Models for Dynamic Networks with Metadata

John Fitzgerald

Mansfield College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2022

Acknowledgements

While this thesis has been my own labour of love, it would not have been possible without the support of a great number of other people.

First and foremost, I would like to thank my supervisors, Pete Grindrod and Neave O'Clery, who have stuck with me through the many twists and turns on which this project has taken me. Their patience and encouragement for me to explore my own areas of interest – and subsequent aid condensing my disparate thoughts down into more manageable (and realistic!) goals – has been crucial in ensuring that this is the complete work that it is, rather than the beginnings of 10 different possible PhD projects. The wider research group of Neave has also been both the inspiration of some good ideas and their interconnections, and the origin of some good friendships — thank you to Mattie and Nils in particular for their help, and their happiness to bounce ideas over this final year.

Another key contact for the development of the technical aspects of this project has been the brilliant Leto Peel. Leto joined as a collaborator just as I was consolidating my efforts into developing my own initial models, and determining how they might be most useful. His insights and questions have been vital in prompting me to explore new, related ideas, and directing me towards prior efforts in the surrounding literature — his assistance provided an inflection point in my work, and without it this thesis would both be much lesser, and the entire process much more arduous.

This doctoral research is based on work supported by the EPSRC Centre for Doctoral Training in Industrially Focused Mathematical Modelling (EP/L015803/1) in collaboration with Elsevier. While the pandemic somewhat threw a spanner in the works and prevented me from visiting them in person as much as I would have liked, my industrial supervisors there – Rachel Herbert and Andrew Plume – have both been fantastic throughout our time together. Our regular videocalls were always a pleasure, and prompted me to regularly consider the real-world applications of my

ideas, without overly constraining my overall research pathway. This work uses Scopus© data provided by Elsevier through the ICSR Lab, which was the basis of most of the empirical networks to which my methods are applied throughout.

Beyond those providing direct intellectual aid for the project, my friends and family have been the crutch without which I would no longer be standing. To my childhood friends, David, Rob, James, Calum, and Jamie, thank you for always providing an escape from academia, and a reminder that not everything has to be so serious all the time. To Alex and Heledd, my thanks for always driving me to deepen my interests and wider understanding outside of mathematics, and the generous pints along the way. Lydia and Aaron, I'm indebted to you for keeping me fed as well as sane, and Yuying and Leon you have my endless gratitude for always bringing my energy up when I most needed it. I'm thankful for Dylan not just for providing a desk at UCL, but for commiseration about PhD life, and a patient ear for me to talk maths with, even when neither of us had much of a clue what the other was talking about. To Chevonne, Aihem, and Omar, my appreciation for somehow making the start of the pandemic not just bearable, but actually a surprisingly pleasant time at points with all else considered. Michael and Giuseppe, you've always helped me to reevaluate what I actually believe is important — and I hope we have more breaks away in future! To my family, I know this journey hasn't always been as smooth as you might have wished, but I'm grateful for your love and patience, and the numerous care packages that have sustained me through some tougher periods. For all of the above, thank you for putting up with me venting practically the same speech most times we've met for the last few years — I hope in future to be a better conversation partner, and a closer friend to you all.

Finally, I would like to thank my partner, Carina. It's hard to put into words all the ways you've helped me over the years, not only materially – though the cups of tea and coffee are countless to say the least – but to grow and learn as an individual. Your love has been the part of my life that I am most grateful for since starting this DPhil in the first place, and the part of the future that I most look forward to. Through all the ups and downs, you (and now Panza too!) have brought me so much joy, and I truly don't want to imagine what it would have been like without you.

Abstract

There is increasing understanding that many complex systems of interest – everything from the global economy, to social group dynamics, to biochemical processes in the brain – require holistic modelling, rather than the consideration of units of the population in isolation. Network science techniques, which commence by viewing the system as a set of vertices or nodes (the units of the population, *e.g.* individual people) and edges (the relationships between them, *e.g.* friendship), are one popular approach to do so.

Naturally, such complex systems express a wide array of important properties that we ought to account for when modelling them, beyond simply the presence or absence of a particular relationship. Most pertinently for this work, they evolve over time – *i.e.* they are *dynamic* – and the units of the population may have distinct properties, or attributes, which further differentiate them from each other. We define any such extra information we might possess outside of the simple node/edge paradigm to be *metadata*.

Despite the potential utility of such metadata, it is only quite recently that methods have begun to jointly model both network and metadata together. In this thesis, we provide a new class of models that do so — specifically, with the purpose of finding groups in networks that change over time. We describe distinct versions of this class of models that allow the networks to be weighted and directed, as well as avoid the potential issue of placing nodes with similar degrees in the same group.

In addition to elaborating such models, we derive novel requirements for the efficient detectability of groups given the presence of metadata — and in the process explain why a recent paper which claims to do the same for a similar static model is flawed. The inference method we leverage to investigate detectability is also highly scalable, and we further accelerate the process by proposing both a 'greedy' scheme, and a recursive procedure that effectively provides a top-down hierarchy of the network groups.

We conclude by using our models as one component of a larger method, that provides an entirely novel means of estimating the influence of an author. We use a causal framing of the problem that to our knowledge has not previously been explored in this context, and depends upon recent ideas from the causal inference literature.

Contents

List of Figures

Chapter 1

Introduction

With the exception of the recent pandemic, for decades it has seemed that every passing year has resulted in intensified connections between all parts of the globe — even during the mandatory spatial separation of lockdowns, the flow of information and interpersonal communication, that helped ameliorate feelings of isolation for so many, continued uninterrupted. Indeed, the deepening integration and inter-relatedness of all the constituent networks (*e.g.* social, knowledge, and economic) that make up our wider society appear undeniable. There is increasing understanding that such complex systems require holistic modelling, rather than the consideration of units of the population in isolation – network science techniques, which account for the relations between units, are one popular approach to do so. As the problems considered are wide-ranging, so too are the backgrounds of those investigating them. This interdisciplinarity has been one of the field's greatest strengths, with key insights being motivated by, and drawing on, subjects from neuroscience and biology to statistical physics and mathematics.

In this thesis, we consider networks to be any formalisation of a system into a set of units (nodes or vertices) with relations between them (edges). This general formulation permits many different setups, each of which may require distinct methods to handle them suitably. For instance, the nodes may be of different classes [124], and the edges may be weighted [116], define separate types of relationship [170], change over time [98, 29], or link more than two nodes [9], to name just several options. Most pertinently for this work, both nodes and edges may also contain additional information about themselves, beyond this vertex/edge paradigm. We define such extra information to be *metadata*.

Our primary empirical focus when applying methods within this work are networks constructed from publication data, as we return to in Sec. 1.2 below. Hence, for a

relevant example consider a co-authorship network, where the nodes are authors, with edges connecting pairs (or groups) of authors who have produced a publication together. Node-wise metadata in such a network could include author profiles – affiliation, duration of publishing *etc.* – while edge-wise metadata might be information about the corresponding publications produced – the topic area, date of publication and so on.

Despite the potential utility of such metadata, and the relatively long history of applying network science approaches to real data – including specifically to academic networks [109] – it is only quite recently that methods have begun to jointly model both network and metadata together [113]. The primary aim of this thesis is to provide a new class of models that do so — specifically, with the purpose of finding groups, or meso-scale structure in networks that change over time.

1.1 Finding meso-scale structure in networks

When considering real-world networks, it is frequently observed that the nodes may intuitively be categorised into similar classes, or groups. For instance, these 'categories' might be friendship groups or online forums in social networks, who are more strongly connected to each other than to the rest of the network — this type of assortative grouping is normally referred to in the literature at a community, and often appears to be a useful way of coarse-graining, and gaining greater insight into innumerable complex systems, ranging from economies to biochemical processes [83, 121, 35]. However, once again our formulation is quite general. Rather than solely being defined by a greater intensity of connectivity, one could seek *e.g.* groups between which the connectivity was similar, *i.e.* where each group plays a distinct functional role in the wider network — this is the focus of the increasingly popular class of stochastic block models (SBMs) [139], where such groups are the 'blocks' of the system. This functional role framing generalises several other common types of network structure, like core-periphery or hierarchical structure, and indeed there are SBMs that explicitly seek such if desired [47, 136]. Beyond the network, the metadata itself may immediately define categories, like year cohorts in networks of students, departments in a workplace, or origin locations of genomes [132].

As such groupings of nodes fall between the micro-scale – the nodes themselves and their immediate neighbourhoods – and the macro-scale – the network itself – we consider them to be *meso-scale* structure. Often, the emphasis has been placed on this structure providing an efficient representation of the network itself [141, 144], but

we are predominantly concerned with simultaneously ensuring the interpretability and utility of groups. For instance, if we discard metadata about author affiliation and consider a global co-authorship network, then we might end up with groups that best represent research communities as a whole, but are less useful to policy-makers than those found emphasising author affiliation, and thus rooting communities more in place. Incorporating such metadata directly into the model also allows us to quantify the influence of this information on the groups inferred, and hence the relative importance for the network structure [62]. This being said, as we demonstrate later in this thesis both empirically and theoretically, interpretability and structural quality of representation are not mutually exclusive — the inclusion of metadata typically simultaneously improves both.

As one might expect from the numerous ways of defining them in the first place, there are naturally a multitude of methods to find such groups. In a general sense, these reduce to (i) a choice of 'quality' function to evaluate a labelling of the nodes, and (ii) a choice of optimisation procedure for this function. From these two simple steps, there emerge methods based on non-negative matrix factorisation (NMF) [178], dynamics or diffusion on the network [82], and many more. In this work, we take the approach of positing a generative model for the network — that is, a joint distribution over the network and the groups. Once this model is defined, we describe several different options for inferring groups and model parameters — specifically, we elaborate options for 'greedy' (or local) inference [69, 152], mean-field variational inference (MFVI) [98], and belief propagation [34, 50].

1.2 Networks constructed from publication data

As suggested above, when applying our novel algorithms, our main focus is to do so on networks constructed from publication data. This is motivated both by our collaboration with the publishing house Elsevier, who have provided full access to their publication databases – Scopus, ScienceDirect, and SciVal – and due to their increased importance, particularly as many countries are transitioning away from economies based upon physical or natural resources towards knowledge intensive activities, *i.e.* a 'knowledge economy' [148].

Indeed, much recent work in economic complexity and economic geography has emphasised the importance of locally available skills and capacities for the growth of cities and regions [59, 46, 58, 55, 120]. Fundamentally, they recognise that much knowledge and many abilities are tacit (*i.e.* learned through experience), and so

diffusion of these capabilities is difficult, resulting in geographic 'stickiness'. As such, better understanding the dynamics of research and researchers themselves is crucial — who publishes what, with whom, in what topic, and why? Without this knowledge, the ability to craft effective policies for economic and technological development is greatly weakened.

As a result, there have understandably been a multitude of papers investigating the 'Science of Science' — particularly through the lens of academic co-authorship or citations, both in network form or through detailed surveys. These range from focused studies on particular topics, such as how individual collaboration patterns change with academic career age (further suggesting this should be an important factor to include in any model, see *e.g.* [179]), to analyses of factors driving collaboration at the regional level, such as investment in R&D [4]. Fewer studies have holistically modelled the evolution of the network, while detecting blocks of similarly connected academics, and incorporating additional information such as the affiliation, research topics, and career age of each author in the process – we aim to contribute to this literature.

Several papers have looked at the importance of institutional affiliation in academic networks, *e.g.* in terms of how collaborations within or between institutions affect the subsequent impact of resulting publications, the significance of institutional prestige, and indeed their utility in predicting future collaborations [147, 181, 173]. One interesting work used topic models based on SBMs to investigate affiliations and research topics to publications in computer science, specifically to try and understand the research focus of different institutions, but do not further investigate the relationship to co-authorship, nor what might be causing observed changes [42, 49].

However, there are a variety of papers effectively regressing links (typically co-authorship or citations) against topic similarity, simultaneously accounting for such links when modelling topics – these are broadly here labelled Relational Topic Models (RTMs), as introduced in [25]. The coupled nature of such models means that as well as using topics to inform the likelihood of observing a link, the observed links influence the topics themselves – effectively assuming such links confer semantic information about the content of the documents alongside the text itself. These models typically do not simultaneously model structure in the network along with topics, with minor exceptions such as [91], which had simultaneous community detection and topic modelling, effectively through combining the RTM with the mixed-membership SBM of [7], but only for static networks. Another more recent paper [56] claims to model community structure and topics simultaneously in dynamic text networks, but their formulation of 'communities' is heavily restrictive (and thus not particularly

structurally informative), they make numerous parameter choices without further justification, and there are a variety of other issues. One particular matter of note is that both of these models assume blocks are assortative, in that within blocks there should be dense connections, and similar documents. This may prevent the discovery of (i) nodes that are similarly connected to the rest of the network, if not each other (*e.g.* peripheral nodes), and/or (ii) groups of authors collaborating on interdisciplinary work (or alternatively the model may produce extraneous 'interdisciplinary' topics rather than identifying what particular mix of topics actually makes up a publication).

While such models have been developed in a multitude of directions, and each new model evaluated against real datasets – typically in terms of their predictive capabilities – they are (i) rarely used to any practical effect, (ii) are designed primarily for document networks (rather than author networks), (iii) typically do not permit other types of metadata, and (iv) often face other issues such as scalability. Within this work, we do not perform simultaneous representation of topics (as 'bags-of-words') and network structure, but we do consider finely-grained pre-computed topics as one type of input metadata — and within a model that completes in time that scales near-linearly with the size of the network. This provides us with groups directly influenced by the distribution of topics contained within, which we believe may be more helpful for understanding research communities, and the inter-relatedness of topics, than purely focusing on topics alone — extension to directly model topics is quite immediate (*e.g.* through the bipartite formulation of [49], or in an iterative process), and would be an interesting direction to pursue in future.

1.3 Papers and software

Papers At the time of submission, one paper has been published, based on exploratory work early in the course of the project, using the same primary database considered herein. Another short paper, that uses a form of the dynamic stochastic block model proposed in this thesis to simulate synthetic networks, was published in NeurIPS workshop proceedings:

- J.A. Fitzgerald, S. Ojanperä, and N. O'Clery. Is academia becoming more localised? The growth of regional knowledge networks within international research collaboration. *Applied Network Science*, 6(1):1–27, 2021 [43];

- J. Dyer, J.A. Fitzgerald, B. Rieck, and S.M. Schmon. Approximate Bayesian Computation for Panel Data with Signature Maximum Mean Discrepancies *NeurIPS 2022 Temporal Graph Learning Workshop*, 2022 [40].

We intend to submit the remaining work contained in this thesis as several papers early in the coming year, in preparation with Leto Peel alongside the supervision team. The primary three papers will be made up of (i) parts of Chap. 3, to introduce the dynamic stochastic block model with metadata (DSBMM), to our knowledge the first SBM to handle generic nodal metadata alongside group dynamics; (ii) Chap. 4, to describe a novel way to understand group detectability in static networks with metadata, and a rebuttal to a previous attempt to do so, accompanied by the extension of the belief propagation inference procedure to the dynamic case, and associated simulated experiments of Chap. 5; and (iii) Chap. 6, where we propose a causal model – and supporting inference procedure – that may be used to estimate author influence in a hitherto unexplored way, alongside the necessary hierarchical, directed, and degree-corrected developments of the DSBMM to do so.

Software

- dynsbmmeta: Dynamic stochastic block model for temporal networks with nodal metadata. R and C++ implementation, with some supporting Python for analysis of results. Available on GitHub; substantially developed after a fork of a previous related model [98].
- dsbmm-bp: Model dynamic networks with metadata using dynamic SBMs, with parameters inferred through belief propagation. Pure Python/numpy implementation, using numba at points for additional acceleration. Available on GitHub; entirely my own work.
- PIF-DSBMM-DPF: Dynamic Poisson Influence Factorisation, using DSBMM and dPF to construct substitutes. Python implementation, calling C++ as subprocesses. Available on GitHub; substantially developed after a fork of a previous related model [166].

1.4 Organisation of the thesis

The remainder of this work is organised as follows. First, in Chap. 2, we introduce some of the prerequisite theory and concepts for network analysis — in particular, we cover the basic formulation of SBMs.

Next, in Chap. 3, we present the basis case of our generative model for dynamic networks with metadata – the Dynamic SBM for networks with Metadata (DSBMM), the first of its kind – along with an extension that allows degree-correction (*i.e.* avoids clustering the nodes by degree), and a mean-field variational inference method for the model parameters. To better elucidate its strengths and weaknesses, we evaluate the performance of this model on simulated networks, and consider insights it can provide on several different real networks.

From this foundation, in Chap. 4 we then introduce the background to belief propagation (BP) inference, and how this can be used to find groups in a static version of our model. We then simplify this model into a toy example, which allows us to utilise the BP message equations to investigate parameter-dependent thresholds for efficient detectability of groups — *i.e.* under what conditions we can expect our model to perform well. In the process, we also explain how a recent paper which claims to do the same for a similar model is incorrect.

Following on from the static case, in Chap. 5 we describe how to perform BP inference for the full DSBMM, including fitting model parameters, and once more evaluate this procedure on both simulated and real networks — now significantly larger than previously possible, due to the greatly improved scaling permitted by BP. We also introduce two ways of further accelerating inference of the model — a greedy method, and a top-down (recursive) hierarchical procedure, which we use for the empirical results.

Finally, having both permitted scalable inference, and demonstrated the effectiveness of the incorporating metadata, in Chap. 6 we elaborate how the DSBMM can be used in conjunction with other models to provide an entirely novel means of estimating the influence of an author. Specifically, we use a causal framing of the problem that to our knowledge has not previously been explored in this context, and depends upon recent ideas from the causal inference literature. The data considered also requires further extensions of the DSBMM: a directed (and optionally degree-corrected) version, and when applying the hierarchical procedure, a method to obtain global model parameters for the groups at the finest level.

Chapter 2

Preliminaries

2.1 Networks 101

As described in Chap. 1, the fundamental quantities of a network are the entities under consideration, and the relationships between them. Within this work, this is considered as a graph, $G = (\mathcal{V}, \mathcal{E})$, where the entities correspond to the set of nodes (or vertices), V, and the relationships are represented by the edges, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, where $e = (i, j) \in \mathcal{E}$ denotes a relationship from node i to node j. In this case, i and j are said to be adjacent, or neighbours. Note edges (i, i) that connect a node to itself are typically referred to as self-loops, and are frequently disallowed/removed to simplify analysis. Throughout this work we denote the number of nodes, $|\mathcal{V}|$, by N, and the number of edges, $|\mathcal{E}|$ by E.

A network may be directed or undirected: undirected if for every edge $(i, j) \in \mathcal{E}$, the reciprocal edge (j, i) is also in $\mathcal{E}$, and directed otherwise. Which of the two to use is typically motivated by the type of relationship considered — for instance, co-authorship is (theoretically) naturally a reciprocal relationship, where $i \to j$ implies $j \to i$, while citation is not, as most authors do not ensure that they cite everyone who cites them. The edges may also have some scalar value associated to them, say w_{ij} signifying *e.g.* the strength of the relationship, in which case the network is considered as weighted – unweighted networks are simply the case where all weights are binary. Within this work, most frequently we take the weight of an edge to be the count of the number of times that relationship has occurred — *e.g.* the number of publications a pair of authors have produced together, or number of times that one has cited the other. A network can then be represented by an adjacency matrix, $A = (A_{ij}) \in \mathbb{R}^{N \times N}$,

where for undirected networks

$$A_{ij} = \begin{cases} w_{ij} & \text{if } (i,j) \in \mathcal{E},\ i \neq j, \\ 2w_{ii} & \text{if } (i,i) \in \mathcal{E}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

If the network is directed, the factor of two in front of w_{ii} can be removed – it is included so that sums of edges for undirected networks are consistent (as then effectively each non self-loop edge is included twice in A). Clearly, A must be symmetric for a network to be undirected, and it is directed if this is not the case.

Common quantities of interest are then the row and column sums of this matrix. If the network is undirected, these are the same, and are referred to as either the degree or strength of the nodes, depending on whether the network is unweighted or weighted respectively – we denote these by the vector $\boldsymbol{d} \in \mathbb{R}^N$, with

$$d_i = \sum_j A_{ij}. \tag{2.2}$$

For co-authorship networks, the unweighted degree of a node corresponds to the total number of collaborators, while the weighted strength could correspond to *e.g.* the total number of publications. As the adjacency matrix of a directed network is no longer symmetric, the sums are no longer equivalent, in which case the column sums correspond to the in-degree/strength, $\boldsymbol{d}_{in}$, and the row sums correspond to the out-degree/strength $\boldsymbol{d}_{out}$, *i.e.*

$$d_i^{\text{in}} = \sum_j A_{ji}, \qquad d_i^{\text{out}} = \sum_j A_{ij}. \tag{2.3}$$

For an unweighted network, the $(i,j)th$ element of the matrix A^p counts the number of walks length p from i to j, where such a walk is defined as a sequence of nodes $\{v_k\}$, with $v_0 = i$ and $v_p = j$, and $(v_k, v_{k+1}) \in \mathcal{E}\ \forall k$. A directed network is then said to be strongly connected if there is a walk between any pair of vertices in the graph, such that $\forall i,\ j,\ \exists p$ such that $A^p_{ij} \neq 0$ – if this is not the case, but the condition holds when edges are made undirected, then it is said to be weakly connected, and otherwise it is disconnected. We will later often restrict ourselves to connected components when considering an empirical dataset.

Beyond the categories described above, real networks exhibit a variety of additional features, primarily that the the same set of vertices may have a multitude of different types of relationship between them (*e.g.* for academic institutions, collaborations in

different disciplines and languages, academic mobility *etc.*), that both nodes and edges may possess additional metadata, and that the network may evolve over time. Within this work, we focus on permitting node-wise metadata, and discrete-time network evolution, where the network is aggregated into 'snapshots' at each timestep — for a more general overview of temporal networks, see *e.g.* [86]. The extensions of such to handle edge categories (*e.g.* via a multilayer framing [77, 22]) or node categories (as a bi-/tri-/...-partite network [185, 124]) are fairly immediate. We now proceed to describe one of the primary notions of structure in networks, which builds from the basic ideas of this section.

2.2 Meso-scale structure

A key concept to introduce for our work is that of meso-scale structure — in broadest terms these are groups of nodes in a network, and so are typically specified by some partition of the graph (though overlapping groups may also be considered [52, 138]). There is no universal agreement as to how such groups should be defined, and the desired shared functional or structural properties often vary depending on the problem at hand. Nonetheless, as previously suggested finding such groups can be useful for a variety of applied problems, and can reveal structural information about the network — hence they are highly popular across a wide array of fields.

Often, this structure is assumed to be constituted of assortative groups of nodes, or *communities*, within which there are many links when compared to links between groups [111] – this view is related to an alternative, dynamic perspective which views communities as regions where random walkers over the graph are more likely to be trapped for some time, which solves some of the issues with classical techniques [85]. However, both of these conceptions of communities neglect groupings which are dissortative, *i.e.* where there may be preference for nodes to connect to others *unlike* themselves, or have more complex patterns of connectivity *e.g.* hierarchy or core-periphery structure. To allow such occurrences, we may instead seek groups of nodes which are stochastically similar, *i.e.* connect to nodes in other groups of the network with the same probability – this is the basis of stochastic block models (SBMs).

In these preliminaries, we present only the classic SBM, and a degree-corrected version (the DC-SBM), as it is upon these as a basis that we construct our own models. We refer the reader to *e.g.* [66] for an overview of alternative methods. In particular, we note that within this work we do not explicitly consider overlapping communities,

in which nodes may be members of multiple groups – see *e.g.* [7, 138]. This being said, the belief propagation inference procedure we utilise from Chap. 4 onwards does provide estimates of the marginal probability of a node belonging to each group, which may be interpreted in a similar manner to these mixed-membership methods to some degree.

2.2.1 Stochastic block models

A principal avenue of investigation for network science is the construction of generative probabilistic models that can describe real data well — *i.e.* generate 'realistic' networks after inferring suitable parameters, where this is often determined via posterior predictive checks, as we describe and perform in Chap. 6. One of the most popular branches of this pursuit are exponential random graph models (ERGMs, see *e.g.* [155]), which model the adjacency matrix of the graph through the likelihood (*i.e.* probability distribution)

$$\mathrm{p}(A \mid \boldsymbol{\theta}) = \frac{1}{Z} \exp(\boldsymbol{\theta}^\top \boldsymbol{t}(A)), \tag{2.4}$$

for some set of parameters $\boldsymbol{\theta}$, and network configurations present in A, $\boldsymbol{t}(A)$. The denominator Z normalises the function so that the sum over all possible graphs comes to one, and it defines a suitable likelihood. Here, as throughout this work, we use $\mathrm{p}(\cdot)$ to denote a probability distribution, rather than *e.g.* model parameters, or factors that make up a distribution.

Examples of network configurations could be the presence of an edge (i, j), a triangle (i, j, k), a link between two nodes in different groups *etc.* — to reduce the number of parameters, typically the parameters are assumed to be homogeneous for a given class of configurations. For instance, if the parameters governing the likelihood of each edge (assumed binary) are taken to be the same and we neglect to model other configurations – effectively the classical Erdös-Rényi model [41] – then

$$\mathrm{p}(A \mid \theta) = \frac{1}{Z} \prod_{(i,j)\in\mathcal{E}} \exp(\theta) = \frac{1}{Z} \exp(E\theta), \tag{2.5}$$

where $E = |\mathcal{E}|$. The popularity of this family then largely stems from the relatively easy interpretation of inferred models, and the apparent simplicity of extending the model to incorporate other features (*e.g.* triangles or other subgraphs). In particular, they are frequently used in modelling social networks, as if the parameter multiplying a particular type of configuration is found to be positive (or negative) at some confidence level, it suggests that that configuration occurs relatively more often (respectively less

often) than expected at random (if the parameter were zero). This can thus allow testing of hypotheses for the underlying processes of link formation, with estimates of uncertainty, and (provided the inferred model reasonably represents the observed network) a way to simulate comparable synthetic datasets with similar structure — which itself has been particularly of interest of late in the field of differential privacy [70]. The basis of the models that we propose later in this work, the stochastic block model (SBM), is part of this family.

In addition to the general benefits of using a generative network model, SBMs are a natural approach for modelling academic networks. As collaboration is often driven by face-to-face meetings, institutions that are closer to each other, are specialised in certain areas, and/or between which academics often move (typically stratified by the rank of the institution [36]) are more likely to collaborate more frequently. Thus we would expect the co-authorship network to be made up of some combination of groups of primarily locally collaborating institutions, perhaps along with other institutions playing a different structural role of connecting to more distant institutions. Likewise, citations are typically strongly grouped into research areas, often with more edges between authors who know – or have at least come into contact with – each other in person [163, 8]. As such, SBMs are ideally placed to determine *e.g.* what role an institution plays in the network, and which other institutions it is most closely related to, or which research community an author belongs to, and subsequently what papers or prospective collaborators might be relevant to them.

There are a multitude of variants of SBMs, as we discuss (and use) throughout the course of this work. The simplest is defined for undirected, unweighted graphs with no self-loops, into a fixed number of groups, Q, by the likelihood

$$\mathrm{p}(A \mid \boldsymbol{\omega}, Z) = \prod_{i<j} \omega_{z_i,z_j}^{A_{ij}} (1 - \omega_{z_i,z_j})^{1-A_{ij}}, \tag{2.6}$$

where $\boldsymbol{\omega} \in [0,1]^{Q\times Q}$ is the matrix with entries ω_{qr} providing the probability of an edge existing between nodes of groups q and r [60], and $Z = \{z_1, \ldots, z_N\}$ denotes the set of group labels, where $z_i = q \in \{1, \ldots, Q\}$ if node i belongs to group q. In other words, each edge is generated by a Bernoulli distribution with probability ω_{z_i,z_j}. In subsequent chapters, when it is more convenient to work with binary indicators of the node label, we use a one-hot encoding of these groups — *i.e.* $\boldsymbol{z}_i \in \{0,1\}^Q$, where $z_{iq} = 1$ if $z_i = q$, and $z_{ir} = 0$ else.

To posit a fully generative model, we must place prior distributions over the model parameters, $\boldsymbol{\omega}$ and Z. In this work, as is commonly performed (see *e.g.* [6]),

for mathematical convenience we assume conjugate priors, for which the posterior distribution over parameters $\boldsymbol{\theta}$ given the data, $\mathcal{D}$,

$$\mathrm{p}(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{\mathrm{p}(\mathcal{D} \mid \boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta})}{\mathrm{p}(\mathcal{D})}, \tag{2.7}$$

by Bayes theorem, remains in the same family of distributions as the prior, $\mathrm{p}(\boldsymbol{\theta})$. We note however, that more complex, structured priors are also popular in the literature [136]. For an exponential family model, like the SBM – where $\boldsymbol{\theta} = \{\boldsymbol{\omega}, Z\}$, and the sufficient statistics are $\boldsymbol{t}(A) = \{m_{qr}, n_q\}$ for m_{qr} the counts of edges between groups q and r, and n_q the number of nodes in q – there is a conjugate prior for $\boldsymbol{\omega}$ in the form

$$\mathrm{p}(\boldsymbol{\omega} \mid \boldsymbol{\chi}) = \frac{g(\boldsymbol{\omega}) \exp(\boldsymbol{\omega}^\top \boldsymbol{\chi})}{Z(\boldsymbol{\chi})}, \tag{2.8}$$

for some hyperparameters $\boldsymbol{\chi}$, for which the resulting posterior given the data simply corresponds to $\mathrm{p}(\boldsymbol{\omega} \mid \boldsymbol{\chi} + \boldsymbol{t}(A))$. As such, we may consider these pseudo-observations — initialise to whatever value, then sequentially update using the observed edge counts between each pair of groups inferred at that iteration of parameters. For the group labels, Z, we use a simple multinomial prior, such that each node independently has $\mathrm{p}(z_i = q) = \alpha_q$, where $\sum_q \alpha_q = 1$. This means that we can now describe a full joint distribution, neglecting the prior for the pseudo-observations $\boldsymbol{\omega}$, as

$$\mathrm{p}(A, Z \mid \alpha, \boldsymbol{\omega}) = \prod_{i,q} \alpha_q^{z_{iq}} \prod_{i<j} \omega_{z_i,z_j}^{A_{ij}} (1 - \omega_{z_i,z_j})^{1-A_{ij}}. \tag{2.9}$$

However, optimising this likelihood typically performs poorly in networks with large degree variability, as the expected degree for all nodes within each group under the model is the same — this means that fitting the model generally clusters node by degree to some degree.

Accordingly, there is a degree-corrected SBM (DC-SBM) variant, which introduces node 'popularity' parameters, θ_i, to address this problem [69]. The corresponding likelihood (for undirected graphs with no self-loops, but now allowing multiple edges between nodes, so $A_{ij} \in \mathbb{N}$) is

$$\begin{aligned} \mathrm{p}(A \mid \boldsymbol{\lambda}, \boldsymbol{\theta}, Z) &= \prod_{i<j} \mathrm{Pois}(\theta_i \theta_j \lambda_{z_i,z_j}), \\ &= \prod_{i<j} \frac{\mathrm{e}^{-\theta_i\theta_j\lambda_{z_i,z_j}} (\theta_i\theta_j\lambda_{z_iz_j})^{A_{ij}}}{A_{ij}!}, \end{aligned} \tag{2.10}$$

i.e. the likelihood of each edge is modelled as a Poisson distribution with parameter $\kappa_i \kappa_j \lambda_{z_i,z_j}$. Here λ_{qr} influences the overall number of edges between groups q and q much as before, but each node is permitted some additional parameter θ_i (fixed only up to some multiplicative constant that could be absorbed into λ_{rs}) affecting the number of edges they receive. For instance, in the original paper, the authors constrain $\boldsymbol{\theta}$ such that

$$\sum_i \theta_i \delta_{z_i,r} = 1 \tag{2.11}$$

for all blocks r. In this case, θ_i becomes the probability that an edge connected to block z_i connects to i itself, and λ_{qr} is the expected number of edges between groups q and r. As this model decouples the degrees from the group memberships, arbitrary degree variability inside the modules is now allowed. The Poisson distribution is used to simplify inference, and in sparse networks where on average $\theta_i \theta_j \lambda_{z_i,z_j} := \mu_{ij} \ll 1$, provides approximately the same likelihood as a Bernoulli formulation with $\boldsymbol{\theta}$, as

$$\begin{aligned}
\mathrm{Pois}(a_{ij}; \mu_{ij}) &= \mathrm{e}^{-\mu_{ij}} (\mu_{ij})^{A_{ij}} / A_{ij}!, \\
&= \mathrm{e}^{-\mu_{ij}} (\delta_{a_{ij},0} + \mu_{ij} \delta_{a_{ij},1} + \mathcal{O}(\mu_{ij}^2)), \\
&\approx (1 - \mu_{ij} + \mathcal{O}(\mu_{ij}^2))(\delta_{a_{ij},0} + \mu_{ij} \delta_{a_{ij},1}), \\
&\approx (1 - \mu_{ij}) \delta_{a_{ij},0} + \mu_{ij} \delta_{a_{ij},1} + \mathcal{O}(\mu_{ij}^2),
\end{aligned} \tag{2.12}$$

where $\mathcal{O}(\mu_{ij}^2)$ means terms of order μ_{ij}^2 or higher — so using this even for binary networks is often still reasonable. Throughout this thesis, to avoid confusion between these two base models, we refer to the SBM of Eqn. (2.9) as either the classical SBM, or the non-degree-corrected SBM (NDC-SBM).

To find suitable groups in a network under these models, we are interested in the posterior distribution given the observed data and other parameters,

$$\mathrm{p}(Z \mid A, \alpha, \boldsymbol{\omega}) = \frac{\mathrm{p}(A \mid Z, \alpha, \boldsymbol{\omega}) \mathrm{p}(Z \mid \alpha)}{\mathrm{p}(A \mid \alpha, \boldsymbol{\omega})}, \tag{2.13}$$

or indeed given hyperpriors, we might want to first marginalise $\alpha, \boldsymbol{\omega}$ for the posterior $\mathrm{p}(Z \mid A)$ [136].

This posterior takes into account prior uncertainty in the labels, via $\mathrm{p}(Z \mid \alpha)$, which also penalises the complexity of the model – *i.e.* importantly for SBMs, the number of groups – and so helps avoid overfitting to the data. Further, as it provides a distribution rather than a point estimate, this allows for estimates of the uncertainty in the labels as previously suggested.

Unfortunately, the denominator of Eqn. (2.13), is generally difficult to calculate, as it requires marginalising $\mathrm{p}(A \mid Z, \alpha, \boldsymbol{\omega})$ over all possible labellings, *i.e.*

$$\mathrm{p}(A \mid \alpha, \boldsymbol{\omega}) = \sum_{Z} \mathrm{p}(A \mid Z, \alpha\boldsymbol{\omega})\mathrm{p}(Z \mid \alpha), \tag{2.14}$$

where the sum contains $\mathcal{O}(Q^N)$ terms — to be avoided calculating if possible!

As such, there are a variety of methods to help tackle this problem. One popular approach is to use Markov chain Monte Carlo (MCMC) methods to sample directly from the posterior [134, 194]. In this work, we instead focus on variational methods, as we introduce in Chap. 3.

In addition to these options, other popular ways to find suitable network groups include local algorithms [84], spectral clustering methods [9], and matrix factorisation procedures [178] among others, but we do not explore these in depth herein. We do, however, explore greedy maximum-likelihood estimator (MLE) approaches for our model in Chap. 5 — this is another popular method for allowing models to scale to large networks, *e.g.* as in [69].

2.2.2 Choosing the number of groups, and comparing partitions

A key hyper-parameter in all models considered herein is the number of groups we believe are in the network, Q. As such, we typically fit models for multiple values of Q, and then perform model selection by comparing these in some way — there are alternative nonparametric procedures that directly infer Q from the data, *e.g.* [136, 154], but we do not explore these in this work.

There are a variety of ways one might compare different models for the network. One option is to compute the likelihood of observing the data under each model, and choose the model with the greatest value — that which most closely represent the data. Variants of this – which generally involve placing hyperpriors then marginalising model parameters – include the integrated completed likelihood (ICL), and the minimum description length (MDL) framings [152, 134]. This is our principal choice throughout this work. We note that even without marginalising model parameters, this is not quite equivalent to applying a likelihood ratio test, where we would solely compute $\mathrm{p}(A \mid Z, \boldsymbol{\omega})$ after finding suitable labels and block parameters, due to the prior α over Z — *i.e.* as we are explicitly modelling the joint distribution $\mathrm{p}(A, Z \mid \alpha, \boldsymbol{\omega})$.

Other simplified options using the likelihood include the Akaike and Bayesian information criterions (AIC and BIC resp.), which correspond to using MLEs for the model parameters rather than placing any priors, then penalising according to some function of the number of parameters and samples. However, the quality of these criterions rely upon assumptions that are typically unrealistic for network models, and thus tend to work poorly when used to determine the number of blocks in SBMs [192, 141].

Alternatively, it may be the case that we have access to some ground-truth labels, or metadata that we believe ought to be correlated with the true groups. In such a scenario, we can directly compare the partition of the network inferred to these 'true' values. Measures of partition similarity that we use throughout this work include the adjusted Rand index (ARI), maximum normalised overlap over permutations, normalised mutual information (NMI), and normalised reduced mutual information (NRMI), though these are but some of the measures available. In the following, we provide details of each of these measures, and note we assume that no groups in the partitions considered are empty, *i.e.* $n_q > 0$ for all q.

Firstly, for two partitions, X, Y over the same set with N elements, the ARI is given by

$$\mathrm{ARI}(X,Y) = \frac{\mathrm{RI}(X,Y) - \mathbb{E}[\mathrm{RI}(X,Y)]}{1 - \mathbb{E}[RI(X,Y)]}, \tag{2.15}$$

where the Rand index is given by

$$\mathrm{RI}(X,Y) = \frac{2\sum_{q,r}\binom{n_{qr}}{2} - \sum_q \binom{a_q}{2} - \sum_r \binom{b_r}{2} + \binom{N}{2}}{\binom{N}{2}}, \tag{2.16}$$

for n_{qr} the (q,r)th position in the contingency table between the partitions – *i.e.* the number of elements in common between group q in X and group r in Y – and $a_q = \sum_r n_{qr}$, $b_r = \sum_q n_{qr}$ the row and column sums respectively (*i.e.* the total size of the corresponding group in the partition). As per convention, $\binom{n}{k} = n!/(n-k)!k!$ is the binomial coefficient, which corresponds to the number of ways to choose k (unordered) items from a set size n. The expectation $\mathbb{E}[\mathrm{RI}(X,Y)]$ is then taken over some random null model for the partitions – for simplicity we follow the norm of correcting using a permutation model for the clusterings, where the group sizes are fixed, despite some known possible problems, as discussed in *e.g.* [48]. We thus obtain

$$\mathrm{ARI}(X,Y) = \frac{\sum_{q,r}\binom{n_{qr}}{2} - \left[\sum_q \binom{a_i}{2}\sum_r \binom{b_r}{2}\right] / \binom{N}{2}}{\frac{1}{2}\left[\sum_q \binom{a_i}{2} + \sum_r \binom{b_r}{2}\right] - \left[\sum_q \binom{a_i}{2}\sum_r \binom{b_r}{2}\right] / \binom{N}{2}}. \tag{2.17}$$

Note that the ARI can take negative values, effectively if the partitions have less overlap than expected in the random null model, and has a maximum value of 1 should the partitions be identical up to label permutations.

Next, as in [50] we consider another useful measure to be the normalised overlap, *i.e.* the proportion of correct labellings, maximised over their $Q!$ permutations (hence not feasible to calculate exactly for Q much more than $\mathcal{O}(10)$), and normalised to take a value of zero if it is no better than random, and one if there is perfect agreement. That is, in the case of uniform group sizes as considered in our use case,

$$\text{overlap}(Z_{true}, Z_{pred}) = \max_{\sigma \in S_Q} \left(\frac{\frac{1}{NT} \sum_i \delta_{x_i, \sigma(y_i)} - 1/Q}{1 - 1/Q} \right), \tag{2.18}$$

where S_Q is the set of permutations over Q elements.

Finally, we describe the two mutual information based measures. The basic idea is that we can consider the partitions as defining simple probability distributions, *i.e.* where the the probability of a random node drawing a label q in X to be $\mathrm{p}_X(q) = a_q/N$, and r in Y is $\mathrm{p}_Y(r) = b_r/N$. Likewise, the probability of a node carrying both labels q in X and r in Y is $\mathrm{p}_{X,Y}(q,r) = n_{qr}/N$.

Given these distributions, we may then appeal to information theoretical measures of similarity. Specifically, the mutual information, $I(X,Y)$, may be considered to be the amount of information saved in communicating one partition given the other. This symmetric measure corresponds to the difference between the entropy of one distribution, $H(X)$ say, and the conditional entropy of $H(X \mid Y)$,

$$I(X,Y) = H(X) - H(X \mid Y), \tag{2.19}$$

where these entropy terms are defined as

$$H(X) = -\sum_q \mathrm{p}_X(q) \log \mathrm{p}_X(q), \qquad H(X \mid Y) = -\sum_{q,r} \mathrm{p}_{X,Y}(q,r) \log \frac{\mathrm{p}_{X,Y}(q,r)}{\mathrm{p}_Y(r)}. \tag{2.20}$$

If the distributions over X and Y are independent, then $\mathrm{p}_{X,Y}(q,r) = \mathrm{p}_X(q)\mathrm{p}_Y(r)$, and hence the conditional entropy takes its maximum value $H(X \mid Y) = H(X)$, and if they are identical then $H(X \mid Y) = 0$, thus $I(X,Y)$ is zero for completely uncorrelated partitions, and $H(X) = H(Y)$ if they are perfectly identical. As such, to rescale this to take values between 0 and 1, and thus avoid dependene on the number of groups *etc.* , we define the normalised mutual information (NMI), where we choose to normalise

using the mean of the entropies,

$$\mathrm{NMI}(X, Y) = \frac{2I(X, Y)}{H(X) + H(Y)} \tag{2.21}$$

There has also been a recent improvement upon this latter measure, normalised reduced mutual information (NRMI) [112], that accounts for information contained in the contingency table of the two partitions as well as (roughly) the conventional mutual information. Recalling that the entropy $H(X) = I(X, X)$, this corresponds to simply replacing the mutual information terms in the NMI equation above with the reduced mutual information,

$$\mathrm{RMI}(X, Y) = I(X, Y) - \frac{1}{N}\Omega(a, b), \tag{2.22}$$

where $\Omega(a, b)$ counts the number of possible contingency tables for the partitions, with row sums $a = \{a_q\}$ and column sums $b = \{b_r\}$. We refer the interested reader to the original paper for details of how to approximate this additional term.

Chapter 3

Dynamic SBMs with general metadata

Now we have covered the necessary preliminaries, in this chapter we describe the principal class of models to be applied in the remainder of this work. Precisely, we elaborate a dynamic SBM that allows us to infer evolving latent groups in a discrete time series of network snapshots, while incorporating temporal metadata defined over the nodes of the network. By metadata, we mean any additional information beyond the (possibly weighted) adjacency of the network at each snapshot. Primarily this could include (a) nodal metadata as considered herein, for instance the institution or subject area of an author, or (b) edge metadata as considered in some of our previous work, for instance the language of a publication, that is then used to define a relationship between authors in some way.

Finding latent groups in network data has been a fertile pursuit for several decades now [114, 119], as such groups often confer information about *e.g.* different functional areas in the network, which individuals are more closely related *etc.*, or simply allow coarse-graining of the network for visualisation purposes without losing some aspects of the overall structure. Within our primary area of interest, academic networks, such techniques have been applied to better understand research communities (groups in author-level networks) and topics (usually as groups in publication- or word-level networks). By including both metadata and time evolution within the model used to infer these latent groups, we hope to improve both quality and interpretability of the resulting clustering.

To produce such a model, we begin from the dynamic SBM model proposed in [98], which has a number of desirable features, and has been found to often provide the best temporal clusterings of models that explicitly aim to capture how groups evolve over time. Our initial contribution is to allow mixed metadata to be defined over the nodes of the network, with a variety of different distributions, as proposed

in Sec. 3.2.2, while leaving unchanged the likelihood function defined over the edges of the network. As we later describe, this results in new equations for some of the parameters in common between both models, along with those necessary for the new parameters describing the metadata.

We then elaborate one way of 'tuning' the relative importance of metadata within our model to improve results, somewhat along the lines of a method proposed for categorical metadata in [132]. Further, along with several other improvements of the codebase, we have implemented inference for Poisson (or rather zero-truncated Poisson (ZTP)) distributed edge weights, and obtained results (see Sec. 3.4) — the equations necessary to do so were discussed to some degree in [98], but were not coded, and so no results were provided.[1]

Beyond these newly available features, we further describe a variety of novel extensions of the base model. Of particular relevance for later chapters, we describe a degree-corrected version, though other suggestions may be found in App. A.

The inclusion of metadata in generative models for networks has been a key focus of much research in recent years, with a particular surge of interest since the pioneering work of [113]. Perhaps the most exciting outcome of this study was the empirical finding that their method outperformed previous techniques for network clustering based on network structure alone, including beyond detectability limits for the classical SBM (see *e.g.* [105] for details of such limits). Further, should either metadata or network structure be non-informative with respect to groups of the nodes, their method successfully ignored one to return estimates based solely on the other, thus apparently providing all of the benefits of both methods without any of the drawbacks. We discuss these desirable features in greater depth in the following sections.

In sum, the aims of our model are to:

- Model the full generation of the temporal network observed – including metadata – rather than either (i) taking metadata as fixed (commonly performed, see *e.g.* [193]), or (ii) discretising into time periods, inferring blocks for each timeslice individually, before post hoc matching using some heuristic. This allows us to better understand uncertainty in metadata in the usual empirical case where it is not necessarily completely reliable (*i.e.* noisy);

[1] Supporting code for this chapter is publicly available as an R package on GitHub. Though substantively modified, this package developed from a fork of the code for [98].

- Allow groups to change over time, rather than finding some optimal single partition, as is the case for methods which aggregate the series of observations into a single network somehow, or the multilayer approach of *e.g.* [62];
- Further explicitly model group transitions to allow understanding of the evolution of the groups, rather than drawing groups from a prior which may evolve (for instance as in [177, 5, 159] and others). To do so, we assume first-order Markov transitions over the latent groups, *i.e.* effectively take a hidden Markov model (HMM) approach, as has been explored in several previous works we comment on in greater detail below, with some noteworthy success. Explicitly, for each node, its group membership forms a Markov chain, independent of the values of the other nodes memberships;
- Allow for different classes of metadata, *i.e.* not solely categorical/discrete but also continuous metadata. Continuous metadata prevents immediate application of models based on some bipartite/labelling formulation (*e.g.* [62, 132], and the principal section of [113]), while the presence of discrete data also prevents the multivariate Gaussian approach of [167] being fully satisfactory.

The remainder of the chapter is organised as follows: first, in Sec. 3.1, we review some of the most closely related literature — models for dynamic networks, and models for networks with metadata, especially those based on the SBM. Next, in Sec. 3.2, we proceed to elaborate the details of the base case of our own model class – the Dynamic SBM with Metadata (DSBMM) – for undirected networks, describe one method to find suitable parameters for the model given observed data, and discuss related technical aspects of the procedure such as its complexity.

Following this, in Sec. 3.3 we apply the model to simulated data, which allows us to explore certain strengths and weaknesses — in particular compared to the analogous model without metadata. Having done so, in Sec. 3.4 we consider some of the insights it can provide on several real network datasets, before concluding the chapter with a discussion.

3.1 Literature review

Before proceeding to our own contributions, we provide a brief survey of previous related work. As there is a dearth of SBMs for dynamic networks with metadata, we separately explore prior work on dynamic models, and models with metadata in the sections that follow.

3.1.1 Dynamic network models

There has been a wealth of knowledge produced over the last decade or so on the topic of statistical models for dynamic networks. In this section, we provide a brief overview of the area, with a particular focus on dynamic SBMs that are similar to that proposed.

3.1.1.1 Dynamic SBMs

The first of these closely related models is an early dynamic SBM introduced in [194]. The authors also considered Markov transitions for latent groups in discrete time network series, though in a Bayesian setting (for simple conjugate priors over parameters), with inference through (collapsed) Gibbs sampling combined with probabilistic simulated annealing. An alternative variational inference procedure is also explored, though their proposed family of variational distributions results in more involved equations than that herein. They allow for weighted edges, online learning, and – though without positing a process governing birth/death – the addition or removal of nodes. The method scales reasonably well, though [98] demonstrate that poor initialisation of clusters cause the MCMC scheme to struggle, as is often the case for local move schemes. We note however, that in the static case merge-split operations have recently been proposed to help overcome this, and such could be extended to the dynamic case [142]. Only the group membership is allowed to vary across time – parameters governing connectivity between groups are assumed constant.

Another similar dynamic SBM model for discrete time networks with Markov transitions between groups is that of [188]. They assume the network edges are binary, while groups transition according to a state-space model. Along with these groups, connectivity parameters also freely vary, which is known to cause identifiability problems [98]. Inference is performed through a combination of label-switching and applying a Kalman filter. As for [194], this paper does not suggest how to select the number of blocks to infer. There has however been some work producing 'non-parametric' Bayesian variants analogous to these models, where the number of blocks is effectively inferred along with the partition itself, *e.g.* [65, 57].

The closest previous model to that proposed, as in fact we use their work as a basis upon which to build, is that of [98]. The paper describes an SBM that allows the study of weighted networks that once again evolve over discrete timesteps according to Markov transitions. They note that the main difference to [194, 188] is that they

allow both groups and parameters to vary through time, but include identifiability conditions for valid statistical inference.

Some studies allow for consideration of continuous rather than discrete time, typically through non-homogeneous Poisson process (NHPP) based approaches – see *e.g.* [38, 31, 99]. [31] allows nodes to change groups over time, but only by clustering time periods, then fixing node groups and edge intensities within these clusters – as such, the transitions between groups are not modelled explicitly. Both [38] and [99] fix node groups over the full time period, but allows the intensities to vary, and [99] discuss identifiability requirements. Extending the NHPP approach is one possibility for SBMs for continuous-time dynamic networks with metadata, but in doing so we lose the explicit modelling of transitions between groups. Such processes are suitable for count-based data in continuous-time, and are thus one avenue for further exploration.

Importantly, [29] allows for textual metadata to further be defined over the edges of the network, effectively through coupling their blocks with the popular latent Dirichlet allocation topic model [16] — the only other real example of a generative SBM for dynamic networks with metadata. Further, they allow for change-point detection in the network, a matter of growing interest [131, 145]. Nonetheless, they do not explicitly model the evolution of groups as they use the NHPP framework, and do not allow for either nodal metadata of any form, nor other types of edge metadata. Scaling issues also appear to occur.

Finally, there are various dynamic SBMs that allow link persistence or correlation in some way, as this is known to be an important effect in many networks. While we model group persistence, we do not explicitly allow for this. A recent avenue of research has begun to explore methods to address such, for instance [187, 94, 123], but we do not pursue the topic further here.

3.1.1.2 Other relevant dynamic models

Other than the SBM approaches above, various dynamic latent factor models have been proposed, for instance a 'latent multilayer' approach [195], a gamma process based approach [3] and similar [74, 103, 45]. The multiple (binary or otherwise) factors are often interpreted as overlapping/mixed groups, but substantively frequently have quite different results, and hence use-cases, to the hard codings inferred through conventional forms of labelling such as that via SBMs. Such models are closer to the 'mixed-membership' SBM framework introduced in [7]), which have also been extended to allow network dynamics, see *e.g.* [186].

Other studies incorporate time by explicitly considering the data as a multilayered network, with time as one of the layer attributes – further allowing for categorical metadata [137]. Nonetheless, such multilayer models do not directly propose a mechanism for network evolution, and thus properly ascertaining the likelihood of future states is a computationally costly process . Indeed, other than via the prior over the groups , there is no direct use of the ordinal nature of the dataset, and thus meaningful information is discarded.

This overview is by no means exhaustive, but broadly covers the avenues of prior investigation most pertinent to the current study. For a recent comprehensive overview of probabilistic models for dynamic networks, we refer the interested reader to [29].

3.1.2 SBMs with metadata

As previously described, in this work we consider metadata to be any information defined over the vertices or edges of a network, beyond the connectivity structure (and possibly associated weights) itself. There are a host of relevant existing extensions to SBMs to incorporate such metadata, and in this section we review only a subset.

Note following convention in the literature, henceforth we describe nodal metadata that falls into specific categories (*e.g.* institutional affiliation for publication data) as node attributes. Many models are suitable for this type of data, but far fewer for ordinal metadata (*e.g.* duration of career) or more generally continuous metadata (for instance an inferred latent embedding given an authors publications). If only a single attribute is permitted for each node, rather than multiple, we refer to such metadata as categorical nodal attributes.

As previously suggested, a pioneering work in this area was [113], where they developed a static SBM allowing for metadata by seeking suitable conditional likelihoods of nodes belonging to a certain class given their metadata. Of particular utility is that the method allows for metadata to be uncorrelated with the revealed community structure, which itself can be worthwhile information. The model is predominantly designed for nodal attribute metadata, taken as a given, and these condition priors over the labels before generating network – effectively learning a linear transformation from metadata to latent groups given network structure. They do however suggest a method to allow continuous, bounded metadata, via the use of Bernstein polynomials. Inference is performed through a combination of belief propagation (BP) and expectation maximisation (EM).

Another important contribution was [132], where they proposed a means to test categorical nodal attribute significance with regards to network structure (BESTest).

They further allow the tuning of importance of metadata to move between the attribute partition, and that of the conventional SBM (the neoSBM), which inspired a development for the DSBMM we outline in Sec. 3.2.8.

A variety of others (*e.g.* [62]) instead form an augmented graph in addition to the underlying network to investigate nodal attributes, adding new attribute nodes with corresponding edges to the relevant original vertices as a separate bipartite graph layer. If the metadata were continuous, we would have to discretise into bins, and thus such methods are only really suitable for networks with little metadata and/or of small size. Note in this setup, while metadata must still be constrained within categories, nodes may have multiple labels associated with them, thus significantly broadening the scope of application relative to [132]. Further, this is effectively an extension of the multilayer framework previously discussed, which would hence also allow the investigation of metadata defined over the edges, provided it is either constrained within categories, or can be binned suitably in this manner (though with considerable computational expense associated).

A more contemporary work, [167], proposed a fully generative model for networks with multiple continuous attributes. To do so, they posit that such metadata is distributed according to a mixture of (multivariate) normals, with unique parameters for each group. As such, it is not particularly suitable for application to datasets with discrete/categorical metadata.

An important recent contribution has been [153], where they describe a SBM modified to account for metadata. To do so, they fix the metadata then modify the probability of an edge by some measure of metadata similarity, and describe an efficient belief propagation inference procedure. Perhaps the most notable aspect is that they appeal to prior examples in the literature to use this belief propagation framework to analytically investigate the detectability of groups in the network, as we discuss in Chap. 4. This is the first work to our knowledge that aims to explicitly quantify the improvement to group detectability, albeit for a toy model, that can be obtained through incorporating metadata. However, as we discuss later, in addition to several errors in algebraic manipulation, they make a more fundamental error that prevents their results from being legitimate.

Finally, we note again that edge metadata can also be considered – for instance [19] deal with textual edge metadata in a similar way to our approach (*i.e.* allowing separate topic distributions for each pair of nodal groups). As we previously mentioned, of particular relevance to our own study is the extension of this model to the dynamic case in [30], effectively a combination of [19] with the NHPP approach introduced in

[31]. However, as such the transitions between groups are not modelled explicitly, rather groups are fixed within each time cluster, and are modelled within each independently.

There are also a variety of other methods to handle networks with metadata in the literature, for instance matrix factorisation based approaches such as [150], or latent feature models as in *e.g.* [193]. As they are less related to our current study, we do not explore such alternative in greater depth herein.

3.2 Methods

3.2.1 The temporal network component

In this section, we elaborate the temporal network component of our model, before discussing how we further incorporate metadata in Sec. 3.2.2. As stated in the introduction to this chapter, this component of our model – prior to further specifications discussed below – is much as in [98].

Specifically, we consider temporal network data, A, to be defined over N vertices at T discrete points in time, or timesteps, such that we may describe the network through a series of $N \times N$ matrices $(A^t)_{t=1,\dots,T}$. Here, A^t_{ij} is a real value that confers some information about the relationship between the vertices i and j at time t. As we describe in following sections, these values do not necessarily have to be binary.

We aim to infer evolving latent groups of nodes in this network, and as is frequently done, we specify the number of groups permitted as a hyper-parameter of the model, Q. We denote the group labels for the set of nodes at each timestep by $Z = (Z^t)_{t=1,\dots,T}$, where $Z^t_i \in \{1, \dots, Q\}$ denotes the group label assigned to node i at time t.

As suggested above, we assume the set of random variables (Z_i) are independent and identically distributed (i.i.d.), and model the evolution of these values, (Z^t_i), as following an irreducible, aperiodic, stationary Markov chain with $Q \times Q$ transition matrix $\pi = (\pi_{qq'})$ and initial distribution $\alpha = (\alpha_1, \dots, \alpha_Q)$.

Within each time period, the network is assumed to follow a (weighted) stochastic blockmodel, *i.e.* the likelihood of an edge between i and j is solely dependent upon Z^t_i, Z^t_j — we explore one immediate improvement upon this assumption, degree-correction, in Sec. 3.2.9 below. To take into account possible sparsity in weighted graphs, as is commonly the case in large empirical networks, we explicitly introduce a Dirac mass at 0, denoted by δ_0, as a component of this distribution. Without specifying the exact form of the weighted distribution, this corresponds to

$$A^t_{ij} \mid \left\{ Z^t_{iq} Z^t_{jl} = 1 \right\} \sim \left(1 - \omega^t_{ql}\right) \delta_0(\cdot) + \omega^t_{ql} F\left(\cdot, \gamma^t_{ql}\right), \tag{3.1}$$

where $\{F(\cdot,\gamma), \gamma \in \Gamma\}$ is a parametric family of distributions with no point mass at 0, and $Z_{iq}^t = 1$ if node i is in group q at time t and zero else. Let $\phi(\cdot;\omega,\gamma)$ denote the density of this distribution. Throughout, we frequently abbreviate the group distributions $\phi(\cdot;\omega_{ql}^t,\gamma_{ql}^t)$ to $\phi_{ql}^t(\cdot)$ or $\phi_{ql}^t(\cdot;\theta)$.

Within this work, when we consider networks with nodes that are not present at all timesteps, we make the same assumption as [98] — that the initial likelihood of belonging to a group q, α_q, remains applicable for new nodes entering the network, *i.e.* given a node i that is not present in the network at time $t-1$, our prior $\mathrm{p}(z_i^t) = \alpha_q$ for any t. Note that permitting the set of nodes to change between each timestep means that the system is no longer truly time homogeneous as assumed, but this is necessary when considering the empirical data of interest — academics do not continue to publish indefinitely, and new academics are trained continuously.

Prior to continuing, we note that there is a further implicit step that must be taken to apply this temporal formulation to empirical data. That is, time is not naturally a discrete process, nor are observations (*e.g.* academic publications) typically sampled at a constant rate. As such, to convert real data to the necessary format for input to the model, numerous choices must be made — perhaps most importantly

(i) The frequency to which data will be aggregated, *e.g.* monthly, annually *etc.* , and

(ii) How this aggregation will occur, *e.g.* strict aggregation within discrete windows, rolling windows over time *etc.*

As these choices fundamentally determine the data to which any subsequent model is fitted, different choices may lead to significantly different groups being inferred at any given time. Within this work, as this is not our primary area of model development, we choose to divide data into regular windows, and strictly aggregate data within these windows. We specify the length of the windows according to intuitive time-scales for the problem. However, numerous heuristic alternatives were considered — for instance, using methods such as that developed in [131], determine change-points in the system, and fitting multiple homogeneous models between these. Another option was to account for the well-known property for regular discrete-time HMMs, that the expected waiting-time (in terms of timesteps), $\mathbb{E}[d_i]$ in a given state, i, given the time-homogeneous transition matrix π, is [151]

$$\mathbb{E}(d_i) = \frac{1}{1-\pi_{ii}}. \tag{3.2}$$

As such, if the expected waiting-time for a reasonable proxy for groups in the empirical network is known — for instance the typical duration of tenure of an academic at a given institution — one could instead specify a desired average probability of switching groups between times (*i.e.* $1 - \pi_{ii}$), then define the window period in order to match the two. Finally, another alternative considered should the network be changing at an increasing rate over time was to increase the sampling rate (*i.e.* decrease window size), in order to maintain similar rates of change. We leave rigorous exploration of this issue to future work.

Given observed temporal relational data — discretised in whichever chosen method — to estimate suitable parameters of our model, in this chapter we perform variational expectation maximisation (VEM), as we describe in Sec. 3.2.3. Before proceeding to do so, in the following section we elaborate the details of the key feature of the DSBMM — how we further incorporate metadata.

3.2.2 Incorporating metadata

As stated above however, we are interested in more than just edge data – we choose to model the generation of all observed data, *i.e.* both network and metadata, and assume the two are conditionally independent given the latent groups. To clarify, in Fig. 3.1 we display two popular graphical frameworks for models that generate node labels Z, networks A, and metadata X. We take the approach on the left, where we assume that the metadata and the network are independent after conditioning upon the labels. The approach on the right is common, particularly when metadata is assumed fixed. This is often taken to be the case, and typically the subsequent modelling assumption is that the conventional SBM probability of an edge is multiplied by a similarity measure between nodal metadata for the given pair. Not shown are alternatives where *e.g.* labels are conditional upon metadata prior to themselves generating the network, *i.e.* $X \longrightarrow Z \longrightarrow A$ as in [113], or combined approaches (*e.g.* weakly/strongly connected graphical models).

Specifically, to more readily allow us to consider how (possibly evolving) metadata defined over the vertices of the network, denoted by $X = (X_i^t)$, might influence the latent groups, we make some stronger further assumptions about the dependency relations in our model:

1. Much as for edges in the SBM, the metadata of a given node at a particular time is conditionally independent of all others, at all other times, given its latent group at that point.

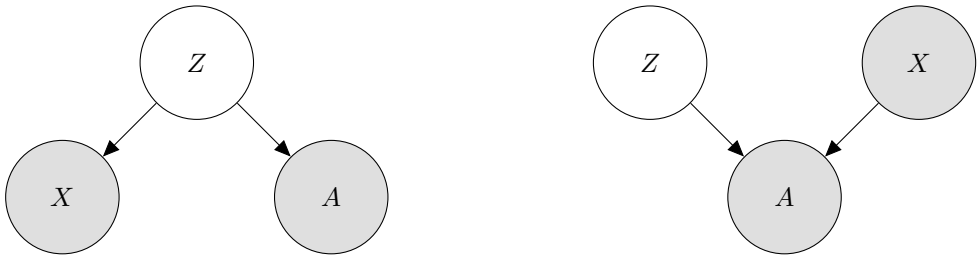


FIGURE 3.1: In this figure we display in simplest terms two possible graphical model frameworks for inferring latent groups, Z, for observed networks with (weighted/temporal *etc.*) adjacency matrices denoted by A, and associated (temporal) metadata denoted X. On the left, the network itself and nodal metadata are independent given the node labels, while on the right the network is generated according to both labels and metadata. We take the approach on the left.

2. Should there be multiple pieces of metadata over each node, these too are conditionally independent given its latent group.

With these considerations, we assume that given some parameters for the distribution within each group at a certain time, ϑ_q^t, the metadata is generated according to

$$X_i^t \mid Z_{iq} = 1 \sim \mathrm{p}(\cdot, \vartheta_q^t), \tag{3.3}$$

for some distribution with density $\mathrm{p}(\cdot, \vartheta_q^t)$, which – abbreviating as before to $p_q^t(\cdot)$ – would be the case should

$$X_i^t \sim \prod_q p_q^t(\cdot)^{Z_{iq}^t}. \tag{3.4}$$

Should we have multiple pieces of metadata, say S different types such that $\boldsymbol{x}_i^t = \{x_{i1}^t, \ldots, x_{iS}^t\}$, our further conditional independence assumption suggests we may simply model these using the product

$$\boldsymbol{X}_i^t \sim \prod_q \left(\prod_s p_{qs}^t(\cdot) \right)^{Z_{iq}^t}. \tag{3.5}$$

Note that choosing this form of distribution over the metadata leads to immediately interpretable group-level parameters, as we demonstrate in Sec. 3.4. The trade-off resulting from this interpretability benefit is the introduction of possible problems when metadata has meso-scale structure within the network, but in a different manner to edge meso-scale structure. We discuss this in greater depth below, and display examples of this behaviour (and a remedy to it) in Sec. 3.3.2.

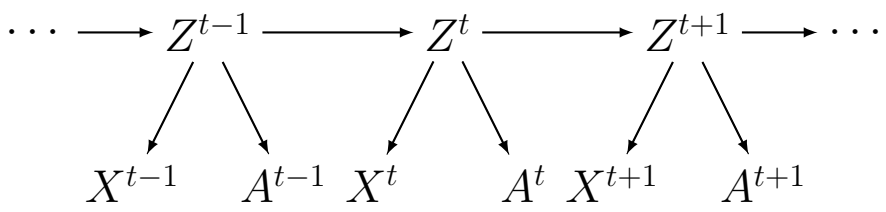


FIGURE 3.2: The high-level graphical model proposed

Now we may elaborate the full data log likelihood of our model, neglecting self-edges and new nodes entering the system for clarity, as

$$\begin{aligned}\log \mathrm{p}_\theta(X, A, \mathbf{Z}) = &\sum_{i=1}^{N}\sum_{q=1}^{Q} Z_{iq}^{1} \log \alpha_q + \sum_{t=2}^{T}\sum_{i=1}^{N}\sum_{1\le q,q'\le Q} Z_{iq}^{t-1} Z_{iq'}^{t} \log \pi_{qq'} \\ &+ \sum_{t=1}^{T}\sum_{1\le i<j\le N}\sum_{1\le q,l\le Q} Z_{iq}^{t} Z_{jl}^{t} \log \phi\left(A_{ij}^{t}; \omega_{ql}^{t}, \gamma_{ql}^{t}\right) \\ &+ \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{q=1}^{Q} Z_{iq}^{t} \log \mathrm{p}(X_i^t; \vartheta_q^t). \end{aligned} \tag{3.6}$$

The high-level overview of the proposed graphical model is shown in Fig. 3.2.

3.2.3 Mean-field variational inference for the model

Of course, the immediate question we now face, as for any probabilistic model, is how to fit parameters such that they best match observed data. With regards to SBMs in general, the principal difficulty arises from trying to choose the block labels, Z — the reason being that they are not independent when conditioning on the observed network, as the likelihood of each edge is jointly dependent on the labels of the two nodes involved. To find these parameters, there are several main approaches: (i) local algorithms (*e.g.* greedy MLE approaches as in [69] *etc.*), (ii) spectral clustering methods (*e.g.* [50]), (iii) belief propagation (*e.g.* [153]), (iv) Markov chain Monte Carlo (MCMC) procedures like [134, 142], or (v) mean-field variational inference (MFVI) approaches (as in [98]). Within this chapter, we focus on MFVI — in subsequent chapters, we describe alternatives, in particular belief propagation, to overcome some of the downsides of this method.

The key step when performing variational inference is deciding upon a variational family of distributions, which are sufficiently complex to reasonably fit the model and hence describe the data, while remaining sufficiently simple such that they allow us

to optimise the new parameters introduced. In the current work, we use the same variational family as [98], parameterised by τ:

$$\begin{aligned}\mathbb{Q}_\tau(\mathbf{Z}) &= \prod_{i=1}^{N} \mathbb{Q}_\tau\left(Z_i\right) = \prod_{i=1}^{N} \mathbb{Q}_\tau\left(Z_i^1\right) \prod_{t=2}^{T} \mathbb{Q}_\tau\left(Z_i^t \mid Z_i^{t-1}\right), \\ &= \prod_{i=1}^{N} \left[\prod_{q=1}^{Q} \tau(i,q)^{Z_{iq}^1}\right] \times \prod_{t=2}^{T} \prod_{1 \leq q,q' \leq Q} \tau\left(t,i,q,q'\right)^{Z_{iq}^{t-1} Z_{iq'}^{t}},\end{aligned} \tag{3.7}$$

where for any values (t,i,q,q'), we have $\tau(i,q)$ and $\tau(t,i,q,q') \in [0,1]$, with the constraints $\sum_q \tau(i,q) = 1$ and $\sum_{q'} \tau(t,i,q,q') = 1$. This class of probability distributions $\mathbb{Q}_\tau$ corresponds to considering independent laws through individuals (*i.e.* assuming we can factorise the likelihood over the nodes), while maintaining that for each i the distribution of Z_i under $\mathbb{Q}_\tau$ follows a Markov chain, now with inhomogeneous transition $\tau(t,i,q,q') = \mathbb{Q}_\tau(Z_i^t = q' \mid Z_i^{t-1} = q)$, and initial distribution $\tau(i,q) = \mathbb{Q}_\tau(Z_i^1 = q)$.

Key quantities are thus the marginal components of $\mathbb{Q}_\tau$, namely $\tau_{\text{marg}}(t,i,q) := \mathbb{Q}_\tau(Z_i^t = q)$, as this corresponds to our estimation of the marginal likelihood of node i belonging to group q at time t. These quantities are computed recursively by

$$\tau_{\text{marg}}(1,i,q) = \tau(i,q) \text{ and } \forall t \geq 2, \tau_{\text{marg}}(t,i,q) = \sum_{q'=1}^{Q} \tau_{\text{marg}}\left(t-1,i,q'\right) \tau\left(t,i,q',q\right). \tag{3.8}$$

To decide how to choose the 'best' variational parameters, consider the following. One logical way of defining the optimal true distribution, given our model, is the set of parameters that maximises the overall likelihood of observing the data – the model evidence $\mathrm{p}_\theta(X,A) = \sum_Z \mathrm{p}_\theta(X,A,Z)$. However, given the combinatorial enormity of the size of the partition space, this sum is intractable, hence we instead use our variational approximation and Jensen's inequality [67] to find our optimisation objective, the so-called Evidence Lower Bound (ELBO):

$$\begin{aligned}\log \mathrm{p}_\theta(X,A) &= \log \sum_Z \mathrm{p}_\theta(X,A,Z), \\ &= \log \sum_Z \mathrm{p}_\theta(X,A,Z) \frac{\mathbb{Q}_\tau(Z)}{\mathbb{Q}_\tau(Z)}, \\ &= \log \left(\mathbb{E}_{\mathbb{Q}_\tau}\left[\frac{\mathrm{p}_\theta(X,A,Z)}{\mathbb{Q}_\tau(Z)}\right]\right), \\ \text{(Jensen's ineq.)} &\geq \mathbb{E}_{\mathbb{Q}_\tau}[\log \mathrm{p}_\theta(X,A,Z)] - \mathbb{E}_{\mathbb{Q}_\tau}[\log \mathbb{Q}_\tau(Z)].\end{aligned} \tag{3.9}$$

As the model evidence does not depend on our variational parameters, to make the bound as tight as possible, we want find τ that maximises the RHS. Indeed, it can be easily verified that maximising this term is equivalent to minimising the Kullback-Leibler divergence between the variational distribution and the posterior over the labels given the data, $D_{\mathrm{KL}}(Q(Z)\|\mathrm{p}_\theta(Z \mid X, A)) = \mathbb{E}_{\mathbb{Q}_\tau}[\log(Q(Z)/\mathrm{p}(Z \mid X, A))]$, further supporting this choice.

As the Z_{iq}^t variables are truly just indicator functions for the event $Z_i^t = q$, taking the expectation over the variational distribution is trivial – we may effectively substitute $\tau_{\mathrm{marg}}(t, i, q)$ for Z_{iq}^t when this is a single observation, and $\tau_{\mathrm{marg}}\,(t-1, i, q)\,\tau\,(t, i, q, q')$ for the joint observation $Z_{iq}^{t-1} Z_{iq'}^t$. Making these substitutions (*i.e.* taking the expectation over $\mathbb{Q}_\tau$), we wish to maximise

$$\begin{aligned} J(\theta, \tau) := &\sum_{i=1}^{N}\sum_{q=1}^{Q} \tau(i,q)\left[\log \alpha_q - \log \tau(i,q)\right] \\ &+ \sum_{t=2}^{T}\sum_{i=1}^{N}\sum_{1\le q,q'\le Q} \tau_{\mathrm{marg}}\,(t-1,i,q)\tau\,(t,i,q,q')\left[\log \pi_{qq'} - \log \tau\,(t,i,q,q')\right] \\ &+ \sum_{t=1}^{T}\sum_{1\le i<j\le N}\sum_{1\le q,l\le Q} \tau_{\mathrm{marg}}(t,i,q)\tau_{\mathrm{marg}}(t,j,l)\log \phi_{ql}^t\left(A_{ij}^t\right) \\ &+ \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{q=1}^{Q} \tau_{\mathrm{marg}}(t,i,q)\log p_q^t(X_i^t). \end{aligned} \tag{3.10}$$

When compared to the objective in [98], the inclusion of metadata has resulted in the addition of the final term. To maximise, we alternate between maximising with respect to the variational parameters τ while holding all others fixed, then using these new values to maximise with respect to model parameters. To do so, we require the partial derivatives of $J(\theta, \tau)$. Most of these are trivial to compute, with the minor exception of needing to note that for $t \ge 2$,

$$\frac{\partial \tau_{ti\ell}^m}{\partial \tau_{tiqq'}} = \frac{\partial}{\partial \tau_{tiqq'}} \sum_k \tau_{t-1,ik}^m \tau_{tik\ell} = \tau_{t-1,iq}^m \delta_{\ell q'}, \tag{3.11}$$

where δ_{ij} is the Kronecker delta – one if $i = j$ else zero – and for parsimony, we have introduced the abbreviations τ_{tiq}^m for $\tau_{\mathrm{marg}}(t, i, q)$, $\tau_{tiqq'}$ for $\tau\,(t, i, q, q')$, and not explicitly stated the range of the sum over the group label, here k. When there is no confusion, we often subsequently perform such abbreviations. With this in mind, we

can use the normalisation constraints for τ to find the fixed point equation

$$\forall t \geq 2, \forall i \geq 1, \forall q, q' \in \mathcal{Q}, \quad \hat{\tau}\left(t, i, q, q'\right) \propto \pi_{qq'} p^t_{q'}(X^t_i) \prod_{j=1}^{N} \prod_{l'=1}^{Q} \left[\phi^t_{q'l'}\left(A^t_{ij}\right)\right]^{\hat{\tau}_{\text{marg}}(t,j,l')}. \tag{3.12}$$

As a result of our conditional independence assumptions, equations for parameters for metadata distributions only depend on the variational parameters, τ, and the specific type of metadata itself. Below we provide examples for several common distributions of interest.

The one parameter that requires more care in deducing suitable equations is the initial variational likelihood of a node belonging to a particular group, τ_{iq}. This is because, as we can see from the recursive definition given in Eqn. (3.8), all subsequent marginal variational likelihoods involve this term. However, the relation is relatively easy to compute — we have that

$$\forall t \geq 2, \forall i \geq 1, \forall q^t, \in \mathcal{Q}, \quad \frac{\partial \tau^m_{tiq^t}}{\partial \tau_{iq}} = \sum_{q^2,\ldots,q^{t-1}} \prod_{t'=2}^{t} \tau_{t'iq^{t'-1}q^{t'}}, \tag{3.13}$$

with the convention that $q^{t-1} = q$ for $t = 2$. Using this equation, we can recover the full fixed point equation for $\tau_{iq} = \tau(i, q)$ – this is much as that presented in [98], with metadata through the introduction of the further RHS factor

$$p^1_q(X^1_i) \prod_{t\geq 2} \prod_{q^2,\ldots,q^t} p^t_{q^t}(X^t_i)^{\tau_{tiq^{t-1}q^t}}. \tag{3.14}$$

We do not include the full equation here, as the computational complexity required to calculate it turns out to be unnecessary – in practice we instead make an approximation, and neglect the latter part of this factor. That is, we approximate the fixed point equation for τ_{iq} as being identical to Eqn. (3.12) for $t = 1$, $q' = q$, only replacing $\pi_{qq'} p^t_{q'}(X^t_i)$ with $\alpha_q p^1_q(X^1_i)$.

The remaining update equations, with the exception of metadata and weighted

edges, are as follows:

$$\pi_{qq'} \propto \sum_{t=2} \sum_{i} \tau^m_{t-1,iq} \tau_{tiqq'}, \tag{3.15}$$

$$\omega^t_{ql} = \frac{\sum_{i \neq j : A^t_{ij} \neq 0} \tau^m_{tiq} \tau^m_{tjl}}{\sum_{i,j} \tau^m_{tiq} \tau^m_{tjl}}, \tag{3.16}$$

$$\omega_{qq} = \frac{\sum_{t, i \neq j : A^t_{ij} \neq 0} \tau^m_{tiq} \tau^m_{tjl}}{\sum_{t,i,j} \tau^m_{tiq} \tau^m_{tjl}}, \tag{3.17}$$

where we use $i, j : A^t_{ij} \neq 0$ to denote a sum over all i, j such that $A^t_{ij} \neq 0$, and we have fixed within-group connectivity parameters for identifiability, as discussed in [98].

3.2.3.1 Positive integer weighted edges

If we allow edges to take values in $\mathbb{N}_{>0}$ rather than being simply binary, one option is to choose our edge block distributions ϕ^t_{ql} to be zero-truncated Poisson (ZTP) distributions. As this is a natural distribution for weighted edges corresponding to the count of times a relationship has occurred, we use such distributions when considering academic networks in Sec. 3.4.

Denoting by λ^t_{ql} the mean parameter for the distribution counting the number of edges between groups q and l at time t, we have the variational update equations

$$\lambda^t_{ql} = \psi^{(-1)} \left(\frac{\sum_{i,j} \tau^m_{tiq} \tau^m_{tjl} A^t_{ij}}{\sum_{i,j : A^t_{ij} \neq 0} \tau^m_{tiq} \tau^m_{tjl}} \right), \tag{3.18}$$

$$\lambda_{qq} = \psi^{(-1)} \left(\frac{\sum_{t,i,j} \tau^m_{tiq} \tau^m_{tjq} A^t_{ij}}{\sum_{t,i,j : A^t_{ij} \neq 0} \tau^m_{tiq} \tau^m_{tjq}} \right), \tag{3.19}$$

where $\psi^{(-1)}(y)$ is the inverse function for $\psi(x) = x\mathrm{e}^x/(\mathrm{e}^x - 1)$. Extending the considerations of this equation beyond those in [98], we note this inverse function may be concisely written as

$$\psi^{(-1)}(y) = W_0(-\mathrm{e}^{-y} y) + y, \tag{3.20}$$

where $W_0(\cdot)$ is the principal branch of the Lambert W function, also known as the product logarithm, which satisfies

$$W_0(y) = x \iff y = x\mathrm{e}^x \qquad \text{for } x \in \mathbb{R}_{\geq 0}. \tag{3.21}$$

While this function cannot be further expressed in terms of elementary functions – though an integral representation in terms of elementary functions does exist for this

principal branch – various efficient implementations of asymptotic approximations are widely available.

3.2.3.2 Discrete edge categories

If rather than ordinal values, we have that there are different types of edges possible between nodes, with each pair of nodes restricted to a single type of edge (if any), it is natural to place a categorical distribution over these possible values for each pair of blocks. That is, for C categories,

$$F\left(\cdot, \gamma_{ql}^{t}\right) = \prod_{c} \gamma_{ql}^{t}(c)^{\delta_{\cdot,c}}. \tag{3.22}$$

In this case, the update equations – once again fixing the within-group parameters across time – are simply

$$\gamma_{ql}^{t}(c) = \frac{\sum_{i \neq j : A_{ij}^{t}=c} \tau_{tiq}^{m} \tau_{tjl}^{m}}{\sum_{c,i,j : A_{ij}^{t}=c} \tau_{tiq}^{m} \tau_{tjl}^{m}}, \tag{3.23}$$

$$\gamma_{qq}(c) = \frac{\sum_{t,i \neq j : A_{ij}^{t}=c} \tau_{tiq}^{m} \tau_{tjl}^{m}}{\sum_{t,c,i,j : A_{ij}^{t}=c} \tau_{tiq}^{m} \tau_{tjl}^{m}}. \tag{3.24}$$

This is the form of edge distribution we use when considering an empirical co-authorship network in Sec. 3.4.

3.2.3.3 Equations for maximising metadata likelihood for different distributions

Having understood how to infer suitable parameters for the edge and time components of our model, in this section, we elaborate the update equations necessary for several popular choices of distribution over the metadata.

Poisson (for positive integers) We commence with the Poisson distribution, which may be a suitable model for the distribution of career ages within a group. This is parameterised by λ, with probability mass function

$$\mathrm{p}(X = k) = \frac{\lambda^{k} \mathrm{e}^{-\lambda}}{k!}, \tag{3.25}$$

from which, if we allow variation between groups and over time, we find the update equation

$$\lambda_q^t = \frac{\zeta_q^t}{\xi_q^t}, \tag{3.26}$$

where $\xi_q^t = \sum_i \tau_{tiq}^m$, and $\zeta_q^t = \sum_i \tau_{tiq}^m x_i^t$. This is effectively just the weighted sample mean, as we might expect.

Categorical (for single-choice categories) If rather than ordinal data we have metadata belonging to particular categories, a intuitive distribution to use is the categorical distribution. That is, suppose that within each group, q, the likelihood of a node belonging to a particular category ℓ out of L options at time t is given by the categorical distribution,

$$p_q^t(\cdot) = \prod_{\ell=1}^{L} p_{q\ell,t}^{x_{i\ell}^t}, \tag{3.27}$$

where $x_{i\ell}^t = 1$ if node i belongs uniquely to category ℓ at time t and zero else. Including a suitable Lagrange multiplier in our objective function so as to ensure that $\sum_\ell p_{q\ell,t} = 1$, we find that

$$p_{q\ell,t} = \frac{\rho_{q\ell}^t}{\sum_\ell \rho_{q\ell}^t}, \tag{3.28}$$

where we have introduced $\rho_{q\ell}^t = \sum_i x_{i\ell}^t \tau_{tiq}^{\mathrm{m}}$. Indeed, in fact as all nodes must belong to one particular category and no others, we must have $\sum_\ell \rho_{q\ell}^t = \xi_q^t$.

Independent Bernoulli (for multiple-choice categories) Now suppose instead that metadata remains discrete categories, but nodes may belong to multiple categories simultaneously — useful for instance to describe the institutional affiliation of academics. An intuitive way of modelling this is to suppose that within each group, q, the likelihood of a node belonging to a particular category ℓ out of L options at time t is simply a Bernoulli distribution. That is, it has probability $x_{i\ell} = 1$ given by $p_{q\ell}^t$, where $x_{i\ell} = 1$ if node i belongs (not necessarily uniquely) to category ℓ and zero else, and thus probability $(1 - p_{q\ell}^t)$ of not belonging to this category. Further assume (quite strongly) that belonging to a particular category does not influence the likelihood of belonging to any other, *i.e.* each Bernoulli distribution is independent of all others.

Working through the details, and now with $\rho_{q\ell}^t = \sum_i \tau_{tiq}^{\mathrm{m}} x_{i\ell}$, we find that the equation is maximised when

$$p_{q\ell}^t = \frac{\rho_{q\ell}^t}{\xi_q^t}, \tag{3.29}$$

as we might predict. Not that while this appears equivalent to Eqn. (3.28), without the constraint that $\sum_\ell p^t_{q\ell} = 1$ we no longer have that $\sum_\ell \rho^t_{q\ell} = \xi^t_q$.

Multinomial (for counts of categories) Finally, if we instead posit a multinomial distribution over metadata, have

$$p^t_q(x^t_i) \sim \frac{n^t_i!}{\prod_\ell x^t_{i\ell}!} \prod_\ell p^{x^t_{i\ell}}_{q\ell}, \tag{3.30}$$

where $n^t_i = \sum_\ell x^t_{i\ell}$ is the total count of the given metadata observed for a particular node i at time t. This provides updates for $p^t_{q\ell}$ effectively identical to the categorical case, *i.e.*

$$p^t_{q\ell} = \frac{\sum_i x^t_{i\ell} \psi^{it}_q}{\sum_{i,\ell} x^t_{i\ell} \psi^{it}_q}. \tag{3.31}$$

Indeed, the counts observed for each individual do not enter the update equation directly, and only introduce a constant factor to the overall log likelihood with respect to the node labels. If desired, one could place *e.g.* a Poisson prior over these within each group, which amounts to considering the counts as additional metadata with the Poisson distribution updates as before. We provide additional details for negative binomial and normal distributed metadata in App. A.

To conclude our consideration of distributions over metadata, we comment that for each node at each time, after marginalising labels, these are effectively a mixture model of whichever type of distribution chosen. As such, previous works have established consequent identifiability – provided unique true parameters – for Gaussian and Poisson distributions [190], while Bernoulli distributions in particular are known to be problematic. However, in practice this typically does not cause issues [23].

3.2.4 Complexity

As currently introduced, the method scales relatively poorly with respect to comparable methods — typically as $\mathcal{O}(TQ^2N^2)$. Indeed, for every iteration of the inference process, the dominant part of the time is spent calculating the time-varying block parameters, of which there are $\mathcal{O}(TQ^2)$. For each of these we must perform sums over all pairs of nodes within that time period, typically $\mathcal{O}(N^2)$, hence an overall $\mathcal{O}(TQ^2N^2)$ as previously stated. This dominant contribution is the same as present for the model of [98].

Supposing we have S types of metadata, each with D_s parameters, to compute the full set of $QT\sum_s D_s$ metadata parameters, each of which has $\mathcal{O}(N)$ calculations, thus

takes $\mathcal{O}(QTN\sum_s D_s)$ time. As such, this term is negligible for large networks with few types of low-dimensional metadata, but can be notable if metadata is high-dimensional or there are many types. For instance, when considering academic networks in Sec. 3.4, we have $\sum_s D_s \approx N/3$, and so a reasonable proportion of the time is spent fitting metadata distributions, despite only involving nodes rather than edges.

The inference procedure is implemented in C++, and largely parallelised, so despite this high complexity, it is feasible to apply for networks with several thousand nodes. However, this is still at least an order of magnitude beneath what is possible using alternatives we describe in following chapters.

This poor scaling can be overcome in various ways, for instance constraining the model as we describe in App. A, or changing the inference approach. One immediate option is stochastic variational inference (SVI), *i.e.* only using a subset of the data to update the parameters in each loop, as performed in the static network case by [52] for instance. Alternatively we could take (i) a greedy approach – we later describe such for the DSBMM in Chap. 5 – or (ii) change inference method completely. Variants of MCMC are popular, as in [194, 94], and a multitude of static network cases, along with belief propagation, as in [50] for example. In later chapters, we elaborate how to perform belief propagation inference, and discuss how this allows us to investigate the detectability of groups.

3.2.5 Initialisation

All previous dynamic SBM methods have demonstrated that the initialisation of the clusters is highly influential on the quality of the final partition found. While popular, starting with completely random labels is typically found to be a bad strategy.

For the results in this chapter, we use the same method as [98]. We first horizontally concatenate the adjacency matrices of each timestep together, then perform K-means clustering on the rows to provide a homogeneous initial clustering across time. This means that our model generally performs better for networks with more stable groups over time, to which this initialisation strategy is better suited to approximating the true labels. We then run our model given this clustering, followed by a given additional number of runs on initialisations where some subset of nodes have their labels randomly permuted.

The labels inferred by the model from this set that has the highest model criterion value – ICL, as discussed in the following section – is then used as the initial clustering for one final full run, with a larger maximum number of iterations permitted, to increase the likelihood of obtaining a final high quality clustering.

There are a multitude of alternatives explored in the literature. For instance, [92] applies a conventional Poisson SBM on a matrix obtained from aggregating the networks at each timestep in several different ways. Once again, this hence results in a homogeneous clustering for nodes across time. This overall approach, of aggregating timesteps together before applying a static method, is popular, but unless we take an (expensive) multiedge-based labelling approach and keep a record of which edges were present at each timestep, the outcome is inevitably a homogeneous initial clustering over time. As stated previously, this means that for networks with lower group stability, any method commencing from such initialisations is likely to struggle.

To overcome this, one option would be to use an inexpensive method to fit static partitions, then apply some label matching procedure providing it is sufficiently cheap – *e.g.* as introduced in [143], which for our setup would take $\mathcal{O}(T(N + Q^3))$ time in the worst case.

In later chapters, we introduce both a greedy method, and a hierarchical inference procedure that provide viable alternatives, and importantly also account for nodal metadata.

3.2.6 Model selection

The principal downside with our method overall is that Q must be provided in advance as a parameter of the model, and thus to determine the 'best' number of groups, a multitude of models must be fitted for different choices and then compared (say by some information criterion) afterwards. Nonetheless, for the desired model features – in particular allowing non-generic transitions between groups – it is not trivial to instead make the model fully non-parametric. The method proposed for static SBMs by [154] is not immediately suitable, partly because changing the number of groups in a single time period in such a manner is unlikely to change the number of groups in the model as a whole.

One alternative is to use a relational extension of the infinite hidden Markov model of [13], as proposed by [65], but this appears to come with additional computational cost and possibly poorer performance. Instead, in this chapter we apply an information criterion like approach, approximating the ICL – Integrated Completed (or Classification) Likelihood – where 'better' models are assumed to have higher values of the criterion. In truth, this is effectively just the negative of the Bayesian Information Criterion (BIC) approach, where for a given model, $\mathcal{M}$, with

log likelihood given the data and inferred parameters $\mathcal{L}$, we take

$$\mathrm{ICL}(\mathcal{M}) = \mathcal{L} - \frac{1}{2}k \log n \tag{3.32}$$

where k is the number of parameters, and there are n pieces of data to be described. As different parts of our model are pertinent to different quantities of data, we consider each in turn.

Firstly, much as in [98], the penalisation term due to the transition matrix is given by

$$\mathrm{pen}_{\pi} = \frac{1}{2}Q(Q-1)\log(NT). \tag{3.33}$$

For binary networks, where we use Bernoulli edge distributions, the further penalisation term from the describing edge blocks is

$$\mathrm{pen}_{edges} = \frac{1}{2}Q \log\left(\frac{N(N-1)T}{2}\right) + \frac{1}{2}\frac{Q(Q-1)}{2}T \log\left(\frac{N(N-1)}{2}\right), \tag{3.34}$$

where the first term is for the within-group block parameters that are fixed across time for identifiability, while the second term is for the rest of the model.

If instead we are using a zero-truncated Poisson distribution, for instance to model count-based data as in Sec. 3.4, , the edge penalisation term becomes

$$\mathrm{pen}_{edges} = Q \log\left(\frac{N(N-1)T}{2}\right) + \frac{Q(Q-1)}{2}T \log\left(\frac{N(N-1)}{2}\right), \tag{3.35}$$

i.e. simply twice the Bernoulli case due to the new parameters.

Finally, if we are considering a fixed number of discrete categories for edges, as for the second empirical dataset considered in Sec. 3.4, we now have

$$\mathrm{pen}_{edges} = \frac{1}{2}Q(1+M) \log\left(\frac{N(N-1)T}{2}\right) + \frac{1}{2}\frac{Q(Q-1)}{2}T(1+M) \log\left(\frac{N(N-1)}{2}\right). \tag{3.36}$$

Nonetheless, if we have a significant number of categories, then this heavily penalises the number of groups, hence as for [98] we may also use the 'elbow' method applied to the complete data log-likelihood, *i.e.* look for points at which the gradient of the line between complete log-likelihood values for different numbers of groups changes significantly.

To account for our metadata distributions, we add

$$\mathrm{metapen} = \frac{1}{2}QT \sum_{s} D_s \log(N), \tag{3.37}$$

to the penalty term, where D_s is the number of parameters required to describe the type of distribution chosen for metadata of type s. This is the case as for each timestep (T total), we assume we have metadata for N nodes, and we have Q different distributions for each piece of metadata – one for each group – with D_s parameters each.

The 'best' model is then that which maximised the overall criterion. This allows us to compare between models with differing numbers of parameters. However, using this as a criterion to compare between our model and that of [98] without metadata is not immediately viable. This is because we have observations (or dependent variables) that we are trying to fit (*i.e.* the metadata) that the model neglecting metadata is not – to compare we have to introduce some penalty to the model of [98] for not describing this additional information.

The simplest way of doing so is to suppose that we do not use any network information to describe the metadata as in the conventional model, but do still propose metadata distributions – for instance, a Poisson distribution for count data, fitted with a maximum likelihood estimate of the mean say either on (i) all metadata of that type available, or (ii) all metadata of that type within each time period. These propositions would result in the new penalty functions

$$\text{metapen}' = \begin{cases} \frac{1}{2}\sum_s D_s \log NT & \text{Case (i)}, \\ \frac{1}{2}T\sum_s D_s \log N & \text{Case (ii)}. \end{cases} \tag{3.38}$$

If we now denote the log likelihood of the best model found with metadata by $\mathcal{L}_{wm}$, and without by $\mathcal{L}_{wom}^{(i),(ii)}$ for each of the cases above, we find that this criterion recommends the model with metadata if

$$\mathcal{L}_{wm} - \mathcal{L}_{wom}^{(i),(ii)} > \begin{cases} \frac{1}{2}\sum_s D_s(QT\log N - \log(NT)) & \text{Case (i)}, \\ \frac{1}{2}T\sum_s D_s \log N(Q-1) & \text{Case (ii)}, \end{cases} \tag{3.39}$$

where we have assumed the same type of distribution is chosen over each of the pieces of metadata whether or not we are accounting for network structure, hence D_s is the same for both models. As we might expect, as the complexity of the necessary model increases with any of the parameters, the quality of the fit of the model (as measured by the log likelihood for the best parameters found) must improve accordingly, else the simpler model is preferred. In particular, for our model to perform well, the likelihood function would have to improve notably when separating the metadata into groups, which themselves must have some relation (whether strong or not) to the observed

network structure. As such, this highlights that misalignment between metadata and edge structure may cause issues for our model, unless addressed suitably.

However, it is not always realistic to use such information criterion estimates to perform model selection, and indeed the underlying assumptions are often unrealistic for network models [141]. Furthermore we have had to pose some arbitrary choice of distribution over the metadata to allow comparison between our model and that of [98] in this metric. Instead, one alternative option for selection – without more fundamental modification of our procedure – is to choose between models based upon their ability to predict links, and/or metadata, as we explore in Sec. 3.4. We can then use *e.g.* estimates of the computational complexity of each model to determine whether the improvement in predictive performance is worth the additional time required.

3.2.7 Missing edges

As for all empirical data, in practice any measurement process or observation procedure used to collect the data in the first place is prone to error — for instance, some edges could be mistakenly recorded as absent, spurious edges may be added, or some nodes could be missing. As such, various recent efforts have sought to account for this suitably in the context of SBMs [62, 115, 174, 140, 168]. For a set of missing or possibly spurious edges, δA, that are assumed to be missing at random (MAR) from the real network, we should calculate

$$\mathrm{p}(\delta A \mid A, X) \propto \sum_{Z} \frac{\mathrm{p}(A \cup \delta A \mid Z)}{\mathrm{p}(A \mid Z)} \mathrm{p}(Z \mid A, X), \tag{3.40}$$

as discussed for the case without metadata in [174]. However, the normalisation constant within the sum, $\mathrm{p}(A \mid Z)$ involves a sum over all possible missing/spurious edges, and so is expensive to obtain in general, even were we to approximate the sum over all Z using a small sample (or even a single inferred labelling), and use our estimated node marginals for $\mathrm{p}(Z \mid A, X)$.

Instead, for the link prediction task performed below in Sec. 3.4, we take two approaches. Firstly, as is common in the literature (see *e.g.* [167]), to simply directly use the inferred parameters, and take $\mathrm{p}(a^t_{ij} \mid z^t_i, z^t_j)$ at some threshold to classify edges/non-edges. Secondly, we instead use our estimates of the node marginals to potentially improve our predictions, *i.e.*

$$\mathrm{p}(a^t_{ij} \mid \mathcal{D}) = \sum_{q,r} \mathrm{p}(a^t_{ij} \mid z^t_{iq} = 1, z^t_{jr} = 1) \tau^m_{tiq} \tau^m_{tjr}. \tag{3.41}$$

3.2.8 Tuning the importance of metadata

Of course, even when metadata is present, this may not always be actually useful when finding groups in the network. As such, rather than detrimentally affecting performance by fully accounting for such un- or mis-informative metadata, we should allow ourselves to tune its relevancy within our model. This is similar to the work of [132], though there they consider only categorical node attributes for the static case.

There are a variety of ways we might go about this. At the simplest level, we can posit that there is a single parameter governing the probability of using each piece of metadata, θ say, and if this switch variable is 'off', then the metadata does not further influence the likelihood function.

We introduce a new latent variable for each node at each time, $\tilde{z}_i^t$, and denote the metadata being directly incorporated as in our previous model by $\tilde{z}_i^t = r$, and switched off by $\tilde{z}_i^t = b$. We may then write the new likelihood of observing the data, A, X given all parameters of the model as

$$\mathrm{p}(A, X \mid \cdot) = \mathrm{p}(A \mid Z, \cdot) \prod_i (\theta \mathrm{p}(x_i^t \mid z_i^t, \cdot))^{\delta_{\tilde{z}_i^t, r}} (1-\theta)^{\delta_{\tilde{z}_i^t, b}}, \tag{3.42}$$

where $\mathrm{p}(A \mid Z, \cdot)$ and $\mathrm{p}(x_i^t \mid z_i^t, \cdot)$ are as previously defined. However, we can immediately see that when maximising this function with respect to model parameters, we would always want to take $\theta = 0$ as $\mathrm{p}(x_i^t \mid z_i^t, \cdot) \leq 1$.

Instead, we take an approach much as we did for our model selection criterion, to allow for comparison between the model without metadata and our own. That is, we can posit that with probability θ we generate the metadata conditioned on the inferred group of the node at that time, while with probability $1-\theta$ we instead generate this according to some null distribution. For instance, the two ungrouped cases previously suggested, *i.e.* either (i) a static distribution over all metadata of that type, or (ii) a distribution over all metadata of that type within that specific timeslice. Using $\mathrm{p}(x_i^t)$ to denote either of these two choices, the new log likelihood term for the metadata now becomes

$$\sum_t \sum_i \delta_{\tilde{z}_i^t, r} [\log \theta + \log \mathrm{p}(x_i^t \mid z_i^t)] + \delta_{\tilde{z}_i^t, b} [\log(1-\theta) + \log \mathrm{p}(x_i^t)]. \tag{3.43}$$

Taking the expectation over our variational distribution as before, we know the

probabilities $\mathrm{p}(\tilde{z}_i^t = r) = \theta$, $\mathrm{p}(\tilde{z}_i^t = b) = 1 - \theta$, so we have the modified term

$$\begin{aligned}
&\sum_t \sum_i \theta[\log\theta + \sum_q \tau_{tiq}^m \log p_q^t(x_i^t)] + (1-\theta)[\log(1-\theta) + \log \mathrm{p}(x_i^t)], \\
&= NT(\theta\log\theta + (1-\theta)\log(1-\theta)) + \sum_{t,i}\left[\theta\left(\sum_q \tau_{tiq}^m \log p_q^t(x_i^t)\right) + (1-\theta)\log \mathrm{p}(x_i^t)\right].
\end{aligned} \tag{3.44}$$

Now the second partial derivative of this term with respect to θ is $1/\theta + 1/(1-\theta)$, which once again is $\geq 0\ \forall\theta \in [0,1]$. Hence we appeal to the boundedness theorem – as the full term is uniformly continuous for θ in the given range, it is bounded on this set and achieves its bounds, and as $\frac{\partial^2}{\partial\theta^2}(\cdot) \geq 0$ throughout we know the maximum is at one of the two boundaries. At these points, we have

$$f(\theta) = \begin{cases} \sum_{t,i} \log \mathrm{p}(x_i^t) & \text{when } \theta = 0, \\ \sum_{t,i,q} \tau_{tiq}^m \log p_q^t(x_i^t) & \text{when } \theta = 1. \end{cases} \tag{3.45}$$

As such, we see that whether we use metadata or not effectively comes down to whether the metadata is more likely to be observed under our model, where it is assumed to be grouped in some way that also affects the network structure, or whether it is homogeneous (either throughout or within each timeslice). By allowing θ to take different values nonetheless, we can tune the groups inferred between that of our original model ($\theta = 1$), and that of [98] (with $\theta = 0$).

Most of the variational update equations remain unchanged, with the exception of that for $\tau_{tiqq'}$. There, we now have

$$\forall t \geq 2, \forall i \geq 1, \forall q, q' \in \mathcal{Q}, \quad \hat{\tau}\,(t,i,q,q') \propto \pi_{qq'}[p_{q'}^t(X_i^t)]^\theta \prod_{j=1}^{N}\prod_{l'=1}^{Q}\left[\phi_{q'l'}^t\left(A_{ij}^t\right)\right]^{\hat{\tau}_{\mathrm{marg}}(t,j,l')}, \tag{3.46}$$

where we highlight the change from Eqn. (3.12). In App. A, we consider the extension to separate tuning parameters for each group, but do not use this herein.

We conclude this section by again recalling that a desirable feature of our model would be to correctly learn whether metadata and/or the network are useful or not for inferring latent groups in the data, as for the method introduced in [113]. The case of no structure in either network or metadata is handled satisfactorily by the proposed method – in each respective case, we learn either a flat distribution over groups for the metadata, or a flat distribution over groups for the network. However, as we previously

suggested – and verify in Sec. 3.3.2 – hard-coding the dependence of metadata on a single group as we have means that the model may struggle in cases where there is meso-scale structure in *both* the network and metadata, but the resulting groups are not aligned. Using the tuning parameter introduced above allows us to lessen the detriment of this effect, as we demonstrate in Sec. 3.3.2.

3.2.9 Degree correction

As described in Chap. 2, one of the problems with using classical SBMs – problematic depending on the features desired from the latent groups given the application – is that the expected degrees, $d_i = \sum_j A_{ij}$, of all the nodes within each block are identical. As such, when inferring parameters, often the resulting clustering groups nodes in the network largely by degree, rather than incorporating any more nuanced structural effects. Conventionally, addressing this concern in the static case – incorporating 'degree-correction' [69] – involves two sets of parameters, a 'promiscuity' parameter for each node, θ_i, and a block parameter, λ_{ql}, such that the (multi-)edge A_{ij} is distributed like

$$A_{ij} \sim \text{Pois}(\theta_i \theta_j \lambda_{z_i, z_j}). \tag{3.47}$$

The formulation of this model does not sufficiently constrain the parameters, as any multiplicative factor of θ could be absorbed in λ, so typically some constraint is placed over θ. In particular, the maximum likelihood estimates for θ_i under this model, given the constraint that $\sum_i \theta_i \delta_{z_i,q} = \kappa_q$, where $\kappa_q = \sum_i d_i \delta_{z_i,q}$ is the sum of degrees within group q, are simply $\theta_i = d_i$, the node degrees.

With this observation, some papers, *e.g.* [113], directly use the degrees instead of leaving θ as a free parameter, and hence only solve for λ_{ql}. For our variational formulation, while leaving θ free might have caused some complications with how to define the constraint, by likewise directly substituting in the degrees we can also describe a degree-corrected version of our previous model.

Beyond substituting the new edge likelihood function above for $\phi_{ql}(A^t_{ij})$, the only new equation we must calculate if we do so is that for the block parameters λ^t_{ql} at time t. For parsimony, we define

$$m^\tau_{tql} = \sum_{i,j} \tau^m_{tiq} \tau^m_{tjl} A^t_{ij}, \qquad \kappa^\tau_{tq} = \sum_i \tau^m_{tiq} d^t_i, \tag{3.48}$$

using which the update equation for λ^t_{ql} is then

$$\lambda^t_{ql} = \frac{m^\tau_{tql}}{\kappa^\tau_{tq}\kappa^\tau_{tl}}. \tag{3.49}$$

Note in the case of perfect information/model fit, where $\tau^m_{tiq} = z^t_{iq}$, this recovers the conventional MLE for λ^t_{ql}, as introduced in the original paper [69].

3.3 Results on simulated data

In this section, we elucidate strengths and weaknesses of the proposed model through evaluating its performance on a variety of simulated networks.

We commence in Sec. 3.3.1 by producing directly analogous simulated networks to [98], so as to allow immediate comparison to their closely related model without metadata, before proposing several further tests of our own to further distinguish likely use-cases for each in Sec. 3.3.2. These highlight situations in which metadata may be helpful or harmful for inferring meso-scale structure within the network edges, as conventionally considered. We then conclude the section by comparing timings of the procedure for the model with and without metadata on networks of various size, in Sec. 3.3.3.

3.3.1 Comparing performance to a similar model without metadata

In this section, we assess the benefits of our model through comparison with that of [98], using tests reproduced from their paper. Note that on the test networks considered, the method of [98] was found to outperform that of other related dynamic SBMs (*i.e.* [194, 188]), hence we do not directly compare to these here.

Specifically, we set $Q = 2$, and generate undirected networks with binary edges directly from the model for $N = 100$ for a variety of time lengths, T, group transition matrices, π, and block connectivity probabilities ω. We take $T \in \{5, 10\}$, set π to be a simple matrix with probability $\in \{0.6, 0.75, 0.9\}$ of staying in the same group, else switching (corresponding to low, medium and high group stability), and ω as given in Table 3.1 – likewise defining tests at various levels of difficulty, *i.e.* more/less easily detectable groups. Probabilities for the initial group labels, α, are set to $1/2$, which is also trivially the stationary distribution of the Markov chain over the groups for all three chÃ¥oices of π, and so group proportions are roughly equal throughout time.

For each combination of parameters, we generate 20 networks to collect a set of results for each model.

We choose to vary π and ω as these are known to affect the detectability of groups in dynamic SBMs [50]. Groups that are less stable over time (as described by π) are more difficult to detect, particularly if they less well defined in terms of intra- *vs.* inter-block connectivity (as described by ω). Indeed, networks simulated with less stable groups cause even more difficulty in terms of inferring parameters accurately, given our current initialisation process – because it is designed on the assumption of relatively coherent groupings. We further vary the length of time to investigate how receiving more data affects the results of our model – we would predict that provided groups are relatively stable, the longer the time series, the higher the quality of results.

We also note that the final set of ω parameter values considered, 'Affiliation', *i.e.* ω defined by only two parameters, p_{in} and p_{out}, is separated from the rest due to known identifiability issues for this case, despite its popularity in the literature [98].

Table 3.1: Block connectivity parameters, ω, to reproduce tests in [98]

Difficulty	ω_{11}	ω_{12}	ω_{22}
Weak	0.2	0.1	0.15
Medium	0.25	0.1	0.2
Strong	0.3	0.1	0.2
Strongest	0.4	0.1	0.2
Aff.	0.3	0.1	0.3

Of course, our model further requires metadata to be defined over the nodes of the network in order to have any additional utility. As such, for tests in this section, we generate two types of metadata over the nodes: Poisson distributed, with parameters for each group at each timestep taken as uniform random integers between 3 and 12, and independent Bernoulli distributed over four categories, with uniform random probabilities of each category for each group at each timestep. We chose not to specify the metadata parameters further so as to allow the information content carried by the metadata with regards to the groups to vary across groups and time, and between each simulated network. However, importantly we have assumed that the partitions of the network into groups pertinent to the edges, and groups pertinent to the metadata, are identical. This is a strong assumption, no matter the information content of the metadata in terms of assisting in distinguishing between groups, and will be addressed in Sec. 3.3.2.

Given these choices, we expect to see an improvement in results given the presence of metadata, and indeed we demonstrate just how significant this improvement may be in the subsequent section.

3.3.1.1 Evaluating clustering quality

As a first means of assessing the performance of the two models, we evaluate their ability to recover the true node labels used to generate the networks at each timestep.

For comparability with [98], in this section we use the adjusted Rand index (ARI) [63] between the predicted and true labellings to quantify the degree of recovery, as introduced in Sec. 2.2.2 in Eqn. (2.17). We also evaluated the performance in terms of maximum overlap as in Eqn. (2.18), and the results were much as follows for ARI.

Importantly, most methods for comparing partitions are not immediately designed for the temporal case under consideration herein. Two dominant approaches to compare temporal partitions nonetheless are to either (i) calculate these values between the partitions at each timestep, and average the result, or (ii) flatten the temporal partition into a single vector, and again then apply whichever method desired.

These may provide significantly different assessments of model quality: the first approach allows labels to permute between each timestep, and so only compares the partitions locally, while the latter is a global measure across time. As such, it is much easier for models to perform well on the first measure, where it is only important to recover blocks within each timestep rather than the latter, where the correct evolution of these groups is also inferred – this is verified in our results below. The former is nonetheless used in many studies, *e.g.* [194, 188].

Proceeding now to our tests on simulated data, we display the results in Fig. 3.3. The upper row of subfigures corresponds to networks simulated with $T = 5$, while $T = 10$ on the bottom row. Low, and medium stability refer to the first two choices of group transition matrix, π, and within each subfigure we display results for each choice of ω, as displayed in Table 3.1. We do not show results for the highest level of group stability, as both models performed similarly, getting near perfect recovery for almost all experiments.

For each test, *i.e.* selected combination of T, π and ω, we display four points: the leftmost two show the global value of ARI, while the rightmost two correspond to the local ARI measure, averaged across 20 realisations of parameters, and 5 full runs of both models for each realisation. Each full run uses the whole procedure described above: we first perform several runs using a partially randomised initialisation, then a final run initialised using the best partition (in terms of ICL) found from these

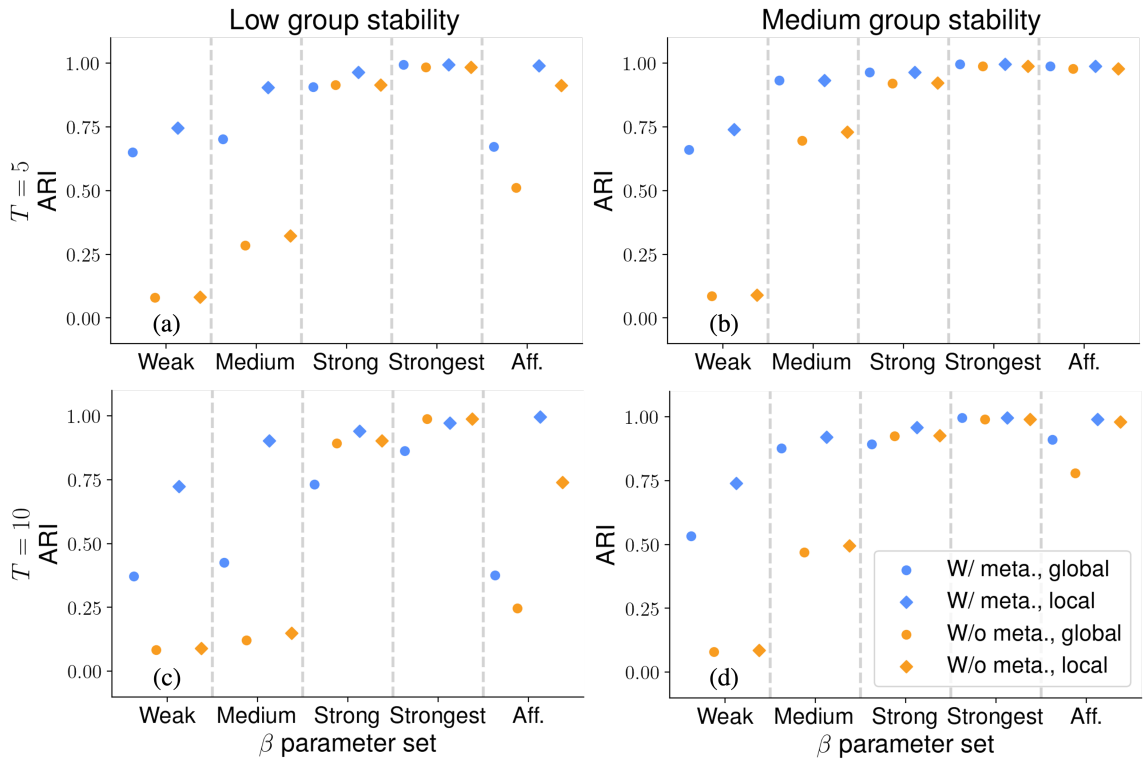


FIGURE 3.3: In this figure we present global (circles) and averaged local (diamonds) ARI results for our model (blue), and that of [98] (orange), with the true labelling. On the top row we display results for simulations with $T = 5$, while on the bottom row $T = 10$.

perturbed runs. Results for our model are shown blue, and those for the model of [98] – *i.e.* neglecting metadata – are shown orange.

The results verify that given alignment between the partition of the metadata and the underlying network, recovery of group labels can typically be significantly improved, particularly for tests where the model of [98] struggles. Taking the subfigure (a) as a case in point, that is results for simulated networks with poor group stability over five timesteps, for each choice of ω, we see clear improvements. This is especially drastic for choices of ω that have a smaller difference in connectivity probabilities between and within blocks, *i.e.* those marked 'Weak' and 'Medium', where metadata often allows high global recovery of labels, while ignoring metadata causes the model of [98] to fail almost completely.

The only two possible exceptions to this general observation are global ARI results for longer time series with better separated groups (*i.e.* easier ω, the choices 'Strong' and 'Strongest'), but low stability, as shown in subfigure (c). For these cases, we find that while including metadata does result in an improvement on average, there is

a much wider range – while ignoring metadata typically allows good recovery, on occasion including it seems to be detrimental.

The nonetheless high average local ARI values for these tests suggest that the metadata helps guide the correct allocation of nodes into local groups rapidly, but this then results in a local optima that is more difficult to escape and infer the correct overall sequence of labels. This is likely due to the current initialisation process being better designed for coherent groups, as labels are initially constant across time. This is a strong incentive for changing initialisation procedure to allow changing groups over time, as discussed at the end of Sec. 3.2.5.

Other results are much as predicted — for easier tests (more stable, better defined groups), both models typically perform well, and longer time series generally reduce the variance in results, as might be expected from the provision of more data. Both models struggle to recover parameters for networks drawn from the affiliation model in the case of low group stability —- as might be expected from identifiability concerns, and likewise found in [98].

Finally, we note that while all parameters used are relatively far from the theoretical detectability limit for groups [50], it has been demonstrated that VI-based methods can often struggle even in these cases, despite often expressing high confidence in their clusterings in terms of inferred node marginals [199]. This is another reason we change to belief propagation later in this work, which typically has the confidence of the model in its labels aligned with its actual accuracy therein, as well as performing better closer to the detectability limit. Of course in practice, empirical networks are often very sparse, with denser modules, so this is not necessarily as much of a concern as the scaling of the method previously discussed.

3.3.1.2 MSE comparison of transition matrices

Beyond solely contrasting the quality of clusterings, however it is measured, we should also evaluate the quality of fit for other parameters. Perhaps most pertinent of those that are shared between the two models are the transition matrices – these provide information about the how the groups evolve, *i.e.* which are more or less stable, which are more closely related (in that there is frequent movement between them), which are core groups and which are transitory *etc.*, all of which is highly valuable for a multitude of real-world applications. Further, the performance of this parameter of the model is less strongly correlated with the clustering quality than that for edge parameters, ω.

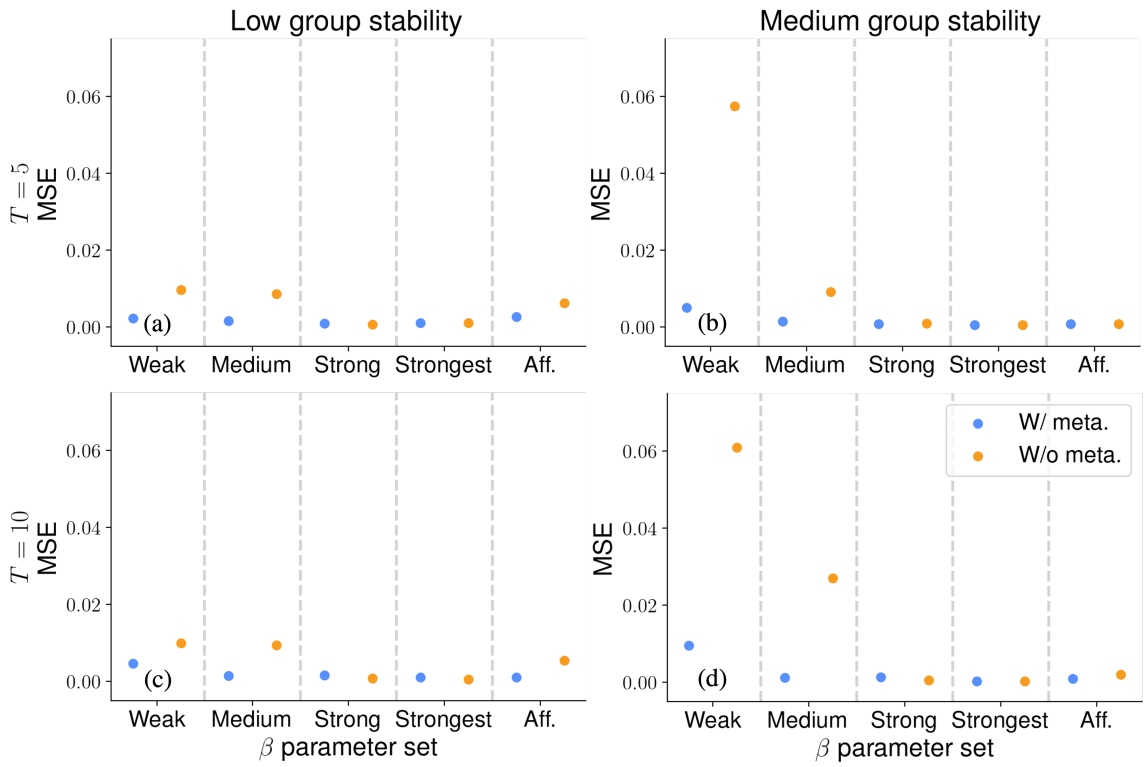


FIGURE 3.4: In this figure we present MSE from the true transition matrix for our model (blue), and that of [98] (orange), for the tests described in the text. On the top row we display results for simulations with $T = 5$, while on the bottom row $T = 10$.

To quantify the accuracy of recovery for the transition matrix, π, we use the mean-squared error (MSE), $\|\pi_{pred} - \pi\|_2/Q^2$.

We present the results for both models in Fig. 3.4, following the same styling as for ARI previously – given the initialisation process previously described, our model results are shown in blue, while that of [98] is given in orange.

The findings are overall as we would expect given the clustering performance displayed previously, in that on average our model outperforms that without metadata, with the possible minor exception (in terms of range of results) of more separated, unstable groups over a longer timescale.

Interestingly, the recovery of the transition matrix, and so the dynamics of the groups for these cases nonetheless appears to be quite good, suggesting any permutation errors that are causing poorer performance in the global labelling have cancelled out in some way. The other observation of note is that for both models, the recovery of poorly separated groups with medium stability (the right two subfigures) appears to be more challenging than the same for low stability. We hypothesise that this originates from nodes changing groups relatively often, but not having sufficient separation to define

the groups well. As such, the model gets stuck incorrectly inferring group dynamics – likely because subsequences of the network time series use different permutations of labels (*e.g.* group A as 1, group B as 2 for the first few timesteps, then switch to 2 and 1 respectively) for their respective SBMs – and struggles to escape this local optima. For less stable groups, such permutation difficulties are more easily overcome, as it is less likely to fit coherent subsequences in the first place, and thus the barrier to escape such optima is lower.

3.3.2 Model performance when metadata is unhelpful

For all tests above, we have assumed that the partition of metadata groups is perfectly aligned with that of network groups. However, as various papers have focused on in recent years, perhaps most prominently [132], this is often not the case for real networks. Indeed, the worst case scenario for our model is well separated groups within the metadata, completely misaligned with the groups in the network. In this situation, the two partitions would each suggest different labellings of the nodes, and so we would expect the quality of the singular labels inferred to be detrimentally affected. As such, in this section we explore how the alignment between the two partitions can influence the results, in this case of well separated metadata groups.

As our model nonetheless performs well for simulations of well-separated, stable groups even with misaligned metadata, we restrict ourselves to presenting four tests that are all more difficult. We now set $Q = 4$, and keep $N = 100$. For each test, once again we generate Poisson metadata, and independent Bernoulli metadata for four categories, simulating 20 realisations of each combination of parameters. We now fix the Poisson parameters to be five times the numeric group label (*i.e.* $[5, 10, 15, 20]$), and the independent Bernoulli parameters to be constant across all categories, then four evenly spaced values between 0.1 and 0.9 inclusive. As such, significant meso-scale information is now carried in the metadata, so we expect that whether the groups are aligned or not will be critical. We qualitatively define the tests as follows:

1. Stable, less well separated groups, with aligned metadata
2. Less stable, less well separated groups, with aligned metadata
3. Less stable, more well separated groups, with misaligned metadata
4. Less stable, less well separated groups, with misaligned metadata

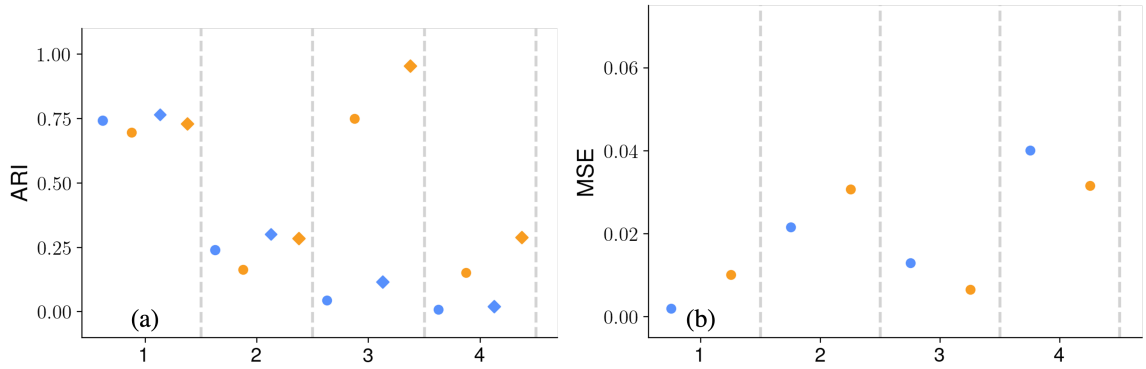


FIGURE 3.5: In subfigure (a) we present global (leftmost two points) and averaged local (rightmost two points) ARI results for our model (blue), and that of [98] (orange), with the true labelling for the metadata alignment tests described in the text. In subfigure (b) we present MSE of the inferred transition matrix to the true matrix for these same tests.

Following the same styling as for previous figures, we display the results of the two models on these tests in Fig. 3.5. In subfigure (a) we present the global and locally averaged ARIs between the partitions inferred and the true labelling, and in subfigure (b) the MSE between the transition matrix inferred and the true matrix. Note, importantly in these initial results we consider the 'true' labelling to be solely that used to generate the network structure, and neglect entirely the partition used to generate the metadata – as such, we are not capturing the extent to which the model is finding a partition suitable for metadata *vs.* edges, which requires more detailed exploration. We return to this shortly.

Much as in the previous set of tests, we see that for test 1, the easiest of the four, using the aligned metadata has helped our model often recover the true parameters near perfectly, while the model without metadata has more difficulty. For the second test, a challenging scenario, the model without metadata generally almost completely fails, while our model can typically recover the true parameters to a good degree of accuracy. However, for test 3 the roles are now reversed – the model without metadata often performs near perfectly, while ours seriously struggles – and for test 4, where the model without metadata fares poorly, using metadata results in near complete failure.

While this clarification of the significance of the alignment between structure in the network, and structure in the metadata, may initially seem alarming, thankfully this is not necessarily a significant problem. This is for two principal reasons: firstly, in empirical networks, given some sensible selection of metadata to consider, it is not hugely common for total misalignment to occur – metadata may be helpful in some

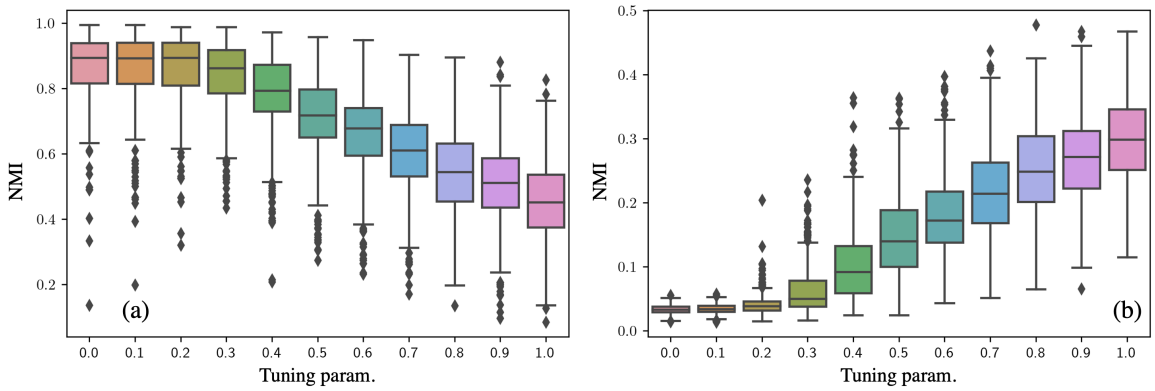


FIGURE 3.6: In subfigure (a) we present ranges of NMI between the true network partition and that inferred for different tuning parameter values. In subfigure (b), we perform the same tests, but compare ourselves to the *product* partition of both metadata and network groups.

areas of the network, and less helpful in others, or it may lead to *different* groups to those that most coherently describe the network structure alone, but this does not necessarily mean such groups are worse. Secondly, and particularly related to the former of these points, we can infer the model for a range of tuning parameters, as permitted through the extension of Sec. 3.2.8. Should it be the case that metadata is genuinely harmful when it comes to inferring groups, we can simply neglect its contribution entirely.

We demonstrate this in Fig. 3.6, where we focus on the third of the experiments above – where our model seriously struggled, while that without metadata performed well – and allow the tuning parameter to vary from zero to one. In subfigure (a), we show boxplots of the NMI between the partition inferred and the true network partition over 20 different realisations, where the box shows the median (inner line) and inter-quartile range (boundaries), and whiskers show the 5-95 percentiles — values outside of this (potential outliers) are shown as points. We observe the desired behaviour — a steady improvement as we decrease the tuning parameter and 'switch off' the metadata contribution. However, in subfigure (b), we show the same tests, but now comparing to the *product* partition of both metadata and network groups, that is the partition where each node is labelled with the tuple $(q_{\text{net}}, q_{\text{meta}})$. We now observe that increasing the tuning parameter makes the model better fit the metadata groups, and so trend towards this product partition — but as there are insufficient groups used to pick this up, instead there is a trade-off and the model results in fitting neither partition well.

As we demonstrate in Sec. 3.4, in general the relation between model performance (however measured) and the tuning parameter is not as smooth as in these simulated examples — we believe that this is due to the varying importance of metadata across the network. That is, if metadata carries information meaningfully pertinent to network structure in some areas but not others, then a lower weighting allows this to nonetheless carry across into the labelling when most helpful, without being too detrimental when it is not.

3.3.3 Scaling evaluation

We conclude this section with an evaluation of how our method as currently implemented scales compared to that of [98]. To do so, we simulated simple planted partition models for a range of network size, N, where for realism we take $Q = \text{floor}(\log(N))$, then time the full runs for each model.

The dimensions of metadata considered are fixed to relatively low numbers rather than scaling with N, and so metadata does not provide a significant contribution to the complexity, though this can occur, as discussed in Sec. 3.2.4. This is a reasonable assumption for most types of metadata considered within this work – *e.g.* career age, institution, subject area – given constraints applied to the network (*i.e.* for academic networks, restricting our focus to a particular geographic region and time period), but this may not always be the case. For instance, without constraints the number of unique institutions typically does grow as the number of unique authors increases.

As both models presently use the same variational inference procedure, recall that the time complexity is dominated by calculations over all possible edges ($\mathcal{O}(N^2)$) for each block pair ($Q(Q+1)/2$) at each timestep, hence the $\mathcal{O}(TN^2Q^2)$ complexity as previously discussed.

Indeed, as demonstrated in Fig. 3.7, we can verify that the two methods scale as anticipated, with the relative increase in time required to incorporate metadata terms diminishing as the size of the network increases, and the quadratic complexity term further dominates.

Finally, ambiguities in timings are the result of performing calculations on a shared machine under heavy workload, hence causing some variance in computation speed between different runs.

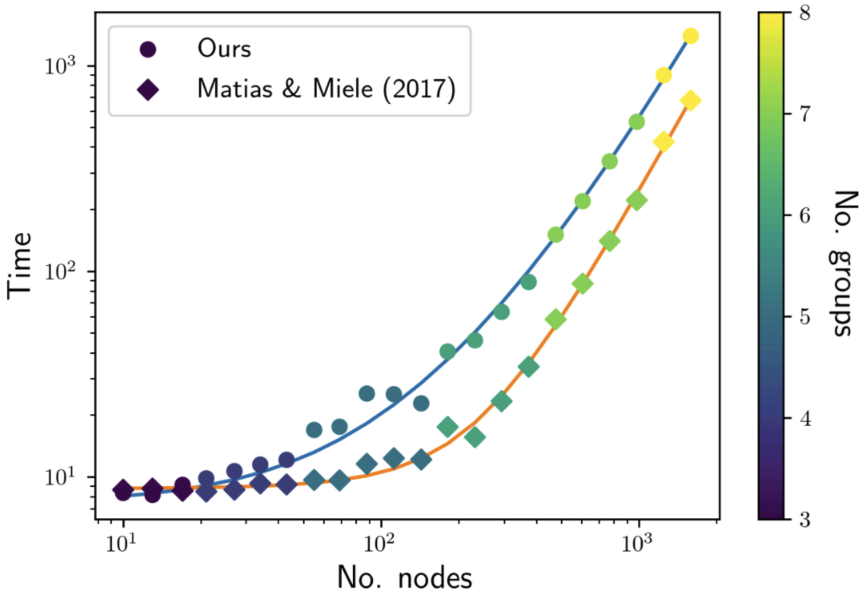


FIGURE 3.7: Here we display a scaling comparison between the two models, with $Q = \text{floor}(\log(N))$ to represent realistic values, and total time as measured on CPU. Curves shown are fitted to quadratic N^2Q^2, linear NQ, and intercept terms, as expected from the overall algorithmic complexity, and show close agreement.

3.4 Application to empirical data

In this section, we apply the proposed model to three different empirical dynamic networks with metadata. First, in Sec. 3.4.1 we consider two networks obtained from a dataset of publications with authors affiliated to institutions in Colombia, then in Sec. 3.4.2 we instead focus on a small network from a longitudinal sociological study of students, as introduced by [175]. We briefly investigate some of the utility our model can provide for each of these cases, but leave more in depth exploration to further work.

Before proceeding, we again note that using a model which accounts for metadata rather than network structure alone to group nodes should not solely be interpreted as providing partitions that are better/worse than models without metadata. Indeed, increasingly it is recognised that networks frequently permit various different partitions that perform well for different models, which can describe different types of organisation within the network [143, 76]. Further, metadata groups may not align well with meso-scale structure present within the edges of the network, as investigated by [132]. As such, it is often better to generally think about different models providing *different*, rather than improved, insights into the empirical data at hand, as there is no real ground-truth. Nonetheless, we can evaluate models performance on various metrics,

for instance through link prediction, collaborative filtering, the human interpretability of resulting groups or other measures. The ideal goal, which our tuning parameter lets us explore in greater depth, is to make the most of metadata when and where it is useful to do so, and discard otherwise – we show how we may perform the first steps towards accomplishing this in Sec. 3.4.4.

3.4.1 Networks of Colombian authors

In this section, we consider two original empirical dynamic networks with metadata constructed from a publication dataset from the digital repository Scopus, obtained through our collaboration with Elsevier. Explicitly, we construct (i) a co-authorship network, where academics are connected by weighted edges according to the number of publications they have produced together in the period considered, and (ii) a citation network, where academics are connected by weighted edges according to the number of times they cite one another. As the weights are positive integers, we use the ZTP distribution over edges as introduced in Sec. 3.2.3.1.

There is a rich literature exploring networks constructed from publication data, particularly since the seminal studies of [108] and [11]. With regards to inferring latent network groups, these have been used to understand changing research communities (as latent groups in author-level co-authorship/co-citation networks), describe research topics (groups in publication-level co-citation/keyword co-occurrency networks), and more. Some previous works have also explored differences in the distribution of various metadata within such groups, for instance gender in co-authorship networks, but have not generally used this additional information to help define the groups in the first place. As such, we expect the blocks of our model to provide more human-interpretable/meaningful coarse-grainings of networks than most previous findings — this is in addition to expected resolution and quality improvements, as found in the previous section.

In the case of academic networks, inferring distributions over the metadata within each group further allows us to better understand how certain factors vary across different research communities. These could help guide the discovery of communities with certain desired characteristics for authors and research policy-makers alike, for instance those active in a certain field, with stronger connections to certain institutions, and with a preference for academics at an earlier stage in their career.

Within this section, specifically we acquired a dataset of several million publications published between 1996 and 2020, where the only further constraint was that

authors were affiliated to institutions in Latin America and the Caribbean. To allow consideration in discrete time intervals, we chose to window this corpus into 3-year time intervals, starting from 1997-1999, so as to include the most recently available data. This window length was chosen so as to increase the likelihood of capturing publications (and so collaborations, citations *etc.*) in each period between authors that occurred specifically for events such as semi-annual conferences, as well as being a common timeframe for *e.g.* PhD student training and subsequent collaboration and similar.

3.4.1.1 Data

It is, however, infeasible to apply algorithms with anything greatly over a linear scaling to such a considerable dataset, and even such algorithms would take a reasonable amount of time to run, thus preventing rapid data exploration as is the aim herein. As such, we applied several further significant constraints:

- Firstly, we required that among the subject areas of the publication destination (journal, conference proceedings *etc.*) – metadata provided by Elsevier – were any of Maths, Physics, or Computer Science. The assumption was that these disciplines have some commonalities in terms of academic behaviour, and are more likely to have interdisciplinary collaborations between them than many other subjects.

- We next removed hyper-authored papers, defined here as papers with 50 or more authors – these are frequent in topics such as high-energy Physics, where large teams collaborate in using *e.g.* colliders or similar, but skew the dataset significantly, particularly when considering co-authorship networks. Some approaches to address this problem considered in the literature are (i) fractional counting, where the contribution to the strength of connections between two authors from co-producing a paper is inversely weighted by the number of authors on the paper, (ii) instead considering bipartite author-publication networks, and (iii) author-level hypergraph models, where this publication would nonetheless only constitute a single hyperedge, but we do not consider such in this exploratory work.

- As we intend to consider author-level networks, we removed authors with fewer than 5 publications in the subsequent dataset, so as to prevent consideration of highly peripheral nodes. Note that this does not mean we discard publications involving these authors, only edges (however defined) connecting to them.

- We then aimed to find a country in the dataset which included a sufficient number of authors to be interesting (*i.e.* at minimum $\mathcal{O}(100)$), but not so many that scaling problems in the model as currently implemented would prevent its application. To construct these country datasets, we required both authors involved in a connection to be affiliated to institutions located in the country in question – as such, a significant number of publication records are discarded. The optimal country appeared to be Colombia, hence its selection. Note that authors with affiliations to multiple institutions, one of which is located in Colombia, are still included.

- We are left with a set of 1954 unique authors. We consider both (i) a (undirected) co-authorship network, where weighted edges correspond to counts of the number of publications a pair of authors have produced together in each time period, and (ii) a (initially directed) citation network, where a weighted connection from author i to author j in time period t counts the number of times author i cited author j in publications produced in that period.

- To further limit the dataset, and improve the quality of results, we finally require that each author considered in the two respective networks has at least 5 connections therein, and take those present in the largest connected component in the most recent time period, 2018-2020. This results in 522 authors in our co-authorship network, and 578 authors in our citation network – small enough for us to proceed.

Within the windowed series of network snapshots, we choose to include the following metadata for each author at each time period, where available:

- Unique subject areas, after discarding frequencies — thus suitably modelled by independent Bernoulli distributions. For both co-authorship and citation networks, we have 24 such subject areas as provided by Elsevier present.

- Career age, defined as the number of years since first publication, and taken as the median value over publications within the period — we choose to model this using a Poisson distribution.

- Unique institutional affiliations, again after discarding frequencies. Within the co-authorship network, there are 149 such institutions, while the citation network has 173 institutions present. This high number of categories relative to the size of the network means that a non-negligible portion of the time fitting the model

is spent calculating metadata parameters, as discussed in Sec. 3.2.4, and so neglecting the metadata would noticeably increase the pace of inference.

These metadata were selected as various previous studies have demonstrated that they play a significant role in both collaboration and citation, as one might expect.

Discarding count data, we find that there is 99% reciprocity in the citation network in the most recent time period, hence we consider the network undirected and use the undirected model instead, to reduce the number of parameters. Furthermore, the density of both networks was incredibly low for early time periods in the dataset. Requiring that there be at least $\mathcal{O}(N)$ edges in each time period, we were left with the four most recent time periods for the co-authorship network (*i.e.* 2009-2020), and only the two most recent periods for the citation network (2015-2020).

To get an immediate sense of the distributions of these pieces of metadata, in Fig. 3.8 we display counts/histograms for each within the full Colombian dataset. Here, a single author present at time period t contributes a count to each piece of metadata they possess, *i.e.* authors present at multiple timesteps have their metadata included multiple times – this is important to allow, as the metadata often changes (in the case of career age, of course always).

Several other potential concerns arise from this figure that we address before continuing. Firstly, there is significant class imbalance in both pieces of categorical metadata – this is not necessarily an issue for smaller categories, as it could mean that those less represented classes are simply more informative about node groups, but it could mean that the larger categories are less helpful. However, in such cases, the *simultaneous* inclusion of multiple categories still varies across the dataset — we anticipate that indeed, while Physics for instance is highly represented in the dataset, perhaps specific combinations of Physics and Healthcare constitute a well-separated group. Secondly, on aggregate the Poisson distribution is not ideally suited for career age — it is both over-dispersed and zero-inflated, and both of these features, particularly zero-inflation, could be addressed more readily by other distributions. The use of negative binomial distributions, as described in App. A, is one option, though it would require additional computational complexity — a cheaper viable alternative as described in the same appendix, would be to use log-normals. Nonetheless, we find that the simpler Poisson is sufficient in practice — that is, the mixture of Poisson distributions resulting from the groups is sufficiently representative even in this case.

Certain other model assumptions were also assessed. For instance, we used the Predictive Power Score to explore the connections between these pieces of metadata

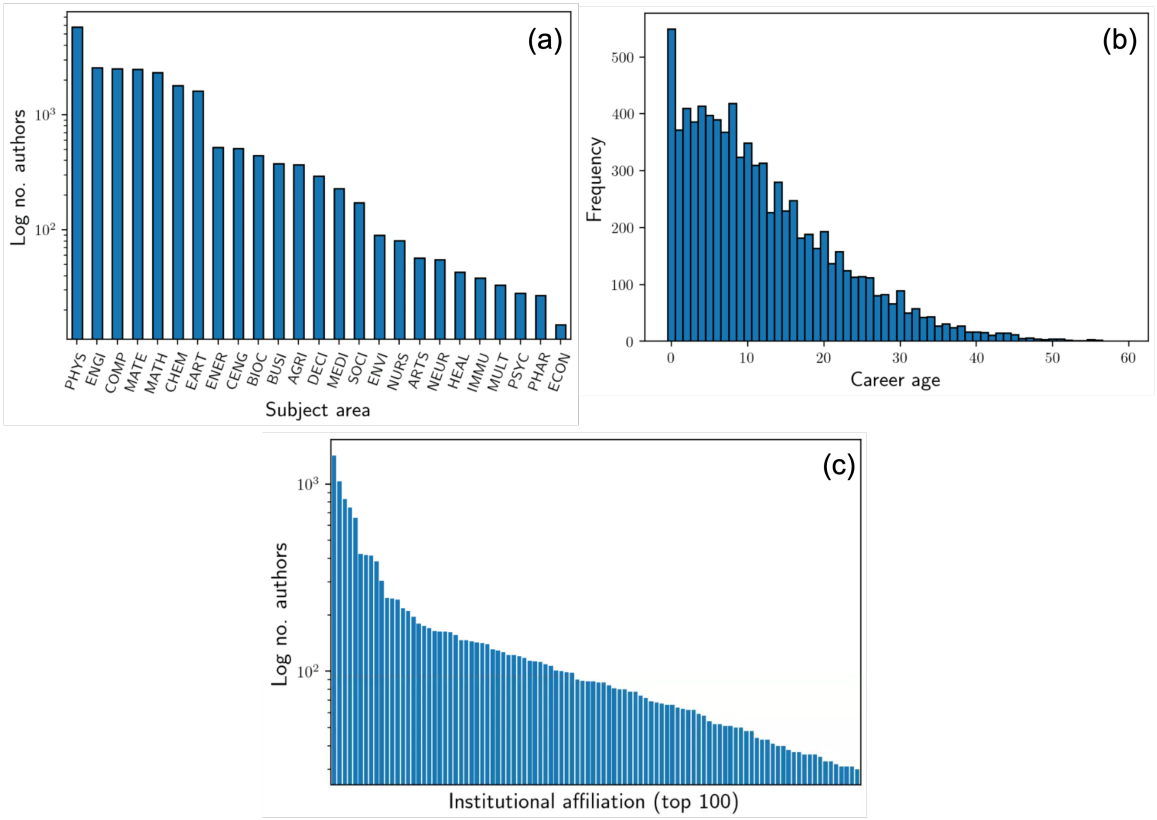


FIGURE 3.8: In this figure we display observed frequencies of each piece of metadata considered in our full Colombian dataset, where the metadata for each author present in each period is counted. For institutional affiliation counts in subfigure (c) we display only the top 100 institutions by count, and neglect (unimportant) identifiers, as more important is the shape of distribution.

[182]. This suggested that our assumption of the independence of metadata does broadly seem to hold true.

3.4.1.2 Results

As the full inference procedure takes a significant amount of time, we wish to restrict ourselves to commencing the search around a reasonable estimate for the number of groups present. As such, we first apply the greedy Leiden algorithm [172] to find groups approximately optimising Newman-Girvan modularity [114], and investigate in the area around the resulting number of communities. The method found 26 communities for the co-authorship network, and 21 communities in the citation network, and hence we explore the ranges 22-27, and 19-24 respectively. The Leiden method prioritises finding assortative communities, rather than stochastically equivalent blocks, and so is not necessarily representative of the types of groups fitted by our model, but we have

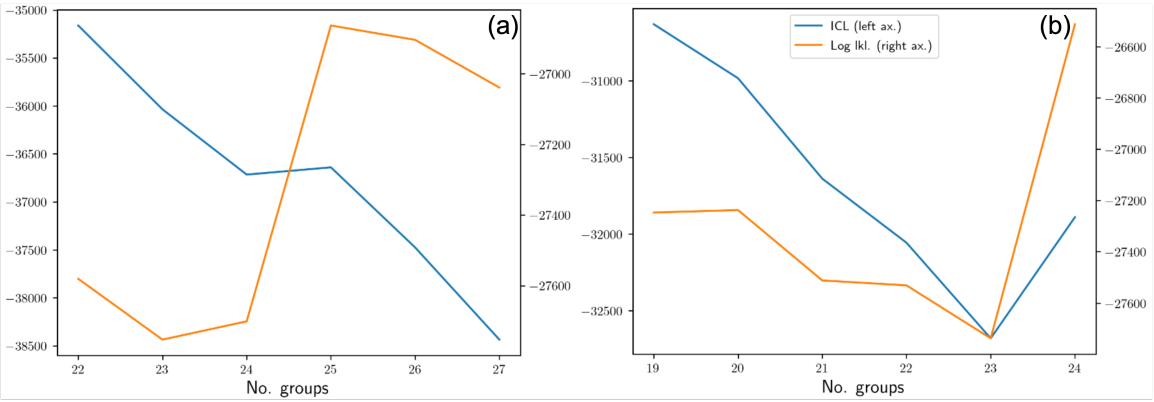


FIGURE 3.9: In this figure we display ICL (with scale as defined by left tick values for each axis, blue) and log-likelihood (scale shown by right tick values, orange) for the best model (maximum ICL) inferred for each value of Q, for each of the Colombian networks. Subfigure (a) is that of the co-authorship network, and subfigure (b) the citation network.

found that nonetheless in practice the number of groups it infers provides a reasonable baseline.

Following our initialisation procedure with 10 randomly perturbed starts, and hence 11 total models for each number of groups permitted, we plot the ICL and log-likelihoods of the model with the highest ICL value for each value of Q in Fig. 3.9, for each network – subfigure (a) for co-authorship, and (b) for citations. Note the vertical axes are at different scales and shifts for ICL and log-likelihood values in both figures, as the important aspects to compare are the shapes of the plots rather than their values.

While not a perfect measure, as previously discussed, we choose the model inferred with the greatest ICL value — as such, we use $Q = 22$ for the co-authorship network, and $Q = 19$ for the citation network. We note that the non-monotonicity observed in the log-likelihood is not typical for such problems — usually, as the number of groups specified at inference time is increased, naturally the model is able to fit itself more closely to the data, as the number of parameters increases quadratically, and thus the model has sufficient flexibility to increasingly perfectly match the observations. However, in doing so, the significance of any groups inferred decreases, along with their utility when generalising to unseen data (for instance predicting new links). Indeed, this risk of overfitting is precisely why alternative metrics such as the ICL are used.

More experiments are required to verify if such behaviour occurs frequently in other scenarios, but the likely origin is in the increased difficulty in converging to something close to the global optimum when simultaneously modelling metadata,

but requiring groups for both network and metadata structure to align. That is, as previously discussed, we have introduced potential conflict within the objective between fitting network and metadata groups, should these not perfectly agree.

As such, now when Q increases, we may find that this leads to an increased likelihood of the group distribution converging to a local optima that prioritises network over metadata structure, which then results in a poorer fit for metadata groups, or *vice versa*, and subsequently an overall decrease in log-likelihood relative to fewer groups. One intuition for why this might occur in practice is if there are Q_{net} groups in the network, and Q_{meta} groups in the metadata, which only overlap slightly. As such, ideally we may wish to fit $Q_{\text{net}}Q_{\text{meta}}$ groups overall — however this might be very large, and thus greatly over-complicates the model (and slows inference). Instead, as we increase the number of groups from a smaller value, $Q_{\text{test}} > \min(Q_{\text{net}}, Q_{\text{meta}})$, it could be that values closer to multiples of either Q_{net} or Q_{meta} lead to a tendency towards prioritising one over the other, without sufficient flexibility to optimise for both. Steps to alleviate this potential concern include (i) improving the optimisation procedure, as performed in following chapters, (ii) use of the tuning parameter introduced, (iii) simply perform more runs with different initialisations for each Q value, and (iv) improving the quality of the initialisation, especially in its ability to jointly optimise for network and metadata structure.

Returning to the current dataset, for each network, using the Q values with highest ICL of the range considered, we subsequently fit the model of [98] – that neglects the metadata present – to the networks for comparison. We present results for the co-authorship network in Fig. 3.10. In subfigure (a) we show the blocks inferred for the network in the final time period using our model, while in (b) we show those for the model of [98]. Subfigure (c) then shows the NRMI [112] between these partitions at each timestep, a measure of their similarity.

Comparing subfigures (a) and (b), as expected we observe that the inclusion of metadata has refined the groups detected. This is particularly clear for more peripheral authors in the network, where ignoring metadata has resulted in all such authors being placed in a single group, while accounting for differences in institution *etc.* has allowed us to better uncover truly similar groups.

We note that a useful measure for understanding this is comparing the effective number of groups inferred, defined as

$$Q_{\text{eff}} = \exp\left(-\sum_{q} \frac{n_q}{N} \log \frac{n_q}{N}\right), \tag{3.50}$$

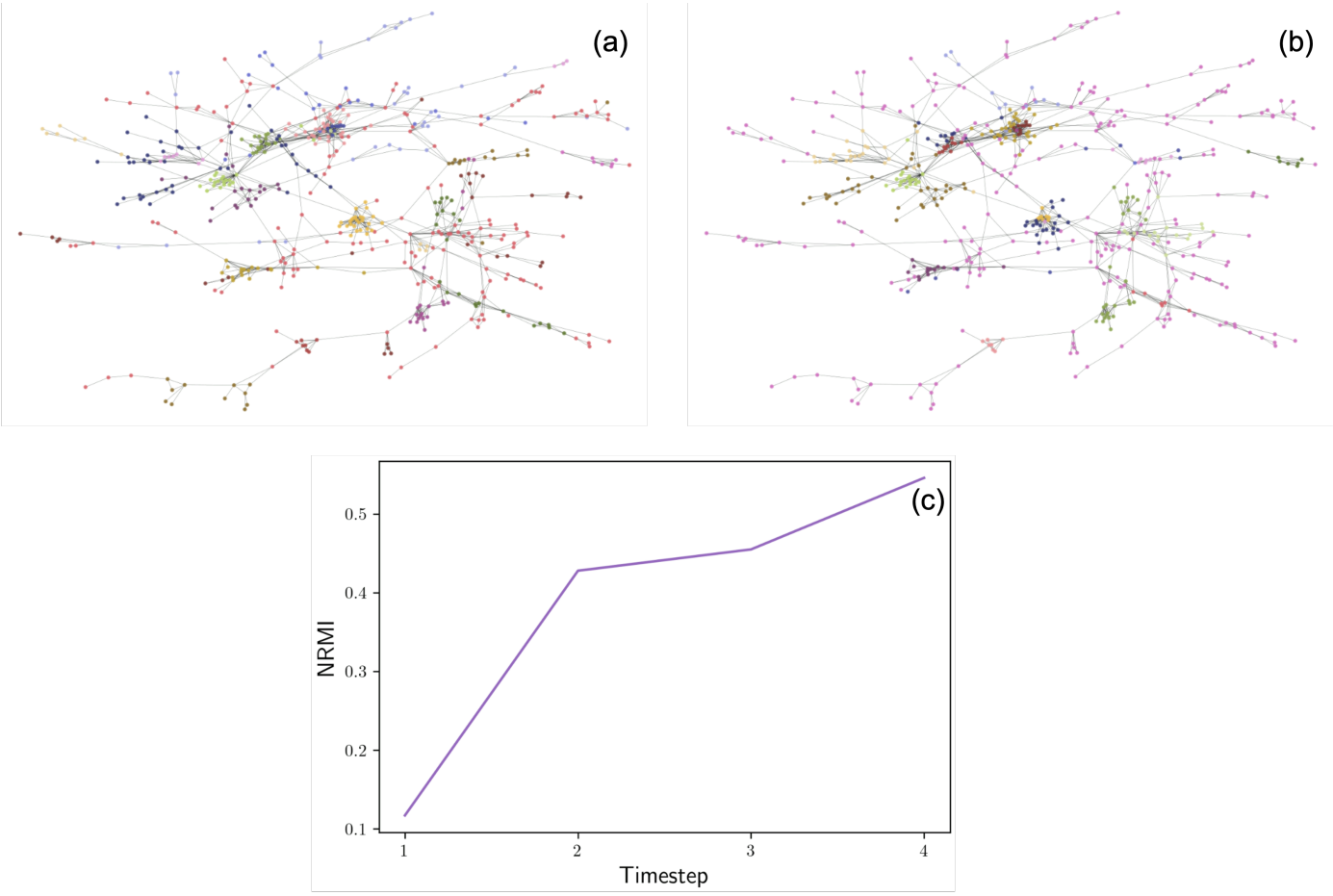


FIGURE 3.10: Subfigures (a) and (b) show the blocks inferred for the network at the final timestep for the co-authorship data, using our model and that of [98] respectively with $Q = 22$. In subfigure (c) we display the NRMI between the partitions inferred by the two models at each timestep.

i.e. the exponential of the entropy of the distribution of group sizes, n_q. Were all groups of equal size, this would be identical to the actual number of groups, Q, but if this is not the case then it better summarises the partition if *e.g.* there are a few large groups that almost all nodes belong to, with remaining groups being near singletons. Throughout time, we find that including metadata results in a larger Q_{eff}, even when Q for both models is identical. This is particularly the case for earlier timesteps, where there is insufficient network data to meaningfully distinguish between authors.

For both models, we may also observe that some clustering is performed due to the degree of the nodes in the network, the 'problem' for which degree-corrected models were introduced [69]. However, in this particular case it is somewhat informative to separate group authors with higher degree, as such authors have a wider range of collaborations and/or citations than average and so may be of particular interest.

Indeed, in Chap. 6, we demonstrate that these non-degree-corrected groups can be more useful for certain tasks.

When metadata is neglected, the real problem is that such degree-based partitions result in groups with members located in disparate locations in the network, but much as for the issue of peripheral nodes, through the inclusion of metadata this is largely resolved. In following chapters, when interested in groups that may define *e.g.* research communities, we typically 'correct' this using the method proposed in Sec. 3.2.9.

Finally, in subfigure (c) we see that the partitions inferred are meaningfully different between the two models throughout time. As we might expect, as more edge data becomes available and so the relative importance of metadata wanes, the two partitions do become more similar, but as described above there are still significant differences.

Now proceeding to the citation network, we present analogous results in Fig. 3.11. The benefits of our model, as found in the co-authorship network, are broadly reproduced in this dataset – peripheral author groupings are somewhat more refined, there is some clustering according to degree, and the partitions are more similar when the network is more dense. However, in this case we actually find that Q_{eff} is comparable over time between the model. While some peripheral groups have been refined, other groups that were distinct in the model without metadata have now merged. This highlights that incorporating metadata does not always provide a finer-grained view of the network, and that when justified may unify groups with some structural differences in the network.

As previously suggested, one of the benefits of our model is that beyond helping guide the choice of latent labels for the nodes, the distributions we fit over the metadata carry useful information themselves. As a brief example of this, in Fig. 3.12 we display the logarithm of the parameters inferred for the distribution over institutions, for two blocks in the most recent time period in our best model for the citation network, each with more than 20 authors (so as to reduce spurious results). Such information is useful for Elsevier as it helps in understanding how research communities are split across particular groups of more closely collaborating institutions.

For clarity, we only display institutions with log values greater than -3.5 – this threshold is chosen as $e^{-3} \approx 0.05$, and so we effectively require more than $1/20$ authors in these groups to be affiliated to the institutions shown in expectation.

We can immediately see that the block on the left is dominated by authors at two universities in Colombia, and particularly centred on the capital Bogotá. The most likely institutions for authors in the block on the right however, are actually in

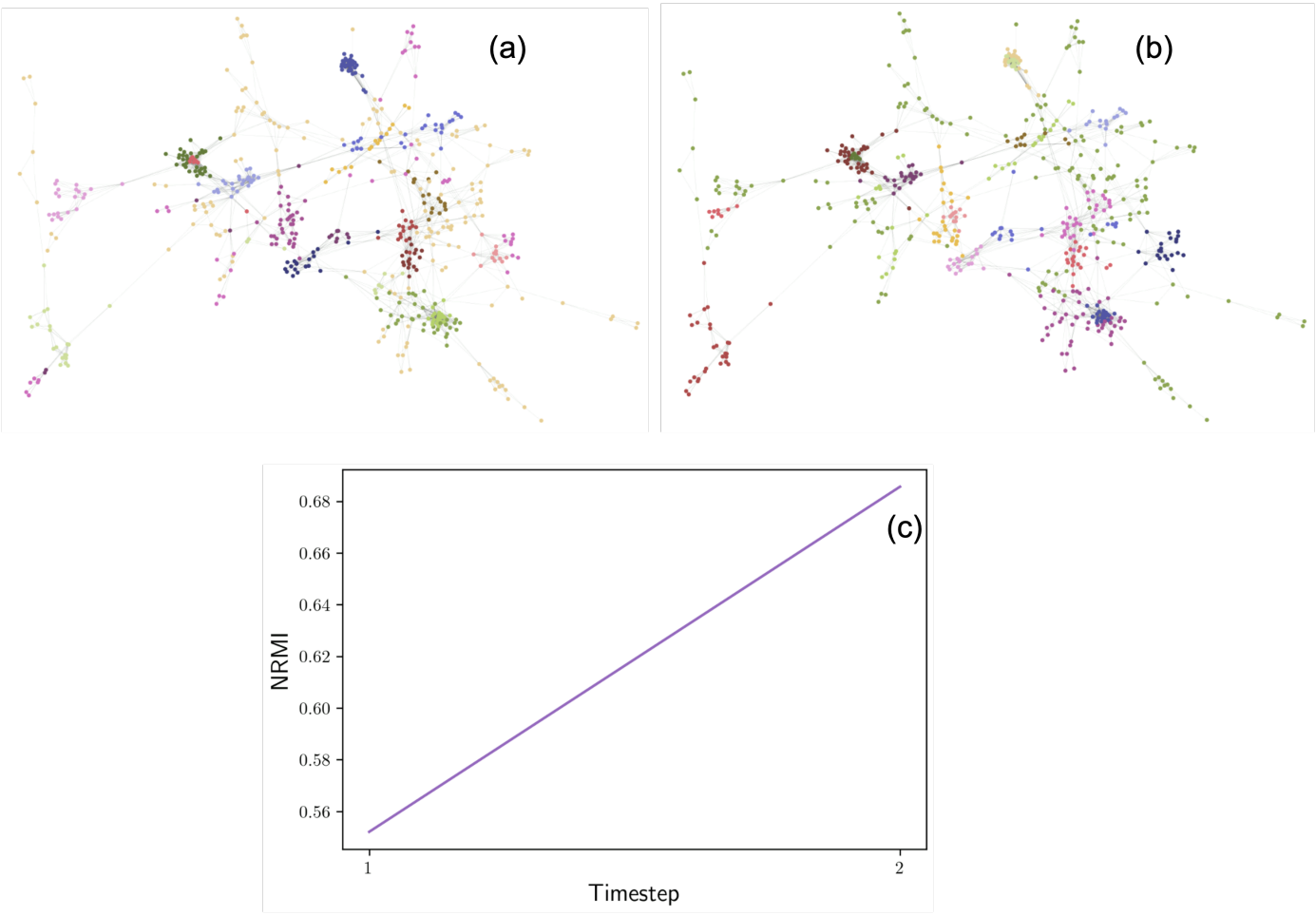


FIGURE 3.11: Subfigures (a) and (b) show the blocks inferred for the network at the final timestep for the citation data, using our model and that of [98] respectively with $Q = 19$. In subfigure (c) we display the NRMI between the partitions inferred by the two models at each timestep.

Mexico, particularly Mexico City, suggesting a possible migrant, or visiting, academic community.

Indeed, beyond the ability to produce such plots, in general knowing these distributions allows another means of comparing two groups, and/or finding similar groups to those of interest – highly valuable when summarising and exploring complex networks.

3.4.2 van de Bunt students dataset

To allow more rapid exploration of insights our model can provide, in this section we consider a significantly smaller network collected by van de Bunt, as introduced in [175].

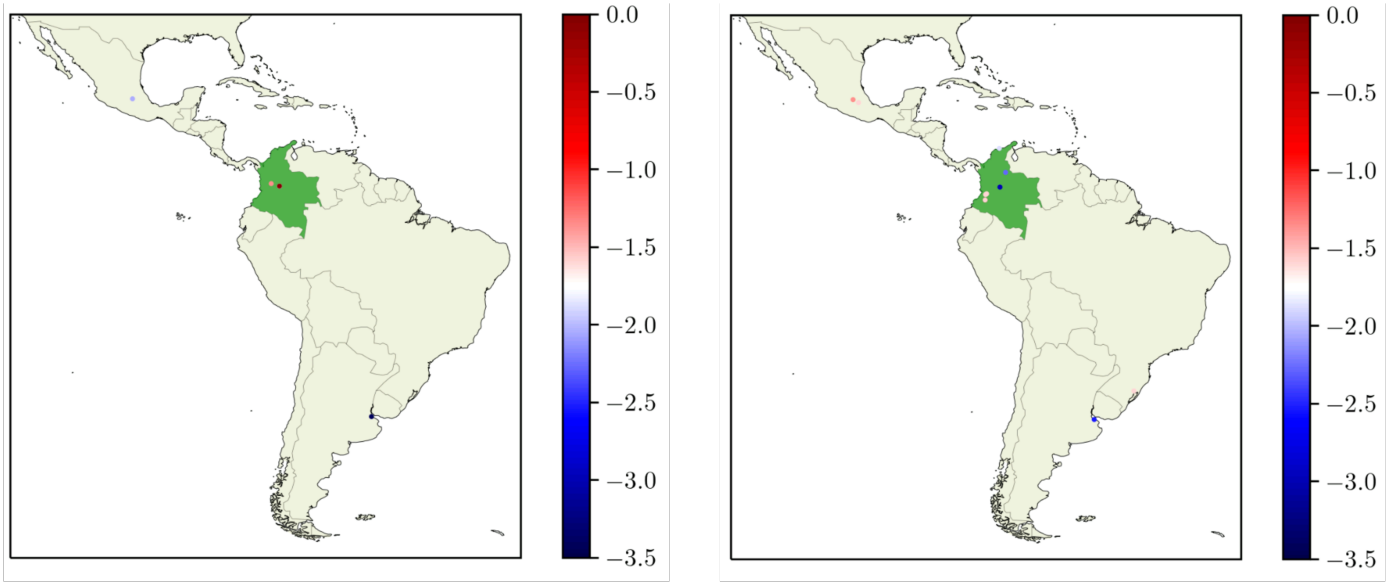


FIGURE 3.12: Here we plot the logarithm of the parameters inferred for the distribution over institutions for two blocks in our best model for the citation network in the most recent time period. Colombia is highlighted in green. For clarity, we only display institutions with log values greater than -3.5.

3.4.3 Data

This corresponds to a snapshot of relationships between 32 university students at 7 time points. The first four time points are three weeks apart, while the last three are six weeks apart – thus this provides an example of non-uniform period durations. Finding evolving groups in the network provides insight into changing social groups, and hence incorporating metadata when shaping these groups can help us to understand the importance of certain attributes to relationship formation.

Relationships were rated on a six point scale, with categories as follows:

1. Best friendship: Persons whom you would call your 'real' friends;
2. Friendship: Persons with whom you have a good relationship, but whom you do not (yet) consider a 'real' friend;
3. Friendly relationship: Persons with whom you regularly have pleasant contact during classes. The contact could grow into a friendship;
4. Neutral relationship: Persons with whom you have not much in common. In case of an accidental meeting the contact is good. The chance of it growing into a friendship is not large;

5. Unknown person: Persons whom you do not know (*i.e.* non-edges);

6. Troubled relationship: Persons with whom you can't get on very well, and with whom you definitely do not want to start a relationship. There is a certain risk of getting into a conflict.

As the questions are asked to each student individually, this is a directed network, and so we apply directed formulations of the edge distributions of our models – as extensions are quite immediate, for brevity we do not display the full set of corresponding equations herein. To model the different types of relationship, we choose to use the discrete edge formulation described in Sec. 3.2.3.2.

Continuing now to the metadata available in this dataset, for each student we know (i) their sex, (ii) their education program, and (iii) whether they smoke or not. The original study notes that smoking was only allowed in special areas, and thus effected different contact opportunities dependent on smoking behavior. For educational program, while all started to study at the same time, there were three groups, following different courses. After a few months, the programs diverged from each other, particularly for the 2-year program, hence again influencing individuals' contact opportunities.

In Fig. 3.13, we display an overview of this metadata. We observe (a) there is strong sex imbalance, with only 8 men, (b) that the standard 4-year program constitutes half of the students, with only 6 and 10 students on the 2- and 3-year programs respectively, and (c) there are 13 smokers. Naturally, we use categorical distributions to describe these pieces of metadata – the frequencies suggest that if groups were forming at random, the probability of women would be 0.75, of program 0.1875, 0.3125, 0.5 respectively, and of smoking 0.40625, hence deviations from these values suggest an importance of metadata in defining social groups. Note unlike Colombian dataset, these are not measured over time, as researchers assumed they were all static attributes – as such, we could more strongly enforce group metadata coherence throughout if desired by fixing metadata distribution parameters across time.

3.4.4 Results

Through evaluating models for $Q = 3$ to 10, we found the maximum ICL at $Q = 4$ – we note that while the presence of 5 edge categories heavily penalises the number of groups, the 'elbow' method applied to the log-likelihood also suggested this was a reasonable choice to take for Q. We first present following results as obtained from this 'best' model.

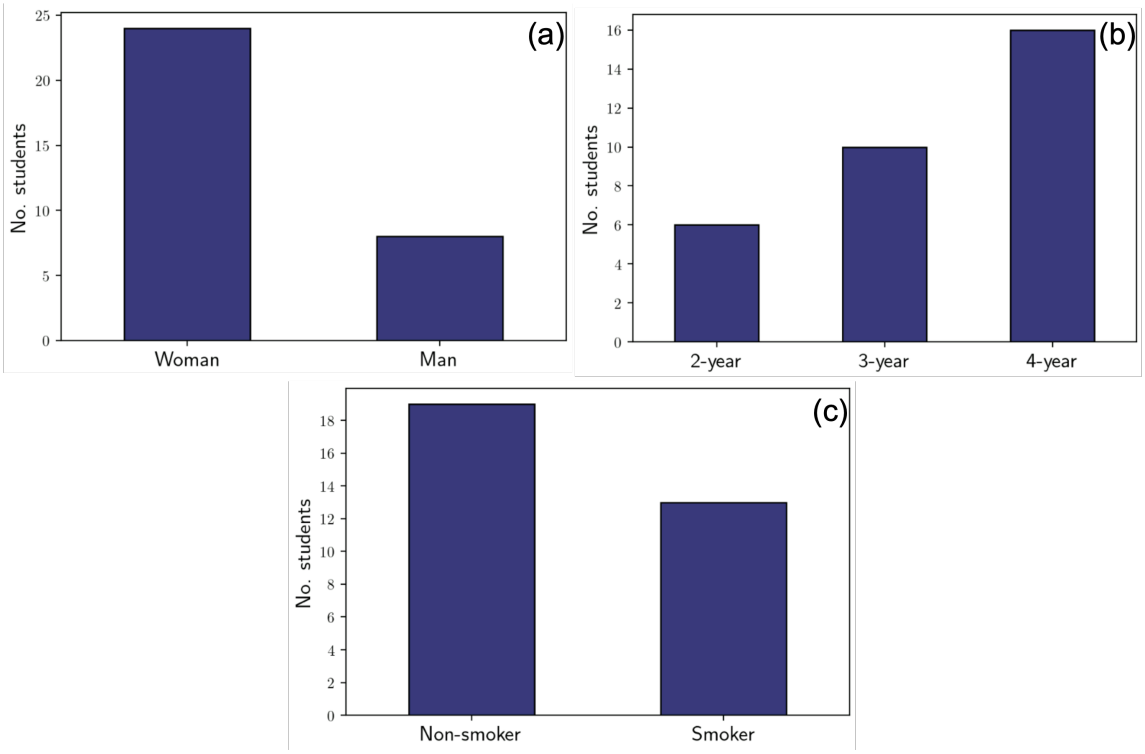


FIGURE 3.13: Counts of metadata considered for the van de Bunt students dataset: subfigure (a) shows sex, subfigure (b) the program of study, and subfigure (c) smokers/non-smokers. Note unlike Colombian dataset, these are not measured over time, as researchers assumed they were all static attributes.

In Fig. 3.14 we display the metadata distribution parameters inferred for each group over time in this model. Subfigure (a) shows p(woman), subfigure (b) p(smoking), and subfigure (c) p(program) for each option.

Keeping in mind the expected values of each should groups be completely random, we may observe that all factors appear to have a significant influence on the groups inferred in the network. The separation of smokers from non-smokers appears to have somewhat diminished over time, and likewise for program to some extent, while for sex we finish with two more mixed groups, and two more segregated than one would expect.

We also assessed the ability of our model to distinguish the true time sequence of the networks *vs.* a series of datasets where we permuted the time orderings. For all evaluated, the maximum ICL on the true dataset was higher than that on the shuffled time series, with higher mean value and smaller standard deviation as well. Further, this was supported by comparing the average entropy of the inferred group marginals (a measure of model confidence), which was lower for the true dataset than

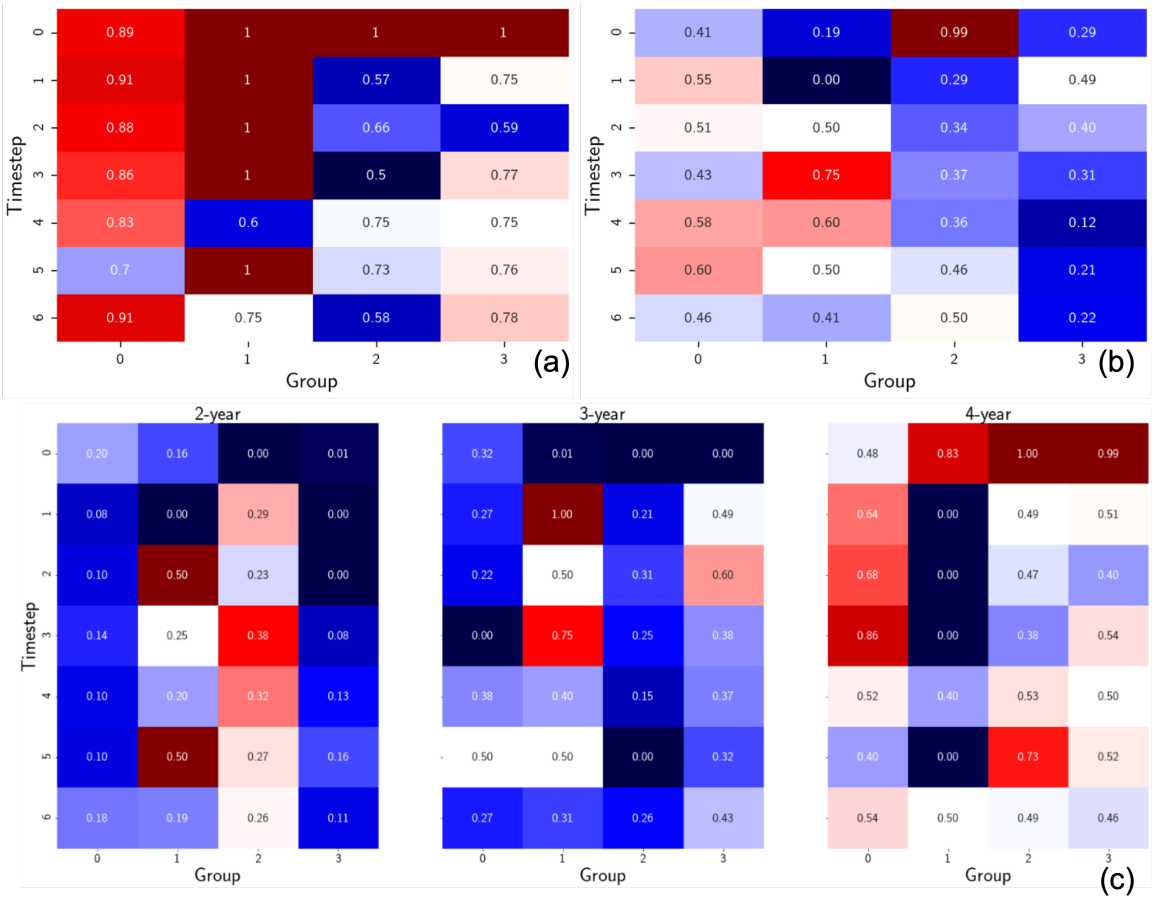


FIGURE 3.14: Inferred metadata parameters for each group over time in our first maximum ICL model, with $Q = 4$: subfigure (a) shows p(woman), subfigure (b) p(smoking), and subfigure (c) p(program) for each option.

for shuffled. This suggests we are indeed capturing something meaningful about the evolution of the network, and hence social groups, through our model.

We will not dwell further on these initial results, though note there is a wealth of further information present, *e.g.* how ω suggests the interactions between social groups, who are more central/consistent figures for each group, how stable the groups are over time *etc.*

We also investigate another possible application of the proposed model: we evaluate our ability to predict edges/non-edges correctly, as suggested in Sec. 3.2.7. Within the dataset, six pairs are marked as missing already, but without ground truth available we neglect these and impute zeros. Note however that three of these six are at a very sparse timestep, and are very likely non-edges, while the remaining three are at a dense timestep and connected to a high-degree node, thus are likely in fact edges.

Rather than considering these edges, to assess our performance, we perturb the

network provided: for each edge, with probability 0.05 we turn it into a non-edge, while with probability 0.01, for each non-edge involving a node present at that timestep we form an edge with value 1 (*i.e.* the weakest strength relationship). Considering separate probabilities of perturbation means that we are applying double-standard sampling to the dataset — further discussion of how this affects the model, and modifications for the static case that could improve accuracy may be found in [168].

This provides a new network, where we have changed the values of 55 pairs over the time period. We then fit our model for the same range of Q to this network, and compare the resulting predictions on these pairs to the true values. In Fig. 3.15 we display the performance that results as measured using the Receiver Operating Characteristic (ROC), where we plot the rate of true positives against false positives as we vary the threshold value of probability at which we predict an edge *vs.* non-edge. The area under this curve may be used as a single score (AUC) to assess performance, and takes maximum value 1 in the case of perfect recovery. Subfigure (a) shows the ROC curve for the model fitted to the perturbed dataset with maximum ICL, now with $Q = 3$. However, in subfigure (b) we display AUC values for this curve for models fitted to this dataset with maximum ICL for each number of groups considered, demonstrating that better predictive performance may be found with lower ICL models. Subfigures (c) and (d) repeat these tests for the 'oracle' model fitted to the true dataset for comparison.

There are several interesting observations from this figure. Firstly, for both the oracle model and that fitted on the perturbed network, the Q value for the best performing model in terms of AUC is not the same as that for maximum ICL. As noted in [174], link prediction performance on held-out data – *i.e.* the predictive power of a model – is a common alternative method to perform model selection, that does not require heuristics. However, as a means for choosing the 'correct' number of groups, link predictive power is consistent only when it agrees with proper Bayesian model selection, otherwise it overfits.

Secondly, the values of Q that perform best for the perturbed model and the true model are not the same. This implies that as one might expect, the specific links that are missing, or spurious, are highly influential on the final groups inferred, and hence the optimal Q to be selected. This reinforces the need to perform cross-validation on multiple perturbations of the network should predictive performance be used as a means for model selection.

Finally, performing prediction using the node marginals, which in effect incorporates influence from the metadata of the node as well as the edges, almost always outperforms

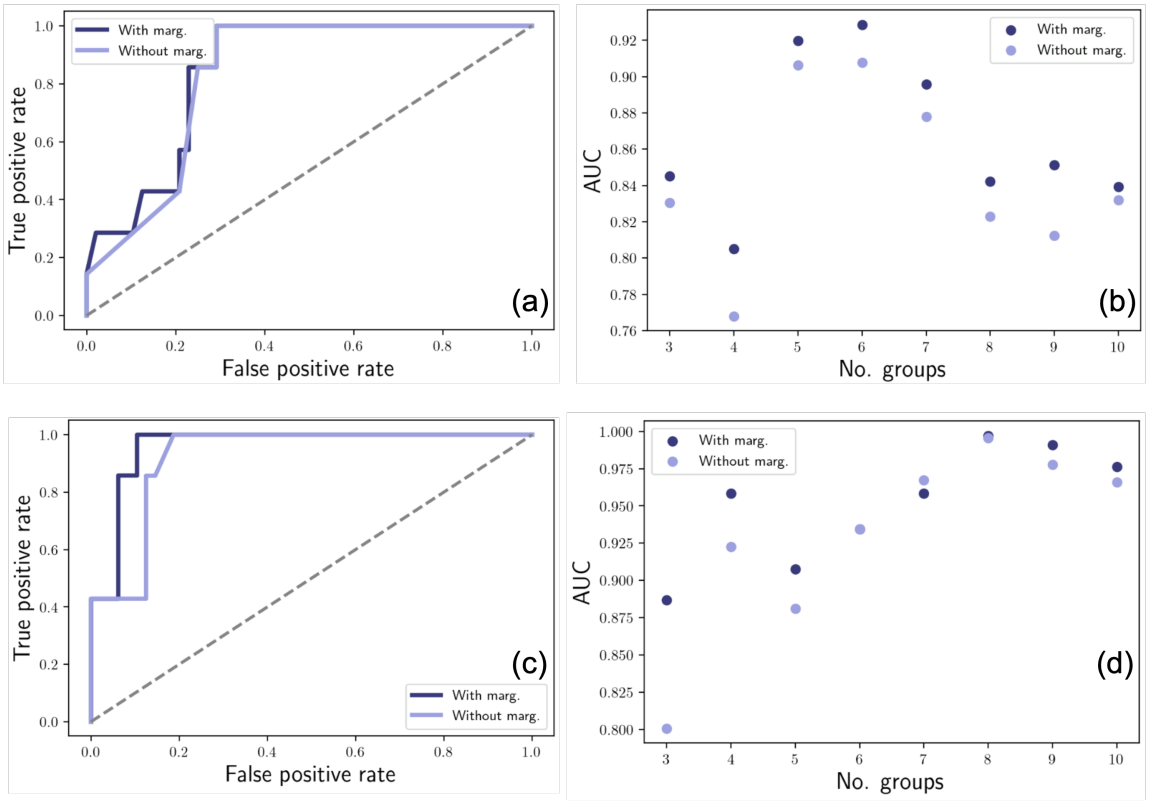


FIGURE 3.15: Link prediction performance for our model. Subfigure (a) shows the ROC curve for the model fitted to the perturbed dataset with maximum ICL, now with $Q = 3$. Subfigure (b) shows AUC values for this curve for models with maximum ICL for each number of groups considered, demonstrating that better performance may be found with lower ICL models. Subfigures (c) and (d) repeat these tests for the 'oracle' model fitted to the true dataset.

the basic version of simply taking the value of $\mathrm{p}(a_{ij}^t \mid z_i^t, z_j^t)$ alone . This suggests that the inferred marginals carry important information about the network, and should not be discarded lightly, as is often the case.

We conclude our presentation of results for this chapter with a brief demonstration of the importance of the tuning parameter, as introduced in Sec. 3.2.8. As previously suggested, in cases where the alignment between a partition of the metadata distributions and that of the structure in the edges is not perfect throughout the network, it may well be the case that we can improve the quality of our model by tuning the relative importance of each.

As such, in this section we consider models fitted to the van de Bunt dataset with $Q = 4$, for values of the global tuning parameter, θ, from 0 to 1 in steps of 0.1. We now find that the model with the maximum value of ICL is actually that with $\theta = 0.7$. On multiple measures, the groups appear to be improved, and more

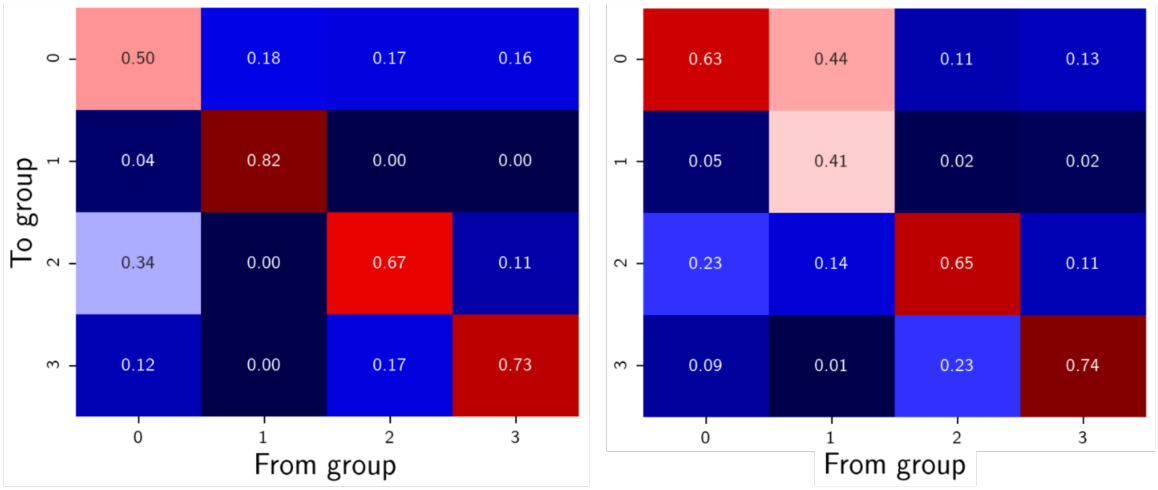


FIGURE 3.16: Inferred transition matrices, π, for the tuned model (left) with tuning parameter value $\theta = 0.7$, and the original model (right), where effectively $\theta = 1$.

stable in this tuned model — for instance with regards to metadata parameters as inferred above, ω values and more. For length considerations, we restrict ourselves to comparing the transition matrices inferred, as the most parsimonious demonstration of this improvement — should groups be more stable, we should immediately see larger diagonal values therein.

We display the two inferred transition matrices in Fig. 3.16 – the tuned model on the left, and the original 'best' model on the right. We can indeed observe that the tuned model has notably stronger diagonal values than the untuned version, and more clearly displays a more transitory group along with one more separated group, and relatively more likely transitions between the other two groups.

3.5 Discussion

In this chapter, we have introduced a dynamic SBM allowing for nodal metadata, a novel contribution to the literature, and the primary focus of this work. Beyond our base model, we also elaborated several immediate extensions to improve upon it.

Some further features of interest not addressed herein could be to *e.g.*

(i) Allow for hierarchical groups, while still including metadata. This is desirable, given we know this type of organisation is important in practice, and can further improve the detectability of groups [160]. We explore one such option in Chap. 6.

(ii) Design a formulation of the model that explicitly seeks separate groups for bipartite networks. This is a relatively immediate extension, see *e.g.* [184] for

a discussion and various implementations of the static case, which has been extended for edge metadata (much as for the textual model of [19]) and model selection considerations by [185]. Further, this would allow the separate grouping of high-cardinality node attributes, rather than using them directly (see *e.g.* [62] for similar prior work) — a partial solution to the scaling problem that emerges for high-dimensional metadata.

(iii) In lieu of edge categories, particularly for data such as that in Sec. 3.4.2 where relationships are definitively positive or negative, allow consideration of signed networks.

We briefly comment that one approach to do would be to take an independent layers multilayer SBM framework as described in *e.g.* [137], where node groups are fixed across layers, each layer is permitted distinct parameters — such that $\mathrm{p}(A \mid Z) = \prod_\ell \mathrm{p}(A^\ell \mid Z)$, for A the full multilayer network, and A^ℓ the network in layer ℓ. Effectively, our equations for $\tau_{tiqq'}$ then just include a product over the edge parameters for the set of layers, ℓ, while individual equations for parameters within layers would be much as before, only now with a shared τ^m_{tiq}. The number of parameters, and hence computational complexity however would then increase by roughly a factor of the number of layers, so application to many-layered large networks could become challenging.

(iv) Extensions to investigate core-periphery SBMs, akin to that introduced in [47].

(v) Account for non-MAR missing data — recent developments in the area suggest a means to do so without substantial modification of our model [168].

(vi) Permit alternative weight distributions in addition to Poisson- or categorical-distributed edges. Similar methods specifically utilising mean-field variational inference in the literature have done so for Gaussian [98], and recently approximations for Gamma distributions [116]. The latter was performed for the static case only, but may be extended analogously for dynamic networks. More generally, if alternative Bayesian inference approaches are taken, then a multitude of other distributions are readily available — particularly if conjugate prior distributions are used, though non-conjugate distributions are also possible [152, 93].

(vii) As an alternative to the tuning parameter, address the possible issue of metadata misalignment by further increasing model complexity, and essentially learning

a linear transformation from metadata (or distributions over non-categorical metadata), much as in [113].

(viii) Explore how the model might be used to determine change-points in the network, as in [29, 131]. This is also pertinent to our model, as such change-points could be defined by times when the group transitions significantly change — thus neglecting them may lead to inaccurate estimates over longer time periods, or systems that change more rapidly, than those considered in this thesis. Perhaps relatedly, it would be desirable to perform rigorous selection of the window length used to define network snapshots.

(ix) Modify the model selection measure, and/or make the model fully nonparametric, such that it infers a suitable number of groups directly from the data without requiring multiple runs. For instance, an alternative emerges if conjugate priors are placed over all parameters, and parameters are then integrated out. The result is a distribution over A and Z that only depends on observed statistics and hyperparameters. This allows the computation of an 'exact' ICL, for some prior hyper-parameters a, b, c, δ, as

$$\mathrm{ICL}_{ex} = \log \mathrm{p}(A, X \mid Z, a, b, c) + \log \mathrm{p}(Z \mid \delta). \tag{3.51}$$

We refer the interested reader to *e.g.* [152] for a use of this for the dynamic SBM case without metadata. Given non-informative priors, this provides a similar term to the more interpretable information criterion approach provided, but the authors use it as a partition quality metric for a greedy inference procedure that scales well.

(x) Account for edge persistence. We know this to be an important factor — for instance, for the networks of Colombian authors considered in Sec. 3.4, these values were extremely high: of edges present in the penultimate period, about 70% persisted in the co-authorship network, and 66% for the citation network. A recent avenue of research has begun to explore methods to address such persistence in dynamic networks, for instance [187, 94, 123, 100], and these could be built upon.

Chapter 4

Belief propagation inference and detectability

In this chapter, we elucidate the necessary theoretical background for the method used to tackle one of the major drawbacks of the previous inference procedure for our Dynamic SBM with Metadata (DSBMM) – quadratic scaling in the number of nodes – as well as explore analytically how the inclusion of metadata affects the recovery of groups in a network. To do so, we introduce the method of belief propagation (BP), which uses 'messages' passed between nodes to update the 'belief' that a node belongs to a particular group — *i.e.* to approximate the marginal distribution at each vertex. We postpone the detailed derivation of the necessary equations for the specific case of the DSBMM to Sec. 5.1 in the following chapter. Instead, here we focus on elaborating the general case, and explaining one way the resulting system of messages between nodes has been leveraged in prior work, to determine conditions on model parameters for efficient recovery of groups in the network.

For certain families of model parameters – discussed in greater depth below – it can be proven that beyond efficient recovery, a variety of SBMs that neglect metadata display important, sharp, transitions in their identifiable behaviour across different parameter regimes, where these regimes are separated by a boundary (or boundaries) past which groups cannot be recovered whatsoever better than at random. In the following, we refer to the conditions that determine such boundaries as *weak detectability* thresholds. Under more general classes of SBM, the issue of weak detectability vanishes, as simple algorithms may be used to generate partitions from the observed data that are guaranteed to be better than random, if the network is truly drawn from the SBM — for instance, by clustering according to degree. Instead, similar arguments as for weak detectability lead to the notion of *efficient* detectability

thresholds, which define conditions that model parameters must satisfy in order for strong recovery (*i.e.* high accuracy, and/or without requiring an exponential number of iterations) to be feasible. In particular, we discuss why the argument of a recent paper [153], that used BP to seemingly determine tractable, clean analytic expressions for a weak detectability threshold – importantly in the context of this thesis, for a static SBM variant *with* metadata – is fundamentally flawed. We then proceed to remediate this by describing approximate conditions for efficient detectability in a static version of our own model.

4.1 Introduction

Scalability of graph clustering algorithms, while maintaining desirable properties such as the statistical significance of groups uncovered, is an increasing focus in the literature, particularly since the advent of focal events for the community such as GraphChallenge [68]. This is considered of vital importance, as the quantity of network data available vastly increases each year – for instance, the networks of Latin American authors we produce from Scopus data [20] and consider in this work contain tens of thousands of authors at each timestep, even solely in their largest connected component. While network backboning techniques have also been pursued (see *e.g.* [32]), that seek to reduce larger networks into smaller, more manageable subgraphs that maintain desired properties of the wider system, discarding the vast quantities of information required for the application of algorithms with quadratic scaling, such as that of the previous chapter, prevent certain insights. For instance, an example in our case is that there may be too few authors in such a subset that publish in a particular area for them to be clustered into a single statistically significant group, and thus the corresponding research community is neglected.

An additional benefit of scalability for network clustering algorithms is that it allows the application of these methods as a component of broader procedures to better understand complex systems of interest. For example, the improved scaling that the techniques in the following two chapters provide for the DSBMM subsequently allows us to investigate author influence using a causal model in Chap. 6, by extending the ideas of a recent paper to the dynamic case, and translating to our particular context [166]. The relations between authors, publications, and citations that must be understood in order to do so may not be well-captured by a model, unless the networks considered are sufficiently large, *i.e.* in the limit of sufficient data. For instance, we would expect that the influence of an author may not be accurately

estimated until at least their own research community, and likely those in adjacent fields are observed. Given the diversity of topics pursued by each author, and the size of research communities in contemporary academia, this immediately necessitates at least $\mathcal{O}(1000)$ authors. As we explain in greater depth in Chap. 6, the DSBMM is particularly well-suited to play an important role in such methods, but with quadratic scaling the necessary search of hyperparameter space, let alone multiple runs for each choice of parameters, quickly becomes infeasible in practice.

However, no matter their speed, blind application of algorithms to large datasets, without understanding when the results are likely to be of any quality, may only be a waste of computational resources. As such, as described above, in the context of SBMs there has been a considerable quantity of research dedicated to understanding the conditions under which convergence to better-than-random node labels may be expected, particularly since the seminal study of [34]. In that study, the authors determine such limits through an argument using the BP framework, for a simplified version of the classical SBM. In brief, this is possible by considering the stability of a so-called factorised fixed point (FFP) for the message-passing system, which corresponds to uniform messages and node marginals. If the FFP is stable, *i.e.* all eigenvalues of the Jacobian of the system at this point have modulus less than one, then as in classical nonlinear dynamics [64], it forms an attractor for any inference procedure — thus inference fails to converge to anything other than a uniformly random partition. The resulting stability conditions are variously known in different fields as the Kesten-Stigum bound on reconstruction on trees [80], Almeida-Thouless local stability conditions that describe 'phase transition' boundaries [71], and the robust reconstruction threshold [104], and correspondingly there are numerous different arguments that have been used to derive the same fundamental conditions on parameters.

Within this chapter, we proceed in several stages. First, in Sec. 4.2, we discuss some of surrounding literature, with a particular focus on works that explore group detectability in different SBM variants. Next, in Sec. 4.3 we elaborate how to derive the message passing system necessary for belief propagation in general factor models. We then proceed in Sec.4.4 to discuss how this has been used to determine weak detectability thresholds in the past, and rebut a recent paper that claims to do so for a SBM incorporating metadata. Finally, in Sec. 4.5, we describe how we may instead approximate efficient detectability thresholds for a static version of our own model.

4.2 Literature review

Beyond the classical SBM, other works that follow a similar argument to the seminal [34] have since explored weak detectability thresholds in a variety of other SBM classes. These include, among others: the greedy modularity quality function [200]; differences for directed networks [183]; for bipartite networks [44]; in dynamic networks with group persistence [50] — of particular relevance to our own model; the subsequent change in dynamic detectability in the presence of link persistence [12]; and even hypergraphs [9].

However, as described in detail by [201], all such examples only hold for especially restrictive families of SBM parameters — for instance, they focus on demonstrating that as soon as the expected size distribution of groups (as governed by α in our model) is not uniform in the classical SBM, the previous FFP no longer applies. Instead, given a particular initialisation, the application of BP beyond the previous detectability limit may be shown to optimally account for the local information available to each node, specifically the degrees of itself and its neighbours at different distances (or 'hops' as they are often known in the literature). In this chapter, we demonstrate a similar property for our own model in the case of metadata.

In addition to these restrictive parameter families, all the previous weak detectability methods described do not account for the presence of metadata. To our knowledge, only two works have claimed to do so, and both in very recent years. The first does so through introducing the Joint SBM (JSBM) [202], which corresponds to the popular framing of the network and metadata as a bipartite system, with different types of inter- and intra-class edges — much as performed in *e.g.* [62]. As such, the model is not immediately amenable to many metadata types, *i.e.* continuous metadata in particular, and more importantly their detectability thresholds necessarily include separate terms for the network-metadata block parameters that they introduce. By introducing a significant number of new parameters, they both risk overfitting, and we believe also provides a less intuitive framing than our own model. For instance, the new eigenvalues defining the threshold are entirely analogous to the classical case, only using the inter-class block parameters, and otherwise the only additional factor that must be accounted for is the permissible paths between nodes of different classes — thus the threshold is effectively a modification of that for bipartite networks, as in [44], where one class is permitted additional intra-class links. Nonetheless, this is a valuable contribution, and excellent paper as a whole – particularly given novel connections with *e.g.* graph neural networks – but we believe does not provide as much clarity

about the specific importance of metadata as that proposed herein. Instead there is more relevance to a generalised bipartite formulation.

The second recent study on metadata SBM detectability [153] appears to be the only work, other than our own, that instead explores detectability in an SBM which directly incorporates general nodal metadata. Unlike our own, it does so by modifying the probability of each edge according to a similarity measure between the metadata of the participating nodes, in a manner reminiscent of some latent factor models, *e.g.* [193]. However, as we discuss in greater detail below, there are several unfortunate errors and misunderstandings in their methodology, and these prevent the legitimacy of their principal results. Indeed, the presence of nodal metadata, unless considered as simply a separate class of nodes as in the JSBM, fundamentally does not permit the same methodological procedure, due to breaking necessary symmetries — this is analogous to the issue of unequal group sizes.

Specifically, we demonstrate that the new message equations do not directly permit the use of any of the typical methods for establishing either weak, or efficient detectability. However, using an extension of the recent proposition to understand the detectability of hierarchical partitions [133], and the signal-to-noise ratio (SNR) threshold for efficient inference of [1] allows us to nonetheless proceed — though we warn in advance that the form of the (approximate) threshold thus determined is not as pleasant as in previous cases.

In addition to efficient scaling, belief propagation provides near-optimal recovery of groups when in a detectable regime — specifically, the node marginals inferred are typically accurate representations of the confidence of the model in the label of each node, while MFVI is known to often greatly over-estimate such confidence [199]. Indeed, the fact that the estimated node marginals are immediately available after the inference procedure completes, without requiring additional computation, is another benefit. This has led to widespread adoption for numerous probabilistic models, for instance: to perform rapid inference and model selection for the DC-SBM [192]; for the influential SBM with metadata of [113]; and for the JSBM of [202], where in addition to detectability results, BP leads them to subsequently propose a novel BP Graph Convolutional Network (BPGCN) for network parameter inference; alongside many other studies.

4.3 Belief propagation inference

As previously described, there are several downsides to the mean-field variational inference approach. Primarily, these are (i) often reduced quality of results compared to the principal alternatives of MCMC or belief propagation (BP), (ii) over-confidence in the accuracy of the labels inferred, and (iii) its quadratic scaling with respect to the number of nodes considered — unless one takes a stochastic approach, and subsamples the system at each iteration, thus introducing further inaccuracy. In this section, we provide the necessary technical background for us to instead infer DSBMM parameters through belief propagation.

4.3.1 Understanding the origins of belief propagation

To commence, suppose we have a general probability distribution over some set of variables, $\boldsymbol{x}$, that factorises according to

$$\mathrm{p}(\boldsymbol{x}) = \frac{1}{Z} \prod_a f_a(\boldsymbol{x}_a), \tag{4.1}$$

where we use $\boldsymbol{x}_a$ to denote the subset of variables in $\boldsymbol{x}$ that participate in the factor a, and $Z = \int \prod_a f_a(\boldsymbol{x}_a)\, \mathrm{d}\boldsymbol{x}$ is a normalisation constant, often known as the partition function, to ensure this defines a proper probability distribution. This defines a bipartite *factor graph*, where there are nodes for variables, and nodes for factors, with edges connecting variables to factors that they participate in.

When computing functions that involve this distribution, for instance should we seek a posterior over some parameters given observed data, we may immediately face a problem — as it requires an integral (or sum) over the entire domain, calculating the value of the partition function, Z, can be prohibitively expensive. As such, numerous methods seek to either approximate, or entirely circumvent the need to calculate, Z. For instance, after convergence (which is not guaranteed, and may take some time), MCMC allows one to directly sample from the posterior without calculating Z, by using *e.g.* proposal distributions with Metropolis-Hastings acceptance probabilities [136]. Alternatively, as described in the previous chapter we may posit a variational distribution that simplifies the computation of the intractable term, then seek to minimise the discrepancy between this and the true distribution, often in terms of the Kullback-Liebler divergence. Belief propagation falls in this latter category, but rather than positing a variational family with only single-variable factors – as in MFVI – it instead further allows pairwise factors.

Belief propagation has a rich history, and in particular was greatly motivated by problems in statistical mechanics and physics, where it is also known as the *cavity method*. The legacy of this history remains, and hence there are particular quantities, and terms used to describe them, that almost always appear in works that utilise BP — these are best understood with some additional context.

Firstly, a fundamental theory in statistical physics is that at thermal equilibrium, a system of particles with discrete states that may be described by $\boldsymbol{x}$ satisfies Boltzmann's law,

$$\mathrm{p}(\boldsymbol{x}) = \frac{1}{Z(T)} \mathrm{e}^{-E(\boldsymbol{x})/T}, \tag{4.2}$$

where $E(\boldsymbol{x})$ is the energy of a state $\boldsymbol{x}$ of the system, T is the temperature, and $Z(T)$ is once again a normalisation constant, the partition function

$$Z(T) = \sum_{\boldsymbol{x} \in S} \mathrm{e}^{-E(\boldsymbol{x})/T}, \tag{4.3}$$

for S the space of all possible states $\boldsymbol{x}$ of the system.

By analogy with Eqn.(4.1), we may observe that if we choose $T = 1$, and assume this law holds for the factorised model, this defines an 'energy' of the factorised probability distribution,

$$E(\boldsymbol{x}) = -\sum_{a} \log f_a(\boldsymbol{x}_a). \tag{4.4}$$

By imposing this equality, we may draw on the considerable literature used to understand the expected behaviour of physical systems, given observable quantities — this insight led to a spate of work in multiple fields, particularly computer science, in the late 1990's and early 2000's.

In this physical context, one such key functional quantity is the *Helmholtz free energy*, which corresponds in this case to the negative log of the partition function,

$$F_H = -\log Z. \tag{4.5}$$

This is of great interest in practice, as if the dependence of F_H on experimentally measurable quantities such as temperature is understood, then predictions about the response of the system under subsequent perturbations can be computed (and models hence validated). In the context of Bayesian inference, this would correspond to the negative log of the model evidence, and is thus naturally an important quantity for statistical inference also.

However, as suggested above, this is frequently intractable to calculate, thus physicists also often resort to the variational approach. As such, recall that the KL-divergence between a proposed variational distribution $\psi(\boldsymbol{x})$, and the desired distribution p is given by

$$\begin{aligned} D_{KL}(\psi \parallel \mathrm{p}) &= \sum_{\boldsymbol{x}} \psi(\boldsymbol{x}) \log \frac{\psi(\boldsymbol{x})}{\mathrm{p}(\boldsymbol{x})}, \\ &= \sum_{\boldsymbol{x}} \psi(x) \left(\log \psi(\boldsymbol{x}) + E(\boldsymbol{x}) + \log Z\right), \\ &= -F_H + \sum_{\boldsymbol{x}} \psi(x) \left(\log \psi(\boldsymbol{x}) + E(\boldsymbol{x})\right). \end{aligned} \tag{4.6}$$

In physical terms, the sum remaining corresponds to an (variational) entropy term,

$$S(\psi) = -\sum_{\boldsymbol{x}} \psi(\boldsymbol{x}) \log \psi(\boldsymbol{x}), \tag{4.7}$$

and the expectation of the energy over the posited variational distribution, *i.e.* the variational average energy

$$U(\psi) = \sum_{\boldsymbol{x}} \psi(\boldsymbol{x}) E(\boldsymbol{x}). \tag{4.8}$$

A final important quantity in statistical mechanics is the *Gibbs free energy*, which is related to the amount of work a system can perform at constant temperature and pressure. At unit temperature, this is given by $G = -S + U$, for entropy S and average internal energy U. Hence it suggests a variational free energy in our case

$$G(\psi) = -S(\psi) + U(\psi), \tag{4.9}$$

and thus we have that we can express the KL divergence above as

$$D_{KL}(\psi \parallel p) = -F_H + G(\psi). \tag{4.10}$$

As such, minimising the variational free energy, given its lower bound of the Helmholtz free energy, is equivalent to minimising the KL divergence and optimising our variational distribution in this measure — *i.e.* an alternative view of the derivation of the ELBO as in Chap. 3. In addition providing further methods, this physical interpretation provides some greater intuition as to the properties of a good variational distribution: there is a clear trade-off between the entropy term, *i.e.* the expressivity of the variational family, and the average energy — that is, a good variational family is

sufficiently expressive to approximate a local minima in the energy functional E, while minimising the quantity of information necessary to do so, as measured by its entropy. This context is why the term 'free energy' occurs so frequently in belief propagation related papers, as it is a fundamental quantity when performing statistical inference for a model using a variational approximation.

4.3.2 The Bethe approximation

If we are now interested in the marginal likelihood of a particular variable, $\mathrm{p}(x_i) = \sum_{\boldsymbol{x}\setminus x_i} \mathrm{p}(\boldsymbol{x})$, it may not be immediately obvious how to approach the problem in general, particularly should the partition function Z be intractable, other than via MCMC. This may take a considerable time to converge to the posterior, and subsequently we would need to approximate the marginals through collecting many independent samples from the posterior — and in the case of the SBM, suitably aligning the sampled partitions, as discussed in [143]. However, if we constrain ourselves to considering distributions for which the factor graph has the form of a tree, *i.e.* no cycles, we have that [33]

$$\mathrm{p}(\boldsymbol{x}) = \frac{\prod_a \mathrm{p}_a(\boldsymbol{x}_a)}{\prod_i (\mathrm{p}_i(x_i))^{d_i - 1}}, \tag{4.11}$$

where d_i is the number of factors in which i participates, *i.e.* the degree of i in the factor graph. In particular, for such acyclic factor graphs, or trees, it is possible to recover *exact* marginals — through application of a message-passing, or belief propagation algorithm [196]. This suggests that we should consider variational distributions with the same factorisation, *i.e.* that we posit a family $\psi_a(\boldsymbol{x}_a)$ and $\psi_i(x_i)$, where $\psi_i(x_i)$ then provides us the desired estimate for the marginal likelihood.

For this variational family, the Gibbs free energy (4.9) exactly takes the form

$$\begin{aligned} F_{\text{Bethe}} = \sum_a \sum_{\boldsymbol{x}_a} \psi_a(\boldsymbol{x}_a) \log \psi_a(\boldsymbol{x}_a) - \sum_i (d_i - 1) \sum_{x_i} \psi_i(x_i) \log \psi_i(x_i) \\ - \sum_a \sum_{\boldsymbol{x}_a} \psi_a(\boldsymbol{x}_a) \log f_a(\boldsymbol{x}_a). \end{aligned} \tag{4.12}$$

Indeed, [198] demonstrate that fixed points in the message-passing algorithm correspond to fixed points of this free energy, known as the *Bethe* free energy. As such, for tree factor graphs where the Bethe free energy and true free energy are identical, when in a detectable, convex region of parameter space, the algorithm converges towards the minimum energy, thus simultaneously minimising the KL-divergence between the approximation and the true marginals [105]. While F_{Bethe} is no longer

exactly equal to the Gibbs free energy when the factor graph is *not* a tree, it is commonly used as an approximation nonetheless, and the subsequent message-passing inference procedure is commonly known as 'loopy' belief propagation [75].

To perform BP, messages are passed from variable to factor nodes, $i \to a$, and from factor to variable nodes, $a \to i$, according to the update equations

$$\psi^{i\to a}(x_i) \propto \prod_{b\in\partial i\setminus a} \hat{\psi}^{b\to i}(x_i), \tag{4.13}$$

$$\hat{\psi}^{a\to i}(x_i) \propto \sum_{\boldsymbol{x}_{a\setminus x_i}} f_a(\boldsymbol{x}_a) \prod_{j\in\partial a\setminus i} \psi^{j\to a}(x_j), \tag{4.14}$$

where ∂i and ∂a to denote the neighbourhood of variable i and factor a in the factor graph respectively, and proportionality is converted to equality by normalising the message over all possible values of x_i. In words, factor nodes receive messages according to the estimates of a node marginal from other neighbouring factors, while variable nodes receive messages according to the marginalised likelihood estimate at the sending factor, in the absence of the variable receiving the message. This exclusion of the receiving node from the equations lends the algorithm the name of the cavity method, by which it is often referred in the physics literature. This operates on the assumption at each step that all correlations between neighbours of the variable x_i in question are mediated via the sending node itself, *i.e.* they are conditionally independent given x_i, and thus take the simple form of a product distribution.

If the factors are constrained to contain a maximum of two variables, as is the case for instance almost all SBM variants, by direct substitution we observe that this message system can in fact be further reduced into effectively passing messages directly between variable nodes only, *i.e.*

$$\psi^{i\to j}(x_i) \propto f_i(x_i) \prod_{k\in\partial i\setminus j} \sum_{x_k} f_{ki}(x_k, x_i)\psi^{k\to i}(x_k). \tag{4.15}$$

Upon convergence, the marginals may then be estimated by

$$\psi^{i}(x_i) \propto f_i(x_i) \prod_{k\in\partial i} \sum_{x_k} f_{ki}(x_k, x_i)\psi^{k\to i}(x_k), \tag{4.16}$$

i.e. by including the neighbourhood factor removed in the message equation.

An additional simplification for the free energy is then also possible [102, 146] —

it reduces to

$$F_{\text{Bethe}} = \sum_i \log Z_i - \sum_{i,j\in\mathcal{E}} Z_{ij}, \tag{4.17}$$

where $\mathcal{E}$ denotes the set of pairwise factors present, and Z_i and Z_{ij} are the normalisation factors for the node, or one-point marginals,

$$Z_i = \sum_{x_i} f_i(x_i) \prod_{k\in\partial i} \sum_{x_k} f_{ki}(x_k, x_i)\psi^{k\to i}(x_k), \tag{4.18}$$

and so-called two-point marginals,

$$\psi^{ij}(x_i, x_j) \propto \psi^{i\to j}(x_i) f_{ij}(x_i, x_j)\psi^{j\to i}(x_j), \tag{4.19}$$

such that

$$Z_{ij} = \sum_{x_i,x_j} \psi^{i\to j}(x_i) f_{ij}(x_i, x_j)\psi^{j\to i}(x_j), \tag{4.20}$$

respectively.

Thus, upon convergence of the message passing system, we immediately have both estimates for the variable marginals, and the ability to quickly calculate the free energy, which may be used to evaluate the quality of the marginals inferred – *i.e.* perform model selection – and so determine the best partition found from different runs of the full algorithm.

While convergence is unfortunately not guaranteed for belief propagation on factor graphs with loops, in practice this 'loopy' belief propagation often nonetheless performs well – we proceed on this basis. For further discussion of BP, we refer the interested reader to *e.g.* [197, 102].

4.4 Analytic expressions for detectability limits

As discussed in the introduction to this chapter, we leave the elaboration of the BP message system for the DSBMM to the subsequent chapter. Instead, in this section we now turn to one of the alternative uses of the BP system: to determine detectability thresholds. To do so, the convention is to assume that the true parameters of the model are known, and only the latent variables of interest remain to be inferred — in statistical physics, this is sometimes referred to as being on the Nishimori line [118].

In this setting, the logic to determine weak detectability thresholds in an SBM variant is typically as follows:

1. Impose that every group has the same average degree c, *i.e.* that

$$\sum_r c_{qr} n_r = \sum_q c_{qr} n_q = c, \tag{4.21}$$

for all q, r. For a uniform prior, the typical case considered, where all $n_q = N/Q$, this corresponds to requiring that the block connectivity matrix is a multiple of a doubly stochastic matrix, with constant row and column sums. If this is not the case, then as [34] note, a positive overlap in the NDC-SBM is always possible simply by labelling nodes based on their degrees, thus there is no detectability limit. If the condition holds, the classical SBM is then referred to as the factorised block model. In the dynamic case, we must also impose that the transition matrix π is doubly stochastic [50].

2. If these conditions hold – with the additional requirement that n_q is uniform in the dynamic case [50] – the message passing system has a trivial fixed point, with $\psi_q^{i\to j} = n_q$ for all q — and likewise for both spatial and temporal messages in the dynamic case. As this does not depend on i, j, this is referred to as a factorised fixed point (FFP).

3. At this FFP, substituting into the equation for the marginals shows that in fact we also have that $\psi_q^i = n_q$, *i.e.* no better than a random guess. As such, if this fixed point is stable, we expect to converge to it, and thus in order to detect non-trivial communities we require instability.

   As n_q is a fixed point, at first order in small perturbations $\boldsymbol{\epsilon}$ a simple Taylor expansion of the message passing equations provides a linear system,

$$\boldsymbol{\epsilon} = M\boldsymbol{\epsilon}, \tag{4.22}$$

   where $M_{((i,j),q),((k,\ell),r)} = \frac{\partial \psi_q^{i\to j}}{\partial \psi_r^{k\to \ell}}$ is the Jacobian of the message system. In fact, for the classical SBM, if we define $M_{((i,j),q),((k,\ell),r)} = B_{(i,j),(k,\ell)} T_{qr}$, with B the non-backtracking matrix such that

$$B_{(i,j),(k,\ell)} = \begin{cases} 1 & \text{if } \ell = i \text{ and } k \neq j \\ 0 & else \end{cases}, \tag{4.23}$$

and

$$T_{qr} = \left.\frac{\partial \psi_q^{i\to j}}{\partial \psi_r^{k\to i}}\right|_{FFP} = \psi_q^{i\to j}\left(\frac{\omega_{qr}}{\sum_{r'}\psi_{r'}^{k\to i}\omega_{qr'}} - \sum_s \frac{\psi_s^{i\to j}p_{sr}}{\sum_{r'}\psi_{r'}^{k\to i}p_{sr'}}\right)\Bigg|_{FFP}, \\ = n_a\left(\frac{c_{qr}}{c} - 1\right), \tag{4.24}$$

a sort of local transfer matrix for the message perturbations, we have $M = B \otimes T$ for $\otimes$ the tensor product [105]. Here this is defined such that for two matrices $A \in \mathbb{R}^{a_1 \times a_2}$ and $C \in \mathbb{R}^{c_1 \times c_2}$, $(A \otimes C) \in \mathbb{R}^{a_1 c_1 \times a_2 c_2}$ is defined by

$$(A \otimes C)_{(i,\ell),(j,m)} = A_{ij} C_{\ell m}, \tag{4.25}$$

i.e. the generalisation of the outer product between vectors to tensors.

As such, we can construct eigenvectors for M from eigenvectors for B and T using the tensor product, with eigenvalues as the product of eigenvalues. To see this, observe that if $\boldsymbol{u}$ is an eigenvector for B, with eigenvalue λ, and $\boldsymbol{v}$ is an eigenvector for T with eigenvalue μ, if the tensor product holds then

$$M(\boldsymbol{u} \otimes \boldsymbol{v}) = \sum_{(k,\ell),r} B_{(i,j),(k,\ell)} T_{qr} u_{(k,\ell)} v_r \tag{4.26}$$

$$= \sum_{(k,\ell)} B_{(i,j),(k,\ell)} u_{(k,\ell)} \sum_r T_{qr} v_r \tag{4.27}$$

$$= \lambda\mu(\boldsymbol{u} \otimes \boldsymbol{v}). \tag{4.28}$$

This means that determining if the stability condition holds – that M at the FFP has no eigenvalue λ with $|\lambda| > 1$ – may be reduced to considering only the spectrum of B and T respectively.

Equivalently to this stability argument, we may instead consider the propagation of perturbations to the fixed point through the network. That is, suppose we introduce a small perturbation at the leaves of the network, such that

$$\psi_q^k = n_q + \epsilon_q^k. \tag{4.29}$$

Now if we take a path length d from a leaf k_d to a root node k_0, we have $\boldsymbol{\epsilon}^{k_0} \approx T^d \boldsymbol{\epsilon}^{k_d}$, which in the limit $d \to \infty$ is dominated by the largest eigenvalue of T, such that $\boldsymbol{\epsilon}^{k_0} \approx \lambda_{\max}^d \boldsymbol{\epsilon}^{k_d}$. As the expected number of nodes at each step is simply the average degree, c, there are around c^d leaves in total contributing perturbations to the root node.

If we assume the perturbations at each leaf are independent, with mean zero, then while the mean expected perturbation on the root is zero, the variance

$$\langle \epsilon^{k_0}, \epsilon^{k_0} \rangle \leq c^d \lambda_{\max}^{2d} \langle \epsilon^{k_d}, \epsilon^{k_d} \rangle \tag{4.30}$$

clearly grows exponentially if $c\lambda_{\max}^2 > 1$ – providing a stability, and thus detectability condition. An identical condition emerges from the tensor product spectrum argument above [105], or from an argument based on robust reconstruction on trees, as used to produce an analogous threshold in both the dynamic case of [50], and in the bipartite node-attribute JSBM of [202].

Indeed, every paper to successfully investigate precise detectability thresholds for a certain class of SBM has effectively followed one of these procedures. The absolutely essential requirement is independence of the local transition matrix, T, of i, j, and k, including in the directed case of [183], where two-way edges are neglected, or at least the ability to determine the expected connections at each step (*i.e.* class symmetry), as for the JSBM [202]. This symmetry is vital for every method to determine detectability thresholds, as (i) in the perturbation propagation argument, it means that the perturbation produced is invariant to the path taken from leaf to root, thus allowing simplification as for Eqn. (4.30); (ii) in the robust reconstruction on trees line of reasoning, the problem must be able to be considered in the limit $N \to \infty$ (and $T \to \infty$ for dynamic networks [50]) as a branching process (or a multi-type branching process [202]), which requires invariance to the index of the node at which the process lands at each step; and finally (iii) in the direct analysis of the spectrum of M, progress is only immediately possible due to the tensor product relation providing correspondence between spectra of B and T and M itself – if symmetry is broken, and dependence on i is introduced, the tensor product relation no longer holds, thus there is no simple way to deduce the spectrum of M without brute force.

4.4.1 A response to Ren *et al.* (2022)

The above requirements are the source of a fundamental misunderstanding by [153], in which the authors aim to analytically determine a weak detectability threshold in an SBM variant with metadata, as previously described. Unfortunately the make several critical errors: most importantly, they assume that it is sufficient to solely consider the maximum eigenvalue of a local transition matrix, T^i, despite noting the dependence on i, *i.e.* that local symmetry does not hold. Furthermore, even beyond this crucial flaw, there are errors in other steps of their working. For instance, while they recognise that

as in our own model, the conventional FFP does not hold, neither does the alternative fixed point they propose (*cf.* Eqn. (17) [153]) — in correspondence, the authors seemed to recognise that this is the case, at least without further strict conditions on permissible parameters, but there is no mention of the fact within the paper. In addition to these errors, the expression they derive for the 'message transfer' matrix T^i in their Eqn. (19) is incorrect — the denominator should have messages from $k \to i$, thus introducing additional dependence on k, *i.e.* $M_{((i,j),q),((k,\ell),r)} = B_{(i,j),(k,\ell)} T^{i,k}_{qr}$, and furthermore, their use of marginals as direct replacements for messages is no longer legitimate, as this only holds for FFPs.

In fact, the problem in their case can instead be approached using the measure of *ease* of inference suggested in [133]. Precisely, the absence of a weak detectability limit in the toy model of [153] should be clear, as their metadata is deterministic, and in both proposed cases, directly informative about the group structure. That is, by labelling the nodes solely according to their observed metadata, you are guaranteed a better-than-random partition.

Having demonstrated the absence of weak detectability, we instead consider a more complex case proposed by the authors, where there are $\tilde{q}$ metadata categories, each of which contain q_b nested 'brother' communities, for $q = q_b\tilde{q}$ total groups. The real question in this setting then becomes: what are the necessary conditions for model parameters in order to be able to further *subdivide* the metadata groups, better than random? This precise question, of the recovery of nested communities in a two-layer hierarchy, was addressed recently by [133].

Precisely, within the sub-networks defined by each metadata group, under the definition of the toy model in [153], the edges are distributed exactly according to the classical SBM. As such, the key quantity in the overall system becomes the detectability of these SBMs, which in the terminology of the authors corresponds to

$$\mathrm{SNR}_{\mathrm{nested}} = \frac{(\omega_{\mathrm{in}} - \omega_{\mathrm{out}})^2}{q_b(\omega_{\mathrm{in}} + (q_b - 1)\omega_{\mathrm{out}})}, \tag{4.31}$$

identical to that found in the hierarchical setting (*cf.* [133]), only with $q_b = k_2$, $a = \omega_{\mathrm{in}}$, $b = \omega_{\mathrm{out}}$, and where it is no longer possible to relate the denominator to expected degrees at the level above. The dependence on metadata thus only enters via $q = q_b\tilde{q}$, rather than the more complex equation proposed.

Prior to elaborating our own threshold for approximate detectability, to better understand the influence of metadata in a static version of the DSBMM, we may consider an approach akin to [201] for SBMs with unequal groups. That is, if we

consider the FFP for the factorised blockmodel, $\psi_q^{i\to j} \propto \alpha_q$, and simplify our model as much as possible —- remove temporal dynamics, and take $\alpha_q = 1/Q$ for all q as well as the constraint on ω_{qr}, such that

$$\sum_q \omega_{qr} = \sum_r \omega_{qr} = p, \tag{4.32}$$

we can consider finite iterations, starting from the previous, uniform, FFP.

At the first iteration, using Eqn. (4.15) to define the messages, we have

$$\begin{aligned}
\psi_q^{i\to j} &\propto \mathrm{p}_q(x_i)\mathrm{e}^{-Np}p^{d_i-1}, && (4.33)\\
\implies \psi_q^{i\to j} &= \frac{\mathrm{p}_q(x_i)}{\sum_r \mathrm{p}_r(x_i)}, && (4.34)\\
\therefore \psi_q^{i} &\propto \psi_q^{i\to j}\sum_r \omega_{rq}\psi_r^{j\to i}, && (4.35)\\
&\propto \mathrm{p}(x_i \mid z_i)\sum_r \mathrm{p}(z_j = r \mid j \in \partial i, z_i = q)\mathrm{p}(x_j \mid z_j), && (4.36)\\
&= \mathrm{p}(x_i \mid z_i)\mathrm{p}(x_j \mid z_i, j \in \partial i), && (4.37)\\
&= \mathrm{p}(z_i \mid x_i, x_j, j \in \partial i), && (4.38)
\end{aligned}$$

i.e. the Bayesian optimal estimate given x_i and x_j, for $j \in \partial i$. The conversion of ω_{qr} in proportionality to $\mathrm{p}(z_j = r \mid j \in \partial i, z_i = q)\mathrm{p}(x_j \mid z_j)$ follows as given the block connectivity constraint and uniform α, we have by Bayes rule that

$$\begin{aligned}
\mathrm{p}(z_j = r \mid j \in \partial i, z_i = q) &= \frac{\mathrm{p}(j \in \partial i \mid z_j = r, z_i = q)\mathrm{p}(z_j = r)}{\sum_{r'}\mathrm{p}(j \in \partial i \mid z_j = r', z_i = q)\mathrm{p}(z_j = r')},\\
&= \frac{\omega_{qr}/Q}{p/Q} = \omega_{qr}/p
\end{aligned} \tag{4.39}$$

We briefly note that this first iteration returns $1/Q$ iff $\mathrm{p}_r(x_i) = \mathrm{p}(x_i)$ say for all r – in this case, clearly the metadata doesn't help distinguish between different groups in the network, and detectability proceeds as originally. On the other hand, if these new messages were themselves a fixed point, we find that we require $p_{\mathrm{in}} = p_{\mathrm{out}}$, *i.e.* no block structure within the network — and thus the groups are entirely determined by the metadata.

Assuming neither of these hold, and that the metadata is sufficiently minimally informative that we may continue to take the non-edge contribution as approximately

e^{-Np}, continuing to another iteration provides

$$\psi_q^{i\to j} \propto \mathrm{p}_q(x_i) \prod_{k\in\partial i\setminus j} \sum_r \omega_{qr} \frac{\mathrm{p}_r(x_k)}{\sum_{r'} p_{r'}(x_k)}, \tag{4.40}$$

$$\propto \mathrm{p}(x_i \mid z_i) \prod_{k\in\partial i\setminus j} \sum_r \mathrm{p}(k \in \partial i \mid z_k = r, z_i = q)\mathrm{p}(z_k \mid x_k), \tag{4.41}$$

$$\propto \mathrm{p}(x_i \mid z_i) \prod_{k\in\partial i\setminus j} \mathrm{p}(z_i \mid x_k, k \in \partial i), \tag{4.42}$$

and with the marginal factor much as before, we find that again we have a Bayesian optimal estimate for $\mathrm{p}(z_i \mid x_i, \{x_j : j \in \partial i\})$ — the message updates optimally incorporate local metadata at each step. As the BP equations amount to simply applying Bayes rule locally this perhaps should not be surprising, but is nonetheless a nice property!

4.5 Requirements for efficient recovery of groups

We now proceed to the main results of this chapter. To do so, we first define a static toy version of our model, where we restrict the spatial connectivity matrix ω to two values,

$$\omega_{qr}^t = \begin{cases} p_{\mathrm{in}} & \text{if } q = r \\ p_{\mathrm{out}} & \text{else} \end{cases} \tag{4.43}$$

and assume $\alpha_q = 1/Q$ for all q, *i.e.* the groups are all of roughly equal size. One value previously found to be of particular relevance in both static and dynamic cases is then the ratio between these, $\epsilon = p_{\mathrm{out}}/p_{\mathrm{in}}$ [34, 50]. We also specify the average degree in the network, $c = N/Q(p_{\mathrm{in}} + (Q-1)p_{\mathrm{out}})$, thus controlling both the strength of the signal for the blocks at each timestep in terms of the network edges, and the overall density of the networks. In the absence of metadata, this is often referred to as the planted partition model (PPM).

Now in the simplest possible case, the metadata corresponds to noisy observations of the group labels. As such, we impose that with probability ρ we assign the true label, else randomly choose an incorrect label — that is we have

$$p_q^t(x_i^t) = \rho^{\delta_{x_i q}} \frac{1}{Q-1} \prod_{q'\neq q} (1-\rho)^{\delta_{x_i q'}} \tag{4.44}$$

for each q.

Suppose for difficulty that we slightly extend this, by having an analogy of the 'brother' communities that the previous static metadata SBM of [153] propose. Otherwise, as in the conventional toy model, the probability of an edge between groups q, r, denoted ω_{qr}, is p_{in} if $q = r$, else p_{out}.

Specifically, assume there are Q_B possible metadata labels. However, rather than these directly being a noisy observation of the true group label – where it was correct with probability ρ else incorrect – assume that instead it is a noisy observation of the group label at the first level in a hierarchical partition of the network, where each of Q_B groups at the first level themselves contain Q_b groups at the next level. That is, if we observe the label $x_i = q_B$, we take that this means that $z_i \in \{q_{q_B,1}, \ldots, q_{q_B,Q_b}\}$ with probability ρ, else with probability $(1-\rho)$ it belongs to the set of the remaining $Q_b(Q_B - 1)$ groups. In the following, we use the shorthand $q \in q_B$ to denote a single group that belongs to the set of groups $\{q_{q_B,1}, \ldots, q_{q_B,Q_b}\}$, and $q \notin q_B$ to denote a single group in the set of the remaining groups — as the metadata, x_i is only a noisy observation of the first level, note that we are not guaranteed that $x_i = q_B$ if $z_i = q \in q_B$.

Equivalently, as our formulation of metadata generation in the DSBMM posits that it is distributed independently, conditioned on the group label of each node, $\mathrm{p}(x_i \mid z_i)$, this imposes that there are Q_b groups within which the metadata distributions are indistinguishable — that is, we have $\mathrm{p}(x_i \mid z_i = q)$ is identical for each such set of Q_b groups, $q \in \{q_{q_B,1}, ..., q_{q_B,Q_b}\}$, where there are Q_B distinct such sets (and thus $Q = Q_B Q_b$ groups in total), so $\mathrm{p}(x_i \mid z_i = q_{q_B,1}) = \mathrm{p}(x_i \mid z_i = q_{q_B,2})$ *etc.* The metadata alone cannot then help you determine which of these subgroups is more likely.

This means that

$$\mathrm{p}(x_i = q_B \mid z_i = q) = \begin{cases} \rho & q \in q_B, \\ (1-\rho)/(Q_B - 1) & q \notin q_B. \end{cases} \tag{4.45}$$

Inspired by the procedure used to determine a limit in the flawed model of [153], we explore the analogous – but now complicated – process. Given the observed metadata, we restrict our focus to the subgraph produced by metadata label q_B say, within which the expected group distribution is now governed by

$$\mathrm{p}(z_i = q \mid x_i = q_B) = \frac{\mathrm{p}(x_i = q_B \mid z_i = q)\mathrm{p}(z_i = q)}{\mathrm{p}(x_i = q_B)}, \tag{4.46}$$

by Bayes rule, where in the full graph, recalling that we denote the prior distribution

$\mathrm{p}(z_i = q) = \alpha_q$,

$$\begin{aligned}\mathrm{p}(x_i = q_B) &= \sum_{q'_B} \sum_{q' \in q'_B} \mathrm{p}(z_i = q')\mathrm{p}(x_i \mid z_i = q') \\ &= \sum_{q'_B} \sum_{q' \in q'_B} \alpha_{q'} \left(\rho \delta_{q_B q'_B} + \frac{(1-\rho)(1-\delta_{q_B q'_B})}{(Q_B - 1)} \right),\end{aligned} \tag{4.47}$$

so we have that

$$\begin{aligned}\mathrm{p}(z_i = q \mid x_i = q_B) &= \frac{(\rho \delta_{q \in q_B} + (1-\rho)(1-\delta_{q \in q_B})/(Q_B - 1))\alpha_q}{\sum_{q'_B} \sum_{q' \in q'_B} \alpha_{q'} (\rho \delta_{q_B q'_B} + (1-\rho)(1-\delta_{q_B q'_B})/(Q_B - 1))}, \\ &= \begin{cases} \rho \alpha_q / \mathrm{p}(x_i = q_B) & q \in q_B, \\ (1-\rho)\alpha_q / (Q_B - 1)\mathrm{p}(x_i = q_B) & q \notin q_B. \end{cases}\end{aligned} \tag{4.48}$$

If all groups are equal size, *i.e.* $\alpha_q = 1/(Q_B Q_b)$ for all q, then

$$\begin{aligned}\mathrm{p}(x_i = q_B) &= 1/(Q_B Q_b) \sum_{q'_B} \sum_{q' \in q'_B} (\rho \delta_{q_B q'_B} + (1-\rho)(1-\delta_{q_B q'_B})/(Q_B - 1)), \\ &= 1/Q_B \sum_{q'_B} (\rho \delta_{q_B q'_B} + (1-\rho)(1-\delta_{q_B q'_B})/(Q_B - 1)), \\ &= 1/Q_B,\end{aligned} \tag{4.49}$$

as we should expect, and so

$$\mathrm{p}(z_i = q \mid x_i = q_B) = \begin{cases} \rho/Q_b & q \in q_B, \\ (1-\rho)/(Q_b(Q_B - 1)) & q \notin q_B, \end{cases} \tag{4.50}$$

i.e. with probability ρ the nodes true group is one of the 'brother' communities of the metadata group q_B, else it was labelled incorrectly and belongs to a group in one of the other communities at the metadata level, as we would expect.

With this as $\tilde{\alpha}_q$ in our subgraph, we can now consider this as a classic SBM, as all

metadata labels are constant. As there are unequal size groups,

$$\begin{aligned}\mathbb{E}[d_i \mid z_i = q] &= \sum_j \mathbb{E}[a_{ij} \mid z_i = q], \\ &= (N_{q_B} - 1) \sum_r \omega_{qr} \tilde{\alpha}_r, \\ &= \begin{cases} \frac{N_{q_B}-1}{Q_b}(\rho p_{\text{in}} + [(Q_b - 1)\rho + Q_b(1-\rho)]p_{\text{out}}) & q \in q_B, \\ \frac{N_{q_B}-1}{Q_b}(\frac{(1-\rho)}{Q_B-1}p_{\text{in}} + [Q_b\rho + Q_b(1-\rho) + (Q_b-1)\frac{(1-\rho)}{Q_B-1}]p_{\text{out}}) & q \notin q_B, \end{cases}\end{aligned} \tag{4.51}$$

i.e. we can use the observed degrees to at least approximately cluster into the metadata groups, so once again we verify there is no true (weak) detectability limit — as we showed before by using the metadata directly.

However, we can appeal to [1], and instead look at the SNR, *i.e.* conditions under which *efficient* inference is possible. We define P to be the diagonal, $Q \times Q$ matrix with $\tilde{\alpha}$ on the diagonal, and Q the corresponding full expected block connectivity matrix, with Np_{in} on the diagonal and Np_{out} elsewhere, where N is the number of nodes in the network. Efficient detectability for the classic SBM is then possible if

$$\text{SNR} = \lambda_2^2/\lambda_1 > 1, \tag{4.52}$$

where λ_1 and λ_2 are the largest and second largest distinct eigenvalues of PQ respectively.

In our case, without loss of generality we may assume that $q \in q_B$ make up the first Q_b groups, such that

$$PQ = \frac{N}{Q_b}\begin{pmatrix} \rho p_{\text{in}} & \rho p_{\text{out}} & \cdots & & \\ \rho p_{\text{out}} & \rho p_{\text{in}} & \cdots & & \\ \vdots & & \ddots & & \vdots \\ \frac{(1-\rho)}{(Q_B-1)}p_{\text{out}} & \cdots & & \frac{(1-\rho)}{(Q_B-1)}p_{\text{in}} & \cdots \\ \vdots & & & \ddots & \end{pmatrix}, \tag{4.53}$$

i.e. with $\tilde{Q}_{Q_b} = (p_{\text{in}} - p_{\text{out}})I^{Q_b \times Q_b} + p_{\text{out}}J_{Q_b}$ the block matrix restricted to Q_b dimensions, where J_{Q_b} is the ones matrix in $Q_b \times Q_b$ dimensions, this is

$$PQ = \frac{N}{Q_b}\begin{pmatrix} \rho\tilde{Q}_{Q_b} & \rho p_{\text{out}}J_{Q_b} & \cdots \\ \frac{(1-\rho)}{(Q_B-1)}p_{\text{out}}J_{Q_b} & \frac{(1-\rho)}{(Q_B-1)}\tilde{Q}_{Q_b} \cdots & \\ \vdots & \ddots & \end{pmatrix}, \tag{4.54}$$

i.e. a block matrix of $Q_B \times Q_B$ blocks, each of size $Q_b \times Q_b$, with the first row of blocks multiplied by ρ, and the others a reordering of this row but multiplied by $(1-\rho)/(Q_B-1)$ instead.

The algebra for the full derivation of the eigenvalues needed for the SNR is rather convoluted, thus within this chapter we provide only a brief sketch of the process — complete working is presented in App. B. The key steps are as follows:

(i) Recognise that the problem is degenerate, in that we may only proceed if $\rho \notin \{0, 1, 1/Q_B\}$. These cases intuitively correspond to the toy model effectively reducing to a (possibly hierarchical) classical SBM: when $\rho \in \{0, 1\}$, the metadata is perfectly informative, and thus may be used to partition the network into Q_B sub-graphs of homogeneous metadata, each of which may be expressed purely using a classical SBM with $Q = \{Q_b(Q_B-1), Q_b\}$ respectively (analogously to [133]); when $\rho = 1/Q_B$, the metadata is perfectly random, and thus provides no information about node groups — the network is simply a classical SBM with $Q = Q_b Q_B$ groups. As such, in the following we assume $0 < \rho < 1$, with $\rho \neq 1/Q_B$.

(ii) Note that given invertible diagonal blocks, the determinant of a 2×2 block matrix may be expressed as

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(A)\det(D - CA^{-1}B). \tag{4.55}$$

In particular, as PQ may be considered as such a block matrix when $\rho \notin \{0, 1, 1/Q_B\}$, we may use this to derive the characteristic polynomial of PQ, $\phi_{PQ}(\lambda) = \det(PQ - \lambda I) = \det(A - \lambda I)\det(D - \lambda I - C(A - \lambda I)^{-1}B)$, of which the eigenvalues are roots.

(iii) In our case, $(A - \lambda I)$ is a rank-one update of a multiple of the identity matrix, thus we may use the Sherman-Morrison formula to compute $(A - \lambda I)^{-1}$.

(iv) This subsequently allows direct calculation of the internal term in the second factor, $D - C(A - \lambda I)^{-1}B$, which remains a rank-one update of a multiple of the identity matrix.

(v) Hence, the matrices for each determinant factor are rank-one updates of an invertible matrix, thus may be derived exactly using the matrix determinant lemma.

(vi) By combining the resulting terms, and finally using the quadratic formula, we find four unique roots — the eigenvalues desired.

As the equations for these eigenvalues in the general case are rather long, for brevity we restrict ourselves herein to elaborating the case of two metadata groups, *i.e.* $Q_B = 2$. For this particular case, the four distinct eigenvalues for PQ are as follows:

$$\lambda_1 = \frac{N}{Q_b}\rho(p_{\text{in}} - p_{\text{out}}), \quad \text{with mult. } (Q_b - 1), \tag{4.56}$$

$$\lambda_2 = \frac{N}{Q_b}(1-\rho)(p_{\text{in}} - p_{\text{out}}), \quad \text{with mult. } Q_b(Q_B - 1) - 1, \tag{4.57}$$

$$\begin{aligned}\lambda_{3,4} = \frac{N}{2Q_b}\Bigg(&p_{\text{in}} + (Q_b - 1)p_{\text{out}} \\ &\pm \sqrt{(2\rho-1)^2(p_{\text{in}} - p_{\text{out}})(p_{\text{in}} + (2Q_b - 1)p_{\text{out}}) + (Q_b p_{\text{out}})^2}\Bigg), \\ &\text{both with mult. 1,}\end{aligned} \tag{4.58}$$

where we assume λ_3 takes the minus and thus $\lambda_3 \leq \lambda_4$.

Within the permissible range of parameters, it always holds that λ_4 is the largest eigenvalue of this set. However, the second largest eigenvalue varies depending on the area of parameter space. In particular, for $0 < \rho < 1/2$, it is clear that $\lambda_2 > \lambda_1$, while the reverse is true for $1/2 < \rho < 1$. Further, in both regions, the maximum of $\lambda_{1,2}$ is greater than λ_3 if and only if $p_{\text{in}} > p_{\text{out}}$. That is, we have three distinct cases to consider:

(i) $0 < \rho < 1/2$, $p_{\text{in}} > p_{\text{out}}$: $\lambda_4 > \lambda_2 > \cdots$;

(ii) $1/2 < \rho < 1$, $p_{\text{in}} > p_{\text{out}}$: $\lambda_4 > \lambda_1 > \cdots$;

(iii) $0 < \rho < 1$, $p_{\text{in}} < p_{\text{out}}$: $\lambda_4 > \lambda_3 > \cdots$.

In practice, as the model in this case is symmetric in ρ about $\rho = 1/2$, cases (i) and (ii) are effectively identical, and mapping $\rho \to (1-\rho)$ will provide the equation for the limit in case (ii) from that for case (i) (or vice versa).

Finally, this allows the calculation of possible SNR thresholds for efficient inference. Introducing the expected block degrees $c_{\text{in,out}} = N p_{\text{in,out}}$, for case (i) we require that

$$\begin{aligned}&\lambda_2^2 > \lambda_4, \\ \textit{i.e. } &(1-\rho)^2(c_{\text{in}} - c_{\text{out}})^2 > \frac{Q_b}{2}\Bigg(c_{\text{in}} + (Q_b - 1)c_{\text{out}} \\ &\qquad + \sqrt{(2\rho-1)^2(c_{\text{in}} - c_{\text{out}})(c_{\text{in}} + (2Q_b - 1)c_{\text{out}}) + (Q_b c_{\text{out}})^2}\Bigg),\end{aligned}$$

and hence similarly for case (ii),

$$\rho^2(c_{\text{in}} - c_{\text{out}})^2 > \frac{Q_b}{2}\Bigg(c_{\text{in}} + (Q_b - 1)c_{\text{out}} + \sqrt{(2\rho - 1)^2(c_{\text{in}} - c_{\text{out}})(c_{\text{in}} + (2Q_b - 1)c_{\text{out}}) + (Q_b c_{\text{out}})^2}\Bigg),$$

simply replacing $(1-\rho)^2$ on the LHS with ρ^2.

For comparability with previous works, following [50], we introduce the expected total number of within-group edges at the first level of the hierarchy — if this hierarchy is perfect — as

$$g = Q_B c_{\text{in}}^1 = Q_B(c_{\text{in}} + (Q_b - 1)c_{\text{out}})/(Q_b Q_B), \tag{4.59}$$

hence we may substitute $(c_{\text{in}} + (Q_b - 1)c_{\text{out}}) = Q_b g$. In terms of g and $\epsilon = c_{\text{out}}/c_{\text{in}}$, the requirement for case (i) corresponds to

$$(1-\rho)^2(1-\epsilon)^2 > \frac{(1 + (Q_b - 1)\epsilon)}{2g}\Bigg(1 + (Q_b - 1)\epsilon + \sqrt{(2\rho - 1)^2(1-\epsilon)(1 + (2Q_b - 1)\epsilon) + (Q_b\epsilon)^2}\Bigg),$$

Comparing to the classical requirement for weak detectability without metadata, of

$$(c_{\text{in}} - c_{\text{out}})^2 > 2Q_b(c_{\text{in}} + (2Q_b - 1)c_{\text{out}}) \tag{4.60}$$

for the same number of total groups, if we have $0 < \epsilon < 1$ and either case (i) or case (ii) applies, we may directly compute that the metadata bound above is less strict. However, if $\epsilon > 1$ and thus case (iii) holds, we find that the requirements on parameters for efficient inference appear to always be more lax in the classical case than our own.

In Fig. 4.1(a) we display the ratio between our estimate for the SNR given metadata, and the classical SNR, for $0 < \rho < 1$, $0 < \epsilon < 2$, $Q_b = 1$, $Q_B = 2$. We observe that as expected, when metadata is weakly informative ($\rho \approx 1/2$), the two are near-identical and the ratio tends to one, while as ρ becomes increasingly informative the relative benefit further increases. However, for $\epsilon > 1$ the SNR suggests that including metadata has *increased* the difficulty of detecting groups. Intuitively, this must be an artifact of the process used to estimate the SNR, as informative metadata ($\rho \neq 1/2$) must always imply some improved capability in detection. The origin of this artifact lies in the subsampling step, where for tractable analysis we restrict our focus to a subgraph

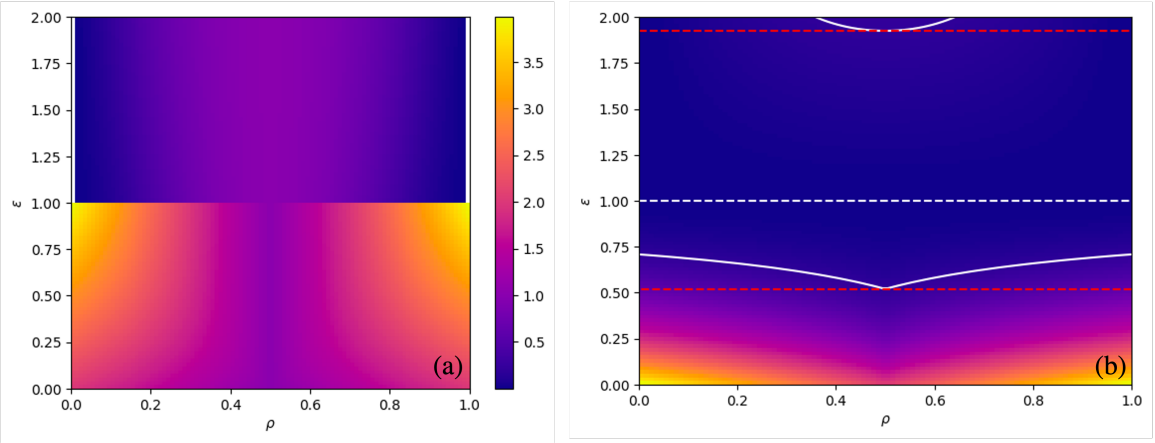


FIGURE 4.1: In subfigure (a), we display the ratio between estimated SNR for groups in the model with metadata to that without (*i.e.* the classical SNR), for $0 < \rho < 1$, $0 < \epsilon < 2$, $Q_b = 1$, $Q_B = 2$. For $\epsilon < 1$, this is always > 1, suggesting improved detectability, while for $\epsilon > 1$ it is always < 1, suggesting *increased* difficulty in detection. However, intuitively this must be an artifact of the process used to estimate the SNR, as informative metadata ($\rho \neq 1/2$) must always allow some improved capability in detection — see main text for discussion. In subfigure (b), we display actual estimated SNR values for average degree $c = 20$, over the same parameter ranges — note the scale is different to (a), but still with lighter meaning higher (better) values. In solid white, we show the contour for the key threshold, $SNR = 1$, when including metadata, in dashed red the same threshold for the classical model, and in dashed white the line $\epsilon = 1$, beyond which the SNR is likely unreliable. Again, if $\epsilon < 1$ including informative metadata clearly improves detectability, while the reverse appears to hold for $\epsilon > 1$.

with homogeneous metadata. This introduces a bias towards associative networks, for which subsampling according to informative metadata results in a relative increase in the density of the important edges between related nodes, and hence simplifies detection of groups. On the other hand, in dis-associative networks, as generated when $\epsilon > 1$, performing such subsampling has a bias towards discarding the most important edges — those between groups. Indeed, more informative metadata discards a larger proportion of such edges, and so actually increases the difficulty of detecting groups within the subgraph, as observed. The solution is to instead subsample a bi-partite network, where all nodes of each class have the same metadata label — we shall not elaborate upon this further herein, but will release the full details as part of a paper in the coming months.

In subfigure (b), we display actual estimated SNR values for average degree $c = 20$, over the same parameter ranges — note the scale is different to subfigure (a), but still with lighter meaning higher (better) values. In solid white, we show the contour

for the key threshold, $SNR = 1$, when including metadata, in dashed red the same threshold for the classical model, and in dashed white the line $\epsilon = 1$, beyond which the SNR presented herein is likely unreliable, as discussed above. We may now observe clearly that, if $\epsilon < 1$, including informative metadata demonstrably improves the ease of detecting groups, while the reverse appears to hold for $\epsilon > 1$.

For reference, in the hierarchical detectability work of [50], they find the requirement

$$\begin{aligned} \frac{(c_{\text{in}} - c_{\text{out}})^2}{Q_b(c_{\text{in}} + (Q_b - 1)c_{\text{out}})} &\geq 1, \\ \therefore \frac{(1-\epsilon)^2}{Q_b(1 + (Q_b - 1)\epsilon)^2} &\geq \frac{1}{Q_b g}, \\ i.e.\ \frac{(1-\epsilon)}{(1 + (Q_b - 1)\epsilon)} &\geq \frac{1}{\sqrt{g}}, \end{aligned} \tag{4.61}$$

and as this requires perfect labelling of the first level in the hierarchy ($\rho = 1$), this should be an upper bound to the metadata detectability case, — *i.e.* $\rho = 1/2$ corresponds to the classical bound (hardest to recover groups), and $\rho = 1$ provides the hierarchical bound (easiest to recover groups), with our bound between the two.

Indeed, as $\rho \to 1$, for $\epsilon < 1$ we have

$$\lambda_4 \to \frac{1}{2}\left(c_{\text{in}} + (Q_b - 1)c_{\text{out}} + \sqrt{(c_{\text{in}} - c_{\text{out}})(c_{\text{in}} + (2Q_b - 1)c_{\text{out}}) + (Q_b c_{\text{out}})^2}\right), \tag{4.62}$$

$$= \frac{1}{2}\left(c_{\text{in}} + (Q_b - 1)c_{\text{out}} + \sqrt{c_{\text{in}}^2 + 2(Q_b - 1)c_{\text{in}}c_{\text{out}} + (Q_b^2 - 2Q_b + 1)c_{\text{out}}^2}\right), \tag{4.63}$$

$$= c_{\text{in}} + (Q_b - 1)c_{\text{out}}, \tag{4.64}$$

i.e. the largest eigenvalue in the strict hierarchical formulation, thus recovering the identical bound.

The classical bound in terms of ϵ, and the average degree $c = (c_{\text{in}} + (Q_b Q_B - 1)c_{\text{out}})/(Q_b Q_B)$, is

$$\frac{(1-\epsilon)^2 c}{(1 + (Q_b Q_B - 1)\epsilon)^2} \geq 1. \tag{4.65}$$

Note that

$$\begin{aligned} \frac{g}{c} &= \frac{Q_B(c_{\text{in}} + (Q_b - 1)c_{\text{out}})}{(c_{\text{in}} + (Q_b Q_B - 1)c_{\text{out}})}, \\ &= \frac{Q_B(1 + (Q_b - 1)\epsilon)}{1 + (Q_b Q_B - 1)\epsilon}, \\ \therefore g &= \frac{Q_B(1 + (Q_b - 1)\epsilon)}{1 + (Q_b Q_B - 1)\epsilon} c. \end{aligned} \tag{4.66}$$

This means that the hierarchical bound reads

$$\begin{aligned} \frac{(1-\epsilon)}{(1+(Q_b-1)\epsilon)} &\geq \frac{1}{\sqrt{g}}, \\ &\geq \frac{1}{\sqrt{c}}\sqrt{\frac{1+(Q_bQ_B-1)\epsilon}{Q_B(1+(Q_b-1)\epsilon)}}, \\ \therefore \frac{(1-\epsilon)}{\sqrt{(1+(Q_b-1)\epsilon)(1+(Q_bQ_B-1)\epsilon)}} &\geq \frac{1}{\sqrt{Q_Bc}}, \\ \therefore \frac{(1-\epsilon)}{\sqrt{Q_b(1+(Q_b-1)\epsilon)(1+(Q_bQ_B-1)\epsilon)}} &\geq \frac{1}{\sqrt{Q_bQ_Bc}}. \end{aligned} \tag{4.67}$$

Now clearly $\sqrt{Q_b(1+(Q_b-1)\epsilon)(1+(Q_bQ_B-1)\epsilon)} \leq (1+(Q_bQ_B-1)\epsilon)$ if $Q_b = 1$ for all $Q_B \geq 2$, in which case the hierarchical bound is more generous — indeed if $\rho = 1$ the first level of hierarchy is perfectly known, and if $Q_b = 1$ then this provides all meso-scale structure present, and thus in this case the bound holds for all ϵ for $c > 1$, *i.e.* always. Otherwise, for this bound to be more generous when $Q_b \geq 2$ we require

$$Q_B > Q_b \text{ and } \frac{Q_b - 1}{Q_b(Q_B - Q_b + 1) - 1} \leq \epsilon < 1, \tag{4.68}$$

and if these requirements do not hold, it turns out that the classical bound is more lax. Effectively, this is informs us that if there are fewer groups at the first level than the second level, then constraining our focus to a single label, before proceeding to seek groups in this subgraph, is harder than just looking for the coarse partition. Even if there are sufficient groups at the coarse level, the second condition suggests that if ϵ is sufficiently small, *i.e.* the blocks are sufficiently clearly defined by their intra-group connections, then once again it is better to seek the groups directly, rather than discarding information by restricting to a subgraph.

Using this same transformation, our bound for $\epsilon < 1$ when $Q_B = 2$ becomes

$$\begin{aligned} (1-\rho)^2(1-\epsilon)^2 > \frac{(1+(2Q_b-1)\epsilon)}{4}c\Big(1+(Q_b-1)\epsilon \\ + \sqrt{(2\rho-1)^2(1-\epsilon)(1+(2Q_b-1)\epsilon)+(Q_b\epsilon)^2}\Big), \end{aligned}$$

and once again we may easily verify that as $\rho \to 1/2$ this matches with the classical SNR.

We explore detectability in practice in the following chapter.

4.6 Discussion

In this chapter, we have addressed an important gap in the literature by providing a means of understanding how metadata affects the efficiency of recovery of groups in a network. In doing so, we redress the mistakes of a prior attempt to do so, and provide the foundations for a scalable inference method for the DSBMM. We build on this in the following chapter, using the BP inference procedure to consider networks orders of magnitude larger than previously possible, and performing empirical dynamic detectability tests to evaluate the success of the method well beyond the previous limit.

With regards to further work, the major aspect of the DSBMM that we neglected in our detectability analysis was of course the group (and correspondingly network) dynamics. The form of SNR used has not had its applicability investigated in this case – it would be a very interesting avenue to pursue in future. Precisely, to determine an alternative measure that might be used to investigate the improvement in the efficiency of detectability from metadata in *dynamic* networks, as considered herein. Given the results of [50], while – for the reasons described above – their argument cannot be followed exactly in the presence of metadata, we would nonetheless expect that such a measure would depend on the second largest eigenvalue of the group transition matrix. Importantly, akin to our own results for metadata, they find that in a toy model where a single parameter governs the likelihood of remaining in a group, any difference from uniformly random transitions makes detecting groups easier. As such, the measure described in this chapter should provide a lower, or worst-case bound for efficient detectability in dynamic networks with metadata.

Chapter 5

Dynamic BP and an application

In this chapter, we extend the belief propagation procedure previously described to the full dynamic case, and perform numerous experiments using this method. Concretely, we describe a way to fit the model using belief propagation (BP), which for sparse networks allows each iteration of the algorithm to be completed in $\mathcal{O}(NTQ^2)$ time, *i.e.* near linear in the size of the network, while providing various additional benefits over comparable alternatives, *e.g.* greedy methods.

We begin in Sec. 5.1 by describing the equations necessary for the DSBMM both where edges are described by the binary SBM, and that for the DC-SBM of [69]. We also briefly describe a greedy inference scheme for the model in Sec. 5.2, which we intend to use in future to help initialise the BP system. We believe such initialisation to be of significant importance to the success of the procedure, and the greedy scheme is proposed as an alternative to the spectral approaches such as [157, 50] obtained from linearisation of the BP system around the trivial fixed point, which no longer holds precisely, as discussed in the previous chapter.

We then proceed to perform numerous experiments on networks simulated from the model in Sec. 5.3, to explore the performance of the BP inference procedure proposed in greater depth. Precisely, in Sec. 5.3.1 we first investigate how performance varies with the tuning parameter – governing the relative contribution of metadata – for simulations where we specify the degree of alignment between metadata and network partitions. This elucidates the importance of the parameter highlighted in the previous chapter. Next, in Sec. 5.3.2, we empirically probe how successfully a simplified version of our model can recover network groups, near and beyond the detectability limit found in the case where metadata is neglected. We conclude our experiments on synthetic networks by demonstrating the improved scaling that BP provides as the number of nodes, N, increases.

Continuing from this, in Sec. 5.4 we present explorative results of the model on a medium-scale ($\mathcal{O}(5 \times 10^4)$) network of Latin American authors, several orders of magnitude larger than that considered in Chap. 3. To further accelerate the inference procedure, we also describe a recursive method that provides a top-down hierarchical partition for the dynamic network, while largely avoiding the quadratic dependence of the method on Q. Additionally, we suggest a heuristic measure that may be used to 'auto-tune' the model, by selecting the tuning parameter in order to approximately balance the average contribution of edge and metadata terms for each node. Finally, we conclude the chapter in Sec. 5.5, with a brief discussion and several avenues for future work.

5.1 Belief propagation for the DSBMM

Following the steps described in the previous chapter, in this section we derive the message equations for BP inference in the DSBMM, and elaborate how to update model parameters given the estimated messages, in an iterative variational expectation maximisation procedure.

With respect to the factorised framing of Eqn. (4.1), the DSBMM has node-wise factors at each timestep

$$f_i^t(z_i^t = q) = \alpha_q^{\delta_{t,0}} \mathrm{p}_q^t(x_i^t), \tag{5.1}$$

where recall α_q is the prior probability of belonging to a group q at $t = 0$, and $\mathrm{p}_q^t(x_i^t)$ is the distribution of nodal metadata, x_i^t of i at t in group q. There are then two different sets of pairwise factors. One of these is spatial, identical to that in static SBMs:

$$f_{ij}^t(z_i^t = q, z_j^t = r) = \phi_{qr}(A_{ij}^t), \tag{5.2}$$

where $\phi_{qr}(A_{ij}^t)$ corresponds to whichever block likelihood chosen to describe the potential edge A_{ij}^t, *e.g.* $\omega_{qr}^{A_{ij}^t}(1-\omega_{qr})^{1-A_{ij}^t}$ for the NDC-SBM, or $\mathrm{Pois}(d_i^t d_j^t \lambda_{qr}^t)$ for the DC-SBM.

The other pairwise factor comes from the group dynamics, specifically how group membership changes over time via the transition matrix π, *i.e.*

$$f_i^{t,t+1}(z_i^t = q, z_i^{t+1} = q') = \pi_{qq'}. \tag{5.3}$$

We may then directly substitute these factors into Eqn. (4.15) to obtain updates for messages acting over both spatial and temporal edges. In the following, $\psi_q^{i\to j}(t)$

denotes spatial messages from i to j at t, carrying an estimate of the probability of i belonging to q. Analogously, $\psi_q^{i(t)\to i(t\pm1)}$ denotes temporal messages carrying an estimate of the probability of i belonging to q at t, sent to itself at adjacent timesteps. Considering for clarity the NDC-SBM to begin with, we find the message updates for spatial messages

$$\begin{aligned}\psi_q^{i\to j}(t) \propto \mathrm{p}_q^t(x_i^t)^\theta &\left(\sum_q \pi_{qr}\psi_q^{i(t-1)\to i(t)}\right)\left(\sum_q \pi_{rq}\psi_q^{i(t+1)\to i(t)}\right)\\ &\times\left(\prod_{\substack{k:(i,k)\in\mathcal{E}\\k\neq j}}\sum_q \psi_q^{k\to i}(t)\omega_{qr}^t\right)\left(\prod_{\substack{k:(i,k)\notin\mathcal{E}\\k\neq j}}\sum_q \psi_q^{k\to i}(t)(1-\omega_{qr}^t)\right),\end{aligned} \tag{5.4}$$

and forward temporal messages

$$\begin{aligned}\psi_r^{i(t)\to i(t+1)} \propto \mathrm{p}_r^t(x_i^t)^\theta &\left(\sum_q \pi_{qr}\psi_q^{i(t-1)\to i(t)}\right)\\ &\times\left(\prod_{k:(i,k)\in\mathcal{E}}\sum_q \psi_q^{k\to i}(t)\omega_{qr}^t\right)\left(\prod_{\substack{k:(i,k)\notin\mathcal{E}\\k\neq j}}\sum_q \psi_q^{k\to i}(t)(1-\omega_{qr}^t)\right).\end{aligned} \tag{5.5}$$

The final equation for backward temporal messages, $\psi_r^{i(t)\to i(t-1)}$, is much as for $\psi_r^{i(t)\to i(t+1)}$, only necessarily involving messages from $i(t+1)$ instead, and π_{rq} rather than π_{qr}. For $t=T$, terms involving $i(t+1)$ are not present, while for $t=1$ (or nodes not present at the previous timestep) we require a factor of α_q in lieu of terms involving $i(t-1)$. We highlight in blue the metadata contribution, as this is the key distinction from otherwise similar update equations previously described [50], alongside the more general form of group transitions permitted.

The astute reader will notice that due to the presence of non-edge factors, each update equation contains $\mathcal{O}(QN^2)$ terms, which immediately suggests at least quadratic scaling of computational complexity with respect to the size of networks considered. Thankfully, for sparse networks we have $\omega_{qr}^t = \mathcal{O}(1/N)$ in general, as the corresponding expected number of edges from a node in block q, $N\alpha_r\omega_{qr}^t = \mathcal{O}(1)$, hence we can make a series of approximations for the term over non-edges to reduce the complexity.

Firstly, normalisation ensures that $\sum_q \psi_q^{k\to i}(t) = 1$, and so $\sum_q \psi_q^{k\to i}(t)(1-\omega_{qr}^t) = 1 - \sum_q \omega_{qr}^t\psi_q^{k\to i}(t)$. Next, we have from the definition of the full marginals that $\psi_q^{k,t} = \psi_q^{k\to i}(t)\sum_r(1-\omega_{qr}^t)\psi_r^{i\to k}(t)$ for k such that $(k,i)\notin\mathcal{E}$, with $\mathcal{E}$ the set of edges, which given the sparsity assumption then means that for such non-adjacent nodes we

have $\psi_q^{k\to i}(t) = \psi_q^{k,t} + \mathcal{O}(1/N)$. Together, these take us to

$$\prod_{\substack{k:(i,k)\notin\mathcal{E}\\ k\neq j}} \sum_r \psi_r^{k\to i}(t)(1-\omega_{rq}^t) \approx \prod_{\substack{k:(i,k)\notin\mathcal{E}\\ k\neq j}} \left(1 - \sum_q \psi_r^k(t)\omega_{rq}^t\right), \tag{5.6}$$

at leading order. Now we may again invoke sparsity, and use that for small x we can approximate $(1-x) \approx \mathrm{e}^{-x}$, to find

$$\prod_{\substack{k:(i,k)\notin\mathcal{E}\\ k\neq j}} \sum_r \psi_r^{k\to i}(t)(1-\omega_{rq}^t) \approx \mathrm{e}^{-h_q^t}, \tag{5.7}$$

where our final approximation to remove dependence on i is to take the product over all k, such that we only need to calculate the 'external fields' for each q, t, given by

$$h_q^t = \sum_k \sum_r \omega_{rq}^t \psi_r^{k,t}. \tag{5.8}$$

In this chapter, we also permit degree-correction, but along the lines of [69]. That is, rather than proceeding as proposed in the previous chapter where there was a parameter controlling sparsity, and a ZTP distribution over non-zero edges, edges are distributed solely according to a Poisson distribution, *i.e.*

$$\mathrm{p}(a_{ij}^t \mid z_i^t, z_j^t) \sim \mathrm{Pois}(\lambda_{z_i^t, z_j^t}^t d_i^t d_j^t), \tag{5.9}$$

where λ_{qr}^t is a parameter for the rate of edges between blocks q and r at t, and z_i^t, d_i^t are the block and degree of node i at time t as before.

For this edge distribution, the argument for the NDC case above is slightly modified, but proceeds analogously. Instead, we have that the corresponding non-edge term

$$\prod_{\substack{k:(i,k)\notin\mathcal{E}\\ k\neq j}} \sum_r \psi_r^{k\to i}(t)\mathrm{e}^{-d_k^t d_i^t \omega_{rq}^t} \approx \prod_k \left(1 - d_k^t d_i^t \sum_r \psi_r^{k,t}\omega_{rq}^t\right), \tag{5.10}$$

on the assumption that $d_k^t d_i^t \omega_{rq}^t \ll 1$, so that $\mathrm{e}^{-d_k^t d_i^t \omega_{rq}^t} \approx (1 - d_k^t d_i^t \omega_{rq}^t)$, and again using that $\sum_r \psi_r^{k,t} = 1$. From here we assume $d_k^t d_i^t \sum_r \psi_r^{k,t}\omega_{rq}^t \ll 1$ so that we may once more use $(1-x) \approx \mathrm{e}^{-x}$, and find that our new external field term (now also dependent on i) is now $\mathrm{e}^{-h_q^{i,t}}$, with

$$h_q^{i,t} = d_i^t \sum_{k,r} d_k^t \omega_{rq}^t \psi_r^{k,t}. \tag{5.11}$$

Returning to the NDC model to avoid notational clutter, using this external field approximation – which may be computed once upon initialisation, then updated at each iteration for only $\mathcal{O}(1)$ cost – our message equations have become

$$\psi_r^{i\to j}(t) \propto \mathrm{p}_r^t(x_i^t)^\theta \left(\sum_q \pi_{qr}\psi_q^{i(t-1)\to i(t)}\right)\left(\sum_q \pi_{rq}\psi_q^{i(t+1)\to i(t)}\right) \mathrm{e}^{-h_r^t} \prod_{\substack{k:(i,k)\in\mathcal{E}\\ k\neq j}} \sum_q \psi_q^{k\to i}(t)\omega_{qr}^t, \tag{5.12}$$

for spatial messages, and

$$\psi_r^{i(t)\to i(t+1)} \propto \mathrm{p}_r^t(x_i^t)^\theta \left(\sum_q \pi_{qr}\psi_q^{i(t-1)\to i(t)}\right) \mathrm{e}^{-h_r^t} \prod_{k:(i,k)\in\mathcal{E}} \sum_q \psi_q^{k\to i}(t)\omega_{qr}^t, \tag{5.13}$$

for forward temporal messages, with the same simple change for backward temporal messages as previously. Calculating all messages in each sweep in sparse networks can thus be performed in $\mathcal{O}(QNTd)$ for d the average degree in the network, *i.e.* near linear time in the size of the network — a significant improvement!

To conclude this section, we use Eqn. (4.17) to derive the free energy of our model. We have that

$$\psi_{qr}^{ijt} = \frac{\omega_{qr}^t(\psi_q^{i\to j}(t)\psi_r^{j\to i}(t) + \psi_r^{i\to j}(t)\psi_q^{j\to i}(t))}{Z^{ijt}}, \tag{5.14}$$

$$\psi_{qq'}^{it} = \frac{\pi_{qq'}\psi_q^{i(t-1)\to i(t)}\psi_{q'}^{i(t)\to i(t-1)}}{Z^{it}}, \tag{5.15}$$

correspond to the spatial and temporal two-point marginals respectively, with normalisation performed over all (q, r) for spatial messages, and (q, q') for temporal messages, Note that in the spatial two-point marginals, Eqn. (5.14), the second term is only present in the undirected case, and for $r \neq q$ — this is used in the following chapter, after modifying the message equations suitably. As the full model has terms for both edges and non-edges, we then have that

$$F_{\text{Bethe}} = \sum_{i,t} \log Z_{i,t} - \sum_{i,j,t\in\mathcal{E}} \log Z_{ijt} - \sum_{i,j,t\notin\mathcal{E}} \log \tilde{Z}_{ijt} - \sum_{i,t=2} \log Z_{i,t-1,t} \tag{5.16}$$

where much as before, $Z_{i,t}$ is the normalisation term for the marginal of i at t, and Z_{ijt} and $Z_{i,t-1,t}$ are the corresponding normalisation terms for the two-point spatial and temporal marginals respectively.

However, for the contribution from spatial *non*-edges, $\sum_{i,j,t\notin\mathcal{E}} \log \tilde{Z}_{ijt}$, as previously performed in the static case [34], to avoid performing $\mathcal{O}(QTN^2)$ calculations, we instead

use the single point marginals. For the NDC-DSBMM, the term then becomes

$$\sum_{i,j,t\notin\mathcal{E}} \log \tilde{Z}_{ijt} = \sum_{i,j,t\notin\mathcal{E}} \log \sum_{q,r} (1-\omega^t_{qr})\psi^{it}_q \psi^{jt}_r, \tag{5.17}$$

$$= \sum_{i,j,t\notin\mathcal{E}} \log \left(1 - \sum_{q,r} \omega^t_{qr}\psi^{it}_q\psi^{jt}_r\right), \tag{5.18}$$

$$\approx -\sum_{i,j,t,q,r} \omega^t_{qr}\psi^{it}_q\psi^{jt}_r, \tag{5.19}$$

$$= -\sum_{q,r,t} \omega^t_{qr} \left(\sum_i \psi^{it}_q\right)\left(\sum_j \psi^{jt}_r\right), \tag{5.20}$$

where in the approximation we have used both that the network is sparse, and so extending the sum to all i and j introduces few errors, and that $\log(1-x) \approx -x$ for x small[1].

A similar set of steps for the degree-corrected version instead provides the approximation

$$\sum_{i,j,t\notin\mathcal{E}} \log \tilde{Z}_{ijt} \approx -\sum_{q,r,t} \lambda^t_{qr} \left(\sum_i d^t_i \psi^{it}_q\right)\left(\sum_j d^t_j \psi^{jt}_r\right). \tag{5.21}$$

5.1.1 Expectation maximisation for parameter inference

Of course, in practice the true parameter values are not usually known. As such, much as in [113], and when performing MFVI in the previous chapter, we proceed by performing variational expectation maximisation (VEM) with BP. That is, we

1. Make an initial (optionally non-informative) guess about the parameter values, then with these fixed use Eqs. (5.12) and (5.13) to update our message system until convergence. In this work, this convergence is measured by the maximum change in any single message.

2. We then hold the variational estimates – *i.e.* the one- and two-point marginals – constant, and maximise the ELBO with respect to other parameters.

3. Repeat from step 1. until convergence, which may be evaluated either by change in parameters, or change in free energy.

[1]Note that in the static case, from the MLE for $\alpha_q = 1/N\sum_i \psi^i_q$, this non-edge term corresponds to $-N^2\sum_{qr}\omega_{qr}\alpha_q\alpha_r = -Nc$ for c the average expected degree in the network, which is used (up to a factor of 1/N that is often used to premultiply the free energy) without comment in various papers *e.g.* [183], but this no longer holds true in the dynamic version.

Following along similar lines to the static SBM [34], we find the parameter update equations

$$\alpha_q = \frac{1}{N}\sum_i \psi_q^{i1}, \tag{5.22}$$

$$\pi_{qq'} = \frac{\sum_{t=2,i}\psi_{qq'}^{it}}{\sum_{t=2,i}\psi_q^{i,t-1}}, \tag{5.23}$$

$$\text{(Bernoulli NDC case)} \quad \omega_{qr}^t = \frac{\sum_{i,j\in\mathcal{E}}\psi_{qr}^{ijt}}{\left(\sum_i \psi_q^{it}\right)\left(\sum_j \psi_r^{jt}\right)}, \tag{5.24}$$

$$\text{(Poisson DC case)} \quad \lambda_{qr}^t = \frac{\sum_{i,j\in\mathcal{E}}\psi_{qr}^{ijt}A_{ij}^t}{\left(\sum_i \psi_q^{it}d_i^t\right)\left(\sum_j \psi_r^{jt}d_j^t\right)}, \tag{5.25}$$

where the important changes to previous updates are the use of two-point marginals, as in Eqs. (5.14) and (5.15).

Note that while technically the denominator in Eqn. (5.23) is correct, this is simply obtained from the Lagrange multiplier enforcing that $\sum_{q'}\pi_{qq'} = 1\ \forall q$. As such, for numerical stability – and to save some unnecessary calculations – we can instead solely calculate the numerator then simply directly ensure that $\sum_{q'}\pi_{qq'}$ indeed equals one, by dividing the numerator by the unnormalised version of this sum. Should the two-point marginal have been correctly computed, $\sum_{q'}\psi_{qq'}^{it} = \psi_q^{i,t-1}$, so this is indeed equivalent. Further, as we allow new nodes to enter the system after the startpoint, also with probability α_q of belonging to group q, we fix $\alpha_q = \frac{1}{NT}\sum_{i,t}\psi_q^{it}$ rather than using the equation above, much as in [98].

Note these update equations are all quite logical — they follow from ensuring the MLE parameter estimates hold in expectation over the variational formulation, known as the Nishimori condition. That is, we have

$$\alpha_q = \frac{1}{N}\langle n_q^1\rangle, \tag{5.26}$$

$$\pi_{qq'} = \frac{\langle\tilde{m}_{qq'}\rangle}{\sum_{t=1}^{T-1}\langle n_q^t\rangle}, \tag{5.27}$$

$$\omega_{qr}^t = \frac{\langle m_{qr}^t\rangle}{\langle n_q^t\rangle\langle n_r^t\rangle}, \tag{5.28}$$

$$\lambda_{qr}^t = \frac{\langle m_{qr}^t\rangle}{\langle \kappa_q^t\rangle\langle \kappa_r^t\rangle}, \tag{5.29}$$

$$\tag{5.30}$$

for n_q^t the number of nodes in group q at time t, $\tilde{m}_{qq'}$ the number of transitions from q to q' – so the update for $\pi_{qq'}$ is the expected number of transitions from q to q' over the expected total number of nodes in q – m_{qr}^t the number of edges between blocks q and r at time t, and κ_q^t the total degree of block q at time t.

The full VEM procedure is then as follows:

1. Initialise labels somehow – for instance by the K-means clustering procedure used in the previous chapter – and either (i) use these to define the initial parameters using the equations above, or (ii) use minimally informative estimates — *e.g.* uniform for α and each row of π, average density of the network at t for p^t, total edge weight over total degree squared for λ^t *etc.*

2. Perform BP inference given these parameters:

   (a) Initialise messages – popular choices are random, uniform, or planted according to the initial partition found with some noise added.

   (b) For each message either

      (i) (in random order, sequentially / asynchronously), use the update equations to update the message, then update the marginal accordingly. Use this new ψ^j to update the external field h by subtracting the old ψ^j and adding the new value, or

      (ii) (synchronous updating) likewise use message equations to update messages and marginals, but delay updating the external field until all messages have been updated – this allows trivial parallelisation of updates, but convergence even for trees in a suitable parameter regime is no longer guaranteed [169]. For speed, this is used in our implementation.

   (c) Stop after convergence or maximum number of iterations, where convergence is assessed by the maximum absolute differences after updating all messages being less than a given threshold.

3. Update parameters given new marginals, using the equations above.

4. Repeat until convergence / maximum number of iterations, where convergence is defined either (i) in the same way as for messages, or (ii) by change in free energy.

5. Repeat the entire process for several runs, and select the run with the smallest final free energy, *i.e.* using this for model selection — as described in Chap. 4, this should correspond to the model with smallest KL divergence from the true posterior distribution of the node labels.

5.2 Greedy approximate inference

Finally, before displaying experimental results for BP, we propose one final option for inference. Primarily as an option to improve scalability of their model, various authors have proposed greedy methods to infer suitable blocks for their flavour of SBM. Here, as described in Chap. 2, we consider a method to be greedy if it proposes an objective (or quality) function to maximise for each local (*i.e.* node-wise) change of groups — for instance moving a specific node i from group q to group r at time t.

Often, the quality function chosen is the local change in log likelihood, given maximum-likelihood estimates for parameters [69]. Clustering is then performed by iterating over the nodes from a random initialisation of clusters, where at each iteration we calculate the value of this objective function for each possible move, then assign the new label of the node as the group with the maximum increase from the move. Intermediary state values as the algorithm proceeds are saved, and the optimal overall state is chosen – *i.e.* we can parallelise search for different node orderings, then take the best at each iteration. Such methods are deterministic given the initialisation and ordering that nodes are visited in, hence multiple runs from different initialisations and/or orderings should be performed, and the best overall final result selected.

Note that these local moves may change the overall number of groups, *i.e.* we might merge the last node of one group into another, or place a node in a new group entirely. As such, we should not solely look at log likelihood, but *e.g.* ICL. Equations are described below considering only the log-likelihood, as in our formulation of ICL as in Chap. 3, the only modification necessary would be a penalty (resp. reward) for placing a node in a new group (resp. merging the last node of a group into another) should such a move be considered.

For brevity, we first consider the static case of the DSBMM, before reintroducing group dynamics. Due to the formulation of the model, in this static, degree-corrected case the change in the log-likelihood due to the edge generation terms remains as demonstrated in the paper introducing the degree-corrected SBM, [69], and so for length considerations we do not reproduce this here. The new terms occur only due to the metadata distributions, $\mathrm{p}(X \mid Z)$. Further, we only here include the necessary

terms for Poisson, independent Bernoulli/categorical, and multinomial distributed metadata, but similar equations are possible for alternative distributions.

For Poisson metadata, we have that the maximum-likelihood estimator for the group parameters λ_q are given by the group means,

$$\lambda_q = \frac{1}{n_q} \sum_{i \in q} x_i, \tag{5.31}$$

with $n_q = |\{i \in q\}|$, and so – neglecting terms that do not change depending on the labelling – the corresponding term in the log-likelihood becomes

$$\mathcal{L}_{meta} = \sum_q X_q (\log X_q - \log n_q - 1), \tag{5.32}$$

where we define $X_q = \sum_{i \in q} x_i$. If we now consider the change in this term upon moving a node i from group q to group r, with $b(x) = x \log x$ as in [69], we have

$$\begin{aligned} \Delta\mathcal{L}_{meta} = {} & b(X_r + x_i) + b(X_q - x_i) - b(X_r) - b(X_q) \\ & + X_r \log \frac{n_r}{n_r + 1} + X_q \log \frac{n_q}{n_q - 1} + x_i \log \frac{n_q - 1}{n_r + 1}, \end{aligned} \tag{5.33}$$

so only $\mathcal{O}(1)$ terms to calculate per node.

If we instead have independent Bernoulli or categorical metadata, we have that the maximum-likelihood estimators for the group probabilities of observing metadata of type ℓ, $p_{q\ell} = X_{q\ell}/n_q$, where much as before we have defined $X_{q\ell} = \sum_{i \in q} x_{i\ell}$. For independent Bernoulli metadata, we now have

$$\mathcal{L}_{meta} = \sum_{q,\ell} n_q \log \left(1 - \frac{X_{q\ell}}{n_q}\right) + X_{q\ell} \log \frac{X_{q\ell}}{n_q - X_{q\ell}}, \tag{5.34}$$

providing

$$\begin{aligned} \Delta\mathcal{L}_{meta} = {} & \sum_\ell (n_r + 1) \log \left(1 - \frac{X_{r\ell} + x_{i\ell}}{n_r + 1}\right) + (X_{r\ell} + x_{i\ell}) \log \left(\frac{X_{r\ell} + x_{i\ell}}{n_r + 1 - X_{r\ell} - x_{i\ell}}\right) \\ & + (n_q - 1) \log \left(1 - \frac{X_{q\ell} - x_{i\ell}}{n_q - 1}\right) + (X_{q\ell} - x_{i\ell}) \log \left(\frac{X_{q\ell} - x_{i\ell}}{n_q - 1 - X_{q\ell} + x_{i\ell}}\right) \\ & - n_r \log \left(1 - \frac{X_{r\ell}}{n_r}\right) - X_{r\ell} \log \left(\frac{X_{r\ell}}{n_r - X_{r\ell}}\right) \\ & - n_q \log \left(1 - \frac{X_{q\ell}}{n_q}\right) - X_{q\ell} \log \left(\frac{X_{q\ell}}{n_q - X_{q\ell}}\right), \end{aligned} \tag{5.35}$$

and similar for categorical metadata provides

$$\begin{aligned}\Delta\mathcal{L}_{meta} = \sum_{\ell} b(X_{q\ell} - x_{i\ell}) - b(X_{q\ell}) + b(X_{r\ell} + x_{i\ell}) - b(X_{r\ell}) \\ - b(n_q - 1) + b(n_q) - b(n_r + 1) + b(n_r).\end{aligned} \tag{5.36}$$

Finally for multinomial distributed metadata, introducing $X_q = \sum_\ell X_{q\ell}$ such that the MLE for $p_{q\ell} = X_{q\ell}/X_q$, we find

$$\begin{aligned}\Delta\mathcal{L}_{meta} = \sum_{\ell} b(X_{q\ell} - x_{i\ell}) - b(X_{q\ell}) + b(X_{r\ell} + x_{i\ell}) - b(X_{r\ell}) \\ - b\Big(X_q - \sum_{\ell} x_{i\ell}\Big) + b(X_q) - b\Big(X_r + \sum_{\ell} x_{i\ell}\Big) + b(X_r).\end{aligned} \tag{5.37}$$

Note for each of these three last distributions we now have $\mathcal{O}(L)$ terms, so if there are a large number of possible categories this could in fact be quite expensive. However, in practice the distribution over categories for most groups (and individual nodes) is highly sparse, *i.e.* $X_{q\ell} = 0$ for most ℓ, in which case the number of terms we must compute for the leading order contribution reduces considerably.

Beyond initialisation of groups, one reason such a static greedy method might appeal is for online inference for the full dynamic model. That is, we can fit suitable parameters for newly observed timesteps sequentially, given previous parameters. In this case, we assume the past groups are fixed, and may then seek the labelling at the current step, Z^t, that maximises the posterior probability

$$\mathrm{p}(Z^t \mid A^t, X^t, Z^{t-1}) \propto \mathrm{p}(A^t \mid Z^t)\mathrm{p}(X^t \mid Z^t)\mathrm{p}(Z^t \mid Z^{t-1}). \tag{5.38}$$

In effect, this replicates the static situation bar the small modification of $\mathrm{p}(Z^t \mid Z^{t-1})$. The MLE for this term is simply $\pi_{qq'} = n_{qq'}/N_q$, where $n_{qq'}$ is defined to be the observed number of transitions from group q to group q', and N_q is the total number of times that nodes have been assigned to group q. As such, if we consider the impact of moving a node i from group q at time t (the new final time in this online scenario)

to group r, all previous changes will be made, along with

$$\begin{aligned}\Delta\mathcal{L}_\pi &= \Delta\sum_{q',q} n_{q'q}(\log n_{q'q} - \log N_{q'}),\\ &= \sum_{q'} b(n_{q'q}-1) + b(n_{q'r}+1) - b(n_{q'q}) - b(n_{q'r})\\ &\quad - \sum_{q'\notin\{q,r\}} (n_{q'q}-1)\log N_{q'} + (n_{q'r}+1)\log N_{q'} - n_{q'q}\log N_{q'} - n_{q'r}\log N_{q'}\\ &\quad - (n_{qq}-1)\log(N_q-1) - (n_{rr}+1)\log(N_r+1) + n_{qq}\log N_q + n_{rr}\log N_r\\ &= \sum_{q'} b(n_{q'q}-1) + b(n_{q'r}+1) - b(n_{q'q}) - b(n_{q'r})\\ &\quad - (n_{qq}-1)\log(N_q-1) + n_{qq}\log N_q - (n_{rr}+1)\log(N_r+1) + n_{rr}\log N_r.\end{aligned} \tag{5.39}$$

We note that recently [152] proposed a similar greedy method for a dynamic SBM for binary networks – though proceeded by placing conjugate priors over non-group parameters to allow marginalising these – and demonstrated good performance, so this may well be a viable means of approximate inference for the full model. The only further modification necessary is considering how the log likelihood term for π changes at other timesteps, but this is fairly immediate – if $t < T$, there is now additionally the backwards term

$$\begin{aligned}&\sum_{q'} b(n_{qq'}-1) + b(n_{rq'}+1) - b(n_{qq'}) - b(n_{rq'})\\ &\quad - \sum_{q'}(n_{qq'}-1)\log(N_q-1) - n_{qq'}\log N_q + (n_{rq'}+1)\log(N_r+1) - n_{rq'}\log N_r,\end{aligned} \tag{5.40}$$

while if $t = 1$ the forward terms included previously are not present.

Importantly, this greedy method – specifically that accounting for metadata contributions – is easily combined with others in the literature. In App. A, we suggest combining with a recent 'hypergraph modularity' method, to extend the DSBMM to hypergraphs with metadata.

5.3 Experiments on simulated networks

With scalable inference now established, in this section we perform a series of experiments on simulated networks, investigating our new BP inference procedure for the model in greater detail.[2] Specifically we explore

[2]Code for BP inference in the DSBMM is available on Github

(i) How does the model performance vary with respect to the tuning parameter, for simulations where we explicitly control the degree of alignment between metadata and network partitions? In Sec. 5.3.1, we hope to see that as alignment increases, so too does the best-performing tuning parameter. If it is the case that optimal tuning parameter values have a strong relation to the degree of alignment, then in empirical networks, the optimal value – as determined through *e.g.* link prediction performance in the absence of ground truth labels – provides better information about the utility of metadata.

(ii) How does the utility of metadata affect detectability of groups in a simple toy model? Given the results of the previous chapter, in Sec. 5.3.2 we explore a region of parameter space found in the network-only case to be difficult for recovery of groups [50]. As the SNR measure found only required the metadata to be minimally informative, we expect to see strong recovery even far from the prior threshold.

(iii) Have we sufficiently improved scaling of the model? Finally, in Sec. 5.3.3 we demonstrate the near-linear scaling now obtained. In Sec. 5.4, we use this new capability to explore empirical networks previously infeasible due to their size.

5.3.1 Exploring misalignment and tuning parameter in detail

As discussed in previous chapters, the tuning parameter of our model is particularly important in the case that the metadata does cluster according to some group structure, but that group structure does not align well with the group structure of the network edges. To investigate this relationship, we may explicitly control the degree of this alignment.

Specifically, starting from the partition used to generate network edges, prior to simulating metadata we produce a new metadata partition with desired alignment according to a specified measure – in the following experiments, we use NMI. To do so, we sequentially introduce small amounts of variation to the network partition by randomly sampling new labels for a subset of nodes, then calculate the alignment between the new partition and the original, and repeat this process until the measure approximates the chosen value.

We simulate 10 tests for $N = 100$ nodes over $T = 5$ timesteps, with $c_{\text{in}} = 30$ and $c_{\text{out}} = 10$, $p_{\text{stay}} = 0.6$, and mildly informative Poisson and independent Bernoulli metadata, where metadata and edges are drawn according to *different* partitions into

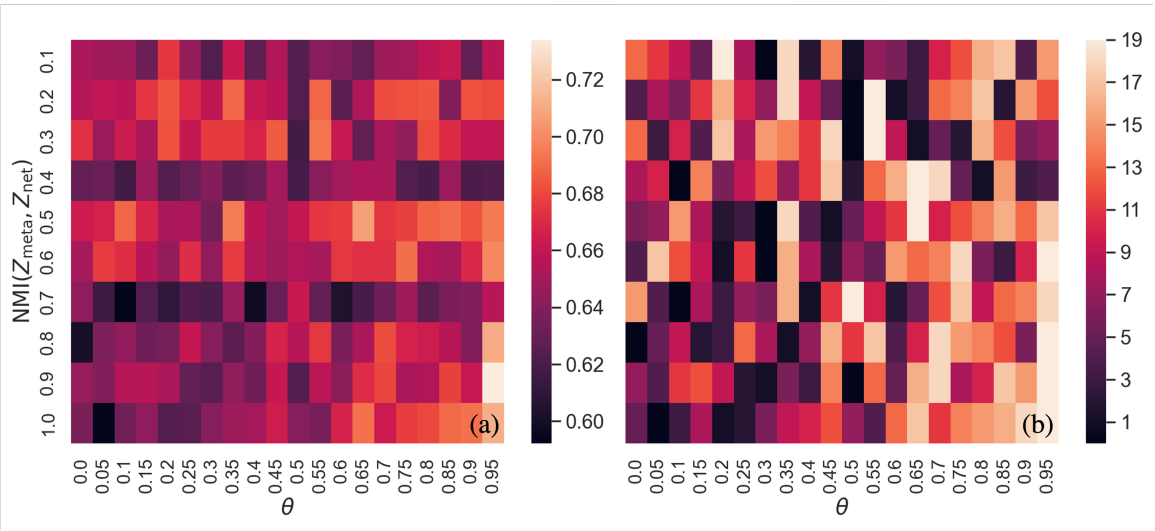


FIGURE 5.1: In subfigure (a), we display ARIs using the BP VEM procedure, for the alignment tests described in text, averaged over 20 simulations. In subfigure (b) we display the corresponding ranks of each tuning parameter within each test.

$Q = 4$ groups. The NMI between these network and metadata partitions vary over the tests, between 0.1 (almost no alignment) and 1.0 (perfect alignment).

For each test, we simulate 20 samples from the model, and then run our full inference procedure on each sample, varying the tuning parameter from 0.0 to 0.95 — the average ARI of the partition inferred using BP VEM to the true network partition for each tuning parameter, test combination are displayed in Fig. 5.1.

We observe that indeed there appears to be some weak dependence on the tuning parameter, where in the case that metadata and network partitions are very poorly aligned, it is better to minimally include metadata, if at all. However, for increased alignment, the inclusion of metadata at suitable tuning parameter weights almost always improves the quality of partition inferred, even for surprisingly low levels of alignment between the two — exactly as we would desire.

However, these results may be contrasted to those obtained using the MFVI procedure of Chap. 3, presented for a coarser range of tuning parameter values in Fig. 5.2.

In this figure, we observe considerably higher average performance than the BP case, and a much clearer relation between the tuning parameter, partition alignment, and corresponding ARI — though note, that the same issue as discussed in Chap. 3 is present, in that this evaluates the ability of the method to recover *only* the network partition, rather than the metadata, or indeed joint partition. In particular, it appears that for the MFVI method, over-weighting the metadata appears to almost always

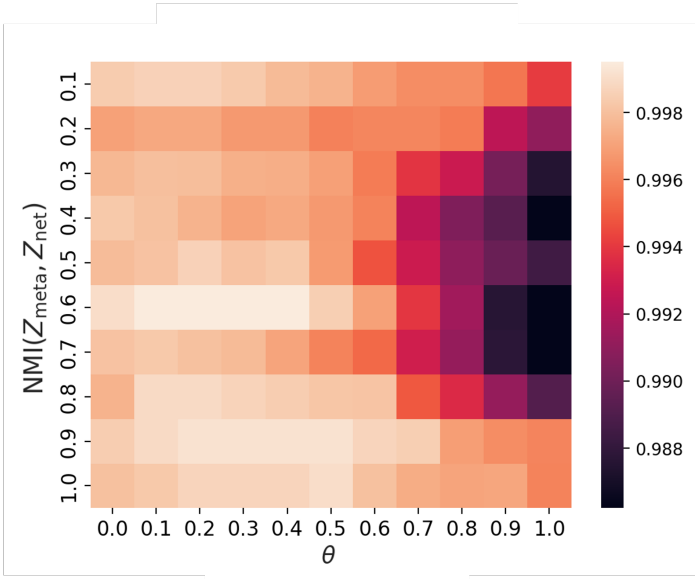


FIGURE 5.2: In this figure, we display ARIs for the alignment tests described in the text, averaged over 20 simulations, where we now use the MFVI procedure of Chap. 3 to perform inference.

result in a deterioration in performance. However, likewise the overall best performance incorporates information from the metadata to some degree, when it is indeed useful.

The comparison of these results highlights some of the strengths of mean-field variational inference over BP — it is guaranteed to converge (even if not necessarily to the optimal partition), and appears to be less likely to get trapped in local, poor performing optima, at least under certain conditions, such as misalignment. Specifically, we found that the *modal* performance of the BP procedure across more challenging experiments typically bested that of MFVI, in addition to completing significantly faster. However, there were some runs which entirely failed to converge, or (more rarely than convergence failure, though more frequently than for MFVI) instead converged to poor local optima.

Importantly, the results of these experiments only demonstrate the relation between (a) the network and metadata partition alignment, and (b) the tuning parameter for a fixed, relatively small amount of metadata at each node. As we discuss in greater depth in Sec. 5.4 below, in practice it is often better to treat θ as a parameter controlling the relative weighting of network and metadata terms in the likelihood, rather than a switching variable for the inclusion of metadata or not — *i.e.* to extend the range from $[0, 1]$ to $[0, \infty)$. Values greater than one may be useful when small quantities of highly informative metadata are available, but in practice the reverse is often true. In particular, a possible issue emerges when large quantities of metadata are available,

where the metadata contribution at each node may dominate the network contribution to such a degree that network information is effectively ignored — even for relatively small θ values. Thus, in Sec. 5.4 we suggest one heuristic to automatically choose the tuning parameter, rather than explore wide ranges as hitherto performed.

5.3.2 Empirical detectability of a toy model

As in previous studies of detectability in the literature, to investigate how the inclusion of metadata affects the ability to recover the true groups defining the network, we simplify the DSBMM as much as possible.

To do so, we consider an analogous dynamic extension of the toy model of the previous chapter, where we restrict the transition matrix controlling the evolution of groups, π, to be governed by a single parameter, η. This is defined as the probability that a node remains in the same group between one timestep and the next, else any group (including the previous group) is chosen uniformly at random. This follows the setting of [50], where the authors determine weak detectability thresholds for the dynamic SBM in the absence of metadata.

This reduces our model to

$$\begin{aligned}\mathrm{p}(A,X,Z) = \frac{1}{Q^N}\prod_{i,t=2}\left(\eta+\frac{1-\eta}{Q}\right)^{\delta_{z_i^{t-1},z_i^t}}\left(\frac{1-\eta}{Q}\right)^{1-\delta_{z_i^{t-1},z_i^t}} \\ \times\prod_{i,j,t} p_{\text{in}}^{A_{ij}\delta_{z_i^t,z_j^t}} p_{\text{out}}^{A_{ij}(1-\delta_{z_i^t,z_j^t})}(1-p_{\text{in}})^{(1-A_{ij})\delta_{z_i^t,z_j^t}}(1-p_{\text{out}})^{(1-A_{ij})(1-\delta_{z_i^t,z_j^t})} \\ \times\prod_{i,t}\rho^{\delta_{z_i^t,x_i^t}}\left(\frac{1-\rho}{Q-1}\right)^{1-\delta_{z_i^t,x_i^t}}.\end{aligned} \tag{5.41}$$

To produce the tests in this section, we initially consider two regions of ϵ, η space previously demonstrated to either be close, or actually cross the detectability threshold (*cf.* Fig. 3 in [50]). We draw 20 samples of network time-series for $N = 512$ nodes over $T = 40$ timesteps, for $Q = 2$ groups, while keeping the average degree $c = 16$ fixed, where we then either

(i) stay within, but close to the previous detectability limit by varying ϵ from 0.3 (well-defined groups) to 0.6 (more poorly-defined groups), and η from 0.4 (low group stability) to 0.6 (improved stability), or

(ii) cross the prior detectability threshold, by varying both ϵ and η from 0.6 to 0.8.

The model is then supplied with the true DSBMM parameters other than the partition, which we freeze to these values before running the full inference procedure for the node messages and marginals. As discussed in the previous chapter, in statistical physics this is sometimes described as assuming we are on the Nishimori line [34]. Upon convergence of the algorithm, as previously we assign each node to the group with greatest marginal density.

For comparable results to those in the literature, we evaluate these estimates using the maximum overlap between the true and predicted labels, over all $Q!$ permutations of the labels, normalised to take a value of zero if it is no better than random, and one if there is perfect agreement — as introduced in Sec. 2.2.2, Eqn. (2.18).

The phase transition in the case with no metadata, beyond which groups could not be recovered better than chance, was found to occur when [50]

$$c\left(\frac{1-\epsilon}{1+(Q-1)\epsilon}\right)^2 < \frac{1-\eta^2}{1+\eta^2}, \tag{5.42}$$

i.e. in our particular situation when

$$\left(\frac{4(1-\epsilon)}{1+\epsilon}\right)^2 < \frac{1-\eta^2}{1+\eta^2}. \tag{5.43}$$

Our second area of tests thus crosses this threshold (shown in dashed grey in subsequent plots where relevant), beyond which previous methods could not perform better than random – maximum normalised overlap values found for over half of the tests considered in this area were close to zero, while further from the transition boundary they increased up to a maximum of around 0.8. On the other hand, the first area is close to the limit, but previous methods still typically managed to recover the true partition well, with values ranging from around 0.2 to close to 1.0.

First, in Fig. 5.3 we display results where we initialise messages as uniform — the previous factorised fixed point. We show plots for two choices each of η and ϵ, fixing these while allowing ρ and the other parameter not fixed to vary — subfigures (a) and (b) show ranges within the detectable area, while (c) and (d) show ranges that cross the detectability threshold. We observe that with uniform message initialisation, ϵ and η in the ranges considered have no effect on the ability to recover the true partition, including beyond the previous detectability limit. Instead, only the level of information carried by the metadata, as determined by ρ is important, and – as for all results in this section – there is a clear improvement of the recovery of results as ρ increases. As we demonstrated in Sec. 4.4, in this particular case of uniform message initialisation,

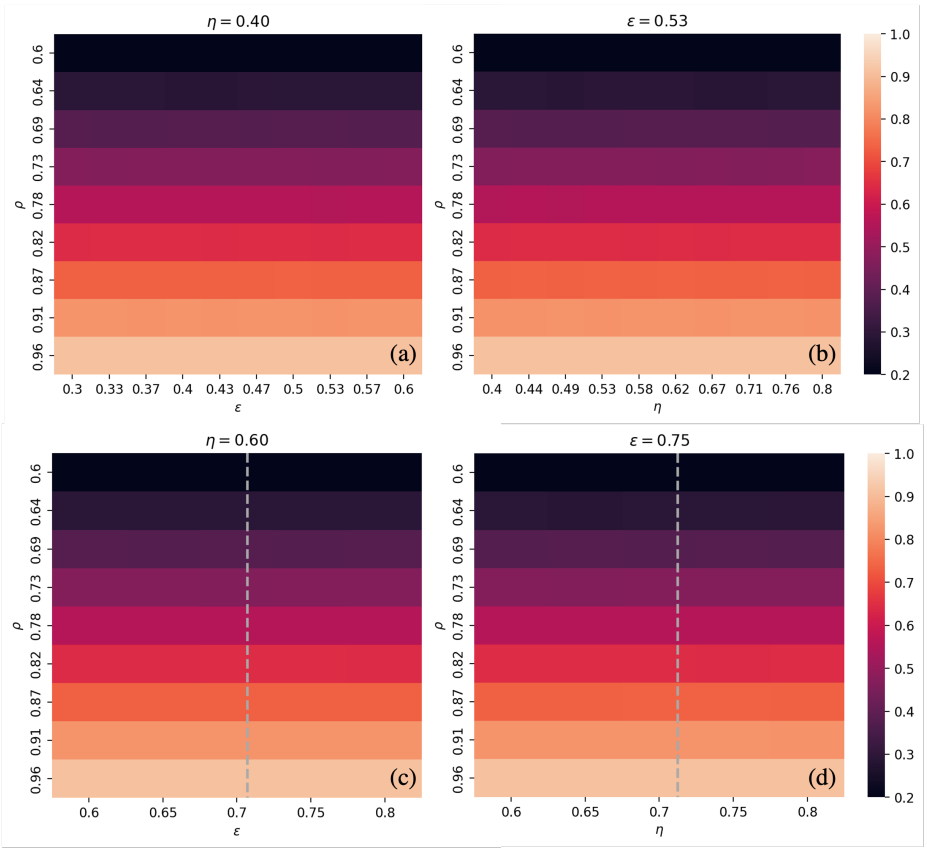


FIGURE 5.3: Normalised overlap for empirical detectability tests discussed in the text, where messages were initialised as uniform. Subfigures (a) and (b) show results within the detectable area tested, where we fix η and ϵ respectively then allow ρ and the other to vary. Subfigures (c) and (d) show analogous results in the area of (ϵ, η) space crossing the previous detectability threshold – marked in dashed grey.

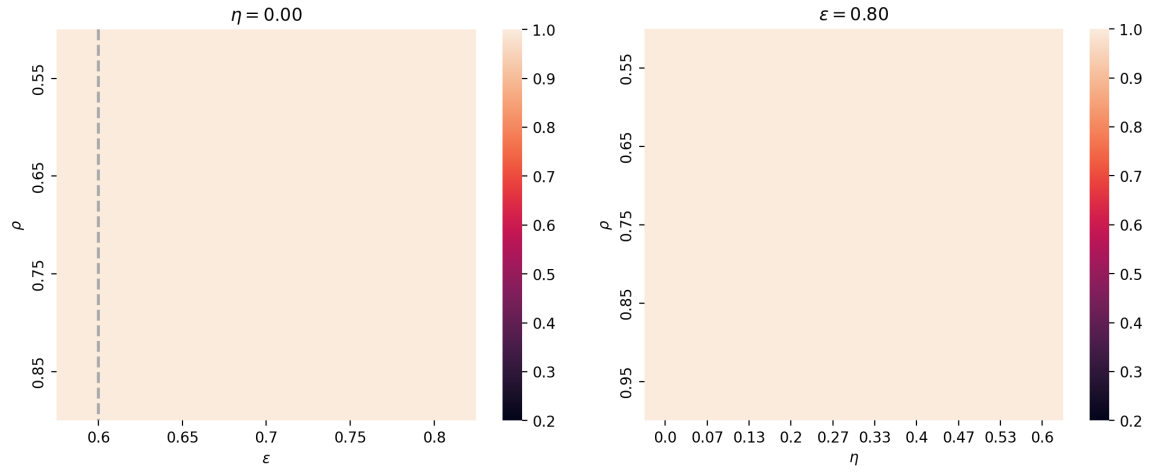
sequential updates approximately correspond to optimal Bayesian estimates given the metadata. Hence such behaviour emphasises how factors beyond the tuning parameter play an important role in how metadata is incorporated in the model.

When we instead initialised using random messages, we found that for some experiments, we observed counter-intuitive behaviour: for larger η values, the true partition was better recovered with *increasing* ϵ, *i.e.* decreased coherence of groups within the network. This occurs as for the toy model described, when we sample X, we fix 'true' group labels to a particular permutation — that is, if $x_i^t = 1$, then with probability ρ, $z_i^t = 1$ also. However, network edges express no such permutation bias, *i.e.* they remain permutation invariant, thus if the initialisation of messages places greater weight within a group, say at $q = 2$ rather than the 'true' $q = 1$, the network parameters may express some preference to increase this weight — the network and metadata are effectively providing conflicting information. With ϵ small, this bias is stronger than with ϵ large, when the network has a much lesser impact on the messages. If η is small, initial preference for one label permutation over another at several timesteps is less easily propagated to the rest of the system, thus information within each timestep dominates, eventually preferring the metadata's 'true' labelling, and henceforth more coherent groups permit better recovery as in the case without metadata. On the other hand, if η is large, this provides further reinforcement of initial label permutation problems across timesteps, hence we expect to observe an increasing importance of ϵ, as we do in practice. However, this problem is artificial — when parameters are not frozen, this problem would not emerge, thus this should not be an issue when considering real networks.

To successfully resolve this issue, for the remaining results in this section, we introduced a bias towards the 'true' (*i.e.* metadata) permutation, by initialising messages according to the metadata of the sending node plus some random noise. We then found that for every range of $0.5 < \rho < 1$, $0 < \epsilon < 1$, $0 < \eta < 1$ considered, the method achieved near perfect recovery — as predicted by the SNR measure of the previous chapter, given the degree threshold is met. We display two such plots in Fig. 5.4, where for each the errors in recovery are indistinguishable.

5.3.3 Scaling

Briefly, we conclude this section on synthetic networks by demonstrating that BP does indeed provide near-linear scaling with respect to the number of nodes for sparse networks, a significant improvement to the previous quadratic scaling.



(a) Fixed $\eta = 0.0$, *i.e.* independent SBMs at each timestep

(b) Fixed $\epsilon = 0.8$, *i.e.* well beyond the previous static limit

FIGURE 5.4: Normalised overlap for empirical detectability tests for two previously completely undetectable regions, as discussed in the text, where messages were initialised according to the metadata of the sending node plus some random noise.

To do so, we simulate networks of varying size, from $N = \lfloor e^5 \rfloor$ to $\lfloor e^8 \rfloor$, with $Q = \lfloor \log(N) \rfloor$ for comparability to real networks. For each N, we sample 20 networks from the model for $T = 5$ timesteps, with fixed average in- and out-degrees $c_{\text{in}} = 20$ and $c_{\text{out}} = 5$ respectively, with probability of staying in the same group at each timestep of $p_{\text{stay}} = 0.8$, else another (different) group is chosen at random. We supply both Poisson and four-dimensional independent Bernoulli metadata.

We display the average run-time of the full inference procedure in Fig. 5.5. Our earlier expectation of near-linear scaling for sparse networks has successfully been met, and this allows feasible application of the algorithm to networks orders of magnitude larger than before. We use this capability in the following section to explore networks with $\mathcal{O}(5 \times 10^4)$ nodes, in a similar amount of compute time to previous MFVI results on only $\mathcal{O}(10^2)$ nodes, thus greatly expanding the utility of the method.

5.4 Application to a medium-scale Latin American co-authorship network

5.4.1 Hierarchical block detection

Before continuing to apply our model to real dynamic network data, we note that as demonstrated in Sec. 5.3.3, the scaling of the BP inference procedure with respect

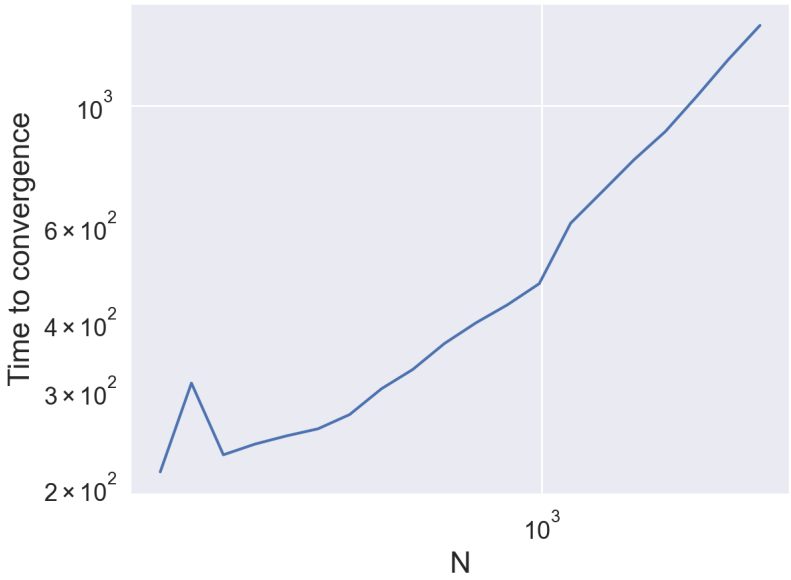


FIGURE 5.5: Time to convergence for the BP algorithm for networks of increasing size. A near-linear scaling is observed, as predicted.

to the number of nodes in the network considered has greatly improved. However, the dependence of the computational complexity on Q^2 means that seeking a large number of groups can still slow completion of the algorithm dramatically. As such, analogously to [200], we may instead apply our method recursively to find a hierarchy of blocks.

Precisely, we may first infer a small number of groups, Q_1 say, then partition the dynamic system accordingly, *e.g.* by assigning each node to the group to which it belongs most frequently,

$$z_i^{\ell=1} = \text{mode}_t(\{z_i^{t,\ell=1}\}). \tag{5.44}$$

Each resulting time-homogeneous group, q^1, that contains more than a specified minimum number of nodes may then be considered as defining a dynamic subgraph, and the method applied once more to find Q_2 groups, $Z^{t,2}_{i\in q^1}$, at the next level. This process may be repeated until (i) no significant groups are found at the current level, (ii) all groups are under the specified minimum size, (iii) a specified maximum number of total groups are found, or (iv) a specified maximum number of levels. As the number of groups at each level is relatively small, and the number of nodes considered when applying the VEM procedure decreases – often very rapidly, like $\mathcal{O}(N\prod_{k<\ell} Q_k^{-1})$ – this can provide significant speedups.

Indeed, if there are $\prod_{k<\ell} Q_k$ groups considered as separate subgraphs at level ℓ, each of which has size approximately $N/\prod_{k<\ell} Q_k$ nodes, the hierarchical method scales like

$\mathcal{O}(NT\sum_\ell Q_\ell^2)$. If we are seeking a fixed number of groups in total, $Q \approx \prod_\ell Q_\ell$ over L total levels say, if $Q_\ell = h_Q$ is constant, then we have $\sum_\ell Q_\ell^2 = Lh_Q^2 \approx \mathcal{O}(Q^{2/L})$. Hence, *ceteris paribus*, if deeper hierarchies are considered, the procedure is considerably accelerated.

In this work, we use the modal option of Eqn. (5.44) to partition the dynamic system at each level of the hierarchy, but we note that this is likely not ideal, especially in certain circumstances. In particular, it means that if nodes frequently change groups between time periods – one could imagine for instance a network of social contacts between people during and outside of working hours, with two clusters for each person that they switch between – then enforcing time-homogeneity to allow the split discards important information for most nodes at many timesteps. As such, in these dynamic networks, applying the hierarchical method above would result in a poor quality partition after the first level or so.

It may nonetheless be one option for initialising a temporal partition, which could subsequently be refined by using the full, slower method, on the assumption that fewer iterations overall would be necessary. However, alternative, cheaper methods to provide such initialisations prior to full BP updates may be preferable in this case, such as the greedy method proposed, or simply the application of static methods to each time period, followed by post hoc alignment between timesteps. In our application to academic networks, the potential problems with the hierarchical method are not as significant a concern, as researchers typically focus their efforts in a particular area, so the groups inferred do not change as rapidly until deeper in the hierarchy — *e.g.* when current institution / research group or similar become more important rather than overall research area and region.

5.4.2 Automatically choosing the tuning parameter

As mentioned in Sec. 5.3.1, a further concern arises if we are considering networks with large quantities of metadata, in that the metadata likelihood term may dominate the network contribution at each iteration of the message equations — thus effectively neglecting network structure when inferring groups. As such, if we treat the tuning parameter as solely a weighting term controlling the relative contribution of each, a simple heuristic for the tuning parameter is as follows: initialise the messages and model parameters according to whichever scheme chosen, then calculate the metadata and spatial message terms for each node. As the metadata term is taken to the power of the tuning parameter, if these two contributions are desired to be of similar order

on average, then we may take the tuning parameter to be the average ratio of the log of spatial message terms to the log of the metadata term, *i.e.*

$$\theta \sim \frac{1}{NTQ} \sum_{i,t,q} \frac{\sum_{k \in \partial i} \log \sum_r \omega_{rq} \psi_r^{k \to i}}{\log \mathrm{p}_q^t(x_i^t)}. \tag{5.45}$$

If desired, this process can be performed iteratively, throughout the course of the BP VEM procedure, or thresholds may be set to constrain the permissible range and/or fix the parameter after a given number of iterations.

5.4.3 Results

With these two further modifications, we now explore the application of the DSBMM to a medium-scale real dynamic network. Specifically, we consider a dynamic network of Latin American authors that publish in journals related to Math, Physics, Computer Science, or Economics (as defined by their Scopus© subject area) between 2009 and 2020 (inclusive), where a weighted connection between two authors corresponds to a count of the number of times they have published together in the period under consideration. We split the full period into four 3-year windows, thus have $T = 4$ in our model.

In our full dataset, there are over 50,000 such authors, but we further restrict those considered to focus on the most important, well-connected authors. First, we take the 5-core network at each timestep, that is the subset of authors with at least five publications in the period, co-authored with other authors within our dataset. We then take the largest connected component at each timestep, rather than permit disjoint subgraphs. In doing so, we are left with 4,959 distinct authors — still more than 10× larger than the largest network considered in Chap. 3.

We also have a considerable quantity of metadata available for each author. Many of these have been studied previously in the literature (see *e.g.* [54, 173, 179]), hence we select the following, that have often been found to be important to collaboration:

- The total number of citations an author has received (at any time) to papers published in the period[3];

[3]Note that this provides a simple measure of the average impact/perceived quality of the output of an author in this particular period, rather than their popularity/renown as might be measured by the total number of citations received in the period to papers published at any prior time. As such, more recent periods will naturally have fewer such citations for all authors. This means it differs to other metadata selected, as all others are linked to the paper/author at the time of publication, while this dimension accrues over time – if we updated the dataset collected, particularly for more recent periods this subset of metadata may change considerably. Nonetheless, we don't expect this to

- The median duration since the authors first publication (their 'career age');
- Counts of (i) Scopus© All Science Journal Classification (ASJC) codes, of which there are 138 different codes present in the network selected, and (ii) journal subject areas of their publications, of which there are 23 different areas present;
- Counts of their institutional/organisational affiliations, of which there are 442 unique institutions or organisations present;
- Weighted versions of (i) ASJC code and (ii) subject area counts – produced by downweighting the contribution of each publication in a period by the number of authors in total on the paper;
- A weighted version of affiliation counts, by downweighting by the total number of affiliations the author themselves has on a publication.

To include these in our model, we place Poisson distributions over the citation count and career age, multinomial distributions over the count vectors, and thresholded independent Bernoulli distributions over the weighted versions, where we consider only the top 10 categories by value to be present – or fewer if the author publishes in fewer than 10 categories overall.

We then use the hierarchical inference procedure described in Sec. 5.4.1, specifying four groups at each level, and three levels in total, for a maximum of 64 possible groups at the lowest level. We use the heuristic described in Sec. 5.4.2 to select the tuning parameter at each application of the inference procedure, thus permitting the relative influence of metadata to vary across the network, depending on how informative it appears to be upon initialising the system.

In Fig. 5.6 we display the hierarchical partition inferred at each timestep, produced using the drawing functionality of the `graph-tool` python package [135]. Each subfigure shows authors placed on the edge of the circle, and coloured according to their community. Edges between authors then approximately follow the hierarchy, hence edges that do not pass through the centre of the circle correspond to edges between blocks that merge at upper levels. We note that this figure does not quite capture the dynamic hierarchical partition actually inferred, as recursive application provides a dynamic partition at each level, while here we display the hierarchy according to the modal block of each higher level – as described above, that which was used to split

dramatically impact the utility of the metadata, as by allowing time-varying metadata distributions, it only matters that there is some distinction in metadata distributions between groups *within* each timestep. We leave investigation of the latter option of counting citations in a period to future work.

the network prior to repeat application – but the difference is not enormous, thus the figures are still useful for interpreting the dynamic partition inferred.

Note that the authors present at each timestep vary, thus there is not an exact correspondence between nodes at a given location between each time. Nonetheless, there is a remarkable consistency in the overall structure over the full period considered. In particular, immediately at the first level of the hierarchy, one group may largely be separated from the rest of the network, while the remainder has a far greater quantity of interconnections.

To investigate this further, in Fig. 5.7, we focus in on the hierarchical partition inferred for the 2018-2020 period, but now label each block with the countries to which a plurality of the authors contained are affiliated. Where multiple countries are listed, they are in descending order of commonality. We may now observe that, while Brazilian authors are present throughout the network – likely as a result of enforcing that the network we consider is connected – the group of authors that separate most strongly are also those most embedded in Brazil specifically, rather than tightly collaborating with academics, or visiting institutions of other nationalities. This may be expected, due to the linguistic difference of Brazil (Portuguese) to the remainder of Latin America.

Finally, we wish to investigate the degree to which each type of metadata included has influenced the partition inferred. To do so, here we choose to compare our partition to another inferred solely using the metadata selected, neglecting all other information – most importantly, the network structure itself. We use the K-Means clustering algorithm [89], and specify the same number of groups as we infer at each level in the hierarchy.

In Fig. 5.8 we display the resulting normalised reduced mutual information of [112] between these two partitions, by each choice of metadata – including both the log of the citations received (given the orders of magnitude over which this varies), and of course all metadata together. We observe that up to the specified depth, and corresponding number of groups, different types of metadata are important at different scales – levels in the hierarchy.

Career age in particular, and the raw counts of institutional affiliations for each author show the lowest correspondence with meso-scale structure in the network. This is not too surprising, as career age only becomes more important at a finer resolution than that considered – *e.g.* within research groups, with lead researchers *etc.* – while the presence of a large number of unique institutions in the network considered means

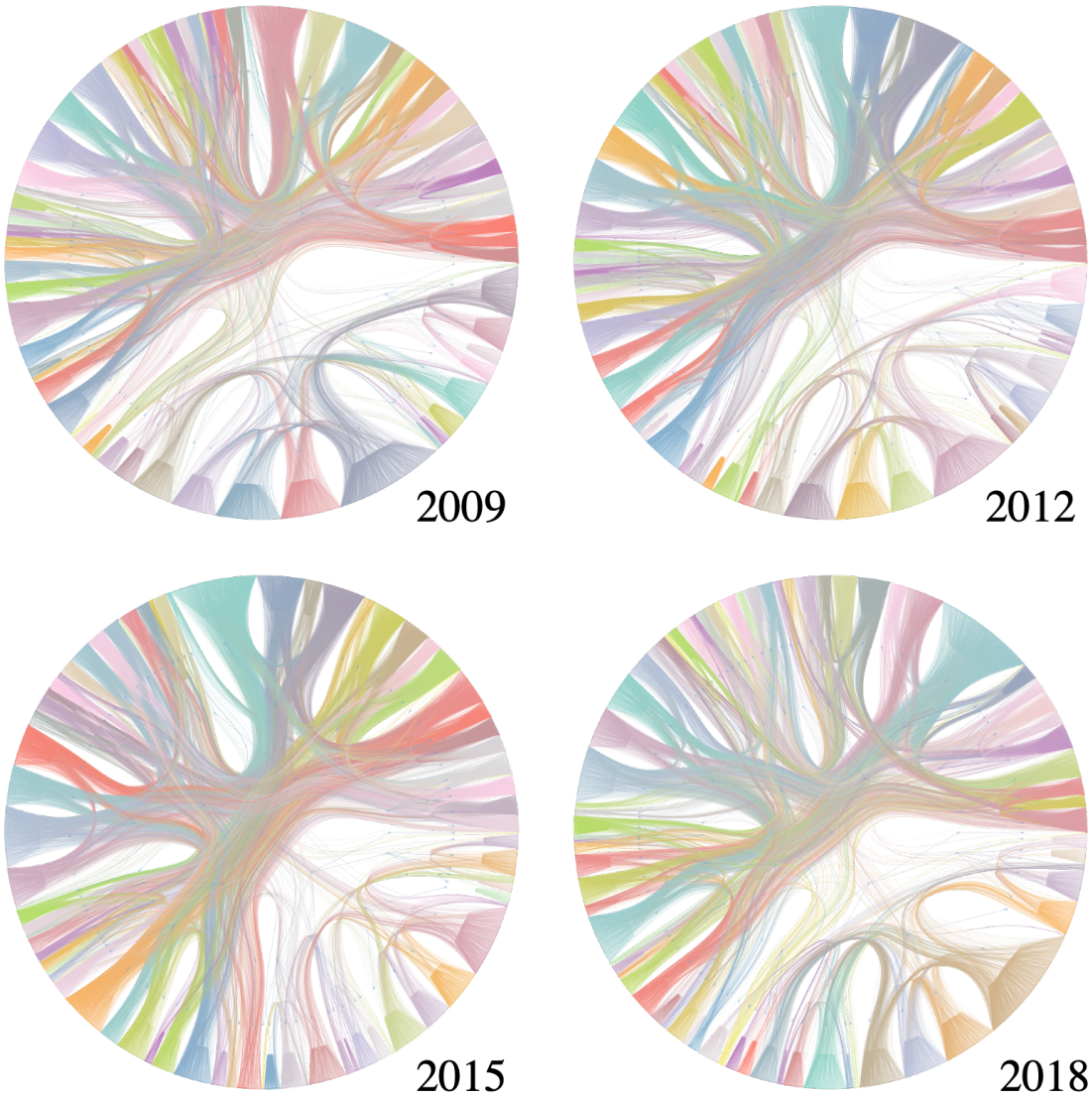


FIGURE 5.6: Hierarchical partition inferred for each period. In each figure, authors are nodes on the edge of the circle, coloured according to their block, then edges between nodes approximately follow the hierarchy inferred.

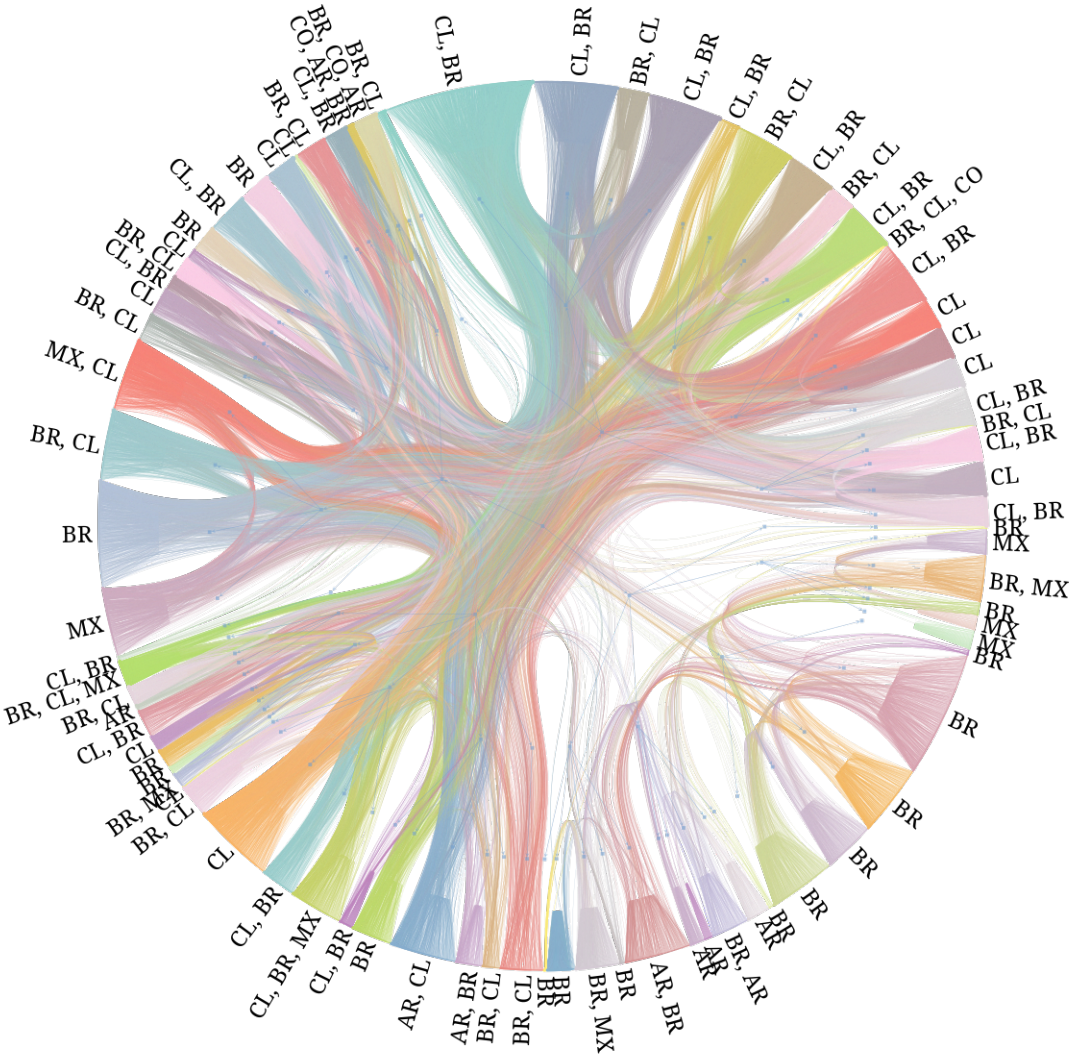


FIGURE 5.7: Hierarchical partition inferred for 2018-2020 period, with primary countries to which authors in each block are affiliated to labelled accordingly. In this figure, authors are nodes on the edge of the circle, coloured according to their block, then edges between nodes approximately follow the hierarchy inferred.

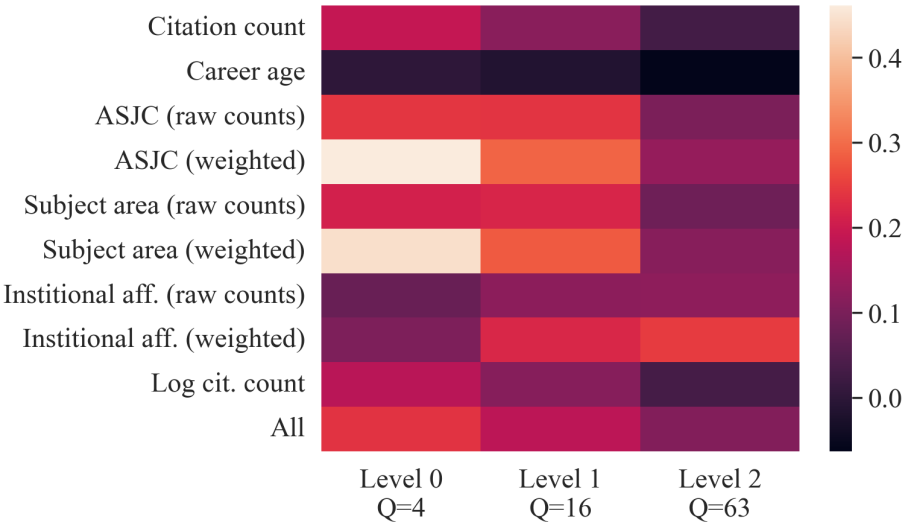


FIGURE 5.8: Normalised RMI between the partition inferred by K-Means on the specified metadata, with the same number of groups, to the partition inferred by hierarchical application of our model, by level in the hierarchy – level 0 corresponds to the initial application of the model, *i.e.* the highest level. We also specify the number of groups inferred, varying from $Q = 4$ to 63. Higher values (lighter shades) correspond to greater agreement between the metadata partition and the partition inferred by our model.

that without introducing further variation, say through the weighting proposed, there isn't a strong enough signal to come through at these scales within the network.

Indeed, all three weighted versions of metadata appear to permit clusterings with much closer correspondence with the partition inferred. The ASJC and subject area codes appear to separate the network well at the highest level, while deteriorating at finer levels of granularity – this is likely due to the increased importance of interdisciplinarity at these scales. On the other hand, the weighted version of institional affiliation becomes more closely aligned at deeper levels of the hierarchy. This is also logical, as it is particularly within each institution that authors often most collaborate, and perhaps with other institutions nearby, but this is at a relatively fine-grained view relative to the network as a whole.

The count of citations an author receives – both raw and the log of this value – once again provides a clustering that decreases in similarity at deeper levels. As a proxy of the prestige of an author, this perhaps suggests that there is some stratification of collaborations at a higher level, but within each of the resulting (broad) bands finer distinctions matters significantly less.

Finally, clusterings obtained from the combination of all metadata considered once

again diminish in similarity as we progress down the hierarchy. This clearly shows that while metadata provides an important contribution to the groups inferred, it by no means dominates them – the network structure is still vital.

This application provides both (i) a case study of how the model performs when many nodes are missing at different timesteps – something that may suggest inhomogeneous group transitions, a potential problem for the model given that is one of the assumptions – and (ii) a possible example of when the tuning parameter should be varied over the network, not necessarily because of poor alignment, but also because of the large quantity of metadata at each node, as previously discussed. Comparison with state-of-the-art methods for static networks (*e.g.* [136], as implemented in `graph-tool`) appears to show relatively good agreement overall in terms of how the blocks capture network structure, while also providing more human-interpretable groups that better represent differences in metadata. In some areas of the network, it also appears to be able to use metadata to further distinguish differences between members of what may otherwise be a larger group, exactly as desired.

5.5 Discussion

In this chapter, we have introduced a highly scalable BP inference scheme for the model previously described, and demonstrated that indeed, as predicted by the analysis of Chap. 4, the inclusion of metadata permits near-perfect recovery of groups well past the previous limit. Note that in general, using labels derived from the metadata to initialise the messages as we performed is an entirely legitimate practical strategy, as the metadata is assumed to be observed. This is somewhat in contrast to most 'semi-observed' settings, *e.g.* as used for label propagation models [191]. As such, given the great boost in recovery observed even with lower levels of information content, it would be interesting to investigate in future how performance in empirical networks changes using the same strategy. For instance, one could perform simple K-means clustering on available node-wise metadata, then use this to initialise the messages and marginals.

We also probed in greater depth the utility of the tuning parameter for networks where metadata and network partitions are misaligned, showing that on average there does appear to be some correlation between the optimal tuning parameter value and the quantity of alignment between the partitions. This also highlighted how for certain networks, the MFVI approach of Chap. 3 may prove more robust, and capable

of overcoming issues that cause the BP method to fail to converge — though we emphasise that for the majority of tests performed, BP still provides superior results.

We concluded with an application of our hierarchical inference procedure to a dynamic co-authorship network of Latin American authors, revealing clear linguistic separation within the network, as well as proving that our method can scale sufficiently well to handle networks with thousands of nodes at each timestep. Further study of how the modification of network edge weights impacts results, for instance through fractional weighting as is popular in the literature [88], would be valuable. A deeper exploration of the wealth of information that our model can provide on these empirical networks is also worthwhile, for instance through use in link prediction, and proper estimation of the relative importance of each type of metadata. For this purpose, a key area of current development is the presentation of results in an interactive format, so non-technical experts and non-specialist users can better interpret the groups inferred, and the model outputs as a whole. We explore some of the predictive capabilities of the model in the subsequent chapter.

More technical avenues worth exploring in future work include determining suitable corrections for the update equations for networks with loops, for instance along the lines suggested in [75], somewhat related to the Kikuchi or cluster variational method [196], and a greater exploration of how the model might be most useful in practice — which may not be through simply determining groups in the network. On this latter point, in the following chapter we pursue the use of inferred node marginals as possible 'deconfounders' for causal inference problems with analogous structure, which we believe may be a particularly promising venture.

We also intend to resolve a potential downside to our implementation, in that we choose to update all messages synchronously rather than sequentially. This is a major difference to many other BP implementations, and means the information available to a node at each timestep is significantly more dated than in an asynchronous framework. Indeed, as stated earlier in the chapter, it is known that parallel updating for BP does not have guaranteed convergence in general, even on trees, unlike the typical serial case [169]. This might prompt adoption of a more modern parallelisation approach, *e.g.* as in [101, 61] should this property be desirable — we instead plan to address it through a subsequent redevelopment of the codebase, by simply allowing asynchronous updates once more.

Chapter 6

Dynamic substitutes for causal inference of author influence

In this chapter, we conclude by exploring one promising application of improved models for networks with metadata — to assist when performing causal inference in the presence of confounders. In addition, we both (i) describe how to model and perform BP inference for directed dynamic networks – including for a degree-corrected model – and (ii) provide one way to extract parameters for the full network after applying the top-down hierarchical fitting procedure previously described. Both of these extensions are stimulated by the empirical data we consider.

Specifically, we investigate how our models can be used to generate substitutes for potential confounders in a dynamic causal model, and apply the proposed procedure to better understand which authors carry the most influence in a network – without conflating influence with popularity, whether of authors or their research areas. As the influence of an author is not something for which accurate quantitative data is available, we evaluate the quality of our method by both generating semi-synthetic data, for which the 'ground-truth' influence is known, and by performing posterior predictive checks. We elucidate precisely what this means in greater detail below.

Despite the estimation of author influence being an area of considerable research interest, to our knowledge we are the first to take an explicitly causal view of the matter, and hence bring methods from the causal inference toolkit to bear on the matter.

6.1 Introduction

In many fields of contemporary academia, when speaking to individual academics in the area about their sources of inspiration, it is common to hear the same names pop up again and again. Identifying these 'thought leaders' – considered here as the key people driving forward a given topic – is something that humans, or at least specialists in the area in question, are quite good at doing given sufficient time to do so. Indeed, the ability to identify these leaders in practice is useful for a variety of reasons, from the immediate author-level purpose of keeping up-to-date with the cutting edge of your domain, to the more holistic aims of those directing research funding *etc.* However, despite its importance, there is no universally agreed-upon procedure to quantify such influence given the data available.

There are several methods that have been used to try and estimate the influence of an academic in the literature. Some common methods include:

- Citation analysis: Citation analysis is a widely used method for estimating the influence of an academic in the literature. This method involves counting the number of times an academic's work has been cited by other researchers, as well as the impact factor of the journals in which their work has been published. This method has been criticised for its reliance on the number of citations, which can be influenced by factors such as the popularity of a particular research topic, rather than the true quality of the work. Additionally, the impact factor of journals is not always a reliable indicator of the quality of the research they publish [18, 27].

- Co-authorship analysis: Co-authorship analysis is another commonly used method for estimating the influence of an academic in the literature. This method involves looking at the number of times an academic has co-authored papers with other researchers, as well as the impact of those papers. This method has been criticised for its reliance on the number of co-authored papers, which can be influenced by factors such as the size of a research team and the popularity of a particular research topic. Additionally, the impact of co-authored papers may not always accurately reflect the contribution of an individual academic [51, 2].

- Network analysis: Network analysis is a method that has been used to try and understand the influence of an academic in the literature by looking at the connections between researchers. This method involves looking at who has

co-authored papers with whom, and using this information to understand the influence of an academic. This method has been criticised for its reliance on co-authorship data, which can be influenced by the factors mentioned above. Additionally, network analysis alone does not necessarily capture the true influence of an academic, as it does not consider factors such as the quality or impact of their work, and is typically restricted somewhat arbitrarily to considering specific domains [107, 110].

- Social media analysis: Social media analysis is a relatively new method that has been used to try and estimate the influence of an academic in the literature. This method involves looking at the number of followers an academic has on social media platforms, as well as the reach and engagement of their posts. This method has been criticised for its reliance on social media metrics, which can be influenced by factors such as the popularity of a particular research topic or the personal brand of the academic. Additionally, social media metrics may not always accurately reflect the true influence of an academic in the literature [78].

As we elucidate in detail through the proposed mathematical framework below, our definition of an author's influence is as the degree to which other – citing – researchers continue to follow their work, and subsequently explore or expand upon their topics themselves, particularly when they may not have been expected to do so otherwise. That is, for an academic to be a true leader in their field (or outside of this) at a given time, it is insufficient for them to have either (a) been influential or produced influential work at some point in the past – their output must continue to be of import – nor (b) simply publish either many papers, or *e.g.* survey papers, and thus receive many citations or a high degree of awareness of their work, while *changing* the research output of others only slightly.

We note that this definition of influence is distinct from the influence propagation/maximisation, or information diffusion literature, which have also received much attention — particularly recently, as they can help identify nodes to target for interventions to identify and/or slow *e.g.* epidemic processes spreading on the network, though some analogies can be drawn (see *e.g.* [53]). In that framework, typically the variable of interest is the probability of a cascade sequence [87, 158], *e.g.* the probability of a series of retweets, but most often outside of a causal framework, which can thus lead to spurious inferred relations at the individual level. As a result, there is generally a focus on network-level properties, such as epidemic thresholds, or average rate of diffusion, rather than estimating the influence of individual nodes (academics)

as we are interested. Additionally, the methods discussed above are generally based on quantitative data, such as the number of citations or co-authorships, while the influence propagation or information diffusion literature is often based on more qualitative data, such as the abstract notion of ideas or opinions being spread, and the characteristics of the people or organisations involved in the diffusion process.

Within this chapter, we place a particular emphasis on the need to approach the evaluation of author influence, as we have defined it, with consideration of causal effects — specifically, through positing a causal model. Causal models are necessary for estimating the influence of an author because they allow us to determine the extent to which a particular factor or treatment causes a specific outcome. In the context of estimating the influence of an author, this means that causal models can help us to identify – and thus account for – the specific factors that contribute to what inspires an author's work.

That is, we may try and tease apart the contributions from *e.g.* the popularity of a topic, or an author's preference for it, compared to what we are really interested in — the specific influence of other authors. Without viewing the problem through a causal lens, it is likely that many of the methods above have produced biased estimates of author influence for precisely this reason. For example, the popularity of a research topic or the reputation of the journal in which an author's work is published may influence the number of citations their work receives, but these factors may not accurately reflect the quality or originality of the work. By using causal models, we can instead explicitly control for these factors and focus on the specific contributions of an author to the research community.

In the existing literature, there is a degree of awareness of this problem – the presence of *confounding* factors – but thus far, attempts to alleviate the issue almost exclusively use one of the simplest methods of causal adjustment — instrumental variables (IV). The basic justification for the necessity of causal methods, including IV, is that if we are interested the effect of a particular 'treatment', X, on an outcome, Y, but we are aware of the presence of other factors, Z, that influence both X and Y, we must account for the resulting spurious correlation between the two. IV effectively takes a two-stage approach to addressing this [156]. First, one seeks 'instruments', W, which affect the treatment X, but *not* the outcome Y. By then modelling X as a function of these instruments, and any other exogenous variables, the idea is that you are left with only the part of X which is independent of Y. As such, as a second stage you may proceed to use these estimates of X given the instruments to estimate Y, and hence estimate the causal effect. For instance, in the Science of Science, these have been

used *e.g.* to estimate the impact of federal funding on university publication output [125], scientific output and citation count given patent authorship [21], scientific output (*i.e.* publication count) given funding and some simple co-authorship network features [14], and numerous other problems — generally focusing on either funding or output based outcomes. However, the requirement of independence of Y to the instruments is often only weakly met, if at all, and furthermore they may only impact X very slightly — each of these, let alone both, mean that the instruments are ineffective in remediating the confounding problem we started with. We elaborate an alternative (though still imperfect) causal adjustment procedure – back-door adjustment – below.

One of the primary reasons that instrumental variables are not particularly well-suited for assessing causal effects at the level of individual academics is the importance of network effects. When considering interpersonal networks, these are often viewed as being constituted by two dominant driving social processes[96]. The first, social influence, is used to refer to the more active effect of interactions breeding conformity with peers. For instance, people trying to 'fit in' by changing their views to more closely match those of their friends (close to the influence propagation / opinion dynamics framing). The second, homophily, refers to more passive processes for why the network has the observed structure in the first place — the selection of those particular connections. The assumption is that if this is important, individuals are more likely to connect to those that *e.g.* already hold similar views to their own. There have been some efforts to separate social influence from homophily in networks, *e.g.* [81], and indeed estimating homophily and social influence were some of the key drivers of the development of certain network models, *e.g.* [165].

In this work, we aim to investigate both of these — the former is effectively what we are interested in estimating, while controlling for the latter. Specifically, in our case authors are likely to cite others in their own field, not necessarily because they are inspired by them, but because their research is related, and further as they are in the same research community, they may well know each other personally — thus they determine that they warrant a citation, via a homophilic effect.

As we describe in detail below, the problem is that the precise factors driving these confounding effects are not typically observed. In the example given of an author's research community, there is no ground-truth label that tells us this information. To try and circumvent this problem, a recent strand of literature has proposed the idea of *substitutes* — in brief, the replacement (*i.e.* substitution) of unobserved confounders by latent factors inferred for a probabilistic model with the same posited dependency structure [180, 166]. While it is only recently that this idea has been formalised to any

degree, we note that other authors have effectively been using substitutes, and just calling them confounders — *e.g.* for understanding influence in social networks [164]. However, these works have generally been quite limited, and as they skip the step of explicitly posing a causal model, they missed the key observation of works utilising substitutes like [166] — that using a *joint* model for factors that influence both links and the outcome of interest would likely be more suitable.

Beyond this idea of substitutes, there have been other interesting pathways proposed for causal adjustment in networks, for instance via 'causal' graph neural networks (GNNs) [90, 95], or through randomisation [171], but we do not explore such herein. We note that given alternative methods to IV for causal effect estimation, including substitutes, have been explored in the (static) network context, it is perhaps surprising that to our knowledge these have not been brought to bear on any type of academic network — particularly as these are often some of the first empirical networks to receive consideration for any new model or method, given the ready availability of numerous datasets.

The remainder of the chapter is as follows. First, in Sec. 6.2, we provide a brief review of some of the key ideas and tools required for our proposed procedure, which we elaborate in Sec. 6.3. We then proceed to describe the dataset considered in our experiments – including how we generate semi-synthetic data for 'ground-truth' influence values by which we can evaluate our models – in Sec. 6.4. The results of these experiments are given in Sec. 6.6[1], and we then conclude in Sec. 6.7 with a brief discussion, and some immediate options for future work.

6.2 A partial review of the causal inference toolkit

To introduce some of the key ideas in causal influence, consider a simplified related example. Suppose we're interested in the number of publications an academic produces, in any topic, given their known collaborators – *i.e.* what is the impact of working alongside those particular people, rather than any of the alternatives?

Initially, this might seem like a fairly immediate problem to address, and indeed numerous papers investigate similar problems by simply looking at summary statistics over some corpus – we could for instance directly compare the rate of publication of each author immediately before and after initiating a collaboration with a new individual, relative to their previous collaborators that did not join this new team

[1]Code necessary to perform these experiments is publicly available on GitHub. This was developed, and since substantially modified, from a fork of the code for PIF [166].

(effectively a kind of difference-in-differences [37]). However, the problem is that this does not necessarily isolate the impact of the collaboration itself from factors that might affect *both* who the author collaborates with, and their propensity for publishing.

For instance, each author is affiliated to a particular institution, which affects both who they are more likely to collaborate with, and the number of publications they produce, due to *e.g.* publication requirements for tenure, the opportunities that come with the prestige of association with that institute *etc.* Additionally, the funding of an academic will likely have a considerable effect on both their collaborations, and the number of publications they produce – perhaps from a combination of saving time applying for further grants, being able to take on PhD students and/or post-docs to explore areas of interest, and simply due to having some more money to enjoy at their leisure.

Variables such as these, which are a common cause to both the 'treatment' of interest (here the author's previous collaborations), and the outcome of interest (their scientific 'productivity' as measured by raw publication count), are called *confounders*. Indeed, both of these confounding factors are themselves of course related to the academic's true ability, but this is unobserved.

Now for simplicity, imagine that given their affiliation and funding, academics care little about their collaborators ability, nor does their ability further directly affect their propensity to publish – plausible perhaps if academic institutions and funding bodies are sufficiently accurate estimators of the true ability of an author, and for these to be known as such, so most academics at comparable institutes with comparable funding are both similarly productive and similarly attractive prospects for collaboration.

As popularised by J. Pearl (see *e.g.* [129, 128]), this set of relations can be expressed in the structural causal model (SCM) shown in Fig. 6.1a, where a directed edge $i \to j$ signifies a causal effect of i on j, and a node is shaded if the corresponding variable is observed, else it is unobserved (or latent). To simplify notation as we formalise this example below, we recreate this same graph in Fig. 6.1b, where we follow the convention [130] of using Y for the outcome variable of interest, given the 'treatment' variable X, Z for observed variables, and U for unobserved variables. Posing the problem in such a way – as a directed acyclic graph (DAG) – immediately clarifies issues that will arise when investigating causal relations between the observations. For instance, even if this model is correct, then as they share a common cause – the academic's ability – there will be a spurious correlation between the affiliation of an author and their funding, even though there is no causal effect of either on the other.

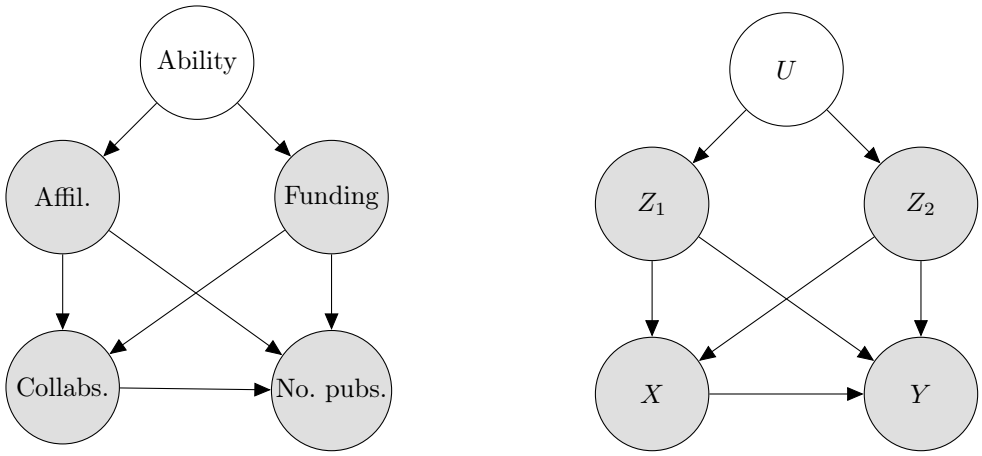


FIGURE 6.1: A simple causal model for the number of publications an academic produces, in text (left) and causal variable (right) form. Shaded variables are observed, unshaded are unobserved (latent).

6.2.1 Formalising causality

We take the 'potential outcomes' approach of causality, along with 'do-calculus' notation [130, 127]. By the former, we mean that for a treatment where a variable X is assigned the value x, we are interested in comparing to *counterfactual* scenarios, where this value were instead set to x' say. For instance, in the example above, we would be interested not in the outcome of the number of publications given an author is collaborating with a particular set of individuals in the abstract, $\mathrm{p}(y \mid x)$, but rather comparing the outcome observed when the author was 'treated' to collaborate with that specific team (effectively *made* to work with those collaborators), *vs.* the potential outcome that would be observed in an imaginary world where they were instead made *not* to collaborate with a different team. That is, a causal effect should emerge from *interventions*.

These interventions are defined mathematically by the aforementioned do-calculus. We correspondingly use the notation $\mathrm{do}(x)$ to mean that we *fix* the variable X to the value x. The effect of this do-calculus has a simple probabilistic interpretation: given a distribution $\mathrm{p}(x_1, x_2, \dots)$ defined by a DAG, G, such that

$$\mathrm{p}(x_1, x_2, \dots) = \prod_{x_i} \mathrm{p}(x_i \mid \mathrm{pa}_G(x_i)), \tag{6.1}$$

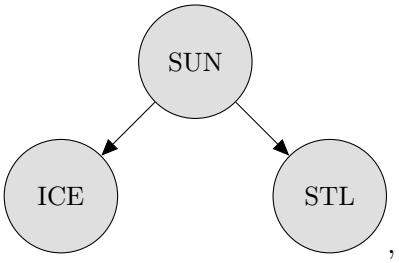
where $\mathrm{pa}_G(X_i)$ corresponds to the parents of X_i in G (*i.e.* the variables that have edges directed into X_i, which may be an empty set), denoting by $\boldsymbol{X}^{-i}$ the set of

variables excluding X_i, we define $\mathrm{p}(\boldsymbol{x}^{-i} \mid \mathrm{do}(x_i))$ as the 'mutilated' distribution

$$\begin{aligned} \mathrm{p}(\boldsymbol{x}^{-i} \mid \mathrm{do}(x_i)) &= \frac{\mathrm{p}(\boldsymbol{x})}{\mathrm{p}(x_i \mid \mathrm{pa}(x_i))}, \\ &= \prod_{x_j \in \boldsymbol{x}^{-i}} \mathrm{p}(x_j \mid \mathrm{pa}(x_j)), \end{aligned} \tag{6.2}$$

where we remove the factor $\mathrm{p}(x_i \mid \mathrm{pa}(x_i))$ as we have set its given value, rather than permitting it to be drawn freely.

This immediately highlights the key difference between an intervention, and a merely observed association between two factors — the former asserts a causal relationship, while the latter only connotes a conditional relationship. For instance, one might observe in sales data that there is a correlation between purchases of sun tan lotion, STL, and ice cream, ICE. As such, the conditional probability $\mathrm{p}(\mathrm{STL} \mid \mathrm{ICE}) \neq \mathrm{p}(\mathrm{STL})$, and this could lead to the naïve belief that ice cream sales were driving those for sun tan lotion. However, the simplest causal model for this situation would be that both were in fact driven by sunny weather, SUN, *i.e.*



,

so $\mathrm{p}(\mathrm{ICE}, \mathrm{STL}, \mathrm{SUN}) = \mathrm{p}(\mathrm{SUN})\mathrm{p}(\mathrm{ICE} \mid \mathrm{SUN})\mathrm{p}(\mathrm{STL} \mid \mathrm{SUN})$. In this case, if we now intervene on ice cream sales (perhaps by lowering their price) — hence removing the term $\mathrm{p}(\mathrm{ICE} \mid \mathrm{SUN})$ from the joint distribution — the distribution over STL is now given by $\mathrm{p}(\mathrm{STL} \mid \mathrm{do}(\mathrm{ICE})) = \sum_{\mathrm{SUN}} \mathrm{p}(\mathrm{STL} \mid \mathrm{SUN})\mathrm{p}(\mathrm{SUN})$, which is clearly independent of whichever intervention was performed on ICE.

Combining these two frameworks, we now have a methodology for quantifying what we mean by the causal effect of a particular intervention. Specifically, for a binary treatment on a variable X, as suggested above (*i.e.* either work with that team, $x = 1$, or do not, $x = 0$), we may then express our goal as determining the average treatment effect (of participating in that team), on some outcome of interest, Y (number of publications)

$$\mathrm{ATE}(x) = \mathbb{E}[Y \mid \mathrm{do}(x = 1)] - \mathbb{E}[Y \mid \mathrm{do}(x = 0)]. \tag{6.3}$$

Before proceeding, we require some intermediary definitions [126]. In the following,

we use $\text{an}_G(X)$ to denote the 'ancestors' of X in a DAG G, that is X itself, and all variables that are reachable from X by following edges in the reverse ($\leftarrow$) direction.

Definition 6.2.1 Let $\mathcal{G}$ be a directed graph and π a path in $\mathcal{G}$. We say that an internal vertex t on π is a *collider* if the edges adjacent to t meet as $\rightarrow t \leftarrow$. Otherwise ($\rightarrow t \rightarrow$, $\leftarrow t \rightarrow$ or $\leftarrow t \leftarrow$) we say t is a non-collider.

Definition 6.2.2 Let π be a path from a to b. We say that π is *open* given (or conditional on) $C \subseteq V \setminus \{a, b\}$ if

- all colliders on π are in $\text{an}_{\mathcal{G}}(C)$;
- all non-colliders on π are outside C.

Recall that $C \subseteq \text{an}_{\mathcal{G}}(C)$. A path which is not open given C is said to be *blocked* by C.

We note that perhaps the simplest interpretation of this definition is to view the model like a plumbed system, with edges as pipes and variables as valves. If we do not interfere (condition on variables) water (potential causality) can flow from one location (variable) to another (*i.e.* the path between the two is open) through the pipes (the edges) unless it is blocked by a closed valve (a collider). If we do 'interfere' by conditioning on variables, then this act switches the state of the valve – if the variable is a collider then the 'valve' becomes open, otherwise it closes it.

Definition 6.2.3 Let A, B, C be disjoint sets of vertices in $\mathcal{G}$ (C may be empty). We say that A and B are *d-separated* given C if every path from $a \in A$ to $b \in B$ is blocked by C.

Now the key question we're interested in for performing causal inference is how we can resolve the distribution $\text{p}(y \mid \text{do}(x))$ even in the presence of confounders. The seminal contribution of Pearl [129] was introducing the method of – and precise requirements for – *back-door adjustment*, as follows:

Definition 6.2.4 We say that a set of variables, C, in a probabilistic graphical model is a back-door adjustment set for the ordered pair (v, w) if

- no vertex in C is a descendant of v
- every path from v to w with an arrow into v (*i.e.* starting $v \leftarrow \cdots$) is blocked by C

Theorem 6.2.1 *Let C be a back-door adjustment set for (v, w). Then*

$$\mathrm{p}(x_w \mid do(x_v)) = \sum_{x_C} \mathrm{p}(x_C)\mathrm{p}(x_w \mid x_v, x_C), \tag{6.4}$$

i.e. *C is a valid adjustment set for the causal distribution.*

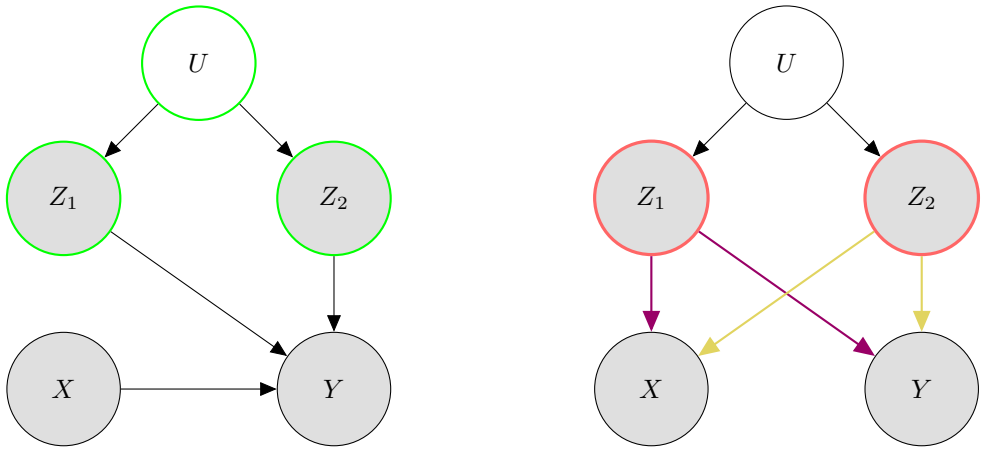
The requirements for a back-door adjustment set can be verified quickly graphically for a given DAG using various subgraphs of G, as follows. We use the expression $(X \perp\!\!\!\perp Y \mid Z)_G$ to denote that the set Z d-separates X from Y in G, then $G_{\overline{X}}$ to be the graph obtained by deleting from G all arrows pointing to nodes into X (*i.e.* those from the parents of X in G), and likewise $G_{\underline{X}}$ the graph obtained by deleting from G all arrows emerging from nodes in X (those to the children of X in G). The requirement of a back-door adjustment set, C, is then simply that $C \subseteq \mathrm{an}_G(X)$, *i.e.* $(W \perp\!\!\!\perp X)_{G_{\overline{X}}}$, and $(X \perp\!\!\!\perp Y \mid W)_{G_{\underline{X}}}$.

Back-door adjustment provides a way to understand simple circumstances under which we can determine a causal effect, *i.e.* resolve a control problem, in which we specify a particular treatment (or set of treatments/plan):

Definition 6.2.5 The expression $\mathrm{p}(y \mid \mathrm{do}(x))$ is said to be *identifiable* in a DAG G if, for every assignment $\mathrm{do}(x)$, the expression can be determined uniquely from the joint distribution of the observables $\{X, Y, Z\}$. A control problem is said to be identifiable whenever $\mathrm{p}(y \mid \mathrm{do}(x))$ is identifiable.

Returning to our toy example, the SCM in Fig. 6.1, we can observe that indeed, if the model is specified correctly, then the effect of collaboration on the number of publications produced is identifiable. We show the two key subgraphs for the problem in Fig. 6.2 – in subfigure (a) we show $G_{\overline{X}}$, and highlight in green the variables that are then d-separated from X, thus candidates for the adjustment set. In subfigure (b) we show $G_{\underline{X}}$, and highlight in purple and yellow the two main paths that need to be closed by the adjustment set – we don't show the path going via U (the academic's ability), as blocking the two highlighted paths will necessarily block that also. Clearly, under the given model the academic's affiliation and funding, Z_1 and Z_2 respectively, highlighted in red, hence provide a back-door adjustment set by which we can properly assess the causal effect of collaborations in question, via Eqn. (6.4).

Of course, the problem in practice is that performing this adjustment requires that the confounders are *observed*. For instance, in our example, if we quite reasonably assumed that potential collaborators are themselves discerning of ability, and that



(a) $G_{\overline{X}}$ (b) $G_{\underline{X}}$

FIGURE 6.2: The two key subgraphs for determining back-door adjustment sets in our toy example. In subfigure (a) we remove edges into X, and highlight in green the variables that are then d-separated from X, thus candidates for the adjustment set. In subfigure (b) we remove edges from X, and highlight in purple and yellow the two main paths that need to be closed by the adjustment set, Z_1, Z_2, which we now highlight in red.

ability also has a direct effect on the number of publications produced, then we could no longer immediately use back-door adjustment to proceed.

6.2.2 From latent confounders to substitutes

To circumvent this problem of latent confounders, we build upon the ideas proposed in Poisson Influence Factorisation (PIF) [166], a method for estimating social influence that addresses precisely this challenge. The key concept in PIF is that of *substitutes*, as formalised in [180] for the 'deconfounder' algorithm for multivariate treatments. This in turn is motivated by a simple insight — even though the confounders are latent, if the posited causal model is accurate, they should express themselves through their common causes – in our case, citations and topics – and so there must be some evidence of them in the observed data. As such, by fitting latent variable models with this same structure to the observations, we should capture some of the same information as the confounders themselves. The latent variables inferred – *i.e.* groups of authors, or general topic factors – may then be used as drop-in substitutes for these confounders in the chosen causal adjustment procedure to uncover the 'treatment' effect. For our purposes, as we elaborate in Sec. 6.3, this treatment will be an author j publishing in a topic k in the preceding time period, and being cited by the author

i, for whom we are trying to determine the influence on the outcome of i subsequently publishing in that same topic.

The general justification for substitutes is as follows. Consider a set of observations – *e.g.* the citations a_i made by an author i – and initially for simplicity assume that there are no traits which uniquely affect these observations a_i, and not our treatment (publication topics), y_i. Then the individual set of confounders ζ_i may be further separated: either a confounder affects multiple observations (citations) in a_i, or it affects only a single observation, $a_{ij} \in a_i$ say. Now the assumption is that the observations a_i are rendered conditionally independent given ζ_i, *i.e.*

$$\mathrm{p}(a_i \mid \zeta_i) = \prod_j \mathrm{p}(a_{ij} \mid \zeta_i), \tag{6.5}$$

but any confounder, b_i say that only impacts a single observation will not modify this equation in any way — *i.e.* even if we could find a substitute that provides perfect conditional independence of the observations, we would not capture the information of b_i. This is one of the key assumptions of the method of substitutes — that all confounders affect multiple observations, as any variable that only influences a single observation leaves no discernible implication in the data.

Given that this is the case, we may propose a generative model that posits this same factorisation of a_i given some latent variables c_i. If such a model well-approximates the empirical distribution of the observations, a_i – *i.e.* the model is well-specified – then the variables c_i must capture the same information as the confounders ζ_i of the causal model.

Formalising the above corresponds to the following [166]:

Proposition 6.2.2 *Let $u_k \subset \{\zeta_j\}$ be the variables of an individual j that affect more than one observation in a_j. Suppose there exists a variable c_j such that the distribution $\mathrm{p}(a_j \mid c_j)$ factorises as*

$$\mathrm{p}(a_j \mid c_j; \eta_{a_j}) = \prod_i \mathrm{p}(a_{ij} \mid c_j; \eta_{a_j}), \tag{6.6}$$

for some parameters η_{a_j}. Then, the variable c_j contains the information that is contained in the set of variables u_j.

Proof 6.2.1 Assume that a variable c_j which satisfies the factorisation above exists. Then, by definition, each pair of variables (a_{ij}, a_{kj}) in the set a_j are d-separated by c_j. Suppose for contradiction that c_j does not contain all the information contained in the

confounders u_j, common causes of multiple observations in a_j. If this were the case, it implies there exists some subset $U \subset u_j$ that is not contained in a_j but which are parents of at least two variables, (a_{ij}, a_{kj}) in a_j say. But we assumed that c_j d-separates all pairs of variables in a_j. Hence, c_j must contain all the information contained in u_j. ■

With this tool at our disposal, we may proceed to probe the question of author influence. We briefly note that there are potential alternatives to the use of substitutes, in particular proxy modelling — the use of 'proxies', *i.e.* noisy observations of, or variables dependent on the latent confounders, that in turn have weaker restrictions on their dependencies than other options like IV or mediators. Indeed, these are an increasing area of interest following the seminal work of [79], particularly since very recently the constraints on the different causal relations posed, linearity of models, and discreteness of variables have begun to be relaxed [97, 189]. We leave exploration of such options for future work.

6.3 Our proposed causal model

As suggested above, the work of [166] – the Poisson Influence Factorisation (PIF) model, which uses network data, as well as information about the past and present behavious of nodes in the network, to understand the influence of each member on the behaviours of others – was a key inspiration for the methodology of this chapter.

The PIF model was designed with the aim of ascertaining how the purchase (or non-purchase) of a product, k, by an individual, j, yesterday affects the (non-)purchase of the same product today by another individual, i, given the presence/absence of a social link, a_{ij} between these individuals. To do so, it assumes that in addition to (i) individual-level factors that influence only the likelihood of social links, z_i, *i.e.* personal traits that do not directly affect purchases, and (ii) individual-level factors that influence only the likelihood of purchases, θ_i, *i.e.* personal preferences for products that do not directly relate to their social behaviours, there exist confounding factors. Specifically, they posit that there are both factors that influence purchases *and* social connections (*e.g.* the individual belongs to a particular subculture), ρ_i, and factors that influence purchases both yesterday and today (perhaps the general popularity of a product), τ_k.

We build upon their work, by incorporating dynamics into our model via first-order Markov processes over latent parameters. To allow comparison, we typically use similar notation for analogous variables. Specifically, to investigate which authors are most

influential, given citation data and publication topics, we posit the following causal relationships between variables of interest:

- If one author influences another, then the topics of the potentially influenced author i at timestep t, denoted $y_{i.}^{t}$, should be partially caused by the topics of the influencing author j at the preceding timestep. Adopting the notation of PIF, we denote these previous topics by $x_{j.}^{t-1}$ — using different notation to 'influenced' topics y, despite identical underlying data (*i.e.* $X = Y$), clarifies the role these topics play in the model, and further permits us to consider binarising these past topics separately, should we assume that the counts of influencing author topics are unimportant;

- The knowledge of another author's work, and acknowledgement of such as relevant to their own work, may be expressed through citation relations, where we use a_{ij}^{t} to denote the number of times author i cited author j in period t. As publications take time to produce, and new knowledge takes time to be fully incorporated, we assume the key citation-mediated dependence in terms of influence on an author's current research topics is with the citations of the previous time period.

  The logic is also that even if the citation is to a publication that the cited author did not produce in the previous period, the citation demonstrates knowledge of that author's work. As such, if we are interested in estimating contemporary author influence at *each* timestep, rather than over the course of their career, the combination of the cited author's most recent publication topics, alongside the citation demonstrating that the citing author is indeed both aware of, and somewhat related to their output, should be the most insightful.

  That is, we expect that if an author, Jane say, is currently producing influential work, then another author, Geoff, that is aware of Jane's work in general should be influenced to publish in similar topics at the next timestep, whether or not there is evidence of them previously citing papers by Jane in that specific topic — if she is influential *currently*, rather than simply producing influential work at some point in the past, then Geoff would keep up with her work, and subsequently be influenced.

- Topics themselves have some attributes which make them more or less attractive — for instance the 'hotness' of a topic, say deep learning, which drives many authors to publish in it, rather than necessarily being influenced by any key

individual. These attributes change over time, dependent on their previous values — hotness does not emerge *ex nihilo*.

- Similarly, there are author attributes which affect both their preferences for particular topics, and who they are most likely to cite — for instance, their research community and academic background;
- Authors may also have further topic preferences, that are presumed to not directly impact citation links – perhaps they have a hobby that suddenly sparks some new connection between topics – and additional attributes that affect citations, but not topics — maybe the propensity to give token citations to important authors.

Formalising these posited relations results in the SCM presented in Fig. 6.3, where the variables are defined as follows:

- $x_{jk}^t \in \{0, 1\}$: Author j publishes in topic k at time t (optionally not binarised);
- $y_{ik}^t \in \mathbb{N}$: Number of papers author i publishes in topic k at time t (optionally binarised);
- $a_{ij}^t \in \{0, 1\}$: Author i cites author j at time t;
- $\zeta_i^t \in \{1, \ldots, Q\}$ or $\mathbb{R}^Q$: The attributes (*e.g.* group) of author i at t, influencing both metadata (including topics) and links;
- $\tau_k^t \in \mathbb{R}^K$: Attributes (*e.g.* group) affecting popularity of topic k at time t;
- $\theta_i^t \in \mathbb{R}^{P_K}$: Author i's preference for topics at time t;
- $z_i^t \in \mathbb{R}^{P_N}$: Author i's traits that drive connections;

In this model, $x_{ik}^t = y_{ik}^t$, but we denote them separately according to whether we are considering them as treatment or outcome.

6.3.1 Translating author influence into an estimable quantity

Using this causal model, we may now proceed to formalise what we mean by author influence. We follow the same argument as described for the toy problem in Sec. 6.2. That is, the crucial quantity for determining author influence is the difference between the number of publications an author, i, produces in a particular topic, k, at time t, given we have evidence of their knowledge of another author, j at the previous

timestep – as i cited j at $t-1$ – under the binary treatment of j publishing in that same topic k at that time, *vs.* the counterfactual world where instead j did not publish in that topic.

Precisely, this is the causal effect

$$\xi_{ijk}^t = \mathbb{E}[y_{ik}^t \mid a_{ij}^{t-1} = 1; \mathrm{do}(x_{jk}^{t-1} = 1)] - \mathbb{E}[y_{ik}^t \mid a_{ij}^{t-1} = 1; \mathrm{do}(x_{jk}^{t-1} = 0)], \tag{6.7}$$

i.e. the average treatment effect of j publishing in k at $t-1$ on i publishing in k at t, given the citation.

As authors may publish in a variety of areas, and are almost certainly cited by a wide variety of other academics, we then consider the main causal effect of interest to be the *average* influence of each academic at each timestep, over all citing authors and publication topics. That is, for $n_j^t = |\partial j^t|$ the total number of unique authors who cite j at time t, and M the total number of topics in the corpus, we wish to estimate

$$\xi_j^t = \frac{1}{n_j^t} \sum_{i \in \partial j^t} \frac{1}{M} \sum_k \xi_{ijk}^{t+1}. \tag{6.8}$$

If this average influence is high for an author, then we should expect to see that many other authors who cite them at some point – demonstrating they are both aware of their work, and in a related area – subsequently absorb their ideas and begin publishing related work.

Much as in our toy example, to proceed in estimating the component influence quantities of Eqn. (6.7), we must attempt to resolve confounding bias between our treatment and outcome, *i.e.* we must block back-door paths that result in spurious non-causal associations. In Fig. 6.4, we highlight these paths. In yellow we show the obvious path via the evolving topic-only attributes, $x_{jk}^{t-1} \leftarrow \tau_k^{t-1} \rightarrow \tau_k^t \rightarrow y_{ik}^t$, while in blue we show the path via the citation collider, $x_{jk}^{t-1} \leftarrow \zeta_j^{t-1} \rightarrow a_{ij}^{t-1} \leftarrow \zeta_i^{t-1} \rightarrow \zeta_i^t \rightarrow y_{ij}^t$, which we only opened by our conditioning on the citation a_{ij}^{t-1}. To block these paths, in red we highlight one choice of back-door adjustment set, τ_k^{t-1} and ζ_i^{t-1}, which thus allow us to block the paths to the outcome, hence shown dashed. As these are latent, these are what we aim to estimate using substitutes.

In our experiments in Sec. 6.6, we test both the use of substitutes for τ_k^{t-1} and ζ_i^{t-1}, and τ_k^t and ζ_i^t. The former choice allows us to hold out entire time periods, and so not pollute our evaluation of the model by directly incorporating information about the observations y^t themselves — this is analogous to the approach of PIF. However, the problem is that to properly adjust for these, we should marginalise out τ_k^t and ζ_i^t,

and this is not necessarily trivial — we avoid this by assuming they may play the role of τ_k^t and ζ_i^t directly, but this is a somewhat naïve assumption. Alternatively, using τ_k^t and ζ_i^t requires us to observe all the data for which we are estimating influence, but then the substitute adjustment procedure is much simplified — this is more akin to another substitute-based causal model, Social Poisson Factorisation (SPF) [24].

Now assuming perfect substitutes, to estimate individual effect of the particular link $i \to j$, and publication k by j at $t-1$, we must first have marginalised out all other connections of i at $t-1$, other than that to j, and all publications of j at $t-1$ in topics other than k. Once again, we use $a_i^{-j,t} = a_i^t \setminus \{a_{ij}^t\}$ to denote such connections excluding that to j at t, and analogously $x_k^{-j,t} = x_k^t \setminus \{x_{jk}^t\}$ for publications of j at t excluding k.

Performing the necessary marginalisation then corresponds to

$$\xi_{ijk}^t = \mathbb{E}_{a_i^{-j,t-1}, x_k^{-j,t-1}}\Big[\mathbb{E}[y_{ik}^t \mid a_{ij}^{t-1} = 1, a_i^{-j,t-1}, x_k^{-j,t-1}; \mathrm{do}(x_{jk}^{t-1} = 1)] \\ - \mathbb{E}[y_{ik}^t \mid a_{ij}^{t-1} = 1, a_i^{-j,t-1}, x_k^{-j,t-1}; \mathrm{do}(x_{jk}^{t-1} = 0)]\Big], \quad (6.9)$$

where the inner distribution is with respect to the number of publications of i in k, y_{ik}^t.

Now to utilise the back-door adjustment formula, Eqn. (6.4) , we first define the conditional expected outcome given the confounders (or substitutes) and pertinent observed data,

$$\mu_{ik}^t(a, x) = \mathbb{E}[y_{ik}^t \mid a_{ij}^{t-1} = a, x_{jk}^{t-1} = x, a_i^{-j,t-1}, x_k^{-j,t-1}, \zeta_i^{\{t-1,t\}}, \tau_k^{\{t-1,t\}}]. \quad (6.10)$$

In terms of $\mu_{ik}^t(a, x)$, we now have that Eqn. (6.9) becomes

$$\xi_{ijk}^t = \mathbb{E}_{\zeta_i^{\{t-1,t\}}, \tau_k^{\{t-1,t\}}}\left[\mathbb{E}_{a_i^{-j,t-1}, x_k^{-j,t-1}}[\mu_{ik}^t(1,1) - \mu_{ik}^t(1,0)]\right], \quad (6.11)$$

And so finally, we have that the average causal effect of author j on those who cite them is

$$\psi_j^{t-1} = \frac{1}{n_j K} \sum_{i \in \partial j,k} \mathbb{E}_{\zeta_i, \tau_k}\left[\mathbb{E}_{a_i^{-j}, x_k^{-j}}[\mu_{ik}(1,1) - \mu_{ik}(1,0)]\right]. \quad (6.12)$$

In general, computing these expectations would not necessarily be a trivial task, but by positing particular distributions for the outcome given its stated dependencies, it is possible to simplify matters. Continuing to follow along the lines of PIF [166], we use a Poisson likelihood to model the outcome at each timestep given the substitutes,

$\hat{\tau}$, $\hat{\zeta}$, such that

$$\mathrm{p}(Y^t \mid A^{t-1}, X^{t-1}, \hat{\tau}, \hat{\zeta}) = \prod_{i,k} \mathrm{Pois}(y_{ik}^t \mid \lambda_{ik}^t), \tag{6.13}$$

$$\text{where} \quad \lambda_{ik}^t = \gamma_k^{t,\top} \hat{\zeta}_i^t + \alpha_i^{t,\top} \hat{\tau}_k^t + \sum_j a_{ij}^{t-1} x_{jk}^{t-1} \beta_j^{t-1}. \tag{6.14}$$

In this model, we presume that the expected number of times person i publishes in a topic k at time t, is solely a linear function of their estimated traits – ζ_i^t and α_i^t, the estimated attributes of topic k, and the all important influence from their peers – controlled by the parameter β. As in PIF, we place conjugate sparse Gamma priors on the unobserved variables γ, α, β, to ensure that the term λ_{ik}^t is non-negative.

Using this model, analogously to PIF we directly have the following interpretation of parameters:

Proposition 6.3.1 *If*

(i) the functions of expected publications, $\mu_{ik}^t(a, x)$, satisfy

$$\mu_{ik}^t(a, x) = \mathbb{E}[y_{ik}^t \mid a_{ij}^t = a, x_{jk}^t = x, a_i^{-j,t-1}, x_k^{-j,t-1}, \hat{\zeta}_i^{\{t-1,t\}}, \hat{\tau}_k^{\{t-1,t\}}], \tag{6.15}$$

and

(ii) the publications y_{ik}^t are drawn from the Poisson model in Eqn. (6.13),

then $\xi_j^{t-1} = \beta_j^{t-1}$.

More verbosely, the first assumption is that the substitutes $\hat{\zeta}_i^{\{t-1,t\}}$ and $\hat{\tau}_k^{\{t-1,t\}}$ are ideally valid — *i.e.* they are perfect replacements for their respective latent confounders, as they contain identical information with respect to the expected publications y_{ik}^t. The second assumption is that the posited model for the outcome, y_{ik}^t, is perfectly specified. Given both of these hold, the average influence ξ_j^t of author j at t is equal to the parameter β_j^t in Eqn. (6.13). This result only holds strictly in the limit of infinite data, where expectations taken over the observations are exact — even with both assumptions, there are no known guarantees on the quality of estimates from finite samples.

The proof of this result is as follows:

Proof 6.3.1 Define $\tilde{\mu}_{ik}^t(a, x) = \mathbb{E}[y_{ik}^t \mid a_{ij}^{t-1} = a, x_{jk}^{t-1} = x, a_i^{-j,t-1}, x_k^{-j,t-1}, \hat{\zeta}_i^{\{t-1,t\}}, \hat{\tau}_k^{\{t-1,t\}}]$, *i.e.* $\mu_{ik}^t(a, x)$ but using the substitutes in lieu of confounders. Now recall from

Eqn. (6.12),

$$\xi_j = \frac{1}{n_j M} \sum_{i \in \partial j^t, k} \mathbb{E}_{\zeta_i^{\{t-1,t\}}, \tau_k^{\{t-1,t\}}} \left[\mathbb{E}_{a_i^{-j,t-1}, x_k^{-j,t-1}} [\mu_{ik}^t(1,1) - \mu_{ik}^t(1,0)] \right]. \tag{6.16}$$

By assumption (ii), we have that the likelihood of publications y_{ik}^t is given by Eqn. (6.13). Using the fact that $\mathbb{E}[Y] = \lambda$ for a random variable $Y \sim \text{Pois}(\lambda)$, the rate λ_{ik}^t for each publication y_{ik}^t corresponds exactly to $\tilde{\mu}_{ik}^t(a, x)$, when we substitute the values a and x for a_{ij}^{t-1} and x_{jk}^{t-1} in the expression Eqn. (6.14).

Now by assumption (i) we have that $\mu_{ik}^t(a, x) = \tilde{\mu}_{ik}^t(a, x) = \lambda_{ik}^t$, and hence

$$\begin{aligned} \mu_{ik}^t(1,1) - \mu_{ik}^t(1,0) &= \gamma_k^{\{t-1,t\},\top} \hat{\zeta}_i^{\{t-1,t\}} + \alpha_i^{\{t-1,t\},\top} \hat{\tau}_k^{\{t-1,t\}} + \sum_{l \neq j} a_{il}^{t-1} x_{lk}^{t-1} \beta_l^{t-1} + \beta_j^{t-1} \\ &\quad - \gamma_k^{t-1,\top} \hat{\zeta}_i^{t-1} - \alpha_i^{t-1,\top} \hat{\tau}_k^{t-1} - \sum_{l \neq j} a_{il}^{t-1} x_{lk}^{t-1} \beta_l^{t-1} \\ &= \beta_j^{t-1}. \end{aligned} \tag{6.17}$$

For clarity, in the first line we separated the sum over citees l into the term that correspond to author j plus the remaining terms — these then immediately cancel.

Finally, substituting the above equation into the definition of ξ_j^t,

$$\xi_j^t = \frac{1}{n_j^t M} \sum_{i \in \partial j^t, k} \mathbb{E}_{a_i^{-j,t-1}, x_k^{-j,t-1}} \left[\mathbb{E}_{\zeta_i^{\{t-1,t\}}, \tau_k^{\{t-1,t\}}} [\beta_j^{t-1}] \right] = \beta_j^{t-1}. \tag{6.18}$$

■

Note that if the proposition assumptions do not hold, *i.e.* either the model for publications is mis-specified (and not necessarily Poisson), and/or the substitutes do not contain identical information to the true latent confounders, then the key steps of equating $\mu_{ik}^t(a, x)$ to $\tilde{\mu}_{ik}^t(a, x)$, and the latter to λ_{ik}^t, no longer hold. In particular, these assumptions make the calculation of the expectations in Eqn. (6.12) trivial, as the variables over which they are being taken have vanished from the equation. This circumvents the requirement of any additional steps to determine author influence — we only have to (i) fit the models used to define the substitutes, then (ii) hold these fixed in Eqn. (6.13), and fit the remaining parameters. Hence the importance of understanding the behaviour under model mis-specification, as the authors of PIF suggest — we return to this below.

To estimate the remaining parameters of Eqn. (6.13), given the substitutes $\hat{\zeta}$, $\hat{\tau}$, we use a lightly modified version of the MFVI procedure described in PIF [166] — we

provide the details in App. C.

We also stress again that for the substitutes to be valid, the generative models used to construct them must capture the empirical data distributes. When we perform our experiments in Sec. 6.6, we thus first evaluate both models ability to represent the observed data, using posterior predictive checks — we describe these in detail below.

Of course, in practice, in addition to the generative models being mis-specified, the causal model for publications itself may be also, leading to biased estimates of influence. We leave analytic exploration of this issue of estimation quality to future work, but provide empirical results that suggest that in such a case, the use of substitutes can have additional benefits.

6.3.2 Chosen models for substitutes

Now the key step to complete our goal of estimating the average influence ξ_j^t of an author j at each timestep, as defined in Eqn. (6.8), is to construct good quality substitutes for the homophily confounders, ζ, and topic-only confounders, τ. As originally proposed in the method of deconfounding [180], these substitutes are taken to be the expected values of the latent variables under the corresponding substitute model — *i.e.* in our case, $\hat{\zeta} = \mathbb{E}[\zeta \mid A, Y]$, and $\hat{\tau} = \mathbb{E}[\tau \mid Y]$.

The principal stimulation for the work of this chapter is that the DSBMM should be especially well-suited to play such a role. To elucidate this, consider what information is captured by the substitute $\hat{\zeta}_i^t$ — chosen to correspond to the node marginal distribution $\psi^{i,t}$ in the DSBMM. In the proposed causal model, there are multiple causes for the citations, A — that is the set of citations for each author a_i^t is driven, and rendered conditionally independent, by *both* the confounder ζ_i^t, and additional author traits z_i^t. As such, a dynamic extension of *e.g.* Social Poisson Factorisation [24], which effectively finds substitutes that render the links in the network, A, conditionally independent, but does not do so *simultaneously* for the observations Y, is unlikely to be ideal. That is, by neglecting Y during the inference procedure, the resulting substitutes are likely to contain a significant amount of information about the link-only variables, z_i^t, which do not directly impact the causal quantity of interest.

Instead, by jointly modelling A and Y – as the DSBMM does – we may better target ζ-specific information, and thus the substitutes should be of higher quality. In our experiments below, we compare results that jointly use network and publications (metadata) to find substitutes for ζ to those where we neglect publications, and if this is an important property, doing so should improve overall performance. In both

cases, we expect that accounting for the evolution of the system over time will also be important if there is a significant relation between present and past states — this means that methods such as PIF which neglect this may not perform optimally. We return to this later, when describing our semi-synthetic data generation procedure.

To find substitutes for the topic-only confounders τ, we may choose any other model with the desired dependency structure — in the following, we use dynamic Poisson factorisation (dPF) [26]. There are of course numerous alternatives one might explore.

Before proceeding, note that a key assumption for the DSBMM is that the network and the metadata are conditionally independent given the network groups, *i.e.* $A \perp\!\!\!\perp X \mid Z$. Furthermore, this is important for the causal inference procedure of this chapter, as in the proposed causal model model we assume $A \perp\!\!\!\perp X \mid \zeta$. As a result, A^t is a non-descendant of X^t, and hence a valid member of the back-door adjustment set. If this were not the case, estimation of the causal effect is still possible, but would require a different procedure. We provide some of the details in App. D.

6.4 Data

Now that the problem has been clarified, in this section we describe the available data in detail. For all results in this chapter, we use real citation information to construct a dynamic network — however, to provide a means to evaluate our model against known ground-truth influence values in any setting, we consider data about the topics in which an author publishes produced in two ways. In the first, for the given network, we generate synthetic topics according to the assumed data generating process, for which we thus know the true influence values for each author at each timestep. As such, the dataset is semi-synthetic — it should reproduce some aspects of the real dataset, while allowing us to use classical supervised metrics to evaluate the quality of our estimates. We elaborate this process in Sec. 6.4.2. In the second scenario, we use the same network, but instead use SciVal topic clusters assigned to each publication, provided by Elsevier[2] — we use this real data, as described in Sec. 6.4.1, to help determine which authors to use to construct the overall citation network we consider.

[2]Further information on SciVal topics may be found on the Elsevier website, here.

6.4.1 Full empirical data

To produce our full dynamic network with metadata, we once again use the Scopus corpus of publications by Latin-American affiliated authors from 1997–2020, aggregated into three-year windows as described in previous chapters. However, unlike our prior case studies, we now construct the network at each time period using the citation information available — a directional relation, rather than the symmetric relationship that we assumed co-authorship to be.

Specifically, a weighted, directed edge a_{ij}^t between authors i and j in time period t now denotes the number of times which author i cited author j during the period. To allow matching of the citation to the author j, the cited publication must be within our dataset, thus from 1997 onwards, but other than this we discard temporal information about the citation — *i.e.* we keep the edge as a triple, (i, j, t), rather than the quadruple $(i, j, t^{\text{pub.}}, t^{\text{cit.}})$ which it truly is. We discuss one option of incorporating this information in future work at the end of this chapter.

Given the resulting dynamic network, we then consider our metadata to be the counts of publications of an author in each SciVal topic cluster during each period. That is, x_{ik}^t now denotes the number of publications that author i produces during time period t in topic k.

We further pre-process this data to reduce it to more prominent authors and topics of interest. First, much as before we discard the initial time periods, for which the network available from our dataset is highly disconnected and sparse. We instead take only the final five periods, from 2006 onwards. During this period, between 68 and 73% of citation relations are not reciprocated, so it is indeed important to consider as a directed network, unlike the small citation case study of Chap. 3.

Next, we require that all topics occur in at least 20 publications in this period, and that all authors publish in at least two of these periods. This leaves us with 104,512 authors and 1,152 topic clusters, with 468,886 unique observations. The corresponding author-publication topic matrix, X, even aggregated over the full period, has a sparsity of around 99.6% — among the higher considered in most collaborative filtering tasks. Around 82% of these observations have a count of two or fewer, so where it may be beneficial – specifically for using dPF to construct the topic-only substitutes – we treat X as binary, which the authors suggest is preferable.

We then proceed to perform one final filtering step, to remove more peripheral authors, and improve overall network connectivity. To focus on better-connected authors, we consider k-core subgraphs at each timestep — the maximal subgraph with minimum total degree of at least k, not counting self-edges [15]. We choose

$k = 18$ when taking the k-core as this is the median total degree over all timesteps – it will constrain the network more at earlier, sparser timesteps, but not too much at later timesteps with more connections. We further restrict ourselves to authors that are present in the largest weakly connected component of this subgraph at any timestep – *i.e.* the union of the nodes in the largest component at each timestep, which thus typically provides several components in the network of each period, with one dominating the others. By doing so we end up with 43,494 authors, who – in particular in the most recent period (that we are most interested in) – are generally well-connected. We note that this final weak-connectedness constraint does not greatly reduce the number of authors considered, as in the 18-core subgraph at each timestep, the largest component accounts for almost all the present authors.

After obtaining this union set, we consider all available data for these authors at each time, rather than maintaining the k-core restriction. The resulting networks are overall then very well-connected — between 97.6% (first period) and more than 99.9% (final period) of the subset of authors chosen are weakly connected by citations in each period, and of course by construction every author belongs strongly to the largest connected component in at least one period. As such, there should be sufficient information to properly identify research communities.

In addition to their publication topic counts, for each author we also have a wealth of further information. In particular, we geocode their affiliations, and extract both the country and the ADM1 region to which they belong — the first administrative level below the nation, typically the state. We then construct a categorical variable for each author at each timestep, by assigning them to the region to which they were affiliated most frequently. We do not provide this information to our models directly, but as discussed in previous chapters this geographical information is known to be an important factor in scientific production. As such, instead we use it as a plausible homophily-related factor to assist in generating synthetic data, as we describe in the following section.

6.4.2 Semi-synthetic data

The problem with evaluating the success of the model previously described is that the actual ground-truth influence of an author is not known. As such, following along the lines of [166], we first construct a semi-synthetic dataset, in which we take real citation network data, and then simulate publications in 1000 artificial topics as follows:

1. Each publication in a particular topic is assumed to depend on the topic's attributes at that time (*i.e.* τ_k^t), and the author's preferences for those attributes (*i.e.* θ_j^t and ζ_j^t).

2. We construct four categorical covariates, two for authors and two for topics.

   For authors, as previously suggested we take one likely real confounder – the ADM1 region to which an author is affiliated at each timestep (if there are multiple in the period, we take the most common), which we denote r_i^t for author i at time t – and then randomly simulate another categorical covariate with 5 categories, v_i^t, which should hence be completely unrelated to the links observed in the network. The idea is that r_i^t is one possible homophily-related confounder, *i.e.* ζ, that might influence both connections and topics — if two authors are based in the same region, they're more likely to know about each other (as well as speak the same language *etc.*), but different regions often also have different preferences for topics, *e.g.* due to funding available, the region's economic specialties *etc.* On the other hand, as v_i^t doesn't affect links but does affect topics, this will play the role of the topic-only confounders, τ.

   We verify that the posited first-order Markov assumption is reasonable for the region data, by comparing Bayesian information criterion (BIC) [106] values between this and a second-order Markov model. The former is better in this metric, so we proceed as planned, though it is plausible that *e.g.* on shorter timescales higher-order terms become more important, and/or that perhaps if we compared to a variable-order model instead, it would demonstrate that specific higher maximum order sequences were important.

   For topics, we construct a region-related covariate by randomly sampling a categorical variable of the same number of categories as countries observed, r_k^t. Rather than try to produce a realistic evolution of such region-related topic attributes over time, we sample these once and then fix to the same values over all timesteps — this may be reasonable, if for instance knowledge is particularly spatially embedded, and capabilities in each region only change slowly. The simulated homophily-related confounding may nonetheless vary considerably for each author over time, as they change affiliations and move between regions. We then again construct another random categorical variable covariate with 5 categories, v_k^t, which we evolve according to a simple first-order Markov process, such that $\mathrm{p}(v_\ell^t \mid v_k^{t-1}) = \eta_v + (1-\delta_{k\ell})(1-\eta_v)/5$. For the experiments in Sec. 6.6.1, we use $\eta_v = 0.8$.

3. We then construct P-dimensional region-related topic attributes, and author preferences for these attributes, according to

$$\gamma_{kp}^t \sim r_k^t \cdot \text{Gam}(a, b) + (1 - r_k^t) \cdot \text{Gamma}\left(\frac{a}{s_\gamma}, b\right), \tag{6.19}$$

$$\zeta_{ip}^t \sim r_i^t \cdot \text{Gam}(a, b) + (1 - r_i^t) \cdot \text{Gamma}\left(\frac{a}{s_\zeta}, b\right), \tag{6.20}$$

along with P-dimensional topic attributes and author preferences related to the random covariate,

$$\tau_{kp}^t \sim v_k^t \cdot \text{Gam}(a, b) + (1 - v_k^t) \cdot \text{Gamma}\left(\frac{a}{s_\tau}, b\right), \tag{6.21}$$

$$\alpha_{ip}^t \sim v_i^t \cdot \text{Gam}(a, b) + (1 - v_i^t) \cdot \text{Gamma}\left(\frac{a}{s_\alpha}, b\right), \tag{6.22}$$

where we use Gamma distributions to ensure non-negativity, as required for the Poisson rate parameter to which they contribute.

4. Given these attributes and corresponding preferences, we assume that topics are dependent on some simple combination of these two covariate-dependent attributes and preferences. That is, we consider whether there is confounding due to homophily (terms related to the region covariates), topic-only confounding (terms related to the random covariates, described as exogenous confounding below), or both together, *i.e.*

$$\mu_{ik}^t \in \{(\zeta_i^t)^\top \gamma_k^t,\ (\alpha_i^t)^\top \tau_k^t,\ (\zeta_i^t)^\top \gamma_k^t + (\alpha_i^t)^\top \tau_k^t\}, \tag{6.23}$$

such that

$$x_{jk}^0 \sim \text{Pois}(\mu_{jk}^0), \tag{6.24}$$

$$y_{ik}^t \sim \text{Pois}(\mu_{ik}^t + \sum_j a_{ij}^{t-1} x_{jk}^{t-1} \beta_j^{t-1}). \tag{6.25}$$

Thus far, the generating process for these synthetic topics matches that posited above. However, the procedure (as described in detail in App. C) assumes that the final remaining parameter to sample, β, is drawn independently at each timestep, from a conjugate Gamma prior. In practice, this is unlikely to hold — logically, an author's influence at time t is likely to depend to some degree on their previous influence.

As such, to provide a more realistic baseline to evaluate different substitute choices, we introduce a mild mis-specification into the model. We assume that the outcome equation, Eqn. (6.13), still holds, but now sample β according to

$$\beta_i^t \sim \begin{cases} \text{Gamma}(0.005, 0.1) \text{ if } t = 1, \\ \text{Normal}(\beta_i^{t-1}, 0.05 \times 0.05) \text{ else,} \end{cases} \tag{6.26}$$

where the standard deviation is chosen for fluctuations of roughly 5% of the initial mean, and we threshold any negative values to zero. Correspondingly, we expect that substitutes that better capture the evolution of the observed data may also capture some of this temporal dependence. Importantly, as the Gamma prior for all β^t is now mis-specified, this means that providing the true confounders as generated above – often called the 'oracle' setting in machine learning literature – may not provide the best results. If this is the case, then it highlights an important potential benefit of using substitutes — if the model used approximates the true data generating process sufficiently well, then it is possible that it may capture information that was not directly considered in the (mis-)specified causal model.

5. Finally the influence is post-processed – any authors with no sampled publications and/or no observed citations (given or received) at that timestep have their influence set to one, to match with the Gamma distribution in this case. We account for this when evaluating the quality of our models in Sec. 6.6.1, so metrics are not artificially inflated by missing data.

6.5 Necessary extensions for the DSBMM

The data described above necessitates two primary further model developments to proceed. Firstly, the low rate of reciprocation of citations observed means that considering the networks as undirected is no longer a reasonable simplification. In Sec. 6.5.1 we resolve this by elaborating the necessary equations for performing belief propagation inference for directed DSBMMs, including permitting degree correction.

Secondly, to perform the posterior predictive checks for the network, we require the parameters fitted that govern edge probabilities, *i.e.* ω_{qr}^t or λ_{qr}^t depending on whether we are performing degree correction or not, while if we wish to approximately evolve our citation-topic substitutes $\hat{\zeta}$ we require the inferred transition matrix, π. However, these are not trivially available if we use the hierarchical procedure described in the

previous chapter to accelerate inference. As such, in Sec. 6.5.2, we provide one method to approximate global parameters for the full network from those inferred at each level of the hierarchy.

6.5.1 Belief propagation for directed DSBMMs

The directed version of the DC-SBM, as introduced in [203], is much as one might expect, *i.e.* making the same assumption as before that we can approximately perform degree correction by simply replacing the node-wise parameters by the degrees – now separated by in- and out-degrees – we have

$$\mathrm{p}(A \mid \omega, Z) = \prod_{i,j} \frac{(d_i^{\mathrm{out}} d_j^{\mathrm{in}} \omega_{z_i z_j})^{A_{ij}}}{A_{ij}!} \exp(-d_i^{\mathrm{out}} d_j^{\mathrm{in}} \omega_{z_i z_j}). \tag{6.27}$$

Belief propagation for directed SBMs has been explored previously in the literature [183], but to our knowledge never for the degree-corrected model, let alone now accounting for group dynamics and nodal metadata. Following the same logic as in the previous chapter, the resulting message equations in this case are

$$\begin{aligned}\psi_q^{i\to j} = \frac{\alpha_q \mathrm{p}_q(x_i)}{Z^{i\to j}} \prod_{k\in\partial i\setminus j} \Bigg[\sum_r \Bigg\{ &\frac{(d_i^{\mathrm{out}} d_k^{\mathrm{in}} \omega_{qr})^{A_{ik}}}{A_{ik}!} \exp(-d_i^{\mathrm{out}} d_k^{\mathrm{in}} \omega_{qr}) \\ &\times \frac{(d_k^{\mathrm{out}} d_i^{\mathrm{in}} \omega_{rq})^{A_{ki}}}{A_{ki}!} \exp(-d_k^{\mathrm{out}} d_i^{\mathrm{in}} \omega_{rq}) \Bigg\} \psi_r^{k\to i} \Bigg] \\ \times \prod_{k\notin\partial i\setminus j} \sum_r &\exp(-d_i^{\mathrm{out}} d_k^{\mathrm{in}} \omega_{qr}) \exp(-d_k^{\mathrm{out}} d_i^{\mathrm{in}} \omega_{rq}) \psi_r^{k\to i},\end{aligned} \tag{6.28}$$

but importantly note that this is not the same as just allowing asymmetry in block parameters. In particular, the external field has an extra term (effectively twice the parameter value when undirected), and the product over both directions is inside the sum in the message equation.

Indeed, the key term for the external field is now

$$\begin{aligned}
&\prod_{k\notin\partial i\setminus j}\sum_r \mathrm{e}^{-d_i^{\mathrm{out}}d_k^{\mathrm{in}}\omega_{qr}}\mathrm{e}^{-d_k^{\mathrm{out}}d_i^{\mathrm{in}}\omega_{rq}}\psi_r^{k\to i}\\
&\qquad\approx \prod_{k\notin\partial i\setminus j}\sum_r (1-d_i^{\mathrm{out}}d_k^{\mathrm{in}}\omega_{qr})(1-d_k^{\mathrm{out}}d_i^{\mathrm{in}}\omega_{rq})\psi_r^k,\\
&\qquad\approx \prod_{k\notin\partial i\setminus j}\sum_r (1-(d_i^{\mathrm{out}}d_k^{\mathrm{in}}\omega_{qr}+d_k^{\mathrm{out}}d_i^{\mathrm{in}}\omega_{rq}))\psi_r^k,\\
&\qquad= \prod_{k\notin\partial i\setminus j}\left(1-\sum_r (d_i^{\mathrm{out}}d_k^{\mathrm{in}}\omega_{qr}+d_k^{\mathrm{out}}d_i^{\mathrm{in}}\omega_{rq})\psi_r^k\right),\\
&\qquad\approx \exp\left(\sum_{k,r}(d_i^{\mathrm{out}}d_k^{\mathrm{in}}\omega_{qr}+d_k^{\mathrm{out}}d_i^{\mathrm{in}}\omega_{rq})\psi_r^k\right),\\
&\qquad= \exp\left(d_i^{\mathrm{out}}\sum_{k,r}d_k^{\mathrm{in}}\omega_{qr}\psi_r^k+d_i^{\mathrm{in}}\sum_{k,r}d_k^{\mathrm{out}}\omega_{rq}\psi_r^k\right),
\end{aligned} \tag{6.29}$$

i.e. we now have

$$h_q^{i,t}=d_i^{t,\mathrm{out}}\sum_{k,r}d_k^{t,\mathrm{in}}\omega_{qr}^t\psi_r^{k,t}+d_i^{t,\mathrm{in}}\sum_{k,r}d_k^{t,\mathrm{out}}\omega_{rq}^t\psi_r^{k,t} \tag{6.30}$$

We performed the final rearrangement into two sums to make it clear that with each update of a marginal for a node, we need only update each sum term of the generic external field with the new value (*i.e.* subtract the old $d_k^{\mathrm{in}}\omega_{qr}\psi_r^k$ term from the first sum, and $d_k^{\mathrm{out}}\omega_{rq}\psi_r^k$ from the second, then add the same terms with the updated values) – the dependence on i then only enters by taking the product of these with the corresponding out- and in-degree respectively. As such, there is minimal slow-down.

The final simplified message equation – once again neglecting the temporal aspect, which remains unchanged – is now as we would expect,

$$\begin{aligned}
\psi_q^{i\to j}\propto \alpha_q \mathrm{p}_q(x_i)\mathrm{e}^{-h_q^i}\prod_{k\in\partial i\setminus j}\Bigg[\sum_r\Bigg\{&\frac{(d_i^{\mathrm{out}}d_k^{\mathrm{in}}\omega_{qr})^{A_{ik}}}{A_{ik}!}\exp(-d_i^{\mathrm{out}}d_k^{\mathrm{in}}\omega_{qr})\\
&\times\frac{(d_k^{\mathrm{out}}d_i^{\mathrm{in}}\omega_{rq})^{A_{ki}}}{A_{ki}!}\exp(-d_k^{\mathrm{out}}d_i^{\mathrm{in}}\omega_{rq})\Bigg\}\psi_r^{k\to i}\Bigg],
\end{aligned} \tag{6.31}$$

that is, an exact analogy of the NDC version only with a suitably updated edge likelihood and corresponding external field.

Now the parameter updates are also slightly changed. Once again following the logic of [34], the suitable values correspond to the expected value of MLEs that depend

on observables (*e.g.* count of edges between groups) under the specified variational formulation.

To find these, we first aggregate Eqn. (6.27) over the groups, to rewrite as

$$\mathrm{p}(A \mid \omega, Z) = \frac{\prod_i (d_i^{\mathrm{out}})^{d_i^{\mathrm{out}}} (d_j^{\mathrm{in}})^{d_j^{\mathrm{in}}} \prod_{q,r} \omega_{qr}^{m_{qr}} \exp(-\kappa_q^{\mathrm{out}} \kappa_r^{\mathrm{in}} \omega_{qr})}{\prod_{i,j} A_{ij}!}, \tag{6.32}$$

up to a constant, where much as before we have $\kappa_q^{\mathrm{out}} = \sum_{i \in q} d_i^{\mathrm{out}}$, $\kappa_q^{\mathrm{in}} = \sum_{i \in q} d_i^{\mathrm{in}}$, and $m_{qr} = \sum_{i \in q, j \in r} A_{ij}$. Ignoring constants, the important log-likelihood term is then

$$\log \mathrm{p}(A \mid \omega, Z) = \sum_{q,r} m_{qr} \log \omega_{qr} - \kappa_q^{\mathrm{out}} \kappa_r^{\mathrm{in}} \omega_{qr}, \tag{6.33}$$

and so we find the MLE

$$\omega_{qr} = \frac{m_{qr}}{\kappa_q^{\mathrm{out}} \kappa_r^{\mathrm{in}}}. \tag{6.34}$$

As such, to satisfy the Nishimori condition we want

$$\omega_{qr} = \frac{\langle m_{qr} \rangle}{\langle \kappa_q^{\mathrm{out}} \rangle \langle \kappa_r^{\mathrm{in}} \rangle} = \frac{\sum_{i,j \in \mathcal{E}} \psi_{qr}^{ij} A_{ij}}{\left(\sum_i d_i^{\mathrm{out}} \psi_q^i\right) \left(\sum_j d_j^{\mathrm{in}} \psi_r^j\right)} \tag{6.35}$$

The two-point marginals are constructed by $\psi_q^{i \to j} f_{ij}(z_i = q, z_j = r) \psi_r^{j \to i}$ for $f_{ij}(z_i = q, z_j = r)$ the *ordered* pairwise factor between i and j given their groups, *i.e.* in this case $f_{ij}(z_i = q, z_j = r) = \frac{(d_i^{\mathrm{out}} d_k^{\mathrm{in}} \omega_{qr})^{A_{ik}}}{A_{ik}!} \exp(-d_i^{\mathrm{out}} d_k^{\mathrm{in}} \omega_{qr})$. In other words, as claimed before, the spatial parameter update equations in the directed case are almost the same as in the undirected case (identical for the NDC case), only without the addition of the reverse term (present due to symmetry in the parameters that no longer holds), and with suitable modifications of the likelihood term for the DC case. Note that in fact as the sum in the undirected case is over unique edges, this is exactly equivalent to the directed case if every edge is reciprocated.

6.5.2 Approximating global parameters from a top-down hierarchical application

We now describe how we might obtain global parameters for the DSBMM from the series of submodels fitted during application of our hierarchical procedure. Focusing on the transition matrix for now, perhaps the simplest way we could construct the overall matrix would be assuming uniform likelihood of transitions to any sub-group

in another branch, *i.e.*

$$\pi^L_{qr} \mid q^{\ell-1} = r^{\ell-1} = a_{q^{\ell-1}} \tilde{\pi}^\ell_{qr}, \tag{6.36}$$

where $\ell - 1$ is the level in the hierarchy at which point q and r merge, $q^{\ell-1}$ corresponds to the group at this level to which q, r belong, and $\tilde{\pi}^\ell_{qr}$ corresponds to the transition matrix actually inferred at level ℓ. The prefactor $a_{q^{\ell-1}}$ ensures overall normalisation of π^L_{qr}. If groups q and r never merge, *i.e.* they split at the first level, then ℓ is defined to be the first level (as at the 'zero' level all groups merge).

As the rows must sum to unity, *i.e.* $\sum_r \pi^L_{qr} = 1$ for each q, our assumption of uniform transitions to groups in other branches translates to

$$\pi^L_{qr} = \left(\prod_{m<\ell} \tilde{\pi}^m_{q^m q^m} \right) \tilde{\pi}^\ell_{q^\ell r^\ell} / |\mathrm{desc}(r^\ell)|, \tag{6.37}$$

where q^m denotes the ancestor group of q at level m, and $|\mathrm{desc}(r^\ell)|$ is the number of descendants of the ancestor of r at level ℓ. When $\ell = 1$ we set the value of the empty product to one, and when $\ell = L$ we set the value of $1/|\mathrm{desc}(r^L)| = 1$.

Verifying this, using $\hat{\pi}^m$ for values of the normalised transition matrix were level m to become the new top level, we would have

$$\begin{aligned}
\sum_r \pi^L_{qr} &= \tilde{\pi}^1_{q^1 q^1} \sum_{r \in q^1} \hat{\pi}^2_{q,r} + \sum_{r \notin q^1} \pi^L_{qr}, \\
&= \tilde{\pi}^1_{q^1 q^1} \sum_{r \in q^1} \hat{\pi}^2_{q,r} + \sum_{r^1 \neq q^1} |\mathrm{desc}(r^1)| \tilde{\pi}^1_{q^1 r^1} / |\mathrm{desc}(r^1)|, \\
&= \tilde{\pi}^1_{q^1 q^1} \sum_{r \in q^1} \hat{\pi}^2_{q,r} + (1 - \tilde{\pi}^1_{q^1 q^1}), \\
\sum_{r \in q^1} \hat{\pi}^2_{q,r} &= \tilde{\pi}^2_{q^2 q^2} \sum_{r \in q^2} \hat{\pi}^3_{q,r} + \sum_{r^2 \neq q^2} |\mathrm{desc}(r^2)| \tilde{\pi}^2_{q^2 r^2} / |\mathrm{desc}(r^2)|, \\
&= \tilde{\pi}^2_{q^2 q^2} \sum_{r \in q^2} \hat{\pi}^3_{q,r} + (1 - \tilde{\pi}^2_{q^2 q^2}), \\
&\vdots \\
\sum_{r \in q^L} \hat{\pi}^L_{qr} &= \tilde{\pi}^L_{qq} + (1 - \tilde{\pi}^L_{qq}) = 1, \\
\therefore \sum_r \pi^L_{qr} &= 1.
\end{aligned} \tag{6.38}$$

For instance, if groups q and r meet immediately in q' at level $L - 1$, with $L = 3$, and

q' was a subgroup of $\tilde{q}$ in the first level, then

$$\pi^L_{qr} = \pi^3_{qr} = \tilde{\pi}^1_{\tilde{q}\tilde{q}} \tilde{\pi}^2_{q'q'} \tilde{\pi}^3_{qr}. \tag{6.39}$$

Analogously, for probabilities of edges between nodes belonging to groups in two different blocks, now necessary for performing our PPCs, we assume that the (Q, Q, T) block matrix, ω if NDC or λ else, may be constructed in the hierarchical case in an analogous manner as for π. Doing so is much simpler than for the transition matrix, as the block parameters are effectively edge-wise parameters (just homogeneous over all nodes in the same block) — as such, we can simply take the pairwise block parameter inferred at the lowest possible level, *i.e.* that the probability of an edge from one node in a group $q^{\ell+1}_t \in q^\ell$ to any node belonging to a group in a different branch at the same level, $r^{\ell+1}_t \in r^\ell$ say, is given by *e.g.* $\omega^{t,\ell}_{qr}$, with no prefactor now present.

Future work might explore incorporating the group distributions α fitted at each level to further tune the transitions and block parameters, for π for instance perhaps via taking the product of α^m_r at each node of the tree as you descend to r from ℓ rather than just imposing uniformity. Of course the issue in general is that in our top-down approach, we are actively discarding data as we descend the hierarchy, so the transition matrices (and block parameters) inferred will be imperfect at best.

6.6 Results

In this section, we finally present results obtained from the procedure outlined above. As previously stated, we closely follow [166] in the experimental process used, suitably translated into the temporal case — we describe these experiments in detail for both synthetic and real topic data in their respective subsections below.

For all experiments, rather than operating on the full dynamic network, we perform snowball sampling on the aggregated citation network. This aggregate network is defined by placing an edge between two authors if they cite each other at any point during the entire time period under consideration. Snowball sampling then proceeds as follows: for each run we initialise the sampler at a random node, then iteratively either (a) add any new neighbours of this node to our sample, then randomly choose one of the neighbours as a new root for the next iteration, or (b) if there are no neighbours which haven't already had their own neighbourhoods yet added to the sample, restart the process at a new random root node. The process terminates once a minimum sample size has been drawn — in the following experiments, we require at least 3,000

authors in the semi-synthetic sample, and 8,000 in the real topic sample. This means that each run of the model acts on a different subnetwork from the full dataset — hence, if a model consistently performs well across different subsamples, it is more likely to generalise better to the full system than those that do not, even if they may both do well on particular samples.

For both semi-synthetic and real data, we compare results obtained by using substitutes produced by different models, where the number of groups (for the network models) or factors (for dPF) are held constant in all cases for direct comparability. These hyperparameters are chosen through posterior predictive checks, as described below. The four principal models considered are as follows:

1. **Unadjusted**: We provide no substitutes whatsoever, and so perform no adjustment — if confounding is non-negligible, this should provide the weakest results;

2. **Network-only $\hat{\boldsymbol{\zeta}}$**: We fit a directed, degree-corrected version of the DSBMM to produce a substitute for the topic-citation confounder ζ, but do not account for topic-only confounding from τ;

3. **Topics-only $\hat{\boldsymbol{\tau}}$**: We use dPF to produce substitutes for τ, but do not provide any substitute for ζ;

4. **Network-only $\hat{\boldsymbol{\zeta}}$, $\hat{\boldsymbol{\tau}}$**: We fit the DSBM to the network data, neglecting topic information, for a substitute for ζ, but now also include substitutes for τ from dPF — we refer to this as 'Ours (no meta)' below;

5. **Joint $\hat{\boldsymbol{\zeta}}$, $\hat{\boldsymbol{\tau}}$**: We use the directed, DSBMM, either degree-corrected or NDC, to fit substitutes for ζ that jointly model both the topics and the links, alongside substitutes for τ from dPF — in the following, we refer to these models as 'Ours' and 'Ours-NDC' respectively.

We evaluate our results on authors and topics that both publish/are published in the final period, and were present before this point. These test sets are split from the training data in different ways depending on the experiment, as we elaborate below.

As emphasised above, substitute confounders will only be valid to any degree when the fitted model sufficiently captures the actual distribution of empirical data. Hence, before blindly proceeding with experiments exploring the use of the DSBMM and dPF to generate substitutes for our causal model, we must validate their ability to represent the real data. To do so, we perform posterior predictive checks (PPCs):

1. For each author, randomly hold out some connections and publications – this provides a new dynamic network, A^{heldout}, and set of author publications, X^{heldout};

2. Fit the DSBMM and dPF to these subsampled datasets;

3. Sample s times from the posterior predictive, or an approximation thereof, of each fitted model to create a replicated dataset, A' and X' respectively. For the DSBMM, we use

$$\mathrm{p}(a_{ij}^t) = \sum_{q,r} \psi_q^{it} \psi_r^{jt} \begin{cases} \omega_{qr}^t & \text{if NDC,} \\ \text{Pois}(d_{it}^{\text{out}} d^{\text{in}} \lambda_{qr}) & \text{else,} \end{cases} \tag{6.40}$$

   then sample replicate networks from either Bernoulli or Poisson distributions, while for the dPF we directly use the rates inferred to sample author publications from the corresponding Poisson distribution;

4. For a chosen discrepancy function, $D(\cdot)$ – we use the log likelihood under each selected model – calculate its value on both the replicated datasets, $D(A^{\text{rep}})$, $D(X^{\text{rep}})$, and the real held-out data, $D(A^{\text{heldout}})$, $D(X^{\text{heldout}})$;

5. Finally, the posterior predictive p-value is defined as $\mathrm{p}(D(A^{\text{rep}}) > D(A^{\text{heldout}}))$, *i.e.* the probability that the discrepancy of the replicated data is higher than that of the real, held-out data. A simple way to empirically estimate this is to use the ratio $\frac{s_R}{s}$, where s_R is the number of replicated datasets in which $D(A^{\text{rep}}) > D(A^{\text{heldout}})$.

If the resulting p-value is close to 0.5, this suggests the ideal scenario — that the model explains the replicated data as well as it explains the heldout data, as the chosen discrepancy function cannot be used to meaningfully discern between the two.

6.6.1 Application to semi-synthetic data

We commence our experiments for this chapter by considering the semi-synthetic dataset, for which we know the simulated ground-truth influence for each author. As for PIF [166], we evaluate the performance of each substitute model under varying degrees of confounding – qualitively described as low, medium, and high – for the three confounding scenarios described in Sec. 6.4.2, *i.e.* exogenous (topic-only) confounding, homophily-related confounding, or both together. Additionally, in these semi-synthetic experiments, the knowledge of the true confounding parameters used to generate the

publication data – *i.e.* ζ and τ – allows us to compare results to 'oracle' models, which receive the true confounders (or a subset thereof) prior to fitting the remaining parameters.

Prior to the experiments of this section, we perform PPCs to select the latent dimension of the respective models — Q for the DSBMM, and K for dPF. We consider $Q \in \{4, 9, 16\}$ – chosen to be a square number, so we can apply the hierarchical procedure for our model with two layers, where the submodels at each layer have two, three or four groups respectively – and $K \in \{5, 8, 10\}$. For each choice of Q or K, we take the average over 100 replicates, for 20 different subsampled networks and sets of held-out data. As performed for the experiments in PIF, this held-out data is constructed by sampling for each author, i, and timestep, t, uniformly at random (i) another author j, and viewing the value of a_{ij}^t, and (ii) a topic k, and viewing y_{ik}^t, then setting all viewed values to zero before passing the data to the model. As such, if the sampled author was not cited, or the sampled topic was not published in, the model receives the true corresponding value (a true negative), else the model is potentially mislead by false negatives — the AUC is then a measure of how well the model can both verify true negatives, and detect false negatives. We provide the results for Q and K in Tables 6.1 and 6.2 respectively.

Q	A
4	0.494
9	0.565
16	0.423

Table 6.1: In this table, we present the average PPC scores for different numbers of author groups in the DSBMM, Q, using the method described in the main text. We run 20 experiments, each on a different, snowball sampled subgraph of 3K authors and 1K topics, where for each experiment we fit the DSBMM to the subsampled data, then perform 100 replicate tests.

We observe that the DSBMM seems to represent its target data quite well across the range of Q considered. As such, because all choices are viable under this metric, we experimented with their performance with respect to other measures, such as AUC, and found that $Q = 16$ was the best value to proceed with for all subsequent experiments.

On the other hand, the results for the dPF under this measure suggest a poor capability to represent this synthetic data. However, this is primarily an artifact of the generative process used, when applied in conjunction with the real networks considered.

K	Y
5	0.994
8	0.987
10	0.968

Table 6.2: In this table, we present the average PPC scores for different numbers of topic factors in dPF, K, using the method described in the main text. We run 20 experiments, each on a different, snowball sampled subgraph of 3K authors and 1K topics, where for each experiment we fit the dPF to the subsampled data, then perform 100 replicate tests.

Specifically, the real citation networks used become more dense over the time period considered, both as more authors enter the system, and the average number of citations to others within the dataset increases. As such, by imposing that author influence remains relatively steady over the full time period considered, the total contribution from the network-influence-topic interaction term in the rate equation, Eqn. (6.14) – $\sum_j a_{ij}^{t-1}\beta_j^{t-1}x_{jk}^{t-1}$ – typically increases, roughly proportionally to the out-degree of the author i. This means that the average number of publications in all topics increases for each author, and hence effectively breaks an assumption of the dPF — that the average rate does not significantly change between timesteps. In practice, we do not expect this to be an issue, as we predict that instead the average influence, $\langle\beta\rangle$, should decrease as the density of A increases. That is, if the average author is expected to both give and receive more citations than in the past, then the likelihood that a randomly chosen citation corresponds to an inspirational source should be expected to decrease, unless their rate of publication itself increases proportionally. As producing a publication takes a significantly greater amount of time than citing another work, we expect that this cannot occur beyond a certain limit, and thus $\langle\beta\rangle$ must decrease.

In the following experiments, nonetheless we proceed with using the dPF, as we found that it still demonstrates some level of ability to *rank* the likelihood of an author publishing in different topics, even if the estimated average rate is inaccurate. We demonstrate this using AUC results for each K in Table 6.3, where we use the estimated rates as a prediction score. This is more important than whether the rates perfectly reproduce the observed data, as the other parameters to be fitted – specifically α – can adjust for issues of scale. As performance under this metric is similar for each K, we chose to proceed with $K = 10$, which we found gave improved results for a wider range of confounding types and strengths compared to when using fewer factors.

Given these choices of Q and K, we may proceed to evaluate the ability of the full

K	AUC
5	0.653
8	0.634
10	0.634

Table 6.3: In this table, we present the average AUC for different numbers of topic factors in dPF, K, using the estimated rates as a prediction score. We run 20 experiments, each on a different snowball sampled subgraph of 3K authors and 1K topics, where for each we fit the dPF to the subsampled data.

procedure to recover the simulated author influence, for different choices of substitute model. As our primary accuracy metric, we present the mean squared error (MSE) between our predicted influence, and the ground truth. As described in Sec. 6.4.2, we generate potential homophily-related confounders of both topics and links (*i.e.* ζ in our model) using true information about the region to which an author is affiliated, alongside random topic-only (τ) confounders that follow a simple Markov process. This allows us to control not only the degree of confounding, but also the type — whether there is only homophily-related confounding, topic-only confounding, or both effects together. Further, recall that now we sample the influence parameter, β, according to a simple evolutionary process, and thus the assumption of a static prior – used for β in the model – is mis-specified. This means that we are not guaranteed that the oracle models will provide the best quality results, as alternatives that better capture how this evolution affects the observed data may prove more suitable when performing causal adjustment.

In Table 6.4, we provide the results of these experiments. We only compute the MSE for β up until the final timestep, as at that point all methods struggle with the issue of high rates due to the increased density of A relative to prior times, as described above — in this case our NDC variant typically performs best, but still does not demonstrate high accuracy. Additionally, for clarity we only display results here using the ADM1 region covariate, and choice of present substitutes, $\hat{\zeta}^t$, $\hat{\tau}^t$ for each t. More complete results including those using the country covariate, and the alternative substitutes, $\hat{\zeta}^{t-1}$, $\hat{\tau}^{t-1}$, may be found in App. D.

In the exogenous tests, where homophily-related confounding is absent, adjusting for τ using substitutes from the dPF – *i.e.* the 'Topic-only' model – performs best of the non-oracle choices, as we should expect. However, once homophily is included, one of the combinations of DSBMM and dPF together is typically the best option. The fact that this does not hold universally true (*i.e.* in particular not for the high homophily

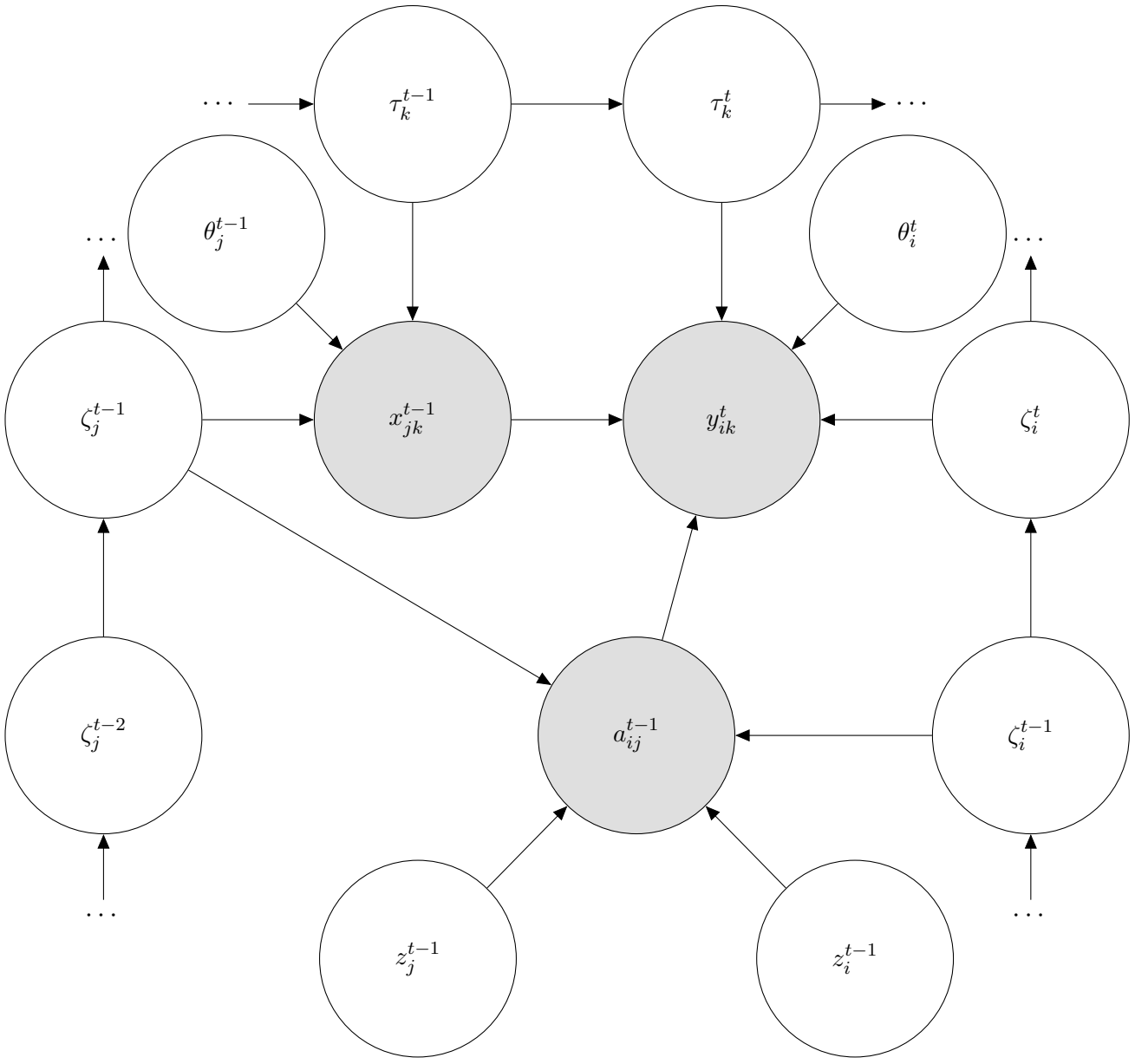


FIGURE 6.3: In this figure, we present a snapshot of our proposed causal model for understanding author influence as a DAG. Shaded variables are observed, while unshaded are latent. Descriptions of the variables are found in the main text.

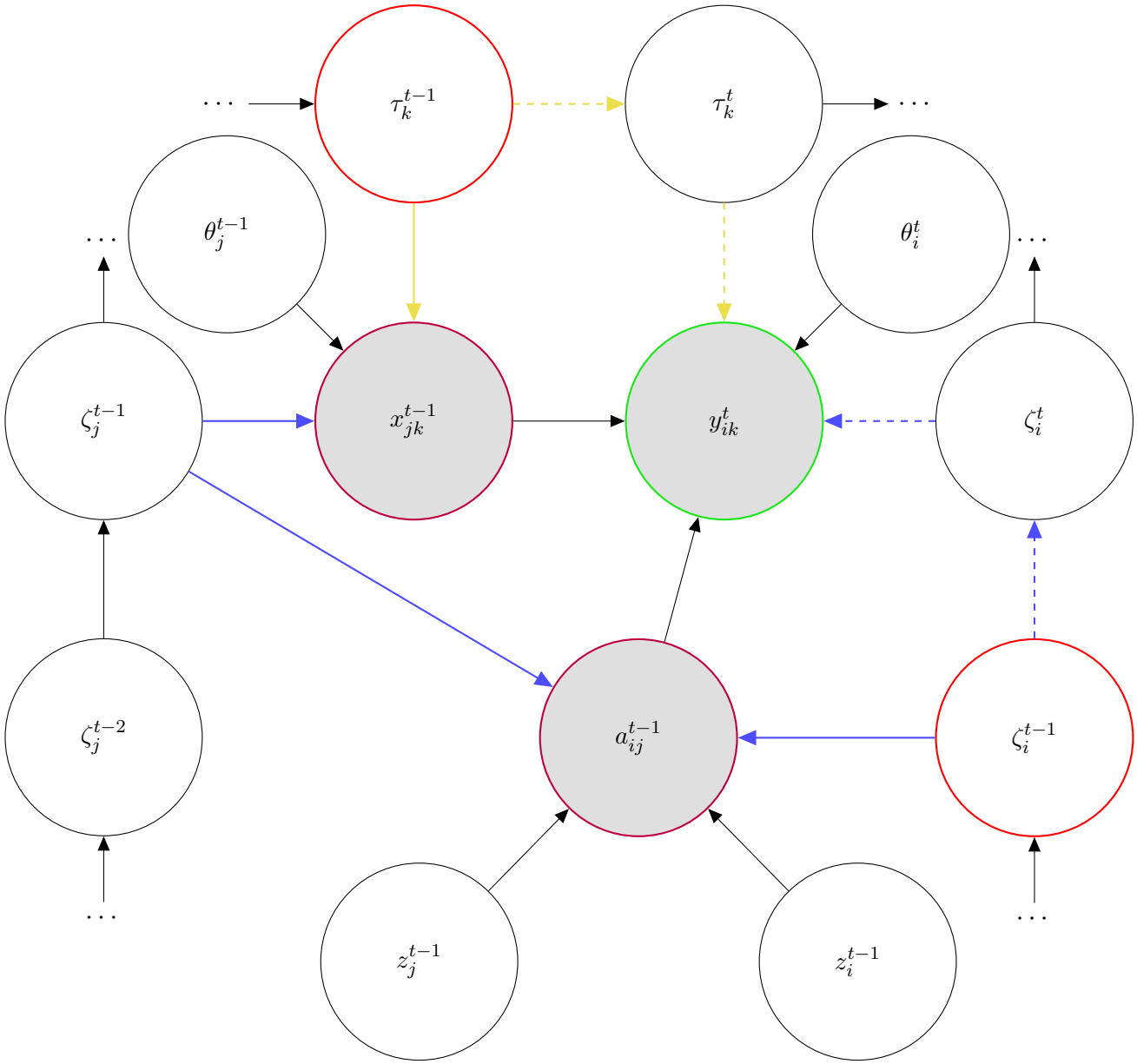


FIGURE 6.4: In this figure, we highlight back-door paths from our treatment (coloured purple) – the number of publications of an author j in topic k in the previous time period, $t-1$, given a citation between them at that time – to the outcome of interest – the number of publications of another author i in the same topic at time t (in green). In yellow we show the obvious path via the evolving topic-only attributes, τ_k, while in blue we show the path via the citation collider, a_{ij}^{t-1}, which we opened by our conditioning. Finally, in red we highlight the main elements of the viable back-door adjustment set that we estimate using substitutes, τ_k^{t-1} and ζ_i^{t-1}, which thus allow us to block the paths to the outcome, hence shown dashed.

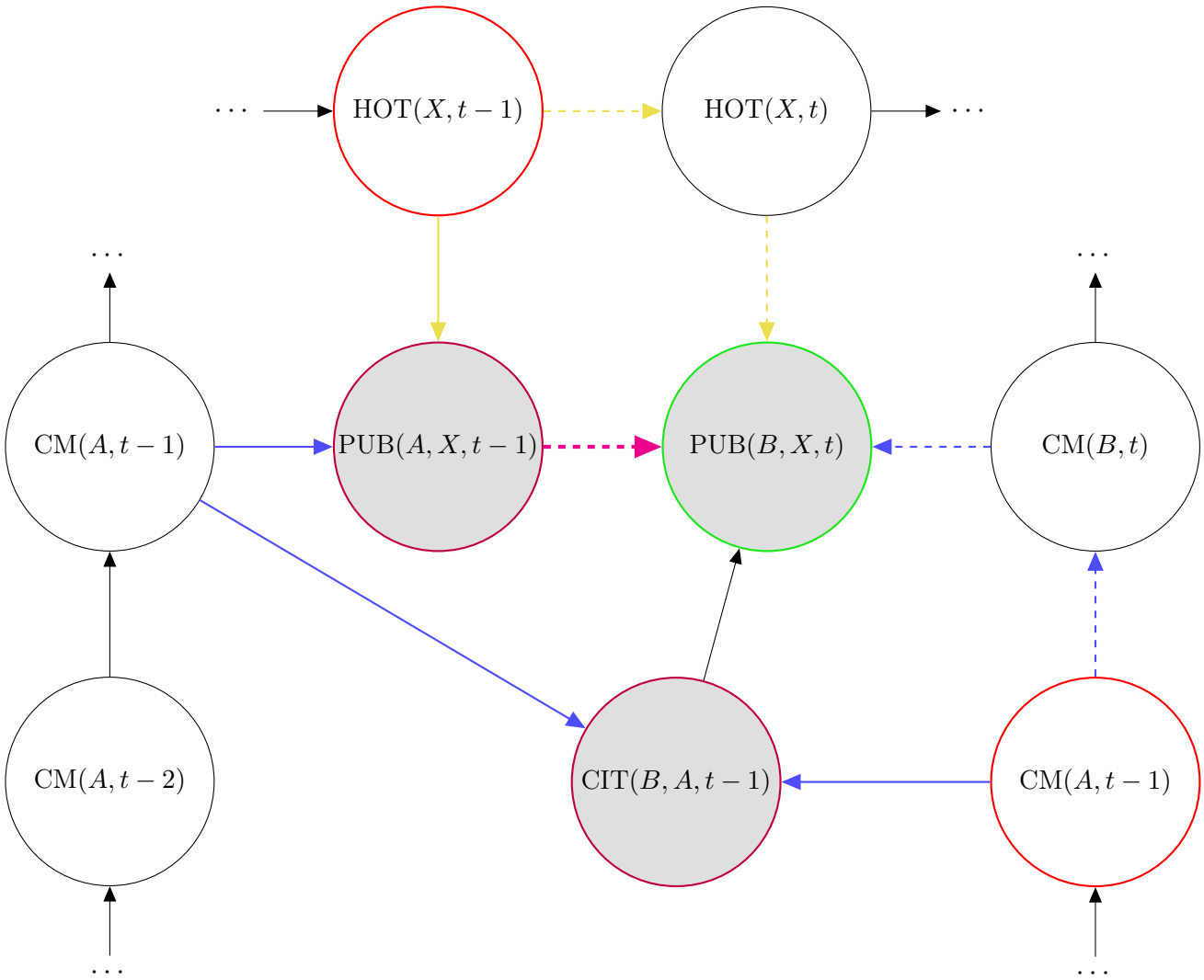


FIGURE 6.5: TMP FOR LAY REPORT, REMOVE

	Exog.			Homophily			Both		
	Low	Med.	High	Low	Med.	High	Low	Med.	High
Oracle	0.32±0.2	0.37±0.18	0.41±0.22	32.1±44.13	13.11±16.71	7.35±7.83	26.83±36.57	13.12±16.76	9.36±10.77
Topic-oracle	**0.25±0.11**	**0.31±0.15**	**0.32±0.15**	14.59±17.63	13.14±14.66	8.97±8.03	25.16±32.52	17.82±21.16	10.8±10.64
Net.-only	2.34±2.22	1.35±0.56	1.43±0.45	22.18±28.39	**7.82±2.16**	19.66±20.61	44.9±28.96	16.28±16.66	8.06±2.7
Ours-NDC	2.03±3.4	2.17±3.69	1.81±3.04	21.43±42.9	9.69±12.33	8.28±10.18	**11.68±16.73**	**9.47±12.91**	9.18±11.02
Ours (no meta)	1.6±2.94	2.43±4.93	2.12±2.99	**8.37±11.31**	11.16±16.32	8.54±9.92	12.37±15.0	9.49±10.9	9.84±11.12
Ours	1.84±2.94	2.49±4.96	2.32±3.0	17.02±28.03	9.84±10.46	8.86±10.54	12.74±14.91	10.21±11.27	10.48±11.63
Topic-only	**0.35±0.17**	**0.66±0.23**	**0.59±0.27**	10.68±14.24	7.98±8.88	**4.36±2.8**	11.71±15.68	10.76±12.8	**6.5±5.89**
Unadjusted	0.7±0.2	1.06±0.35	1.01±0.37	15.81±12.78	14.39±10.58	11.77±6.77	26.39±26.7	17.11±13.94	14.23±9.9

Table 6.4: In this table, we present the accuracy of estimated academic influence in our semi-synthetic data experiments. We evaluate three different levels of confounding (low, medium and high), and three different types of confounding — whether of topics only (Exog., *i.e.* τ only), homophily effects only (ζ only), or both effects together. Entries are the average MSE ($\times 10^3$) of estimated influence β and corresponding std. across 5 repeated simulations on different subsamples, where these simulations use the ADM1 region covariate, and we choose to use the substitutes at the current timestep, *i.e.* $\hat{\zeta}^t$, $\hat{\tau}^t$. Each simulation corresponds to results on a different, snowball sampled subgraph of 3K authors and 1K topics. Bolded entries show which variant has the lowest mean MSE, or the best non-oracle variant. In the presence of homophily confounding, the DSBMM variants typically perform best — see discussion in main text.

confounding experiment) demonstrates how the use of substitutes is not a 'free-lunch', that negates the difficulty of latent confounders: it trades weaker requirements on the adjustment set of variables at the cost of reduced information at the final step, and estimation error in the latent variable models, and these often lead to higher variance in the estimates of the causal effect than traditional causal inference approaches. For instance, in these experiments, 'Ours (no meta)' provides the lowest *median* MSE across all homophily-only experiments, while 'Ours-NDC' provides the lowest median MSE in the presence of both types of confounding, but these models both have high variance — especially poor performance in some simulations skewed their average MSE, as shown. The importance of the degree to the author's influence – which we would also expect to hold true to some extent in real data – is clear, as the NDC version almost always outperforms the DC alternative.

Importantly, in the presence of homophily, the 'oracle' models are always outperformed by the other choices, that better account for the evolution of the data — as previously suggested, this is due to mis-specification of the model now that β is assumed to evolve over time. This provides an interesting insight into how sufficiently strong substitutes may actually improve mis-specified causal inference procedures, by capturing some of the dependencies that were neglected.

6.6.2 Estimating real author influence

Finally, we conclude with an explorative application of this procedure to the real topic data — thus producing estimates of which authors are truly the most influential. We apply the principal models previously described.

As we no longer have any ground-truth to appeal to, we cannot use MSE to understand the performance of each model. Instead, we evaluate the model in a transductive setting, where we mask all publications and citations for a random 20% of authors present in the final period. We then fit each model on the remaining data, before using the held-out data to calculate two performance metrics: the log-likelihood of observing the held-out data under the Poisson regression model, given the most recently available substitutes fitted — denoted HOL below; and the AUC, where we use the estimated rates as a score by which to predict whether an author will publish in a particular topic or not. The logic is that if a method performs well at generalising to unseen data, then it likely better captures the true underlying causal process. We present our results under these metrics, as well as the average author influence inferred at each timestep, in Table 6.5.

	HOL	AUC	$\langle\beta^1\rangle$	$\langle\beta^2\rangle$	$\langle\beta^3\rangle$	$\langle\beta^4\rangle$
Unadjusted	-16.91	0.888	0.08	0.059	0.045	0.03
Network-Only	-16.76	0.888	0.071	0.05	0.038	0.027
Topic-Only	-16.93	0.888	0.089	0.069	0.049	0.031
Ours	-20.29	0.927	0.071	0.051	0.038	0.027
Ours-NDC	**-12.33**	**0.934**	0.074	0.053	0.039	0.027
Ours (no meta)	-20.28	0.93	0.072	0.052	0.039	0.027

Table 6.5: In this table, we present results after applying the procedure to the real publication data. For each method, we show the average influence across all authors at each timestep, $\langle\beta^t\rangle$, and two measures on held-out test data: (i) the average held-out Poisson log likelihood (HOL; higher is better), and (ii) area under the ROC curve (AUC; higher is better) obtained using the estimated rate as a prediction score. Our NDC method provides the best performance (bold) under both measures.

Firstly, we note that as the density of the citation networks increase, so too does the average influence of an author — as predicted in the previous section. All models produce predict average influence in similar ranges, with the unadjusted procedure typically over-estimating the value — a similar finding to PIF. However, the topic-only model, adjusting only for τ, also appears to over-estimate author influence, while both this and the network-only model have lower AUC than combined methods. This suggests that while both types of confounding are important, homophily-related confounding is likely dominant to some degree. In App. D, we provide additional results for the average influence contribution to the rate at each timestep, and demonstrate that in fact this changes at a slower rate than the average influence — *i.e.* authors are 'inspired' to a similar degree over time, but by a wider variety of others.

With respect to both quality measures, the NDC-DSBMM in conjunction with dPF provides the best performance. Interestingly, the other two (DC) combined methods have similar AUC scores, despite considerably worse HOL — this suggests that much as for dPF on synthetic topics, the methods can rank the likelihood of events effectively, but the estimated rates have the incorrect scale. We note that the AUC performance is close to the state-of-the-art for predicting future publications by topic [161], though we do so at the author- rather than publication-level. We believe that this provides considerably greater potential utility to the method — rather than simply estimating the likelihood of a particular topic occurring in future, we can also predict the most likely authors to be involved. We return to this in the subsequent discussion section.

6.7 Discussion and further work

In this chapter, we have described how to extend the recent ideas of Poisson influence factorisation (PIF) [166] to estimate the influence of each author in a large dataset of Latin American authors, by constructing a dynamic citation network, and considering the publication topics of each author at each timestep. To our knowledge, this is the first time author influence has been considered in an explicitly causal framework. Furthermore, by introducing a mild mis-specification into the proposed causal model, we have demonstrated that not only can substitutes be used effectively in such dynamic problems, but additionally have the potential to ameliorate model mis-specification to some degree.

To construct the dynamic substitutes, we used separate latent variable models for the joint topic-citation, and topic-only confounders: an extension of the DSBMM that allows for directed networks, and a recent popular collaborative filtering model, dynamic Poisson factorisation (dPF, [26]) respectively. To determine suitable numbers of groups, Q, and factors, K, in these models, we performed posterior predictive checks, and found reasonable representative capabilities for both models across the range of values considered, in particular for the DSBMM.

We evaluated the accuracy of our procedure for estimating influence on semi-synthetic datasets, and additionally explored real academic influence, using alternative metrics in both cases to compare different substitute models. In both cases, we found that often the combination of the directed NDC-DSBMM with dPF provided one of – if not the best – sets of substitutes for adjustment. This demonstrates that clustering by degree may be beneficial if the subsequent outcome of interest for each node has some direct relation to the degrees, even if weak, as author influence does in the case of the citation network.

There are numerous pathways we believe the work of this chapter opens for future research, some of which include:

- Investigating what disparity between the estimated influence of an author and more conventional measures of impact/influence *e.g.* citations might mean — for instance, if the marginal impact of a paper greatly increase the influence of an author in a topic that is growing, might it suggest a 'sleeping beauty' that will be very popular in future years [176]?

- The high AUC score when using the estimated rates to predict future publication topics of an author suggests a wealth of possible applications. For instance, this

provides a means to estimate the 'surprise', or novelty, of a paper within the context of both wider trends in the community, and the author's specific prior work — similar procedures have recently been explored with some success, but not at the author level [161].

- As described above, assuming an independent Gamma prior for the influence parameter, β, over all time is not necessarily realistic. In addition to incorporating temporal dependence, another desirable feature in the particular context of estimating author influence would be to account for the number of citations they receive, or perhaps for an alternative citation-based measure such as FWCI [149]. This is particularly the case for generating synthetic topics, as we demonstrated that even with temporal dependence, if the density of (or average degree in) the network is not suitably accounted for, the contribution from the influence terms can quickly grow to become unrealistically large. We expect that in practice, to some degree – as found above for temporal dependence – models such as the DSBMM that better capture the observed data may ameliorate this issue, even without explicitly accounting for it, though of course doing so would be preferable.

- Rather than using pre-computed topics, we may wish to directly incorporate the discovery of such from the words of each paper themselves. There is some related work in the literature, for instance [177, 30].

- How do these substitute methods compare to alternative contemporary methods? For instance, one promising avenue could be to use the signature kernel trick to construct dynamic graph kernels, in conjunction with contemporary methods that leverage kernels for causal inference [162]. We recently explored dynamic graph kernels constructed in such a way to some success [40], so believe this warrants further investigation.

- Finally, the codebase used to perform the procedure could also be further improved. For instance, one such pathway for accelerating code would be to impose sparse priors on all parameters, and enforce this *e.g.* through thresholding, so each node assumed to belong to only one of a subset of all groups — this suggests that any initialisation method that can provide such would be particularly useful. The simplest way to do so that we believe may still be effective would be to *e.g.* perform some initial clustering procedure, then perform label propagation on the results — collect the set of labels within a given

neighbourhood of each node, and only permit the node to belong to these groups.

We provide details for several other possible extensions in App. D.

Appendix A

Paths for further developments of the DSBMM

In this appendix, we briefly elaborate several possible variants of the DSBMM to develop further functionality, or pathways towards such that may be worth exploring.

A.1 Additional metadata distributions

Negative binomial (for positive integers, with overdispersion) An alternative choice to the Poisson disrtibution for modelling career ages, that may be more suitable in general for over-dispersed data, is the more complex negative binomial distribution. This has the probability mass function

$$f(k; r, p) \equiv \mathrm{p}(X = k) = \frac{\Gamma(k + r)}{k!\Gamma(r)}(1 - p)^k p^r \quad \text{for } k = 0, 1, 2, \ldots \tag{A.1}$$

for r a real, positive number, where $\Gamma(x) = \int_0^\infty y^{x-1} e^{-x}\,\mathrm{d}y$ is the Gamma function. Denoting the within-group parameters at time t by p_q^t and r_q^t respectively, we find

$$p_q^t = \frac{\xi_q^t r_q^t}{\xi_q^t r_q^t + \zeta_q^t}, \tag{A.2}$$

which we can substitute into the equation for r_q^t,

$$0 = \sum_i \tau_{tiq}^m (\psi(x_i^t + r_q^t) - \psi(r_q^t) + \log p_q^t), \tag{A.3}$$

with $\psi(\cdot) = \frac{\Gamma'(\cdot)}{\Gamma(\cdot)}$ the digamma function (the derivative of the log of the gamma function), and solve for r_q^t *e.g.* using Newton's method.

Multivariate Gaussian (for real vectors) If instead we want to consider generic continuous metadata, which are real vectors, we might decide to turn to the multivariate Gaussian. For k dimensions this has the probability density function given by

$$\mathrm{p}(\boldsymbol{x}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k \mid \boldsymbol{\Sigma} \mid}}, \tag{A.4}$$

where $\boldsymbol{\mu}$ is the mean, and $\boldsymbol{\Sigma}$ is the (positive definite) covariance matrix. Once more neglecting the details, we find

$$\boldsymbol{\mu}_q^t = \frac{\rho_{q\ell}^t}{\xi_q^t}, \qquad \boldsymbol{\Sigma} = \frac{1}{\xi_q^t}\sum_i \tau_{tiq}^{\mathrm{m}}(\boldsymbol{x}_i - \boldsymbol{\mu}_q^t)(\boldsymbol{x}_i - \boldsymbol{\mu}_q^t)^{\top}. \tag{A.5}$$

However, while the latter of these is correct within our system of equations, by analogy with weighted sampling, this would provide a biased estimate of the covariance matrix – to correct if desired we may further divide by

$$1 - \frac{\sum_i \left(\tau_{tiq}^{\mathrm{m}}\right)^2}{\left(\xi_q^t\right)^2}. \tag{A.6}$$

Note that if we had $\tau_{tiq}^{\mathrm{m}} = 1/N$ for all i, this would reduce to the familiar $(N-1)/N$ correction factor.

Of course, allowing for general covariance greatly increases the complexity of the problem, as we would need to perform $\mathcal{O}(N^2)$ calculations every iteration. Instead, we may wish to assume that either (i) the probability is simply the product of independent normal distributions, each with their own variance, or (ii) likewise the probability is the product of independent normal distributions, but they now share a single variance parameter. Both cases immediately follow from the equation above – for (i) we have

$$(\sigma_{q\ell}^t)^2 = \frac{\sum_i \tau_{tiq}^{\mathrm{m}}(x_{i\ell}^t - \mu_{q\ell}^t)^2}{\xi_q^t}, \tag{A.7}$$

while for (ii) the numerator is simply summed over ℓ for $(\sigma_q^t)^2$. In both cases, to correct for bias we should use the same inverse factor as in Eqn. (A.6).

As for efficiency we might not want to loop over all data again after finding $\mu_{q\ell}^t$ in order to calculate the variance, for these latter cases we introduce the new variable

$$\psi_{q\ell}^t = \sum_i \tau_{tiq}^{\mathrm{m}}(x_{i\ell}^t)^2, \tag{A.8}$$

so that we now have

$$(\sigma_{q\ell}^t)^2 = \frac{\psi_{q\ell}^t - 2\mu_{q\ell}^t \rho_{q\ell}^t + \xi_q^t (\mu_{q\ell}^t)^2}{\xi_q^t}. \tag{A.9}$$

We note that we can use these equations to infer log-normal distributions over metadata, by applying the above equations for the logarithm of the metadata. The only additional factor is the resulting scaling from the transformation on the metadata probability density function, *i.e.* we must now use

$$\mathrm{p}(x \mid z) = \mathrm{p}(\log x \mid z) \frac{\mathrm{d}(\log x)}{\mathrm{d}x} = \mathrm{p}(\log x \mid z) \frac{1}{x} \tag{A.10}$$

when calculating our ICL, and in the update equations for τ. In general, we could of course experiment with other transformations of metadata onto these base distributions above provided we scale the density suitably.

A.2 Constraining the model, for scalable MFVI

As previously discussed, one of the problems with our model as currently formulated is a somewhat prohibitive computational complexity, scaling as $\mathcal{O}(TN^2Q^2)$. One natural way to reduce this complexity is to simplify our model, say by constraining it. In particular, we make two modifications:

(i) Rather than consider general group connectivities ϕ_{ql}, we can limit ourselves to allowing unique distributions governing connectivity within each group, ϕ_{qq}, and a single further distribution for all inter-group connectivity (*i.e.* $q \neq l$), ϕ_{out}.

(ii) Instead of allowing general transitions between each pair of groups, $\pi_{qq'}$, we can consider restricting each group to its own persistence probability, π_{qq}, with a uniformly random choice among the remaining groups otherwise such that $\pi_{qq'} = (1 - \pi_{qq})/(Q - 1)$ for $q \neq q'$.

With these constraints, we reduce the number of connectivity parameters we need to infer at each timestep to $Q + 1$, and the number of transition parameters to Q. Somewhat similar constraints have been performed in the literature, for instance in [92] for a very different model and inference procedure.

Further, this new formulation suggests an alternative variational family $\mathbb{Q}_\tau(Z)$,

where

$$\mathbb{Q}_\tau(Z) = \prod_i \prod_q \tau(i,q)^{Z_{iq}^1} \times \prod_{t=2} \prod_q \tau(t,i,q,q)^{Z_{iq}^{t-1} Z_{iq}^t} \prod_{q' \neq q} \left(\frac{1 - \tau(t,i,q,q)}{Q-1} \right)^{Z_{iq}^{t-1} Z_{iq'}^t}. \tag{A.11}$$

Note that in this family, by construction we immediately have that $\sum_q' \tau(t,i,q,q') = 1$. When performing variational inference, the equations previously derived for parameters other than τ and π remain unchanged, while for π we now have

$$\pi_{qq} = \frac{\sum_{t=2} \sum_i \tau_{t-1,i,q}^m \tau_{tiqq}}{\sum_{t=2} \sum_i \tau_{t-1,i,q}^m}, \tag{A.12}$$

where previously we would have only had the numerator, and the fixed point equation for τ_{tiqq} is now

$$\begin{aligned} \tau_{tiqq} &= \frac{f_{tiqq}}{\langle f_{tiq\cdot} \rangle_q + f_{tiqq}}, \\ \text{where} \quad f_{tiqq'} &= p_{q'}^t(x_i^t) \pi_{qq'} \prod_j \phi_{q'q'}(A_{ij}^t)^{\tau_{tjq'}^m}, \end{aligned} \tag{A.13}$$

and we use $\langle x. \rangle_q$ to denote the geometric mean of some set of values $x_{q'}$, that depend on the group labels $q' \in \{1, \ldots, Q\}$, over $q' \neq q$, *e.g.*

$$\langle \pi_{q\cdot} \rangle_q = \left(\prod_{q' \neq q} \frac{1 - \pi_q}{Q-1} \right)^{\frac{1}{Q-1}} = \frac{1 - \pi_q}{Q-1}. \tag{A.14}$$

With these equations in hand, we have reduced the number of parameters to update from $\mathcal{O}(TN^2Q^2)$ to $\mathcal{O}(TNQ)$, and further require fewer computations for each τ_{tiqq} value then before – hence we should expect a corresponding significant speedup to be observable. The necessary modifications for our model selection criterion for this case are immediate, given the reduced number of parameters – note the resulting reduced penalisation will likely increase the 'optimal' number of groups inferred.

A.3 Group-wise tuning parameters

As an extension of the current tuning parameter, at one further level of complexity, we could allow each group q to have their own probability of using data or not, θ_q. As this would require conditioning first on the group labels before drawing the new metadata 'switch' latent variables, with regards to the MFVI approach of Chap. 3,

the metadata term in the variational objective function would instead become

$$\sum_{t,i,q} \tau^m_{tiq} \left[\theta_q \left(\log \theta_q + \log \mathrm{p}^t_q(x^t_i)\right) + (1-\theta_q)(\log(1-\theta_q) + \log \mathrm{p}(x^t_i))\right]. \tag{A.15}$$

Neglecting contributions from other groups, we would then find the corresponding boundary cases

$$f(\theta) = \begin{cases} \sum_{t,i} \tau^m_{tiq} \log \mathrm{p}(x^t_i) & \text{when } \theta_q = 0, \\ \sum_{t,i} \tau^m_{tiq} \log \mathrm{p}^t_q(x^t_i) & \text{when } \theta_q = 1, \end{cases} \tag{A.16}$$

and so the requirement for using metadata locally is now that specifically the distribution for that group is an improvement upon the average expected. Similarly however, we could manually tune the importance of metadata for each group — useful if we expected in advance that the saliency of metadata would vary across the network, though label permutation issues would have to be addressed in this case, *i.e.* we would have to restrict the permitted labels in such areas of the network.

Once again highlighting the change in blue, the new update equation for $\tau_{tiqq'}$ would now be

$$\forall t \geq 2, \forall i \geq 1, \forall q, q' \in \mathcal{Q}, \quad \hat{\tau}\left(t, i, q, q'\right) \propto \pi_{qq'} [\theta_{q'} p^t_{q'}(X^t_i)]^{\theta'_q} [(1-\theta_{q'})\mathrm{p}(X^t_i)]^{(1-\theta_{q'})} \times \prod_{j=1}^{N} \prod_{l'=1}^{Q} \left[\phi^t_{q'l'}\left(A^t_{ij}\right)\right]^{\hat{\tau}_{\mathrm{marg}}(t,j,l')}. \tag{A.17}$$

We also note that the single parameter version of θ enters the update equation in a way that is reminiscent of simulated annealing type schemes — it may be the case that varying θ in some way over the course of inference could improve results for some datasets.

A.4 Missing data

In real data, it is often the case that various items – metadata or edges – might be unobserved or unreliable. As such, it is important to understand how we might handle these cases, and indeed infer likely values given our model. In this section, we prioritise describing how to impute missing metadata (also known as collaborative filtering), and edges for missing nodes, and only touch briefly upon how to handle more general missing data.

A.4.1 Missing metadata

Given the model assumptions, inferring missing metadata is simple. First, we neglect terms in the model likelihood for missing metadata, which is reasonable providing the data is missing at random (MAR) [168], and fit our model to the observations. Next, thus given our estimate of the most likely group of a node with missing metadata, most trivially we can impute using the mean/mode of the corresponding group. Alternatively, we can use our estimated node marginals, τ^m_{tiq} (or ψ^{it}_q in the BP case), to marginalise the labels and take the mean/mode instead of

$$\begin{aligned} \mathrm{p}(x_i^t) &= \sum_q \mathrm{p}(x_i^t \mid z_{iq}^t)\mathrm{p}(z_{iq}^t), \\ &\approx \sum_q \mathrm{p}(x_i^t \mid z_{iq}^t)\tau^m_{tiq}. \end{aligned} \tag{A.18}$$

A.4.2 Missing nodes

Following the same process, we can infer all edges incident upon a new node that enters the system simultaneously. Without metadata, at best we would have to solely be guided by the distribution of group sizes, but metadata can help us place the node, much as in [62].

That is, for a new node i entering the system at time t, denoting the edges incident by the vector $\boldsymbol{a}_i^t$, we have

$$\mathrm{p}(\boldsymbol{a}_i^t \mid A_{-i,t}, Z_{-i,t}, x_i^t) = \sum_q \mathrm{p}(\boldsymbol{a}_i^t \mid Z_{-i}^t, z_{iq}^t)\mathrm{p}(z_{iq}^t \mid x_i^t), \tag{A.19}$$

where $A_{-i,t}$ refers to all observed network data excluding that for node i at time t, and similar for Z, and

$$\mathrm{p}(z_{iq}^t \mid x_i^t) = \frac{\mathrm{p}(x_i^t \mid z_{iq}^t)\mathrm{p}(z_{iq}^t)}{\sum_q \mathrm{p}(x_i^t \mid z_{iq}^t)\mathrm{p}(z_{iq}^t)}. \tag{A.20}$$

We can then quantify the relative predictive improvement provided through using the metadata through the predictive likelihood ratio $\lambda_i^t \in [0, 1]$, with

$$\lambda_i^t = \frac{\mathrm{p}(\boldsymbol{a}_i^t \mid A_{-i,t}, Z_{-i,t}, x_i^t)}{\mathrm{p}(\boldsymbol{a}_i^t \mid A_{-i,t}, Z_{-i,t}, x_i^t) + \mathrm{p}(\boldsymbol{a}_i^t \mid A_{-i,t}, Z_{-i,t})}. \tag{A.21}$$

Note that for the term without metadata, we can only use some estimate of $\mathrm{p}(z_i^t)$ (*i.e.* α_q) rather than $\mathrm{p}(z_i^t \mid x_i^t)$. This relative predictive improvement provides us one

method to evaluate the importance of different pieces of metadata to latent labels, and hence structure in the network – both globally (*i.e.* averaged across all tested nodes), and locally for specific nodes.

This also suggests that a two-step approach could be explored for incorporating nodes that join the system after the initial timestep, rather than the current approach, as in [98], of assuming that these nodes have the same universal prior for their group, α. However, this would increase inference time, without any guarantee of improved results — indeed, given this would require initially discarding all nodes that join at a later stage, and their associated edges, it may be detrimental overall. Instead, explicitly modelling

A.4.3 Missing data that is *not* missing at random

Throughout this work, we have frequently restricted our focus by reducing a larger network down to something more manageable, or of particular interest. However, doing so has resulted in a considerable amount of data that is now missing from our model when performing inference, in a manner that is clearly *not* at random. This missing data could have a strong impact on the results should it be included, for instance by the introduction of nodes and edges that connect two previously separate components together, or simply by changing the apparent functional roles of previously observed nodes.

A recent seminal work explored how one might address this problem to some degree when performing inference in SBMs [168]. Combining their mean-field variational inference procedure with that of Chap. 3 to account for dynamics and metadata would be one approach, though it is also possible to extend our belief propagation method. However, in doing so we would need to account for new latent variables, corresponding to the unobserved edges (or absence thereof), and introduce new factor nodes involving the latent groups of *e.g.* i, j, and the unobserved potential edge between them, a_{ij}. As such, the reduction to the direct node-node system as we perform, that relied on only pairwise or single factors, is no longer possible. Nonetheless, the message-passing system emerging over the factors should still have additional benefits, and likely improved performance compared to the mean-field approach, and we believe is worth pursuing in future.

A.5 Considering dynamic hypergraphs with metadata

One of the most popular topics in network science at the moment is hypergraphs — in brief, an extension of graphs that allows (hyper)edges to connect more than two nodes at once. These are a much more natural choice to model *e.g.* co-authorship networks, where a single hyperedge can link all authors of a publication, rather than projecting this into many dyadic connections as we perform in this work.

With regards to the SBM, a variety of authors have proposed different extensions that allow application to hypergraphs. For instance, the work of [9] posits a spectral method inspired by a simplified belief propagation system for a hypergraph version of the toy model. This only applies to uniform hypergraphs, but allows them to conjecture the existence of sharp detectability thresholds that were later proved rigorously in [122]. However, even given their simplifications, two clear scaling problems are present — as for all models that follow along similar lines:

(i) Calculating the initial external field term for non-edges involves $\mathcal{O}(N^{k-1}Q^k$ terms, where k is the (uniform) hyper-edge degree (*i.e.* the number of nodes each edge connects), thus very quickly becoming infeasible for large networks, or even small hyper-edge degree. This first problem could be circumvented by initialising messages in a known way, *e.g.* all uniform — which is not a problem in the case with metadata, nor if we are performing parameter updating.

(ii) A greater issue is that each message involves a sum over Q^{k-1} choices for labels of nodes present in adjacent hyperedges, within each of which there is a product of $k-1$ terms — messages into the hyperedge from each label. As such, if the average node participates in d hyperedges, performing the full message update for a single node – for all possible groups, so they may be properly normalised – involves $\mathcal{O}((d-1)Q^k(k-1)$ terms. Hence a full iteration for the network takes $\mathcal{O}(NdQ^kk)$ time, which grows rapidly if many groups are sought. Allowing Q^k parameters governing inter-group connections also greatly increases the risk of overfitting, so the most immediate options are to either (i) greatly reduce the number of groups sought (*e.g.* without reducing the overall number by applying the method recursively as previously), or (ii) impose constraints on the block connectivity parameters – as is most commonly performed.

As such, appealing to belief propagation to accelerate inference is not a fruitful path to follow.

While there are numerous other options, *e.g.* the stochastic gradient descent for maximum-likelihood inference used to fit effectively a simple dynamic mixed-membership SBM [161], viewing the hypergraph as bipartite node-hyperedge networks as in [117], or more specific simplifications that permit inference [73, 72, 122, 39], we believe that the best immediate approach for the DSBMM is a greedy one.

Specifically, the recent work of [28] propose a greedy procedure to find groups from a simple hypergraph SBM, in a manner analogous to the classic Louvain method [17] — and thus suitably described as 'hypergraph Louvain' or modularity. The key step is the reduction of the model to a form in which the optimal local updates at each node become obvious — as this has already been performed for static hypergraphs, the extension to the dynamic case with metadata is nearly immediate. That is, we may effectively use a lightly modified version of the greedy method proposed in Sec. 5.2, with identical metadata-related contributions. It would be interesting to see how this simple modification may further improve performance in this case, as some of the downsides of the necessary constraints on block parameters – specifically reduced expressivity in the groups, and imposition of group assortativity – may be ameliorated by allowing (and accounting for) more generic metadata contributions.

Appendix B

Detailed derivation of the key eigenvalues for determining efficient detectability

In this appendix we provide full details for the derivation of the eigenvalues necessary to determine the efficient detectability threshold, as described in Chap. 4. Specifically, recall that efficient detectability for a classic SBM is possible if

$$\mathrm{SNR} = \lambda_2^2/\lambda_1 > 1, \tag{B.1}$$

where λ_1 and λ_2 are the largest and second largest distinct eigenvalues of PQ respectively, where P is a diagonal matrix corresponding to the prior over groups, and Q is the block connectivity matrix [1].

In the case of our static toy model, without loss of generality we may assume that $q \in q_B$ make up the first Q_b groups, such that

$$PQ = \frac{N}{Q_b}\begin{pmatrix} \rho p_{\mathrm{in}} & \rho p_{\mathrm{out}} & \cdots & & \\ \rho p_{\mathrm{out}} & \rho p_{\mathrm{in}} & \cdots & & \\ \vdots & & \ddots & & \vdots \\ \frac{(1-\rho)}{(Q_B-1)}p_{\mathrm{out}} & \cdots & & \frac{(1-\rho)}{(Q_B-1)}p_{\mathrm{in}} & \cdots \\ \vdots & & & \ddots & \end{pmatrix}, \tag{B.2}$$

i.e. with $\tilde{Q}_{Q_b} = (p_{\mathrm{in}} - p_{\mathrm{out}})I^{Q_b \times Q_b} + p_{\mathrm{out}} J_{Q_b}$ the block matrix restricted to Q_b dimensions,

where J_{Q_b} is the ones matrix in $Q_b \times Q_b$ dimensions, this is

$$PQ = \frac{N}{Q_b} \begin{pmatrix} \rho\tilde{Q}_{Q_b} & \rho p_{\text{out}} J_{Q_b} & \cdots \\ \frac{(1-\rho)}{(Q_B-1)} p_{\text{out}} J_{Q_b} & \frac{(1-\rho)}{(Q_B-1)} \tilde{Q}_{Q_b} \cdots & \\ \vdots & \ddots & \end{pmatrix}, \tag{B.3}$$

i.e. a block matrix of $Q_B \times Q_B$ blocks, each of size $Q_b \times Q_b$, with the first row of blocks multiplied by ρ, and the others a reordering of this row but multiplied by $(1-\rho)/(Q_B-1)$ instead.

Note that $\tilde{Q} = (p_{\text{in}} - p_{\text{out}})I + p_{\text{out}}J$, so this has a particularly simple structure. Indeed, when seeking eigenvalues, we can use that for a block matrix with invertible diagonal blocks,

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(A)\det(D - CA^{-1}B). \tag{B.4}$$

By definition, eigenvalues for a matrix M are roots of the characteristic polynomial, $\det(M - \lambda I) = 0$. Thus in our case, neglecting the constant factor N/Q_b which will just scale each eigenvalue (thus providing a N/Q_b term to our SNR threshold), we can construct a suitable block matrix using

$$A = \rho\tilde{Q}_b \in Q_b \times Q_b, \tag{B.5}$$

$$B = \rho p_{\text{out}}(J_{Q_b}, ..., J_{Q_b}) \in Q_b \times Q_b(Q_B - 1), \tag{B.6}$$

$$C = \frac{(1-\rho)}{(Q_B-1)} p_{\text{out}}(J_{Q_b}, ..., J_{Q_b})^\top \in Q_b(Q_B - 1) \times Q_b, \tag{B.7}$$

$$D = \frac{(1-\rho)}{(Q_B-1)} \tilde{Q}_{Q_b(Q_B-1)} \in Q_b(Q_B - 1) \times Q_b(Q_B - 1), \tag{B.8}$$

then calculate the characteristic polynomial using

$$\det(PQ - \lambda I) = \det(A - \lambda I)\det(D - \lambda I - C(A - \lambda I)^{-1}B), \tag{B.9}$$

so long as $A - \lambda I$ is invertible. The first term provides factors $(\lambda - N\rho(p_{\text{in}} + (Q_b - 1)p_{\text{out}})/Q_b)$ with multiplicity one, and $(\lambda - N\rho(p_{\text{in}} - p_{\text{out}})/Q_b)$ with multiplicity $(Q_b - 1)$, while the contribution of the latter term $\det(D - \lambda I - C(A - \lambda I)^{-1}B)$ is less immediately obvious.

This block formulation makes the formal degeneracy of the problem with respect to ρ immediately clear: *i.e.* key requirements do not hold if

1. $\rho = 0$, in which case $A, B \equiv 0$, so A is not invertible;

2. $\rho = 1$, in which case $C, D \equiv 0$, so D is not invertible;

logically as there is zero probability of these groups occurring in the corresponding subgraph. In the latter case we expect the results of [133] to hold, while in the former a simple modification of the results therein follows, as instead the metadata would tell you which groups were *not* present, thus reducing the number of groups to detect in the subgraph to $Q_b(Q_B - 1)$. The problem is of course also degenerate if

3. $\rho = 1/Q_B$, in which case $A = D$ (up to a change in dimensions, *i.e.* importantly in eigenspectrum ignoring multiplicities), $B = C^\top$, and so $(A - \lambda I)$ and $(D - \lambda I)$ are not simultaneously invertible

and we have a classic SBM with corresponding (weak) detectability threshold.

Now assuming $0 < \rho < 1$, with $\rho \neq 1/Q_B$, by the Sherman-Morrison formula, that if $M' = M + uv^\top$ is a rank-one update of the invertible matrix M,

$$M'^{-1} = M^{-1} - \frac{M^{-1}uv^\top M^{-1}}{1 + v^\top M^{-1}u}, \tag{B.10}$$

– we may use that $A - \lambda I = (\rho(p_{\text{in}} - p_{\text{out}}) - \lambda)I_{Q_b} + \rho p_{\text{out}} J_{Q_b}$ is simply a rank-one update of a multiple of the identity matrix. Hence, we have that

$$\begin{aligned}(A - \lambda I)^{-1} &= 1/(\rho(p_{\text{in}} - p_{\text{out}}) - \lambda)I - \frac{\rho p_{\text{out}} J}{(\rho(p_{\text{in}} - p_{\text{out}}) - \lambda)^2(1 + Q_b \rho p_{\text{out}}/(\rho(p_{\text{in}} - p_{\text{out}}) - \lambda))}, \\ &= 1/(\rho(p_{\text{in}} - p_{\text{out}}) - \lambda)I - \frac{\rho p_{\text{out}} J}{(\rho(p_{\text{in}} - p_{\text{out}}) - \lambda)(\rho(p_{\text{in}} + (Q_b - 1)p_{\text{out}}) - \lambda)},\end{aligned} \tag{B.11}$$

and so

$$\begin{aligned}C(A - \lambda I)^{-1} &= \frac{(1-\rho)}{(Q_B - 1)} p_{\text{out}} \left(\frac{(\rho(p_{\text{in}} + (Q_b - 1)p_{\text{out}}) - \lambda) - Q_b \rho p_{\text{out}}}{(\rho(p_{\text{in}} - p_{\text{out}}) - \lambda)(\rho(p_{\text{in}} + (Q_b - 1)p_{\text{out}}) - \lambda)} \right) J, \\ &= \frac{(1-\rho)}{(Q_B - 1)} \frac{p_{\text{out}}}{(\rho(p_{\text{in}} + (Q_b - 1)p_{\text{out}}) - \lambda)} J, \\ \therefore C(A - \lambda I)^{-1} B &= \frac{\rho(1-\rho)}{(Q_B - 1)} \frac{p_{\text{out}}^2 Q_b}{(\rho(p_{\text{in}} + (Q_b - 1)p_{\text{out}}) - \lambda)} J,\end{aligned} \tag{B.12}$$

thus

$$\begin{aligned} D - \lambda I - C(A - \lambda I)^{-1}B &= \left(\frac{(1-\rho)}{(Q_B - 1)}(p_{\text{in}} - p_{\text{out}}) - \lambda \right) I \\ &\quad + \frac{(1-\rho)}{(Q_B-1)} p_{\text{out}} \left(1 - \frac{p_{\text{out}} \rho Q_b}{(\rho(p_{\text{in}} + (Q_b - 1)p_{\text{out}}) - \lambda)} \right) J, \\ &= \left(\frac{(1-\rho)}{(Q_B - 1)}(p_{\text{in}} - p_{\text{out}}) - \lambda \right) I \\ &\quad + \frac{(1-\rho)}{(Q_B-1)} p_{\text{out}} \left(\frac{\rho(p_{\text{in}} - p_{\text{out}}) - \lambda}{\rho(p_{\text{in}} + (Q_b - 1)p_{\text{out}}) - \lambda} \right) J. \end{aligned} \tag{B.13}$$

Now we can use the matrix determinant lemma, that the determinant of a rank-one update of a matrix M is given by

$$\det(M + uv^\top) = (1 + v^\top M^{-1} u)\det(M), \tag{B.14}$$

with these matrices to find the characteristic polynomial. Neglecting $\det(A - \lambda I)$ which we know, *i.e.* defining $\tilde{\phi}_{PQ}(x) = \phi_{PQ}(x)/\det(A - xI)$, this means

$$\begin{aligned} \tilde{\phi}_{PQ}(x) &= \left(1 + \frac{Q_b(Q_B - 1)\frac{(1-\rho)}{(Q_B-1)} p_{\text{out}} \left(\frac{\rho(p_{\text{in}} - p_{\text{out}}) - x}{\rho(p_{\text{in}} + (Q_b-1)p_{\text{out}}) - x} \right)}{\left(\frac{(1-\rho)}{(Q_B-1)}(p_{\text{in}} - p_{\text{out}}) - x \right)} \right) \\ &\quad \times \left(\frac{(1-\rho)}{(Q_B-1)}(p_{\text{in}} - p_{\text{out}}) - x \right)^{Q_b(Q_B-1)}, \\ &= \left(\frac{(1-\rho)}{(Q_B-1)}(p_{\text{in}} - p_{\text{out}}) - x \right. \\ &\quad \left. + Q_b(1-\rho)p_{\text{out}} \left(\frac{\rho(p_{\text{in}} - p_{\text{out}}) - x}{\rho(p_{\text{in}} + (Q_b-1)p_{\text{out}}) - x} \right) \right) \\ &\quad \times \left(\frac{(1-\rho)}{(Q_B-1)}(p_{\text{in}} - p_{\text{out}}) - x \right)^{Q_b(Q_B-1)-1}, \\ &= \Big[\left((1-\rho)(p_{\text{in}} - p_{\text{out}}) - (Q_B - 1)x \right) \left(\rho(p_{\text{in}} + (Q_b-1)p_{\text{out}}) - x \right) \\ &\quad + Q_b(Q_B-1)(1-\rho)p_{\text{out}} \left(\rho(p_{\text{in}} - p_{\text{out}}) - x \right) \Big] \\ &\quad \times \frac{\left(\frac{(1-\rho)}{(Q_B-1)}(p_{\text{in}} - p_{\text{out}}) - x \right)^{Q_b(Q_B-1)-1}}{(Q_B-1)(\rho(p_{\text{in}} + (Q_b-1)p_{\text{out}}) - x)}. \end{aligned} \tag{B.15}$$

Importantly, note that in fact the denominator of the final factor will cancel the root of single multiplicity from $\det(A - xI)$, so that in total we find the correct number of

roots for the characteristic polynomial. Now focusing on the first factor, collecting terms as $ax^2 + bx + c$, we have

$$
\begin{aligned}
a &= Q_B - 1, && \text{(B.16)}\\
b &= -\Big[\rho(Q_B-1)(p_{\text{in}} + (Q_b-1)p_{\text{out}}) + (1-\rho)(p_{\text{in}} - p_{\text{out}}) \\
&\qquad + Q_b(Q_B-1)(1-\rho)p_{\text{out}}\Big], \\
&= -\Big[(Q_B-1)(\rho(p_{\text{in}} + (Q_b-1)p_{\text{out}}) + (1-\rho)Q_b p_{\text{out}}) + (1-\rho)(p_{\text{in}} - p_{\text{out}})\Big], \\
&= -\Big[(\rho(Q_B-2)+1)(p_{\text{in}} - p_{\text{out}}) + Q_b(Q_B-1)p_{\text{out}}\Big], && \text{(B.17)}\\
c &= \rho(1-\rho)(p_{\text{in}} - p_{\text{out}})\Big[(p_{\text{in}} + (Q_b-1)p_{\text{out}}) + Q_b(Q_B-1)p_{\text{out}}\Big], \\
&= \rho(1-\rho)(p_{\text{in}} - p_{\text{out}})\Big[(p_{\text{in}} - p_{\text{out}}) + Q_b Q_B p_{\text{out}}\Big]. && \text{(B.18)}
\end{aligned}
$$

Finally, setting $y = (p_{\text{in}} - p_{\text{out}})$, we can use quadratic formula: *i.e.* roots are $\frac{-b \pm \sqrt{b^2-4ac}}{2a}$, where

$$
\begin{aligned}
b^2 - 4ac &= (\rho(Q_B-2)+1)^2 y^2 + 2(\rho(Q_B-2)+1)Q_b(Q_B-1)p_{\text{out}} y \\
&\quad + Q_b^2(Q_B-1)^2 p_{\text{out}}^2 \\
&\quad - 4(Q_B-1)\rho(1-\rho)(y^2 + Q_b Q_B p_{\text{out}} y), \\
&= \Big[\rho^2(Q_B^2 - 4(Q_B-1) + 4(Q_B-1)) \\
&\quad + 2\rho(Q_B-1)(1-2) + 1 - 2\rho\Big] y^2 \\
&\quad + 2Q_b(Q_B-1)p_{\text{out}} y[\rho(Q_B-2) + 1 - 2Q_B\rho(1-\rho)] \\
&\quad + Q_b^2(Q_B-1)^2 p_{\text{out}}^2, \\
&= (\rho Q_B - 1)^2 y^2 - 2Q_b(Q_B-1)p_{\text{out}} y(\rho Q_B - 1)(1-2\rho) \\
&\quad + Q_b^2(Q_B-1)^2 p_{\text{out}}^2, \\
&= ((\rho Q_B - 1)y - Q_b(Q_B-1)p_{\text{out}})^2 + 4\rho Q_b(Q_B-1)p_{\text{out}} y(\rho Q_B - 1).
\end{aligned}
\qquad \text{(B.19)}
$$

As we are seeking eigenvalues for a real symmetric matrix, all eigenvalues must be real, thus this equation must be positive. The first term is clearly positive, while for the second to be positive, we must have

(i) $p_{\text{in}} > p_{\text{out}}$ and $\rho Q_B > 1$; or

(ii) $p_{\text{in}} < p_{\text{out}}$ and $\rho Q_B < 1$.

If the second term is not positive, it is less immediately clear that this holds in general. However, in the following, as the equations for the general case are rather long, for brevity we restrict ourselves to the case of two metadata groups, *i.e.* $Q_B = 2$. For this particular case, the above reduces to

$$((2\rho-1)y-Q_b p_{\text{out}})^2+4\rho(2\rho-1)Q_b p_{\text{out}}y = (2\rho-1)^2 y(y+2Q_b p_{\text{out}})+(Q_b p_{\text{out}})^2, \quad \text{(B.20)}$$

for which we can easily verify that if $y < 0$, there is a maxima for ρ when $\rho = 1/2$, and is decreasing in ρ in either direction away from this, as $Q_b \geq 1$ and thus $y + 2Q_b p_{\text{out}} \geq p_{\text{in}} + p_{\text{out}} > 0$. Hence we can simply consider the extremal values $\rho \in \{0, 1\}$, for which this takes the same value,

$$y(y + 2Q_b p_{\text{out}}) + (Q_b p_{\text{out}})^2 = (y + Q_b p_{\text{out}})^2 > 0, \quad \text{(B.21)}$$

and so indeed we obtain real roots.

Proceeding for this case, and reintroducing the factor $\frac{N}{Q_b}$ and p_{in}, this allows us to find four distinct eigenvalues for PQ, as stated in Chap. 4:

$$\lambda_1 = \frac{N}{Q_b}\rho(p_{\text{in}} - p_{\text{out}}), \quad \text{with mult. } (Q_b - 1), \quad \text{(B.22)}$$

$$\lambda_2 = \frac{N}{Q_b}(1 - \rho)(p_{\text{in}} - p_{\text{out}}), \quad \text{with mult. } Q_b(Q_B - 1) - 1, \quad \text{(B.23)}$$

$$\lambda_{3,4} = \frac{N}{2Q_b}\Bigg(p_{\text{in}} + (Q_b - 1)p_{\text{out}} \pm \sqrt{(2\rho - 1)^2(p_{\text{in}} - p_{\text{out}})(p_{\text{in}} + (2Q_b - 1)p_{\text{out}}) + (Q_b p_{\text{out}})^2}\Bigg), \quad \text{(B.24)}$$

both with mult. 1,

where we assume λ_3 takes the minus and thus $\lambda_3 \leq \lambda_4$.

Appendix C

MFVI for the causal model given substitutes

To fit the Poisson model in Eqn. (6.13) given whichever substitutes chosen, we use a lightly modified version of that in PIF, approximating the posterior over parameters using mean-field variational inference (MFVI) [166]. As such, the following proceeds much as in their work. The process of deriving the necessary equations for MFVI is much as in Chap. 3.

That is, given the set of author-topic publication counts, $Y = \{Y^1, \ldots, Y^T\}$, the author citation networks A, and our substitutes for both the topic-citation confounders, $\hat{\zeta}$, and topic-only confounders, $\hat{\tau}$, we want to calculate the posterior distribution over the remaining parameters. These are the author preferences for the topics at each timestep, α^t, attributes of the topic at that time, γ^t, and the main parameter of interest, each author's influence, β^t — that is, we target $\mathrm{p}(\alpha, \gamma, \beta \mid Y, A, \hat{\zeta}, \hat{\tau})$.

To do so, MFVI posits a fully factorised 'variational' family of distributions, then optimises this to be as close to the true posterior as possible, usually measured in terms of Kullback-Liebler (KL) divergence. Much as for Gibbs sampling, the important factor when choosing the overall family of distributions to use to do so is the complete conditional of each latent variable, *i.e.* its distribution conditioned on all other variables, both latent and observed. This is somewhat simplified if a model is conditionally conjugate – that is the posterior distribution of the latent variables is in the same family as the prior distribution of latent variables – as complete conditionals are then in same family, and iterative updating is easy. As such, we can sequentially update the variational parameters for each latent variable in turn, holding the others fixed, and repeat until convergence.

Formally, we seek to infer parameters for

(i) the author interest profiles at each timestep, α_i^t, for the topic attributes τ_k^t, which we estimate with our corresponding substitutes (held fixed during this stage of inference) from the dPF, $\hat{\tau}$;

(ii) the topic attributes γ_k^t for the author-topic-link factors ζ_i^t, which we estimate using our substitutes from the DSBM(M), $\hat{\zeta}$;

(iii) author influence at each timestep, β_i^t.

To do so, we use SVI almost exactly as in [166]. That is, we seek the posterior over these parameters given the observed data,

$$\mathrm{p}(\alpha, \gamma, \beta \mid Y, A, \hat{\zeta}, \hat{\tau}) = \frac{\mathrm{p}(Y \mid \alpha, \gamma, \beta, A, \hat{\zeta}, \hat{\tau})\mathrm{p}(\alpha, \gamma, \beta \mid A, \hat{\zeta}, \hat{\tau})}{\mathrm{p}(Y \mid A, \hat{\zeta}, \hat{\tau})}. \tag{C.1}$$

Recall we assume that

$$\mathrm{p}(y_{ik}^t \mid \alpha_i^t, \gamma_k^t, \tau_k^t, \zeta_i^t, \beta^{t-1}, a_i^{t-1}, y^{t-1}) \sim \mathrm{Pois}\left(\alpha_i^{t,\top}\tau_k^t + \gamma_k^{t,\top}\zeta_i^t + \sum_j \beta_j^{t-1} a_{ij}^{t-1} y_{jk}^{t-1}\right), \tag{C.2}$$

and so if we introduce the auxiliary variables $\psi_{ik\ell}^t$, ξ_{ikp}^t and ϑ_{ikq}^t, such that

$$\begin{aligned} \vartheta_{ikq}^t &\sim \mathrm{Pois}(\gamma_{kq}^t \hat{\zeta}_{iq}^t), & \xi_{ikp}^t &\sim \mathrm{Pois}(\alpha_{ip}^t \hat{\tau}_{kp}^t), \\ \psi_{ik\ell}^t &\sim \mathrm{Pois}(\beta_l^{t-1} a_{i\ell}^{t-1} y_{\ell k}^{t-1}), & y_{ik}^t &= \sum_{qp\ell} \vartheta_{ikq} + \xi_{ikp} + \psi_{ik\ell}, \end{aligned} \tag{C.3}$$

we have the complete conditionals

$$\alpha_{ip}^t \mid y^t, \hat{\tau}^t, \xi^t \sim \mathrm{Gamma}\left(a + \sum_k \xi_{ikp}^t, b + \sum_k \hat{\tau}_{kp}\right), \tag{C.4}$$

$$\gamma_{kq}^t \mid y^t, \hat{\zeta}^t, \vartheta^t \sim \mathrm{Gamma}\left(a + \sum_i \vartheta_{ikq}^t, b + \sum_i \hat{\zeta}_{iq}\right), \tag{C.5}$$

$$\beta_\ell^{t-1} \mid y^t, a^{t-1}, y^{t-1}, \psi^t \sim \mathrm{Gamma}\left(c + \sum_{i,k} \psi_{ik\ell}^t, d + \sum_{i,k} a_{i\ell}^{t-1} y_{\ell k}^{t-1}\right), \tag{C.6}$$

$$\vartheta_{ik}^t \mid y^t, \gamma^t, \hat{\zeta}^t \sim \mathrm{Mult}\left(y_{ik}^t, \frac{\hat{\zeta}_i^t \odot \gamma_k^t}{Z_{ik}^t}\right), \tag{C.7}$$

$$\xi_{ik}^t \mid y^t, \alpha^t, \hat{\tau}^t \sim \mathrm{Mult}\left(y_{ik}^t, \frac{\alpha_i^t \odot \hat{\tau}_k^t}{Z_{ik}^t}\right), \tag{C.8}$$

$$\psi_{ik}^t \mid y^t, a^{t-1}, y^{t-1}, \beta^{t-1} \sim \mathrm{Mult}\left(y_{ik}^t, \frac{\beta^{t-1} \odot a_i^{t-1} \odot y_k^{t-1}}{Z_{ik}^t}\right), \tag{C.9}$$

where the scalar values a, b, c and d are prior shape and rate parameters, $\odot$ denotes element-wise multiplication, and

$$Z_{ik}^t = \hat{\zeta}_i^{t,\top} \gamma_k^t + \alpha_i^{t,\top} \hat{\tau}_k^t + \beta^{t-1,\top} (a_i^{t-1} \odot y_k^{t-1}). \tag{C.10}$$

The conditionals for α, γ, and β follow immediately, as the Gamma distribution is a conjugate prior for the Poisson distribution — this means that given the variational parameters, updates for the parameters of interest are immediate. The conditionals for the variational parameters are slightly less immediate, and use that if a random variable Y is made up of the sum of independent Poisson random variables, $X_1, \ldots, X_R$, then the conditional distribution over $\{X\}$ given that Y takes a particular value, n, is multinomial:

$$\begin{aligned}
\mathrm{p}(\{X\} \mid Y = n) &= \frac{\mathrm{p}(\{X\}, Y = n)}{\mathrm{p}(Y = n)}, \\
&= \frac{\mathrm{p}(\{X\})}{\mathrm{p}(Y = n)}, \\
&= \frac{\prod_i \mathrm{Pois}(X_i; \lambda_i)}{\mathrm{Pois}(n; \sum_i \lambda_i)}, \\
&= \left(\prod_i \mathrm{e}^{-\lambda_i} \lambda_i^{X_i} / X_i! \right) \mathrm{e}^{\sum_i \lambda_i} (\sum_i \lambda_i)^{-n} n!, \\
&= \frac{n!}{\prod_i X_i!} \prod_i \left(\frac{\lambda_i}{\sum_i \lambda_i} \right)^{X_i}, \\
&= \mathrm{Mult}\left(n, \frac{\lambda_i}{\sum_i \lambda_i} \right)
\end{aligned} \tag{C.11}$$

As such, if we know the value of y_{ik}^t and a_i^{t-1}, and fix the other variables – and thus the rate for each variational parameter – then we have that

$$\mathrm{p}(\vartheta_{ik}^t, \xi_{ik}^t, \psi_{ik}^t \mid \ldots) \sim \mathrm{Mult}\left(y_{ik}^t, \frac{\{\zeta_i^t \odot \gamma_k^t, \alpha_i \odot \tau_k^t, \beta^{t-1} \odot a_i^{t-1} \odot y_k^{t-1}\}}{Z_{ik}^t} \right). \tag{C.12}$$

Given this is the case, the true variational parameters of interest are the 'class probabilities' of these multinomial distributions, as the variational parameters themselves are then just these scaled by y_{ik}^t. We denote these by ϕ_{ikpt}^{ϑ}, ϕ_{ikqt}^{ξ}, $\phi_{ik\ell t}^{\psi}$ respectively. Substituting these conditionals into the ELBO for the model and

differentiating then gives us the updates

$$\tilde{\phi}^{\vartheta}_{ikpt} \propto \exp\left(\Psi\left(a + \sum_{k} y^{t}_{ik}\phi^{\vartheta}_{ikpt}\right) - \log\left(b + \sum_{k}\hat{\tau}^{t}_{kp}\right)\right) + \hat{\tau}^{t}_{kp}, \tag{C.13}$$

$$\tilde{\phi}^{\xi}_{ikqt} \propto \exp\left(\Psi\left(a + \sum_{i} y^{t}_{ik}\phi^{\xi}_{ikqt}\right) - \log\left(b + \sum_{i}\hat{\zeta}^{t}_{iq}\right)\right) + \hat{\zeta}^{t}_{iq}, \tag{C.14}$$

$$\tilde{\phi}^{\psi}_{ik\ell t} \propto \exp\left(\Psi\left(c + \sum_{\ell,k} y^{t}_{\ell k}\phi^{\psi}_{\ell kit}\right) - \log\left(d + \sum_{\ell,k} a^{t-1}_{\ell i}y^{t-1}_{ik}\right)\right) + a^{t-1}_{\ell i}y^{t-1}_{ik}, \tag{C.15}$$

where $\Psi(\cdot)$ is the digamma function.

Appendix D

Additional results for the causal inference procedure

D.1 Substitutes for an altered causal model

As we described in Sec. 6.3.2, a key assumption for the DSBMM is that the network and the metadata are conditionally independent given the network groups, *i.e.* $A \perp\!\!\!\perp X \mid Z$. Furthermore, this is important for the causal inference procedure of Chap. 6, as in the proposed causal model model we assume $A \perp\!\!\!\perp X \mid \zeta$. As a result, A^t is a non-descendant of X^t, and hence a valid member of the back-door adjustment set. If this were not the case, estimation of the causal effect is still possible, but would need to do so in two stages.

Precisely, we could use that

$$\mathrm{p}(Y \mid A, \mathrm{do}(X)) = \mathrm{p}(Y, A \mid \mathrm{do}(X))/\mathrm{p}(A \mid \mathrm{do}(X)), \tag{D.1}$$

and subsequently both outcome sets in this equation are estimable through back-door adjustment (with substitutes). The key difference is that we would need to adjust with $\{\zeta_j^{t-1}, \tau_k^t\}$ for the numerator, and just ζ_j^{t-1} for the denominator, rather than the previous set of terms $\{\zeta_i^t, \tau_k^t, a_{ij}^{t-1}\}$ directly.

While this formula holds in the causal model proposed in Chap. 6 also, as $(Y \perp\!\!\!\perp A \mid \zeta)_{G_{\overline{X}\underline{Z}}}$, we can treat $\mathrm{p}(Y \mid A, \mathrm{do}(X), \zeta)$ as $\mathrm{p}(Y \mid \mathrm{do}(A), \mathrm{do}(X), \zeta)$ – hence the notation used for μ, as $\mu(a, x)$ rather than just $\mu(x)$ – then simply seek the necessary back-door adjustment set. This can be seen directly through the definition of the

corresponding causal quantities, *i.e.*

$$
\begin{aligned}
\mathrm{p}(Y \mid \mathrm{do}(A), \mathrm{do}(X)) &= \sum_{\zeta,\theta,\tau,Z} \frac{\mathrm{p}(Y, A, X, \zeta, \theta, \tau, Z)}{\mathrm{p}(A \mid \zeta)\mathrm{p}(X \mid \zeta, \theta, \tau)}, \\
&= \sum_{\zeta,\theta,\tau,Z} \frac{\mathrm{p}(Y \mid A, X, \zeta, \theta, \tau)}{\mathrm{p}(A \mid \zeta, Z)\mathrm{p}(X \mid \zeta, \theta, \tau)} \\
&\qquad \times \mathrm{p}(A \mid \zeta, Z)\mathrm{p}(X \mid \zeta, \theta, \tau)\mathrm{p}(\zeta)\mathrm{p}(\theta)\mathrm{p}(\tau)\mathrm{p}(Z), \\
&= \sum_{\zeta,\theta,\tau,Z} \mathrm{p}(Y \mid A, X, \zeta^t, \theta^t, \tau^t)\mathrm{p}(\zeta^t \mid \zeta^{t-1}) \\
&\qquad \times \mathrm{p}(\zeta^{t-1})\mathrm{p}(\theta)\mathrm{p}(\tau^t \mid \tau^{t-1})\mathrm{p}(\tau^{t-1})\mathrm{p}(Z), \\
&= \sum_{\zeta^t,\tau^t} \mathrm{p}(Y \mid A, X, \zeta^t, \tau^t)\mathrm{p}(\zeta^t)\mathrm{p}(\tau^t), \\
\mathrm{p}(Y \mid A, \mathrm{do}(X)) &= \frac{\mathrm{p}(Y, A \mid \mathrm{do}(X))}{\mathrm{p}(A \mid \mathrm{do}(X))}, \\
&= \frac{\mathrm{p}(Y, A \mid \mathrm{do}(X))}{\mathrm{p}(A)}, \\
&= \sum_{\zeta,\theta,\tau,Z} \frac{\mathrm{p}(Y \mid A, X, \zeta, \theta, \tau)\mathrm{p}(A \mid \zeta, Z)\mathrm{p}(X \mid \zeta, \theta, \tau)\mathrm{p}(\zeta)\mathrm{p}(\theta)\mathrm{p}(\tau)\mathrm{p}(Z)}{\mathrm{p}(X \mid \zeta, \theta, \tau)\sum_{\zeta,Z}\mathrm{p}(A \mid \zeta, Z)\mathrm{p}(\zeta)\mathrm{p}(Z)}, \\
&= \sum_{\zeta,\theta,\tau,Z} \frac{\mathrm{p}(Y \mid A, X, \zeta, \theta, \tau)\mathrm{p}(A \mid \zeta^{<t}, Z)\mathrm{p}(\zeta)\mathrm{p}(\theta)\mathrm{p}(\tau)\mathrm{p}(Z)}{\sum_{\zeta,Z}\mathrm{p}(A \mid \zeta, Z)\mathrm{p}(\zeta)\mathrm{p}(Z)}, \\
&= \sum_{\zeta^t,\tau^t} \frac{\mathrm{p}(Y \mid A, X, \zeta^t, \tau^t)\mathrm{p}(A)\mathrm{p}(\zeta^t)\mathrm{p}(\tau^t)}{\mathrm{p}(A)} \\
&= \sum_{\zeta^t,\tau^t} \mathrm{p}(Y \mid A, X, \zeta^t, \tau^t)\mathrm{p}(\zeta^t)\mathrm{p}(\tau^t).
\end{aligned}
\tag{D.2}
$$

■

This condition for equivalence of passive observation and external intervention amounts to ζ^t blocking all back-door (*i.e.* spurious) paths from A to Y in $G_{\overline{X}}$.

Assuming that the groups could render the citations and topics conditionally independent is not necessarily unreasonable, as given two researchers belong to particular research communities that are partially defined by their topics, it is plausible that the likelihood of one citing another only depends on the set of topics, or broad focus of those communities. That is, the cited paper is in a particular topic presumably of interest to the citing author, but the *reason* that that particular paper was cited – rather than any other in the same topic – could be the particular contextualisation used by and/or prestige of the author *etc.*, which may be directly determined by the research community to which they belong.

D.2 Additional results on semi-synthetic data

In Table D.2 we provide additional author influence MSE results on semi-synthetic data. We evaluate DSBMM recovery both in the case where the region-based homophily confounder is generated using ADM1 regions of affiliations (marked -A), and using countries (-C) — we expect for the use of countries to result in synthetic influence that is easier to account for, as there the latent covariate is lower dimensional.

D.3 Average total influence contribution to the rate in the real data

	$\langle\tilde{\beta}^1\rangle$	$\langle\tilde{\beta}^2\rangle$	$\langle\tilde{\beta}^3\rangle$	$\langle\tilde{\beta}^4\rangle$
Unadjusted	1.30	1.19	1.20	0.94
Network-Only	1.12	0.89	0.87	0.74
Topic-Only	1.14	1.07	1.09	0.86
Ours	1.12	0.95	0.89	0.78
Ours-NDC	1.13	0.95	0.89	0.77
Ours (no meta)	1.13	0.96	0.90	0.75

Table D.1: In this table, we present results after applying the procedure to the real publication data. For each method, we show the average influence contribution to the publication rate, $\sum_j a_{ij}^t \beta_j^t x_{jk}^t$ ($\times 10^3$), across all authors at each timestep, denoted $\langle\tilde{\beta}^t\rangle$ — this changes at a slower rate than the average influence, *i.e.* authors are 'inspired' to a similar degree over time, but by a wider variety of others.

D.4 The effect of sequential citation 'treatments'

A key extension for properly understanding influence is to extend the idea of 'treatment' citations to permit any previously measured exposure to the other author's ideas, then test how *sequences* of citations change future publications observed. For instance, we might expect that the longer the duration for which an author has cited another, and/or the wider the variety of topics, the higher the likelihood that they are a key inspiration or colleague, and thus the larger the influence.

Indeed, if one were to operate on the same assumption as used in Chap. 6, that substitutes may be used in lieu of observing the relevant confounders, this can actually

	Exog.			Homophily			Both		
	Low	Med.	High	Low	Med.	High	Low	Med.	High
Oracle-old-A	0.28±0.2	0.3±0.2	0.33±0.22	20.31±28.2	12.56±17.45	6.33±8.72	26.21±36.32	12.7±17.61	8.67±12.01
Oracle-pres-A	0.32±0.2	0.37±0.18	0.41±0.22	32.1±44.13	13.11±16.71	7.35±7.83	26.83±36.57	13.12±16.76	9.36±10.77
Topic-oracle-old-A	**0.09±0.04**	**0.1±0.1**	**0.13±0.06**	18.64±23.34	13.12±14.65	8.96±7.98	24.88±32.21	17.46±20.82	10.79±10.71
Topic-oracle-pres-A	0.25±0.11	0.31±0.15	0.32±0.15	14.59±17.63	13.14±14.66	8.97±8.03	25.16±32.52	17.82±21.16	10.8±10.64
Net.-only-pres-A	2.34±2.22	1.35±0.56	1.43±0.45	22.18±28.39	7.82±2.16	19.66±20.61	44.9±28.96	16.28±16.66	8.06±2.7
Ours-NDC-old-A	2.03±3.4	2.11±3.73	1.88±3.29	23.63±51.03	9.21±12.08	8.18±10.22	11.54±16.69	9.29±12.88	8.96±10.94
Ours-NDC-old-C	0.85±0.99	1.15±1.25	1.39±1.61	**6.3±9.49**	**5.49±6.42**	5.32±4.86	**6.56±7.34**	**5.15±5.17**	**5.39±5.04**
Ours-NDC-pres-A	2.03±3.4	2.17±3.69	1.81±3.04	21.43±42.9	9.69±12.33	8.28±10.18	11.68±16.73	9.47±12.91	9.18±11.02
Ours-NDC-pres-C	0.83±1.0	2.16±3.91	2.04±2.94	8.64±16.11	5.7±6.51	5.47±4.94	6.77±7.51	5.37±5.26	5.56±5.11
Ours-no-meta-old-A	1.6±2.94	2.43±4.93	2.12±2.99	8.14±10.87	11.17±16.29	8.58±9.92	12.37±14.99	9.52±10.9	9.87±11.12
Ours-no-meta-old-C	1.43±2.47	1.66±2.18	2.12±2.49	13.23±14.92	7.8±9.05	6.56±5.71	13.85±15.96	8.08±8.29	8.49±8.26
Ours-no-meta-pres-A	1.6±2.94	2.43±4.93	2.12±2.99	8.37±11.31	11.16±16.32	8.54±9.92	12.37±15.0	9.49±10.9	9.84±11.12
Ours-no-meta-pres-C	1.37±2.49	1.65±2.18	2.12±2.49	6.54±5.81	10.37±12.03	6.49±5.71	13.77±15.94	8.01±8.27	8.4±8.24
Ours-old-A	1.6±2.93	2.47±4.91	2.29±2.97	11.71±14.6	8.9±10.44	8.6±10.01	12.43±15.02	9.61±11.04	9.89±11.23
Ours-old-C	1.61±2.96	2.49±4.96	2.31±2.99	12.68±14.75	9.18±10.36	8.87±9.95	12.73±14.86	9.87±10.91	10.17±11.15
Ours-pres-A	1.84±2.94	2.49±4.96	2.32±3.0	17.02±28.03	9.84±10.46	8.86±10.54	12.74±14.91	10.21±11.27	10.48±11.63
Ours-pres-C	1.79±2.95	2.46±4.86	2.31±2.98	15.43±22.59	10.59±13.63	9.67±11.93	12.84±15.06	10.38±11.64	10.96±12.93
Topic-only-pres-A	0.35±0.17	0.66±0.23	0.59±0.27	10.68±14.24	7.98±8.88	**4.36±2.8**	11.71±15.68	10.76±12.8	6.5±5.89
Unadjusted-old-A	0.7±0.2	1.07±0.35	1.01±0.37	14.34±10.3	14.39±10.58	11.77±6.77	26.39±26.7	17.11±13.94	14.23±9.9
Unadjusted-pres-A	0.7±0.2	1.06±0.35	1.01±0.37	15.81±12.78	14.39±10.58	11.77±6.77	26.39±26.7	17.11±13.94	14.23±9.9

Table D.2: In this table, we present the accuracy of estimated academic influence in our semi-synthetic data experiments. Entries are the average MSE ($\times 10^3$) of estimated influence β across 5 repeated simulations, each on a different subsampled network. Variants marked '-old' correspond to using substitutes inferred for the previous timestep, while '-pres' uses the current values. Variants marked '-A' correspond to simulations using the ADM1 region covariate, while '-C' denotes simulations using the country covariate. Bolded entries show which variant performed best. Each simulation corresponds to results on a different, snowball sampled subgraph of 3K authors and 1K topics. We evaluate three different levels of confounding (low, medium and high), and three different types of confounding — whether of topics only (Exog., *i.e.* τ only), homophily effects only (ζ only), or both effects together.

be done immediately. That is, if all confounders in our system were observed, but we now assume that

$$\mathrm{p}(y_{ik}^t \mid \{a_{ij}^{<}t\}, \zeta_i^t, \tau_k^t, \theta_i^t), \tag{D.3}$$

we may exactly follow the process for sequential plans elaborated in [130], as the prerequisite conditions are met. One might then define the overall influence of one author on another as the cumulative effect of exposure to their ideas, as measured by citations, by comparing to the expected number of publications were these citations / exposures never to have occurred, *i.e.*

$$\xi_{ijk}^t = \mathbb{E}[y_{ik}^t \mid \{a_{ij}^{<t} = 1\}]. \tag{D.4}$$

A more immediately available option for incorporating cumulative effects may appear to be to construct the citation graph at each timestep from some accumulation (perhaps with some forgetting) of citations in previous timesteps, in addition to the new edges present. However, in such a construction, the importance of edge persistence is immediately obvious, as many edges present will solely be so due to presence at a previous timestep, rather than inherently due to some grouping of the nodes. As such, we should first develop a model that accounts for this, as suggested at the end of Chap. 4.

D.5 Treating citations as temporal quadruples

As suggested in Chap. 6, we discard temporal information about the cited paper when constructing our citation networks. However, in truth every edge is a quadruple — not a triple like most temporal networks viewed over discrete timesteps. Nor is this a quadruple where the event has some duration, which may be treated as (i, j, t_s, t_e) for an event recorded from i to j starting at t_s and finishing at t_e, which has received a reasonable amount of attention in the literature with respect to continuous time dynamic networks (see *e.g.* [30, 10]). Instead, we ought to consider (i, j, t_i, t_j), for author i producing a publication at time t_i, that cites a paper by j published at time t_j.

In fact, this is more in line with a *multilayer* view of the dynamic network, which suggests one could view the system instead as a dynamic bipartite network of authors and publications, with citations being potentially inter-time period (/layer) directed edges from paper to paper. This could be analysed using our model by viewing the

bipartite network as the primary network, but *e.g.* further incorporating a separate probability governing connections between groups at different lags, for instance like

$$\begin{aligned} \mathrm{p}(m_k^t \to m_\ell^{t'} &= m_{k\ell}^{tt'} \mid z_k^t, z_\ell^{t'}, t' \leq t), \\ &= \eta_{z_k^t, z_\ell^{t'}, (t-t')}^{m_{k\ell}^{tt'}} (1 - \eta_{z_k^t, z_\ell^{t'}, (t-t')})^{1 - m_{k\ell}^{tt'}}. \end{aligned} \tag{D.5}$$

As this would solely be an additional pairwise factor, for inclusion in the DSBMM, it would only necessitate the inclusion of (i) two additional factors in the BP message equations, $\prod_{(t-t')} \prod_{\ell \in \partial m_i^{(t-t')}} \sum_r \eta_{qr(t-t')} \psi_r^{\ell, t' \to k, t}$ for backwards (sending) citations, and $\prod_{(t'-t)} \prod_{\ell \in \partial m_i^{(t'-t)}} \sum_r \eta_{rq(t'-t)} \psi_r^{\ell, t' \to k, t}$ for forwards (receiving) citations, (ii) a modification of the external field, and (iii) the new set of citation-carried messages, $\psi_q^{k, t \to \ell, t'}$ for publication k at time t either citing an earlier or contemporary publication ($t' \leq t$), or receiving a citation from a later/contemporary publication ($t' \geq t$). However, this would of course drastically increase both the number of nodes that must be considered, and the number of messages necessary to be computed at each iteration, as well as adding a further $Q^2(T-1)(T-2)/2$ parameters to estimate in η.

Bibliography

[1] Emmanuel Abbe and Colin Sandon. Achieving the ks threshold in the general stochastic block model with linearized acyclic belief propagation. *Advances in Neural Information Processing Systems*, 29, 2016.

[2] Francisco José Acedo, Carmen Barroso, Cristóbal Casanueva, and José Luis Galán. Co-Authorship in Management and Organizational Studies: An Empirical and Network Analysis*. *Journal of Management Studies*, 43(5):957–983, July 2006.

[3] Ayan Acharya, Avijit Saha, Mingyuan Zhou, Dean Teffer, and Joydeep Ghosh. Nonparametric dynamic network modeling. In *KDD Workshop on Mining and Learning from Time Series*, 2015.

[4] Manuel Acosta, Daniel Coronado, Esther Ferrándiz, and M Dolores León. Factors affecting inter-regional academic scientific collaboration within europe: The role of economic distance. *Scientometrics*, 87(1):63–74, 2011.

[5] Amr Ahmed and Eric Xing. Dynamic non-parametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary clustering. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 219–230. SIAM, 2008.

[6] Christopher Aicher, Abigail Z Jacobs, and Aaron Clauset. Adapting the stochastic block model to edge-weighted networks. *arXiv preprint arXiv:1305.5782*, 2013.

[7] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of machine learning research*, 9(Sep):1981–2014, 2008.

[8] Zafar Ali, Irfan Ullah, Amin Khan, Asim Ullah Jan, and Khan Muhammad. An overview and evaluation of citation recommendation models. *Scientometrics*, 126(5):4083–4119, 2021.

[9] Maria Chiara Angelini, Francesco Caltagirone, Florent Krzakala, and Lenka Zdeborová. Spectral detection on sparse hypergraphs. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 66–73. IEEE, 2015.

[10] Makan Arastuie, Subhadeep Paul, and Kevin S Xu. Chip: A hawkes process model for continuous-time networks with scalable and consistent estimation. In *NeurIPS*, 2020.

[11] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590–614, 2002.

[12] Paolo Barucca, Fabrizio Lillo, Piero Mazzarisi, and Daniele Tantari. Disentangling group and link persistence in dynamic stochastic block models. *Journal of Statistical Mechanics: Theory and Experiment*, 2018(12):123407, 2018.

[13] Matthew J Beal, Zoubin Ghahramani, and Carl Edward Rasmussen. The infinite hidden markov model. *Advances in neural information processing systems*, 1:577–584, 2002.

[14] Catherine Beaudry and Sedki Allaoui. Impact of public and private research funding on scientific production: The case of nanotechnology. *Research Policy*, 41(9):1589–1606, 2012.

[15] Allan Bickle. *The k-cores of a graph*. Western Michigan University, 2010.

[16] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[17] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[18] Katy Börner, Luca Dall'Asta, Weimao Ke, and Alessandro Vespignani. Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams. *arXiv:cond-mat/0502147*, February 2005. arXiv: cond-mat/0502147.

[19] Charles Bouveyron, Pierre Latouche, and Rawya Zreik. The stochastic topic block model for the clustering of vertices in networks with textual edges. *Statistics and Computing*, 28(1):11–31, 2018.

[20] Frances Boyle and Damien Sherman. Scopus™: The product and its development. *The Serials Librarian*, 49(3):147–153, 2006.

[21] Stefano Breschi, Francesco Lissoni, and Fabio Montobbio. University patenting and scientific productivity: a quantitative study of italian academic inventors. *European Management Review*, 5(2):91–109, 2008.

[22] Jane Carlen, Jaume de Dios Pont, Cassidy Mentus, Shyr-Shea Chang, Stephanie Wang, and Mason A Porter. Role detection in bicycle-sharing networks using multilayer stochastic block models. *arXiv preprint arXiv:1908.09440*, 2019.

[23] Miguel A Carreira-Perpinán and Steve Renals. Practical identifiability of finite mixtures of multivariate bernoulli distributions. *Neural Computation*, 12(1):141–152, 2000.

[24] Allison JB Chaney, David M Blei, and Tina Eliassi-Rad. A probabilistic model for using social networks in personalized item recommendation. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 43–50, 2015.

[25] Jonathan Chang and David Blei. Relational topic models for document networks. In *Artificial Intelligence and Statistics*, pages 81–88, 2009.

[26] Laurent Charlin, Rajesh Ranganath, James McInerney, and David M Blei. Dynamic poisson factorization. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 155–162, 2015.

[27] Yang Chen, Cong Ding, Jiyao Hu, Ruichuan Chen, Pan Hui, and Xiaoming Fu. Building and Analyzing a Global Co-Authorship Network Using Google Scholar Data. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 1219–1224, Perth, Australia, April 2017. International World Wide Web Conferences Steering Committee.

[28] Philip S Chodrow, Nate Veldt, and Austin R Benson. Generative hypergraph clustering: From blockmodels to modularity. *Science Advances*, 7(28):eabh1303, 2021.

[29] Marco Corneli. *Dynamic stochastic block models, clustering and segmentations in dynamic graphs*. PhD thesis, Université Paris 1-Panthéon Sorbonne, 2017.

[30] Marco Corneli, Charles Bouveyron, Pierre Latouche, and Fabrice Rossi. The dynamic stochastic topic block model for dynamic networks with textual edges. *Statistics and Computing*, 29(4):677–695, 2019.

[31] Marco Corneli, Pierre Latouche, and Fabrice Rossi. Modelling time evolving interactions in networks through a non stationary extension of stochastic block models. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1590–1591. IEEE, 2015.

[32] Michele Coscia and Frank MH Neffke. Network backboning with noisy data. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pages 425–436. IEEE, 2017.

[33] Robert Cowell. Advanced inference in bayesian networks. In *Learning in graphical models*, pages 27–49. Springer, 1998.

[34] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

[35] Antoine Delmotte, Edward W Tate, Sophia N Yaliraki, and Mauricio Barahona. Protein multi-scale organization through graph partitioning and robustness analysis: application to the myosin? myosin light chain interaction. *Physical biology*, 8(5):055010, 2011.

[36] Pierre Deville, Dashun Wang, Roberta Sinatra, Chaoming Song, Vincent D Blondel, and Albert-László Barabási. Career on the move: Geography, stratification, and scientific impact. *Scientific reports*, 4:4770, 2014.

[37] Stephen G Donald and Kevin Lang. Inference with difference-in-differences and other panel data. *The review of Economics and Statistics*, 89(2):221–233, 2007.

[38] Christopher DuBois, Carter Butts, and Padhraic Smyth. Stochastic blockmodeling of relational event dynamics. In *Artificial intelligence and statistics*, pages 238–246. PMLR, 2013.

[39] Ioana Dumitriu, Haixiao Wang, and Yizhe Zhu. Partial recovery and weak consistency in the non-uniform hypergraph stochastic block model. *arXiv preprint arXiv:2112.11671*, 2021.

[40] Joel Dyer, John Fitzgerald, Bastian Rieck, and Sebastian M Schmon. Approximate bayesian computation for panel data with signature maximum mean discrepancies. In *NeurIPS 2022 Temporal Graph Learning Workshop*, 2022.

[41] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[42] Josemar Faustino, Nandini Iyer, Juan Mendonza, and Ronaldo Menezes. Characterizing the dynamics of academic affiliations: A network science approach. In *Complex Networks XI*, pages 393–404. Springer, 2020.

[43] John Fitzgerald, Sanna Ojanperä, and Neave O'Clery. Is academia becoming more localised? the growth of regional knowledge networks within international research collaboration. *Applied Network Science*, 6(1):1–27, 2021.

[44] Laura Florescu and Will Perkins. Spectral thresholds in the bipartite stochastic block model. In *Conference on Learning Theory*, pages 943–959. PMLR, 2016.

[45] James Foulds, Christopher DuBois, Arthur Asuncion, Carter Butts, and Padhraic Smyth. A dynamic relational infinite feature model for longitudinal social networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 287–295. JMLR Workshop and Conference Proceedings, 2011.

[46] Koen Frenken and Ron A Boschma. A theoretical framework for evolutionary economic geography: Industrial dynamics and urban growth as a branching process. *Journal of Economic Geography*, 7:635–649, 2007.

[47] Ryan J Gallagher, Jean-Gabriel Young, and Brooke Foucault Welles. A clarified typology of core-periphery structure in networks. *Science Advances*, 7(12):eabc9800, 2021.

[48] Alexander J Gates and Yong-Yeol Ahn. The impact of random models on clustering similarity. *arXiv preprint arXiv:1701.06508*, 2017.

[49] Martin Gerlach, Tiago P Peixoto, and Eduardo G Altmann. A network approach to topic models. *Science advances*, 4(7):eaaq1360, 2018.

[50] Amir Ghasemian, Pan Zhang, Aaron Clauset, Cristopher Moore, and Leto Peel. Detectability thresholds and optimal algorithms for community structure in dynamic networks. *Physical Review X*, 6(3):031005, 2016.

[51] Wolfgang Glänzel and András Schubert. Analysing Scientific Networks Through Co-Authorship. In Henk F. Moed, Wolfgang Glänzel, and Ulrich Schmoch, editors, *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems*, pages 257–276. Springer Netherlands, Dordrecht, 2005.

[52] Prem K Gopalan and David M Blei. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, 110(36):14534–14539, 2013.

[53] Amit Goyal, Francesco Bonchi, and Laks VS Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 241–250, 2010.

[54] Noriko Hara, Paul Solomon, Seung-Lye Kim, and Diane H Sonnenwald. An emerging view of scientific collaboration: Scientists' perspectives on collaboration and factors that impact collaboration. *Journal of the American Society for Information science and Technology*, 54(10):952–965, 2003.

[55] Ricardo Hausmann and César A Hidalgo. The network structure of economic output. *Journal of Economic Growth*, 16(4):309–342, 2011.

[56] Teague R Henry, David Banks, Derek Owens-Oas, and Christine Chai. Modeling community structure and topics in dynamic text networks. *Journal of Classification*, 36(2):322–349, 2019.

[57] Tue Herlau, Morten Mørup, and Mikkel Schmidt. Modeling temporal evolution and multiscale structure in networks. In *International Conference on Machine Learning*, pages 960–968. PMLR, 2013.

[58] César A Hidalgo and Ricardo Hausmann. The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, 106(26):10570–10575, 2009.

[59] César A Hidalgo, Bailey Klinger, Albert-László Barabási, and Ricardo Hausmann. The product space conditions the development of nations. *Science*, 317(5837):482–487, 2007.

[60] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[61] Junteng Hou, Chengxiang Si, Shupeng Wang, Guangjun Wu, and Lei Zhang. Parallel belief propagation optimized by coloring on gpus. In *International Conference on Algorithms and Architectures for Parallel Processing*, pages 645–660. Springer, 2020.

[62] Darko Hric, Tiago P Peixoto, and Santo Fortunato. Network structure, metadata, and the prediction of missing nodes and annotations. *Physical Review X*, 6(3):031038, 2016.

[63] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[64] Richard Lee Ingraham. *A Survey of Nonlinear Dynamics: "Chaos Theory"*. Number 1 in 91. World scientific, 1992.

[65] Katsuhiko Ishiguro, Tomoharu Iwata, Naonori Ueda, and Joshua Tenenbaum. Dynamic infinite relational model for time-varying relational data analysis. *Advances in Neural Information Processing Systems*, 23:919–927, 2010.

[66] Muhammad Aqib Javed, Muhammad Shahzad Younis, Siddique Latif, Junaid Qadir, and Adeel Baig. Community detection in networks: A multidisciplinary review. *Journal of Network and Computer Applications*, 108:87–111, 2018.

[67] Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30(1):175–193, 1906.

[68] Edward Kao, Vijay Gadepally, Michael Hurley, Michael Jones, Jeremy Kepner, Sanjeev Mohindra, Paul Monticciolo, Albert Reuther, Siddharth Samsi, William Song, et al. Streaming graph challenge: Stochastic block partition. In *2017*

IEEE High performance extreme computing conference (HPEC), pages 1–12. IEEE, 2017.

[69] Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

[70] Vishesh Karwa, Aleksandra B Slavković, and Pavel Krivitsky. Differentially private exponential random graphs. In *International Conference on Privacy in Statistical Databases*, pages 143–155. Springer, 2014.

[71] Helmut G Katzgraber and A Peter Young. Probing the almeida-thouless line away from the mean-field model. *Physical Review B*, 72(18):184416, 2005.

[72] Zheng Tracy Ke, Feng Shi, and Dong Xia. Community detection for hypergraph networks via regularized tensor power iteration. *arXiv preprint arXiv:1909.06503*, 2019.

[73] Chiheon Kim, Afonso S Bandeira, and Michel X Goemans. Stochastic block model for hypergraphs: Statistical limits and a semidefinite programming approach. *arXiv preprint arXiv:1807.02884*, 2018.

[74] Myunghwan Kim and Jure Leskovec. Nonparametric multi-group membership model for dynamic networks. *arXiv preprint arXiv:1311.2079*, 2013.

[75] Alec Kirkley, George T Cantwell, and MEJ Newman. Belief propagation for networks with loops. *Science Advances*, 7(17):eabf1211, 2021.

[76] Alec Kirkley and MEJ Newman. Representative community divisions of networks. *arXiv preprint arXiv:2105.04612*, 2021.

[77] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of complex networks*, 2(3):203–271, 2014.

[78] Kyle N Kunze, Evan M Polce, Amar Vadhera, Brady T Williams, Benedict U Nwachukwu, Shane J Nho, and Jorge Chahla. What is the predictive ability and academic impact of the altmetrics score and social media attention? *The American journal of sports medicine*, 48(5):1056–1062, 2020.

[79] Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

[80] Thomas Kurtz, Russell Lyons, Robin Pemantle, and Yuval Peres. A conceptual proof of the kesten-stigum theorem for multi-type branching processes. In *Classical and modern branching processes*, pages 181–185. Springer, 1997.

[81] Timothy La Fond and Jennifer Neville. Randomization tests for distinguishing social influence and homophily effects. In *Proceedings of the 19th international conference on World wide web*, pages 601–610, 2010.

[82] R Lambiotte, JC Delvenne, and M Barahona. Dynamics and modular structure in networks. *arXiv preprint arXiv:0812.1770*, 2008.

[83] R. Lambiotte and P. Panzarasa. Communities, knowledge creation, and information diffusion. *Journal of Informetrics*, 3(3):180–190, 2009.

[84] Renaud Lambiotte, J-C Delvenne, and Mauricio Barahona. Laplacian dynamics and multiscale modular structure in networks. *arXiv preprint arXiv:0812.1770*, 2008.

[85] Renaud Lambiotte, Jean-Charles Delvenne, and Mauricio Barahona. Random walks, markov processes and the multiscale modular organization of complex networks. *IEEE Transactions on Network Science and Engineering*, 1(2):76–90, 2014.

[86] Renaud Lambiotte and Naoki Masuda. *A guide to temporal networks*, volume 4. World Scientific, 2016.

[87] Jure Leskovec, Ajit Singh, and Jon Kleinberg. Patterns of influence in a recommendation network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 380–389. Springer, 2006.

[88] L Leydesdorff, HW Park, et al. Full and fractional counting in bibliometric networks. *Journal of Informetrics*, 11, 2017.

[89] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461, 2003.

[90] Wanyu Lin, Hao Lan, and Baochun Li. Generative causal explanations for graph neural networks. In *International Conference on Machine Learning*, pages 6666–6679. PMLR, 2021.

[91] Yan Liu, Alexandru Niculescu-Mizil, and Wojciech Gryc. Topic-link lda: joint models of topic and author community. In *proceedings of the 26th annual international conference on machine learning*, pages 665–672, 2009.

[92] Matthew Ludkin. *The autoregressive stochastic block model with changes in structure*. Lancaster University (United Kingdom), 2017.

[93] Matthew Ludkin. Inference for a generalised stochastic block model with unknown number of blocks and non-conjugate edge models. *Computational Statistics & Data Analysis*, 152:107051, 2020.

[94] Matthew Ludkin, Idris Eckley, and Peter Neal. Dynamic stochastic block models: parameter estimation and detection of changes in community structure. *Statistics and Computing*, 28(6):1201–1213, 2018.

[95] Yunpu Ma and Volker Tresp. Causal inference under networked interference and intervention policy enhancement. In *International Conference on Artificial Intelligence and Statistics*, pages 3700–3708. PMLR, 2021.

[96] Winter A Mason, Frederica R Conrey, and Eliot R Smith. Situating social influence processes: Dynamic, multidirectional flows of influence within social networks. *Personality and social psychology review*, 11(3):279–300, 2007.

[97] Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt Kusner, Arthur Gretton, and Krikamol Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *International Conference on Machine Learning*, pages 7512–7523. PMLR, 2021.

[98] Catherine Matias and Vincent Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1119–1141, 2017.

[99] Catherine Matias, Tabea Rebafka, and Fanny Villers. A semiparametric extension of the stochastic block model for longitudinal networks. *Biometrika*, 105(3):665–680, 2018.

[100] Piero Mazzarisi, Paolo Barucca, Fabrizio Lillo, and Daniele Tantari. A dynamic network model with persistent links and node-specific latent variables, with an application to the interbank market. *European Journal of Operational Research*, 281(1):50–65, 2020.

[101] Alexander Mendiburu, Roberto Santana, Jose A Lozano, and Endika Bengoetxea. A parallel framework for loopy belief propagation. In *Proceedings of the 9th annual conference companion on Genetic and evolutionary computation*, pages 2843–2850, 2007.

[102] Marc Mezard and Andrea Montanari. *Information, physics, and computation.* Oxford University Press, 2009.

[103] Kurt Miller, Michael Jordan, and Thomas Griffiths. Nonparametric latent feature models for link prediction. *Advances in neural information processing systems*, 22:1276–1284, 2009.

[104] Ankur Moitra, William Perry, and Alexander S Wein. How robust are reconstruction thresholds for community detection? In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 828–841, 2016.

[105] Cristopher Moore. The computer science and physics of community detection: Landscapes, phase transitions, and hardness. *arXiv preprint arXiv:1702.00467*, 2017.

[106] Andrew A Neath and Joseph E Cavanaugh. The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):199–203, 2012.

[107] M. Newman. Scientific collaboration networks. I. Network construction and fundamental results. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 64(1 Pt 2):016131, 2001.

[108] M. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, January 2001.

[109] Mark EJ Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E*, 64(1):016132, 2001.

[110] Mark EJ Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the national academy of sciences*, 101(suppl 1):5200–5205, 2004.

[111] Mark EJ Newman. Detecting community structure in networks. *The European physical journal B*, 38(2):321–330, 2004.

[112] Mark EJ Newman, George T Cantwell, and Jean-Gabriel Young. Improved mutual information measure for clustering, classification, and community detection. *Physical Review E*, 101(4):042304, 2020.

[113] Mark EJ Newman and Aaron Clauset. Structure and inference in annotated networks. *Nature communications*, 7(1):1–11, 2016.

[114] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[115] MEJ Newman. Network structure from rich but noisy data. *Nature Physics*, 14(6):542–545, 2018.

[116] Tin Lok James Ng and Thomas Brendan Murphy. Weighted stochastic block model. *Statistical Methods & Applications*, pages 1–34, 2021.

[117] Tin Lok James Ng and Thomas Brendan Murphy. Model-based clustering for random hypergraphs. *Advances in Data Analysis and Classification*, 16(3):691–723, 2022.

[118] Hidetoshi Nishimori. *Statistical physics of spin glasses and information processing: an introduction*. Number 111 in 1. Clarendon Press, 2001.

[119] Krzysztof Nowicki and Tom A B Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American statistical association*, 96(455):1077–1087, 2001.

[120] Neave O'Clery, Rafael Prieto Curiel, and Eduardo Lora. Commuting times and the mobilisation of skills in emergent cities. *Applied Network Science*, 4(1):118, 2019.

[121] Sean O'Connor, Eleanor Doyle, and Stephen Brosnan. Clustering in ireland: development cycle considerations. *Regional Studies, Regional Science*, 4(1):263–283, 2017.

[122] Soumik Pal and Yizhe Zhu. Community detection in the sparse hypergraph stochastic block model. *Random Structures & Algorithms*, 59(3):407–463, 2021.

[123] A Roxana Pamfil, Sam D Howison, and Mason A Porter. Inference of edge correlations in multilayer networks. *Physical Review E*, 102(6):062307, 2020.

[124] Adina Roxana Pamfil. *Communities in annotated, multilayer, and correlated networks*. PhD thesis, University of Oxford, 2018.

[125] A Abigail Payne, Aloysius Siow, et al. *Does federal research funding increase university research output?* Citeseer, 1999.

[126] Judea Pearl. Graphs, causality, and structural equation models. *Sociological Methods & Research*, 27(2):226–284, 1998.

[127] Judea Pearl. Causal inference. In *Proceedings of the 2008th International Conference on Causality: Objectives and Assessment-Volume 6*, pages 39–58, 2008.

[128] Judea Pearl. *Causality*. Cambridge university press, 2009.

[129] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2), 2000.

[130] Judea Pearl and James M Robins. Probabilistic evaluation of sequential plans from causal models with hidden variables. In *UAI*, volume 95, pages 444–453. Citeseer, 1995.

[131] Leto Peel and Aaron Clauset. Detecting change points in the large-scale structure of evolving networks. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[132] Leto Peel, Daniel B Larremore, and Aaron Clauset. The ground truth about metadata and community detection in networks. *Science advances*, 3(5):e1602548, 2017.

[133] Leto Peel and Michael T Schaub. Detectability of hierarchical communities in networks. *arXiv preprint arXiv:2009.07525*, 2020.

[134] Tiago P Peixoto. Parsimonious module inference in large networks. *Physical review letters*, 110(14):148701, 2013.

[135] Tiago P. Peixoto. The graph-tool python library. *figshare*, 2014.

[136] Tiago P Peixoto. Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X*, 4(1):011047, 2014.

[137] Tiago P Peixoto. Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. *Physical Review E*, 92(4):042807, 2015.

[138] Tiago P Peixoto. Model selection and hypothesis testing for large-scale network models with overlapping groups. *Physical Review X*, 5(1):011033, 2015.

[139] Tiago P Peixoto. Nonparametric bayesian inference of the microcanonical stochastic block model. *Physical Review E*, 95(1):012317, 2017.

[140] Tiago P Peixoto. Reconstructing networks with unknown and heterogeneous errors. *Physical Review X*, 8(4):041011, 2018.

[141] Tiago P Peixoto. Bayesian stochastic blockmodeling. *Advances in network clustering and blockmodeling*, pages 289–332, 2019.

[142] Tiago P Peixoto. Merge-split markov chain monte carlo for community detection. *arXiv preprint arXiv:2003.07070*, 2020.

[143] Tiago P Peixoto. Revealing consensus and dissensus between network partitions. *arXiv preprint arXiv:2005.13977*, 2020.

[144] Tiago P Peixoto. Disentangling homophily, community structure, and triadic closure in networks. *Physical Review X*, 12(1):011004, 2022.

[145] Tiago P Peixoto and Laetitia Gauvin. Change points, memory and epidemic spreading in temporal networks. *Scientific reports*, 8(1):1–10, 2018.

[146] Henry D Pfister. The gibbs free energy, the bethe free entropy, and the sum-product algorithm. *Supp. Material for Graphical Models and Inference*, 2014.

[147] Maksym Polyakov, Serhiy Polyakov, and Md Sayed Iftekhar. Does academic collaboration equally benefit impact of research across topics? the case of agricultural, resource, environmental and ecological economics. *Scientometrics*, 113(3):1385–1405, 2017.

[148] Walter W. Powell and Kaisa Snellman. The Knowledge Economy. *Annu. Rev. Sociol.*, 30(1):199–220, 2004.

[149] Amrita Purkayastha, Eleonora Palmaro, Holly J Falk-Krzesinski, and Jeroen Baas. Comparison of two article-level, field-independent citation metrics: Field-weighted citation impact (fwci) and relative citation ratio (rcr). *Journal of Informetrics*, 13(2):635–642, 2019.

[150] Meng Qin, Di Jin, Kai Lei, Bogdan Gabrys, and Katarzyna Musial-Gabrys. Adaptive community detection incorporating topology and content in social networks. *Knowledge-Based Systems*, 161:342–356, 2018.

[151] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[152] Riccardo Rastelli, Pierre Latouche, and Nial Friel. Choosing the number of groups in a latent stochastic blockmodel for dynamic networks. *Network Science*, 6(4):469–493, 2018.

[153] Ren Ren, Jinliang Shao, Adrian N Bishop, and Wei Xing Zheng. Modeling and detecting communities in node attributed networks. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[154] Maria A Riolo, George T Cantwell, Gesine Reinert, and Mark EJ Newman. Efficient method for estimating the number of communities in a network. *Physical review e*, 96(3):032310, 2017.

[155] Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher. An introduction to exponential random graph (p*) models for social networks. *Social networks*, 29(2):173–191, 2007.

[156] Joshua L Rosenbloom, Donna K Ginther, Ted Juhl, and Joseph A Heppert. The effects of research & development funding on scientific productivity: Academic chemistry, 1990-2009. *PloS one*, 10(9):e0138176, 2015.

[157] Alaa Saade, Florent Krzakala, and Lenka Zdeborová. Spectral clustering of graphs with the bethe hessian. *Advances in Neural Information Processing Systems*, 27, 2014.

[158] Kazumi Saito, Ryohei Nakano, and Masahiro Kimura. Prediction of information diffusion probabilities for independent cascade model. In *International conference on knowledge-based and intelligent information and engineering systems*, pages 67–75. Springer, 2008.

[159] Purnamrita Sarkar and Andrew W Moore. Dynamic social network analysis using latent space models. In *Advances in Neural Information Processing Systems*, pages 1145–1152, 2006.

[160] Michael T Schaub and Leto Peel. Hierarchical community structure in networks. *arXiv preprint arXiv:2009.07196*, 2020.

[161] Feng Shi and James Evans. Science and technology advance through surprise. *arXiv preprint arXiv:1910.09370*, 2019.

[162] Rahul Singh, Liyuan Xu, and Arthur Gretton. Kernel methods for multistage causal inference: Mediation analysis and dynamic treatment effects. *arXiv preprint arXiv:2111.03950*, 2021.

[163] Henry Small, Kevin W Boyack, and Richard Klavans. Identifying emerging topics in science and technology. *Research policy*, 43(8):1450–1467, 2014.

[164] Steven T Smith, Edward K Kao, Danelle C Shah, Olga Simek, and Donald B Rubin. Influence estimation on social media networks using causal inference. In *2018 IEEE Statistical Signal Processing Workshop (SSP)*, pages 328–332. IEEE, 2018.

[165] Tom Snijders, Christian Steglich, and Michael Schweinberger. Modeling the coevolution of networks and behavior. In *Longitudinal models in the behavioral and related sciences*, pages 41–71. Routledge, 2017.

[166] Dhanya Sridhar, Caterina De Bacco, and David Blei. Estimating social influence from observational data. *Proceedings of Machine Learning Research vol*, 140:1–22, 2022.

[167] Natalie Stanley, Thomas Bonacci, Roland Kwitt, Marc Niethammer, and Peter J Mucha. Stochastic block models with multiple continuous attributes. *Applied Network Science*, 4(1):1–22, 2019.

[168] Timothée Tabouy, Pierre Barbillon, and Julien Chiquet. Variational inference for stochastic block models from sampled data. *Journal of the American Statistical Association*, 115(529):455–466, 2020.

[169] Peng Hui Tan and Lars K Rasmussen. The serial and parallel belief propagation algorithms. In *Proceedings. International Symposium on Information Theory, 2005. ISIT 2005.*, pages 729–733. IEEE, 2005.

[170] Dane Taylor, Saray Shai, Natalie Stanley, and Peter J Mucha. Enhanced detectability of community structure in multilayer networks through layer aggregation. *Physical review letters*, 116(22):228301, 2016.

[171] Panos Toulis and Edward Kao. Estimation of causal peer influence effects. In *International conference on machine learning*, pages 1489–1497. PMLR, 2013.

[172] Vincent A Traag, Ludo Waltman, and Nees Jan van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

[173] Chun-Hua Tsai and Yu-Ru Lin. Tracing and predicting collaboration for junior scholars. In *Proceedings of the 25th international conference companion on world wide web*, pages 375–380, 2016.

[174] Toni Vallès-Català, Tiago P Peixoto, Marta Sales-Pardo, and Roger Guimerà. Consistencies and inconsistencies between model selection and link prediction in networks. *Physical Review E*, 97(6):062316, 2018.

[175] Gerhard G Van de Bunt, Marijtje AJ Van Duijn, and Tom AB Snijders. Friendship networks through time: An actor-oriented dynamic statistical network model. *Computational & Mathematical Organization Theory*, 5(2):167–192, 1999.

[176] Anthony FJ Van Raan. Sleeping beauties in science. *Scientometrics*, 59(3):467–472, 2004.

[177] Eric Wang, Jorge Silva, Rebecca Willett, and Lawrence Carin. Dynamic relational topic model for social network analysis with noisy links. In *2011 IEEE Statistical Signal Processing Workshop (SSP)*, pages 497–500. IEEE, 2011.

[178] Fei Wang, Tao Li, Xin Wang, Shenghuo Zhu, and Chris Ding. Community discovery using nonnegative matrix factorization. *Data Mining and Knowledge Discovery*, 22(3):493–521, 2011.

[179] Wei Wang, Shuo Yu, Teshome Megersa Bekele, Xiangjie Kong, and Feng Xia. Scientific collaboration patterns vary with scholars' academic ages. *Scientometrics*, 112(1):329–343, 2017.

[180] Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

[181] Samuel F Way, Allison C Morgan, Daniel B Larremore, and Aaron Clauset. Productivity, prominence, and the effects of academic environment. *Proceedings of the National Academy of Sciences*, 116(22):10729–10733, 2019.

[182] Florian Wetschoreck, Tobias Krabel, and Surya Krishnamurthy. 8080labs/ppscore: zenodo release, 2020.

[183] Mateusz Wilinski, Piero Mazzarisi, Daniele Tantari, and Fabrizio Lillo. Detectability of macroscopic structures in directed asymmetric stochastic block model. *Physical Review E*, 99(4):042310, 2019.

[184] Jason Wyse and Nial Friel. Block clustering with collapsed latent block models. *Statistics and Computing*, 22(2):415–428, 2012.

[185] Jason Wyse, Nial Friel, and Pierre Latouche. Inferring structure in bipartite networks using the latent blockmodel and exact icl. *Network Science*, 5(1):45–69, 2017.

[186] Eric P Xing, Wenjie Fu, and Le Song. A state-space mixed membership blockmodel for dynamic network tomography. *The Annals of Applied Statistics*, pages 535–566, 2010.

[187] Kevin Xu. Stochastic block transition models for dynamic networks. In *Artificial Intelligence and Statistics*, pages 1079–1087. PMLR, 2015.

[188] Kevin S Xu and Alfred O Hero. Dynamic stochastic blockmodels for time-evolving social networks. *IEEE Journal of Selected Topics in Signal Processing*, 8(4):552–562, 2014.

[189] Liyuan Xu, Heishiro Kanagawa, and Arthur Gretton. Deep proxy causal learning and its application to confounded bandit policy evaluation. *Advances in Neural Information Processing Systems*, 34:26264–26275, 2021.

[190] Sidney J Yakowitz and John D Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968.

[191] Yuto Yamaguchi and Kohei Hayashi. When does label propagation fail? a view from a network generative model. In *IJCAI*, pages 3224–3230, 2017.

[192] 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. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(5):P05007, 2014.

[193] Jaewon Yang, Julian McAuley, and Jure Leskovec. Community detection in networks with node attributes. In *2013 IEEE 13th International Conference on Data Mining*, pages 1151–1156, 2013.

[194] Tianbao Yang, Yun Chi, Shenghuo Zhu, Yihong Gong, and Rong Jin. Detecting communities and their evolutions in dynamic social networks—a bayesian approach. *Machine learning*, 82(2):157–189, 2011.

[195] Yasser Yasami and Farshad Safaei. A novel multilayer model for missing link prediction and future link forecasting in dynamic complex networks. *Physica A: Statistical Mechanics and its Applications*, 492:2166–2197, 2018.

[196] Jonathan S Yedidia, William T Freeman, and Yair Weiss. Bethe free energy, kikuchi approximations, and belief propagation algorithms. *Advances in neural information processing systems*, 13:689, 2001.

[197] Jonathan S Yedidia, William T Freeman, and Yair Weiss. Exploring artificial intelligence in the new millennium, chapter understanding belief propagation and its generalizations. *Science & Technology Books*, 2003.

[198] Jonathan S Yedidia, William T Freeman, and Yair Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on information theory*, 51(7):2282–2312, 2005.

[199] Pan Zhang, Florent Krzakala, Jörg Reichardt, and Lenka Zdeborová. Comparative study for inference of hidden classes in stochastic block models. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(12):P12021, 2012.

[200] Pan Zhang and Cristopher Moore. Scalable detection of statistically significant communities and hierarchies, using message passing for modularity. *Proceedings of the National Academy of Sciences*, 111(51):18144–18149, 2014.

[201] Pan Zhang, Cristopher Moore, and MEJ Newman. Community detection in networks with unequal groups. *Physical review E*, 93(1):012303, 2016.

[202] Pengfei Zhou, Tianyi Li, and Pan Zhang. Phase transitions and optimal algorithms for semisupervised classifications on graphs: From belief propagation to graph convolution network. *Physical Review Research*, 2(3):033325, 2020.

[203] Yaojia Zhu, Xiaoran Yan, and Cristopher Moore. Oriented and degree-generated block models: generating and inferring communities with inhomogeneous degree distributions. *Journal of Complex Networks*, 2(1):1–18, 2014.