Typology of Mixed-Membership Models: Towards a Design Method

Gregor Heinrich

University of Leipzig + vsonix GmbH
Darmstadt, Germany
gregor@arbylon.net

Abstract. Presents an analysis of the structure of mixed-membership models into elementary blocks and their numerical properties. By associating such model structures with structures known or assumed in the data, we propose how models can be constructed in a controlled way, using the numerical properties of data likelihood and Gibbs full conditionals as predictors of model behavior. To illustrate this "bottom-up" design method, example models are constructed that may be used for expertise finding from labeled data.

1 Introduction

In many areas of data mining, it is of interest to re-enact the structure that exists or is assumed in the data by a model that then quantifies this structure for analysis purposes. A good example is knowledge discovery in social community data. Such data often exist in the form of text in documents, which have associated with them meta-data like annotations and ratings, comments and tags, as well as or relational information like authorship, citation and linkage on the Web.

Analysis of such data (that in similar structure arise in other fields, from bioinformatics to computer vision) has often been associated with latent-variable models, and one specific type of such models has empirically led to robust results in the presence of sparsity and noise in the data and especially with complex interrelations between items of different modalities. These latent-variable models cover mixed membership, that is, each document etc. may be a member of a mixture of latent variables, which themselves may be interpreted as a "topic", a group of items/words etc. of similar meaning. Simple models of this model family (also referred to as topic models) were based on handling co-occurrence between words in documents, as in latent Dirichlet allocation (LDA) [1], or words associated to authors, as in the author–topic model (ATM) [2], and these seminal approaches have been extended into various directions.

Typically, in the literature such models are designed by assuming generative processes to re-enact observations, for example each word in LDA is thought to be generated by sampling a topic indicator from a document-specific topic multinomial and a word from a topic-specific vocabulary multinomial.

While this viewpoint of generative processes is intuitive in the sense of explaining models, it remains somewhat "short-sighted" in terms of the connection of model structures and behavior to data structures: The actual behavior of data likelihood (as an

D. Gunopulos et al. (Eds.): ECML PKDD 2011, Part II, LNAI 6912, pp. 32–47, 2011.

essential objective measure of model quality given trained parameters) is not directly found from the model structure. So is the structure of the inference equations necessary to find the optimum model parameters, typically by running approximative EM-type optimization.

On the other hand, meanwhile generic formulations of mixed-membership/topic models have been proposed, such as [3] and [4], that derive numerical properties across particular models and may allow some deeper look into correspondence between model and data structure.

Objectives and Outline. In this article, we complement the pure Bayesian-network viewpoint adopted in the literature by simplifying model structures. We aim at using these structures as building blocks to construct models in a principled way, keeping track of the model behavior when assembling the pieces to fit to data structures in question.

In particular, we will give a deeper introduction of latent-variable models in Sec. 2, reviewing a generic formulation. Based on this, we present a typology of model substructures in Sec. 3 that will serve as the basis for model construction in Sec. 4. Finally, we present a brief empirical study of the proposed approach in Sec. 5.

2 Networks of Mixed Membership

In [3], a generic view on topic models has been taken that formulates their structure as what we may call here "networks of mixed membership" (NoMMs). A NoMM is a directed acyclic graph whose nodes represent sets of mixture components and whose edges transmit variables to child nodes. Selection of components is achieved as a function of the incoming edge values, and values sampled from selected components are transmitted to child nodes. This process ultimately leads to observations at one or more terminal nodes.

Graphically, a simple NoMM structure is shown in Fig. 1, using the seminal LDA model as an example and introducing generic quantities. By default, NoMMs are Bayesian mixture models whose nodes include component parameters along with their prior distributions, and in the typical case, components, with index $k^\ell \in [1, K^\ell]$, have multinomial parameters, $\vec{\vartheta}^\ell_k$, with conjugate Dirichlet priors with hyperparameters $\vec{\alpha}^\ell$. Edges represent variables, $x^\ell_i \in [1, T^\ell]$, that run along sequences, $i^\ell \in I^\ell$. In this notation, a superscript like $\cdot^\ell$ indicates a "mixture level" in the network (a node with its direct child edges) and by convention also extends to variable indices, that is, $x^\ell_i \equiv x^\ell_{i^\ell}$.

Opposed to Bayesian networks (BNs [5]), the NoMM representation strictly distinguishes model variables (that grow with the data) and model parameters (that control variable generation), representing them as edges and nodes, respectively. In connection to this, two types of BN plates are distinguished: On one hand, "sequence plates" run over the data points and correspond to NoMM sequence indices i^ℓ as part of the data "streamed" along edges, x^ℓ_i. On the other, "component plates" index mixture components. In NoMMs, they correspond to the indices k^ℓ in nodes, which depend on incoming information as arguments of component selection functions, $k^\ell = f^\ell_k(\text{parents}(x^\ell_i), i^\ell)$.

The resulting structure is a domain-specific compact alternative to BNs that directly visualises the flow of (typically discrete) information through the generative model,

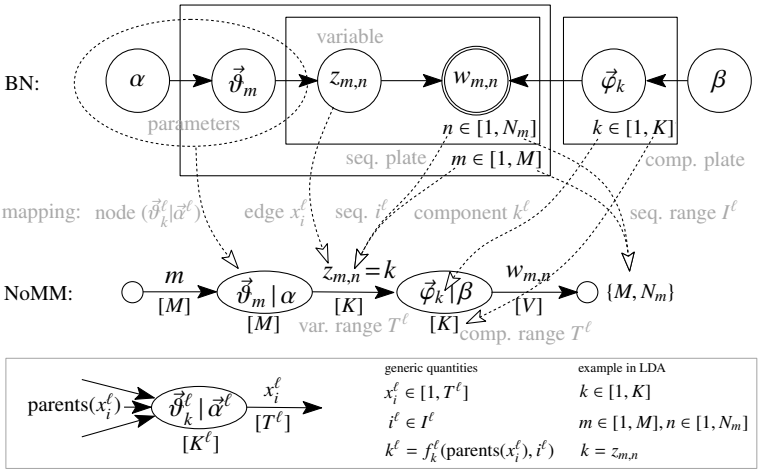


Fig. 1. NoMM notation and correspondence to Bayesian network for example model LDA and generically

mimicking a "systems view" with the nodes representing sub-systems and the edges signals processed by them. Clearly, the NoMM representation is focussed on the domain of mixture models, and especially such models that use complex interactions between different mixtures, such as mixed-membership models and topic models.

2.1 Numerical Properties

With conjugate distributions in NoMM nodes, there exist closed-form solutions for approximate Bayesian inference that lead to good empirical results [1,6], and for collapsed Gibbs sampling [3] and variational inference [4] meta-algorithms have been proposed that may be specialised to a wide range of models. Due to its relatively simple forms [3], especially Gibbs sampling may bear some intuitive meaning.

For the following considerations, let upper-case symbols denote sets of their lower-case counterparts introduced above. That is, Θ, A, and X correspond to all component parameters, hyperparameters and variables of a given model. If a superscript $\cdot^\ell$ is given, symbols are specific to a level. Among variables, X, we further distinguish the sets of hidden and visible variables, H and V.

Posterior. Aside from being the key to model training, the posterior, $p(H, \Theta|V, A)$, may provide insight into the expected behaviour of a model. In a collapsed Gibbs sampler as used for LDA-like models [6,7], the posterior is represented by *full conditional distributions*, the marginals of hidden variables at single data points i, $p(h_i|V, H_{-i}, A)$, given all other information except the parameters Θ, which are integrated out in collapsed inference. Here the index $-i$ denotes exclusion of i. The set of these distributions forms the transition matrix of a Markov chain, and round-robin sampling through i over time leads to a stationary state that simulates the true posterior. Full conditionals may therefore be seen as a low-dimensional representations of the true posterior.

In NoMMs, full conditionals have the following form [3]:[1]

$$p(h_i \mid V, H_{-i}, A) \propto \prod_{\ell} \prod_{k^\ell} \frac{\mathrm{B}(\vec{n}_k^\ell + \vec{\alpha}^\ell)}{\mathrm{B}(\vec{n}_{k,-i}^\ell + \vec{\alpha}^\ell)} \tag{1}$$

where $\mathrm{B}(\vec{x})$ is the multidimensional beta function [8] and $\vec{n}_k^\ell = \{n_{kt}\}_t^\ell$ the "co-occurrence" count vector between component index k^ℓ and node output values t^ℓ. Note that h_i are dependent hidden variables for an observation v_i across different levels ℓ.

To illustrate the principle that underlies (1), it may be noted that its factors reduce to simple quotients of sums if the exclusion of the current sample (with $-i$) from the vectors corresponds to a unit difference between numerator and denonimator:

$$p(h_i \mid V, H_{-i}, A) \propto \frac{n_{kt,-i}^a + \alpha_t^a}{\sum_t n_{kt,-i}^a + \alpha_t^a} \cdot \frac{n_{kt,-i}^b + \alpha_t^b}{\sum_t n_{kt,-i}^b + \alpha_t^b} \cdots , \tag{2}$$

that is, the normalised and smoothed co-occurrence counts reinforce the respective sampling weights in a "rich-get-richer" manner, which is known from Pólya urn sampling schemes associated with the Dirichlet–multinomial compound distribution.

Likelihood. Another descriptive property is the likelihood under the set of trained model parameters. Node parameters, $\Theta^\ell = \{\{\vartheta_{kt}^\ell\}_t\}_k$, themselves are simply the expectations of the Dirichlet priors given the co-occurrences, $\vartheta_{kt} \propto n_{kt} + \alpha_t$, and based on this, the likelihood of observations under the model, Θ, may be expressed as:[1]

$$p(v_i|\Theta) = \sum_{h_i} \prod_{\ell} \vartheta_{kt}^\ell \tag{3}$$

where the summation over h_i refers to all configurations of values of the dependent hidden variables.

Model Structure Influence. In the full conditional and the likelihood, the structure of the NoMM and its component selection functions f_k^ℓ control the association of values k^ℓ and t^ℓ of each level with model variables h_i and v_i, corresponding to paths over levels ℓ that assemble the products in (1) and (3). This and the appearance of the intuitive co-occurrence counts in both likelihood and full conditional may be a key for model design. Consequently, we consider building models from network sub-structures.

3 Model Structure

In the following, we study the decomposition of NoMMs into sub-structures, first taking a look at how models are generically decomposed and then at specific sub-structures.

Notation. For notational simplicity, we define a shorthand for the factors in the generic full conditional (1):

$$q(k,t) \triangleq \frac{\mathrm{B}(\vec{n}_k^\ell + \vec{\alpha}^\ell)}{\mathrm{B}(\vec{n}_{k,-i}^\ell + \vec{\alpha}^\ell)} \overset{\text{case of (2)}}{=} \frac{n_{kt,-i} + \alpha_t}{\sum_t n_{kt,-i} + \alpha_t} , \tag{4}$$

[1] This is a simplifying view for clarity, see Appendix A for details.

emphasizing the interrelation of indices that is expected to play a vital role in designing models. We introduce other conventions: Indexes added up with $\oplus$ refer to sums of the indexed counts, for instance $q(a, b \oplus c)$ contains $n_{ab} + n_{ac}$. Furthermore, if hyperparameters, α, are considered explicitly, the notation is augmented to $q(k, t \,|\, \alpha)$; if this information is clear from context, it is omitted to avoid notational clutter.

3.1 Model Decomposition

As a prerequisite to analyzing models, it is of interest to know how they decompose into sub-structures in terms of their full conditional and likelihood functions.

Full Conditional. Decomposing (1), it may be seen that partial full conditionals of sub-structures, $w(\cdot)$, can be factored with other parts of the model: $p(\cdot) = \prod_c w_c(\cdot) = \prod_c \prod_d q_d(\cdot)$. This enables us to indeed look at the sub-structures separately. For example, considering two sub-structures with hidden variables x and y to be connected with a hidden variable b, constructing a full conditional term $w(x, y, b|a, c)$ from the sub-terms $w(x|a, b)$ and $w(y|b, c)$ just multiplies their "q-terms" $q(a, x)q(x, b)$ and $q(b, y)q(y, c)$.

Likelihood. The data likelihood as an indicator of expected model performance (i.e., the best result it can in principle achieve) may like the full conditional be partitioned into substructures. From (3), it may be inferred that the inner terms of the likelihood of the complete model factor into that of sub-models. If two dependent substructures are joined, the marginal sums $\sum_h$ need to be taken care of, which is done by summing over hidden variables that connect the sub-structures. For example, considering two substructures with hidden variables x and y that are to be connected with a hidden variable b, the likelihood becomes: $p(c|a) = \sum_b p(c|b)p(b|a) = \sum_b (\sum_y \vartheta_{b,y}\vartheta_{y,c} \sum_x \vartheta_{a,x}\vartheta_{x,b})$.

3.2 Typology of Sub-structures

In order to analyze the structure of NoMMs usable in practice, we adopted an inductive approach based on an extensive study of the state of the art in topic models. This study resulted in a set of primitive structures that NoMMs are topologically composed of, in particular characterizing these structures (1) according to probability distributions their nodes use, (2) the way how models branch node information, that is, distribute samples of a given node, and finally (3) the way how models merge information, that is, how incoming data index components of a node. For reference, an overview of the described structures is given in Fig. 2, along with the numerical behavior of the Gibbs full conditional and the likelihood of observations. In these quantities, dependencies on A and Θ have been omitted. Although this set of structures is not considered a complete one, it fully covers all of the example models mentioned in this article, except nonparametric ones.

1. Node Types. Besides the standard Dirichlet–multinomial node with hidden parameters used in models like LDA [1](N1; we introduce alphanumeric structure class labels), there are special types that use alternative parameter distributions. First of all, N1 types may have different variants: While N1a uses a single hyperparameter, N1b introduces a selection function for $\vec{\alpha}_j$ that may be used to add an additional level of grouping among components (see App. A).

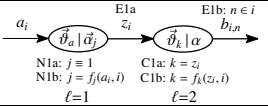
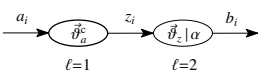
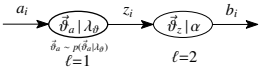
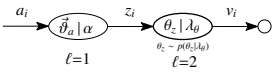
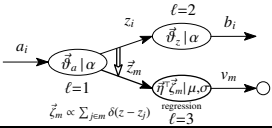
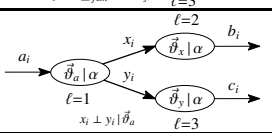
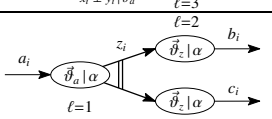
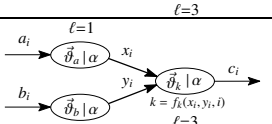
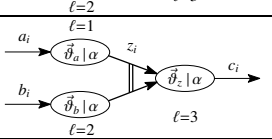
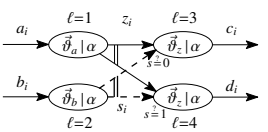
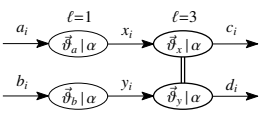
ID. Name	Structure diagram	Gibbs sampler weight w, Likelihood p for single token i Modelled aspect, Example models
N1, E1, C1. Dir–Mult nodes, unbranched		$w(z\|a,b) = q(a,z)q(z,b)$ E1b:$q(a,z)q(z,b_1 \oplus b_2 \oplus \dots b_{N_i})$ $p(b\|a) = \sum_z \vartheta_{a,z}\vartheta_{z,b}$ *Mixture/admixture*: LDA [1], PAM [9]; LDCC [10] (E1b)
N2. Observed parameters		$w(z\|a,b) = \vartheta^{\mathrm{c}}_{a,z}\, q(z,b)$ $p(b\|a) = \sum_z \vartheta^{\mathrm{c}}_{a,z}\vartheta_{z,b}$ *Label distribution*: ATM [2]
N3. Non-Dirichlet prior		$w(z\|a,b;\vec{\vartheta}_a) = p(z_i\|a_i,\vec{\vartheta}_a)q(z,b)$; M-step: estimate $\vec{\vartheta}_a$[11] $p(b\|a) = \sum_z \vartheta_{a,z}\vartheta_{z,b}$ *Alternative distributions on the simplex*: CTM [11]: $\vec{\vartheta}_a \propto \exp\vec{\eta}$, $\vec{\eta}\sim\mathcal{N}(\vec{\mu},\underline{\Sigma})$; TLM [12]: hierarchy of Dirichlet priors
N4. Non-discrete output		$w(z\|a,v;\theta) = q(a,z)p(v_i\,\|\,\theta_z)$; M-step: estimate θ_z $p(v\|a) = \sum_z \vartheta_{a,z}p(v\,\|\,\theta_z)$ *Non-multinomial observ.*: Corr-LDA [13], GMM [14]: $p(v\|\theta) = \mathcal{N}(\vec{x}\|\vec{\mu},\underline{\Sigma})$
N5, E4. Regression		$w(z\|\vec{z}_m,v_m,a,b) = q(a,z)\,q(z,w)\mathcal{N}(v_m\,\|\,\vec{\eta}_v^{\top}\vec{\zeta}_m,\sigma^2)$; M-step: estimate $\vec{\eta}_v,\sigma^2\,\|\,\vec{z},\vec{v}$ (for linear regression) prediction: $v_m = \vec{\eta}_v^{\top}\vec{\zeta}_m$ *Regression/supervised learning*: Supervised LDA [15]
E2. Independent edges		$w(x,y\|a,b,c) = q(a,x\oplus y)q(x,b)q(y,c)$ $p(b,c\|a) = \sum_x \vartheta_{a,x}\vartheta_{x,b}\sum_y \vartheta_{a,y}\vartheta_{y,c}$ *Common mixture of causes*: Multimodal LDA [16]
E3. Coupled edges		$w(z\|a,b,c) = q(a,z)q(z,b)q(z,c)$ $p(b,c\|a) = \sum_z \vartheta_{a,z}\vartheta_{z,b}\vartheta_{z,c}$ *Common cause*: Hidden relational model (HRM) [17], Link-LDA [18]
C2. Combined indices		$w(x,y\|a,b,c) = q(a,x)q(b,y)q(k,c)$ $p(c\|a,b) = \sum_x[\vartheta_{a,x}\sum_y \vartheta_{b,y}\vartheta(k,c)]\,, k = f_k(x,y,i)$ *Different correlated causes, relation*: hPAM [19], HRM [17], Multi-LDA [20]
C3. Coupled indices		$w(z\|a,c) = q(a,z)q(z,c\oplus\tilde{c})$, $w(z\|b,c) = q(b,z)q(z,\tilde{c}\oplus c)$ $p(c\|a,b) = \sum_z(\vartheta_{a,z}+\vartheta_{b,z})/2\cdot\vartheta_{z,c}$ *Different causes, same effect*: (proposed here)
C4. Switch		$w(z,s\|a,b,c,d) = q(a,z)[q(b,\mathrm{c})q(z,c)]^{\delta(s,1)} \cdot [q(b,\mathrm{d})q(z,d)]^{\delta(s,2)}$ $p(c,d\|a,b) = \sum_z \vartheta_{a,z}[\vartheta_{b,s=0}\vartheta_{z,c} + \vartheta_{b,s=1}\vartheta_{z,d}]$ *Select complex submodels*: Multi-grain LDA [21], Entity-topic models [22]
C5. Node coupling		$w(x,y\|a,b,c,d) = q(a,x)q(b,y)[q(x,c\oplus d)]^{\delta(x,y)} \cdot [q(x,c\oplus\tilde{d})q(y,\tilde{c}\oplus d)]^{1-\delta(x,y)}$ $p(c,d\|a,b) = \sum_x \vartheta_{a,x}\vartheta_{x,c}\sum_y \vartheta_{b,y}\vartheta_{y,d}$ *Correlation of submodels, relations*: Simple relational component model [23], Relational topic model [24]

Fig. 2. NoMM sub-structure types. Notation (also see (4)): $a \oplus b$ adds counts $n_{\cdot a} + n_{\cdot b}$; $\tilde{c}$ means that the count is not decremented with $-i$ in (1).

Other node types vary the distributions they represent: One important type uses observed parameters to accommodate label information, as authorship metadata in the author–topic model [2]. In this case, the prior is lost (N2). Furthermore, there are models with alternative prior distributions (N3), such as "structured" Dirichlet distributions (N3a) [12], and non-conjugate priors for the parameters, such as logistic-normal (N3b) [11]. Varying the output distributions allows modelling of non-discrete observations (N4), for instance using Gaussian with conjugate Gaussian and inverse-Wishart priors, esp. in media mining [13,25]. Another type of non-discrete output may be produced by regression nodes (N5). In connection with aggregation edges (E4, below), N5 nodes apply a regression model to subsets $\vec{z}_m$ of hidden values, $\vec{z}$, allowing supervised learning approaches within the framework of mixed-membership models.

2. Forks / Edge Structures. When connecting nodes of the network, there are different configurations if nodes that receive the output of a given node. The most frequent type of connection is an unbranched edge to a single node (E1). Beside standard unbranched edges (E1a), type E1b incorporates aggregation of a subsequence, such as words as part of sentences: From a single sample of the parent node z_i, e.g., a sentence topic, a whole sub-sequence of tokens $\vec{b}_i$ is produced, as in [10]. To sample z_i, the corresponding Gibbs full conditional term becomes $q(z_i, \vec{b}_i)$, which causes (1) to deviate from its standard form, (2).

Apart from E1 edges, branching edges to several nodes is a common structure. Here either the samples are generated independently (E2), or both children are forced to the same latent variable (E3).

As an interpretation, branching may be seen as a common cause to several observed modalities. Note that the $\oplus$ in the E2 Gibbs weighting function in Fig. 2 expands to:

$$q(a, x \oplus y) = \frac{\mathrm{B}(\vec{n}_a^{(x)} + \vec{n}_a^{(y)} + \alpha)}{\mathrm{B}(\vec{n}_{a,-i}^{(x)} + \vec{n}_{a,-i}^{(y)} + \alpha)} = \frac{(\hat{n}_{ax,-i} + \alpha - \delta(x,y))(\hat{n}_{ay,-i} + \alpha)}{(\sum_t \hat{n}_{at,-i} + \alpha)((\sum_t \hat{n}_{at,-i} + \alpha) + 1)} \tag{5}$$

where $\hat{n}_{at}$ corresponds to the added contributions of both branches and $\delta(x, y)$ the Kronecker delta.

The last edge type (E4) converts a sequence of values to a vector that may be used for instance for regression, thus complementing the regression node type (N5).

3. Joins / Component Selectors. The dual structure type to a fork is a join, a structure in the network that collects edges at the input of a node and computes an index k from the set of incoming values. Such index structures may trivially collect a single edge value as in LDA (C1a), use an edge value and a sequence index as in pachinko allocation models (PAM) [9] (C1b) or be constructed out of several hidden values (C2), as in [19]. Such multi-inputs are made dependent by observed node output and may be used to merge several influences. It is illustrative to verify this by using the information in Fig. 2. The Gibbs weighting term for the C2 structure is $w(x, y|a, b, c) = q(a, x)q(b, y)q(k, c)$, and if the indices x and y simply refer to the dimensions of k, i.e., $k = f_k(x, y) = (x, y)$, one component exists for each combination of input values. With c an observed edge, x and y become dependent, and the full conditional tends to be high wherever $n_{(x,y)c}$ as part of $q(k, c)$ is high. According to the clustering property of the Dirichlet, sampling (x, y) jointly with c further increases $n_{(x,y)c}$.

While branching structures have coupled and independent variants (E3 and E2), so far there seems to exist no structure in literature that does the same for component indices. A desirable structure that complements C2 may take values from multiple inputs and map them into the same variable range. As an approach to this, we propose a substructure that duplicates the sampling process for incoming branches, with the merging node collecting counts from both, and Fig. 2) shows this as structure C3.

Looking at the Gibbs weighting term in Fig. 2 is illustrative. The shorthand $w(z|a,c) = q(a,z)q(z,c \oplus \tilde{c})$ corresponds to:

$$w(z|a,c) = \frac{n_{az,-i} + \alpha}{\sum_z n_{az,-i} + \alpha} \frac{n^{(a)}_{zc,-i} + n^{(b)}_{zc} + \alpha}{\sum_z n^{(a)}_{zc,-i} + n^{(b)}_{zc} + \alpha} \tag{6}$$

and analogously for $w(z|b,c)$.[2] The contributions of both incoming edges, $n^{(a)}_{zc}$ and $n^{(b)}_{zc}$, are summed in the second quotient, effectively superimposing their influences. Compared to C2, this behavior is slightly different: While the effect of a C2 structure is comparable to an intersection of the co-occurrences between pairs (x,c) and (y,c), the C3 structure is likely to behave closer to a union operator.

Another variant of component selector structures is to switch input edges according to the value of a parent node (C4) [21,22]. This allows control of the influence of more complex sub-models in the branches switched.

The final component selector structure results from sharing parameters among different nodes (C5). This allows coupling of different sub-models without additional need for edges and has been of particular interest in analyzing relational data, see, e.g., [23].

4 Towards a Model Design Method

Different to approaches to design models representable by NoMMs, including mixed-membership/topic models, we propose a design method based the "library" of model structures collected in Fig. 2. This method directly takes into account model assumptions and formalizes the actual steps to reach viable structures in a straight-forward workflow. In the following, we outline the general method and subsequently illustrate it with an example design.

4.1 Designing a Design Method

From Sec. 3, we have discussed how the likelihood reflects the general potential of reaching some model quality (likelihood of held-out data is a standard metric in topic modeling), and that Gibbs full conditionals may serve as an indicator of how such an optimization may be achieved: As a low-dimensional "excerpt" of the posterior, a Gibbs sampler will maximize the weights for those latent dimensions that lead to the best model given the data in a Bayesian sense.

Along with any special metrics to capture model quality or computational complexity, these two measures may also be used as predictors of model behavior. For NoMMs, one may develop a design method from them, a strategy to design models is proposed as follows:

[2] For simplicity, we assume they are sampled in two distinct sweeps.

1. Define data *input*: modalities (type of documents, metadata, relational structure) and dimensions available.
2. Define model *output*: results expected, under which *metrics*. This may include retrieval measures and computational complexity.
3. Make *assumptions*: e.g., "topics ⇔ document semantics", "labels ⇔ topics", where "⇔" refers to a correspondence via correlation or co-occurrence. This will be elaborated below.
4. *Structure* model: with artefacts from Fig. 2, map assumptions to structures, often a correspondence,"⇔", leads to a node in the model.
5. *Predict* behaviour: Gibbs + likelihood (Fig. 2) and metrics.
6. *Iterate* model: optionally go to Step 3 or 4.

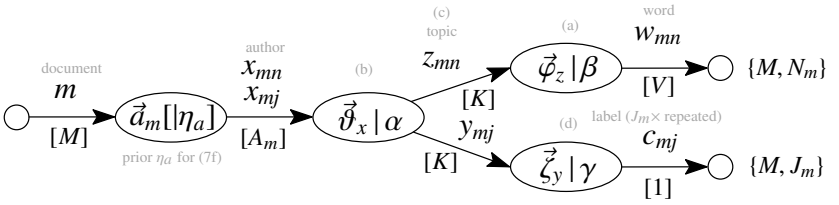
For the design, Step 3 is central, and a rule of thumb is to gather as many assumptions as are needed to "connect" all modalities and hidden assumed structures (like topics). Typically assumptions are qualitative and may be based on correlation or, on the token/item-level, co-occurrence, for instance, terms co-occur in documents and in topics, etc. If topics co-occur in other topics, we obtain a hierarchy of topics. Alternatively, viewpoints of mixing or generative processes may be adopted: Mixing assumes that the output of a node is a mixture of its components, and the generative-process perspective is that used in traditional topic modelling with Bayesian networks. By analogy between BN and NoMM representations (cf. Fig. 1), assumptions can be transformed between them.

Ideally, there exists a one-by-one correspondence between data modalities and assumed hidden structures on one side and model structures on the other. Empirically, the strategy works best with one assumption on each data structure in question, excluding relationships that are transitive, like document ⇔ label = (document ⇔ topic) ∘ (topic ⇔ label). Under-determination of structure by assumptions leaves more structures arbitrary, increasing room for experimentation. For over-determination, assumptions may be prioritized.

For Step 4, some intuition is required to map assumptions to models. Currently, a set of rules is being developed to formalize this process. The next step in this direction is to develop a clearer mapping between the structure types and assumptions, complementing the considerations undertaken on likelihood and full conditional properties in Sec. 3. Furthermore, criteria like scalability are important, as adding any dependent hidden variables increases model complexity considerably: Computational load is on the order of $O(\prod_{h^\ell} T^\ell)$ for dependent h^ℓ.

4.2 Example: Expert–Tag–Topic Model

To illustrate model design, we consider an example scenario: For expert finding, a community of authors is to be indexed to recommend the best expert given a term query or a subject descriptor. While the former may be solved using the author–topic model (ATM) [2], the latter is special to our scenario: Subject descriptors, such as ACM CCS or Medline MeSH, have a controlled vocabulary and are added to the documents authored by experts. For such a scenario, we construct an "expert–tag–topic" (ETT) model using the method above:



$$p(x_{mn}{=}x, z_{mn}{=}z \,|\, w_{mn}{=}w, \{\vec{x}, \vec{z}, \vec{w}\}_{-mn}, \vec{y}) \propto a_{m,x} q(x, z \oplus y) q(z, w) \tag{7a}$$

$$p(x_{mj}{=}x, y_{mj}{=}y \,|\, c_{mj}{=}c, \{\vec{x}, \vec{y}, \vec{c}\}_{-mj}, \vec{z}) \propto a_{m,x} q(x, y \oplus z) q(y, c) \tag{7b}$$

$$p(w_{mn}|\vec{a}_m, c_m, \Theta) = \textstyle\sum_x a_{m,x} \sum_z \vartheta_{x,z} \varphi_{z,w} \qquad w_{mn} \perp c_{mj} \,|\, \Theta \tag{7c}$$

$$p(c_m|a, \Theta) = \textstyle\sum_y \vartheta_{a,y} \zeta_{y,c} \tag{7d}$$

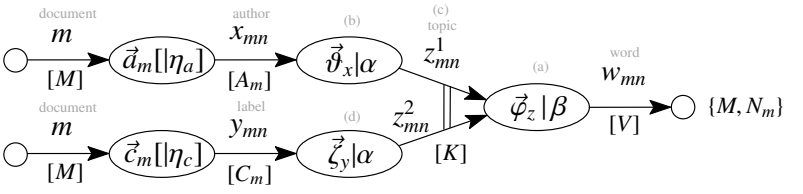
$$p(a|c_q, \Theta) \propto p(c_q|a, \Theta) p(a|\Theta)\,, \quad p(a|\Theta) \propto \textstyle\sum_{mn} \delta(x_{mn} - a) \tag{7e}$$

$$p(a|\vec{w}_q, \Theta) \propto \textstyle\sum_n \delta(x_{qn} - a)\,,\ x_{qn} \sim \text{Gibbs (7a) with } a(\cdot) = q(m, x \,|\, \eta_a) \tag{7f}$$

Fig. 3. Expert–Tag–Topic model, iteration 1

1. *Input*: document text $\vec{w} = \{\{w_{mn}\}_{n=1}^{N_{m=1}}\}_m^M$, subject labels $\vec{c} = \{c_m\}_{m=1}^M$, authorship $\vec{a} = \{\{a_{mx}\}_{x=1}^{A_m}\}_{m=1}^M$.
2. *Output*: expert recommendations $p(a|\vec{w}_q)$ and $p(a|c_q)$ for queries $\vec{w}_q$ and subject labels c_q; *metric*: subjective consistency of topics estimated.
3. *Assumptions*: (a) topics z_{mn} ⇔ document text w_{mn} (semantic similarity of items represented by topics), (b) authors a ⇔ topics $z_{mn} \forall m : a \in \vec{a}_m$, (c) authors a_{mx} ⇔ document text w_{mn}, (d) topics z_{mn} ⇔ labels c_{mj}.
4. *Structure*: Topics z_{mn} appear as the central variable, and we combine on one hand the author–topic model, which fulfills assumptions (a)–(c) and introduces an author–word association x_{mn}, and on the other hand a branch with an N1 node that generates an observable label from a topic, corresponding to assumption (d). The resulting model, "ETT1", is shown in Fig. 3, with the author–topic model in the upper branch and the label branch below. With the E2 branching structure in center, the Gibbs sampling term of the author–topic distribution, $\vec{\vartheta}_x$, becomes $q(x, z \oplus y)$, and it can be seen that words and categories influence the association of topics to authors directly. By setting the number of label samples per document, J_m, we can control their influence on the topics.
5. *Prediction*: Model properties are derived from Fig. 2 and shown in Fig. 3, with full conditionals (7a–b) and likelihoods (7c–d), as well as recommendation tasks in (7e–f). The Gibbs sampler in (7f) makes the node $\vec{a}_m$ unsupervised and starts with a fair distribution $a_{qx} = 1/A$, updating for $\vec{w}_q$.

ETT1 has the disadvantage that it only supports a single label per document and that it does not directly model the dependence between words and labels (see (7c)). Therefore, we iterate the structure:



$$p(\{x_{mn}{=}x, z^1_{mn}{=}k \,|\, w_{mn}{=}w, \{\vec{x},\vec{y},\vec{z}^1,\vec{w}\}_{-mn}, \vec{z}^2) \propto a_{m,x} q(x,k) q(k, w \oplus \tilde{w}) \tag{8a}$$

$$p(\{y_{mn}{=}y, z^2_{mn}{=}k \,|\, w_{mn}{=}w, \{\vec{c},\vec{x},\vec{z}^2,\vec{w}\}_{-mn}, \vec{z}^1) \propto c_{m,y} q(y,k) q(k, \tilde{w} \oplus w) \tag{8b}$$

$$p(w_{mn}|\vec{a}_m, \vec{c}_m, \Theta) = \textstyle\sum_z (\sum_x a_{m,x}\vartheta_{x,z} + \sum_y c_{m,y}\zeta_{y,z})/2 \cdot \varphi_{z,w} \tag{8c}$$

$$p(a|c_q, \Theta) = \textstyle\sum_z p(a|z,\Theta) p(z|c_q,\Theta) \propto \sum_z \vartheta_{a,z} p(a|\Theta) \zeta_{c_q,z}\,, \quad p(a|\Theta) \propto \textstyle\sum_{mn} \delta(x_{mn}{-}a) \tag{8d}$$

$$p(a|\vec{w}_q, \Theta) \propto \textstyle\sum_n \delta(x_{qn}{-}a)\,,\ x_{qn} \sim \text{Gibbs (8a,b) with } a_{m,x} = q(m,x|\eta_a) \tag{8e}$$

Fig. 4. Expert–Tag–Topic model, iteration 2

6. *Iteration*: To allow multiple labels, we may actually use the same structure as the ATM for labels and merge both with a C2 or C3 structure. Here label and author-generated topics merge. The "ETT2" model is shown in Fig. 4 with properties (8a–e). For unseen documents with unknown labels or authors, the sampler is run using (8e) analogous to (7f), using unsupervised $\vec{a}_q$ and $\vec{c}_q$ with priors $\eta_.$.

Beyond these variants of an Expert–Tag–Topic model, there are various alternatives, for instance, instead of the central E2 and C3 structures we may use the E3 and C2 ones and may obtain a more straight-forward recommendation rule than (8d) in Fig. 4. As has been discussed above, the basic approach of model design needs to be complemented with guidelines for the selection of model structures given a task and dataset at hand, so the best structure types may be identified from the outset.

Furthermore, actual model performance is likely to depend on the finer details of co-occurrence structure in the data and the questions asked about them, and with these details the mileage of different models may vary. Looking at the summing structure in likelihoods (7c) and (8c) indicates that at least the models are in principle able to reach the likelihood of LDA, while full conditionals (7a–b) and (8a–b) seem to create the right "gradient" in this direction, increasing co-occurrences in hidden nodes where they are assumed in the data.

5 Empirical Analysis

Testing a complete framework of models like the one in question is the necessary step to prove the applicability of the design method. However, this larger task is ongoing work. In this paper, we limit ourselves to a verification of the results of the design approach taken and performed a proof-of-concept test of the ETT models.

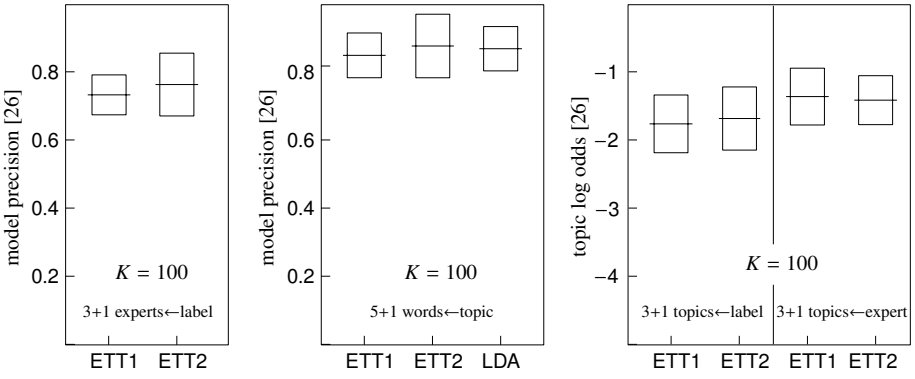


Fig. 5. Semantic coherence of ETT output

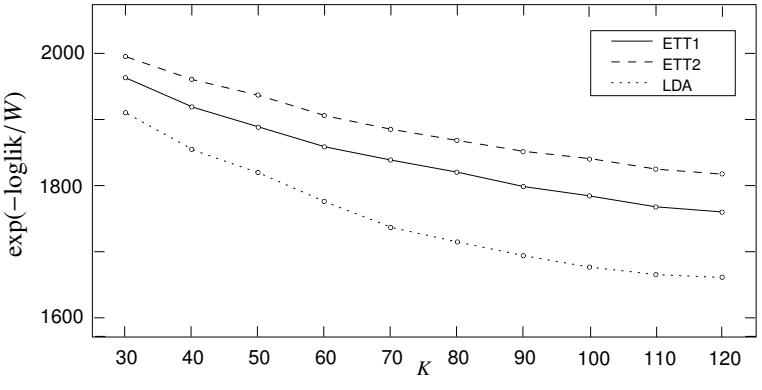


Fig. 6. ETT likelihood against baseline

As data set, we consider the NIPS corpus commonly used in topic model research, with $M = 1740$ documents, $V = 13649$ unique terms, $W = 2301375$ words, $A = 2037$ authors with $W_A = 3990$ authorship relations, and $C = 50$ categories with $M_C = 1254$ labelled documents (conference tracks and manual labels). For both models ETT1 and ETT2, Gibbs sampling was run over a range of K and the tasks performed on the trained parameters for 20 topics, 10 labels and 10 experts.

Precision of recommendations was measured by human judgements of five expert voters in the spirit of a topic-coherence experiment [26]: Voters are presented with groups of (a) 6 topic words and (b) 4 document topics and asked to detect the least consistent item. In every question, items presented have high probability according to the model, except for an unlikely "intrusion" item that participants may easier identify in semantically coherent groups. For our scenario, beside topic coherence we adapted the experiment to test associations of experts with labels (showing titles and frequent words of experts' papers), as well as topics for both experts and labels to check topic coherence. We measured the model precision and topic log odds from [26]. Results are

shown in Fig. 5, and it generally can be seen that both ETT models produce coherent output with low intrusion votings (higher values better), validating the model.

We also tested the log likelihood of held-out documents (prediction of second half of test documents, as proposed by [2]) as a "control metric" and compared against the baseline model LDA, and the result is given in Fig. 6. The ETT models are slightly inferior, but this is in line with results of [26] that report some deviation of human perception of topic coherence from model generalizability measured by likelihood. Our model creates *different* topics; it is not designed to compete with LDA that can freely adapt to the data available, taking constraints from author and label contexts. Notably, for the scenario considered here, the proposed NoMM structure C3 turns out as a viable alternative with low model complexity.

6 Conclusions and Future Work

In this article, we have shown how mixed-membership models can be separated into sub-structures, and how the sub-structures may be used as a "library" to create models according to a straight-forward design workflow. The method proposed is based on mapping qualitative assumptions to model structures and allows to stay aware of quantities like full conditional distributions of Gibbs samplers and data likelihood, important predictors of the model performance to be expected.

We have applied the design method successfully to an example scenario of expertise finding from labeled documents, but more work needs to be done in order to refine and validate the method itself. Ongoing work [27] applies the method to other scenarios and looks into refined mapping rules between data properties and model structures in order to obtain clearer modeling guidelines. A more extensive validation will be based on synthetic data with controlled properties that benchmark the different model structures, and a special aspect to look at in this context is the relation between model structures and higher-order co-occurrences in multimodal data, analogous to language data [28].

Finally, we will study how "networks of mixed membership", the model representation used here, may be used as a generalized representation of the finite models discussed here and non-parametric variants with Dirichlet or Pitman-Yor process priors [29].

References

1. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
2. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: 20th Conference on Uncertainty in Artificial Intelligence (2004)
3. Heinrich, G.: A generic approach to topic models. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009. LNCS, vol. 5781, pp. 517–532. Springer, Heidelberg (2009)
4. Heinrich, G., Goesele, M.: Variational bayes for generic topic models. In: Mertsching, B., Hund, M., Aziz, Z. (eds.) KI 2009. LNCS, vol. 5803, pp. 161–168. Springer, Heidelberg (2009)
5. Pearl, J.: Bayesian networks: A model of self-activated memory for evidential reasoning. In: Proc. 7th Conf. of the Cognitive Science Society, pp. 329–334 (1985)

6. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences 101, 5228–5235 (2004)
7. Liu, J.: The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problems. Journal of the American Statistical Association 89(427), 958–966 (1994)
8. Andrews, G.E., Askey, R., Roy, R.: Special functions. Cambridge University Press, Cambridge (1999)
9. Li, W., McCallum, A.: Pachinko allocation: DAG-structured mixture models of topic correlations. In: ICML 2006: Proceedings of the 23rd International Conference on Machine Learning, pp. 577–584. ACM, New York (2006)
10. Shafiei, M.M., Milios, E.E.: Latent Dirichlet co-clustering. In: ICDM 2006: Proceedings of the Sixth International Conference on Data Mining, pp. 542–551. IEEE Computer Society, Washington, DC, USA (2006)
11. Blei, D., Lafferty, J.: A correlated topic model of Science. Annals of Applied Statistics 1, 17–35 (2007)
12. Wallach, H.M.: Structured Topic Models for Language. PhD thesis, University of Cambridge (2008)
13. Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D., Jordan, M.: Matching words and pictures. JMLR – Special Issue on Machine Learning Methods for Text and Images 3, 1107–1136 (2003)
14. McLachlan, G., Peel, D.: Finite Mixture Models. Wiley, Chichester (2000)
15. Blei, D., McAuliffe, J.: Supervised topic models. In: Advances in Neural Information Processing Systems (2007)
16. Ramage, D., Heymann, P., Manning, C.D., Garcia-Molina, H.: Clustering the tagged web. In: Proc. WSDM (2009)
17. Xu, Z., Tresp, V., Yu, K., Kriegel, H.P.: Infinite hidden relational models. In: Proc. 22nd Conference in Uncertainty in Artificial Intelligence UAI (2006)
18. Erosheva, E., Fienberg, S., Lafferty, J.: Mixed membership models of scientific publications. PNAS 101, 5220–5227 (2004)
19. Li, W., Blei, D., McCallum, A.: Mixtures of hierarchical topics with pachinko allocation. In: International Conference on Machine Learning (2007)
20. Porteous, I., Bart, E., Welling, M.: Multi-HDP: A non-parametric Bayesian model for tensor factorization. In: Proc. AAAI (2008)
21. Titov, I., McDonald, R.: Modeling online reviews with multi-grain topic models. In: Proc. 17th International World Wide Web Conference (WWW 2008), Beijing, China (2008)
22. Newman, D., Chemudugunta, C., Smyth, P.: Statistical entity-topic models. In: KDD 2006: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 680–686. ACM, New York (2006)
23. Sinkkonen, J., Parkkinen, J., Aukia, J., Kaski, S.: A simple infinite topic mixture for rich graphs and relational data. In: Proc. NIPS Workshop on Analyzing Graphs: Theory and Applications (2008)
24. Chang, J., Blei, D.M.: Relational topic models for document networks. In: AISTATS (2009)
25. Cao, L., Fei-Fei, L.: Spatially coherent latent topic model for concurrent object segmentation and classification. In: Proc. ICCV (2007)
26. Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., Blei, D.: Reading tea leaves: How humans interpret topic models. In: Proc. Neural Information Processing Systems, NIPS (2009)
27. Heinrich, G.: Typology of mixed-membership models: Applications to community data. Technical note TN2011/2, arbylon.net (2011)
28. Heyer, G., Bordag, S.: A Structuralist Framework for Quantitative Linguistics. In: Aspects of Automatic Text Analysis. Studies in Fuzziness and Soft Computing. Springer, Heidelberg (2007)

29. Teh, Y.W., Jordan, M.I.: Hierarchical Bayesian nonparametric models with applications. In: Hjort, N., Holmes, C., Müller, P., Walker, S. (eds.) To appear in Bayesian Nonparametrics: Principles and Practice. Cambridge University Press, Cambridge (2009)

A NoMM Gibbs Sampling

To keep this paper self-contained, Gibbs sampling in NoMMs is outlined for the case of multinomial levels with Dirichlet priors.

NoMMs estimate latent variables using two assumptions: (1) Models consist of a set of discrete mixtures whose multinomial parameters are generated from conjugate Dirichlet distributions. Each discrete mixture is governed by the following generative process (omitting level superscripts ℓ):

$$x_i \sim \mathrm{Mult}(x_i|\vec{\vartheta}_k) \qquad \vec{\vartheta}_k \sim \mathrm{Dir}(\vec{\vartheta}_k|\vec{\alpha}_j) \tag{9}$$

where x_i is one discrete data point (token), latent or observed, $\vec{\vartheta}_k$ is a vector of multinomial parameters with k the mixture component index and $\vec{\alpha}_j$ the parameter of the Dirichlet prior distribution (scalar or vector). These discrete mixtures are (2) coupled by discrete variables x_i to choose a component k:

$$k = f_k(\mathrm{parents}(x_i), i)\ . \tag{10}$$

This results in dependencies between the x_i (and $\vec{\vartheta}_k$) of different levels, which allows modelling of complex co-occurrences in the data. Further, components may be grouped by drawing them from different hyperparameters $\vec{\alpha}_j$, where the group indicator j may be a function of values known at the time of generating $\vec{\vartheta}_k$ due to (9):

$$j = f_j(\mathrm{known\_parents}(x_i), i)\ . \tag{11}$$

The conjugacy between the multinomial and Dirichlet distributions of model levels leads to a simple complete-data likelihood:

$$p(X, \Theta|A) = \prod_\ell \prod_i \mathrm{Mult}(x_i^\ell|\vec{\vartheta}^\ell, k^\ell) \prod_k \mathrm{Dir}(\vec{\vartheta}_k^\ell|\vec{\alpha}_j^\ell) \tag{12}$$

$$= \prod_\ell \left[\prod_k \frac{\mathrm{B}(\vec{n}_k + \vec{\alpha}_j)}{\mathrm{B}(\vec{\alpha}_j)} \mathrm{Dir}(\vec{\vartheta}_k|\vec{n}_k + \vec{\alpha}_j) \right]^\ell \tag{13}$$

where brackets $[\cdot]^\ell$ enclose a particular level ℓ.

Gibbs full conditionals are derived for groups of dependent hidden edges, $H^d \subset X$ (with dependent tokens $h_i^d \in H^d$) and their "surrounding" edges S^d (with $s_i^d \in S^d$) considered observed. We also define the set of all tokens co-located with a particular observation, $x_i^d = \{h_i^d, s_i^d\}$ where i (actually i^d) is the sequence of token indices for group d. For each of the dependency groups thus defined, a full conditional is created, using (13) with Θ integrated out:

$$
\begin{aligned}
p(h_i^d \mid X \backslash h_i^d, A) &= \frac{p(h_i^d, s_i^d \mid X \backslash \{h_i^d, s_i^d\}, A)}{p(s_i^d \mid X \backslash \{h_i^d, s_i^d\}, A)} \\
&\propto p(x_i^d \mid X \backslash x_i^d, A) = \frac{p(X \mid A)}{p(X \backslash x_i^d \mid A)} \\
&= \prod_\ell \left[\prod_k \frac{\mathrm{B}(\vec{n}_k + \vec{\alpha}_j)}{\mathrm{B}(\vec{n}_k \backslash x_i^d + \vec{\alpha}_j)} \right]^\ell \\
&\propto \prod_{\ell \in L(H^d)} \left[\frac{\mathrm{B}(\vec{n}_k + \vec{\alpha}_j)}{\mathrm{B}(\vec{n}_k \backslash x_i^d + \vec{\alpha}_j)} \right]^\ell
\end{aligned} \tag{14}
$$

where $L(H^d)$ is the set of all levels ℓ whose variables interact with the edges in the set H^d. For a single hidden variable set H^d, this leads to (1). Note that $\backslash x_i^d \equiv \backslash x_{i^d}^d$ excludes more than a single token from a particular edge if a token with index i^d at node input corresponds to multiple tokens $c \in i^\ell$ at its output, which leads more complex terms than (2). This occurs in Fig. 2 for structures E2 and when data aggregations are explicitly modelled using E1b structures (e.g. [10]). In (3), this case of "sub-tokens" is excluded for simplicity, and a complete formulation may redefine the parameter ϑ_{kt} as a product of sub-token likelihoods: $\vartheta_{kt} = \prod_{c \in i^\ell} \vartheta_{kt_c}$. This expands to a hierarchy of products for recursive sub-sequences.