mathbf{b})$. The marginal likelihood is usually $P(\mathbf{A}|\mathbf{b}) = \sum_{\mathbf{e}} P(\mathbf{A}|\mathbf{k}, \mathbf{e}, \mathbf{b})P(\mathbf{k}|\mathbf{e}, \mathbf{b})P(\mathbf{e}|\mathbf{b})$, but due to the 'hard' constraints of the microcanonical SBM, there is only one non-zero element in this sum [45].

It turns out that this is where we can re-introduce information theoretical interpretation. In particular, the posterior can be rewritten as

$$P(\mathbf{b}|\mathbf{A}) = \frac{P(\mathbf{A}|\mathbf{k}, \mathbf{e}, \mathbf{b})P(\mathbf{k}|\mathbf{e}, \mathbf{b})P(\mathbf{e}|\mathbf{b})P(\mathbf{b})}{P(\mathbf{A})} = \frac{e^{-\Sigma}}{P(\mathbf{A})} \tag{3}$$

where $\Sigma = -\ln P(\mathbf{A}|\mathbf{k}, \mathbf{e}, \mathbf{b}) - \ln P(\mathbf{k}, \mathbf{e}, \mathbf{b})$ is the description length. Equivalently to the more general introduction to description length above, the first component is the amount of information required to describe the network under the SBM and the given parameters and the second component is the amount of information needed to describe the parameters themselves. Due to this definition, finding the partition $\mathbf{b}$ that minimises the description length is equivalent to finding the partition that maximises the posterior, i.e. to finding the most likely partition of the network.

The description length can be used to identify the most likely model given the observed data, upon comparing multiple competing models. One way of interpreting a 'model' in this context is as a particular partition $\mathbf{b}_1$ under an SBM variant m_1. We can identify for which of two partitions $\mathbf{b}_1$ and $\mathbf{b}_2$ (under models m_1 and m_2 respectively) there is more evidence

in the data, by calculating their posterior odds ratio

$$\begin{aligned}\Lambda &= \frac{P(\mathbf{b}_1, m_1|\mathbf{A})}{P(\mathbf{b}_2, m_2|\mathbf{A})} \\ &= \frac{P(\mathbf{A}|\mathbf{k}, \mathbf{e}_1, \mathbf{b}_1, m_1)P(\mathbf{k}|\mathbf{e}_1, \mathbf{b}_1, m_1)P(\mathbf{e}_1|\mathbf{b}_1, m_1)P(\mathbf{b}_1|m_1)P(m_1)}{P(\mathbf{A}|\mathbf{k}, \mathbf{e}_2, \mathbf{b}_2, m_2)P(\mathbf{k}|\mathbf{e}_2, \mathbf{b}_2, m_2)P(\mathbf{e}_2|\mathbf{b}_2, m_2)P(\mathbf{b}_2|m_2)P(m_2)} \\ &= e^{-\Delta\Sigma},\end{aligned} \tag{4}$$

where $\Delta\Sigma = \Sigma_1 - \Sigma_2$ and $\Sigma_i = -\ln P(\mathbf{A}|\mathbf{k}, \mathbf{e}_i, \mathbf{b}_i, m_i) - \ln P(\mathbf{k}, \mathbf{e}_i, \mathbf{b}_i, m_i)$ is the description length of model i (e.g. of the network under SBM m_i with partition $\mathbf{b}_i$). Here we assume that both variants are equally likely i.e., $P(m_1) = P(m_2)$[5]. The first component of Σ_i is the amount of information required to describe the network under model i and the given parameters; the second component is the amount of information needed to describe the parameters themselves. One can therefore identify the more likely model (in terms of the specific parameters) by calculating the description lengths of the network under each model and partition. For $\Lambda = 1$ or, equivalently, $\Sigma_1 = \Sigma_2$, the models are equally likely and for $\Lambda > 1$ ($\Sigma_1 < \Sigma_2$) model m_1 is more likely than model m_2.

Description length calculation

As described above, the description length Σ of a network $\mathbf{A}$ under an SBM with a set of parameters θ can be directly derived from its full joint distribution, since $\Sigma = -\ln P(\mathbf{A}, \theta) = -\ln P(\mathbf{A}|\theta) - \ln P(\theta)$. Here, we detail the description length calculation for each of the SBM variants discussed in the main text, by providing their joint distributions, made up of model likelihood and priors. We will see that the microcanonical framework makes it possible to derive the likelihood and priors through combinatorics. Specifically, to be as parsimonious as possible, the priors tend to be uniform distributions over the number of possible realisations of a particular parameter under the given modelling assumptions.

Note that in the formulas of the description length calculations, we use the notation introduced for the SBM parameters. However, in the metablox measure, the description length is not only used as part of the inference of the optimal partition but it is also calculated for the metadata partition. To calculate the metadata description length, we simply replace the parameters $\mathbf{b}$, $\mathbf{e}$, $\mathbf{n}$, B by the respective metadata quantities that can be directly induced by $\mathbf{d}$. For example, $\mathbf{e}'$ is the number of edges within and between the node sets with shared node attributes, $\mathbf{n}'$ is the number of nodes that share each type of node attributes, and B' is the number of unique node attributes in $\mathbf{d}$.

The formulations we provide here for the likelihood and priors for the non-degree-corrected (NDC) and degree-corrected (DC) SBM are based on the work in Refs. [5, 33], those for the assortative planted partition (PP) SBM are from Ref. [38].

For NDC, the only model parameters are the edge counts e_{rs} between blocks r and s and the block assignment vector $\mathbf{b}$, so the model is fully described by $P(\mathbf{A}, \mathbf{e}, \mathbf{b}) = P(\mathbf{A}|\mathbf{e}, \mathbf{b})P(\mathbf{e}|\mathbf{b})P(\mathbf{b})$,

where $P(\mathbf{A}|\mathbf{e},\mathbf{b})$ is the model likelihood of NDC, $P(\mathbf{e}|\mathbf{b})$ is the prior on edge counts and $P(\mathbf{b})$ is the prior on the partition. For DC, we need to consider the additional prior on the degree sequence $\mathbf{k}$, and we thus have $P(\mathbf{A},\mathbf{e},\mathbf{k},\mathbf{b}) = P(\mathbf{A}|\mathbf{e},\mathbf{k},\mathbf{b})P(\mathbf{k}|\mathbf{e},\mathbf{b})P(\mathbf{e}|\mathbf{b})P(\mathbf{b})$ [5], where $P(\mathbf{k}|\mathbf{e},\mathbf{b})$ is the probability of the degree sequence. For PP, the prior on the edge counts serves as a constraint on the network to favour assortative structure. The full joint distribution can be written as $P(\mathbf{A},\mathbf{e},\mathbf{k},\mathbf{b}) = P(\mathbf{A}|\mathbf{e},\mathbf{k},\mathbf{b})P(\mathbf{k}|\mathbf{e},\mathbf{b})P(\mathbf{e}|e_{\text{in}},e_{\text{out}},\mathbf{b})P(e_{\text{in}},e_{\text{out}}|E,\mathbf{b})P(E)P(\mathbf{b})$, where E is the total number of edges in the network, e_{in} and e_{out} are the number of edges within and between blocks respectively, and where we use the model likelihood of DC [38].

We start with the elements of the joint distribution of NDC,

$$P(\mathbf{A},\mathbf{e},\mathbf{b}) = P(\mathbf{A}|\mathbf{e},\mathbf{b})P(\mathbf{e}|\mathbf{b})P(\mathbf{b}) \tag{5}$$

The model likelihood of NDC is given by

$$P(\mathbf{A}|\mathbf{e},\mathbf{b}) = \frac{\prod_{r<s} e_{rs}! \prod_r e_{rr}!!}{\prod_r n_r^{e_r} \prod_{i<j} A_{ij}! \prod_i A_{ii}!!} \tag{6}$$

The prior for the block matrix e_{rs} is a uniform distribution over the total possible number of symmetric block matrices given B, with the constraint that the sum of all elements must equal $2E$:

$$P(\mathbf{e}|\mathbf{b}) = \left(\!\!\binom{B(B+1)/2}{E}\!\!\right)^{-1} \tag{7}$$

The prior on the partition is defined as

$$\begin{aligned} P(\mathbf{b}) &= P(\mathbf{b}|\mathbf{n})P(\mathbf{n}|B)P(B) \\ &= \frac{\sum_r n_r!}{N!}\binom{N-1}{B-1}^{-1}\frac{1}{N} \end{aligned} \tag{8}$$

Here, $P(\mathbf{b})$ and $P(\mathbf{n}|B)$ are hyperpriors on the number of blocks B and on the block sizes n_r respectively, to be as parsimonious as possible about these parameters.

In the case of DC, the model likelihood includes terms for the degree sequence $\mathbf{k}$, so that:

$$P(\mathbf{A}|\mathbf{e},\mathbf{k},\mathbf{b}) = \frac{\prod_{r<s} e_{rs}! \prod_r e_{rr}!! \prod_i k_i!}{\prod_r e_r! \prod_{i<j} A_{ij}! \prod_i A_{ii}!!} \tag{9}$$

Additionally, the DC case also includes a prior on the degree sequence k, namely

$$P(\mathbf{k}|\mathbf{e},\mathbf{b}) = \prod_r \frac{\prod_k \eta_k^r!}{n_r!} \prod_r q(e_r,n_r)^{-1} \tag{10}$$

where η_k denotes the number of degree-k nodes in group r and $q(x,y)$ is the number of times an integer x can be partitioned into a maximum of y parts [5].

For PP, the prior on the block matrix needs to be defined differently, to encode the constraint that is put on the structural arrangement [46]. In fact, Zhang and Peixoto [38] proposed two different versions of this probability: one which assumes that *uniform* expected number of edges within each community and one that allows the number of expected edges to vary across communities (*non-uniform*). Here, we give the formulation for both versions, since in our analysis, we use the uniform version in the case of synthetic networks (since they are generated with equal size blocks) and the non-uniform version in the analysis of the metablox vector on real networks. The uniform version of the prior on the edge counts in PP is described by

$$P(\mathbf{e}|e_{\text{in}}, e_{\text{out}}, \mathbf{b})P(e_{\text{in}}, e_{\text{out}}|E, \mathbf{b}) \tag{11}$$

with

$$P(\mathbf{e}|e_{\text{in}}, e_{\text{out}}, \mathbf{b}) = \frac{e_{\text{in}}!e_{\text{out}}!}{B^{e_{\text{in}}}\prod_r (e_{rr}/2)!\binom{B}{2}^{e_{\text{out}}}\prod_{r<s} e_{rs}!}. \tag{12}$$

This is equivalent to the product of two uniform multinomial distributions: one for the elements of the block matrix that correspond to the within-block edge counts and one for those that correspond to the between-block edge counts, given e_{in} and e_{out}. The second part is then the hyperprior on e_{in} and e_{out}:

$$P(e_{\text{in}}, e_{\text{out}}|E, \mathbf{b}) = \left(\frac{1}{E+1}\right)^{1-\delta_{B,1}} \tag{13}$$

For the non-uniform version, the prior is also made up of two probabilities

$$P(\mathbf{e}|\{e_{rr}\}, e_{\text{out}}, \mathbf{b})P(\{e_{rr}\}, e_{\text{out}}|\mathbf{b}, E) \tag{14}$$

where

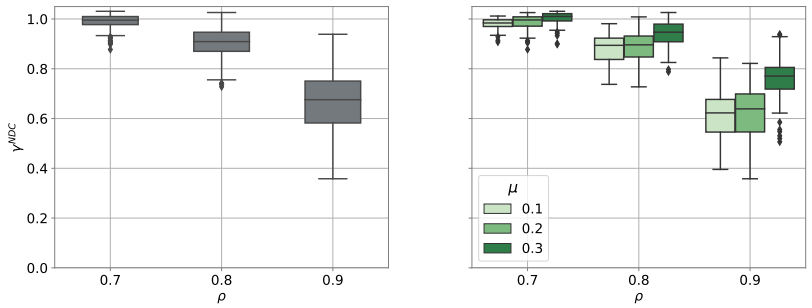
$$P(\mathbf{e}|\{e_{rr}\}, e_{\text{out}}, \mathbf{b}) = \frac{e_{\text{out}}!}{\binom{B}{2}^{e_{\text{out}}}\prod_{r<s} e_{rs}!} \tag{15}$$

corresponds to a uniform multinomial distribution for the off-diagonal elements of the block matrix given e_{out}. The second component is made out of a uniform distribution over all possible values e_{in} from E, and a uniform distribution over all ways of choosing the set of diagonal block matrix values $\{e_{rr}\}$, given e_{in}:

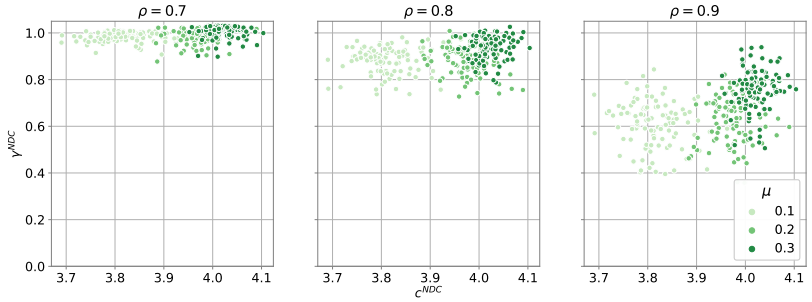
$$\begin{aligned} P(\{e_{rr}\}, e_{\text{out}}|\mathbf{b}, E) &= P(\{e_{rr}\}|e_{\text{in}}, \mathbf{b})P(e_{\text{in}}|E, \mathbf{b}) \\ &= \binom{B + e_{\text{in}} - 1}{e_{\text{in}}}^{-1}\left(\frac{1}{E+1}\right)^{1-\delta_{B,1}} \end{aligned} \tag{16}$$

The role of network compressibility

In the main text, we discussed the limitations of our measure and the way in which they can be mitigated by using the compression of the network under the *optimal* per network edge, $c^m = \Sigma^m_{opt}/E$, as a second dimension.



(a) Distributions of γ_d^{NDC} values for a total of 300 networks, each with three sets of metadata. The values of γ_d^{NDC} are shown for each correlation ρ on the left-hand side and then disaggregated by block structure signal strength μ on the right-hand side.



(b) Distributions of the same networks, showing the edge compression c_{NDC} along the x-axis, separately for each correlation value ρ.

Figure 7: Using edge compression to disentangle block structure signal strength and metadata correlation. The center line of the box plots refers to the median, box limits to upper and lower quartiles and whiskers to 1.5 times the interquartile range; remaining points are outliers.

Recall that we considered two limitations: one related to identifying the best SBM variant as part of a comparison of different dimensions of γ and another regarding the conflation of block structure signals and metadata correlation. In Figures 4, 6, and 5, we used edge compression to highlight the particular SBM variant whose optimal partition provided the best partition of the network and therefore illustrated how this second partition can be used to address the first limitation. Including this additional information enabled a distinction between the cases in which the metadata is closest – in terms of how well it compresses the network – to *the* optimal partition among the variants included in the analysis, or merely to the particular variant being measured. In other words, when the strongest relevance is measured under DC and DC is also the model whose optimal partition provides the best fit overall, we know that the metadata partition is strongly relevant not just under the particular model but in comparison to the most optimal partition we have for this network.

Here, we provide a concrete example for the second limitation: on the ability of the second dimension to disentangle the signal of the block structure and the metadata correlation. For the purpose of this demonstration, we generate a total of 300 synthetic networks with community structure, for each of which we fix size $N = 200$, expected degree $k = 10$ and number of blocks B. We create the networks in this collection such that we end up with three sets of 100 networks with differently 'strong' community structure. The stronger the signal of the community structure, the more pronounced the difference between the within- and between-block edge probability. Specifically, we generate the networks using an SBM with block matrix $\boldsymbol{\theta}_{\text{BC}} = 2E\left(\begin{smallmatrix} 1-\mu & \mu \\ \mu & 1-\mu \end{smallmatrix}\right)$, with $\mu = 0.1$ for the strongest signal, $\mu = 0.2$ for a medium signal and $\mu = 0.3$ for weak community structure. For each network, we generate three sets of categorical node metadata, correlating with the planted block structure with correlation $\rho = 0.7; 0.8; 0.9$. In Figure 7, we visualise an analysis of the NDC element of metablox for these networks; we refrain from including it for other dimensions as those yielded very similar results. In the panel on the left-hand side of Figure 7a, we see that increasing metadata correlation leads to lower values of γ. On the right-hand side, we disaggregate the data by signal strength. We observe that for each correlation value ρ, the γ values for the networks with strong and medium community structure are essentially indistinguishable, while the networks with weak community structure have a larger value of γ for the same ρ. While γ on the whole does a good job at differentiating between the correlation strength of the different network-metadata pairs, the differences we observe upon separating by community strength call for further analysis. Especially in light of working with real networks, whose generative process is unknown to us, these insights demand a tool to disentangle the different impacting factors on the metadata relevance. In Figure 7b, we demonstrate the way in which the edge compression mitigates this limitation: As expected, the stronger the community structure, the 'better' (i.e. smaller c) the compression of the optimal partition per network edge.

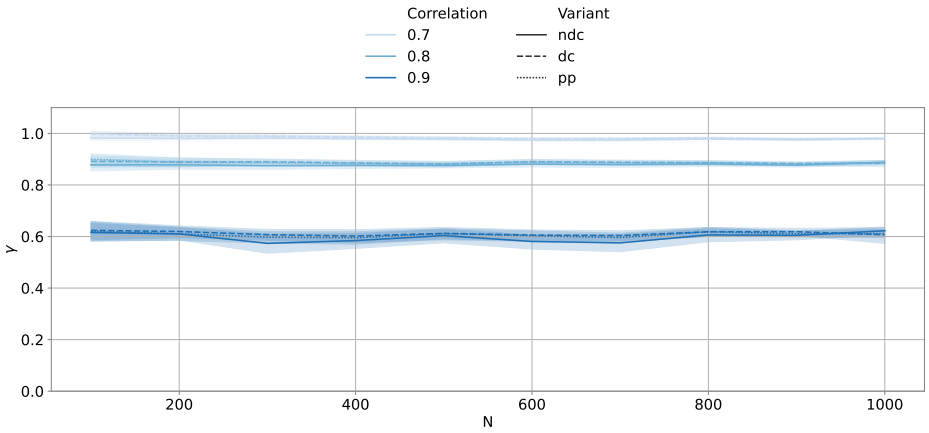


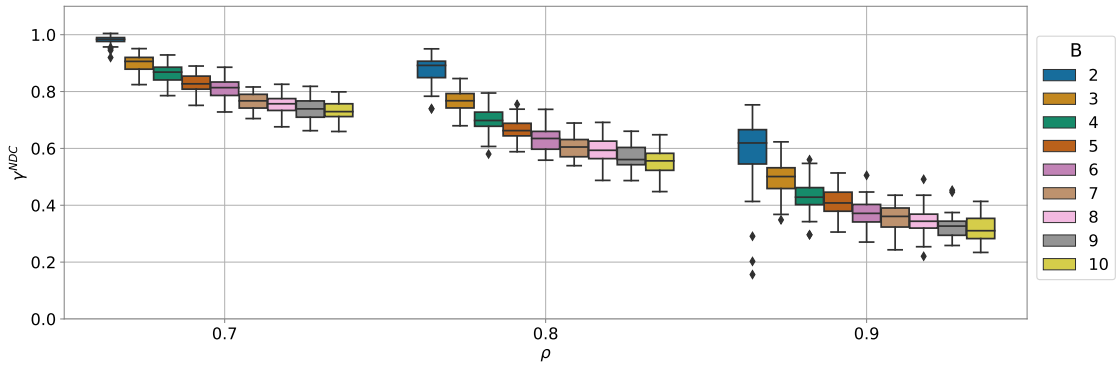
Figure 8: Metablox values for networks of varying size N, for three different correlation values ρ and three variants. The lines show the mean metablox values across 50 networks, with 95% confidence intervals.

Robustness analysis

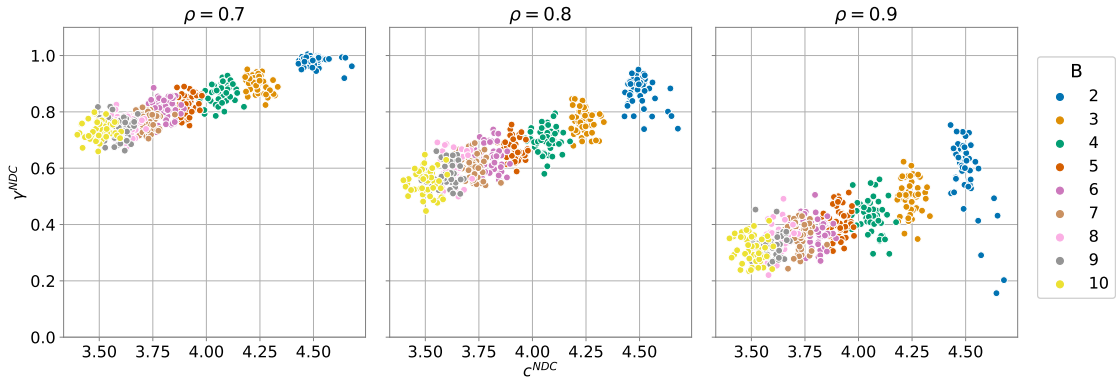
To understand the suitability of metablox to be used as part of comparative studies, we analyse the sensitivity of the measure to networks of different sizes and topologies. We first discuss the parameters to which we do and do not expect the measure to be robust. For an inter-network comparison, it is clearly important that our measure is robust with respect to varying network size N, which is not a trivial question since the description length of a network increases with growing N. For structural differences between two networks that are related to the compressibility under an SBM, such as the number of blocks, the edge density, or the signal strength of the block structure, the question becomes more nuanced. In fact, we have already discussed in that – due to our definition of metadata-block structure relevance – we *expect* the measure to be impacted by the compressibility of a network: metadata is more relevant to a network with stronger block structure signal. We therefore expect metablox to yield stronger relevance for network-metadata pairs for which the SBM variant provides better compression (e.g. for increasing numbers of blocks.

To test these assumptions, we generate a number of synthetic networks. Specifically, we increase N from $N = 100$ to $N = 1000$ at steps of 100, and generate 50 networks for each N, all with expected degree $k = 10$ and with planted bicommunity structure ($B = 2$). For the within- and between block edge counts, we use the same block matrix $\boldsymbol{\theta}_{\mathrm{BC}}$ as above, with $\mu = 0.1$. In Figure 8, we plot the mean description lengths and metablox (plus 95% confidence intervals) for each value of N and three SBM variants, for correlation values of $\rho = 0.7, 0.8, 0.9$. We clearly see the intended normalising effect, as γ remains stable for growing N for all three variants.

To test the measure's behaviour for varying numbers of blocks, we fix the network size at $N = 400$ and leave all other parameters as above, increasing B from 2 to 10, again generating



(a)



(b)

Figure 9: Panel (a): Metablox values for networks with varying number of blocks B, for three different correlation values ρ and variant NDC. Panel (b): Metablox values for the same data, as a function of edge compression.

50 networks at each step. In Figure 9a we show the results for γ_d^{NDC} (the results for the other two variants are similar). We observe that, for each value of ρ, γ decreases with the number of blocks, confirming our expectation of the measure capturing stronger metadata relevance for more compressible networks. In Figure 9b, we show how the edge compression of the network can be used as a second dimension, if a distinction between metadata correlation and block structure signal strength is required. The three figures show the edge compression c on the x-axis and γ on the y-axis. A lower value of c implies better compression, and we can see that the second dimension separates the compressibility from the metadata correlation by placing the most compressible networks low on the x-axis and low on the y-axis. In a network comparison, in which disentangling these two factors is important, researchers can plot the two dimensions in this way to draw the correct conclusions.

Heterogeneous partition landscape

For the synthetic network used to illlustrate the measure in the Results section, we demonstrate here that this network does, in fact, exhibit a heterogeneous partition landscape – as designed by the SCBM [28]. Recall that we created a network with $N = 100$ nodes, $B = 2$ blocks and expected degree $k = 10$, generated in a way such that upon fitting an SBM to the network and sampling from the posterior distribution, one recovers at least two 'clusters' of partitions – a bicommunity partition (BC) and a second one dividing the network into a core and a periphery (CP). We choose $\boldsymbol{\theta}_{\mathrm{BC}} = 2E\left(\begin{smallmatrix} 1-\mu & \mu \\ \mu & 1-\mu \end{smallmatrix}\right)$ and $\boldsymbol{\theta}_{\mathrm{CP}} = 2E\left(\begin{smallmatrix} 1-\lambda & \frac{1}{2} \\ \frac{1}{2} & \lambda \end{smallmatrix}\right)$ as block matrices, with $\mu = 0.25$ and $\lambda = 0.05$ and where E denotes the total number of edges. The choice of μ and λ is motivated by the findings in Ref [28], where it was demonstrated that the partition landscape inferred by a degree-corrected SBM on a network with two partitions planted with these parameters does, in fact, uncover both the planted partitions. The probability of the existence of edges in this network depends solely on the block membership of nodes and on the edge probability given by these two matrices.

For the purpose of demonstrating that this network's partition landscape does have multiple competing explanations, we fit an SBM and sample from the posterior distribution by using the methods from the graph-tool library [26]. In particular, we use the degree-corrected SBM variant (DC), which is meant to account for heterogeneous degree distributions within blocks [47]. The description lengths Σ^{DC} of the network according to these partitions are plotted in grey dots on the x-axis in Figure 10. We also calculate the description lengths Σ^{PP} of the network according to each partition under the planted partition SBM (PP), the variant of the model that assumes assortativity [38]; these are shown on the y-axis of the same figure. Note that, unsurprisingly, we find that $\Sigma^{\mathrm{PP}} \geq \Sigma^{\mathrm{DC}}$ for all partitions, since partitions were found by DC and corresponding description lengths were then calculated for the same partitions under PP. If we take a closer look at the partitions that we sampled, we find groups of similar partitions, here marked by the dashed-line circles, that look like the two partitions we planted. We show representative partitions for each partition cluster in the two network visualisations on the left-hand side of the plot, in which nodes are painted according to block assignments. We clearly see a strong similarity between these inferred partitions and each of the two planted partitions in Figure 1 and we therefore conclude that we have successfully created a network with the desired diverse partition landscape. We also observe that for those inferred partitions that are similar to the planted bicommunity partition, the PP offers a similarly good encoding of the network as the DC, since $\Sigma^{\mathrm{PP}} \approx \Sigma^{\mathrm{DC}}$ for those partitions. In contrast, the description lengths of the network are considerably higher under the same partitions but according to PP. This illustrates that calculating the description length of a network under different models can help us probe its partition landscape.

In line with the objective of our measure, we use this example network to support our hypothesis that using description length is a suitable tool to understand the way in which metadata is relevant for different parts of the partition landscape. For this purpose, we calculate Σ_d^{DC} and Σ_d^{PP} for each of the 202 sets of metadata that we generated for this

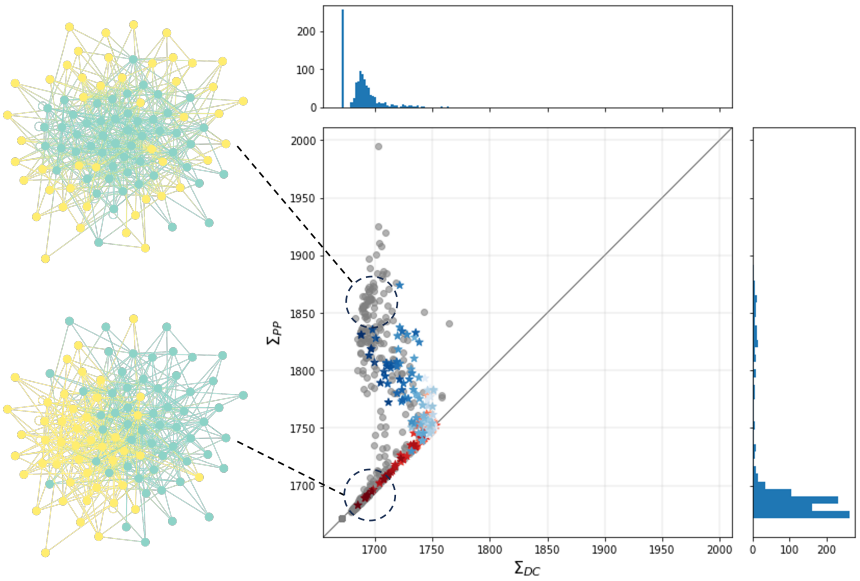


Figure 10: Description lengths under the DC and PP of the partitions sampled from DC (grey dots) and description lengths of the same network under DC and PP with metadata partitions (blue and red stars). The darker the colour of the stars representing the metadata partitions, the higher the correlation of the metadata labels with the respective planted structure. The network is visualised on the left-hand side, with nodes painted according to two of the inferred partitions, similar to the *planted* core-periphery (top) and bicommunity structures.

network, as introduced in the Results section and plot the resulting description length values alongside the grey dots (i.e. alongside the description length values of the *inferred* partitions) on Figure 10. The red stars represent the metadata partitions that are correlated with the planted bicommunity structure, the blue stars represent those that are similar to the core-periphery structure, with darker colours depicting higher values of ρ. As expected, we observe that under the more general of the two models, the partitions with the highest values of ρ have the lowest description length: the stronger the correlation of the metadata with the planted structure, the lower the description length, since the metadata partition is similar to the two ground truth partitions that were responsible for generating the network. The description lengths under the PP on the y-axis illustrate that additional to measuring the extent to which metadata is related to structure, we can also probe the type of structural arrangement: the bicommunity-like metadata are encoded as well under PP as under DC, indicating that assortativity was a prominent feature of the network generation process [38]. The core-periphery-like metadata, however, yield much higher description lengths under PP compared to DC, suggesting that – as we know is true – if we were to assume that this metadata was responsible for generating this network, assortativity was much less likely the prominent structure compared to some other more general structure.

Acknowledgements

This work was supported by the "Socsemics" Consolidator grant from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 772743).

Data and code availability A Python library to use the *metablox* measure can be downloaded and installed from `https://github.com/lenafm/metablox` The data needed to evaluate the conclusions in the paper are available from the corresponding author upon request.

References

[1] Francois Lorrain and Harrison C White. "Structural equivalence of individuals in social networks". In: *The Journal of mathematical sociology* 1.1 (1971), pp. 49–80.

[2] Harrison C White, Scott A Boorman, and Ronald L Breiger. "Social structure from multiple networks. I. Blockmodels of roles and positions". In: *American journal of sociology* 81.4 (1976), pp. 730–780.

[3] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. "Stochastic Blockmodels: First Steps". In: *Social Networks* 5.2 (June 1983), pp. 109–137. ISSN: 03788733. DOI: 10.1016/0378-8733(83)90021-7. (Visited on 09/11/2022).

[4] Brian Karrer and M. E. J. Newman. "Stochastic Blockmodels and Community Structure in Networks". In: *Physical Review E* 83.1 (Jan. 2011), p. 016107. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.83.016107. (Visited on 12/31/2021).

[5] Tiago P. Peixoto. "Nonparametric Bayesian Inference of the Microcanonical Stochastic Block Model". In: *Physical Review E* 95.1 (Jan. 2017), p. 012317. ISSN: 2470-0045, 2470-0053. DOI: 10.1103/PhysRevE.95.012317. (Visited on 02/21/2022).

[6] Tiago P Peixoto. "Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models". In: *Physical Review E* 89.1 (2014), p. 012804.

[7] Tiago P Peixoto. "Merge-split Markov chain Monte Carlo for community detection". In: *Physical Review E* 102.1 (2020), p. 012305.

[8] Mark EJ Newman and Aaron Clauset. "Structure and Inference in Annotated Networks". In: *Nature communications* 7.1 (2016), pp. 1–11.

[9] Darko Hric, Tiago P. Peixoto, and Santo Fortunato. "Network Structure, Metadata, and the Prediction of Missing Nodes and Annotations". In: *Physical Review X* 6.3 (Sept. 2016), p. 031038. ISSN: 2160-3308. DOI: 10.1103/PhysRevX.6.031038. (Visited on 03/06/2023).

[10] Leto Peel, Daniel B Larremore, and Aaron Clauset. "The Ground Truth about Metadata and Community Detection in Networks". In: *Science Advances* 3.5 (2017), e1602548. DOI: 10.1126/sciadv.1602548.

[11] Lada A Adamic and Natalie Glance. "The political blogosphere and the 2004 US election: divided they blog". In: *Proceedings of the 3rd international workshop on Link discovery*. 2005, pp. 36–43.

[12] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. "Political Polarization on Twitter". In: *Proceedings of the International Aaai Conference on Web and Social Media*. Vol. 5. 2011, pp. 89–96.

[13] Pablo Barberá, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. "Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?" In: *Psychological Science* 26.10 (Oct. 2015), pp. 1531–1542. ISSN: 0956-7976. DOI: 10.1177/0956797615594620. (Visited on 02/10/2022).

[14] Henry Small and Belver C Griffith. "The structure of scientific literatures I: Identifying and graphing specialties". In: *Science studies* 4.1 (1974), pp. 17–40.

[15] Wayne W. Zachary. "An Information Flow Model for Conflict and Fission in Small Groups". In: *Journal of Anthropological Research* 33.4 (Dec. 1977), pp. 452–473. ISSN: 0091-7710, 2153-3806. DOI: 10.1086/jar.33.4.3629752. (Visited on 08/16/2022).

[16] Amanda L Traud, Peter J Mucha, and Mason A Porter. "Social structure of facebook networks". In: *Physica A: Statistical Mechanics and its Applications* 391.16 (2012), pp. 4165–4180.

[17] Jaewon Yang and Jure Leskovec. "Defining and evaluating network communities based on ground-truth". In: *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*. 2012, pp. 1–8.

[18] Tanmoy Chakrabort, Sandipan Sikdar, Vihar Tammana, Niloy Ganguly, and Animesh Mukherjee. "Computer science fields as ground-truth communities: Their impact, rise and fall". In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2013, pp. 426–433.

[19] Jaewon Yang, Julian McAuley, and Jure Leskovec. "Community Detection in Networks with Node Attributes". In: *2013 IEEE 13th International Conference on Data Mining*. Dec. 2013, pp. 1151–1156. DOI: 10.1109/ICDM.2013.167.

[20] Cecile Bothorel, Juan David Cruz, Matteo Magnani, and Barbora Micenková. "Clustering Attributed Graphs: Models, Measures and Methods". In: *Network Science* 3.3 (Sept. 2015), pp. 408–444. ISSN: 2050-1242, 2050-1250. DOI: 10.1017/nws.2015.9. (Visited on 03/29/2023).

[21] Norbert Binkiewicz, Joshua T Vogelstein, and Karl Rohe. "Covariate-assisted spectral clustering". In: *Biometrika* 104.2 (2017), pp. 361–377.

[22] Martina Contisciani, Eleanor A Power, and Caterina De Bacco. "Community detection with node attributes in multilayer networks". In: *Scientific reports* 10.1 (2020), p. 15736.

[23] Oscar Fajardo-Fontiveros, Roger Guimerà, and Marta Sales-Pardo. "Node metadata can produce predictability crossovers in network inference problems". In: *Physical Review X* 12.1 (2022), p. 011010.

[24] Darko Hric, Richard K Darst, and Santo Fortunato. "Community detection in networks: Structural communities versus ground truth". In: *Physical Review E* 90.6 (2014), p. 062805.

[25] Tracy M Sweet and Qiwen Zheng. "Estimating the effects of network covariates on subgroup insularity with a hierarchical mixed membership stochastic blockmodel". In: *Social Networks* 52 (2018), pp. 100–114.

[26] Tiago P. Peixoto. "The graph-tool python library". In: *figshare* (2014). DOI: 10.6084/m9.figshare.1164194. URL: http://figshare.com/articles/graph_tool/1164194 (visited on 09/10/2014).

[27] Tiago P. Peixoto. "Revealing Consensus and Dissensus between Network Partitions". In: *Physical Review X* 11.2 (Apr. 2021), p. 021003. ISSN: 2160-3308. DOI: 10.1103/PhysRevX.11.021003. (Visited on 02/10/2022).

[28] Lena Mangold and Camille Roth. "Generative Models for Two-Ground-Truth Partitions in Networks". In: *Physical Review E* 108.5 (Nov. 2023), p. 054308. DOI: 10.1103/PhysRevE.108.054308. (Visited on 11/16/2023).

[29] Natalie Stanley, Marc Niethammer, and Peter J. Mucha. *Testing Alignment of Node Attributes with Network Structure Through Label Propagation*. 2018. arXiv: 1805.07375 [cs.SI].
[30] Peter D Grünwald. *The minimum description length principle*. MIT press, 2007.
[31] William M Rand. "Objective criteria for the evaluation of clustering methods". In: *Journal of the American Statistical association* 66.336 (1971), pp. 846–850.
[32] Marina Meilă. "Comparing clusterings by the variation of information". In: *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*. Springer. 2003, pp. 173–187.
[33] Tiago P. Peixoto. "Entropy of Stochastic Blockmodel Ensembles". In: *Physical Review E* 85.5 (May 2012), p. 056122. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.85.056122. (Visited on 02/14/2023).
[34] Ginestra Bianconi. "Entropy of Network Ensembles". In: *Physical Review E* 79.3 (Mar. 2009), p. 036114. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.79.036114. (Visited on 02/20/2023).
[35] Hirotugu Akaike. "A new look at the statistical model identification". In: *IEEE transactions on automatic control* 19.6 (1974), pp. 716–723.
[36] Gideon Schwarz. "Estimating the dimension of a model". In: *The annals of statistics* (1978), pp. 461–464.
[37] Xiaoran Yan, Cosma Shalizi, Jacob E Jensen, Florent Krzakala, Cristopher Moore, Lenka Zdeborová, Pan Zhang, and Yaojia Zhu. "Model selection for degree-corrected block models". In: *Journal of Statistical Mechanics: Theory and Experiment* 2014.5 (2014), P05007.
[38] Lizhi Zhang and Tiago P. Peixoto. "Statistical Inference of Assortative Community Structures". In: *Physical Review Research* 2.4 (Nov. 2020), p. 043271. ISSN: 2643-1564. DOI: 10.1103/PhysRevResearch.2.043271. (Visited on 09/30/2022).
[39] Anne Condon and Richard M. Karp. "Algorithms for Graph Partitioning on the Planted Partition Model". In: *Random Structures & Algorithms* 18.2 (2001), pp. 116–140. ISSN: 1098-2418. DOI: 10.1002/1098-2418(200103)18:2<116::AID-RSA1001>3.0.CO;2-2. (Visited on 08/15/2022).
[40] Emmanuel Lazega. *The collegial phenomenon: The social mechanisms of cooperation among peers in a corporate law partnership*. Oxford University Press, USA, 2001.
[41] Eve Chiapello and Lisa Knoll. "Social finance and impact investing. Governing welfare in the era of financialization". In: *Historical Social Research/Historische Sozialforschung* 45.3 (2020), pp. 7–30.
[42] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. "Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship". In: *Proceedings of the 2018 world wide web conference*. 2018, pp. 913–922.

[43] Marilena Hohmann, Karel Devriendt, and Michele Coscia. "Quantifying Ideological Polarization on a Network Using Generalized Euclidean Distance". In: *Science Advances* 9.9 (Mar. 2023), eabq2044. DOI: 10.1126/sciadv.abq2044. (Visited on 03/03/2023).

[44] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. "The Echo Chamber Effect on Social Media". In: *Proceedings of the National Academy of Sciences* 118.9 (Mar. 2021), e2023301118. DOI: 10.1073/pnas.2023301118. (Visited on 03/03/2023).

[45] Tiago P. Peixoto. "Parsimonious Module Inference in Large Networks". In: *Physical review letters* 110.14 (2013), p. 148701.

[46] Jean-Gabriel Young, Guillaume St-Onge, Patrick Desrosiers, and Louis J. Dubé. "Universality of the Stochastic Block Model". In: *Physical Review E* 98.3 (Sept. 2018), p. 032309. ISSN: 2470-0045, 2470-0053. DOI: 10.1103/PhysRevE.98.032309. (Visited on 05/02/2022).

[47] Brian Karrer, Elizaveta Levina, and Mark EJ Newman. "Robustness of Community Structure in Networks". In: *Physical review E* 77.4 (2008), p. 046119.