REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
	Technical Report	-

4. TITLE AND SUBTITLE

When are Overcomplete Representations Identifiable? Uniqueness of Tensor Decompositions Under Expansion Constraints

5a. CONTRACT NUMBER	W911NF-12-1-0404
5b. GRANT NUMBER	
5c. PROGRAM ELEMENT NUMBER	611102
5d. PROJECT NUMBER	
5e. TASK NUMBER	
5f. WORK UNIT NUMBER	

6. AUTHORS

Animashree Anandkumar, Daniel Hsu, Majid Janzamin, Sham Kakade

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES

University of California - Irvine
5171 California Avenue, Suite 150

Irvine, CA 92697 -7600

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
62594-NS-II.2

12. DISTRIBUTION AVAILIBILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. 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. While general overcomplete admixtures are not identifiable, we establish {\em generic} identifiability under a constraint, referred to as topic persistence. Our sufficient conditions for identifiability involve a novel set of expansion conditions on the

15. SUBJECT TERMS

Overcomplete representation, admixture models, generic identifiability, tensor decomposition.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Animashree Anandkumar
UU	UU	UU	UU		19b. TELEPHONE NUMBER: 949-824-9072

Report Title

When are Overcomplete Representations Identifiable? Uniqueness of Tensor Decompositions Under Expansion Constraints

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. While general overcomplete admixtures are not identifiable, we establish {\em generic} identifiability under a constraint, referred to as topic persistence. Our sufficient conditions for identifiability involve a novel set of expansion conditions on the population structure (i.e. the topic-word matrix) of the persistent topic model. Specifically, we require the existence of a perfect matching from hidden variables to higher order observed variables, and we can thus, incorporate overcomplete models. We establish that random structured topic models
are identifiable w.h.p. in the overcomplete regime. Our analysis allows for general (non-degenerate) distributions for modeling the topic proportions. Our proof techniques incorporate a novel class of tensor decompositions which falls in between the well-known candecomp/parafac (CP) and the more general Tucker decomposition.

When are Overcomplete Representations Identifiable? Uniqueness of Tensor Decompositions Under Expansion Constraints

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

June 16, 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. While general overcomplete admixtures 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 expansion conditions on the *population structure* (i.e. the topic-word matrix) of the persistent topic model. Specifically, we require the existence of a perfect matching from hidden variables to higher order observed variables, and can thus, incorporate overcomplete models. In particular, we establish that random models are identifiable w.h.p. in the overcomplete regime. Moreover, our analysis allows for general (non-degenerate) distributions for modeling the topic proportions. Our proof techniques incorporate a novel class of tensor decompositions which fall in between the well-known candecomp/parafac (CP) and Tucker decompositions, and provide novel conditions for unique tensor decomposition.

Keywords: Overcomplete representation, admixture 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 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 often posits latent or hidden variables that can provide a good explanation to the observed data. Overcomplete probabilistic models can have

*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

a latent space dimensionality, which can be far exceed 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 of the actors in a social networks (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, which 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 documents consisting of a tuple of observed variables (i.e. words). In this paper, we characterize conditions under which the topic models are identified through their observed moments in the overcomplete regime. To this end, we introduce an additional constraint on the model, referred to as *topic persistence*. Intuitively, this captures the "locality" effect among the observed words, and goes beyond the usual "bag-of-words" or *exchangeable* topic models. Such local dependencies among observations abound in applications such as text, images and speech, and can lead to more faithful representation. In addition, we establish that the presence of topic persistence is central to obtaining model identifiability in the overcomplete regime, and we provide an in-depth analysis of this phenomenon in this paper.

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., a certain number of words in each document). We introduce a novel constraint, referred to as *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 (observed variables). Finally, we prove that random models satisfy these criteria, and are thus identifiable in the overcomplete regime.

We first introduce the n-persistent topic model, where the parameter n determines the so-called persistence level of a common topic in a sequence of n successive words, as seen in Figure 1. 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 to be identifiable in the overcomplete regime. On the other hand, n-persistent overcomplete topic models with $n \geq 2$ are *generically* identifiable, and we provide a set of transparent

[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.

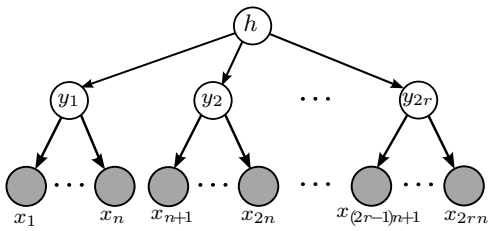


Figure 1: Hierarchical structure of the n-persistent topic model. $2rn$ number of words (views) are shown for some integer $r \geq 1$. A single topic $y_j, j \in [2r]$, is chosen for each n successive views $\{x_{(j-1)n+1}, \ldots, x_{(j-1)n+n}\}$.

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, and thus it is impossible to have expansion from topics to words. Instead, we impose a novel expansion constraint from topics to "higher order" words, which allows us to handle overcomplete models. We establish that this condition translates to the presence of a novel combinatorial object, referred to as *perfect n-gram matching*, on the bipartite graph from topics to words, which encodes the sparsity pattern of the topic-word matrix. Intuitively, this condition implies "diversity" of the word support for different topics which leads to identifiability. In addition, we present trade-offs between the topic and word space dimensionality, topic persistence level, the order of the observed moments at hand, the maximum degree of any topic in the bipartite graph, and the *Kruskal rank* [6] of the topic-word matrix, for identifiability to hold. We also show that ℓ_1-based optimization can efficiently recover the model under some additional conditions.

We then explicitly characterize the regime of identifiability for the random case, where each topic is randomly supported on a set of d words, i.e. the bipartite graph is a random d-regular graph. For this d-random model with q topics, p-dimensional word vocabulary, and topic persistence level n, when $q = O(p^n)$ and $\Theta(\log p) \leq d \leq \Theta(p^{1/n})$, the topic-word matrix is identifiable from $2n^{\text{th}}$ order observed moments with high probability. Furthermore, we establish that the size condition $q = O(p^n)$ is tight for identifiability. Thus, we prove that random-structured topic models are identifiable in the overcomplete regime.

To the best of our knowledge, this is the first work to provide expansion-based conditions for characterizing identifiability of overcomplete admixture models. We prove these results by characterizing the tensor algebra underlying the observed moments of the topic model. We establish that model identifiability for persistent topic models reduces to establishing uniqueness for a new class of tensor decompositions. For the special case of the bag-of-words model (with persistence level 1), this tensor decomposition reduces to the *Tucker* decomposition [7], while for the single topic model (with infinite persistence), it reduces to the *candecomp/parafac* (CP) decomposition. Thus, our in-depth analysis provides novel identifiability results for overcomplete tensor decomposition under expansion conditions.

1.2 Related works

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. [1,8–10], 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 (through the analysis of the resulting overcomplete CP tensor decomposition) has been considered in [11–14]. However, their results are not directly applicable to ad-mixture models since they result in tensor decompositions which are more general than the CP decomposition used in these works. Moreover, we explicitly characterize the effect of the sparsity pattern of the mixing matrix (i.e., the topic-word matrix) on model identifiability, while the above works assume fully dense (generic) mixing matrices.

There are a number of works which analyze conditions for generic identifiability of a variety of overcomplete latent variable models such as the phylogenetic tree models [15, 16]. These results provide conditions for strict identifiability of the model, and here, the dimensionality of the latent space has to be of the same order as the observed space dimensionality. In contrast, we use the weaker notion of *generic* identifiability and can therefore, allow for the latent space dimensionality to scale polynomially in the observed space dimensionality. The above works primarily utilize the Kruskal's result on uniqueness of tensor CP decompositions [6,17] to derive the identifiability results. Recently, an extension of these identifiability results to the robust setting has been considered in [18]. A number of recent works also analyze generic identifiability of overcomplete tensors [19–21] by utilizing tools from algebraic geometry. For a general overview of the algebraic geometry behind tensor decompositions, see [7].

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 prevents the latent space dimensionality from exceeding the dimensionality of the observed space.

The works of Anandkumar et. al. [22–24] 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 or the popular latent Dirichlet allocation (LDA). However, these works cannot handle general admixture models, where the distribution of the topic proportions is not limited to these classes. In addition, the approach can handle a variety of latent variable models such as Gaussian mixtures, hidden Markov models (HMM) and community models [25]. The use of simultaneous diagonalization based approaches for learning HMM's has been considered in a number of earlier works, e.g. [26, 27].

Our work is closely related to the work of Anandkumar et. al. [28] which considers identifiability and learning of topic models under expansion conditions on the topic-word matrix. The work of Spielman et. al [29] considers a similar model in the context of dictionary learning, but in addition

assumes that the coefficient matrix is random. However,these works [28, 29] can handle only the under-determined setting, where the number of topics is less than the dimensionality of the word vocabulary. We extend these results to the overcomplete setting by proposing novel higher order expansion conditions on the topic-word matrix.

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, 30, 31] However, the heuristics employed in these works have no performance guarantees. The work of Spielman et. al [29] considers learning undercomplete dictionaries and provide guaranteed learning under the assumption that the coefficient matrix is random (distributed as Bernoulli-Gaussian variables). Recent works [32, 33] 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 question of recovering the underlying latent representation.

2 Model

Notation: The set $\{1, 2, \ldots, n\}$ is denoted by $[n] := \{1, 2, \ldots, n\}$. Given set $X = \{1, \ldots, p\}$, set $X^{(n)}$ denotes all ordered n-tuples generated from X. The cardinality of set S is denoted by $|S|$. For any vector u (or matrix U), the support denoted by $\mathrm{Supp}(u)$ corresponds to the location of its non-zero entries and the ℓ_0 norm denoted by $\|u\|_0$ 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 u on its main diagonal. The column space of a matrix A is denoted by $\mathrm{Col}(A)$. We refer to matrix $R \in \mathbb{R}^{m\times r}$ as a square root of matrix $M \in \mathbb{R}^{m\times m}$ if $RR^T = M$.
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 [34]

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

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 which imposes an additional constraint, known as topic persistence on the popular admixture model [4, 5, 35]. The n-persistent topic model reduces to the bag-of-words admixture model in the case of $n = 1$.

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 *topic-word matrix* [35]. For instance, a_i represents the conditional distribution of words given topic i. The latent variable h is a q dimensional random vector $h := [h_1, \ldots, h_q]^T$ 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 distinct q topics.

The structure of 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 with size n which is the persistence level, but are not globally exchangeable.

We now describe a linear representation for the n-persistent topic model on lines of [24], 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 different possibilities $\{1, \ldots, q\}$, i.e., $y_j \in [q]$ for $j \in [2r]$. In the topic modeling, a single common topic is chosen for a sequence of n words $\{x_{(j-1)n+1}, \ldots, x_{(j-1)n+n}\}, j \in [2r]$. For notational purposes, we equivalently assume that variables $y_j, j \in [2r]$, are encoded by the basis vectors $e_i,\ i \in [q]$, where e_i is the i-th basis vector in $\mathbb{R}^q$ with the i-th entry equal to 1 and all the others equal to zero. Thus, the variable $y_j, j \in [2r]$, can be interpreted as

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

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 $\Pr\big[y_j = e_i | h\big] = h_i, j \in [2r], i \in [q]$. Note that for each sequence of n observed variables, the same hidden variable y_j is assumed in the n-persistent topic model, i.e., the topic is persistent for n different views.

Finally, at the bottom layer, each observed variable x_l for $l \in [2rn]$, is a discrete-valued p-dimensional random variable (word) where p is the size of 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 called *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) [2]

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

For the n-persistent topic model with $2rn$ number of observations (words) $x_l, l \in [2rn]$, the corresponding moment is denoted by $M_{2rn}^{(n)}(x)$.
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 $A = [a_1|a_2|\cdots|a_q] \in \mathbb{R}^{p \times q}$ is identifiable? This is answered in Section 3.
Remark 1. *Note that, we can alternatively introduce the linear generative model $x_l = Ay_j$ (more precisely, $x_{(j-1)n+k} = Ay_j, j \in [2r], k \in [n]$) instead of the conditional probabilistic model proposed above. In this new model, 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 should be still discrete. It is crucial to notice that the derivation of moments, mentioned later in Section 4.1, still holds for this new model since each vector $y_j, j \in [2r]$, takes the basis vectors as its values. Hence, the proposed identifiability results in Section 3 also hold.*

3 Sufficient Conditions for Generic Identifiability

In this section, the identifiability result for the n-persistent topic model with access to $(2rn)$-th order moment of words is provided. For a n-persistent topic model, it suffices to only have the $(2n)$-th order moment ($r = 1$ case) of words in order to be able to uniquely recover population structure A under proposed sufficient conditions.

The identifiability conditions and results under deterministic and random cases are provided in this section. First, sufficient deterministic conditions on the population structure A are provided which lead to identifiability result in Theorem 1. Next, according to the deterministic analysis, the identifiability result for a random model is provided in Theorem 2 under reasonable size and degree conditions on the bipartite graph which encodes sparsity pattern of A.

We 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 of the columns for all valid A. Instead, we consider a more relaxed notion of identifiability, known as generic identifiability.
Definition 1 (Generic identifiability, [16])**.** *Assume that the population structure A is* generic*, which means that the sparsity pattern of A is fixed and then the nonzero entries are drawn from a distribution (over those entries) that is absolutely continuous with respect to Lebesgue measure* [3]*. The generic population structure (parameters) A is* generically identifiable *if all the non-identifiable parameters 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

[2] Vector x is the vector generated by concatenating all vectors $x_l, l \in [2rn]$.

[3] As an equivalent definition, if the non-zero entries of an arbitrary matrix are randomly independently perturbed (continuous perturbation) to generate matrix A, then A is called generic.

as

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

The following natural non-degeneracy condition is assumed.

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.

Furthermore, note that we can only hope to identify the population structure A up to scaling. This is because the columns of A can be scaled by some arbitrary amount and, the hidden variables can be also scaled appropriately such that the observed variables does not change. Therefore, we can identify A up to some 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 deterministic sparsity pattern on the population structure A and establish generic identifiability, i.e., when non-zero entries are generically identifiable. Before providing the main result, a generalized notion of (perfect) matching for bipartite graphs is defined and its properties are proposed. We need these notions to establish identifiability.

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. Furthermore, for any subset $S \subseteq Y$, the set of neighbors of vertices in S with respect to the edge matrix A is defined as $N_A(S) := \{i \in X : A_{ij} \neq 0 \text{ for some } j \in S\}$. Equivalently, it can be also defined by the corresponding edge set E as $N_E(S) := \{i \in X : (j, i) \in E \text{ for some } j \in S\}$ with respect to the edge set E.

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

Definition 3 (n-gram Matching)**.** *A n-*gram matching *M for a bipartite graph $G(Y, X; E)$ is a subset of edges $M \subseteq E$ for which 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). More concretely, let $N_M(j)$ denote the set of neighbors of vertex $j \in Y$ according to the edge subset $M \subseteq E$. Then, the following conditions should be satisfied in order to call M as a n-gram matching. 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 $\min\{|N_M(j_1)|, |N_M(j_2)|\} > |N_M(j_1) \cap N_M(j_2)|$.*

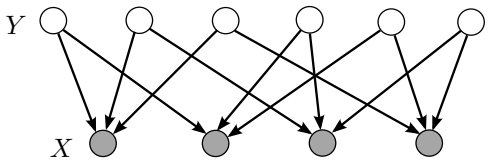


Figure 2: A bipartite graph $G(Y, X; E)$ with $|X| = 4$ and $|Y| = 6$ where the edge set E itself is a perfect 2-gram matching.

The perfect n-gram matching is also defined as follows.

Definition 4 (Perfect n-gram Matching)**.** *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 for which each vertex in Y is the end-point of exactly n edges in M.*

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.

Remark 2. *For special case $n = 1$, the (perfect) n-gram matching reduces to the regular (perfect) matching for bipartite graphs.*

In the following remark, which is proved in Appendix A.3, a deterministic necessary bound is provided on the size of a bipartite graph which has a perfect n-gram matching.

Remark 3. *For a bipartite graph $G(Y, X; E)$ with $|Y| = q$ and $|X| = p$ which has a perfect n-gram matching, we have necessarily $q \leq \binom{p}{n}$.*

Finally, note that the existence of perfect n-gram matching does not necessarily result in the existence of perfect $(n-1)$-gram matching or any other lower order matchings, e.g., the bipartite graph $G(Y, X; E)$ with $|X| = 4$ and $|Y| = \binom{4}{2} = 6$ constructed as explained in the proof of above remark and sketched in Figure 2, have a perfect 2-gram matching, but obviously it does not have perfect (1-gram) matching (since $6 > 4$). But the reverse statement is true. If the degree of each node (on matching side Y) is at least n, then, the existence of perfect $(n-1)$-gram matching results in the existence of perfect n-gram matching, which is easily seen from the definition.

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. The following identifiability conditions impose some combinatorial structure on A.

Condition 2 (Perfect n-gram matching)**.** *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 for 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 dimension of hidden variables can be more than the dimension of observed variables and still have perfect n-gram matching in the deterministic

case. It is seen later in Section 3.2 that this bound (in the order $q = \Theta(p^n)$) can be also achieved in the random case. Note that this overcomplete regime is not identifiable for $n = 1$ which is also further discussed in Remark 4.

Definition 5 (Kruskal rank, [17]). *The* Kruskal rank[4] *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.*

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

In the overcomplete regime, it is not possible to have matrix A full column rank and krank is necessarily less than $|V_h| = q$. 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. The above krank condition, imposes that krank is large enough compared to the degree. Later, it is seen that the above krank condition can be also satisfied with some sufficient random combinatorial conditions.

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

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

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 of $n = 1$. This is the usual bag-of-words admixture model. Identifiability of this model has been studied earlier [36] and we recall it below.

Remark 4 (Bag-of-words admixture model). *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 [36]*

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

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 without imposing additional conditions. Furthermore, the result for the bag-of-words admixture model requires full column rank of A for identifiability which is more stringent than our krank condition 3.

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

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 different nodes in set V_o.

Condition 4 (Size condition)**.** *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 $q \leq \left(c\frac{p}{n}\right)^n$ 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 satisfying deterministic condition 2). It is shown in Section 5.2 that the necessary size constraint $q = O(p^n)$ proposed for the deterministic case in Remark 3, is achieved in the random case. Thus, similar to the deterministic case, the above constraint allows for the overcomplete regime where $q \gg p$ for $n \geq 2$.

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 satisfies the following lower and upper bounds:*

- Lower bound: $d \geq \max\{\alpha \log p, 4 + \beta \log p\}$ *for some constants* $\alpha > n^2/2, \beta > n - 1$.
- Upper bound: $d \leq (\frac{1}{e}p)^{\frac{1}{n}}$.

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 $d_{\max}(A)^n \leq \text{krank}(A)$.

It is important to see that, for $n \geq 2$, the above condition on degree d 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 6 (**whp**)**.** *A sequence of events $\mathcal{E}_p$ occurs* with high probability *(**whp**) if* $\lim_{p\to\infty} \Pr(\mathcal{E}_p) = 1$.

The main random identifiability result for the model described in Section 2 is stated in the following theorem for $n \geq 2$, while $n = 1$ case is addressed in Remark 5. The identifiability result relies on having the $(2rn)$-th order moment of observed variables $x_l, l \in [2rn]$, defined in equation (2) as

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

for some integer $r \geq 1$.

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**, 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.

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

Remark 5 (Bag-of-words admixture model)**.** *The identifiability result for the random bag-of-words admixture model is comparable with the result in [37] which is about exact recovery of sparsely-used dictionaries. They assume that $Y = AX$ is given for some unknown arbitrary dictionary A and unknown random sparse coefficient matrix X. They establish that if the random sparse coefficient matrix X 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 identifibility result for $n \geq 2$ requires more stringent upper bound on the degree, while more relaxed condition on the size which allows to identifiability in the overcomplete regime.*

3.3 Algorithm

According to the proof of identifiability result provided in Appendix A.1, columns of the n-gram matrix $A^{(n\text{-gram})}$, defined in Definition 7, are the sparsest and rank-1 (in the tensor form) vectors in $\mathrm{Col}\Big(M_{2n}^{(n)}(x)\Big)$. This identifiability result can be used to recover the columns of A by an exhaustive search which is not efficient. More efficient algorithm provided in [36, 37] for the special case of $n = 1$, can be used to recover population structure A with appropriate slight modifications. The proposed algorithm is a convex optimization program which requires some sufficient conditions to succeed in recovering A. In our setting, the proposed sufficient conditions for exact recovery needs to be imposed on $A^{(n\text{-gram})}$ instead of A. The main condition imposes that each column of $A^{(n\text{-gram})}$ contains at least one entry that has the maximum absolute value in its row. Then, it is shown that under some additional sufficient conditions, the algorithm succeeds. See [36, 37] for details.

4 Relationship to Tensor Decomposition

In this section, we first 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. Then, we discuss the special cases of this model, viz., the single topic model (infinite-persistent topic model) and the bag-of-words admixture model (1-persistent topic model). Then, we obtain tensor forms for the moments of the topic model and discuss the relationship to the popular CP and Tucker tensor decompositions.

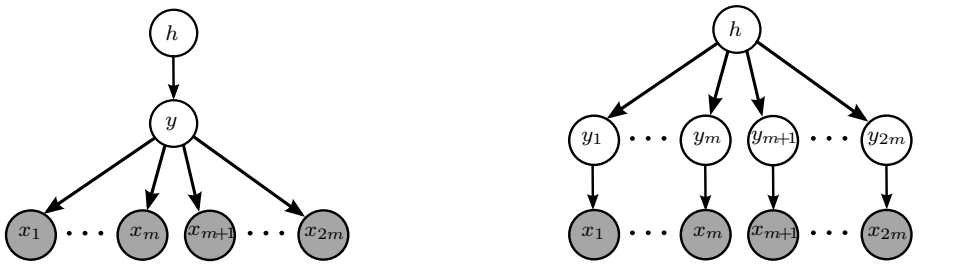
4.1 Moment characterization of the persistent topic model

The n-gram matrix is defined as follows.

Definition 7 (n-gram Matrix)**.** *For any matrix $A \in \mathbb{R}^{p \times q}$, n-gram matrix $A^{(n\text{-gram})} \in \mathbb{R}^{p^n \times q}$ is defined as the matrix whose $((i_1, \ldots, i_n), j)$-th entry is given by*

$$A^{(n\text{-gram})}((i_1, \ldots, i_n), j) := A_{i_1,j} A_{i_2,j} \cdots A_{i_n,j}$$

for all $(i_1, \ldots, i_n) \in [p]^n$ and $j \in [q]$.



(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).

That is, $A^{(n\text{-gram})}$ is the column-wise n-th order Kronecker product (or the Khatri-Rao product) [34] of n copies of A.

We now characterize the observed moments of a persistent topic model. Throughout this section, the number of observed variables 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:*

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

$$M_{2m}^{(n)}(x) = \Big(\overbrace{A^{(n\text{-gram})} \otimes \cdots \otimes A^{(n\text{-gram})}}^{r \text{ times}}\Big) M_{2r}(h) \Big(\overbrace{A^{(n\text{-gram})} \otimes \cdots \otimes A^{(n\text{-gram})}}^{r \text{ times}}\Big)^T, \quad (5)$$

  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*

$$M_{2m}^{(n)}(x) = \Big(\overbrace{A \odot \cdots \odot A}^{m \text{ times}}\Big) M_1(h) \Big(\overbrace{A \odot \cdots \odot A}^{m \text{ times}}\Big)^T \quad (6)$$

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

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

In this section, the proposed n-persistent topic model in Section 2.1 is compared with the single topic model ($n \to \infty$) in Figure 3a and the bag-of-words admixture model ($n = 1$) in Figure 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 moment of single topic model where $n \to \infty$ is characterized by equation (6). As expected, this moment form is more "structured" than the moment of n-persistent topic model in equation (5). Note that the involved moment of hidden variables in the

single topic model, is diagonal. Moreover, the observed moment of the single topic model only involves Khatri-Rao products of the population structure A, while the observed moment of the n-persistent topic model also involves Kronecker products of the n-gram matrix $A^{(n\text{-gram})}$. Therefore, the n-persistent topic model is more general than the single topic model, and is still identifiable in the overcomplete regime, which is important.

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

$$M_{2m}^{(1)}(x) = \Big(\overbrace{A \otimes \cdots \otimes A}^{m \text{ times}}\Big) M_{2m}(h) \Big(\overbrace{A \otimes \cdots \otimes A}^{m \text{ times}}\Big)^T, \tag{7}$$

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).
Why persistence helps in identifiability of overcomplete models? Comparing equations (7) and (5), it is seen that, the n-persistent topic model has a more succinct representation of the $(2m)$-th order moment of the observed variables which is crucial for providing identifiability in the overcomplete regime. More number of Kronecker products are involved in the bag-of-words admixture model in contrast to the n-persistent topic model.

We now give a simple example to illustrate how persistence helps in providing identifiability in the overcomplete regime. Consider the instances $r = 1, n = 2$, a 2-persistent topic model and $r = 2, n = 1$, a bag-of-words admixture model. From equations (5) and (7), the moments of these instances are respectively characterized as

$$M_4^{(2)}(x) = (A \odot A)\mathbb{E}\big[hh^T\big](A \odot A)^T,$$
$$M_4^{(1)}(x) = (A \otimes A)\mathbb{E}\big[(h \otimes h)(h \otimes h)^T\big](A \otimes A)^T.$$

In the 2-persistent model, by the Khatri-Rao product, the number of columns of the resulting matrix $A \odot A \in \mathbb{R}^{p^2 \times q}$ is the same as the number of columns of the original matrix A, while the number of rows is increased. The columns of $A \odot A$ are indexed by the first order hidden variables and the rows are indexed by the second order observed variables. Therefore, the Khatri-Rao product expands the effect of hidden variables to higher order observed variables. In general, it is done by retaining the order of involved hidden variables (retaining the number of columns in the resulting matrix $A^{(n\text{-gram})}$) while increasing the order of involved observed variables (increasing the number of rows in the resulting matrix $A^{(n\text{-gram})}$). This kind of expansion on the higher order observed variables in the persistent models is the key which helps to identify the model in the overcomplete regime. In other words, the original overcomplete representation becomes determined by expanding the effect of (retained order) hidden variables to higher order observed variables.
On the other hand, in the bag-of-words admixture model, this interesting expansion property does not happen where the Kronecker product $A \otimes A \in \mathbb{R}^{p^2 \times q^2}$ is incorporated. Kronecker product increases both the order of involved hidden variables and observed variables by the same amount. Therefore, for the regular admixture model (with persistence level 1), it is not possible to identify its population structure A in the overcomplete regime.

The above discussion can be also generalized to the general case of moments in equations (5) and (7).

4.2 Tensor representation of the model

In this section, we derive the tensor algebra of the moments derived in Section 4.1 for the persistent topic model. We compare the tensor form with the well-known Tucker and CP decompositions.

Tensor algebra preliminaries

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.
The tensor $A \in \bigotimes^{n} \mathbb{R}^{p}$, is stacked in 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 [34]

$$A = a_1 \circ a_2 \circ \cdots \circ a_n \in \bigotimes_{i=1}^{n} \mathbb{R}^{p_i} \Leftrightarrow A(i_1, i_2, \ldots, i_n) := a_1(i_1)a_2(i_2)\cdots a_n(i_n). \tag{8}$$

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 [5].
In general, the outer product operation is a way to combine lower order tensors to construct higher order tensors, e.g., for $B \in \mathbb{R}^{p_1 \times p_2}, C \in \mathbb{R}^{p_3 \times p_4}$, the 4-th order tensor $A = B \circ C \in \bigotimes_{i=1}^{4} \mathbb{R}^{p_i}$ is defined as $A(i_1, i_2, i_3, i_4) := B(i_1, i_2)C(i_3, i_4)$.
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

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

The second power is usually called the n-the order *tensor power* of vector a.
Finally, the CP (CANDECOMP/PARAFAC) and Tucker representations and the *Kruskal form* notation are defined as follows [34].

[5] This type of rank is called CP (CANDECOMP/PARAFAC) tensor rank in the literature [34].

Definition 8 (CP representation and Kruskal form). *Given $\lambda \in \mathbb{R}^r, U_i \in \mathbb{R}^{p_i \times r}, i \in [n]$, the n-th order tensor $A \in \bigotimes_{i=1}^n \mathbb{R}^{p_i}$ is defined in the Kruskal form as*

$$A = [[\lambda; U_1, U_2, \ldots, U_n]] := \sum_{i=1}^{r} \lambda_i U_1(:, i) \circ U_2(:, i) \circ \cdots \circ U_n(:, i), \tag{9}$$

where $U_j(:, i)$ denotes the i-th column of matrix U_j. The above representation of tensor A is called the CP representation (decomposition) where the tensor A is written as a weighted sum of rank-1 tensors.

More generally, the Tucker representation is defined as follows.
Definition 9 (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 n-th order tensor $A \in \bigotimes_{i=1}^n \mathbb{R}^{p_i}$ is*

$$A = \sum_{i_1=1}^{r_1} \sum_{i_2=1}^{r_2} \cdots \sum_{i_n=1}^{r_n} S(i_1, i_2, \ldots, i_n) U_1(:, i_1) \circ U_2(:, i_2) \circ \cdots \circ U_n(:, i_n), \tag{10}$$

where $U_j(:, i_j)$ denotes the i_j-th column of matrix U_j. With a slight abuse of notation, the above Tucker representation can be also denoted in the form $A = [[S; U_1, U_2, \ldots, U_n]]$.

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

Tensor representation of moments under topic model

The $(2m)$-th order moment of the words $x_l, l \in [2m]$, is defined as (in the tensor form)

$$T_{2m}(x)_{(i_1, i_2, \ldots, i_{2m})} := \mathbb{E}[x_1(i_1) x_2(i_2) \cdots x_{2m}(i_{2m})], \quad i_1, i_2, \ldots, i_{2m} \in [p], \tag{11}$$

where $T_{2m}(x) \in \bigotimes^{2m} \mathbb{R}^p$. For the n-persistent topic model with $2m$ number of observations (words) $x_l, l \in [2m]$, the corresponding moment is denoted by $T_{2m}^{(n)}(x)$, which is the tensor form of moment $M_{2m}^{(n)}(x)$ characterized in Lemma 1. This tensor is characterized in the following lemma, 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 (11), for the n-persistent topic model is characterized as:*

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

$$T_{2m}^{(n)}(x) = \sum_{i_1=1}^{q} \sum_{i_2=1}^{q} \cdots \sum_{i_{2r}=1}^{q} \mathbb{E}[h_{i_1} h_{i_2} \cdots h_{i_{2r}}] a_{i_1}^{\circ n} \circ a_{i_2}^{\circ n} \circ \cdots \circ a_{i_{2r}}^{\circ n}. \tag{12}$$

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

$$T_{2m}^{(n)}(x) = \sum_{i \in [q]} \mathbb{E}[h_i] a_i^{\circ 2m}. \tag{13}$$

The tensor representation (12) is a specific type of tensor decomposition which is a special case of the Tucker representation, but more general than the CP representation.

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

The tensor representation of our model provided in equation (12) is a special case of the Tucker representation but more general than the symmetric CP representation. In order to have a fair comparison, the number of observed variables is fixed to $2m$ and the persistence level is varied.

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 (13) as

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

where the Kruskal notation, defined in Definition 8, is used in the last equality. This representation is exactly the symmetric CP representation (decomposition) of $T_{2m}^{(\infty)}(x)$ where $\lambda_i = \mathbb{E}[h_i], i \in [q]$, and $U_i = A, i \in [2m]$.

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

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

where the Kruskal notation defined in Definition 9, is used in the last equality. This representation is exactly Tucker representation (decomposition) of $T_{2m}^{(1)}(x)$ where the core tensor $S = \mathbb{E}\big[h^{\circ(2m)}\big]$ is the tensor form of the $(2m)$-th order hidden moment $M_{2m}(h)$, defined in equation (3). Furthermore, the inverse factors $U_i = A, i \in [2m]$, correspond to the population structure A.

On lines of discussion in Section 4.1, above general Tucker decomposition is not identifiable in the overcomplete regime, while the proposed tensor decomposition in equation (12) is identifiable under the sufficient conditions provided in Section 3.

5 Proof Techniques and Auxiliary Results

The main identifiability results are provided for both deterministic and random cases of topic-word graph structure, in Sections 3.1 and 3.2 respectively. In this section, we first provide the proof sketch of these results and then, we propose two auxiliary results on the existence of perfect n-gram matching for random bipartite graphs and lower bound on the Kruskal rank of random matrices.

5.1 Proof sketch

The deterministic analysis is primarily based on conditions on the n-gram matrix $A^{(n\text{-gram})}$; but since these conditions (mainly expansion condition on $A^{(n\text{-gram})}$, provided in condition 7) are opaque, this analysis is postponed to Appendix A.1, where the identifiability result is stated in Theorem 6. In the following, first, we provide a summary of the hierarchical relationships among all of these identifiability results and the corresponding conditions. Then, according to this hierarchy, a proof sketch of each result is stated.

Summary of relationships among different conditions: To summarize, there exists a hierarchy among the proposed conditions as follows. 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 4 and 5. 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 6. Note that, the genericity of A is also required for these results to hold.

Primary deterministic analysis in Theorem 6: The deterministic analysis in Theorem 6, is described here for the case when $2n$ number of words are available under the n-persistent topic model. From equation (5), the $(2n)$-th order moment of the observed variables under the n-persistent topic model can be written as

$$M_{2n}^{(n)}(x) = \Big(A^{(n\text{-gram})}\Big)\mathbb{E}\big[hh^T\big]\Big(A^{(n\text{-gram})}\Big)^T. \tag{14}$$

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. Without such non-degeneracy assumptions, there is no hope for identifiability. Assuming these two conditions, we have from (14) that

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

Therefore, the problem of recovering A from $M_{2n}^{(n)}(x)$ reduces to finding $A^{(n\text{-gram})}$ in $\mathrm{Col}\big(A^{(n\text{-gram})}\big)$. Then, it is shown that under the following expansion condition on $A^{(n\text{-gram})}$ and the genericity property, matrix A is identifiable from $\mathrm{Col}\big(A^{(n\text{-gram})}\big)$. The expansion condition (refer to condition 7 for a more detailed statement), imposes the following property on the bipartite graph $G\big(V_h, V_o^{(n)}; A^{(n\text{-gram})}\big)$ [6],

$$\Big|N_{A^{(n\text{-gram})}_{\text{Rest.}}}(S)\Big| \geq |S| + d_{\max}\Big(A^{(n\text{-gram})}\Big), \quad \forall S \subseteq V_h,\ |S| > \mathrm{krank}(A), \tag{15}$$

where $d_{\max}\big(A^{(n\text{-gram})}\big)$ is the maximum node degree in set V_h, and the restricted version of n-gram matrix, denoted by $A^{(n\text{-gram})}_{\text{Rest.}}$, is defined in Definition 10. The identifiability claim is proved by showing that the columns of $A^{(n\text{-gram})}$ are the sparsest and rank-1 (in the tensor form) vectors in $\mathrm{Col}\big(A^{(n\text{-gram})}\big)$ under the sufficient expansion in (15) and genericity conditions. This finishes the proof sketch for the deterministic identifiability result based on $A^{(n\text{-gram})}$, proposed in Theorem

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

6. Note that the expansion condition (15) is a more relaxed condition compared to expansion condition proposed in [36, 37] for identifiability in the undercomplete regime. For a more detailed comparison, refer to Remark 8 in Appendix A.1.

Deterministic analysis in Theorem 1: Expansion and rank conditions in Theorem 6 are imposed on the n-gram matrix $A^{(n\text{-gram})}$. 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 $A^{(n\text{-gram})}$ are satisfied. This is shown in Lemma 5 using the observation in the following lemma.
In the following lemma which is proved in Appendix A.3, we state an interesting property which relates the existence of a perfect matching in $A^{(n\text{-gram})}$ 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 $G(Y, X^{(n)}; A^{(n\text{-gram})})$ has a perfect matching. In the other direction, if $G(Y, X^{(n)}; A^{(n\text{-gram})})$ has a perfect matching $M^{(n\text{-gram})}$, then $G(Y, X; A)$ has a perfect n-gram matching under the following condition on $M^{(n\text{-gram})}$. All the matching edges $(j, (i_1, \dots, i_n)) \in M^{(n\text{-gram})}$ 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.*

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

Random analysis in Theorem 2: Finally, the identifiability result for a random matrix A is provided in Theorem 2 in Section 3.2. 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, it is proved in Theorem 4 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 5 in Section 5.3 which helps to satisfy 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 Existence of perfect n-gram matching for random bipartite graphs

The result of this section is used in the proof of Theorem 2, but since the result is interesting and useful by itself, we also propose it independently. In this section, it is shown 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**.

In the proof of the necessary size condition for the existence of perfect n-gram matching proposed in Remark 3, we provide an analysis which is also constructive, i.e., we provide a deterministic greedy method to construct a bipartite graph which has a perfect n-gram matching when satisfying $q \leq \binom{p}{n}$. Now, the question is under what conditions a random bipartite graph has a perfect n-gram matching. In this section, this question is answered in Theorem 4, where it is seen that size bound

$q = O(p^n)$ is also sufficient for the existence of perfect n-gram matching in a random bipartite graph.

Before proposing our result on the existence of perfect n-gram matching in random bipartite graphs, the existing results on the existence of perfect matching in random bipartite graphs are reviewed [38–40]. Here, we recap the result of [39] ,which is used to prove the existence of perfect n-gram matching in random bipartite graphs. Let z_1 and c^* satisfy [39]

$$\begin{aligned} z_1 &= \frac{e^{z_1} - 1}{d - 1}, \\ c^* &= \frac{z_1}{d(1 - e^{-z_1})^{d-1}}, \end{aligned} \tag{16}$$

where each node $i \in Y$ in the random bipartite graph $G(Y, X; E)$, is randomly connected to d different nodes in set X.

Theorem 3 (Existence of perfect matching for random bipartite graphs, [39])**.** *Consider a random bipartite graph $G(Y, X; E)$ with node size ratio $c := \frac{|Y|}{|X|}$ and $d \geq 3$. If $c \leq c^*$, then **whp**, there exists a perfect matching in the random bipartite graph $G(Y, X; E)$.*

Theorem 4 (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. 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 (degree of nodes in Y) $d \geq \alpha \log p$ for some $\alpha > n^2/2$ (lower bound in condition 5). Then, **whp**, there exists a perfect (Y-saturating) n-gram matching in the bipartite graph $G(Y, X; E)$.*

Remark 6 (Necessity of the proposed size bound)**.** *It is crucial to see that the size bound $q = O(p^n)$ in the proposed random result for the existence of perfect n-gram matching achieves the necessary size bound $q \leq \binom{p}{n} = O(p^n)$, proposed in Remark 3.*

Remark 7 (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 4, the latter has a better size scaling while the former 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 to at least have $(4n)$-th order moment according to the union bound arguments, while it is only required to know the $(2n)$-th order moment, according to our more involved arguments.*

5.3 Lower bound on the Kruskal rank of random matrices

The result of this section is used in the proof of Theorem 2. 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 5 (Lower bound on the Kruskal rank of random matrices)**.** *Consider a random matrix $A \in \mathbb{R}^{p \times q}$ for which there exist d (which is called degree) number of random non-zero entries in each column. Assume that it satisfies size condition $q \leq \left(c\frac{p}{n}\right)^n$ (condition 4) and degree condition*

$d \geq 4 + \beta \log p$ *for some* $\beta > n-1$ *(lower bound in condition 5) and in addition* A *is generic. Then,* ***whp****,* $\mathrm{krank}(A) \geq \frac{1}{e}p$.

Acknowledgements

The authors acknowledge useful discussions with Sina Jafarpour, Moses Charikar, and Kamalika Chaudhuri. A. Anandkumar is supported in part by NSF Career award CCF-1254106, NSF Award CCF-1219234, AFOSR Award FA9550- 10-1-0310, and ARO Award W911NF-12-1-0404. M. Janzamin is supported by NSF Award CCF-1219234 and ARO Award W911NF-12-1-0404.

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, $A^{(n\text{-gram})}$, and then, we show that the conditions of Theorem 1 are sufficient for this alternative result.

A.1 Deterministic analysis based on $A^{(n\text{-gram})}$

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

In the n-gram matrix, $A^{(n\text{-gram})} \in \mathbb{R}^{p^n \times q}$, redundant rows exist. Precisely, if some row of $A^{(n\text{-gram})}$ is indexed by n-tuple $(i_1, \ldots, i_n), i_l \in [p]$, then another row indexed by any permutation of tuple $(i_1, \ldots, i_n)$ has exactly the same entries. In other words, since multiplication is commutative, the number of distinct rows of $A^{(n\text{-gram})}$ is at most the number of (potentially) different products $A_{i_1,j} A_{i_2,j} \cdots A_{i_n,j}$ in a column $j \in [q]$. Therefore, the number of distinct rows of $A^{(n\text{-gram})}$ 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 10 (Restricted n-gram matrix)**.** *For any matrix* $A \in \mathbb{R}^{p \times q}$*, restricted* n*-gram matrix* $A^{(n\text{-gram})}_{\text{Rest.}} \in \mathbb{R}^{s \times q}$*,* $s = \binom{p+n-1}{n}$*, is defined as the restricted version of* n*-gram matrix* $A^{(n\text{-gram})} \in \mathbb{R}^{p^n \times q}$*, where the redundant rows of* $A^{(n\text{-gram})}$ *are removed, as explained above.*

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

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

$$\left| N_{A^{(n\text{-gram})}_{\text{Rest.}}}(S) \right| > |S| + d_{\max}\left(A^{(n\text{-gram})}\right), \quad \forall S \subseteq V_h,\ |S| > \mathrm{krank}(A), \tag{17}$$

where $d_{\max}\Big(A^{(n\text{-gram})}\Big)$ *is the maximum node degree in set* V_h.

Remark 8. *The expansion condition for the bag-of-words admixture model is provided in* (4)*, introduced in [36]. The proposed expansion condition in* (17) *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,* $A^{(n\text{-gram})}$*, as stated in* (14)*. Second, the expansion property* (4)*, proposed in [36], needs to be satisfied for all subsets* S *with size* $|S| \geq 2$*, which is a much stricter condition than the one proposed here in* (17)*, since we can have* $\text{krank}(A) \gg 2$*. Note that because of the* $d_{\max}$ *term in the expansion property in* (17)*, it is hard to satisfy* (17) *for small sets.*

The deterministic identifiability result for the model described in Section 2, based on the conditions on $A^{(n\text{-gram})}$, is stated in the following theorem for $n \geq 2$, while $n = 1$ case is addressed in Remarks 4 and 8. This is actually the basic result from which the main deterministic and random identifiability results respectively proposed in Theorems 1 and 2, are concluded. The identifiability result relies on having the $(2n)$-th order moment of observed variables $x_l, l \in [2n]$, defined in equation (2) as

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

Theorem 6 (Generic identifiability under deterministic conditions on $A^{(n\text{-gram})}$)**.** *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 $B := A^{(n\text{-gram})} \in \mathbb{R}^{p^n \times q}$. Then, the moment characterized in equation (14) can be written as $M_{2n}^{(n)}(x) = B\mathbb{E}\left[hh^T\right]B^T$. Since both matrices $\mathbb{E}\left[hh^T\right]$ and B have full column rank (from conditions 1 and 6), the rank of $B\mathbb{E}\left[hh^T\right]B^T$ is q where $q = O(p^n)$, and furthermore $\text{Col}(B\mathbb{E}\left[hh^T\right]B^T) = \text{Col}(B)$. Let $\mathcal{U} := \{u_1, \ldots, u_q\} \in \mathbb{R}^{p^n}$ be any basis of $\text{Col}(B\mathbb{E}\left[hh^T\right]B^T)$ 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., $U_i := \text{ten}(u_i), i \in [q]$, 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 $\text{Col}(B\mathbb{E}\left[hh^T\right]B^T)$) are rank-1 in the n-th order tensor form (since $\text{ten}(b_i) = a_i^{\circ n}$) 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

$$\max_i \text{Rank}(\text{ten}(u_i)) = 1 \quad \text{and} \quad \max_i \|u_i\|_0 \leq d_{\max}(B). \tag{18}$$

The above bounds are concluded from the fact that $b_i \in \text{Col}(B\mathbb{E}\left[hh^T\right]B^T),\ i \in [q]$, 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

$$\text{ten}(u_i) = \sum_{j=1}^{q} v_i(j)\,\text{ten}(b_j) = \sum_{j=1}^{q} v_i(j)\,\text{ten}(a_j^{\otimes n}) = \sum_{j=1}^{q} v_i(j) a_j^{\circ n} = [[v_i; \overbrace{A, \ldots, A}^{n \text{ times}}]], \tag{19}$$

where the last equality is based on the Kruskal form 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 (19) reduces to

$$\text{ten}(u_i) = [[\widetilde{v}_i; \overbrace{\widetilde{A}_i, \ldots, \widetilde{A}_i}^{n \text{ times}}]] = \sum_{j=1}^{r} \widetilde{v}_i(j)[(\widetilde{a}_i)_j]^{\circ n}, \tag{20}$$

which is derived by removing columns of A corresponding to the zero entries in v_i.
Next, we rule out the case that $\|v_i\|_0 \geq 2$ under two cases ($2 \leq \|v_i\|_0 \leq \text{krank}(A)$ and $\text{krank}(A) < \|v_i\|_0 \leq q$), for $u_i = Bv_i$ equality 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 $\text{krank}(A)$ and therefore it is full column rank. From Fact 1, rank-1 tensors $[(\widetilde{a}_i)_j]^{\circ n}, j \in [r]$, are linearly independent. Hence, for any $n \geq 2$, [7] from equation (20), 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 (18).

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 10. Let $u_i' = B_{\text{Rest.}}v_i$. Since u_i' is the restricted version of u_i, we have

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

where the second inequality is from the genericity of A used in 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 (18).

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

The following lemma is useful in the proof of Theorem 6. The result proposed in this lemma is similar to the parameter genericity condition in [36], but generalized for the n-gram matrix, $A^{(n\text{-gram})}$. The lemma can be also proved on lines of the proof of Remark 2.2 in [36].

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

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

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

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

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

Definition 12. *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$) = 1$.*

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 $\widetilde{C} = [\tilde{c}_1|\tilde{c}_2|\cdots|\tilde{c}_s]^T$, where $\tilde{c}_i^T$ is the i-th row of $\widetilde{C}$. Also, let $C := [c_1|c_2|\cdots|c_s]^T$ and $W := [w_1|w_2|\cdots|w_s]^T$ be the corresponding $|S| \times |S|$ submatrices of $M^{(2\text{-gram})}_{\text{Rest.}}$ and U, respectively. For each $i \in [s]$, denote by $\mathcal{N}_i$ the null space of the matrix $\widetilde{C}_i = [\tilde{c}_1|\tilde{c}_2|\cdots|\tilde{c}_i]^T$. 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 $M^{(2\text{-gram})}_{R,S}$, every row c_{i+1}^T of C contains at least one non-zero entry. Therefore,

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

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 $w_{i+1}(j)$,

$$\Pr\left[v \in \mathcal{N}_{i+1}\middle|\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_i\right] = \Pr\left[\sum_{j\in[s]:c_{i+1}(j)\neq 0} v(j)(c_{i+1}(j) + w_{i+1}(j)) = 0\middle|\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_i\right] = 0. \quad (21)$$

Consequently,

$$\Pr\left[\dim(\mathcal{N}_{i+1}) < \dim(\mathcal{N}_i)\middle|\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_i\right] = 1 \tag{22}$$

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

$$\|Hu\|_0 > |R| - |S|. \tag{23}$$

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

Proof of Lemma 1: First, 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

$$U^{\otimes r} := \overbrace{U \otimes U \otimes \cdots \otimes U}^{r \text{ times}} \in \mathbb{R}^{p^r \times q^r}. \tag{24}$$

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

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

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

$$\mathbb{E}[x_1 \otimes x_2 \otimes \cdots \otimes x_m | y_1, y_2, \ldots, y_r] = \left(A^{(n\text{-gram})}\right)^{\otimes r} (y_1 \otimes y_2 \otimes \cdots \otimes y_r), \tag{25}$$

where the $(\cdot)^{\otimes r}$ notation is defined in equation (24). Now for the $(2m)$-th order moment, we have

$$\begin{aligned}
M_{2m}^{(n)}(x) &:= \mathbb{E}\Big[(x_1 \otimes x_2 \otimes \cdots \otimes x_m)(x_{m+1} \otimes x_{m+2} \otimes \cdots \otimes x_{2m})^T\Big] \\
&= \mathbb{E}_{(y_1,y_2,\ldots,y_{2r})}\Big[\mathbb{E}\left[(x_1 \otimes \cdots \otimes x_m)(x_{m+1} \otimes \cdots \otimes x_{2m})^T | y_1, y_2, \ldots, y_{2r}\right]\Big] \\
&\overset{(a)}{=} \mathbb{E}_{(y_1,y_2,\ldots,y_{2r})}\Big[\mathbb{E}\big[(x_1 \otimes \cdots \otimes x_m)|y_1, \ldots, y_{2r}\big]\mathbb{E}\big[(x_{m+1} \otimes \cdots \otimes x_{2m})^T|y_1, \ldots, y_{2r}\big]\Big] \\
&\overset{(b)}{=} \mathbb{E}_{(y_1,y_2,\ldots,y_{2r})}\Big[\mathbb{E}\big[(x_1 \otimes \cdots \otimes x_m)|y_1, \ldots, y_r\big]\mathbb{E}\big[(x_{m+1} \otimes \cdots \otimes x_{2m})^T|y_{r+1}, \ldots, y_{2r}\big]\Big] \\
&\overset{(c)}{=} \mathbb{E}_{(y_1,y_2,\ldots,y_{2r})}\left[\left(\left[A^{(n\text{-gram})}\right]^{\otimes r}\right)(y_1 \otimes \cdots \otimes y_r)(y_{r+1} \otimes \cdots \otimes y_{2r})^T\left(\left[A^{(n\text{-gram})}\right]^{\otimes r}\right)^T\right] \\
&= \left(\left[A^{(n\text{-gram})}\right]^{\otimes r}\right)\mathbb{E}\left[(y_1 \otimes \cdots \otimes y_r)(y_{r+1} \otimes \cdots \otimes y_{2r})^T\right]\left(\left[A^{(n\text{-gram})}\right]^{\otimes r}\right)^T \\
&\overset{(d)}{=} \left(\left[A^{(n\text{-gram})}\right]^{\otimes r}\right)M_{2r}(y)\left(\left[A^{(n\text{-gram})}\right]^{\otimes r}\right)^T, \qquad (26)
\end{aligned}$$

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 (25) is used in (c) and finally, the $(2r)$-th order moment of $(y_1, \ldots, y_{2r})$ is defined as $M_{2r}(y) := \mathbb{E}\left[(y_1 \otimes \cdots \otimes y_r)(y_{r+1} \otimes \cdots \otimes y_{2r})^T\right]$ in (d).

On the other hand, for $M_{2r}(y)$, we have by the law of total expectation

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

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 (26), 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 $B := \big[A^{(n\text{-gram})}\big]^{\otimes r} \in \mathbb{R}^{p^{rn} \times q^r}$, 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 (5) can be written as

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

Let $b_{(i_1,\ldots,i_r)} \in \mathbb{R}^{p^{rn}}$ denote the corresponding column of B indexed by r-tuple $(i_1, \ldots, i_r), i_k \in$

$[q], k \in [r]$. Then, the above matrix equation can be expanded as

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

where relation $b_{(i_1,\ldots,i_r)} = a_{i_1}^{\otimes n}\otimes\cdots\otimes a_{i_r}^{\otimes n}, i_1,\ldots,i_r \in [q]$, 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

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

Then, we have the following equivalent tensor form for the original model proposed in equation (5)

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

□

A.3 Proof of generalized matching properties

Proof of Lemma 3: We show that if $G(Y, X; A)$ has a perfect n-gram matching, then $G(Y, X^{(n)}; A^{(n\text{-gram})})$ 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 $E^{(n\text{-gram})}$ denote the edge set of the bipartite graph $G(Y, X^{(n)}; A^{(n\text{-gram})})$. Assume $G(Y, X; A)$ has a perfect n-gram matching $M \subseteq E$. For any $j \in Y$, let set $N_M(j)$ 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 $A^{(n\text{-gram})}(N'_M(j), j) \neq 0$ for all $j \in Y$. Hence, $(j, N'_M(j)) \in E^{(n\text{-gram})}$ for all $j \in Y$ which together with the fact that all $N'_M(j)$'s tuples are distinct, it results that $M^{(n\text{-gram})} := \{(j, N'_M(j)) | j \in Y\} \subseteq E^{(n\text{-gram})}$ is a perfect matching for $G(Y, X^{(n)}; A^{(n\text{-gram})})$.
□

Proof of Remark 3: In order to show this, we fix the dimension of vertex set X to p and see what the maximum number of vertices in set Y could be such that the resulting bipartite graph still has

a perfect n-gram matching. Therefore, assume we have p vertices in X and an empty vertex set Y on the other side. We want to introduce vertices in set Y with degree n such that the resulting bipartite graph has a perfect n-gram matching. In order to satisfy this property, for any subset of vertices $S \subseteq X$ with $|S| = n$, we introduce a new vertex in set Y to ensure it has a perfect n-gram matching. Hence, we can introduce up to $\binom{p}{n}$ vertices in Y. □

A.4 Sufficient matching properties for satisfying rank and graph expansion conditions

In the following lemma, it is shown that under having a perfect n-gram matching and additional genericity and krank conditions, the rank and graph expansion conditions 6 and 7 on $A^{(n\text{-gram})}$, 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 $A^{(n\text{-gram})}$:*

1) *If A is generic, $A^{(n\text{-gram})}$ is full column rank (condition 6) with Lebesgue measure one (almost surely).*

2) *If krank condition 3 holds, $A^{(n\text{-gram})}$ satisfies the proposed expansion property in condition 7.*

Proof: Let M indicate the perfect n-gram matching of the bipartite graph $G(V_h, V_o; A)$. From Lemma 3, there exists a perfect matching $M^{(n\text{-gram})}$ for the bipartite graph $G(V_h, V_o^{(n)}; A^{(n\text{-gram})})$. 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 $M^{(n\text{-gram})}$. Note that $\text{Supp}(A_M) \subseteq \text{Supp}(A)$ and $\text{Supp}(B_M) \subseteq \text{Supp}(A^{(n\text{-gram})})$.

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 [8]), $A^{(n\text{-gram})}$ 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

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

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

$$|N_A(S')| \geq \text{krank}(A) \quad \text{for } S' \subseteq V_h, |S'| = \text{krank}(A). \tag{27}$$

Then, we have

$$|N_A(S)| \geq |N_A(S')| \quad \text{for } S' \subset S \subseteq V_h, |S| > \text{krank}(A), |S'| = \text{krank}(A),$$

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

$$\geq \operatorname{krank}(A)$$
$$> d_{\max}(A)^n, \tag{28}$$

where (27) is used in the second inequality and the last inequality is from krank condition 3.

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

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

where the first inequality is concluded from the existence of a perfect n-gram matching in A, and the bound (28) is used in the second inequality. Since $d_{\max}\big(A^{(n\text{-gram})}\big) = d_{\max}(A)^n$, the proof of part 2 is also completed.

□

Remark 9. *The second result of above lemma is similar to the necessity argument of (Hall's) Theorem 7 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 which is expected since the expansion condition proposed here is stricter than the one in Hall's theorem.*

A.5 (Auxiliary) lemmata and facts

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 $\operatorname{Supp}(\widetilde{C}) \subseteq \operatorname{Supp}(C)$ 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. □

Fact 1. *If vectors $w_i \in \mathbb{R}^p, i \in [r]$ are linearly independent, then the n-th order tensor powers $W_i := w_i^{\circ n} \in \bigotimes^n \mathbb{R}^p, i \in [r]$, are also linearly independent.*

Proof: For the sake of contradiction, let the tensors $W_i, i \in [r]$, be linearly dependent. Therefore, there exist coefficients $\alpha_i \in \mathbb{R}, i \in [r]$, not all zero, such that

$$\sum_{i=1}^{r} \alpha_i W_i = 0. \tag{29}$$

Without loss of generality, assume that $\alpha_1 \neq 0$. Since w_i's are linearly independent, we have $w_i \neq 0$ for all $i \in [r]$ and therefore, there exists some $k \in [p]$ such that $w_1(k) \neq 0$.
Consider the fibers $f_i := W_i(1 : p, k, k, \ldots, k), i \in [r]$, corresponding to the vectors obtained by fixing all but first indices of W_i's. From tensor outer product definition in (8), we have $f_i = \beta_i w_i$ where $\beta_i := w_i(k)^{n-1} \in \mathbb{R}$ for all $i \in [r]$. Furthermore, according to the special selection of k mentioned earlier we have $\beta_1 \neq 0$.
Restricting equality (29) to the subset indexed by $(1 : p, k, k, , \ldots, k)$, results

$$\sum_{i=1}^{r} \alpha_i f_i = \sum_{i=1}^{r} \gamma_i w_i = 0,$$

where $\gamma_i := \alpha_i \beta_i$. Since at least one of the scalar coefficients $\gamma_i, i \in [r]$, is non-zero ($\gamma_1 := \alpha_1 \beta_1 \neq 0$), it is concluded from above equality that vectors $w_i, i \in [r]$, are linearly dependent. This contradicts the assumption of lemma and completes the proof. □

Finally, Theorem 1 is proved by combining the results of Theorem 6 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 6 are satisfied almost surely and this completes the proof. □

B Proof of Random Identifiability Result (Theorem 2)

According to the proof sketch provided in Section 5.1, the steps for the proof of Theorem 2 are provided in the following subsections.

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

Proof of Theorem 4: Define $J := c\frac{p}{n}$. Divide set X randomly (uniform) into n different partitions with (almost) equal size[9] denoted by $X_l^{(2)}, l \in [n]$. Define sets $X_l^{(1)} := \cup_{i=1}^{l} X_i^{(2)}, l \in [n]$. Furthermore, divide set Y randomly (uniform) into $O(p^{n-1})$ partitions with size at most $J = c\frac{p}{n}$. Applying Lemma 8 and Theorem 3, **whp**, there exists a perfect matching from each of these partitions of Y to set $X_1^{(1)}$. Then, we combine every J number of these partitions (on Y side) creating $O(p^{n-2})$ new bipartite graphs. Therefore, Lemma 7 can be applied which results that **whp**, there exists a perfect 2-gram matching from each of these combined partitions of Y (with size less than or equal to $J^2 = \left(c\frac{p}{n}\right)^2 = O(p^2)$) to set $X_2^{(1)}$. This combining procedure is performed iteratively; in step l,

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

every J number of partitions on Y are combined to create $O(p^{n-l})$ new bipartite graphs. Applying Lemma 7, **whp**, there exists a perfect l-gram matching from corresponding partitions of Y (with size less than or equal to $J^l = \left(c\frac{p}{n}\right)^l = O(p^l)$) to set $X_l^{(1)}$. Finally at iteration n, **whp**, we have a perfect n-gram matching from Y (with size less than or equal to $J^n = \left(c\frac{p}{n}\right)^n = O(p^n)$) to set $X_n^{(1)} = X$.

Above discussion is the main part of the proof, but in order to complete the proof, it is required to argue on the total number of times that perfect matching result proposed in Theorem 3 is used in the above random discussion. In this way, we want to ensure that high probability rate proposed in Theorem 3 still holds after several times of its exploitation. Let $N^{(\text{hp})}$ denote the total number of times that perfect matching result proposed in Theorem 3 is used in the above random discussion. Similarly, let $N_l^{(\text{hp})}, l \in \{2, \ldots, n\}$, denote the total number of times that perfect matching result proposed in Theorem 3 is used in step l to ensure that there exists a perfect l-gram matching from corresponding partitions of Y to set $X_l^{(1)}$, **whp** (Note that is is done through Lemma 7 as explained in the above discussion).
As mentioned in Lemma 7, let $P_{l-1}\Big(X_{l-1}^{(1)}\Big)$ denote the set of all subsets of $X_{l-1}^{(1)}$ with cardinality $l-1$ with has the size

$$\Big|P_{l-1}\Big(X_{l-1}^{(1)}\Big)\Big| = \binom{\big|X_{l-1}^{(1)}\big|}{l-1} = \binom{\frac{l-1}{n}p}{l-1}.$$

According to the construction method of l-gram matching proposed in Lemma 7, $\Big|P_{l-1}\Big(X_{l-1}^{(1)}\Big)\Big|$ is the number of times Theorem 3 is used in order to ensure that there exists a perfect l-gram matching for each partition on 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

$$N_l^{(\text{hp})} \leq J^{n-l}\ \Big|P_{l-1}\Big(X_{l-1}^{(1)}\Big)\Big| = J^{n-l}\binom{\frac{l-1}{n}p}{l-1}. \tag{30}$$

Since in the first step, J^{n-1} number of perfect matchings needs to exist in the above discussion, we have

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

where inequality (30) 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. Therefore, $N^{(\text{hp})} = O(p^{n-1})$ and Theorem 3 is used polynomial number of times. □

Proof of Theorem 5: 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 $S \subseteq Y$. Then

$$\Pr(|N(S)| \leq |S|) \leq \sum_{\substack{T \subseteq X: \\ |T|=|S|}} \Pr(N(S) \subseteq T) \leq \binom{p}{|S|} \left[\binom{|S|}{d} \Big/ \binom{p}{d} \right]^{|S|} \leq \binom{p}{|S|} \left(\frac{|S|}{p} \right)^{d|S|},$$

where the bound $\binom{|S|}{d} \Big/ \binom{p}{d} \leq \left(\frac{|S|}{p} \right)^d$ is used in the last inequality.
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

$$\begin{aligned} \Pr(\mathcal{E}^c) = \Pr(\exists S \subseteq Y \,\text{s.t.}\, 1 \leq |S| \leq r \wedge |N(S)| < |S|) &\leq \sum_{s=1}^{r} \binom{q}{s} \binom{p}{s} \left(\frac{s}{p} \right)^{ds} \\ &\leq \sum_{s=1}^{r} \left(e\frac{q}{s} \right)^s \left(e\frac{p}{s} \right)^s \left(\frac{s}{p} \right)^{ds} \\ &\leq \sum_{s=1}^{r} \left(\frac{e^2 q r^{d-2}}{p^{d-1}} \right)^s, \end{aligned}$$

where the bound $\binom{n}{k} \leq \left(e\frac{n}{k} \right)^k$ is used in the second inequality.
For $r := \frac{1}{e}p$, the above inequality reduces to

$$\begin{aligned} \Pr(\mathcal{E}^c) &\leq \sum_{s=1}^{r} \left(\frac{q/p}{e^{d-4}} \right)^s \\ &\leq \sum_{s=1}^{r} \left(\frac{q/p}{p^{\beta}} \right)^s \\ &= \sum_{s=1}^{r} \left(\frac{q}{p^{\beta+1}} \right)^s \\ &\leq \sum_{s=1}^{r} \left(\frac{c'}{p^{\beta-n+1}} \right)^s, \end{aligned}$$

where the degree condition assumed in the theorem is used in the second inequality and the size condition is exploited in the last inequality by defining $c' := \left(\frac{c}{n} \right)^n$. Since c', β and n are constants and $\beta - n + 1 > 0$ by the theorem assumption, it is concluded that

$$\lim_{p \to \infty} \Pr(\mathcal{E}^c) = 0,$$

which results that event $\mathcal{E}$ happens **whp**. Therefore, Lemma 9 can be applied concluding that $\text{krank}(A) \geq r = \frac{1}{e}p$, **whp**. □

Proof of Remark 7: Consider a random bipartite graph $G(Y, X; E)$ where for each node $i \in X$:

1. Neighbors $N(i) \subseteq X$ is 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.

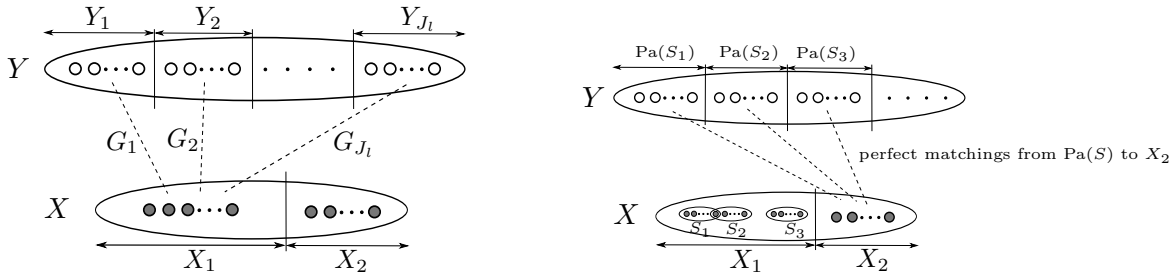
□

B.2 (Auxiliary) lemmata

Lemma 7. *Consider a bipartite graph $G(Y, X; E)$ with $|Y| = r$ and $|X| = s$ where $r = O(s^l)$ and each node $i \in Y$ is randomly connected to d_l different nodes in set X. Divide the nodes in set Y randomly (uniform) to $J_l := c\frac{s}{l}$ partitions $Y_1, \ldots, Y_{J_l}$ with (almost) equal size for some constant $c < 1$. In addition, divide the nodes in set X randomly (uniform) to two partitions X_1 and X_2 with sizes $|X_1| = \frac{l-1}{l}s$ and $|X_2| = \frac{s}{l}$. Next, create J_l different bipartite graphs $G_i(Y_i, X_1; E_i), i \in [J_l]$, by considering partitions Y_i and X_1 and the corresponding subset of edges $E_i \subset E$ incident to them. Refer to Figure 4a. Furthermore, assume that $d_l \geq \beta \log s$ for some $\beta > l^2/2$. Then, if each of the corresponding J_l bipartite graphs $G_i(Y_i, X_1; E_i), i \in [J_l]$, has a perfect $(l-1)$-gram matching, then **whp**, the original bipartite graph $G(Y, X; E)$ has a perfect l-gram matching.*

Proof: Let us denote the corresponding perfect $(l-1)$-gram matching of $G_i(Y_i, X_1; E_i)$ by M_i. Furthermore, the set of all subsets of X_1 with cardinality $l-1$ are denoted by $P_{l-1}(X_1)$, i.e., $P_{l-1}(X_1)$ includes the sets with $(l-1)$ elements in the power set [10] of X_1. For each set $S \in P_{l-1}(X_1)$, take the set of all nodes in Y which are connected to all members of S according to the union of matchings $\cup_{i=1}^{J_l} M_i$. Call this set as the parents of S denoted by $\mathrm{Pa}(S)$. According to the definition of perfect $(l-1)$-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 $|\,\mathrm{Pa}(S)| \leq J_l = c\frac{s}{l}$. 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 $S \in P_{l-1}(X_1)$. This is because of the perfect $(l-1)$-gram matchings considered for sets $Y_i, i \in [J_l]$. Now, a perfect l-gram matching for the original bipartite graph is constructed as follows. For any $S \in P_{l-1}(X_1)$, consider the set of parents $\mathrm{Pa}(S)$. Create the bipartite graph $G(\mathrm{Pa}(S), X_2; E_S)$ where $E_S \subset E$ is the subset of edges incident to partitions $\mathrm{Pa}(S) \subset Y$ and $X_2 \subset X$. Denote by

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



(a) Original partitioning of sets Y and X proposed in Lemma 7.

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

Figure 4: Auxiliary figures for proof of Lemma 7. (a) Original partitioning of sets Y and X proposed in the lemma where set Y is partitioned to $J_l := c\frac{s}{l}$ partitions $Y_1, \ldots, Y_{J_l}$ with (almost) equal size for some constant $c < 1$. In addition, set X is partitioned to two partitions X_1 and X_2 with sizes $|X_1| = \frac{l-1}{l}s$ and $|X_2| = \frac{s}{l}$. The bipartite graphs $G_i(Y_i, X_1; E_i), i \in [J_l]$, are also shown in the figure. (b) Set Y is partitioned to subsets $\mathrm{Pa}(S), S \in P_{l-1}(X_1)$, which is generated through perfect $(l-1)$-gram matchings $M_i, i \in [J_l]$. S_1, S_2 and S_3 are three different sets in $P_{l-1}(X_1)$ shown as samples. In addition, the perfect matchings from $\mathrm{Pa}(S), S \in P_{l-1}(X_1)$, to X_2 proposed in the proof are also pointed in the figure.

d_S the minimum degree of nodes in set $\mathrm{Pa}(S)$ in the bipartite graph $G(\mathrm{Pa}(S), X_2; E_S)$. Applying Lemma 8, it is concluded that [11]

$$\Pr[d_S \geq 3] \geq 1 - J_l \exp\left(-\frac{2}{l^2}\frac{(d_l - 2l)^2}{d_l}\right).$$

Furthermore, we have $|\mathrm{Pa}(S)| \leq c\frac{s}{l} = c|X_2|$. Now, we can apply Theorem 3 concluding that **whp** there exists a perfect matching from $\mathrm{Pa}(S)$ to X_2 within the bipartite graph $G(\mathrm{Pa}(S), X_2; E_S)$. Refer to Figure 4b for a schematic picture. The edges of this perfect matching are combined with the corresponding edges of existing perfect $(l-1)$-gram matchings $M_i, i \in [J_l]$, to provide l incident edges to each node $i \in \mathrm{Pa}(S)$. It is easy to see that this proposes a perfect l-gram matching from $\mathrm{Pa}(S)$ to X.

We perform the same steps for all sets $S \in P_{l-1}(X_1)$ to propose a perfect l-gram matching from any $\mathrm{Pa}(S)$ to X. Finally, according to the construction, the union of all of these matchings is a perfect l-gram matching from $\cup_{S \in P_{l-1}(X_1)} \mathrm{Pa}(S) = Y$ to X and the result is proved. □

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 different nodes in set X. Let $Y' \subset Y$ be any subset*[12] *of nodes in Y with size $|Y'| = q'$ and $X' \subset X$ be a random (uniformly chosen) subset of nodes in X with size $p' := |X'| = p/n$. Create the new bipartite graph $G(Y', X'; E')$ 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. Define $d' := \min_{i \in Y'} d'_i$. Then, if*

[11] Note that in the context of Theorem 4, this bound can be written as $\Pr[d_S \geq 3] \geq 1 - J_l \exp\left(-\frac{2}{n^2}\frac{(d-2n)^2}{d}\right)$. It is concluded from the fact that when the application of this lemma in Theorem 4 is considered, we have $s = \frac{l}{n}p$ and therefore $|X_2| = \frac{p}{n}$.

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

$d > rn$ for a non-negative integer r, we have

$$\Pr[d' \geq r+1] \geq 1 - q' \exp\left(-\frac{2}{n^2}\frac{(d-rn)^2}{d}\right).$$

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

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

where the inner term of summation is a hypergeometric distribution with parameters p (population size), p' (number of success states in the population), d(number of draws) and j is the hypergeometric random variable denoting number of successes. The following tail bound for the hypergeometric distribution is provided [41, 42]

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

for $t > 0$ given by $r = \left(\frac{p'}{p} - t\right)d$. Note that assumption $d > rn$ in the lemma is equivalent to having $t > 0$. Substituting t from this equation gives the following bound

$$\Pr[d'_i \leq r] \leq \exp\left(-\frac{2}{n^2}\frac{(d-rn)^2}{d}\right). \tag{31}$$

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

$$\begin{aligned}
\Pr[d' \geq r+1] =& \Pr[\cap_{i=1}^{q'}\{d'_i \geq r+1\}] \\
=& 1 - \Pr\left[\overline{\cap_{i=1}^{q'}\{d'_i \geq r+1\}}\right] \\
=& 1 - \Pr[\cup_{i=1}^{q'}\{d'_i \leq r\}] \\
\geq& 1 - \sum_{i=1}^{q'} \Pr[d'_i \leq r] \\
\geq& 1 - \sum_{i=1}^{q'} \exp\left(-\frac{2}{n^2}\frac{(d-rn)^2}{d}\right) \\
=& 1 - q' \exp\left(-\frac{2}{n^2}\frac{(d-rn)^2}{d}\right),
\end{aligned} \tag{32}$$

where the union bound is applied in the first inequality and the second inequality is concluded from (31). □

Note that more strict degree condition 5 proposed in Section 3.2, implies that the probability bound proposed in the above lemma goes to one with the rate proportional to inverse polynomial function of q. Therefore, the lower bound on degree in the above lemma holds with **whp**.

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*[13] *expansion property $|N(S)| \geq |S|$ for any subset $S \subseteq Y$ with $|S| \leq r$, then* $\operatorname{krank}(A) \geq r$*, almost surely.*

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

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 7 (Hall's theorem, [43])**.** *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 4 and 5.

Proof of Theorem 2: It is shown that if random conditions 4 and 5 are satisfied then deterministic conditions 2 and 3 also hold. Then Theorem 1 can be applied and the proof is done.
According to the theorem assumptions, size and degree conditions required for Theorem 4 hold and therefore by applying this theorem, the perfect n-gram matching condition 2 is satisfied **whp**. The conditions required for Theorem 5 also hold and by applying this theorem we have the bound $\text{krank}(A) \geq \frac{1}{e}p$, **whp**. Combining this inequality with the upper bound on degree d in condition 5, concludes that krank condition 3 is also satisfied **whp**. Hence, all the conditions required for Theorem 1 are satisfied **whp**, and this completes the proof. □

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] 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.

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

[8] 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.

[9] 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.

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

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

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

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

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

[28] A. Anandkumar, D. Hsu, and A. Javanmard S. M. Kakade. Learning Bayesian Networks with Latent Variables. In *Proc. of Intl. Conf. on Machine Learning*, June 2013.

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

[30] 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.

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

[32] 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.

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

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

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

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

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

[38] Nikolaos Fountoulakis and Konstantinos Panagiotou. Sharp load thresholds for cuckoo hashing. *Random Struct. Algorithms*, 41(3):306–333, 2012.

[39] Alan M. Frieze and Páll Melsted. Maximum matchings in random bipartite graphs and the space utilization of cuckoo hash tables. *Random Struct. Algorithms*, 41(3):334–364, 2012.

[40] Martin Dietzfelbinger, Andreas Goerdt, Michael Mitzenmacher, Andrea Montanari, Rasmus Pagh, and Michael Rink. Tight thresholds for cuckoo hashing via xorsat. *Arxiv preprint arXiv:0912.0287*, Dec. 2010.

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

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

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