Master DMKM Report

Cold Start Recommendations: A Non-negative Matrix Factorization Approach

Martin SAVESKI

June 25, 2013

Supervision: Amin Mantrach, Yahoo! Labs,
Daniele Quercia, Yahoo! Labs,
Ricard Gavaldà, Universitat Politècnica de Catalunya.

Location: Yahoo! Labs, Barcelona, Spain.

Abstract:

Recommender systems suggest to users items that they might like (e.g., news articles, songs, movies) and, in doing so, they help users deal with information overload and enjoy a personalized experience. One of the main problems of these systems is the cold-start, i.e., when a new item or user is introduced in the system and no past information is available, no effective recommendation can be produced. The cold-start is a very common problem in practice: modern online platforms have hundreds of new items and millions of visits from logged-out users every day. Despite the importance of this problem not many solutions have been proposed in the literature. We contribute to closing this gap by studying whether it could be overcome without sacrificing performance. We do so by exploiting two aspects: the combination of the content and collaborative information, and the users' location.

First, we combine the properties of the items and past user preferences by introducing Joint NMF, a novel recommender system based on Non-negative Matrix Factorization (NMF). We also propose a variation, Joint NMF with Graph Regularization, which accounts for the local geometric structure of the data. We present two training strategies, based on multiplicative update rules and alternating least squares. The experimental results show that the two proposed methods outperform the existing content-based recommenders, which are often used for recommending newly introduced items (item cold-start).

Second, to exploit user location, we test whether specific interests are linked to specific locations. To this end, we study the geography of user engagement in online news platforms on a random sample of 200K news articles and the corresponding 41M comments posted on the Yahoo! News website for the USA. We find that time zones play an important role on the user engagement: users from the same time zone preferentially engage with each other on the same articles about the same topics. Thus, we propose a time-zone-aware recommender that suggests the most popular articles not in the whole USA but in the users' time zones, when no past information of the user preferences is available (user cold-start). We find that suggesting recommendations that are specific to time zones improves the recommendation accuracy by a factor of one point five.

Résumé:

Les systèmes de recommandations en ligne suggèrent aux utilisateurs le ou les articles qu'ils seraient susceptibles d'apprécier (par exemple une chanson, un article de presse ou un film). De cette manière, les systèmes en ligne offrent à l'utilisateur une expérience unique et personnalisée, plus en ligne avec ses attentes, réduisant ainsi la surcharge d'informations non relevantes. L'un des problème majeures des systèmes de recommandation et ce que l'on appelle communément le démarrage à froid, c'est à dire lorsque le système fait face à nouvel élément (de type article ou utilisateur) pour lequel, donc, aucun historique d'utilisation ne peut-être utilisé pour faire la moindre recommandation. Le démarrage à froid est un problème très courant en pratique: les plateformes modernes en lignes ont chaque jour des centaines de nouveaux articles et des millions de visiteurs non-enregistrés. Malgré son importance, peu de solutions spécifiques au problème ont été jusqu'à alors présentées. Afin de contribuer à l'avancement du problème, nous étudions s'il est possible d'y apporter une solution tout en conservant une performance acceptable. Nous exploitons deux aspect de l'informations pour répondre au problème: La combinaison du contenu et de l'information collaborative d'une part, et l'utilisation des informations géographiques utilisateurs d'autre part.

Premièrement, nous exploitons les propriétés des articles et des préférences utilisateurs passées qui leurs sont associées en introduisant un NMF Joint, un nouveau système de recommandation basé sur une factorisation matricielle non-négative. Nous proposons également une variante, qu'est le NMF joint avec une contrainte de régularisation graphe, qui tient compte de la structure géométrique locale des données. Nous présentons deux techniques d'apprentissages, l'une basée sur des règles de mise à jour multiplicatives, et l'autre sur des moindres carrés alternés. Les résultats expérimentaux montrent que les deux techniques proposées surpassent les systèmes de recommandations basés sur le contenu qui sont souvent utilisés dans le cas de nouveaux articles (c.à.d. lors d'un démarrage à froid).

Deuxièmement, afin d'exploiter l'information géographique des utilisateurs, nous nous intéressons à savoir si les zones géographiques peuvent être identifiées par des intérêts particuliers. Pour cela, nous étudions la géographie de l'engagement utilisateurs sur un échantillon de 200 milles articles de presses et de leurs 41 millions de commentaires associés publiés sur la plateforme de presses en ligne Yahoo! News US. Nous avons découvert, que les fuseaux horaires jouent un rôle important dans l'engagement utilisateur: les utilisateurs d'un même fuseau horaire ont une tendance à plus s'engager sur les même articles de presse couvrant des sujets similaires. Dés lors, nous proposons un système de recommandations tenant compte du fuseau horaire, il suggère l'article le plus populaire non pas sur l'ensemble des US, mais dans le fuseau horaire utilisateur, lorsqu'aucun historique n'est encore disponible sur celui-ci (c.à.d. dans le cas d'un démarrage à froid). De cette façon, nous améliorons la performance de recommandation d'un facteur 1.5.

Contents

Notation Glossary

Table of Symbols

Symbol	Description
$\mathbb{R}$	field of real numbers
$\mathbf{X}$	matrix $\mathbf{X}$
$\mathbf{X}^{\mathrm{T}}$	transpose of $\mathbf{X}$
$[\mathbf{X}]_{ij}$	element in row i and column j of $\mathbf{X}$
trace($\mathbf{X}$)	trace of a square matrix $\mathbf{X}$
$\|\|\mathbf{X}\|\|$	Frobenius norm of a matrix $\mathbf{X}$
$\|\|x\|\|$	l_2 norm of a vector x
$\frac{\partial \mathbf{X}}{\partial \mathbf{Y}}$	partial derivative of $\mathbf{X}$ with respect to $\mathbf{Y}$
$\mathbf{X} \odot \mathbf{Y}$	Hadamard, element-wise, product between matrices $\mathbf{X}$ and $\mathbf{Y}$
$\mathbf{X} \oslash \mathbf{Y}$	Hadamard, element-wise, division between matrices $\mathbf{X}$ and $\mathbf{Y}$
$\mathbf{X} \otimes \mathbf{Y}$	Kronecker product between matrices $\mathbf{X}$ and $\mathbf{Y}$
vec($\mathbf{X}$)	vectorization of a matrix $\mathbf{X}$ obtained by stacking the columns of the matrix
$reshape(x, m, n)$	converts a vector x in matrix of size $m \times n$
$Diag(x)$	diagonal matrix with vector x in the diagonal

Table of Aberrations

Aberration	Description
NMF	Non-negative Matrix Factorization
JNMF	Joint Non-negative Matrix Factorization
JNMF-GR	Joint Non-negative Matrix Factorization with Graph Regularization
SVD	Singular Value Decomposition

Hosting institution

Yahoo! Labs Barcelona is the first Yahoo! research center established outside of the USA. The lab has over twenty researchers and research engineers, fifteen postdoctoral researchers and frequent visitors including visiting professors and graduate students. It has five research groups led by senior researchers: Web Mining, Social Media Engagement, Scalable Computing, Web Retrieval and Semantic Search. Main conferences targeted by the members of the lab include: WWW, RecSys, ICDM, ECML, ICML, KDD, SIGIR, CSCW, CHI, and others.

The *Web Mining* research group aims at developing novel methods for analyzing web-scale data. Their research mainly focuses on obtaining actionable knowledge about people and content on the web, and using this knowledge to efficiently support innovative applications. The expertise of the group spans a wide range of areas and applications in data mining. Current work of the group includes projects on: sequence and graph mining (pattern discovery, distance indexing, probabilistic graphs, mining evolving graphs, community detection), and analysis of influence propagation (finding "leaders and followers", learning influence strengths, finding the "backbone" of social networks).

The *Social Media Engagement* research group aims to advance the state-of-the-art in social media and to transfer the results to have a strong, positive impact on the experience of customers using Yahoo! products. The group focuses on two main areas: (1) deep analysis and understanding of users (segments, actions, and interactions) as well as social content (creation, sharing, and general use); and (2) increase user engagement by creating next generation social media experiences based on new functionalities and services that leverage Yahoo! network content and other data sources. The research carried out is at the intersection of large-scale data analysis (*i.e.*, mining) and interdisciplinary insights on individual and group behavior that takes into account social, economic, and cultural factors.

The *Scalable Computing* research group focuses on cloud technologies, distributed algorithms, dependability, and storage systems. Their primary goal is to develop systems that improve the existing software infrastructure of Yahoo!. They address a number of aspects of distributed systems, but focus mostly on replication, concurrency, and performance. The recent work of the group includes: (1) algorithms for dependable systems, (2) dependable logging and stream processing, (3) distributed systems for web search, (4) systems and algorithms to support social media applications, and (5) transaction support for key-value stores (*e.g.*, HBase). The group is actively involved in a number of current projects, internal and external, including ZooKeeper, BookKeeper, HDFS, HBase and S4.

The *Web Retrieval* research group mainly focuses on web search, information retrieval, and distributed search systems. More specifically, the group takes interest in web search efficiency optimizations, information retrieval architectures, and search result ranking. Their research also covers a number of topics that are indirectly related to search, such as user engagement, sentiment analysis, financial cost optimizations, and web crawling.

The *Semantic Search* group focuses on fundamental and applied research in processing natural language and semantic metadata to go beyond current keyword-bound search. The team includes researchers with expertise in information retrieval, natural language processing and semantic web, and research engineers with expertise in large-scale data analysis and processing. They contribute prominently to the Web of Objects project and other initiatives exploiting structured data inside and outside Yahoo!

Acknowledgments

I would like to thank my supervisors: Amin Mantrach, Daniele Quercia and Ricard Gavaldà for their advice and support. Thanks to Amin for giving me directions in this short journey in the world of matrix factorizations, it has been a joy. To Daniele for sharing his wisdom and giving me crush courses on writing, research, and academic life in general.

Thanks to Gianmarco De Francisci Morales and Nicola Barbieri who found time read and comment on my work, even on Sant Joan.

Yahoo! Labs Barcelona has been a great place to meet many like-minded people. Thanks to: Carmen (I will never forget camping in the lab to write papers); Rosa, Sandra, Guillem and David (for all the Fussball games); Antonio and Arinto (for being there for the thesis writing all-nighters).

DMKM has been a long journey which brought many life-lasting friendships. Thanks to: Hana Susak (for being an awesome flatmate), Xinyu, Nazanin, Vishwanatan, Yabebal, Ghufran, Brindusa, Revekka, Diana (the real and the Unreal), Dominik, Pavel, and Agata. It has been a pleasure to meet you all.

Many thanks to *her*, for being there and supporting me all the time, it will never be forgotten.

Finally, I would like to thank my parents for their endless love, encouragement, advice and support. Bez vashata podrshka nishto nemashe da bide kako shto e! Fala vi!

1 Introduction

It is often necessary to make choices without sufficient personal experience of the alternatives. In everyday life, we rely on recommendations from other people either by word of mouth, recommendation letters, movie reviews or general surveys such as Lonely Planet's travelling guides. Recommender systems assist and augment this natural social process. These systems are aimed to help users of online platforms deal with the large volumes of information and to provide them a personalized experience. This achieved by suggesting items of interest to the users based on their explicit and implicit preferences.

Recommender systems use a number of different technologies, but may be broadly classified into two groups: content-based and collaborative filtering systems. *Content-based systems* examine the properties of the items and recommend items which are similar to the ones the user preferred in the past. They model the taste of a user by building a user profile based on the properties of the items the user liked, and use the profile to compute the similarity with new items. Items which are most similar to the user's profile are recommended to the user. *Collaborative filtering systems*, on the other hand, ignore the properties of the items and base their recommendations on community preferences. They recommend items that users with similar tastes and preferences liked in the past. Two users are considered similar if they have many items in common. Thus, the system can use the preferences of the similar users to predict the items a user may like.

One of the main problems for recommender systems is the *cold-start* problem, *i.e.*, when a new item or user is introduced in the system. We distinguish two types of cold-starts: (1) when a new item is introduced, and (2) when a new user enters the system. The collaborative filtering approaches suffer from both, item and user cold-start, as they rely on the previous ratings of the users. The item cold-start, on the other hand, does not appear in content-based recommenders (as the properties of the new item can be used for recommendation), but still they suffer from the user cold-start.

Despite the fact that these problems are very frequent in practice, they have not been widely studied in the literature. This work aims to contribute to close this gap by proposing solutions for both: the item and user cold-start. We consider a large collection of articles and user comments posted on the Yahoo! News website. The cold start commonly occurs on the website as there are hundreds of new articles published every day and millions of users that visit the website without being logged in. Thus, addressing this problem is of great importance for improving the user experience on the platform.

Item cold-start. The state-of-the-art collaborative filtering systems are based on dimensionality reduction techniques, such as Singular Value Decomposition (SVD) or UV decomposition [18]. The goal of these techniques is to approximate the user-item matrix (*i.e.*, which user liked which item) with several low-rank matrices. On the other hand, there has been a large body of work on factorization of the item-property matrix (*e.g.*, document-term matrix, in the case of textual items). Non-negative matrix factorization (NMF) is one such approach which aims at factorizing the document-term matrix in two non-negative, low-rank matrices, where one matrix corresponds to the topics that occur in the collection, and the other represents the extent to which each document belongs to these topics. Although, the two techniques, collaborative filtering based on dimensionality reduction and topic modeling, address different problems, they are very similar in nature. We propose to combine them in a hybrid recommendation system, *Joint NMF*, that exploits both the properties of the items and the similarity of user preferences. The main idea is to jointly factorize the item-property matrix and the user-item matrix in a common low-dimensional space. As the system makes use of both the content and user preferences, it does not suffer from item cold-start. Given the description of a new item (*e.g.*, the content of a news article), we may project it in the common low-dimensional space and infer the users which are most likely to be interested in it. We also propose an extension of the model, which exploits the local geometric structure of the data in order to discover better low-dimensional space. We evaluate the two methods and we show that both outperform a pure content-based approach and variations thereof.

User cold-start. When logged out users visit the website, there is no information about the past users' preferences available to the system. Thus, the most common way to serve these users is by recommending

the items that currently receive the majority of the attention (*i.e.*, the top-k most popular items), leading to one-size-fit-all recommendations. However, the user's requests come with some contextual information. For instance, using a user's IP address, we may infer the geographical location of the user and their time zone. One may exploit this information to understand the needs of the users in specific regions or in specific time of the day. In this work, we study the geography of the engagement of users in news articles published on the Yahoo! US website in a period of two years. We find that users engage with each other depending on where they live. More specifically, users in the same time zone preferentially engage with each other. We also find that this phenomenon is due to (1) the interests of the users in specific topics, and (2) the socio-demographic conditions and personality traits of the users. Based on these findings, we propose a recommendation system that partially overcomes the user cold-start problem. We find that suggesting what is popular in the time zone of the user rather than what is popular in the whole USA significantly improves the recommendation accuracy.

Contributions. We make a number of contributions to addressing both the item and user cold-start. In the context of *item cold-start*:

- We introduce a new method for recommendation, *Joint NMF*, that combines the content and collaborative information in a unified matrix factorization framework. We propose two training strategies, based on multiplicative update rules and alternating least squares, and we show their convergence properties.
- We develop a variation of the proposed method, *Joint NMF with Graph Regularization*, that accounts for the local geometric structure of the data, and we provide algorithms for training it.
- We conduct an extensive experimental study and we show that the proposed methods outperform the existing content-based recommenders, which are often used to address the item cold-start.

In the context of *user cold-start*:

- We study the extent to which certain locations are associated with certain interests. We find that users engage with each other depending on where they live. More specifically, users in the same time zone preferentially engage with each other.
- We find that this phenomenon is linked to the users' interests in specific topics, their socio-demographic conditions and personality traits.
- We propose a *time-zone-aware recommender system* that partially overcomes the user cold-start problem. Experimental results show that recommending items that are specific to time zones improves the recommendation accuracy by a factor of one and a half.

Road map. The remainder of this document is organized as follows. In Chapter 2, we present the related work and spell out our contributions.

In Chapter 3, we study the problem of *item cold-start* recommendations. In Section 3, we formally define the problem. In Section 3.1, we investigate the properties of NMF as a basis for our proposal. In Section 3.2, we address the item cold-start problem by proposing a model which exploits both, the content information and the user preferences. In Section 3.3, we present a variant of the model which also takes into account the local geometric structure of the data. Finally, in Section 3.4, we evaluate these two models and present experimental results.

In Chapter 4, we study the problem of *user cold-start* recommendations. In Section 4.2, we test if users who live in the same time zone are more likely to engage in the same articles than users further away. In Section 4.3, we investigate how much this phenomenon is due to the better alignment of certain time zones with the publishing cycle. In Section 4.4, we study the influence of the topical interests of the users and in Section 4.5 the influence of the socio-demographic situation and the personality traits. Based on the findings, in Section 4.6, we propose a time-zone-aware recommender system and we evaluate its performance. Finally, in Section 4.7, we outline the practical and theoretical implications of the results of this analysis.

2 Related Work

In this section, we highlight the previous studies that are relevant to this work. The remainder of this chapter is organized in four major sections: Sections 2.1 to 2.3 cover work related to our approaches for items cold-start, while Section 2.4 covers studies that investigate the engagement of users in online platforms. As the methods proposed for item cold-start is based on NMF we defer its discussion to Section 3.1, where we provide an extensive overview of the problem and algorithms for its optimization.

2.1 NMF on Multiple Sources of Information

Badea [3] proposes a variation of NMF called *Simultaneous NMF* for the purpose of clustering. The goal of the method is to find two factorization sharing a common low-dimensional representation. The method has been applied on two gene expression datasets with the aim of uncovering gene regulatory programs that are common to the two phenotypes. The model incorporates offset variables which have specific role in the problem considered. The offsets absorb the constant expression levels of the genes and allow better defined clusters to be extracted.

Liu *et al.* [35] introduce a multi-view clustering based on NMF. They propose to cluster the data from each view separately, but enforcing a clustering close to the consensus one, computed on the clustering of each view in the current iteration. To form the consensus clustering from the multiple view they enforce that the l_1 norm of each factor is equal to 1. Experimental results show that the method outperforms the single-view NMF or performing NMF on the concatenation of the features from each view.

Akata and Thurau [1] propose a NMF model similar to the Joint NMF proposed in this work in the context of tag prediction of Flickr images. Unlike our approach, Tikhonov regularization is not imposed on the factors. They show anecdotal examples where the model predicts the correct tags, but no systematic evaluation or comparison with other methods is performed.

2.2 NMF with Graph Regularization

The idea of exploiting the local geometric structure of the data for discovering better low-dimensional representations in NMF has been first proposed by Cai *et al.* [11]. Inspired by the success of using the nearest neighbor graph for label propagation in semi-supervised learning, they propose a clustering technique based on NMF. The algorithm favors factorizations for which similar instances have similar low-dimensional representations. The authors show that, by imposing this constraint, they outperform NMF and classical clustering algorithms.

A semi-supervised technique based on the similar ideas has been proposed by Wang *et al.* [50]. In addition to the traditional NMF objective (*i.e.*, that the product of the low-rank matrices is close to the original data matrix), they enforce two other constraints. First, they increase the discriminative power of the factorization by imposing that the low-dimensional representations of instances from different classes should be different, while the representations of instances from the same class should be similar. Second, they impose a label propagation constraint, *i.e.*, similar instances should have similar labels, and this allows propagating label information from labeled to unlabeled instances.

2.3 Hybrid Recommender Systems

Melville *et al.* [36] propose a hybrid recommender system for enriching collaborative filtering with content information, leading to so-called *content-boosted collaborative filtering*. The approach is motivated by the sparsity of the collaborative filtering matrix, *i.e.*, the fact that the majority of the users rate only few items. They propose to enhance the existing user data by incorporating a content-based recommender. The collaborative user profiles are built with (1) the actual ratings for the items rated by the user, and (2) the pseudo ratings (as predicted by a content-based recommender) for the items not rated by the user. The profiles are then used as an input to a nearest neighbour collaborative filtering algorithm. The experimental results show that content-boosted collaborative filtering outperforms both pure content-based and pure collaborative filtering recommenders. The authors point out that the system can be used to overcome the items cold-start problem, but no empirical result is provided.

Soboroff [48] proposes a technique based on Latent Semantic Indexing (LSI) for combining the collaborative filtering input and the document content for recommendation of textual items. The method builds a content profile for each user as a linear combination of the preferred documents. LSI is then applied to the user profiles to discover topics in the collection and implicitly learn commonalities among the user profiles. Incoming documents are projected into the LSI space and compared to user profiles. The documents are recommended to the users who have the most similar profiles. The authors argue that applying LSI on the user profiles instead of the documents allows one to take into account the collaborative input and consequently improves the recommendation performance.

Schein *et al.* [46] propose a method for recommending items that combines content and collaborative data in a single probabilistic model and overcomes the item cold-start problem. They focus on the problem of movie recommendation, where the actors of the movie are considered as content information. They suppose that there exists some latent cause that motivates users to like certain movies and actors. Based on the actors appearing in the movies watched by the users, they estimate the probability that: (1) a certain user will be motivated by some latent cause, and (2) that a given latent cause will favour some actors. At query time, the actors of a new movie are used as an evidence for placing the movie in the latent space. Based on those estimates, the probability that a certain user will like the movie is inferred. As opposed to our approach, this method explicitly uses the actors as a proxy for the movies.

Rosen *et al.* [44] introduce the *author-topic model*, a generative model which extends Latent Dirichlet Allocation (LDA) to include authorship information. They associate each author with a multinomial distribution over topics, and each topic with multinomial distribution over words. Thus, a document with multiple authors can be modeled as a distribution over topics that is a mixture of the distribution associated with the authors. The model may be used to answer a range of interesting queries including: which topics an author writes about or who may be the authors of an unobserved document. Notice the similarity of the problem of author-topic modeling and the item cold-start recommendation. One may associate the documents to the users who showed interest in them, instead of their authors. Thus, by using the same model, one may predict which users may be interested in a new document.

2.4 User Engagement in Online Platforms

Golder and Macy [20] examined how the expression of emotion words on Twitter changed over the course of one day and they found that it was regularly shifted along time zones. That is similar to what Mislove *et al.* [38] independently reported when comparing the usage of Twitter in the west cost with that in the east.

Kwak *et al.* [30] studied to which extent Twitter is used as a news sharing platform and found that indeed the majority of trending topics (over 85%) are headline news. They also found that reciprocal relations (75% of them) tend to be between users who live no more than three time zones away. Recent studies have also examined the geographic spread of topics in Twitter by investigating the adoption of hashtags across locations around the world [28]. They found that physical distance between locations constrained the spreading of hashtags: the adoption of the same hashtag by two locations was inversely proportional to their geographical distance.

Jones and Altadonna [27] examined the introduction of badges (*i.e.*, awards for users with frequent posting) to encourage user engagement on the Huffington Post website. They found that longer threads do not come from badges but from the desirability of news articles. Himelboim *et al.* [21] studied 20 Usenet newsgroups on politics to investigate around which content considerable discussion is generated. They found that the most discussed content comes from traditional high quality news outlets.

2.5 Discussion

From this brief literature review, one concludes that we hitherto lack a detailed understanding of: (1) how to combine the content and collaborative filtering in a unified framework and to what extent such model can be used to overcome the item cold-start; and (2) how the geography impacts the engagement of user in online platforms and whether it may be used to address the user cold-start. We thus set out to close this gap by: (1) proposing a model for combining content and collaborative information based on NMF (Chapter 3) and (2) studying the geographical processes that impact user to show interest in different items (Chapter 4).

3 Item Cold-Start

The scenario we consider is the cold-start recommendation and, in particular, the recommendation of new items, *i.e.*, item cold-start. Given a new item and its corresponding description, we would like to find users which may be interested in it. If we adopt a collaborative filtering approach, as a state-of-the-art in recommendation systems, then we are interested in finding users who are similar to those who already showed interest in the item. However, since it is a new item, none of the users had a chance to examine it yet, and thus this approach is not applicable. We may, on the other hand, consider the description of the item. Thus, given the description of the item, we would like to find users that expressed interested in similar items before. In the context of the particular problem we consider, items represent news articles and their description are the words that appear in the news. We also have information of which users commented on which news articles in the past. Thus, when a new article arrives in the system, we would like to find the users who are more likely to comment.

To have a better understanding of this class of problems, we also consider two other tasks that are very similar to the item cold-start recommendations: email recipient and author prediction.

Email recipient prediction. When people write emails, they do not necessarily fill in the destination address first, but might rather start by writing the content. Moreover, they tend to write about specific topics with specific people. For instance, emails to employers may contain reports and discuss work activities, while emails to friends are more likely to be informal and discuss fun and personal issues. Thus, given the content of the email, we would like to predict the most likely recipients and propose them to the user. Similar to the recommendation of news articles, emails are described by words and one might predict the most likely recipients based on those words.

Author Prediction. We also consider a collection of scientific articles and their authors. Given a new article, based on the previous publications of the authors, we would like to predict who are the most likely authors. Similar to the previous two tasks, scientific articles are described by their words and by their authors.

Problem Statement. More formally, we can define the problem as follows. At training time, we are given a collection of n items described by: (1) a set of m properties stored in a matrix $\mathbf{X_s} \in \mathbb{R}^{n \times m}$, where a row corresponds to an item and a column to an item property; and (2) a set of u users stored in a matrix $\mathbf{X_u} \in \mathbb{R}^{n \times u}$, where a cell (i, j) indicates if the user j has shown interest in item i. At test time, we are given a new item q with the corresponding description $q_s \in \mathbb{R}^{1 \times m}$, and our goal is to predict $q_u \in \mathbb{R}^{1 \times u}$, *i.e.*, how likely is a user to show interest in the new item.

In this chapter, we address this problem by proposing two hybrid recommender systems, based on Non-negative Matrix Factorization (NMF), that exploit both the properties of the items ($\mathbf{X_s}$) and preferences of the users ($\mathbf{X_u}$). We preceded as follows:

- We provide a short overview of NMF and the algorithms used for its optimization, which we later use as foundations for proposing new algorithms applicable to the problem at hand (Section 3.1).
- We introduce an extension of NMF which jointly factorizes multiple matrices in a common basis, *Joint NMF* (JNMF). This allows us to link the properties of the items and preferences of the users. We develop two algorithms for the optimization of the new formulation and we study their convergence properties (Section 3.2).
- We also propose a variation of JNMF, *Joint NMF with Graph Regularization*, that takes into account the local geometric structure of the data and enforces smoothness of the solutions (Section 3.3).
- Finally, we perform an extensive experimental evaluation of the two methods and we compare their performance with three other approaches. We also study the stability of the methods under different parameter settings. (Section 3.4).

3.1 Non-negative Matrix Factorization

As our approach builds upon the foundations of the Non-negative Matrix Factorization (NMF), we start with a short overview of the NMF and of the algorithms for its optimization.

NMF aims at decomposing a matrix $\mathbf{X}$ in two non-negative matrices $\mathbf{W}$ and $\mathbf{H}$ (Figure 1). Unlike other factorization techniques, such as Singular Value Decomposition, it imposes non-negativity constants on the resulting matrices. These constraints result in an additive effect, that leads to a, so-called, "additive parts-based" representation of the data. For instance, in the context of text processing, $\mathbf{X}$ may be the document-term matrix (as is $\mathbf{X_s}$ in the case of news article prediction) and the factorization may represent the topics that appear in the documents ($\mathbf{H}$) and the extent to which each document belongs to them ($\mathbf{W}$). Another interesting example is the decomposition of the collaborative matrix (item-user, as $\mathbf{X_u}$ in news prediction), where the factorization matrices may be interpreted as communities of users with similar interests ($\mathbf{H}$) and the extent to which each item is preferred by each community ($\mathbf{W}$). These examples motivate us to think of NMF, not only as a technique appealing for exploratory analysis of the data, but also for prediction.

More formally the problem of Non-negative Matrix Factorization can be defined as follows. Given a non-negative matrix $\mathbf{X}_{n\times m}$, NMF aims at finding two positive, low-rank matrices ($\mathbf{W}_{n\times k}$ and $\mathbf{H}_{k\times m}$) whose product approximates $\mathbf{X}$. This is achieved by solving the following non-linear optimization problem:

$$\min \; J = \mathcal{L}(\mathbf{X}_{n\times m} - \mathbf{W}_{n\times k}\mathbf{H}_{k\times m}), \quad s.t. \;\; \mathbf{W} \geq 0, \mathbf{H} \geq 0,$$

where $\mathcal{L}$ is a loss function measuring the discrepancy between the original matrix and its approximation, and k is a parameter set by the user. The idea is graphically illustrated in Figure 1. Common choices for $\mathcal{L}$ include Frobenius Norm and Kullback–Leibler divergence, although there are other formulations using Relative Entropy, Bose–Einstein divergence, or Jensen–Shannon divergence [13]. For simplicity in the derivations, throughout this report, we use the Frobenius Norm formulation, resulting in the following optimization problem:

$$\min \; J = \frac{1}{2}||\mathbf{X} - \mathbf{WH}||_{\mathrm{F}}^2, \quad s.t. \;\; \mathbf{W} \geq 0, \mathbf{H} \geq 0. \tag{3.1}$$

By computing the Hessian, it may be verified that the problem is convex in either $\mathbf{W}$ or $\mathbf{H}$, but not in both. Therefore, there is no global minimum and the optimization algorithms can only guarantee convergence to a local minimum. Furthermore, it may be easily shown that there is no unique solution. If two matrices $\mathbf{W}$ and $\mathbf{H}$ are a solution, then $\mathbf{WD}$ and $\mathbf{D}^{-1}\mathbf{H}$ will also be a solution, for any positive diagonal matrix $\mathbf{D}$. To avoid this uncertainty, in practice, depending on the optimization approach (as we will see in the following sections), one may further normalize the Euclidean length of each row of $\mathbf{W}$ or $\mathbf{H}$ is equal to 1 to make solutions comparable, or introduce a Tikhonov regularization term to the original objective function.

3.1.1 Optimization techniques

We start by taking partial derivatives of J with respect to $\mathbf{W}$ and $\mathbf{H}$, which we use for later reference:

$$\frac{\partial J}{\partial \mathbf{W}} = \mathbf{WHH}^{\mathrm{T}} - \mathbf{XH}^{\mathrm{T}}, \qquad \frac{\partial J}{\partial \mathbf{H}} = \mathbf{W}^{\mathrm{T}}\mathbf{WH} - \mathbf{W}^{\mathrm{T}}\mathbf{X}. \tag{3.2}$$

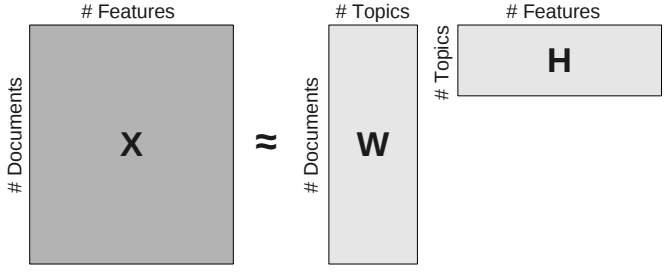


Figure 1: Graphical illustration of NMF.

Multiplicative update rules. Lee and Seung have introduced the, so-called, multiplicative update rules for solving Equation (3.1) [16]. Both matrices ($\mathbf{W}$ and $\mathbf{H}$) are randomly initialized to any positive values and the following updates are iteratively applied until convergence:

$$[\mathbf{W}]_{ij} \leftarrow [\mathbf{W}]_{ij} \cdot \frac{[\mathbf{X}\mathbf{H}^{\mathrm{T}}]_{ij}}{[\mathbf{W}\mathbf{H}\mathbf{H}^{\mathrm{T}}]_{ij}}, \tag{3.3a}$$

$$[\mathbf{H}]_{ij} \leftarrow [\mathbf{H}]_{ij} \cdot \frac{[\mathbf{W}^{\mathrm{T}}\mathbf{X}]_{ij}}{[\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{H}]_{ij}}. \tag{3.3b}$$

In practice, a small constant ϵ is added to denominator to avoid division by zero. The non-negativity constraint is implicitly enforced. Negative values cannot be reached, since $\mathbf{W}$ and $\mathbf{H}$ are initialized to positive values and the updates include only multiplications.

The rules are derived by applying the Karush–Kuhn–Tucker (KKT) first-order optimality conditions on J, stating the following:

$$\mathbf{W} \geq \mathbf{0} \qquad \mathbf{H} \geq \mathbf{0} \tag{3.4a}$$

$$\frac{\partial J}{\partial \mathbf{W}} \geq \mathbf{0} \qquad \frac{\partial J}{\partial \mathbf{H}} \geq \mathbf{0} \tag{3.4b}$$

$$\mathbf{W} \odot \frac{\partial J}{\partial \mathbf{W}} = \mathbf{0} \qquad \mathbf{H} \odot \frac{\partial J}{\partial \mathbf{H}} = \mathbf{0} \tag{3.4c}$$

where $\odot$ ($\oslash$) refers to the Hadamard, element-wise, product (division). Substituting the partial derivatives from Equation (3.2) in Equation (3.4) we obtain:

$$\mathbf{W} \odot (\mathbf{W}\mathbf{H}\mathbf{H}^{\mathrm{T}}) = \mathbf{W} \odot (\mathbf{X}\mathbf{H}^{\mathrm{T}}),$$
$$\mathbf{H} \odot (\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{H}) = \mathbf{H} \odot (\mathbf{W}^{\mathrm{T}}\mathbf{X}).$$

Hence, we obtain the multiplicative update rules from Equations (3.3a) and (3.3b). It can be proved (by means of auxiliary functions) that the objective function J (Equation (3.1)) is nonincreasing under the above update rules [16]. Moreover, from the KKT conditions it directly follows that every limit point found by these update rules is a stationary point.

Once the algorithm has converged, it is common to normalize each row of $\mathbf{W}$ to an Euclidean length of 1 and adjust the $\mathbf{H}$ matrix accordingly, so that the product $\mathbf{WH}$ does not change. By doing so, we assure that solutions are comparable, although they may not be the same before normalization, due to the problem of solution non-uniqueness. This can be achieved by applying the following rules:

$$[\mathbf{W}]_{ik} \leftarrow [\mathbf{W}]_{ik} / \sqrt{\sum_i [\mathbf{W}]_{ik}^2}, \qquad [\mathbf{H}]_{jk} \leftarrow [\mathbf{H}]_{jk} \cdot \sqrt{\sum_i [\mathbf{W}]_{ik}^2}.$$

The full procedure is summarized in Algorithm 1.

Algorithm 1 NMF, Multiplicative Update Rules.

1: Initialize $\mathbf{W}$ and $\mathbf{H}$ with random positive values
2: **repeat**
3: $\quad \mathbf{W} \leftarrow \mathbf{W} \odot [(\mathbf{X}\mathbf{H}^{\mathrm{T}}) \oslash (\mathbf{W}\mathbf{H}\mathbf{H}^{\mathrm{T}} + \epsilon)]$
4: $\quad \mathbf{H} \leftarrow \mathbf{H} \odot [(\mathbf{W}^{\mathrm{T}}\mathbf{X}) \oslash (\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{H} + \epsilon)]$
5: **until** stopping condition
6: $[\mathbf{W}]_{ik} \leftarrow [\mathbf{W}]_{ik} / \sqrt{\sum_i [\mathbf{W}]_{ik}^2} \qquad \forall i, k$
7: $[\mathbf{H}]_{jk} \leftarrow [\mathbf{H}]_{jk} \cdot \sqrt{\sum_i [\mathbf{W}]_{ik}^2} \qquad \forall j, k$

Alternating Least Squares. The first algorithm proposed for solving the NMF problem is the Alternating Least Squares method (ALS) [41]. As mentioned in the previous paragraphs, the problem is convex in either $\mathbf{W}$ or $\mathbf{H}$, but not in both simultaneously. Fixing one of them ($\mathbf{W}$ or $\mathbf{H}$), the problem becomes a least squares problem with a non-negativity constraint. Thus, the most natural way to approach the problem is applying an ALS method, leading to Algorithm 2.

Algorithm 2 NMF, Alternating Least Squares.

1: Initialize $\mathbf{W}$ with random positive values
2: **repeat**
3: Solve: $\min_{\mathbf{H} \geq \mathbf{0}} ||\mathbf{X} - \mathbf{WH}||^2$
4: Solve: $\min_{\mathbf{W} \geq \mathbf{0}} ||\mathbf{X}^{\mathrm{T}} - \mathbf{H}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}||^2$
5: **until** stopping condition

Since the least squares problems in Algorithm 2 (lines 3 and 4) can be perfectly decoupled into smaller sub-problems corresponding to the columns or rows of $\mathbf{X}$, we can directly apply non-negative least squares (NNLS) to each sub-problem. There exist algorithms specifically designed for solving the NNLS problem and are commonly used in practice. For instance, the NNLS algorithm of Lawson and Hanson [32] appears as a built-in function in MATLAB (`lsqnonneg`).

Alternating Constraint Least Squares. Although, the algorithms for solving NNLS are well developed, they are rather slow in practice. They rely on the active set optimization method, implying that only one variable can be swapped from the basis at a time. Even faster implementations of these algorithms (*e.g.*, [10]) are not fast enough, and the NNLS step still remains a computational bottleneck. To address this issue, a modified version of the ALS algorithm, called Alternating Constraint Least Squares (ACLS), has been proposed [31]. Instead of solving the constrained NNLS problem, ACLS solves the unconstrained version of the same problem and projects the negative elements to zero (Algorithm 3). The names of the algorithms are counter intuitive, but this is due to historical reasons. The ACLS algorithm is also commonly refereed as Inexact Alternating Least Squares. Uniqueness of the solution is imposed by introducing a Tikhonov regularization (Frobenius norm) on both $\mathbf{W}$ and $\mathbf{H}$ weighted by hyper-parameters λ_W and λ_H, respectively:

$$J = \frac{1}{2}||\mathbf{X} - \mathbf{WH}||^2 + \frac{\lambda_W}{2}||\mathbf{W}||^2 + \frac{\lambda_H}{2}||\mathbf{H}||^2.$$

Although, this ad-hock enforcement of non-negativity is not theoretically appealing, it works very well in practice. It completes the same number of iterations two orders of magnitude faster then ALS, while providing comparable solutions.

Discussion. ALS is the first algorithm proposed for solving the NMF introduced by Pattero and Tapper in 1994 [41]. However, the proposal of the multiplicative update rules by Lee and Seung (1999) [16] has been the key finding that made this technique appealing. As the rules are simple and easy to implement they have attracted tremendous attention in the literature. ACLS is a more recent algorithm proposed in 2006 [31] and commonly used as an alternative for the update rules (as it is also simple to implement).

Algorithm 3 NMF, Alternating Constraint Least Squares.

Input: λ_W, λ_H
1: Initialize $\mathbf{W}$ with random positive values
2: **repeat**
3: Solve for $\mathbf{H}$: $\mathbf{W}^{\mathrm{T}}\mathbf{X} = (\mathbf{W}^{\mathrm{T}}\mathbf{W} + \lambda_H\mathbf{I})\ \mathbf{H}$ % $\mathbf{W}$ fixed
4: Set all negative elements in $\mathbf{H}$ to 0
5: Solve for $\mathbf{W}$: $\mathbf{X}\mathbf{H}^{\mathrm{T}} = \mathbf{W}\ (\mathbf{H}\mathbf{H}^{\mathrm{T}} + \lambda_W\mathbf{I})$ % $\mathbf{H}$ fixed
6: Set all negative elements in $\mathbf{W}$ to 0
7: **until** stopping condition

In what follows, we compare the three algorithms from different perspectives and point out some of the advantages and disadvantages of each.

Locking phenomenon. It happens when some entries of the matrices $\mathbf{W}$ and $\mathbf{H}$ are set to zero in some of the iterative steps. Due to the nature of the multiplicative update rules, these entries will never be set to any other value in the later iterations, thus will be locked to zero. On the other hand, the ALS and ACLS do not suffer from this phenomenon.

Sparsity. The notion of sparsity refers to a representation scheme where only a few features are effectively used to represent the data vectors. Sparsity is desired because of storage and interpretability. Applying the projection to zero in every step of ACLS implicitly enforces sparsity. Thus, compared to the solutions obtained using ALS or the multiplicative update rules, it tends to provide sparser solutions.

Time to converge. In practice, the ALS algorithm is seldom used due to its inefficiency. In terms of the number of iterations needed for convergence, ACLS usually requires fewer iterations then the multiplicative rules. In our experience ACLS tends converge faster while providing sparser solutions [31].

Convergence issues. Some researchers have also questioned the convergence of the multiplicative update rules to a stationary point. While the proposed updates fulfill KKT optimality condition of Equation (3.4), they do not guarantee that Equation (3.4b) holds. This theoretical weakness does not seem to strongly affect the algorithm in practice. Lin [33] proposes a slight modification of the rules for which the convergence to a stationary point is guaranteed. However, this modification increases the computational requirements of the algorithm.

3.2 Joint Non-negative Matrix Factorization

In the previous section, we have described the classical NMF methods that factorize a single data matrix. In this section, we propose a new NMF formulation that jointly factorizes two data matrices and establishes a link between the two factorizations (Figure 2).

Given the problem defined, items are associated with a description and a set of users who consumed them. In the case of news, each news is explained by the set of words in the article and all the users who commented on it. This information is then represented with two matrices, a document-term matrix $\mathbf{X_s} \in \mathbb{R}^{n \times v}$, and a document-user matrix $\mathbf{X_u} \in \mathbb{R}^{n \times u}$, where n is the number of documents, v is the vocabulary size and u is the number of users. The document-term matrix ($\mathbf{X_s}$) may be a boolean matrix or may represent the TF-IDF score of the words in the document. On the other hand, the entries of document-user matrix ($\mathbf{X_u}$) reflect whether a given user commented on a given article. As both matrices are non-negative, NMF can be applied to each of them. As we have seen in the previous section, a factorization of the document-term matrix will represent the topics that appear in the documents ($\mathbf{H}$), and the extent to which each document belongs to them ($\mathbf{W}$). In a similar way, a factorization of the document-user matrix can represent the communities (groups) of users ($\mathbf{H}$), and how much each document is preferred by each group ($\mathbf{W}$). The topics and the communities may also be seen as hidden (unobserved) variables which describe the documents. If decomposed separately each factorization will represent a different hidden space, one for users and one for words. The idea of our approach is that

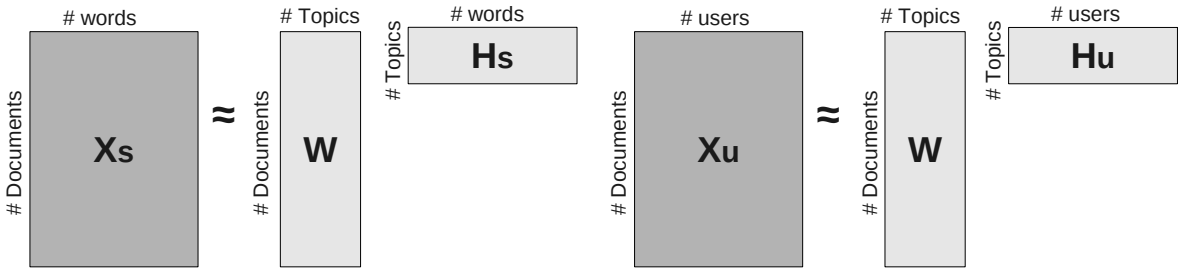


Figure 2: Graphical representation of the NMF model defined. Notice that the matrix $\mathbf{W}$ is common to both factorizations.

both, documents and users, may be represented in a common hidden space (Figure 2). To achieve this, we have to factorize both $\mathbf{X_s}$ and $\mathbf{X_u}$ simultaneously and enforce a low-dimensional representation in common space.

More formally, given the matrices $\mathbf{X_s}$ and $\mathbf{X_u}$, we define the following optimization problem:

$$\min : J = \frac{1}{2}(\alpha||\mathbf{X_s} - \mathbf{W}\mathbf{H_s}||^2 + (1-\alpha)||\mathbf{X_u} - \mathbf{W}\mathbf{H_u}||^2 + \lambda_W||\mathbf{W}||^2 + \lambda_{H_s}||\mathbf{H_s}||^2 + \lambda_{H_u}||\mathbf{H_u}||^2)$$
$$s.t. \quad \mathbf{W} \geq 0, \mathbf{H_s} \geq 0, \mathbf{H_u} \geq 0 \tag{3.5}$$

The first and the second term correspond to the factorization of the matrices $\mathbf{X_s}$ and $\mathbf{X_u}$, respectively, and $\alpha \in [0, 1]$ is a hyper-parameter that controls the importance of each factorization. Setting $\alpha = 0.5$ gives equal importance to both factorizations, while values of $\alpha > 0.5$ (or $\alpha < 0.5$) give more importance to the factorization of $\mathbf{X_s}$ (or $\mathbf{X_u}$). The remaining terms are *Tikhonov* (Frobenius norm) regularization of $\mathbf{W}$, $\mathbf{H_u}$, and $\mathbf{H_s}$, controlled by the hyper-parameters λ_W, λ_W, and λ_W, respectively. They are used to enforce smoothness of the solution. Note that the model requires that $\mathbf{X_s}$ and $\mathbf{X_u}$ have the same number of rows, but not necessary the same number of columns: each factorization will result in an $\mathbf{H}$ matrix of the corresponding size. This is of great importance as the multiple representations of the objects usually contain different number of features (*e.g.*, in case of the news articles the size of the vocabulary is different then the number of users considered).

3.2.1 Optimization Algorithms

We start by computing the partial derivatives with respect to each optimization variable:

$$\frac{\partial J}{\partial \mathbf{W}} = \alpha\mathbf{W}\mathbf{H_s}\mathbf{H_s}^{\mathrm{T}} - \alpha\mathbf{X_s}\mathbf{H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{W}\mathbf{H_u}\mathbf{H_u}^{\mathrm{T}} - (1-\alpha)\mathbf{X_u}\mathbf{H_u}^{\mathrm{T}} + \lambda_W\mathbf{W} \tag{3.6a}$$
$$\frac{\partial J}{\partial \mathbf{H_s}} = \alpha\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{H_s} - \alpha\mathbf{W}^{\mathrm{T}}\mathbf{X_s} + \lambda_{H_s}\mathbf{H_s} \tag{3.6b}$$
$$\frac{\partial J}{\partial \mathbf{H_u}} = (1-\alpha)\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{H_u} - (1-\alpha)\mathbf{W}^{\mathrm{T}}\mathbf{X_u} + \lambda_{H_u}\mathbf{H_u} \tag{3.6c}$$

Similar to the classical NMF problem, the proposed objective function is convex in $\mathbf{W}$, $\mathbf{H_s}$, and $\mathbf{H_u}$, but not in all of them simultaneously. Thus, one can only guarantee to find a local minimum, but not a global one. We propose two algorithms for optimizing the objective function of Equation (3.2) based on (1) multiplicative update rules, and (2) alternating constrained least squares.

Multiplicative Update Rules. Applying the Karush–Kuhn–Tucker (in the spirit of Equation (3.4)), we derive the following multiplicative update rules:

$$[\mathbf{W}]_{ij} \leftarrow [\mathbf{W}]_{ij} \cdot \frac{[\alpha\mathbf{X_s}\mathbf{H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{X_u}\mathbf{H_u}^{\mathrm{T}}]_{ij}}{[\alpha\mathbf{W}\mathbf{H_s}\mathbf{H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{W}\mathbf{H_u}\mathbf{H_u}^{\mathrm{T}} + \lambda_W\mathbf{W}]_{ij}}, \tag{3.7a}$$
$$[\mathbf{H_s}]_{ij} \leftarrow [\mathbf{H_s}]_{ij} \cdot \frac{[\alpha\mathbf{W}^{\mathrm{T}}\mathbf{X_s}]_{ij}}{[\alpha\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{H_s} + \lambda_{H_s}\mathbf{H_s}]_{ij}}, \tag{3.7b}$$
$$[\mathbf{H_u}]_{ij} \leftarrow [\mathbf{H_u}]_{ij} \cdot \frac{[(1-\alpha)\mathbf{W}^{\mathrm{T}}\mathbf{X_u}]_{ij}}{[(1-\alpha)\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{H_u} + \lambda_{H_u}\mathbf{H_u}]_{ij}}. \tag{3.7c}$$

THEOREM 1: The objective function J (Equation (3.2)) in non-increasing under the above update rules.

Please see the Appendix A for a detailed proof of the theorem. Our proof essentially follows the idea of the proof in the Lee and Seung's paper [16] for the original NMF.

The full algorithm is summarized in Algorithm 4. Note that a small constant ϵ is added to the denominators of the update rules to avoid division by zero.

Algorithm 4 Joint NMF, Multiplicative Update Rules.

Input: k, α, λ_W, λ_{H_s}, λ_{H_u}

1: Initialize $\mathbf{W}$, $\mathbf{H_s}$ and $\mathbf{H_u}$ with random positive values
2: **repeat**
3: $\quad \mathbf{W} \leftarrow \mathbf{W} \odot [(\alpha \mathbf{X_s H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{X_u H_u}^{\mathrm{T}}) \oslash (\alpha \mathbf{W H_s H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{W H_u H_u}^{\mathrm{T}} + \lambda_W \mathbf{W} + \epsilon)]$
4: $\quad \mathbf{H_s} \leftarrow \mathbf{H_s} \odot [(\alpha \mathbf{W}^{\mathrm{T}} \mathbf{X_s}) \oslash (\alpha \mathbf{W}^{\mathrm{T}} \mathbf{W H_s} + \lambda_{H_s} \mathbf{H_s} + \epsilon)]$
5: $\quad \mathbf{H_s} \leftarrow \mathbf{H_s} \odot [((1-\alpha)\mathbf{W}^{\mathrm{T}} \mathbf{X_u}) \oslash ((1-\alpha)\mathbf{W}^{\mathrm{T}} \mathbf{W H_u} + \lambda_{H_u} \mathbf{H_u} + \epsilon)]$
6: **until** stopping condition

Alternating Constrained Least Squares. One may notice that if one of the parameters is fixed ($\mathbf{W}$, $\mathbf{H_s}$, or $\mathbf{H_u}$) then the objective to solve becomes a least squares problem. In the spirit of the alternating constrained least squares (ACLS) for the classical NMF, we propose an algorithm for solving the Joint NMF formulation (Algorithm 5). Instead of solving the constrained non-negative least squares problem, we solve without considering the non-negativity constraints and we project the negative values to zero. The projection step significantly reduces the computation time and provides sparser solutions.

Algorithm 5 Joint NMF, Alternating Constraint Least Squares.

Input: k, α, λ_W, λ_{H_s}, λ_{H_u}

1: Initialize $\mathbf{W}$ with random positive values
2: **repeat**
3: $\quad$ Solve for $\mathbf{H_s}$: $\alpha \mathbf{W}^{\mathrm{T}} \mathbf{X_s} = (\alpha \mathbf{W}^{\mathrm{T}} \mathbf{W} + \lambda_{H_s} \mathbf{I})\ \mathbf{H_s}$ $\quad$ % $\mathbf{W}$, $\mathbf{H_u}$ fixed
4: $\quad$ Set all negative elements in $\mathbf{H_s}$ to 0
5: $\quad$ Solve for $\mathbf{H_u}$: $(1-\alpha)\mathbf{W}^{\mathrm{T}} \mathbf{X_u} = [(1-\alpha)\mathbf{W}^{\mathrm{T}} \mathbf{W} + \lambda_{H_u} \mathbf{I}]\ \mathbf{H_u}$ $\quad$ % $\mathbf{W}$, $\mathbf{H_s}$ fixed
6: $\quad$ Set all negative elements in $\mathbf{H_u}$ to 0
7: $\quad$ Solve for $\mathbf{W}$: $\alpha \mathbf{X_s H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{X_u H_u}^{\mathrm{T}} = \mathbf{W}\ [\alpha \mathbf{H_s H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{H_u H_u}^{\mathrm{T}} + \lambda_W \mathbf{I})$ % $\mathbf{H_s}$, $\mathbf{H_u}$ fixed
8: $\quad$ Set all negative elements in $\mathbf{W}$ to 0
9: **until** stopping condition

Stopping conditions. We impose two stopping conditions: (1) maximum number of iterations, and (2) minimum relative change in the objective function. The first criterion is not a mathematically appealing way to control the number of iterations, as it is problem dependent, but it is useful in the cases when the user is willing to spend only a limited amount of time or resources. The second criterion, on the other hand, checks if the difference in the objective function (J) from one iteration to another is below a threshold ϵ:

$$|J^{(i-1)} - J^{(i)}| \leq \epsilon.$$

If the changes in the objective function are extremely small it is very likely that the algorithm has reached a local minimum. Finally, if any of the stopping conditions is fulfilled the algorithm stops.

Convergence. Figure 3 shows typical case of the convergence behavior of the two algorithms. We run both algorithms, multiplicative update rules and ACLS, on the NIPS data set setting $k = 50$, $\alpha = 0.5$, $\lambda = 0.5$ and stopping condition with $\epsilon = 0.1$. The multiplicative update rules take small steps and require 60 iterations to converge, while ACLS converges in only 30 iterations. Moreover, the ACLS iterations take less time to complete, and the algorithm requires 5.2 seconds to converge, while the multiplicative update rules require 29.1 seconds.

As a comparison, we run the ACLS (Section 3.1, Algorithm 3) only on the document-term matrix ($\mathbf{X_s}$) and the algorithm converges in 5.8 seconds and takes 34 iterations. The reason for the slight improvement in the computational performance of the JNMF over the classical NMF is that: (1) the problem is more constrained, specifically in terms of $\mathbf{W}$, and thus the states space is reduced; (2) more observed information is introduced in the model, *i.e.*, both $\mathbf{X_s}$ and $\mathbf{X_u}$. Thus, although we have introduced a more complex model, its optimization algorithms *do not suffer from reduced computational performance.*

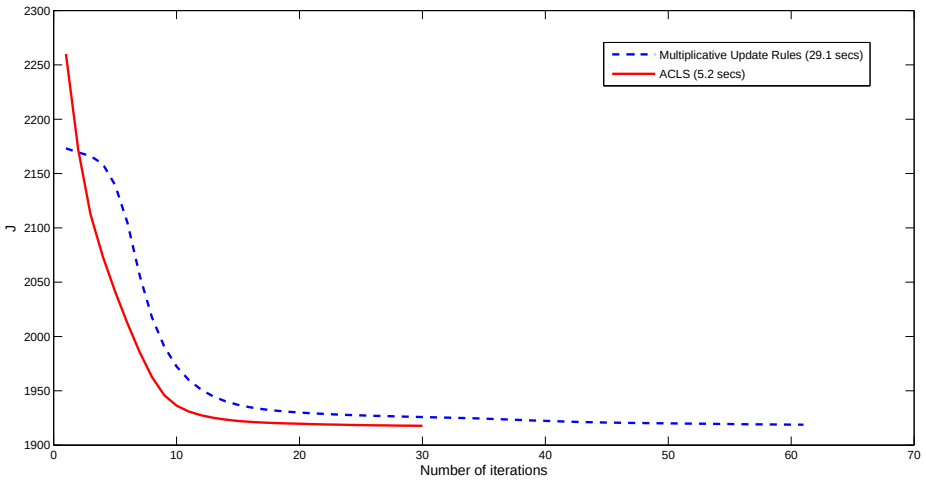


Figure 3: Change of the objective function J as a function of the number of iterations for both algorithms, run on the NIPS data set with $k = 50$, $\alpha = 0.5$, $\lambda = 0.5$ and stopping condition with $\epsilon = 0.1$. Both algorithms converge to similar values of J. However, the multiplicative update rules require more iterations and time to converge.

3.2.2 Implementation Tricks

A careful implementation significantly reduces the computational expense of the algorithms. In this section, we present several simple implementation details which allow us to achieve the computational performance reported.

Computing the objective function is an expensive step which has to performed in every iteration. The most efficient way to compute it is by using the fact that the Frobenius norm can be expressed in terms of traces[1]. Thus, the objective function of Equation (3.2) can be written as:

$$\begin{aligned} J &= \frac{\alpha}{2}[\text{trace}(\mathbf{X_s}^{\mathrm{T}}\mathbf{X_s}) - 2\ \text{trace}(\mathbf{H_s}^{\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\mathbf{X_s})) + \text{trace}(\mathbf{H_s}^{\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{H_s}))] \\ &+ \frac{(1-\alpha)}{2}[\text{trace}(\mathbf{X_u}^{\mathrm{T}}\mathbf{X_u}) - 2 \cdot \text{trace}(\mathbf{H_u}^{\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\mathbf{X_u})) + \text{trace}(\mathbf{H_u}^{\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{H_u}))] \\ &+ \frac{1}{2}[\lambda_W \text{trace}(\mathbf{W}^{\mathrm{T}}\mathbf{W}) + \lambda_{H_s}\text{trace}(\mathbf{H_s}^{\mathrm{T}}\mathbf{H_s}) + \lambda_{H_u}\text{trace}(\mathbf{H_u}^{\mathrm{T}}\mathbf{H_u})]. \end{aligned}$$

Notice that $\text{trace}(\mathbf{X_s}^{\mathrm{T}}\mathbf{X_s})$ and $\text{trace}(\mathbf{X_u}^{\mathrm{T}}\mathbf{X_u})$ are constant and need to be computed only once. Also, this expression allows us to adjust the order of matrix multiplication (notice the additional brackets) to achieve maximum efficiency. Moreover, some terms are already computed for the updates and thus may be stored and reused. For instance, $\mathbf{W}^{\mathrm{T}}\mathbf{W}$, $\mathbf{W}^{\mathrm{T}}\mathbf{X_s}$ and $\mathbf{W}^{\mathrm{T}}\mathbf{X_u}$ are computed in both algorithms, ACLS and the multiplicative update rules.

Finally, recalling the definition of trace, one may notice that when computing the trace of a product of two matrices, the full product of the matrices not necessarily needs to be computed. As the trace is the sum of the diagonal elements of the matrix, it is sufficient to compute just these elements. More specifically, when the two matrices are of the same size, we only need to compute the sum of element-wise product. Let's consider two matrices $\mathbf{X}_{n\times k}$ and $\mathbf{Y}_{n\times k}$ and suppose that we want to compute the trace of their product $\mathbf{X}\mathbf{Y}^{\mathrm{T}}$, then we may compute it as:

$$\text{trace}(\mathbf{X}\mathbf{Y}^{\mathrm{T}}) = \sum_i [\mathbf{X}\mathbf{Y}^{\mathrm{T}}]_{ii} = \sum_{ij} [\mathbf{X} \odot \mathbf{Y}]_{ij}.$$

We realize that this observation greatly reduces both the number of multiplications and the memory required to compute the traces in the objective function, especially when k is large.

[1] $||\mathbf{X}||_{\mathrm{F}}^2 = \text{trace}(\mathbf{X}^{\mathrm{T}}\mathbf{X})$

3.2.3 Inference

Once the model is trained and $\mathbf{W}$, $\mathbf{H_s}$ and $\mathbf{H_u}$ are obtained, we may use them for prediction. Given a query document with representation in only one view, we can predict the other. For instance, given the content of a new news article q_s, we can predict the users which are most likely to leave a comment, *i.e.*, q_u. We project the document to the common hidden space by solving for w in $q_s = w\mathbf{H_s}$. Then using the low dimensional representation w we may approximate the representation in other view as: $q_u = w\mathbf{H_u}$. Each element of q_u represents a score of how likely it is that the user will comment the new article. Given these scores we may rank the users. The complete procedure is summarized in Algorithm 6.

Algorithm 6 Joint NMF, Inference.

Input: q_s, $\mathbf{H_s}$, and $\mathbf{H_u}$

1: Solve for w: $q_s = w\mathbf{H_s}$
2: Set all negative elements in w to 0
3: Compute q_u: $q_u = w\mathbf{H_u}$

3.3 Joint NMF with Graph Regularization

In the previous section, we have introduced Joint NMF, an NMF formulation that allows us to jointly factorize two data matrices. In this section, we extend JNMF by adding an additional term that takes into account the local geometric structure of the data.

Recall that when performing JNMF factorization, we attempt to find a common low-dimensional space that is optimized for the linear approximation of the data from both views. We also (implicitly) suppose that the data from both views are drawn from a common distribution P. One may hope that additional knowledge of the distribution P can be exploited for a better discovery of the low-dimensional space. A natural assumption could be that *if two data points x_i and x_j (in any view) are close in the intrinsic geometry of the distribution, then the representations of these two data points in the low-dimensional space should also close to each other.* This assumption is commonly referred to as *manifold assumption* and plays an essential role in algorithms for dimensionality reduction [6] and semi-supervised learning [7, 53, 54]. Figure 4 shows an example which motivates this assumption. The data is distributed in a form of a "Swiss roll" and taking into account the local geometrical structure of the data will lead to a discovery of a better low dimensional representation.

In reality the geometric structure of the distribution P is not known and cannot be directly used. However, recent studies on spectral graph theory [12] and manifold learning [5] have demonstrated that the local geometric structure can be effectively modeled through a nearest neighbor graph on a scatter of data points. Consider a graph with n nodes where each node represents a data point. For each point we find the p nearest neighbors and we connect the corresponding nodes in the graph. The edges may be binary (1 if one of the nearest neighbors, 0 otherwise) or may be weighted (e.g., cosine similarity). This results in a matrix $\mathbf{A}$ which can later be used to measure the local closeness of two points x_i and x_j.

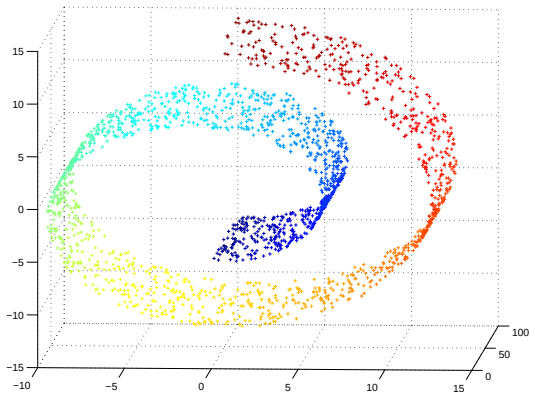


Figure 4: Data distribution in a form of a "Swiss roll". Taken from [5].

Recall that the Joint NMF maps each data point x_i into a low-dimensional representation w_i (a row of the matrix $\mathbf{W}$). A natural way to measure the distance between two low dimensional representations, given the choice of a loss function, is by computing the Euclidean distance: $||w_i - w_j||^2$. Using the above defined weight matrix $\mathbf{A}$ we may measure the smoothness of the low dimensional representation as follows:

$$\begin{aligned} S &= \frac{1}{2}\sum_{i,j=1}^{n} ||w_i - w_j||^2 \mathbf{A}_{ij} \\ &= \sum_{i=1}^{n}(w_i^T w_i)\mathbf{D}_{ij} - \sum_{i,j=1}^{n}(w_i^T - w_j)\mathbf{A}_{ij} \\ &= \text{trace}(\mathbf{W}^{\mathrm{T}}\mathbf{D}\mathbf{W}) - \text{trace}(\mathbf{W}^{\mathrm{T}}\mathbf{A}\mathbf{W}) = \text{trace}(\mathbf{W}^{\mathrm{T}}\mathbf{L}\mathbf{W}), \end{aligned}$$

where $\mathbf{D}$ is a diagonal matrix whose entries are the row (or column, as $\mathbf{A}$ is symmetric) sums of $\mathbf{A}$, *i.e.*, $\mathbf{D}_{ii} = \sum_i \mathbf{A}_{ij}$. $\mathbf{L} = \mathbf{D} - \mathbf{A}$ is the called the Laplacian matrix of the graph [12].

New objective function. Given the above, we may modify the Joint NMF model as to enforce smooth low-dimensional representations of the data. This leads to the Joint NMF with Graph Regularization:

$$\begin{aligned} \min : J = \frac{1}{2}(\alpha||\mathbf{X_s} - \mathbf{W}\mathbf{H_s}||^2 + (1-\alpha)||\mathbf{X_u} - \mathbf{W}\mathbf{H_u}||^2 + \beta\text{trace}(\mathbf{W}^{\mathrm{T}}\mathbf{L}\mathbf{W}) \\ +\lambda_W||\mathbf{W}||^2 + \lambda_{H_s}||\mathbf{H_s}||^2 + \lambda_{H_u}||\mathbf{H_u}||^2) \\ s.t. \quad \mathbf{W} \geq 0, \mathbf{H_s} \geq 0, \mathbf{H_u} \geq 0, \end{aligned} \tag{3.8}$$

where $\mathbf{L}$ is the Laplacian matrix of the graph, and β is a hyper-parameter which controls the extent to which smoothness is enforced. It is easy to check that when $\beta = 0$ the formulation is equivalent to JNMF.

3.3.1 Optimization algorithms

The partial derivatives with respect to $\mathbf{H_s}$ and $\mathbf{H_u}$ remain the same as in the Joint NMF formulation (Equations (3.6b) and (3.6c), respectively), while the partial derivative with respect to $\mathbf{W}$ becomes:

$$\frac{\partial J}{\partial \mathbf{W}} = \alpha\mathbf{W}\mathbf{H_s}\mathbf{H_s}^{\mathrm{T}} - \alpha\mathbf{X_s}\mathbf{H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{W}\mathbf{H_u}\mathbf{H_u}^{\mathrm{T}} - (1-\alpha)\mathbf{X_u}\mathbf{H_u}^{\mathrm{T}} + \beta\mathbf{L}\mathbf{W} + \lambda_W\mathbf{W}. \tag{3.9}$$

The properties of the problem remain the same in terms of convexity and uniqueness of the solution. We propose two algorithms for solving the problem based on multiplicative update rules and alternating constrained least squares.

Multiplicative Update Rules. Applying the Karush–Kuhn–Tucker first-order optimality conditions, we derive multiplicative update rules. The rules for $\mathbf{H_s}$ and $\mathbf{H_u}$ remain the same as in the Joint NMF formulation (Equations (3.7b) and (3.7c), respectively), while the update rule for $\mathbf{W}$ changes to the following:

$$[\mathbf{W}]_{ij} \leftarrow [\mathbf{W}]_{ij} \cdot \frac{[\alpha\mathbf{X_s}\mathbf{H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{X_u}\mathbf{H_u}^{\mathrm{T}} + \beta\mathbf{A}\mathbf{W}]_{ij}}{[\alpha\mathbf{W}\mathbf{H_s}\mathbf{H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{W}\mathbf{H_u}\mathbf{H_u}^{\mathrm{T}} + \beta\mathbf{D}\mathbf{W} + \lambda_W\mathbf{W}]_{ij}}. \tag{3.10}$$

THEOREM 2: The objective function J (Equation (3.8)) in non-increasing under the above update rules.

Please see the Appendix B for a detailed proof of the theorem. Our proof essentially follows the idea of the proof in the paper of Cai *et al.* [11] for the Graph Regularized NMF.

The full procedure is given in Algorithm 7.

Algorithm 7 Joint NMF with Graph Regularization, Multiplicative Update Rules.

Input: $\mathbf{A}$, k, α, β, λ_W, λ_{H_s}, λ_{H_u}

Compute the diagonal matrix $\mathbf{D}$ as: $\mathbf{D}_{ii} = \sum_i \mathbf{A}_{ij}$
Initialize $\mathbf{W}$, $\mathbf{H_s}$ and $\mathbf{H_u}$ with random positive values
repeat
  $\mathbf{W} \leftarrow \mathbf{W} \odot [(\alpha \mathbf{X_s H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{X_u H_u}^{\mathrm{T}} + \beta \mathbf{AW})$
    $\oslash (\alpha \mathbf{W H_s H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{W H_u H_u}^{\mathrm{T}} + \beta \mathbf{DW} + \lambda_W \mathbf{W} + \epsilon)]$
  $\mathbf{H_s} \leftarrow \mathbf{H_s} \odot [(\alpha \mathbf{W}^{\mathrm{T}} \mathbf{X_s}) \oslash (\alpha \mathbf{W}^{\mathrm{T}} \mathbf{W H_s} + \lambda_{H_s} \mathbf{H_s} + \epsilon)]$
  $\mathbf{H_s} \leftarrow \mathbf{H_s} \odot [((1-\alpha)\mathbf{W}^{\mathrm{T}} \mathbf{X_u}) \oslash ((1-\alpha)\mathbf{W}^{\mathrm{T}} \mathbf{W H_u} + \lambda_{H_u} \mathbf{H_u} + \epsilon)]$
until stopping condition

Alternating Constrained Least Squares. Similar to the JNMF formulation, fixing $\mathbf{H_s}$ or $\mathbf{H_u}$ leads to a least squares problem. However, setting to zero the partial derivative with respect to $\mathbf{W}$, we obtain:

$$\alpha \mathbf{X_s H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{X_u H_u}^{\mathrm{T}} = \alpha \mathbf{W H_s H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{W H_u H_u}^{\mathrm{T}} + \beta \mathbf{LW} + \lambda_W \mathbf{W}.$$

Notice that in the new term ($\beta \mathbf{LW}$), $\mathbf{W}$ appears after the other elements, unlike in all other terms where it appears before. Because of this, we are unable to separate $\mathbf{W}$ and solve for it.

Fortunately, this is a specific kind of matrix equation which have been well-studied in the field of control theory [4]. The equation of the form $\mathbf{AX} + \mathbf{XB} = \mathbf{C}$ is commonly referred to as the *Sylvester equation* and may be solved for $\mathbf{X}$ by using the following property:

$$\mathrm{vec}(\mathbf{AXB}) = (\mathbf{B}^{\mathrm{T}} \otimes \mathbf{A})\mathrm{vec}(\mathbf{X}),$$

where $\otimes$ is the Kronecker product, and **vec** is the vectorization operator. Applying this property to the partial derivative with respect to $\mathbf{W}$ (Equation (3.9)) we obtain:

$$\begin{aligned} &\mathrm{vec}(\alpha \mathbf{X_s H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{X_u H_u}^{\mathrm{T}}) = \\ &\quad = \alpha \mathrm{vec}(\mathbf{W H_s H_s}^{\mathrm{T}}) + (1-\alpha)\mathrm{vec}(\mathbf{W H_u H_u}^{\mathrm{T}}) + \beta \mathrm{vec}(\mathbf{LW}) + \lambda_W \mathrm{vec}(\mathbf{W}) \\ &\quad = \alpha(\mathbf{H_s H_s}^{\mathrm{T}} \otimes \mathbf{I})\mathrm{vec}(\mathbf{W}) + (1-\alpha)(\mathbf{H_u H_u}^{\mathrm{T}} \otimes \mathbf{I})\mathrm{vec}(\mathbf{W}) + \beta(\mathbf{I} \otimes \mathbf{L})\mathrm{vec}(\mathbf{W}) + \lambda_W \mathrm{vec}(\mathbf{W}) \\ &\quad = [\alpha(\mathbf{H_s H_s}^{\mathrm{T}} \otimes \mathbf{I}) + (1-\alpha)(\mathbf{H_u H_u}^{\mathrm{T}} \otimes \mathbf{I}) + \beta(\mathbf{I} \otimes \mathbf{L}) + \lambda_W \mathbf{I}]\mathrm{vec}(\mathbf{W}). \end{aligned} \tag{3.11}$$

Thu, we are able to separate $\mathbf{W}$ (*i.e.*, $\mathrm{vec}(\mathbf{W})$) and to obtain a least squares problem. Based on this finding we propose Algorithm 8. Notice that once that once we solve for the vectorization of $\mathbf{W}$ ($\mathrm{vec}(\mathbf{W})$), we have to use the *reshape* operator to transform the matrix $\mathbf{W}$ to its original size.

Algorithm 8 Joint NMF with Graph Regularization, Alternating Constraint Least Squares.

Input: $\mathbf{A}$, k, α, β, λ_W, λ_{H_s}, λ_{H_u}

Compute the diagonal matrix $\mathbf{D}$ as: $\mathbf{D}_{ii} = \sum_i \mathbf{A}_{ij}$, and the Laplacian matrix $\mathbf{L}$ as: $\mathbf{L} = \mathbf{D} - \mathbf{A}$
Initialize $\mathbf{W}$ with random positive values
repeat
  Solve for $\mathbf{H_s}$: $\alpha \mathbf{W}^{\mathrm{T}} \mathbf{X_s} = (\alpha \mathbf{W}^{\mathrm{T}} \mathbf{W} + \lambda_{H_s} \mathbf{I})\ \mathbf{H_s}$ % $\mathbf{W}$, $\mathbf{H_u}$ fixed
  Set all negative elements in $\mathbf{H_s}$ to 0
  Solve for $\mathbf{H_u}$: $(1-\alpha)\mathbf{W}^{\mathrm{T}} \mathbf{X_u} = [(1-\alpha)\mathbf{W}^{\mathrm{T}} \mathbf{W} + \lambda_{H_u} \mathbf{I}]\ \mathbf{H_u}$ % $\mathbf{W}$, $\mathbf{H_s}$ fixed
  Set all negative elements in $\mathbf{H_u}$ to 0
  Set $\mathbf{C} = \alpha \mathbf{X_s H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{X_u H_u}^{\mathrm{T}}$
  Set $\mathbf{B} = \alpha(\mathbf{H_s H_s}^{\mathrm{T}} \otimes \mathbf{I}) + (1-\alpha)(\mathbf{H_u H_u}^{\mathrm{T}} \otimes \mathbf{I}) + \beta(\mathbf{I} \otimes \mathbf{L}) + \lambda_W \mathbf{I}$
  Solve for $\mathrm{vec}(\mathbf{W})$: $\mathrm{vec}(\mathbf{C}) = \mathbf{B} \cdot \mathrm{vec}(\mathbf{W})$ % $\mathbf{H_s}$, $\mathbf{H_u}$ fixed
  Reshape $\mathrm{vec}(\mathbf{W})$ to the original size of $\mathbf{W}$: $\mathbf{W} = reshape(\mathrm{vec}(\mathbf{W}), n, k)$
  Set all negative elements in $\mathbf{W}$ to 0
until stopping condition

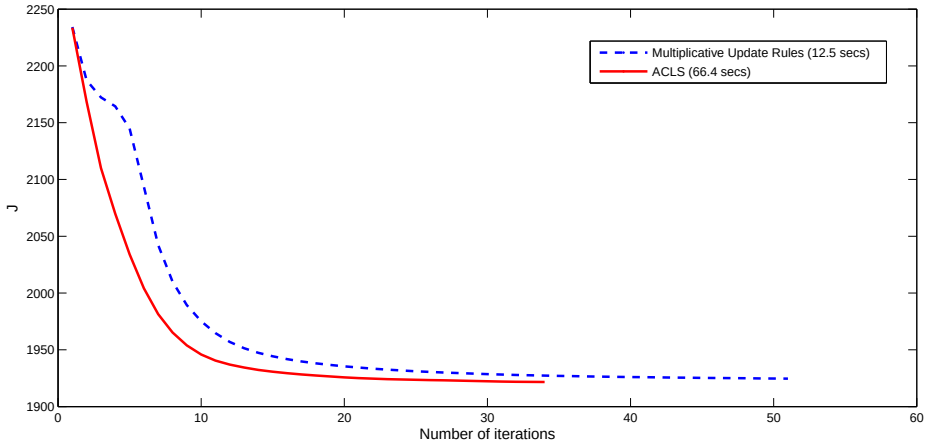


Figure 5: Change of the objective function J as a function of the number of iterations for both algorithms, run on the NIPS data set with $k = 50$, $\alpha = 0.5$, $\beta = 0.25$, $\lambda = 0.5$ and stopping condition with $\epsilon = 0.1$. The ACLS algorithms requires less iterations but also more time to perform each iteration, and thus needs more time to converge.

Convergence. Figure 5 shows a typical case of the convergence behavior of the two algorithms. We run both algorithms on the NIPS data set setting $k = 50$, $\alpha = 0.5$, $\beta = 0.25$, $\lambda = 0.5$ and stopping condition with $\epsilon = 0.1$. We compute the regularization graph by considering the first nearest neighbor of each document, based on the document content and weight corresponding to the cosine similarity. The multiplicative update rules take small steps and require 51 iterations to converge, while ACLS converges in 34 iterations. However, each iteration of ACLS takes more time to complete and consequently the algorithm needs more time to converge, *i.e.*, 66.4 seconds compared to the multiplicative update rules which converge in 12.5 seconds. Thus, unlike in the case of the Joint NMF formulation, the multiplicative update rules outperform the ACLS algorithm. The reason for the reduced performance of the ACLS is the size of the system of linear equations that needs to be solved for $\mathbf{W}$ in each iteration. In the Joint NMF formulation, one has to solve a linear system of size $k \times k$ to infer $\mathbf{W}$, where k is the number of topics. However, in the new formulation, due to the Sylvester equation, a system of nk equations and nk variables needs to be solved, where n is the number of documents. Although, the system can be solved in $O(n^3)$ time using specialized techniques [4], the computational cost of the algorithm still depends on the number of documents, making the multiplicative update rules preferable when running the model on large data sets.

3.4 Experimental Results

In this section, we evaluate the performance of the two approaches, Joint NMF and JNMF with Graph Regularization, on three tasks: item cold-start, author prediction, and email-recipient prediction. The goal is to test whether the proposed methods outperform the existing content-based approaches.

3.4.1 Data sets

For the purpose of the experiments we consider three data sets. (1) A collection of Yahoo! News articles and user comments for the task of item cold-start recommendations, (2) collection of scientific articles from NIPS for author prediction, and (3) the Enron data set for prediction of email recipients.

Yahoo! News. We consider a random sample of news articles and the corresponding comments posted on the Yahoo! News website during a period of 40 days. The data set contains >41K articles, >3.5M comments posted by >650K users. The size of the vocabulary is ∼60K (*i.e.*, number of unique tokens in all articles) and >9M tokens. The content of the articles has been preprocessed such that all tokens are converted to lower case, and stop-words, digits, punctuation, short (< 3 characters) and infrequent (appearing < 3 times) tokens are removed.

NIPS. The data set contains papers from the NIPS conferences between 1987 and 1999 (13 years). The conference is characterized by contributions from a number of different research communities in

	#documents	#users	|vocabulary|
Yahoo! News	41K	650K	60K
NIPS	1.7K	2K	13K
Enron	36K	5K	12K-56K

Table 1: Size of the different data sets.

the general area of learning algorithms. The collection contains 1,740 articles written by 2,037 distinct authors. There are 2,307,375 tokens and a vocabulary size of 13,649 unique tokens. The articles have been preprocessed by converting all tokens to lower case, removing punctuation and stop-words.

Enron. The data is composed of email messages released during investigation of the Federal Energy Regulatory Commission against the Enron Corporation. We consider the 10 largest mailboxes and within each mailbox only the emails sent by the owner. The total number of emails is 36010 sent to 4984 recipients. The size of the vocabularies for each mailbox ranges from 12,375 to 56,193 unique tokens. The messages have been preprocessed by removing the headers (from/to/cc fields), converting all tokens to lower case and removing numbers, stop-words and infrequent (appearing < 1 time) tokens.

3.4.2 Metrics

The output of the proposed methods is a ranking of which items are most preferred by the user (in the item cold-start recommendation), or who are the most likely authors/recipients of a document/email (in the author and email recipient prediction task). It is important to realize the that in the tasks of author and email recipient prediction we have reliable feedback of the ranking result, *i.e.*, we explicitly know who are the authors/recipients and who are not. Therefore, precision based metrics more suitable. On the other hand, when making recommendations we do not have a reliable feedback of which news articles were undesired, as not commenting an article may stem from multiple different reasons. Thus, precision based metrics are not appropriate, as they require knowing which articles were undesired to a user. However, commenting an article is an indication of interest in it, making recall oriented metrics appropriate. Taking this into account, we evaluate the author and email prediction task using the state-of-the-art metrics from Information Retrieval: Micro and Macro F1, MAP, and NDCG; while we use the Ranking percentile/accuracy at 3, 5, 7, 10 to evaluate the item cold-start recommendations.

F1-score is defined as the harmonic mean of recall (R) and precision (P) *i.e.*, $F1 = 2PR/(P+R)$. We combine the F1 scores across different documents by computing the *micro* (combining predictions from documents and computing F1-score) and *macro* (averaging the F1-scores across all documents) average. The micro F1 is an unweighted average, *i.e.*, all authors/recipients are equally weighted, while the macro F1 is weighted average, where active authors/recipients are given more importance.

MAP (*Mean Average Precision*) summarizes the ranking of each query document by averaging the precision values from the rank positions where a relevant item is retrieved. The scores across different documents are aggregated by computing the mean (*i.e.*, average).

NDCG (*Normalized Discounted Cumulative Gain*) is based on two assumptions (1) highly relevant items are more useful and (2) the lower the ranked position of a relevant item the less useful it is for the user. Thus, highly relevant items appearing lower in a result are penalized as the graded relevance value is reduced logarithmically proportional to the position of the result. Formally the $NDCG$ is defined as:

$$NDCG = \frac{DCG}{iDCG}, \qquad DCG = rel_1 + \sum_{i=2}^{n} \frac{rel_i}{\log_2 i},$$

where rel_i is the relevance of the document at rank i, and $iDCG$ is the ideal DCG, *i.e.*, the prefect ranking of the results.

Percentile-ranking/Ranking accuracy is recall based measure and is suitable when no reliable feedback of the recommendations is available [22]. The ranking $rank_{u,i}$ for user u of each recommended

article i is 0%, if i is predicted to be the most desirable article for u, while it is 100% if i is predicted to be the least desirable. We average each $rank_{u,i}$ across all users u's and all articles i's:

$$\overline{rank} = \frac{\sum_{u,i} comment_{u,i} \cdot \ rank_{u,i}}{\sum_{u,i} comment_{u,i}},$$

where $comment_{u,i}$ is an indicator function that equals to: 1 if the user u commented on article i; and 0 otherwise. The lower $\overline{rank}$, the better the quality of the ranking. For random predictions, the expected value of $rank_{u,i}$ is 50%. Thus, if $\overline{rank} < 50\%$, then the algorithm is better than random. To ease illustration, we covert the percentile ranking into ranking accuracy. This is 1 (best/ideal predictions), if the percentile ranking is 0%; and it is 0 (random predictions), if the percentile ranking is 50%:

$$accuracy = \frac{50\% - \overline{rank}}{50\%}.$$

3.4.3 Baselines for comparison

We compare our models, Joint NMF (JNMF) and Joint NMF with Graph Regularization (JNMF-GR), to three other approaches: pure content-based recommender, content-topic-based recommender and a matrix factorization version of the author topic-model.

Content-based recommender (CB). We build a profile of each user based on the properties of the items preferred in the past. We find that weighting each item inversely proportional to the number of times the item was liked leads to improved performance. Thus, we build the user profile U as: $U = \sum_{i \in I}(\vec{v_i}/freq_i)$, where I is the set of items the user liked and $\vec{v_i}$ is the description of item i and $freq_i$ is the number of times the item i was liked in overall. In test time, we rank the items by computing the cosine similarity between the description of the new items and the user profiles.

Content-topic-based recommender (CTB). We extract topics from the content of the items by applying NMF and we describe each item as a mixture of the topics. We then build a topic profile for each user based on the topics of the items liked in the past. In test time, we infer the topics of the new items and we rank the items based on the cosine similarity between the item's topics and the user topic profiles. The CTB recommender allows us to investigate the importance of performing joint factorization of both, the content and collaborative matrix, instead of factorizing only the content matrix.

Author topic-model (ATM). Although, the author topic-model [44] is based on a generative probabilistic model, the authors also provide a matrix factorization interpretation. They propose to factorize the content matrix $\mathbf{X_s}$ as: $\mathbf{X_s}_{(n \times v)} \approx \mathbf{X_u}_{(n \times u)}\mathbf{Z}_{(u \times k)}\mathbf{H}_{(k \times v)}$, where n is the number of documents, v is the vocabulary size, u is the number of users, and k is the number of topics. The matrix $\mathbf{H}$ corresponds to the topics and the matrix $\mathbf{Z}$ corresponds to the topic vector associated to each user. During training, we compute $\mathbf{Z}$ and $\mathbf{H}$, and keep $\mathbf{X_u}$ fixed. At test time, given the new articles $\mathbf{X_s}^{test}$ we keep $\mathbf{Z}$ and $\mathbf{H}$ fixed and we compute $\mathbf{X_u}^{test}$, leading to scores for each user-item pair.

3.4.4 Evaluation Protocol

In all tasks the data is intrinsically influenced by the time, thus we sort the data chronologically and produce train/test subsets by shifting a time window, instead of sampling by random. In the NIPS data set we consider training period of 8 years and we predict the next year, shifting for one year every time, resulting in 5 folds. In Enron, we divide each of the 10 mailboxes in 80% training and 20% testing. On the Yahoo! News data set we train using the past 30 days and we predict on the comments in the next day, shifting for one day every time, resulting in 10 folds. In the test sets we consider only the authors/recipients/users that appear in the train set. We tune the hyper-parameters of each model on an independent validation set (20% of the training). Finally, we report the average performance over all folds and we evaluate the statistical significance of the differences in performance by using a paired t-test.

3.4.5 Results

Author prediction. We compute the average performance of each method in all 5 folds (Figure 6 and Table 2). Joint NMF (JNMF) and Joint NMF with Graph Regularization (JNMF-GR) achieve the best results. They outperform the CB and CTB recommenders as well as the ATM. All differences are statistically significant ($p < 0.001$). The JNMF performs slightly better then the JNMF-GR with performance improvement ranging from 1.7% to 2.4%, across different measures. However, the difference is not statistically significant, indicating that exploiting the local geometric structure of the data does not lead to an improved performance.

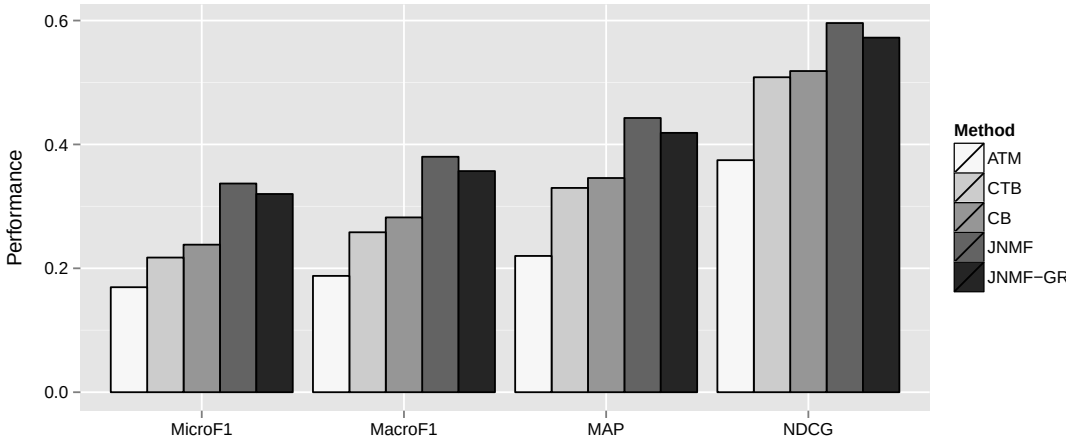


Figure 6: Comparison of the different methods on the NIPS dataset.

	MicroF1	MacroF1	MAP	NDCG
JNMF	**33.69%**	**38.01%**	**44.27%**	**59.60%**
JNMF-GR	32.00%	35.70%	41.87%	57.24%
CB	23.82%	28.22%	34.58%	51.85%
CTB	21.74%	25.82%	32.98%	50.85%
ATM	16.95%	18.78%	22.00%	37.46%

Table 2: Results on for the task of author prediction obtained on the NIPS data set.

Email recipient prediction. We evaluate the methods on each of the 10 mailboxes and we compute the average performance (Figure 7 and Table 3). JNMF and JNMF-GR perform better then the other methods in all measures with differences ranging from 4%-15%. All differences are statistically significant ($p < 0.005$). JNMF and JNMF-GR achieve similar results, one performing better then the other with <1% depending on the evaluation measure considered. However, the differences in the results are not statistically significant.

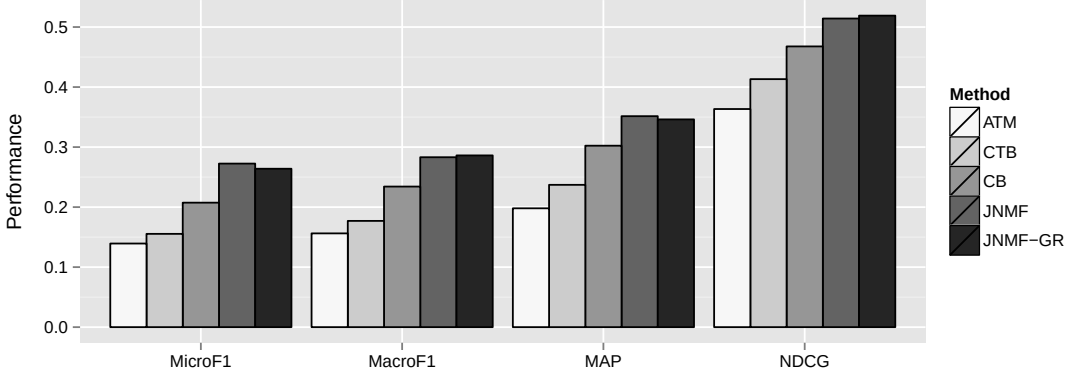


Figure 7: Comparison of the different methods on the Enron dataset.

	MicroF1	MacroF1	MAP	NDCG
JNMF	**27.25%**	28.31%	**35.15%**	51.42%
JNMF-GR	26.39%	**28.61%**	34.61%	**51.90%**
CB	20.74%	23.42%	30.23%	47.77%
CTB	15.54%	17.70%	23.71%	41.32%
ATM	13.93%	15.62%	19.80%	36.34%

Table 3: Results on for the task of email recipient prediction on the Enron data set.

Item cold-start recommendation. We evaluate the different methods in each of the 10 testing days and we compute the average performance (Figure 8 and Table 4). We test all methods expect the ATM, which due to the large footprint of the data required a great amount of memory. All algorithm perform better then random (RA $> 0\%$). JNMF and JNMF-GR outperform all other methods with statistically significant differences ($p < 0.001$). JNMF-GR achieves better ranking accuracy then JNMF in all positions, however the difference diminishes as we consider larger lists. The difference is statistically significant only for RA@3 ($p < 0.05$). This indicates that considering the local geometric structure of the data allow the algorithm to push the relevant items towards the top of the list.

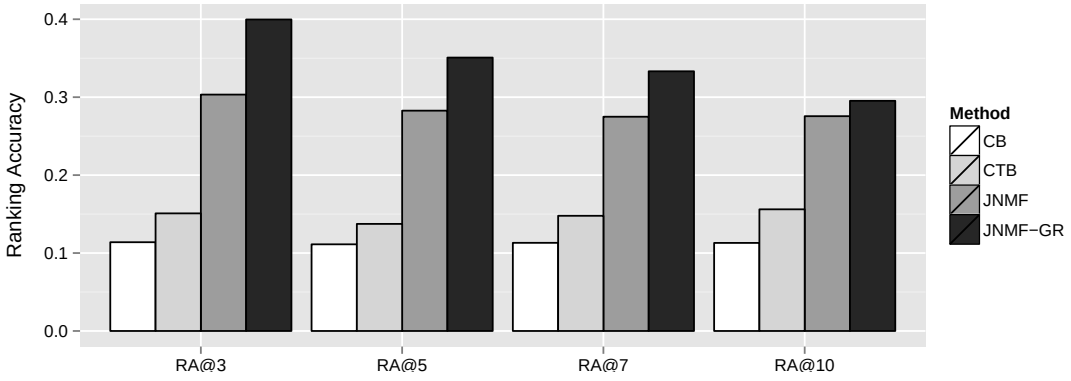


Figure 8: Comparison of the different methods on the Yahoo! News dataset.

	RA@3	RA@5	RA@7	RA@10
JNMF	30.33%	28.27%	27.49%	27.56%
JNMF-GB	**39.97%**	**35.08%**	**33.32%**	**29.53%**
CB	11.39%	11.12%	11.31%	11.30%
CTB	15.09%	13.74%	14.78%	15.61%

Table 4: Results of item cold-start recommendation on the Yahoo! News website.

3.4.6 Choices of Regularization Graphs in JNMF-GR

There are many choices of how to define the regularization graph $\mathbf{A}$ in JNMF-GR. During the course of the experiments, we consider p-nearest neighbor graphs based on the content of the data ($\mathbf{X_s}$) and the collaborative information ($\mathbf{X_u}$). For each, we use two weighting schemes: *binary*, setting to 1 the weights to the p nearest neighbors and to 0 all other; and *cosine*, where the weights to the p nearest neighbors are set to the cosine similarity between the vectors and to 0 for all other nodes. We test for $p = \{1, 2, 3\}$. The performance of the different choices is depicted in Figure 9.

On the NIPS data set, building the graph based on the content information is slightly more useful then using the collaborative information. Best results are obtained setting $p = 1$ (*i.e.*, considering only the first nearest neighbor) when using the content and $p = 2$ when using the collaborative information. For any choice of the regularization graph the performance of JNMF-GR is lower then JNMF, but higher then CB recommender.

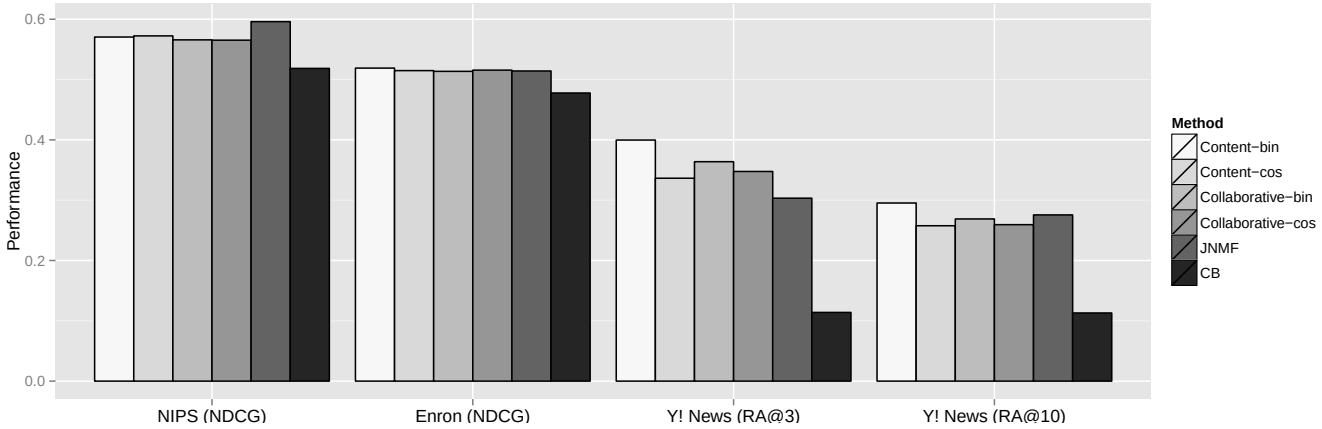


Figure 9: Comparison between performance of JMNF-GR with different regularization graphs. The performance of the Joint NMF and the Content-based recommender are included for reference.

On the Enron data set, all regularization graphs lead to similar performance. The best choice for p is $p = 2$ for the binary and $p = 1$ for the cosine weighted graphs. The performance of JNMF-GR is slightly better then JNMF, but difference is not statistically significant.

On the Yahoo! News data set, the best performance is achieved when building the graph based on content information and using binary weights. Setting $p = 1$ works best for the binary and $p = 2$ for the cosine weighting schemes. If we consider the Ranking Accuracy at 3 (RA@3), any choice of the regularization graph leads to performance then JNMF. On the other hand, if we consider RA@10 the differences diminish and only the content-based binary regularization graph leads to improved performance.

3.4.7 Parameter analysis

JNMF has three essential parameters: k, the number of latent variables; α, weight of the importance of each view of the data; and λ ($\lambda = \lambda_W = \lambda_{H_s} = \lambda_{H_u}$), controlling the smoothness of the solution. Figure 10 shows a typical behavior of the algorithm for different values of the parameters.

The parameter k controls the complexity of the model, small values of k (simple models) under-fit whereas large values of k over-fit the data and lead to poor performance (Figure 10). Thus, one has to find a balance between the two that fits best the problem at hand. As the latent variables are mainly used for prediction and do not need to be interpreted (as in topic modeling), one may use large number of hidden variables without sacrificing the usefulness of the model. On the other hand, balancing the two views of the data, *i.e.*, $\alpha \approx 0.5$ tends to achieve the best performance. Figure 10 also suggests that giving slightly more importance to the collaborative information (*e.g.*, $\alpha \in [0.2, 0.5]$) may be helpful. Finally, imposing the smoothness of the solutions helps. However, imposing it too strongly, *i.e.* large values of λ, decreases the performance. Setting α between 0 and 1 leads to stable and high performance.

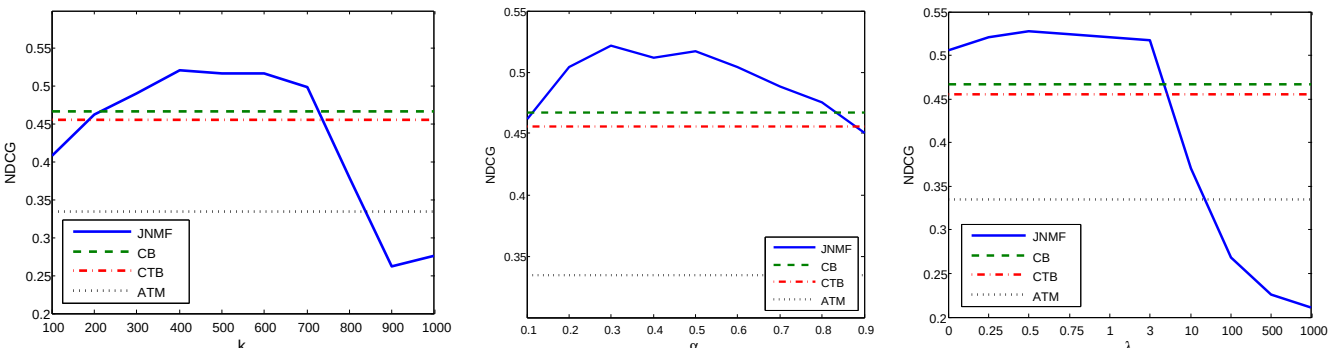


Figure 10: Behavior of the JNMF hyper-parameters.

4 User Cold-Start

Up to now, we have considered the item cold-start scenario, when no collaborative information for an item is available (*i.e.*, no users have rated the item yet). In this chapter, we address the problem of user cold-start, *i.e.*, when no previous information about the user is available. In the specific problem we consider, recommendation of news articles, the user cold-start is a very common scenario, the majority of the visits on the online news platforms are by users who are not logged in, *i.e.*, no user profile is available. Due to its frequency, the problem is of great importance, as even small improvements may result in big increase of the overall satisfaction of the users.

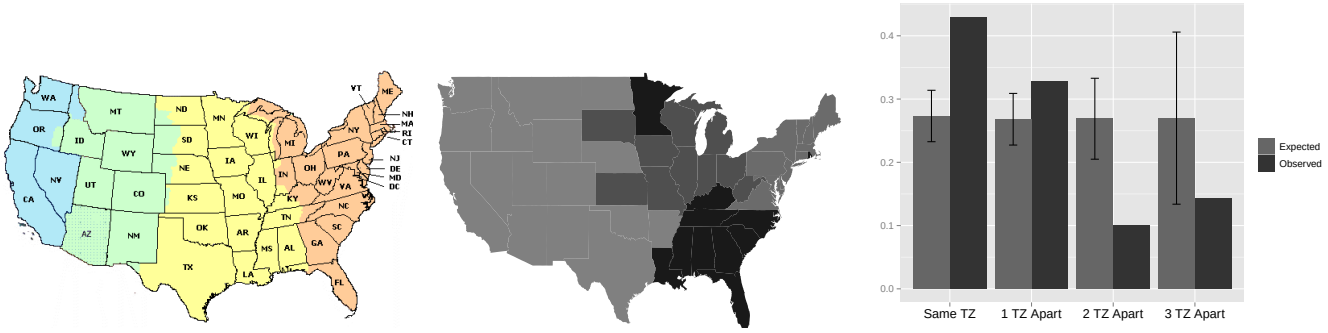
As no profiles are available for these users, the most common way to serve them is by recommending the items which are currently most popular, leading to one-size-fits-all recommendations. However, although we do not have any previous information about the users, their requests come with some contextual information. For instance, using a user's IP address, we may infer the geographical location of the user and his/her time zone. If we have a better understanding of the needs of the user in specific regions or in specific times of the day, we may use the contextual information from the request to partially overcome the user cold-start problem.

Recent studies have shown that our body clocks do not only regulate our actions actions offline, but also online. Golder and Macy examined status updates on Twitter to analyze how users' mood (which is regulated by their internal clocks) changes during a day [20]. They found that changes of the use of emotion words were predictable and mainly based on different daytimes (*e.g.*, work, sleep, daylight). Mislove *et al.* [38] analyzed how the posting of tweets changes during a week, and how tweets spread across time zones. In so doing, they observed geographic variations of the mood intensity, with those in the west coast posting happier tweets consistently three hours behind those in the east coast.

Online actions whose geographic processes have been well studied include not only posting status updates on Twitter [52, 23, 25], but also uploading pictures on Flickr [15, 51], and visiting Foursquare venues [40, 34].

However, the geographic processes of online engagement on news platforms has not been widely studied. In quest to overcome the user cold-start problem, we set out to close this gap by studying the geographical processes on the Yahoo! News website for the USA. We consider a dataset containing articles and user comments posted on the website for more than two years, and we make the following contributions:

- We find that users engage with each other depending on where they live (Section 4.2). More specifically, users in the same time zone preferentially engage with each other, producing what we call "the time zone bubble".
- We test the extent to which such an engagement bubble is created also by their interests in specific topics (Section 4.4). We find that, based on the articles they comment, users living in the same time zone or one time zone apart tend to have common (topical) interests, while those two or three time zones apart tend to have different ones.
- Since one's interests have been linked to one's socio-demographic conditions and personality traits, we test whether this is also the case at geographic level, and we do so by combing our online data with census data (Section 4.5).
- Based on those findings, we build a recommender system to partly overcome the user cold-start problem (Section 4.6). When one does not know anything about users, one could still recommend the most popular articles in the users' time zones. We find that suggesting what is popular in a zone rather than what is popular in the whole USA improves the recommendation accuracy by a factor of one and a half.



(a) USA Map of time zones.

(b) US Map of like-minded states. These are states engaging on the same articles. Four different clusters of like-minded states emerge.

(c) The probability that two users who are k time zone (TZ) apart engage on the same news article.

Figure 11: The time zone bubble.

4.1 Initial Analysis

4.1.1 Data Description

The dataset consists of a random sample of 200K news articles and corresponding 41M comments, published from August 2010 to February 2013. Yahoo! News features articles from a variety of news publishers including: Reuters, ABC News, Associated Press, The Atlantic Wire and other. The content of each article comes with its publication time and comments. Each comment comes with a timestamp, the commenter's anonymous user identifier and *IP* address (which we translate into the corresponding city name using the Yahoo! Places Web service).

4.1.2 State Commenting Graph

To understand whether any geographical process shapes user engagement, we build a graph whose nodes are US states and whose links are weighted with the number of times two users in states i and j comment on the same article. To see the extent to which different states show similar commenting patterns (whether they are like-minded, in that, they tend to engage with the same articles), we apply a community detection algorithm on the graph. We use the Louvain community detection algorithm [9], whose main advantages are both the automatic detection of the number of communities and being one of the most competitive state-of-the-art approaches [17]. After running the algorithm, four main clusters of like-minded states are detected and mapped in Figure 11b. Interestingly, we see that the four detected groups are geographically clustered (*i.e.*, cover contiguous regions). Furthermore, one readily sees a similarity between this map and the USA Map of time zones (Figure 11a). This suggests that a hidden time zone effect drives user engagement. Next, we test whether this is true.

4.2 The Time Zone Bubble

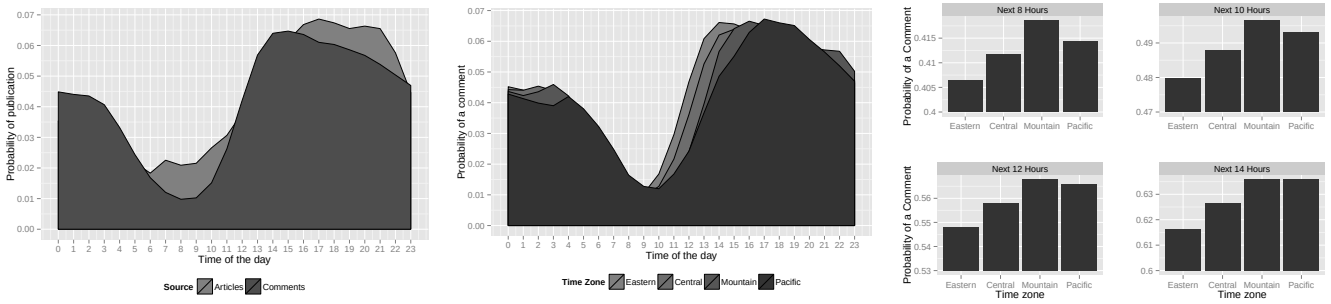
To quantify whether time zone affects engagement, we test this hypothesis:

[H1] *Users in the same time zone preferentially engage with the same articles, while users in different time zones engage with different articles.*

To this end, we perform an experiment in three steps (which we shall detail): (1) We measure the observed engagement among users in the same time-zone, 1 time zone, ..., k time zones apart; (2) By keeping all factors constant except the time zone which are randomly permuted, we measure again the user engagement due to chance; and (3) we compare both engagement measures to assess if time zone affects engagement.

(1) Engagement in k-time zone apart. To measure engagement, we associate users with their time zones[2] and count the number of times users from k-time zone apart engage in the same articles. More

[2]States that belong to more than one time zone are assigned to the time zone in which the majority of the territory belongs to. We considered only the continental states, Alaska and Hawaii have been excluded from the analysis.



(a) Article publishing cycle and global commenting cycle (across all time zones), according to EST time.

(b) Commenting cycle of each time zone, aligned to EST time.

(c) Time advantage for the different time zones, if users comment within the next 8, 10, 12 and 14 hours.

Figure 12: The time advantage.

formally, we measure the probability p_k two users in k time zones apart engage in the same article:

$$p_k = \frac{\sum_{i \in S} \sum_{j \in S} I_k(i,j) \cdot \ interaction_{ij}}{n},$$

where S is the set of all states; I_k is an indicator function that equals to 1 if states i and j are k time zones apart, or 0 otherwise; $interaction_{ij}$ is the number of times users from states i and j have engaged in the same article; and n normalizes the numerator for the total number of interactions across all time zones.

(2) Engagement due to chance. To test whether what we observe is not due to chance, we resort to a null (random) model [47]. We reshuffle the assignment of time zones by associating each user to a random zone, and repeat this procedure 2000 times to obtain accurate estimates. The random model removes the time zone effect and keeps all other factors constant. Thus, the difference between the engagement values that are observed and those in the random model depend only on effects strictly related to time zones. If there is no difference, then what we observe does not depend on time zone.

(3) Compare the two engagements. By comparing the observed engagement with the engagement under the random model (Figure 11c), we find that users in the same time zone and (to a lesser extent) those one time zone away engage with the same articles (first two dark bars) more than expected by chance (light bars). By contrast, those in three and four-time zone away engage less than chance. We perform a t-test (as the distribution of levels of engagement follows a normal distribution) to verify whether the differences between observed values and those in the random model are statistically significant. We find that all differences are significant at p-value less than 0.001. Similar results are obtained when performing non-parametric hypothesis test.

4.3 Time Advantage

One reason why users in the same time zone preferentially engage with each other might be that the publishing cycle of articles might match the commenting cycle of their time zones. We plot the publishing cycle and the commenting cycle across all time zones (Figure 12a); and the commenting cycle for each time zone (Figure 12b). We find that the commenting cycles for the different time zones all follow the same pattern, in that, they are consistently shifted by one hour.

We then quantify the advantage A a time zone might have within the next n hours. We do so by computing the probability of commenting in that time zone (commenting cycle) given the availability of articles (publishing cycle):

$$A = \sum_{i=0}^{n} \sum_{t=0}^{23} (p_t^a \ \cdot \ p_{t+i}^c) \ ,$$

where p_t^a is the probability of a new article being published at time t (with $t \in [0, 23]$), and p_{t+i}^c is the probability of a comment at time $t + i$. By computing the advantage of each time zone within the next 8, 10, 12, and 14 hours (Figure 12c), we find that The Mountain and Pacific time zones have slightly more advantage over the Eastern and Central time zones.

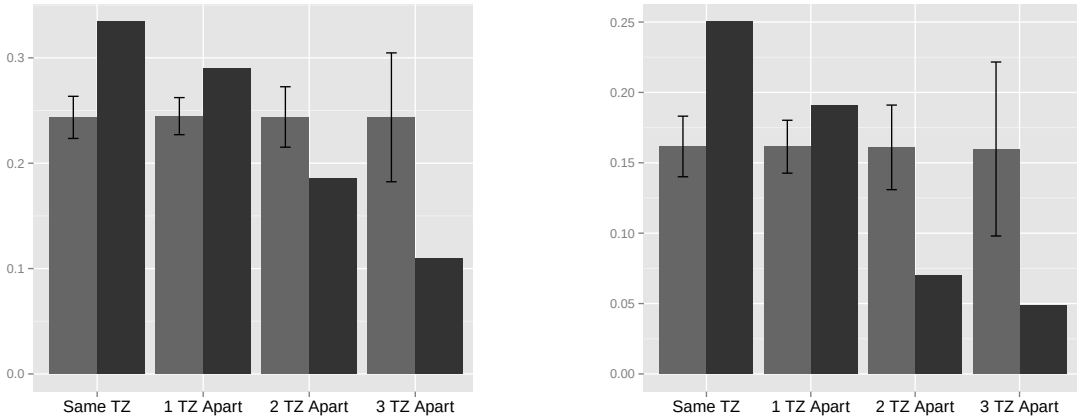


Figure 13: Topical similarities (dark bars) between states at k-time zone away as opposed to similarities in the random model (light bars) for articles (left panel) and comments (right). These results are for number of topics $t = 20$, yet similar results are obtained for $t = \{30, 50\}$.

4.4 Time Zone Interests

We have shown that users across different time zones engage with different articles, but that does not necessarily mean they are interested in different topics: different articles might well be about the same topics. Since previous experimental studies have shown that residents in the same area tend to be like-minded (a tendency often called "geographic sorting" [8]), we hypothesize:

[H2] *Users in the same time zone are interested in the same topics, while those in different time zones are interested in different topics.*

To test it, we need to separately assign topics to articles and comments, and aggregate those topics at the level of state. In so doing, given two states, we could consider the corresponding two topical vectors and measure their similarity. We could then track the average similarity of the states that are k time zone apart, and compare the results to a null model (in which we reshuffle topical associations across states at random). The most important step is the assignment of topics to articles and comments. To do so, we perform topical modeling using *Non-negative Matrix Factorization* (Section 3.1), using different number of topics $t = \{20,\ 30,\ 50\}$. We find that, for both articles and comments (Figure 13), users in the same time zone and one time zone away (first two dark bars) read and engage with the same topics; by contrast, those two and three time zones apart (last two dark bars) engage on the same topics less than chance. We perform a t-test to test the statistical significance of the difference, and we find that all differences are significant at p-value less than 0.001. Similar results are obtained when performing non-parametric hypothesis test.

4.5 What Makes the Bubble

We have seen that users across time zones engage with different articles, and that is partly because of different interests. Since one's interests have been often linked to one's socioeconomic conditions and personality traits [43], we test whether that is true not only at individual level but also at US state level. We opt for US state and not time zone because socioeconomic and personality indicators are defined at the former level.

To begin with, we assign topics to both articles and comments. Since we need explicit topic labels (previously we just needed to compute similarity measures), we cannot use unsupervised techniques such as NMF as we have done so far. Instead, we opt for studying a subset (13.8%) of the articles that have been editorially labeled with topical categories from the IPTC news subject taxonomy [14]. The taxonomy consists of 1400 topics and is organized into three levels, according to the specificity of the topics. To have the finest-grained topical view, we use the lowest level of the taxonomy. The number of labels associated with each article ranges from 1 to 25, where the average number of labels per article

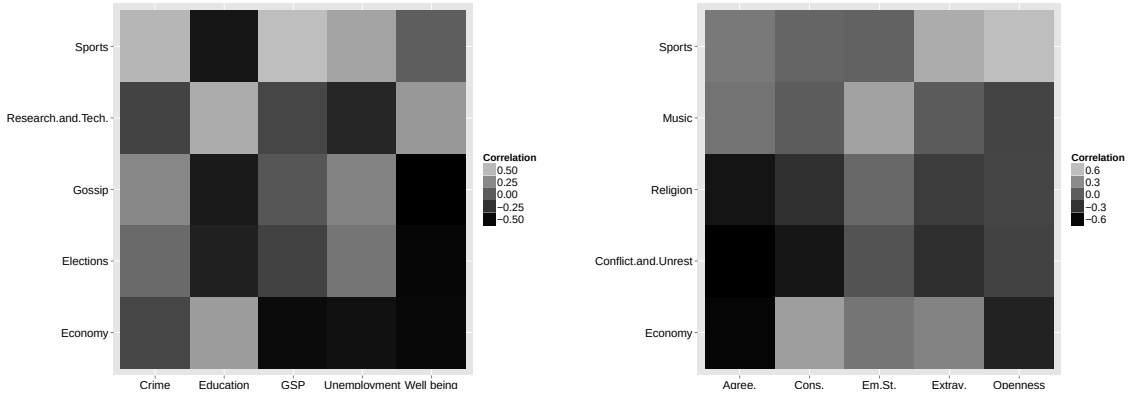


Figure 14: Correlation between state's topics of interest and: socioeconomic indicators (left panel) and personality traits (right panel).

is 5. We aggregate these topics at state level by considering the number of times users from a given state commented on articles with a certain tag, and the number of times the tag appears in the data set (to avoid the bias of dominant topics).

Socioeconomic Indicators. We analyze the correlations between a state's assigned topics and the five most studied socioeconomic indicators: well-being index[3], crime level[4], rate of unemployment[5], Gross State Product[6], and education level[7] (number of people with higher education). We report only the correlations for which the test for association between paired samples is significant (p-value < 0.01). In what follows r denotes the correlation coefficient observed.

As illustrated in Figure 14, states with high levels of well-being (satisfaction with life) do not engage with articles about economy&business&finance ($r = -0.50$), about elections ($r = -0.53$), or about gossip&celebrities ($r = -0.53$). Economy is also not popular in states with unemployment ($r = -0.46$). Sport, instead, is popular in states with high levels of crime ($r = 0.48$), unemployment ($r = 0.39$), and low gross state product ($r = 0.52$); it is, instead, not very popular in states with high levels of education ($r = -0.43$) whose residents prefer to engage with articles about research&technology ($r = 0.43$) and avoid those on celebrities ($r = -0.40$). States with high levels of education also tend to be interested in diverse topics (as the Shannon diversity of states topical vector, $r = 0.44$).

The Big Five Personality Traits. The five-factor model of personality, or the big five, is the most comprehensive, reliable and useful set of personality concepts [19]. An individual is associated with five scores that correspond to the five main personality traits and that form the acronym of *OCEAN* (Table 5 collates a brief explanation). Imaginative, spontaneous, and adventurous individuals are high in **O**peness. Ambitious, resourceful and persistent individuals are high in **C**onscientiousness. Individuals who are sociable and tend to seek excitement are high in **E**xtraversion [2, 49]. Those high in **A**greeableness are trusting, altruistic, tender-minded, and are motivated to maintain positive relationships with others [26]. Finally, emotionally liable and impulsive individuals are high in **N**euroticism [29].

These big five traits have been studied not only at individual level but also at geographic level [42]. Rentfrow *et al.* [43] have examined the personality scores of half a million US residents and found clear patterns of regional variation across the country, and they have also strong relationships between state-level personality and socioeconomic indicators.

[3] `http://www.thewellbeingindex.com`
[4] `http://www.ucrdatatool.gov`
[5] `http://www.bls.gov/web/laus/lauhsthl.htm`
[6] `http://www.usgovernmentspending.com`
[7] `http://www.census.gov`

Personality trait	High scorers	Low scorers
Openness	Imaginative	Conventional
Conscientiousness	Organized	Spontaneous
Extraversion	Outgoing	Solitary
Agreeableness	Trusting	Competitive
Neuroticism	Prone to stress and worry	Emotionally stable

Table 5: The big five personality traits.

We now correlate state-level personality scores with engagement with articles about specific topics. Economy is popular in states with conscientious residents ($r = 0.42$), and unpopular in states with residents who tend to be agreeable ($r = -0.61$) and open ($r = -0.42$). Sport articles are popular in states whose residents tend to be both extroverts ($r = 0.49$) and open to new experiences ($r = 0.50$). As one might expect, agreeable states avoid articles about religion ($r = -0.53$) and war&unrest ($r = -0.63$). The latter category is also avoided by conscientious states ($r = -0.49$). States with prevalence of neuroticism (emotional instability) tend to avoid article about music&theater ($r = 0.44$). Finally, states with low levels of neuroticism (*i.e.*, emotional stability) show interest in diverse topics ($r = -0.44$)

4.6 User Cold-Start Recommendations

As described in the beginning of this section, one way to partially overcome the user cold-start problem is to exploiting the contextual information of the users. Understanding the patterns of engagement and interests of the users in specific regions may allow us to recommend them suitable items. Given the significance of the time zone effect, we propose a *time-zone-aware recommender system*. Thus, when no past information about the user is available, the system suggests the most popular articles not in the whole USA but in the user's time zone. Next, test whether this way of recommending articles leads to better recommendations then the top-k recommender, which does not take the time zone into account.

Experimental Setup. We consider a random sample of articles and corresponding comments published during a whole month (>6.5M comments, >35K articles and >500K users). Since we have the publishing trough at 9am and the publishing peak at 3pm (Figure 12a), we train our time zone-aware model from 9am to 3pm, recommend articles after 3pm (*i.e.*, we recommend the top k articles most popular either globally or in the user's time zone), and test the accuracy of these two strategies. As the feedback of the recommendations is only implicit (*i.e.*, users' comments show their interest in the article, but the absence of comments does not necessarily mean that they were not interested in it), we measure the performance of each strategy by using the Ranking Accuracy defined in Section 3.4.2.

Results. Across all time zones, both ways of recommending articles (top-k vs. time zone aware) are better than random: their accuracy values are well above 0 (Figure 15). The time zone aware recommender (dark bars) performs (one and half times) better than the the top-k one (light bars), if we average across all time zones (last pair of bars in Figure 15). It performs comparatively well in the Eastern time zone and slightly worse in the Central. However, it outperforms the top-k recommender in two time zones (i.e., Mountain, Pacific). That is because, as we have seen in Figure 12c, the commenting cycles in those two time zones tend to be slightly more aligned with the publishing cycle of the news site, resulting in a more representative representation of the users' taste.

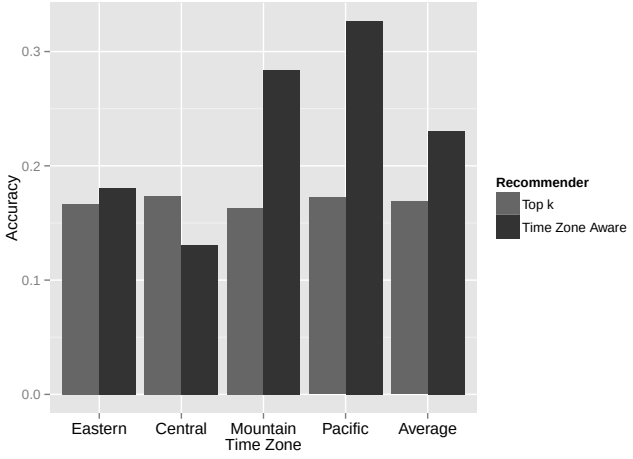


Figure 15: Comparison of the accuracies achieved by the two approaches across all time zones.

4.7 Discussion

Limitations. Our study suffers from two main limitations. First, we have used the users' IP addresses to localize them. So users on the move might be associated with different IP addresses and consequently with different locations. While it might happen to associate the same user to different cities, it has been extremely rare to associate them to different states, let alone to different time zones. Second, our study does not establish any casual relationship. To that end, one would need to apply our methodology to different snapshots over a long period of time.

Theoretical Implications. Our study has made three main contributions to the existing literature. First, it has confirmed that homophily (i.e., one's tendency to engage with like-minded others) holds not only on Twitter (e.g., as shown by Dhiraj Murthy [39]) but also on a news platform when commenting on articles. Second, at state level, we have found evidence for selective exposure [37]: one's tendency to favor information that reinforces pre-existing views, as socio-economic indicators are associated with topics of interest in expected ways. Third, one consequence of selective exposure in our context is that like-minded users comment on the same articles, creating fertile ground for group polarization [24]: as a by-product of commenting together (i.e., of engaging with each other), those like-minded users, the theory goes, might develop views that are more extreme than their initial inclinations. For the future, it might be beneficial to explore how geo-temporal patterns of news engagement impact a country's opinion formation.

Practical implications. Based on our findings, one could further explore:
Tailoring news to time zone. We have shown that recommending news that are popular in a zone is reasonable in a cold-start situation. Given that this situation is the norm for online news readers (who do not tend to log in), those reasonable suggestions translate into considerable improvement (e.g., substantial gain in monetization) for the entire site.

A time machine for volatile user content. Social-networking users tend to miss status updates coming from those contacts who are in different countries or even continents. Therefore, these services could adjust their users' timelines by offsetting the time difference between producers and consumers of updates. For instance, users living in New York are likely to miss updates that come from their contacts who live in Paris and that were posted in the morning (as of Paris time), as New Yorkers tend to sleep at that time. However, if those updates were to be shown after six hours, they would less likely missed. One may argue that a time line such as Facebook's "top stories" can show updates from distant users. However, such a real-time tool intrinsically enforces *exploitation* of current updates, while an adjusted time line is more likely to encourage *exploration* of experiences outside the bubble.

5 Conclusions and Future Work

This thesis has focused on the problem of *cold-start recommendations*, both for new items (item cold-start) and new users (user cold-start).

To overcome the item cold-start, we have proposed *Joint NMF*, a recommender system that combines content and collaborative information in a unified matrix factorization framework. We have also introduced a variation of it, *Joint NMF with Graph Regularization*, that takes into account the local geometric structure of the data and enforces smoothness of the solutions. We have presented two training algorithms, based on multiplicative update rule and alternating least squares, and we have analysed their convergence. Finally, we have experimentally shown that the two proposed methods outperform the existing content-based recommenders.

To address the user cold-start, we have proposed to exploit the user's location. To test the extent to which specific locations are linked to specific interests, we studied the geography of user engagement on the Yahoo! News website. We found that time zones play an important role on the user engagement: users who live in the same time zone tend to preferentially engage with each other on the same articles about the same topics. Based on these findings, we have proposed a *time-zone-aware recommender* that in the user cold-start suggests the most popular articles not in the whole USA, but in the users' time zone. We found that making time zone specific recommendations improves the recommendation accuracy by a factor of one and a half.

Our vision for future work includes the following:

- In the Yahoo! News data set we find that, both words and users are power law distributed. We would like to investigate how one may exploit this phenomenon to improve the accuracy of the proposed models. Recent studies [45] in matrix completion have shown that introducing a *weighted trace norm regularization* leads to significant improvement of the performance when entries of the matrix are sampled non-uniformly.

- In the item cold-start experiments, we have used only one view of the data, either the collaborative or the content information, to construct the regularization graph for the Joint NMF with Graph Regularization algorithm. We would like to investigate whether considering both views, for example by taking the maximum similarity, will result in an improvement of the accuracy of JNMF-GR.

- We would like to extent our study of the geography of user engagement in online news platforms by decoupling the geographic proximity effect. The effect of geographic sorting (*i.e.* the tendency of people to group according to their interests) implies the time-zone effect, but not the other way around. We would like to investigate to what extent the time-zone effect is due solely to the time zones, and not to the geographic sorting.

References

[1] Zeynep Akata, Christian Thurau, and Christian Bauckhage. Non-negative matrix factorization in multimodality data for segmentation and label prediction. In *Computer Vision Winter Workshop*, 2011.

[2] Cameron Anderson, Oliver John, et al. Who attains social status? Effects of personality and physical attractiveness in social groups. *Journal of Personality and Social Psychology*, 2001.

[3] Liviu Badea. Extracting gene expression profiles common to colon and pancreatic adenocarcinoma using simultaneous nonnegative matrix factorization. In *Pacific Symposium on Biocomputing*, 2008.

[4] Richard Bartels. Solution of the matrix equation $ax + xb = c$. *Communications of the ACM*, 1972.

[5] Mikhail Belkin. *Problems of learning on manifolds*. PhD thesis, The University of Chicago, 2003.

[6] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in Neural Information Processing Systems*, 2001.

[7] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research*, 2006.

[8] Bill Bishop. *The Big Sort: Why the Clustering of Like-Minded America Is Tearing Us Apart*. Houghton Mifflin, 2008.

[9] Vincent Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

[10] Rasmus Bro and Sijmen De Jong. A fast non-negativity-constrained least squares algorithm. *Journal of chemometrics*, 1997.

[11] Deng Cai, Xiaofei He, Xiaoyun Wu, and Jiawei Han. Non-negative matrix factorization on manifold. In *International Conference on Data Mining*, 2008.

[12] Fan Chung. *Spectral Graph Teory*. AMS, 1997.

[13] Andrzej Cichocki, Shun-ichi Amari, Rafal Zdunek, Raul Kompass, Gen Hori, and Zhaohui He. Extended smart algorithms for non-negative matrix factorization. In *Artificial Intelligence and Soft Computing*, 2006.

[14] International Press Telecommunications Council. Descriptive NewsCodes. Subject Code. `http://www.iptc.org/site/NewsCodes/View_NewsCodes/`.

[15] Andrew Cox, Paul Clough, and Jennifer Marlow. Flickr: a first look at user behaviour in the context of photography as serious leisure. *Information Research*, 2008.

[16] Lee Daniel and Seung Sebastian. Algorithms for non-negative matrix factorization. In *Neural Information Processing Systems*, 2000.

[17] Santo Fortunato. Community detection in graphs. *Physics Reports*, 2010.

[18] Alexander Tuzhilin Gediminas Adomavicius. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 2005.

[19] Lewis Goldberg et al. The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 2006.

[20] Scott Golder and Michael Macy. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science*, 2011.

[21] Itai Himelboim, Eric Gleave, and Marc Smith. Discussion catalysts in online political discussions: Content importers and conversation starters. *Journal of Computer-Mediated Communication*, 2009.

[22] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative Filtering for Implicit Feedback Datasets. In *2008 Eighth IEEE International Conference on Data Mining*, 2008.

[23] Bernardo Huberman, Daniel Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 2008.

[24] Daniel Isenberg. Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology*, 1986.

[25] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why We Twitter: An Analysis of a Microblogging Community. In *Advances in Web Mining and Web Usage Analysis*. 2009.

[26] Lauri Jensen-Campbell and William Graziano. Agreeableness as a moderator of interpersonal conflict. *Journal of Personality*, 2004.

[27] Julie Jones and Nathan Altadonna. We don't need no stinkin' badges: examining the social role of badges in the huffington post. *Computer Supported Cooperative Work and Social Computing*, 2012.

[28] Krishna Kamath, James Caverlee, Kyumin Lee, and Zhiyuan Cheng. Spatio-temporal dynamics of online memes: a study of geo-tagged tweets. In *World Wide Web Conference*, 2013.

[29] Benjamin Karney and Thomas Bradbury. The longitudinal course of marital quality and stability: a review of theory, methods and research. *Psychological Bulletin*, 1995.

[30] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a Social Network or a News Media? Categories and Subject Descriptors. *World Wide Web Conference*, 2010.

[31] Amy Langville, Carl Meyer, Russell Albright, James Cox, and David Duling. Algorithms, initializations, and convergence for the nonnegative matrix factorization. *preprint*, 2006.

[32] Charles Lawson and Richard Hanson. *Solving least squares problems*. Society for Industrial and Applied Mathematics, 1974.

[33] Chih-Jen Lin. On the Convergence of Multiplicative Update Algorithms for Nonnegative Matrix Factorization. *IEEE Transactions on Neural Networks*, 2007.

[34] Janne Lindqvist et al. I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application. In *Human factors in computing systems*, 2011.

[35] Jialu Liu, Chi Wang, Jing Gao, and Jiawei Han. Multi-view clustering via joint nonnegative matrix factorization. In *SIAM Data Mining Conference*, 2013.

[36] Prem Melville, Raymod Mooney, and Ramadass Nagarajan. Content-boosted collaborative filtering for improved recommendations. In *National Conference on Artificial Intelligence*, 2002.

[37] Solomon Messing and Sean Westwood. Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online. *Communication Research*, 2012.

[38] Alan Mislove and Sune Lehmann. Pulse of the Nation: U.S. Mood Throughout the Day inferred from Twitter. `http://www.ccs.neu.edu/home/amislove/twittermood/`.

[39] Dhiraj Murthy. *Twitter: Social Communication in the Twitter Age*. Polity Press, 2013.

[40] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. An Empirical Study of Geographic User Activity Patterns in Foursquare. *Artificial Intelligence*, 2010.

[41] Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 1994.

[42] Daniele Quercia. Don't Worry, Be Happy: The Geography of Happiness on Facebook. *Web Science*, 2013.

[43] Peter Rentfrow, Samuel Gosling, and Jeff Potter. A Theory of the Emergence, Persistence, and Expression of Geographic Variation in Psychological Characteristics. *Perspectives on Psychological Science*, 2008.

[44] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Uncertainty in Artificial Intelligence*, 2004.

[45] Ruslan Salakhutdinov and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Neural Information Processing Systems*, 2012.

[46] Andrew Schein, Alexandrin Popescul, Lyle Ungar, and David Pennock. Methods and metrics for cold-start recommendations. In *Special Interest Group on Information Retrieval*, 2002.

[47] David Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall, 4 edition, 2007.

[48] Ian Soboroff. Combining content and collaboration in text filtering. In *IJCAI Workshop on Machine Learning for Information Filtering*, 1999.

[49] Rhonda Swickert, Christina Rosentreter, James Hittner, and Jane Mushrush. Extraversion, social support processes, and stress. *Personality and Individual Differences*, 2002.

[50] Changhu Wang, Shuicheng Yan, Lei Zhang, and Hong-Jiang Zhang. Non-negative semi-supervised learning. In *International Conference on Artificial Intelligence and Statistics*, 2009.

[51] Chen Ye. Analysis of Participation in an Online Photo-Sharing Community : A Multidimensional Perspective. *Journal of the American Society for Information Science*, 2010.

[52] Dejin Zhao and Mary Beth Rosson. How and why people Twitter: the role that micro-blogging plays in informal communication at work. In *International conference on Supporting group work*, 2009.

[53] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. *Advances in Neural Information Processing Systems*, 2004.

[54] Xiaojin Zhu and John Lafferty. Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. In *International Conference on Machine learning*, 2005.

ANNEXES

A Proof of Theorem 1

The Joint NMF objective function J (Equation 3.2) is certainly bounded from below by zero. To prove Theorem 1, we need to show that J is non-increasing under the updating steps in Equations 3.7a, 3.7b, and 3.7c. The multiplicative update rules for $\mathbf{H_s}$ and $\mathbf{H_u}$ are exactly the same as in the original NMF, thus we can use the convergence proof of NMF to show that J is non-increasing under the update steps in Equations 3.7b and 3.7c (see [16] for details). Thus, we only need to prove that J is non-increasing under the update step for $\mathbf{W}$ (Equation 3.7a).

Since the objective function J can be decoupled to considering only one instance at time, *i.e.*, one row of $\mathbf{X_s}$, $\mathbf{X_u}$ and $\mathbf{W}$, we can write J as:

$$\min : J = \frac{1}{2}(\alpha||x_s^{\mathrm{T}} - w^{\mathrm{T}}\mathbf{H_s}||_{\mathrm{F}}^2 + (1-\alpha)||x_u^{\mathrm{T}} - w^{\mathrm{T}}\mathbf{H_u}||_{\mathrm{F}}^2 + \lambda_W||w^{\mathrm{T}}||_2^2 + \lambda_{H_s}||\mathbf{H_s}||_{\mathrm{F}}^2 + \lambda_{H_u}||\mathbf{H_u}||_{\mathrm{F}}^2),$$

minimizing it with respect to each of the rows of $\mathbf{W}$ separately.

We consider a current approximation $\hat{w}^{\mathrm{T}}$ of the solution and we formulate the following problem:

$$\min : \hat{J}(w^{\mathrm{T}}) = J(w^{\mathrm{T}}) + \frac{1}{2}(w^{\mathrm{T}} - \hat{w}^{\mathrm{T}})^{\mathrm{T}} S (w^{\mathrm{T}} - \hat{w}^{\mathrm{T}}),$$

where $S = Diag(x) - (\alpha\mathbf{H_s}\mathbf{H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{H_u}\mathbf{H_u}^{\mathrm{T}} + \lambda_W)$, with $x = \frac{\hat{w}^{\mathrm{T}}(\alpha\mathbf{H_s}\mathbf{H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{H_u}\mathbf{H_u}^{\mathrm{T}} + \lambda_W)}{\hat{w}^{\mathrm{T}}}$.

Since S is positive semi-definite matrix [16], we have that $\hat{J}(w^{\mathrm{T}}) \geq J(w^{\mathrm{T}})$ for all w^{T} and specifically $\hat{J}(\hat{w}^{\mathrm{T}}) = J(\hat{w}^{\mathrm{T}})$.

Furthermore, the function is also convex. We set the derivative of $\hat{J}(\hat{w}^{\mathrm{T}})$ to zero, *i.e.*

$$\frac{\partial \hat{J}}{\partial w^{\mathrm{T}}} = \alpha w^{\mathrm{T}}\mathbf{H_s}\mathbf{H_s}^{\mathrm{T}} - \alpha x_s^{\mathrm{T}}\mathbf{H_s}^{\mathrm{T}} + (1-\alpha)w^{\mathrm{T}}\mathbf{H_u}\mathbf{H_u}^{\mathrm{T}} - (1-\alpha)x_u^{\mathrm{T}}\mathbf{H_u}^{\mathrm{T}} + \lambda_W\hat{w}^{\mathrm{T}} + (w^{\mathrm{T}} - \hat{w}^{\mathrm{T}})S = 0$$

in order to obtain the minimizer $w^{\mathrm{T}*}$:

$$w^{\mathrm{T}*}(\alpha\mathbf{H_s}\mathbf{H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{H_u}\mathbf{H_u}^{\mathrm{T}} + \lambda_W + S) = \alpha x_s^{\mathrm{T}}\mathbf{H_s}^{\mathrm{T}} + (1-\alpha)x_u^{\mathrm{T}}\mathbf{H_u}^{\mathrm{T}} + \hat{w}^{\mathrm{T}}S$$

notice that $\hat{w}^{\mathrm{T}}S = 0$ (by the construction of S).

Expanding S and cancelling terms, we obtain:

$$w^{\mathrm{T}*} \cdot diag^{-1}(\hat{w}^{\mathrm{T}}) \cdot diag(\hat{w}^{\mathrm{T}}(\alpha\mathbf{H_s}\mathbf{H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{H_u}\mathbf{H_u}^{\mathrm{T}} + \lambda_W)) = \alpha x_s^{\mathrm{T}}\mathbf{H_s}^{\mathrm{T}} + (1-\alpha)x_u^{\mathrm{T}}\mathbf{H_u}^{\mathrm{T}}.$$

Notice that multiplying vector by a diagonal matrix formed by another vector, corresponds to performing an element-wise product between the two vectors. Thus, we obtain:

$$w^{\mathrm{T}*} = \hat{w}^{\mathrm{T}} \odot (\alpha x_s^{\mathrm{T}}\mathbf{H_s}^{\mathrm{T}} + (1-\alpha)x_u^{\mathrm{T}}\mathbf{H_u}^{\mathrm{T}}) \oslash (\alpha\hat{w}^{\mathrm{T}}\mathbf{H_s}\mathbf{H_s}^{\mathrm{T}} + (1-\alpha)\hat{w}^{\mathrm{T}}\mathbf{H_u}\mathbf{H_u}^{\mathrm{T}} + \lambda_W\hat{w}^{\mathrm{T}}).$$

Leading to the multiplicative update rule of Equation 3.7a.

Since, $w^{\mathrm{T}*}$ is the global minimizer of $\hat{J}(w^{\mathrm{T}})$, we have $\hat{J}(w^{\mathrm{T}*}) \leq \hat{J}(\hat{w}^{\mathrm{T}})$. Moreover, $\hat{J}(w^{\mathrm{T}})$ is constructed to satisfy $\hat{J}(w^{\mathrm{T}}) \geq J(w^{\mathrm{T}})$ for all w^{T}. This implies that $J(w^{\mathrm{T}*}) \leq \hat{J}(w^{\mathrm{T}*}) \leq \hat{J}(\hat{w}^{\mathrm{T}}) = J(\hat{w}^{\mathrm{T}})$ *i.e.* we have a decrease of the objective function. □

B Proof of Theorem 2

Similar to Theorem 1, J (Equation 3.8) is bounded from below by zero and the update rules for $\mathbf{H_s}$ and $\mathbf{H_u}$ are the same as in the original NMF formulation. Thus, we only need to prove that J is non-increasing under the update step for $\mathbf{W}$ (Equation 3.10). We will follow a procedure based on auxiliary functions, similar to the one described in [11].

Definition. $G(w, w')$ is an auxiliary function for $J(w)$ if the conditions:

$$G(w, w') \geq J(w), \quad G(w, w) = J(w)$$

are satisfied. The auxiliary function is very useful because of the following lemma.

Lemma. If G is an auxiliary function of J, then J is non-increasing under the update:

$$w^{(t+1)} = \arg\min_{w} G(w, w^{(t)}) \tag{B.1}$$

Proof.

$$J(w^{(t+1)}) \leq G(w^{(t+1)}, w^{(t)}) \leq G(w^{(t)}, w^{(t)}) = J(w^{(t)}).$$

We rewrite the objective function of JNMF-GR in Equation 3.8 as follows:

$$\min : J = \frac{1}{2}(\alpha \sum_{i=1}^{N}\sum_{j=1}^{M}(x_{ij}^s - \sum_{k=1}^{K} w_{ik}h_{kj}^s)^2 + (1-\alpha)\sum_{i=1}^{N}\sum_{j=1}^{F}(x_{ij}^u - \sum_{k=1}^{K} w_{ik}h_{kj}^u)^2 +$$
$$+\beta \sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{l=1}^{N} w_{jk}[\mathbf{L}]_{jl}w_{lk} + \lambda_W \sum_{i=1}^{N}\sum_{j=1}^{K} w_{ij}^2 + \lambda_{H_s}\sum_{i=1}^{K}\sum_{j=1}^{M}(h_{ij}^s)^2 + \lambda_{H_u}\sum_{i=1}^{K}\sum_{j=1}^{F}(h_{ij}^u)^2.$$

Considering any element w_{ab} of $\mathbf{W}$, we use J_{ab} to denote the part of J which is only relevant to w_{ab}. It is easy to check that:

$$J'_{ab} = \left[\frac{\partial J}{\partial \mathbf{W}}\right]_{ab} = [\alpha \mathbf{W}\mathbf{H_s}\mathbf{H_s}^{\mathrm{T}} - \alpha \mathbf{X_s}\mathbf{H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{W}\mathbf{H_u}\mathbf{H_u}^{\mathrm{T}} - (1-\alpha)\mathbf{X_u}\mathbf{H_u}^{\mathrm{T}} + \beta \mathbf{L}\mathbf{W} + \lambda_W \mathbf{W}]_{ab}.$$
$$J''_{ab} = \alpha[\mathbf{H_s}\mathbf{H_s}^{\mathrm{T}}]_{bb} + (1-\alpha)[\mathbf{H_u}\mathbf{H_u}^{\mathrm{T}}]_{bb} + \beta[\mathbf{L}]_{aa} + \lambda_W$$

Since our update is essentially element-wise, it is sufficient to show that each J_{ab} is non-increasing under the update step of Equation 3.10.

We define:

$$G(w, w_{ab}^{(t)}) = J_{ab}(w_{ab}^{(t)}) + J'_{ab}(w_{ab}^{(t)})(w - w_{ab}^{(t)}) +$$
$$+\frac{\alpha[\mathbf{W}\mathbf{H_s}\mathbf{H_s}^{\mathrm{T}}]_{ab} + (1-\alpha)[\mathbf{W}\mathbf{H_u}\mathbf{H_u}^{\mathrm{T}}]_{ab} + \beta[\mathbf{D}\mathbf{W}]_{ab} + \lambda_W[\mathbf{W}]_{ab}}{w_{ab}^{(t)}}(w - w_{ab}^{(t)})^2$$

is an auxiliary function for J_{ab}, the part of J which is only relevant to w_{ab}.

Since $G(w, w) = J_{ab}(w)$ is obvious, we need only show that $G(w, w_{ab}^{(t)}) \geq J_{ab}(w)$. To do this, we compare the Tylor series expansion of $J_{ab}(w)$:

$$J_{ab}(w) = J_{ab}(w_{ab}^{(t)}) + J'_{ab}(w_{ab}^{(t)})(w - w_{ab}^{(t)}) + (\alpha[\mathbf{H_s}\mathbf{H_s}^{\mathrm{T}}]_{bb} + (1-\alpha)[\mathbf{H_u}\mathbf{H_u}^{\mathrm{T}}]_{bb} + \beta[\mathbf{L}]_{aa} + \lambda_W)(w - w_{ab}^{(t)})^2,$$

with $G(w, w_{ab}^{(t)})$ to find that $G(w, w_{ab}^{(t)}) \geq J_{ab}(w)$ is equivalent to:

$$\frac{\alpha[\mathbf{W}\mathbf{H_s}\mathbf{H_s}^{\mathrm{T}}]_{ab} + (1-\alpha)[\mathbf{W}\mathbf{H_u}\mathbf{H_u}^{\mathrm{T}}]_{ab} + \beta[\mathbf{D}\mathbf{W}]_{ab} + \lambda_W[\mathbf{W}]_{ab}}{w_{ab}^{(t)}} \geq$$
$$\geq \alpha[\mathbf{H_s}\mathbf{H_s}^{\mathrm{T}}]_{bb} + (1-\alpha)[\mathbf{H_u}\mathbf{H_u}^{\mathrm{T}}]_{bb} + \beta[\mathbf{L}]_{aa} + \lambda_W.$$

We have:

$$\alpha[\mathbf{W}\mathbf{H_s}\mathbf{H_s}^{\mathrm{T}}]_{ab} = \alpha \sum_{l=1}^{k} w_{al}^{(t)} [\mathbf{H_s}\mathbf{H_s}^{\mathrm{T}}]_{lb} \geq \alpha w_{ab}^{(t)} [\mathbf{H_s}\mathbf{H_s}^{\mathrm{T}}]_{bb},$$

$$(1-\alpha)[\mathbf{W}\mathbf{H_u}\mathbf{H_u}^{\mathrm{T}}]_{ab} = (1-\alpha) \sum_{l=1}^{k} w_{al}^{(t)} [\mathbf{H_u}\mathbf{H_u}^{\mathrm{T}}]_{lb} \geq (1-\alpha) w_{ab}^{(t)} [\mathbf{H_u}\mathbf{H_u}^{\mathrm{T}}]_{bb},$$

$$\beta[\mathbf{D}\mathbf{W}]_{ab} = \beta \sum_{j=1}^{N} [\mathbf{D}]_{aj} w_{jb}^{(t)} \geq \beta [\mathbf{D}]_{aa} w_{ab}^{(t)} \geq \beta [\mathbf{D}-\mathbf{A}]_{aa} w_{ab}^{(t)} = \beta [\mathbf{L}]_{aa}.$$

Thus, $G(w, w_{ab}^{(t)}) \geq J_{ab}(w)$ holds.

Replacing $G(w, w_{ab}^{(t)})$ in B.1 results in the update rule:

$$\begin{aligned} w_{ab}^{(t+1)} &= w_{ab}^{(t)} - w_{ab}^{(t)} \frac{J'_{ab}(w_{ab})}{\alpha[\mathbf{W}\mathbf{H_s}\mathbf{H_s}^{\mathrm{T}}]_{ab} + (1-\alpha)[\mathbf{W}\mathbf{H_u}\mathbf{H_u}^{\mathrm{T}}]_{ab} + \beta[\mathbf{D}\mathbf{W}]_{ab} + \lambda_W [\mathbf{W}]_{ab}} \\ &= w_{ab}^{(t)} \frac{[\alpha \mathbf{X_s}\mathbf{H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{X_u}\mathbf{H_u}^{\mathrm{T}} + \beta \mathbf{A}\mathbf{W}]_{ab}}{[\alpha \mathbf{W}\mathbf{H_s}\mathbf{H_s}^{\mathrm{T}} + (1-\alpha)\mathbf{W}\mathbf{H_u}\mathbf{H_u}^{\mathrm{T}} + \beta \mathbf{D}\mathbf{W} + \lambda_W \mathbf{W}]_{ab}}. \end{aligned}$$

Since G is an auxiliary function, J_{ab} is non-increasing under this update rule. □