Title
A Computational Study of Story Narratives and Dynamics in On-line Social Media

Author
Ebrahimzadeh Houlasou, Ehsan

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

A Computational Study of Story Narratives and Dynamics in On-line Social Media

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Electrical and Computer Engineering

by

Ehsan Ebrahimzadeh Houlasou

2018

ABSTRACT OF THE DISSERTATION

A Computational Study of Story Narratives and Dynamics
in On-line Social Media

by

Ehsan Ebrahimzadeh Houlasou
Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2018
Professor Vwani P. Roychowdhury, Chair

Social media has changed manners in which people access information, form their opinion and act in real life. Therefore, there is an urgent need to design information retrieval systems to turn large scale unstructured users' data into structured knowledge. Traditional text summarization techniques and co-occurrence based topic models, however, cannot capture the complex social dynamics that drive individual and group behaviors. On the other hand, well-known models of narratives and legends that have been proven to be effective in capturing story dynamics do not have a scalable machine learning formulation. The main goal of this dissertation is to develop computational and statistical tools that can efficiently and accurately extract multi-scale narrative structures from large-scale social media datasets.

In particular, a narrative is modeled as a "Story Narratives Networks" comprised of nodes that represent primary actants, which interact via a sequence of actions that define the links in the network. One of the contributions of this dissertation is to determine distinct actant groups in an unsupervised manner from contextual unstructured data. Each such group consists of actors that have the same contextual role in the narrative. In order to cluster actors, we construct low-dimensional sparse vector embeddings using dimensionality reduction techniques such as Non-Negative Matrix Factorization (NMF). We propose an exterior point method to solve the NMF problem, which constructs a solution based on a suitably rotated optimal solution of the unconstrained matrix factorization problem. We

evaluate the performance of our proposed algorithm and embedding-based clustering scheme on two datasets, namely data from a discussion forum on parenting issues and a corpus of tweets on user experience with contact-less payment methods.

Finally, towards understanding the dynamics in the evolution of stories, we study the problem of detecting changes in the temporal evolution of the user activities. We formulate this problem in a transient change point detection setting and design a statistical test to detect the change based on the number of user activities observed so far, with minimum expected delay under a controlled measure of false alarm. We evaluate the change detection method on a corpus of tweets related to Super Bowl 2015.

The dissertation of Ehsan Ebrahimzadeh Houlasou is approved.

Timothy R. Tangherlini

Arash Ali Amini

Lieven Vandenberghe

Vwani P. Roychowdhury, Committee Chair

University of California, Los Angeles

2018

To my parents...

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to sincerely thank my advisor Professor Vwani Roychowdhury for his support and mentorship. Our long discussions always lit up the road ahead. It was through his mentorship that I could find my passion for data analytics, which indeed changed my career path. He was always supportive throughout this journey. I would also like to express my gratitude to the members of my doctoral committee for taking the time to serve on this committee and providing feedback on this work. Specifically, I would like to thank Professor Timothy Tangherlini, with whom I had wonderful collaborations and whose charm and sense of humor has always been refreshing, Professor Lieven Vandenberghe, whose insightful comments improved the quality of the presentation a lot, and Professor Arash Amini, with whom I shared the joy of theoretical research and whose patience and research attitude has always been inspiring. I would also like to thank my collaborator, Professor Aslan Tchamkerten, for his continued support. That last part of this dissertation is due to a collaboration with Aslan on Summer 2014, when I was in Paris.

I wanted to thank my friends and loved ones for all the support they have given me in various ways during these years. This would have not been possible without you. Finally, I would like to thank my family for their immeasurable love and support. There is nothing I could say that would express how grateful I am for what you have done for me.

VITA

2011	B.S. (Electrical Engineering), Isfahan University of Technology University, Iran.
2011	B.S. (Electrical Engineering), Isfahan University of Technology University, Iran.
2013	M.S. (Electrical Engineering), University of Waterloo, Canada
2013–2018	PhD Student, Electrical and Computer Engineering Department, University of California, Los Angeles.

CHAPTER 1

Introduction

1.1 Overview

The growing adoption of on-line social media as the primary source of obtaining and sharing information, on the one hand, and recent advancements in large scale data analytics, on the other hand, have overhauled the way large scale sociological studies are performed. Social media users are no longer just silent observers of the "Stories" but rather, through expressing their own take of the story in social media, they add a new dimension to how stories are developed. As such, this collective social behavior has contributed to the recent advent of notions like "trending stories" or "fake news" in the vocabulary of social network analytics.

On-line social media is indeed a new lens to study, at Internet scale, how people's opinions are formed, expressed, and evolved over time. Users have different, sometimes opposing, viewpoints about the same subject, event or story and this is reflected in the subjective tone of their expressions. Moreover, in contrast to expert-generated opinion pieces or summaries that narrate the story in a standard and structured fashion, social media users use a variety of different expressions to refer to the same semantic concept/event or they may even ignore some aspects of the story and express themselves in a much less structured fashion.

In this dissertation, we seek to address the following question: Is it possible to find consensus models for a holistic semantic analysis of an on-line story based on aggregation of social media posts? In other words, can we find a unified view of all the partial information surfaced in unstructured social media posts about an on-line story in a structured form? We address these questions with an affirmative answer by developing models towards characterizing a structured information network where nodes, which represent objects with different entity

types, are linked via edges that represent relationships with different relation types, which we refer to as story narrative network. Supervised machine learning that require large amount of human annotated corpus specific data fall short in solving this problem. This defines a wide array of challenges for unsupervised or distantly supervised approaches that leverage the structures present in the problem, and richness of the data in terms of the redundancy of information. In this work, we develop information retrieval algorithms that take advantage of such structures and information redundancies by transforming unstructured data to models suitable for machine learning algorithms.

A common theme that appears in all of the problems we study in this work is a form of "sparse" structure in the problem. The following observations capture such sparse structures in different scenarios:

- Entities appear only with a small set of relation phrases
- User networks are sparsely connected
- A social media post is comprised of only a small number of narrative contexts/topics
- The changes in the statistical behavior of the user activities are rare
- There are a only small number of terms that contribute to determining the alignment of a post with respect to the narrative

Throughout the dissertation, we make some important assumptions in terms of temporal evolution of the underlying ground truth and the data. In the first part of the dissertation, when dealing with characterizing the aggregate story network or the major entities, a.k.a. actants, we assume that the ground-truth story network is static, thereby neglecting the temporal changes in the story narrative. Moreover, we are oblivious to the sequencing of the relationships in the story network and the temporal order in which comments have appeared. In the second part, however, we touch upon characterizing some aspects of the temporal development of the story. Namely, we are interested in understanding the temporal changes in the evolution of activities around a story. Another major aspect of this work,

which is particularly relevant to these sequential problems, is sample efficiency. That is, in an adaptive sequential algorithm for a specific learning task, we can basically use less data samples as we learn about the structure of the problem over time.

When possible, we show how the theoretical frameworks and the computational tools we develop for the story analysis problems branch out to a wider range of problems in information retrieval. For example, we demonstrate how factorization techniques we develop for the entity resolution problem can be used for designing personalized search engines. The running examples throughout the dissertation, however, are problems in narrative analysis for on-line stories.

1.2 Motivation and Problem Statement

In this section, we briefly motivate the problems studied in this work within the framework of story narrative analysis and discuss our solutions in general terms. We leave the formal problem statement and detailed discussions to the respective chapters.

1.2.1 Characterizing Significant Entity Groups in a Story

One of the major challenges towards characterizing the story narrative and understanding the underlying dynamics in a corpus of text is the task of entity recognition. Entity resolution is the problem of identifying, matching, and merging references corresponding to the same entity within a dataset. Traditional entity recognition rely on external knowledge bases, fully or partially, to resolve the type of candidate mentions. The standard approach is to first define a set of types that the entities should be mapped to, then map a fraction of the candidate entities to the corresponding types using an external knowledge base, and finally disambiguate the type of the other entities based on the relationships in which they co-occur with already tagged entities. However, a fundamental challenge in domain-specific text corpora where entities are not necessarily mappable to knowledge base entries, is that the type categories might not be known a priori or the learner does not have access to a set of seed entities with known type labels. Therefore, the learner has to group entities of

similar types by essentially dropping the knowledge base supervision.

We formulate this problem in a multi-view clustering framework. We first generate candidate entity and relation mentions based on semantic and syntactic features of the tokens in text. Specifically, we extract relation triplets that connect a pair of entity mentions with a relation verb. We then find low dimensional distributed representations, aka embeddings, for entity mentions by factorizing suitably constructed matrices from extracted relation triplets.

The embedding idea is proved promising in a variety of prediction tasks in text mining, including relational learning. In this work, however, we use vector embeddings of the entity mentions to cluster semantically similar entities in the same group. This requires the embedded vectors to have certain structural properties, namely sparsity. To impose sparse structure on entity embeddings, we consider non-negativity constraint on the factor matrices in the matrix factorization objective. To solve the non-negative matrix factorization(NMF) problem, we propose an exterior point method, which constructs a solution for the NMF problem based on a suitably rotated optimal solution of the unconstrained matrix factorization problem. We evaluate the performance of our proposed NMF algorithm and embedding-based clustering scheme on two datasets, namely vaccination related data from a discussion forum on parenting issues and a corpus of tweets on user experience with contact-less payment methods.

1.2.2 Detecting Statistical Changes in the Evolution of User Activities

Characterizing the underlying dynamics in the evolution of the story narratives in a social media setting is not a well-defined problem in general. To make the problem concrete, let us assume that we are only interested in detecting "major changes" in the temporal evolution of the story. These changes can reflect in the temporal evolution of the story narrative in terms of the textual content of the users' posts as well as the frequency of the activities irrespective of the textual content. Let us focus only on the latter.

Consider an on-line micro-blog setting, where users post textual pieces about an event. As our running example, let us take Twitter as the on-line platform and let us refer to each post

as a tweet. Each tweet is indexed by the time it is posted. Let us also assume that each tweet contains a number of hash-tags, which serve as textual signatures that contextualize the subject of the tweet. Therefore, the total number of tweets that contain a particular hash-tag that is related to the event can be regarded a measure of user activity on the subject. Therefore, a problem of interest in characterizing the temporal evolution of the user activities is to quickly detect statistical changes in the distribution of the number of activities on the hash-tag. This change in the distribution can reflect a major event related to the subject of the study.

A major challenge in this problem is the transient nature of this change, meaning that the change has to be identified before change period, which is assumed to be short compared to the whole observation window, is over. Thus, we can view this problem as a characterizing a trade-off between quick detection of the change versus reliability of the decision.

We formulate this problem in a transient change point detection setting where the objective is to design a statistical test to detect the change, if present, based on the sequence observed so far with minimum expected delay and a controlled measure of false alarm. Since obtaining all the tweets in every single time instance is costly, we add an additional constraint on the number of time instances that the statistic of interest is observed. We evaluate the change detection framework on a corpus of tweets related to Super Bowl 2015. In chapter 5, we discuss this problem in details.

1.2.3 Organization and Summary of Contributions

In the first part of the dissertation, we develop models towards characterizing story narrative network. In particular, in order to characterize groups of entities with contextually similar role in the story narrative network, we attribute distributed vector representations, aka embeddings, to entity mentions that appear in extracted relations from users' posts, and then by clustering the embedded vectors, we partition the entities into groups.

In chapter 2, we propose an embedding approach based on explicit factorization of suitably generated entity-relation matrices that capture the contextual role of an entity mention

as a subject and an object in a relationship. In order to obtain interpretable embedding vectors with improved clustering behavior, we impose sparse structure on the embeddings by considering a non-negativity constraint on the factor matrices in the matrix factorization formulation.

In chapter 3, we propose a new exterior point method to solve the Non-Negative Matrix Factorization(NMF) problem, based on the geometry on the optimization landscape of the unconstrained matrix factorization problem. Finally, we apply K-means clustering on the obtained embedded vectors to cluster the entities into groups.

In chapter 4, we evaluate the performance of our proposed algorithm and embedding-based clustering scheme on two datasets, namely data from a discussion forum on parenting issues and a corpus of tweets on user experience with contact-less payment methods. It is shown that our exterior point method has a significantly better sparsity properties over the considered models as well as better prediction performance over the celebrated multiplicative update rules method for solving NMF. Moreover, we show that the clusters obtained by our method can very well recover the underlying ground truth groups in the studied datasets and it is computationally verified that our NMF-based embedding approach has superior clustering performance over embeddings obtained by the optimal matrix completion approach based on SVD.

In the second part of the dissertation, we switch gears to study problems related to the temporal evolution of story narratives. In chapter 5, we study the problem of detecting changes in the temporal evolution of the user activities and formulate this problem in a transient change point detection setting and applied a statistical test to detect the change based on the number of user activities observed so far, with minimum expected delay under a controlled measure of false alarm. We evaluate the change detection method on a corpus of tweets related to Super Bowl 2015. We show that our method is able to detect the start of the game reliably and effectively within less than a minute from the start of the game.

Finally, in chapter 6 we make some concluding remarks and point out some of the future directions that we would like to follow up on.

CHAPTER 2

Entity Resolution Problem

2.1 Motivation

In recent years, social media has turned into the primary source of data for much of our insight to the society, from trending topics to behavioral patterns of small and large groups such as circles of friends or a city's residents. One particular area of interest is the analysis of the underlying stories and interactions amongst real world entities through their trace in the social media. Given the scale of such data, it is practically impossible to curate a set of cherry picked user posts to represent a holistic view of the interactions amongst the entities involved in a story. Moreover, summarizing large text corpora in the form term co-occurrence topics does not provide a refined view of the interactions in a story. Therefore, there is an indispensable need for large scale methods that can aggregate pieces of information present in different posts and provide a holistic view of the story narratives in a structured form. In this work, we address this problem by introducing a network structure that describes interactions among major entities involved in a story. Formally, a story narrative network is a structured information network where nodes, which represent objects with different entity types, are linked via edges that represent relationships with different relation types. Before we start our formal discussion let us start with defining some terminology that we will frequently be using all over this work.

- **Entity**: An entity is a real world object, such as a person, place, organization, etc. that is recognizable by a human agent. Entities can also be abstract, such as entities in a novel or word senses in the context of linguistics. In the context of story narratives, we are interested particularly in entities that take part in forming the narrative. The

entities may be known for some domains, while in others, they need to be discovered.

- **Mention**: A mention, aka reference, is a token span surfaced in the data, which refers to an entity. Usually, entity mentions need to be extracted from textual documents in the corpus. Surface words that refer to the actual entities of the story are examples of a mention(reference). They are likely to appear in certain syntactic forms, such as subject or object of the sentences in text corpora. Mapping such references to entities is a fundamental challenge that we deal with in this work, that we will formally define as the entity resolution task. A fundamental assumption that we make in this work is that mentions with the same surface word refer to the same entity. This assumption may not be true in certain text corpora where there exist different entities that might appear in text with the same surface token. For example, two entities that have the same family name that are mentioned in different documents via their family name only.

- **Relationship**: When multiple mentions(references) are observed together in a context that forms an interaction or link, specified for the task, such co-occurrence is called a relationship between those references. In the context of stories, a relationship refers to a tuple that describes how two or more entity mentions are related. Co-occurrences usually happen as a result of ties or links between the underlying entities. We sometimes use the term relationship to refer to these ties between entities as well.

- **Attribute**: An attribute is an observed property of an individual mention, for example the word tokens that co-occur with a mention in the same sentence. These attributes can be used as complimentary information about entities or relationships that can be used for the resolution task.

- **Actant**: A group, aka an actant, is a collection of entities that serve the same or similar purpose in the setting we study and have close ties between themselves. In the context of story narrative models, we specifically refer to these groups of entities that have the same contextual meaning as actants. Such groups are only observed indirectly through co-occurrences that mostly happen between references to entities that belong

to the same group. The observed co-occurrence relations in the data provide evidence for discovering the group structures among the entities, and the group evidence in turn helps in improved resolution of the references.

One of the major challenges towards characterizing the story narrative and understanding the underlying dynamics in a corpus of text is the task of entity recognition. Entity resolution is the problem of identifying, matching, and merging references corresponding to the same entity within a dataset. In the context of story narrative analysis, entity resolution is the task of mapping entity mentions that have a similar contextual role in the story to their corresponding actant group. In this chapter, we formalize this problem and take a machine learning approach based on finding distributed vector representations for entities to cluster similar entities that belong to the same group. The chapter is organized as follows. We first formally define the type of relations that we deal with in this work. We then formalize the entity resolution problem in the context story narrative analysis followed by our data model as suitably defined entity relation matrices. We then describe our entity and relation embedding approach as explicit factorization of the defined matrices and show how such learned representations can be used for clustering entity mentions with contextually similar roles in the story. Finally we provide a comprehensive overview of the prior art in entity resolution and embedding-based approaches to that.

2.2 Entity and Relation Mentions

Given a corpus of unstructured text, we generate candidate mentions based on semantic and syntactic features of the words in the text. Specifically, we extract semantic structures in the form of certain paths in dependency parse trees of all sentence tokens in the corpus. The main type of dependency paths that we use in this work, which in turn reflect action-based relationships between the entities, starts with the token that is tagged as the subject of the sentence, if present, linked to the verb token that it is connected to and finally the object token that is connected to the verb. We describe this structure in more details in Chapter 4. Some other works use Semantic Role Labeling(SRL) [1] or use segmentation methods that

adopt a probabilistic approach based on a set of tagged instances [2].

To start our formal discussion, we define the specific form of relationships that we consider in this work. An entity mention m is a token span in text which represents an entity e. An assertion or a relation instance is a tuple $(m_1, m_2, \cdots, m_S, v)$ that represents some type of relation between multiple entities $\{m_1, \cdots, m_k\}$ through some relation phrase v.

In this work, we focus only on binary relations, i.e., (m_s, m_o, v). Thus, assertions will be expressed as (ordered) triplets

$$\mathcal{T} = \{(m_s^{(i)}, m_o^{(i)}, v^{(i)})\}_{i=1}^{N},$$

of entities mentions $m_s^{(i)}$ and $m_o^{(i)}$ and the relation phrase $r^{(i)}$. Let us also define the set of potential entities $\mathcal{M}$ as

$$\mathcal{M} = \{m \mid (m, m', v) \in \mathcal{T} \text{ or } (m', m, v) \in \mathcal{T} \text{ for some entity mention } m' \text{ and some relation verb } v\},$$

and the set of all relation verbs as

$$\mathcal{V} = \{v \mid (m, m', v) \in \mathcal{T} \text{ for some entity mentions } m, m'\}.$$

2.3 Problem Statement

In this section, we formally define the entity resolution problem in narrative discovery setting. Suppose that a corpus of text along with a set of relationship triplets extracted from the corpus in the form $\mathcal{T} = \{(m_s^{(i)}, m_o^{(i)}, v^{(i)})\}_{i=1}^{N}$ are given; with entity mention set $\mathcal{M}$ and relation mentions $\mathcal{V}$ as defined above. Suppose also that the set of actual entities that the mentions refer to also belongs to the entity set $\mathcal{M}$. Then, given a portioning of the entities

$$\mathcal{E} = \cup_{i=1}^{k} \mathcal{E}_i$$

into k ground truth groups $\mathcal{E}_i \in \mathcal{M}$, the objective is to find a clustering

$$\mathcal{C} = \cup_{i=1}^{k'} \mathcal{C}_i$$

of $\mathcal{M}$ into k' groups such that a clustering measure $D(\mathcal{E}, \mathcal{C})$, which evaluates a divergence between the two partitions, is minimized. Note that the set the set of actual entities $\mathcal{E}$ and the underlying ground truth grouping of them is unknown to the learner. This is why the performance is measured in terms of the divergence between the learner's clustering and the underlying ground truth. Note also that we are interested in a hard clustering of the entity mentions into groups, while in some applications an entity can be long to a number of different cluster groups. Moreover, in our problem, entity mentions with the same surface name are all mapped to the same entity.

Note that a parallel problem objective is to design an evaluation scheme to determine how likely it is for a given triplet (m_s, m_o, v) with $m_s, m_o \in \mathcal{M}$ and $v \in \mathcal{V}$ to be a valid relation. In fact, most of the literature in multi-relational learning focus on this objective. Generally, there is a trade-off between the accuracy in the prediction model and interoperability of the model, as recognized by [3].

2.4 Data Representation

Recall that a triplet (m_i, m_j, v_k) simply means that the i-th entity mention and the j-th entity mention have the k-th relation. In our data model, we aim to represent each mention with two representations, one based on its appearance in the subject role and another one for its representation in the object role in the extracted triplets

$$\mathcal{T} = \{(m_s^{(i)}, m_o^{(i)}, v^{(i)})\}_{i=1}^{|\mathcal{T}|}.$$

In order to do so, we construct two matrices, namely left(resp. right) entity relation matrix, which encode the number of occurrences of a subject(resp. object) entity with all the verbs in $\mathcal{V}$. Let us denote the right entity relation matrix by $\boldsymbol{X}_R \in \mathbb{R}^{|\mathcal{M}| \times N_R}$, and the left entity relation matrix by $\boldsymbol{X}_L \in \mathbb{R}^{|\mathcal{M}| \times N_L}$. Formally, the left and right relation matrices can be defined as follows.

$$\begin{aligned} \boldsymbol{X}_L[i,j] &= \Big| \{(m_i, m, v_j) \mid (m_i, m, v_j) \in \mathcal{T}, \text{ such that } m \in \mathcal{M}\} \Big|, \\ \boldsymbol{X}_R[i,j] &= \Big| \{(m, m_i, v_j) \mid (m, m_i, v_j) \in \mathcal{T}, \text{ such that } m \in \mathcal{M}\} \Big| \end{aligned} \tag{2.1}$$

The left entity relation matrix only carries the information about the co-occurrence of subject entity mentions with all the possible relationship mentions and this is regardless of the entities that appear on the right hand side of the corresponding relations as the object entities. By the same token, the right entity relation matrix only carries the information about the co-occurrence of object entity mentions with the possible relationship mentions.

These matrix representations can be regarded as projections of a 3-way tensor $\mathcal{X}$ whose dimensions represent the left hand side entities, the right hand side entities and the relationships, explicitly defined

$$\mathcal{X}_{i,j,k} = \text{number of occurrences of } (m_i, m_j, v_k)$$

Although the tensor model encodes all the information present in the set of entity triplets $\mathcal{T}$, we argue that such projection, which throws away the information about the co-occurring entities can serve as an inherent regularization in the model, yielding feature vectors for entities that capture how likely is an entity to co-occur with a relation phrase.

In construction of the right and left matrix, we note that due of the imbalance in the occurrence of different relation mentions with an entity mention, the reconstruction of the matrix comprise of plain co-occurrence frequencies might lead to skewed representations for the embedding that do not capture the essential distributed characteristics of the entity and relation mentions.

In the literature of co-occurrence based context representation, in order to put more emphasis on the objects that are more representative of a context, the idea of Term Frequency-Inverse Document Frequency (TF-IDF) is the standard technique to assign higher values to significant objects in a context that do not occur in many other contexts. Specifically, TF-IDF measures seek to find keywords of a document by evaluating each word's frequency in the document and its frequency in all documents of the corpus. For document set D, a word t, and a document $d \in D$, if $f_{t,d}$ is the frequency of t in d and $n_{t,D}$ is the number of documents in D that contain t, a popular pair of "term frequency" and "inverse document frequency"

functions used to calculate TF-IDF are defined as:

$$TF_{t,d} = \frac{f_{t,d}}{\max_{d \in D} t, d}$$

$$IDF_t = \log \frac{|D|}{n_{t,D}}$$

Then $TF\text{-}IDF_{t,d}$ is simply equal to the product $TF_{t,d}.IDF_{t,d}$. The "term frequency" measure assigns higher value to words that are more frequent in the current document. The "inverse document frequency" seeks to diminish the importance of common words that occur frequently in every document. The result is that TF-IDF boosts significant words of a document by finding relatively rare words that are frequent in the current document. These words carry the most information about the document's meaning. One may use any other pair of functions that imply the same idea. For instance, another definition of TF and IDF can be written as:

$$TF_{t,d} = 1 + \log f_{t,d}$$

$$IDF_t = \log \left(1 + \frac{|D|}{\log n_{t,D}}\right)$$

Inspired by this observation we define a refined measure for entity-relation co-occurrence that better captures the significance of a relation mention for an entity mention and vice versa.

Before introducing our proposed entity-relationship significance measure, let us define some notations to describe co-occurrence between entity mentions and relations. We define the set of relation mentions that co-occur with a subject entity mention as

$$\mathcal{V}_m^L = \{v \mid (m, m', v) \in \mathcal{T} \text{ for some entity mention } m'\}$$

and the set of subject entities that concur with a relation mention as

$$\mathcal{M}_v^L = \{m \mid (m, m', v) \in \mathcal{T} \text{ for some entity mention } m'\}$$

We treat each entity as a bag of relation mentions that co-occur with it in some relation in $\mathcal{T}$. Therefore for each relation phrase $v \in \mathcal{V}_m^L$, we can define a notion of term frequency as

$$TF_{v,m}^L = \log(1 + s_v^L(m)) \tag{2.2}$$

where $s_v^L(m)$ is the co-occurrence score of an entity mention m with respect to the relation mention v as

$$s_v^L(m) = \sum_{m' \in \mathcal{M}} \mathbf{1}_{(m,m',v) \in \mathcal{T}} \tag{2.3}$$

Based on this score, we define the collection of all left co-occurrence scores of the relation mention v as

$$\mathcal{S}_v^L = [s_v^L(m)]_{m \in \mathcal{M}_v^L}$$

Given this collection, we can define a notion of inverse document frequency(IDF) in terms of a rank attributed to an entity mention with respect to a relation mention. A natural choice for the rank function is the usual rank of an entity with respect to the $s_v^L(m)$ score; that is

$$\text{Rank}_v^L(m) = \text{rank of } m \in \mathcal{M}_v \text{ in } \mathcal{S}_v^L, \tag{2.4}$$

which amounts to the overall TF-IDF score

$$\text{TF-IDF}^L(v, m) = \frac{TF_{v,m}^L}{\text{Rank}_v^L(m)}$$

However, this notion of rank may dump the low ranked entities in the collection $\mathcal{S}_v^L$ too much. Thus, we define a quantized rank, which scales down the rank of an entity based on its position in the ranked collection $\mathcal{S}_v^L$. To exemplify such a function, we consider

$$\text{Q-Rank}_v^L(m) = \sqrt{1 + \text{Rank}_v^L(m) \ / \ w} \tag{2.5}$$

for some fixed, window size w, which can depend on the size of the collection $\mathcal{S}_v^L$. This amounts to a final TF-IDF scoring

$$\text{TF-IDF}^L(v, m) = \frac{TF_{v,m}^L}{\text{Q-Rank}_v^L(m)} \tag{2.6}$$

Note that we can define all these measure for the right entity-relation matrix in a similar fashion.

Thus, from now on, when we refer to the left matrix $\boldsymbol{X}_L \in \mathbb{R}^{|\mathcal{M}| \times N_L}$, we mean the matrix that is populated with values defined in 2.7. Similarly for the right matrix $\boldsymbol{X}_R \in \mathbb{R}^{|\mathcal{M}| \times N_R}$, we have

$$\boldsymbol{X}_R[i,j] = \text{TF-IDF}^R(v_j, m_i), \tag{2.7}$$

for $m_i \in \mathcal{M}$ and $v_j \in \mathcal{V}$.

2.5 Latent Factor Model: Explicit Matrix Factorization

In order to cluster entity mentions into groups that contextually have the same a similar meaning, we adopt an approach based on finding low dimensional latent embeddings for the mentions. In essence, the idea is to capture complex structural properties of the objects of interest by finding distributed vector representations, a suitable function of which describes the observed data. Similar mentions can then be clustered by clustering the respective latent vector representations. The embedding idea is proved efficient, both in terms of prediction accuracy, outperforming supervised models, and efficiency in terms of the scalability of the learning process.

Given the matrix data model, presented in the previous section, we define a latent factor model that describes the observed data based on the latent embeddings for the constituent entity and relation mentions. Specifically, we assume that the entries in the the above matrices are *noisy samples* from some ground truth value generated as a function of the embedding vectors. Let the entity mention m_i be associated with some latent representations $\boldsymbol{u}^L_{m_i}, \boldsymbol{u}^R_{m_i} \in \mathbb{R}^r$. Note that $\boldsymbol{u}^L_{m_i}$ captures the properties of m_i as a subject mention in $\mathcal{T}$ and $\boldsymbol{u}^R_{m_i}$ captures the properties of m_i as an object mention in $\mathcal{T}$. By the same token, the relation verb v_j is associated with some latent factors $\boldsymbol{v}^L_{v_j}, \boldsymbol{v}^R_{v_j} \in \mathbb{R}^r$ which capture the properties of the token as combined with subject and object mentions respectively.

Then, the ground truth value for entry $\boldsymbol{X}_L[i,j]$ is generated as a noisy version of $f(\boldsymbol{u}^L_{m_i}, \boldsymbol{v}^L_{v_j})$,

for some function f. Likewise, the entries of the right entity relation matrix are functions of the corresponding object embeddings of the entity mentions and right hand side distributed representations of the relation mentions. The following plate representation demonstrates how such latent representations amount to generating the observed values in the left and right relation matrices.

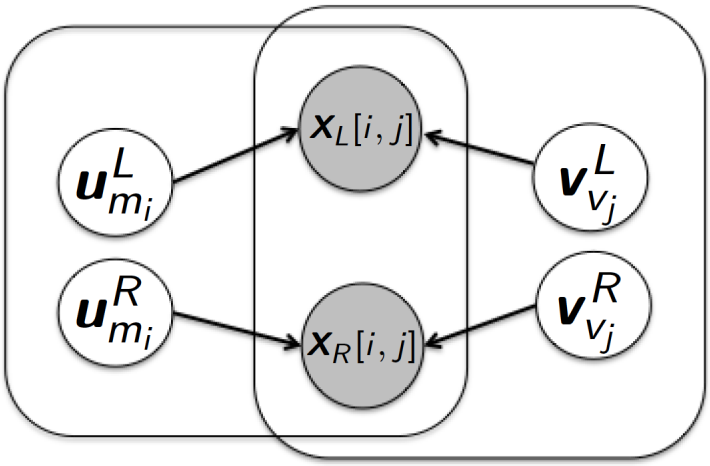

Figure 2.1: Plate Representation for the Latent Matrix Model

Next, we specify the choice of the function $f(\cdot)$ to be the inner product of the entity and relation embedding arguments, that is

$$\begin{aligned} \boldsymbol{X}_L[i,j] &= \langle \boldsymbol{u}^L_{m_i}, \boldsymbol{v}^L_{v_j} \rangle + \epsilon^L_{ij} \\ \boldsymbol{X}_R[i,j] &= \langle \boldsymbol{u}^R_{m_i}, \boldsymbol{v}^R_{v_j} \rangle + \epsilon^R_{ij} \end{aligned} \tag{2.8}$$

An important assumption that we make about unobserved pairs is that the corresponding embedding vectors for the entity and relation mentions should estimate a zero value. Following this assumption, for any triplet (m_i, m_j, v_k) with $m_i, m_j \in \mathcal{M}$ and $v_k \in \mathcal{V}$, we can define a scoring function of the following form

$$\mathcal{S}((m_i, m_j, v_k)) = \ell(\boldsymbol{X}_L[i,j], \langle \boldsymbol{u}^L_{m_i}, \boldsymbol{v}^L_{v_j} \rangle) + \ell(\boldsymbol{X}_R[i,j], \langle \boldsymbol{u}^R_{m_i}, \boldsymbol{v}^R_{v_j} \rangle), \tag{2.9}$$

for some loss function $\ell(\cdot,\cdot) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ Note that, this scoring function can in turn be used as a confidence measure for a relation triplet to be a valid relation.

Note that such latent factor model can be understood as factorizing each of the right and left entity relation matrices into two lower dimensional matrices, as demonstrated in the following figure.

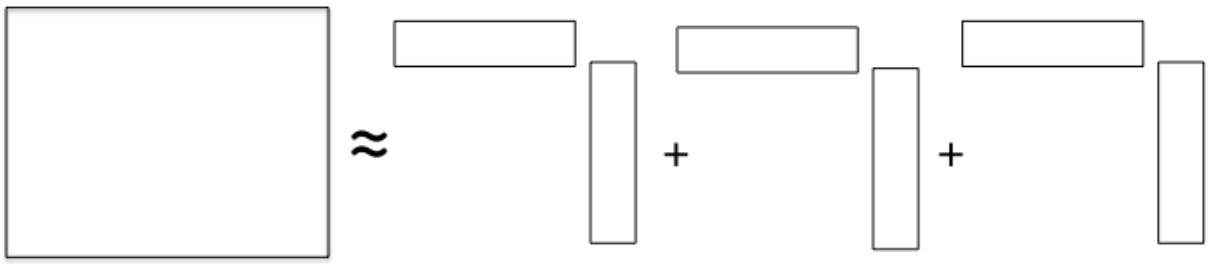

Figure 2.2: Latent Embedding Model as Matrix Factorization

2.6 Learning Latent Embeddings: Non-Negative Matrix Factorization

We now describe how to learn the parameters of the proposed latent factor model. Note that, for each relation and entity mention, we have to learn two vector representations; that is the parameters of the model are

$$\theta = \{\boldsymbol{u}_m^L, \boldsymbol{u}_m^R, \boldsymbol{v}_v^L, \boldsymbol{v}_v^R | m \in \mathcal{M}, v \in \mathcal{V}\}, \tag{2.10}$$

where $\boldsymbol{u}_m^L, \boldsymbol{u}_m^R, \boldsymbol{v}_v^L, \boldsymbol{v}_v^R \in \mathbb{R}^r$. Following a matrix representation,

$$\theta = \{\boldsymbol{U}_L, \boldsymbol{U}_R \in \mathbb{R}^{|\mathcal{M}| \times r}, \boldsymbol{V}_L, \boldsymbol{V}_R \in \mathbb{R}^{|\mathcal{V}| \times r}\},$$

where each row of the matrices above represents right/left hand side representation of an entity/relation mention.

In order to learn the parameters of the model, we define an optimization objective based on the scoring function defined earlier. Specifically we learn the parameters of the model by solving the following regularized matrix factorization objective. We only specify the optimization for the left entity relation matrix and the left hand side representation of the

entity and relation mentions. The formulation for the right hand side follows similarly. In order to learn the parameters $(\boldsymbol{U}_L, \boldsymbol{V}_L)$, we solve

$$\min_{\substack{\boldsymbol{U}\in\mathbb{R}^{|\mathcal{M}|\times r}\\ \boldsymbol{V}\in\mathbb{R}^{|\mathcal{V}|\times r}}} \quad \mathcal{L}(\boldsymbol{X}_L, \boldsymbol{U}\boldsymbol{V}^T) + \rho(\boldsymbol{U}, \boldsymbol{V}), \tag{2.11}$$

where $\mathcal{L}(\cdot,\cdot)$ is a strongly convex loss function and $\rho(\cdot,\cdot)$ is a regularization function to impose structural assumptions on the embeddings.

Here, we note that the underlying entity relation matrices are very sparse. Therefore, since we aim to use the obtained embeddings of the entity mentions to cluster semantically similar entities, we require the embedded vectors to have certain structural properties, namely sparsity. In order to impose sparse structure on entity embeddings, we consider non-negativity constraint on the factor matrices in the matrix factorization objective. The reason why, such non-negativity constraint can be used as a surrogate to sparsity is described in details in the next chapter. Therefore, the learning objective simply becomes

$$\min_{\substack{\boldsymbol{U}\in\mathbb{R}_+^{|\mathcal{M}|\times r}\\ \boldsymbol{V}\in\mathbb{R}_+^{|\mathcal{V}|\times r}}} \quad \mathcal{L}(\boldsymbol{X}_L, \boldsymbol{U}\boldsymbol{V}^T). \tag{2.12}$$

To solve the non-negative matrix factorization(NMF) problem, we propose an exterior point method, which constructs a solution for the NMF problem based on a suitably rotated optimal solution of the unconstrained matrix factorization problem, which we detail in the next chapter.

2.7 Using Embeddings for Clustering

Let $\{\boldsymbol{u}_m\}_{m\in\mathcal{M}}$ be the set of entity mention embeddings. Note that in our methods, we have two distinct embeddings for an entity mention that characterize its behavior as a subject and an object in relationships separately. Therefore, the overall distributed representation of an entity will be the concatenation of the two representations, that is

$$\boldsymbol{u}_m = \boldsymbol{u}_m^L || \boldsymbol{u}_m^R$$

where $\boldsymbol{u}_m^L$ and $\boldsymbol{u}_m^R$ represent the right hand side and the left hand side representations for entity $m \in \mathcal{M}$ and the concatenation operator is denoted by $||$.

Given these embeddings, one can use a variety of different clustering algorithms to assign similar entity mentions to the same group. Since the obtained embeddings are supposedly sparse vectors in a Euclidean space, we use K-means clustering as our method of choice. Defining $\boldsymbol{U} = \boldsymbol{U}_L||\boldsymbol{U}_R$, the K-means clustering problem can be expressed as

$$\min_{\substack{\boldsymbol{G}\geq 0 \\ \boldsymbol{G}^T\boldsymbol{G}=I}} \ \mathrm{TR}(\boldsymbol{G}^T\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{G}).$$

Finally, we note here that the distributed representations learned by our method to cluster entity mentions are only based on the relationships that the mentions appear in. However, one can incorporate other embeddings of the entity mentions that capture other types of information that we have about the entities.

For example in order to incorporate the co-occurrence information of the mentions from text, one can characterize topic/co-occurrence embeddings of the entities similarly to the relation embeddings. Specifically one can use word vector representations [4, 5]as co-occurrence embeddings or we can use a topic modeling algorithm to embed the mentions in a topic space. We also note that in order to combine different embeddings of the entity mentions we have to use appropriate normalization and and weighting.

2.8 Evaluating the clustering results

Given a set $\mathcal{M}$ of N elements $S = \{m_1, m_2, \cdots m_N\}$, consider two partitions of a set of $\mathcal{M}$, namely $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$ with K clusters, and $\mathcal{E} = \{E_1, E_2, \ldots, E_J\}$ with J clusters.

Note that we are considering a hard clustering of the set $\mathcal{M}$, which means that the groups in each partition are pairwise disjoint; $U_i \cap U_j = V_i \cap V_j = \varnothing$, for all $i \neq j$.

We also assume that the portions are complete; that is $\cup_{i=1}^{K} C_i = \cup_{j=1}^{J} E_j = \mathcal{M}$.

The mutual information of cluster overlap between $\mathcal{C}$ and $\mathcal{E}$ can be summarized in a matrix $\boldsymbol{N} \in \mathbb{R}^{K\times J}$, where $\boldsymbol{N}[i, j]$ denotes the number of objects that are common to clusters C_i

and E_j.

Suppose an object is picked at random from $\mathcal{M}$; the probability that the object falls into cluster C_i is $P_i = \frac{|C_i|}{N}$ Then, the entropy associated with the partitioning $\mathcal{C}$ is:

$$H(\mathcal{C}) = -\sum_{i=1}^{K} P_i \log P_i$$

Similarly, the entropy of the clustering $\mathcal{E}$ can be calculated as:

$$H(\mathcal{E}) = -\sum_{i=1}^{J} P'_i \log P'_i$$

where $P'_j = |E_j|/N$.

The Mutual Information (MI) between two partitions can be computed as

$$I(\mathcal{C};\mathcal{E}) = \sum_{i=1}^{K}\sum_{j=1}^{J} P_{i,j} \log \frac{P_{i,j}}{P_i P'_j}$$

where $P_{i,j}$ denotes the probability that a point belongs to both the cluster C_i in $\mathcal{C}$ and cluster E_j in $\mathcal{E}$; that is $P_{i,j} = \frac{|C_i \cap E_j|}{N}$. Mutual Information is a non-negative quantity upper bounded by the entropies $H(\mathcal{C})$ and $H(\mathcal{E})$. It quantifies the information shared by the two partitions and thus can be employed as a clustering measure.

We can also define the Conditional Entropy of a partition given another one as

$$H(\mathcal{C}|\mathcal{E}) = -\sum_{i=1}^{K}\sum_{j=1}^{J} P_{i,j} \log \frac{P_{i,j}}{P'_j}$$

The Homogeneity score of a partition $\mathcal{C}$ given the ground truth partitions $\mathcal{E}$ is then defined as

$$h(\mathcal{C};\mathcal{E}) = 1 - \frac{H(\mathcal{C}|\mathcal{E})}{H(C)}.$$

By the same token, the completeness score of a partition $\mathcal{C}$ given the ground truth partitions $\mathcal{E}$ is defined as

$$c(\mathcal{C};\mathcal{E}) = 1 - \frac{H(\mathcal{E}|\mathcal{C})}{H(C)}.$$

Rosenberg and Hirschberg [6] further define V-measure as the harmonic mean of homogeneity and completeness

$$v(\mathcal{C};\mathcal{E}) = \frac{2h(\mathcal{C};\mathcal{E})c(\mathcal{C};\mathcal{E})}{h(\mathcal{C};\mathcal{E}) + c(\mathcal{C};\mathcal{E})}.$$

2.9 Related work and Future Directions

In this section, we review the related works to our entity resolution problem. We first provide a brief survey on the other information retrieval problems. We then review the generic approaches to the problem, namely i) the supervised approach, where the objective is to predict the type of an entity or relation mention based on a training set, ii) the generative approach, where the objective is to describe the relation generating procedure in a Bayesian paradigm, iii) Similarity clustering based approach, where the relation information is used as feature vectors based on which a similarity score can be defined between two entity mentions, and iv) embedding-based approach, where the objective is to associate low dimensional latent distributed representations to entity and relation mentions that capture their behavior in relationships. Finally, we provide a structured review of the prior art in embedding based relational learning, and show how our approach can be viewed as an instance of such formulation. We conclude by a few concluding remarks.

2.9.1 Entity Resolution in Other Domains

Entity resolution is the problem of identifying, matching, and merging references corresponding to the same entity within a dataset. It lies in the core of many other information retrieval problems in a number of different domains. In the following we summarize some of the major domains in which the entity resolution comes forth as an essential task for understanding the underlying structures in the problem.

- In Data Base systems, Entity Resolution (ER) is the task of identifying all records in a database that refer to the same underlying entity. A related problem that is extensively studied in the literature on relation discovery is to identify whether a relation can be

added to the data-base. Given a predefined database schema, the traditional supervised approach to determine whether candidate entities extracted from the text corpora have a particular relations is to learn a classifier based on clues in the textual data in the form of patterns between the occurrences of two candidate mentions in the documents. In order to incorporate the existing known facts in the data based in the learning process a variety of distant supervision methods are proposed [7–9]. A more recent approach is to bring in also the existing facts from an external knowledge to build upon a richer set of seed relationships [2, 10, 11]. Such approaches can guide using a similar approach in the entity resolution problem in the context of story narrative analysis. For example, when the some of the actant groups, type of some of the entity mentions or the type of their relationships are known a priori.

- Knowledge-Base Construction: Web-scale knowledge bases (KBs) provide a structured representation of human knowledge in a variety of domains. Popular examples of knowledges bases include DBPedia [12], Freebase [13] and the Google Knowledge Vault [14]. Knowledge-bases are used in a number of information retrieval applications such as recommender systems, question answering, and query rewriting for search. Identifying the relationship between entities from free text is key to acquiring new facts to increase the coverage of a structured knowledge base. As we will discuss in more details, the predominant approach for understanding knowledge base graphs is to find low dimensional distributed representations(embeddings) for nodes(and links in the graph). A comprehensive review on knowledge graph embedding techniques can be found in [15].

- Link Prediction: In statistical relational learning, the link prediction problem is to determine whether two entities in an information network are connected. Knowledge-base construction can be viewed as an instance of link prediction. In social networks, the link prediction problem is to infer which new interactions among users are likely to form in the near future, given a snapshot of a social network in the previous times [16].

- Sense Disambiguation: Sense identification and disambiguation are long standing problems in natural language processing. Sense Identification is an essential aspect of the

of synonymy analysis, where different words can be used to refer to the same sense. Sense Disambiguation, however, deals with polysemous words that can correspond to multiple senses.

- Citation Networks: Document retrieval, specifically in the context of academic papers, is an important information retrieval problem, where the objective is to rank documents with respect to a query by estimating the probability of relevance for each document. Methods based on citation counting has been the primary approach for academic paper retrieval. In [17], a relation-based mathod is proposed for building structural retrieval results for academic literatures in order to uncover the relationships of the retrieval results.

- Familial Networks: Familial networks consist of the members of a family together with their relations. Such networks are prevalent in family health history applications, genealogy, areal administrative records. Given partial views of a familial network as described from the point of view of different people in the network the entity resolution problem in a familial network is to reconstruct the underlying familial network from these perspective partial views [18].

2.9.2 Generic Approaches to Entity Resolution

The entity resolution problem is framed as a classification problem with a given a set of training data in a supervised setting. Formally, the entity classification(typing) problem can be summarized as follows: given a sentence S with the annotated pairs of entity mentions m_1 and m_2, the problem is to identify the type of the entity mentions m_1 and m_2. By the same token relation classification problem is to determine whether there is relation (and if so, of what type) between entity mentions m_1 and m_2. In the supervised setting, this problem is studied as a multi-classification problem and it has yield relatively high performance [7, 19] For such methods to perform well, complex features should be extracted from the text. A variety of techniques are used to design lexical level features, such as parsing tree features and POS tags, sentence level features, such as contextual information about co-occurrence

of the words, and word features, which are a combination of word's vector representation learned from external datasets [4, 5] and the vector representations of the words in context, in a certain window. [20–22] . Feature-based methods suffer from the problem of choosing a suitable feature set when converting the structured representation into feature vectors.

The major drawback of supervised models though is the necessity of labeled data. Knowledge-bases are usually domain specific and may not provide much useful information about the entities in the corpus. Moreover, in many applications entity types are highly context specific and may not necessarily correspond to the entity types preset in knowledge bases. Therefore, it is inevitable to develop unsupervised models for this task.

Yao et al. [23] propose generative probabilistic models that model the relation phrase in a relation triplet by the surface syntactic dependency path between the pair of relation mentions. In their model, entity type constraints extracted from a knowledge base are used along with features on the dependency path between the entity mentions.

The final objective is to find a clustering over observed relation paths such that expressions in the same cluster have the same semantic relation type. Specifically, they proposed variants of Latent Dirichlet Allocation(LDA) [24], the celebrated generative topic model to describe how documents are generated as bags of words from a set of latent distributions on words, called topics. At the document level a multinomial distribution is drawn over a fixed number of relation types $|\boldsymbol{V}|$, denoted by θ, from a pripor $\mathrm{Dir}(\alpha)$. A document consists of a bag of relation tuples. Each relation tuple is drawn from a relation type topic distribution $\mathrm{Multi}(\theta)$ selected by a latent relation type indicator variable. Specifically, relation tuples are generated using a collection of independent features $\{f\}$ drawn from the underlying relation type distribution $\phi_{r,f}$. The model is describe via the following plate notation.

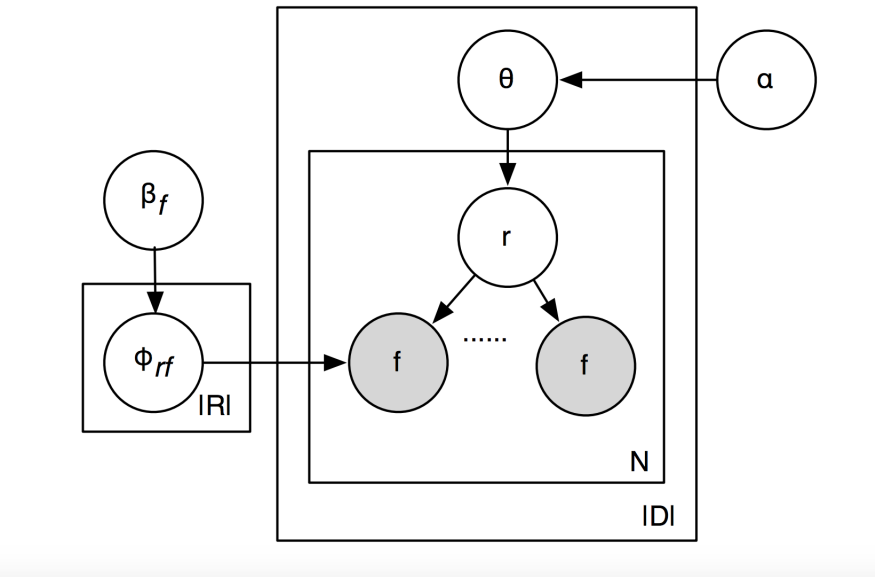

Figure 2.3: Plate Representation for REL-LDA model

A related relation extraction system is proposed in [25], referred to as Discovery of Inference Rules from Text (DIRT), which aims to discover different representations of the same semantic relation using distributional similarity [26] of dependency paths.

The problem of such generative models is their scalability, which becomes prohibitive in large scale applications. Also due to the large number of parameters of the model, the data requirements of such methods are also prohibitive. The modern approach to entity resolution is to associate the entity and relation mentions with low dimensional distributed representations that capture the complex behavior of such objects. In a data-rich setting, by taking the ambient dimension of the distributed representations(aka embeddings) large enough, many of the complex structural properties, such as hierarchical relationships between different objects can be captured in the flattened real representations. The embedding approach is proved promising in a variety of different predictive tasks in natural language processing. The underlying principle of most of these methods is to find a factorization of the matrix/tensor in a low dimensional space and use the obtained factors as low dimensional embeddings of the entities. Models based on tensor factorization [27, 28] are proved efficient in terms of scalability. In a predictive setting, factorization based methods are outperformed

by energy based models [29–31]. These methods represent entity mentions as low-dimensional embeddings and relations as linear or bilinear operators on them. Another big advantage of such methods is that the models can be trained using stochastic approximation methods, which speeds up the optimization and makes the task practically online. Such scalability properties are desirable to high dimensional applications.

While such latent representations are usually used to predict unknown relationships between new pairs of entities and relation phrases, in this work, we take advantage of the obtained embeddings to cluster similar entities.

2.9.3 Embedding-based Models

2.9.3.1 Data Model and Latent Factor Model: Tensors

Before we start the discussion, we present the data model that most of the embedding-based approaches have adopted. In order to encode the set of all triples

$$\mathcal{T} = \{(m_s^{(i)}, m_o^{(i)}, v^{(i)})\}_{i=1}^{|\mathcal{T}|},$$

a 3-way tensor $\mathcal{X}$ captures all the information in $\mathcal{T}$, where the dimensions of the tensor represent the left hand side entities, the right hand side entities and the relationships.

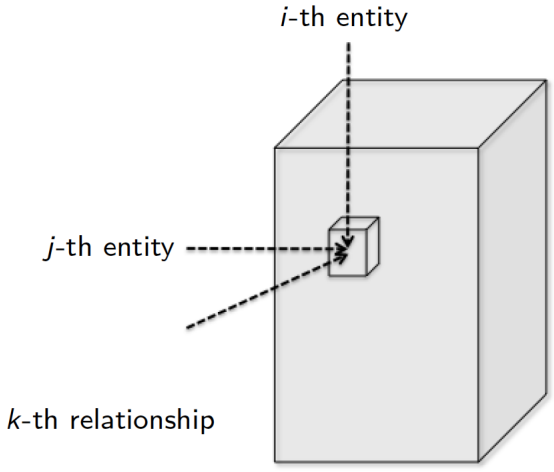

Figure 2.4: Tensor Representation of the Triplets

Particularly,

$$\mathcal{X}_{i,j,k} = \text{number of occurrences of } (m_i, m_j, v_k).$$

Depending on the application and the assumptions on the underlying ground truth, different tensor latent tensor models are defined. For example, in a knowledge base construction or link prediction application, the entries of the tensor will be binary encoding of whether the triplet exits or not.

In a latent factor model, the entries in the tensor encoding of the data can be regarded as *noisy samples* from some underlying ground truth, which is in turn generated by latent factors corresponding to the constituent entity and relation mentions. Formally, let $\boldsymbol{a}_{m_i}$, $\boldsymbol{a}_{m_j}$, and $\boldsymbol{W}_{v_k}$ encode latent representations for i-th left entity mention, j-th right entity mention, and k-th relation respectively. Then, $\mathcal{X}_{i,j,k}$ is a noisy observation of $f(\boldsymbol{a}_{m_i}, \boldsymbol{a}_{m_j}, \boldsymbol{W}_{v_k})$, for some positive valued function f.

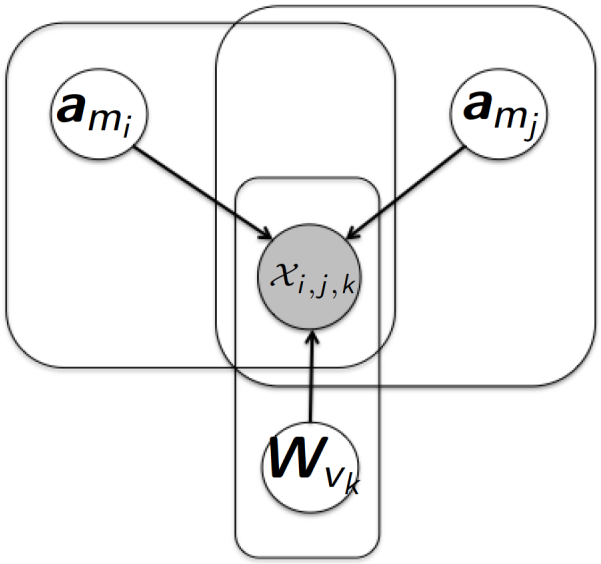

Figure 2.5: Plate Representation for the Latent Tensor Model

2.9.3.2 Explicit Matrix factorization

A natural way, similar to our approach, to learn latent embeddings for entity and relation mentions is through factorizing the the data tensor. PARAFAC model is one of the oldest models for tensor factorization, which can be regarded as a simple generalization of matrix factorization [28, 32, 33]. In this model, a three dimensional tensor $\mathcal{X} \in \mathbb{R}^{|\mathcal{M}|\times|\mathcal{M}|\times|\mathcal{V}|}$ is factorized via three matrices $\boldsymbol{A} \in \mathbb{R}^{|\mathcal{M}|\times d}, \boldsymbol{B} \in \mathbb{R}^{|\mathcal{M}|\times d}$, and $\boldsymbol{C} \in \mathbb{R}^{|\mathcal{V}|\times d}$, so that every entry of the data tensor is generated by the following generalized dot product of the factor matrices,

$$\widehat{\mathcal{X}}_{i,j,k} = \sum_{r}^{d} \boldsymbol{A}_{i,r}\boldsymbol{B}_{j,r}\boldsymbol{C}_{k,r}.$$

Particularly, PARAFAC solves the following optimization problem

$$\min_{\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}} D(\mathcal{X}; \widehat{\mathcal{X}}),$$

for some appropriate divergence measure $D(\cdot;\cdot)$.

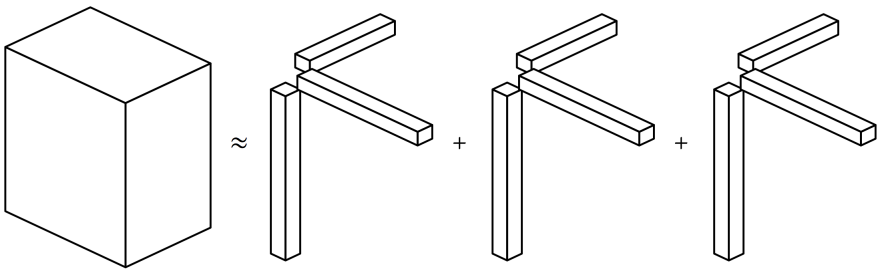

Figure 2.6: PARAFAC model

There are a number of different ways to explicitly factorize a tensor into latent components. Nickel et al. propose RESCAL [27], a tensor factorization model where the tensor factorization can be reduced to a number of low rank matrix factorizations that share a common factor matrix that encodes low dimensional distributed representations for entities. More precisely let $\mathcal{X}_k$ be the k-th frontal slice of the tensor(that is the slice corresponding to relation v_k). The RESCAL model seeks to find factorizations in the following form

$$\mathcal{X}_k \sim \boldsymbol{A}\boldsymbol{R}_k\boldsymbol{A}^T,$$

where $\boldsymbol{A} \in \mathbb{R}^{|\mathcal{M}| \times r}$ represents r dimensional latent representations for entity mentions, and $\boldsymbol{R}_k \in \mathbb{R}^{r \times r}$ is an asymmetric square matrix that represents the interactions between the latent representations of the entities corresponding to the k-th relation.

In order to learn the latent representations for entities and the relationships the RESCAL model solves the following optimization problem

$$\min_{\boldsymbol{A}, \boldsymbol{R}_k} \sum_{k}^{|\mathcal{V}|} ||\mathcal{X}_k - \boldsymbol{A}\boldsymbol{R}_k\boldsymbol{A}^T||_F^2 + \lambda(||\boldsymbol{A}||_F^2 + \sum_k ||\boldsymbol{R}_k||_F^2)$$

Note that the first term is the reconstruction error for predicting the known values in the tensor and the regularization terms correspond to smoothness penalties for the embedding matrices.

RESCAL is a special form of Tucker decomposition [34] operating on a 3-dimensional tensors. It can also be regarded as a relaxed form of DEDICOM [35].

Note that although the problem is strongly convex on each of the optimization variables, it is a non-convex optimization problem when all the variables $\boldsymbol{A}$ and all $\boldsymbol{R}$'s are considered together.

Using block coordinate descent by fixing $\boldsymbol{A}$ and $\boldsymbol{R}_k$ alternatively, the following iterative update equations are can be obtained

$$\boldsymbol{A} \leftarrow \left[\sum_k \mathcal{X}_k \boldsymbol{A}\boldsymbol{R}_k^T + \mathcal{X}_k^T \boldsymbol{A}\boldsymbol{R}_k\right] \left[\sum_k \boldsymbol{B}_k + \boldsymbol{C}_k + \lambda \boldsymbol{I}\right]^{-1}$$

where $\boldsymbol{B}_k = \boldsymbol{R}_k\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{R}_k^T$ and $\boldsymbol{C}_k = \boldsymbol{R}_k^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{R}_k$

and

$$\text{vec}(\boldsymbol{R}_k) \leftarrow \left([\boldsymbol{Z}^T\boldsymbol{Z} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{Z}^T \text{vec}(\mathcal{X}_k),$$

where $\boldsymbol{Z}_t = \boldsymbol{A} \bigotimes \boldsymbol{A}$, with $\bigotimes$ denoting Kronecker product.

Since each relation phrase only connects certain subsets of entities that belong to some "compatible types" with respect to the relationship, [36] proposes to incorporate the type information of the entities and relationships into the factorization process. They argue that incompatible entity-relations incurs unnecessary computations in each optimization step and by choosing values for the incompatible entries, the quality of training is reduced. Therefore,

they propose the following alternative optimization which only deals with compatible entities with the relationships. let $\mathcal{L}_{v_k}$ and $\mathcal{R}_{v_k}$ be the set of left hand side and right hand side entities with a compatible type to the k-th relation. That is, (m_i, m_j, v_k) is a feasible triple if and only if $m_i \in \mathcal{L}_{v_k}$ and $m_j \in \mathcal{R}_{v_k}$. Correspondingly, let $\boldsymbol{A}_k^{\mathcal{L}}$ and $\boldsymbol{A}_k^{\mathcal{R}}$ denote the sub-matrices of $\boldsymbol{A}$ that consists of rows associated with $\mathcal{L}_{v_k}$ and $\mathcal{R}_{v_k}$, respectively. By the same token, let $\mathcal{X}_k^{\mathcal{L},\mathcal{R}}$ denote the sub-matrix of $\mathcal{X}_k$ that corresponds only to entities in $\mathcal{L}_{v_k}$ and $\mathcal{R}_{v_k}$. [36] solves the following optimization problem

$$\min_{\boldsymbol{A},\boldsymbol{R}_k} \sum_{k}^{|\mathcal{V}|} ||\mathcal{X}_k^{\mathcal{L},\mathcal{R}} - \boldsymbol{A}_k^{\mathcal{L}} \boldsymbol{R}_k \boldsymbol{A}_k^{\mathcal{R}^T}||_F^2 + \lambda(||\boldsymbol{A}||_F^2 + \sum_k ||\boldsymbol{R}_k||_F^2)$$

Note that RESCAL scales linearly with the number of entities, linearly with the number of relations, and linearly with the number of known facts, but it scales cubical with regard to the rank of the factorization.

In [37], theoretical bounds are developed on the factorization rank and they propose a similar model to RESCAL with an extra additive term to address the scalability problem.

2.9.3.3 Neural embedding approach

One can view learning the embeddings, as learning weights of a shallow neural network, where the first set of weights projects a pair of input entity mentions to low dimensional vectors in the middle layer, and in the last layer the latent representations in the middle layer are combined to a scalar for comparison via a scoring function with relation-specific parameters. Each input entity mention is then represented as a high-dimensional vector, usually a "one-hot" index vector while it can be a dense high dimensional feature vector. Specifically, let x_m be a one-hot encoding representation of the entity m. Entity mention embeddings then can be represented as

$$\begin{aligned} \boldsymbol{a}_s &= f(\boldsymbol{W}_{\mathcal{M}} \boldsymbol{x}_{m_s}) \\ \boldsymbol{a}_o &= f(\boldsymbol{W}_{\mathcal{M}} \boldsymbol{x}_{m_o}), \end{aligned} \tag{2.13}$$

where f can be a linear or non-linear function, and $\boldsymbol{W}_{\mathcal{M}}$ is a parameter matrix, which can

be randomly initialized or initialized using pre-trained vectors, like word vectors obtained by word2vec [4] or [5].

The choice of the embedding for the relation mention becomes explicit in the form of the scoring function for each relation triplet (m_s, m_o, r). The scoring functions adopted in the literature, such as DistMul [38], Neural tensor network (NTN) model [31], TransE [30] and Distance [29], can be unified based on a basic linear transformation or a bilinear transformation or a combination of both. When describing each algorithm, we will specify the choice of representation for the relation phrase and the corresponding scoring function.

In [38], the following bilninear scoring function is adopted

$$g^{\mathrm{DM}}(m_s, m_o, v) = \boldsymbol{a}_s \boldsymbol{W}_v \boldsymbol{a}_o \tag{2.14}$$

where $\boldsymbol{W}_v$ is a diagonal matrix that represents the distributed embedding of the relation phrase v. This model can be viewed as special case of the RESCAL [27] scoring function where the relation matrix $\boldsymbol{W}_v$ is not constrained to be diagonal. Considering the diagonal constraint on the relation matrix $\boldsymbol{W}_v$, this scoring function is essentially equivalent to

$$g^{\mathrm{DM}}(m_s, m_o, v) = \boldsymbol{a}_v^T (\boldsymbol{a}_s \odot \boldsymbol{a}_o), \tag{2.15}$$

with $\boldsymbol{a}_v$ being a vector representation for the relation phrase v. Note that this scoring function is similar to the PARAFAC model [32].

The neural network parameters of can be learned by minimizing a margin-based ranking objective, which encourages the scores of positive relation triplets to be higher than the scores of any negative relation triplets. In most of the applications, only positive triplets are observed in the data. Given a set of positive triplets $\mathcal{T}$, negative examples can be constructed by corrupting either one of the relation arguments. The set of negative examples can be defined as

$$\begin{aligned} \mathcal{T}' = &\{(m', m_o, v) \mid m' \in \mathcal{M}, (m_s, m_o, v) \in \mathcal{T}\} \\ &\cup \{(m', m_o, v) \mid m' \in \mathcal{M}, (m_s, m_o, v) \in \mathcal{T}\} \end{aligned} \tag{2.16}$$

The margin-based ranking objective can then be defined as

$$\mathcal{J}(\theta, \mathcal{T}, \mathcal{T}') = \sum_{(m_s, m_o, v) \in \mathcal{T}} \sum_{(m'_s, m'_o, v) \in \mathcal{T}'} \max(0, 1 - g(m_s, m_o, v) + g(m'_s, m'_o, v)) \tag{2.17}$$

where θ is the set of parameters of the model. Note that in the case of the DistMul algorithm

$$\theta^{\mathrm{DM}} = \{\{\boldsymbol{W}_v \mid v \in \mathcal{V}\}, \boldsymbol{W}_{\mathcal{M}}\} = \{\boldsymbol{a}_m, \boldsymbol{a}_v || m \in \mathcal{M}, v \in \mathcal{V}\} \tag{2.18}$$

Neural tensor network (NTN) model is introduced in [31] to predict new relationship entries that can be added to a database, given a set of existing relations.

The scoring function for a relation triplet (m_s, m_o, v) in NTN is defined as

$$g^{\mathrm{NTN}}(m_s, m_o, v) = \boldsymbol{u}^T f(\boldsymbol{a}_s^T \mathcal{W}_v^{[1:k]} \boldsymbol{a}_o + \boldsymbol{V}_v \begin{bmatrix} \boldsymbol{a}_s \\ \boldsymbol{a}_o \end{bmatrix} + \boldsymbol{b}_v), \tag{2.19}$$

where $f(\cdot)$ is a standard non-linear function like hyperbolic tangent function, $\mathcal{W}_r^{[1:k]} \in \mathbb{R}^{d \times d \times k}$ is a tensor that represents weights specific to relation v and the bilinear tensor product $\boldsymbol{a}_s^T \mathcal{W}_v^{[1:k]} \boldsymbol{a}_o$ results in a vector $\boldsymbol{h}$, where

$$\boldsymbol{h}[i] = \boldsymbol{a}_s^T \mathcal{W}_v^{[i]} \boldsymbol{a}_o$$

with $\mathcal{W}_v^{[i]}$ being the i-th frontal slice of the tensor $\mathcal{W}_v^{[1:k]} \in \mathbb{R}^{d \times d}$. The remaining parameters for relation v are the standard form of a neural network, with $\boldsymbol{u} \in \mathbb{R}^k$ and $\boldsymbol{V} \in \mathbb{R}^{k \times 2d}$ and $\boldsymbol{b}_v \in \mathbb{R}^k$.

In order to train the parameters of the model

$$\theta^{\mathrm{NTN}} = \{\{\boldsymbol{u}, \mathcal{W}_v, \boldsymbol{V}_v, \boldsymbol{b}_v\}_{v \in \mathcal{V}}, \{\boldsymbol{a}_m\}_{m \in \mathcal{M}}\},$$

they minimize the contrastive max-margin objective as defined in 2.17. The model is then trained by taking gradients with respect to the parameters and is solve by a first order method, specifically mini-batch L-BFGS.

Bordes et al. [29] propose a model to learn to representations for elements of a knowledge base in a low dimensional embedding vector space. For a relationship triplet (m_s, m_o, v), their scoring function is defined as

$$g^{\mathrm{EM}}(m_s, m_o, v) = ||\boldsymbol{W}_{v,L} \boldsymbol{a}_s - \boldsymbol{W}_{v,R} \boldsymbol{a}_s||_1 \tag{2.20}$$

where for a given relation type v, matrices $\boldsymbol{W}_{v,L}, \boldsymbol{W}_{v,R} \in \mathbb{R}^{d\times d}$ are specific similarity measure that captures the characteristics of the relation. The above loss function can in fact be modeled as a neural network, specifically a generalization of a siamese network [39], conventionally takes a pair of inputs learns a similarity measure between them.

In order to train the parameters of the model

$$\theta^{\text{EM}} = \{\{\boldsymbol{W}_{v,L}, \boldsymbol{W}_{v,R}\}_{v\in\mathcal{V}}, \{\boldsymbol{a}_m\}_{m\in\mathcal{M}}\},$$

they minimize the contrastive max-margin objective as defined in 2.17 using stochastic gradient descent.

In [40], in attempt to maintain the scalability while preserving the capability of handling asymmetry in relationship scoring, they argue that the standard dot product between embeddings can be a very effective composition function if complex embeddings are used.

Specifically they propose the following scoring function

$$g^{\text{CE}}(m_s, m_o, v) = \mathcal{R}e(\boldsymbol{a}_s \boldsymbol{a}_v, \bar{\boldsymbol{a}_o}) \tag{2.21}$$

where $\bar{\boldsymbol{a}_o}$ is the conjugate of the complex vector $\boldsymbol{a}_o$.

In order to train the parameters of the model

$$\theta^{\text{CE}} = \{\{a_v\}_{v\in\mathcal{V}}, \{\boldsymbol{a}_m\}_{m\in\mathcal{M}}\},$$

they minimize the contrastive max-margin objective as defined in 2.17.

Another novel way to handle antisymmetry is via the Holographic Embeddings (HolE) model by [41]. In HolE the circular correlation is used for combining entity embeddings, measuring the covariance between embeddings at different dimension shifts. This generally suggests that other composition functions than the classical tensor product can be helpful as they allow for a richer interaction of embeddings. Specifically, they define their scoring function as

$$g^{\text{HolE}}(m_s, m_o, v) = \boldsymbol{a}_v^T(\mathcal{F}^{-1}[\bar{\mathcal{F}}[\boldsymbol{a}_s] \odot \mathcal{F}[\boldsymbol{a}_o]]), \tag{2.22}$$

where $\mathcal{F}[\cdot]$ and $\mathcal{F}^{-1}[\cdot]$ denote the Fourier transform and its inverse respectively. In order to train the parameters of the model

$$\theta^{\text{HolE}} = \{\{a_v\}_{v\in\mathcal{V}}, \{\boldsymbol{a}_m\}_{m\in\mathcal{M}}\},$$

they minimize the contrastive max-margin objective as defined in 2.17.

2.9.3.4 Bayesian Clustered Tensor Factorization

Unlike most of the other approaches that aim only at making predictions whether particular unobserved relations are likely to be true, Sutskever et al. [3] propose a model to discover interpretable structures in the data. Although, they are not particularly interested in clustering similar entities,

Specifically, they introduce the Bayesian Clustered Tensor Factorization (BCTF) model, which embeds a factorized representation of relations in a nonparametric Bayesian clustering framework. They define a joint distribution over the truth values of all conceivable relations.

Formally, for each entity mention $m \in \mathcal{M}$, similar to our model, the model maintains two vectors $\boldsymbol{a}_{m,L}, \boldsymbol{a}_{m,R} \in \mathbb{R}^d$, and for each relation $v \in \mathcal{V}$ it maintains a matrix $\boldsymbol{V}_v \in \mathbb{R}^{d\times d}$, where d is the dimensionality of the model.

Then, for any triplet (m_s, m_o, v), the probability that such triplet is an actual relationship can be expressed as

$$\mathbb{P}(T(m_s, m_o, v) = 1|\theta) = \frac{1}{1 + \exp(-\boldsymbol{a}_{m_s,L}^T \boldsymbol{V}_v \boldsymbol{a}_{m_o,R})} \tag{2.23}$$

where θ encodes the set of all the parameters of the model. θ can then be learned by minimizing a penalized log-likelihood

$$\sum_{(m_s,m_o,v)\in\mathcal{T}} -\log \mathbb{P}(T(m_s, m_o, v) = 1|\theta) + \rho(\theta) \tag{2.24}$$

In order to make the model fully Bayesian, they next define a prior over the vectors $\boldsymbol{a}_{m,L}, \boldsymbol{a}_{m,R} \in \mathbb{R}^d$ for $m \in \mathcal{M}$ and the matrices $\boldsymbol{V}_v$ for $v \in \mathcal{V}$. In fact the prior distribution is defined over partitions of objects and partitions of relations via the Chinese Restaurant Process. For a

given partition, each cluster C of entity mentions or relation verbs has a different mean and covariance, which implies that objects within a cluster have similar distributed representations. Therefore the joint distribution of data and model variables is given as

$$\mathbb{P}(\mathcal{T}, \theta, \mathcal{C}, \alpha, \gamma) = \mathbb{P}(\mathcal{T}|\theta)\mathbb{P}(\theta|\mathcal{C}, \alpha)\mathbb{P}(\mathcal{C}|\gamma)\mathbb{P}(\alpha, \gamma) \tag{2.25}$$

where the observed data $\mathcal{T}$ is a set of triples with ground truth value 1; the variable $\mathcal{C} = \{\mathcal{C}_m, \mathcal{C}_v\}$ contains the cluster assignments (partitions) of the objects and the relations; the model variables $\theta = \{\boldsymbol{a}_{m,L}, \boldsymbol{a}_{m,R}, \boldsymbol{V}_v\}$ for $m \in \mathcal{M}$ and $vin\mathcal{V}$ consists of the distributed representations of the objects and the relations and $\{\alpha, \gamma\}$ are model parameters.

Finally they specify the distributions in 2.25 and infer the model parameters via Gibbs Sampling.

2.9.4 Matrix Factorization based Approach

The most similar embedding method in the literature is the universal schema approach in [10] based on matrix factorization Their goal is to develop a model that can estimate, for a given relation a triplet (m_s, m_o, v) if the relationship holds, that is for some binary truth function T, $T(m_s, m_o, v) = 1$. In order to do so, they define another representation of the relationship triplets $(m_s, m_o, v) \in \mathcal{T}$ as

$$\mathcal{O} = \{< t, v > \mid t = (m_s, m_o) \text{ for all } (m_s, m_o, v) \in \mathcal{T}\},$$

and adopt a matrix factorization based approach. Specifically, like [3], they define a likelihood for each entity triplet

$$\mathbb{P}(T(< t, v >) = 1|\theta_{t,v}) = \frac{1}{1 + \exp(-\theta_{t,v})}, \tag{2.26}$$

where $\theta_{t,v}$ is defined as a superposition of different terms specified as follows.

Letting $\boldsymbol{a}_t$ and $\boldsymbol{a}_v$ represent latent vectors in a lower dimensional space for the entity pair t and relation verb v, the first component of θ can be defined as the dot product of the two latent representations, that is

$$\theta^F_{<t,v>} = \langle \boldsymbol{a}_t, \boldsymbol{a}_v \rangle \tag{2.27}$$

Notice that there is no per-relation weight in the dot product above. We will remark that the multiplicity of the relations can be incorporated here as weights for each term.

Inspired by the item-based collaborative filtering idea [42], the second component is defined as

$$\theta^N_{<t,v>} = \sum_{<t,v'>\in\mathcal{T}} w_{v,v'}, \tag{2.28}$$

where $w_{v,v'}$ corresponds to a directed association strength between relations v and v'.

Finally, given a relationship triplet (m_s, m_o, v), the third component of θ is comprised of two terms that account for latent representations for each entity as well as separate latent representation for the right hand side and left hand side effect of the relation verb v. Specifically,

$$\theta^E_{(m_s,m_o,v)} = \langle \boldsymbol{a}_{m_s}, \boldsymbol{a}_{v,L} \rangle + \langle \boldsymbol{a}_{m_o}, \boldsymbol{a}_{v,R} \rangle, \tag{2.29}$$

where $\boldsymbol{a}_{m_s}$ and $\boldsymbol{a}_{m_o}$ are respectively latent representations for the left and right entity mentions; and $\boldsymbol{a}_{v,L}$ and $\boldsymbol{a}_{v,R}$ are representations for the right hand side and left hand side.

Finally by setting

$$\theta = \theta^F + \theta^N + \theta^E,$$

they aim to learn the model, parametrized by weights and latent component vectors, using a maximum likelihood approach. In order to avoid only learning positive facts, one has to bring in negative samples, which is usually done through sampling a set of unobserved facts similar to distant supervision approaches. In [10], they define the objective as follows. For every observed assertion $< t^+, v > \in \mathcal{T}$, they choose all(or sample a number of) unobserved assertions $< t^-, v > \notin \mathcal{T}$. Since the objective is to have

$$\mathbb{P}(T(< t^-, v >) = 1) < \mathbb{P}(T(< t^+, v >) = 1)$$

for all relation verbs $v \in \mathcal{V}$ and positive and negative assertions, they aim to maximize the following objective

$$\sum_{<t^+,v>\in\mathcal{T}} \sum_{<t^-,v>\notin\mathcal{T}} \log(\sigma(\theta_{t^+,v} - \theta_{t^-,v})) \tag{2.30}$$

This objective can simply be learned using (stochastic) gradient descent.

2.9.5 Future Directions

In nearly all the prior work that we described in the previous sub-section in great details, the embedded representations are used towards a prediction task as to whether a given relation triplet (m_s, m_o, v) is a valid relation. In the entity resolution problem, however, we are interested in clustering similar mentions based on their corresponding distributed representations. There is, however, a fundamental trade-off between the predictive ability of a model and its interoperability, which translates to structural properties of the embeddings, which in turn ease the clustering task. The problem of interest that we seek to investigate in the future work is to take advantage of the efficiency of the neural embedding approaches, both in terms of the freedom to have nonlinear scoring function that leads to a better predictive performance and the scalability of the learning procedure due to its online nature, while brining some structure to the learned embeddings.

For example, given a neural-embedding approach with parameters

$$\theta = \{\{\boldsymbol{W}_v\}_{v\in\mathcal{V}}, \{\boldsymbol{a}_m\}_{m\in\mathcal{M}}\},$$

and some scoring function $g(m_s, m_o, v)$ for a relation triplet, and a set of ground truth relation triplet $\mathcal{T}$, we can learn the parameters of the model with an appropriately regularized optimization objective in the following form

$$\mathcal{J}(\theta) = \sum_{(m_s,m_o,v)\in\mathcal{T}} \sum_{(m'_s,m'_o,v)\in\mathcal{T}^C} \max(0, 1 - g(m_s, m_o, v) + g(m'_s, m'_o, v)) + \lambda\rho(\theta), \qquad (2.31)$$

where $\rho(\theta)$ is an appropriately chosen regularization function to impose a desired property on the embeddings. For example an ℓ_1 norm regularization on the entity mention embeddings can lead to sparse representations for the entity embeddings as desired in our application.

CHAPTER 3

Structured Matrix Completion

3.1 Introduction

3.1.1 Motivation

As discussed in the previous chapter, our approach to clustering entity mentions with similar contextual role is based on finding distributed representations(embeddings) for entity mentions that carry the relational information in the extracted relation triplets. We then proposed to identify such representations via reconstructing the left and right entity relation matrices, which capture the role of each mention as a subject and object, receptively. We argued that an approach based on explicit factorization of the ground truth matrices that takes into account the sparsity of the resulting factors serves this purpose. We then argued to that containing the factors to be non-negative and formulating the factorization problem as Non-Negative Matrix Factorization(NMF) can lead to sparse representations for the embeddings. In this chapter, we study the problem of structured matrix factorization and its non-negativity constrained variant (NMF problem) in details. We start our formal discussion by noting the structural assumptions on the ground truth matrices.

Traditionally, the only structural assumption that is taken into account in matrix factorization problems is the low rank property of the ground truth matrix, which is explicitly modeled in the factorization objective. However, in many practical problems, specifically in the context of relational learning, the ground truth matrices meet stronger structural properties, for example sparsity. Moreover, we are interested in promoting a structure geared towards the clustering task that we are primarily interested in, on the factor matrices.

In order to impose structures such as sparsity or smoothness on the factor matrices, there are several approaches that impose a norm constraint or a regularization penalty on the factor matrices along with the approximation loss [43–45]. Another commonly used approach to obtain sparse and interpretable factor matrices is to consider a non-negativity constraint on the factor matrices. This approach popularized by the observations in [46] that such constraints lead to a part-based decomposition of a dataset of facial images.

In this chapter we study the non-negative matrix factorization problem in detail and propose a new exterior point method to solve this problem based on the recent results in characterizing the global landscape of the unconstrained matrix factorization problem [47]. Our proposed method leads to sparser representations than the state of the art NMF methods while maintaining a competitive or better prediction accuracy. The factor matrices obtained from solving NMF will in turn be used as low dimensional representations for a subsequent clustering task. before, we start our formal discussion, we define some notations that we will use throughout this chapter.

3.1.2 Notation

Vectors and matrices are denoted by bold-face lowercase and uppercase letters, respectively. The i-th row and j-th column of a matrix $\boldsymbol{X}$ are represented by $\boldsymbol{X}_{i*}$ and $\boldsymbol{X}_{*j}$, respectively. By the same token, the set of rows and columns of a matrix $\boldsymbol{X}$ corresponding to an index set Ω are represented by $\boldsymbol{X}_{\Omega *}$ and $\boldsymbol{X}_{*\Omega}$, respectively. The Euclidean norm of any vector $\boldsymbol{v}$ is denoted by $\|\boldsymbol{v}\|_2$. For any arbitrary matrix $\boldsymbol{X}$, we use $\|\boldsymbol{X}\|_2$ and $\|\boldsymbol{X}\|_F$ to denote its spectral and Frobenius norms, respectively. Moreover, given a positive number $p \geq 1$, $\|\boldsymbol{X}\|_{2,p}$ will denote the row-wise $\ell_{2,p}$ norm of $\boldsymbol{X}$, which is defined as $\|\boldsymbol{X}\|_{2,p} = (\sum_{i=1}^{p} \|\boldsymbol{X}_{i*}\|_2^p)^{1/p}$. Assuming that $\boldsymbol{X}$ is a rank-r matrix, $\sigma_1(\boldsymbol{X})$ and $\sigma_r(\boldsymbol{X})$ will denote its maximum and minimum singular values, respectively. We denote a function with two matrix arguments, $f(\boldsymbol{U}, \boldsymbol{V})$, with $\boldsymbol{U} \in \mathbb{R}^{n\times m}$ and $\boldsymbol{U} \in \mathbb{R}^{n\times m}$, also by the lifted variable $f(\boldsymbol{W})$, where $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}$. For a scalar function $f(\cdot)$ with a matrix variable $\boldsymbol{X} \in \mathbb{R}^{n\times m}$, the gradient is defined as $\nabla f(\boldsymbol{X})[i,j] = \frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}[i,j]}$, for $i \in \{1, 2, \cdots, n\}$ and $j \in \{1, 2, \cdots, m\}$. The Hessian of $f(\cdot)$

can be expressed in multiple ways: $\nabla^2 f(\boldsymbol{X})$ can be viewed as an $mn \times mn$ matrix with $[\nabla^2 f(\boldsymbol{X})](i,j) = \frac{\partial^2 f(\boldsymbol{X})}{\partial x[i] \partial x[j]}$, for $i, j \in \{1, 2, \cdots, mn\}$, where $x[i]$ is the i-th coordinate of the vectorization of $\boldsymbol{X}$. Equivalently, it can be expressed as the bilinear form $[\nabla^2 f(\boldsymbol{X})](\boldsymbol{A}, \boldsymbol{B}) = \sum_{i,j,k,l} \frac{\partial^2 f(\boldsymbol{X})}{\partial \boldsymbol{X}[i,j] \partial \boldsymbol{X}[k,l]} \boldsymbol{A}[i,j] \boldsymbol{B}[k,l]$, for any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times m}$. For a scalar function with a vector argument $f(\boldsymbol{x})$, the directional derivative of $f(\cdot)$ in the direction of the column vector $\boldsymbol{u}$ is defined as $D_{\boldsymbol{u}} f = \langle \boldsymbol{u}, \nabla f \rangle$. Moreover $D_{\boldsymbol{v}}(D_{\boldsymbol{u}} f) = D_{\boldsymbol{v}}(\langle \boldsymbol{u}, \nabla f \rangle) = \boldsymbol{v}^T \nabla^2 f \boldsymbol{u}$; thereby $[\nabla^2 f(\boldsymbol{x})](\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{v}^T \nabla^2 f \boldsymbol{u} = D_{\boldsymbol{v}}(D_{\boldsymbol{u}} f)$.

3.2 Non-Negative Matrix Factorization: A Brief Overview

Consider the regularized matrix factorization objective with the non-negativity constraint on the factors, that is

$$\begin{aligned} \min_{\substack{\boldsymbol{U} \in \mathbb{R}^{n \times r} \\ \boldsymbol{V} \in \mathbb{R}^{m \times r}}} \quad & \mathcal{L}(\boldsymbol{X}^*, \boldsymbol{U}\boldsymbol{V}^T) + \rho(\boldsymbol{U}, \boldsymbol{V}), \\ \text{subject to} \quad & \boldsymbol{U} \geq 0 \ \text{ and } \ \boldsymbol{V} \geq 0, \end{aligned} \tag{3.1}$$

where $\boldsymbol{X}^*$ is the ground truth matrix, to be estimated by multiplication of the low dimensional factor matrices $\boldsymbol{U} \in \mathbb{R}^{n \times r}$ and $\boldsymbol{V} \in \mathbb{R}^{m \times r}$. The choice of the loss function $\mathcal{L}(\cdot, \cdot)$ depends on the prior knowledge about the data and the desired statistical interpretation of the approximation noise.

A popular choice, as discussed in the unconstrained matrix completion problem, is to adopt Frobenius norm, which due to its quadratic form and strong convexity leads to more tractable optimization problems, that is

$$\mathcal{L}(\boldsymbol{X}^*, \boldsymbol{U}\boldsymbol{V}^T) = \frac{1}{2} \left\| \boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{X}^* \right\|_F^2.$$

Statistically, minimizing this loss function can be seen as a maximum likelihood estimator where the approximation error is due to additive Gaussian noise. In other words, the generative model for the data can be expressed as

$$\boldsymbol{X}^* = \boldsymbol{U}^* \boldsymbol{V}^{*T} + \mathcal{E},$$

where $\mathcal{E} \in \mathbb{R}^{m \times n}$ is a matrix with i.i.d Gaussian entries.

A major drawback of this loss function, as we will extensively discuss in a later section is that the gradients of the loss depend on the scales of factor matrices, which leads to technical difficulties in the optimization, including a large number of iterations.

Another widely used loss function in the NMF literature is the Generalized Kullback-Leibler Divergence defined as follows

$$\mathcal{L}(\boldsymbol{X}^*, \boldsymbol{U}\boldsymbol{V}^T) = \boldsymbol{X}^* \odot \ln\left(\boldsymbol{X}^* \oslash \boldsymbol{U}\boldsymbol{V}^T\right) + \boldsymbol{X}^* - \boldsymbol{U}\boldsymbol{V}^T$$

where $\odot$ and $\oslash$ are Hadamard(element-wise) multiplication and division respectively. Minimizing this loss function is equivalent to using Expectation Maximization (EM) algorithm for a maximum likelihood problem on Poisson processes [48].

In our study, we do not consider any regularization function in 3.1, and our objective has the following simple form

$$\begin{aligned} \min_{\substack{\boldsymbol{U}\in\mathbb{R}^{n\times r}\\ \boldsymbol{V}\in\mathbb{R}^{m\times r}}} \quad & \frac{1}{2}\left\|\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{X}^*\right\|_F^2, \\ \text{subject to} \quad & \boldsymbol{U} \geq 0 \ \text{ and } \ \boldsymbol{V} \geq 0, \end{aligned} \tag{3.2}$$

recognizing that the problem does not necessarily admit a unique solution.

Suppose that $(\boldsymbol{U}^*, \boldsymbol{V}^*)$, is an optimal solution for the NMF problem. As discussed in the earlier sections, without constraining the geometry of the problem, any pair $(\boldsymbol{U}^*\boldsymbol{R}, \boldsymbol{V}^*\boldsymbol{R}^{-1})$ will have the same objective value. In order for the transformed pair to be a solution for NMF, $\boldsymbol{U}^*\boldsymbol{R}$, and $\boldsymbol{V}^*\boldsymbol{R}^{-1}$ both have to be non-negative.

If a nonsingular matrix and its inverse are both nonnegative, then the matrix is a generalized permutation. That is, $\boldsymbol{R}$ can be expressed as $\boldsymbol{R} = \boldsymbol{P}\boldsymbol{S}$, where $\boldsymbol{P}$ is a permutation matrix and $\boldsymbol{S}$ is a scaling matrix. Note, however, that for the transformed matrices to be non-negative the transformation matrix $\boldsymbol{R}$ need not to be generalized permutation or even nonnegative. Thus, we cannot conclude that the NMF problem has a unique solution up to permutation and scaling and in this sense the non-negativity constrain alone does not guarantee uniqueness of the solution unless the data satisfies additional structural properties.

A necessary and sufficient condition for the uniqueness of NMF solution is called boundary close condition, detailed discussions about which can be found in [49–51]. Geometrically , the NMF problem can be viewed as finding a simplicial cone encompassing all the data points in the positive orthant. In [52], Vavasis defined exact NMF, where it is assumed that underlying ground truth matrix $\boldsymbol{X}^*$ is rank r matrix with non-negative values and the problem is to determine whether it admits a factorization $\boldsymbol{X}^* = \boldsymbol{U}^*\boldsymbol{V}^{*T}$ for some non-negative matrices $\boldsymbol{U}^* \in \mathbb{R}_+^{n\times r}$ and $\boldsymbol{V}^* \in \mathbb{R}_+^{m\times r}$ or not. By showing the equivalence of this problem with a problem in polyhedral combinatorics, he proves that exact NMF it is NP-hard, although there exists a polynomial-time local search heuristic for it. Tied to the exact NMF problem, is the notion of Non-negative Rank, which refers to the rank r of the ground truth matrix $\boldsymbol{X}^*$ if it admits an exact NMF factorization.

The standard interpretation of the NMF problem is to view data points represented in columns of the ground truth matrix $\boldsymbol{X}^*$ and understand columns of the first factor matrix $\boldsymbol{U}$ as basis vectors or latent components, a linear combination of which with coefficients taken from rows of the factor matrix $\boldsymbol{V}$ can reconstruct the data.

It is in fact due to the non-negativity of the factors that one can interpret basis elements(parts), the columns of $\boldsymbol{U}$, in the same way as the data and interpret weights, in the rows of $\boldsymbol{V}$ as activation coefficients.

In our study, however, we view the ground truth matrix as a relational mapping between two sets of objects represented in rows and columns respectively. That is, each row object(entity mention) has a representation in terms of column objects(relation mentions) and each column object has a representation in terms of row objects. Each factor in turn yields lower dimensional representations for one of the sets of objects, which together reconstruct the observed relational matrix and can in turn be used for finding similar objects in each set.

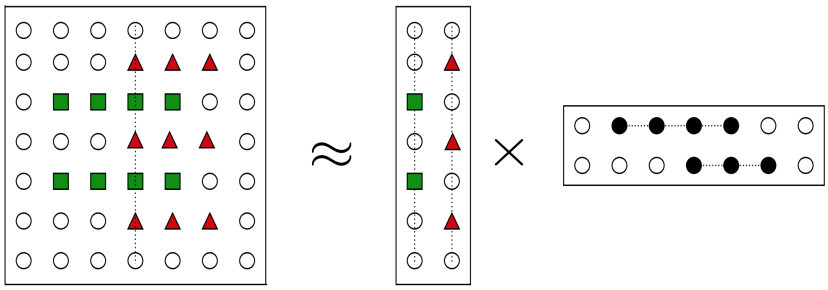

Figure 3.1: NMF viewed as a data reduction scheme with part-based interpretations and a dimensionality reduction embedding technique

To start the discussion on solving the NMF problem , let us first consider a compact form by defining the vertical concatenation of the factor matrices $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}$ as a new variable. In this light, can be written as

$$\begin{aligned} \min_{\boldsymbol{W} \in \mathbb{R}^{(n+m)\times r}} \quad & f(\boldsymbol{W}), \\ \text{subject to} \quad & \boldsymbol{W} \geq 0, \end{aligned} \tag{3.3}$$

where $f(\boldsymbol{W}) = f(\boldsymbol{U}, \boldsymbol{V}) = \frac{1}{2} \left\| \boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{X}^* \right\|_F^2$. In order to get a general idea on the behavior of the optimization problem 3.3, we first consider a similar one dimensional problem

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{x} \geq 0, \end{aligned}$$

The first order optimality conditions(a.k.a. Karush–Kuhn–Tucker (KKT) conditions) for this problem are

$$\begin{aligned} x_i &\geq 0 \\ \nabla f_i(\boldsymbol{x}) &\geq 0 \\ x_i f_i(\boldsymbol{x}) &= 0, \end{aligned}$$

for all $1 \leq i \leq n$. By the same the K. K. T. conditions the non-negative matrix factorization

can be expressed as

$$\begin{aligned} \nabla f(\boldsymbol{W}) &\geq 0 \\ \boldsymbol{W} &\geq 0 \\ \nabla f(\boldsymbol{W}) \odot \boldsymbol{W} &= 0, \end{aligned} \tag{3.4}$$

where $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}$, and $\odot$ represents Hadamard product and the gradient of $f(\boldsymbol{W})$ is given as

$$\nabla f(\boldsymbol{W}) = \begin{bmatrix} (\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{X}^*)\boldsymbol{V} \\ (\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{X}^*)^T\boldsymbol{U} \end{bmatrix}$$

This observation partially justifies why the NMF problem inherently leads to sparse solutions.

As opposed to the unconstrained matrix completion problem, NMF does not admit a convexification approach [53]. There is, however, a convexification approach via non-negative nuclear norms to compute lower bounds for the nonnegative rank [54]. Non-negative matrix factorization, like the unconstrained counterpart, is a non-convex problem. The global landscape of this problem, however, is not characterized and it may have several local minima that are not global.

Therefore, iterative methods that sequentially solve the problem for disjoint blocks of variables are deployed. These blocks are chosen such that if the rest of the variables are fixed, the problem turns into a tractable convex problem. This strategy is known as Block Coordinate Descent(BCD) in the context of bound constrained optimization problems.

The most commonly adopted BCD approach to NMF is to take $\boldsymbol{U}$ and $\boldsymbol{V}$ as the underlying blocks and solve

Algorithm 1 Block Coordinate Descent with two blocks

Input: the ground truth matrix $\boldsymbol{X}^*$

initial factors $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{U}_0 \\ \boldsymbol{V}_0 \end{bmatrix}$ with $\boldsymbol{U}_0 \geq 0$ and $\boldsymbol{V}_0 \geq 0$

repeat

solve $\boldsymbol{U}_{t+1} = \arg\min_{\boldsymbol{U} \geq 0} \left\| \boldsymbol{U}\boldsymbol{V}_t^T - \boldsymbol{X}^* \right\|_F^2$

solve $\boldsymbol{V}_{t+1} = \arg\min_{\boldsymbol{V} \geq 0} \left\| \boldsymbol{U}_{t+1}\boldsymbol{V}^T - \boldsymbol{X}^* \right\|_F^2$

Increment t

until some convergence criterion is met

return $(\boldsymbol{U}_t, \boldsymbol{V}_t)$

Each of the sub-problems in the above routine, e.g.

$$\boldsymbol{U}_{t+1} = \arg\min_{\boldsymbol{U} \geq 0} \left\| \boldsymbol{U}\boldsymbol{V}_t^T - \boldsymbol{X}^* \right\|_F^2, \tag{3.5}$$

are known as Non-Negative Least Square, which can be solved either exactly or approximately. If solved exactly, the convergence analysis of the block coordinate descent methods guarantees that this iterative approach converges to a stationary point of the original problem, if each sub-problem has a unique solution [55]. Even if the solution of the sub-problems is not unique, the convergence to a stationary point is guaranteed in [56] for two block problems.

It is worth noting that solving the subproblems 3.5 exactly, which might actually be computationally expensive, does not necessarily guarantee a faster convergence than heuristic iterative approach that efficiently reduces the cost function in each iteration. In fact an alternative iterative approach to solve 3.2, can be summarized simply as

Algorithm 2 Heuristic iterative algorithm with two blocks

Input: the ground truth matrix $\boldsymbol{X}^*$

initial factors $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{U}_0 \\ \boldsymbol{V}_0 \end{bmatrix}$ with $\boldsymbol{U}_0 \geq 0$ and $\boldsymbol{V}_0 \geq 0$

repeat

Find $\boldsymbol{U}_{t+1} \geq 0$ such that $\left\|\boldsymbol{U}_{t+1}\boldsymbol{V}_t^T - \boldsymbol{X}^*\right\|_F^2 \leq \left\|\boldsymbol{U}_t\boldsymbol{V}_t^T - \boldsymbol{X}^*\right\|_F^2$

Find $\boldsymbol{V}_{t+1} \geq 0$ such that $\left\|\boldsymbol{U}_{t+1}\boldsymbol{V}_{t+1}^T - \boldsymbol{X}^*\right\|_F^2 \leq \left\|\boldsymbol{U}_{t+1}\boldsymbol{V}_t^T - \boldsymbol{X}^*\right\|_F^2$

Increment t

until some convergence criterion is met

return $(\boldsymbol{U}_t, \boldsymbol{V}_t)$

The most popular approach to solve 3.2 in the NMF literature, which instantiates algorithm 2 with a heuristic that guarantees the non-negativity of a multiplicative update coefficient for each entry of the factor matrices at each iteration, is the so called multiplicative update rules proposed by Lee and Seung in [46]. This work was the first to draw attentions to NMF as a dimensionality reduction technique with part-based interpretations. The Pseudo code of the multiplicative updates procedure, which we henceforth call NMF-MUL, can be summarized as follows

Algorithm 3 NMF-MUL to Solve NMF

Input: the ground truth matrix $\boldsymbol{X}^*$

initial factors $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{U}_0 \\ \boldsymbol{V}_0 \end{bmatrix}$ with $\boldsymbol{U}_0 \geq 0$ and $\boldsymbol{V}_0 \geq 0$

repeat

Set $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t \odot \left[\boldsymbol{X}^*\boldsymbol{V}_t \oslash \boldsymbol{U}_t\boldsymbol{V}_t^T\boldsymbol{V}_t\right]$

Set $\boldsymbol{V}_{t+1} = \boldsymbol{V}_t \odot \left[\boldsymbol{X}^{*T}\boldsymbol{U}_t \oslash \boldsymbol{V}_t\boldsymbol{U}_t^T\boldsymbol{U}_t\right]$

Increment t

until some convergence criterion is met

return $(\boldsymbol{U}_t, \boldsymbol{V}_t)$

Lee and Seung proved that Algorithm 3 is indeed an instance of the meta-algorithm 2

and claimed that the sequence of the solutions converges to a stationary point. It is later shown [57], however, that the fixed point of the algorithm 3 is not necessarily a stationary point, i.e. it does not necessary meet all the K.K.T conditions 3.4.

Lee and Seung also mention that their algorithm is equivalent to the following gradient updates

$$\begin{aligned} \boldsymbol{U}_{t+1} &= \boldsymbol{U}_t - \left[\boldsymbol{U}_t \oslash \boldsymbol{U}_t \boldsymbol{V}_t^T \boldsymbol{V}_t\right] \odot \nabla_U f(\boldsymbol{U}_t, \boldsymbol{V}_t) \\ \boldsymbol{V}_{t+1} &= \boldsymbol{V}_t - \left[\boldsymbol{V}_t \oslash \boldsymbol{V}_t \boldsymbol{U}_t^T \boldsymbol{U}_t\right] \odot \nabla_V f(\boldsymbol{U}_t, \boldsymbol{V}_t). \end{aligned} \tag{3.6}$$

A modified form of the above update rules were later used in [58] to prove convergence to a stationary point, which we refer to as NMF-GRAD. In fact, the updates proposed by Lin [58] are summarized in the following pseudo-code.

Algorithm 4 NMF-GRAD algorithm to solve NMF

Input: the ground truth matrix $\boldsymbol{X}^*$

initial factors $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{U}_0 \\ \boldsymbol{V}_0 \end{bmatrix}$ with $\boldsymbol{U}_0 \geq 0$ and $\boldsymbol{V}_0 \geq 0$

positive constants $\sigma > 0$ and $\delta > 0$

repeat

Set $\bar{\boldsymbol{W}}_t[i,j] = \begin{cases} \boldsymbol{W}_t[i,j] & \text{if } \nabla_{\boldsymbol{W}} f(\boldsymbol{W}_t)[i,j] \geq 0 \\ \max(\sigma, \boldsymbol{W}_t[i,j]) & \text{if } \nabla_{\boldsymbol{W}} f(\boldsymbol{W}_t)[i,j] < 0 \end{cases}$

for $1 \leq i \leq m+n$, $1 \leq j \leq r$

Set $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \left\{\bar{\boldsymbol{U}}_t \oslash [\bar{\boldsymbol{U}}_t \boldsymbol{V}_t^T \boldsymbol{V}_t + \delta]\right\} \odot \nabla_{\boldsymbol{U}} f(\boldsymbol{U}_t, \boldsymbol{V}_t)$

Set $\boldsymbol{V}_{t+1} = \boldsymbol{V}_t - \left\{\bar{\boldsymbol{V}}_t \oslash [\bar{\boldsymbol{V}}_t \boldsymbol{U}_t^T \boldsymbol{U}_t + \delta]\right\} \odot \nabla_{\boldsymbol{V}} f(\boldsymbol{U}_t, \boldsymbol{V}_t)$

Increment t

until $(\boldsymbol{U}_t, \boldsymbol{V}_t)$ is a stationary point

return $(\boldsymbol{U}_t, \boldsymbol{V}_t)$

There is huge literature on developing algorithms to solve NMF that we extensively discuss in the appendix of this section. The proposed algorithms either improve the convergence behavior of the NMF problem, for example guaranteeing convergence to a stationary point or solving the sub-problems 3.5 to a better precision or encourage certain structures in the

solution. In the next section, we propose a new algorithm for solving NMF that has improved sparsity properties over the popular NMF-MUL and NMF-GRAD algorithms and even has a slightly better convergence behavior, as verified by our experimental results in the next chapter.

3.3 An Exterior Point Method for solving NMF

The premise of the non-negativity constraint in the NMF formulation is to serve as a surrogate to sparsity in the factor matrices. The intuition behind this can be explained by considering the K.K.T conditions and noting that the solutions should lie on the boundary of the region. In other words, considering the geometry of the unconstrained matrix factorization problem, the solution to the NMF problem tends to converge to a stationary point of the unconstrained problem except that it would hit the boundaries of the feasible region.

In the light of this intuition, it is interesting to explore optimization methods that build an NMF solution based on the stationary solutions for the unconstrained problem, based on SVD, as characterized in Theorem 3.4.1 from [47] that is stated in the appendix section of this chapter. In fact, one of the most common ideas in theoretical analysis of the non-convex procedures is to start with a careful initialization that is already close to optimum. It is shown [59] that after a clever initialization the problem is effectively strongly-convex, implying that the problem be analyzed by standard convex optimization techniques.

The idea of using the optimal SVD solution for the unconstrained matrix factorization to construct an initialization point for the NMF problem was first introduced in [60]. The construction of the initial point in the feasible region is based on taking the positive part of the products of the constituent right and left singular matrices. Our work, however, is based on the recent insights on the global optimization landscape of the unconstrained problem [47] and the intuition that the solutions of the NMF problem lie close to the boundary of the positive orthant. Figure 3.2 demonstrates our approach for constructing an initialization point for the NMF algorithm that lies in the positive orthant. It is based on finding a suitably rotated stationary point of the unconstrained problem that is close to the positive

orthant, in a sense defined later on, and then moving the obtained stationary point to the feasible region(positive orthant), using a simple gradient method with constant step size. In the following, we describe each step in details.

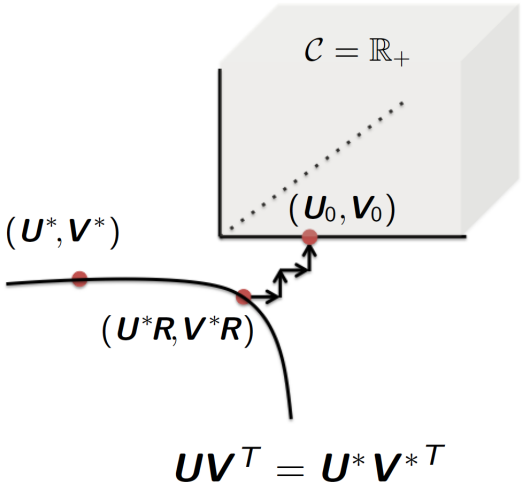

Figure 3.2: Illustration of the exterior point initialization scheme

3.3.1 Closets Optimal Solution to the Positive Orthant

Given a stationary point for the unconstrained factorization problem

$$\min_{\substack{\boldsymbol{U}\in\mathbb{R}^{n\times r}\\ \boldsymbol{V}\in\mathbb{R}^{m\times r}}} \quad \left\|\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{X}^*\right\|_F^2,$$

recall that the set of equally-footed factorizations for $\boldsymbol{X}^*$, i.e. all orthogonal transformations of $(\boldsymbol{U}^*, \boldsymbol{V}^*)$, as defined in (3.21)

$$\mathcal{X}^* := \{(\boldsymbol{U}, \boldsymbol{V}) | \boldsymbol{U} = \boldsymbol{U}^*\boldsymbol{R} \text{ and } \boldsymbol{V} = \boldsymbol{V}^*\boldsymbol{R}, \text{ for } \boldsymbol{R} \in \mathcal{O}_r\},$$

with $\mathcal{O}_r := \{\boldsymbol{O} \in \mathbb{R}^{r\times r} : \boldsymbol{O}^T\boldsymbol{O} = \boldsymbol{I}_r\}$, are also stationary.

In order to find a suitably close solution to the positive orthant, we aim to find a pair $(\boldsymbol{U}, \boldsymbol{V})$, that is also a stationary point of the unconstrained problem, with the least sum-aggregated coordinate-wise distance to the positive orthant. Let $(\boldsymbol{U}^*, \boldsymbol{V}^*)$ be a stationary point of the unconstrained problem, as characterized in Theorem 3.4.1, based on the SVD of the ground

truth matrix $\boldsymbol{X}^*$. Recognizing that all rotations of the stationary points are also stationary, we can find the closest solution to the feasible region, in the sense of total element-wise distance, by solving the following the following optimization problem

$$\begin{aligned} \min_{\substack{\boldsymbol{\Upsilon_U}\in\mathbb{R}^{n\times r}\\ \boldsymbol{\Upsilon_V}\in\mathbb{R}^{m\times r}\\ \boldsymbol{R}\in\mathbb{R}^{r\times r}}} \quad & \mathbf{1}_n^T\boldsymbol{\Upsilon_U}\mathbf{1}_r + \mathbf{1}_m^T\boldsymbol{\Upsilon_V}\mathbf{1}_r, \\ \text{subject to} \quad & \boldsymbol{U}^*\boldsymbol{R} + \boldsymbol{\Upsilon_U} \geq \mathbf{0} \text{ and } \boldsymbol{V}^*\boldsymbol{R} + \boldsymbol{\Upsilon_V} \geq \mathbf{0} \\ & \boldsymbol{\Upsilon_U} \geq \mathbf{0} \text{ and } \boldsymbol{\Upsilon_V} \geq \mathbf{0} \\ & \boldsymbol{R} \in \mathcal{O}_r \end{aligned} \tag{3.7}$$

Note that in the optimization problem (3.7), the utility function is linear in the variables, as well as all the constraints except the last one, concerning orthogonality of the transformation matrix $\boldsymbol{R}$, i.e.

$$\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}_r,$$

which is quadratic. In order to solve (3.7), we propose a sequential greedy approach that solves a set of linear programs for columns of $\boldsymbol{R}$ sequentially. We recognize this greedy solution is not necessarily optimal, however, we adopt this approximate solution for the next steps.

In the following, we describe the LP approximation to solve (3.7) sequentially for columns of $\boldsymbol{R}$. Let $\xi_j^{\boldsymbol{U}} := \boldsymbol{\Upsilon_U}[:, j] \in \mathbb{R}^n$, $\xi_j^{\boldsymbol{V}} := \boldsymbol{\Upsilon_V}[:, j] \in \mathbb{R}^m$, and $\boldsymbol{r}_j := \boldsymbol{R}[:, j] \in \mathbb{R}^r$ represent the j-th column of $\boldsymbol{\Upsilon_U}$ and $\boldsymbol{\Upsilon_V}$ respectively. Note that the utility function is separable on columns of $\boldsymbol{\Upsilon_U}$ and $\boldsymbol{\Upsilon_V}$. Also, in the first set of constraints, the $\boldsymbol{r}_j$, the j-th column of $\boldsymbol{R}$ is only related to the corresponding columns in $\boldsymbol{\Upsilon_U}$ and $\boldsymbol{\Upsilon_V}$; that is $\xi_j^{\boldsymbol{U}}$ and $\xi_j^{\boldsymbol{V}}$. In order to account for the quadratic constraint concerning orthogonality of $\boldsymbol{R}$, when solving for the j-th column of $\boldsymbol{R}$ we only have to make sure that it is orthogonal to all the previously obtained columns. Finally, we have to add a regularity constraint to avoid trivial solutions, like an all zeros vector. In order to do so, we generate a fixed random vector $\boldsymbol{s} \in \mathbb{R}^r$ and constrain each column of $\boldsymbol{R}$ to have a positive inner product with that vector.

The following linear program that is solved sequentially for $j \in \{1, 2, \cdots, r\}$ captures our sequential formulation.

$$\begin{aligned} \min_{\substack{\xi_j^{\boldsymbol{U}} \in \mathbb{R}^n \\ \xi_j^{\boldsymbol{V}} \in \mathbb{R}^m \\ \boldsymbol{r}_j \in \mathbb{R}^r}} \quad & \mathbf{1}_n^T \xi_j^{\boldsymbol{U}} + \mathbf{1}_m^T \xi_j^{\boldsymbol{V}}, \\ \text{subject to} \quad & \boldsymbol{U}^* \boldsymbol{r}_j + \xi_j^{\boldsymbol{U}} \geq \mathbf{0} \text{ and } \boldsymbol{V}^* \boldsymbol{r}_j + \xi_j^{\boldsymbol{V}} \geq \mathbf{0} \\ & \xi_j^{\boldsymbol{U}} \geq \mathbf{0} \text{ and } \xi_j^{\boldsymbol{V}} \geq \mathbf{0} \\ & \boldsymbol{r}_j^T \boldsymbol{s} > 0 \\ & \boldsymbol{r}_j^T \boldsymbol{r}_i = 0 \text{ for all } i < j \end{aligned} \tag{3.8}$$

Note that in formulating the optimization problem to find an optimal factor pair close to the positive orthant, we could adopt another measure of closeness. For example, One could aim for the closest solution to the positive orthant in the sense of maximum coordinate wise distance. The corresponding set of linear programs in our sequential approach would look the following

$$\begin{aligned} \min_{\substack{\xi_j^{\boldsymbol{U}} \in \mathbb{R}^n \\ \xi_j^{\boldsymbol{V}} \in \mathbb{R}^m \\ \boldsymbol{r}_j \in \mathbb{R}^r}} \quad & t_{\boldsymbol{U}} + t_{\boldsymbol{V}}, \\ \text{subject to} \quad & \boldsymbol{U}^* \boldsymbol{r}_j + \xi_j^{\boldsymbol{U}} \geq \mathbf{0} \text{ and } \boldsymbol{V}^* \boldsymbol{r}_j + \xi_j^{\boldsymbol{V}} \geq \mathbf{0} \\ & t_{\boldsymbol{U}} \geq \xi_j^{\boldsymbol{U}} \geq \mathbf{0} \text{ and } t_{\boldsymbol{V}} \geq \xi_j^{\boldsymbol{V}} \geq \mathbf{0} \\ & \boldsymbol{r}_j^T \boldsymbol{s} > 0 \\ & \boldsymbol{r}_j^T \boldsymbol{r}_i = 0 \text{ for all } i < j \end{aligned} \tag{3.9}$$

3.3.2 Projection into the Positive Orthant

Let us first introduce a simple ascent algorithm, to perturb a variable based on the direction of its corresponding gradient vector, which will appear frequently throughout this section.

Algorithm 5 Flat-Ascent/Descent Algorithm

Input: $\boldsymbol{W} \in \mathbb{R}^{(m+n)\times r}$

$\nabla f(\boldsymbol{W}) \in \mathbb{R}^{(m+n)\times r}$

ascent constant $\delta \in \mathbb{R}$

Set $\bar{\boldsymbol{W}}[i,j] = \begin{cases} \boldsymbol{W}[i,j] - \min\{\delta, \boldsymbol{W}[i,j]\} & \text{if } \nabla f(\boldsymbol{W})[i,j] \geq 0 \text{ and } \boldsymbol{W}[i,j] \geq 0 \\ \boldsymbol{W}[i,j] + \delta & \text{otherwise} \end{cases}$

for $1 \leq i \leq m+n$, $1 \leq j \leq r$

return $\widehat{\boldsymbol{W}}$

In words, this flat ascent algorithm takes constant steps in the opposite direction of the gradient when the point lies in the feasible region. When a point lies outside of the feasible region, the algorithm pushes the point into the feasible region with constant steps, regardless of the direction of the gradient.

Once a close enough optimal point to the positive orthant is found, by running a few iterations of the flat ascent algorithm, we can obtain a projection of that optimal point onto the positive orthant. As discussed earlier, by considering the K.K.T conditions of the NMF problem, such a point should have a reasonably small prediction error.

After a small perturbation to push the point, which lies on the boundary of the feasible region, inside the positive orthant, it can then be used as good initialization for an NMF algorithm. The following pseudo-code describes the procedure to project an optimal solution from outside the feasible region into the positive orthant via flat gradient updates.

Algorithm 6 Project into Positive Orthant via Flat Ascent

Input: the ground truth matrix $\boldsymbol{X}^*$

initial factors $\boldsymbol{W}_0 = \begin{bmatrix} \boldsymbol{U}_0 \\ \boldsymbol{V}_0 \end{bmatrix}$

Set $\delta = \min_{i,j} |\boldsymbol{W}_0[i,j]|$

repeat

Compute $\nabla f(\boldsymbol{W}_t) = \begin{bmatrix} \nabla_{\boldsymbol{U}} f \\ \nabla_{\boldsymbol{V}} f \end{bmatrix} = \begin{bmatrix} (\boldsymbol{U}_t \boldsymbol{V}_t^T - \boldsymbol{X}^*)\boldsymbol{V}_t \\ (\boldsymbol{U}_t \boldsymbol{V}_t^T - \boldsymbol{X}^*)^T \boldsymbol{U}_t \end{bmatrix}$

$\boldsymbol{W}_{t+1} = \textbf{Flat-Ascent}(\boldsymbol{W}_t, \nabla f(\boldsymbol{W}_t), \delta)$

Increment t

until $\boldsymbol{W}_t \geq 0$

return $\boldsymbol{W}_t$

3.3.3 Gradient updates on the initial point

Once a good initialization point based on projecting an optimal point from the unconstrained factorization problem onto the boundary of the feasible region is obtained, we apply a few gradient steps to construct an NMF solution from this point.

Virtually any of the gradient algorithms discussed in the previous section can be used to optimize the NMF cost using this initialization. We propose a very simple gradient algorithm with adaptive flat step sizes to obtain a reasonably good NMF solution starting from this initialization. Our algorithm leverages the inherent imbalance in the gradient of the loss function. Consider the gradient of the Frobenius norm loss; that is

$$\nabla f(\boldsymbol{W}) = \begin{bmatrix} (\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{X}^*)\boldsymbol{V} \\ (\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{X}^*)^T \boldsymbol{U} \end{bmatrix}.$$

Clearly if one of the factors is scaled up by some factor $\kappa > 1$ and the other factor is scaled down so that $\boldsymbol{U}\boldsymbol{V}^T$ is fixed, it can immediately observed that this scaling is reflected inversely in the gradient terms corresponding to the other factor. This provides an opportunity to take loner steps by the gradient algorithm. The following figure provides a visualization of this step.

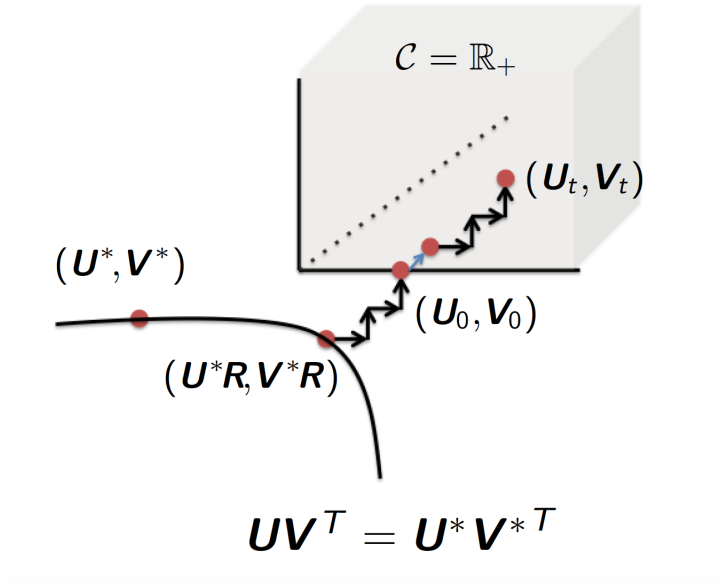

Figure 3.3: Constructing an NMF solution from the obtained initialization

A pseudo-code for the algorithm is provided in the following.

Algorithm 7 Solving NMF with Flat Ascent

Input: the ground truth matrix $\boldsymbol{X}^*$

initial factors $\boldsymbol{W}_0 = \begin{bmatrix} \boldsymbol{U}_0 \\ \boldsymbol{V}_0 \end{bmatrix}$

initial ascent constant δ

Perturbation constant ϵ

Scaling constant κ

Number of outer iterations n_1

Number of inner iterations n_2

Step parameters β, σ

Set $\widehat{\boldsymbol{W}} = \boldsymbol{W}_0$

for $i \in \{1, \ldots, n_1\}$ **do**

  if $i \bmod 2 = 0$ **then**

    $s = \kappa$

  else

    $s = \frac{1}{\kappa}$

  end if

  $\boldsymbol{W}_0 = \textbf{PerturbScale}(\widehat{\boldsymbol{W}}, \epsilon, s)$

  $\delta_0 = \delta$

  for $t \in \{0, \ldots, n_2 - 1\}$ **do**

    Set $\boldsymbol{W}_{t+1} = \textbf{Flat-Descent}(\boldsymbol{W}_t, \nabla f(\boldsymbol{W}_t), \delta_t)$,

      where $\delta_t = \delta_{t-1}\beta^M$, with M being the smallest integer that

      $f(\boldsymbol{W}_{t+1}) - f(\boldsymbol{W}_t) \leq \sigma \nabla f(\boldsymbol{W}_t)^T (\boldsymbol{W}_{t+1} - \boldsymbol{W}_t)$,

  end for

  $\widehat{\boldsymbol{W}} = \boldsymbol{W}_{n_2}$

end for

return $\widehat{\boldsymbol{W}}$

In the above algorithm the perturb and scale step refers to simple operations of thresholding the variable from below to the given value ϵ, and scaling the factors U and V by s and

$\frac{1}{s}$ respectively. Note that in the algorithm, the ascent parameter is chosen in such a way to guarantee a function decrease by a certain amount proportional to the gradient. This guarantees the convergence of the algorithm to a stationary point using standard first order arguments.

In the next chapter, we evaluate the performance of the proposed exterior point method algorithm along with some of the other popular NMF optimization computationally on real data.

3.4 Appendix

3.4.1 Global Landscape of the Unconstrained Matrix Factorization

Let us consider the regularized matrix factorization problem

$$\min_{\substack{\boldsymbol{U}\in\mathbb{R}^{n\times r}\\ \boldsymbol{V}\in\mathbb{R}^{m\times r}}} \quad \mathcal{L}(\boldsymbol{X}^*, \boldsymbol{U}\boldsymbol{V}^T) + \rho(\boldsymbol{U}, \boldsymbol{V}), \tag{3.10}$$

for a low rank matrix $\boldsymbol{X}^*$, where $\mathcal{L}(\boldsymbol{X}^*, \boldsymbol{U}\boldsymbol{V}^T)$ is a strongly convex loss function in $\boldsymbol{U}$ and $\boldsymbol{V}$ individually and $\rho(\boldsymbol{U}, \boldsymbol{V})$ is a regularization term that is convex in both $\boldsymbol{U}$ and $\boldsymbol{V}$. Specifically, let us focus on the sum of squared entry-wise estimation error

$$\mathcal{L}(\boldsymbol{W}) = \mathcal{L}(\boldsymbol{X}^*, \boldsymbol{U}\boldsymbol{V}^T) = \frac{1}{2}\left\|\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{X}^*\right\|_F^2. \tag{3.11}$$

Let us also define the *balanced* factors

$$\boldsymbol{U}^* := \boldsymbol{A}^*\boldsymbol{\Sigma}^{*1/2} \in \mathbb{R}^{n\times r}, \text{and}$$
$$\boldsymbol{V}^* := \boldsymbol{B}^*\boldsymbol{\Sigma}^{*1/2} \in \mathbb{R}^{m\times r},$$

where

$$\boldsymbol{A}^*\boldsymbol{\Sigma}^*\boldsymbol{B}^{*T}$$

is the (truncated) singular value decomposition (SVD) of the low-rank ground truth matrix $\boldsymbol{X}^*$.

Following this representation, the ground matrix $\boldsymbol{X}^*$ can be represented with a factorization $\boldsymbol{X}^* = \boldsymbol{U}^*\boldsymbol{V}^{*T}$, with $\boldsymbol{U}^* \in \mathbb{R}^{n\times r}$ and $\boldsymbol{V}^* \in \mathbb{R}^{m\times r}$ to reflect the low rank assumption on $\boldsymbol{X}^*$.

We note that for an optimal solution $(\boldsymbol{U}^*, \boldsymbol{V}^*)$ of the un-regularized problem

$$\min_{\substack{\boldsymbol{U}\in\mathbb{R}^{n\times r}\\ \boldsymbol{V}\in\mathbb{R}^{m\times r}}} \quad \frac{1}{2}\left\|\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{X}^*\right\|_F^2, \tag{3.12}$$

any transformation of the form

$$(\boldsymbol{U}^*\boldsymbol{R}_{\boldsymbol{U}}, \boldsymbol{V}^*\boldsymbol{R}_{\boldsymbol{V}})$$

where $\boldsymbol{R}_{\boldsymbol{U}}, \boldsymbol{R}_{\boldsymbol{V}} \in \mathbb{R}^{r\times r}$ and $\boldsymbol{R}_{\boldsymbol{U}}\boldsymbol{R}_{\boldsymbol{V}}^T = \boldsymbol{I}_r$ is also an optimal solution of 3.12. As such, $\boldsymbol{R}_{\boldsymbol{U}} = c\boldsymbol{I}_r$ and $\boldsymbol{R}_{\boldsymbol{V}} = \frac{1}{c}\boldsymbol{I}_r$, where c can be arbitrarily large. To address this imbalance, the standard approach is to add a regularizer to the utility function with to force the difference between the Gram matrices of $\boldsymbol{U}$ and $\boldsymbol{V}$ as small as possible. In particular,

$$\rho(\boldsymbol{W}) = \rho(\boldsymbol{U}, \boldsymbol{V}) = \frac{\lambda}{4}\left\|\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}\right\|_F^2, . \tag{3.13}$$

where $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}$.

Thus, the overall regularized objective that we consider in the rest of this section can be expressed as

$$\min_{\substack{\boldsymbol{U}\in\mathbb{R}^{n\times r}\\ \boldsymbol{V}\in\mathbb{R}^{m\times r}}} \quad \frac{1}{2}\left\|\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{X}^*\right\|_F^2 + \frac{\lambda}{4}\left\|\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}\right\|_F^2, \tag{3.14}$$

which can alternatively be written as

$$\min_{\boldsymbol{W}\in\mathbb{R}^{(n+m)\times r}} f(\boldsymbol{W}) \qquad , \tag{3.15}$$

$f(\boldsymbol{W}) = f(\begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}) = \frac{1}{2}\left\|\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{X}^*\right\|_F^2 + \frac{\lambda}{4}\left\|\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}\right\|_F^2$

The gradient of $f(\boldsymbol{W})$ is then given as

$$\begin{aligned} \nabla_{\boldsymbol{U}} f(\boldsymbol{U}, \boldsymbol{V}) &= (\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{X}^*)\boldsymbol{V} + \lambda\boldsymbol{U}(\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}) \\ \nabla_{\boldsymbol{V}} f(\boldsymbol{U}, \boldsymbol{V}) &= (\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{X}^*)^T\boldsymbol{U} - \lambda\boldsymbol{V}(\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}), \end{aligned} \tag{3.16}$$

which can equivalently be written as

$$\nabla f(\boldsymbol{W}) = \begin{bmatrix} (\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{X}^*)\boldsymbol{V} \\ (\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{X}^*)^T\boldsymbol{U} \end{bmatrix} + \lambda \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^T\boldsymbol{W}, \tag{3.17}$$

where $\widehat{\boldsymbol{W}} = \begin{bmatrix} \boldsymbol{U} \\ -\boldsymbol{V} \end{bmatrix}$.

The Hessian quadrature $[\nabla^2 f(\boldsymbol{W})](\boldsymbol{\Delta}, \boldsymbol{\Delta})$, with $\boldsymbol{\Delta} = \begin{bmatrix} \boldsymbol{\Delta_U} \\ \boldsymbol{\Delta_V} \end{bmatrix} \in \mathbb{R}^{(n+m)\times r}$(where $\boldsymbol{\Delta_U} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{\Delta_V} \in \mathbb{R}^{m\times r}$) can be expressed as [47]

$$[\nabla^2 f(\boldsymbol{W})](\boldsymbol{\Delta}, \boldsymbol{\Delta}) = \left\|\boldsymbol{\Delta_U}\boldsymbol{V}^T + \boldsymbol{U}\boldsymbol{\Delta_V}^T\right\|_F^2 + 2\langle \boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{X}^*, \boldsymbol{\Delta_U}\boldsymbol{\Delta_V}^T\rangle + [\nabla^2\rho(\boldsymbol{W})](\boldsymbol{\Delta}, \boldsymbol{\Delta}), \tag{3.18}$$

where

$$[\nabla^2\rho(\boldsymbol{W})](\boldsymbol{\Delta}, \boldsymbol{\Delta}) = \lambda\langle \widehat{\boldsymbol{W}}\boldsymbol{W}, \widehat{\boldsymbol{\Delta}}^T\boldsymbol{\Delta}\rangle + \lambda\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{\Delta}}^T, \boldsymbol{\Delta}\boldsymbol{W}^T\rangle + \lambda\langle \widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^T, \boldsymbol{\Delta}\boldsymbol{\Delta}^T\rangle. \tag{3.19}$$

Let W^* be a global minimizer of the un-regularized objective. We note that W^* is still a global minimizer of the regularized problem 3.14. The global minimum of $\rho(\boldsymbol{W})$ is 0, which is achieved when $\boldsymbol{U}$ and $\boldsymbol{V}$ have the same Gram matrices, i. e. when $\boldsymbol{W}$ belongs to

$$\mathcal{E} = \left\{\boldsymbol{W} \mid \boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V} = 0\right\}. \tag{3.20}$$

The following lemma from [47] shows that any critical point $\widehat{\boldsymbol{W}}$ of the regularized objective 3.13 belongs to $\mathcal{E}$, meaning that $\boldsymbol{U}$ and $\boldsymbol{V}$ are balanced factors of their product.

Lemma 3.4.1. *[47] Let $\widehat{\boldsymbol{W}}$ be a critical point of $f(\boldsymbol{W}) = \mathcal{L}(\boldsymbol{W}) + \rho(\boldsymbol{W})$, i.e. $\nabla f(\widehat{\boldsymbol{W}}) = 0$; then $\widehat{\boldsymbol{W}} \in \mathcal{E}$.*

We next note that, even by considering a regularized objective, all rotations of an arbitrary pair of factors $(\boldsymbol{U}, \boldsymbol{V}) \in \mathbb{R}^{n\times r} \times \mathbb{R}^{m\times r}$ will attain the same objective value in 3.14. Thus the set of global minimizers of the regularized objective can be expressed as

$$\mathcal{X}^* := \{(\boldsymbol{U}, \boldsymbol{V}) | \boldsymbol{U} = \boldsymbol{U}^*\boldsymbol{R} \text{ and } \boldsymbol{V} = \boldsymbol{V}^*\boldsymbol{R}, \text{ for } \boldsymbol{R} \in \mathcal{O}_r\}, \tag{3.21}$$

where

$$\mathcal{O}_r := \{\boldsymbol{O} \in \mathbb{R}^{r\times r} : \boldsymbol{O}^T\boldsymbol{O} = \boldsymbol{I}_r\} \tag{3.22}$$

In the following we state the results from [47] that fully characterize the optimization landscape of the regularize matrix factorization objective 3.14 In order to characterize th set of stationary points of 3.14, we first note that, without loss of generality, one can only focus on stationary points where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal. To see this, suppose that $\boldsymbol{W}$ is a stationary point of $f(\boldsymbol{W})$. By applying the Gram-Schmidt process, one can construct $\widetilde{\boldsymbol{U}} = \boldsymbol{U}\boldsymbol{R}$ for some $\boldsymbol{R} \in \mathcal{O}_r$ where $\widetilde{\boldsymbol{U}}$ is orthogonal. By the same token, define $\widetilde{\boldsymbol{V}} = \boldsymbol{V}\boldsymbol{R}$. The necessary condition for stationary points in Lemma 3.4.1,$\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{U}}^T = \widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{V}}^T$, is satisfied because $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{V}\boldsymbol{V}^T$. Also, it can be directly shown that $\widetilde{\boldsymbol{W}}$ is a critical point of $f(\boldsymbol{W})$, since

$$\nabla_{\widetilde{\boldsymbol{U}}} f(\widetilde{\boldsymbol{W}}) = \nabla_{\boldsymbol{U}} f(\boldsymbol{W})\boldsymbol{R} = 0$$

and

$$\nabla_{\widetilde{\boldsymbol{V}}} f(\widetilde{\boldsymbol{W}}) = \nabla_{\boldsymbol{V}} f(\boldsymbol{W})\boldsymbol{R} = 0.$$

Moreover, for any $\boldsymbol{\Delta} \in \mathbb{R}^{(n+m)\times r}$, we have

$$[\nabla^2 f(\boldsymbol{W})](\boldsymbol{\Delta}, \boldsymbol{\Delta}) = [\nabla^2 f(\widetilde{\boldsymbol{W}})](\boldsymbol{\Delta}\boldsymbol{R}, \boldsymbol{\Delta}\boldsymbol{R}),$$

implying that the Hessian information is preserved.

The following lemma characterizes all the critical points of $f(\boldsymbol{W})$.

Lemma 3.4.2. *[47] Let* $\boldsymbol{X}^* = \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Psi} = \sum_{i=1}^{r} \boldsymbol{\sigma}_i(\boldsymbol{X}^*)\boldsymbol{\phi}_i\boldsymbol{\psi}_i$ *be the reduced SVD of* $\boldsymbol{X}^*$ *and* $f(\boldsymbol{W})$ *as defined in 3.15. Any* $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}$ *is a critical point of* $f(\boldsymbol{W})$ *if and only if* $\boldsymbol{W} \in \mathcal{C}$*, where*

$$\mathcal{C} := \Bigg\{\boldsymbol{W} = \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \mid \boldsymbol{U} = \boldsymbol{\Phi}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{R} \text{ and } \boldsymbol{V} = \boldsymbol{\Psi}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{R}, \boldsymbol{R} \in \mathcal{O}_r$$
$$\boldsymbol{\Lambda} \text{ is diagonal}, \boldsymbol{\Lambda} \geq \boldsymbol{0}, (\boldsymbol{\Sigma} - \boldsymbol{\Lambda})\boldsymbol{\Sigma} = 0\Bigg\}, \tag{3.23}$$

The above lemma implies that a critical point $\widetilde{\boldsymbol{W}}$ of $f(\boldsymbol{W})$ is such that $\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{U}}$ is a low rank approximation of $\boldsymbol{X}^*$, where the diagonal matrix $\boldsymbol{\Lambda}$ is formed as

$$\lambda_i \in \{0, \sigma_i(\boldsymbol{X}^*)\}, \tag{3.24}$$

for $i \in \{1, 2, \cdots, r\}$. Moreover, the set of optimal solutions is

$$\mathcal{X}^* := \left\{ \boldsymbol{W} = \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \mid \boldsymbol{U} = \boldsymbol{\Phi}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{R} \text{ and } \boldsymbol{V} = \boldsymbol{\Psi}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{R}, \boldsymbol{R} \in \mathcal{O}_r \right\},$$

as defined in 3.21.

Theorem 3.4.1. *[47] Let $f(\boldsymbol{W})$ be as defined in 3.15 and let $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}$ be any critical point of $f(\boldsymbol{W})$, i.e. $\boldsymbol{W} \in \mathcal{C}$. Any $\boldsymbol{W} \in \mathcal{C} \setminus \mathcal{X}^*$ is a strict saddle point of $f(\boldsymbol{W})$ satisfying*

$$\lambda_{min}(\nabla^2 f(\boldsymbol{W})) \leq -\frac{1}{2}\left\|\boldsymbol{W}\boldsymbol{W}^T - \boldsymbol{W}^*\boldsymbol{W}^{*T}\right\|_F^2 \leq -\sigma_r(\boldsymbol{X}^*) \tag{3.25}$$

This theorem implies that $f(\boldsymbol{W})$ has no spurious local minima and obeys the strict saddle property, that is $f(\boldsymbol{W})$ has a directional negative curvature at all of the critical points except local minima, which are in turn shown to be global optima. These two properties imply that gradient methods with random initialization converges to a global minimizer almost surely [61, 62].

Next, we briefly mention two similar results that extend Theorem 3.4.1 to under-parametrized(over-parametrized) matrix completion problem, where the ground truth matrix has a lower(higher) rank than the factor matrices.

For over-parametrized scenario, where $\text{rank}(\boldsymbol{X}^*) \leq r$, we have

Theorem 3.4.2. *[47] Let $\boldsymbol{X}^* = \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Psi} = \sum_{i=1}^{r'} \boldsymbol{\sigma}_i(\boldsymbol{X}^*)\boldsymbol{\phi}_i\boldsymbol{\psi}_i$ be the reduced SVD of $\boldsymbol{X}^*$ with $r' \leq r$ and $f(\boldsymbol{W})$ as defined in 3.15. Any $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}$ is a critical point of $f(\boldsymbol{W})$ if and only if $\boldsymbol{W} \in \mathcal{C}$, where*

$$\mathcal{C} := \left\{ \boldsymbol{W} = \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \mid \boldsymbol{U} = \boldsymbol{\Phi}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{R} \text{ and } \boldsymbol{V} = \boldsymbol{\Psi}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{R}, \boldsymbol{R} \in \mathcal{O}_{r'} \right.$$
$$\left. \boldsymbol{\Lambda} \text{ is diagonal}, \boldsymbol{\Lambda} \geq \boldsymbol{0}, (\boldsymbol{\Sigma} - \boldsymbol{\Lambda})\boldsymbol{\Sigma} = 0 \right\}, \tag{3.26}$$

Moreover, all the local minima (which are also global) belong to the following set

$$\mathcal{X}^* := \Big\{ \boldsymbol{W} = \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \mid \boldsymbol{U} = \boldsymbol{\Phi}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{R} \text{ and } \boldsymbol{V} = \boldsymbol{\Psi}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{R}, \boldsymbol{R} \in \mathcal{O}_{r'} \Big\},$$

and finally, any $\boldsymbol{W} \in \mathcal{C} \setminus \mathcal{X}^*$ is a strict saddle point of $f(\boldsymbol{W})$ satisfying

$$\lambda_{min}(\nabla^2 f(\boldsymbol{W})) \leq -\frac{1}{2} \left\| \boldsymbol{W}\boldsymbol{W}^T - \boldsymbol{W}^*\boldsymbol{W}^{*T} \right\|_F^2 \leq -\sigma_{r'}(\boldsymbol{X}^*) \tag{3.27}$$

Similarly for the under-parametrized scenario, where $\text{rank}(\boldsymbol{X}^*) \geq r$, we have

Theorem 3.4.3. *[47] Let $\boldsymbol{X}^* = \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Psi} = \sum_{i=1}^{r'} \boldsymbol{\sigma}_i(\boldsymbol{X}^*)\boldsymbol{\phi}_i\boldsymbol{\psi}_i$ be the reduced SVD of $\boldsymbol{X}^*$ with $r' \geq r$ and $\sigma_r(\boldsymbol{X}^*) > \sigma_{r+1}(\boldsymbol{X}^*)$; and $f(\boldsymbol{W})$ as defined in 3.15. Any $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}$ is a critical point of $f(\boldsymbol{W})$ if and only if $\boldsymbol{W} \in \mathcal{C}$, where*

$$\mathcal{C} := \Big\{ \boldsymbol{W} = \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \mid \boldsymbol{U} = \boldsymbol{\Phi}_{*\Omega}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{R} \text{ and } \boldsymbol{V} = \boldsymbol{\Psi}_{*\Omega}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{R}, \boldsymbol{\Lambda} = \boldsymbol{\Sigma}[\Omega, \Omega], \\ \boldsymbol{R} \in \mathcal{O}_\ell, \Omega \subset \{1, 2, \cdots, r'\}, |\Omega| = \ell \leq r \Big\}, \tag{3.28}$$

where $\boldsymbol{\Phi}_{*\Omega}$, as defined in the notations, is a sub-matrix of $\boldsymbol{\Phi}$ obtained by keeping the columns indexed by Ω and $\boldsymbol{\Sigma}[\Omega, \Omega]$ is a matrix obtained by taking the elements of $\boldsymbol{\Sigma}$ in rows and columns indexed by Ω. Moreover, all the local minima (which are also global) belong to the following set

$$\mathcal{X}^* := \Big\{ \boldsymbol{W} = \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \mid \boldsymbol{U} = \boldsymbol{\Phi}_{*\underline{r}}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{R} \text{ and } \boldsymbol{V} = \boldsymbol{\Psi}_{*\underline{r}}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{R}, \boldsymbol{\Lambda} = \boldsymbol{\Sigma}[\underline{r}, \underline{r}], \boldsymbol{R} \in \mathcal{O}_r \Big\},$$

where $\underline{r} = \{1, 2, \cdots, r\}$. Finally, any $\boldsymbol{W} \in \mathcal{C} \setminus \mathcal{X}^*$ is a strict saddle point of $f(\boldsymbol{W})$ satisfying

$$\lambda_{min}(\nabla^2 f(\boldsymbol{W})) \leq -(\sigma_r(\boldsymbol{X}^*) - \sigma_{r+1}(\boldsymbol{X}^*)) \tag{3.29}$$

These three theorems fully characterize the optimization landscape of the regularized matrix factorization objective 3.14 for the low rank matrix $\boldsymbol{X}^*$. Our exterior point method relies on this characterization for constructing the initialization point from a stationary point of the unconstrained problem.

3.4.2 An Extended Summary of NMF Algorithms

In this sub-section, we give an extended overview of the algorithms proposed for solving the NMF problem. Specifically, we review interior point methods based on projection along with the gradient based algorithms. Next we consider NMF optimization with explicit sparsity constraints. The NMF problem can be viewed as a bound-constrained optimization problem where projection methods can be applied. As such, [63] proposed Projected Non-Negative Least Squares method, which solves subproblems in algorithm 1 by solving an unconstrained least squares problem, and projecting the solution back to the positive orthant. The unconstrained Least Squares problem can be solved as follows

Lemma 3.4.3. *Given a pair of matrices $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{R}^{k\times k}$, The optimal transformation matrix that minimizes*

$$\min_{\boldsymbol{R}\in\mathbb{R}^{k\times k}} |\boldsymbol{X}_1 - \boldsymbol{X}_2\boldsymbol{R}|_F. \tag{3.30}$$

is $\boldsymbol{R} = \boldsymbol{X}_2^{\dagger}\boldsymbol{X}_1$.

Therefore, Projected Non-Negative Least Squares method is performed by iteratively applying

$$\begin{aligned} \boldsymbol{U}_{t+1} &= \mathcal{P}\left(\boldsymbol{X}^{*}\boldsymbol{V}_t^{T^{\dagger}}\right) \quad \text{and} \\ \boldsymbol{V}_{t+1} &= \mathcal{P}\left(\boldsymbol{X}^{*T}\boldsymbol{U}_t^{T^{\dagger}}\right), \end{aligned} \tag{3.31}$$

where $\mathcal{P}(\cdot)$ is the non-negative projection operator that projects all the negative components of its argument to zero, and $\boldsymbol{M}^{\dagger}$ denotes the Moore-Penrose inverse of the matrix $\boldsymbol{M}$. The drawback of such approach is that after applying the projection operator, the cost function does not necessarily decrease.

The same projection operator can be applied on gradient updates to solve each of the sub-problems in algorithm 1 using a gradient approach. where projected gradient methods could be used [64]

$$\boldsymbol{W}_{t+1} = \mathcal{P}\left(\boldsymbol{W}_t - \mu_t \nabla f(\boldsymbol{W}_t)\right), \tag{3.32}$$

for some appropriate step size μ_t, where $\boldsymbol{W}_t = \begin{bmatrix} \boldsymbol{U}_t \\ \boldsymbol{V}_t \end{bmatrix}$.

Choosing the step size μ_t in an adaptive way lied in the heart of projected gradient methods. One can for example aim for optimizing for the steps size in the sense of maximizing the decrease in the function value along the chosen direction that is

$$\mu_t = \arg\min_{\mu} f\left(\mathcal{P}\left[\boldsymbol{W}_t - \mu \nabla f(\boldsymbol{W}_t)\right]\right), \tag{3.33}$$

A simple yet popular gradient projection scheme widely used in many other bounded optimization algorithm is the so called *Armijo rule along the projection arc* [65], a pseudo code for which is given in the following

Algorithm 8 Armijo rule along the projection arc

Input: any feasible initial factors $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{U}_0 \\ \boldsymbol{V}_0 \end{bmatrix}$

positive constants $0 \leq \sigma \leq 1$ and $0 \leq \beta \leq 1$

repeat

Set $\boldsymbol{W}_{t+1} = \mathcal{P}\left(\boldsymbol{W}_t - \mu_t \nabla f(\boldsymbol{W}_t)\right)$,

where $\mu_t = \beta^M$, with M being the smallest integer that

$f(\boldsymbol{W}_{t+1}) - f(\boldsymbol{W}_t) \leq \sigma \nabla f(\boldsymbol{W}_t)^T \left(\boldsymbol{W}_{t+1} - \boldsymbol{W}_t\right)$,

Increment t

until some convergence criterion is met

return $(\boldsymbol{U}_t, \boldsymbol{V}_t)$

Since finding the step size for the projection operator is very time consuming, various methods have been proposed to initialize the search for the next step based on the previous step size(s).

Several other gradient approaches have been proposed for solving the sub-problems in Algorithm 1. For example, [66] uses Nesterov's acceleration method [67] to achieve fast rates for the convergence of the sub-problems. In particular, at each iteration round, the factors are updated using the projected gradient method performed on an adaptively chosen search

point, with a step size determined by the Lipschitz constant of the optimization objective. Authors in [68] also consider sub-problems in Algorithm 1 as bound-constrained optimization problems and use a quasi-Newton method to them by efficiently computing the Hessian. Another second order to solved the sub-problems is proposed in [69]. The method is based on the quasi-Newton type algorithm, by using symmetric rank-one matrices and finding proper negative curvature directions to approximate the Hessian matrix.

By considering the component-wise square of the factor matrices as the optimization variables, [70] frame the NMF problem as an unconstrained problem. A nice survey of the various optimization algorithms for solving NMF can be found in [71]. As has already been discussed so far, the premise of the non-negativity constraint in the NMF formulation is to serve as a surrogate for sparsity in the factor matrices so as to allow part-based interpretation of the factors. However, it is reported in some experimental studies, including [72, 73], that the mere non-negativity constraint does not yield sparse representations for the factor matrices and NMF does not necessarily learn localized features. Therefore, a remarkable body of research is developed around methods that impose sparsity constraint in a suitable form on one(or both) of the factor matrices along with the non-negativity constraint. Following the part based interpretation of NMF, a sparsity constraint on the columns of $\boldsymbol{U}$ implies that each part being sparse should represent a small part of the data. On the other hand, imposing a sparsity constraint on rows of $\boldsymbol{V}$ implies that each data point is approximated by a linear combination of a limited number of basis elements. If columns of $\boldsymbol{V}$ are sparse, then each basis vector is used to approximate a limited number of data points, which is particularly relevant when NMF is regarded as a soft clustering tool.

Adopting sparse constraints can also be seen as restricting the geometry of the problem to alleviate the issue of non-uniqueness of the solutions.

The Non-negative sparse coding algorithm proposed by Hoyer [74] is the first to incorporate a sparsity constraint in the NMF objective. Let us consider the regularized objective 3.1

with a sparsity constraint on rows of $\boldsymbol{V}$

$$\begin{aligned} \min_{\substack{\boldsymbol{U}\in\mathbb{R}^{n\times r}\\ \boldsymbol{V}\in\mathbb{R}^{m\times r}}} \quad & \frac{1}{2}\left\|\boldsymbol{U}\boldsymbol{V}^T-\boldsymbol{X}^*\right\|_F^2+\lambda||\boldsymbol{V}||_{1,1}, \\ \text{subject to} \quad & \boldsymbol{U}\geq 0 \text{ and } \boldsymbol{V}\geq 0, \end{aligned} \tag{3.34}$$

where $||\boldsymbol{V}||_{1,1}=\sum_{i=1}^{m}||\boldsymbol{V}_{i*}||_1$ is the sum of the ℓ_1 norm of the rows of $\boldsymbol{V}$, which is in fact equivalent to summing up all the entries of $\boldsymbol{V}$. Adopting the two block coordinate descent algorithm 1, Hoyer proves that the solution to the following sub-problem involving $\boldsymbol{V}$ by fixing $\boldsymbol{U}$

$$\begin{aligned} \boldsymbol{V}_{t+1}=\arg\min_{\boldsymbol{V}\in\mathbb{R}^{m\times r}} \quad & \frac{1}{2}\left\|\boldsymbol{U}_t\boldsymbol{V}^T-\boldsymbol{X}^*\right\|_F^2+\lambda||\boldsymbol{V}||_{1,1} \\ \text{subject to} \quad & \boldsymbol{V}\geq 0, \end{aligned} \tag{3.35}$$

can be found by multiplicative updates in the following form

$$\boldsymbol{V}_{t+1}=\boldsymbol{V}_t\odot\left[\boldsymbol{X}^{*T}\boldsymbol{U}_t\oslash\left(\boldsymbol{V}_t\boldsymbol{U}_t^T\boldsymbol{U}_t+\lambda\right)\right]. \tag{3.36}$$

In order to find the update equations for the subproblem involving $\boldsymbol{U}$, he adopts projected gradient updates

$$\boldsymbol{U}_{t+1}=\mathcal{P}\left(\boldsymbol{U}_t-\mu\nabla_{\boldsymbol{U}}\frac{1}{2}\left\|\boldsymbol{U}_t\boldsymbol{V}^T-\boldsymbol{X}^*\right\|_F^2\right). \tag{3.37}$$

In summary, the Pseudo-code for this algorithm is summarized in

Algorithm 9 Non-negative Sparse Coding

Input: any feasible initial factors $\boldsymbol{W}=\begin{bmatrix}\boldsymbol{U}_0\\ \boldsymbol{V}_0\end{bmatrix}$

positive constant $\lambda\geq 0$

repeat

Set $\boldsymbol{U}_{t+1}=\mathcal{P}\left(\boldsymbol{U}_t-\mu\nabla_{\boldsymbol{U}}\frac{1}{2}\left\|\boldsymbol{U}_t\boldsymbol{V}^T-\boldsymbol{X}^*\right\|_F^2\right)$

Normalize $\boldsymbol{U}_{t+1}$

Set $\boldsymbol{V}_{t+1}=\boldsymbol{V}_t\odot\left[\boldsymbol{X}^{*T}\boldsymbol{U}_t\oslash\left(\boldsymbol{V}_t\boldsymbol{U}_t^T\boldsymbol{U}_t+\lambda\right)\right]$

Increment t

until some convergence criterion is met

return $(\boldsymbol{U}_t,\boldsymbol{V}_t)$

In view of Hoyer's update equations 3.36 for the factor matrix $\boldsymbol{V}$, the gradient approach of [58] in algorithm 4 can be seen as solving a sparse constrained problem with a step size specified in their algorithm, that is

$$\begin{aligned} \min_{\substack{\boldsymbol{U}\in\mathbb{R}^{n\times r}\\ \boldsymbol{V}\in\mathbb{R}^{m\times r}}} \quad & \frac{1}{2}\left\|\boldsymbol{U}\boldsymbol{V}^T-\boldsymbol{X}^*\right\|_F^2+\lambda||\boldsymbol{U}||_{1,1}+\lambda||\boldsymbol{V}||_{1,1}, \\ \text{subject to} \quad & \boldsymbol{U}\geq 0 \ \text{ and } \ \boldsymbol{V}\geq 0, \end{aligned} \tag{3.38}$$

In a later work [72], Hoyer proposed a constrained optimization problem to address sparsity of the factors in the NMF problem. Specifically, by defining a heuristic measure for sparsity of a vector $\boldsymbol{x}\in\mathbb{R}^n$defined as

$$\mathcal{S}(\boldsymbol{x})=\frac{\sqrt{n}-\frac{||\boldsymbol{x}||_1}{||\boldsymbol{x}||_2}}{\sqrt{n}-1}, \tag{3.39}$$

Hoyer proposed the following constrained optimization formulation for sparse NMF

$$\begin{aligned} \min_{\substack{\boldsymbol{U}\in\mathbb{R}^{n\times r}\\ \boldsymbol{V}\in\mathbb{R}^{m\times r}}} \quad & \frac{1}{2}\left\|\boldsymbol{U}\boldsymbol{V}^T-\boldsymbol{X}^*\right\|_F^2, \\ \text{subject to} \quad & \boldsymbol{U}\geq 0 \ \text{ and } \ \boldsymbol{V}\geq 0 \\ & \mathcal{S}(\boldsymbol{U}_{*i})=s_{\boldsymbol{U}} \ \text{ and } \ \mathcal{S}(\boldsymbol{V}_{*i})=s_{\boldsymbol{V}} \ \ \forall 1\leq i\leq r \end{aligned} \tag{3.40}$$

In order to solve 3.40, a projected gradient algorithm with a projection that at each step enforces the columns of the factors to be non-negative, have unchanged ℓ_2 norm, but ℓ_1 norm set in a such a way that the sparseness constraint is met.

An alternative view on sparse NMF was later proposed in consequent papers of Kim and Park [75, 76], where they add a Frobenius norm regularization on $\boldsymbol{U}$

$$\begin{aligned} \min_{\substack{\boldsymbol{U}\in\mathbb{R}^{n\times r}\\ \boldsymbol{V}\in\mathbb{R}^{m\times r}}} \quad & \frac{1}{2}\left\|\boldsymbol{U}\boldsymbol{V}^T-\boldsymbol{X}^*\right\|_F^2+\eta||\boldsymbol{U}||_F^2+\lambda||\boldsymbol{V}||_{1,2}^2, \\ \text{subject to} \quad & \boldsymbol{U}\geq 0 \ \text{ and } \ \boldsymbol{V}\geq 0, \end{aligned} \tag{3.41}$$

where $||\boldsymbol{V}||_{1,2}^2=\sum_{i=1}^n||\boldsymbol{V}_{i*}||_1$ is a sparsity constraint on the rows of the factor matrix $\boldsymbol{V}$. Using the two block coordinate descent in algorithm 1, they reformulated the sub-problems

as standard non-negative least square problems in the following form. The sub-problem involving $\boldsymbol{V}$ can be written as

$$\begin{aligned} \min_{\boldsymbol{V}\in\mathbb{R}^{m\times r}} \quad & \frac{1}{2}\left\| \begin{pmatrix} \boldsymbol{U} \\ \sqrt{\lambda}\mathbf{1}_{1\times r} \end{pmatrix} \boldsymbol{V}^T - \begin{pmatrix} \boldsymbol{X}^* \\ \mathbf{0}_{1\times n} \end{pmatrix} \right\|_F^2, \\ \text{subject to} \quad & \boldsymbol{V} \geq 0, \end{aligned} \tag{3.42}$$

where $\mathbf{1}_{1\times r}$ is an all ones vector of size r and $\mathbf{0}_{1\times n}$ is an all zeros vector of size n.The sub-problem involving $\boldsymbol{U}$ can be written as

$$\begin{aligned} \min_{\boldsymbol{U}\in\mathbb{R}^{n\times r}} \quad & \frac{1}{2}\left\| \begin{pmatrix} \boldsymbol{V} \\ \sqrt{\eta}\mathbf{I}_r \end{pmatrix} \boldsymbol{U}^T - \begin{pmatrix} \boldsymbol{X}^{*T} \\ \mathbf{0}_{r\times n} \end{pmatrix} \right\|_F^2, \\ \text{subject to} \quad & \boldsymbol{U} \geq 0, \end{aligned} \tag{3.43}$$

This approach is closely related to an earlier work [77], where Frobenious norm regularization was adopted to impose sparse constraint on the factor matrices

$$\begin{aligned} \min_{\substack{\boldsymbol{U}\in\mathbb{R}^{n\times r} \\ \boldsymbol{V}\in\mathbb{R}^{m\times r}}} \quad & \frac{1}{2}\left\|\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{X}^*\right\|_F^2 + \eta||\boldsymbol{U}||_F^2 + \lambda||\boldsymbol{V}||_F^2, \\ \text{subject to} \quad & \boldsymbol{U} \geq 0 \ \text{ and } \ \boldsymbol{V} \geq 0, \end{aligned} \tag{3.44}$$

which yields similar update rules

$$\begin{aligned} \min_{\boldsymbol{U}\in\mathbb{R}^{n\times r}} \quad & \frac{1}{2}\left\| \begin{pmatrix} \boldsymbol{V} \\ \sqrt{\eta}\mathbf{I}_r \end{pmatrix} \boldsymbol{U}^T - \begin{pmatrix} \boldsymbol{X}^{*T} \\ \mathbf{0}_{r\times n} \end{pmatrix} \right\|_F^2, \\ \text{subject to} \quad & \boldsymbol{U} \geq 0, \end{aligned}$$

and

$$\begin{aligned} \min_{\boldsymbol{V}\in\mathbb{R}^{n\times r}} \quad & \frac{1}{2}\left\| \begin{pmatrix} \boldsymbol{U} \\ \sqrt{\eta}\mathbf{I}_r \end{pmatrix} \boldsymbol{V}^T - \begin{pmatrix} \boldsymbol{X}^{*} \\ \mathbf{0}_{r\times n} \end{pmatrix} \right\|_F^2, \\ \text{subject to} \quad & \boldsymbol{V} \geq 0, \end{aligned}$$

Although the solutions to this problems have fairly small values due to the shrinkage effect of the ℓ_2 norms used for the regularization, they are however not necessarily sparse in the strict sense.

CHAPTER 4

Evaluation and Computational Results

In this chapter, we evaluate our entity resolution model in two case studies; namely a discussion forum on parenting issues and a transactional setting on Twitter.

4.1 Mothering Discussion Forums

4.1.1 Motivation

Over the past decade and a half, the explosion in social media and the concomitant rise in informational websites has changed the manner in which people access health care information. Various sites dedicated to conversations about child rearing and parenting, colloquially referred to as "mommy blogs," attract millions of users. Although straightforward data mining techniques such as topic modeling exist for determining what parents are talking about on these sites and other similar sites, few techniques exist for determining how they are talking about those topics.

Among the many topics discussed on these parenting sites, few topics garner as much attention and vigorous discussion as childhood vaccination. Despite the fact that safe and effective vaccines exist, sporadic outbreaks of vaccine preventable diseases (VPDs) point to the continuing tension between public programs intended to make these vaccinations easily accessible and broadly adapted and parents who resist vaccination based largely on ideological principles.

For example, Measles was officially eliminated in the United States with no continuous transmission for twelve months in the year 2000. However, public health officials were worried

that the pronouncement of the demise of measles might contribute to a false sense of security among parents, thereby diminishing their commitment to vaccinating their children. This concern was realized in the events of late 2014 and early 2015, when over 120 cases of measles across the United States were linked to an infected visitor at a Southern Californian amusement park.

Reduced rates of vaccination have jeopardized the elimination of diseases that have been on the cusp of such elimination for decades and, as recent outbreaks attest, threaten the hard-won herd immunity developed through long-term vaccination programs. In the last decades of the 20th century, even while many childhood diseases were disappearing from the disease landscape of America because of successful vaccination programs, certain communities, particularly those that resisted vaccination based on ideological principles (largely communities of faith), continued their use of exemptions as an expression of these principles.

The role of exemptions in precipitating outbreaks in vaccine-communicable disease is increasingly being considered, although little evidence is currently available to directly support this link. Although simple inspection of parenting sites and standard text mining approaches can confirm that vaccination is a topic of frequent discussion on these sites, such methods cannot determine the structure of those discussions.

Importantly, these communities were easy to monitor, and schools or school districts with high exemption rates were uncommon. Conversations about vaccinations were largely confined to interactions between parents and medical professionals.

Both the recent adoption of social media forums like "mommy blogs" changed fundamentally the nature and reach of these conversations. Over the course of a relatively short time span, these sites became a locus for discussions not only about the safety and efficacy of vaccinations, but also for sharing strategies for obtaining exemptions. Unlike the relatively circumscribed groups that circulated this information in the pre-social media era, these virtual communities were geographically dispersed and visible in many more social strata. Importantly, these online conversations became clearinghouses for information about the local regulatory regimes governing exemptions. In this study we are interested in characterizing

the narratives in these forums that underly the discussions [78].

4.1.2 Objective

We analyzed 300 thousand posts contributed by 12,376 users and viewed a few million times indexed from a popular parenting site, "mothering.com", over almost 8 years ending in 2012. Beyond simply identifying the main topics of discussion on the site, our objective is to identify the underlying narrative frameworks that explain the stories circulating in these various discussions. In addition to delineating the narrative framework that parents activate in their storytelling, we provide a fine-grained view of actant interactions and relationships in these stories, offering insight into individuals' shifting attitudes toward vaccination. Figure 4.1 shows the pipeline describing the steps of this study.

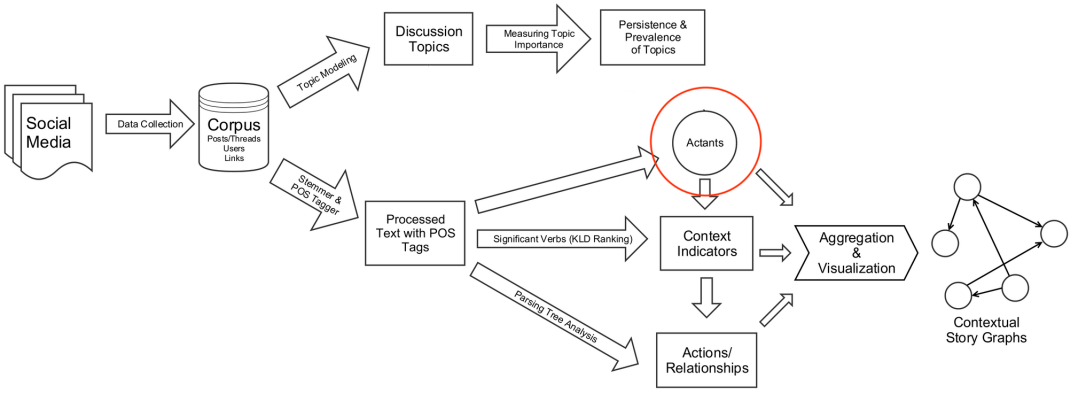

Figure 4.1: Pipeline of the study

In this dissertation, we only focus on automatic detection of actant groups, which we pose as an entity clustering problem.

4.1.3 Data

Data for this study were obtained from a popular social media site dedicated to parenting, mothering.com. This website is particularly popular among a group of user who self-identify as mothers. As mothers are on the "front line" of discussions about the health of their

infant children, the site offers important information about how they approach decisions related to vaccination. The site draw members from a wide range of backgrounds with broad geographic diversity, largely in the United States and Canada, but it is know to have an anti-vaccination tone. The language of the discussion forum is English. We indexed posts that appeared in the forums related to childhood vaccination, recursively visiting and storing all publicly available discussion threads, and date-time–data, while creating an anonymized index of any accessible user data, resulting in a corpus of 299,778 posts from 12,376 users, based on 105 months of indexed data (2004-2012). These posts comprised the corpus for analyses.

4.1.4 Ground Truth Actants

We use the notion of an actant to refer to a set of entities that serve the same or similar purpose in the decision making ecosystem we are exploring. We captured each actant by an associated set of words that most commonly referred to that group in the forums.

Within we identified three main categories of actants:

- *individual actants:*
  comprised of parents, children, and medical professionals
- *corporate actants:*
  comprised of government institutions, religious institutions, and schools
- *objects:*
  comprised of vaccines, exemptions, VPDs, and adverse effects.

The words associated with an actant consist of both synonyms for the actant and entities that have the actant as a super-category. For example, the actant 'government' includes the colloquial synonym 'the Feds' as well as the government institution, the 'CDC', where 'government' is the super-category for CDC. This list of actants is summarized in Table 4.1.

Table 4.1: Ground Truth actants

Entities (Nodes)	Associated Word Set
Individuals	
• Parents	• parents, parent, i, we, us, you
• Children	• child, kid, kids, children, daughter, daughters, son, sons, toddler, toddlers, kiddo, boy, d[ear]d[aughter], d[ear]s[on]
• Medical Prof.	• doctor, doctors, pediatrician, pediatricians, nurse, nurses, ped, md, dr
Corporates	
• Government	• government, cdc, federal, feds, center for disease control, officials, politician, official, law
• Religious Inst.	• faith, religion, pastor, pastors, parish, parishes, church, churches, congregation, congregations, clergy
• Schools	• teacher, teachers, preschools, preschool, school, schools, class, daycare, daycares, classes
Objects	
• Vaccines	• vaccines, vax, vaccine, vaccination, vaccinations, shots, shot, vaxed, unvax, unvaxed, nonvaxed, vaccinate, vaccinated, vaxes, vaxing, vaccinating, substances, ingredients
• Exemptions	• exemption, exempt
• VPDs	• varicella, chickenpox, flu, whooping cough, tetanus, pertussis, hepatitis, polio, mumps, measles, diphtheria
• Adverse Effects	• autism, autistic, fever, fevers, reaction, reactions, infection, infections, inflammation, inflammations, pain, pains, bleeding, bruising, diarrhea, diarrhoea

4.1.5 Sample Posts from the Discussion Forum

In the posts below all actants are in bold while significant ones are also printed in red. The relational verb between the significant actants is italicized and underlined.

- My hubby and **I** are presently concerned about the ingredients - our **faith** prevents us from *allowing* certain **substances** into our **son**'s body.
- **I** have come to the conclusion that there is more of a likelihood that my **child** would *have* a bad **reaction** to certain **vaccines** than getting the actual disease itself .
- Every **vaccine** *has* a very small chance of more serious **side effects** (generally linked to allergic reactions , about which **parents** could not be expected to *know* before **vaccination**, because the **child** is so young), and a larger chance of less serious **reactions** (soreness , fever , redness) .
- He mentioned piggy backing off other 's **immunizations** and how **autism** isn't *caused* by **vaxes** and a bunch of other things just to try to convince me to vax my **son** .
- It is ironic that the pro-vax crowd will laugh off the **autism-vaccines** link as if it is somehow preposterous - yet here is just one example of a **child** who was *diagnosed* with **autism** and the **government** *conceded* his **vaccine injury**.
- **I** certainly wouldn't want **my kids** to *suffer* socially because of our **decision not to vax** , but OTOH if I say something that can save one family from the pain of a serious **reaction** or death of a child , then maybe it 's worth it .
- This concern is reinforced by a study which revealed that 1 in 175 **children** who completed the full DPT series *suffered* "severe **reactions**", and a **Dr**'s report for attorneys which found that 1 in 300 **DPT immunizations** *resulted in* **seizures** .
- If a **parent** wants to *exempt* their **child** only from the **MMR , Hep A and varicella vaccines** because of the aborted fetal tissue , the **religious exemption** would *be invalid* in almost every state .
- Even if the **Church** *told* all Catholic **parents** not to let their **child** *get* the **MMR** for instance , most **parents** would have to still be required to *submit* a **religious exemption** which would *exempt* all **vaccines**.

- **I** think , first of all , that since **my kids** never *had* a **reaction** to **vaccines**, I just would never be able to forgive myself if they got a really horrible disease that would have been preventable through taxes .

- Here is the Hawaii **immunization** brochure , which states the **exemption** forms can also be *obtained* from the **school**: **Immunization** and TB code : Surprisingly , **I** don't see anything about religiously exempting a **child** from the TB screening requirement in the code .

- My **kids** are partially vaxed and **we** just *submitted* a **religious exemption** to the **school** she will be attending this fall.

- My younger **daughter** may be going to **preschool** next year and **I** am either going to have to *immunize* or *claim* a **religious exemption**, and I want my thoughts together on the topic so **I** can advocate for her .

- The **CDC** here told me they wouldn't *accept* the any **kids** w/out **shots**, even with a signed **exemption**.

- If the **school** has funding by the **government** then it is my understanding they must *accept* the **exemption**

- For Peace Corps , the US **Government** *requires* **vaccinations** BUT there is a **religious exemption** and disclaimer, I 'm sure of it- anyway, there was in 1998.

4.1.6 Relation Extraction

In this section, we briefly describe our relation extraction method. A dependency relationship [79] is an asymmetric binary relationship between a word called **head**, and another word called **modifier**. The structure of a sentence can be represented by a set of dependency relationships that form a tree. A word in the sentence may have several modifiers, but each word may modify at most one word. The root of the dependency tree, which is called the head of the sentence, does not modify any word. For example, Figure 4.2 shows the

dependency tree for the sentence “the doctor signed our exemption form.”, generated by the Stanford parser [80]. The links in the diagram represent dependency relationships.

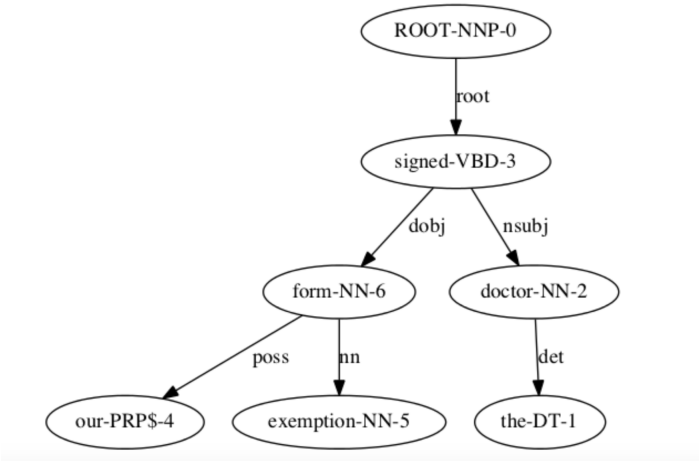

Figure 4.2: Parsing tree for the sentence “the doctor signed our exemption form.”, obtained by the Stanford parser

The direction of a link is from the head to the modifier in the relationship. Labels associated with the links represent types of dependency relations. In the dependency trees, each link between two words represents a direct semantic relationship. A path allows us to represent indirect semantic relationships between two content words. We name a path by concatenating dependency relationships and words along the path, excluding the words at the two ends. For example, the main parsing structure(path), that we are interested in is

$$N : subj : V : obj : N$$

Given a corpus of text, after cleaning the data and tokenizing the documents into sentences. We scan every tokenized sentence from the corpus and extract relation triplets of the form (m_s, m_o, v), where m_s is the head word of the subject node in the parsing tree, m_o is the headword of the object node in the parsing tree, and v is the verb that connects the two. The following table provides a summary of the extractions $\mathcal{T} = \{(m_s, m_o, v)\}$ with our method.

Table 4.2: Summary statistics for relation extraction

Extraction	Number
Relation triplets($\lvert\mathcal{T}\rvert$)	226949
Unique subject entity mentions($\lvert\mathcal{M}_S\rvert$)	15727
Unique object entity mentions($\lvert\mathcal{M}_O\rvert$)	37829
$\lvert\mathcal{M}_S \cap \mathcal{M}_O\rvert$	3955
Unique relation phrases($\lvert\mathcal{V}\rvert$)	33818

4.1.7 Entity-Relationship Matrices

Recall Chapter 3 that we construct the left entity relation matrix $\boldsymbol{X}_L \in \mathbb{R}^{|\mathcal{M}| \times N_L}$ and the right entity relation matrix $\boldsymbol{X}_R \in \mathbb{R}^{|\mathcal{M}| \times N_R}$ as follows

$$\boldsymbol{X}_L[i,j] = \text{TF-IDF}^L(v_j, m_i), \text{ and}$$
$$\boldsymbol{X}_R[i,j] = \text{TF-IDF}^R(v_j, m_i).$$

We review the computation of $\text{TF-IDF}^L(v_j, m_i)$ and specify the parameters used in the construction. The construction of the right entity relation matrix is similar. In the above construction, the TF-IDF is defined as

$$\text{TF-IDF}^L(v, m) = \frac{TF^L_{v,m}}{\text{Q-Rank}^L_v(m)},$$

where the term-frequency component in the nominator is computed as

$$TF^L_{v,m} = \log(1 + s^L_v(m)),$$

where $s^L_v(m) = \sum_{m' \in \mathcal{M}} \mathbf{1}_{(m,m',v) \in \mathcal{T}}$.

The IDF component in the denominator is computed by

$$\text{Q-Rank}^L_v(m) = \sqrt{1 + \text{Rank}^L_v(m) \ / \ w},$$

where $\text{Rank}_v^L(m)$ = rank of $m \in \mathcal{M}_v$ in $\mathcal{S}_v^L$, in which $\mathcal{S}_v^L$ representing the collection of all co-occurrence counts of the relation phrase $v \in \mathcal{V}$ with all entity mentions that co-occur with it in $\mathcal{M}_v^L = \{m \mid (m, m', v) \in \mathcal{T} \text{ for some } m' \in \mathcal{M}\}$. In our experiments we set

$$w = \max(5, |\mathcal{S}_v^L|/10).$$

Also, in construction of the left and right entity-relation matrices, we only keep the entity mentions that appear both as a subject in some relation tuple and an object in some other relation tuple. Furthermore, in order to improve the quality of the results, we filter out the entities that appear in less than 4 relation tuples in $\mathcal{T}$.

After this filtering, we have

$$\boldsymbol{X}_L \in \mathbb{R}^{286\times 1242}, \text{ and}$$
$$\boldsymbol{X}_R \in \mathbb{R}^{286\times 1528},$$

which means that $|\mathcal{M}| = 286, |\mathcal{V}_L| = 1242$ and $|\mathcal{V}_R| = 1528$.

4.1.8 Matrix Estimation results

In this section, we compare the performance of the algorithms studied in the previous chapter to solve the matrix completion optimization on the ground truth matrices. The optimization problem for the right entity relation matrix is given as follows. The optimization problem for the left entity relation matrix is likewise.

$$\begin{aligned} \min_{\substack{\boldsymbol{U}_R\in\mathbb{R}^{|\mathcal{M}|\times d}\\ \boldsymbol{V}_R\in\mathbb{R}^{N_R\times d}}} \quad & ||\boldsymbol{X}_R - \boldsymbol{U}_R\boldsymbol{V}_R^T||_F^2 \\ \text{subject to} \quad & \boldsymbol{U}_R \geq 0 \text{ and } \boldsymbol{V}_R \geq 0, \end{aligned} \tag{4.1}$$

In our evaluations, we study the following three algorithms to solve 4.1. In all the algorithms, the ambient dimension of the embedded space is set to $d = 20$, implying that

$$\boldsymbol{U}_R \in \mathbb{R}^{286\times 20}, \text{ and}$$
$$\boldsymbol{V}_R \in \mathbb{R}^{1528\times 20}.$$

We specify the parameters of each algorithm in the pseudo-codes.

- Multiplicative Updates Algorithm(NMF-MUL):

  Multiplicative update rules [46] is the most commonly used algorithm for solving 4.1. It is very fast, as it does not require computing the gradient in each steps and the multiplicative updates are easy to implement. It is almost parameter free, except that a stopping criterion based on the number of iterations or convergence should be designed.

Algorithm 10 Solving NMF with NMF-MUL

Input: the ground truth matrix $\boldsymbol{X}_R$

Initialize factors with $\boldsymbol{U}_0 \sim \mathcal{U}[0,1]$ and $\boldsymbol{V}_0 \sim \mathcal{U}[0,1]$

repeat

Set $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t \odot \left[\boldsymbol{X}^* \boldsymbol{V}_t \oslash \boldsymbol{U}_t \boldsymbol{V}_t^T \boldsymbol{V}_t\right]$

Set $\boldsymbol{V}_{t+1} = \boldsymbol{V}_t \odot \left[\boldsymbol{X}^{*T} \boldsymbol{U}_t \oslash \boldsymbol{V}_t \boldsymbol{U}_t^T \boldsymbol{U}_t\right]$

Increment t

until $t > 3000$ **or** $f(\boldsymbol{W}_{t+1}) - f(\boldsymbol{W}_t) < 10^{-4}$

return $(\boldsymbol{U}_t, \boldsymbol{V}_t)$

- Gradient-Based Algorithm(NMF-GRAD):

  This algorithm is designed to guarantee convergence to a stationary point with a small modification of the NMF-MUL. As mentioned in chapter 3, the parameter δ in this algorithm can be regarded as a sparsity controlling parameter. But the algorithm is very sensitive to this parameter and it is usually set to be very small. In the following we specify the parameters of the algorithm.

Algorithm 11 Solving NMF with NMF-GRAD

Input: the ground truth matrix $\boldsymbol{X}_R$

positive constants $\sigma = 0.01$ and $\delta = 10^{-10}$

Initialize factors with $\boldsymbol{U}_0 \sim \mathcal{U}[0,1]$ and $\boldsymbol{V}_0 \sim \mathcal{U}[0,1]$

repeat

Set $\bar{\boldsymbol{W}}_t[i,j] = \begin{cases} \boldsymbol{W}_t[i,j] & \text{if } \nabla_{\boldsymbol{W}} f(\boldsymbol{W}_t)[i,j] \geq 0 \\ \max(\sigma, \boldsymbol{W}_t[i,j]) & \text{if } \nabla_{\boldsymbol{W}} f(\boldsymbol{W}_t)[i,j] < 0 \end{cases}$

for $1 \leq i \leq m+n$, $1 \leq j \leq r$

Set $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \left\{\bar{\boldsymbol{U}}_t \oslash [\bar{\boldsymbol{U}}_t \boldsymbol{V}_t^T \boldsymbol{V}_t + \delta]\right\} \odot \nabla_{\boldsymbol{U}} f(\boldsymbol{U}_t, \boldsymbol{V}_t)$

Set $\boldsymbol{V}_{t+1} = \boldsymbol{V}_t - \left\{\bar{\boldsymbol{V}}_t \oslash [\bar{\boldsymbol{V}}_t \boldsymbol{U}_t^T \boldsymbol{U}_t + \delta]\right\} \odot \nabla_{\boldsymbol{V}} f(\boldsymbol{U}_t, \boldsymbol{V}_t)$

Increment t

until $t > 3000$ **or** $f(\boldsymbol{W}_{t+1}) - f(\boldsymbol{W}_t) < 10^{-4}$

return $(\boldsymbol{U}_t, \boldsymbol{V}_t)$

- The Exterior Point Method(NMF-EXT) As fully described in chapter 3, NMF-EXT solves NMF by first finding an optimal solution to the unconstrained problem and then moving to the feasible region from that point by flat gradient updates. Once in the feasible region, any NMF algorithm can be used. Here, we specify the parameters of the flat descent algorithm for solving NMF.

Algorithm 12 Solving NMF with NMF-EXT

Input: the ground truth matrix $\boldsymbol{X}_R$

initial factors $\boldsymbol{W}_0 = \begin{bmatrix} \boldsymbol{U}_0 \\ \boldsymbol{V}_0 \end{bmatrix}$ obtained as explained in Chapter 3

initial descent constant $\delta = 0.1$

Perturbation constant $\epsilon = 10^{-6}$

Scaling constant $\kappa = 0.1$

Number of outer iterations $n_1 = 30$

Number of inner iterations $n_2 = 100$

Step parameters $\beta = 0.5$, $\sigma = 10^{-1}$

Set $\widehat{\boldsymbol{W}} = \boldsymbol{W}_0$

for $i \in \{1, \ldots, n_1\}$ **do**

  if $i \bmod 2 = 0$ **then**

    $s = \kappa$

  else

    $s = \frac{1}{\kappa}$

  end if

  $\boldsymbol{W}_0 = \textbf{PerturbScale}(\widehat{\boldsymbol{W}}, \epsilon, s)$

  $\delta_0 = \delta$

  for $t \in \{0, \ldots, n_2 - 1\}$ **do**

    Set $\boldsymbol{W}_{t+1} = \textbf{Flat-Descent}(\boldsymbol{W}_t, \nabla f(\boldsymbol{W}_t), \delta_t)$,

      where $\delta_t = \delta_{t-1}\beta^M$, with M being the smallest integer that

      $f(\boldsymbol{W}_{t+1}) - f(\boldsymbol{W}_t) \leq \sigma \nabla f(\boldsymbol{W}_t)^T (\boldsymbol{W}_{t+1} - \boldsymbol{W}_t)$,

  end for

  $\widehat{\boldsymbol{W}} = \boldsymbol{W}_{n_2}$

end for

return $\widehat{\boldsymbol{W}}$

As for measures of evaluation, we consider the approximation error and the sparsity of the

factor matrices. That is

$$\epsilon_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}}, \boldsymbol{V}_L^{\mathcal{A}}) = ||\boldsymbol{X}_R - \boldsymbol{U}_R^{\mathcal{A}}\boldsymbol{V}_R^{\mathcal{A}^T}||_F^2 \text{ and}$$
$$\kappa_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}}) = \frac{||\boldsymbol{U}_R^{\mathcal{A}}||_0}{|\mathcal{M}| \times d},$$

where $(\boldsymbol{U}_R^{\mathcal{A}}, \boldsymbol{V}_R^{\mathcal{A}})$ are the solutions of 4.1 given by algorithm $\mathcal{A}$.

We run these algorithms $n_s = 200$ times and obtain a sample of size n_s for the statistics of interest. Let $\bar{\epsilon}_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}}, \boldsymbol{V}_R^{\mathcal{A}}) = \mathbf{mean}(\{\epsilon_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}}, \boldsymbol{V}_R^{\mathcal{A}})\})$ and $\bar{\kappa}_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}}) = \mathbf{mean}(\{\kappa_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}})\})$.

Tables below represents $\bar{\epsilon}_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}}, \boldsymbol{V}_R^{\mathcal{A}})$, minimum value, maximum value and a 95% confidence interval length for $\{\epsilon_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}}, \boldsymbol{V}_R^{\mathcal{A}})\}$. The confidence interval

$$\left[\bar{\epsilon}_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}}, \boldsymbol{V}_R^{\mathcal{A}}) \pm \frac{1.96\sigma}{\sqrt{n_s}}\right]$$

is obtained through a standard application of the central limit theorem, where σ is the empirical standard deviation of $\{\epsilon_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}}, \boldsymbol{V}_R^{\mathcal{A}})\}$.

By the same token, we have represented $\bar{\kappa}_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}})$, minim value, maximum value and a 95% confidence interval length for $\{\kappa_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}})\}$.

Table 4.3: Summary statistics for prediction accuracy and sparsity of NMF-MUL algorithm for the right entity relation matrix X_R

NMF-MUL	Prediction Error	Sparsity
mean	184.89	18.709
confidence interval size	0.0551	0.1398
minimum	184.41	17.304
maximum	185.79	20.336

Table 4.4: Summary statistics for prediction accuracy and sparsity of NMF-GRAD algorithm for the right entity relation matrix X_R

NMF-GRAD	Prediction Error	Sparsity
mean	184.86	20.463
confidence interval size	0.0520	0.1129
minimum	184.38	18.865
maximum	185.59	22.322

Table 4.5: Summary statistics for prediction accuracy and sparsity of NMF-EXT algorithm for the right entity relation matrix X_R

NMF-EXT	Prediction Error	Sparsity
mean	184.74	45.762
confidence interval size	0.0394	3.9155
minimum	184.38	0.0
maximum	185.23	55.921

In order to get a sense of how well the NMF-optimization algorithms do in terms of reconstructing the matrix, we have to compare the prediction errors in the above table with the optimal SVD error. In this case

$$\min_{\substack{\boldsymbol{U}_R \in \mathbb{R}^{|\mathcal{M}| \times d} \\ \boldsymbol{V}_R \in \mathbb{R}^{N_R \times d}}} ||\boldsymbol{X}_R - \boldsymbol{U}_R \boldsymbol{V}_R^T||_F^2 = 176.116$$

Clearly, the factor matrices obtained by the NMF-EXT are much sparser than those of the other algorithms. Moreover, the prediction error of the NMF-EXT is lower than that of the other two algorithms. It can also be observed that NMF-GRAD performs a little bit better than NMF-MUL in terms of prediction accuracy but "significantly" better in terms of sparsity. To test the statistical significance of our observations, we perform a two sample

one-sided t-test with the following hypotheses on the samples we have obtained from each pair of algorithms

$$\mathcal{H}_0\text{: } \mathbf{mean}(\{\epsilon_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}}, \boldsymbol{V}_R^{\mathcal{A}})\}) \leq \mathbf{mean}(\{\epsilon_{\mathcal{A}'}(\boldsymbol{U}_R^{\mathcal{A}'}, \boldsymbol{V}_R^{\mathcal{A}'})\})$$

$$\mathcal{H}_1\text{: } \mathbf{mean}(\{\epsilon_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}}, \boldsymbol{V}_R^{\mathcal{A}})\}) > \mathbf{mean}(\{\epsilon_{\mathcal{A}'}(\boldsymbol{U}_R^{\mathcal{A}'}, \boldsymbol{V}_R^{\mathcal{A}'})\}).$$

The corresponding p-value of the test is then

$$p_{\mathcal{A},\mathcal{A}'} = \mathbb{P}(T > \frac{\bar{\epsilon}_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}}, \boldsymbol{V}_R^{\mathcal{A}}) - \bar{\epsilon}_{\mathcal{A}'}(\boldsymbol{U}_R^{\mathcal{A}'}, \boldsymbol{V}_R^{\mathcal{A}'})}{\sqrt{\frac{\sigma_{\mathcal{A}}^2 + \sigma_{\mathcal{A}'}^2}{n_s}}}),$$

where T is the Student's random variable with $n_s - 1$ degrees of freedom and $\sigma_{\mathcal{A}}$ is the empirical standard deviation of the sample $\{\epsilon_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}}, \boldsymbol{V}_R^{\mathcal{A}})\}$. Table below gives the values of $p_{\mathcal{A},\mathcal{A}'}$ for all pairs of algorithms

Table 4.6: p-values of the t-test

$(\mathcal{A}, \mathcal{A}')$	$p_{\mathcal{A},\mathcal{A}'}$
(NMF-MUL,NMF-GRAD)	0.228
(NMF-MUL,NMF-EXT)	2.85×10^{-5}
(NMF-GRAD,NMF-EXT)	0.0004

We repeated the same experiments on the left entity-relation matrix and the results are summarized in the following tables

Table 4.7: Summary statistics for prediction accuracy and sparsity of NMF-MUL algorithm for the left entity relation matrix X_L

NMF-MUL	Prediction Error	Sparsity
mean	190.59	20.536
confidence interval size	0.0332	0.2074
minimum	190.28	17.748
maximum	191.02	23.333

Table 4.8: Summary statistics for prediction accuracy and sparsity of NMF-GRAD algorithm for the left entity relation matrix X_L

NMF-GRAD	Prediction Error	Sparsity
mean	190.46	22.268
confidence interval size	0.0293	0.2701
minimum	190.24	19.893
maximum	190.87	26.684

Table 4.9: Summary statistics for prediction accuracy and sparsity of NMF-EXT algorithm for the left entity relation matrix X_L

NMF-EXT	Prediction Error	Sparsity
mean	190.51	51.995
confidence interval size	0.0331	3.0066
minimum	190.25	0.0
maximum	191.12	57.765

In order to compare the reconstruction error of the NMF algorithms with that of the optimal

SVD, we note that

$$\min_{\substack{\boldsymbol{U}_L \in \mathbb{R}^{|\mathcal{M}| \times d} \\ \boldsymbol{V}_L \in \mathbb{R}^{N_L \times d}}} \quad ||\boldsymbol{X}_L - \boldsymbol{U}_L \boldsymbol{V}_L^T||_F^2 = 179.786$$

It can be clearly observed that the NMF-EXT algorithm beats the other two in terms of sparsity. Although, it is interesting to note that, there are instances for which the solution of the NMF-EXT algorithm is not sparse at all. In terms of prediction accuracy, it can be observed that the NMF-GRAD algorithm does better than the other two; while NMF-EXT performs better than the NMF-MUL.

4.1.9 Entity Clustering results

As discussed earlier, our primary objective is to find clusters of entity mentions that refer to the same group of objects in the stories, aka actants. In fact, the main intuition behind seeking sparse embeddings for the entity mentions, which was achieved by imposing the non-negativity of the factor matrices in the matrix estimation problem as a surrogate to sparsity of the factors, is that sparse embeddings lead to better clustering results. In this section we verify this intuition by evaluating the clustering performance of the above algorithms on the dataset based on the actant groupings in table 4.1.

In order to show the importance of the sparsity of the entity mention embeddings in the performance of the clustering, we compare the clustering performance of the NMF-based algorithms with that of the embeddings obtained from optimal SVD factorization. Note that for a fair comparison between the different algorithms, we fix the clustering method that groups the obtained embeddings by the different algorithms. Specifically, we use K-means clustering with $K = 20$ clusters.

Let us recall the measures of clustering performance from chapter 2. Suppose that Given a set $\mathcal{M}$ of entity mentions with the underlying ground truth actant groupings $\mathcal{E} = \{E_1, E_2, \ldots, E_J\}$, we want to measure the quality of a clustering $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$ with respect to $\mathcal{E}$. The measures of clustering performance that we consider in our experiments are

- Mutual Information $I(\mathcal{C}; \mathcal{E}) = \sum_{i=1}^{K} \sum_{j=1}^{J} P_{i,j} \log \frac{P_{i,j}}{P_i P'_j}$

- Homogeneity $h(\mathcal{C};\mathcal{E}) = 1 - \frac{H(\mathcal{C}|\mathcal{E})}{H(C)}$
- Completeness $c(\mathcal{C};\mathcal{E}) = 1 - \frac{H(\mathcal{E}|\mathcal{C})}{H(C)}$
- V-measure $v(\mathcal{C};\mathcal{E}) = \frac{2h(\mathcal{C};\mathcal{E})c(\mathcal{C};\mathcal{E})}{h(\mathcal{C};\mathcal{E})+c(\mathcal{C};\mathcal{E})}$,

where $P_i = \frac{|C_i|}{N}$, $P'_j = \frac{|E_j|}{N}$, $P_{i,j} = \frac{|C_i \cap E_j|}{N}$, are relative sizes of the groups and their intersection and $H(\mathcal{C}) = -\sum_{i=1}^{K} P_i \log P_i$, and

$$H(\mathcal{C}|\mathcal{E}) = -\sum_{i=1}^{K}\sum_{j=1}^{J} P_{i,j} \log \frac{P_{i,j}}{P'_j},$$

are the cluster entropy and conditional entropy respectively. For each algorithm, we have collected an independent sample of size $n_s = 200$ on the statistics of interest.

For each measure $D(\mathcal{C};\mathcal{E})$, let $\{D^{(i)}(\mathcal{C};\mathcal{E})\}_{i=1}^{n_s}$ be the sample. Moreover, let $\bar{D}(\mathcal{C};\mathcal{E}) = \mathbf{mean}(\{D^{(i)}(\mathcal{C};\mathcal{E})\}_{i=1}^{n_s})$.

Tables below represents $\bar{D}(\mathcal{C};\mathcal{E})$, minim value, maximum value and a 95% confidence interval length for $\{D^{(i)}(\mathcal{C};\mathcal{E})\}_{i=1}^{n_s}$ for the various measures defined above.

Table 4.10: Summary of K-means clustering results for NMF-MUL algorithm with K=20

NMF-MUL	homogeneity	completeness	V-measure	Mutual Information
mean	0.523	0.428	0.471	0.473
confidence interval size	0.002	0.001	0.002	0.002
minimum	0.490	0.403	0.442	0.444
maximum	0.554	0.450	0.497	0.500

Table 4.11: Summary of K-means clustering results for NMF-GRAD algorithm with K=20

NMF-GRAD	homogeneity	completeness	V-measure	Mutual Information
mean	0.526	0.427	0.471	0.474
confidence interval size	0.002	0.001	0.002	0.002
minimum	0.488	0.405	0.448	0.450
maximum	0.565	0.451	0.502	0.505

Table 4.12: Summary of K-means clustering results for NMF-EXT algorithm with K=20

NMF-EXT	homogeneity	completeness	V-measure	Mutual Information
mean	0.523	0.425	0.469	0.471
confidence interval size	0.003	0.002	0.002	0.002
minimum	0.488	0.401	0.446	0.448
maximum	0.563	0.456	0.504	0.506

Table 4.13: Summary of K-means clustering results for SVD algorithm with K=20

SVD	homogeneity	completeness	V-measure	Mutual Information
mean	0.511	0.424	0.463	0.466
confidence interval size	0.003	0.002	0.002	0.003
minimum	0.467	0.390	0.426	0.428
maximum	0.560	0.455	0.499	0.501

It can be observed that the clustering accuracy is more or less the same on embeddings obtained from different NMF algorithms and it is significantly better than that of the SVD embeddings.

In order to visualize the obtained embeddings from these methods, we use a popular method

to visualize high dimensional embeddings, called t-distributed Stochastic Neighbor Embedding (t-SNE) [81]. t-SNE is a nonlinear dimensionality reduction technique to visualize high dimensional vectors on a 2D plane. It models each high-dimensional object by a two-dimensional point such that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability.

In fact it constructs two probability distributions, one over pairs of high-dimensional vectors and another one over the points in the low-dimensional space, and then minimizes the Kullback–Leibler(KL) divergence between the two distributions with respect to the locations of the points in the map.

In the following, we visualize the entity mention embeddings obtained by the NMF-EXT algorithm via the t-SNE mapping. It can be clearly seen that groups of entity mentions that refer to the same actant group tend to cluster together.

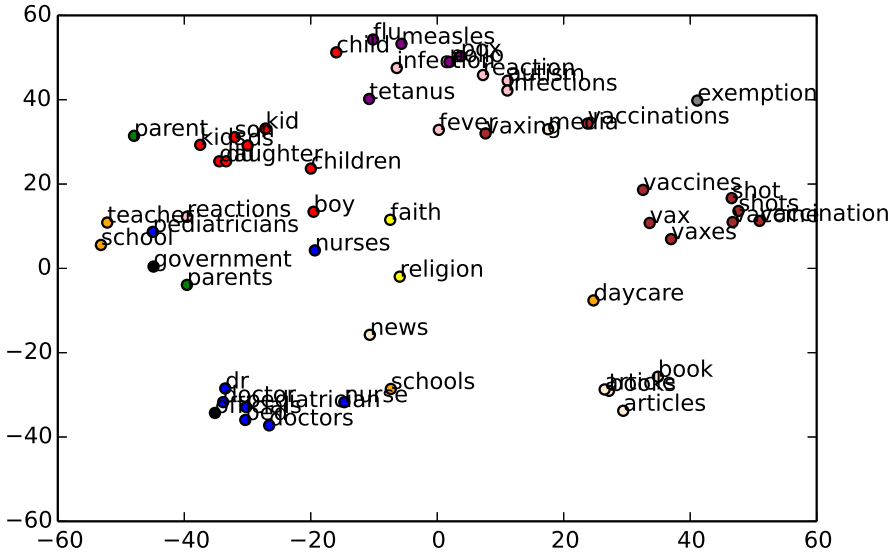

Figure 4.3: Entity Embeddings for Mothering data

In the following table, the clusters by the K-means algorithm with $k = 20$ applied to embeddings obtained by NMF-EXT algorithm are listed. It can be clearly observed that clusters 1 and 13 represent care-receiver actant groups. Cluster 3 represents the actant group related

to big pharma. Cluster 4, 11 and 15 represents media that parents use as sources of obtaining and sharing information. Cluster 5 concerns the actant group related to vaccination. Cluster 6 represents an actant group related to causes of the diseases. Cluster 12 represents the parents actant in the story narrative. Cluster 14 represents the actant that concerns adverse effects allegedly due to vaccines. Clusters 16 and 19 represent care providers. Cluster 17 represents that actant group related to authorities and finally cluster 18 represents the actant group concerning vaccination preventable diseases.

cluster 1	infant women groups generation girl boy infants months girls adult cat patients today adults year member majority niece babies students grandmother persons wife dog organization
cluster 2	vaxers decision mamas point news lol statement case report sense
cluster 3	brain group pharma workers government world condition companies stuff man judge idea society lawyer manufacturer public body members faith ladies question system author bodies life poster company mine individual systems researchers
cluster 4	books info story link thread information book
cluster 5	shot vaccinations dose doses vaccines shots vaxes vaccine vaccination vax antibiotics
cluster 6	thimerosal bacteria strains amount viruses mercury cells antibodies
cluster 7	records sources medicine schedule care type treatment form drugs exemptions sort exemption test blood letter exposure thought
cluster 8	years day program tests couple bit luck research state reason times place lot kind work matter thing daycare
cluster 9	breastfeeding results conditions bill factors experience media cases problem evidence breastmilk immunity status levels response
cluster 10	authors risks issues folks risk religion issue side number vaxing rates part science situation beliefs numbers things posters fact laws
cluster 11	paper list articles page posts article post study studies stories links data
cluster 12	mother father parents dad woman dh moms lady mothers brother husband people parent sister family friends mama friend families physician nurses midwife authorities person
cluster 13	oldest children ds kids daughter son kid boys child baby babe dd
cluster 14	increase symptoms effects reactions autism damage problems reaction
cluster 15	board forum website places site countries
cluster 16	guys docs guy peds officials lots scientists patient profession community district manufacturers population providers experts pediatricians age
cluster 17	rest clinic schools school practice country hospital states hospitals law
cluster 18	pox disease cough tetanus infection flu illnesses illness diseases polio measles fever virus infections
cluster 19	office vet dr ped mom pedi nurse teacher pediatrician doc doctor doctors
cluster 20	dogs animals foods diet

4.2 Transactional Relations on Twitter

4.2.1 Motivation

Over the last few years, there has been a growing public and enterprise interest in social media. Service Providers, manufacturers and merchants allow customers to write reviews and express their experience with their service or product in an on-line platform, being it a designated public page in a social media platform like Facebook or an e-merchant website like Amazon.

This has lead to a paradigm shift in ways parties in an online transactional environment interact. From the perspective of the service provider, social media data can dramatically improve business intelligence, for branding and awareness, customer engagement, user experience analysis and improving customer service. From customers' perspective, social media can help making informed choices based on a more comprehensive view of the pros and cons of the service or product. It helps also as a venue to compare with the product or service similar competitors. Moreover, social media can be used as a troubleshooting platform as other customers with similar issues might have shared their experience. Finally, from the perspective of the investors and new players, monitoring customer reviews in social media helps with understanding potential failure situations, product popularity problems, as well as to be alerted against potential threats for Investment.

This new paradigm has opened up opportunities for understanding interactions in a transactional setting to guide building socially-aware systems. For example, in behavioral economics studies the correlation between public mood, financial rumors or news stories and economic indicators. In Data analytics and business intelligence, it provides opportunities for developing computational methods for monitoring marketing activities, consumer opinion, influencers, competitors, brands, investment, market prediction, and aggregation of events from user-generated content.

The main challenge in turning such data into business intelligence is scalability with customer engagement. It is practically impossible to curate a set of hand picked reviews to

represent a holistic view of customers' opinion. Moreover, it is not clear how to distinguish between objective comments versus subjective ones that involve experiences of individuals that may not necessarily reflect the overall opinion of the customers. Therefore, there is an indispensable need for large scale methods that can aggregate pieces of information present in different reviews and provide a holistic view of the customers' opinion.

Such summarization can help with opinion mining with various levels of granularity. As such, one can obtain a general assessment of sentiment polarity regarding a particular product or service, which can be invaluable for marketing or reputation management. A more granular objective would be to target specific query-based information, such as "Which particular features do customers like best about a given product?" Therefore, having a structured representation of the transactional relationships in the form of a summary network between the involved entities with connections that instantiate transactional relationships is necessary.

Most of the studies on opinion mining in microblogging settings involve limited views of the data. For example, Turney [82] proposes an unsupervised method for classifying reviews in a consumer platform. Popescu and Etzioni [83] proposed a method for opinion phrase extraction based on the semantic orientation of words. This system can then be used for entity-level classification with extraction rules based on sentence level rules. Hu et al. [84] develop an extraction method that identifies sentences that contain one or more product features and characterize the polarity of the opinion sentences using the adjective set from WordNet. Along the same line of work [85] aim to identify noun phrases referred to the problem target in the sentence by first identifying what phrases potentially contain information about the problem then finding possible targets using the set of nouns for a given problem expression.

In this work, we introduce an automated and scalable machine learning framework which uses entity/relationship extractions to summarize customer reviews. In a transactional setting, where there only a limited types of entities involved in the story and their interactions are rather limited to a set of known relationship types, the summarization can be cast as an instance of entity/relation typing problem, where the problem is to determine the type of unknown entities rather than the harder problem of grouping similar entities into clusters

without taking into consideration that there are only a small number of types involved in the problem.

In order to show the merits of our entity resolution method, we focus on the harder problem of clustering the entities based on the extracted relationships. In the typing problem, we need to have a set of seed entities with known type labels to disseminate the type information to the unknown entities, while in the clustering problem, the partitioning is independent of any known types and the cluster labels are up to interpretation. Note that a clustering partition of the entity mentions can be turned into a typing result in presence of partial type labels by taking the majority type in the cluster.

4.2.2 Data

The data mainly consist of objective tweets which are describing interactions among banks and mobile payments, or merchants and mobile payments. Tweets such as:

- "Barclays works with #Apple- Pay", "Tried using Apple Pay at McDonalds. Didn't work. Rubbish."
- "Starbucks will soon accept Apple Pay #news #tech".
- "chrisdrackett I wish ApplePay would work with my paypal business card. :("
- "roymartin gav Just tried ApplePay in McDonaldsUK. Amex card didn't work so had to pick a different card from Passbook instead.."
- "McDonaldsUK Twice in 1 wk I've tried to use ApplePay. Both times its registered on my device but not acknowledged on your machines. WHY?!?!"

Total number of 527K posts (including retweets) are crawled from twitter using keywords and hashtags around mobile payments such as Apple Pay and Samsung Pay. Among them, 202K tweets are unique. We applied typical text cleaning such as removing hashtags, links, non-ascii characters, fixing encoding issues, separating sentences and etc. Then, using nltk and stanford parser [80] we obtained POS tags and dependency parse trees that are used,

following the same approach described in the previous section, to extract relation triplets of the form (m_s, m_o, v), where m_s is the head word of the subject node in the parsing tree, m_o is the headword of the object node in the parsing tree, and v is the verb that connects the two. The following table provides a summary of the extractions $\mathcal{T} = \{(m_s, m_o, v)\}$ with our method.

Table 4.14: Summary statistics for relation extraction

Extraction	Number
Relation triplets($\|\mathcal{T}\|$)	70699
Unique subject entity mentions($\|\mathcal{M}_S\|$)	11354
Unique object entity mentions($\|\mathcal{M}_O\|$)	18602
$\|\mathcal{M}_S \cap \mathcal{M}_O\|$	1850
Unique relation phrases($\|\mathcal{V}\|$)	16188

4.2.3 Ground Truth Actants

In this study, we use the notion of entity type to refer to a set of entities that serve the same or similar purpose in the transactional ecosystem we are exploring. This is similar to the notion of actant in the story model, discussed in the previous section

Within our corpus three main categories of types can be identified:

- *Banks*
- *Contact-less Payment Methods*
- *Merchants*

In order to evaluate the performance of the clustering method, we manually curate a set of ground truth entities that belong to each type. This set is generated by querying the type of the first 200 most frequent word tokens from an external knowledge-base(specifically, Google's knowledge graph)

We complement this list by extracting all word tokens that contain along with their frequencies. We believe that frequent tokens of this type correspond to real-world entities of our interest. e.g. McDonalds, ApplePay. But our observation is that twitter accounts with very low number of mentions in the "main body" of tweets are often not entities and are simply individuals tweeting here and there. Therefore, in order to discard mentions in the "main body" of tweets that refer to individual accounts, we filter out those with a frequency below a threshold, namely 5. To be more accurate, we don't discard entities with just low frequencies though, instead we discard those with low total frequencies of their entire entity cluster. Read about our definition of entity tree in the following section. After some basic cleaning, such as discarding certain prefix and suffix tokens like country names, we query a knowledge-base to determine the type of most frequent tokens with character.

The following table provides a list of entity mentions with their associate type, obtained with the method explained above, which we will use as our ground truth set to evaluate our entity resolution method.

Table 4.15: Ground Truth Type Assignment

Entity types (Nodes)	Associated Word Set
Banks	barclaysuk, ulster, natwestbusiness, bankofamerica, simple, td, ns, bmo, arvest, barclays, hsbc, lloyds, ally, mbna, rbsgroup, rbs, amex, first-direct, natwest, nationwide, mastercard, halifax, discover, barclaycard, santanderuk, tsb, visa, americanexpress, chase
Payment Methods	worldpay, judopayments, paypal, changeit, izettle, coin, metro, apple-payinfo, ovchipkaart, samsungmobile, applepay, square, wocketwallet, bpay, googlewallet
Merchants	starbucks, costacoffee, pret, mcdonalds, chilis, subway, justeat, kfc, pub-lix, homedepot, meijer, greggsofficial, sparinthe, tesco, cvs, cooperative-food, morrisons, target, sainsburys, wholefoods, bestbuy, riteaid, wal-greens, coop, asda, lovewilko, bloomandwild, gpdb, snapdeal, waitrose, boots, paytm, mobikwik, verifone, applemusic, justpark, youtube, green-gro, tfl, arrayit, linkedin, verizon, freecharge, affinorgrowers, marksand-spencer, familyroomfilm

4.2.4 Entity-Relationship Matrices

Matrix constructions follow the exact same steps as in the previous section. Again, in construction of the left and right entity-relation matrices, we only keep the entity mentions that appear both as a subject in some relation tuple and an object in some other relation tuple and filter out the entities that appear in less than 2 relation tuples in $\mathcal{T}$.

After this filtering, we have

$$\boldsymbol{X}_L \in \mathbb{R}^{92\times 629}, \text{ and}$$
$$\boldsymbol{X}_R \in \mathbb{R}^{92\times 562},$$

which means that $|\mathcal{M}| = 92, |\mathcal{V}_L| = 629$ and $|\mathcal{V}_R| = 562$.

4.2.5 Matrix Estimation results

We compare the performance of the three algorithms studied in the previous section, namely NMF-MUL, NMF-GRAD, and NMF-EXT, to solve the non-negative matrix factorization problem 4.1 on both the left and right entity-relation ground truth matrices. In all the algorithms, the ambient dimension of the embedded space is set to $d = 20$.

Again, we use matrix estimation error $\epsilon_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}}, \boldsymbol{V}_R^{\mathcal{A}})$ and the sparsity of the factor matrices $\kappa_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}})$ as measures of evaluation. We run each algorithm $\mathcal{A}$ in $n_s = 200$ instances and obtain a sample of size n_s for the statistics of interest; namely $\epsilon_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}}, \boldsymbol{V}_R^{\mathcal{A}})\}$ and $\{\kappa_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}})\}$.

Tables below represent mean value, minimum value, maximum value and a 95% confidence interval length for the samples $\{\epsilon_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}}, \boldsymbol{V}_R^{\mathcal{A}})\}$ and $\{\kappa_{\mathcal{A}}(\boldsymbol{U}_R^{\mathcal{A}})\}$.

Table 4.16: Summary statistics for prediction accuracy and sparsity of NMF-MUL algorithm for the right entity relation matrix X_R

NMF-MUL	Prediction Error	Sparsity
mean	17.936	0.4655
confidence interval size	0.0427	0.0041
minimum	17.560	0.4207
maximum	18.545	0.5264

Table 4.17: Summary statistics for prediction accuracy and sparsity of NMF-GRAD algorithm for the right entity relation matrix X_R

NMF-GRAD	Prediction Error	Sparsity
mean	17.911	0.4891
confidence interval size	0.0356	0.0037
minimum	17.583	0.4443
maximum	18.383	0.5405

Table 4.18: Summary statistics for prediction accuracy and sparsity of NMF-EXT algorithm for the right entity relation matrix X_R

NMF-EXT	Prediction Error	Sparsity
mean	17.741	0.6615
confidence interval size	0.0274	0.0408
minimum	17.558	0.0
maximum	18.146	0.7415

To compare the approximation error of the NMF-optimization algorithms, we compare the prediction errors in the above table with the optimal SVD error. In this case

$$\min_{\substack{\boldsymbol{U}_R \in \mathbb{R}^{|\mathcal{M}| \times d} \\ \boldsymbol{V}_R \in \mathbb{R}^{N_R \times d}}} \quad ||\boldsymbol{X}_R - \boldsymbol{U}_R \boldsymbol{V}_R^T||_F^2 = 15.943$$

Again, we observe that the factor matrices obtained by the NMF-EXT are by far sparser than those of the other algorithms. Moreover, the prediction error of the NMF-EXT is "significantly" lower than that of the other two algorithms. It can also be observed that NMF-GRAD performs negligibly better than NMF-MUL in terms of prediction accuracy but "significantly" better in terms of sparsity.

Next, we repeated the same experiments on the left entity-relation matrix and the results are summarized in the following tables.

Table 4.19: Summary statistics for prediction accuracy and sparsity of NMF-MUL algorithm for the left entity relation matrix X_L

NMF-MUL	Prediction Error	Sparsity
mean	17.137	0.1928
confidence interval size	0.0210	0.0023
minimum	16.942	0.1688
maximum	17.435	0.2169

Table 4.20: Summary statistics for prediction accuracy and sparsity of NMF-GRAD algorithm for the left entity relation matrix X_L

NMF-GRAD	Prediction Error	Sparsity
mean	17.072	0.2082
confidence interval size	0.0206	0.0018
minimum	16.904	0.1886
maximum	17.380	0.2349

Table 4.21: Summary statistics for prediction accuracy and sparsity of NMF-EXT algorithm for the left entity relation matrix X_L

NMF-EXT	Prediction Error	Sparsity
mean	17.092	0.5425
confidence interval size	0.0211	0.0429
minimum	16.914	0.0
maximum	17.493	0.6433

In order to compare the reconstruction error of the NMF algorithms with that of the optimal SVD, we note that

$$\min_{\substack{\boldsymbol{U}_L \in \mathbb{R}^{|\mathcal{M}| \times d} \\ \boldsymbol{V}_L \in \mathbb{R}^{N_L \times d}}} ||\boldsymbol{X}_L - \boldsymbol{U}_L \boldsymbol{V}_L^T||_F^2 = 14.466.$$

It can be clearly observed that the factors obtained by NMF-EXT algorithm are sparser than those of the other two algorithms. In terms of prediction accuracy, it can be observed that the NMF-GRAD algorithm does better than the other two; while NMF-EXT performs "significantly" better than the NMF-MUL.

4.2.6 Entity Clustering results

In this section we demonstrate the quality of the embeddings obtained by the matrix estimation algorithms by evaluating whether clustering the entity mentions with those embeddings can recover the the actant groupings in table 4.15. We note that a smaller subset(of size 54) of the entities present in the actant groupings are captured in the relationships we have extracted.

We compare the clustering performance of the NMF-based algorithms as well as the SVD factorization. As for the clustering method, We use K-means clustering with $K = 20$ clusters. For each algorithm, we have collected an independent sample of size $n_s = 200$ on the statistics of interest, that is set divergence measures $D(\mathcal{C}; \mathcal{E})$ reviewed in the previous section.

Tables below represents the mean value, minimum value, maximum value and a 95% confidence interval length for $\{D^{(i)}(\mathcal{C}; \mathcal{E})\}_{i=1}^{n_s}$ for the various divergence measures discussed in the previous section.

Table 4.22: Summary of K-means clustering results for NMF-MUL algorithm with K=20

NMF-MUL	homogeneity	completeness	V-measure	Mutual Information
mean	0.364	0.139	0.202	0.225
confidence interval size	0.008	0.003	0.004	0.005
minimum	0.260	0.099	0.143	0.161
maximum	0.468	0.188	0.268	0.296

Table 4.23: Summary of K-means clustering results for SVD algorithm with K=20

SVD	homogeneity	completeness	V-measure	Mutual Information
mean	0.253	0.093	0.136	0.153
confidence interval size	0.006	0.002	0.003	0.004
minimum	0.179	0.064	0.095	0.108
maximum	0.342	0.128	0.187	0.209

Table 4.24: Summary of K-means clustering results for NMF-GRAD algorithm with K=20

NMF-GRAD	homogeneity	completeness	V-measure	Mutual Information
mean	0.384	0.146	0.212	0.237
confidence interval size	0.007	0.002	0.004	0.004
minimum	0.290	0.108	0.157	0.177
maximum	0.466	0.178	0.257	0.287

Table 4.25: Summary of K-means clustering results for NMF-EXT algorithm with K=20

NMF-EXT	homogeneity	completeness	V-measure	Mutual Information
mean	0.352	0.135	0.195	0.218
confidence interval size	0.007	0.003	0.004	0.004
minimum	0.280	0.108	0.156	0.175
maximum	0.492	0.190	0.274	0.305

In the following, we visualize the entity mention embeddings obtained by the NMF-EXT algorithm via the t-SNE mapping. It can be clearly seen that groups of entity mentions that refer to the same actant group tend to cluster together.

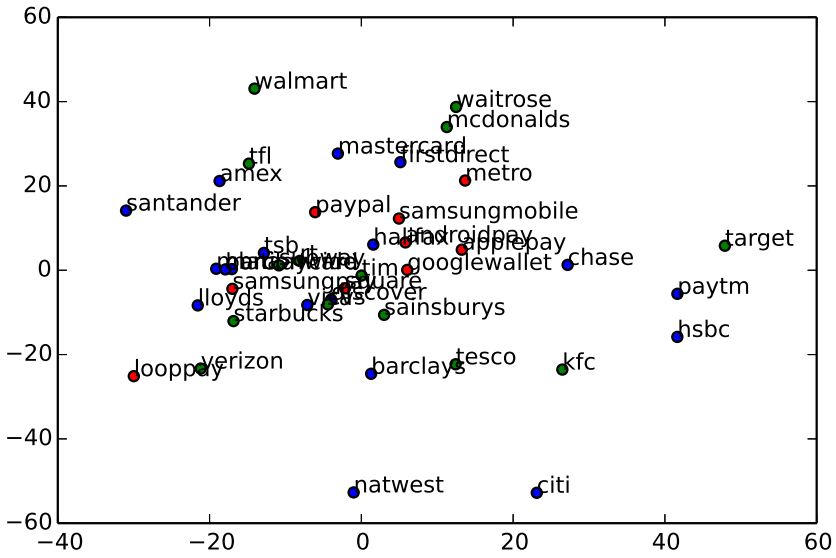

Figure 4.4: Entity Embeddings for twitter data

In the following table, the clusters by the K-means algorithm with $k = 15$ applied to embeddings obtained by NMF-EXT algorithm are listed. It can be clearly observed that clusters 1, 4 and 9 represent banks, clusters 2, 5, 11, 12, 13, and 14 represent merchants, cluster 3 represents mobile payment methods and the other clusters are mixtures of different actant groups, usually with an actant group dominating.

cluster 1	arvest santanderuk chase coin bpay
cluster 2	walgreens mcdonalds costacoffee starbucks subway
cluster 3	samsungpay googlewallet paypal applepay
cluster 4	barclays halifax lloyds hsbc
cluster 5	boots mbna safeway
cluster 6	citi waitrose amex wholefoods staples target
cluster 7	publix paytm td kroger samsungmobile
cluster 8	ulster simple chipotletweets wawa americanexpress asda morrisons verifone verizon santander firstdirect cvs barclaycard ally greggs coop walmart pret bestbuy metro rbs androidpay mastercard sainsburys
cluster 9	square discover visa nationwide mobikwik
cluster 10	gpdb arrayit greengro
cluster 11	kfc chilis
cluster 12	freecharge cooperativefood
cluster 13	tsb natwest
cluster 14	tfl
cluster 15	looppay tesco marksandspencer

CHAPTER 5

Detecting Changes in temporal Dynamics of the Stories

5.1 Motivation

So far, we have studied online stories through a holistic view of all the posts, trying to aggregate the partial information present in different posts, regardless of their timing. While such static aggregate view helps understanding an overview summary of the story, oblivious to the temporal dynamics in data. Characterizing the underlying dynamics in the evolution of the story narratives in a social media setting is not a well-defined problem in general. To make the problem concrete, let us assume that we are interested in detecting "major changes" in the temporal evolution of the story. These changes can reflect in the temporal evolution of the story narrative in terms of the textual content of the users' posts as well as the frequency of the activities irrespective of the textual content.

Consider, for example, an online micro-blog setting, where users post textual pieces about a specific subject or an event with a textual signature specific to it. As our running example, let us take Twitter as the online platform and let us refer to each post as tweet. Each tweet is indexed by the time it is posted. Let us also assume that each tweet contains a number of hashtags, which serve as textual signatures that contextualize the subject of the tweet. Therefore, the total number of tweets that contain a particular hashtag related to the subject or event can be regarded a measure of user activity. Monitoring the temporal evolution of this measure reveals different aspects of the dynamics of user behavior. Specifically, in this work, we aim at studying the statistical changes in the distribution of the number of activities on particular hashtags. This change in distribution can reflect occurrence of a major event

in the story.

In dealing with this problem, we can no longer assume that the whole batch of data is available to the learner for decision making; rather a short term decision on whether a change has occurred in the distribution of data is demanded upon observing new data points. Therefore, the problem should be cast as an online learning problem so that data can be analyzed sequentially. In a statistical framework, sequential analysis, which is also referred to as online hypothesis testing, is a form of statistical analysis where the sample size of the learning task is not fixed in advance and data is processed and evaluated as it is observed over time. Sampling further observations is then performed based on past observations up to a pre-defined stopping time that evaluates if a significant result is observed. In such a framework the goal is to optimize an objective function, while meeting constraints on the computational cost of the algorithm, precision of the learning task and the amount of data to be observed.

Another major challenge in detecting changes in the temporal evolution of the activities is the transient nature of such changes, meaning that the change has to be identified before change period, which is assumed to be short compared to the whole observation window, is over. Thus, we can view this problem as a characterizing a trade-off between quick detection of the change versus reliability of the decision.

We formulate this problem in a transient change point detection setting where the objective is to design a statistical test to detect the change, if present, based on the sequence observed so far with minimum expected delay and a controlled measure of false alarm. This problem finds applications in a variety of other fields as diverse as industrial quality control, intrusion detection, and on-line fault detection.

We first develop a theory for transient change point detection and in the last section, we show how this theory can be used to monitor the temporal evolution of user activities in twitter. In our case study, we analyze tweets containing hashtags related Super bowl 2015 for a span of 2 weeks before the game to a week after the game. We show how changes in the distribution of activities, which reflects major events in the game, can be detected quickly

and reliably.

5.2 Summary of prior art

In the classical quickest change detection problem, a sequence of random variables $\{X_i\}_{i\geq 1}$, monitored sequentially, undergoes a change in distribution at some unknown point ν. It is typically assumed that the random variables X_i are independent with a common probability density function f_0 for $i < \nu$ and with another common density f_1 for $i \geq \nu$. Both f_0 and f_1 are known to the observer. Framing the problem as a sequential hypothesis testing, a natural approach is to consider a nonrandomized stopping time with respect to the observed sequence so far. The objective is to design a statistical test to detect the change, if present, based on the sequence observed so far with minimum expected delay and a controlled measure of false alarm. In this setting, Lorden [86] formulated the problem considering minimization a measure of worst case expected delay under the so called *average run length*(ARL) constraint that the mean time to false alarm is bounded from below by a parameter γ. He established an asymptotic lower bound, in the asymptote of γ, on the worst case expected delay for all stopping times satisfying the ARL constraint and showed that the CuSum statistic proposed earlier by Page [87] achieves this lower bound. Moustakides [88] proved the optimality of CuSum rule beyond the asymptotic setting considered by Lorden, casting the problem into an optimal stopping time formulation. Later, Lai in his seminal paper [89] extended the asymptotic results of Lorden to the non i.i.d setting employing a change of measure argument. He also suggested new performance criteria and false alarm constraints, better suited for variations of the problem, as well as new detection procedures, namely window limited versions of CuSum and generalized likelihood ratio rules to attain the fundamental performance bounds he developed for different variations of the problem.

Yet another common formulation of the change point detection problem is to assume a prior distribution on the change time ν and cast the problem into a Bayesian setting. Shiryayev [90] formulated the problem in a Bayesian framework assuming a geometric prior distribution. He showed that a procedure based on threshold comparison of the posteriori probability that

a change has occurred is optimal in this setting. Inspired by [89], the asymptotic optimality of this procedure was extended to the non- i.i.d. case [91]. A survey of the results on different variations of the problem can be found in [92].

In many practical applications there are multiple data streams to be monitored and the changes are quite rare. Moreover, there might be a cost associated with taking observations. We refer to the observations taken from the sequence $\{X_i\}_{i\geq 1}$ as *samples* and the fraction of samples taken in a sampling scheme as its *sampling rate*. In a series of papers, Banerjee and Veeravalli [93, 94] formulated the quickest change point detection within both Bayesian and Minimax frameworks, considering an additional constraint on the sampling rate before the change time. They show in this setting, which they refer to as data-efficient change point detection, that natural variations of the optimal procedures under full sampling are optimal under sparse sampling.

While the problem is well-studied in the non-transient setting, much less is known when the change is transient. In the transient setting, $\{X_i\}_{i\geq 1}$ is a sequence of independent random variables where all random variables have the same density function f_0 except for a possible subsequence of length L starting at an unknown point ν, i.e. $\{X_i\}_{\nu}^{\nu+L-1}$, along which the random variables have the common density f_1. Inspired by the alternative criteria proposed in [89], the problem of quickest detection of transient changes is formulated in [95] as minimizing the worst-case probability of missed detection under a constraint on the false alarm rate in a given period. In [96] and [97], the problem is formulated within the framework of partially observed Markov decision processes under several performance criteria. The main challenge in this setting is to design the statistical test in such a way that it reacts to the change before it disappears. In an attempt to address the problem of detectability of a transient change with a given duration, the probability of detection under Page's test is examined in [98].

We study the problem of quickest detection of transient changes under a Minimax formulation [99]. A fundamental problem of interest is to determine the smallest duration of a change detectable with an ARL constrained sequential test. That is, given the constraint that the expected time to false alarm be at least γ, what is the minimum duration of a change that can

be detected *reliably*, when γ tends to infinity. Next, given a transient change with duration greater than the asymptotic minimum duration specified earlier, we seek to determine the smallest sampling rate under which a transient change can be detected as quick and reliable as in the full sampling regime. We address these two questions leveraging known results for the non-transient setting.

5.3 Non-Transient Change Detection under Full Sampling

5.3.1 Problem Statement

Let us first review the classical change point detection problem for an independent random process $\{X_i\}_{i\geq 1}$ defined over a finite alphabet $\mathcal{X}$, where all the random variables before an unknown instant ν, so called the change point, have the common density function f_0, while all random variables $\{X_i\}_{i\geq\nu}$ have the common density function f_1; that is

$$X_i \sim \begin{cases} f_0 & \text{if } 1 \leq i < \nu \\ f_1 & \text{if } \nu \leq i \ . \end{cases} \tag{5.1}$$

The problem of interest is to detect the change point with a possibly small detection delay and a controlled false alarming reaction.

5.3.2 Characterizing Minimum Delay

Framing the problem as a sequential hypothesis testing, a natural approach is to consider a non-randomized stopping time τ with respect to the observed sequence so far. In the setting where no assumption is made on the prior distribution of the change point, Lorden [86] proposed a Minimax formulation of the problem as to minimize a measure of worst case expected delay, while constraining expected time to false alarm being large; that is minimizing

$$\bar{\mathbb{E}}_1[\tau] = \sup_{\nu\geq 1} \operatorname{ess\,sup} \mathbb{E}_\nu[(\tau-\nu+1)^+ | X_1, X_2, \cdots, X_{\nu-1}], \tag{5.2}$$

over all stopping times τ satisfying

$$\mathbb{E}_\infty[\tau] \geq \gamma, \tag{5.3}$$

where $\mathbb{E}_\nu[\cdot]$ denotes expectation under $\mathbb{P}_\nu$ which is the probability measure when the change occurs at time ν. Lorden showed that asymptotically as $\gamma \to \infty$, for any $\epsilon > 0$ and for any stopping time τ satisfying (5.3),

$$\bar{\mathbb{E}}_1[\tau] \geq \sup_{\nu \geq 1} \mathbb{E}_\nu[\tau - \nu | \tau \geq \nu] \geq (1-\epsilon)\frac{\log \gamma}{D(f_1||f_0)} \tag{5.4}$$

where $D(f_1||f_0) = \mathbb{E}_1 \log \frac{f_1(X_i)}{f_0(X_i)}$. Then, he showed that the stopping time based on the so called CuSum statistic, defined in the following, achieves this lower bound asymptotically.

Definition 5.3.1. *The CuSum procedure is defined as*

$$\tau_\gamma^* = \inf\{n | \max_{k \leq n} \sum_{i=k}^{n} Z_i \geq c\}, \tag{5.5}$$

where $Z_i = \log \frac{f_1(X_i)}{f_0(X_i)}$*, and* c *is chosen appropriately such that* $\mathbb{E}_\infty[\tau^*] \geq \gamma$*.*

Thus,

$$\inf\{\bar{\mathbb{E}}_1[\tau] | \mathbb{E}_\infty[\tau] \geq \gamma\} \sim \mathbb{E}_1 \tau_\gamma^* \sim \frac{\log \gamma}{D(f_1||f_0)}. \tag{5.6}$$

Later, Lai extended this result for the non independent identically distributed setting using a natural generalization of the CuSum rule in [89], where he introduced new performance criteria which also provide insights on the other variations of the problem. The main ingredient in [89] to prove (5.6) is the following important observations which stands as the core of asymptotic optimality of the CuSum rule for minimizing (5.2). For any stopping time τ satisfying (5.3) and any $\epsilon > 0$,

$$\mathbb{P}_\nu(\tau - \nu > (1-\epsilon)\frac{\log \gamma}{D(f_1||f_0)} | \tau > \nu) = 1 + o(1). \tag{5.7}$$

where the $o(1)$ term is uniform over all ν.Moreover, for the stopping time τ_γ^* defined in (5.35),

$$\mathbb{P}_\nu(\tau_\gamma^* - \nu \leq (1+\epsilon)\frac{\log \gamma}{D(f_1||f_0)} | \tau_\gamma^* > \nu) = 1 + o(1), \tag{5.8}$$

for all ν.

These observations provide fundamental insight on detectability of transient changes, suggesting an asymptotic threshold on the smallest duration of the change detectable reliably by a stopping random variable satisfying (5.3).

5.4 Transient Change Point Detection Under Sparse Sampling

5.4.1 Problem Statement

Suppose now that the random process $\{X_i\}_{i\geq 1}$ is distributed as follows

$$X_i \sim \begin{cases} f_0 & \text{if } 1 \leq i < \nu \text{ or } i \geq \nu + L \\ f_1 & \text{if } \nu \leq i < \nu + L \ . \end{cases} \tag{5.9}$$

The sequence $\{X_i\}_{i\geq 1}$ is observed sequentially according to a sampling strategy

Definition 5.4.1. *A Sampling Strategy with respect to a random process $\{X_i\}_{i\geq 1}$ is an ordered collection of random time indices $\mathcal{S} = \{S_1, S_2, \cdots\}$, at which samples are taken from the sequence. In general, the decision as to whether take a sample at a certain time instance or not depends on the past samples. Since the change time is unknown, without loss of generality we can start sampling from the first time instance of the sequence, that is $S_1 = 1$. For any $n \geq 2$,*

$$S_n = \Phi_n(\{X_{S_i}\}_{\{i<n\}}) \tag{5.10}$$

where $\Phi_n : \mathcal{X}^{n-1} \to \{S_{n-1} + 1, S_{n-1} + 2, \cdots\}$ is a decision function at time n.

The objective is to detect the change efficiently, in a suitable sense, with a so called decision policy Ψ which consists of a sampling strategy $\mathcal{S}$ and a stopping time τ with respect to the sampled sequence so far, that is $\Psi = (\mathcal{S}, \tau)$. A decision policy is evaluated with a measure of detection delay and its sampling rate constrained on a measure of false alarm. As in the non-transient setting, we consider the class of all decision policies satisfying the ARL constraint.

Definition 5.4.2 (False Alarm Constraint)**.** *An ARL constrained decision policy $\Psi = (\mathcal{S}, \tau)$ is such that*

$$\mathbb{E}_\infty[\tau] \geq \gamma. \tag{5.11}$$

Definition 5.4.3 (Detection Delay)**.** *For a decision policy $\Psi = (\mathcal{S}, \tau)$, and $\epsilon > 0$, the worst case minimum delay in probability is defined as*

$$d(\Psi, \epsilon) = \min \left\{ \ell : \sup_{\nu \geq 1} \mathbb{P}_\nu(\tau - \nu > \ell | \tau \geq \nu) \leq \epsilon \right\} \tag{5.12}$$

We will later argue that in the transient setting a measure of worst case delay in probability is more appropriate to be adopted compared to the measure of worst case expected delay defined earlier in (5.2).

Definition 5.4.4 (Sampling Rate)**.** *For a decision policy $\Psi = (\mathcal{S}, \tau)$, the pre-change sampling rate is defined as*

$$\begin{aligned} \rho(\Psi) &= \limsup_n \mathbb{E}_n\Big[\frac{|\mathcal{S}^{(n)}|}{n}\Big|\tau \geq n\Big] \\ &= \limsup_n \mathbb{E}_\infty\Big[\frac{|\mathcal{S}^{(n)}|}{n}\Big|\tau \geq n\Big], \end{aligned} \tag{5.13}$$

where $|\mathcal{S}^{(n)}| \triangleq \{i \in \mathcal{S} | i \leq n\}$ is the number of samples taken up to time n.

Definition 5.4.5 (Achievable Sampling Rate)**.** *Let $\{\rho_\gamma\}_{\gamma>0}$ be an indexed family with $0 \leq \rho_\gamma \leq 1$. Sampling rates $\{\rho_\gamma\}_{\gamma>0}$ are achievable with respect to an indexed family of change durations $\{L_\gamma\}_{\gamma>0}$, if there exists an indexed family of decision rules $\{\Psi_\gamma = (\mathcal{S}_\gamma, \tau_\gamma)\}_{\gamma>0}$, such that for γ large enough,*

1. *The ARL constraint $\mathbb{E}_\infty[\tau_\gamma] \geq \gamma$ is satisfied,*
2. *The sampling rates satisfy $\rho(\Psi_\gamma) \leq \rho_\gamma$,*
3. *The delay satisfies*

$$d(\Psi_\gamma, \epsilon_\gamma) \leq L_\gamma,$$

for some indexed family $\{\epsilon_\gamma\}_{\gamma>0}$ such that $\lim_{\gamma\to\infty} \epsilon_\gamma = 0$.

5.4.2 Notational Convention

When clear from the context, we represent an indexed family with its representative element, e.g. we denote $\{\rho_\gamma\}_{\gamma>0}$ simply as ρ_γ. Moreover, we will use d_γ instead of $d(\Psi_\gamma, \epsilon_\gamma)$, leaving out any explicit reference to the decision policy Ψ_γ and to the indexed family $\{\epsilon_\gamma\}_{\gamma>0}$ which we assume satisfies $\lim_{\gamma\to\infty} \epsilon_\gamma = 0$, unless it is necessary to make the sampling strategy or the stopping time explicit.

5.4.3 Minimum Duration of a Reliably Detectable Change

Theorem 5.4.1 (Transient Change under Full Sampling)**.** *Under full sampling ($\rho_\gamma = 1$),*

(i) Let $\alpha > 1$. suppose $L_\gamma \geq \alpha \frac{\log \gamma}{D(F_1||F_0)}$. Then ρ_γ is achievable with respect to L_γ. Moreover, for decision policies Ψ^*_γ with stopping time τ^*_γ defined in (5.35) and full sampling strategy, we have

$$d(\Psi^*_\gamma) \sim \frac{\log \gamma}{D(f_1||f_0)} \tag{5.14}$$

(ii) Let $0 \leq \alpha < 1$. Suppose that $L_\gamma \leq \alpha \frac{\log \gamma}{D(f_1||f_0)}$. Then ρ_γ is not achievable with respect to L_γ. Moreover, in this case

$$\liminf_{\gamma \to \infty} \frac{d_\gamma}{\gamma^{1-\alpha}} \geq 1 \tag{5.15}$$

Remark 1. *Theorem 1 establishes an asymptotic threshold on the minimum duration of a change that can be detected reliably. Specifically, for γ large enough, if the duration of the change is above $\frac{\log \gamma}{D(f_1||f_0)}$, the delay is as short as if the change had infinite duration. Henceforth we call such transient changes asymptotically detectable. For transient changes with duration below this threshold, delay grows as a polynomial function of γ. The lower bound on the asymptotic worst case delay in probability in (5.17) can be converted to a lower bound on the worst case expected delay defined in (5.2). Note, however, that the guarantee provided on the asymptotic worst case delay in (5.14) cannot necessarily be translated to a guarantee on the worst case expected delay. This is because when the event $\{\tau^*_\gamma > \nu + L_\gamma\}$ occurs, although happening with a vanishing probability, the delay can be arbitrarily large, as the rest of the observations are f_0 distributed, which leads the expected delay to grow unbounded.*

5.4.4 Minimum Sampling Rate

The following two theorems characterize the minimum achievable sampling rates with respect to duration of detectable transient changes. Theorem 3 is proved using the so called DE-CuSum decision policy proposed in [94], which was used to achieve any constant sampling rate and asymptotically the same worst case expected delay as under full sampling in a non-transient scenario [94]. A brief description of the DE-CuSum rule is provided in the next

section right before proof of Theorem 3. For a more detailed discussion on the description of the algorithm and the guarantees it provides we refer the reader to [94].

Theorem 5.4.2 (Minimum Asymptotic Achievable Rate [94])**.** *Let $\alpha > 1$. suppose that $L_\gamma \geq \alpha \frac{\log \gamma}{D(f_1||f_0)}$. Sampling rates $\rho_\gamma = \omega(\frac{1}{\log \gamma})$ are achievable with respect to any L_γ. Moreover, as $\gamma \to \infty$, the delay satisfies*

$$d(\widehat{\Psi}_\gamma) \sim \frac{\log \gamma}{D(f_1||f_0)}, \tag{5.16}$$

where $\widehat{\Psi}_\gamma$ is the DE-CuSum procedure.

Remark 2. *Note that in [94], authors are only interested in constant sampling rates while we are interested in sampling rates that scale with γ as does the duration of the change. In this case, the step size of the DE-CuSum procedure in the idle regime is not a constant but a function of γ the same scaling behavior as the sampling rate ρ_γ.*

In the following theorem, we assume that the indexed family L_γ corresponds to durations of some asymptotically detectable changes and show that sampling rates $\rho_\gamma = o(\frac{1}{L_\gamma})$ are not achievable.

Theorem 5.4.3 (Converse to Theorem 5.4.2)**.** *Let $\alpha > 1$. suppose that $L_\gamma \geq \alpha \frac{\log \gamma}{D(f_1||f_0)}$. Consider sampling rates $\rho_\gamma = o(\frac{1}{L_\gamma})$ with respect to L_γ. Then, for any decision policy Ψ_γ, satisfying the false alarm and the sampling rate constraints in Definition 6, Ψ_γ only takes samples from f_0(completely misses the change), with probability bounded away from zero. Moreover, in this case*

$$\liminf_{\gamma \to \infty} \frac{d_\gamma}{\gamma} \geq 1 \tag{5.17}$$

Corollary 5.4.1. *Considering detectable changes with $L_\gamma = \theta(\log \gamma)$, sampling rates $o(\frac{1}{\log \gamma})$ are not achievable.*

5.5 Proofs

1. Suppose that there exists some $\delta > 0$ such that $L_\gamma \geq (1+\delta)\frac{\log\gamma}{D(f_1||f_0)}$ for some $\delta > 0$, and consider the CuSum rule

$$\tau_\gamma^* = \inf\{n | \max_{k\leq n} \sum_{i=k}^{n} Z_i \geq \log\gamma\}. \tag{5.18}$$

It follows that for any $0 < \epsilon \leq \delta$

$$\begin{aligned} & \mathbb{P}_\nu(\tau_\gamma^* - \nu > (1+\epsilon)\frac{\log\gamma}{D(f_1||f_0)} | \tau_\gamma^* > \nu) \\ \leq \; & \mathbb{P}_\nu\left(\bigcap_{\nu\leq n\leq \nu+(1+\epsilon)\frac{\log\gamma}{D(f_1||f_0)}} \left\{\max_{k\leq n}\sum_{i=k}^{n} Z_i < \log\gamma\right\}\right) \\ \leq \; & \mathbb{P}_\nu\left(\sum_{i=\nu}^{(1+\epsilon)\frac{\log\gamma}{D(f_1||f_0)}} Z_i < \log\gamma\right) \underset{\gamma\to\infty}{\longrightarrow} 0, \end{aligned} \tag{5.19}$$

where the first inequality follows from the definition of the CuSum rule in (5.18) and the last step follows from applying the law of large numbers to the sequence of i.i.d. random variables $\{Z_i\}_{i\geq\nu}$ with mean $D(f_1||f_0)$.

Note that (5.19) establishes an upper bound on the detection delay. Moreover, by causality of the stopping time random variables, (5.7) gives a lower bound on the detection delay, which combined with the upper bound yields the desired result.

2. Now suppose that $L_\gamma \leq (1-\delta)\frac{\log\gamma}{D(f_1||f_0)}$ for some $\delta > 0$. Let $\Psi_\gamma = (\mathcal{S}_\gamma, \tau)$ be any decision policy satisfying the false alarm constraint $\mathbb{E}_\infty[\tau] \geq \gamma$. Since $\mathbb{E}_\infty[\tau] \geq \gamma$, it follows [Proof of Theorem 1 in [89]] that for any integer $m < \gamma$, there is some $\nu \geq 1$ such that

$$\mathbb{P}_\infty(\tau \geq \nu) > 0, \text{ and } \mathbb{P}_\infty(\tau < \nu + m | \tau \geq \nu) \leq \frac{m}{\gamma}. \tag{5.20}$$

Let m be the largest integer less than $2\gamma^{\delta-\epsilon}$ for some $0 \leq \epsilon < \delta$. Define the events

$$\mathcal{C}_\epsilon = \left\{0 \leq \tau - \nu \leq \gamma^{\delta-\epsilon}, \quad \sum_{i=\nu}^{\min\{\tau,\nu+L_\gamma-1\}} Z_i < (1-\epsilon)\log\gamma\right\},$$

and

$$\mathcal{C}'_{\epsilon} = \left\{ 0 \leq \tau - \nu \leq \gamma^{\delta-\epsilon}, \quad \sum_{i=\nu}^{\min\{\tau,\nu+L_\gamma-1\}} Z_i \geq (1-\epsilon)\log\gamma \right\}.$$

Following the same lines as that of Lai's change of measure argument (Proof of [89, Theorem 1]), we have:

Claim 1. *As long as* $\delta > 2\epsilon > 0$,

$$\mathbb{P}_\nu(\mathcal{C}_\epsilon | \tau \geq \nu) \underset{\gamma\to\infty}{\longrightarrow} 0. \tag{5.21}$$

Claim 2.

$$\mathbb{P}_\nu \left\{ \mathcal{C}'_{\epsilon} \mid \tau \geq \nu \right\} \underset{\gamma\to\infty}{\longrightarrow} 0. \tag{5.22}$$

Combining Claims 1 and 2, we get $\mathbb{P}_\nu \left\{ \tau - \nu \leq \gamma^{\delta-\epsilon} | \tau \geq \nu \right\} \to 0$, as γ tends to infinity, for ν given in (5.20). Since ϵ can be made arbitrarily small, it follows that

$$\sup_{\nu>0} \mathbb{P}_\nu \left\{ \tau - \nu > \gamma^\delta | \tau \geq \nu \right\} \to 1, \tag{5.23}$$

which in turn implies

$$\liminf_{\gamma\to\infty} \frac{d_\gamma}{\gamma^\delta} \geq 1,$$

as desired.

5.5.1 Description of DE-CuSum decision rule

Let us briefly review the DE-CuSum detection rule $\widehat{\Psi}$ proposed in [94]. Define the sampling indicator random variable M_i as being 1 when the time instance i is sampled, and zero otherwise. Start with $D_0 = 0$ and fix $\gamma > 0$, $\mu(\gamma) > 0$ and $h > 0$. For $n \geq 0$, define the following sampling strategy

$$M_{n+1} = \begin{cases} 1 & \text{if } D_n \geq 0 \\ 0 & \text{Otherwise.} \end{cases}$$

Also, define the stopping time as

$$\widehat{\tau}(\gamma) = \inf\{n \geq 1 | D_n > \log\gamma\}. \tag{5.24}$$

At each step, the statistic D_n is being updated as follows

$$D_{n+1} = \begin{cases} \min\{D_n + \mu, 0\} & \text{if } M_{n+1} = 0 \\ (D_n + Z_{n+1})^{h+} & \text{if } M_{n+1} = 1 \end{cases} \tag{5.25}$$

where $(x)^{h+} = \max\{x, -h\}$. In fact the algorithm naturally performs a hypothesis test between the distributions f_1 and f_0 and skips the samples while the statistic is below 0. As long as $D_n < 0$, which depends on the last undershoot from 0, samples are skipped and D_n is being updated by adding the deterministic increment μ to D_n. The truncation to $-h$ is just to avoid large undershoots and is imposed for the ease of analysis.

Note that the DE-CuSum rule consists of a sequence of two sided tests; where each two-sided test contains a sequential probability ratio test (SPRT) and a possible sojourn below zero. Therefore, the stopping time of a two sided test in DE-CuSum rule is

$$\Lambda_\gamma = \lambda_\gamma + \lceil (D_{\lambda_\gamma})^{h+} \rceil \mathbf{1}_{\{D_{\lambda_\gamma} < 0\}} \tag{5.26}$$

where, $\lambda_\gamma = \inf\{n \geq 1 | \sum_{i=k}^{n} Z_i \notin [0, \log \gamma]\}$.

5.5.2 Proof of Theorem 2

We prove this theorem by the DE-CuSum procedure $\widehat{\Psi}_\gamma$ described earlier. It is shown in [94] that the sampling constraint $\rho(\widehat{\Psi}_\gamma) \leq \rho_\gamma$ is met if

$$\mu_\gamma < K \frac{\rho_\gamma}{1 - \rho_\gamma}, \tag{5.27}$$

where $K = \frac{\mathbb{E}_\infty[|Z_1^{h+}||Z_1<0]\ \mathbb{P}_\infty(Z_1<0)^2}{\mathbb{E}_\infty[\lambda_\infty]}$ is a constant that does not scale with γ [94].

Using standard arguments, as it is shown in [94], the asymptotic worst case delay in probability of the DE-CuSum rule is bounded from above as follows

$$\limsup_{\gamma \to \infty} d_\gamma(\widehat{\Psi}) \leq \frac{\log \gamma}{D(f_1||f_0)} + K' \frac{1}{\mu_\gamma} + K'', \tag{5.28}$$

where $K' = \frac{\mathbb{E}_\infty[|D_{\lambda_\infty}^{h+}|]}{\mathbb{P}_1(D_{\lambda_\infty}>0)\mathbb{P}_1(Z_1<0)} + h$ and $K'' = 2 + \frac{1}{\mathbb{P}_1(D_{\lambda_\infty}>0)}$ are constants which do not scale with γ [94].

Given a sampling rate constraint $\rho_\gamma = \omega(\frac{1}{\log \gamma})$, by setting the step parameter for the sojourn time in the DE-CuSum procedure as $\mu_\gamma = \theta(\rho_\gamma)$ such that (5.27) is satisfied, the desired result follows by considering (5.28) and the lower bound (5.7).

5.6 Proof of Theorem 3

We show that for any decision policy $\Psi_\gamma = \{S_\gamma, \tau\}$ satisfying the false alarm constraint $\mathbb{E}_\infty[\tau] \geq \gamma$ and the sampling rate constraint $\rho(\Psi_\gamma) \leq \rho_\gamma$ with $\rho_\gamma = o(\frac{1}{L_\gamma})$, there exists a time interval of duration L_γ such that, with probability bounded away from zero, no point within this interval is sampled by Ψ_γ, for γ sufficiently large.

First note that since Ψ_γ satisfies the false alarm constraint $\mathbb{E}_\infty[\tau] \geq \gamma$, we get

$$\begin{aligned} \mathbb{P}_\infty(\tau \geq \frac{\gamma}{2}) &\geq \mathbb{P}_\infty(\tau \geq \frac{\gamma}{2} + \nu) \\ &> \frac{1}{2}\mathbb{P}_\infty(\tau \geq \nu) > 0, \end{aligned} \tag{5.29}$$

where the last inequalities hold for some $\nu \geq 1$ by (5.20) with $m = \frac{\gamma}{2}$.

The proof is based on the fact that as long as the samples are drawn from the distribution f_0, the sampling rate constraint guarantees existence of sufficiently long gaps among the sampling times. Note that if no change occurs at all, the sampling rate constraint implies

$$\mathbb{E}_\infty\Big[\frac{|\mathcal{S}^{(\gamma/2)}|}{\gamma/2}\Big|\tau \geq \frac{\gamma}{2}\Big] = o(\frac{1}{L_\gamma}). \tag{5.30}$$

Divide the time frame up to time $\nu = \frac{\gamma}{2}$ into consecutive intervals of size L_γ. For each time interval, define the following indicator random variable

$$K_i = \begin{cases} 1 & \text{If at least one point from } i^{th} \text{ interval is sampled} \\ 0 & \text{Otherwise.} \end{cases}$$

We show that there exists an interval from which no sample is taken with probability bounded away from zero. In fact, we show that

$$\limsup_{\gamma \to \infty} \max_j \mathbb{P}_\infty(K_j = 0 | \tau \geq \frac{\gamma}{2}) = 1. \tag{5.31}$$

Otherwise,

$$\liminf_{\gamma\to\infty} \min_i \mathbb{P}_\infty(K_i = 1|\tau \geq \frac{\gamma}{2}) > 0, \tag{5.32}$$

which implies

$$\begin{aligned}
\mathbb{E}_\infty\Big[\frac{|\mathcal{S}^{(\gamma/2)}|}{\gamma/2}|\tau \geq \frac{\gamma}{2}\Big] &\overset{(a)}{\geq} \mathbb{E}_\infty\Big[\frac{2}{\gamma}\sum_{j=1}^{\lfloor\gamma/2L_\gamma\rfloor} K_j|\tau \geq \frac{\gamma}{2}\Big] \\
&= \frac{2}{\gamma}\sum_{j=1}^{\lfloor\gamma/2L_\gamma\rfloor} \mathbb{P}_\infty(K_j = 1|\tau \geq \frac{\gamma}{2}) \\
&= \theta(\frac{1}{L_\gamma}),
\end{aligned} \tag{5.33}$$

where (a) holds because more than one sample can be taken from a given interval. Note that (5.33) contradicts (5.30), establishing that (5.31) holds.

Combining (5.29) with (5.31) yields $\mathbb{P}_\infty(K_j = 0) > 0$ for some $j \geq 1$. Therefore, for any decision policy, there exists some interval of size L_γ from which no sample is taken with probability bounded away from zero.

Suppose now that a change of duration L_γ occurs along the sequence within the jth interval for which $\mathbb{P}_\infty(K_j = 0) > 0$. In such a case, with probability bounded away from zero, no point of the change period is sampled, meaning that only samples from the distribution f_0 are observed. In this case, since $\mathbb{E}_\infty[\tau] \geq \gamma$, it follows using an argument similar to part (ii) of theorem 1 that

$$\liminf_{\gamma\to\infty} d(\Psi_\gamma) \geq \frac{\gamma}{2}, \tag{5.34}$$

as desired.

5.7 Evaluation Results on Twitter Data

We analyzed a total of 3,138,823 tweets collected by querying popular hash-tags related to the 2015 Super Bowl spanning a period starting from almost 3 weeks before the game to a week after the game. We are interested in detecting "major changes" in the temporal evolution of the story, specifically changes in the distribution of number of activities. In our

study, we are oblivious to the textual content of the tweets except for the hash-tags that contextualize the tweets. Let $\mathcal{H}$ denote the set of hash-tags related to the event; that is

$$\mathcal{H} = \{\#\text{gopatriots}, \#\text{gohawks}, \#\text{nfl}, \#\text{patriots}, \#\text{sb49}, \#\text{superbowl}\}$$

Let N_t^h denote the number of tweets that contain hashtag $h \in \mathcal{H}$ at time t. In our study, based on the scope of the changes that we want to capture, the unit of time can vary. The following table summarizes some basic statistics on the hashtags in $\mathcal{H}$, with per hour time intervals

Table 5.1: Statistics on Super Bowl 2015 tweets Data

Hashtag	#gopatriots	#gohawks	#nfl	#patriots	#sb49	#superbowl
# tweets	26,232	188,136	259,024	489,713	826,951	1,348,767
Avg. # tweets/hr	38.38	193.54	279.55	499.42	1419.88	1401.24
Avg. # followers/twt	1618.53	2477.06	4864.89	3759.78	10496.06	10136.34
Avg. # followers/user	1594.14	1831.94	4346.95	2061.58	2507.25	4490.15
Avg. # retweets/hr	0.0268	0.2092	0.051	0.091	0.178	0.137

Both the average number of followers per tweet and the average number of followers per user are shown in Table 5.1. The difference between these values is that the former does not account for the fact that users could have posted multiple tweets for the given time period. The latter therefore only considers the number of followers for each unique user in the dataset. The follower per user to the follower per tweet ratio is therefore an indication of how many of the tweets were generated by unique authors. For #gopatriots, the ratio is high and the difference between the followers per tweet and followers per user is relatively small. However, for #sb49, the ratio is very low indicating that popular twitter authors posted multiple tweets for this hashtag which in turn skewed the average towards a higher value. The following histograms demonstrate the temporal distribution of the tweets on the per hour basis.

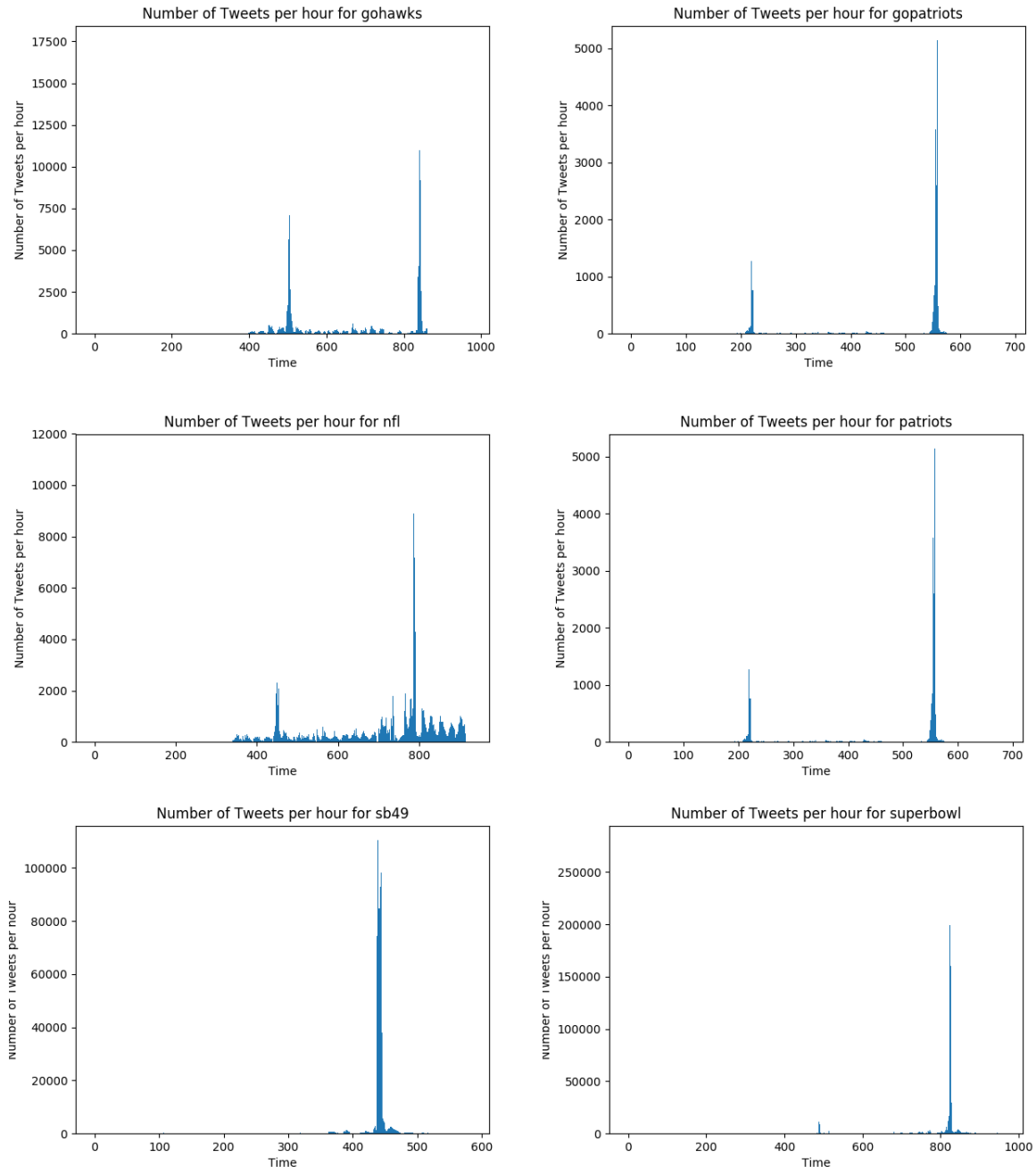

Figure 5.1: Histogram of the activities on different hashtags

As can be seen in the histograms corresponding to #gohawks, and #superbowl, there is a large spike around the 800-th hour mark, corresponds to when the Super Bowl game was played. Interestingly the first small spike in the #nfl histogram corresponds to one of the conference final games. The second spike in the histogram again corresponds to the when

the Super Bowl was played. With a small difference in time scaling we, see that observe a spike between 500 and 600-th hour in hash-tags #gopatripots, #patriots, and a similar one between 400 and 500-th hour in #nfl. This show as asynchrony in our collected data. It should be note that our change detection framework is oblivious to the asynchronism in the timing of the events.

As it can be observed from the histograms above, the per hour time resolutions is a little low to capture more granular events. Like a touch-down in the game. Thus we zoom in and define a more granular time step in the order of minute or seconds to be able to capture the events during the game because the timespan over which a trend occurs during the super bowl is minutes, not hours (e.g. a touchdown or the halftime show). Below is a plot of tweets with the keyword "touchdown" for both Patriots-friendly and Seahawks-friendly hash-tags.

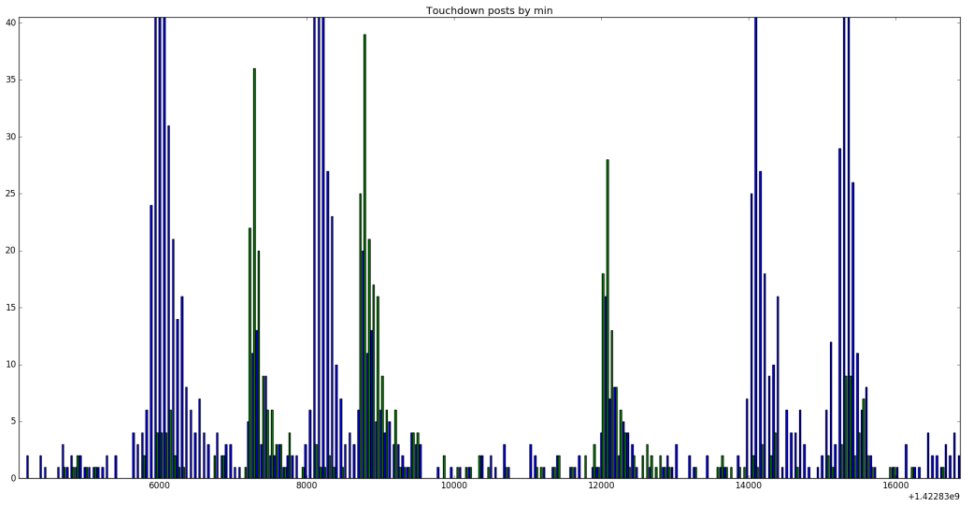

Figure 5.2: Number of touchdown posts per minute for Patriots (BLUE) and Seahawks (GREEN). Halftime show in the middle

It is clear that there are four peaks for the Patriots and three for the Seahawks corresponding to their respective touchdowns. The timing matches up with the actual scoring of Super Bowl XLIX and shows the central gap for the halftime show. The final score was 28-24 for the Patriots, which matches up with four Patriots touchdowns and three Seahawks touchdowns plus one Seahawks field goal, which does not appear with the "touchdown" filter.

Now, let us focus on applying our change detection framework to detecting events during the

game; namely, the beginning of the game and the first touchdown. Formally, the objective is to detect the start of the game only from the evolution of the activities around hashtags. To this end, we choose to observe the hash-tag #gopatriots in the hours leading to the game. In order to have a meaningfully fast detection, we change the time scale in 10 seconds and we aim to determine the start of the game only from the activities on the tweet #gopatriots. Thus we are observing the stochastic process $\{N_t^h\}$ for $h = \#\text{gopatriots}$, with time stamps corresponding to 10 second time interval. The time series in Figure 5.3 is a realization of the process.

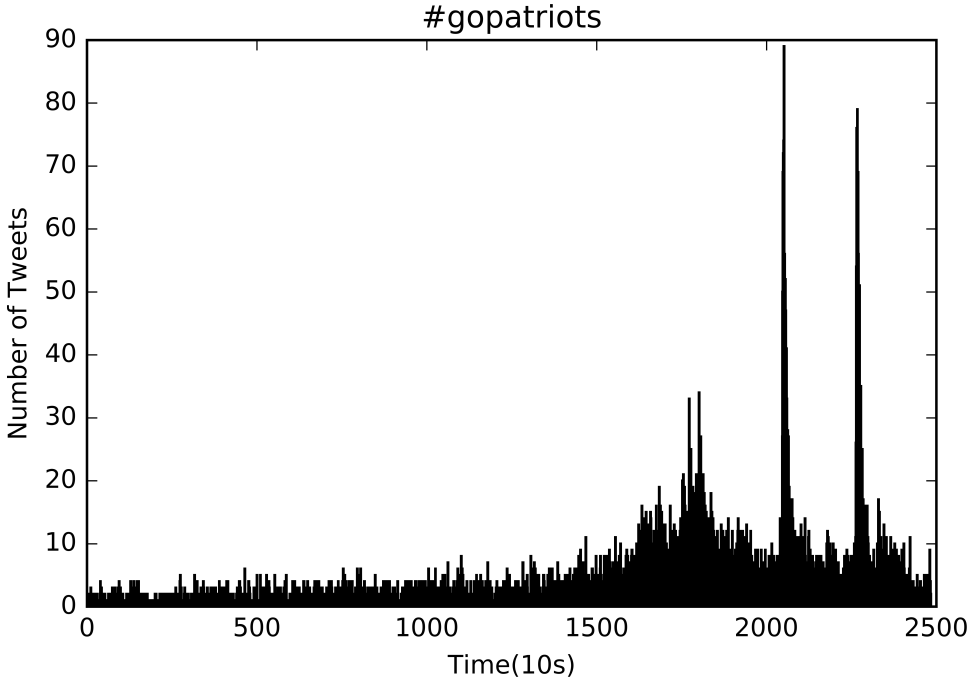

Figure 5.3: The realization of the process $\{N_t^h\}$

It can be clearly noted that the pre-change distribution of the process can approximated as a Poisson distribution

$$N_t^h \sim poi(\lambda_0) \quad t < \nu$$

with $\lambda_0 = 3$. Note that ν is the unknown time that the game starts and the pmf of the

Poisson distribution is defined as

$$f_0(x;\lambda_0) = \frac{e^{-\lambda_0}\lambda_0^x}{x!}$$

Let us assume that the change distribution that we are seeking to detect is again a Poisson distribution with parameter $\lambda_1 = 3\lambda_0$, that is

$$f_1(x;\lambda_1) = \frac{e^{-(\lambda_1)}(\lambda_1)^x}{x!}.$$

In order to detect the change in distribution from f_0 to f_1, we sequentially observe the log-likelihood of the data, which amounts to observing the behavior of the CUSUM procedure as defined in Definition 5.3.1.

$$\tau_\gamma^* = \inf\{n | \max_{k\leq n} \sum_{t=k}^{n} Z_t \geq \log\gamma\},$$

where $Z_t = \log\frac{f_1(N_t^h)}{f_0(N_t^h)}$.

Luckily, the log-likelihood ration of two Poisson distributions can be easily computed. Specifically,

$$\log(\frac{f_1(X)}{f_0(X)}) = X\log(\frac{\lambda_1}{\lambda_0}) + \lambda_0 - \lambda_1.$$

Note also that by taking the expectation of the log-likelihood, it follows that

$$D(f_1||f_0) = \mathbb{E}_{f_1}[\log(\frac{f_1(X)}{f_0(X)})] = \lambda_1\log(\frac{\lambda_1}{\lambda_0}) + \lambda_0 - \lambda_1;$$

Moreover, let us also assume that the Average Run Length(ARL) constraint for false alarm be a large number so as to guarantee the reliability of the detected changes; specifically we set $\gamma = 10^6$.

The CuSUM procedure stops when the CUSUM statistic, which is the cumulative log likelihood ratio, when positive, hits the pre-specified threshold $\log\gamma \sim 20$. Such a scheme, as shown earlier, guarantees that if no change occurs in the time window of interest will likely not mistakenly declare a change.

In order to measure how quickly the algorithm is able to detect the change, we follow the temporal evolution of the statistic as shown in Figure 5.4.

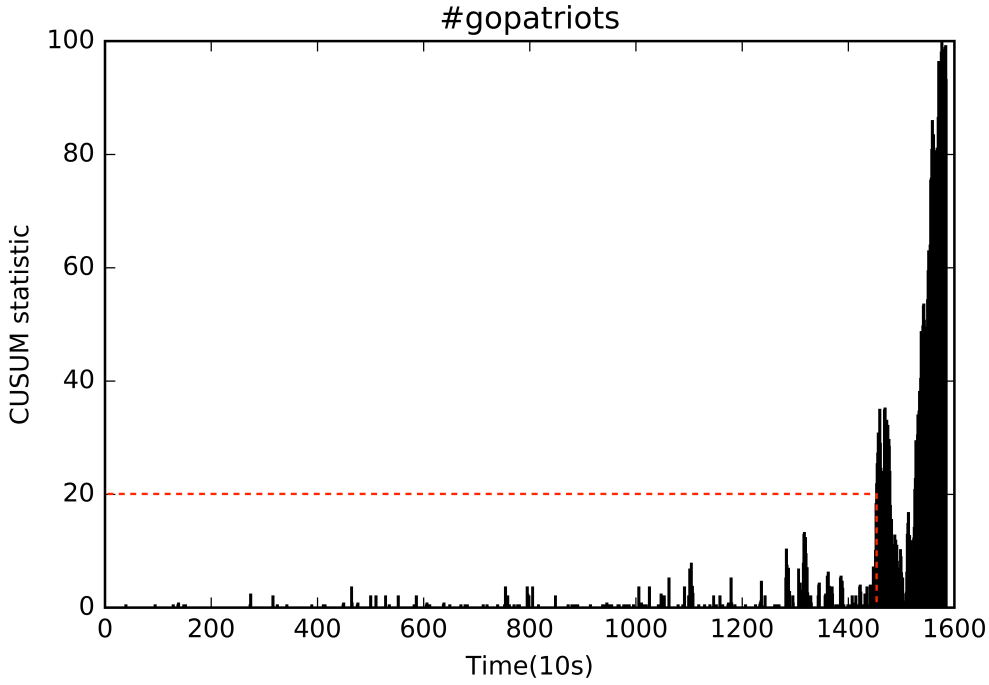

Figure 5.4: Evolution of the CUSUM statistic: As shown in the figure, the statistic captures the change as soon the statistic hits the pre-specified threshold.

It turns out that in time $t = 1486$, the algorithm can declare that a change has occurred in the distribution of the data, which is less than a minute after the change starts. This delay should be compared with the average delay $\frac{\log \gamma}{D(f_1||f_0)}$.

We recognize that in order to have a more solid way to evaluate our algorithm we need more realizations form the process so that we can make in probability or in expectation statements. However, this small experiments shows that the natural CUSUM statistic is promising in terms of a quick, yet reliable change detection scheme.

CHAPTER 6

Summary of Contributions and Future Work

In this dissertation, we propose computational models for a holistic semantic analysis of an on-line story based on aggregation of social media posts.

In the first part of the dissertation, we develop models towards characterizing a structured information network, referred to as story narrative network, where nodes, which represent groups of entities with contextually similar roles in the story, are linked via edges that represent relationships with different relation types. In particular, in order to characterize the actants, which are nodes in the story narrative network, we attribute distributed vector representations, aka embeddings, to entity mentions that appear in extracted relations from users' posts, and then by clustering the embedded vectors, we partition the entities into actant groups.

In chapter 2, we propose an embedding approach based on explicit factorization of suitably generated entity-relation matrices that capture the contextual role of an entity mention as a subject and an object in a relationship. In order to obtain interpretable embedding vectors with improved clustering behavior, we impose sparse structure on the embeddings by considering a non-negativity constraint on the factor matrices in the matrix factorization formulation. In chapter 3, we propose a new exterior point method to solve NMF, based on the results on the optimization landscape of the unconstrained matrix factorization problem. Finally, we apply K-means clustering on the obtained embedded vectors to cluster the entities into actant groups. In chapter 4, we evaluate the performance of our proposed algorithm and embedding-based clustering scheme on two datasets, namely data from a discussion forum on parenting issues and a corpus of tweets on user experience with contact-less payment methods. It is shown that our exterior point method has a significantly better sparsity

properties over the considered models as well as better prediction performance over the celebrated multiplicative update rules method for solving NMF. Moreover, we show that the clusters obtained by our method can very well recover the underlying ground truth actant groups in the studied datasets and it is computationally shown that our NMF-based embedding approach has superior clustering performance over embeddings obtained by the optimal matrix completion approach based on SVD.

In the second part of the dissertation, we touched upon characterizing the dynamics in development of the online stories. In chapter 5, we study the problem of detecting changes in the temporal evolution of the user activities and formulate this problem in a transient change point detection setting and applied a statistical test to detect the change based on the number of user activities observed so far, with minimum expected delay under a controlled measure of false alarm. We evaluate the change detection method on a corpus of tweets related to Super Bowl 2015. We show that our method is able to detect the start of the game reliably and effectively within less than a minute from the start of the game.

In the future work, we expand our approach in a number of directions itemized below:

- Joint embedding and clustering of relation and entity mentions: In our entity resolution approach, although we jointly embed both the entity and relation mentions in the same low dimensional space via entity relation matrix factorization, we are only using entity embeddings for the clustering task. An interesting future direction is to perform a joint clustering of the entity and relation embeddings.

- Adopting tensor-based models: In our current data model, the left entity relation matrix only carries the information about the co-occurrence of subject entity mentions regardless of the entities that appear as the object in the corresponding relations as the object entities. Likewise, the right entity relation matrix only carries the information about the co-occurrence of object entity mentions with the possible relationship mentions. These matrix representations can be regarded as projections of a 3-way tensor whose entries represent the number of occurrences of a triplet. Such model encodes all the information present in the set of entity relation triplets and potentially yields

superior performance on larger datasets. By defining appropriate scoring functions, as detailed in Chapter 2, we can characterize embeddings tailored to the clustering task.

- Exploring other regularization techniques to encourage sparsity structure: For a given embedding scheme, in order to impose desired structures on the parameters of the model, we have to regularize the optimization objective with penalty terms that encourage the structure of interest in the data. Specifically, in the future work, we aim at exploring imposing $\ell 1$ penalty on the embedded vectors on the tensor models presented in Chapter 2.

- Characterizing the trade-off between prediction power and interpretability of the embeddings in a model: As discussed in chapter 2, there is a fundamental trade-off between prediction power of a model versus the presence of certain structures in the learned parameters. Understanding such trade-off helps developing more efficient interpretable, yet efficient models, for predicting whether a triplet is a valid relation.

- Exploring sparsity encouraging methods for non-negative matrix factorization and understanding its optimization landscape: The primary embedding technique that we studied in this work was based on explicit factorization of the entity relation matrices. For this purpose, we used non-negative matrix factorization in order to get sparse factors in the factorization. Although, our proposed exterior point method to solve NMF inherently enforces more structured sparsity than the popular NMF methods, it would be interesting to investigate further the effect of sparsity constraint in the NMF optimization and how it changes the optimization landscape of the problem.

- Non-parametric detection schemes: A major drawback to our transient change detection approach is that the pre-change and post-change distributions should be known. In the future work, it would be interesting to develop non-parametric methods that estimate the distributions upon observing the new points.

REFERENCES

[1] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288, 2002.

[2] Xiang Ren, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R Voss, and Jiawei Han. Clustype: Effective entity recognition and typing by relation phrase-based clustering. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 995–1004. ACM, 2015.

[3] Ilya Sutskever, Joshua B Tenenbaum, and Ruslan R Salakhutdinov. Modelling relational data using bayesian clustered tensor factorization. In *Advances in neural information processing systems*, pages 1821–1828, 2009.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[5] Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics, 2010.

[6] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, 2007.

[7] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2009.

[8] Alan Ritter, Luke Zettlemoyer, Oren Etzioni, et al. Modeling missing data in distant supervision for information extraction. *Transactions of the Association for Computational Linguistics*, 1:367–378, 2013.

[9] Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer, 2010.

[10] Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–84, 2013.

[11] Jason Weston, Antoine Bordes, Oksana Yakhnenko, and Nicolas Usunier. Connecting language and knowledge bases with embedding models for relation extraction. *arXiv preprint arXiv:1307.7973*, 2013.

[12] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

[13] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM, 2008.

[14] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems*, pages 926–934, 2013.

[15] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2016.

[16] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *journal of the Association for Information Science and Technology*, 58(7):1019–1031, 2007.

[17] Danping Liao and Yuntao Qian. Paper evolution graph: Multi-view structural retrieval for academic literature. *arXiv preprint arXiv:1711.08913*, 2017.

[18] Pigi Kouki, Jay Pujara, Christopher Marcum, Laura Koehly, and Lise Getoor. Collective entity resolution in familial networks. *Under Review*, 2017.

[19] Zhou GuoDong, Su Jian, Zhang Jie, and Zhang Min. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 427–434. Association for Computational Linguistics, 2005.

[20] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, 2014.

[21] Fabian M Suchanek, Georgiana Ifrim, and Gerhard Weikum. Combining linguistic and statistical analysis to extract relations from web documents. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 712–717. ACM, 2006.

[22] Nanda Kambhatla. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive*

poster and demonstration sessions, page 22. Association for Computational Linguistics, 2004.

[23] Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. Structured relation discovery using generative models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1456–1466. Association for Computational Linguistics, 2011.

[24] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[25] Dekang Lin and Patrick Pantel. Dirt- discovery of inference rules from text. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–328. ACM, 2001.

[26] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[27] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *ICML*, volume 11, pages 809–816, 2011.

[28] Rasmus Bro. Parafac. tutorial and applications. *Chemometrics and intelligent laboratory systems*, 38(2):149–171, 1997.

[29] Antoine Bordes, Jason Weston, Ronan Collobert, Yoshua Bengio, et al. Learning structured embeddings of knowledge bases. In *AAAI*, volume 6, page 6, 2011.

[30] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795, 2013.

[31] Danqi Chen, Richard Socher, Christopher D Manning, and Andrew Y Ng. Learning new facts from knowledge bases with neural tensor networks and semantic word vectors. *arXiv preprint arXiv:1301.3618*, 2013.

[32] J Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika*, 35(3):283–319, 1970.

[33] Richard A Harshman. Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis. 1970.

[34] Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[35] Brett W Bader, Richard A Harshman, and Tamara G Kolda. Temporal analysis of semantic graphs using asalsan. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 33–42. IEEE, 2007.

[36] Kai-Wei Chang, Scott Wen-tau Yih, Bishan Yang, and Chris Meek. Typed tensor decomposition of knowledge bases for relation extraction. 2014.

[37] Maximilian Nickel, Xueyan Jiang, and Volker Tresp. Reducing the rank in relational factorization models by including observable patterns. In *Advances in Neural Information Processing Systems*, pages 1179–1187, 2014.

[38] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

[39] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, pages 1735–1742. IEEE, 2006.

[40] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, pages 2071–2080, 2016.

[41] Maximilian Nickel, Lorenzo Rosasco, Tomaso A Poggio, et al. Holographic embeddings of knowledge graphs. In *AAAI*, pages 1955–1961, 2016.

[42] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434. ACM, 2008.

[43] V.Q. Vu and J. Lei. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013.

[44] J. D. Lee, B. Recht, N. Srebro, J. Tropp, and R. R. Salakhutdinov. Practical large-scale optimization for max-norm regularization. In *Advances in Neural Information Processing Systems*, pages 1297–1305, 2010.

[45] Mojtaba Kadkhodaie Elyaderani. A computational and statistical study of convex and nonconvex optimization with applications to structured source demixing and matrix factorization problems. In *University of Minnesota, ProQuest Dissertations Publishing*, 2017.

[46] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 556–562. MIT Press, 2001.

[47] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin. The global optimization geometry of non-symmetric matrix factorization and sensing. *arXiv preprint arXiv:1703.01256*, 2017.

[48] Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun-ichi Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[49] Saïd Moussaoui, David Brie, and Jérôme Idier. Non-negative source separation: range of admissible solutions and conditions for the uniqueness of the solution. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 5, pages v–289. IEEE, 2005.

[50] Scott Rickard and Andrzej Cichocki. When is non-negative matrix decomposition unique? In *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*, pages 1091–1092. IEEE, 2008.

[51] Hans Laurberg, Mads Græsbøll Christensen, Mark D Plumbley, Lars Kai Hansen, and Søren Holdt Jensen. Theorems on positive data: On the uniqueness of nmf. *Computational intelligence and neuroscience*, 2008, 2008.

[52] Stephen A Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2009.

[53] Nicolas Gillis. Introduction to nonnegative matrix factorization. *arXiv preprint arXiv:1703.00663*, 2017.

[54] Hamza Fawzi and Pablo A Parrilo. Lower bounds on nonnegative rank via nonnegative nuclear norms. *Mathematical Programming*, 153(1):41–66, 2015.

[55] Michael JD Powell. On search directions for minimization algorithms. *Mathematical programming*, 4(1):193–201, 1973.

[56] Luigi Grippo and Marco Sciandrone. On the convergence of the block nonlinear gauss–seidel method under convex constraints. *Operations research letters*, 26(3):127–136, 2000.

[57] Edward F Gonzalez and Yin Zhang. Accelerating the lee-seung algorithm for nonnegative matrix factorization. *Dept. Comput. & Appl. Math., Rice Univ., Houston, TX, Tech. Rep. TR-05-02*, pages 1–13, 2005.

[58] Chih-Jen Lin. On the convergence of multiplicative update algorithms for nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 18(6):1589–1596, 2007.

[59] R. Sun and Z.-Q. Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

[60] Christos Boutsidis and Efstratios Gallopoulos. Svd based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362, 2008.

[61] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[62] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent converges to minimizers. *arXiv preprint arXiv:1602.04915*, 2016.

[63] Michael W Berry, Murray Browne, Amy N Langville, V Paul Pauca, and Robert J Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational statistics & data analysis*, 52(1):155–173, 2007.

[64] Chih-Jen Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.

[65] Dimitri Bertsekas. On the goldstein-levitin-polyak gradient projection method. *IEEE Transactions on automatic control*, 21(2):174–184, 1976.

[66] Naiyang Guan, Dacheng Tao, Zhigang Luo, and Bo Yuan. Nenmf: An optimal gradient method for nonnegative matrix factorization. *IEEE Transactions on Signal Processing*, 60(6):2882–2898, 2012.

[67] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[68] Rafal Zdunek and Andrzej Cichocki. Non-negative matrix factorization with quasi-newton optimization. In *International conference on artificial intelligence and soft computing*, pages 870–879. Springer, 2006.

[69] Shu-Zhen Lai, Hou-Biao Li, and Zu-Tao Zhang. A symmetric rank-one quasi-newton method for nonnegative matrix factorization. *ISRN Applied Mathematics*, 2014, 2014.

[70] Moody Chu, Fasma Diele, Robert Plemmons, and Stefania Ragni. Optimality, computation, and interpretation of nonnegative matrix factorizations. In *SIAM Journal on Matrix Analysis*. Citeseer, 2004.

[71] Yu-Xiong Wang and Yu-Jin Zhang. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1336–1353, 2013.

[72] Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(Nov):1457–1469, 2004.

[73] Stan Z Li, Xin Wen Hou, Hong Jiang Zhang, and Qian Sheng Cheng. Learning spatially localized, parts-based representation. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2001.

[74] Patrik O Hoyer. Non-negative sparse coding. In *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*, pages 557–565. IEEE, 2002.

[75] Hyunsoo Kim and Haesun Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM journal on matrix analysis and applications*, 30(2):713–730, 2008.

[76] Hyunsoo Kim and Haesun Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.

[77] V Paul Pauca, Jon Piper, and Robert J Plemmons. Nonnegative matrix factorization for spectral data analysis. *Linear algebra and its applications*, 416(1):29–47, 2006.

[78] Timothy R Tangherlini, Vwani Roychowdhury, Beth Glenn, Catherine M Crespi, Roja Bandari, Akshay Wadia, Misagh Falahi, Ehsan Ebrahimzadeh, and Roshan Bastani. "mommy blogs" and the vaccination exemption narrative: results from a machine-learning approach for story aggregation on parenting social media sites. *JMIR public health and surveillance*, 2(2), 2016.

[79] David G Hays. Dependency theory: A formalism and some observations. *Language*, 40(4):511–525, 1964.

[80] Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. Universal stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–4592, 2014.

[81] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[82] Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

[83] Ana-Maria Popescu and Orena Etzioni. Extracting product features and opinions from reviews. In *Natural language processing and text mining*, pages 9–28. Springer, 2007.

[84] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[85] Elena Tutubalina and Vladimir Ivanov. Unsupervised approach to extracting problem phrases from user reviews of products. In *Proceedings of the First AHA!-Workshop on Information Discovery in Text*, pages 48–53, 2014.

[86] Gary Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, pages 1897–1908, 1971.

[87] ES Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–527, 1955.

[88] George V Moustakides. Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, pages 1379–1387, 1986.

[89] Tze Leung Lai. Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information Theory*, 44(7):2917–2929, 1998.

[90] Albert N Shiryaev. *Optimal stopping rules*, volume 8. Springer Science & Business Media, 2007.

[91] Alexander G Tartakovsky and Venugopal V Veeravalli. General asymptotic bayesian theory of quickest change detection. *Theory of Probability & Its Applications*, 49(3):458–497, 2005.

[92] Venugopal V Veeravalli and Taposh Banerjee. Quickest change detection. In *Academic Press Library in Signal Processing*, volume 3, pages 209–255. Elsevier, 2014.

[93] Taposh Banerjee and Venugopal V Veeravalli. Data-efficient quickest change detection with on–off observation control. *Sequential Analysis*, 31(1):40–77, 2012.

[94] Taposh Banerjee and Venugopal V Veeravalli. Data-efficient quickest change detection in minimax settings. *IEEE Transactions on Information Theory*, 59(10):6917–6931, 2013.

[95] Blaise Kévin Guépié, Lionel Fillatre, and Igor Nikiforov. Sequential detection of transient changes. *Sequential Analysis*, 31(4):528–547, 2012.

[96] K Premkumar, Anurag Kumar, and Venugopal V Veeravalli. Bayesian quickest transient change detection. In *Proc. Fifth International Workshop on Applied Probability (IWAP)*, pages 1–3, 2010.

[97] Vikram Krishnamurthy. Bayesian sequential detection with phase-distributed change time and nonlinear penalty—a pomdp lattice programming approach. *IEEE Transactions on Information Theory*, 57(10):7096–7124, 2011.

[98] Chunming Han, Peter K Willett, and Douglas A Abraham. Some methods to evaluate the performance of page's test as used to detect transient signals. *IEEE transactions on signal processing*, 47(8):2112–2127, 1999.

[99] Ehsan Ebrahimzadeh and Aslan Tchamkerten. Sequential detection of transient changes in stochastic systems under a sampling constraint. In *Information Theory (ISIT), 2015 IEEE International Symposium on*, pages 156–160. IEEE, 2015.