

**User Profiling and Privacy Preserving from
Multiple Social Networks**

Xuemeng Song

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

SCHOOL OF COMPUTING

NATIONAL UNIVERSITY OF SINGAPORE

2016

©2016

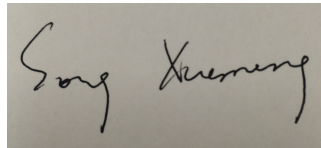
Xuemeng Song

All Rights Reserved

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

A rectangular box containing a handwritten signature in black ink. The signature appears to read 'Song Xueming'.

.....
Xueming Song

29 September 2016

Publications

- ***Xuemeng Song***, Zhao-Yan Ming, Liqiang Nie, Yi-Liang Zhao, Tat-Seng Chua. Volunteerism Tendency Prediction via Harvesting Multiple Social Networks. ACM Transactions on Information Systems, 2016. Full journal paper.
- Liqiang Nie, ***Xuemeng Song***, Tat-Seng Chua. Learning from Multiple Social Networks. Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers, 2016. Book.
- ***Xuemeng Song***, Liqiang Nie, Luming Zhang, Maofu Liu, Tat-Seng Chua. Interest Inference via Structure-Constrained Multi-Source Multi-Task Learning. In Proceedings of the International Joint Conference on Artificial Intelligence, 2015. Full conference paper.
- ***Xuemeng Song***, Liqiang Nie, Luming Zhang, Mohammad Akbari, Tat Seng Chua. Multiple Social Network Learning and Its Application in Volunteerism Tendency Prediction. In Proceedings of the International ACM SIGIR Conference, 2015. Full conference paper.

Contents

List of Tables	xiii
List of Figures	xv
Chapter 1 Introduction	1
1.1 Social Media	1
1.2 User Profiling across Multiple Social Networks	4
1.3 Challenges	6
1.4 Assumptions and Limitations	8
1.5 Strategies	8
1.5.1 Multiple Social Account Alignment	9
1.5.2 Missing Data Completion	9
1.5.3 Multiple Social Network Learning	10
1.6 Contributions	10
1.7 Outline of the Thesis	12
Chapter 2 Literature Review	15
2.1 User Profiling	15
2.2 Multi-view Learning	23
2.3 Multi-task Learning	27
2.4 Multi-view Multi-task Learning	28

2.5	Summary	29
-----	-------------------	----

Chapter 3 User Profiling via Multi-Source Mono-task Learning: Vol-

	unteerism Tendency Prediction	31
3.1	Introduction	31
3.2	Related Work	35
3.2.1	Volunteerism	35
3.3	Multi-source Mono-task Learning	36
3.3.1	Notation	36
3.3.2	Problem Formulations	37
3.3.3	Optimization	38
3.3.3.1	Computing α with \mathbf{w}_s fixed	39
3.3.3.2	Computing \mathbf{w}_s with α fixed	40
3.4	Missing Data Completion	42
3.4.1	Optimization	44
3.5	Application: Volunteerism Tendency Prediction	45
3.5.1	Necessity of Multiple Social Networks	46
3.5.2	Social Accounts Alignment	46
3.5.3	Ground Truth Construction	47
3.5.4	Features	49
3.5.4.1	Demographic Characteristics	49
3.5.4.2	Linguistic Features	50
3.5.4.3	Behavior-based Features	53
3.6	Experiments	54
3.6.1	Data Preprocessing	55
3.6.2	On Model Comparison	55
3.6.3	On Data Completion Comparison	57
3.6.4	On Feature Comparison	58

3.6.5	On Source Comparison	60
3.6.6	Size Varying of Positive Samples	61
3.6.7	Complexity Discussion	62
3.7	Summary	63

Chapter 4 User Profiling via Multi-source Multi-task Learning: User

	Interest Inference	65
4.1	Introduction	65
4.2	Related Work	68
4.2.1	User Interest Inference	68
4.3	User Interest Inference	69
4.3.1	Notation	70
4.3.2	Problem Formulations	70
4.3.3	Optimization	73
4.3.4	Construction of Interest Tree Structure	75
4.3.5	Complexity Discussion	77
4.4	Experiments	77
4.4.1	Dataset Construction	78
4.4.2	Feature Extraction	79
4.4.3	On Tree Construction	81
4.4.4	On Evaluation Metrics	81
4.4.5	On Model Comparison	82
4.4.6	On Source Comparison	84
4.5	Summary	84

Chapter 5 A Personal Privacy Detection Framework **87**

5.1	Introduction	87
5.2	Related Work	90

5.2.1	Privacy	90
5.3	Data and Description	91
5.3.1	Taxonomy Induction	92
5.3.2	Data Collection	92
5.3.3	Ground Truth Construction	93
5.3.4	Features	94
5.3.4.1	LIWC	94
5.3.4.2	Privacy Dictionary	94
5.3.4.3	Sentiment Analysis	95
5.3.4.4	Sentence2Vector	95
5.3.4.5	Meta-features	95
5.4	Prediction	96
5.4.1	Notation	96
5.4.2	Model Formulations	96
5.4.3	Optimization	98
5.4.3.1	Computing \mathbf{L} with \mathbf{S} fixed	99
5.4.3.2	Computing \mathbf{S} with \mathbf{L} fixed	99
5.5	Experiments	102
5.5.1	Data Preprocessing	102
5.5.2	Experimental Setting	102
5.5.3	Evaluation of Description	103
5.5.4	Evaluation of Prediction	107
5.5.5	Case Study	109
5.5.5.1	Example Study	109
5.5.5.2	Failure Study	111
5.6	Summary	113

Chapter 6	Conclusions and Future Research	115
6.1	Conclusions	115
6.2	Future Work	117

Abstract

User profiling, which aims to infer users' unobservable information based on observable information such as individual's behavior or utterances, is the basis for many applications, such as personalized recommendation and expert finding. Traditional user profiling conducted with traditional media, such as document records, is often hindered by limited data sources. In recent years, the proliferation of social media has opened new opportunities for user profiling. Moreover, as different social networks provide different services, an increasing number of people are involved in multiple social networks, in which different aspects of users can be revealed by different social networks. Therefore, to comprehensively learn users' profiles, it is time to shift from a single social network to multiple social networks. Therefore, this thesis aims to investigate user profiling across multiple social networks. In particular, it covers studies in general scenarios of user profiling, in which a single task and multiple tasks are involved, respectively. Meanwhile, as user profiling would potentially put users at high privacy risks, this thesis also proposes a framework for privacy preservation.

In general, multi-social network learning involves two main steps: 1) social account mapping, and 2) multi-source learning. The first step aims to identify the same users across different social networks, while the second step targets at effectively aggregating multiple sources. This thesis will not address the social account mapping problem, and concentrate instead on the second step.

This thesis first proposes a novel scheme for multi-source mono-task learning to infer users' attributes, such as volunteerism tendency, which involves a single task. In particular, this proposed scheme is able to tackle the missing data problem, which is due to the fact that users may not be active enough in certain social networks. In addition, this scheme is capable of modeling both the source confidence and source consistency simultaneously. This thesis then proposes a multi-source

multi-task learning scheme to infer users attributes, such as interest, where multiple related tasks can be involved. The proposed scheme jointly regularizes two important aspects: source consistency and task relatedness. Finally, this thesis also develops a framework for privacy preserving to reduce users' privacy risks on social media. In particular, it proposes a taxonomy to comprehensively characterize users' personal aspects. With the guidance of such a taxonomy, we correspondingly propose a multi-task learning scheme to identify the potential privacy leakage.

Extensive experiments have been conducted on the real-world datasets. The experimental results enable us to draw the following key findings. First, utilizing multiple social networks does improve the performance of user profiling problems. Second, it is important to take source consistency and source confidence into consideration when dealing with multiple social networks. Third, in the context of user profiling with multiple tasks, taking task relatedness into account is plausible. Fourth, LIWC and Sentence2Vector features are the most discriminating features regarding privacy leakage detection. Last, the privacy leakage via user-generated content holds certain temporal patterns and distinct behavior patterns.

List of Tables

1.1	Various forms of social media.	2
2.1	Category summarization of LIWC directories.	18
2.2	Summarization of literature findings regarding the correlations between personality traits and LIWC features.	20
2.3	Summarization of related works on user profiling.	21
2.4	The summarization of related works on multi-view learning.	25
3.1	Statistics of our dataset.	49
3.2	Performance of different models(%).	56
3.3	Performance of different models over different data completion strategies.	58
3.4	Performance of different features(%).	59
3.5	Hot topics discussed by volunteers. Followee and retweeting: contextual topics; Self: user topics.	59
3.6	Comparison of the value of LIWC features among volunteers and non-volunteers. (%)	60
3.7	Performance of different social network combinations (%). Facebook* and LinkedIn* both refer to the complete data, whose missing data is pre-inferred. F1: F1-measure.	61

4.1	Top interest-pairs based on the tree constructed by the external source and internal source, respectively.	82
4.2	Performance comparison among various models.	83
4.3	Contribution of individual social network and their various combinations.	85
5.1	Performance comparison of our LG-MTL model trained with different feature configurations (%).	103
5.2	Ten representative word categories in LIWC, that can capture the personal aspects comprehensively.	103
5.3	Top categories regarding the percentage of tweets that containing images.	105
5.4	Top categories regarding the percentage of tweets that containing user mentions.	106
5.5	Sentiment Ranking.	106
5.6	Performance comparison between our LG-MTL model and the baselines in S@K and P@K (%).	107
5.7	Examples of some categories.	108
5.8	Keywords or phrases for each category.	110
5.9	Poorly classified tweets.	112

List of Figures

1.1	Illustration of the three dimensions of an identify on social media websites.	3
1.2	Optional caption for list of figures	9
1.3	Illustration of users' presence on multiple social networks.	10
2.1	Aggregation and enrichment of profile data with Mypes.	22
3.1	Illustration of our proposed scheme. We first collect and align users' distributed data from multiple social networks. We then jointly infer the block-wise missing data based on the available data. We finally apply MSNL to the complete data. SN_i , x_j , and y_l refer to the i -th social network, j -th user sample, and the l -th corresponding label, respectively.	33
3.2	Illustration of the incomplete data from three sources. $\mathbf{X}_s^{\mathcal{C}_i}$ denotes the samples generated from social network s that are only available in the social network combination of \mathcal{C}_i	42
3.3	Optional caption for list of figures	45
3.4	Statistics of the incomplete data. Tw: Users with Twitter data only; Tw+Fb: Users with Twitter and Facebook data only; Tw+In: Users with Twitter and LinkedIn data only; Tw+Fb+In: Users without missing data.	50

3.5	Perplexity values varying over the number of topics in Twitter. . . .	51
3.6	Optional caption for list of figures	52
3.7	Failure sample distribution.	57
3.8	F1-measure at different fraction of volunteer samples.	62
4.1	Illustration of inter-interests relatedness in a tree structure.	72
4.2	Distribution of user frequency distribution with respect to the number of interests over our dataset.	78
4.3	Optional caption for list of figures	80
5.1	Illustration of the proposed scheme for privacy leakage detection. In the first component, we build a comprehensive taxonomy of the personal aspects, collect a benchmark dataset and extract a rich set of features to describe the UGC. The second component presents a taxonomy-constrained model to detect whether the given post leaks certain personal aspects.	89
5.2	Illustration of our pre-defined taxonomy.	93
5.3	Optional caption for list of figures	105

Chapter 1

Introduction

In this chapter, we introduce social media as well as its characteristics, highlight the motivation for user profiling across multiple social networks, and describe the challenges we need to address.

1.1 Social Media

With the booming of Web 2.0 technologies, the last couple of years has witnessed the unprecedented prosperity of social media websites. Social media has evolved from a service for simple broadcasting (e.g. blogs) to a rich multimedia service for maintaining social connections. Table 1.1 lists assorted forms of social media websites, including social networking websites, social tagging websites, wikis, media sharing websites, social news websites, blogs, microblogging platforms, location-based social networks (LBSNs), event-based social networks (EBSNs) and forums. Taking advantage of Web 2.0 technologies, they all share a common feature that distinguishes them from the conventional web and traditional media: the “consumers” of content or information online are also the “producers”. Essentially, everybody in social media can be an information outlet, resulting in a huge amount of user-

Categories	Social media
Social Networking	Facebook, MySpace, LinkedIn, Orkut
Social Tagging	Del.icio.us, Stumpupon
Wikis	Wikipedia, Scholarpedia, Ganfyd, AskDrWiki
Media Sharing	Pinterest, Vine, Instagram, Flickr, YouTube, Ustream, Scribd
Social News	Digg, Reddit
Microblogging	Twitter, Wordpress, Blogspot, LiveJournal, BlogCatalog
LBSNs	Foursquare, Gowalla, Brightkite, Whrrl
EBSNs	Meetup, Plancast
Forums	Yahoo! Answers, StackOverflow, Epinions

Table 1.1: Various forms of social media.

generated content (UGC). In other words, social media websites act as services for content sharing and social networking, where people can build social connections with others and freely contribute and share contents [1]. On joining a social media website, users usually create an identity with three major dimensions: profile, content, and network, as illustrated in Figure 1.1. The details for each dimension are illustrated as follows.

- **Profile.** It is composed of a set of attributes that describe the identity's persona, which usually consists of name, age, gender, location and so on.
- **Content.** It is composed of a set of posts created or shared by the user.
- **Network.** It is composed of a set of user connections, which depicts the network a user creates to connect with other users.

Due to the tremendous popularity of social media, surfing in social media has become a daily routine for many users. According to the GWI¹ report of 2015, the average time spent by a user on social platforms is 1.72 hours, which contributes

¹<http://insight.globalwebindex.net/social>

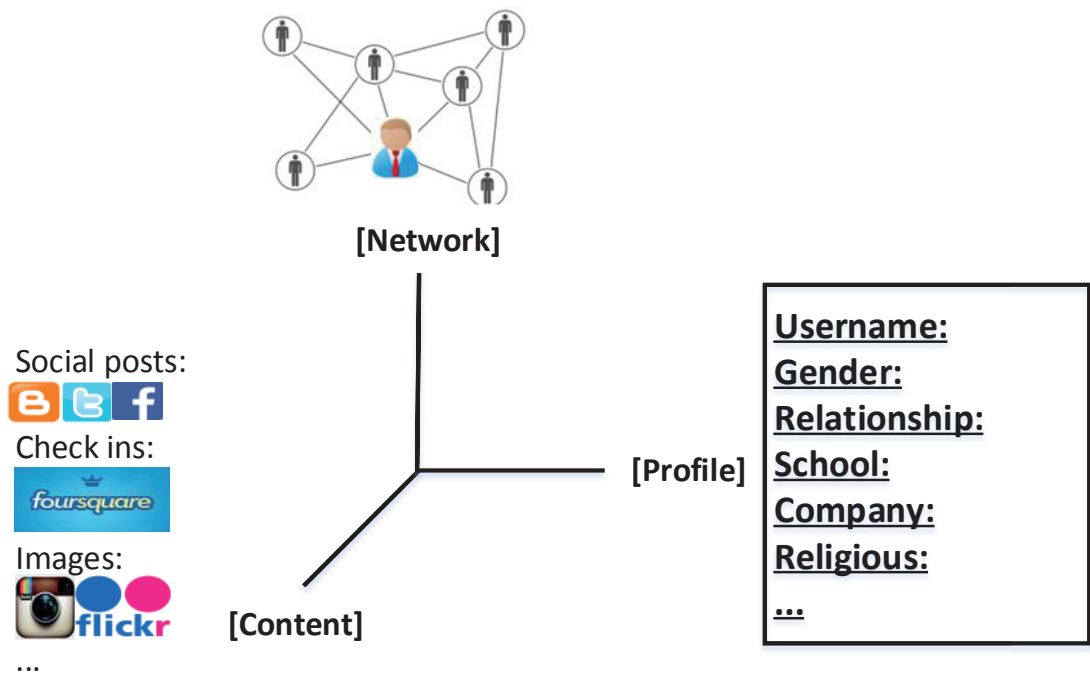


Figure 1.1: Illustration of the three dimensions of an identify on social media websites.

28% of one's all online activity. Notably, Facebook², Twitter³, and LinkedIn⁴ hold the top three places stably in the ranking list of the most popular social networking websites, respectively. Facebook and Twitter are the most successful social networks with 1.28 billion and 0.26 billion monthly active users by 2014, respectively. Due to the different intrinsic mechanisms (e.g. social connection structure) on Facebook and Twitter, users may prefer to use Facebook to keep social connecting while using Twitter to exchange information. Unlike the first two social networks, LinkedIn with more than 250 million members, has become a brilliant star in the eyes of professional users. LinkedIn offers users a platform to construct an abbreviated resume, which usually include sections of summary, education, experiences, skill, and interest.

²<http://facebook.com/>.

³<http://twitter.com/>.

⁴<http://linkedin.com/>.

1.2 User Profiling across Multiple Social Networks

A user profile can be treated as a description of user's personal data, such as age, location, personality traits and interests. In the era of the Internet, an accurate and comprehensive user profile usually facilitate others to gain a good understanding of this user, and hence enables many promising services, such as personalized recommendation [143], target advertisement [17], expert finding [11] and planning social service of governments [36]. In fact, user profiles can be either explicitly given by users or implicitly inferred from the data. However, users always feel reluctant to explicitly provide their personal attributes, which makes the intelligent inference highly desired. User profiling is such an intelligent technique that aims to infer users' unobservable information based on observable information such as individual's behavior or utterances [148].

Traditional user profiling is conducted with traditional media, such as document records. Garera et. al [48] presented a novel partner-sensitive model to predict individuals' biographic attributes over two corpora of conversation records. In addition, Bocklet et. al [21] investigated the potential of age determination of children in preschool and primary school from the speech records by conventional machine learning models. Although promising results have been demonstrated by these efforts, the limited data sources hinder, to a large extent, the impact and extensibility of these studies.

With the proliferation of social media, everybody in social media, essentially, can be an information outlet, resulting in a huge amount of UGC. Consequently, such rich social media opens up new opportunities for user profiling, and has attracted many research efforts [90, 96, 101, 103]. The existing efforts successfully demonstrate that users' attributes can be inferred from their generated contents on social media. Nevertheless, most existing works failed to learn multiple social networks together to profile users more comprehensively.

In fact, as different social networks provide different services, more and more people are involved in multiple social networks [52, 76]. It is reported by the GWI that Internet users have an average of 5.54 social media accounts with 2.82 being used actively. In general, different aspects of users can be revealed by different social networks due to their different functional emphasis. For instance, people frequently post their personal opinions in Twitter, share their casual activities in Facebook, and reveal their career experiences in LinkedIn. Meanwhile, these aspects are usually complementary to each other and essentially characterize the same users from different perspectives. Therefore, the appropriate aggregation of users' footprints on multiple social networks provides us a unique opportunity to understand the users more comprehensively. However, such significant gap thus far remains largely untapped.

On the other hand, risk always co-exists with opportunities. The proliferation of social media not only provides unique opportunities for user profiling, but also puts users at high privacy risk. As validated by many previous studies [90, 103], a lot of users' personal aspects can be extracted from the UGC. It is reported that 66% of users' micro-posts are about themselves [58]. On the other hand, people are usually connected with heterogeneous circles on social networks, such as family members, casual friends and even strangers. Users are thus easier than ever before to leak their personal information to those who are not appropriate to see it. Take a real story as an example. A video podcaster's home was broken into and several video equipments were stolen during his travel. It is ultimately found out that the break-in was caused by his detailed tweets regarding his leave [58]. Consequently, it is highly expected to investigate privacy preserving techniques to avoid users' privacy leakage from the UGC. However, most of the existing work focused on structured data, such as users' privacy settings on social media, but failed to pay attention to unstructured data, namely, UGC. To date, this significant research gap

has still not been bridged well.

In this work, we aim to investigate user profiling from multiple social networks as well as study how to protect users from the privacy leakage on social media. In a sense, user profiling across multiple social networks in nature can be treated as multi-source learning. Furthermore, according to the nature of users' attributes, user profiling across multiple social networks can be both framed by a multi-source mono-task learning scheme or a multi-source multi-task learning scheme. For example, users' gender or volunteerism tendency can be learned by the multi-source mono-task learning scheme, where only one binary classification (task) is involved. When it comes to learning users' interests, which usually involves a set of binary classifications (tasks), it should be appropriate to frame it in the multi-source multi-task learning scheme. This is due to the fact that multi-task learning works by jointly solving a task together with other related tasks simultaneously using a shared representation, which often leads to a better model for the research problem [25]. Consequently, we first proposed a multi-source mono-task learning scheme for user profiling on multiple social networks, and applied it to a practical scenario of volunteerism tendency prediction. Sequentially, we moved from the mono-task scenario to the multi-task context, proposing a multi-source multi-task learning scheme, and applied it to the application of user interest inference. Based on the insights obtained from these two works, we further proposed a framework for privacy leakage detection.

1.3 Challenges

People's public presence provides abundant free data for us to approach the problem of user attribute inference in new ways. In particular, aggregating and exploring users' footprints casually left on all of these OSNs is a promising approach to generate more comprehensive summaries of users' profiles [91]. Meanwhile, the

boom of social media services also introduces new challenges for the problem of user attribute inference.

The first challenge lies in how to collect users' distributed social contents on multiple social networks. Essentially, the main problem we need to solve is the "social account alignment", which aims to identify the same user across different social networks by linking one's multiple social accounts [3, 76]. As a consequence, how to track users' distributed data on different social networks is the first challenge need to be addressed.

The second challenge is the missing data problem. Although some users have social accounts on multiple social networks, generally they are active on only a few of them. One simple approach to address this challenge is to discard all incomplete subjects. It is apparent that this method will dramatically reduce the training size, resulting in overfitting in the model learning stage. Therefore, accurately completing missing data by jointly utilizing multiple sources is a necessity to enhance the learning performance.

Another challenge we face is how to effectively integrate users' heterogeneous distributed data from multiple social networks. The heterogeneous data structure makes user profiling across multiple social networks more challenging. One naïve approach is to concatenate the feature spaces generated from different sources into a unified feature space, and employ the traditional machine learning models to tackle the problems. However, this method simply treats the confidence of all data sources equally and may lead to the curse of dimensionality. Moreover, it ignores two important facts: 1) different aspects of users can be revealed in different social networks and are thus distributed in different feature spaces; and 2) all these aspects tend to characterize the same users. In particular, data from multi-sources describes the same user and thus the results predicted by different sources should be similar. Therefore, it is expected to take the source confidence and source consistency into

consideration to achieve better performance regarding data fusion.

1.4 Assumptions and Limitations

In this work, we focus more on the effective multi-source learning rather than the alignment of the social accounts of users on multiple social networks. The problem of social account alignment should be treated as another interesting research line—entity matching, where several content-based and social connection-based methods have been proposed [77, 116, 137, 141]. Since this is not the focus of this thesis, we just assume that we have a set of users, whose multiple social accounts are available. On the other hand, to ensure the quality of user profiling, we only consider those relatively active users in social networks. In other word, they are active in at least one social network.

1.5 Strategies

To address these problems, we present a scheme for multiple social network learning (MSNL), which co-regulates the source confidence and source consistency. The proposed scheme comprises of three components. Given a set of users, we first crawl their historical contents and all social connections. The first component extracts the multi-faceted information cues to describe a given user, including demographic information, practical behaviors, historical posts, and profiles of social connections. To deal with the block-wise missing data, the second component attempts to infer the block-wise missing data by learning a latent space shared by different social networks, achieving a complete input to the next component. We finally use the last component to conduct MSNL on the completed data. Particularly, we model the confidence of different data sources and the consistency among them by unifying two regularization terms into our model.

1.5.1 Multiple Social Account Alignment

To collect users' distributed social contents on multiple social networks, we need to first tackle the problem of "social account alignment". Since this is not the focus of this thesis, we just take advantage of the social services, such as About.me⁵ and Quora⁶, that encourage users to explicitly list their multiple social accounts on one homepage. Figure 1.2 shows the screenshots of a user's About.me profile and Quora profile, respectively. From these screenshots we can see that the bottom of each profile displays a list of external links to this user's other social network profiles. This functionality greatly facilitates the process of accurately harvesting users' distributed social contents from multiple social networks.



(a) About.me (b) Quora
Figure 1.2: Screenshot of a user's About.me profile and Quora profile.

1.5.2 Missing Data Completion

Although some users have social accounts on multiple social networks, generally they are active on only a few of them. To deal with this realistic problem, we utilize Non-negative Matrix Factorization (NMF) to explore the latent spaces that are shared by different social networks, and further infer the missing data based upon these latent spaces. The underlying assumption is that users' data extracted from multiple social networks shares certain latent features. In particular, we use

⁵<https://about.me/>.

⁶<http://quora.com/>.

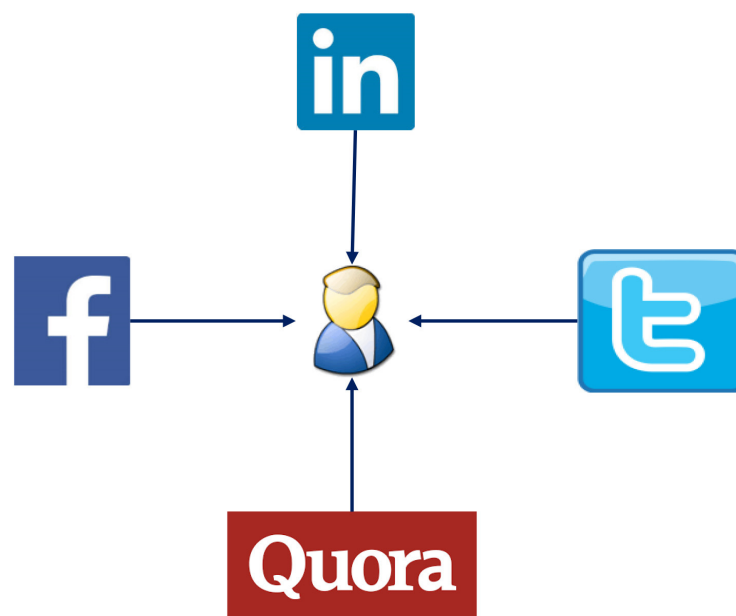


Figure 1.3: Illustration of users' presence on multiple social networks.

NMF to map multiple social networks (views) to latent spaces, where users' latent representations should be similar.

1.5.3 Multiple Social Network Learning

Different aspects of users are disclosed on different social networks due to their different emphasis. Given a user, each social network presents a view of him/her. Figure 1.3 illustrates one's presence on multiple social networks. Essentially, these views are complementary to each other and essentially characterize the same user from different perspectives.

1.6 Contributions

Our main contributions are threefold:

- We propose a novel multi-source mono-task learning scheme to handle user

profiling problem where a single task is involved. This scheme is able to model both the source confidence and source consistency. Moreover, this scheme is able to handle block-wise missing data in multiple social networks. We empirically evaluate our proposed scheme on the application of volunteerism tendency prediction, where we develop a rich set of volunteer-oriented features to characterize users' volunteerism tendency. We have released our compiled dataset⁷ to facilitate other researchers to repeat our experiments and verify their proposed approaches.

- We propose a novel multi-source multi-task learning scheme to tackle the problem of user profiling where multiple tasks are involved. This scheme is able to jointly regularize two important aspects: source consistency and task relatedness. Regarding the task relatedness, two kinds of prior knowledge are introduced: external knowledge and internal knowledge. We practically applied the proposed multi-source multi-task learning scheme in the context of user interest inference.
- We propose a novel learning scheme to detect users' privacy leakage in social media, consists of two components: description and prediction. Regarding the description component, we build a comprehensive taxonomy to characterize users' personal aspects, construct a benchmark dataset, and develop a set of privacy-oriented features. In terms of the prediction component, we propose a taxonomy-guided multi-task learning model to categorize users' social posts, which is able to learn both group-sharing and aspect-specific features simultaneously.

⁷The compiled dataset is currently publicly accessible via: <http://multiplesocialnetworklearning.azurewebsites.net/>.

1.7 Outline of the Thesis

The remainder of this thesis is organized as follows:

Chapter 2 provides the literature survey on the user profiling and related research domains: multi-view learning and multi-task learning.

Chapter 3 presents a novel scheme for multiple social network learning in the context of users' volunteerism tendency prediction [112]. This scheme takes the source confidence and source consistency into consideration by introducing regularization to the objective function. We further demonstrate that the proposed scheme, designed for complete data, is also able to handle the real and more challenging cases where there exists block-wise missing data. In particular, before feeding the data into the proposed MSNL model, we infer the missing data via Non-negative Matrix Factorization (NMF) technique. Furthermore, we practically evaluate the proposed scheme with extensive experiments.

Chapter 4 proposes a structure-constrained multi-source multi-task learning scheme in the context of user interest inference [113]. This scheme is able to co-regularize the source consistency and the tree-guided task relatedness. Meanwhile, it is capable of jointly learning task-sharing and task-specific features. We evaluate the proposed scheme with comprehensive experiments on a real-world dataset.

Chapter 5 presents a framework for privacy leakage detection, consists of two components: description and prediction. In the description component, we pre-define a comprehensive taxonomy, construct a benchmark dataset, and develop a set of privacy-oriented features. The prediction component then proposes a taxonomy-guided multi-task learning model, which is able to learn the latent group-sharing and aspect-specific features simultaneously. We further theoretically relax the proposed non-smooth model to a smooth one and derive the closed-form solution. Finally, we comprehensively evaluate the proposed scheme on a real-world dataset.

Chapter 6 concludes the thesis, highlights the limitations, and points the

future potential research directions.

Chapter 2

Literature Review

In this chapter, we review previous work on user profiling. Since our work involves multiple social networks, which shares the spirit of multi-view learning, we also review the corresponding literature.

2.1 User Profiling

It has been claimed that user profiling refers to inferring unobservable information about an individual based on observable information such as his/her behavior or utterances [148]. In general, user profiling can be helpful in multiple application scenarios, including advertising targeting, personalized recommendation, expert finding, user mobility and planning of social services or governments.

Beyond the world of social media, much attention has been paid to the inference of user attributes from conversational discourse [21, 42, 48]. Fischer et. al [42] first investigated the effects of personal attributes over the morphological features such as the preference between the *-in* and the *-ing* variants of participle ending of verbs. Garera et. al [48] presented a novel partner-sensitive model to predict individuals' biographic attributes based on the sociolinguistic differences

exist in conversations between mixed-gender and same-gender. Bocklet et. al [21] investigated the potential of age determination of children in preschool and primary school from the speech records by machine learning models. Although these work achieves promising results regarding user profiling, they are always hindered by the limited data sources.

With the boom of Web 2.0, most recent state-of-the-art work has focused on investigating the inference of user attributes from social media [101, 90, 6]. Especially, gender and age are the most popular personal attributes being investigated [90, 96, 101, 103]. Rao et. al [101] first proposed to discover author-property such as gender, age, regional origin and political views from microblogs. Four separate support vector machine (SVM) [32] based binary classifications were conducted over two rich set of features: sociolinguistic features and n-gram features. Otterbacher et. al [90] showed that the gender of movie reviewers can be predicted based on stylistic, content, and metadata features. The authors employed the statistical regression model to predict users' gender. Bi et. al [16] demonstrated that utilizing users' historical search queries can promote the inference of user demographic characteristics such as age, gender, and political views. The authors took advantage of the publicly available myPersonality¹ dataset to train the predictive model and applied this model to predict users' demographic characteristics based on their search query logs. Moreover, the authors utilized the Dmoz² to bridge the gap between training samples and test samples that were derived from different sources. The dmoz open directory maintains a hierarchy of conceptual classes for the categorization of web pages. In another work, Pennacchiotti et al. [91] described a general machine learning framework for user classification in three scenarios: political affiliation detection, ethnicity identification and favor prediction for a particular business. Specifically, a set of user-centric features were developed first

¹<http://mypersonality.org/wiki>

²<http://www.dmoz.org>

and then social graph information was taken into consideration to boost the performance achieved by solely user-centric features. Experimental results showed that the boost in performance is limited. Recently, Choudhury [35] studied the potential signals for prediction of depression from social media, ranging from the decrease in social activity, raised negative affect, to greater expression of religious involvement. The authors also employed the SVM classifier to do the prediction. In addition to predicting individual's attributes, Zhao et al. [145] mined the location-based social networks, such as Foursquare, to understand users' profiles at community level.

Additionally, as personality has been verified to be of high relevance to the voluntary behaviors [5, 26], which is related to the application of volunteerism tendency prediction investigated by this thesis, we particularly explore the literature about personality prediction. The widely approved "Big Five" personality model in psychology was first systematically introduced by McCrae [82], which represents an individual's personality at five broad dimensions of: Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness.

- **Extraversion.** Extraversion refers to showing a higher degree of sociability, assertiveness and talkativeness.
- **Agreeableness.** Agreeableness refers to being cooperative, helpful and sympathetic towards other people.
- **Conscientiousness.** Conscientiousness refers to being disciplined, organized and achievement-oriented.
- **Neuroticism.** Neuroticism refers to the degree of emotional stability, impulse control and anxiety.
- **Openness.** Openness refers to a strong intellectual curiosity and a preference for novelty and variety.

Categories	Examples	Categories	Examples
Word count per essay		Affective processes	happy, cried
Words per sentence		Positive emotions	love, nice, sweet
Question marks		Negative emotions	hurt, ugly, nasty
Dictionary words		Anxiety	worried, nervous
Words of more than 6 letters	technology	Anger	hate, kill, annoyed
First person singular	I, me, mine	Sadness	crying, grief, sad
First person plural	my, us, our	Cognitive processes	cause, know, ought
Second person singular	you , your	Causation	because, hence
Third person singular	she, her, him	Insight	think, consider
Negations	no, not, never	Discrepancy	should, could
Swear words	dame, piss, fuck	Inhibition	block, constrain
Articles	a, an, the	Tentativeness	maybe, guess
Prepositions	to, with, above	Certainty	always, never
Numbers	one, two	Social	mate, child
Third person plural	they, their	Motion	arrive, car, go
Auxiliary verbs	am, will, have	Future tense	will, gonna
Present tense	is, does, hear	Past tense	went, ran, had
Adverbs	very, really	Conjunctions	and, but
Quantifiers	few, many	Assent	ok, yes, agree
Nonfluencies	er, hm, umm	Fillers	Blah, I mean
Hearing	listen, hearing	Perceptual process	observing, heard
Job	job, majors	Feeling	feels, touch
Money	audit, cash, owe	Friends	buddy, friend
Achievement	earn, hero, win	Family	daughter
Leisure	cook, chat	Humans	adult, baby, boy
Home	apartment, family	Seeing	view, saw, seen
Sports	and, with, include	Time	end, until, season
Religion	altar, church	Past tense	went, ran, had
Death	bury, coffin, kill	Present tense	is, does, hear
Body	cheek, hands, spit	Future tense	will, gonna
Sexuality	horny, love	Space	down, in, thin
Health	clinic, flu, pill	Inclusive	and, with, include
Biological processes	eat, blood, pain	Exclusive	but, without
Ingestion	dish, eat, pizza	Motion	arrive, car, go

Table 2.1: Category summarization of LIWC directories.

Pennebaker et al. [93] analyzed the linguistic features for each personality trait and developed a transparent text analysis tool in psychology—Linguistic Inquiry and Word Count (LIWC). The underlying idea behind LIWC is that language is the most common and reliable way for people to translate their internal thoughts and emotions into a form that others can understand. The LIWC program consists of two central components listed as follows.

- **Dictionaries.** The core component of LIWC is the dictionaries, which refers to the collection of words that are pre-defined in a particular category and is constructed by certain professionals. Table 2.1 shows the total categories contained in the current LIWC program. These categories go across a variety OF contextS, including linguistic processes, psychological processes, relativity, personal concern and spoken processes. Several language categories are straightforward. For example, the category of “First person singular” characterizing the linguistic processes is made up of three words: ‘I’, ‘me’ and ‘mine’. The category “work” indicates individual’s personal concern and the category “assent” expresses one’s spoken processes.
- **Processing components.** This component goes through each file word by word, where each word is compared with the dictionary file. Once the processing component goes through all the words, LIWC will calculate the percentage of words for each LIWC category. Therefore, each file’s LIWC feature can be denoted as a vector, where each dimension represents a category and the value corresponds to the percentage of words in this text that belonging to this category.

Recently, lots of studies have been conducted to examine personality traits over a variety of social media, including blogs [59, 133], OSNs [10, 81, 98, 108], and even the community question and answering forums [13]. Similar to the inference of user attributes such as age, gender, researchers spend considerable efforts

Personality traits	Positively correlated	Negatively correlated
Extraversion	Personal pronouns; First person singular; Social; Positive emotion;	Death; Negative emotion; Tentative;
Agreeableness	Positive emotion	Negative emotion;Articles; Death; Swear; Anger; Anxiety;
Conscientiousness	Positive emotion	Negations; Negative emotion;
Neuroticism	Anger; Anxiety; Negative emotion;	Positive emotion;Prepositions;
Openness	Tentativeness; Words with more than 6 letters;	Present tense; Causation;

Table 2.2: Summarization of literature findings regarding the correlations between personality traits and LIWC features.

to investigate user-centric features from users’ social media contents, behavior and egocentric social networks for characterizing the individual difference. It is worth mentioning that the LIWC features extracted from users’ textual information are employed in most of the literature regarding personality prediction. It also has been claimed in several works that there do exist individual linguistic difference among people with different personality. Several related finding [55, 93] are listed in Table 2.2. Although many researchers have achieved huge success in user profiling of a single OSN, shown in Table 2.3, they tend to overlook the advantages of aggregating UGC from multiple different functional OSNs. The state-of-the-art in user profiling has shifted from the traditional single OSN to multiple perspectives. Abel et al. [4] exploited users’ professional interests from their social web profiles, including Twitter profile, LinkedIn profile and Delicious profile. Experimental results confirmed that professional interests can be inferred from users’ casual social posts and also showed the high dependence of performance on the sizes of user social posts. In other words, the more active the user is, the better performance of

Related Works	Specific domain	Text	Behavior	Relation	Data source	Data scale
[Fischer et al. 1958]	social influence on linguistic features	✓			text records	24 children
[Bocklet et al. 2008]	age				speech records	212 children
[Garera et al. 2009]	gender, age, native language	✓			tele-phone conversation corpus;	10,000 users; 543 users
[Rao et al. 2010]	age, gender, regional origin, political views	✓	✓	✓	Twitter	1,000 users for gender, 2,000 users for age, 1,000 users for regional origin, 400 users for political views
[Popescu et al. 2010]	home location, gender	✓			Flickr ³	30,000 users
[Otterbacher et al. 2010]	gender	✓			IMDB ⁴	21,012 users (31,300 reviews)
[Bi et al. 2013]	gender, age, religious, political view	✓			Facebook	457,000 users' Facebook data & 3.3 million users' search query logs
[Pennacchiotti et al. 2011]	user classification; democrats, republicans and starbucks	✓	✓	✓	Twitter	10,338 users for democrats task; 6,000 users for republicans task; 10,000 users for Starbucks task
[Choudhury et al. 2013]	depression	✓	✓	✓	Twitter	476 users
[Zhao et al. 2013]		✓				
[Quercia et al. 2012]	Personality			✓	Twitter	335 users
[Markovikj et al. 2013]	Personality	✓			Facebook	250 users (10,000 statuses)
[Bai et al. 2012]	Personality	✓	✓	✓	Renren ⁵	335 users
[Bazelli et al. 2013]	Personality	✓			Stack-Overflow ⁶	total posts on StackOverflow between Aug. 2008-Aug. 2012

Table 2.3: Summarization of related works on user profiling.

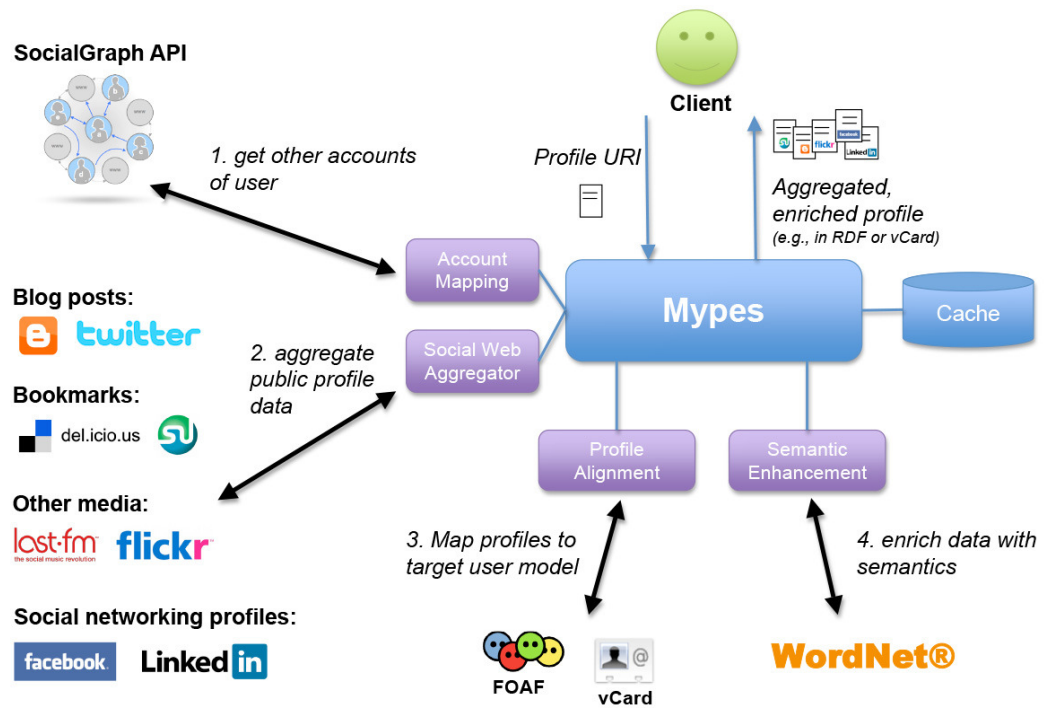


Figure 2.1: Aggregation and enrichment of profile data with Mypes.

inference of professional interests can be achieved. Moreover, Abel et al. [3] presented a service “*Mypes*” that aggregates distributed user profile information from a variety of online services and provides an overview of unified profiles to end-users in ad-hoc manner. As shown in Figure 2.1, *Mypes* consists of four general components: *Account Mapping*, *Social Web Aggregator*, *Profile Alignment* and *Semantic Enrichment*. Tang et. al [117] proposed a combination approach to deal with the profiling tasks with several subtasks: profile extraction, profile integration and user interest discovery. They focused on investigating researchers’ interests from the publications using a probabilistic Topic-Conference-Topic (TCT) model. One distinct limitation of existing work regarding user profiling from multiple sources is that they all fail to take the source relationship into consideration to enhance the performance. It is worth highlighting that, as far as we known, little work has published regarding user profiling from multiple social networks, especially from

the perspective of multi-view learning.

2.2 Multi-view Learning

With the development of technology, data in many research domains, such as image processing, computer vision, and social computing, are growing not only larger but also more complex. These data can be collected from diverse sources, different modalities, and even various feature generators. In a sense, the complex data exhibit a heterogeneous property, and thus can be grouped into different views, where each view describes the sample from particular aspect. For example, in the domain of computer vision, video summarization usually investigates the data, which always involve visual, acoustic and textual modalities. Therefore, the features extracted from one modality comprise a view. Similarly, in our work, we treat that the UGC on each social network as a view to characterize users.

To deal with data with multiple views, conventional machine learning algorithms can be roughly classified into two major categories: early fusion and late fusion. Early fusion methods concatenate all feature spaces from different views into a joint feature space [44], over which further machine learning algorithms can be applied [130]. However, they may suffer from several limitations. First, they are unable to differentiate the discrimination power of different views. Second, it may lead to the curse of dimensionality since the joint view may be of rather high dimensionality, which will further cause the overfitting when the training dataset is not large-scale. Third, it also lacks of physical meaning as each view holds distinct statistical properties [130]. On the other hand, late fusion methods learn each view separately and then integrates all the results. Obviously, these methods overlooked the relationship among different views and thus can only obtain the suboptimal results. Consequently, multi-view learning is a highly desired new paradigm, which is designed to solve such shortcomings and improving the learning performance by

introducing a function to model each view and jointly optimizing all functions. Existing work follows this line can be roughly classified into two categories: co-training and subspace learning.

Co-training was one of the earliest schemes for multi-view learning [20]. In essence, the co-training style algorithms usually train separate learners on distinct views, which are then imposed to be consistent across views. To date, many variants have been developed. Sindhwani et. al. [110] introduced a co-regularization framework for multi-view semi-supervised learning, as an extension of supervised regularization algorithms. Christoudias et al. [30] proposed an approach for multi-view learning in the context of views with corruption, taking the view disagreement into consideration. Considering the existence of incomplete data that miss certain views, Yuan et al. [136] presented an incomplete multi-source feature learning method. In particular, the incomplete data are split into disjoint groups, where feature learning can be conducted independently. However, such a mechanism constrains us to conduct source level analysis. Later, Xiang et al. [129] investigated multi-source learning with block-wise missing data with an application of Alzheimer’s Disease prediction and proposed the iSFS model. Apart from feature-level analysis, the authors also conducted source-level analysis by introducing the weights for the models obtained from different sources. However, ignoring the consistency relationships among different models seems inappropriate. In addition, the authors also adapted the model to handle cases where block-wise missing data exist, which makes it less generalizable to different scenarios.

Subspace learning approaches hold the general assumption that different views are generated from a latent view. Chaudhuri et al. [29] first employed canonical correlation analysis (CCA) to learn an efficient subspace, on which traditional machine learning algorithms can be applied. In particular, the proposed approach is applied to the context of clustering. It is worth noting that the proposed

Related Works	Specific domain	Category	Data source	Data scale
[Sindhwani et al. 2005]	Hypertext document categorization	Co-training	Web documents	1,051 documents
[Christoudias et al. 2012]	User agreement recognition from speech and head gesture	Co-training	User study	15 subjects
[Yuan et al. 2012]	Alzheimer's Disease Prediction	Co-training	Medical records	780 subjects
[Xiang et al. 2013]	Alzheimer's Disease Prediction	Co-training	Medical records	780 subjects
[Chaudhuri et al. 2009]	Audio-visual speaker clustering	Subspace learning	VidTIMIT (audio+visual)	41 users
[Yu et al. 2012]	Image Processing	Subspace learning	Cartoon videos	1,500 characters
[Salzmann et al. 2010]	Pose estimation from monocular images	Subspace learning	HumanEva [109]	72 images
[Zhai et al. 2012]	Pose estimation and facial expression recognition	Subspace learning	COIL-20, private dataset created by the authors, JAFFE	1,440 images, 1,011 images, 213 images
[Gao et al. 2015]	Computer vision	Subspace learning	Caltech101-7, MSRCV1, eth-80, Caltech101-20	441 images, 240 images, 400 images, 1230 images
[Yin et al. 2015]	Document classification	Subspace learning	UCI Handwritten Digit, Cora, BBC, WebKB	2,000 samples, 2,708 publications, 2,012 documents, 1,051 webpages

Table 2.4: The summarization of related works on multi-view learning.

method performs the multi-view learning and clustering independently. Different from this, Gao et al. [46] later introduced a novel multi-view subspace clustering method, which is able to simultaneously perform clustering on the subspace of each view and guarantee the consistency among multiple views by a common clustering structure. Beside the consistency among views, Salzmann et al. [105] further explored the private latent space of each view. The authors introduced a robust approach, where the latent space can be factorized into shared and private spaces by imposing orthogonality constraints among latent spaces. Apart from the case of supervised learning, several efforts have been dedicated to the semi-supervised context, where the problem of insufficient training data can be addressed. Yu et al. [135] proposed a semi-supervised multi-view distance metric learning (SSM-DML) approach, which aims to seek an effective metric to accurately measure the distance between samples and thus promote the learning performance. In addition, Zhai et al. [138] investigated the multi-view metric learning problem under the umbrella of the semi-supervised learning setting. The proposed approach—Multi-view Metric Learning with Global consistency and Local smoothness (MVML-GL), aims to seek a latent feature space, where global consistency and local smoothness are considered. Recent, Yin et al. [134] particularly investigated multi-view learning with the incomplete multi-view data in the context of clustering. The authors employed unified latent representations and projection matrices to deal with the incomplete data. Existing work related to multi-view learning is summarized in Table 2.4. To the best of our knowledge, limited efforts have been dedicated to taking advantage of multi-view learning in the user profiling domain. Furthermore, different from existing work, we not only take the source (view) consistency and source (view) confidence into consideration simultaneously, but also infer the missing data by making full use of the available data before applying multi-view learning, which is more generalizable to other applications.

2.3 Multi-task Learning

Since we aim to learn users' privacy leakage from social media, another important literature is on multi-task learning [25, 63, 131, 37, 132, 146]. Multi-task learning works by jointly solving a problem together with other related problems simultaneously, using a shared representation. This often leads to a better model for the research problem, because it allows the learner to use the commonality among the tasks [25]. Hence, precisely identifying and modeling the task relatedness are of importance. Several regularization-style methods have been proposed in the literature to model task relatedness [37, 7]. Argyriou et al. [7] proposed a framework of multi-task feature learning, which learns shared features among all tasks with convex optimization. The philosophy behind this framework is that all tasks are related, while it may be too restrictive and may adversely affect performance due to the existence of outlier tasks. Towards this end, several approaches have been proposed to discover the relationship among different tasks. One prominent research line is the clustered multi-task learning [146, 118, 61]. Such approaches assume that all tasks can be clustered into several groups, which are usually unknown. Tasks within one group are hence assumed to be closer and share more similar representation. However, the assumption of such approaches is still relatively restrictive in practice, since it only focuses on the grouping structure over task-level but ignores that over feature-level. To address this issue, Xu et al. [131] proposed to formulate multi-task learning with task-feature co-clusters to investigate more comprehensive task-feature relationship.

Beyond them, we manually pre-define a taxonomy to structure the task relatedness, and utilize such taxonomy to guide a novel multi-task learning model, which is capable of learning group-sharing and aspect-specific features. Moreover, we assume that tasks within a group should share certain latent features. On the other hand, MTL has been applied to solve many problems, including social behav-

ior prediction [41], image annotation [18, 38, 39], and web search [9, 28]. However, to the best of our knowledge, limited efforts have been dedicated to applying MTL in the privacy domain, which is the major concern of our work.

2.4 Multi-view Multi-task Learning

The problem of user interest inference from multiple social networks exhibits dual-heterogeneities: each task (interest) corresponds to features from multiple sources. Towards this end, the most related work lies in the area of multi-view multi-task learning. [53] proposed a graph-based iterative framework for multi-view multi-task learning (*IteM²*) in the context of text classification. Given task pairs, *IteM²* projects them to a new Reproducing Kernel Hilbert Space based upon the common views they share. However, this is a transductive model, which fails to generate predictive models on independent and unknown samples. To deal with the intrinsic trouble of transductive models, [140] presented an inductive multi-view multi-task learning model (*regMVMT*). It employs a co-regularization term to achieve model consistency on unlabeled samples from different views. Meanwhile, another regularization function is utilized across multiple tasks to guarantee that the learned models are similar. Noticeably, the implicit assumption that all tasks are uniformly related without prior knowledge might be inappropriate. Realizing this limitation, the authors proposed a revised model (*regMVMT+*) that incorporates a component to automatically infer the task relatedness. As a generalized model of *regMVMT*, an inductive convex shared structure learning algorithm for multi-view multi-task problem (*CSL-MTMV*) was developed in [62]. *CSL-MTMV* considers the shared predictive structure among multiple tasks.

Notably, only a limited number of works have been published regarding multi-view multi-task learning and few of them have been applied to user interest inference. Distinguished from these existing methods which maximize the agree-

ment between views using unlabeled data, our model works towards supervised learning with two advantages: 1) our model consider source consistency and tree-guided relatedness among tasks simultaneously; 2) our model allows the learning of task-sharing features and task-specific features using weighted group lasso, where the weights can be learned from prior knowledge.

2.5 Summary

In this chapter, we review the literature regarding user profiling, multi-view learning and multi-task learning. Literature shows that limited efforts have been dedicated to the user profiling across multiple social networks. Moreover, advanced machine learning techniques such as multi-view learning and multi-task learning have not been applied well to the user profiling domain.

Chapter 3

User Profiling via Multi-Source Mono-task Learning: Volunteerism Tendency Prediction

In this chapter, we aim to propose a multi-source mono-task learning scheme for user profiling, especially, where only a single task would be involved. In particular, we apply the proposed scheme to predict users' volunteerism tendency. Extensive experiments have demonstrated the effectiveness of the proposed scheme.

3.1 Introduction

Volunteerism was defined in [94] as long-term, planned, prosocial behaviors that occur within organizational settings and can benefit strangers. Persons exhibiting volunteerism are the so-called volunteers, serving socially and economically as an important work force in modern society. According to [102], society would face a major crisis without volunteers, especially for nonprofit organizations (NPOs), since they are always in urgent need of volunteers to sustain their daily operations.

Traditionally, it is expensive and time consuming for NPOs to aimlessly recruit volunteers from the huge crowd. It is thus highly desirable to develop an automatic volunteerism tendency prediction system to alleviate the dilemma that a number of NPOs are facing.

In fact, several social researchers have paid attention to volunteerism analysis before the Web 2.0 era. These efforts are mainly based on survey data or related records of individual's volunteer activities [128]. Although great success has been achieved, these approaches suffer from the following two limitations. First, such approaches are hindered by limited and isolated samples as well as constrained individual characteristics. In particular, the experimental data are collected via questionnaires or face-to-face interviews, only small scale dataset and certain basic demographic information, such as gender, marital status and income are available. Second, they mainly focus on the correlation analysis between volunteerism and such characteristics without quantitative volunteerism tendency prediction. For instance, [95] found that users' volunteerism tendency can be affected by four factors: demographic characteristics, personal attributes, volunteer activators and social pressure.

The proliferation of social media has opened a unique opportunity for the volunteerism analysis. In particular, it is a promising approach to predict users' volunteerism tendency by exploring users' distributed UGC of multiple social networks. In a sense, the volunteerism tendency prediction can be treated as the user profiling problem where only a single binary classification (task) is involved. Therefore, this thesis first tackles the user profiling across multiple social networks in the mono-task scenario, where only a single task is involved.

However, integration of multiple sources is non-trivial [139]. The first tough challenge lies in how to fuse users' heterogeneous distributed data from multiple social networks effectively. One naïve approach is to concatenate the feature spaces

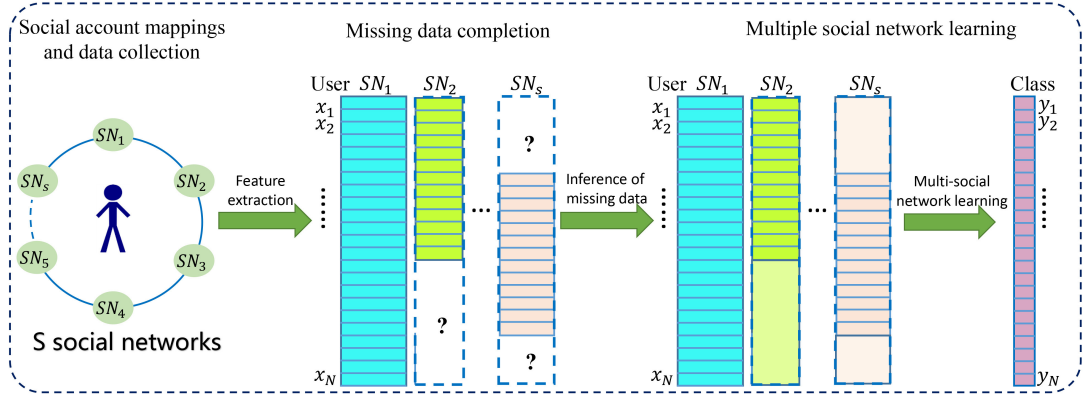


Figure 3.1: Illustration of our proposed scheme. We first collect and align users’ distributed data from multiple social networks. We then jointly infer the block-wise missing data based on the available data. We finally apply MSNL to the complete data. SN_i , x_j , and y_l refer to the i -th social network, j -th user sample, and the l -th corresponding label, respectively.

generated from different sources into a unified feature space. Thereby, traditional machine learning models can be further applied. However, this method simply treats the confidence of all data sources equally and may also lead to the curse of dimensionality. Moreover, it ignores two important facts: 1) different aspects of users are revealed in different social networks and are thus distributed in different feature spaces; and 2) all these aspects tend to characterize the same users. In particular, data from multi-sources describe the same user and thus the results predicted by different sources should be similar. Therefore, it is expected to take the source confidence and source consistency into consideration. Another challenge we face is the missing data problem. Although some users have social accounts on multiple social networks, generally they are active on only a few of them. One simple approach to address this challenge is to discard all incomplete subjects. It is apparent that this method will dramatically reduce the training size, thereby result in overfitting in the model learning stage. Therefore, accurately completing missing data by jointly utilizing multiple sources is a necessity to enhance the learning performance.

To address these problems, we present a multi-source mono-task learning scheme (MSNL), which co-regulates the source confidence and source consistency. Figure 3.1 shows our proposed scheme comprising of three components. Given a set of users, we first crawl their historical contents and all social connections. The first component extracts the multi-faceted information cues to describe a given user, including demographic information, practical behaviors, historical posts, and profiles of social connections. To deal with the block-wise missing data, the second component attempts to infer the block-wise missing data by learning a latent space shared by different social networks, achieving a complete input to the next component. We finally use the last component to conduct MSNL on the complete data. Particularly, we model the confidence of different data sources and the consistency among them by unifying two regularization terms into our model.

Our main contributions can be summarized in threefold:

- We propose a novel MSNL model, which is able to model both the source confidence and source consistency. Specifically, we can obtain a closed-form solution by taking the inverse of a linear system, which has been mathematically proven to be invertible.
- We propose an approach to deal with missing data in multiple social networks, which first learns a common latent subspace shared by different sources [71] and the original missing data can then be derived in turn.
- We empirically evaluate our proposed scheme on the application of volunteerism tendency prediction. In addition, we develop a set of volunteer-oriented features to characterize users' volunteerism tendency. We have released our compiled dataset¹ to facilitate other researchers to repeat our experiments and verify their proposed approaches.

¹The compiled dataset is currently publicly accessible via: <http://multiplesocialnetworklearning.azurewebsites.net/>.

The remainder of this chapter is structured as follows, Section 3.2 briefly reviews the related work. Section 3.3 describes the proposed MSNL model. Missing data completion is introduced in Section 3.4. Section 3.5 mainly presents the dataset and the set of volunteer-oriented features we developed. Section 3.6 details the experimental results and analysis, followed by our concluding remarks in Section 3.7.

3.2 Related Work

Our cross-discipline work is related to a broad spectrum of previous literature, including volunteerism analysis in social science study and multi-view learning.

3.2.1 Volunteerism

Volunteerism analysis has gained tremendous attention from scholars in social science in the past few years. The efforts mainly focus on exploring the motivations and factors that affect volunteering decision [128, 33, 122, 24, 95]. Carlo et al. [24] demonstrated that personality traits, such as extraversion and agreeableness are positively associated with volunteerism. Extraversion characterizes people who are talkative, active and keen on social, while agreeableness characterizes people who are cooperative, helpful and sympathetic to others [12]. Another work in [95] presented an advanced conceptual model of factors that contribute to the decision of volunteering. The proposed factors are *Demographic Characteristics*, *Personal Attributes*, *Volunteer Activators* and *Social Pressure*. Recently, an ongoing project for implementing a volunteer-matching service was introduced in [54]. This project aims to match students' specialties as well as interests with the needs of the local nongovernmental organizations. It also enhances the "Town and Gown Relation" that exists between universities and the towns they reside in.

In spite of the compelling success achieved by these social science researchers, far too little attention has been paid to identifying volunteers from social media. Moreover, most of the existing efforts [95, 24] employ survey or face-to-face interview with samples for data collection, which limits the scalability of their approaches. To bridge the gap, we propose this novel cross-discipline research, aiming to enhance social welfare by exploring the large-scale information in social media.

3.3 Multi-source Mono-task Learning

This section details our proposed MSNL model and derives an analytic solution by solving the inverse of a linear system, whose invertibility is proved rigorously.

3.3.1 Notation

We first declare some notations. In particular, we use bold capital letters (e.g. \mathbf{X}) and bold lowercase letters (e.g. \mathbf{x}) to denote matrices and vectors, respectively. We employ non-bold letters (e.g. x) to represent scalars, and Greek letters (e.g. λ) as parameters. By default, all vectors are in column forms.

Suppose we have a set of N labeled data samples and $S \geq 2$ social networks. We compile the S social networks with an index set $\mathcal{C} = \{1, 2, \dots, S\}$. Let D_s and N_s denote the number of features and samples in the s -th social network, $s \in \mathcal{C}$, respectively. Let $\mathbf{X}_s \in \mathbb{R}^{N \times D_s}$ denote the feature matrix extracted from the s -th social network. Each row represents a user sample. Then the dimension of features extracted from all these social networks is $D = \sum_{s=1}^S D_s$. The whole feature matrix can be written as $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_S\} \in \mathbb{R}^{N \times D}$ and $\mathbf{y} = \{y_1, y_2, \dots, y_N\}^T \in \{1, -1\}^{N \times 1}$ is the corresponding label vector.

3.3.2 Problem Formulations

Based on a set of data samples with S social networks, we can learn S predictive models, where each model is individually and independently trained on a social network. The final predictive model can be strengthened via linear combination of these S models. Mathematically, we learn one linear mapping function \mathbf{f}_s for the s -th social network. In addition, we assume that the mapping functions learned from all social networks agree with one another as much as possible. Particularly, we can formalize this assumption using regularization function. As reported in [88], the squared loss usually yields good performance as other complex ones. We thus adopt the least square loss function for simplicity and have the following objective function,

$$\min_{\mathbf{f}_s} \frac{1}{2N} \left\| \mathbf{y} - \mathbf{f}(\mathbf{X}) \right\|^2 + \frac{\mu}{2N} \sum_{s=1}^S \sum_{s' \neq s} \left\| \mathbf{f}_s(\mathbf{X}_s) - \mathbf{f}_{s'}(\mathbf{X}_{s'}) \right\|^2 + \frac{\lambda}{2} \left\| \mathbf{f} \right\|^2, \quad (3.1)$$

where $\mathbf{f}(\mathbf{X})$ is the final predictive model. $\mathbf{f}_s(\mathbf{X}_s)$ is the prediction results generated from data \mathbf{X}_s . λ and μ are the nonnegative regularization parameters that regulate the sparsity of the solution regarding \mathbf{f}_s and the disagreement among models learned from different social networks, respectively. If we just treat the confidence of different social networks equally, the final predictive model can be formalized as follows,

$$\mathbf{f}(\mathbf{X}) = \frac{1}{S} \sum_{s=1}^S \mathbf{f}_s(\mathbf{X}_s). \quad (3.2)$$

However, in reality, different social networks always have different confidence to the final prediction, and we consider modeling the weights of multiple sources instead of treating all sources equally by introducing the weight vector: $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_S]^T \in \mathbb{R}^{S \times 1}$, where α_s controls the weight of model learned from s -

th social network. Then the final model is defined as follows,

$$\begin{aligned} \mathbf{f}(\mathbf{X}) &= \sum_{s=1}^S \alpha_s \mathbf{f}_s(\mathbf{X}_s) \\ \text{subject to } & \mathbf{e}^T \boldsymbol{\alpha} = 1, \end{aligned} \quad (3.3)$$

where $\mathbf{e} = [1, 1, \dots, 1]^T \in \mathbb{R}^{S \times 1}$. It is worth mentioning that we do not impose the constraint of $\alpha_s \geq 0$, as we want to keep both positive and negative weights. Positive weights indicate the positive correlations of social networks with the final results, while negative weights reflect negative correlations between the given task and different sources, which may contain unreliable and noisy data.

For the s -th social network, we learn a linear mapping function indexed by a model $\mathbf{w}_s \in \mathbb{R}^{D_s \times 1}$. Then the objective function can be rewritten as follows,

$$\begin{aligned} \min_{\mathbf{w}_s, \boldsymbol{\alpha}} \frac{1}{2N} \left\| \mathbf{y} - \sum_{s=1}^S \alpha_s \mathbf{X}_s \mathbf{w}_s \right\|^2 &+ \frac{\mu}{2N} \sum_{s=1}^S \sum_{s' \neq s} \left\| \mathbf{X}_s \mathbf{w}_s - \mathbf{X}_{s'} \mathbf{w}_{s'} \right\|^2 \\ &+ \frac{\lambda}{2} \sum_{s=1}^S \left\| \mathbf{w}_s \right\|^2 + \frac{\beta}{2} \left\| \boldsymbol{\alpha} \right\|^2, \end{aligned} \quad (3.4)$$

where $\mathbf{e}^T \boldsymbol{\alpha} = 1$ and β is the regularization parameter, controlling the sparsity of the solution regarding $\boldsymbol{\alpha}$.

3.3.3 Optimization

We adopt the alternating optimization strategy to solve the two variables $\boldsymbol{\alpha}$ and \mathbf{w}_s in Eqn. (3.4). In particular, we optimize one variable while fixing the other one in each iteration. We keep this iterative procedure until the objective function converges.

3.3.3.1 Computing α with \mathbf{w}_s fixed

We denote the objective function as Γ . For simplicity, we replace \mathbf{y} in Eqn. (3.4) by $\mathbf{y}\mathbf{e}^T\alpha$, as $\mathbf{e}^T\alpha = 1$. With the help of Lagrangian, Γ can be rewritten as follows,

$$\min_{\alpha} \frac{1}{2N} \left\| \mathbf{y}\mathbf{e}^T\alpha - \mathbf{X}\mathbf{W}\alpha \right\|^2 + \frac{\beta}{2} \left\| \alpha \right\|^2 + \delta(1 - \mathbf{e}^T\alpha), \quad (3.5)$$

where δ is the nonnegative Lagrange multiplier and $\mathbf{W} = \text{diag}(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_S) \in \mathbb{R}^{D \times S}$. Taking derivative of Γ with respect to α , we have,

$$\frac{\partial \Gamma}{\partial \alpha} = \frac{1}{N} (\mathbf{y}\mathbf{e}^T - \mathbf{X}\mathbf{W})^T (\mathbf{y}\mathbf{e}^T - \mathbf{X}\mathbf{W})\alpha + \beta\alpha - \delta\mathbf{e}. \quad (3.6)$$

Setting Eqn. (3.6) to zero, it can be derived that,

$$\alpha = \delta \mathbf{M}^{-1} \mathbf{e}, \quad (3.7)$$

where

$$\mathbf{M} = \frac{1}{N} (\mathbf{y}\mathbf{e}^T - \mathbf{X}\mathbf{W})^T (\mathbf{y}\mathbf{e}^T - \mathbf{X}\mathbf{W}) + \beta \mathbf{I}. \quad (3.8)$$

Since $\mathbf{e}^T\alpha = 1$, we can obtain that,

$$\delta = \frac{1}{\mathbf{e}^T \mathbf{M}^{-1} \mathbf{e}}, \quad \alpha = \frac{\mathbf{M}^{-1} \mathbf{e}}{\mathbf{e}^T \mathbf{M}^{-1} \mathbf{e}}. \quad (3.9)$$

Obviously, $\mathbf{M} \in \mathbb{R}^{S \times S}$ is positive definite and invertible, according to the definition. We thus can obtain the analytic solution of α as Eqn. (3.9). Moreover, we note that when the prediction results learned from all social networks are equal, where $\mathbf{X}_1\mathbf{w}_1 = \mathbf{X}_2\mathbf{w}_2 = \dots = \mathbf{X}_S\mathbf{w}_S$, then same weights will be assigned, i.e., $\alpha_1 = \alpha_2 = \dots = \alpha_S$. In addition, Eqn. (3.9) tends to assign higher weight α_s , if smaller difference exists between \mathbf{y} and $\mathbf{X}_s\mathbf{w}_s$.

3.3.3.2 Computing \mathbf{w}_s with $\boldsymbol{\alpha}$ fixed

When $\boldsymbol{\alpha}$ is fixed, we compute the derivative of $\boldsymbol{\Gamma}$ regarding \mathbf{w}_s as follows,

$$\begin{aligned}
\frac{\partial \boldsymbol{\Gamma}}{\partial \mathbf{w}_s} &= \frac{1}{N} \alpha_s \mathbf{X}_s^T \left(\sum_{s=1}^S \alpha_s \mathbf{X}_s \mathbf{w}_s - \mathbf{y} \right) + \frac{\mu}{N} \mathbf{X}_s^T \sum_{s=1}^S \sum_{s' \neq s} (\mathbf{X}_s \mathbf{w}_s - \mathbf{X}_{s'} \mathbf{w}_{s'}) + \lambda \mathbf{w}_s \\
&= \left[\lambda \mathbf{I} + \frac{\alpha_s^2}{N} \mathbf{X}_s^T \mathbf{X}_s + \frac{\mu(S-1)}{N} \mathbf{X}_s^T \mathbf{X}_s \right] \mathbf{w}_s \\
&\quad + \sum_{s=1}^S \sum_{s' \neq s} \frac{1}{N} (\alpha_s \alpha_{s'} - \mu) \mathbf{X}_s^T \mathbf{X}_{s'} \mathbf{w}_{s'} - \frac{\alpha_s}{N} \mathbf{X}_s^T \mathbf{y},
\end{aligned} \tag{3.10}$$

where \mathbf{I} is a $D_s \times D_s$ identity matrix. Setting Eqn. (3.10) to zero and rearranging the terms, all \mathbf{w}_s 's can be learned jointly by the following linear system,

$$\begin{aligned}
&\mathbf{L} \mathbf{w} = \mathbf{t} \\
&\begin{bmatrix} \mathbf{L}_{11} & \mathbf{L}_{12} & \mathbf{L}_{13} & \cdots & \mathbf{L}_{1S} \\ \mathbf{L}_{21} & \mathbf{L}_{22} & \mathbf{L}_{23} & \cdots & \mathbf{L}_{2S} \\ \mathbf{L}_{31} & \mathbf{L}_{32} & \mathbf{L}_{33} & \cdots & \mathbf{L}_{3S} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{L}_{S1} & \mathbf{L}_{S2} & \mathbf{L}_{S3} & \cdots & \mathbf{L}_{SS} \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \mathbf{w}_3 \\ \vdots \\ \mathbf{w}_S \end{bmatrix} = \begin{bmatrix} \mathbf{t}_1 \\ \mathbf{t}_2 \\ \mathbf{t}_3 \\ \vdots \\ \mathbf{t}_S \end{bmatrix},
\end{aligned} \tag{3.11}$$

where $\mathbf{L} \in \mathbb{R}^{D \times D}$ is a sparse block matrix with $S \times S$ blocks, $\mathbf{w} = [\mathbf{w}_1^T, \mathbf{w}_2^T, \dots, \mathbf{w}_S^T]^T \in \mathbb{R}^{D \times 1}$ and $\mathbf{t} = [\mathbf{t}_1^T, \mathbf{t}_2^T, \dots, \mathbf{t}_S^T]^T \in \mathbb{R}^{D \times 1}$ are both sparse block vectors with $S \times 1$ blocks. \mathbf{t}_s , \mathbf{L}_{ss} and $\mathbf{L}_{ss'}$ are defined as follows,

$$\begin{cases} \mathbf{t}_s &= \frac{\alpha_s}{N} \mathbf{X}_s^T \mathbf{y}, \\ \mathbf{L}_{ss} &= \lambda \mathbf{I} + \frac{\alpha_s^2 - \mu}{N} \mathbf{X}_s^T \mathbf{X}_s + \frac{\mu S}{N} \mathbf{X}_s^T \mathbf{X}_s, \\ \mathbf{L}_{ss'} &= \frac{\alpha_s \alpha_{s'} - \mu}{N} \mathbf{X}_s^T \mathbf{X}_{s'}. \end{cases} \tag{3.12}$$

Technically, \mathbf{t} can be treated as a constant matrix as $\boldsymbol{\alpha}$ is fixed. It is worth noting that \mathbf{L} is symmetric as $\mathbf{L}_{ss'} = \mathbf{L}_{s's}^T$. If we can prove that \mathbf{L} is invertible, then we can derive the closed-form solution of \mathbf{w} as follows,

$$\mathbf{w} = \mathbf{L}^{-1} \mathbf{t}. \tag{3.13}$$

We now show \mathbf{L} is invertible by proving that \mathbf{L} is a positive-definite matrix. Let $\mathbf{h} = [\mathbf{h}_1^T, \mathbf{h}_2^T, \dots, \mathbf{h}_S^T]^T \in \mathbb{R}^{D \times 1} \neq \mathbf{0}$ be an arbitrary block vector, where $\mathbf{h}_i \in \mathbb{R}^{D_i \times 1}, i \in \mathcal{C}$. Then we need to prove that $\mathbf{h}^T \mathbf{L} \mathbf{h}$

$$\begin{aligned} &= \sum_{i=1}^S \sum_{j=1}^S \mathbf{h}_i^T \mathbf{L}_{ij} \mathbf{h}_j = \lambda \|\mathbf{h}\|^2 + \frac{1}{N} \left[\sum_{i=1}^S \|\alpha_i \mathbf{X}_i \mathbf{h}_i\|^2 + \mu(S-1) \sum_{i=1}^S \|\mathbf{X}_i \mathbf{h}_i\|^2 \right. \\ &\quad \left. + \sum_{i=1}^S \sum_{j \neq i} \alpha_i \mathbf{h}_i^T \mathbf{X}_i^T \alpha_j \mathbf{X}_j \mathbf{h}_j - \mu \sum_{i=1}^S \sum_{j \neq i} \mathbf{h}_i^T \mathbf{X}_i^T \mathbf{X}_j \mathbf{h}_j \right], \end{aligned} \quad (3.14)$$

is always larger than zero. In fact, given an arbitrary vector \mathbf{b}_i , we have,

$$\begin{aligned} &\|\mathbf{b}_1 - \mathbf{b}_2\|^2 + \dots + \|\mathbf{b}_{(S-1)} - \mathbf{b}_S\|^2 + \|\mathbf{b}_S - \mathbf{b}_1\|^2 \geq 0 \\ &\sum_{i=1}^S \|\mathbf{b}_i\|^2 \geq \sum_{i=1}^S \sum_{j \neq i} \mathbf{b}_i^T \mathbf{b}_j. \end{aligned} \quad (3.15)$$

Therefore, as $S \geq 2$, we have the following inequality,

$$\mu(S-1) \sum_{i=1}^S \|\mathbf{X}_i \mathbf{h}_i\|^2 \geq \mu \sum_{i=1}^S \|\mathbf{X}_i \mathbf{h}_i\|^2 \geq \mu \sum_{i=1}^S \sum_{j \neq i} (\mathbf{X}_i \mathbf{h}_i)^T \mathbf{X}_j \mathbf{h}_j. \quad (3.16)$$

Besides, we know that,

$$\sum_{i=1}^S \|\alpha_i \mathbf{X}_i \mathbf{h}_i\|^2 + \sum_{i=1}^S \sum_{j \neq i} \alpha_i \mathbf{h}_i^T \mathbf{X}_i^T \alpha_j \mathbf{X}_j \mathbf{h}_j = \frac{1}{2} \sum_{i=1}^S \|\alpha_i \mathbf{X}_i \mathbf{h}_i\|^2 + \frac{1}{2} \left\| \sum_{i=1}^S \alpha_i \mathbf{X}_i \mathbf{h}_i \right\|^2 \geq 0. \quad (3.17)$$

Based upon Eqn. (3.16) and Eqn. (3.17), we have that,

$$\mathbf{h}^T \mathbf{L} \mathbf{h} \geq \lambda \|\mathbf{h}\|^2. \quad (3.18)$$

As $\mathbf{h} \neq \mathbf{0}$, $\mathbf{h}^T \mathbf{L} \mathbf{h}$ is always larger than zero. Consequently, \mathbf{L} is invertible. As each iteration can decrease Γ , whose lower bound is zero, we can guarantee the convergence [47, 89].

3.4 Missing Data Completion

In this section, we deal with a more challenging and realistic situation, where block-wise missing data exists, and propose an approach for multiple social network data completion (MSNDC). In such situations, user samples may not be active in all social networks, which leads to the block-wise missing data. Suppose we have S

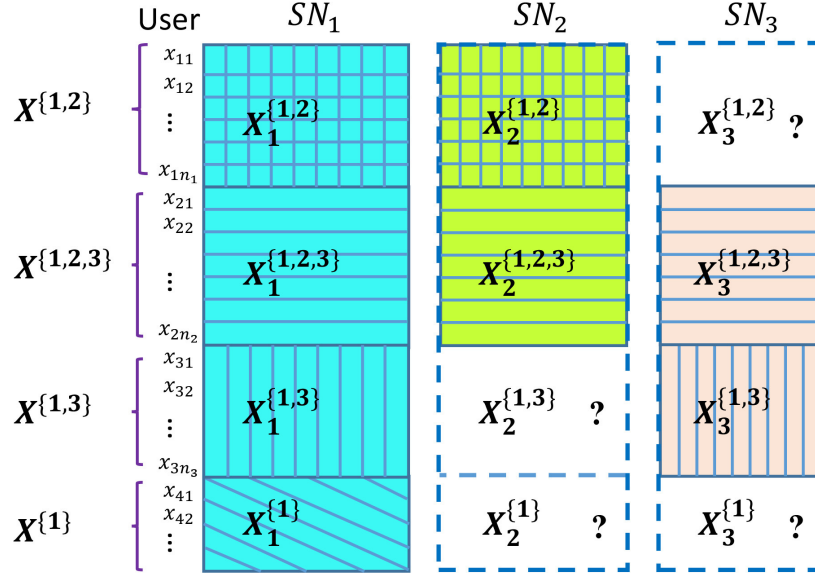


Figure 3.2: Illustration of the incomplete data from three sources. $X_s^{C_i}$ denotes the samples generated from social network s that are only available in the social network combination of C_i .

data sources in total and each sample has at least one data source available. We employ the subset $C_i \subseteq \mathcal{C}$ to indicate the presence of each source and the signature of a specific social network combination. Based on these combinations, all the data samples can be split into multiple exclusive sets, where each set corresponds to a combination. Figure 3.2 illustrates the incomplete data in our dataset. As can be seen, all users have complete features from SN_1 , while some users miss data in SN_2 or SN_3 . Therefore, our dataset can be split by four exclusive social network combinations: $C_1 = \{1, 2\}$, $C_2 = \{1, 2, 3\}$, $C_3 = \{1, 3\}$, $C_4 = \{1\}$.

Inspired by [74], we use Non-negative Matrix Factorization (NMF) to explore

the latent spaces that are shared by different social networks, and further infer the missing data based upon these latent spaces. It is reasonable to assume that the data from different social networks about the same user shares certain latent features. We employ $\mathbf{X}_s^{\mathcal{C}_i} \in \mathbb{R}^{N_{\mathcal{C}_i} \times D_s}$ to denote the samples generated from the s -th social network. It only contains samples that are available in the set of social networks \mathcal{C}_i , where $N_{\mathcal{C}_i}$ stands for the number of these samples. We use $\mathbf{U}_s \in \mathbb{R}^{z \times D_s}$ to represent the latent basis matrix for the s -th social network, and $\mathbf{P}_s^{\mathcal{C}_i} \in \mathbb{R}^{N_{\mathcal{C}_i} \times z}$ to denote the corresponding latent representation of feature matrix $\mathbf{X}_s^{\mathcal{C}_i}$. z is the dimension of the shared latent space of different social networks. The intuitive assumption is that for the samples available in both the s -th and s' -th social networks, their corresponding latent representations should also be similar. In particular, we impose this constraint to NMF as follows,

$$\mathbf{P}_s^{\mathcal{C}_i} = \mathbf{P}_{s'}^{\mathcal{C}_i} = \mathbf{P}^{\mathcal{C}_i}, \quad (3.19)$$

where $s \neq s'$, $s \in \mathcal{C}_i$, and $s' \in \mathcal{C}_i$. We thus learn the shared subspaces by the following objective function,

$$\begin{aligned} \min_{\substack{\mathbf{U}_s \geq \mathbf{0} \\ \mathbf{P}_s \geq \mathbf{0}}} & \left\| \begin{bmatrix} \mathbf{X}_1^{\{1\}} \\ \mathbf{X}_1^{\{1,2\}} \\ \mathbf{X}_1^{\{1,3\}} \\ \mathbf{X}_1^{\{1,2,3\}} \end{bmatrix} - \begin{bmatrix} \mathbf{P}^{\{1\}} \\ \mathbf{P}^{\{1,2\}} \\ \mathbf{P}^{\{1,3\}} \\ \mathbf{P}^{\{1,2,3\}} \end{bmatrix} \mathbf{U}_1 \right\|_F^2 + \nu \|\mathbf{P}_1\|_1 + \eta \|\mathbf{U}_1\|_1 \\ & + \left\| \begin{bmatrix} \mathbf{X}_2^{\{1,2\}} \\ \mathbf{X}_2^{\{1,2,3\}} \end{bmatrix} - \begin{bmatrix} \mathbf{P}^{\{1,2\}} \\ \mathbf{P}^{\{1,2,3\}} \end{bmatrix} \mathbf{U}_2 \right\|_F^2 + \nu \|\mathbf{P}_2\|_1 + \eta \|\mathbf{U}_2\|_1 \\ & + \left\| \begin{bmatrix} \mathbf{X}_3^{\{1,3\}} \\ \mathbf{X}_3^{\{1,2,3\}} \end{bmatrix} - \begin{bmatrix} \mathbf{P}^{\{1,3\}} \\ \mathbf{P}^{\{1,2,3\}} \end{bmatrix} \mathbf{U}_3 \right\|_F^2 + \nu \|\mathbf{P}_3\|_1 + \eta \|\mathbf{U}_3\|_1, \end{aligned} \quad (3.20)$$

where ν and η are the nonnegative tradeoff parameters for the regularizations. Similarly, we employ the alternating optimization strategy to solve the optimization in Eqn. (3.20). To be more specific, we first initialize \mathbf{U}_s and compute the optimal

\mathbf{P}_s . Afterwards, \mathbf{P}_s is updated based on the computed \mathbf{U}_s . We keep this iterative procedure until the objective function converges.

The proposed approach differs from [74] in the following three aspects. First, MSNDC is generalized to handle the more challenging scenario where data samples are extracted from more than two social networks. Second, apart from regulating the latent representation matrix, we also incorporate the regularization on the latent basis matrix. Third, we further derive the original missing data from the latent representation, where the authors in [74] just apply cluster algorithms directly to the latent representation of data instead of the original data. This is due to two considerations. One is that we believe the value of original known data is higher than the latent representation. The other one is that we need to preserve the heterogeneity among data from different sources to fit the MSNL model.

3.4.1 Optimization

In order to increase the efficiency of the iterative procedure, we initialize \mathbf{U}_s by optimizing the following objective function,

$$\begin{aligned} \min_{\mathbf{U}_s \geq \mathbf{0}} & \left\| \mathbf{X}_1^{\{1,2,3\}} - \mathbf{P}^{\{1,2,3\}} \mathbf{U}_1 \right\|^2 + \nu \left\| \mathbf{P}^{\{1,2,3\}} \right\|_1 + \eta \left\| \mathbf{U}_1 \right\|_1 \\ & + \left\| \mathbf{X}_2^{\{1,2,3\}} - \mathbf{P}^{\{1,2,3\}} \mathbf{U}_2 \right\|^2 + \eta \left\| \mathbf{U}_2 \right\|_1 \\ & + \left\| \mathbf{X}_3^{\{1,2,3\}} - \mathbf{P}^{\{1,2,3\}} \mathbf{U}_3 \right\|^2 + \eta \left\| \mathbf{U}_3 \right\|_1. \end{aligned} \quad (3.21)$$

We then alternatively optimize \mathbf{U}_s and \mathbf{P}_s until the objective function converges. Specifically, we employ the greedy coordinate descent (GCD) approach [57], which has been proven to be tremendously fast to solve NMF decomposition with L1-norm regularization. Finally, we obtain $\mathbf{P}_s, \mathbf{U}_s, s \in \mathcal{C}$, based on which we can infer the missing data as follows,

$$\hat{\mathbf{X}}_s^{\mathcal{C}_i} = \mathbf{P}^{\mathcal{C}_i} \mathbf{U}_s, \quad \forall s \notin \mathcal{C}_i. \quad (3.22)$$

3.5 Application: Volunteerism Tendency Prediction

In this work, we cast the problem of volunteerism tendency prediction as a user binary classification. If the predicted tendency score of a given user is larger than a pre-defined threshold γ , we regard this user as a volunteer. In this work, we explore three popular social networks: Twitter, Facebook and LinkedIn, as they are representative of a public, private, and professional social network, respectively. Besides, it is known that users exhibit different aspects on different social networks [3], and the combination of these three social networks would help to better characterize user behaviors on social platforms.

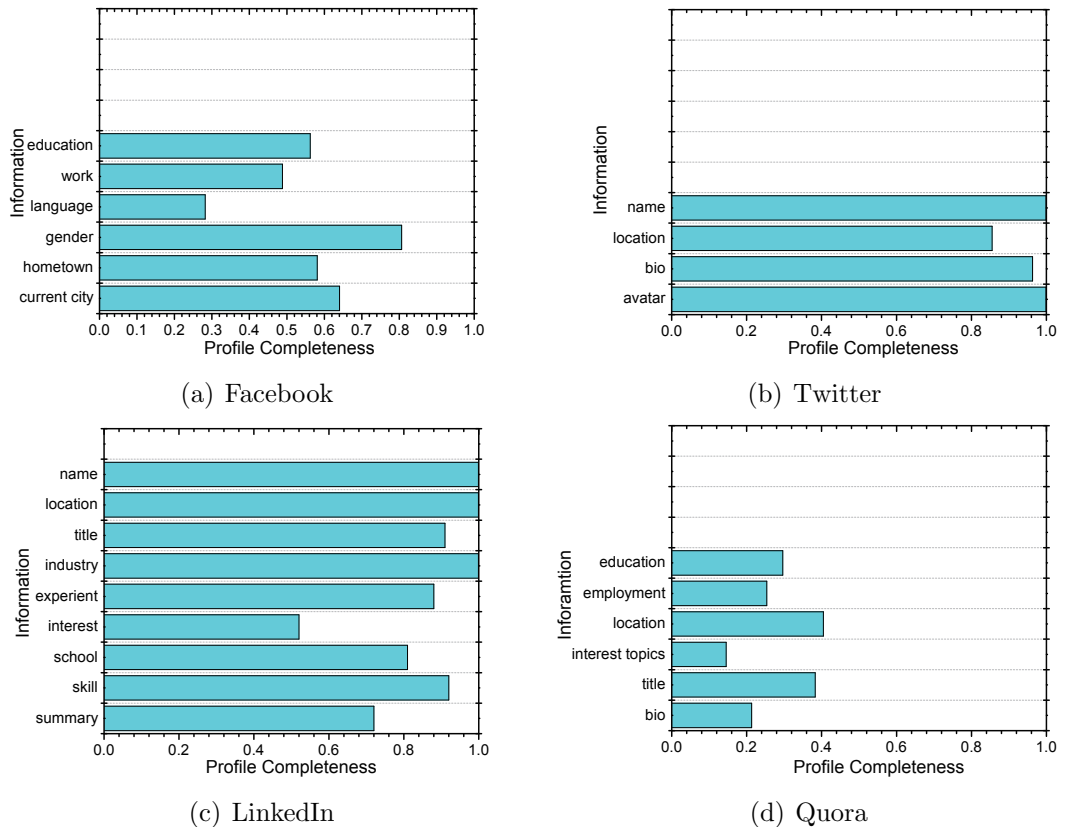


Figure 3.3: Statistics of profile completeness of users over various social networks.

3.5.1 Necessity of Multiple Social Networks

First, we provide the quantitative evidence to validate the necessity of collecting data from multiple social networks. We show the statistics of profile completeness of users over various social networks in Figure 3.3, based on our pilot study of 172,235 users. We have the following observations: 1) 56.2% users provide their education in Facebook profile, while 81% LinkedIn users provide their school information. The incompleteness hinders the effective similarity estimation based on users' profile data; 2) the data distributed in different social networks is complementary. For example, Facebook profiles provide users' gender information but fail to present the bio descriptions for users, which is alternatively given by Twitter profiles. Hence, integration of users' information distributed in various social networks is essential to derive complete user profiles. As a by-product, leveraging multiple sources increases the robustness, helps to handle the cold start problem [106] and may be beneficial to other applications, such as recommendations.

3.5.2 Social Accounts Alignment

To represent the same users with multiple sources, we need to first tackle the problem of "social account alignment", which aims to align the same users across different social networks by linking their multiple social accounts [3]. To accurately establish this mapping, we employ the emerging social services such as About.me and Quora, where they encourage users to explicitly list their multiple social accounts on one profile.

We proposed two strategies to collect data from About.me.

- **Keyword search:** We searched About.me with the keyword "volunteer" and obtained 4,151 volunteer candidates.

- **Random select:** We employed Random API², provided by About.me, to collect non-volunteers. This API returns a specified number of random user profiles. Finally, we harvested 1,867 non-volunteer candidates. It is worth mentioning that volunteers may be present in these random users.

To enlarge our dataset, we also collected candidates from Quora by the breadth-first-search method. Particularly, we took advantage of both the follower and followee³ relations provided by Quora. Initially, we selected two popular users as the seed users and then explored all their neighboring connected users. We applied similar exploration approach to all other non-seed users. In the end, we collected 172,235 users' profiles and only retained those who have accounts in Facebook, Twitter and LinkedIn.

3.5.3 Ground Truth Construction

Based on these candidates, we launched a crawler to collect their historical social contents, including their basic profiles, social posts and relations. However, the traditional web-based crawler is not applicable to Facebook due to its dynamic loading mechanism. We thus resorted to the Selenium⁴ to simulate users' click and scroll operations on a FireFox browser and load users' publicly available information. We limited the access rate to one request per second to avoid being blocked by the robot checkers. It is worth mentioning that the data we collected are all publicly available. On the other hand, due to the privacy constraint, we could not access users' social relations in Facebook and LinkedIn. We hence only collected users' followee relations in Twitter.

In order to improve the quality of our dataset, we employed three annotators from the department of computer science, National University of Singapore, to

²<http://about.me/developer/api/docs/>.

³If A follows B, then A is B's follower and B is A's followee.

⁴<http://docs.seleniumhq.org/download/>.

finalize our ground truth. As users tend to provide more complete and reliable profiles in LinkedIn, we guided the annotators to study the LinkedIn profiles of candidate users, and determine whether they are “volunteers” by majority votes. To ensure a uniformly labeling procedure, we provided them a piece of guideline. Given a user’s LinkedIn profile, we classified the user as a volunteer if and only if this user lists his/her volunteer experiences in the section “Volunteer experience & Causes” or section “Experience”. Candidates who do not satisfy the above two criteria were tagged as non-volunteers. We focused on LinkedIn to determine whether users are volunteers because the volunteer experiences in LinkedIn are the most straightforward evidence to identify volunteers. It should be noted that those who do not mention their volunteer experiences in LinkedIn are not necessarily classified as “non-volunteers”. However, the absence of these mentions, at least, reveals their limited interests and low enthusiasm in volunteerism. Therefore, in our work, we broadly defined users as “non-volunteers” if they do not mention their relevant volunteerism experiences in LinkedIn.

We focus on LinkedIn to obtain volunteers due to this fact: the volunteer experiences in LinkedIn are the most straightforward evidence to identify volunteers. It should be noted that those who do not mention their volunteer experiences in LinkedIn are not necessarily classified as “non-volunteers”. However, the absence of these mentions, at least, suggests their limited interests and low enthusiasm in volunteerism. Therefore, in our work, we broadly define users as “non-volunteers” if they do not mention their relevant volunteerism experiences in LinkedIn.

Table 3.1 lists the statistics of our dataset. We obtained the data for 1,425 volunteers and 4,011 non-volunteers according to the aforementioned strategies. The crawling was conducted between 22nd August to 11th September, 2013. Here we only selected a subset of non-volunteer data and made the dataset balanced to avoid the training bias. To facilitate this line of research, this dataset has been

released after certain privacy preservation processing.

Table 3.1: Statistics of our dataset.

Data	Volun- teer	Non- volunteer
Twitter profiles	$\sim 1.5k$	$\sim 4k$
Twitter posts	$\sim 559k$	$\sim 1m$
Twitter followees' profiles	$\sim 902k$	$\sim 3m$
Facebook profiles	$\sim 1.5k$	$\sim 4k$
Facebook posts	$\sim 83k$	$\sim 338k$
LinkedIn profiles	$\sim 1.5k$	$\sim 4k$

However, in reality, not all users are active enough on all social networks. To ensure the data quality, we treated those inactive users as missing with respect to a specific social network. Therefore, there exists block-wise missing data in our dataset. In particular, we treated a user as missing in Twitter or Facebook, if this user has less than 10 historical social posts. In addition, due to the absence of social post mechanism in LinkedIn, we treated a user as missing⁵ in LinkedIn if the word count of this user's profile is less than 50. Figure 3.4 shows the statistics of our incomplete data. As can be seen, about 50% of users have complete data from all three social networks. 1% and 47% of users only miss the data either from Facebook and LinkedIn, while 2% of users miss the data from both of them.

3.5.4 Features

To capture users' volunteerism tendency, we extracted a rich set of volunteer-oriented features [111].

3.5.4.1 Demographic Characteristics

The study in [95] reported that some demographic characteristics, such as education and income level, are strong indicators for volunteerism. This study inspires us

⁵Here we exclude the contents of section "Volunteer experience & Causes" and section "Experience".

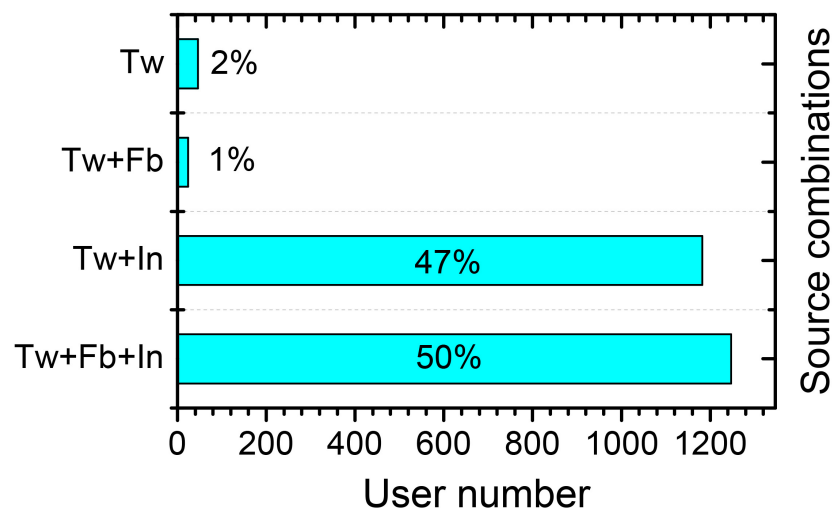


Figure 3.4: Statistics of the incomplete data. Tw: Users with Twitter data only; Tw+Fb: Users with Twitter and Facebook data only; Tw+In: Users with Twitter and LinkedIn data only; Tw+Fb+In: Users without missing data.

to extract demographic characteristics from users’ profiles, especially the Facebook and LinkedIn profiles. In our work, we explored users’ demographic characteristics, including *Gender*, *Relationship status*, *Education level*, and *Number of social connections*.

3.5.4.2 Linguistic Features

We also extracted linguistic features, including Linguistic Inquiry and Word Count (LIWC) features, user topics and contextual topics.

LIWC features. LIWC is widely-used to analyze the psycho-linguistic features in texts. It plays an important role in predicting users’ personality [13, 81]. The main component of LIWC is a directory which contains the mapping from words to 72 categories⁶. Given a document, LIWC computes the percentage of words in each category and represents it as a vector of 72 dimensions. To capture the key aspects of LIWC features, we selected the top 5 dimensions as the representative LIWC features according to the information gain ratio. Considering that the emotions for

⁶<http://www.liwc.net/>.

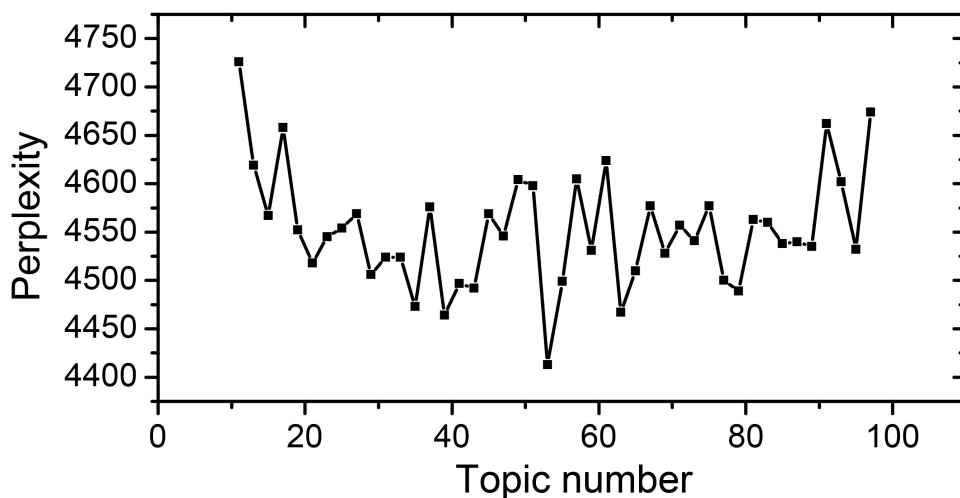


Figure 3.5: Perplexity values varying over the number of topics in Twitter.

individuals may also affect users’ volunteerism tendency, we additionally selected two categories from LIWC: positive emotion and negative emotion.

User topics. According to our observation, volunteers may have, on average, a higher probability of talking about topics such as social caring or giving back, while the non-volunteers may mention other topics more often. This motivates us to explore the topic distributions of users’ social posts to identify volunteers. We generated topic distributions using the Latent Dirichlet Allocation (LDA) model [19], which has been widely found to be useful in latent topic modeling [49, 124]. Based on perplexity [73] metric frequently utilized to find the optimal number of hidden topics. Figure 3.5 shows the perplexity over different topic numbers on users’ historical contents in Twitter. Owing to the noisy nature of UGC, the perplexity distribution can only roughly monotonically decrease as approaching to the lowest point from both ends. Consequently, it is advisable to set the topic number for Twitter as 53 based on the perplexity metric. Following the similar manner, we ultimately obtain 26, 3 dimensional topic-level features over users’ social contents in Facebook and LinkedIn⁷, respectively.

⁷The posts in LinkedIn refer to the section of user summary.

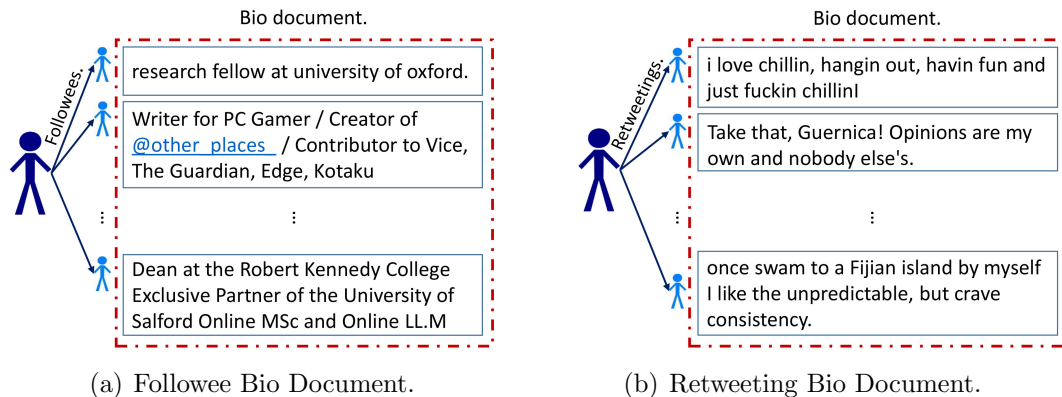


Figure 3.6: Two bio documents of a user. Each document consists of a set of bio descriptions of a user’s specific social connections.

Contextual topics. We define users’ contextual topics as the topics of users’ connections. We believe that the contextual topics intuitively reflect the contexts of users. “He that lies down with dogs must rise up with fleas” tells us that the context significantly affects a user’s tendency. On the other hand, users’ bio descriptions on Twitter are usually employed to briefly introduce users and may indicate users’ interests to some extent. Therefore, we investigate the bio descriptions of a users’ social connections to characterize his/her context. In particular, we studied two kinds of social connections: followees and retweeting⁸ connections on Twitter because of their intuitive reflection of topics that users concern. Consequently, for each user, we thus integrated the bios of his/her followees or retweeting connections into two bio documents, as shown in Figure 3.6. We then further applied LDA model to each kind of documents. We utilized the perplexity to fix the dimensions of topic-level features over followees’ bio documents and retweetings’ bio documents as 40 and 20, respectively. In this work, we only explored the contextual topics in Twitter, since we were unable to crawl the connections’ profiles in LinkedIn and the bio descriptions are usually missing in Facebook.

⁸If A broadcasts a tweet posted by B, then B is A’s a retweeting user.

3.5.4.3 Behavior-based Features

This kind of features is characterized by users’ posting behavior patterns and networking behavior patterns. The former focuses on the written style of users’ social posts, while the latter captures their egocentric network features.

Posting behavior patterns. Posting behavior patterns have been investigated in many scenarios, spanning from age estimation to social spammers discovery [15, 72]. These patterns can be used to depict users’ participation in information diffusion, which correlates with volunteerism tendency much.

On one hand, we employed the fraction of users’ posts containing certain behaviors, including emoticons, slang words⁹, hashtags¹⁰, URLs, and user mentions¹¹, to intuitively reflect users’ engagement in topic discussion and social interaction. On the other hand, we observed that users’ posting behaviors in social networks can be classified into a few categories. For example, posts in Twitter can be classified into two categories, $C_{tw} = \{tweets, retweets\}$, while posts in Facebook can roughly be split into eight types: $C_{fb} = \{share_link, share_video, share_status, share_photo, change_photo, repost, post, tagged\}$. The distributions over users’ posts on these categories also reflect their participation in information diffusion, revealing whether a given user tend to share information in social networks. When it comes to LinkedIn, we utilized the profile completeness to characterize users’ behaviors. Based on our observation, we found that volunteers tend to provide more information for all the sections. This not only reflects volunteers’ active participation in LinkedIn but also signals their self-confidence and openness to public.

⁹Slang words refer to the variety of slang languages coined by Internet users, such as “lol”, “omg” and “asap”.

¹⁰A hashtag refers to a specially designated word prefixed with a ‘#’, which usually represents the topic of this tweet.

¹¹A user mention is a specially designated word in a tweet, prefixed with a “@”, which usually refers to other users.

Profile completeness is defined as a Boolean vector over six dimensions to denote the presence of the six common sections in LinkedIn profiles: `summary`, `interest`, `language`, `education`, `skill` and `honor`. We excluded the sections on `experience` and `volunteer experience & causes`, because the ground truth is built on these two sections.

Egocentric network patterns. We also studied users’ social behaviors from their egocentric networks. Intuitively, we believe that users belong to certain class tend to be connected with several class-specific accounts, as it goes for that “birds of a feather flock together”. Therefore, volunteers should interact with some typical accounts in social media. The set of typical accounts is denoted as \mathcal{TC} . Inspired by [91], we measured the degree of a user’s correlation with volunteerism by three features: the frequency and fraction of a user’s “friends” that belong to \mathcal{TC} as well as the total number of “friends”. In particular, we treated both the followees and retweetings as the “friends” of users in Twitter.

To construct the \mathcal{TC} , we utilized the Twitter profile repository Wefollow¹², which allows us to find the most prominent people given a particular category. By crawling prominent users falling into categories of *Nonprofit*, *Charity*, *Volunteer*, *NGO*, *Community Service*, *Social Welfare* and *Christian* from Wefollow, we obtained 23,285 accounts.

3.6 Experiments

We conducted extensive experiments to comparatively verify our proposed scheme from various angles. Since we have framed the problem of user volunteerism tendency prediction as a standard binary classification, we employed the F_β measure to evaluate the performance [75]. Note that F_β measure considers both precision and recall, where β regulates the importance of recall over precision. In this work,

¹²<http://wefollow.com/>.

we considered precision and recall equally important, and selected F1 measure as the evaluation metric. Furthermore, we launched 10-fold cross validation for each experiment, and reported the average performance. Each fold involves 2,249 training and 250 testing samples. All these experiments were conducted with a server equipped with Intel(R) Xeon(R) CPU X5650 at 2.67GHZ on 48GB RAM, 24 cores and 64-bit CentOS 5.4 operating system.

3.6.1 Data Preprocessing

We first remove the obviously noisy contents by using some filtering rules. Here are a few rules we used: remove sentences that contain fewer than five words; remove sentences that contain more than four punctuation marks; remove sentences that contain fewer than two nouns plus verbs. For the remaining sentences that may contain a lot of noisy terms, such as URLs, user mentions and Internet slangs, we did the following editing: 1) we removed the embedded URLs as well as user mentions; 2) we replaced each slang with its corresponding formal expression. To be more specific, we first constructed a local slang dictionary containing 5,374 words by crawling the Internet Slang Dictionary & Translator¹³, where terms are originated from various sources such as Chat Rooms and Cell Phone Text. Given a UGC, we then transformed each slang to their formal expression by looking up this dictionary; and 3) we also performed lemmatization using *Stanford NLP tool*¹⁴ to link the word variants.

3.6.2 On Model Comparison

We compared MSNL with four baselines. Before that, we completed the data by MSNDC. We also performed the one-way analysis of variance to validate the

¹³<http://www.noslang.com/>.

¹⁴<http://nlp.stanford.edu/software/tmt/tmt-0.4/>.

effectiveness of **MSNL**.

SVM: We chose the learning formulation with the kernel of the radial-basis function. We implemented this method based on LIBSVM [27].

RLS: Regularized least squares model [65] aims to minimize the objective function of $\frac{1}{2N} \left\| \mathbf{y} - \mathbf{X}\mathbf{w} \right\|^2 + \frac{\lambda}{2} \left\| \mathbf{w} \right\|^2$. In fact, the **RLS** model can be deduced from **MSNL** via the settings of $\alpha = [\frac{1}{S}, \frac{1}{S}, \dots, \frac{1}{S}]^T$, $\mu = 0$ and $\beta = 0$.

iSFS: The third baseline is the incomplete source-feature selection model proposed in [129]. This model only assigns weights to models learned from different social networks but ignores the relationships among them. We can derive **iSFS** from **MSNL** by making $\mu = 0$.

regMVMT: The fourth baseline is the regularized multi-view multi-task learning model [140]. This model only regulates the relationships among different views, but fails to take the source confidence into account. We can derive **regMVMT** from **MSNL** by making $\alpha = [\frac{1}{S}, \frac{1}{S}, \dots, \frac{1}{S}]^T$.

Table 3.2: Performance of different models(%).

Approaches	F1-measure	P-value
SVM	83.11	0.038
RLS	82.82	0.025
regMVMT	84.07	0.173
iSFS	84.72	0.281
MSNL	85.59	-

Table 3.2 shows the performance comparison between baselines and our proposed **MSNL**. We noticed that **MSNL** significantly outperforms the **SVM** and **RLS**. This implies that the information on multiple social networks are complementary and characterize users' volunteerism tendency consistently. This also proves that the correlations of different social networks with the task of volunteerism tendency prediction cannot be treated equally. In addition, **MSNL** achieves better performance, as compared with **iSFS** and **regMVMT**, which are the derivations of **MSNL**. This demonstrates that both the source confidence and the source con-

sistency deserve particular attention.

To get better insights about such performance, we also conducted the failure case study. In particular, we compared the failure sample distribution by SVM and MSNL in Figure 3.7. As can be seen, overall, our model shows great superiority than SVM regarding the testing users that involve multiple social networks. This again demonstrates that our model is more applicable to cope with multiple social network learning domain, compared to the single learning method (i.e., SVM). However, we also noticed that SVM outperforms MSNL pertaining to users who only have missing data on LinkedIn. One possible explanation is that the limited data samples (1%) with LinkedIn missing ($\mathbf{x}^{\{1,2\}}$) lower their data completion performance.

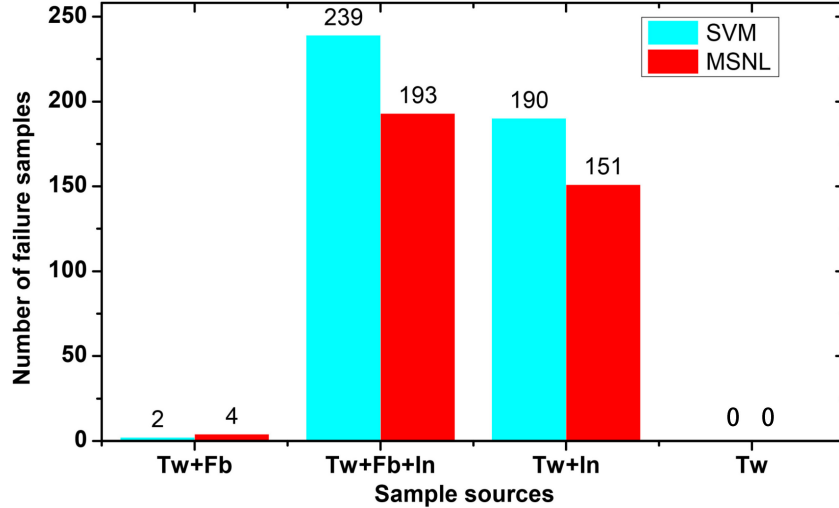


Figure 3.7: Failure sample distribution.

3.6.3 On Data Completion Comparison

We further evaluated the component for missing data completion with the following three baseline methods.

Remove: This method eliminates all data samples that are not complete.

Average: This method imputes the missing features with the average values of the corresponding feature items.

KNN: The missing data are inferred by averaging its K-nearest neighbors. K is experimentally set as 1.

Table 3.3: Performance of different models over different data completion strategies.

Approaches	SVM	RLS	MSNL
Remove	74.91	74.66	81.81
Average	82.09	81.99	85.43
KNN	82.60	82.22	85.55
MSNDC	83.11	82.82	85.59

Table 3.3 shows the performance of different models over different data completion strategies. It can be seen that **MSNDC** outperforms the other strategies. Additionally, removing all incomplete data samples achieves the worst performance, which may be caused by the fact that it introduces training bias, making the dataset unbalanced and reduces the size of training dataset. We found that the percentage of volunteer samples decreases from 50% to 40% after filtering out all incomplete data samples.

3.6.4 On Feature Comparison

To examine the discriminative features we extracted, we conducted experiments over different kinds of features using **MSNL**. We also performed the one-way analysis of variance to validate the advantage of combining multiple social networks. Table 3.4 comparatively shows the performance of **MSNL** in terms of different feature configurations. It can be seen that the linguistic features achieves the best performance, as compared against demographic characteristics and behavior-based features. This reveals that volunteerism tendency is better reflected by their social contents, including their own social posts and the self-descriptions of their social connections. This also implies that users with volunteerism tendency may talk about related topics and follow or retweet related social accounts. In addition, we found that contextual topics are more discriminative as compared to users' own

Table 3.4: Performance of different features(%).

Features	F1-measure
Demographic characteristics	68.43
Linguistic features	80.06
User topics	75.04
Contextual topics	78.14
LIWC	68.48
Behavior-based features	78.52
Posting behavior patterns	69.83
Egocentric network patterns	75.91

topics. This may be due to the fact that users’ self-descriptions are of more value and contain less noise than users’ tweets. Some hot topics discussed by volunteers are given in Table 3.5. Besides, the egocentric network patterns also play a dominant role in our task. This implies that one’s social connections indeed reflect the user’s personal concerns to a large extent.

Table 3.5: Hot topics discussed by volunteers. Followee and retweeting: contextual topics; Self: user topics.

Data source	Topic words
Followee	• public, politics, rights, development
	• editor, global, journalist, university
Retweeting	• global, nonprofit, change, community
	• health, education, learning, university
Self	• woman, help, education, child
	• volunteer, nonprofit, support

However, LIWC, which is also extracted from social posts, does not contribute much compared to the other two personal attribute features. To figure out the underlying logic, we have a close look at the comparison between users belonging to different classes. Table 3.6 comparatively lists the average values of these features among volunteers and non-volunteers. According to [55], Extraversion [82] was much positively associated with the usage of personal pronouns, especially the first person singular. This offers a good explanation of volunteers’ larger adop-

tion of category ‘first person singular’ that volunteers tend to be more open than non-volunteers. Additionally, we can infer that volunteers are more concerned with health than non-volunteers from their larger reference words belong to categories ‘health’ and ‘body’. Moreover, words from the sensory category ‘see’ occur more in volunteers’ posts. This may be due to the fact that volunteers’ active participation in activities and willingness to propagate information in social networks. After checking volunteers’ posts, we found that volunteers do frequently share posts in the following patterns: “... *glad to see...*” and “... *see this proposal: URL*”. Nevertheless, we observed that the difference among people of two classes is not significant.

Table 3.6: Comparison of the value of LIWC features among volunteers and non-volunteers. (%)

	Category	Example	Volunteer	Non-volunteers
1	see	view, seen	1.00	0.95
2	health	clinic, flu	0.48	0.37
3	family	daughter, son	0.22	0.17
4	first person singular	I	2.52	2.26
5	body	hands, spit	0.43	0.40
6	positive	love, great	4.76	4.53
7	negative	hurt, ugly	1.36	1.37
8	PN_emo	-	7.37	6.84

3.6.5 On Source Comparison

To demonstrate the descriptiveness of multiple social network integration, we conducted experiments over various source combinations. Notably, data from Facebook and LinkedIn is incomplete and we need to infer the block-wise missing data first taking advantage of the complete data samples from Twitter.

Table 3.7 shows the performance of **MSNL** over different social network combinations. We noted that the more sources we incorporate, the better the performance we can achieve. This implies the complementary relationships rather than mutual conflicting relationships among the sources. Moreover, we found that aggregating data from all these three social networks can achieve significantly better performance as compared to each of the single source. Additionally, as the performance obtained from different single social networks are not the same, this validates that incorporating the confidence of different social networks to **MSNL** is reasonable. Interestingly, we observed that **MSNL** over Twitter alone achieves the much better performance, as compared to that over LinkedIn or Facebook alone. This may be caused by the fact that the most discriminative features evaluated by Section 3.6.4 are all extracted from Twitter.

Table 3.7: Performance of different social network combinations (%). Facebook* and LinkedIn* both refer to the complete data, whose missing data is pre-inferred. F1: F1-measure.

Social network combinations	F1	p-value
Twitter	82.35	4.2e-2
Facebook*	73.53	5.0e-7
LinkedIn*	74.49	3.1e-7
Twitter+Facebook*	83.67	1.1e-1
Twitter+LinkedIn*	83.84	1.4e-1
Facebook*+LinkedIn*	76.29	6.0e-6
Twitter+Facebook*+LinkedIn*	85.59	-

3.6.6 Size Varying of Positive Samples

In order to verify the usefulness of our model on real world dataset, where the volunteers should account for a minority portion of the user population, we tuned the fraction of volunteer samples in our dataset. In particular, we fed $x\%$, $x \in [5, 50]$, of volunteer samples to our model with stepsize 5%. Figure 3.8 shows the F1-measure with respect to the different fraction of volunteer samples of different

models. As can be seen, our model can achieve satisfactory performance even when volunteer samples only accounts for 5% of the whole samples. This demonstrates that the proposed **MSNL** model is not sensitive to the percentage of positive samples. However, **SVM** and **RLS** are relatively more sensitive to the fraction of volunteer samples in the dataset.

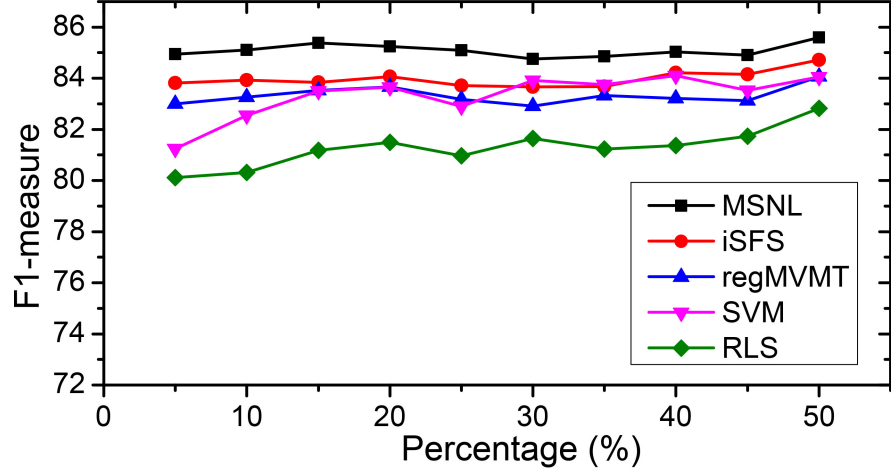


Figure 3.8: F1-measure at different fraction of volunteer samples.

3.6.7 Complexity Discussion

In order to analyze the complexity of **MSNL**, we need to solve the time complexity in terms of constructing **M**, **L** and **t** as defined in Eqn. (3.8) and Eqn. (3.12), and computing the inverse of **M** and **L**. Assume $D \gg S$, the construction of matrix **M** has a time complexity of $\mathcal{O}(NDS)$, and the construction of matrix **L** has a time complexity of $\mathcal{O}(ND^2)$. Due to the fact that the cost of matrix multiplications ($\mathbf{X}_s^T \mathbf{X}_{s'}$) and that of constructing **t** involved in Eqn. (3.12) remain the same for all iterations and **L** is symmetric, we can save much practical time cost. Also, using the standard method, computing the inverse of two core matrices, **M** and **L**, has the complexity of $\mathcal{O}(S^3)$ and $\mathcal{O}(D^3)$, respectively. Furthermore, using the method of Coppersmith and Winograd, the time cost can be bounded by $\mathcal{O}(S^{2.376})$ and $\mathcal{O}(D^{2.376})$ [138], respectively. We note that the speed bottleneck lies in the

number of features and the number of social networks instead of the number of data samples. As S and D are usually small, especially S , **MSNL** should be efficient in time complexity.

3.7 Summary

This chapter presented a novel scheme for multi-source mono-task learning. This scheme takes the source confidence and source consistency into consideration by introducing regularization to the objective function. We further demonstrated that the proposed scheme, designed for complete data, is also able to handle the real and more challenging cases where there exists block-wise missing data. In particular, before feeding the data into the proposed MSNL model, we inferred the missing data via NMF technique. Furthermore, we practically evaluated the proposed scheme in an interesting scenario of volunteerism tendency prediction. Experimental results demonstrated the effectiveness of our proposed scheme and verified the advantages of utilizing multiple social network over a single source.

Chapter 4

User Profiling via Multi-source Multi-task Learning: User Interest Inference

In this chapter, we propose a multi-source mono-task learning scheme for user profiling in situations in which multiple tasks would be involved. In particular, we apply the proposed scheme to infer users' interests. Extensive experiments have demonstrated the effectiveness of the proposed scheme.

4.1 Introduction

User interest inference is the basis for many applications, such as adaptive E-learning [2] and personalized service [92, 97, 125]. Take target advertisement as an example. It is naturally to market cosmetics to ladies, whom are keen on beauty. On the other hand, in a sense, multiple social networks comprehensively convey users' interests from different views. For instance, users may update their daily interests in Facebook, follow their interested accounts in Twitter, and ask or answer

questions they are interested in Quora. Thus, fusing cues from multiple sources can potentially boost the performance of user interest inference by a large margin. In the context of user interest inference, each interest is usually aligned with one task. Given a set of interests, the inference of users' interests can hence be cast as a set of binary classifications. Moreover, these interests (tasks) maybe correlated at different levels. Therefore, it is essential to propose an effective multi-source multi-task learning scheme to model relations between interests (tasks) for user interest inference.

Inferencing user interests from multiple social networks, however, is non-trivial due to the following reasons. (a) **Source Integration.** Although users' footprints on heterogeneous social networks describe their interests from different views, they should characterize a same interest preference consistently. Therefore, how to effectively and comprehensively fuse them is one tough challenge. (b) **Interest Relatedness Characterization.** Interests are usually not independent but correlated in a nonuniform way. For example, given a set of interests $\mathcal{I} = \{basketball, football, travel, cooking\}$, the relatedness between *basketball* and *football* may be stronger than that between *basketball* and *cooking*. Given that in our dataset, most users who like to play basketball are more likely to spend their spare time on the football than cooking. Consequently, the second challenge is how to capture and characterize the relatedness among tasks and how to incorporate this into the multi-task learning. (c) **Discriminant Feature Selection.** The discrimination of features is different from task to task. Learning task-sharing features and task-specific features effectively is significant to user interest inference. This thus poses another crucial challenge for us.

It is noticeable that there are three lines of research dedicated to the problem of user interest inference. One is the single source single task learning [92]. In this context, neither the relatedness among tasks nor the complementary information

across sources is explored. Another line of effort is multi-task learning [132]. They take the task relatedness into account to boost the learning performance and alleviate the problem of insufficient training samples that the traditional single task learning is faced with. It has been observed that learning multiple related tasks simultaneously can improve the modeling accuracy and lead to a better learning performance, especially in cases where only a limited number of positive training samples exist for each task [40]. The third category of approaches is the multi-source learning [3, 4]. Instead of sticking to a single source, they propose to aggregate multiple sources to infer users’ interests. It should be noted that the last two categories of approaches have the weakness of: existing multi-task learning explores the relatedness among tasks, but overlooks the consistency among different sources of a single task; whereas existing multi-source learning ignores the value of the label information of the other related tasks.

As an improvement of the existing works, we propose a structure-constrained multi-source multi-task learning (SM^2L) scheme to infer users’ interests. In particular, our scheme jointly regularizes two important aspects. One is the source consistency. The rationale is that interests reflected by different social networks for the same person should be similar, and hence the disagreement among the prediction results should be penalized. The other is the tree-guided task relatedness modeling. Due to the fact that tree structure has been proven to be capable of characterizing different levels of task relatedness [66], we organize all these interests (tasks) into a tree structure based on our prior knowledge. Specifically, the tree structure settles all tasks in leaf nodes and characterizes the relatedness among them by internal nodes. Moreover, the higher level the internal node is located, the weaker the relatedness imposed on its children tasks is. This is accomplished by a tree-guided group lasso regularizer. Meanwhile, SM^2L learns representative features for individual task and groups of related tasks. A potential benefit of sharing

training instances among tasks is that the data scarcity problem can be alleviated. Extensive experiments on a real-world dataset validate our scheme well. We have released our compiled dataset¹, which will facilitate other researchers to repeat our approach and to comparatively verify their own ideas.

The remainder of this chapter is structured as follows, Section 4.2 briefly reviews the related work. Section 4.3 introduces the proposed structure-constrained multi-source multi-task learning scheme. Section 4.4 details the experimental results and analysis, followed by our concluding remarks in Section 4.5.

4.2 Related Work

4.2.1 User Interest Inference

User interest inference has attracted a lot of researchers' attention. Existing approaches to solving this problem can be roughly classified into three major categories [100]: term vector approaches, ontological approaches and machine learning approaches. First, term vector approaches [117, 127] aim to represent a user's interests by a vector of weighted keywords. For example, Wu et al. [127] applied tf-idf ranking [104] and TextRank [85] to extract keywords and built user interest profiles from Twitter messages. Later, this work was extended by Vu et al. [121], where more advanced techniques were utilized in keyphrase extraction. Although such approaches provide intuitive representation of users' interests, the major limitation they suffer from is the problem of word sparseness and semantic gap. Second, to bridge this semantic gap, ontological approaches [84, 123] were proposed, which attempt to take advantage of existing knowledge bases, such as Wikipedia² and dmoz, to construct users' profiles. Michelson et al. [84] investigated the problem

¹The compiled dataset is currently publicly accessible via: <http://msmt.farbox.com/>.

²<https://en.wikipedia.org/wiki/>

of the discovery of users’ topics of interest. The authors first extracted entities mentioned in tweets, then took advantage of the knowledge base Wikipedia to disambiguate and categorize the entities, and thus constructed the topic profiles for users. However, the evaluation, which always involves user study, constrains the scale of the experimental dataset to a large extent. Third, machine learning approaches [91] infer users’ interests based on training positive and negative samples of users’ interests. Essentially, such approaches are always hindered by the limited labeled samples. Nevertheless, the proliferation of social media, where a huge volume of UGC exists, breaks this dilemma. For example, users’ may edit their bio descriptions and even interests in the profiles. In a sense, such UGC can be utilized as labelled data. Pennacchiotti et al. [91] focused on constructing user interest profiles regarding three aspects: political affiliation, ethnicity and affinity for a particular business, from Twitter by proposing a machine learning framework. Although this work shows great potential of applying machine learning techniques to social media to investigate user interest profiles, the authors overlook the relatedness among users’ interests. Beyond that, in our work, we aim to take advantage of machine learning techniques to perform user interest profiling across multiple social networks. Furthermore, taking the task relatedness into consideration, we embed the problem of user interest inference in the multi-task context to boost the learning performance.

4.3 User Interest Inference

This section details the proposed SM^2L scheme for user interest inference.

4.3.1 Notation

Suppose we have a set of N labeled data samples, $S \geq 2$ sources and $T \geq 2$ tasks. Let D_s denote the number of features extracted from the s -th source. Let $\mathbf{X}_s \in \mathbb{R}^{N \times D_s}$ denote the feature matrix generated from source s , and each row represents a user sample. The feature dimension extracted from all these sources is thus $D = \sum_{s=1}^S D_s$. The whole feature matrix can be written as $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_S\} \in \mathbb{R}^{N \times D}$. The label matrix can be represented as $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T\} \in \mathbb{R}^{N \times T}$, where $\mathbf{y}_t = (y_t^1, y_t^2, \dots, y_t^N)^T \in \mathbb{R}^N$ corresponds to the label vector regarding the t -th task.

4.3.2 Problem Formulations

For each task, we can learn S predictive models, each of which is generated from one source and defined as follows,

$$\mathbf{f}_{st}(\mathbf{X}_s) = \mathbf{X}_s \mathbf{w}_{st}, \quad (4.1)$$

where $\mathbf{w}_{st} = (w_{st}^1, w_{st}^2, \dots, w_{st}^{D_s})^T \in \mathbb{R}^{D_s}$ represents the linear mapping function for the t -th task with respect to the s -th source. The final predictive model for task t can be reinforced via linear combination of these S models. Without the prior knowledge of source confidence, we treat all sources equally as follows,

$$\mathbf{f}_t(\mathbf{X}) = \sum_{s=1}^S \frac{1}{S} \mathbf{f}_{st}(\mathbf{X}_s). \quad (4.2)$$

In multi-class problems, tasks are usually inter-correlated. Multi-source multi-task learning is thus proposed to model their relatedness while seamlessly integrating multiple sources. To select discriminant features, group lasso is considered in the component of multi-task learning. Let $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_T) \in \mathbb{R}^{D \times T}$ denote the linear mapping block matrix, where $\mathbf{w}_t = (\mathbf{w}_{1t}^T, \mathbf{w}_{2t}^T, \dots, \mathbf{w}_{St}^T)^T \in \mathbb{R}^D$. The multi-

source multi-task learning with group lasso can be formalized as follows,

$$\Gamma = \frac{1}{2N} \sum_{t=1}^T \left\| \mathbf{y}_t - \sum_{s=1}^S \frac{1}{S} \mathbf{X}_s \mathbf{w}_{st} \right\|^2 + \frac{\lambda}{2} \sum_{s=1}^S \sum_{d=1}^{D_s} \left\| \mathbf{w}_s^d \right\|, \quad (4.3)$$

where $\mathbf{w}_s^d = (w_{s1}^d, w_{s2}^d, \dots, w_{sT}^d)$, $\sum_{s=1}^S \sum_{d=1}^{D_s} \left\| \mathbf{w}_s^d \right\| = \left\| \mathbf{W} \right\|_{2,1}$ and λ is the nonnegative regularization parameter that regulates the sparsity of the solution regarding \mathbf{W} . When $T \geq 2$, the weights of one feature across all tasks are first grouped by the L_2 norm, and all features are then grouped by the L_1 norm. Thus, the $L_{2,1}$ norm penalty is able to select features based on their strength over all tasks. In this way, we can simultaneously learn the task-sharing features and task-specific features. Obviously, when $T = 1$, this formulation reduces to Lasso [119].

However, the above optimization problem simply assumes that all the tasks share a common set of relevant input features, which might be unrealistic in many real world scenarios. For example, in our work, the tasks “basketball” and “football” tend to share a common set of relevant input features, which are less likely to be useful for the task “cooking”. This consideration propels us to assume that the relatedness among different tasks can be characterized by a tree \mathcal{T} with a set of nodes \mathcal{V} . In particular, the leaf nodes represent all the tasks, while the internal nodes denote the groupings of leaf nodes. Intuitively, each node $v \in \mathcal{V}$ of the tree \mathcal{T} can be associated with group G_v , which consists of all the leaf nodes (tasks) belonging to the subtree rooted at node v . Moreover, the higher level the internal node is located at, the weaker relatedness it controls. The root of \mathcal{T} is assigned the highest level. To characterize such strength of relatedness among tasks, we assign a weight e_v to each node $v \in \mathcal{V}$ according to the prior knowledge via a hierarchical agglomerative clustering algorithm [107]. As illustrated in Figure 4.1, it is apparent that the tasks “basketball” and “football” are more correlated as compared to the task “cooking”. Thus, in Figure 4.1, the tasks “basketball” and “football” are first grouped in node v_4 with a weight $e_{v_4} = 0.6$. Then these two tasks are grouped

in a higher level internal node v_5 , whose weight $e_{v_5} = 0.4$, together with the task “cooking”.

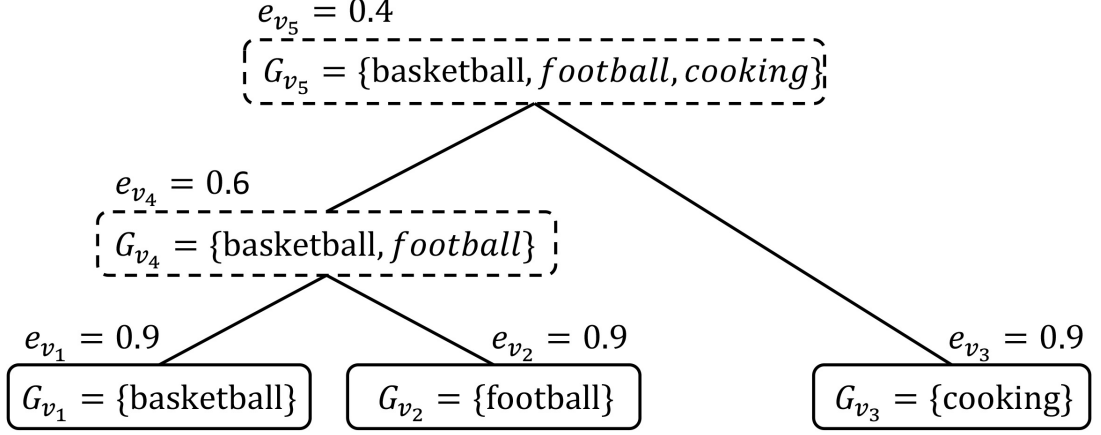


Figure 4.1: Illustration of inter-interests relatedness in a tree structure.

We mathematically formulate the source integration and tree-constrained group lasso into one unified model,

$$\mathbf{\Gamma} = \frac{1}{2N} \sum_{t=1}^T \left\| \mathbf{y}_t - \sum_{s=1}^S \frac{1}{S} \mathbf{X}_s \mathbf{w}_{st} \right\|^2 + \frac{\lambda}{2} \sum_{s=1}^S \sum_{d=1}^{D_s} \sum_{v \in \mathcal{V}} e_v \left\| \mathbf{w}_{sG_v}^d \right\|, \quad (4.4)$$

where $\mathbf{w}_{sG_v}^d$ is a vector of coefficients $\{w_{st}^d : t \in G_v\}$. In addition, we assume that the mapping functions from all sources agree with one another as much as possible. Therefore, we introduce the regularization term to model the result consistency among different sources. The final objective function $\mathbf{\Gamma}$ is restated as follows,

$$\begin{aligned} & \frac{1}{2N} \sum_{t=1}^T \left\| \mathbf{y}_t - \sum_{s=1}^S \frac{1}{S} \mathbf{X}_s \mathbf{w}_{st} \right\|^2 + \frac{\lambda}{2} \sum_{s=1}^S \sum_{d=1}^{D_s} \sum_{v \in \mathcal{V}} e_v \left\| \mathbf{w}_{sG_v}^d \right\| \\ & + \frac{\mu}{2N} \sum_{t=1}^T \sum_{s=1}^S \sum_{s' \neq s} \left\| \mathbf{X}_s \mathbf{w}_{st} - \mathbf{X}_{s'} \mathbf{w}_{s't} \right\|^2, \end{aligned} \quad (4.5)$$

where μ is the nonnegative regularization parameter that regulates the disagreement among models learned from different sources.

4.3.3 Optimization

Considering that the second term in Eqn. (4.5) is not differentiable, we use an equivalent formulation of it, which has been proven by [8], to facilitate the optimization as follows,

$$\frac{\lambda}{2} \left(\sum_{s=1}^S \sum_{d=1}^{D_s} \sum_{v \in \mathcal{V}} e_v \left\| \mathbf{w}_{sG_v}^d \right\| \right)^2. \quad (4.6)$$

Still, the $L_{2,1}$ norm in the above formulation gives rise to a non-convex function, which makes it intractable to solve directly. Therefore, we further resort to another variational formulation [7] of Eqn. (4.6). According to the Cauchy-Schwarz inequality, given an arbitrary vector $\mathbf{b} \in \mathbb{R}^M$ such that $\mathbf{b} \neq \mathbf{0}$, we have,

$$\sum_{i=1}^M |b_i| = \sum_{i=1}^M \theta_i^{\frac{1}{2}} \theta_i^{-\frac{1}{2}} |b_i| \leq \left(\sum_{i=1}^M \theta_i \right)^{\frac{1}{2}} \left(\sum_{i=1}^M \theta_i^{-1} b_i^2 \right)^{\frac{1}{2}} \leq \left(\sum_{i=1}^M \theta_i^{-1} b_i^2 \right)^{\frac{1}{2}}, \quad (4.7)$$

where θ_i 's are introduced variables that should satisfy $\sum_{i=1}^M \theta_i = 1, \theta_i > 0$ and the equality holds for $\theta_i = |b_i| / \left\| \mathbf{b} \right\|_1$. Based on this preliminary, we can derive the following inequality,

$$\begin{aligned} \left(\sum_{s=1}^S \sum_{d=1}^{D_s} \sum_{v \in \mathcal{V}} e_v \left\| \mathbf{w}_{sG_v}^d \right\| \right)^2 &\leq \sum_{s=1}^S \frac{\left(\sum_{d=1}^{D_s} \sum_{v \in \mathcal{V}} e_v \left\| \mathbf{w}_{sG_v}^d \right\| \right)^2}{q_s} \\ &\leq \sum_{s=1}^S \sum_{d=1}^{D_s} \frac{\left(\sum_{v \in \mathcal{V}} e_v \left\| \mathbf{w}_{sG_v}^d \right\| \right)^2}{q_{s,d}} \leq \sum_{s=1}^S \sum_{d=1}^{D_s} \sum_{v \in \mathcal{V}} \frac{e_v^2 \left\| \mathbf{w}_{sG_v}^d \right\|^2}{q_{s,d,v}}, \end{aligned} \quad (4.8)$$

where we introduce the variable $q_{s,d,v}$. The equality can be attained if $q_{s,d,v}$ satisfies that,

$$q_{s,d,v} = \frac{e_v \left\| \mathbf{w}_{sG_v}^d \right\|}{\sum_{s=1}^S \sum_{d=1}^{D_s} \sum_{v \in \mathcal{V}} e_v \left\| \mathbf{w}_{sG_v}^d \right\|}. \quad (4.9)$$

Consequently, minimizing Γ is equivalent to minimizing the following convex objective function,

$$\begin{aligned} & \frac{1}{2N} \sum_{t=1}^T \left\| \mathbf{y}_t - \sum_{s=1}^S \frac{1}{S} \mathbf{X}_s \mathbf{w}_{st} \right\|^2 + \frac{\lambda}{2} \sum_{s=1}^S \sum_{d=1}^{D_s} \sum_{v \in \mathcal{V}} \frac{e_v^2 \left\| \mathbf{w}_{sG_v}^d \right\|^2}{q_{s,d,v}} \\ & + \frac{\mu}{2N} \sum_{t=1}^T \sum_{s=1}^S \sum_{s' \neq s} \left\| \mathbf{X}_s \mathbf{w}_{st} - \mathbf{X}_{s'} \mathbf{w}_{s't} \right\|^2. \end{aligned} \quad (4.10)$$

To facilitate the computation of the derivative of objective function Γ with respect to \mathbf{w}_{st} , we define a diagonal matrix $\mathbf{Q}_{st} \in \mathbb{R}^{D_s \times D_s}$ as follows,

$$Q_{st}(d, d) = \sum_{v: t \in G_v} \frac{e_v^2}{q_{s,d,v}}. \quad (4.11)$$

Finally, we have the following objective function,

$$\begin{aligned} & \frac{1}{2N} \sum_{t=1}^T \left\| \mathbf{y}_t - \sum_{s=1}^S \frac{1}{S} \mathbf{X}_s \mathbf{w}_{st} \right\|^2 + \frac{\lambda}{2} \sum_{t=1}^T \sum_{s=1}^S \mathbf{w}_{st}^T \mathbf{Q}_{st} \mathbf{w}_{st} \\ & + \frac{\mu}{2N} \sum_{t=1}^T \sum_{s=1}^S \sum_{s' \neq s} \left\| \mathbf{X}_s \mathbf{w}_{st} - \mathbf{X}_{s'} \mathbf{w}_{s't} \right\|^2. \end{aligned} \quad (4.12)$$

We adopt the alternating optimization strategy to solve Eqn. (4.12) [66]. Particularly, we alternatively optimize \mathbf{w}_{st} and $q_{s,d,v}$, where we optimize one variable with the other one fixed in each iteration and keep this iterative procedure until the objective value converges.

When $q_{s,d,v}$ is fixed, we take the derivative of objective function Γ regarding \mathbf{w}_{st} as follows,

$$\begin{aligned} \frac{\partial \Gamma}{\partial \mathbf{w}_{st}} &= \frac{1}{NS} \mathbf{X}_s^T \left(\sum_{s=1}^S \frac{1}{S} \mathbf{X}_s \mathbf{w}_{st} - \mathbf{y}_t \right) + \lambda \mathbf{Q}_{st} \mathbf{w}_{st} \\ &+ \sum_{s \neq s'} \frac{\mu}{N} \mathbf{X}_s^T (\mathbf{X}_s \mathbf{w}_{st} - \mathbf{X}_{s'} \mathbf{w}_{s't}). \end{aligned} \quad (4.13)$$

Setting Eqn. (4.13) to zero and rearranging the terms, we derive that all \mathbf{w}_{st} 's can

be learned jointly by the following linear system given a task t ,

$$\mathbf{L}_t \mathbf{w}_t = \mathbf{b}_t,$$

$$\begin{bmatrix} \mathbf{L}_{11} & \mathbf{L}_{12} & \mathbf{L}_{13} & \cdots & \mathbf{L}_{1S} \\ \mathbf{L}_{21} & \mathbf{L}_{22} & \mathbf{L}_{23} & \cdots & \mathbf{L}_{2S} \\ \mathbf{L}_{31} & \mathbf{L}_{32} & \mathbf{L}_{33} & \cdots & \mathbf{L}_{3S} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{L}_{S1} & \mathbf{L}_{S2} & \mathbf{L}_{S3} & \cdots & \mathbf{L}_{SS} \end{bmatrix} \begin{bmatrix} \mathbf{w}_{1t} \\ \mathbf{w}_{2t} \\ \mathbf{w}_{3t} \\ \vdots \\ \mathbf{w}_{St} \end{bmatrix} = \begin{bmatrix} \mathbf{b}_{1t} \\ \mathbf{b}_{2t} \\ \mathbf{b}_{3t} \\ \vdots \\ \mathbf{b}_{St} \end{bmatrix}, \quad (4.14)$$

where $\mathbf{L}_t \in \mathbb{R}^{D \times D}$ is a sparse block matrix with $S \times S$ blocks, $\mathbf{w}_t \in \mathbb{R}^D$ and $\mathbf{b}_t \in \mathbb{R}^D$ are both sparse block matrices with S blocks. \mathbf{L}_{ss} , $\mathbf{L}_{ss'}$ and \mathbf{b}_{st} are defined as,

$$\begin{cases} \mathbf{L}_{ss} &= \frac{1}{NS^2} \mathbf{X}_s^T \mathbf{X}_s + \frac{\mu(S-1)}{N} \mathbf{X}_s^T \mathbf{X}_s + \lambda \mathbf{Q}_{st}, \\ \mathbf{L}_{ss'} &= \frac{1}{NS^2} \mathbf{X}_s^T \mathbf{X}_{s'} - \frac{\mu}{N} \mathbf{X}_s^T \mathbf{X}_{s'}, \\ \mathbf{b}_{st} &= \frac{1}{NS} \mathbf{X}_s^T \mathbf{y}_t. \end{cases} \quad (4.15)$$

According to the definition of positive-definite matrix, \mathbf{L}_t can be easily proven to be positive definite and invertible. Then we can derive the closed-form solution of \mathbf{w}_t as follows,

$$\mathbf{w}_t = \mathbf{L}_t^{-1} \mathbf{b}_t. \quad (4.16)$$

Furthermore, we notice that \mathbf{w}_t can be computed individually, which saves considerable space and time cost. On the other hand, we optimize $q_{s,d,v}$ according to Eqn. (4.9) with fixed \mathbf{w}_t .

4.3.4 Construction of Interest Tree Structure

We aim to employ the hierarchical agglomerative clustering algorithm to construct the tree structure. One challenge is that an interest is usually represented by a single concept, which makes it hard to measure the similarities among interests

and apply the hierarchical agglomerative clustering algorithm. Towards this end, two types of prior knowledge are utilized.

1) **External source.** We exploit an external source—the Web, where a huge amount of prior knowledge about interests is encoded implicitly. We transform each interest into a query and submit it to Google search engine. We collect the top 10 webpages, and then employ the library of BoilerPipe³ [67] to extract clean main contents from the returned webpages. Therefore, each interest can be represented by a document, based on which the Bag-of-words model [87] with TF-IDF term weighting scheme [104] can be applied and the similarities among interests can be evaluated.

2) **Internal source.** Although the external source provides us the general prior knowledge, we believe that the internal prior knowledge stored in our dataset also plays a vital role in user interest inference. Driven by this consideration, we propose to measure the similarities among interests based on their co-occurrence in users’ LinkedIn profiles in our dataset⁴. It deserves attention that we exploit all available LinkedIn profiles that exhibit users’ personal interests rather than that of the subset of users selected for the task of interest inference. Suppose we have a set of interests $\mathcal{I} = \{In_1, In_2, \dots, In_T\}$, and a set of documents $\mathcal{DD} = \{d_1, d_2, \dots, d_N\}$, where d_l contains all interests of user l . Let $c(j, k, l) = 1$ if and only if interests In_j and In_k both occur in d_l , and otherwise $c(j, k, l) = 0$. Then the co-occurrence matrix \mathbf{H} is defined as follows,

$$H(j, k) = \begin{cases} \frac{\sum_l c(j, k, l)}{\sum_j \sum_l c(j, k, l)} & \text{if } j \neq k; \\ 1 & \text{otherwise.} \end{cases} \quad (4.17)$$

Each row of \mathbf{H} corresponds to the co-occurrence of an interest with others. Then we use the JensenShannon divergence [22] to measure the similarities among interests.

Then it is suggested to apply the hierarchical agglomerative clustering algo-

³<https://code.google.com/p/boilerpipe/>.

⁴Users may list a set of personal interests in their LinkedIn profiles.

rithm on these enriched interests and build the tree structure. To assign appropriate weights to nodes, we choose to utilize the normalized height h_v of subtree rooted at node v to characterize its weight e_v , where $e_v = 1 - h_v$. Such assignment guarantees the aforementioned condition that the higher node corresponds to the weaker relatedness. It is worth noting that we normalize the heights for all nodes such that the root node is at height 1. We thus derive two models SM^2L-e and SM^2L-i based on two types of prior knowledge, respectively.

4.3.5 Complexity Discussion

To analyze the complexity of SM^2L , we need to solve the time cost in terms of constructing \mathbf{Q} , \mathbf{L}_t and \mathbf{b}_t , defined in Eqn. (4.11) and Eqn. (4.15), as well as computing the inverse of \mathbf{L}_t . Assuming $D \gg S$, the construction of diagonal matrix \mathbf{Q} has a time complexity of $O(DT)$, and the construction of matrix \mathbf{L}_t has a time complexity of $O(ND^2)$. Due to the fact that the time cost of matrix multiplication $\mathbf{X}_s^T \mathbf{X}_{s'}$ and that of constructing \mathbf{b}_t involved in Eqn. (4.15) remain the same for all iterations and \mathbf{L}_t is symmetric, we can reduce the practical time consumption remarkably. In addition, computing the inverse of \mathbf{L}_t has the complexity of $O(D^3)$ by the standard method. Then the total complexity should be $O(D^3T)$. We notice that the speed bottleneck lies in the number of features and the number of tasks instead of the number of data samples. As D is usually small, SM^2L should be computationally efficient.

4.4 Experiments

In this work, we cast the problem of user interest inference as the structure constrained multi-source multi-task learning problem. In particular, we explored four popular social networks: Twitter, Facebook, Quora and LinkedIn, where the fea-

tures were extracted from the first three sources and the ground truth was constructed based on the last one.

4.4.1 Dataset Construction

To construct the benchmark dataset, we need to first tackle the problem of “social account alignment”, which aims to identify the same users across different social networks by linking their multiple social accounts [3]. To accurately establish this mapping, we employed the emerging social service—Quora, which encourages users to explicitly list their multiple social accounts in their Quora profiles⁵. We collected candidates from Quora by the breadth-first-search method. In the end, we harvested 172,235 Quora user profiles and only retained those who provided their Facebook, Twitter and LinkedIn accounts in their Quora profiles. Based on these mappings, we launched a crawler to collect their historical social contents, including their basic profiles, social posts and relations. To build the ground truth, we employed the

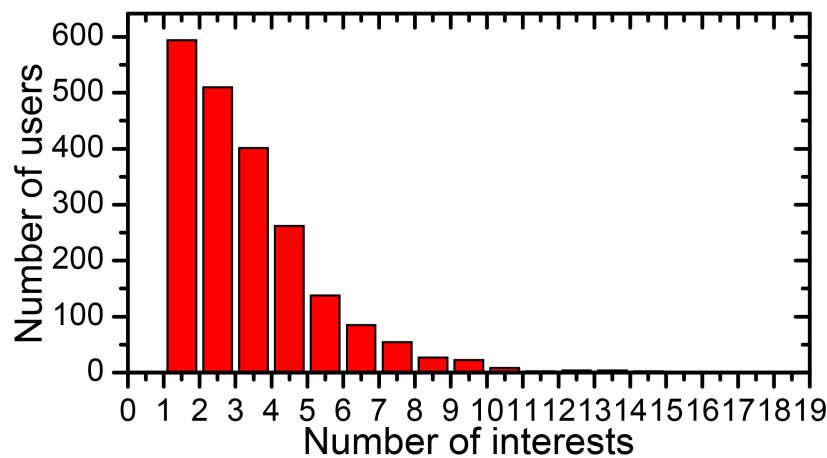


Figure 4.2: Distribution of user frequency distribution with respect to the number of interests over our dataset.

structural information of users’ LinkedIn profiles: “Additional Information”, which usually contains information about users’ personal interests. Users’ interests listed

⁵One representative example can be seen via <https://www.quora.com/Martijn-Sjoorda>.

in their LinkedIn profiles are usually represented by phrases, separated by commas, which facilitates the ground truth construction to a large extent. To obtain the representative interests, we filtered out the interests that are liked by less than 15 users. Finally, we obtained 74 interests⁶. Then we only retained those users who expressed these interests in their LinkedIn profiles and obtained 1,607 users ultimately. Figure 4.2 shows the user frequency distribution with respect to the number of interests over our dataset. The average number of users’ interests is 2.9. In addition, Figure 4.3 shows the detailed user distribution of each interest in our dataset. As we can see, some interests gain more users, while other interests such as ‘cricket’ and ‘open source’ have limited fans.

4.4.2 Feature Extraction

To informatively describe users, we extracted two kinds of features: user topics and contextual topics.

User topics. We explored the topic distributions of users’ social posts to infer users’ interests. We generated topic distributions using the LDA model, which has been widely found to be useful in latent topic modeling [31, 60]. Based on perplexity [73], we ultimately obtained 89, 24, 119 dimensional topic-level features respectively over users’ Twitter⁷, Facebook⁸ and Quora⁹ data.

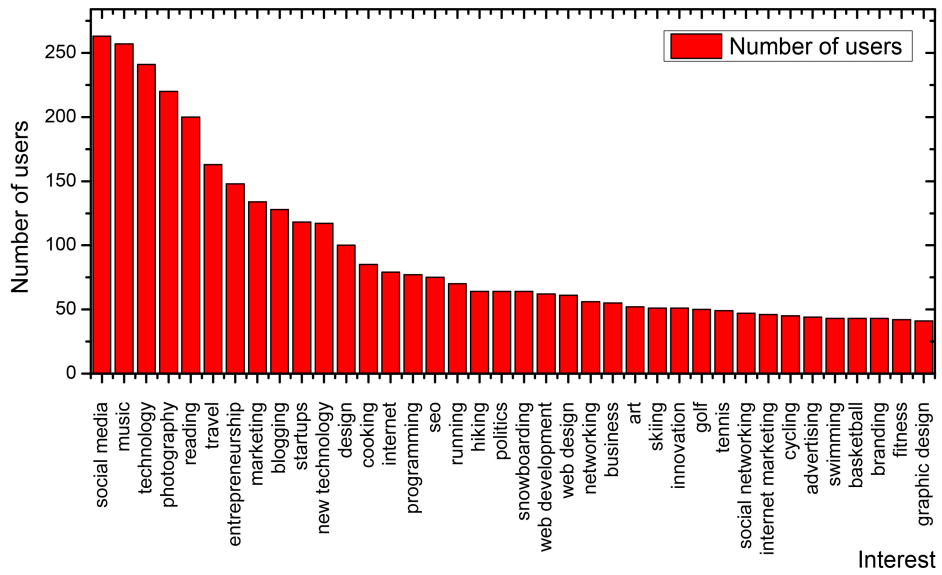
Contextual topics. We define users’ contextual topics as the topics of users’ connections. As it goes that “birds of a feather flock together”, we believe that the contextual topics intuitively reflect the contexts of users and further disclose users’ interests. Particularly, we studied followee connections in Twitter because of their intuitive reflection of topics that users are concerned with. As the bio descriptions

⁶These interests are available at <http://msmt.farbox.com/>.

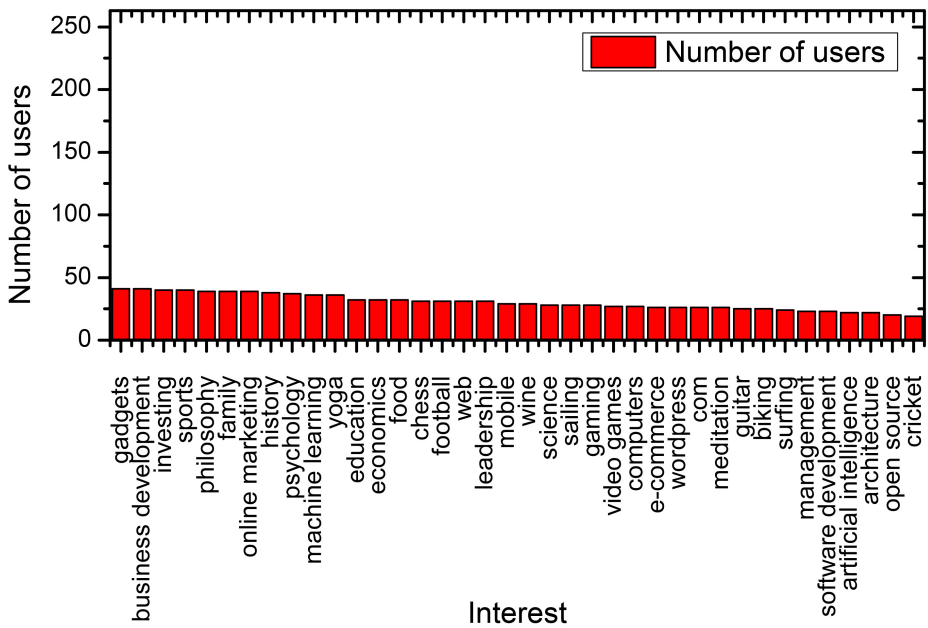
⁷Users’ Twitter data refers to users’ historical tweets.

⁸Users’ Facebook data refers to users’ historical timelines.

⁹Users’ Quora data refers to users’ historical questions and answers.



(a) Part: I



(b) Part: II

Figure 4.3: User distribution of each interest in our dataset. Due to the limited space, we separate the distribution into two parts according to the number of users.

are usually provided by users to briefly express themselves and may indicate users' summarized interests, we merged the bio descriptions of a user's followees into a document, on which we further applied the LDA model. We utilized the perplexity to tune the dimensions of topic-level features over these bio documents and obtained a 64 dimensional feature space. In this work, we only explored the contextual topics in Twitter, since the bio descriptions are usually missing in Facebook and Quora.

4.4.3 On Tree Construction

We list the top interest-pairs based on the tree constructed by different sources in Table 4.1. As can be seen, in general, the interest-pairs obtained from internal sources are more fine-grained than that from external sources. To a certain extent, this shows the superiority of using internal source over external source. In addition, after checking the webpages returned by the search engine, we found that the reason why 'Food' is close to 'Economics' lies in the top returned webpage of food is its Wikipedia page, which talks a lot about the commercial trade of food. We also found that the pages of 'Chess' and 'Cycling', in a sense, both mention words like 'competitors', 'rule' and 'improvement' frequently. This shows that taking advantage of the external search engines may not be appropriate to characterize the relatedness among interests.

4.4.4 On Evaluation Metrics

For the task of user interest inference, precision is of more importance as compared to recall. We thus validated our scheme via two metrics: $S@K$ and $P@K$.

S@K stands for the mean probability that a correct interest is captured within the top K recommended interests.

P@K is the proportion of the top K recommended interests that are correct.

Table 4.1: Top interest-pairs based on the tree constructed by the external source and internal source, respectively.

No.	Interest Pairs	
	Internal Source	External Source
1	<i>Surfing and Sailing</i>	<i>Food and Economics</i>
2	<i>Guitar and Chess</i>	<i>Chess and Cycling</i>
3	<i>Computer and Gadgets</i>	<i>Computer and Video games</i>
4	<i>Family and Fitness</i>	<i>Guitar and Reading</i>
5	<i>Open Source and Gaming</i>	<i>Networking and Startups</i>
6	<i>E-commerce and Business development</i>	<i>Technology and Web development</i>

4.4.5 On Model Comparison

We compared SM^2L with the following five baselines.

SVM: The first baseline is a traditional single source single task learning method—SVM [32], which simply concatenates the features generated from different sources into a single feature vector and learns each task individually. We chose the learning formulation with the kernel of the radial-basis function, implemented based on LIBSVM [27].

RLS: The second baseline is the regularized least squares (RLS) model [65], which also learns each task individually and aims to minimize the objective function of $\frac{1}{2N} \left\| \mathbf{y}_t - \sum_{s=1}^S \frac{1}{S} \mathbf{X}_s \mathbf{w}_{st} \right\|^2 + \frac{\lambda}{2} \left\| \mathbf{w}_t \right\|^2$.

regMVMT: The third baseline is the regularized multi-view multi-task learning model, introduced in [140]. This model regulates both the source consistency and the task relatedness. However, it simply assumes the uniform relatedness among tasks.

SM²L-eu: The fourth baseline is a derivation of SM^2L-e . This method constructs the tree structure based on external source in the same manner as SM^2L-e but assigns uniform weights to all nodes.

SM²L-iu: The fifth baseline is a derivation of SM^2L-i , which constructs the

tree structure using internal source but weights all nodes uniformly.

We adopted the grid search strategy to determine the optimal values for the regularization parameters among the values $\{10^r : r \in \{-12, \dots, -1\}\}$. Experimental results reported in this work are the average values over 10-fold cross validation.

Table 4.2: Performance comparison among various models.

Approaches	$P@1$ (%)	$S@10$ (%)
<i>SVM</i>	8.69	54.69
<i>RLS</i>	24.32	73.86
<i>regMVMT</i>	24.69	74.54
SM^2L - <i>eu</i>	25.50	73.80
SM^2L - <i>iu</i>	24.56	74.11
SM^2L - <i>e</i>	25.72	74.57
SM^2L - <i>i</i>	26.50	74.85

Table 4.2 shows the performance comparison between baselines and our proposed scheme. We observed that SM^2L -*i* and SM^2L -*e* both outperform the single source single task learning *SVM* and *RLS*. This verifies the significance of considering source consistency and task relatedness simultaneously. Moreover, it is not unexpected that *SVM* achieves the worst performance. A possible explanation might be the insufficient positive training samples for certain interests. For example, only 24 positive training samples are available for the interest “surfing”. In addition, the less satisfactory performance of *regMVMT*, as compared to SM^2L -*i* and SM^2L -*e*, confirms that it is advisable to characterize the task relatedness in a tree structure instead of correlating all tasks uniformly. Besides, SM^2L -*i* and SM^2L -*e* show superiority over SM^2L -*iu* and SM^2L -*eu* respectively, which enables us to draw a conclusion that modeling the relatedness strength among tasks merits our particular attention. Last but not least, SM^2L -*i* performs better than SM^2L -*e*. This finding demonstrates the importance of prior knowledge extracted from our internal source.

Based on the practical results, the time complexity of *regMVMT* is remarkably higher than that of *SM²L*. In particular, *regMVMT* costs about 562 seconds to execute, 114 times of that taken by *SM²L* for each iteration. This is mainly attributed to the computation of the inverse of a matrix with the dimension of *DT*, which requires a time complexity of $O(D^3T^3)$. Compared to *SM²L*, it is rather time consuming using *regMVMT*.

4.4.6 On Source Comparison

To shed light on the descriptiveness of multiple social network integration, we conducted experiments over various source combinations.

Table 4.3 shows the performance of *SM²L-i* over individual social network and their various combinations. We noted that the more sources we incorporate, the better the performance we can achieve. This suggests the complementary relationships instead of mutual conflicting relationships among the sources. Moreover, we found that aggregating data from all these three social networks can achieve better performance as compared to each of the single source. Interestingly, we observed that *SM²L* over Twitter alone achieves a much better performance, as compared to that using Quora or Facebook alone. This may be caused by that we additionally extracted contextual topics apart from user topics in Twitter, which can reveal users' interests more directly. It is comprehensible that *SM²L* would degenerate to multi-task learning when the context problem involves only one single source.

4.5 Summary

This chapter presented a structure-constrained multi-source multi-task learning scheme in the context of user interest inference. In particular, this scheme takes both the source consistency and the tree-guided task relatedness into considera-

Table 4.3: Contribution of individual social network and their various combinations.

Social network combinations	$P@1$ (%)	$S@10$ (%)
Twitter	24.75	73.05
Facebook	19.59	69.74
Quora	20.97	68.19
Twitter+Facebook	25.51	74.98
Twitter+Quora	24.89	74.41
Facebook+Quora	22.52	71.80
Twitter+Facebook+Quora	26.50	74.85

tion by introducing two regularizations to the objective function. Moreover, the proposed model is able to effectively select the task-sharing features and task-specific features by employing the weighted group lasso. Notably, the weights can be learned from two kinds of prior knowledge: external source and internal source. Experimental results demonstrated the effectiveness of our proposed scheme.

Chapter 5

A Personal Privacy Detection Framework

5.1 Introduction

Apparently our previous work on user profiling across multiple social networks has confirmed the potential of social networks in user attribute inference. On one hand, it can facilitate many applications as aforementioned. On the other hand, however, it also puts users at high risks on privacy leakage. It is reported that 66% of users' micro-posts are about themselves [58]. Moreover, due to the complicated social connections of users, ranging from close friends to strangers, users are much easier than ever before to leak their personal information to inappropriate audience. Consequently, privacy leakage via UGC in social networks deserves our special attention. In fact, according to the report [99], 50% of Internet users are concerned with the information disclosed about themselves online, up from about 30% in 2009. Therefore, it is highly desired to detect users' privacy leakage on social media to facilitate the corresponding prescription actions, such as gentle alerting to users when they are tweeting.

However, privacy leakage detection is non-trivial due to the following reasons. First, posts in social media may explicitly or implicitly convey different aspects of users. These aspects are usually not independent but can be organized into certain structures, such as groups, according to their relatedness. For example, given a set of aspects $\mathcal{I} = \{age, current\ location, places\ planning\ to\ go\}$, aspects “current location” and “places to go” are more correlated and should be learned together in one group. More often than not, such structure can impose certain constraints to the feature space and enhance the performance of aspect detection. Consequently, the main challenge is how to construct and leverage such structure to learn shared features and specific features. Another challenge lies in the lack of benchmark dataset and the way to extract a set of privacy-oriented features. This is because it is hard to distinguish personal posts from non-personal posts and some posts are too short to provide sufficient contexts for feature extraction. To address the aforementioned challenges, we present a novel scheme for privacy leakage detection, comprising of two components: description and prediction. As illustrated in Figure 5.1, in the first component, we pre-define a comprehensive taxonomy composed of 32 categories, where each category corresponds to one personal aspect of users. To build a benchmark dataset, we then feed a list of keywords to Twitter Search Service¹ for each category. A set of privacy-oriented features, including linguistic and meta features are extracted to describe the given UGC. We choose the real-time sharing website Twitter as the study platform due to the following facts: 1) users in Twitter are keen to share their personal events on various topics; and 2) the followers are broadly mixed and disorderly. Based on these features, the second component then endeavors to discover which personal aspect has been uncovered by the given post. The pre-defined structure in the first component has organized the 32 categories into eight groups, spanning from personal attributes to

¹<https://twitter.com/search-home>

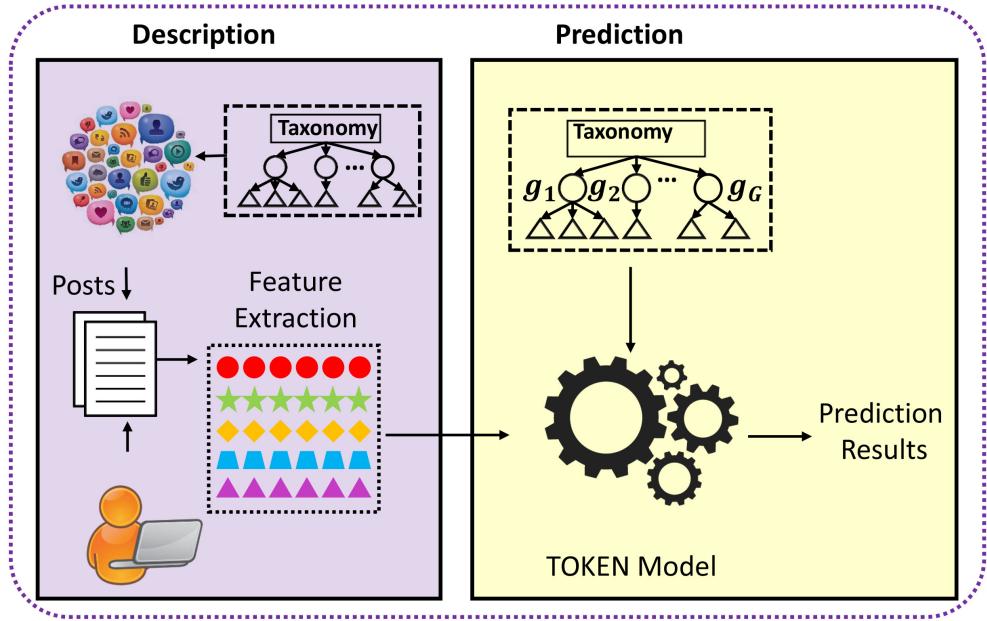


Figure 5.1: Illustration of the proposed scheme for privacy leakage detection. In the first component, we build a comprehensive taxonomy of the personal aspects, collect a benchmark dataset and extract a rich set of features to describe the UGC. The second component presents a taxonomy-constrained model to detect whether the given post leaks certain personal aspects.

life milestones. The categories within each group hold both group-sharing features and aspect-specific features. Meanwhile, we assume that there exists a low dimensional latent feature space that is capable of capturing the higher-level semantics of UGC as compared to the original features. To learn the latent feature space and further boost the aspect detection performance, we treat each personal aspect as a task and propose a Latent Group Multi-Task Learning (LG-MTL) model that is able to leverage the pre-defined structure to learn latent group-sharing features and aspect-specific features simultaneously.

Our main contributions can be summarized in threefold:

- We established a taxonomy to comprehensively characterize users' personal aspects, which consists of 32 categories under eight groups.
- Guided by this taxonomy, we proposed a LG-MTL model to uncover the

personal aspects disclosed by the given posts. The model is capable of learning both latent group-sharing and aspect-specific features simultaneously. We theoretically relaxed the non-smooth model to a smooth one and derived its closed-form solution.

- We collected a representative dataset via Twitter Search Service and developed a rich set of privacy-oriented features. We have released such data to facilitate others to repeat our experiments and verify their own ideas².

The remainder of this paper is structured as follows, Section 2 briefly reviews the related work. Sections 3, and 4 present the description and prediction components of LG-MTL model, respectively. Section 6 details the experimental results and analysis, followed by our concluding remarks and future work in Section 7.

5.2 Related Work

Privacy leakage detection and multi-task learning are related to this work.

5.2.1 Privacy

In the past decades, great efforts have been dedicated to privacy study, including data mining domain [51, 64, 78], and social media domain [126, 142, 147]. In particular, existing work investigating the privacy from the perspective of social media can be broadly divided into two directions [50, 79, 83, 114]. One is investigating privacy issues from structured data, such as users' structured profiles, and their privacy settings. Song et al. [114] studied the re-identification problem from users' trajectory records with a human mobility dataset. Besides, Liu et al. [79] proposed a framework for computing privacy scores for users in online social networks based on the sensitivity and visibility of certain profile items. Han et al. [50] further

²http://sigir16_privacy.farbox.com/

studied in-depth the privacy issues in people search by simulating different privacy settings in a public social network. In spite of the compelling success achieved by these works with different scenarios, far too little attention has been paid to investigate users' unstructured data, whereby the data volume is larger, information is richer, and privacy issues are more prominent, as compared to structured data.

The other direction is learning privacy issues from unstructured data [126, 142], which mainly refers to UGC. Approaches following this direction usually focus on training effective classifiers to predict whether the given UGC is sensitive or not in terms of general or specific user aspects. Mao et al. [80] studied privacy leakage on Twitter by automatically detecting tweets about vacation plans, drunk tweets, and tweets about diseases. Caliskan et al. [23] proposed an approach to detecting sensitive content from Twitter users' timelines and associating each user with a privacy score. Although great success has been achieved, they overlooked the relatedness among personal aspects and fed data into traditional machine learning models, such as Naïve Bayes [87] and AdaBoost! [43]. To bridge this gap, we pre-define a comprehensive taxonomy to capture users' structural personal aspects and based on which we propose a novel multi-task learning method which considers the relatedness among different personal aspects. In fact, MTL has been applied to solve many problems, including social behavior prediction [41], image annotation [38], and web search [9]. However, to the best of our knowledge, limited efforts have been dedicated to applying MTL in the privacy domain, which is the major concern of our work.

5.3 Data and Description

In this section, we detail the procedures for taxonomy induction, data collection, ground truth construction, as well as feature extraction.

5.3.1 Taxonomy Induction

In fact, Caliskan et al. [23] introduced nine categories: location, medical, drug/alcohol, emotion, personal attacks, stereotyping, family or other associations, personal details, and personally identifiable information, for privacy detection. These categories are relatively coarse-grained and hence fail to provide more detailed privacy leakage. In addition, they overlooked the life milestones of individuals, which are also privacy related [34]. Therefore, in this work, we pre-defined a comprehensive taxonomy consisting of 32 fine-grained privacy categories. These categories correspond to users' various personal aspects from different perspectives. As shown in Figure 5.2, these categories can be organized into eight groups, namely, *personal attributes*, *relationship*, *activities*, *location*, *emotion*, *healthcare*, *life milestones* and *neutral statements*. Except the *neutral statements* group, categories in the other seven groups are all related to personal issues to some extent. It is noted that, in our work, the neutral statements refer to those social posts that tell nothing about the post owner with regard to personal aspects of the other seven personal groups. Consequently, based on this taxonomy, given a social post, we can categorize it to at least one category.

5.3.2 Data Collection

To build our benchmark dataset, considering that most of the users' private tweets are extremely sparse, we hence did not collect data follow the user-centric policy. Instead, we collected the social posts for each category in the pre-defined taxonomy by keywords, respectively. In particular, we leveraged Twitter Search Service. We initially compiled a list of seed keywords³ for each category and fed them to Twitter Search Service. In the light of this, we obtained 269,090 raw tweets. To

³These keywords for each category can be available via http://aaai17_privacy.farbox.com/.

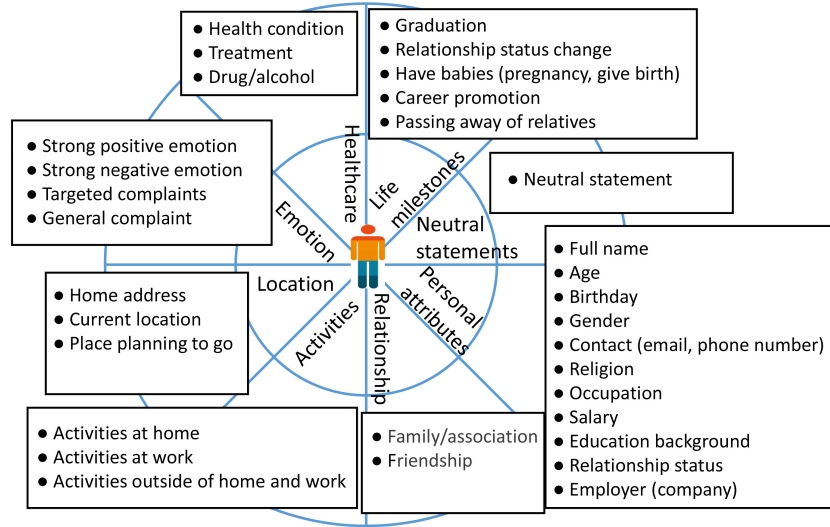


Figure 5.2: Illustration of our pre-defined taxonomy.

improve the quality of the dataset, we then developed several filter modules for different categories to remove the noise. We filtered out tweets that contain external URLs except those refer to users’ other social networks’ (e.g., Instagram) posts. In addition, as we studied the first-order privacy leakage, we ignored retweets in the dataset. Besides, we only retained tweets consisting of more than 50 characters. We ultimately obtained 11,370 tweets for all categories.

5.3.3 Ground Truth Construction

In our work, we constructed the ground truth about what has been revealed by a given post via AMT. We required workers to categorize each post into multiple categories. It is noted that we only focus on first-order privacy leakage. Particularly, we instructed the AMT workers to annotate a tweet as neutral if it reveals nothing about the tweet owner even it may refer to other people’s personal aspects. To ensure the quality of our ground truth, we only employed AMT masters instead of common workers. AMT masters achieve the “master” distinction by completing work requests with a high degree of accuracy. Moreover, we only accepted the

submissions whereby the workers labeled the privacy category correctly at 80% or above based on our sampling validation. To alleviate the problem of subjectivity, we employed three different workers for each post.

At last, we performed majority voting to establish the final labels for each post and obtained 11,368 labeled posts. To uncover insights of labeling quality, we used the Fleiss’ kappa statistic, a variant of Cohen’s kappa, to measure the inter-worker reliability. Considering that the number of categories assigned to each tweet varies, we treated such problem as a set of binary classification. For each binary classification, we counted the number of workers who assigned this category to the given tweet and those who did not. We finally got the average Fleiss’ kappa coefficient as 0.49, which shows a moderate agreement of our workers [70].

5.3.4 Features

To capture users’ personal leakage, we extracted a rich set of privacy-oriented features.

5.3.4.1 LIWC

Considering that users’ personality traits significantly affect their behaviors, including privacy perceptions [68], we adopted the LIWC feature to capture the sensitivity of a given UGC. Moreover, we noticed that the some categories in LIWC dictionary, such as “job” and “home”, just cover users’ personal aspects comprehensively.

5.3.4.2 Privacy Dictionary

The privacy dictionary [120] is a new linguistic resource for automated content analysis on privacy related texts. We believe that sensitive UGC should contain some representative privacy related keywords. We hence employed this dictionary to discriminate sensitive and non-sensitive UGC. This dictionary consists of eight cat-

egories⁴, derived from a wide range of privacy-sensitive empirical materials. With the help of this dictionary, we can generate similar output as LIWC.

5.3.4.3 Sentiment Analysis

Different personal aspects are frequently conveyed with different sentiments. For example, we observed that people usually broadcast their graduation and becoming parents in a more positive way, while describe their treatments in a more negative way. Inspired by this, we utilized the *Stanford NLP sentiment classifier*⁵ to judge tweets' polarity. In particular, we assigned each tweet with a value ranging from 0 to 4, corresponding to *very negative*, *negative*, *neutral*, *positive*, *very positive*.

5.3.4.4 Sentence2Vector

Considering the short-length nature of tweets, to perform content analysis, we employed the state-of-the-art textual feature extraction tool Sentence2Vector⁶. Sentence2Vector is developed based on the word embedding algorithm Word2Vector [86, 144], which has been found to be effective to alleviate the semantic problems of word sparseness [45]. Given a UGC, Word2Vector would project it to a fixed dimensional space, where similar words are encoded spatially. In our work, we treated each tweet as a sentence, and utilized the Sentence2Vector tool to generate the vector representation of each tweet.

5.3.4.5 Meta-features

Apart from the above linguistic features, we extracted several meta-features, which have also been verified to be effective in topic detection [115]. These features

⁴They are the Law, OpenVisible, OutcomeState, NormsRequisites, Restriction, NegativePrivacy, Intimacy, and PrivateSecret.

⁵<http://stanfordnlp.github.io/CoreNLP/>

⁶<https://github.com/klb3713/sentence2vec>

include the presence of hashtags, slang words, images, emojis⁷, and user mentions. In particular, to count the number of slang words, we constructed a local slang dictionary, which consists of 5,374 words by crawling the Internet Slang Dictionary & Translator. Moreover, we also incorporated the timestamp as an important feature, as we observed that users would post activities at work in the daytime while posting their drug/alcohol aspect in the evening. In particular, we just utilized posts’ created-time at the hour level.

5.4 Prediction

In this section, we detail the prediction component.

5.4.1 Notation

In our work, each task is aligned with one personal aspect, and we hence have $Q = 32$ tasks, which have been pre-organized into $G = 8$ groups, according to the proposed taxonomy. Meanwhile, we are given N users and each is represented by a D -dimensional vector. Let $\mathbf{X} \in \mathbb{R}^{N \times D}$ stand for the input matrix and $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_Q\} \in \mathbb{R}^{N \times Q}$ denote the corresponding label matrix, where $\mathbf{y}_q = \{y_1, y_2, \dots, y_N\}^T \in \{1, -1\}^N$ corresponds to the label vector for the q -th task.

5.4.2 Model Formulations

For each task, we can learn a predictive model, which is defined as follows,

$$\mathbf{f}_q(\mathbf{X}) = \mathbf{X}\mathbf{w}_q, \quad (5.1)$$

where $\mathbf{w}_q = (w_q^1, w_q^2, \dots, w_q^D)^T \in \mathbb{R}^D$ represents the linear mapping function for the q -th task. Let $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_Q\} \in \mathbb{R}^{D \times Q}$. We adopt the least square loss

⁷An emoji refers to a “picture character” to express facial expressions, concepts, activities and so on.

function to measure the errors,

$$L(\mathbf{W}) = \frac{1}{2N} \left\| \mathbf{Y} - \mathbf{X}\mathbf{W} \right\|_F^2, \quad (5.2)$$

where $\left\| \cdot \right\|_F$ denotes the Frobenius norm of matrix. $l_{2,1}$ -norm has been proven to be effective to select the relevant features for at least one task. In particular, the multi-task learning with $l_{2,1}$ -norm is defined as follows,

$$\Gamma = L(\mathbf{W}) + \frac{\beta}{2} \left\| \mathbf{W} \right\|_{2,1}, \quad (5.3)$$

where β is a nonnegative regularization parameter, $\left\| \mathbf{W} \right\|_{2,1} = \sum_{d=1}^D \left\| \mathbf{w}^d \right\|$ is the $l_{2,1}$ -norm of \mathbf{W} , $\mathbf{w}^d = (w_1^d, w_2^d, \dots, w_Q^d)$, and $\left\| \mathbf{w}^d \right\|$ represents the Euclidean norm of vector \mathbf{w}^d . The hidden assumption behind $l_{2,1}$ -norm is that all tasks are related and share the common set of relevant features. However, such assumption is not realistic and makes the multi-task learning not robust to the outlier tasks. Beyond that, as aforementioned, all the tasks in our work have been pre-organized into eight groups according to the proposed taxonomy. It is thus more reasonable to assume that tasks belonging to the same group would be more likely to share a common set of relevant features. For example, tasks “places planning to go” and “current location” belonging to the location group of the taxonomy may share a common set of location-relevant features. Let \mathcal{C}_g stand for the index set of tasks belonging to the g -th group and the diagonal matrix $\mathbf{V}_g \in \mathbb{R}^{Q \times Q}$ denote the corresponding group assignment. $\mathbf{V}_g(q, q) = 1$ if $q \in \mathcal{C}_g$, and 0 otherwise. Thereafter, the objective function in Eqn.(5.3) can be strengthened as,

$$\Gamma = L(\mathbf{W}) + \frac{\beta}{2} \sum_{g=1}^G \sum_{d=1}^D \left\| (\mathbf{W}\mathbf{V}_g)^d \right\|. \quad (5.4)$$

It is worth noting there exist two special cases. When the number of groups $G = 1$, where all tasks are learned jointly in one group, it reduces to the traditional

multi-task feature learning [7]. On the other hand, when $G = Q$, where all tasks are learned separately, it reduces to the traditional supervised machine learning. Besides, we also argue that tasks of the same group in the taxonomy may not share the common set of low-level relevant features but have a common set of high-level latent features. We assume that there are J , where $J \leq D$, latent features. Each task is defined as a linear combination of a subset of these latent features. Formally, let us define $\mathbf{W} = \mathbf{L}\mathbf{S}$, where $\mathbf{L} \in \mathbb{R}^{D \times J}$ and $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_Q\} \in \mathbb{R}^{J \times Q}$. Each column of \mathbf{L} stands for a latent feature, and each row of \mathbf{S} represents the linear weights of latent features. We hence impose the $l_{2,1}$ -norm on \mathbf{S} instead of \mathbf{W} to learn the group-sharing latent features. On the other hand, apart from the group-sharing latent features, we also assume each task should be related to a few specific latent features, which is implemented by the l_1 norm of \mathbf{S} . Putting them together, we have the following objective function Γ ,

$$\min_{\mathbf{L}, \mathbf{S}} L(\mathbf{L}, \mathbf{S}) + \frac{\beta}{2} \sum_{g=1}^G \sum_{j=1}^J \|(\mathbf{S}\mathbf{V}_g)^j\| + \frac{\gamma}{2} \|\mathbf{S}\|_1 + \frac{\mu}{2} \|\mathbf{L}\|_F^2, \quad (5.5)$$

where $\|\mathbf{S}\|_1$ is the entry-wise l_1 norm of matrix \mathbf{S} , while μ and γ are nonnegative regularization parameters.

5.4.3 Optimization

We adopt the alternative optimization strategy to solve \mathbf{S} and \mathbf{L} . In particular, we optimize one variable while fixing the other in each iteration. We keep this iterative procedure until the objective function converges.

5.4.3.1 Computing \mathbf{L} with \mathbf{S} fixed

We first fix \mathbf{S} and take derivative of the objective function with respect to \mathbf{L} . We have,

$$\frac{1}{N}\mathbf{X}^T\mathbf{X}\mathbf{L}\mathbf{S}\mathbf{S}^T + \mu\mathbf{L} = \frac{1}{N}\mathbf{X}^T\mathbf{Y}\mathbf{S}^T. \quad (5.6)$$

Inspired by the Lemma 4.3.1 in [56], we transform the above equation to the following linear system,

$$\begin{cases} \mathbf{A}\mathbf{Vec}(\mathbf{L}) = \mathbf{B}, \\ \mathbf{A} = [\frac{1}{N}\mathbf{S}\mathbf{S}^T \otimes \mathbf{X}^T\mathbf{X} + \mu\mathbf{I}], \\ \mathbf{B} = \mathbf{Vec}(\frac{1}{N}\mathbf{X}^T\mathbf{Y}\mathbf{S}^T), \end{cases} \quad (5.7)$$

where \otimes denotes the Kronecker product, $\mathbf{I} \in \mathbb{R}^{(D \times J) \times (D \times J)}$ is an identity matrix, and $\mathbf{Vec}(\cdot)$ stands for stacking columns of a matrix into a single column vector. It is easy to prove that \mathbf{A} is always positive definite [56] and invertible.

5.4.3.2 Computing \mathbf{S} with \mathbf{L} fixed

Fixing \mathbf{L} to optimize \mathbf{S} , we encounter two non-smooth terms, $l_{2,1}$ -norm and l_1 norm, which are intractable to solve directly. To convert the $l_{2,1}$ -norm, we resort to another variational formulation [7, 113] of the $l_{2,1}$ -norm in Eqn.(5.5) as follows,

$$\Gamma = L(\mathbf{L}, \mathbf{S}) + \frac{\beta}{2} \left(\sum_{g=1}^G \sum_{j=1}^J \left\| (\mathbf{S}\mathbf{V}_g)^j \right\| \right)^2 + \frac{\gamma}{2} \left\| \mathbf{S} \right\|_1. \quad (5.8)$$

According to the Cauchy-Schwarz inequality, given an arbitrary vector $\mathbf{b} \in \mathbb{R}^M$ such that $\mathbf{b} \neq \mathbf{0}$, we have,

$$\begin{aligned} \sum_{i=1}^M |b_i| &= \sum_{i=1}^M \theta_i^{\frac{1}{2}} \theta_i^{-\frac{1}{2}} |b_i| \\ &\leq \left(\sum_{i=1}^M \theta_i \right)^{\frac{1}{2}} \left(\sum_{i=1}^M \theta_i^{-1} b_i^2 \right)^{\frac{1}{2}} = \left(\sum_{i=1}^M \theta_i^{-1} b_i^2 \right)^{\frac{1}{2}}, \end{aligned} \quad (5.9)$$

where θ_i 's are introduced variables that should satisfy $\sum_{i=1}^M \theta_i = 1$, and $\theta_i > 0$. The equality holds for $\theta_i = |b_i| / \|\mathbf{b}\|_1$. Based on this, we derive the following inequality,

$$\begin{aligned} \left(\sum_{g=1}^G \sum_{j=1}^J \|(\mathbf{S}\mathbf{V}_g)^j\| \right)^2 &\leq \sum_{g=1}^G \frac{\left(\sum_{j=1}^J \|(\mathbf{S}\mathbf{V}_g)^j\| \right)^2}{\theta_k} \\ &\leq \sum_{g=1}^G \sum_{j=1}^J \frac{\|(\mathbf{S}\mathbf{V}_g)^j\|^2}{\theta_{k,g}}, \end{aligned} \quad (5.10)$$

where we introduce the variable $\theta_{k,g}$. The equality can be attained if $\theta_{k,g}$ satisfies that,

$$\theta_{k,g} = \frac{\|(\mathbf{S}\mathbf{V}_g)^j\|}{\sum_{g=1}^G \sum_{j=1}^J \|(\mathbf{S}\mathbf{V}_g)^j\|}. \quad (5.11)$$

Consequently, fixing \mathbf{L} and minimizing Γ is equivalent to minimizing the following convex objective function,

$$\Gamma = L(\mathbf{L}, \mathbf{S}) + \frac{\beta}{2} \sum_{q=1}^Q \sum_{j=1}^J \frac{\|(\mathbf{S}\mathbf{V}_g)^j\|^2}{\theta_{k,g}} + \frac{\gamma}{2} \|\mathbf{S}\|_1. \quad (5.12)$$

To facilitate the computation of the derivative of objective function Γ with respect to \mathbf{S} , we define a diagonal matrix $\Theta_g \in \mathbb{R}^{J \times J}$ as follows,

$$\Theta_g(j, j) = \frac{1}{\theta_{j,g}}. \quad (5.13)$$

The final objective function Γ can be rewritten as follows,

$$\Gamma = L(\mathbf{X}, \mathbf{Y}) + \frac{\beta}{2} \sum_{g=1}^G \text{tr} \left[(\mathbf{S}\mathbf{V}_g)^T \Theta_g \mathbf{S}\mathbf{V}_g \right] + \frac{\gamma}{2} \|\mathbf{S}\|_1. \quad (5.14)$$

where $\text{tr}(\mathbf{A})$ is the trace of matrix \mathbf{A} . To optimize the L_1 norm, we use the fast iterative shrinkage-thresholding algorithm (FISTA) [14] as follows,

$$\begin{cases} \Gamma_q &= h(\mathbf{s}_q) + p(\mathbf{s}_q), \\ h(\mathbf{s}_q) &= L(\mathbf{L}, \mathbf{s}_q) + \frac{\beta}{2} \sum_{q \in \mathcal{C}_g} \text{tr}(\mathbf{s}_q^T \Theta_g \mathbf{s}_q), \\ p(\mathbf{s}_q) &= \frac{\gamma}{2} \|\mathbf{s}_q\|_1. \end{cases} \quad (5.15)$$

The key iteration step of FISTA is to calculate $s_q^{(k)}$ by minimizing the following function,

$$\min_{\mathbf{s}_q} \left\{ p(\mathbf{s}_q) + \frac{R_q^{(k)}}{2} \left\| \mathbf{s}_q - \left(\mathbf{z}_q^{(k)} - \frac{1}{R_q^{(k)}} \nabla h(\mathbf{z}_q^{(k)}) \right) \right\|_F^2 \right\}, \quad (5.16)$$

where $R_q^{(k)}$ is the Lipschitz constant of $\nabla h(\mathbf{s}_q)$, $\mathbf{z}_q^{(k)}$ is a linear combination of $\mathbf{s}_q^{(k-1)}$ and $\mathbf{s}_q^{(k-2)}$, and $\nabla h(\mathbf{s}_q)$ is,

$$\nabla h(\mathbf{s}_q) = \frac{1}{N} \mathbf{L}^T \mathbf{X}^T (\mathbf{X} \mathbf{L} \mathbf{s}_q - \mathbf{y}_q) + \beta \sum_{g \in \mathcal{C}_g} \Theta_g \mathbf{s}_q. \quad (5.17)$$

We solve Eqn.(5.16) by the following soft-threshold step,

$$\mathbf{s}_q^{(k)} = \mathcal{T}_{\frac{\gamma}{2R_q^{(k)}}}(\mathbf{e}_q) = \max(0, 1 - \frac{\gamma/2R_q^{(k)}}{\|\mathbf{e}_q\|_1}) \mathbf{e}_q, \quad (5.18)$$

where \mathcal{T} is a shrinkage operator [14] and \mathbf{e}_q is defined as,

$$\mathbf{e}_q = \mathbf{z}_q^{(k)} - \frac{1}{R_q^{(k)}} \nabla h(\mathbf{z}_q^{(k)}). \quad (5.19)$$

Based on the sub-multiplicative property of spectral norm, we easily derive that $\|\nabla h(\mathbf{s}_{q_1}) - \nabla h(\mathbf{s}_{q_2})\|$ equals to,

$$\begin{aligned} & \left\| \beta \sum_{g \in \mathcal{C}_g} \Theta_g (\mathbf{s}_{q_1} - \mathbf{s}_{q_2}) + \frac{1}{N} \mathbf{L}^T \mathbf{X}^T \mathbf{X} \mathbf{L} (\mathbf{s}_{q_1} - \mathbf{s}_{q_2}) \right\| \\ & \leq \left(\beta \sum_{g \in \mathcal{C}_g} \|\Theta_g\| + \frac{1}{N} \|\mathbf{L}^T \mathbf{X}^T \mathbf{X} \mathbf{L}\| \right) \|\mathbf{s}_{q_1} - \mathbf{s}_{q_2}\| \\ & \leq R_q \|\mathbf{s}_{q_1} - \mathbf{s}_{q_2}\|, \end{aligned} \quad (5.20)$$

whereby we enforce $R_q^{(1)} = R_q^{(2)} = \dots = R_q$, and $\|\cdot\|$ denotes the spectral norm of matrix as well of Euclidean norm of vector. As Θ_g and $\mathbf{L}^T \mathbf{X}^T \mathbf{X} \mathbf{L}$ are both positive-semidefinite matrices, simple algebra computation gives that,

$$R_q = \beta \sum_{g \in \mathcal{C}_g} \lambda_{\max}(\Theta_g) + \frac{1}{N} \lambda_{\max}(\mathbf{L}^T \mathbf{X}^T \mathbf{X} \mathbf{L}). \quad (5.21)$$

where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue of a matrix.

5.5 Experiments

In this section, we conducted extensive experiments to verify the effectiveness of our proposed scheme.

5.5.1 Data Preprocessing

To boost the performance of content-related features, we first sanitized the noisy tweets by following steps: 1) We removed the user mentions. 2) We replaced the Internet slangs with their corresponding formal expressions. To be more specific, we first constructed a local slang dictionary containing 5,374 words by crawling the Internet Slang Dictionary & Translator. Given a post, we then transformed each slang to their formal expression by looking up this dictionary. 3) We also performed lemmatization using *Stanford NLP tool* to link the word variants. And 4) we further corrected words that contain repeated sequential letters by removing the extra letters (e.g., “coooooool” was changed to “cool”).

5.5.2 Experimental Setting

For the task of privacy leakage detection, precision is more important than recall. We hence measured the proposed LG-MTL model and its competitors via two metrics: $S@K$ and $P@K$.

We employed the grid search strategy to obtain the optimal regularization parameters among the values $\{10^r : r \in \{-8, -7, \dots, 2, 3\}\}$ regarding $P@1$. Experimental results reported in this paper are the average values over 10-fold cross validation.

Table 5.1: Performance comparison of our LG-MTL model trained with different feature configurations (%).

Features	$S@1$	$S@3$	$S@5$	$P@3$	$P@5$	p -value
Privacy-dic	8.56	18.38	54.26	6.33	11.28	$5.9e-22$
Sentiment	30.48	52.23	63.10	17.44	13.32	$1.6e-20$
Meta-features	30.31	52.28	63.12	17.38	13.10	$9.9e-21$
Sentence2Vec	33.29	59.06	70.91	20.66	15.54	$2.0e-21$
LIWC	37.13	67.98	78.65	24.72	17.44	$3.1e-10$
Total	44.37	74.67	84.66	28.42	19.86	-

5.5.3 Evaluation of Description

To examine the discriminative features we extracted, we conducted experiments over different kinds of features using **LG-MTL**. In particular, we also performed the one-way analysis of variance to validate the effectiveness of all the features regarding $S@5$. Table 5.1 comparatively shows the performance of **LG-MTL** in terms of different feature configurations. Note that $S@1$ equals to $P@1$, and we thus exclude the column $P@1$ from the table. First, it can be seen that our model based on LIWC feature achieves the best performance, while the features extracted based on the privacy dictionary are the least powerful ones. This shows that users’ privacy is better characterized by the LIWC dictionary, as compared to the privacy dictionary. One possible explanation is that the 70 categories of LIWC dictionary, whose representative categories are listed in Table 5.2, capture users’ personal aspects more comprehensively. On the other hand, although the privacy dictionary

Table 5.2: Ten representative word categories in LIWC, that can capture the personal aspects comprehensively.

Category	Example	Category	Example
Home	apartment, family	Social	mate, child
Job	job, majors	Feeling	feels, touch
Money	audit, cash	Friends	buddy, friend
Biological processes	eat, blood, pain	Family	daughter, husband
Ingestion	dish, eat, pizza	Motion	arrive, car, go

is not much powerful when $K = 1$ and $K = 3$, its performance boosts sharply when K increases to 5. Second, we observed that the performance derived from Sentence2Vector features is also satisfactory. This verifies that the semantics of different personal aspects are usually different. Third, although meta-features only account for six dimensions and sentiment feature is only one-dimensional, they also yield compelling performance. In particular, we believe that the meta-feature timestamps (hour) of UGC should play an important role regarding privacy leakage detection. We thus had a close look at the comparison among the time distributions of several representative categories in Figure 5.3. As can be seen from Figures 3 (a), (b), and (c), categories related to activities show prominent temporal patterns. For example, tweets related to users’ activities at home reach peaks around 12pm and 20pm, while those related to users’ activities outside are more likely to be posted by users around 20pm. In addition, Figure 5.3(d) shows that users are more likely to post tweets revealing their drug/alcohol aspects. Moreover, to some extent, this also reflects the fact that users are more likely to get drunk after their activities outside. On the other hand, some categories related to users’ life milestones are more time-dependent (Figure 5.3(e) and Figure 5.3(g)) while others are not (Figure 5.3(f) and Figure 5.3(h)). For example, users would post that they become parents or they graduate at anytime, while users prefer to post their status change in the evening and post their relatives’ death after noon.

Apart from the timestamps, we also studied several other meta-features. Table 5.3 shows the top categories regarding the percentage of tweets that containing images. One reasonable explanation for these categories’ high rankings maybe due to the fact that most of them can be used to reflect what is going on, such as “current location”, “friendship”, “status change”, and “activities outside”. Moreover, users would like to take photos to record what is happening, such as, who they are with, where they are, and what event they are joining. Regarding the category

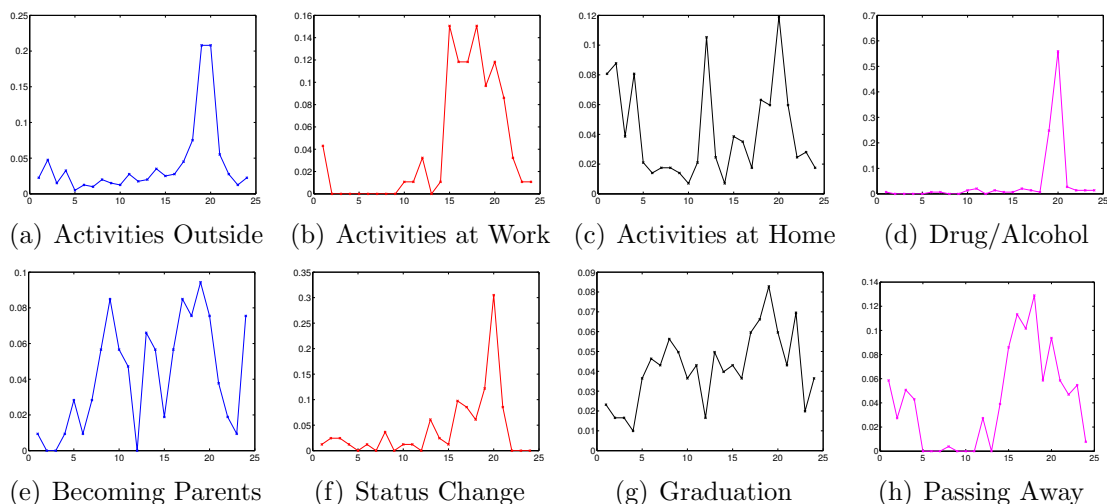


Figure 5.3: Illustration of temporal patterns regarding personal aspects. X axis: Time (Hour); Y axis: Temporal distribution of tweets.

Table 5.3: Top categories regarding the percentage of tweets that containing images.

Rank	Category	Percentage (%)
1	Current location	22
2	Friendship	21
3	Birthday	16
4	Positive emotion	15
5	Status change	13
6	Activities outside	11

“birthday”, it maybe because that users are more likely to hold birthday parties, receive presents, take photos to memorize and further upload to social media.

Table 5.4 shows the top categories regarding the percentage of tweets that containing user mentions. User mentions in a tweet are usually meant to directly reply certain user or to refer specific users who maybe related to the tweet. Tweets belonging to categories “contact”, “home address” and “full name” are much informative and tend to be replied to certain users, while categories “friendship” and “status change” are more likely to refer related users to the given tweets.

Last, we studied the sentiment feature and show the top categories with either positive or negative sentiment in Table 5.5. As can be seen, apart from the “positive emotion” category, categories “friendship”, “birthday” and “career pro-

Table 5.4: Top categories regarding the percentage of tweets that containing user mentions.

Rank	Category	Percentage (%)
1	Friendship	69
2	Status change	57
3	Contact	55
4	Home address	53
5	Full name	47
6	Current location	47

Table 5.5: Sentiment Ranking.

Rank	Positive	Negative
1	Positive emotion	Direct complaints
2	Friendships	General complaints
3	Birthday	Negative
4	Career promotion	Graduation
5	Employer	Treatment

motion” also have relatively positive sentiment. This can be explained by these categories are always associated with positive events. However, unexpectedly, category “employer” also has positive sentiment. After a careful check, we found that user can talk about their career promotion (e.g., “So happy I got promoted at...”), advertise for their company (e.g., “Hi everyone, please follow my company @CrossConnMedia! We hope to have some more exciting projects coming in the near future. #Diversity”) or broadcast their companies’ celebration parties (e.g., “Enjoying my company EnSiteUSA, Inc. Christmas Party!!”). On the other hand, category “graduation” which gets negative sentiment, attracts our attention. After a close look at the tweets, we found that users may feel worried about their future or miss their school life and friends. Therefore, they may tweet like “Shit I know I just graduatedbut for some real still feel empty” and “@Sierraa`Grace yes I just graduated. We all go through shit just mine gets thrown in my face all the time lol. I miss you to. Text me”.

5.5.4 Evaluation of Prediction

To verify the effectiveness of our proposed model, we compared **LG-MTL** with the following five baselines.

Pop_K: The first basic baseline utilizes the prior probability of each category and simply selects the K most common categories as the prediction results.

SVM: We chose the learning formulation with the kernel of the radial-basis function. We implemented this method with the help of LIBSVM [27].

MTL_Lasso: The third baseline is the multi-task learning with Lasso [119]. This model also does not take advantage of prior knowledge about task relatedness.

MTFL: The four baseline is the multi-task feature learning [7], which takes advantage of the group lasso to jointly learn features for different tasks. However, this model assumes that all tasks are relevant like organizing all tasks in a single group.

GO-MTL (without taxonomy): The five baseline is the grouping and overlap in multi-task learning proposed in [69]. This model does not take any advantage of the prior knowledge about tasks relation, as there is no taxonomy constructed to guide the learning. It is worth mentioning that we can derive **GO-MTL (without taxonomy)** from **LG-MTL** by making $\beta = 0$.

Table 5.6: Performance comparison between our LG-MTL model and the baselines in S@K and P@K (%).

Methods	$S@1$	$S@3$	$S@5$	$P@3$	$P@5$	p -value
Pop_K	30.63	52.68	63.41	17.59	13.39	$2.3e-20$
SVM	2.65	52.15	72.01	17.80	16.53	$2.3e-16$
MTL_Lasso	43.99	73.02	82.26	27.35	19.34	$6.9e-7$
MTFL	43.75	73.98	83.69	27.63	19.70	$3.1e-3$
GO-MTL	43.92	73.93	83.45	27.25	19.40	$2.9e-3$
LG-MTL	44.37	74.67	84.66	28.42	19.86	-

For each method mentioned above, the involved parameters were carefully tuned, and the parameters with the best performance in $S@5$ were used to report

Table 5.7: Examples of some categories.

Category	Examples
Occupation	“just got a job offer at an eye laser clinic debating if I should take it”
	“Working at plaza is gonna get me so much more money than what I get now I’m so excited!!”
	“I used to be a swimmer...now I’m a coach. And I love torturing my kids. #evilmutantswimcoach”
	“I felt more control of my work as a T. Even more patience is needed as a coach. ”
	“I’m Barry Bennett. I gave \$1000 to @user1. I live in Alexandria, VA. I’m a Consultant.”
Gender	“I seriously going to buy tacos, but the laziness took over. I am my father’s daughter. ”
	“My girlfriend broke up with me...”
	“@user2 I would disappear if my wife tried to grab MY prunes in the supermarket!”
	“The worst thing you do is piss me off while I’m on my period.”
Current location	“Get to stay in Washington DC tonight...too bad I have to sleep in the airport”
	“At the Bell Performing Arts Centre for the LTS Jazz Band Concert #sweet”
	“She told the doctor tomorrow is my birthday I can’t be in the hospital”
Place to go	“In exactly one month I will be headed to the airport to depart for Cambodia... #WhatIsLife”
	“Good morning friends..preparing for my trip to Sweden..im driving to Kiruna through Riksgrnsen and Abisko to Kiruna airport..”
	“Going to SF this weekend for the Beenzino concert! I can’t wait to get my picture with”
General complaint	“dude if you’re going to cough every 20 seconds in the library can u leave”
	“Sometimes being single sucks but then again I remember the reason why I’m single .”
	“being in a relationship is stressful i wanna take a nap”
Age	“It’s still sinking in how next month I’ll be 30.... Never married but feel damn near divorced and no kids. Wow.”
	“...when I told him I’m only 24”
	“Can it be June so I can be drunk off my ass in Vegas for my 21st birthday”
	“Hey @user3 its my birthday tomorrow. I am turning 12! ”
Neutral statement	“Chelsea look like they got promoted last season..”
	“Do you want my home address and social security too?”

the final comparison results. Table 5.6 shows the performance comparison between the baselines and our proposed **LG-MTL**. First, as we can see, the superiority of **LG-MTL** over **Pop_K** suggests that the prior probability of each category is not reliable due to the limited dataset. Second, we noticed that **LG-MTL** outperforms the single task learning **SVM**. This verifies that there do exist relationships among tasks. This also shows the superiority of our work over other similar privacy detection works [23, 80]. In particular, it is not unexpected that **SVM** achieves the worst performance. This may be due to insufficient positive training samples for certain categories. For example, there are only 52 positive training samples available for category “home address”. Multi-task learning is able to alleviate the unbalanced training sample problems by borrowing some samples from related tasks. In addition, **LG-MTL** shows superiority over **MTL_Lasso** and **MTFL**, respectively, which enables us to draw a conclusion that it is reasonable to learn tasks by groups, defined by the taxonomy. Besides, the less satisfactory performance of **GO-MTL**, as compared to **LG-MTL**, also demonstrates the importance to incorporate the prior grouping knowledge of tasks. Moreover, we also performed the one-way analysis of variance over the 10-fold cross validation and found that **LG-MTL** can significantly outperform the baselines regarding $S@5$.

5.5.5 Case Study

5.5.5.1 Example Study

In order to get a more intuitive understanding of each category, we had a close look at the content of each category. We listed several examples of selected categories in Table 5.7. We found that users’ occupations are mainly revealed by tweeting their new jobs, their feeling or understanding about their occupations, or just self-promotion. Users’ gender information can be embedded in their roles in relationships (e.g., daughter, wife.) or the distinct gender characteristic (e.g.,

Table 5.8: Keywords or phrases for each category.

Category	keywords or phrases
acoutside	playing badminton, going/go/went to, jogging, excited, playing, to see, with my, library, concert
acwork	my company, party, working at, company holiday
age	th birthday, yr/years old, just graduated
becomeparents	it is, a boy/girl, got pregnant, boy or girl
birthday	happy, birthday, will be, going to, got, thank you, th
careerpromotion	got promoted, just got, at work, my job
contact	contact me at, mobile/phone number, email address, send/call/reply me, please contact/reply, looking for
currentloc	just landed, live in, landed in, the airport, I just
directcomp	I hate, cannot, I need, trying to, have to
dragalcohol	get/getting/was/be drunk
education	just graduated, high school, a undergraduate/graduate/professor, bachelor/master degree, going/go to, college
employer	working at, my company, as a
family_asso	passed away, my brother/sister/dad/mum/daughter, love you
friendships	best/good friend, birthday, love
fullname	my full name, my nickname/name, call me, last name
gender	my period, aunt flo, my husband/wife, got pregnant
general_comp	do not, my period, have to, wish, feel like, hate, want to
graduation	graduated, high school, college, lol
healthcondition	got fever, my period, take medicine, got cough, aunt flo, have to, need to, see doctor, medicine, hospital
homeaddr	live in, home address, address is
negative emotion	passed away, do not, my period, cannot, hate, feel
neutral	it is, full name, to be, I think, I can, I will, I have
occupation	working/work at, as a, software developer, designer, writer, editor, photographer, nurse, consultant, artist, got promoted,
passaway	my dad/grandma/grandpa/mom/grandmother/father/uncle, cancer, passed away, I miss, thank you, cannot
placetogo	leave for, fly/going/go to, will, the airport, so excited, cannot wait
positive emotion	love, promoted, best, excited, thank you, happy, cannot wait
relationstatus	my husband/wife/boyfriend/bf, a relationship, broke up with, got married, a housewife, am single,
religion	Christian, Buddhist, agnostic, Jewish, Muslim, bible, lord, god
salary	I make/earn, as a, talkpay ⁸ , less than
statuschange	got divorced, got married, just got, got engaged, just broke, my life/husband/bf/boyfriend
treatment	take medicine, have to, cough, I hate, surgery, need

period for women.). In addition, users' current locations are usually discussed with sharing their current feelings or the current events they are joining, while users' places to go can be tweeted when they are preparing for the trips, or to express their eagerness to the their trips. As to general complaint, it is reasonable to see that frequent cough in the library and unsatisfactory relationship are likely to be complained. Users may mention their age more when their birthdays are coming. Last but not least, although the neutral statements may also talk about "career promotion", "full name" and "my home" and other personal aspects, they are usually revealing others' privacy or providing no detailed personal information. We further listed the representative keywords or phrases of each category in Table 5.8.

5.5.5.2 Failure Study

We found that there are some tweets that are not properly predicted, as shown in Table 5.9. The failure cases are roughly caused by three reasons. First, the semantics of certain tweets are not well characterized. One possible explanation is that the most effective features in our work, LIWC features, are extracted statistically. LIWC features rely on a dictionary and the count of words, especially the noun words. Consequently, it is not effective to cope with complex tweets, such as Tweets 4, 9, and 10. Second, as illustrated by Tweet 8, some categories are subtly correlated, such as 'careerpromotion' and 'occupation'. Therefore, it is hard to precisely predict tweets' categories. Third, the manual annotation is not reliable for certain tweets. Although we have employed three AMT masters and performed the majority voting, there still exist certain tweets, whose ground truth is still not reliable. In particular, for those tweets that revealing multiple personal aspect, AMT annotators may overlook certain weak aspects, such as 'positive emotion'.

Table 5.9: Poorly classified tweets.

No.	Tweet	Ground Truth	Predicted Results
1	“@user4 my Wife is pregnant and due in about 2 weeks in tired of the hospital too #thestrugglesreal least I don’t have to give birth”	gender, becomeparents.	family-association, relastatus, health condition.
2	“@user5 ARE THEY KIDDING IM BROKE UNTIL MY 18 TH BIRTHDAY ON 4TH OF JANUARY FUCK MY LIFE”	age, birthday.	negative emotion.
3	“2 years since I got my dick scar from working at McDonald’s x good memories x pic.twitter.com/B9Hii7wfbC”	employer.	occupation.
4	“Like I’ve wanted to become a nurse since I was 8 and now it’s happening.”	occupation	health condition, treatment.
5	“I guess I’ll go to the hospital to see what’s going on I don’t like the fact that my face & hands look like this”	placetogo.	health condition, acoutside, treatment.
6	“@user6 I was teaching but I got married and my husband doesn’t want me to work right now so im modeling full time finally !”	occupation.	family_asso, relastatus.
7	“Got first tattoo w/ my husband today! It’s a sketch I drew of my dog who passed away thank you @user7 pic.twitter.com/26VYRIhjFF”	relastatus, acoutside.	friendships, passaway.
8	“i just be so busy ever since I got promoted & on the weekends is when I try to catch up on sleep”	careerpromotion.	occupation, positive emotion.
9	“ I got horrible service last night. Little do they know, my company is paying the bill & I’m liable to leave a 30+% tip!!”	general complaint.	occupation, education.
10	“ - lol my nigga , we used to be in church acting a fool in high school you done grew all up”	religion.	education.

5.6 Summary

In this chapter, we studied the problem of privacy leakage detection by presenting a scheme, consists of two components: description and prediction. As to description, we built a comprehensive taxonomy, constructed a benchmark dataset, and developed a set of privacy-oriented features. Experimental results showed that LIWC and Sentence2Vector features are the most discriminative features regarding privacy leakage detection. Meanwhile, we found that the privacy leakage via UGC holds certain temporal patterns. Regarding prediction, we proposed a taxonomy-guided multi-task learning model to categorize social posts, which is able to learn both latent group-sharing and aspect-specific features simultaneously. Experimental results also verified the advantages of taking the proposed taxonomy into consideration in multi-task learning.

Chapter 6

Conclusions and Future Research

6.1 Conclusions

This thesis focused on investigating user profiling across multiple social networks. Considering that user profiling can be framed in either mono-task learning or multi-task learning scheme, based on the nature of user attributes to be inferred, this thesis first proposed two multi-source learning schemes: multi-source mono-task learning scheme and multi-source multi-task learning scheme, respectively.

This thesis first explores a multi-source mono-task learning scheme to infer users' attributes, such as volunteerism tendency, which involves a single task. The proposed scheme is able to model both the source confidence and source consistency. Considering that block missing data may exist, it also proposed a novel approach to fix the problem of missing data and feed the complete data to the proposed model. The data completion approach is closer in spirit of NMF. This thesis applies the proposed multi-source mono-task learning scheme to the application of user volunteerism tendency prediction. The experimental results enable us to draw the following conclusions. First, utilizing multiple social networks does promote the performance regarding the user profiling problem of volunteerism ten-

dency prediction. In other words, the more sources are effectively incorporated, the better performance can be achieved. Second, the information on multiple social networks are complementary to each other and characterize users' volunteerism tendency consistently. Third, it was also demonstrated that the correlations of different social networks with the task of volunteerism tendency prediction cannot be treated equally. Last but not least, among the three kinds of features characterizing users' volunteerism tendency, linguistic features are the most discriminative features regarding the volunteerism tendency prediction. This reveals that volunteerism tendency is better reflected by their social contents, including their own social posts and the self-descriptions of their social connections.

This thesis next develops a multi-source multi-task learning scheme to infer users' attributes, such as interest, which involves multiple related tasks. The proposed scheme takes jointly regularizes two important aspects: source consistency and task relatedness. Regarding the task relatedness, two kinds of prior knowledge are introduced: external knowledge and internal knowledge. These two kinds of knowledge are encoded by the external source such as the Web and our internal dataset, respectively. We practically applied the proposed multi-source multi-task learning scheme in the context of user interest inference. The proposed scheme shows superiority over other baselines regarding the user profiling application—user interest inference. This confirms to the significance of taking both the source consistency and task relatedness into consideration in the multi-source multi-task context. In addition, the internal knowledge is found to be more powerful as compared to external knowledge. This demonstrates the importance of prior knowledge extracted from our internal source.

In addition, noting users' high privacy risks on social media from the aforementioned work, this thesis further studies the problem and privacy leakage detection. Framing such a problem as a set of multiple binary classification, this thesis

proposes a novel learning scheme, which consists of two components: description and prediction. For the description component, we first pre-defined a comprehensive taxonomy, consisting of 32 subcategories under 8 categories (groups). According to such taxonomy, we then constructed a benchmark dataset, which consists of 11,370 tweets. In addition, we developed a set of privacy-oriented features. As for the prediction component, a taxonomy-guided multi-task learning model is proposed to categorize users' social posts, which is capable of learning both latent group-sharing and aspect-specific features simultaneously. Experimental results show the advantages of the proposed learning scheme over other baselines. Additionally, LIWC and Sentence2Vector features are found to be the most discriminative features for privacy leakage.

6.2 Future Work

This main limitation of this thesis is that the essential step for user profiling across multiple social networks—social account mapping, is not investigated deeply. This thesis only utilizes social services that encourage users to explicitly list their multiple social accounts on one profile, such as About.me and Quora, to obtain users with multiple social accounts. As a consequent, the set of users studied in this thesis, are relatively much more active than the average users, in that they are more likely to share their multiple social accounts publicly. In reality, however, the majority of online users are less active or are less inclined to use multiple social account management. Furthermore, several cautious users may not want others to link their multiple social accounts and thus protect themselves by intentionally keeping their multiple social accounts anonymous, by, for example, using different or obscure user names on different social networks. Therefore, further research is needed towards this end.

In addition, the schemes for the user interest inference or privacy detection

proposed in this thesis can only provide a ranking list of the label candidates. The number of labels that should be assigned to each sample still has not been handled well. In the future, we will investigate the practical problem of how to accurately determine the number of user interests or privacy leaks for each user or tweet.

Moreover, despite the value of UGC in facilitating user profiling, they can also place users at high privacy risks, which has thus far still remained largely untapped. Currently, we only propose a general framework for privacy preserving from the perspective of user profiling. In particular, we mainly focus on the general detection of privacy leakage but ignore the subjectivity of privacy. Therefore, considering that people usually hold different privacy perception, further efforts should be dedicated to the development of personalized privacy preserving technique. On the other hand, we only study the first-order privacy leakage, where privacy is usually revealed by user themselves. Nevertheless, users' privacy may sometime be revealed by others, which gives rise to the second-order leakage. In the future, we will extend our work towards this end.

Bibliography

- [1] A. Abdel-Hafez and Y. Xu. A survey of user modelling in social media web-sites. *Computer and Information Science*, 2013.
- [2] F. Abel, I. Celik, C. Hauff, L. Hollink, and G.-J. Houben. U-sem: Semantic enrichment, user modeling and mining of usage data on the social web. *arXiv preprint arXiv:1104.0126*, 2011.
- [3] F. Abel, E. Herder, G.-J. Houben, N. Henze, and D. Krause. Cross-system user modeling and personalization on the social web. *User Modeling and User-Adapted Interaction*, 23(2-3):169–209, 2013.
- [4] F. Abel, E. Herder, and D. Krause. Extraction of professional interests from social web profiles. In *Proceedings of the ACM Workshop of User Modeling, Adaptation and Personalization*, pages 1–6. ACM, 2011.
- [5] S. Adali and J. Golbeck. Predicting personality with social behavior. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, pages 302–309. IEEE Computer Society, 2012.
- [6] M. Akbari, X. Hu, L. Nie, and T. Chua. From tweets to wellness: Wellness event detection from twitter streams. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 87–93. AAAI Press, 2016.

- [7] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.
- [8] F. R. Bach. Consistency of the group lasso and multiple kernel learning. *The Journal of Machine Learning Research*, 9:1179–1225, 2008.
- [9] J. Bai, K. Zhou, G. Xue, H. Zha, G. Sun, B. Tseng, Z. Zheng, and Y. Chang. Multi-task learning for learning to rank in web search. In *Proceedings of the International ACM Conference on Information and Knowledge Management*, pages 1549–1552. ACM, 2009.
- [10] S. Bai, T. Zhu, and L. Cheng. Big-five personality prediction based on user behaviors at social network sites. *arXiv preprint arXiv:1204.4809*, 2012.
- [11] K. Balog and M. De Rijke. Determining expert profiles (with an application to expert finding). In *Proceedings of Joint Conference on Artificial Intelligence*, volume 7, pages 2657–2662, 2007.
- [12] M. R. Barrick and M. K. Mount. The big five personality dimensions and job performance: a meta-analysis. *Personnel Psychology*, 44(1):1–26, 1991.
- [13] B. Bazelli, A. Hindle, and E. Stroulia. On the personality traits of stackoverflow users. In *Proceedings of the IEEE International Conference on Software Maintenance*, pages 460–463. IEEE, 2013.
- [14] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.
- [15] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves. Detecting spammers and content promoters in online video social networks.

- In *Proceedings of the International ACM SIGIR Conference*, pages 620–627. ACM, 2009.
- [16] B. Bi, M. Shokouhi, M. Kosinski, and T. Graepel. Inferring the demographics of search users: social data meets search queries. In *Proceedings of the International Conference on World Wide Web*, pages 131–140. International World Wide Web Conferences Steering Committee, 2013.
- [17] M. Bilenko and M. Richardson. Predictive client-side profiles for personalized advertising. In *Proceedings of the International ACM SIGKDD Conference*, pages 413–421. ACM, 2011.
- [18] A. Binder, W. Samek, K.-R. Müller, and M. Kawanabe. Enhanced representation and multi-task learning for image annotation. *Computer Vision and Image Understanding*, 117(5):466–478, 2013.
- [19] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.
- [20] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.
- [21] T. Bocklet, A. Maier, and E. Nöth. Age determination of children in preschool and primary school age with gmm-based supervectors and support vector machines/regression. In *Text, Speech and Dialogue*, pages 253–260. Springer, 2008.
- [22] S. Bordag. A comparison of co-occurrence and similarity measures as simulations of context. In *Computational Linguistics and Intelligent Text Processing*, pages 52–63. Springer, 2008.

- [23] A. Caliskan Islam, J. Walsh, and R. Greenstadt. Privacy detective: Detecting private information and collective privacy behavior in a large social network. In *Proceedings of the Workshop on Privacy in the Electronic Society*, pages 35–46. ACM, 2014.
- [24] G. Carlo, M. A. Okun, G. P. Knight, and M. R. T. de Guzman. The interplay of traits and motives on volunteering: Agreeableness, extraversion and prosocial value motivation. *Personality and Individual Differences*, 38(6):1293–1305, 2005.
- [25] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.
- [26] Z. Cemalcilar. Understanding individual characteristics of adolescents who volunteer. *Personality and Individual Differences*, 46(4):432–436, 2009.
- [27] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.
- [28] O. Chapelle, P. Shivaswamy, S. Vadrevu, K. Weinberger, Y. Zhang, and B. Tseng. Multi-task learning for boosting with application to web search ranking. In *Proceedings of the International ACM SIGKDD Conference*, pages 1189–1198. ACM, 2010.
- [29] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the International Conference on Machine Learning*, pages 129–136. ACM, 2009.
- [30] C. Christoudias, R. Urtasun, and T. Darrell. Multi-view learning in the presence of view disagreement. *arXiv preprint arXiv:1206.3242*, 2012.
- [31] P. Cimiano, A. Schultz, S. Sizov, P. Sorg, and S. Staab. Explicit versus latent concept models for cross-language information retrieval. In *Proceedings*

of *Joint Conference on Artificial Intelligence*, volume 9, pages 1513–1518. Citeseer, 2009.

- [32] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 1995.
- [33] M. H. Davis, K. V. Mitchell, J. A. Hall, J. Lothert, T. Snapp, and M. Meyer. Empathy, expectations, and situational preferences: Personality influences on the decision to participate in volunteer helping behaviors. *Journal of Personality*, 67(3):469–503, 1999.
- [34] M. De Choudhury, S. Counts, and E. Horvitz. Major life changes and behavioral markers in social media: case of childbirth. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pages 1431–1442. ACM, 2013.
- [35] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting depression via social media. In *Proceedings of the AAAI Conference on Web and Social Media*, page 2. AAAI Press, 2013.
- [36] P. De Meo, G. Quattrone, and D. Ursino. A decision support system for designing new services tailored to citizen profiles in a complex and distributed e-government scenario. *Data & Knowledge Engineering*, 67(1):161–184, 2008.
- [37] T. Evgeniou and M. Pontil. Regularized multi-task learning. In *Proceedings of the International ACM SIGKDD Conference*, pages 109–117. ACM, 2004.
- [38] J. Fan, Y. Gao, and H. Luo. Hierarchical classification for automatic image annotation. In *Proceedings of the International ACM SIGIR Conference*, pages 111–118. ACM, 2007.

- [39] J. Fan, Y. Gao, and H. Luo. Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation. *IEEE Transactions on Image Processing*, 17(3):407–426, 2008.
- [40] H. Fei and J. Huan. Structured feature selection and task relationship inference for multi-task learning. *Knowledge and Information Systems*, 35(2):345–364, 2013.
- [41] H. Fei, R. Jiang, Y. Yang, B. Luo, and J. Huan. Content based social behavior prediction: a multi-task learning approach. In *Proceedings of the International ACM Conference on Information and Knowledge Management*, pages 995–1000. ACM, 2011.
- [42] J. L. Fischer. Social influences on the choice of a linguistic variant. *Word*, 14(1):47–56, 1958.
- [43] Y. Freund, R. E. Schapire, et al. Experiments with a new boosting algorithm. In *Proceedings of the International Conference on Machine Learning*, volume 96, pages 148–156. ACM, 1996.
- [44] Y. Fu, L. Cao, G. Guo, and T. S. Huang. Multiple feature fusion by subspace learning. In *Proceedings of the International Conference on Content-based Image and Video Retrieval*, pages 127–134. ACM, 2008.
- [45] D. Ganguly, D. Roy, M. Mitra, and G. J. Jones. Word embedding based generalized language model for information retrieval. In *Proceedings of the International ACM SIGIR Conference*, pages 795–798, 2015.
- [46] H. Gao, F. Nie, X. Li, and H. Huang. Multi-view subspace clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4238–4246, 2015.

- [47] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, and X. Wu. Visual-textual joint relevance learning for tag-based social image search. *IEEE Transactions on Image Processing*, 22(1):363–376, 2013.
- [48] N. Garera and D. Yarowsky. Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of ACL and the International Joint Conference on Natural Language Processing*, pages 710–718. Association for Computational Linguistics, 2009.
- [49] J. Guo, G. Xu, X. Cheng, and H. Li. Named entity recognition in query. In *Proceedings of the International ACM SIGIR Conference*, pages 267–274. ACM, 2009.
- [50] S. Han, D. He, and Z. Yue. Benchmarking the privacy-preserving people search. In *Proceeding of the ACM Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security*, page 13. ACM, 2014.
- [51] Z. Hao, S. Zhong, and N. Yu. A privacy-preserving remote data integrity checking protocol with data dynamics and public verifiability. *IEEE Transactions on Knowledge and Data Engineering*, 23(9):1432–1437, 2011.
- [52] A. G. Hauptmann, C. Ngo, X. Xue, Y. Jiang, C. Snoek, and N. Vasconcelos, editors. *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, June 23-26, 2015*. ACM, 2015.
- [53] J. He and R. Lawrence. A graph-based framework for multi-task multi-view learning. In *Proceedings of the International Conference on Machine Learning*, pages 25–32, 2011.
- [54] J. Hitchen. *Implementing a Volunteer-Match Service*. PhD thesis, Al Akhawayn University in Ifrane, 2013.

- [55] T. Holtgraves. Text messaging, personality, and the social context. *Journal of Research in Personality*, 45(1):92–99, 2011.
- [56] R. A. Horn and C. R. Johnson. Topics in matrix analysis. *Cambridge University Press, Cambridge*, 37:39, 1991.
- [57] C.-J. Hsieh and I. S. Dhillon. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *Proceedings of the International ACM SIGKDD Conference*, pages 1064–1072. ACM, 2011.
- [58] L. Humphreys, P. Gill, and B. Krishnamurthy. How much is too much? privacy issues on twitter. In *Conference of International Communication Association, Singapore*, 2010.
- [59] F. Iacobelli, A. J. Gill, S. Nowson, and J. Oberlander. Large scale personality classification of bloggers. In *Affective Computing and Intelligent Interaction*, pages 568–577. Springer, 2011.
- [60] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda. Topic tracking model for analyzing consumer purchase behavior. In *Proceedings of Joint Conference on Artificial Intelligence*, volume 9, pages 1427–1432, 2009.
- [61] L. Jacob, J.-p. Vert, and F. R. Bach. Clustered multi-task learning: A convex formulation. In *Advances in Neural Information Processing Systems*, pages 745–752, 2009.
- [62] X. Jin, F. Zhuang, S. Wang, Q. He, and Z. Shi. Shared structure learning for multiple tasks with multiple views. In *Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases*, pages 353–368. Springer, 2013.

- [63] Z. Kang, K. Grauman, and F. Sha. Learning with whom to share in multi-task feature learning. In *Proceedings of the International Conference on Machine Learning*, pages 521–528, 2011.
- [64] M. Kantarcioglu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE Transactions on Knowledge and Data Engineering*, (9):1026–1037, 2004.
- [65] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. A method for large-scale l_1 -regularized least squares problems with applications in signal processing and statistics. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):606–617, 2007.
- [66] S. Kim and E. P. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In *Proceedings of the International Conference on Machine Learning*. ACM, 2010.
- [67] C. Kohlschütter, P. Fankhauser, and W. Nejdl. Boilerplate detection using shallow text features. In *Proceedings of the International ACM Conference on Web Search and Data Mining*, pages 441–450. ACM, 2010.
- [68] M. L. Korzaan and K. T. Boswell. The influence of personality traits and information privacy concerns on behavioral intentions. *Journal of Computer Information Systems*, 48(4):15–24, 2008.
- [69] A. Kumar and H. D. III. Learning task grouping and overlap in multi-task learning. In *Proceedings of the International Conference on Machine Learning*, pages 1383–1390, 2012.
- [70] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.

- [71] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.
- [72] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: social honeypots+ machine learning. In *Proceedings of the International ACM SIGIR Conference*, pages 435–442. ACM, 2010.
- [73] D. Li, B. He, Y. Ding, J. Tang, C. Sugimoto, Z. Qin, E. Yan, J. Li, and T. Dong. Community-based topic modeling for social tagging. In *Proceedings of the International ACM Conference on Information and Knowledge Management*, pages 1565–1568. ACM, 2010.
- [74] S.-Y. Li, Y. Jiang, and Z.-H. Zhou. Partial multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 2014.
- [75] Y. Li, X. Zhou, P. Bruza, Y. Xu, and R. Y. Lau. A two-stage text mining model for information filtering. In *Proceedings of the International ACM Conference on Information and Knowledge Management*, pages 1023–1032. ACM, 2008.
- [76] B. H. Lim, D. Lu, T. Chen, and M.-Y. Kan. # mytweet via instagram: Exploring user behaviour across multiple social networks. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 113–120. IEEE, 2015.
- [77] J. Liu, F. Zhang, X. Song, Y.-I. Song, C.-Y. Lin, and H.-W. Hon. What’s in a name?: an unsupervised approach to link users across communities. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 495–504. ACM, 2013.

- [78] K. Liu, H. Kargupta, and J. Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):92–106, 2006.
- [79] K. Liu and E. Terzi. A framework for computing the privacy scores of users in online social networks. *ACM Transactions on Knowledge Discovery from Data*, 5(1):6, 2010.
- [80] H. Mao, X. Shuai, and A. Kapadia. Loose tweets: an analysis of privacy leaks on twitter. In *Proceedings of the ACM workshop on Privacy in the Electronic Society*, pages 1–12. ACM, 2011.
- [81] D. Markovikj, S. Gievska, M. Kosinski, and D. Stillwell. Mining facebook data for predictive personality modeling. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*. AAAI Press, 2013.
- [82] R. R. McCrae and O. P. John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992.
- [83] F. McSherry and I. Mironov. Differentially private recommender systems: building privacy into the net. In *Proceedings of the International ACM SIGKDD Conference*, pages 627–636. ACM, 2009.
- [84] M. Michelson and S. A. Macskassy. Discovering users’ topics of interest on twitter: a first look. In *Proceedings of the Workshop on Analytics for Noisy Unstructured Text Data*, pages 73–80. ACM, 2010.
- [85] R. Mihalcea and P. Tarau. Textrank: Bringing order into text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 404–411, 2004.

- [86] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.
- [87] T. M. Mitchell. Machine learning. *Burr Ridge, IL: McGraw Hill*, 1997.
- [88] L. Nie, L. Zhang, Y. Yang, M. Wang, R. Hong, and T.-S. Chua. Beyond doctors: future health prediction from multimedia and multimodal observations. In *Proceedings of the ACM conference on multimedia conference*, pages 591–600. ACM, 2015.
- [89] L. Nie, Y.-L. Zhao, X. Wang, J. Shen, and T.-S. Chua. Learning to recommend descriptive tags for questions in social forums. *ACM Transactions on Information Systems*, 32(1):5, 2014.
- [90] J. Otterbacher. Inferring gender of movie reviewers: exploiting writing style, content and metadata. In *Proceedings of the International ACM Conference on Information and Knowledge Management*, pages 369–378. ACM, 2010.
- [91] M. Pennacchiotti and A.-M. Popescu. Democrats, republicans and starbucks aficionados: user classification in twitter. In *Proceedings of the International ACM SIGKDD Conference*, pages 430–438. ACM, 2011.
- [92] M. Pennacchiotti and A.-M. Popescu. A machine learning approach to twitter user classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 281–288. AAAI Press, 2011.
- [93] J. W. Pennebaker and L. A. King. Linguistic styles: language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296, 1999.

- [94] L. A. Penner. Dispositional and organizational influences on sustained volunteerism: An interactionist perspective. *Journal of Social Issues*, 58(3):447–467, 2002.
- [95] L. A. Penner. Volunteerism and social problems: Making things better or worse? *Journal of Social Issues*, 60(3):645–666, 2004.
- [96] A. Popescu, G. Grefenstette, et al. Mining user home location and gender from flickr tags. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*. AAAI Press, 2010.
- [97] X. Qian, H. Feng, G. Zhao, and T. Mei. Personalized recommendation combining user interest and social circle. *IEEE Transactions on Knowledge and Data Engineering*, 26(7):1763–1777, 2014.
- [98] D. Quercia, R. Lambiotte, D. Stillwell, M. Kosinski, and J. Crowcroft. The personality of popular facebook users. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pages 955–964. ACM, 2012.
- [99] L. Rainie, S. Kiesler, R. Kang, M. Madden, M. Duggan, S. Brown, and L. Dabbish. Anonymity, privacy, and security online. *Pew Research Center*, 5, 2013.
- [100] K. Ramanathan and K. Kapoor. Creating user profiles using wikipedia. In *Conceptual Modeling-ER 2009*, pages 415–427. Springer, 2009.
- [101] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in twitter. In *Proceedings of the ACM Workshop on Search and Mining User-generated Contents*, pages 37–44. ACM, 2010.
- [102] R. J. Renes. Sustained volunteerism: justification, motivation and management. *Amsterdam, Kurt Lewin Instituut*, 3:2005, 2005.

- [103] S. Rosenthal and K. McKeown. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 763–772. Association for Computational Linguistics, 2011.
- [104] G. Salton and M. J. McGill. Introduction to modern information retrieval. *McGraw Hill*, 1983.
- [105] M. Salzman, C. H. Ek, R. Urtasun, and T. Darrell. Factorized orthogonal latent spaces. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 701–708, 2010.
- [106] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the International ACM SIGIR Conference*, pages 253–260. ACM, 2002.
- [107] V. Schickel-Zuber and B. Faltings. Using hierarchical clustering for learning theontologies used in recommendation systems. In *Proceedings of the International ACM SIGKDD Conference*, pages 599–608. ACM, 2007.
- [108] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS One*, 8(9):e73791, 2013.
- [109] L. Sigal and M. J. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. *Brown University TR*, 120, 2006.

- [110] V. Sindhwani, P. Niyogi, and M. Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of the International Conference on Machine Learning*, pages 74–79. Citeseer, 2005.
- [111] X. Song, Z. Ming, L. Nie, Y. Zhao, and T. Chua. Volunteerism tendency prediction via harvesting multiple social networks. *ACM Transactions on Information Systems*, 34(2):10, 2016.
- [112] X. Song, L. Nie, L. Zhang, M. Akbari, and T.-S. Chua. Multiple social network learning and its application in volunteerism tendency prediction. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 213–222. ACM, 2015.
- [113] X. Song, L. Nie, L. Zhang, M. Liu, and T. Chua. Interest inference via structure-constrained multi-source multi-task learning. In *Proceedings of the IJCAI*, pages 2371–2377, 2015.
- [114] Y. Song, D. Dahlmeier, and S. Bressan. Not so unique in the crowd: a simple and effective algorithm for anonymizing location data. In *Proceedings of the International ACM SIGIR Conference*, page 19. ACM, 2014.
- [115] D. Spina, J. Gonzalo, and E. Amigó. Learning similarity functions for topic detection in online reputation monitoring. In *Proceedings of the International ACM SIGIR Conference*, pages 527–536. ACM, 2014.
- [116] S. Tan, Z. Guan, D. Cai, X. Qin, J. Bu, and C. Chen. Mapping users across networks by manifold alignment on hypergraph. In *AAAI*, volume 14, pages 159–165, 2014.
- [117] J. Tang, L. Yao, D. Zhang, and J. Zhang. A combination approach to web user profiling. *ACM Transactions on Knowledge Discovery from Data*, 5(1):2, 2010.

- [118] S. Thrun and J. O’Sullivan. Discovering structure in multiple learning tasks: The tc algorithm. In *Proceedings of the International Conference on Machine Learning*, volume 96, pages 489–497, 1996.
- [119] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society.*, pages 267–288, 1996.
- [120] A. Vasalou, A. J. Gill, F. Mazanderani, C. Papoutsis, and A. Joinson. Privacy dictionary: A new resource for the automated content analysis of privacy. *Journal of the American Society for Information Science and Technology*, 62(11):2095–2105, 2011.
- [121] T. Vu and V. Perez. Interest mining from user tweets. In *Proceedings of International ACM Conference on Information and Knowledge Management*, pages 1869–1872. ACM, 2013.
- [122] J. Warburton and T. Crosier. Are we too busy to volunteer?: the relationship between time and volunteering using the 1997 abs time use data. *Australian Journal of Social Issues, The*, 36(4):295, 2001.
- [123] M. Wasim, I. Shahzadi, Q. Ahmad, and W. Mahmood. Extracting and modeling user interests based on social media. In *Proceedings of the IEEE International Conference on Multitopic*, pages 284–289. IEEE, 2011.
- [124] X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of International ACM SIGIR conference*, pages 178–185. ACM, 2006.
- [125] Z. Wen and C.-Y. Lin. Improving user interest inference from social neighbors. In *Proceedings of the International ACM Conference on Information and Knowledge Management*, pages 1001–1006. ACM, 2011.

- [126] S. S. Woo and H. Manjunatha. Empirical data analysis on user privacy and sentiment in personal blogs.
- [127] W. Wu, B. Zhang, and M. Ostendorf. Automatic generation of personalized annotation tags for twitter users. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 689–692. Association for Computational Linguistics, 2010.
- [128] W. W. Wymer Jr and S. Samu. Volunteer service as symbolic consumption: Gender and occupational differences in volunteering. *Journal of Marketing Management*, 18(9-10):971–989, 2002.
- [129] S. Xiang, L. Yuan, W. Fan, Y. Wang, P. M. Thompson, and J. Ye. Multi-source learning with block-wise missing data for alzheimer’s disease prediction. In *Proceedings of the International ACM SIGKDD Conference*, pages 185–193. ACM, 2013.
- [130] C. Xu, D. Tao, and C. Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.
- [131] L. Xu, A. Huang, J. Chen, and E. Chen. Exploiting task-feature co-clusters in multi-task learning. In *AAAI*, pages 1931–1937, 2015.
- [132] Y. Xue, X. Liao, L. Carin, and B. Krishnapuram. Multi-task learning for classification with dirichlet process priors. *The Journal of Machine Learning Research*, 8:35–63, 2007.
- [133] T. Yarkoni. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3):363–373, 2010.

- [134] Q. Yin, S. Wu, and L. Wang. Incomplete multi-view clustering via subspace learning. In *Proceedings of the International ACM Conference on Information and Knowledge Management*, pages 383–392. ACM, 2015.
- [135] J. Yu, M. Wang, and D. Tao. Semisupervised multiview distance metric learning for cartoon synthesis. *IEEE Transactions on Image Processing*, 21(11):4636–4648, 2012.
- [136] L. Yuan, Y. Wang, P. M. Thompson, V. A. Narayan, and J. Ye. Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data. In *Proceedings of the International ACM SIGKDD Conference*, pages 1149–1157. ACM, 2012.
- [137] R. Zafarani and H. Liu. Connecting users across social media sites: a behavioral-modeling approach. In *Proceedings of the International ACM SIGKDD Conference*, pages 41–49. ACM, 2013.
- [138] D. Zhai, H. Chang, S. Shan, X. Chen, and W. Gao. Multiview metric learning with global consistency and local smoothness. *ACM Transactions on Intelligent Systems and Technology*, 3(3):53, 2012.
- [139] D. Zhang, F. Wang, and L. Si. Composite hashing with multiple information sources. In *Proceedings of the International ACM SIGIR Conference*, pages 225–234. ACM, 2011.
- [140] J. Zhang and J. Huan. Inductive multi-task learning with multiple view data. In *Proceedings of the International ACM SIGKDD Conference*, pages 543–551. ACM, 2012.
- [141] J. Zhang and S. Y. Philip. Integrated anchor and social link predictions across social networks. In *Proceedings of the International Conference on Artificial Intelligence*, pages 2125–2131. AAAI Press, 2015.

- [142] S. Zhang, H. Yang, and L. Singh. Increased information leakage from text. In *Proceedings of the International ACM SIGIR Conference*. ACM, 2014.
- [143] Y. Zhang and J. Koren. Efficient bayesian hierarchical user modeling for recommendation system. In *Proceedings of the International ACM SIGIR Conference*. ACM, 2007.
- [144] W. X. Zhao, S. Li, Y. He, E. Chang, J.-R. Wen, and X. Li. Connecting social media to e-commerce: Cold-start product recommendation on microblogs. *IEEE Transactions on Knowledge and Data Engineering*, 28(5):1147 – 1159, 2016.
- [145] Y.-L. Zhao, Q. Chen, S. Yan, T.-S. Chua, and D. Zhang. Detecting profilable and overlapping communities with user-generated multimedia contents in lbsns. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 10(1):3, 2013.
- [146] J. Zhou, J. Chen, and J. Ye. Clustered multi-task learning via alternating structure optimization. In *Advances in Neural Information Processing Systems*, pages 702–710, 2011.
- [147] X. Zhou, X. Liang, H. Zhang, and Y. Ma. Cross-platform identification of anonymous identical users in multiple social media networks. *IEEE Transactions on Knowledge and Data Engineering*, 28(2):411–424, 2016.
- [148] I. Zukerman and D. W. Albrecht. Predictive statistical models for user modeling. *User Modeling and User-Adapted Interaction*, 11(1-2):5–18, 2001.