When are Overcomplete Topic Models Identifiable? Uniqueness of Tensor Tucker Decompositions with Structured Sparsity

Animashree Anandkumar, Daniel Hsu, Majid Janzamin and Sham Kakade*

October 20, 2013

Abstract

Overcomplete latent representations have been very popular for unsupervised feature learning in recent years. In this paper, we specify which overcomplete models can be identified given observable moments of a certain order. We consider probabilistic admixture or topic models in the overcomplete regime, where the number of latent topics can greatly exceed the size of the observed word vocabulary. While general overcomplete topic models are not identifiable, we establish *generic* identifiability under a constraint, referred to as *topic persistence*. Our sufficient conditions for identifiability involve a novel set of "higher order" expansion conditions on the *topic-word matrix* or the *population structure* of the model. This set of higher-order expansion conditions allow for overcomplete models, and require the existence of a perfect matching from latent topics to higher order observed words. We establish that random structured topic models are identifiable w.h.p. in the overcomplete regime. Our identifiability results allow for general (non-degenerate) distributions for modeling the topic proportions, and thus, we can handle arbitrarily correlated topics in our framework. Our identifiability results imply uniqueness of a class of tensor decompositions with structured sparsity which is contained in the class of *Tucker* decompositions, but is more general than the *Candecomp/Parafac* (CP) decomposition.

Keywords: Overcomplete representations, topic models, generic identifiability, tensor decomposition.

1 Introduction

The performance of many machine learning methods is hugely dependent on the choice of data representations or features. Overcomplete representations, where the number of features can be greater than the dimensionality of the input data, have been extensively employed, and are arguably critical in a number of applications such as speech and computer vision [1]. Overcomplete

*A. Anandkumar and M. Janzamin are with the Center for Pervasive Communications and Computing, Electrical Engineering and Computer Science Dept., University of California, Irvine, USA 92697. Email: a.anandkumar@uci.edu,mjanzami@uci.edu. Daniel Hsu and Sham Kakade are with Microsoft Research New England, 1 Memorial Drive, Cambridge, MA 02142. Email: dahsu@microsoft.com, skakade@microsoft.com

representations are known to be more robust to noise, and can provide greater flexibility in modeling [2]. Unsupervised estimation of overcomplete representations has been hugely popular due to the availability of large-scale unlabeled samples in many applications.

A probabilistic framework for incorporating features posits latent or hidden variables that can provide a good explanation to the observed data. Overcomplete probabilistic models can incorporate a much larger number of latent variables compared to the observed dimensionality. In this paper, we characterize the conditions under which overcomplete latent variable models can be identified from their observed moments.

For any parametric statistical model, identifiability is a fundamental question of whether the model parameters can be uniquely recovered given the observed statistics. Identifiability is crucial in a number of applications where the latent variables are the quantities of interest, e.g. inferring diseases (latent variables) through symptoms (observations), inferring communities (latent variables) via the interactions among the actors in a social network (observations), and so on. Moreover, identifiability can be relevant even in predictive settings, where feature learning is employed for some higher level task such as classification. For instance, non-identifiability can lead to the presence of non-isolated local optima for optimization-based learning methods, and this can affect their convergence properties, e.g. see [3].

In this paper, we characterize identifiability for a popular class of latent variable models, known as the *admixture* or *topic* models [4, 5]. These are hierarchical mixture models, which incorporate the presence of multiple latent states (i.e. topics) in each document consisting of a tuple of observed variables (i.e. words). Previous works have established that the model parameters can be estimated efficiently using low order observed moments (second and third order) under some non-degeneracy assumptions, e.g. [6–8]. However, these non-degeneracy conditions imply that the model is undercomplete, i.e., the latent dimensionality (number of topics) cannot exceed the observed dimensionality (word vocabulary size). In this paper, we remove this restriction and consider overcomplete topic models, where the number of topics can far exceed the word vocabulary size.

It is perhaps not surprising that general topic models are not identifiable in the overcomplete regime. To this end, we introduce an additional constraint on the model, referred to as *topic persistence*, which roughly means that topics (i.e. latent states) persist locally in a sequence of observed words (but not necessarily globally). This "locality" effect among the observed words is not present in the usual "bag-of-words" or *exchangeable* topic model. Such local dependencies among observations abound in applications such as text, images and speech, and can lead to a more faithful representation. In addition, we establish that the presence of topic persistence is central towards obtaining model identifiability in the overcomplete regime, and we provide an in-depth analysis of this phenomenon in this paper.

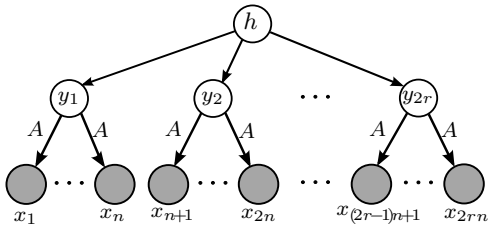


Figure 1: Hierarchical structure of the n-persistent topic model. $2rn$ number of words (views) are shown for some integer $r \geq 1$. A single topic $y_j, j \in [2r]$, is chosen for each n successive views $\{x_{(j-1)n+1}, \ldots, x_{(j-1)n+n}\}$. Matrix A is the population structure or topic-word matrix.

1.1 Summary of results

In this paper, we provide conditions for *generic*[1] model identifiability of overcomplete topic models given observable moments of a certain order (i.e., having a certain number of words in each document). We introduce the notion of *topic persistence*, and analyze its effect on identifiability. We establish identifiability in the presence of a novel combinatorial object, referred to as *perfect n-gram matching*, in the bipartite graph from topics to words. Finally, we prove that random structured topic models satisfy these criteria, and are thus identifiable in the overcomplete regime.

Persistent Topic Model: We first introduce the n-persistent topic model, where the parameter n determines the persistence level of a common topic in a sequence of n successive words. For instance, in Figure 1, the sequence of successive words $x_1, \ldots, x_n$ share a common topic y_1, and similarly, the words $x_{n+1}, \ldots, x_{2n}$ share topic y_2, and so on. The n-persistent model reduces to the popular "bag-of-words" model, when $n = 1$, and to the single topic model (i.e. only one topic in each document) when $n \to \infty$. Intuitively, topic persistence aids identifiability since we have multiple *views* of the common hidden topic generating a sequence of successive words. We establish that the bag-of-words model (with $n = 1$) is too non-informative about the topics in the overcomplete regime, and is therefore, not identifiable. On the other hand, n-persistent overcomplete topic models with $n \geq 2$ can become identifiable, and we establish a set of transparent conditions for identifiability.

Deterministic Conditions for Identifiability: Our sufficient conditions for identifiability are in the form of expansion conditions from the latent topic space to the observed word space. In the overcomplete regime, there are more topics than words in the vocabulary, and thus it is impossible to have expansion on the bipartite graph from topics to words, i.e., the graph encoding the sparsity pattern of the topic-word matrix. Instead, we impose an expansion constraint from topics to "higher order" words, which allows us to incorporate overcomplete models. We establish that this condition translates to the presence of a novel combinatorial object, referred to as the *perfect n-gram matching*, on the topic-word bipartite graph. Intuitively, the perfect n-gram matching condition implies "diversity" among the higher-order word supports for different topics which leads to identifiability. In addition, we present trade-offs among the following quantities: number of

[1] A model is generically identifiable, if all the parameters in the parameter space are identifiable, almost surely. Refer to Definition 1 for more discussion.

topics, size of the word vocabulary, the topic persistence level, the order of the observed moments at hand, the minimum and maximum degrees of any topic in the topic-word bipartite graph, and the *Kruskal rank* [9] of the topic-word matrix, under which identifiability holds. To the best of our knowledge, this is the first work to provide conditions for characterizing identifiability of overcomplete topic models with structured sparsity.

Identifiability of Random Structured Topic Models: We explicitly characterize the regime of identifiability for the random setting, where each topic i is randomly supported on a set of d_i words, i.e. the bipartite graph is a random graph. For this random model with q topics, p-dimensional word vocabulary, and topic persistence level n, when $q = O(p^n)$ and $\Theta(\log p) \leq d_i \leq \Theta(p^{1/n})$, for all topics i, the topic-word matrix is identifiable from $2n^{\text{th}}$ order observed moments with high probability. Intuitively, the upper bound on the degrees d_i is needed to limit the overlap of word supports among different topics in the overcomplete regime: as the number of topics q increases (i.e., n increases in the above degree bound), the degree needs to be correspondingly smaller to ensure identifiability, and we make this dependence explicit. Intuitively, as the extent of overcompleteness increases, we need sparser connections from topics to words to ensure sufficient diversity in the word supports among different topics. The lower bound on the degrees is required so that there are enough edges in the topic-word bipartite graph so that various topics can be distinguished from one another. Furthermore, we establish that the size condition $q = O(p^n)$ for identifiability is tight.

Implications on Uniqueness of Overcomplete Tucker and CP Tensor Decompositions: We establish that identifiability of an overcomplete topic model is equivalent to uniqueness of decomposition of the observed moment tensor (of a certain order). Our identifiability results for persistent topic models imply uniqueness of a structured class of tensor decompositions, which is contained in the class of *Tucker* decompositions, but is more general than the *candecomp/parafac* (CP) decomposition [10]. This sub-class of Tucker decompositions involves structured sparsity and symmetry constraints on the *core tensor*, and sparsity constraints on the *inverse factors* of the Tucker decomposition. The structural constraints on the Tucker tensor decomposition are related to the topic model as follows: the sparsity and symmetry constraints on the core tensor are related to the persistence property of the topic model, and the sparsity constraints on the inverse factors are equivalent to the sparsity constraints on the topic-word matrix. For n-persistent topic model with $n = 1$ (bag-of-words model), the tensor decomposition is a general Tucker decomposition, where the core tensor is fully dense, while for $n \to \infty$ (single-topic model), the tensor decomposition reduces to a CP decomposition, i.e. the core tensor is a *diagonal tensor*. For a finite persistence level n, in between these two extremes, the core tensor satisfies certain sparsity and symmetry constraints, which becomes crucial towards establishing identifiability in the overcomplete regime.

1.2 Overview of Techniques

We now provide a short overview of the techniques employed in this paper.

Recap of Identifiability Conditions in Under-complete Setting (Expansion Conditions on Topic-Word Matrix): Our approach is based on the recent results of [7], where conditions for identifiability of topic models are derived, given pairwise observed moments (specifically, co-occurrence of word-pairs in documents). Consider a topic model with q topics and observed word vocabulary of size p. Let $A \in \mathbb{R}^{p \times q}$ denote the topic-word matrix. Expansion conditions are imposed in [7] on the topic-word bipartite graph which imply that (generically) the sparsest vectors in the column span of A, denoted by $\mathrm{Col}(A)$, are the columns of A themselves. Thus the topic-word matrix A is identifiable from pairwise moments under expansion constraints. However, these expansion conditions constrain the model to be under-complete, i.e., the number of topics $q \leq p$, the size of the word vocabulary. Therefore, the techniques derived in [7] are not directly applicable here since we consider overcomplete models.

Identifiability in Overcomplete Setting and Why Topic-Persistence Helps: Pairwise moments are thus not sufficient for identifiability of overcomplete models, and the question is whether higher order moments can yield identifiability. We can view the higher order moments as pairwise moments of another equivalent topic model, which enables us to apply the techniques of [7]. The key question is whether we have expansion in the equivalent topic model, which implies identifiability. For a general topic model (without any topic persistence constraints), it can be shown that for identifiability, we require expansion of the n^{th}-order *Kronecker product* of the original topic-word matrix A, denoted by $A^{\otimes n} \in \mathbb{R}^{p^n \times q^n}$, when given access to $(2n)^{\text{th}}$-order moments, for any integer $n \geq 1$. In the overcomplete regime where $q > p$, $A^{\otimes n}$ cannot expand, and therefore, overcomplete models are not identifiable in general. On the other hand, we show that imposing the constraint of topic persistence can lead to identifiability. For a n-persistent topic model, given $(2n)^{\text{th}}$-order moments, we establish that identifiability occurs when the n^{th}-order *Khatri-Rao product* of A, denoted by $A^{\odot n} \in \mathbb{R}^{p^n \times q}$, expands. Note that the Khatri-Rao product $A^{\odot n}$ is a sub-matrix of the Kronecker product $A^{\otimes n}$, and the Khatri-Rao product $A^{\odot n}$ can expand as long as $q \leq p^n$. Thus, the property of topic persistence is central towards achieving identifiability in the overcomplete regime.

First-Order Approach for Identifiability of Overcomplete Models (Expansion of n-gram Topic-Word Matrix): We refer to $A^{\odot n} \in \mathbb{R}^{p^n \times q}$ as the n-gram topic-word matrix, and intuitively, it relates topics to n-tuple words. Imposing the expansion conditions derived in [7] on $A^{\odot n}$ implies that (generically) the sparsest vectors in $\mathrm{Col}(A^{\odot n})$, are the columns of $A^{\odot n}$ themselves. Thus, the topic-word matrix A is identifiable from $(2n)^{\text{th}}$-order moments for a n-persistent topic model. We refer to this as the "first-order" approach since we directly impose the expansion conditions of [7] on $A^{\odot n}$, without exploiting the additional structure present in $A^{\odot n}$.

Why the First-Order Approach is not Enough: Note that $A^{\odot n} \in \mathbb{R}^{p^n \times q}$ matrix relates topics to n-tuples of words. Thus, the entries of $A^{\odot n}$ are highly correlated, even if the original topic-word matrix A is assumed to be randomly generated. It is non-trivial to derive conditions on A, so that $A^{\odot n}$ expands. Moreover, we establish that $A^{\odot n}$ fails to expand on "small" sets, as

required in [7], when the degrees are sufficiently different [2]. Thus, the first-order approach is highly restrictive in the overcomplete setting.

Incorporating Rank Criterion: Note that $A^{\odot n}$ is highly structured: the columns of $A^{\odot n}$ matrix possess a tensor [3] rank of 1, when $n > 1$. This can be incorporated in our identifiability criteria as follows: we provide conditions under which the sparsest vectors in $\mathrm{Col}(A^{\odot n})$, which also possess a tensor rank of 1, are the columns of $A^{\odot n}$ themselves. This implies identifiability of a n-persistent topic model, when given access to $(2n)^{\text{th}}$-order moments. Note that when a small number of columns of $A^{\odot n}$ are combined, the resulting vector cannot possess a tensor rank of 1, and thus, we can rule out that such sparse combinations of columns using the rank criterion. The maximum such number is at least the *Kruskal rank* [4] of A. Thus, sparse combinations of columns of A (up to the Kruskal rank) can be ruled out using the rank criterion, and we require expansion on $A^{\odot n}$ only on large sets of topics (of size larger than the Kruskal rank). This agrees with the intuition that when the topic-word matrix A has a larger Kruskal rank, it should be easier to identify A, since the Kruskal rank is related to the *mutual incoherence* [5] among the columns of A, see [11].

Notion of Perfect n-gram Matching and Final Identifiability Conditions: Thus, we establish identifiability of overcomplete topic models subject to expansion conditions $A^{\odot n}$ on sets of size larger than the Kruskal rank of the topic-word matrix A. However, it is desirable to impose transparent and interpretable conditions directly on A for identifiability. We introduce the notion of *perfect n-gram matching* on the topic-word bipartite graph, which ensures that each topic can be uniquely matched to a n-tuple word. This combined with a lower bound on the Kruskal rank provides the final set of deterministic conditions for identifiability of the overcomplete topic model. Intuitively, we require that the columns of A be sparse, while still maintaining a large enough Kruskal rank; in other words, the topics have to be sparse and have sufficiently diverse word supports. Thus, we establish identifiability under a set of transparent conditions on the topic-word matrix A, consisting of perfect n-gram matching condition and a lower bound on the Kruskal rank of A.

Analysis under Random-Structured Topic-Word Matrices: Finally, we establish that the derived deterministic conditions are satisfied when the topic-word bipartite graph is randomly generated, as long as the degrees satisfy certain lower and upper bounds. Intuitively, a lower bound on the degrees of the topics is required to have degree concentration on various subsets so that expansion can occur, while the upper bound is required so that the Kruskal rank of the

[2] For $A^{\odot n}$ to expand on a set of size $s \geq 2$, it is necessary that $s \cdot \binom{d_{\min}+n-1}{n} \geq s + \binom{d_{\max}+n-1}{n}$, where $d_{\min}$ and $d_{\max}$ are the minimum and maximum degrees, and n is the extent of overcompleteness: $q = \Theta(p^n)$. When the model is highly overcomplete (large n) and we require small set expansion (small s), the degrees need to be nearly the same. Thus, it is desirable to impose expansion only on large sets, since it allows for more degree diversity.

[3] When any column of $A^{\odot n} \in \mathbb{R}^{p^n \times q}$ (of length p^n) is reshaped as a n^{th}-order tensor $T \in \mathbb{R}^{p \times p \times \cdots \times p}$, the tensor T is rank 1.

[4] The Kruskal rank is the maximum number k such that every k-subset of columns of A are linearly independent. Note that the Kruskal rank is equal to the rank of A, when A has full column rank. But this cannot happen in the overcomplete setting.

[5] It is easy to show that $\mathrm{krank}(A) \geq (\max_{i \neq j} |a_i^\top a_j|)^{-1}$, where a_i, a_j are any pair of columns of A. Thus, higher incoherence leads to a larger kruskal rank.

topic-word matrix is large enough compared to the sparsity level. Here, the main technical result is establishing the presence of a perfect n-gram matching in a random bipartite graph with a wide range of degrees. We present a greedy and a recursive mechanism for constructing such a n-gram matching for overcomplete models, which can be relevant even in other settings. For instance, our results imply the presence of a perfect matching when the edges of a bipartite graph are correlated in a structured manner, as given by the Khatri-Rao product.

1.3 Related works

We now summarize some recent related works in the area of identifiability and learning of latent variable models.

Identifiability, learning and applications of overcomplete latent representations: Many recent works employ unsupervised estimation of overcomplete features for higher level tasks such classification, e.g. [1,12–14], and record huge gains over other approaches in a number of applications such as speech recognition and computer vision. However, theoretical understanding regarding learnability or identifiability of overcomplete representations is far more limited.

Overcomplete latent representations have been analyzed in the context of the independent components analysis (ICA), where the sources are assumed to be independent, and the mixing matrix is unknown. In the overcomplete or under-determined regime of the ICA, there are more sources than sensors. Identifiability and learning of the overcomplete ICA reduces to the problem of finding an overcomplete candecomp/parafac (CP) tensor decomposition. The classical result by Kruskal provides conditions for uniqueness of a CP decomposition [9, 15], with recent extensions to the notion of robust identifiability [16]. These results provide conditions for strict identifiability of the model, and here, the dimensionality of the latent space is required to be of the same order as the observed space dimensionality. In contrast, a number of recent works analyze *generic* identifiability of overcomplete CP decomposition, which is weaker than strict identifiability, e.g. [17–23]. These works assume that the factors (i.e. the components) of the CP decomposition are generically drawn and provide conditions for uniqueness. They allow for the latent dimensionality to be much larger (polynomially larger) than the observed dimensionality. These results on the uniqueness of CP decompositions also lead to identifiability of other latent variable models, such as latent tree models, e.g. [24, 25], and the single-topic model, or more generally latent Dirichlet allocation (LDA). Recently Goyal et. al. [26] proposed an alternative framework for overcomplete ICA models based on the eigen-decomposition of the reweighted covariance matrix (or higher order moments), where the weights are the Fourier coefficients. However, their approach requires independence of sources (i.e. latent topics in our context), which is not imposed here.

In contrast to the above works dealing with the CP tensor decomposition, we require uniqueness for a more general class of tensor decompositions, in order to establish identifiability of topic models with arbitrarily correlated topics. We establish that our class of tensor decomposition is contained in the class of *Tucker* decompositions which is more general than CP decomposition. Moreover, we explicitly characterize the effect of the sparsity pattern of the factors (i.e., the topic-word matrix) on model identifiability, while all the previous works based on generic identifiability assume fully

dense factors (since sparse factors are not generic). For a general overview of tensor decompositions, see [10, 27].

Identifiability and learning of undercomplete/over-determined latent representations: Much of the theoretical results on identifiability and learning of the latent variable models are limited to non-singular models, which implies that the latent space dimensionality is at most the observed dimensionality. We outline some of the recent works below.

The works of Anandkumar et. al. [6,28,29] provide an efficient moment-based approach for learning topic models, under constraints on the distribution of the topic proportions, e.g. the single topic model, and more generally latent Dirichlet allocation (LDA). In addition, the approach can handle a variety of latent variable models such as Gaussian mixtures, hidden Markov models (HMM) and community models [30]. The high-level idea is to reduce the problem of learning of the latent variable model to finding a CP decomposition of the (suitably adjusted) observed moment tensor. Various approaches can then be employed to find the CP decomposition. In [6], a tensor power method approach is analyzed and is shown to be an efficient guaranteed recovery method in the non-degenerate (i.e. undercomplete) setting. Previously, simultaneous diagonalization techniques have been employed for solving the CP decomposition, e.g. [28, 31, 32]. However, these techniques fail when the model is overcomplete, as considered here. We note that some recent techniques, e.g. [20], can be employed instead, albeit at a cost of higher computational complexity for overcomplete CP tensor decomposition. However, it is not clear how the sparsity constraints affect the guarantees of such methods. Moreover, these approaches cannot handle general topic models, where the distribution of the topic proportions is not limited to these classes (i.e. either single topic or Dirichlet distribution), and we require tensor decompositions which are more general than the CP decomposition.

There are many other works which consider learning mixture models when multiple views are available. See [28] for a detailed description of these works. Recently, Rabani et. al. [33] consider learning discrete mixtures given a large number of "views", and they refer to the number of views as the *sampling aperture*. They establish improved recovery results (in terms of ℓ_1 bounds) when sufficient number of views are available ($2k-1$ views for a k-component mixture). However, their results are limited to discrete mixtures or single-topic models, while our setting can handle more general topic models. Moreover, our approach is different since we incorporate sparsity constraints in the topic-word distribution. Another series of recent works by Arora et. al. [8, 34] employ approaches based on non-negative matrix factorization (NMF) to recover the topic-word matrix. These works allow models with arbitrarily correlated topics, as considered here. They establish guaranteed learning when every topic has an *anchor* word, i.e. the word is uniquely generated from that topic, and does not occur under any other topic. Note that the anchor-word assumption cannot be satisfied in the overcomplete setting.

Our work is closely related to the work of Anandkumar et. al. [7] which considers identifiability and learning of topic models under expansion conditions on the topic-word matrix. The work of Spielman et. al [35] considers the problem of dictionary learning, which is closely related to the setting of [7], but in addition assumes that the coefficient matrix is random. However, these works [7,35] can handle only the under-complete setting, where the number of topics is less than the dimensionality of the word vocabulary (or the number of dictionary atoms is less than the number

of observations in [35]). We extend these results to the overcomplete setting by proposing novel higher order expansion conditions on the topic-word matrix, and also incorporate additional rank constraints present in higher order moments.

Dictionary learning/sparse coding: Overcomplete representations have been very popular in the context of dictionary learning or sparse coding. Here, the task is to jointly learn a dictionary as well as a sparse selection of the dictionary atoms to fit the observed data. There have been Bayesian as well as frequentist approaches for dictionary learning [2,36,37]. However, the heuristics employed in these works [2, 36, 37] have no performance guarantees. The work of Spielman et. al [35] considers learning (undercomplete) dictionaries and provide guaranteed learning under the assumption that the coefficient matrix is random (distributed as Bernoulli-Gaussian variables). Recent works [38, 39] provide generalization bounds for predictive sparse coding, where the goal of the learned representation is to obtain good performance on some predictive task. This differs from our framework since we do not consider predictive tasks here, but the task of recovering the underlying latent representation. Hillar and Sommer [40] consider the problem of identifiability of sparse coding and establish that when the dictionary succeeds in reconstructing a certain set of sparse vectors, then there exists a unique sparse coding, up to permutation and scaling. However, our setting here is different, since we do not assume that a sparse set of topics occur in each document.

2 Model

Notation: The set $\{1, 2, \ldots, n\}$ is denoted by $[n] := \{1, 2, \ldots, n\}$. Given a set $X = \{1, \ldots, p\}$, set $X^{(n)}$ denotes all ordered n-tuples generated from X. The cardinality of a set S is denoted by $|S|$. For any vector u (or matrix U), the support is denoted by $\mathrm{Supp}(u)$, and the ℓ_0 norm is denoted by $\|u\|_0$, which corresponds to the number of non-zero entries of u, i.e., $\|u\|_0 := |\mathrm{Supp}(u)|$. For a vector $u \in \mathbb{R}^q$, $\mathrm{Diag}(u) \in \mathbb{R}^{q\times q}$ is the diagonal matrix with vector u on its diagonal. The column space of a matrix A is denoted by $\mathrm{Col}(A)$. Vector $e_i \in \mathbb{R}^q$ is the i-th basis vector, with the i-th entry equal to 1 and all the others equal to zero. For $A \in \mathbb{R}^{p\times q}$ and $B \in \mathbb{R}^{m\times n}$, the *Kronecker* product $A \otimes B \in \mathbb{R}^{pm\times qn}$ is defined as [41]

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1q}B \\ a_{21}B & a_{22}B & \cdots & a_{2q}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1}B & a_{p2}B & \cdots & a_{pq}B \end{bmatrix},$$

and for $A = [a_1|a_2|\cdots|a_r] \in \mathbb{R}^{p\times r}$ and $B = [b_1|b_2|\cdots|b_r] \in \mathbb{R}^{m\times r}$, the *Khatri-Rao* product $A \odot B \in \mathbb{R}^{pm\times r}$ is defined as

$$A \odot B = [a_1 \otimes b_1|a_2 \otimes b_2|\cdots|a_r \otimes b_r].$$

2.1 Persistent topic model

In this section, the *n-persistent topic model* is introduced and this imposes an additional constraint, known as topic persistence on the popular admixture model [4,5,42]. The n-persistent topic model reduces to the bag-of-words admixture model when $n = 1$.

An admixture model specifies a q-dimensional vector of topic proportions $h \in \Delta^{q-1} := \{u \in \mathbb{R}^q : u_i \geq 0, \sum_{i=1}^q u_i = 1\}$ which generates the observed variables $x_l \in \mathbb{R}^p$ through vectors $a_1, \ldots, a_q \in \mathbb{R}^p$. This collection of vectors $a_i, i \in [q]$, is referred to as the *population structure* or the *topic-word matrix* [42]. For instance, a_i is the conditional distribution of words given topic i. The latent variable h is a q dimensional random vector $h := [h_1, \ldots, h_q]^\top$ known as proportion vector. A prior distribution $P(h)$ over the probability simplex Δ^{q-1} characterizes the prior joint distribution over the latent variables $h_i,\ i \in [q]$. In the topic modeling, this is the prior distribution over the q topics.

The n-persistent topic model has a three-level multi-view hierarchy in Figure 1. $2rn$ number of words (views) are shown in the model for some integer $r \geq 1$. In this model, a common hidden topic is persistent for a sequence of n words $\{x_{(j-1)n+1}, \ldots, x_{(j-1)n+n}\}, j \in [2r]$. Note that the random observed variables (words) are exchangeable within groups of size n, where n is the persistence level, but are not globally exchangeable.

We now describe a linear representation of the n-persistent topic model, on lines of [6], but with extensions to incorporate persistence. Each random variable $y_j, j \in [2r]$, is a discrete valued random variable taking one of the q possibilities $\{1, \ldots, q\}$, i.e., $y_j \in [q]$ for $j \in [2r]$. In the n-persistent model, a single common topic is chosen for a sequence of n words $\{x_{(j-1)n+1}, \ldots, x_{(j-1)n+n}\}, j \in [2r]$, i.e., the topic is persistent for n successive views. For notational purposes, we equivalently assume that variables $y_j, j \in [2r]$, are encoded by the basis vectors $e_i,\ i \in [q]$. Thus, the variable $y_j, j \in [2r]$, is

$$y_j = e_i \in \mathbb{R}^q \iff \text{the topic of } j\text{-th group of words is } i.$$

Given proportion vector h, topics $y_j, j \in [2r]$, are independently drawn according to the conditional expectation

$$\mathbb{E}\big[y_j | h\big] = h, \quad j \in [2r],$$

or equivalently $\Pr\big[y_j = e_i | h\big] = h_i, j \in [2r], i \in [q]$.

Finally, at the bottom layer, each observed variable x_l for $l \in [2rn]$, is a discrete-valued p-dimensional random variable, where p is the size of word vocabulary. Again, we assume that variables x_l, are encoded by the basis vectors $e_k,\ k \in [p]$, such as

$$x_l = e_k \in \mathbb{R}^p \iff \text{the } l\text{-th word in the document is } k.$$

Given the corresponding topic $y_j, j \in [2r]$, words $x_l, l \in [2rn]$, are independently drawn according to the conditional expectation

$$\mathbb{E}\big[x_{(j-1)n+k} | y_j = e_i\big] = a_i,\ i \in [q], j \in [2r],\ k \in [n], \tag{1}$$

where vectors $a_i \in \mathbb{R}^p,\ i \in [q]$, are the conditional probability distribution vectors. The matrix $A = [a_1|a_2|\cdots|a_q] \in \mathbb{R}^{p\times q}$ collecting these vectors is the *population structure* or *topic-word matrix.*

The $(2rn)$-th order moment of observed variables $x_l, l \in [2rn]$, for some integer $r \geq 1$, is defined as (in the matrix form) [6]

$$M_{2rn}(x) := \mathbb{E}\left[(x_1 \otimes x_2 \otimes \cdots \otimes x_{rn})(x_{rn+1} \otimes x_{rn+2} \otimes \cdots \otimes x_{2rn})^\top\right] \in \mathbb{R}^{p^{rn}\times p^{rn}}. \tag{2}$$

For the n-persistent topic model with $2rn$ number of observations (words) $x_l, l \in [2rn]$, the corresponding moment is denoted by $M_{2rn}^{(n)}(x)$. Note that to estimate the $(2rn)^{\text{th}}$ moment, we require a minimum of $2rn$ words in each document. We can select the first $2rn$ words in each document, and average over the different documents to obtain a consistent estimate of the moment. In this paper, we consider the problem of identifiability when exact moments are available.
The moment characterization of the n-persistent topic model is provided in Lemma 1 in Section 4.1. Given $M_{2rn}^{(n)}(x)$, what are the sufficient conditions under which the population structure A is identifiable? This is answered in Section 3.
Remark 1. *Note that our results are valid for the more general linear model $x_l = Ay_j$ (more precisely, $x_{(j-1)n+k} = Ay_j, j \in [2r], k \in [n]$), i.e., each column of matrix A does not need to be a valid probability distribution. Furthermore, the observed random variables x_l, can be continuous while the hidden ones y_j are assumed to be discrete.*

3 Sufficient Conditions for Generic Identifiability

In this section, the identifiability result for the n-persistent topic model with access to $(2n)$-th order observed moment is provided. First, sufficient deterministic conditions on the population structure A are provided for identifiability in Theorem 1. Next, the deterministic analysis is specialized to a random structured model in Theorem 2.

We now make the notion of identifiability precise. As defined in literature, (strict) identifiability means that the population structure A can be uniquely recovered up to permutation and scaling for all $A \in \mathbb{R}^{p\times q}$. Instead, we consider a more relaxed notion of identifiability, known as generic identifiability.
Definition 1 (Generic identifiability)**.** *We refer to a matrix $A \in \mathbb{R}^{p\times q}$ as generic, with a fixed sparsity pattern when the nonzero entries of A are drawn from a distribution which is absolutely continuous with respect to Lebesgue measure* [7]*. For a given sparsity pattern, the class of population structure matrices is said to be* generically identifiable *[25], if all the non-identifiable matrices form a set of Lebesgue measure zero.*

The $(2r)$-th order moment of hidden variables $h \in \mathbb{R}^q$, denoted by $M_{2r}(h) \in \mathbb{R}^{q^r\times q^r}$, is defined as

$$M_{2r}(h) := \mathbb{E}\left[\Big(\overbrace{h\otimes\cdots\otimes h}^{r \text{ times}}\Big)\Big(\overbrace{h\otimes\cdots\otimes h}^{r \text{ times}}\Big)^\top\right] \in \mathbb{R}^{q^r\times q^r}. \tag{3}$$

[6] Vector x is the vector generated by concatenating all vectors $x_l, l \in [2rn]$.
[7] As an equivalent definition, if the non-zero entries of an arbitrary sparse matrix are independently perturbed with noise drawn from a continuous distribution to generate A, then A is called generic.

We now provide a set of sufficient conditions for generic identifiability of structured topic models given $(2rn)$-th order observed moment. We first start with a natural assumption on the hidden variables.

Condition 1 (Non-degeneracy)**.** *The $(2r)$-th order moment of hidden variables $h \in \mathbb{R}^q$, defined in equation* (3)*, is full rank (non-degeneracy of hidden nodes).*

Note that there is no hope of distinguishing distinct hidden nodes without this non-degeneracy assumption. We do not impose any other assumption on hidden variables and can incorporate arbitrarily correlated topics.

Furthermore, we can only hope to identify the population structure A up to scaling and permutation. Therefore, we can identify A up to a canonical form defined as:

Definition 2 (Canonical form)**.** *Population structure A is said to be in* canonical form *if all of its columns have unit norm.*

3.1 Deterministic conditions for generic identifiability

In this section, we consider a fixed sparsity pattern on the population structure A and establish generic identifiability when non-zero entries of A are drawn from some continuous distribution. Before providing the main result, a generalized notion of (perfect) matching for bipartite graphs is defined. We subsequently impose these conditions on the bipartite graph from topics to words which encodes the sparsity pattern of population structure A.

Generalized matching for bipartite graphs

A bipartite graph with two disjoint vertex sets Y and X and an edge set E between them is denoted by $G(Y, X; E)$. Given the bi-adjacency matrix A, the notation $G(Y, X; A)$ is also used to denote a bipartite graph. Here, the rows and columns of matrix $A \in \mathbb{R}^{|X|\times|Y|}$ are respectively indexed by X and Y vertex sets. For any subset $S \subseteq Y$, the set of neighbors of vertices in S with respect to A is defined as $N_A(S) := \{i \in X : A_{ij} \neq 0 \text{ for some } j \in S\}$, or equivalently, $N_E(S) := \{i \in X : (j, i) \in E \text{ for some } j \in S\}$ with respect to edge set E.

Here, we define a generalized notion of matching for a bipartite graph and refer to it as n-gram matching.

Definition 3 ((Perfect) n-gram matching)**.** *A* n-gram matching *M for a bipartite graph $G(Y, X; E)$ is a subset of edges $M \subseteq E$ which satisfies the following conditions. First, for any $j \in Y$, we have $|N_M(j)| \leq n$. Second, for any $j_1, j_2 \in Y, j_1 \neq j_2$, we have $\min\{|N_M(j_1)|, |N_M(j_2)|\} > |N_M(j_1) \cap N_M(j_2)|$.*

A perfect n-gram matching *or* Y-saturating n-gram matching *for the bipartite graph $G(Y, X; E)$ is a n-gram matching M in which each vertex in Y is the end-point of exactly n edges in M.*

In words, in a n-gram matching M, each vertex $j \in Y$ is at most the end-point of n edges in M and for any pair of vertices in Y $(j_1, j_2 \in Y, j_1 \neq j_2)$, there exists at least one non-common neighbor in set X for each of them (j_1 and j_2).

As an example, a bipartite graph $G(Y, X; E)$ with $|X| = 4$ and $|Y| = 6$ is shown in Figure 2 for which the edge set E itself is a perfect 2-gram matching.

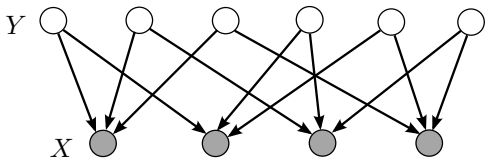


Figure 2: A bipartite graph $G(Y, X; E)$ with $|X| = 4$ and $|Y| = 6$ where the edge set E itself is a perfect 2-gram matching.

Remark 2 (Relationship to other matchings)**.** *The relationship of n-gram matching to other types of matchings is discussed below.*

- *Regular matching: For special case $n = 1$, the (perfect) n-gram matching reduces to the usual (perfect) matching for bipartite graphs.*
- *b-matching: A b-matching for a bipartite graph $G(Y, X; E)$ (with equal vertex sizes $|X| = |Y|$) is a subset of edges $M_b \subseteq E$, where each vertex is connected to b edges. Comparing with the proposed perfect n-gram matching, b-matching does not enforce that the set of neighbors be different, and furthermore, it requires that $X = Y$, which is not possible under the overcomplete setting.*

Remark 3 (Necessary size bound)**.** *Consider a bipartite graph $G(Y, X; E)$ with $|Y| = q$ and $|X| = p$ which has a perfect n-gram matching. Note that there are $\binom{p}{n}$ n-combinations on X side and each combination can at most have one neighbor (a node in Y which is connected to all nodes in the combination) through the matching, and therefore we necessarily have $q \leq \binom{p}{n}$.*

Finally, note that the existence of perfect n-gram matching results the existence of perfect $(n+1)$-gram matching [8], but the reverse is not true. For example, the bipartite graph $G(Y, X; E)$ with $|X| = 4$ and $|Y| = \binom{4}{2} = 6$ in Figure 2, has a perfect 2-gram matching, but not a perfect (1-gram) matching (since $6 > 4$).

Identifiability conditions based on existence of perfect n-gram matching in topic-word graph

Now, we are ready to propose the identifiability conditions and result.

Condition 2 (Perfect n-gram matching on A)**.** *The bipartite graph $G(V_h, V_o; A)$ between hidden and observed variables, has a perfect n-gram matching.*

The above condition implies that the sparsity pattern of matrix A is appropriately scattered in the mapping from hidden to observed variables to be identifiable. Intuitively, it means that every hidden node can be distinguished from another hidden node by its unique set of neighbors under the corresponding n-gram matching.

Furthermore, condition 2 is the key to be able to propose identifiability in the overcomplete regime. As stated in the size bound in Remark 3, for $n \geq 2$, the number of hidden variables can be more than the number of observed variables and we can still have perfect n-gram matching.

Definition 4 (Kruskal rank, [15])**.** *The* Kruskal rank *or the* krank *of matrix A is defined as the maximum number k such that every subset of k columns of A is linearly independent.*

[8] Note that the degree of each node (on matching side Y) in the original bipartite graph should be at least $n + 1$.

Note that krank is different from the general notion of matrix rank and it is a lower bound for the matrix rank, i.e., $\text{Rank}(A) \geq \text{krank}(A)$.

Condition 3 (Krank condition on A)**.** *The Kruskal rank of matrix A satisfies the bound* $\text{krank}(A) \geq d_{\max}(A)^n$*, where $d_{\max}(A)$ is the maximum node degree of any column of A.*

In the overcomplete regime, it is not possible for A to be full column rank and $\text{krank}(A) < |V_h| = q$. However, note that a large enough krank ensures that appropriate sized subsets of columns of A are linearly independent. For instance, when $\text{krank}(A) > 1$, any two columns cannot be collinear and the above condition rules out the collinear case for identifiability. In the above condition, we see that a larger krank can incorporate denser connections between topics and words.

The main identifiability result under a fixed graph structure is stated in the following theorem for $n \geq 2$, where n is the topic persistence level. The identifiability result relies on having access to the $(2rn)$-th order moment of observed variables $x_l, l \in [2rn]$, defined in equation (2) as

$$M_{2rn}(x) := \mathbb{E}\left[(x_1 \otimes x_2 \otimes \cdots \otimes x_{rn})(x_{rn+1} \otimes x_{rn+2} \otimes \cdots \otimes x_{2rn})^\top\right] \in \mathbb{R}^{p^{rn} \times p^{rn}},$$

for some integer $r \geq 1$.

Theorem 1 (Generic identifiability under deterministic topic-word graph structure)**.** *Let $M_{2rn}^{(n)}(x)$ in equation (2) be the $(2rn)$-th order observed moment of the n-persistent topic model for some integer $r \geq 1$. If the model satisfies conditions 1, 2 and 3, then, for any $n \geq 2$, all the columns of population structure A are* generically identifiable *from $M_{2rn}^{(n)}(x)$. Furthermore, the $(2r)$-th order moment of the hidden variables, denoted by $M_{2r}(h)$, is also* generically identifiable*.*

The theorem is proved in Appendix A. It is seen that the population structure A is identifiable, given any observed moment of order at least $2n$. Increasing the order of observed moment results in identifying higher order moments of the hidden variables.

The above theorem does not cover the case when the persistence level $n = 1$. This is the usual bag-of-words admixture model. Identifiability of this model has been studied earlier [7] and we recall it below.

Remark 4 (Bag-of-words admixture model, [7])**.** *Given $(2r)$-th order observed moments with $r \geq 1$, the structure of the popular bag-of-words admixture model and the $(2r)$-th order moment of hidden variables are identifiable, when A is full column rank and the following expansion condition holds [7]*

$$|N_A(S)| \geq |S| + d_{\max}(A), \quad \forall S \subseteq V_h,\ |S| \geq 2. \tag{4}$$

Our result for $n \geq 2$ in Theorem 1, provides identifiability in the overcomplete regime with weaker matching condition 2 and krank condition 3. The matching condition 2 is weaker than the above expansion condition which is based on the perfect matching and hence, does not allow overcomplete models. Furthermore, the above result for the bag-of-words admixture model requires full column rank of A which is more stringent than our krank condition 3.

Remark 5 (Kruskal rank and degree diversity)**.** *Condition 3 requires that the Kruskal rank of the topic-word matrix be large enough compared to the maximum degree of the topics. Intuitively, a larger Kruskal rank ensures enough diversity in the word supports among different topics under a higher level of sparsity. This Kruskal rank condition also allows for more degree diversity among the topics, when the topic persistence level $n > 1$. On the other hand, for the bag-of-words model ($n = 1$), using (4) implies that $2d_{\min} > d_{\max}$, where $d_{\min}, d_{\max}$ are the minimum and maximum degrees of the topics. Thus, we provide identifiability results with more degree diversity when higher order moments are employed.*

Remark 6 (Recovery using ℓ_1 optimization)**.** *It turns out that our conditions for identifiability imply that the columns of the n-gram matrix $A^{\odot n}$, defined in Definition 6, are the sparsest vectors in $\mathrm{Col}\Big(M_{2n}^{(n)}(x)\Big)$, having a tensor rank of one. See Appendix A. This implies recovery of the columns of A through exhaustive search, which is not efficient. Efficient ℓ_1-based recovery algorithms have been analyzed in [7,43] for the undercomplete case ($n = 1$). They can be employed here for recovery from higher order moments as well. Exploiting additional structure present in $A^{\odot n}$, for $n > 1$, such as rank-1 test devices proposed in [20] are interesting avenues for future investigation.*

3.2 Analysis under random topic-word graph structures

In this section, we specialize the identifiability result to the random case. This result is based on more transparent conditions on the size and the degree of the random bipartite graph $G(V_h, V_o; A)$. We consider the random model where in the bipartite graph $G(V_h, V_o; A)$, each node $i \in V_h$ is randomly connected to d_i different nodes in set V_o. Note that this is a heterogeneous degree model.

Condition 4 (Size condition)**.** *The random bipartite graph $G(V_h, V_o; A)$ with $|V_h| = q, |V_o| = p$, and $A \in \mathbb{R}^{p\times q}$, satisfies the size condition $q \leq \big(c\frac{p}{n}\big)^n$ for some constant $0 < c < 1$.*

This size condition is required to establish that the random bipartite graph has a perfect n-gram matching (and hence satisfies deterministic condition 2). It is shown in Section 5.2.1 that the necessary size constraint $q = O(p^n)$ stated in Remark 3, is achieved in the random case. Thus, the above constraint allows for the overcomplete regime, where $q \gg p$ for $n \geq 2$, and is tight.

Condition 5 (Degree condition)**.** *In the random bipartite graph $G(V_h, V_o; A)$ with $|V_h| = q, |V_o| = p$, and $A \in \mathbb{R}^{p\times q}$, the degree d_i of nodes $i \in V_h$ satisfies the following lower and upper bounds ($d_i \in [d_{\min}, d_{\max}]$):*

- Lower bound: $d_{\min} \geq \max\{1+\beta \log p, \alpha \log p\}$ *for some constants* $\beta > \frac{n-1}{\log 1/c}, \alpha > \max\big\{2n^2\big(\beta \log \frac{1}{c} + 1\big), 2\beta n\big\}$.
- Upper bound: $d_{\max} \leq (cp)^{\frac{1}{n}}$.

Intuitively, the lower bound on the degree is required to show that the corresponding bipartite graph $G(V_h, V_o; A)$ has sufficient number of random edges to ensure that it has perfect n-gram matching with high probability. The upper bound on the degree is mainly required to satisfy the krank condition 3, where $d_{\max}(A)^n \leq \mathrm{krank}(A)$.

It is important to see that, for $n \geq 2$, the above condition on degree covers a range of models from sparse to intermediate regimes and it is reasonable in a number of applications that each topic does not generate a very large number of words.

Definition 5 (**whp**)**.** *A sequence of events $\mathcal{E}_p$ occurs* with high probability *(**whp**) if* $\Pr(\mathcal{E}_p) = 1 - O(p^{-\epsilon})$ *for some $\epsilon > 0$.*

The main random identifiability result is stated in the following theorem for $n \geq 2$, while $n = 1$ case is addressed in Remark 8. The identifiability result relies on having access to the $(2rn)$-th order moment of observed variables $x_l, l \in [2rn]$, defined in equation (2) as

$$M_{2rn}(x) := \mathbb{E}\left[(x_1 \otimes x_2 \otimes \cdots \otimes x_{rn})(x_{rn+1} \otimes x_{rn+2} \otimes \cdots \otimes x_{2rn})^\top\right] \in \mathbb{R}^{p^{rn}\times p^{rn}},$$

for some integer $r \geq 1$.

Probability rate constants: The probability rate of success in the following random identifiability result is specified by constants $\beta' > 0$ and $\gamma = \gamma_1 + \gamma_2 > 0$ as

$$\beta' = -\beta \log c - n + 1, \tag{5}$$

$$\gamma_1 = e^{n-1}\Big(\frac{c}{n^{n-1}} + \frac{e^2}{1-\delta_1} n^{\beta'+1}\Big), \tag{6}$$

$$\gamma_2 = \frac{c^{n-1}e^2}{n^n(1-\delta_2)}, \tag{7}$$

where δ_1 and δ_2 are some constants satisfying $e^2\Big(\frac{p}{n}\Big)^{-\beta \log 1/c} < \delta_1 < 1$ and $\frac{c^{n-1}e^2}{n^n} p^{-\beta'} < \delta_2 < 1$.

Theorem 2 (Random identifiability)**.** *Let $M_{2rn}^{(n)}(x)$ in equation (2) be the $(2rn)$-th order observed moment of the n-persistent topic model for some integer $r \geq 1$. If the model with random population structure A satisfies conditions 1, 4 and 5, then **whp** (with probability at least $1-\gamma p^{-\beta'}$ for constants $\beta' > 0$ and $\gamma > 0$, specified in (5)-(7)), for any $n \geq 2$, all the columns of population structure A are identifiable from $M_{2rn}^{(n)}(x)$. Furthermore, the $(2r)$-th order moment of hidden variables, denoted by $M_{2r}(h)$, is also identifiable, **whp**.*

The theorem is proved in Appendix B. Similar to the deterministic analysis, it is seen that the population structure A is identifiable given any observed moment with order at least $2n$. Increasing the order of observed moment results in identifying higher order moments of the hidden variables.

Remark 7 (Trade-off between topic-word size ratio and degree)**.** *When the number of hidden variables increases, i.e. c increases, but the order n is kept fixed, the bounds on degree in condition 5 also needs to grow. Intuitively, a larger degree is needed to provide more flexibility in choosing the subsets of neighbors for hidden nodes to ensure the existence of a perfect n-gram matching in the bipartite graph, which in turn ensures identifiability. Note that as c grows, the parameter β, which is the lower bound on d also grows, and the probability rate (i.e., the term $-\beta \log c$) remains constant. Hence, the probability rate does not change as c increases, since the increase in the degree d compensates the additional "difficulty" arising due to a larger number of hidden variables.*

The above identifiability theorem only covers for $n \geq 2$ and the $n = 1$ case is addressed in the following remark.

Remark 8 (Bag-of-words admixture model)**.** *The identifiability result for the random bag-of-words admixture model is comparable to the result in [43], which considers exact recovery of sparsely-used dictionaries. They assume that $Y = DX$ is given for some unknown arbitrary dictionary $D \in \mathbb{R}^{q\times q}$ and unknown random sparse coefficient matrix $X \in \mathbb{R}^{q\times p}$. They establish that if $D \in \mathbb{R}^{q\times q}$ is full rank and the random sparse coefficient matrix $X \in \mathbb{R}^{q\times p}$ follows the Bernoulli-subgaussian model with size constraint $p > Cq \log q$ and degree constraint $O(\log q) < \mathbb{E}[d] < O(q \log q)$, then the model is identifiable, whp. Comparing the size and degree constraints, our identifiability result for $n \geq 2$ requires more stringent upper bound on the degree ($d = O(p^{1/n})$), while more relaxed condition on the size ($q = O(p^n)$) which allows to identifiability in the overcomplete regime.*

Remark 9 (The size condition is tight)**.** *The size bound $q = O(p^n)$ in the above theorem achieves the necessary condition that $q \leq \binom{p}{n} = O(p^n)$ (see Remark 3), and is therefore tight. The sufficiency*

is argued in Theorem 3, where we show that the matching condition 2 holds under the above size and degree conditions 4 and 5.

4 Identifiability via Uniqueness of Tensor Decompositions

In this section, we characterize the moments of the n-persistent topic model in terms of the model parameters, i.e. the topic-word matrix A and the moment of hidden variables. We relate identifiability of the topic model to uniqueness of a certain class of tensor decompositions, which in turn, enables us to prove Theorems 1 and 2. We then discuss the special cases of the persistent topic model, viz., the single topic model (infinite-persistent topic model) and the bag-of-words admixture model (1-persistent topic model).

4.1 Moment characterization of the persistent topic model

The moment characterization requires the following definition of a n-gram matrix.

Definition 6 (n-gram Matrix)**.** *Given a matrix $A \in \mathbb{R}^{p\times q}$, its n-gram matrix $A^{\odot n} \in \mathbb{R}^{p^n\times q}$ is defined as the matrix whose $(\mathbf{i}, j)$-th entry is given by, for $\mathbf{i} := (i_1, i_2, \ldots, i_n) \in [p]^n$ and $j \in [q]$,*

$$A^{\odot n}(\mathbf{i}, j) := A_{i_1,j}A_{i_2,j}\cdots A_{i_n,j}, \quad \text{or} \quad A^{\odot n} := \overbrace{A \odot \cdots \odot A}^{n \text{ times}}.$$

That is, $A^{\odot n}$ is the column-wise n^{th} order Kronecker product of n copies of A, and is known as the Khatri-Rao product [41].

In the following lemma, which is proved in Appendix A.2, we characterize the observed moments of a persistent topic model. Throughout this section, the order of the observed moment is fixed to $2m$.

Lemma 1 (n-persistent topic model moment characterization)**.** *The $(2m)$-th order moment of observed variables, defined in equation (2), for the n-persistent topic model is characterized as*[9]*:*

- *if $m = rn$, for some integer $r \geq 1$, then*

$$M_{2m}^{(n)}(x) = \left(\overbrace{A^{\odot n} \otimes \cdots \otimes A^{\odot n}}^{r \text{ times}}\right) M_{2r}(h) \left(\overbrace{A^{\odot n} \otimes \cdots \otimes A^{\odot n}}^{r \text{ times}}\right)^\top, \tag{8}$$

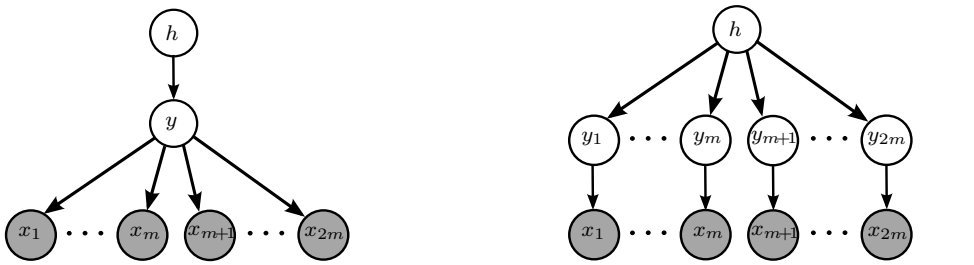
  where $M_{2r}(h) \in \mathbb{R}^{q^r\times q^r}$ is the $(2r)$-th order moment of hidden variables $h \in \mathbb{R}^q$, defined in equation (3).

- *If $n \geq 2m$, then*

$$M_{2m}^{(n)}(x) = \left(A^{\odot m}\right) M_1(h) \left(A^{\odot m}\right)^\top, \tag{9}$$

  where $M_1(h) := \text{Diag}(\mathbb{E}[h]) \in \mathbb{R}^{q\times q}$ is the first order moment of hidden variables $h \in \mathbb{R}^q$, stacked in a diagonal matrix.

[9] The other cases not covered in Lemma 1 are deferred to Appendix A.2. See Remark 12.



(a) Single topic model (infinite-persistent topic model)

(b) Bag-of-words admixture model (1-persistent topic model)

Figure 3: Hierarchical structure of the single topic model and bag-of-words admixture model shown for $2m$ number of words (views).

Thus, we see that the observed moments can be expressed in terms of the hidden moments $M(h)$ and the Kronecker products of the n-gram matrices. In the special case, when the persistence level is large enough compared to the order of the moment ($n \geq 2m$), the moment form reduces to a Khatri-Rao product form in (9). Moreover, in (9), we have a diagonal matrix $M_1(h)$ instead of a general (dense) matrix $M_{2r}(h)$ in (8), when $n < 2m = 2rn$. Thus, we have a more succinct representation of the moments in (9) when the persistence level of the topics is large enough.

In the following, we contrast the special cases when the persistence level n is $n \to \infty$ (single topic model) and $n = 1$ (bag of words admixture model), as shown in Fig.3a and Fig.3b. In order to have a fair comparison, the number of observed variables is fixed to $2m$ and the persistence level is varied.

Single topic model ($n \to \infty$): The condition in (9) ($n \geq 2m$) is always satisfied for the single-topic model, since $n \to \infty$ in this case, and we have

$$M_{2m}^{(\infty)}(x) = \left(A^{\odot m}\right) M_1(h) \left(A^{\odot m}\right)^\top . \tag{10}$$

Note that $M_1(h)$ is a diagonal matrix.

Bag-of-words admixture model ($n = 1$): From Lemma 1, the $(2m)$-th order moment of observed variables $x_l, l \in [2m]$, for the bag-of-words admixture model (1-persistent topic model), shown in Figure 3b, is given by

$$M_{2m}^{(1)}(x) = \Big(\overbrace{A \otimes \cdots \otimes A}^{m \text{ times}}\Big) M_{2m}(h) \Big(\overbrace{A \otimes \cdots \otimes A}^{m \text{ times}}\Big)^\top , \tag{11}$$

where $M_{2m}(h) \in \mathbb{R}^{q^m \times q^m}$ is the $(2m)$-th order moment of hidden variables $h \in \mathbb{R}^q$, defined in (3). Note that $M_{2m}(h)$ is a full matrix in general.

Contrasting single topic ($n \to \infty$) and bag of words models ($n = 1$): Comparing equations (10) and (11), it is seen that the moments under the single topic model in (10) are more "structured" compared to the bag of words model in (11). In (11), we have Kronecker products of the topic-word matrix A, while (10) involves Khatri-Rao products of A. This forms a crucial criterion in determining of whether overcomplete models are identifiable, as discussed below.

Why persistence helps in identifiability of overcomplete models? For simplicity, let the order of the moment $2m = 4$. The equations (10) and (11) reduce to

$$M_4^{(\infty)}(x) = (A \odot A)\,\mathrm{Diag}\left(\mathbb{E}[h]\right)(A \odot A)^\top, \tag{12}$$

$$M_4^{(1)}(x) = (A \otimes A)\mathbb{E}\big[(h \otimes h)(h \otimes h)^\top\big](A \otimes A)^\top. \tag{13}$$

Note that for the single topic model in (12), the Khatri-Rao product matrix $A \odot A \in \mathbb{R}^{p^2 \times q}$ has the same as the number of columns (i.e. the latent dimensionality) of the original matrix A, while the number of rows (i.e. the observed dimensionality) is increased. Thus, the Khatri-Rao product "expands" the effect of hidden variables to higher order observed variables, which is the key towards identifying overcomplete models. In other words, the original overcomplete representation becomes determined due to the 'expansion effect' of the Khatri-Rao product structure of the higher order observed moments.

On the other hand, in the bag-of-words admixture model in (13), this interesting 'expansion property' does not occur, and we have the Kronecker product $A \otimes A \in \mathbb{R}^{p^2 \times q^2}$, in place of the Khatri-Rao products. The Kronecker product operation increases both the number of the columns (i.e. latent dimensionality) and the number of rows (i.e. observed dimensionality), which implies that higher order moments do not help in identifying overcomplete models.

An example is provided in Figure 4 which helps to see how the matrices $A \odot A$ and $A \otimes A$ behave differently in terms of mapping topics to word tuples.

Note that for the n-persistent model, for $n = 2$, the 4^{th} order moment reduces to

$$M_4^{(2)}(x) = (A \odot A)\mathbb{E}\big[hh^\top](A \odot A)^\top. \tag{14}$$

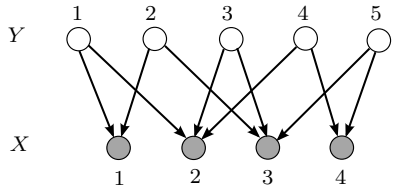
Contrasting the above equation with (12) and (13), we find that the 2-persistent model retains the desirable property of possessing Khatri-Rao products, while being more general than the form for single topic model in (12). This key property enables us to establish identifiability of topic models with finite persistence levels.

4.2 Tensor algebra of the model

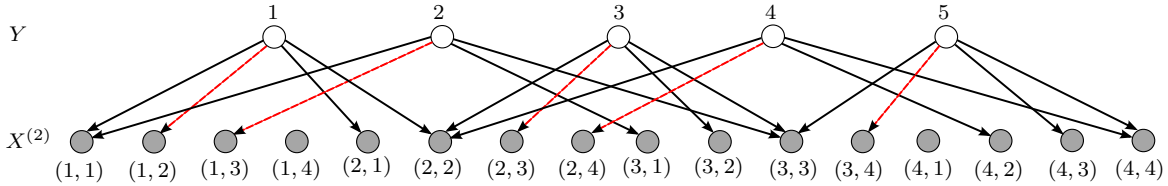
In Section 4.1, we provided a representation of the moment forms in the matrix form. We now provide the equivalent tensor representation of the moments. The tensor representation is more compact and transparent, and allows us to compare the topic models under different levels of persistence. We compare the derived tensor form with the well-known Tucker and CP decompositions. We first introduce some tensor notations and definitions.

4.2.1 Tensor notations and definitions

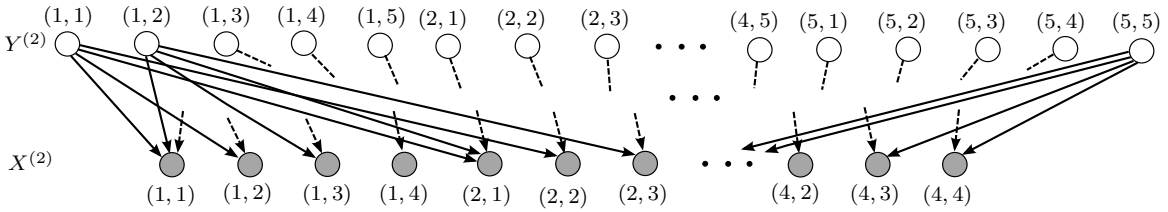
A real-valued order-n tensor $A \in \bigotimes_{i=1}^{n} \mathbb{R}^{p_i} := \mathbb{R}^{p_1 \times \cdots \times p_n}$ is a n dimensional array $A(1 : p_1, \ldots, 1 : p_n)$, where the i-th mode is indexed from 1 to p_i. In this paper, we restrict ourselves to the case that $p_1 = \cdots = p_n = p$, and simply write $A \in \bigotimes^n \mathbb{R}^p$. A *fiber* of a tensor A is a vector obtained by fixing all indices of A except one, e.g., for $A \in \bigotimes^4 \mathbb{R}^3$, the vector $f = A(2, 1 : 3, 3, 1)$ is a fiber.



(a) Structure of an overcomplete matrix $A \in \mathbb{R}^{4\times 5}$ having a perfect 2-gram matching.



(b) Structure of $A \odot A \in \mathbb{R}^{16\times 5}$ having a perfect (Y-saturating) matching, highlighted by dashed red edges.



(c) Structure of $A \otimes A \in \mathbb{R}^{16\times 25}$. For simplicity, only a few edges and nodes are shown and the dashed edges denote the bunch of edges connected to each node, not specifically shown.

Figure 4: An example of an overcomplete matrix A and the matrices $A \odot A$ and $A \otimes A$. The corresponding bipartite graphs encode the sparsity pattern of each of the matrices. $A \odot A$ expands the effect of hidden variables to second order observed variables which is crucial for overcomplete identifiability, while in the $A \otimes A$, the order of both the hidden and observed variables are increased.

For a vector $u \in \mathbb{R}^p$, $\mathrm{Diag}_n(u) \in \bigotimes^n \mathbb{R}^p$ is the n-th order diagonal tensor with vector u on its diagonal. The tensor $A \in \bigotimes^n \mathbb{R}^p$, is stacked as a vector $a \in \mathbb{R}^{p^n}$ by the $\mathrm{vec}(\cdot)$ operator, defined as

$$a = \mathrm{vec}(A) \Leftrightarrow a\big((i_1-1)p^{n-1} + (i_2-1)p^{n-2} + \cdots + (i_{n-1}-1)p + i_n\big) = A(i_1, i_2, \ldots, i_n).$$

The inverse of $a = \mathrm{vec}(A)$ operation is denoted by $A = \mathrm{ten}(a)$.
For vectors $a_i \in \mathbb{R}^{p_i}, i \in [n]$, the tensor outer product operator "$\circ$" is defined as [41]

$$A = a_1 \circ a_2 \circ \cdots \circ a_n \in \bigotimes_{i=1}^{n} \mathbb{R}^{p_i} \Leftrightarrow A(i_1, i_2, \ldots, i_n) := a_1(i_1)a_2(i_2)\cdots a_n(i_n). \tag{15}$$

The above generated tensor is a rank-1 tensor. The *tensor rank* is the minimal number of rank-1 tensors into which a tensor can be decomposed. This type of rank is called CP (Candecomp/Parafac) tensor rank in the literature [41].
According to above definitions, for any set of vectors $a_i \in \mathbb{R}^{p_i}, i \in [n]$, we have the following pair of equalities:

$$\mathrm{vec}(a_1 \circ a_2 \circ \cdots \circ a_n) = a_1 \otimes a_2 \otimes \cdots \otimes a_n,$$
$$\mathrm{ten}(a_1 \otimes a_2 \otimes \cdots \otimes a_n) = a_1 \circ a_2 \circ \cdots \circ a_n.$$

For any vector $a \in \mathbb{R}^p$, the power notations are also defined as

$$a^{\otimes n} := \overbrace{a \otimes a \otimes \cdots \otimes a}^{n \text{ times}} \in \mathbb{R}^{p^n},$$
$$a^{\circ n} := \overbrace{a \circ a \circ \cdots \circ a}^{n \text{ times}} \in \bigotimes^{n} \mathbb{R}^{p}.$$

The second power is usually called the n-th order *tensor power* of vector a.
Finally, the Tucker and CP (Candecomp/Parafac) representations are defined as follows [10,41].
Definition 7 (Tucker representation)**.** *Given a core tensor $S \in \bigotimes_{i=1}^{n} \mathbb{R}^{r_i}$ and inverse factors $U_i \in \mathbb{R}^{p_i \times r_i}, i \in [n]$, the Tucker representation of the n-th order tensor $A \in \bigotimes_{i=1}^{n} \mathbb{R}^{p_i}$ is*

$$A = \sum_{i_1=1}^{r_1} \sum_{i_2=1}^{r_2} \cdots \sum_{i_n=1}^{r_n} S(i_1, i_2, \ldots, i_n) U_1(:, i_1) \circ U_2(:, i_2) \circ \cdots \circ U_n(:, i_n) =: [[S; U_1, U_2, \ldots, U_n]], \tag{16}$$

where $U_j(:, i_j)$ denotes the i_j-th column of matrix U_j. The tensor S is referred to as the core tensor.
Definition 8 (CP representation)**.** *Given $\lambda \in \mathbb{R}^r, U_i \in \mathbb{R}^{p_i \times r}, i \in [n]$, the CP representation of the n-th order tensor $A \in \bigotimes_{i=1}^{n} \mathbb{R}^{p_i}$ is*

$$A = \sum_{i=1}^{r} \lambda_i U_1(:, i) \circ U_2(:, i) \circ \cdots \circ U_n(:, i) =: [[\mathrm{Diag}_n(\lambda); U_1, U_2, \ldots, U_n]], \tag{17}$$

where $U_j(:, i)$ denotes the i-th column of matrix U_j.

Note that the CP representation is a special case of the Tucker representation when the core tensor S is square and diagonal.

4.2.2 Tensor representation of moments under topic model

We now provide a tensor representation of the moments.

For the n-persistent topic model, the $2m$-th observed moment is denoted by $T_{2m}^{(n)}(x)$, which is the tensor form of the moment matrix $M_{2m}^{(n)}(x)$, characterized in Lemma 1. It is given by

$$T_{2m}(x)_{(i_1,i_2,\ldots,i_{2m})} := \mathbb{E}[x_1(i_1)x_2(i_2)\cdots x_{2m}(i_{2m})], \quad i_1,i_2,\ldots,i_{2m} \in [p], \tag{18}$$

where $T_{2m}(x) \in \bigotimes^{2m} \mathbb{R}^p$.

This tensor is characterized in the following lemma, and is proved in Appendix A.2.

Lemma 2 (n-persistent topic model moment characterization in tensor form)**.** *The $(2m)$-th order moment of words, defined in equation* (18)*, for the n-persistent topic model is characterized as*[10]*:*

- *if $m = rn$ for some integer $r \geq 1$, then*

$$\begin{aligned} T_{2m}^{(n)}(x) &= \sum_{i_1=1}^{q}\sum_{i_2=1}^{q}\cdots\sum_{i_{2r}=1}^{q} \mathbb{E}[h_{i_1}h_{i_2}\cdots h_{i_{2r}}]a_{i_1}^{\circ n}\circ a_{i_2}^{\circ n}\circ\cdots\circ a_{i_{2r}}^{\circ n} \\ &= \Big[\Big[S_r; \overbrace{A, A, \ldots, A}^{2m \text{ times}}\Big]\Big], \end{aligned} \tag{19}$$

  where $S_r \in \bigotimes^{2rn} \mathbb{R}^q$ is the core tensor in the above Tucker representation with the sparsity pattern as

$$S_r(\mathbf{i}) = \begin{cases} M_{2r}(h)_{\left((i_n,i_{2n},\ldots,i_{rn}),(i_{(r+1)n},i_{(r+2)n},\ldots,i_{2rn})\right)} & , i_1 = i_2 = \cdots = i_n, i_{n+1} = i_{n+2} = \cdots = i_{2n}, \ldots \\ 0 & , \text{o. w.}, \end{cases}$$

  where $\mathbf{i} := (i_1, i_2, \ldots, i_{2rn})$.

- *If $n \geq 2m$, then*

$$T_{2m}^{(n)}(x) = \sum_{i\in[q]} \mathbb{E}[h_i]a_i^{\circ 2m} = \Big[\Big[\mathrm{Diag}_{2m}(\mathbb{E}[h]); \overbrace{A, A, \ldots, A}^{2m \text{ times}}\Big]\Big]. \tag{20}$$

The tensor representation in (19) is a specific type of tensor decomposition which is a special case of the Tucker representation (since S_r is not fully dense), but more general than the CP representation. The tensor representation in (20) has a CP form.

Comparison with single topic model and bag-of-words admixture model

We now provide the tensor form for the special cases single topic model and bag-of-words admixture model. In order to have a fair comparison, the number of observed variables is fixed to $2m$ and the persistence level is varied.

[10] The other cases not covered in Lemma 2 are deferred to Appendix A.2. See Remark 12.

CP representation of the single topic model: The $(2m)$-th order moment of the words for the single topic model (infinite-persistent topic model) is provided in equation (20) as

$$T_{2m}^{(\infty)}(x) = \sum_{i\in[q]} \mathbb{E}[h_i] a_i^{\otimes 2m} = \big[\big[\mathrm{Diag}_{2m}(\mathbb{E}[h]); \overbrace{A, A, \ldots, A}^{2m \text{ times}}\big]\big]. \tag{21}$$

This representation is the symmetric CP representation [11] of $T_{2m}^{(\infty)}(x)$.

Tucker representation of the bag-of-words admixture model: From Lemma 2, the tensor form of the $(2m)$-th order moment of observed variables $x_l, l \in [2m]$, for the bag-of-words admixture model (1-persistent topic model) is given by

$$\begin{aligned} T_{2m}^{(1)}(x) &= \sum_{i_1=1}^{q} \sum_{i_2=1}^{q} \cdots \sum_{i_{2m}=1}^{q} \mathbb{E}[h_{i_1} h_{i_2} \cdots h_{i_{2m}}] a_{i_1} \circ a_{i_2} \circ \cdots \circ a_{i_{2m}} \\ &= \Big[\Big[\mathbb{E}\big[h^{\otimes(2m)}\big]; \overbrace{A, A, \ldots, A}^{2m \text{ times}}\Big]\Big]. \end{aligned} \tag{22}$$

This representation is the Tucker representation (decomposition) of $T_{2m}^{(1)}(x)$ where the core tensor $S = \mathbb{E}\big[h^{\otimes(2m)}\big]$ is the tensor form of the $(2m)$-th order hidden moment $M_{2m}(h)$, defined in equation (3), and the inverse factors correspond to the population structure A.

Comparing the tensor forms for the n-persistent topic model (19), single topic model (21), and bag of words admixture model (22), we find that all of them involve Tucker decompositions, where the inverse factors correspond to the topic-word matrix A, and the only difference is in the sparsity level of the core tensor S. For the bag of words model, with $n = 1$, the core tensor is fully dense in general, while for the single topic model, with $n \to \infty$, the core tensor is diagonal which reduces to the CP decomposition. For a general topic model with persistence level n, the core tensor is in between these two extremes and has structured sparsity. This sparsity property of the core tensor is crucial towards establishing identifiability in the overcomplete regime. The bag-of-words model is not identifiable in the overcomplete regime since the core tensor is fully dense in this case, while an overcomplete n-persistent topic model can be identified under certain constraints provided in Section 3, since the core tensor has structured sparsity and symmetry.

5 Proof Techniques and Auxiliary Results

The main identifiability results are given in Theorems 1 and 2 for deterministic and random cases of topic-word graph structures. In this section, we provide a proof sketch of these results, and then, we propose auxiliary results on the existence of perfect n-gram matching for random bipartite graphs and a lower bound on the Kruskal rank of random matrices.

[11] In Appendix C, we provide a more detailed comparison between our approach and some of the previous identifiability results for the (overcomplete) CP decomposition.

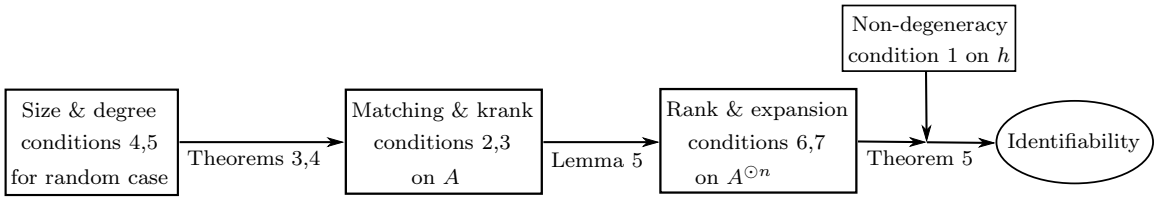


Figure 5: Hierarchy among the proposed conditions and results.

5.1 Proof sketch

Summary of relationships among different conditions: To summarize, there exists a hierarchy among the proposed conditions as follows. See Figure 5. First, in the random analysis, the size and the degree conditions 4 and 5 are sufficient for satisfying the perfect n-gram matching and the krank conditions 2 and 3, shown by Theorems 3 and 4. Then, these conditions 2 and 3 ensure that the rank and the expansion conditions 6 and 7 hold, shown by Lemma 5. And finally, these conditions 6 and 7 together with non-degeneracy condition 1 conclude the primary identifiability result in Theorem 5. Note that the genericity of A is also required for these results to hold.

Primary deterministic analysis in Theorem 5: The deterministic analysis is primarily based on conditions on the n-gram matrix $A^{\odot n}$; but since these conditions are opaque (mainly expansion condition on $A^{\odot n}$, provided in condition 7), this analysis is related to conditions on matrix A itself. See Theorem 5 in Appendix A.1 for the identifiability result based on $A^{\odot n}$. We briefly discuss it below for the case when $2n$ number of words are available under the n-persistent topic model. From equation (8), the $(2n)$-th order moment of the observed variables under the n-persistent topic model can be written as

$$M_{2n}^{(n)}(x) = \Big(A^{\odot n}\Big)\mathbb{E}\big[hh^\top\big]\Big(A^{\odot n}\Big)^\top. \tag{23}$$

The question is whether we can recover A, given the $M_{2n}^{(n)}(x)$. Obviously, the matrix A is not identifiable without any further conditions. First, non-degeneracy and rank conditions (conditions 1 and 6) are required. Assuming these two conditions, we have from (23) that

$$\mathrm{Col}\Big(M_{2n}^{(n)}(x)\Big) = \mathrm{Col}\Big(A^{\odot n}\Big).$$

Therefore, the problem of recovering A from $M_{2n}^{(n)}(x)$ reduces to finding $A^{\odot n}$ in $\mathrm{Col}\big(A^{\odot n}\big)$.
Then, we show that under the following expansion condition on $A^{\odot n}$ and the genericity property, matrix A is identifiable from $\mathrm{Col}\big(A^{\odot n}\big)$. The expansion condition (refer to condition 7 for a more detailed statement), imposes the following property on the bipartite graph $G\big(V_h, V_o^{(n)}; A^{\odot n}\big)$ [12],

$$\Big|N_{A^{\odot n}_{\text{Rest.}}}(S)\Big| \geq |S| + d_{\max}\Big(A^{\odot n}\Big), \quad \forall S \subseteq V_h,\ |S| > \mathrm{krank}(A), \tag{24}$$

where $d_{\max}\big(A^{\odot n}\big)$ is the maximum node degree in set V_h, and the restricted version of n-gram matrix, denoted by $A^{\odot n}_{\text{Rest.}}$, is obtained by removing its redundant (identical) rows (see Definition

[12] $V_o^{(n)}$ denotes all ordered n-tuples generated from set $V_o := \{1, \ldots, p\}$ which indexes the rows of $A^{\odot n}$.

9). The identifiability claim is proved by showing that the columns of $A^{\odot n}$ are the sparsest and rank-1 vectors (in the tensor form) in $\mathrm{Col}\big(A^{\odot n}\big)$ under the expansion condition in (24) and genericity conditions. Note that since we only require expansion on sets larger than Kruskal rank, the expansion condition (24) is a more relaxed condition compared to expansion condition proposed in [7, 43] for identifiability in the undercomplete regime. For a more detailed comparison, refer to Remark 11 in Appendix A.1.

Deterministic analysis in Theorem 1: Expansion and rank conditions in Theorem 5 are imposed on the n-gram matrix $A^{\odot n}$. According to the generalized matching notions, defined in Section 3.1, sufficient combinatorial conditions on matrix A (conditions 2 and 3) are introduced which ensure that the expansion and rank conditions on $A^{\odot n}$ are satisfied. The following lemma is employed to establish these results, where we state an interesting property which relates the existence of a perfect matching in $A^{\odot n}$ to the existence of a perfect n-gram matching in A.

Lemma 3. *If $G(Y, X; A)$ has a perfect n-gram matching, then $G(Y, X^{(n)}; A^{\odot n})$ has a perfect matching. In the other direction, if $G(Y, X^{(n)}; A^{\odot n})$ has a perfect matching $M^{\odot n}$, then $G(Y, X; A)$ has a perfect n-gram matching under the following condition on $M^{\odot n}$. All the matching edges $(j, (i_1, \ldots, i_n)) \in M^{\odot n}$ should satisfy $i_1 \neq i_2 \neq \cdots \neq i_n$ for all $j \in Y$. In words, the matching edges should be connected to nodes in $X^{(n)}$, which are indexed by tuples of distinct indices.*

See Appendix A.4 for a proof. Using this lemma, condition 2 implies that $G(Y, X^{(n)}; A^{\odot n})$ has a perfect matching. Then, it is straightforward to argue that the expansion and rank conditions on $A^{\odot n}$ are satisfied, which is shown in Lemma 5 in Appendix A.3. This leads to the generic identifiability result stated in Theorem 1.

5.2 Analysis of Random Structures

The identifiability result for a random structured matrix A is provided in Theorem 2. Sufficient size and degree conditions 4 and 5 on the random matrix A are proposed such that the deterministic combinatorial conditions 2 and 3 on A are satisfied. The details of these auxiliary results are provided in the following two subsequent sections. In Section 5.2.1, it is proved in Theorem 3 that a random bipartite graph satisfying reasonable size and degree constraints, has a perfect n-gram matching (condition 2), **whp**. Then, a lower bound on the Kruskal rank of a random matrix A under size and degree constraints is provided in Theorem 4 in Section 5.2.2, which implies the krank condition 3. Intuitions on why such size and degree conditions are required, are mentioned in Section 3.2 where these conditions are proposed.

5.2.1 Existence of perfect n-gram matching for random bipartite graphs

We show in the following theorem that a random bipartite graph satisfying reasonable size and degree constraints, proposed earlier in conditions 4 and 5, has a perfect n-gram matching **whp**.

Theorem 3 (Existence of perfect n-gram matching for random bipartite graphs)**.** *Consider a random bipartite graph $G(Y, X; E)$ with $|Y| = q$ nodes on the left side and $|X| = p$ nodes on the right side, and each node $i \in Y$ is randomly connected to d_i different nodes in X. Let $d_{\min} := \min_{i \in Y} d_i$. Assume that it satisfies the size condition $q \leq \big(c\frac{p}{n}\big)^n$ (condition 4) for some constant $0 < c < 1$ and the degree condition $d_{\min} \geq \max\{1 + \beta \log p, \alpha \log p\}$ for some constants $\beta > \frac{n-1}{\log 1/c}, \alpha >$*

$\max\{2n^2(\beta \log \frac{1}{c}+1), 2\beta n\}$ *(lower bound in condition 5). Then, there exists a perfect (Y-saturating) n-gram matching in the random bipartite graph* $G(Y, X; E)$*, with probability at least* $1 - \gamma_1 p^{-\beta'}$ *for constants* $\beta' > 0$ *and* $\gamma_1 > 0$*, specified in* (5) *and* (6).

Note that the sufficient size bound $q = O(p^n)$ in the above theorem is also necessary (see Remark 3), and is therefore tight.

Remark 10 (Insufficiency of the union bound argument)**.** *It is easier to exploit the union bound arguments to propose random bipartite graphs which have a perfect n-gram matching* ***whp****. It is proved in Appendix B.1 that if* $d \geq n$ *and the size constraint* $|Y| = O(|X|^{\frac{n}{2}-\delta})$ *for some* $\delta > 0$ *is satisfied, then* ***whp****, the random bipartite graph has a perfect n-gram matching. Comparing this result with ours in Theorem 3, our approach has a better size scaling while the union bound approach has a better degree scaling. The size scaling limitation in the union bound argument makes it unattractive. In order to identify the population structure A in the overcomplete regime where* $|Y| = O(|X|^n)$*, we need access to at least* $(4n)$*-th order moment under the union bound argument, while only the* $(2n)$*-th order moment is required under our argument.*

5.2.2 Lower bound on the Kruskal rank of random matrices

In the following theorem, a lower bound on the Kruskal rank of a random matrix A under dimension and degree constraints is provided, which is proved in Appendix B.1.

Theorem 4 (Lower bound on the Kruskal rank of random matrices)**.** *Consider a random matrix* $A \in \mathbb{R}^{p \times q}$*, where for any* $i \in [q]$*, there are* d_i *number of random non-zero entries in column i. Let* $d_{\min} := \min_{i \in [q]} d_i$*. Assume that it satisfies the size condition* $q \leq \left(c\frac{p}{n}\right)^n$ *(condition 4) for some constant* $0 < c < 1$ *and the degree condition* $d_{\min} \geq 1 + \beta \log p$ *for some constant* $\beta > \frac{n-1}{\log 1/c}$ *(lower bound in condition 5) and in addition A is generic. Then,* $\operatorname{krank}(A) \geq cp$*, with probability at least* $1 - \gamma_2 p^{-\beta'}$ *for constants* $\beta' > 0$ *and* $\gamma_2 > 0$*, specified in* (5) *and* (7).

Acknowledgements

The authors acknowledge useful discussions with Sina Jafarpour, Adel Javanmard, Alex Dimakis, Moses Charikar, Sanjeev Arora, Ankur Moitra and Kamalika Chaudhuri. A. Anandkumar is supported in part by Microsoft Faculty Fellowship, NSF Career award CCF-1254106, NSF Award CCF-1219234, ARO Award W911NF-12-1-0404, and ARO YIP Award W911NF-13-1-0084. M. Janzamin is supported by NSF Award CCF-1219234, ARO Award W911NF-12-1-0404 and ARO YIP Award W911NF-13-1-0084.

Appendix

A Proof of Deterministic Identifiability Result (Theorem 1)

First, we show the identifiability result under an alternative set of conditions on the n-gram matrix, $A^{\odot n}$, and then, we show that the conditions of Theorem 1 are sufficient for these conditions to hold.

A.1 Deterministic analysis based on $A^{\odot n}$

In this section, the deterministic identifiability result based on conditions on the n-gram matrix, $A^{\odot n}$, is provided.

In the n-gram matrix, $A^{\odot n} \in \mathbb{R}^{p^n \times q}$, redundant rows exist. If some row of $A^{\odot n}$ is indexed by n-tuple $(i_1, \ldots, i_n) \in [p]^n$, then another row indexed by any permutation of the tuple $(i_1, \ldots, i_n)$ has the same entries. Therefore, the number of distinct rows of $A^{\odot n}$ is at most $\binom{p+n-1}{n}$. In the following definition, we define a non-redundant version of n-gram matrix which is restricted to the (potentially) distinct rows.

Definition 9 (Restricted n-gram matrix)**.** *For any matrix $A \in \mathbb{R}^{p \times q}$, restricted n-gram matrix $A^{\odot n}_{\text{Rest.}} \in \mathbb{R}^{s \times q}$, $s = \binom{p+n-1}{n}$, is defined as the restricted version of n-gram matrix $A^{\odot n} \in \mathbb{R}^{p^n \times q}$, where the redundant rows of $A^{\odot n}$ are removed, as explained above.*

Condition 6 (Rank condition)**.** *The n-gram matrix $A^{\odot n}$ is full column rank.*

Condition 7 (Graph expansion)**.** *Let $G(V_h, V_o^{(n)}; A^{\odot n})$ denote the bipartite graph with vertex sets V_h corresponding to the hidden variables (indexing the columns of $A^{\odot n}$) and $V_o^{(n)}$ corresponding to the n-th order observed variables (indexing the rows of $A^{\odot n}$) and edge matrix $A^{\odot n} \in \mathbb{R}^{|V_o^{(n)}| \times |V_h|}$. The bipartite graph $G(V_h, V_o^{(n)}; A^{\odot n})$ satisfies the following* expansion property *on the restricted version specified by $A^{\odot n}_{\text{Rest.}}$,*

$$\left| N_{A^{\odot n}_{\text{Rest.}}}(S) \right| \geq |S| + d_{\max}\left(A^{\odot n}\right), \quad \forall S \subseteq V_h,\ |S| > \text{krank}(A), \tag{25}$$

where $d_{\max}\left(A^{\odot n}\right)$ is the maximum node degree in set V_h.

Remark 11. *The expansion condition for the bag-of-words admixture model is provided in* (4)*, introduced in [7]. The proposed expansion condition in* (25) *is inherited from* (4)*, with two major modifications. First, the condition is appropriately generalized for our model which involves a graph with edges specified by the n-gram matrix, $A^{\odot n}$, as stated in* (23)*. Second, the expansion property* (4)*, proposed in [7], needs to be satisfied for all subsets S with size $|S| \geq 2$, which is a stricter condition than the one proposed here in* (25)*, since we can have* $\text{krank}(A) \gg 2$.

The deterministic identifiability result based on the conditions on $A^{\odot n}$, is stated in the following theorem for $n \geq 2$, while $n = 1$ case is addressed in Remarks 4 and 11. The identifiability result relies on access to the $(2n)$-th order moment of observed variables $x_l, l \in [2n]$, defined in equation (2) as

$$M_{2n}(x) := \mathbb{E}\left[(x_1 \otimes x_2 \otimes \cdots \otimes x_n)(x_{n+1} \otimes x_{n+2} \otimes \cdots \otimes x_{2n})^\top\right] \in \mathbb{R}^{p^n \times p^n}.$$

Theorem 5 (Generic identifiability under deterministic conditions on $A^{\odot n}$)**.** *Let $M_{2n}^{(n)}(x)$ (defined in equation (2)) be the $(2n)$-th order moment of the n-persistent topic model described in Section 2. If the model satisfies conditions 1, 6 and 7, then, for any $n \geq 2$, all the columns of population structure A are* generically identifiable *from $M_{2n}^{(n)}(x)$.*

Proof: Define $B := A^{\odot n} \in \mathbb{R}^{p^n \times q}$. Then, the moment characterized in equation (23) can be written as $M_{2n}^{(n)}(x) = B\mathbb{E}\left[hh^\top\right]B^\top$. Since both matrices $\mathbb{E}\left[hh^\top\right]$ and B have full column rank (from conditions 1 and 6), the rank of $B\mathbb{E}\left[hh^\top\right]B^\top$ is q where $q = O(p^n)$, and furthermore $\mathrm{Col}(B\mathbb{E}\left[hh^\top\right]B^\top) = \mathrm{Col}(B)$. Let $\mathcal{U} := \{u_1, \ldots, u_q\} \in \mathbb{R}^{p^n}$ be any basis of $\mathrm{Col}(B\mathbb{E}\left[hh^\top\right]B^\top)$ satisfying the following two properties:

1) u_i's have the smallest ℓ_0 norms.
2) u_i's have q smallest (tensor) ranks in the n-th order tensor form, i.e., $U_i := \mathrm{ten}(u_i), i \in [q]$, have q smallest ranks.

Let the columns of matrix B be b_i for $i \in [q]$. Since all the b_i's (which belong to $\mathrm{Col}(B\mathbb{E}\left[hh^\top\right]B^\top)$) are rank-1 in the n-th order tensor form (since $\mathrm{ten}(b_i) = a_i^{\circ n}$) and the number of non-zero entries in each of b_i's is at most $d_{\max}(B) = d_{\max}(A)^n$, we conclude that

$$\max_i \mathrm{Rank}(\mathrm{ten}(u_i)) = 1 \quad \text{and} \quad \max_i \|u_i\|_0 \leq d_{\max}(B). \tag{26}$$

The above bounds are concluded from the fact that $b_i \in \mathrm{Col}(B\mathbb{E}\left[hh^\top\right]B^\top),\ i \in [q]$, and therefore the ℓ_0 norm and the rank properties of b_i's are upper bounds for the corresponding properties of basis vectors u_i's (according to the proposed conditions for u_i's).
Now, exploiting these observations and also the genericity of A and the expansion condition 7, we show that the basis vectors u_i's are scaled columns of B. Since u_i for $i \in [q]$, is a vector in the column space of B, it can be represented as $u_i = Bv_i$ for some vector $v_i \in \mathbb{R}^q$. Equivalently, for any $i \in [q]$, $u_i = \sum_{j=1}^q v_i(j) b_j$ where $b_j = a_j^{\otimes n}$ is the j-th column of matrix B and $v_i(j)$ is a scalar which is the j-th entry of vector v_i. Then, the tensor form of u_i can be written as

$$\mathrm{ten}(u_i) = \sum_{j=1}^q v_i(j)\,\mathrm{ten}(b_j) = \sum_{j=1}^q v_i(j)\,\mathrm{ten}(a_j^{\otimes n}) = \sum_{j=1}^q v_i(j) a_j^{\circ n} = [[\mathrm{Diag}_n(v_i); \overbrace{A, \ldots, A}^{n \text{ times}}]], \tag{27}$$

where the last equality is based on the notation defined in Definition 8. We define $\widetilde{v}_i := [v_i(j)]_{j:v_i(j)\neq 0}$ as the vector which contains only the non-zero entries of v_i, i.e., $\widetilde{v}_i$ is the restriction of vector v_i to its support. Therefore, $\widetilde{v}_i \in \mathbb{R}^r$, where $r := \|v_i\|_0$. Furthermore, the matrix $\widetilde{A}_i := \{a_j : v_i(j) \neq 0\} \in \mathbb{R}^{p \times r}$ is defined as the restriction of A to its columns corresponding to the support of v_i. Let $(\widetilde{a}_i)_j$ denote the j-th column of $\widetilde{A}_i$. According to these definitions, equation (27) reduces to

$$\mathrm{ten}(u_i) = [[\mathrm{Diag}_n(\widetilde{v}_i); \overbrace{\widetilde{A}_i, \ldots, \widetilde{A}_i}^{n \text{ times}}]] = \sum_{j=1}^r \widetilde{v}_i(j)[(\widetilde{a}_i)_j]^{\circ n}, \tag{28}$$

which is derived by removing columns of A corresponding to the zero entries in v_i.
Next, we rule out that $\|v_i\|_0 \geq 2$ under two cases ($2 \leq \|v_i\|_0 \leq \mathrm{krank}(A)$ and $\mathrm{krank}(A) < \|v_i\|_0 \leq q$), to conclude that u_i's vectors are scaled columns of B.

Case 1: $2 \leq \|v_i\|_0 \leq \text{krank}(A)$**.** Here, the number of columns of $\widetilde{A}_i \in \mathbb{R}^{p \times \|v_i\|_0}$ is less than or equal to krank(A) and therefore it is full column rank. Since, all the components of CP representation in equation (28) are full column rank [13], for any [14] $n \geq 2$, we have $\text{Rank}(\text{ten}(u_i)) = r = \|v_i\|_0 > 1$, which contradicts the fact that $\max_i \text{Rank}(\text{ten}(u_i)) = 1$ in (26).

Case 2: $\text{krank}(A) < \|v_i\|_0 \leq q$**.** Here, we first restrict the n-gram matrix B to distinct rows, denoted by $B_{\text{Rest.}}$, as defined in Definition 9. Let $u'_i = B_{\text{Rest.}} v_i$. Since u'_i is the restricted version of u_i, we have

$$\begin{aligned} \|u_i\|_0 \geq \|u'_i\|_0 &= \|B_{\text{Rest.}} v_i\|_0 \\ &> \left|N_{B_{\text{Rest.}}}(\text{Supp}(v_i))\right| - |\,\text{Supp}(v_i)| \\ &\geq d_{\max}(B), \end{aligned}$$

where the second inequality is from Lemma 4, and the third inequality follows from the graph expansion property (condition 7). This result contradicts the fact that $\max_i \|u_i\|_0 \leq d_{\max}(B)$ in (26).

From above contradictions, $\|v_i\|_0 = 1$ and hence, columns of $B := A^{\odot n}$ are the scaled versions of u_i's. □

The following lemma is useful in the proof of Theorem 5. The result proposed in this lemma is similar to the parameter genericity condition in [7], but generalized for the n-gram matrix, $A^{\odot n}$. The lemma is proved on lines of the proof of Remark 2.2 in [7].

Lemma 4. *If $A \in \mathbb{R}^{p \times q}$ is generic, then the n-gram matrix $A^{\odot n} \in \mathbb{R}^{p^n \times q}$ satisfies the following property with Lebesgue measure one. For any vector $v \in \mathbb{R}^q$ with $\|v\|_0 \geq 2$, we have*

$$\left\|A^{\odot n}_{\text{Rest.}} v\right\|_0 > \left|N_{A^{\odot n}_{\text{Rest.}}}(\text{Supp}(v))\right| - |\,\text{Supp}(v)|,$$

where for a set $S \subseteq [q]$, $N_{A^{\odot n}}(S) := \{i \in [p]^n : A^{\odot n}(i,j) \neq 0 \text{ for some } j \in S\}$.

Here, we prove the result for the case of $n = 2$. The proof can be easily generalized to larger n.

Let $A := M + Z$ be generic, where M is an arbitrary matrix, perturbed by random continuous perturbations Z. Consider the 2-gram matrix $B := A \odot A \in \mathbb{R}^{p^2 \times q}$. It is shown that the restricted version of B, denoted by $\widetilde{B} := B_{\text{Rest.}} \in \mathbb{R}^{\frac{p(p+1)}{2} \times q}$, satisfies the above genericity condition. We first establish some definitions.

Definition 10. *We call a vector* fully dense *if all of its entries are non-zero.*

Definition 11. *We say a matrix has the* Null Space Property (NSP) *if its null space does not contain any fully dense vector.*

Claim 1. *Fix any $S \subseteq [q]$ with $|S| \geq 2$, and set $R := N_{M^{(2\text{-gram})}_{\text{Rest.}}}(S)$. Let $\widetilde{C}$ be a $|S| \times |S|$ submatrix of $\widetilde{B}_{R,S}$. Then* $\Pr(\widetilde{C}$ *has the NSP*$) = 1$.

[13] Note that for $n \geq 3$, this full rank condition can be relaxed by Kruskal's condition for uniqueness of CP decomposition [15] and its generalization to higher order tensors [44]. Precisely, instead of saying $\text{Rank}(\widetilde{A}_i) = \text{krank}(\widetilde{A}_i) = r$, it is only required to have $\text{krank}(\widetilde{A}_i) \geq (2r + n - 1)/n$ to argue the result of case 1. This only improves the constants involved in the final result.

[14] Note that for $n = 1$, since the (tensor) rank of any vector is 1, this analysis does not work.

Proof of Claim 1: First, note that $\widetilde{B}$ can be expanded as

$$\widetilde{B} := (A \odot A)_{\text{Rest.}} = (M \odot M)_{\text{Rest.}} + \underbrace{(M \odot Z + Z \odot M)_{\text{Rest.}} + (Z \odot Z)_{\text{Rest.}}}_{:=U}.$$

Let $s = |S|$ and let $\widetilde{C} = [\tilde{c}_1|\tilde{c}_2|\cdots|\tilde{c}_s]^\top$, where $\tilde{c}_i^\top$ is the i-th row of $\widetilde{C}$. Also, let $C := [c_1|c_2|\cdots|c_s]^\top$ and $W := [w_1|w_2|\cdots|w_s]^\top$ be the corresponding $|S| \times |S|$ submatrices of $M_{\text{Rest.}}^{(2\text{-gram})}$ and U, respectively. For each $i \in [s]$, denote by $\mathcal{N}_i$ the null space of the matrix $\widetilde{C}_i = [\tilde{c}_1|\tilde{c}_2|\cdots|\tilde{c}_i]^\top$. Finally let $\mathcal{N}_0 = \mathbb{R}^s$. Then, $\mathcal{N}_0 \supseteq \mathcal{N}_1 \supseteq \cdots \supseteq \mathcal{N}_s$. We need to show that, with probability one, $\mathcal{N}_s$ does not contain any fully dense vector.

If one of $\mathcal{N}_i, i \in [s]$, does not contain any full dense vector, the result is proved. Suppose that $\mathcal{N}_i$ contains some fully dense vector v. Since C is a submatrix of $M_{R,S}^{(2\text{-gram})}$, every row $c_{i+1}^\top$ of C contains at least one non-zero entry. Therefore,

$$\begin{aligned} v^\top \tilde{c}_{i+1} &= \sum_{j \in [s]} v(j)\tilde{c}_{i+1}(j) \\ &= \sum_{j \in [s]: c_{i+1}(j) \neq 0} v(j)(c_{i+1}(j) + w_{i+1}(j)), \end{aligned}$$

where $\{w_{i+1}(j) : j \in [s] \text{ s.t. } c_{i+1}(j) \neq 0\}$ are independent random variables, and moreover, they are independent of $\tilde{c}_1, \ldots, \tilde{c}_i$ and thus of v. By assumption on the distribution of the $w_{i+1}(j)$,

$$\Pr\left[v \in \mathcal{N}_{i+1} \middle| \tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_i\right] = \Pr\left[\sum_{j \in [s]: c_{i+1}(j) \neq 0} v(j)(c_{i+1}(j) + w_{i+1}(j)) = 0 \middle| \tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_i\right] = 0. \quad (29)$$

Consequently,

$$\Pr\left[\dim(\mathcal{N}_{i+1}) < \dim(\mathcal{N}_i) \middle| \tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_i\right] = 1 \quad (30)$$

for all $i = 0, \ldots, s-1$. As a result, with probability one, $\dim(\mathcal{N}_s) = 0$. □

Now, we are ready to prove Lemma 4.

Proof of Lemma 4: It follows from Claim 1 that, with probability one, the following event holds: for every $S \subseteq [q], |S| \geq 2$, and every $|S| \times |S|$ submatrix $\widetilde{C}$ of $\widetilde{B}_{R,S}$ where $R := N_{M_{\text{Rest.}}^{(2\text{-gram})}}(S)$, then $\tilde{C}$ has the NSP.

Now fix $v \in \mathbb{R}^q$ with $\|v\|_0 \geq 2$. Let $S := \text{Supp}(v)$ and $H := \widetilde{B}_{R,S}$. Furthermore, let $u \in (\mathbb{R} \setminus \{0\})^{|S|}$ be the restriction of vector v to S; observe that u is fully dense. It is clear that $\|\widetilde{B}v\|_0 = \|Hu\|_0$, so we need to show that

$$\|Hu\|_0 > |R| - |S|. \quad (31)$$

For the sake of contradiction, suppose that Hu has at most $|R| - |S|$ non-zero entries. Since $Hu \in \mathbb{R}^{|R|}$, there is a subset of $|S|$ entries on which Hu is zero. This corresponds to a $|S| \times |S|$ submatrix of $H := \widetilde{B}_{R,S}$ which contains u in its null space. It means that this submatrix does not have the NSP, which is a contradiction. Therefore we conclude that Hu must have more than $|R| - |S|$ non-zero entries, which finishes the proof. □

A.2 Proof of moment characterization lemmata

Remark 12. *In Lemmata 1 and 2, a specific case of order and persistence ($m = rn$) was considered. Here, we provide the moment form for a more general case. Assume that $m = rn + s$ for some integers $r \geq 1, 1 \leq s \leq \frac{n}{2}$, then*

$$M_{2m}^{(n)}(x) = \Big(\overbrace{A^{\odot n} \otimes \cdots \otimes A^{\odot n}}^{r \text{ times}} \otimes A^{(s\text{-gram})}\Big) \widetilde{M}_{2r}(h) \Big(A^{((n-s)\text{-gram})} \otimes \overbrace{A^{\odot n} \otimes \cdots \otimes A^{\odot n}}^{r-1 \text{ times}} \otimes A^{(2s\text{-gram})}\Big)^\top,$$

where $\widetilde{M}_{2r}(h) \in \mathbb{R}^{q^{r+1} \times q^{r+1}}$ is the hidden moment as

$$\widetilde{M}_{2r}(h)_{\big((i_1,\ldots,i_{r+1}),(j_1,\ldots,j_{r+1})\big)} := \begin{cases} \mathbb{E}[h_{i_1} \cdots h_{i_r} h_{i_{r+1}}^2 h_{j_2} \cdots h_{j_{r+1}}] & \text{if } i_{r+1} = j_1, \\ 0 & \text{o. w .} \end{cases}$$

The tensor form is also characterized as

$$T_{2m}^{(n)}(x) = \Big[\Big[\widetilde{S}_r; \overbrace{A, A, \ldots, A}^{2m \text{ times}}\Big]\Big],$$

where $\widetilde{S}_r \in \bigotimes^{2m} \mathbb{R}^q$ is the core tensor in the above Tucker representation with the sparsity pattern as follows. Let $\mathbf{i} := (i_1, i_2, \ldots, i_{2m})$. If

$$i_1 = i_2 = \cdots = i_n, i_{n+1} = i_{n+2} = \cdots = i_{2n}, \cdots, i_{(2r-1)n+1} = i_{(2r-1)n+2} = \cdots = i_{2rn},$$
$$i_{2(m-s)+1} = i_{2(m-s)+2} = \cdots = i_{2m},$$

we have

$$\widetilde{S}_r(\mathbf{i}) = \widetilde{M}_{2r}(h)_{\big((i_n, i_{2n}, \ldots, i_{rn}, i_m), (i_{(r+1)n}, i_{(r+2)n}, \ldots, i_{2rn}, i_{2m})\big)}.$$

Otherwise, $\widetilde{S}_r(\mathbf{i}) = 0$.

Proof of Lemma 1: In order to simplify the notation, similar to tensor powers for vectors, the tensor power for a matrix $U \in \mathbb{R}^{p \times q}$ is defined as

$$U^{\otimes r} := \overbrace{U \otimes U \otimes \cdots \otimes U}^{r \text{ times}} \in \mathbb{R}^{p^r \times q^r}. \tag{32}$$

First, consider the case $m = rn$ for some integer $r \geq 1$. One advantage of encoding $y_j, j \in [2r]$, by basis vectors appears in characterizing the conditional moments. The first order conditional moment of words $x_l, l \in [2m]$, in the n-persistent topic model can be written as

$$\mathbb{E}\big[x_{(j-1)n+k} | y_j\big] = A y_j, \ j \in [2r], \ k \in [n],$$

where $A = [a_1 | a_2 | \cdots | a_q] \in \mathbb{R}^{p \times q}$. Next, the m-th order conditional moment of different views $x_l, l \in [m]$, in the n-persistent topic model can be written as

$$\mathbb{E}[x_1 \otimes x_2 \otimes \cdots \otimes x_m | y_1 = e_{i_1}, y_2 = e_{i_2}, \ldots, y_r = e_{i_r}] = a_{i_1}^{\otimes n} \otimes a_{i_2}^{\otimes n} \otimes \cdots \otimes a_{i_r}^{\otimes n},$$

which is derived from the conditional independence relationships among the observations $x_l, l \in [m]$, given topics $y_j, j \in [r]$. Similar to the first order moments, since vectors $y_j, j \in [r]$, are encoded by the basis vectors $e_i \in \mathbb{R}^q$, the above moment can be written as the following matrix multiplication

$$\mathbb{E}[x_1 \otimes x_2 \otimes \cdots \otimes x_m | y_1, y_2, \ldots, y_r] = \left(A^{\odot n}\right)^{\otimes r} (y_1 \otimes y_2 \otimes \cdots \otimes y_r), \tag{33}$$

where the $(\cdot)^{\otimes r}$ notation is defined in equation (32). Now for the $(2m)$-th order moment, we have

$$\begin{aligned}
M_{2m}^{(n)}(x) &:= \mathbb{E}\Big[(x_1 \otimes x_2 \otimes \cdots \otimes x_m)(x_{m+1} \otimes x_{m+2} \otimes \cdots \otimes x_{2m})^\top\Big] \\
&= \mathbb{E}_{(y_1,y_2,\ldots,y_{2r})}\Big[\mathbb{E}\left[(x_1 \otimes \cdots \otimes x_m)(x_{m+1} \otimes \cdots \otimes x_{2m})^\top | y_1, y_2, \ldots, y_{2r}\right]\Big] \\
&\overset{(a)}{=} \mathbb{E}_{(y_1,y_2,\ldots,y_{2r})}\Big[\mathbb{E}\big[(x_1 \otimes \cdots \otimes x_m)|y_1, \ldots, y_{2r}\big]\mathbb{E}\big[(x_{m+1} \otimes \cdots \otimes x_{2m})^\top | y_1, \ldots, y_{2r}\big]\Big] \\
&\overset{(b)}{=} \mathbb{E}_{(y_1,y_2,\ldots,y_{2r})}\Big[\mathbb{E}\big[(x_1 \otimes \cdots \otimes x_m)|y_1, \ldots, y_r\big]\mathbb{E}\big[(x_{m+1} \otimes \cdots \otimes x_{2m})^\top | y_{r+1}, \ldots, y_{2r}\big]\Big] \\
&\overset{(c)}{=} \mathbb{E}_{(y_1,y_2,\ldots,y_{2r})}\left[\left(\left[A^{\odot n}\right]^{\otimes r}\right)(y_1 \otimes \cdots \otimes y_r)(y_{r+1} \otimes \cdots \otimes y_{2r})^\top \left(\left[A^{\odot n}\right]^{\otimes r}\right)^\top\right] \\
&= \left(\left[A^{\odot n}\right]^{\otimes r}\right)\mathbb{E}\left[(y_1 \otimes \cdots \otimes y_r)(y_{r+1} \otimes \cdots \otimes y_{2r})^\top\right]\left(\left[A^{\odot n}\right]^{\otimes r}\right)^\top \\
&\overset{(d)}{=} \left(\left[A^{\odot n}\right]^{\otimes r}\right) M_{2r}(y) \left(\left[A^{\odot n}\right]^{\otimes r}\right)^\top,
\end{aligned} \tag{34}$$

where (a) results from the independence of $(x_1, \ldots, x_m)$ and $(x_{m+1}, \ldots, x_{2m})$ given $(y_1, y_2, \ldots, y_{2r})$ and (b) is concluded from the independence of $(x_1, \ldots, x_m)$ and $(y_{r+1}, \ldots, y_{2r})$ given $(y_1, \ldots, y_r)$ and the independence of $(x_{m+1}, \ldots, x_{2m})$ and $(y_1, \ldots, y_r)$ given $(y_{r+1}, \ldots, y_{2r})$. Equation (33) is used in (c) and finally, the $(2r)$-th order moment of $(y_1, \ldots, y_{2r})$ is defined as $M_{2r}(y) := \mathbb{E}\left[(y_1 \otimes \cdots \otimes y_r)(y_{r+1} \otimes \cdots \otimes y_{2r})^\top\right]$ in (d).

For $M_{2r}(y)$, we have by the law of total expectation

$$\begin{aligned}
M_{2r}(y) &:= \mathbb{E}\big[(y_1 \otimes \cdots \otimes y_r)(y_{r+1} \otimes \cdots \otimes y_{2r})^\top\big] \\
&= \mathbb{E}_h\Big[\mathbb{E}\big[(y_1 \otimes \cdots \otimes y_r)(y_{r+1} \otimes \cdots \otimes y_{2r})^\top | h\big]\Big] \\
&= \mathbb{E}_h\left[\Big(\overbrace{h \otimes \cdots \otimes h}^{r \text{ times}}\Big)\Big(\overbrace{h \otimes \cdots \otimes h}^{r \text{ times}}\Big)^\top\right] \\
&= M_{2r}(h),
\end{aligned}$$

where the third equality is concluded from the conditional independence of variables $y_j, j \in [2r]$, given h and the model assumption that $\mathbb{E}\big[y_j|h\big] = h, j \in [2r]$. Substituting this in equation (34), finishes the proof for the n-persistent topic model. Similarly, the moment of single topic model (infinite persistence) can be also derived. □

Proof of Lemma 2: Defining $\Lambda := M_{2r}(h) \in \mathbb{R}^{q^r \times q^r}$ and $B := \big[A^{\odot n}\big]^{\otimes r} \in \mathbb{R}^{p^{rn} \times q^r}$, the $(2rn)$-th order moment $M_{2rn}^{(n)}(x) \in \mathbb{R}^{p^{rn} \times p^{rn}}$ of the n-persistent topic model proposed in equation (8) can be

written as

$$M_{2rn}^{(n)}(x) = B\Lambda B^\top.$$

Let $b_{(i_1,\dots,i_r)} \in \mathbb{R}^{p^{rn}}$ denote the corresponding column of B indexed by r-tuple $(i_1,\dots,i_r), i_k \in [q], k \in [r]$. Then, the above matrix equation can be expanded as

$$\begin{aligned} M_{2rn}^{(n)}(x) &= \sum_{\substack{i_1,\dots,i_r\in[q]\\ j_1,\dots,j_r\in[q]}} \Lambda\big((i_1,\dots,i_r),(j_1,\dots,j_r)\big) b_{(i_1,\dots,i_r)} b_{(j_1,\dots,j_r)}^\top \\ &= \sum_{\substack{i_1,\dots,i_r\in[q]\\ j_1,\dots,j_r\in[q]}} \Lambda\big((i_1,\dots,i_r),(j_1,\dots,j_r)\big) [a_{i_1}^{\otimes n} \otimes \cdots \otimes a_{i_r}^{\otimes n}][a_{j_1}^{\otimes n} \otimes \cdots \otimes a_{j_r}^{\otimes n}]^\top, \end{aligned}$$

where relation $b_{(i_1,\dots,i_r)} = a_{i_1}^{\otimes n} \otimes \cdots \otimes a_{i_r}^{\otimes n}, i_1,\dots,i_r \in [q]$, is used in the last equality. Let $m_{2rn}^{(n)}(x) \in \mathbb{R}^{p^{2rn}}$ denote the vectorized form of $(2rn)$-th order moment $M_{2rn}^{(n)}(x) \in \mathbb{R}^{p^{rn}\times p^{rn}}$. Therefore, we have

$$\begin{aligned} m_{2rn}^{(n)}(x) &:= \mathrm{vec}\Big(M_{2rn}^{(n)}(x)\Big) \\ &= \sum_{\substack{i_1,\dots,i_r\in[q]\\ j_1,\dots,j_r\in[q]}} \Lambda\big((i_1,\dots,i_r),(j_1,\dots,j_r)\big) a_{i_1}^{\otimes n} \otimes \cdots \otimes a_{i_r}^{\otimes n} \otimes a_{j_1}^{\otimes n} \otimes \cdots \otimes a_{j_r}^{\otimes n}. \end{aligned}$$

Then, we have the following equivalent tensor form for the original model proposed in equation (8)

$$\begin{aligned} T_{2rn}^{(n)}(x) &:= \mathrm{ten}\Big(m_{2rn}^{(n)}(x)\Big) \\ &= \sum_{\substack{i_1,\dots,i_r\in[q]\\ j_1,\dots,j_r\in[q]}} \Lambda\big((i_1,\dots,i_r),(j_1,\dots,j_r)\big) a_{i_1}^{\circ n} \circ \cdots \circ a_{i_r}^{\circ n} \circ a_{j_1}^{\circ n} \circ \cdots \circ a_{j_r}^{\circ n}. \end{aligned}$$

□

A.3 Sufficient matching properties for satisfying rank and graph expansion conditions

In the following lemma, it is shown that under a perfect n-gram matching and additional genericity and krank conditions, the rank and graph expansion conditions 6 and 7 on $A^{\odot n}$, are satisfied.

Lemma 5. *Assume that the bipartite graph $G(V_h, V_o; A)$ has a perfect n-gram matching (condition 2 is satisfied). Then, the following results hold for the n-gram matrix $A^{\odot n}$:*

1) *If A is generic, $A^{\odot n}$ is full column rank (condition 6) with Lebesgue measure one (almost surely).*
2) *If krank condition 3 holds, $A^{\odot n}$ satisfies the proposed expansion property in condition 7.*

Proof: Let M denote the perfect n-gram matching of the bipartite graph $G(V_h, V_o; A)$. From Lemma 3, there exists a perfect matching $M^{\odot n}$ for the bipartite graph $G(V_h, V_o^{(n)}; A^{\odot n})$. Denote the corresponding bi-adjacency matrix to the edge set M as A_M. Similarly, B_M denotes the corresponding bi-adjacency matrix to the edge set $M^{\odot n}$. Note that $\mathrm{Supp}(A_M) \subseteq \mathrm{Supp}(A)$ and $\mathrm{Supp}(B_M) \subseteq \mathrm{Supp}(A^{\odot n})$.

Since B_M is a perfect matching, it consists of $q := |V_h|$ rows, each of which has only one non-zero entry, and furthermore, the non-zero entries are in q different columns. Therefore, these rows form q linearly independent vectors. Since the row rank and column rank of a matrix are equal, and the number of columns of B_M is q, the column rank of B_M is q or in other words, B_M is full column rank. Since A is generic, from Lemma 6 (with a slight modification in the analysis [15]), $A^{\odot n}$ is also full column rank with Lebesgue measure one (almost surely). This completes the proof of part 1.

Next, the second part is proved. From krank definition, we have

$$|N_A(S')| \geq |S'| \quad \text{for } S' \subseteq V_h, |S'| \leq \mathrm{krank}(A),$$

which is concluded from the fact that the corresponding submatrix of A specified by S' should be full column rank. From this inequality, we have

$$|N_A(S')| \geq \mathrm{krank}(A) \quad \text{for } S' \subseteq V_h, |S'| = \mathrm{krank}(A). \tag{35}$$

Then, we have

$$\begin{aligned} |N_A(S)| &\geq |N_A(S')| \quad \text{for } S' \subset S \subseteq V_h, |S| > \mathrm{krank}(A), |S'| = \mathrm{krank}(A), \\ &\geq \mathrm{krank}(A) \\ &\geq d_{\max}(A)^n, \end{aligned} \tag{36}$$

where (35) is used in the second inequality and the last inequality is from krank condition 3.

In the restricted n-gram matrix $A^{\odot n}_{\mathrm{Rest.}}$, the number of neighbors for a set $S \subseteq V_h, |S| > \mathrm{krank}(A)$, can be bounded as

$$\begin{aligned} \left| N_{A^{\odot n}_{\mathrm{Rest.}}}(S) \right| &\geq |N_A(S)| + |S| \\ &\geq d_{\max}(A)^n + |S| \quad \text{for } |S| > \mathrm{krank}(A), \end{aligned}$$

where the first inequality is due to the fact that the set $N_{A^{\odot n}_{\mathrm{Rest.}}}$ consists of rows indexed by the following two subsets: n-tuples $(i, i, \ldots, i)$ where all the indices are equal and n-tuples $(i_1, \ldots, i_n)$ with distinct indices, i.e., $i_1 \neq i_2 \ldots \neq i_n$. The former subset is exactly $N_A(S)$ while the size of the latter subset is at least $|S|$ due to the existence of a perfect n-gram matching in A. The bound (36) is used in the second inequality. Since $d_{\max}\big(A^{\odot n}\big) = d_{\max}(A)^n$, the proof of part 2 is also completed.

□

Remark 13. *The second result of above lemma is similar to the necessity argument of (Hall's) Theorem 6 for the existence of perfect matching in a bipartite graph, but generalized to the case of perfect n-gram matching and with additional krank condition.*

[15] Lemma 6 result is about the column rank of A itself, but here it is about the column rank of $A^{\odot n}$ for which the same analysis works. Note that the support of B_M (which is full column rank here) is within the support of $A^{\odot n}$ and therefore Lemma 6 can still be applied.

A.4 (Auxiliary) lemma

Proof of Lemma 3: We show that if $G(Y, X; A)$ has a perfect n-gram matching, then $G(Y, X^{(n)}; A^{\odot n})$ has a perfect matching. The reverse can be also immediately shown by reversing the discussion and exploiting the additional condition stated in the lemma.
Let $E^{\odot n}$ denote the edge set of the bipartite graph $G(Y, X^{(n)}; A^{\odot n})$. Assume $G(Y, X; A)$ has a perfect n-gram matching $M \subseteq E$. For any $j \in Y$, let $N_M(j)$ denote the set of neighbors of vertex j according to edge set M. Since M is a perfect n-gram matching, $|N_M(j)| = n$ for all $j \in Y$. It can be immediately concluded from Definition 3 that sets $N_M(j)$ are all distinct, i.e., $N_M(j_1) \neq N_M(j_2)$ for any $j_1, j_2 \in Y, j_1 \neq j_2$. For any $j \in Y$, let $N'_M(j)$ denote an arbitrary ordered n-tuple generated from the elements of set $N_M(j)$. From the definition of n-gram matrix, we have $A^{\odot n}(N'_M(j), j) \neq 0$ for all $j \in Y$. Hence, $(j, N'_M(j)) \in E^{\odot n}$ for all $j \in Y$ which together with the fact that all $N'_M(j)$'s tuples are distinct, it results that $M^{\odot n} := \{(j, N'_M(j)) | j \in Y\} \subseteq E^{\odot n}$ is a perfect matching for $G(Y, X^{(n)}; A^{\odot n})$. □

Lemma 6. *Consider matrix $C \in \mathbb{R}^{m \times r}$ which is generic. Let $\widetilde{C} \in \mathbb{R}^{m \times r}$ be such that* $\text{Supp}(\widetilde{C}) \subseteq \text{Supp}(C)$ *and the non-zero entries of $\widetilde{C}$ are the same as the corresponding non-zero entries of C. If $\widetilde{C}$ is full column rank, then C is also full column rank, almost surely.*

Proof: Since $\widetilde{C}$ is full column rank, there exists a $r \times r$ submatrix of $\widetilde{C}$, denoted by $\widetilde{C}_S$, with non-zero determinant, i.e., $\det(\widetilde{C}_S) \neq 0$. Let C_S denote the corresponding submatrix of C indexed by the same rows and columns as $\widetilde{C}_S$.
The determinant of C_S is a polynomial in the entries of C_S. Since $\widetilde{C}_S$ can be derived from C_S by keeping the corresponding non-zero entries, $\det(C_S)$ can be decomposed into two terms as

$$\det(C_S) = \det(\widetilde{C}_S) + f(C_S),$$

where the first term corresponds to the monomials for which all the variables (entries of C_S) are also in $\widetilde{C}_S$ and the second term corresponds to the monomials for which at least one variable is not in $\widetilde{C}_S$. The first term is non-zero as stated earlier. Since C is generic, the polynomial $f(C_S)$ is non-trivial and therefore its roots have Lebesgue measure zero. It implies that $\det(C_S) \neq 0$ with Lebesgue measure one (almost surely), and hence, it is full (column) rank. Thus, C is also full column rank, almost surely. □

Finally, Theorem 1 is proved by combining the results of Theorem 5 and Lemma 5.

Proof of Theorem 1: Since conditions 2 and 3 hold and A is generic, Lemma 5 can be applied which results that rank condition 6 is satisfied almost surely and expansion condition 7 also holds. Therefore, all the required conditions for Theorem 5 are satisfied almost surely and this completes the proof. □

B Proof of Random Identifiability Result (Theorem 2)

We provide detailed proof of the steps stated in the proof sketch of random result in Section 5.2.

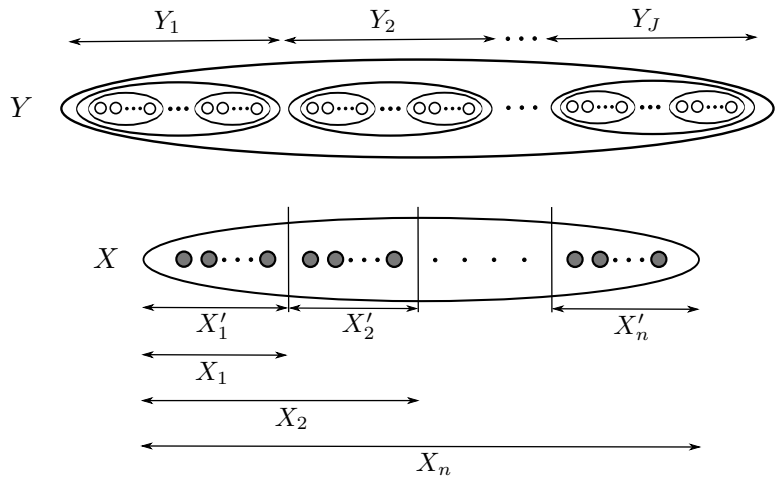


Figure 6: Partitioning of sets Y and X, proposed in the proof of Theorem 3. Set X is randomly (uniform) partitioned into n sets of (almost) equal size, denoted by $X'_l, l \in [n]$. Set Y is also randomly partitioned in a recursive manner. In each step, it is partitioned to $J = c\frac{p}{n} = O(p)$ number of sets. These smaller sets are again partitioned, recursively. This partitioning process is performed until reaching sets with size $O(p)$. The first two steps are shown in this figure.

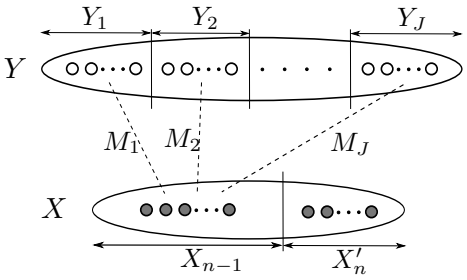
B.1 Proof of existence of perfect n-gram matching and Kruskal results

Proof of Theorem 3: Vertex sets X and Y are partitioned, described as follows (see Figure 6). Define $J := c\frac{p}{n}$. Partition set X uniformly at random into n sets of (almost) equal size [16], denoted by $X'_l, l \in [n]$. Define sets $X_l := \cup_{i=1}^{l} X'_i, l \in [n]$. Furthermore, partition set Y uniformly at random, hierarchically as follows. First, partition into J sets, each with size at most $\left(c\frac{p}{n}\right)^{n-1}$, and denote them by $Y_i, i \in [J]$. Next, partition each of these new smaller sets Y_i further into J sets, each with size at most $\left(c\frac{p}{n}\right)^{n-2}$. Do it iteratively up to $n-1$ steps, where at the end, set Y is partitioned into sets with size at most $c\frac{p}{n}$. The first two steps are shown in Figure 6.

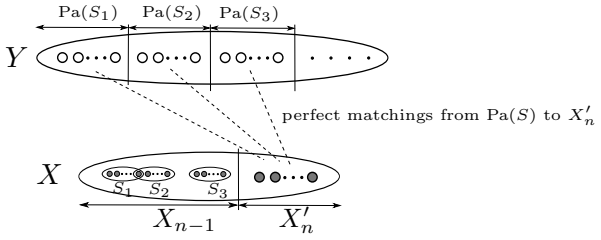
Proof by induction: The existence of perfect n-gram matching from set Y to set X is proved by an induction argument. Consider one of intermediate sets in the hierarchical partitioning of Y with size $O(p^l)$ and its further partitioning into $J := c\frac{p}{n}$ sets, each with size $O(p^{l-1})$, for any $l \in \{2, \ldots, n\}$. In the induction step, it is shown that if there exists a perfect $(l-1)$-gram matching from each of these subsets of Y with size $O(p^{l-1})$ to X_{l-1}, then there exists a perfect l-gram matching from the original set with size $O(p^l)$ to set X_l. Specifically, in the last induction step, it is shown that if there exists a perfect $(n-1)$-gram matching from each set $Y_l, l \in [J]$, to set X_{n-1}, then there exists a perfect n-gram matching from Y to $X_n = X$.

Base case: The base case of induction argument holds as follows. By applying Lemma 8 and Lemma 7, there exists a perfect matching from each partition in Y with size at most $c\frac{p}{n} = O(p)$ to set X_1, **whp**.

[16] By almost, we mean the maximum difference in the size of partitions is 1 which is always possible.



(a) Partitioning of sets Y and X proposed for the induction step.



(b) Partitioning of set Y through perfect $(n-1)$-gram matchings $M_i, i \in [J]$.

Figure 7: Auxiliary figures for proof of induction step. (a) Partitioning of sets Y and X proposed in the proof, where set Y is partitioned to $J := c\frac{p}{n}$ partitions $Y_1, \ldots, Y_J$ with (almost) equal size, for some constant $c < 1$. In addition, set X is partitioned to two partitions X_{n-1} and X'_n with sizes $|X_{n-1}| = \frac{n-1}{n}p$ and $|X'_n| = \frac{p}{n}$. The perfect $(n-1)$-gram matchings $M_i, i \in [J]$, through bipartite graphs $G_i(Y_i, X_{n-1}; E_i), i \in [J]$, are also highlighted in the figure. (b) Set Y is partitioned to subsets $\mathrm{Pa}(S), S \in P_{n-1}(X_{n-1})$, which is generated through perfect $(n-1)$-gram matchings $M_i, i \in [J]$. S_1, S_2 and S_3 are three different sets in $P_{n-1}(X_{n-1})$ shown as samples. In addition, the perfect matchings from $\mathrm{Pa}(S), S \in P_{n-1}(X_{n-1})$, to X'_n, proposed in the proof, are also highlighted in the figure.

Induction step: Consider J different bipartite graphs $G_i(Y_i, X_{n-1}; E_i), i \in [J]$, by considering sets Y_i and X_{n-1} and the corresponding subset of edges $E_i \subset E$ incident to them. See Figure 7a. The induction step is to show that if each of the corresponding J bipartite graphs $G_i(Y_i, X_{n-1}; E_i), i \in [J]$, has a perfect $(n-1)$-gram matching, then **whp**, the original bipartite graph $G(Y, X; E)$ has a perfect n-gram matching.

Let us denote the corresponding perfect $(n-1)$-gram matching of $G_i(Y_i, X_{n-1}; E_i)$ by M_i. Furthermore, the set of all subsets of X_{n-1} with cardinality $n-1$ are denoted by $P_{n-1}(X_{n-1})$, i.e., $P_{n-1}(X_{n-1})$ includes the sets with $(n-1)$ elements in the power set [17] of X_{n-1}. For each set $S \in P_{n-1}(X_{n-1})$, take the set of all nodes in Y which are connected to all members of S according to the union of matchings $\cup_{i=1}^{J} M_i$. Call this set as the parents of S, denoted by $\mathrm{Pa}(S)$. According to the definition of perfect $(n-1)$-gram matching, there is at most one node in each set Y_i which is connected to all members of S through the matching M_i and therefore, $|\mathrm{Pa}(S)| \leq J = c\frac{p}{n}$. In addition, note that sets $\mathrm{Pa}(S)$ impose a partitioning on set Y, i.e., each node $j \in Y$ is exactly included in one set $\mathrm{Pa}(S)$ for some $S \in P_{n-1}(X_{n-1})$. This is because of the perfect $(n-1)$-gram matchings considered for sets $Y_i, i \in [J]$.

Now, a perfect n-gram matching for the original bipartite graph is constructed as follows. For any $S \in P_{n-1}(X_{n-1})$, consider the set of parents $\mathrm{Pa}(S)$. Create the bipartite graph $G_S(\mathrm{Pa}(S), X'_n; E_S)$, where $E_S \subset E$ is the subset of edges incident to partitions $\mathrm{Pa}(S) \subset Y$ and $X'_n \subset X$. Denote by d_S the minimum degree of nodes in set $\mathrm{Pa}(S)$ in the bipartite graph $G_S(\mathrm{Pa}(S), X'_n; E_S)$. Applying Lemma 8, we have

$$\begin{aligned} \Pr[d_S \geq 1 + \beta \log(p/n)] &\geq 1 - J \exp\left(-\frac{2}{n^2}\frac{(d_{\min} - \beta n \log(p/n))^2}{d_{\min}}\right) \\ &\geq 1 - \frac{c}{n} p^{-\beta \log 1/c} = 1 - O(p^{-\beta \log 1/c}), \end{aligned} \tag{37}$$

[17] The power set of any set S is the set of all subsets of S.

where $\beta \log 1/c > n-1$, and the last inequality is concluded from the degree bound $d_{\min} \geq \alpha \log p$. Furthermore, we have $|\operatorname{Pa}(S)| \leq c\frac{p}{n} = c|X'_n|$. Now, we can apply Lemma 7 concluding that there exists a perfect matching from $\operatorname{Pa}(S)$ to X'_n within the bipartite graph $G_S(\operatorname{Pa}(S), X'_n; E_S)$, with probability at least $1 - O(p^{-\beta \log 1/c})$. Refer to Figure 7b for a schematic picture. The edges of this perfect matching are combined with the corresponding edges of the existing perfect $(n-1)$-gram matchings $M_i, i \in [J]$, to provide n incident edges to each node $i \in \operatorname{Pa}(S)$. It is easy to see that this provides a perfect n-gram matching from $\operatorname{Pa}(S)$ to X.
We perform the same steps for all sets $S \in P_{n-1}(X_{n-1})$ to obtain a perfect n-gram matching from any $\operatorname{Pa}(S), S \in P_{n-1}(X_{n-1})$, to X. Finally, according to this construction, the union of all of these matchings is a perfect n-gram matching from $\cup_{S \in P_{n-1}(X_{n-1})} \operatorname{Pa}(S) = Y$ to X. This finishes the proof of induction step. Note that here we analyzed the last induction step where the existence of perfect n-gram matching is concluded from the existence of corresponding perfect $(n-1)$-gram matchings. The earlier induction steps, where the existence of perfect l-gram matching is concluded from the existence of corresponding perfect $(l-1)$-gram matchings for any $l \in \{2, \dots, n\}$, can be similarly proven.

Probability rate: We now provide the probability rate of the above events. Let $N_l^{(\text{hp})}, l \in [n]$, denote the total number of times that perfect matching result of Lemma 7 is used in step l in order to ensure that there exists a perfect l-gram matching from corresponding partitions of Y to set X_l, **whp**. Let $N^{(\text{hp})} = \sum_{l \in [n]} N_l^{(\text{hp})}$. As earlier, let $P_{l-1}\big(X_{l-1}\big)$ denote the set of all subsets of X_{l-1} with cardinality $l-1$. We have

$$\big|P_{l-1}\big(X_{l-1}\big)\big| = \binom{\big|X_{l-1}\big|}{l-1} = \binom{\frac{l-1}{n}p}{l-1}, \quad l \in \{2, \dots, n\}.$$

According to the construction method of l-gram matching from $(l-1)$-gram matchings, proposed in the induction step, $\big|P_{l-1}\big(X_{l-1}\big)\big|$ is the number of times Lemma 7 is used in order to ensure that there exists a perfect l-gram matching for each partition on the Y side. Since at most J^{n-l} number of such l-gram matchings are proposed in step l, the number $N_l^{(\text{hp})}$ can be bounded as

$$N_l^{(\text{hp})} \leq J^{n-l} \ \big|P_{l-1}\big(X_{l-1}\big)\big| = J^{n-l}\binom{\frac{l-1}{n}p}{l-1}, \quad l \in \{2, \dots, n\}. \tag{38}$$

Since in the first step, $N_1^{(\text{hp})} = J^{n-1}$ number of perfect matchings needs to exist in the above discussion, we have

$$\begin{aligned}
N^{(\text{hp})} &= J^{n-1} + \sum_{l=2}^{n} N_l^{(\text{hp})} \\
&\leq J^{n-1} + \sum_{l=2}^{n} J^{n-l}\binom{\frac{l-1}{n}p}{l-1} \\
&\leq \Big(c\frac{p}{n}\Big)^{n-1} + \sum_{l=2}^{n}\Big(c\frac{p}{n}\Big)^{n-l}\Big(e\frac{p}{n}\Big)^{l-1} \\
&\leq n\Big(e\frac{p}{n}\Big)^{n-1} = O(p^{n-1}),
\end{aligned}$$

where inequality (38) is used in the first inequality and $J := c\frac{p}{n}$ and inequality $\binom{n}{k} \leq \left(e\frac{n}{k}\right)^k$ are exploited in the second inequality.
Since the result of Lemma 7 holds with probability at least $1 - O(p^{-\beta \log 1/c})$ and it is assumed that $\beta \log 1/c > n-1$, by applying union bound, we have the existence of perfect n-gram matching with probability at least $1 - O(p^{-\beta'})$, for $\beta' = \beta \log \frac{1}{c} - (n-1) > 0$.
Furthermore, note that the degree concentration bound in (37) is also used $O(p^{n-1})$ times. Since the bound in (37) holds with probability at least $1-O(p^{-\beta \log 1/c})$ and it is assumed that $\beta \log 1/c > n-1$, this also reduces to the same probability rate.
The coefficient of the above polynomial probability rate is also explicitly computed, saying that the perfect n-gram matching exists with probability at least $1 - \gamma_1 p^{-\beta'}$, with

$$\gamma_1 = e^{n-1}\Big(\frac{c}{n^{n-1}} + \frac{e^2}{1-\delta_1} n^{\beta'+1}\Big),$$

where δ_1 is a constant satisfying $e^2\left(\frac{p}{n}\right)^{-\beta \log 1/c} < \delta_1 < 1$. □

Proof of Theorem 4: Let $G(Y, X; A)$ denote the corresponding bipartite graph to matrix A where node sets $Y = [q]$ and $X = [p]$ index the columns and rows of A respectively. Therefore, $|Y| = q$ and $|X| = p$. Fix some $S \subseteq Y$ such that $|S| \leq p$. Then

$$\begin{aligned}
\Pr(|N(S)| \leq |S|) &\leq \sum_{\substack{T \subseteq X: \\ |T|=|S|}} \Pr(N(S) \subseteq T) \\
&= \sum_{\substack{T \subseteq X: \\ |T|=|S|}} \prod_{i \in S} \binom{|S|}{d_i} \Big/ \binom{p}{d_i} \\
&\leq \sum_{\substack{T \subseteq X: \\ |T|=|S|}} \prod_{i \in S} \left(\frac{|S|}{p}\right)^{d_i} \\
&\leq \sum_{\substack{T \subseteq X: \\ |T|=|S|}} \prod_{i \in S} \left(\frac{|S|}{p}\right)^{d_{\min}} \\
&= \binom{p}{|S|} \left(\frac{|S|}{p}\right)^{d_{\min}|S|}, \qquad (39)
\end{aligned}$$

where the bound $\binom{|S|}{d_i}\Big/\binom{p}{d_i} \leq \left(\frac{|S|}{p}\right)^{d_i}$ is used in the second inequality, and the last inequality is concluded from the fact that $\frac{|S|}{p} \leq 1$.
Let $\mathcal{E}$ denote the event that for any subset $S \subseteq Y$ with $|S| \leq r$, we have $|N(S)| \geq |S|$, i.e.,

$$\mathcal{E} := \text{“}\forall S \subseteq Y \wedge 1 \leq |S| \leq r : |N(S)| \geq |S|\text{”}.$$

Then, by the union bound and inequality (39), we have

$$\Pr(\mathcal{E}^c) = \Pr(\exists S \subseteq Y \,\text{s.t.}\, 1 \leq |S| \leq r \wedge |N(S)| < |S|) \leq \sum_{s=1}^{r} \binom{q}{s} \binom{p}{s} \left(\frac{s}{p}\right)^{d_{\min} s}$$

$$\leq \sum_{s=1}^{r} \left(e\frac{q}{s}\right)^s \left(e\frac{p}{s}\right)^s \left(\frac{s}{p}\right)^{d_{\min} s}$$
$$\leq \sum_{s=1}^{r} \left(\frac{e^2 q r^{d_{\min}-2}}{p^{d_{\min}-1}}\right)^s,$$

where the bound $\binom{n}{k} \leq \left(e\frac{n}{k}\right)^k$ is used in the second inequality. For $r = cp$, the above inequality reduces to

$$\begin{aligned}
\Pr(\mathcal{E}^c) &\leq \sum_{s=1}^{r} \left(e^2 c^{d_{\min}-2} \frac{q}{p}\right)^s \\
&\leq \sum_{s=1}^{r} \left(e^2 c' c^{d_{\min}-1} p^{n-1}\right)^s \\
&\leq \sum_{s=1}^{r} \left(e^2 c' c^{\beta \log p} p^{n-1}\right)^s \\
&= \sum_{s=1}^{r} \left(e^2 c' p^{n-1-\beta \log 1/c}\right)^s \\
&\leq \frac{e^2 c'}{p^{\beta'} - e^2 c'} = O(p^{-\beta'}), \quad \text{for } \beta' = \beta \log \frac{1}{c} - (n-1) > 0,
\end{aligned}$$

where the size condition assumed in the theorem is used in the second inequality with $c' := \frac{1}{c}\left(\frac{c}{n}\right)^n$, and the degree condition is exploited in the third inequality. The last inequality is concluded from the geometric series sum formula for large enough p.
Then, Lemma 9 can be applied concluding that $\operatorname{krank}(A) \geq r = cp$, with probability at least $1 - \gamma_2 p^{-\beta'}$ for constants $\beta' = \beta \log \frac{1}{c} - (n-1) > 0$ and $\gamma_2 > 0$ as

$$\gamma_2 = \frac{c^{n-1} e^2}{n^n (1 - \delta_2)},$$

where δ_2 is a constant satisfying $c' e^2 p^{-\beta'} < \delta_2 < 1$. □

Proof of Remark 10: Consider a random bipartite graph $G(Y, X; E)$ where for each node $i \in X$:

1. Neighbors $N(i) \subseteq X$ are picked uniformly at random among all size d subsets of X.
2. Matching $M(i) \subseteq N(i)$ is picked uniformly at random among all size n subsets of $N(i)$.

Note that as long as $n \leq d$, the distribution of $M(i)$ is uniform over all size n subsets of X.
Fix some pair $i, i' \in Y$. Then

$$\Pr(M(i) = M(i')) = \binom{|X|}{n}^{-1}.$$

By the union bound,

$$\Pr\Big(\exists i, i' \in Y, i \neq i' \,\text{s.t.}\, M(i) = M(i')\Big) \leq \binom{|Y|}{2} \binom{|X|}{n}^{-1},$$

which is $\Theta(|Y|^2/|X|^n)$ when n is constant. Therefore, if $d \geq n$ and the size constraint $|Y| = O(|X|^s)$ for some $s < \frac{n}{2}$ is satisfied, then **whp**, there is no pair of nodes in set Y with the same random n-gram matching. This concludes that the random bipartite graph has a perfect n-gram matching **whp**, under these size and degree conditions.

□

B.2 (Auxiliary) lemmata

Lemma 7 (Existence of perfect matching for random bipartite graphs)**.** *Consider a random bipartite graph $G(W, Z; E)$ with $|W| = w$ nodes on the left side and $|Z| = z$ on the right side, and each node $i \in W$ is randomly connected to d_i different nodes in set Z. Let $d_w := \min_{i \in W} d_i$. Assume that it satisfies the size condition $w \leq cz$ for some constant $0 < c < 1$ and the degree condition $d_w \geq 1 + \beta \log z$ for some constant $\beta > 0$. Then, there exists a perfect matching in the random bipartite graph $G(W, Z; E)$ with probability at least $1 - O(z^{-\beta \log 1/c})$ where $\beta \log \frac{1}{c} > 0$.*

Proof: From Hall's theorem (Theorem 6), the existence of perfect matching for a bipartite graph is equivalent to occurrence of the following event

$$\widetilde{\mathcal{E}} := \text{“}\forall S \subseteq W : |N(S)| \geq |S|\text{”}.$$

Similar to the analysis in the proof of Theorem 4, it is concluded from union bound

$$\begin{aligned}
\Pr\big(\widetilde{\mathcal{E}}^c\big) = \Pr(\exists S \subseteq W \,\text{s.t.}\, |N(S)| < |S|) &\leq \sum_{s=1}^{w} \binom{w}{s}\binom{z}{s}\left(\frac{s}{z}\right)^{d_w s} \\
&\leq \sum_{s=1}^{w} \left(e\frac{w}{s}\right)^{s}\left(e\frac{z}{s}\right)^{s}\left(\frac{s}{z}\right)^{d_w s} \\
&\leq \sum_{s=1}^{w} \left(\frac{e^2 w^{d_w - 1}}{z^{d_w - 1}}\right)^{s} \\
&\leq \sum_{s=1}^{w} \Big(e^2 c^{d_w - 1}\Big)^{s},
\end{aligned}$$

where the bound $\binom{n}{k} \leq \big(e\frac{n}{k}\big)^k$ is used in the second inequality. From the assumed lower bound on the degree d_w and the fact that $0 < c < 1$, we have

$$\Pr\big(\widetilde{\mathcal{E}}^c\big) \leq \sum_{s=1}^{w} \Big(e^2 c^{\beta \log z}\Big)^{s} = \sum_{s=1}^{w} \Big(e^2 z^{\beta \log c}\Big)^{s} \leq \frac{e^2}{z^{\beta \log \frac{1}{c}} - e^2} \leq \frac{e^2}{1 - \delta_1} z^{-\beta \log 1/c},$$

where the second inequality is concluded from the geometric series sum formula for large enough z, and δ_1 is a constant satisfying $e^2 z^{-\beta \log 1/c} < \delta_1 < 1$. □

Lemma 8 (Degree concentration bound)**.** *Consider a random bipartite graph $G(Y, X; E)$ with $|Y| = q$ and $|X| = p$, where each node $i \in Y$ is randomly connected to d_i different nodes in set X. Let $Y' \subset Y$ be any subset[18] of nodes in Y with size $|Y'| = q'$ and $X' \subset X$ be a random (uniformly chosen) subset of nodes in X with size $|X'| = p'$. Create the new bipartite graph $G(Y', X'; E')$*

[18] Note that Y' need not to be uniformly chosen and the result is valid for any subset of nodes $Y' \subset Y$.

where edge set $E' \subset E$ is the subset of edges in E incident to Y' and X'. Denote the degree of each node $i \in Y'$ within this new bipartite graph by d'_i. Let $d_{\min} := \min_{i\in Y} d_i$ and $d'_{\min} := \min_{i\in Y'} d'_i$. Then, if $d_{\min} > r\frac{p}{p'}$ for a non-negative integer r, we have

$$\Pr[d'_{\min} \geq r+1] \geq 1 - q' \exp\left(-2(p'/p)^2 \frac{(d_{\min} - (p/p')r)^2}{d_{\min}}\right).$$

Proof: For any $i \in Y'$, we have

$$\Pr[d'_i \leq r] = \sum_{j=0}^{r} \binom{p'}{j}\binom{p-p'}{d_i - j} \Big/ \binom{p}{d_i},$$

where the inner term of summation is a hypergeometric distribution with parameters p (population size), p' (number of success states in the population), d_i (number of draws) and j is the hypergeometric random variable denoting number of successes. The following tail bound for the hypergeometric distribution is provided [45, 46]

$$\Pr[d'_i \leq r] \leq \exp(-2t_i^2 d_i),$$

for $t_i > 0$ given by $r = \left(\frac{p'}{p} - t_i\right)d_i$. Note that assumption $d_{\min} > \frac{p}{p'}r$ in the lemma is equivalent to having $t_i > 0, i \in Y$. Considering the minimum degree, for any $i \in Y'$, we have

$$\Pr[d'_i \leq r] \leq \exp(-2t^2 d_{\min}),$$

for $t > 0$ given by $r = \left(\frac{p'}{p} - t\right)d_{\min}$. Substituting t from this equation gives the following bound

$$\Pr[d'_i \leq r] \leq \exp\left(-2(p'/p)^2 \frac{(d_{\min} - (p/p')r)^2}{d_{\min}}\right). \tag{40}$$

Finally, applying the union bound, we can prove the result as follows

$$\begin{aligned}
\Pr[d'_{\min} \geq r+1] =& \Pr[\cap_{i=1}^{q'}\{d'_i \geq r+1\}] \\
\geq& 1 - \sum_{i=1}^{q'} \Pr[d'_i \leq r] \\
\geq& 1 - \sum_{i=1}^{q'} \exp\left(-2(p'/p)^2 \frac{(d_{\min} - (p/p')r)^2}{d_{\min}}\right) \\
=& 1 - q' \exp\left(-2(p'/p)^2 \frac{(d_{\min} - (p/p')r)^2}{d_{\min}}\right),
\end{aligned}$$

where the union bound is applied in the first inequality and the second inequality is concluded from (40). □

A lower bound on the Kruskal rank of matrix A based on a sufficient relaxed expansion property on A is provided in the following lemma.

Lemma 9. *If A is generic and the bipartite graph $G(Y, X; A)$ satisfies the relaxed*[19] *expansion property $|N(S)| \geq |S|$ for any subset $S \subseteq Y$ with $|S| \leq r$, then* $\text{krank}(A) \geq r$, *almost surely.*

Before proposing the proof, we state the marriage or Hall's theorem which gives an equivalent condition for having a perfect matching in a bipartite graph.

Theorem 6 (Hall's theorem, [47])**.** *A bipartite graph $G(Y, X; E)$ has Y-saturating matching if and only if for every subset $S \subseteq Y$, the size of the neighbors of S is at least as large as S, i.e., $|N(S)| \geq |S|$.*

Proof of Lemma 9: Denote the submatrix $A_{N(S),S}$ by $\widetilde{A}_S$, i.e., $\widetilde{A}_S := A_{N(S),S}$. Exploiting marriage or Hall's theorem, it is concluded that the bipartite graph $G(S, N(S); \widetilde{A}_S)$ has a perfect matching M_S for any subset $S \subseteq Y$ such that $|S| \leq r$. Denote by $\widetilde{A}_{M_S}$ the corresponding matrix to this perfect matching edge set M_S, i.e., $\widetilde{A}_{M_S}$ keeps the non-zero entries of $\widetilde{A}_S$ on edge set M_S and everywhere else, it is zero. Note that the support of $\widetilde{A}_{M_S}$ is within the support of $\widetilde{A}_S$. According to the definition of perfect matching, the matrix $\widetilde{A}_{M_S}$ is full column rank. From Lemma 6, it is concluded that $\widetilde{A}_S$ is also full column rank almost surely. This is true for any $\widetilde{A}_S$ with $S \subseteq Y$ and $|S| \leq r$, which directly results that $\text{krank}(A) \geq r$, almost surely. □

Finally, Theorem 2 is proved by exploiting the random results on the existence of perfect n-gram matching and Kruskal rank, provided in Theorems 3 and 4.

Proof of Theorem 2: We claim that if random conditions 4 and 5 are satisfied, then deterministic conditions 2 and 3 hold **whp**. Then Theorem 1 can be applied and the proof is done.
From size and degree conditions, Theorem 3 can be applied, which implies that the perfect n-gram matching condition 2 is satisfied with probability at least $1 - \gamma_1 p^{-\beta'}$ for $\beta' = \beta \log \frac{1}{c} - (n-1) > 0$. The conditions required for Theorem 4 also hold and by applying this theorem we have the bound $\text{krank}(A) \geq cp$, with probability at least$1-\gamma_2 p^{-\beta'}$. Combining this inequality with the upper bound on degree d in condition 5, we conclude that krank condition 3 is also satisfied **whp**. Hence, all the conditions required for Theorem 1 are satisfied with probability at least $1 - \gamma p^{-\beta'}$, where

$$\gamma = \gamma_1 + \gamma_2 = e^{n-1}\Big(\frac{c}{n^{n-1}} + \frac{e^2}{1-\delta_1} n^{\beta'+1}\Big) + \frac{c^{n-1}e^2}{n^n(1-\delta_2)},$$

and this completes the proof. □

C Relationship to CP Decomposition Uniqueness Results

In this section, we provide a more detailed comparison with some uniqueness results of overcomplete CP decomposition. Here, the following CP decomposition for the third order tensor $T \in \mathbb{R}^{p \times s \times q}$ is considered,

$$T = \sum_{i=1}^{r} a_i \circ b_i \circ c_i, \tag{41}$$

[19] There is no $d_{\max}$ term in contrast to the expansion property proposed in condition 7.

where $A = [a_1|\dots|a_r] \in \mathbb{R}^{p\times r}, B = [b_1|\dots|b_r] \in \mathbb{R}^{s\times r}$ and $C = [c_1|\dots|c_r] \in \mathbb{R}^{q\times r}$.
The most important and general uniqueness result of CP, called Kruskal's condition, is provided in [15], where it is guaranteed that the above CP decomposition is unique if

$$\operatorname{krank}(A) + \operatorname{krank}(B) + \operatorname{krank}(C) \geq 2r + 2.$$

Since then, several works have analyzed the uniqueness of CP decomposition. One set of works assume that one of the components, say C, is full column rank [17, 18]. it is shown in [18], for generic (fully dense) components A, B and C, if $r \leq q$ and $r(r-1) \leq p(p-1)s(s-1)/2$, then the CP decomposition in (41) is generically unique.
Now, we demonstrate how this CP uniqueness result can be adapted to our setting. First, consider the matrix $M \in \mathbb{R}^{ps\times q}$ which is obtained by stacking the entries of T as

$$M_{(i-1)s+j,k} = T_{ijk}.$$

Then, we have

$$M = (A \odot B)C^\top. \tag{42}$$

On the other hand, for the 2-persistent topic model with 4 words ($n = 2, m = 2$), the moment can be written as

$$M_4^{(2)}(x) = (A \odot A)\mathbb{E}\big[hh^\top\big](A \odot A)^\top,$$

for $A \in \mathbb{R}^{p\times q}$. The following matrix has the same column span of $M_4^{(2)}(x)$,

$$M' = (A \odot A)C'^\top,$$

for some full rank matrix $C' \in \mathbb{R}^{q\times q}$. Our random identifiability result in Theorem 2 provides the uniqueness of A and C', given M', under the size condition $q \leq \left(c\frac{p}{2}\right)^2$ and the additional degree condition 5. Note that as discussed in the previous section, this identifiability argument is the same as the unique decomposition of the corresponding tensor.
Thus, in equation (42), by setting $A = B$ and a full rank square matrix C, we obtain the 2-persistent topic model, under consideration in this paper. Thus, the identifiability results of [18] are applicable to our setting, if we assume generic (i.e. fully dense) matrix A. However, we incorporate a sparse matrix A, and therefore, require different techniques to provide identifiability results. We note that the size bound specified in [18] is comparable to the size bound derived in this paper (for random structured matrices), but we have additional degree considerations for identifiability. Analyzing the regime where the uniqueness conditions of [18] are satisfied under sparsity constraints is an interesting question for future investigation.

References

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *arXiv preprint arXiv:1206.5538*, 2012.

[2] Michael S. Lewicki, Terrence J. Sejnowski, and Howard Hughes. Learning overcomplete representations. *Neural Computation*, 12:337–365, 1998.

[3] André Uschmajew. Local convergence of the alternating least squares algorithm for canonical tensor approximation. *SIAM Journal on Matrix Analysis and Applications*, 33(2):639–652, 2012.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

[6] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor Methods for Learning Latent Variable Models. *Under Review. J. of Machine Learning. Available at arXiv:1210.7559*, Oct. 2012.

[7] A. Anandkumar, D. Hsu, A. Javanmard, and S. M. Kakade. Learning Linear Bayesian Networks with Latent Variables. *ArXiv e-prints*, September 2012.

[8] Sanjeev Arora, Rong Ge, Yoni Halpern, David M. Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. *ArXiv 1212.4777*, 2012.

[9] J.B. Kruskal. More factors than subjects, tests and treatments: an indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika*, 41(3):281–293, 1976.

[10] Tamara Kolda and Brett Bader. Tensor decompositions and applications. *SIREV*, 51(3):455–500, 2009.

[11] Silvia Gandy, Benjamin Recht, and Isao Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, 2011.

[12] Adam Coates, Honglak Lee, and Andrew Y. Ng. An analysis of single-layer networks in unsupervised feature learning. *Journal of Machine Learning Research - Proceedings Track*, 15:215–223, 2011.

[13] Quoc V. Le, Alexandre Karpenko, Jiquan Ngiam, and Andrew Y. Ng. ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning. In *NIPS*, pages 1017–1025, 2011.

[14] Li Deng and Dong Yu. *Deep Learning for Signal and Information Processing*. NOW Publishers, 2013.

[15] J.B. Kruskal. Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138, 1977.

[16] A. Bhaskara, M. Charikar, and A. Vijayaraghavan. Uniqueness of Tensor Decompositions with Applications to Polynomial Identifiability. *ArXiv 1304.8087*, April 2013.

[17] Tao Jiang and Nicholas D Sidiropoulos. Kruskal's permutation lemma and the identification of candecomp/parafac and bilinear models with constant modulus constraints. *Signal Processing, IEEE Transactions on*, 52(9):2625–2636, 2004.

[18] Lieven De Lathauwer. A Link between the Canonical Decomposition in Multilinear Algebra and Simultaneous Matrix Diagonalization. *SIAM J. Matrix Analysis and Applications*, 28(3):642–666, 2006.

[19] Alwin Stegeman, Jos M.F. Ten Berge, and Lieven De Lathauwer. Sufficient conditions for uniqueness in candecomp/parafac and indscal with random component matrices. *Psychometrika*, 71(2):219–229, June 2006.

[20] L. De Lathauwer, J. Castaing, and J.-F Cardoso. Fourth-order cumulant-based blind identification of underdetermined mixtures. *IEEE Tran. on Signal Processing*, 55:2965–2973, June 2007.

[21] Luca Chiantini and Giorgio Ottaviani. On generic identifiability of 3-tensors of small rank. *SIAM Journal on Matrix Analysis and Applications*, 33(3):1018–1037, 2012.

[22] Cristiano Bocci, Luca Chiantini, and Giorgio Ottaviani. Refined methods for the identifiability of tensors. *arXiv preprint arXiv:1303.6915*, 2013.

[23] Luca Chiantini, Massimiliano Mella, and Giorgio Ottaviani. One example of general unidentifiable tensors. *arXiv preprint arXiv:1303.6914*, 2013.

[24] E.S. Allman, C. Matias, and J.A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009.

[25] Elizabeth S. Allman, John A. Rhodes, and Amelia Taylor. A semialgebraic description of the general markov model on phylogenetic trees. *Arxiv preprint arXiv:1212.1200*, Dec. 2012.

[26] Navin Goyal, Santosh Vempala, and Ying Xiao. Fourier pca. *ArXiv 1306.5825*, 2013.

[27] Joseph M Landsberg. *Tensors: Geometry and applications*, volume 128. American Mathematical Soc., 2012.

[28] A. Anandkumar, D. Hsu, and S.M. Kakade. A Method of Moments for Mixture Models and Hidden Markov Models. In *Proc. of Conf. on Learning Theory*, June 2012.

[29] A. Anandkumar, D. P. Foster, D. Hsu, S. M. Kakade, and Y. K. Liu. A Spectral Algorithm for Latent Dirichlet Allocation. In *Proc. of Neural Information Processing (NIPS)*, Dec. 2012.

[30] A. Anandkumar, R. Ge, D. Hsu, and S. M. Kakade. A Tensor Spectral Approach to Learning Mixed Membership Community Models. In *Conference on Learning Theory (COLT)*, June 2013.

[31] E. Mossel and S. Roch. Learning nonsingular phylogenies and hidden markov models. *The Annals of Applied Probability*, 16(2):583–614, 2006.

[32] J.T. Chang. Full reconstruction of markov models on evolutionary trees: identifiability and consistency. *Mathematical Biosciences*, 137(1):51–73, 1996.

[33] Yuval Rabani, Leonard Schulman, and Chaitanya Swamy. Learning mixtures of arbitrary distributions over large discrete domains. *arXiv preprint arXiv:1212.1527*, 2012.

[34] Saneev Arora, Rong Ge, and Ankur Moitra. Learning topic models—going beyond svd. In *Symposium on Theory of Computing*, 2012.

[35] Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Proc. of Conf. on Learning Theory*, 2012.

[36] Kenneth Kreutz-Delgado, Joseph F. Murray, Bhaskar D. Rao, Kjersti Engan, Te-Won Lee, and Terrence J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural Computation*, 15:349–396, February 2003.

[37] B. Rao and K. Kreutz-Delgado. An affine scaling methodology for best basis selection. *IEEE Tran. Signal Processing*, 47:187–200, January 1999.

[38] Nishant A. Mehta and Alexander G. Gray. Sparsity-based generalization bounds for predictive sparse coding. In *Proc. of the Intl. Conf. on Machine Learning (ICML)*, Atlanta, USA, June 2013.

[39] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. Sparse coding for multitask and transfer learning. *ArxXiv preprint*, abs/1209.0738, 2012.

[40] Christopher J Hillar and Friedrich T Sommer. Ramsey theory reveals the conditions when sparse coding on subsampled data is unique. *arXiv preprint arXiv:1106.3616*, 2011.

[41] G.H. Golub and C.F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, Maryland, 2012.

[42] XuanLong Nguyen. Posterior contraction of the population polytope in finite admixture models. *arXiv preprint arXiv:1206.0068*, 2012.

[43] Daniel A. Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. *ArxXiv preprint*, abs/1206.5882, 2012.

[44] Nicholas D. Sidiropoulos and Rasmus Bro. On the uniqueness of multilinear decomposition of N-way arrays. *Journal of Chemometrics*, 14(3):229–239, 2000.

[45] V. Chvátal. The tail of the hypergeometric distribution. *Discrete Mathematics*, 25(3):285–287, 1979.

[46] Matthew Skala. Hypergeometric tail inequalities: ending the insanity. `http://ansuz.sooke.bc.ca/professional/hypergeometric.pdf`.

[47] Philip Hall. On representatives of subsets. *J. London Math. Soc.*, 10(1):26–30, 1935.