Thesis for the Degree of Ph.D.

Topic Models for Finding Social Interaction Patterns Using Calls and Proximity Logs

School of Computer Science and Engineering
The Graduate School

Han Yong-Jin

December 2015

The Graduate School
Kyungpook National University

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Understanding Social Interactions

Social interactions are indispensable to modern daily activities. Interpersonal relationships evolve from social interactions, and information is shared through these interactions. In sociology, social interactions have been studied for decades now [66]. For example, in the 1960s, the basic assumption of the consumer theory was that consumers behave independently of each other. However, Duesenberry [22] empirically demonstrated that consumption patterns have a social character. Miller [51] ethnographically inferred the role played by shopping in building relationships with friends and family. These studies provide interesting insights into the workings of human dynamics. However, studies of this kind have relied mainly on manual reports from the participants being studied or by human observers. Current technologies enable us to record massive amounts of data on human activities. Subsequently, it becomes a challenge to understand social interactions from a large-scale hu-

man activity log over time, and this research field is emerging rapidly as the so-called computation social science (CSS) [43, 62].

Voice call is a conventional resource of CSS [12] since large amounts of call detail records (CDRs) have been collected by mobile phone operators and shared by anonymizing customers for research purposes. Face-to-face communication is emerging as a new resource recently [43]. Wearable devices such as sociometer [16] have facilitated the use of face-to-face interactions over time. Currently, mobile phones with various built-in sensors allow both types of interactions to be captured together. For example, Bluetooth provides an imperfect, yet reasonable approximation of a face-to-face meeting. It senses whether other mobile phone users are in close proximity with high probability. In addition, a mobile phone accurately logs phone calls and their duration. Another advantage of mobile phones is that they are carried spontaneously as compared to other wearable devices for sensing and collecting data. Thus, it is possible to capture mobile logs over a sufficiently long period naturally and these logs offer some intuitions about when people meet or call others, and for what activities. A study on social networks has reported that a user might make a different decision depending on whom she is currently staying with [55]. This has motivated the use of social interactions sensed by mobile phones for context-aware recommendation [40], and various ways of using social interactions have been studied for analyzing social relationships [24, 48], spending behaviors [66], users' profiles [5], and opinion diffusion [26].

1.2 Existing Approaches and Challenges

Many existing approaches in CSS focus on a single type of data. For example, Onnela et al. [58] analyzed CDRs from 3.9 million users and found evidence supporting the weak ties hypothesis [29], a central concept in social network analysis. By analyzing CDRs for 2 million users, Hidalgo and Rodriguez-Sickert [34] found that persistent links tend to be reciprocal and associated with low degree nodes. It is a basic principle in these studies to construct a social network among users by regarding the number of phone calls between two users as the strength of their relationship. Thus, in these studies, an individual's calls are aggregated and the individual's call patterns over time are ignored, whereas many studies using face-to-face interactions have paid attention to finding individuals' interaction patterns over time. A preferred way for finding the patterns is to utilize topic models [19, 25]. Farrahi and Gatica-Perez [25] proposed a method to recognize daily routines as a probabilistic combination of interaction patterns. In this work, the interaction patterns are regarded as topics that generate a mobile log. Subsequently, a mobile log is summarized as a mixture of topics by Latent Dirichlet Allocation (LDA) [6]. Do and Gatica-Perez [19] proposed a new probabilistic topic model to infer interaction patterns that predict which users are likely to be in close proximity at a certain time. However, these studies ignored call logs and only focused on proximities to represent social interactions.

There have also been many studies that employ both proximity and call logs for their own tasks [21, 24, 48, 66]. An average proximity frequency within a specific time slot, such as weekday daytime or weekend evening, is often em-

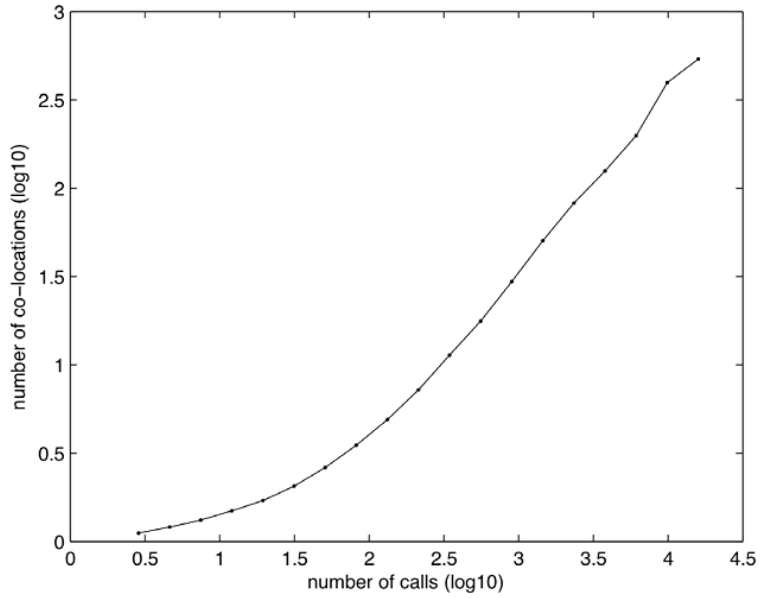


Figure 1.1: Average co-locations according to the number of calls [11].

ployed for social relationship classification [21, 24, 48]. However, these studies utilized call and proximity logs independently. That is, call and proximity logs were separately managed, even though proximities often follow a call. Therefore, these studies lost the information obtainable when considering calls and proximities simultaneously.

There are two challenges to be addressed for a better understanding of social interactions. The first challenge is to find social interaction patterns from calls and proximities simultaneously. Figure 1.1, which is borrowed from Calabrese et al. [11], shows that the more people call each other, the more they co-locate. Thus, it is natural to analyze calls and proximities simultaneously by assuming some influences between the two interaction types. However, it is non-trivial to reveal the influences since we do not know which call depends on which

proximity or vice versa.

The second challenge is to show how useful the discovered interaction patterns are. Most existing studies have utilized an aggregate call, whereas how to find individuals' call patterns over time has not been studied well. It is also the first time to find call and proximity patterns simultaneously. Thus, it is necessary to show how useful the found patterns are for other tasks such as social relationship analysis [24, 48], spending behavior classification [66], app installation prediction [59], and information diffusion [48].

1.3 Approach of the Dissertation

This dissertation proposes a topic-based method to identify call and proximity patterns from a mobile log. A set of calls from a target user to another user is represented as a set of predefined time slots, where a time slot records the frequency of calls during the time slot. As a result, this set can be understood as a document in which words denote the time slots. A number of such documents exist, because the target user places calls to many other users. Therefore, the latent topics of the user's calls are identified through LDA [6] and can be regarded as call patterns of the user. In the same manner, the proximity patterns of the user can be determined.

In order to consider the call patterns and the proximity patterns simultaneously, the proposed method regards the calls and proximities as a homogeneous information type. Thus, call-LDA and proximity-LDA are combined into a single topic model. There are two factors to be considered when they are combined. One is that the number of proximities in a mobile log usually

overwhelms that of calls, and the other is that proximities are observed more regularly than calls [24]. As a result, the proposed method assumes that call topics depend on proximity topics. Therefore, the final topic model is a combined model of call-LDA and proximity-LDA in which calls are dependent on proximities.

1.4 Contribution Summary and Organization of the Dissertation

The proposed model is evaluated by comparing it with several existing topic models as well as the model that considers the calls and the proximities independently. The experimental results based on three different data sets from the Massachusetts Institute of Technology's (MIT's) Reality Mining project showed that the proposed model outperforms all compared models, when performance is measured with a perplexity measure. In addition, to demonstrate the effectiveness of the topic model, it is also shown that the topic model distinguishes friendship relations among users correctly. These results prove that the proposed model is appropriate for identifying social interaction patterns.

Contributions of this dissertation can be summarized as follows.

(1) **Single directional influence modeling from proximities to calls**: This dissertation proposed a single directional influence topic model using call and proximity logs simultaneously. The findings of this dissertation suggest that this single directional influence modeling is reasonable for finding social interaction patterns. The basic idea is published in [31]

and a detailed journal version is published in [32].

(2) **Analyzing social relationship using interaction patterns**: This dissertation proposed a method to utilize the proposed topic model for analyzing social relationships. The experimental results demonstrate that inferred patterns are appropriate for describing social relationships discriminately. The results were published in [32]. In addition, the proposed topic model is applied to the classification of various social relationships successively.

The remainder of this dissertation is organized as follows. Chapter 2 presents the related works in three aspects. This chapter first summarizes studies on employing proximity and call logs and then, explains LDA-based topic models in terms of identifying social interaction patterns. Lastly, the chapter discusses the related works on social relationship classification. Chapter 3 suggests topic models using call and proximity logs simultaneously for finding social interaction patterns. This chapter first describes the background on topic modeling and then, discusses the justification of the proposed model by comparing three different topic models. Then, this chapter explains the estimation process for the parameters of the proposed model. Chapter 4 presents experiments from three aspects. This chapter first demonstrates the justification of the suggested assumption for finding social interaction patterns and evaluates the proposed method. Then, the chapter shows the usefulness of the proposed topic model in terms of discriminating social relationships, and lastly, topic models are actually applied to social relationship classification. Finally, Chapter 5 concludes the dissertation.

Chapter 2

Related Works

This chapter reviews related works in three sub-chapters. Chapter 2.1 presents existing studies on understanding social interactions in terms of computational social science (CSS). Especially, approaches using call and proximity logs are reviewed. Chapter 2.2 summarizes topic model-based approaches and discusses methods to use existing topic models for understanding social interactions. This section also introduces previously published papers [31, 32] on the work of this dissertation. Lastly, Chapter 2.3 discusses related works on social relationship classification.

2.1 Computational Social Science Using Call and Proximity Logs

The related works are summarized in three groups according to data to be used. First two groups are approaches using call and proximity logs, respec-

tively. Here, a proximity log indicates face-to-face communication captured by a wearable device or a mobile phone. The last group utilize the two logs together.

Most earlier studies [11, 12, 34, 58] in computational social science belong to the first group. The studies have utilized call detailed records (CDRs), which contain an enormous amount of information on communications between millions of people. For example, Hidalgo and Rodriguez-Sickert [34] analyzed CDRs for 2 million users for investigating the correlations between the structure of a mobile phone network and the persistence of its links. They found that persistent links tend to be reciprocal and associated with low degree nodes. Onnela et al. [58] construct a weighted phone network of 3.9 million users from CDRs for social network analysis. They regard the network as a proxy for the underlying human communication network. Studies of this kind aggregate calls between users and determine the strength of their relationship using this frequency. In recent, some studies utilize not only phone calls but also locations where the phone calls occur. Calabrese et al. [11] showed that the number of reciprocal calls between users is propositional to the average number of their co-locations within 1 year. Candia et al. [12] grouped users according to number of calls and analyzed spatio-temporal call patterns for each group. Since locations were captured at the moments when calls occur, co-locations can not reflect daily face-to-face meetings exactly. However, these studies inspire the idea that phone calls and face-to-face interactions are related with each other.

The second group which utilizes proximity logs among users have emerged newly with the advent of wearable devices. Sociometer [15] detects face-to-face

interactions between two individuals by an infra-red (IR) transceiver. An IR transceiver sends out unique ID for one individual and receives ID from the other individual within their proximity. Similarly, a Bluetooth embedded in a mobile phone senses nearby devices including phones. Since mobile phones are carried spontaneously, they are particularly suitable for research in Social Signal Processing (SSP), the domain aimed at automatic understanding of social interactions sensing non-verbal behavior [72]. Madan et al. [50] showed that proximities captured by mobile phones explain individual political opinions better than self-reported social ties. They also investigated the activities of people who changed their preferred party versus those that did not. Their results present that the former people that changed preferred party often discuss face-to-face with their democrat political discussants whereas the latter often interact with people that have little or no interest in politics. Epidemiological behavior change also have been studied using proximities sensed by mobile phones [49]. This study demonstrated that symptoms including runny nose, sore throat, and fever affect face-to-face interactions. For example, a person's total proximities and entropy decreased after she had experienced a high fever.

Lastly, there have been many studies on employing proximity and call logs for various social interaction analysis tasks [24, 48, 66]. Eagle et al. [24] employed total number of calls and an average proximity frequency within a specific time slot for social relationship classification, and Madan and Pentland [48] subdivided the number of calls to address the same problem. They both used an average proximity within a specific time slot as well as the number of calls. On the other hand, Singh et al. [66] employed call and proximity statistics to classify human spending behavior. In this work, the number of calls and

proximities becomes a feature to characterize a target human behavior. All these studies commonly assumed that calls and proximities are independent of each other. This independent assumption allows proximity and call logs to be analyzed easily in understanding interactions. That is, the most straightforward method to use both proximity and call logs in understanding interactions is to manage them separately. However, many proximities usually follow a call in a real situation. As a result, these previous studies lost the information that is obtainable when calls and proximities are considered simultaneously. To the best of our knowledge, no previous studies analyze proximities and calls simultaneously for understanding social interactions.

2.2 Understanding Social Interactions Using Topic Models

The proposed method is based on LDA [6] which is a powerful generative model for managing text documents. LDA models a document as a mixture of topics, where a topic is characterized by a distribution over words. Blei et al. [6] showed that text classification performance is improved by using LDA topics as features, instead of using word frequencies. Wei and Croft [75] also demonstrated that an LDA-based document retrieval model provides higher precision than either a cluster-based model or a traditional bag-of-words model.

There have been also some recent studies that utilize LDA for processing non-textual data [4, 8]. A few of them used a mobile log to understand human

activities [19, 25]. For instance, Farrahi and Gatica-Perez [25] adopted LDA to investigate the proximity patterns of mobile phone users. In this study, it was assumed that a proximity log would be generated from LDA topics. Thus, they regarded the LDA topics as proximity patterns. Huynh et al. [39] used LDA to identify daily routines as a probabilistic combination of activity patterns. They used wearable sensor data annotated with 34 activities such as having dinner, walking freely, and so on. However, these studies are also limited to single interaction type.

Recently, the polylingual topic model (PLTM) has been proposed as an extension of LDA to describe the joint distribution of loosely equivalent documents written in different languages [52]. In PLTM, the documents in each language are modeled as an LDA. Subsequently, LDAs for different languages are combined into a single topic model. When the LDAs are combined, it is assumed that loosely equivalent documents share the same topics. A mobile log can be modeled with PLTM by regarding proximities and calls as two types of documents written in two different languages. In this modeling of a mobile log, it should also be assumed that proximities and calls share the same topics as in natural language documents.

Unlike PLTM, the proposed method models a single directional influence from proximities to calls. The proposed model is inspired by recent image annotation topic models [7, 60]. Blei and Jordan [7] proposed the correspondence LDA (Corr-LDA) to model a dependency from an image to captions of the image. Putthividhy et al. [60] then proposed topic regression multi-modal LDA (tr-mmLDA), a generalized version of Corr-LDA. These two topic models generate an image first, and then generate the caption for the image by

conditioning on the topics used in the image. Since the dependency between an image and its caption is observed explicitly, a dependency exists between them. However, the dependency between calls and proximities in a mobile phone log is not observed explicitly. Therefore, this type of dependency is not adequate for analyzing calls and proximities.

The basic idea of this dissertation is published in [31]. The conference paper investigated the justification for modeling a single directional influence from proximities to calls. As a result, the superiority of the proposed method have been shown with data from 44 users in the Reality Mining data set [23]. This initial work is extended significantly in three aspects and the extended version is published in [32]. First, our methodology is applied to two additional data sets - the Social Evolution data set which includes data from 72 users [26], and the Friends & Family data set containing data from 114 users [1]. Because the three data sets (including the Reality Mining [23]) contain data from different user groups, justification for our modeling is thoroughly analyzed by evaluating the proposed method on the data sets. Second, social interaction patterns are represented at a more specific level, compared to the initial work. That is, the initial work described social interaction patterns on weekdays and weekends, whereas the days are subdivided in this dissertation. Lastly, in order to determine the effectiveness of the proposed method, the extended version investigates its ability to discriminate friendships among users by analyzing the interaction patterns that the proposed method has identified.

2.3 Social Relationship Classification

This dissertation additionally demonstrates the usefulness of topic models by applying the models to classifying various social relationships. Social relationship classification can be invoked or applicable to two core problems on social relationship analysis [37]. One problem is link prediction first introduced by Liben-Nowell and Kleinberg [44]. Given a snapshot of a social network at time t, the problem is to predict the edges that will be added to the network during the interval from time t to a given future time t'. To solve the problem, Liben-Nowell and Kleinberg [44] adopt an unsupervised approach using various proximity features including common neighbors, paths between nodes, and PageRank scores [10]. Lichtenwalter et al. [45] modeled the problem as a binary classification and discussed the class imbalance problem that positive links are rarely observed compared with negative links. To solve the problem, they utilized ensemble methods with over/under sampling methods [14, 68]. More recently, Scellato et al. [64] proposed a supervised approach using place features in link prediction on location-based social networks, and Backstrom and Leskovec [2] proposed supervised random works which combines the classification approach and a node ranking approach based on random works. These studies are commonly based on features arising from social ties and regard the features independently, whereas, in this dissertation, social ties are just regarded as class labels and latent call and proximity patterns are simultaneously inferred for describing a feature space.

Another core problem on social relationship analysis is to infer the meanings of social relationships such as the manager-subordinate relationships [18], the

advisor-advisee relationships [74] and the friendship [17]. Eagle et al. [24] employed calls and proximities sensed by mobile phones for inferring friendship and Crandall et al. [17] shows how temporal and spatial co-occurrences between people help to infer social ties among mobile users. These two studies are related with this dissertation in terms of utilizing social interactions. However Crandall et al. [17] only considered face-to-face interactions and Eagle et al. [24] deal with calls and proximities independently whereas this dissertation is concerned with using call and proximity logs simultaneously.

The studies on inferring social relationships also have suffered from the class imbalance problem [17, 24, 48, 50]. The related studies have used existing methods such as the cost sensitive learning [20] to address the problem. The cost sensitive learning is an algorithmic solution whereas this dissertation proposes a solution in terms of extracting features. The details are explained in Chapter 4.3.

Another related work related to social relationship classification is multilayer network explored in various contexts [13, 69, 70]. Kivelä et al. [41] reviewed a comprehensive review on multilayer network models. A call and a proximity can be regarded as links of different types and then call and proximity logs can be understood as a two-layer network. Recently, Hristova et al. [38] showed that the more communication channels utilized, the stronger the tie by using mobile logs including calls and proximities. However, they aggregate calls and proximities to form networks and thus ignore interaction patterns over time. In this dissertation, calls and proximities are regarded as mixtures of patterns, and the mixtures are utilized to distinguish social relationships of different types.

Chapter 3

Topic Models for Finding Social Interaction Patterns

This chapter presents methods using topic models for finding social interaction patterns and also suggests a novel topic model which models single directional influences from proximity to call logs. Chapter 3.1 first describes the focus of this dissertations in terms of computational social science and then provides a primer on topic models and data representation for modeling social interactions. A way to apply a topic model is also introduced. Chapter 3.2 discusses topic models using call and proximity logs simultaneously. Strengths and weaknesses of the models are explained step-by-step. To overcome the weaknesses of the models, Chapter 3.3.1 proposes a novel topic model, single direction influence LDA (sdiLDA). Chapter 3.4 shows examples of patterns obtained from the suggested topic models and discusses the results in qualitative. Lastly, perplexity for each topic model are suggested as an evaluation measure in Chapter 3.5.

3.1 Backgrounds

3.1.1 The Focus of the Dissertation

Figure 3.1 depicts an overview on computational social science (CSS) [43] for understanding social interactions and highlights the focus of the dissertation by the shaded rectangles. The overall flow of CSS is represented on the top of the figure and instances of the flow are represented by rectangles.

Social interactions are observed in various forms such as face-to-face communication, phone calls, and online communication. Hence, the data collection is a key factor in computational social science [9]. For example, social interactions are dynamic and affected by psychology and thus a manual survey is inevitable to investigate relationships of interactions with the inner world of people. The interactions also need to be monitored continuously and intensely [71]. Call detailed records (CDRs) provides a tremendous amount of calls among users. Sociometer [15] is a wearable device to capture face-to-face communication. Now, mobile phone with built-in sensors enables us to collect face-to-face interactions and phone calls together. Online interactions also can be collected automatically by crawling and by using various Web APIs.

This dissertation is concerned with face-to-face communication and phone calls which are reported as the top two preferred mediums for communication with friends and family [57]. Interactions of these two types are modeled using topic models for finding interaction patterns. Found patterns can be utilized in various application domains such as social relationship analysis [34, 61], app installation prediction [59] and spending behavior prediction [66]. Among them, this dissertation demonstrates the usefulness of the proposed approach

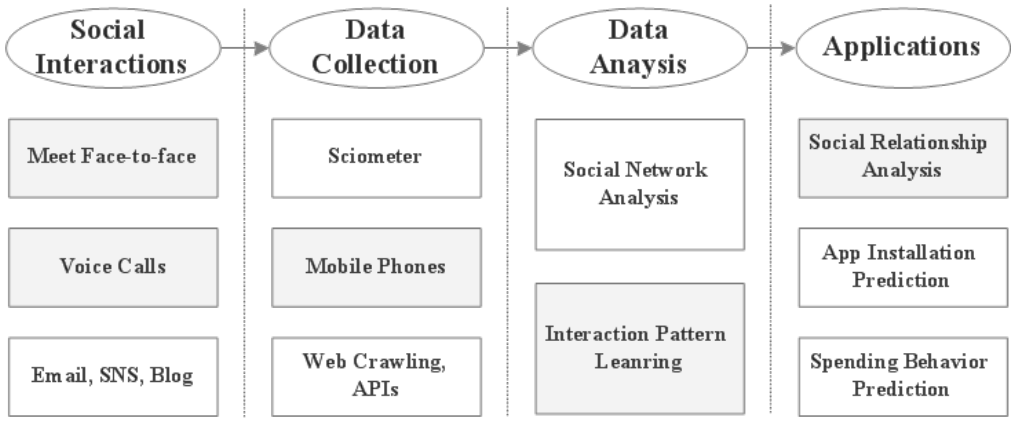


Figure 3.1: The focus of the dissertation in computational social science.

in social relationship analysis.

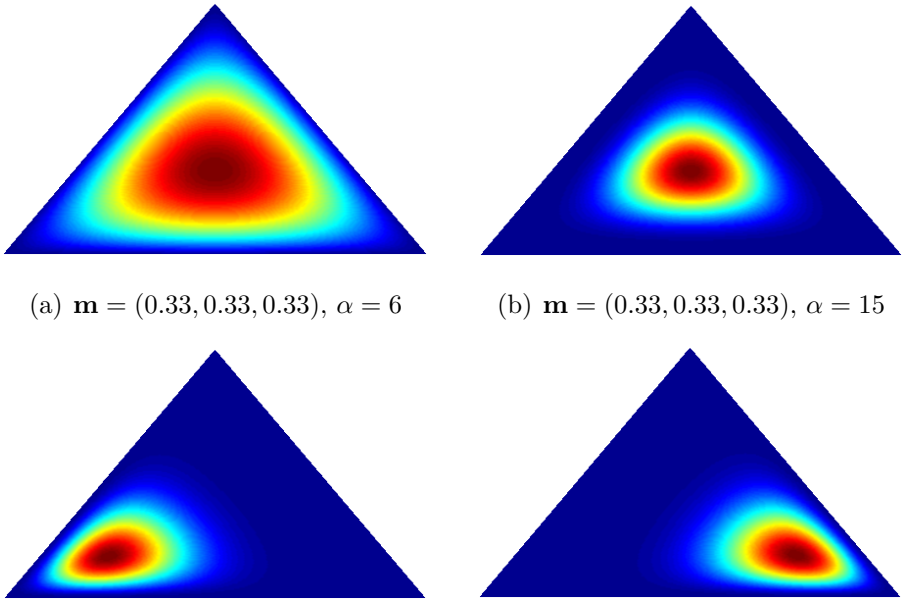
3.1.2 Topic Models

Preliminaries: Dirichlet, Multinomial, and Dirichlet-multinomial distributions

The Dirichlet distribution is defined over the $(K-1)$-dimensional simplex $\{\boldsymbol{\theta} \in \mathbb{R}^K\}$ given by the constrains $x_i \geq 0$, $\sum_{i=1}^{K} \boldsymbol{\theta}_i = 1$. The distribution is parameterized by a vector $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)$, where $\alpha_i > 0$. Its probability density function is given as follows.

$$f(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} \boldsymbol{\theta}_i^{\alpha_i - 1},$$

where Γ is the gamma function, which for non-negative real number is defined as $\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx$. A random variable following the Dirichlet distribu-



(a) $\mathbf{m} = (0.33, 0.33, 0.33)$, $\alpha = 6$ (b) $\mathbf{m} = (0.33, 0.33, 0.33)$, $\alpha = 15$

(c) $\mathbf{m} = (0.6, 0.2, 0.2)$, $\alpha = 15$ (d) $\mathbf{m} = (0.2, 0.6, 0.2)$, $\alpha = 15$

Figure 3.2: Four Dirichlet distributions over 2-simplex.

tion is denoted as follows.

$$\boldsymbol{\theta} = (\theta_1, ..., \theta_K) \sim Dir(\boldsymbol{\alpha}).$$

Let $\alpha = \sum_{i=1}^{K} \alpha_i$ and $\boldsymbol{\theta}'$ be the mean of a $Dir(\boldsymbol{\alpha})$. Then, $\boldsymbol{\theta}'$ is $(\frac{\alpha_i}{\alpha}, ..., \frac{\alpha_K}{\alpha})$, and thus it is proportional to $\boldsymbol{\alpha}$. The α decreases monotonically with the inverse variance of the distribution and thus describes how tightly concentrated the density is around the mean. Hence, the $\mathbf{m}$ and α are often called the base measure and concentration parameter respectively. The distribution $Dir(\boldsymbol{\alpha})$ then can be denoted alternatively as $Dir(\alpha\boldsymbol{\theta}')$. If all components of $\boldsymbol{\theta}'$ (or $\boldsymbol{\alpha}$) are the same, then the distribution is called a symmetric Dirichlet distribution. Otherwise, it is called an asymmetric one.

Figure 3.2 depicts Dirichlet distributions over 2-simplex with various parameters. Figure 3.2(a) and Figure 3.2(b) are symmetric Dirichlet distributions. Both distributions are the same as the center of the simplex. However, since the concentration parameter of Figure 3.2(b) is larger than that of Figure 3.2(a), the density of Figure 3.2(b) is concentrated more around the mean. Figure 3.2(c) and Figure 3.2(d) are asymmetric distributions with the same concentration parameter. As shown in these figures, their base measures lie at the left and right corner of the simplex respectively.

The multinomial distribution describes a probability distribution over histograms of a fixed size. Let $\mathbf{n} \in \mathbb{N}_{\geq 0}^{V}$ denote a V-dimensional vector of counts and the probability of the i-th element to be counted becomes θ_i, such that $\theta_i \geq 0$ and $\sum_{i=1}^{V} \theta_i = 1$. All random variables $\mathbf{n}$'s from a multinomial distribution have the same total count, such that $N = \sum_{i=1}^{V} n_i$, where N is a constant positive integer. The probability of a particular histogram $\mathbf{n}$ is then given as follows.

$$p(\mathbf{n}|\boldsymbol{\theta}) = \frac{\Gamma\left(\left(\sum_{i=1}^{V} n_i\right) + 1\right)}{\prod_{i=1}^{V} \Gamma(n_i + 1)} \prod_{i=1}^{V} \theta_i^{n_i},$$

where Γ is the gamma function, which for integers is defined as $\Gamma(x) = (x-1)!$.

The multinomial distribution, denoted as $Multi(\boldsymbol{\theta})$, is often combined with a Dirichlet distribution prior. The compound distribution $p(\mathbf{n}|\alpha)$ is the marginal distribution of the counts $\mathbf{n}$ conditioned on the $Dir(\alpha)$ prior, such that $p(\mathbf{n}|\alpha) = \int_{\theta} p(\mathbf{n}|\theta)p(\theta|\alpha)d\theta$. This distribution is called the Dirichlet-multinomial distribution, or sometimes the Dirichlet compound multinomial distribution or the

multivariate Pólya distribution. It is given as follows.

$$p(\mathbf{n}|\boldsymbol{\alpha}) = \int_{\theta} p(\mathbf{n}|\theta)p(\theta|\alpha)d\theta = \frac{\Gamma\left(\sum_{i=1}^{V}\alpha_i\right)}{\Gamma\left(\sum_{i=1}^{V}n_i+\alpha_i\right)}\prod_{i=1}^{V}\frac{\Gamma(n_i+\alpha_i)}{\Gamma(\alpha_i)}.$$

The joint distribution $p(\theta_i, n_i|\boldsymbol{\alpha})$ is propositional to $\theta_i^{n_i}\theta_i^{\alpha_i-1} = \theta_i^{n_i+\alpha_i-1}$. From these joint and marginal distributions, the posterior distribution $p(\theta|\mathbf{n}, \boldsymbol{\alpha})$ is solved as follows.

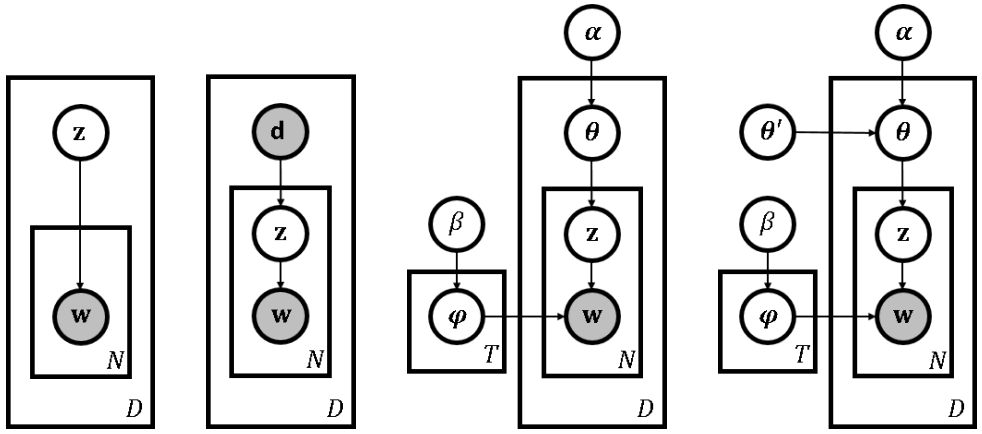
$$\begin{aligned} p(\theta|\mathbf{n}, \boldsymbol{\alpha}) = \frac{p(\theta, \mathbf{n}|\boldsymbol{\alpha})}{p(\mathbf{n}|\boldsymbol{\alpha})} &= \frac{\Gamma\left(\sum_{i=1}^{V}n_i+\alpha_i\right)}{\prod_{i=1}^{V}\Gamma\left(n_i+\alpha_i\right)}\prod_{i=1}^{V}\theta_i^{n_i+\alpha_i-1} \\ &= Dir(\mathbf{n}+\boldsymbol{\alpha}). \end{aligned}$$

Note that the posterior distribution is also a Dirichlet distribution whose parameters are the sum of both the count $\mathbf{n}$ and the Dirichlet parameter α. This property is a key factor for the topic modeling explained in the next chapter. Because the posterior is the same type of distribution as the prior, the Dirichlet distribution is called a conjugate prior to the multinomial distribution [63].

Topic Modeling

Topic models were originally proposed to capture latent topics in text corpora. The topics are shaped differently by different model assumptions. Figure 3.3 presents graphical representations of four different topic models, mixture of unigrams model, probabilistic latent semantic index (pLSI) [36], latent Dirichlet allocation (LDA) [6], and an alternative version of LDA.

The mixture model assumes that all words from a document are correlated by a single topic. As shown in Figure 3.3(a), this assumption is expressed by



(a) MUM (b) pLSI (c) LDA (d) LDA$_{asym(\theta)}$

Figure 3.3: Four topic models with different assumptions.

the dependencies from a topic z to words w's in a document. The outermost plate represents all the variables related to a specific document. The M in the corner of the plate indicates that the variables inside are repeated M times. The inner plate represents N words belonging to each document. That is, there are M documents and each document has a single topic z affecting N words of each document. Here, z is a distribution over words. Thus N words are drawn independently from the conditional multinomial $p(w|z)$. The probability of the document $\mathbf{w}_d$ is then defined as follows.

$$p(\mathbf{w}_d) = \sum_z p(z) \prod_{i=1}^{N} p(w_{id}|z).$$

The mixture model makes each document exhibit exactly one topic. However, this constraint is often too rigid to effectively model a large collection of documents.

The pLSI model relaxes the simplifying assumption in the mixture of uni-

grams model. Instead of assuming that each document is generated from only one topic, it allows a document to contain multiple topics. As shown in Figure 3.3(b), each document has N topics, z's corresponding to w's. Thus words in a document can be drawn from different topics. The probability of a document is given as follows.

$$p(\mathbf{w}_d) = p(d) \sum_z \prod_{i=1}^{N} p(w_{id}|z)p(z|d).$$

The probability, $p(z|d)$ becomes the mixture weights of topics for a document d. Since d is a dummy index into the list of documents in the training set, the model learns the topic mixtures only for those documents used for training. Hence, there is no natural way to assign probability to a previously unseen document.

Another difficulty with pLSI is that the number of parameters which must be estimated grows linearly by the number of training documents. Let V be the number of vocabularies. The parameters for a T-topic pLSI model are then T multinomial distributions of size V and M mixtures over the T topics. Since the number of parameters becomes $TV + TM$, it grows linearly with M. The linear growth in parameters is prone to lead to overfitting.

LDA overcomes both of the problems by treating the model parameters as hidden random variables. In LDA, $p(w|z)$ becomes a set of T multinomial distributions φ over the V words, such that $p(w|z = j)=\varphi_w^{(j)}$, and $p(z|d)$ is a set of T multinomial distributions θ over the T topics, such that for a word in document d, $p(z = j|d) = \theta_j^{(d)}$. To provide a complete generative model for documents, it combines pLSI with symmetric Dirichlet prior distributions on θ and φ as shown in Figure 3.3(c). The α and β are hyperparameters for the

priors on θ and φ respectively. This generative model enables new documents to be produced from just a set of topics φ and it allows φ to be estimated without requiring the estimation of θ. Hence, only T topics should be learned from training documents. The number of parameters required then becomes TV and the parameters do not grow with M.

LDA is often described as a generative process. The generative process can be interpreted as a pseudo code for an algorithm that randomly generates data according to the model. The generative process for LDA is given as follows.

1. For each topic t,
   (a) Sample $\varphi_t|\beta \sim Dir(\beta)$.
2. For each document d,
   (a) Sample
   $$\theta_d|\alpha \sim Dir(\alpha). \tag{3.1}$$
   (b) For each word $w_{di}^p \in \mathbf{w}_d$,
      i. Sample $z \sim Multi(\theta_d)$.
      ii. Sample $w_{di} \sim Multi(\varphi_z)$.

The probability of a document is then defined as follows.

$$p(\mathbf{w}_d|\alpha,\beta) = \iint p(\varphi|\beta)p(\theta_d^p|\alpha)\prod_{i=1}^{N}\sum_{z=1}^{T}p(w_{id}|\varphi,z)p(z|\theta_d)\mathrm{d}\varphi\mathrm{d}\theta_d.$$

The last model in Figure 3.3(d) is an alternative version of LDA, which is denoted as $\text{LDA}_{asym(\theta)}$. The model has asymmetric Dirichlet priors for θ instead of symmetric priors. The generative process of $\text{LDA}_{asym(\theta)}$ is thus defined by replacing $Dir(\alpha)$ in Equation 3.1 with $Dir(\theta'\alpha)$. Wallach et al. [73] showed that an asymmetric Dirichlet prior over θ increase the robustness of topic modeling to variations in the number of topics and to the highly skewed word frequency distributions common in natural language. This dissertation employs $\text{LDA}_{asym(\theta)}$ to model single directional influences from one data source to another.

Collapsed Gibbs Sampling

A set of topics φ for LDA becomes one of maximizing $p(\mathbf{w}|\varphi)$ which is given as follows.

$$p(\mathbf{w}|\varphi) = \prod_{d=1}^{D} \int_{\theta_d} \prod_{i=1}^{N} \sum_{z=1}^{K} p(w_{id}|\varphi_z) p(z|\theta_d) d\theta_d,$$

where $p(\varphi_z)$ and $p(\theta_d)$ are $Dir(\alpha)$ and $Dir(\beta)$ distributions respectively. However, direct maximization of $p(\mathbf{w}|\varphi)$ is computationally intractable due to coupling between φ_z and θ_d in the summation over latent topics z's. The φ is thus estimated by using approximations such as variational inference [6], expectation propagation [54], or collapsed Gibbs sampling [30]. This dissertation chooses the collapsed Gibbs sampling for estimating topics because it is easy to implement and it converges efficiently [30].

The collapsed Gibbs sampling determines $\mathbf{z}$, the assignments of words to topics first. Estimations of φ and θ are obtained using the topic assignments. Algorithm 1 describes the sampling procedure to determine $\mathbf{z}$. The algorithm

Algorithm 1 Collapsed Gibbs sampling algorithm for LDA

1: **procedure** CGS($\mathbf{w}$,K)
2: $\mathbf{w}$: a set of documents, $\{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_D\}$
3: K: the number of desired topics
4: **begin**
5: **for all** documents $\mathbf{w}_d \in \mathbf{w}$ **do**
6: **for all** words $w_{id} \in \mathbf{w}_d$ **do**
7: $z_{id} = k \sim Multi(\frac{1}{K})$
8: **end for**
9: **end for**
10: **while** More samples are needed **do**
11: **for all** documents $\mathbf{w}_d \in \mathbf{w}$ **do**
12: **for all** words $w_{id} \in \mathbf{w}_d$ **do**
13: $z_{id} \sim p(z_{id}|\mathbf{w}, \mathbf{z}_{-id})$
14: **end for**
15: **end for**
16: **end while**
17: **end procedure**

returns $\mathbf{z}$, topic assignments by the conditional distribution $p(z_{id}|\mathbf{w}, \mathbf{z}_{-id})$, where $\mathbf{z}_{-id}$ is the assignments of all z's except z_{id}. First, the algorithm initializes $\mathbf{z}$ by sampling uniformly from Line 5 to Line 9. Then, $\mathbf{z}$ is updated by sampling its elements repeatedly until all the elements converge. This is described from Line 10 to Line 16 in Algorithm 1.

The conditional posterior distribution for z_{id} is given by

$$p(z_{id} = k|\mathbf{w}, \mathbf{z}_{-id}) \propto p(w_{id}|z_{id} = k, \mathbf{z}_{-id}, \mathbf{w}_{-id})p(z_{id}|\mathbf{z}_{-id}). \tag{3.2}$$

This is an application of Bayes' rule, where the first term on the right hand side is a likelihood, and the second is a prior.

The first term of Equation 3.2 can be rewritten by integrating over φ_k, the multinomial distribution over words associated with topic k. This is given by

$$p(w_{id}|z_{id} = k, \mathbf{z}_{-id}, \mathbf{w}_{-id}) = \int_{\varphi_k} p(w_{id}|z_{id} = k, \varphi_k)p(\varphi_k|\mathbf{z}_{-id}, \mathbf{w}_{-id})d\varphi_k. \tag{3.3}$$

Here, the first term on the right hand side is just $\varphi_{w_{id},k}$. The rightmost term is obtained from Bayes' rule as follows.

$$p(\varphi_k|\mathbf{z}_{-id}, \mathbf{w}_{-id}) \propto p(\mathbf{w}_{-id}|\varphi_k, \mathbf{z}_{-id})p(\varphi_k).$$

Since $p(\varphi_k)$ is $Dir(\beta)$ and conjugate to $p(\mathbf{w}_{-id}|\varphi_k, \mathbf{z}_{-id})$, the posterior distribution $p(\varphi_k|\mathbf{z}_{-id}, \mathbf{w}_{-id})$ becomes $Dir(n^{(w)}_{-id,k} + \beta)$, where $n^{(w)}_{-id,k}$ is the frequency of word w assigned to topic k, not including the current word w_{id}. Then, Equation 3.3 is given as follows.

$$p(w_{id}|z_{id} = k, \mathbf{z}_{-id}, \mathbf{w}_{-id}) = \frac{n^{(w_{id})}_{-id,k} + \frac{\beta}{W}}{n^{(\cdot)}_{-id,k} + \beta},$$

where $n^{(\cdot)}_{-id,k}$ is the total number of words assigned to topic k, not including the current one and W is the number of vocabularies.

The second term $p(z_{id} = k|\mathbf{z}_{-id})$ of Equation 3.2 can be derived in the same way. By integrating over θ_d, the topic proportion for document d, $p(z_{id} =$

$k|\mathbf{z}_{-id})$ is given as follows.

$$\begin{aligned} p(z_{id} = k|\mathbf{z}_{-id}) &= \int_{\theta_d} p(z_{id} = k|\theta_d) p(\theta_d|\mathbf{z}_{-id}) d\theta_d \\ &= \frac{n^{(d)}_{-id,k} + \frac{\alpha}{T}}{n^{(d)}_{-id,\cdot} + \alpha}. \end{aligned} \tag{3.4}$$

Here, $n^{(d)}_{-id,k}$ is the number of words from document d assigned to topic j, not including the current one, and $n^{(d)}_{-id,\cdot}$ is the total number of words in document d, not including the current one. Then, from Equation 3.3 and Equation 3.4, the conditional probability for z_{id}'s is obtained as follows.

$$\begin{aligned} p(z_{id} = k|\mathbf{w}, \mathbf{z}_{-id}) &\propto \frac{n^{(w_{id})}_{-id,k} + \frac{\beta}{W}}{n^{(\cdot)}_{-id,k} + \beta} \times \frac{n^{(d)}_{-id,k} + \frac{\alpha}{T}}{n^{(d)}_{-id,\cdot} + \alpha} \\ &\propto \frac{n^{(w_{id})}_{-id,k} + \frac{\beta}{W}}{n^{(\cdot)}_{-id,k} + \beta} \times \left(n^{(d)}_{-id,k} + \frac{\alpha}{T} \right). \end{aligned}$$

Since the denominator of Equation 3.4 is common for all k's, the last term is induced. Finally, φ and θ are estimated by

$$\begin{aligned} p(\varphi_k|\mathbf{z}, \mathbf{w}, \beta) &= \frac{1}{\mathcal{N}_{\varphi_k}} p(\mathbf{w}|\varphi_k) p(\varphi_k|\beta) = Dir(n_k + \beta), \\ p(\theta_d|\mathbf{z}, \alpha) &= \frac{1}{\mathcal{N}_{\theta_d}} p(\mathbf{z}|\theta_d) p(\theta_d|\alpha) = Dir(n^{(d)} + \alpha), \end{aligned}$$

where n_k provides the counts of words assigned topic k and $n^{(d)}$ gives the topic counts assigned to words in document d. $\mathcal{N}_{\varphi_k}$ and $\mathcal{N}_{\theta_d}$ are normalization factors. In detail, φ and θ is given by

$$\begin{aligned} \varphi_{k,w} &= \frac{n^{(w)}_k + \frac{\beta}{W}}{n^{(\cdot)}_k + \beta}, \\ \theta_{d,k} &= \frac{n^{(d)}_k + \frac{\alpha}{T}}{n^{(d)}_\cdot + \alpha}, \end{aligned}$$

where $n_k^{(w)}$ is the count of the word w assigned to topic k and $n_k^{(\cdot)}$ is the total number of words assigned to topic k. $n_k^{(d)}$ is the number of words assigned to topic k in the document d and $n_{\cdot}^{(d)}$ is the total number of words in the document d.

3.1.3 Data Representation

Data representation is one of key factors in understanding social interactions. We follow the common representation of some recent work on human activity and interaction modeling [19, 25, 39, 77]. That is, the interaction log of a user is represented as a bag of words. Thus, the interaction order is not considered in this representation, and a set of specific interactions between a user and another user is simply expressed as a vector, in which elements represent the frequency of interactions within a specific time slot. That is, an interaction within a specific time slot is regarded as a word, and thus a bag of interactions becomes a document.

Let w_{ji} be the j-th word that represents an interaction with another user i. Then, an interaction document with user i is denoted by $\mathbf{w}_i = \{w_{1i}, w_{2i}, ..., w_{Ni}\}$, where N is the total number of interactions with user i. Therefore, the mobile log of a user is given as

$$L = \{\mathbf{w}_1^p, \mathbf{w}_1^c, \mathbf{w}_2^p, \mathbf{w}_2^c, \ldots, \mathbf{w}_D^p, \mathbf{w}_D^c\}\,, \tag{3.5}$$

where p and c indicate proximity and call respectively, and D is the number of other users with which the target user interacts. This log is divided into two

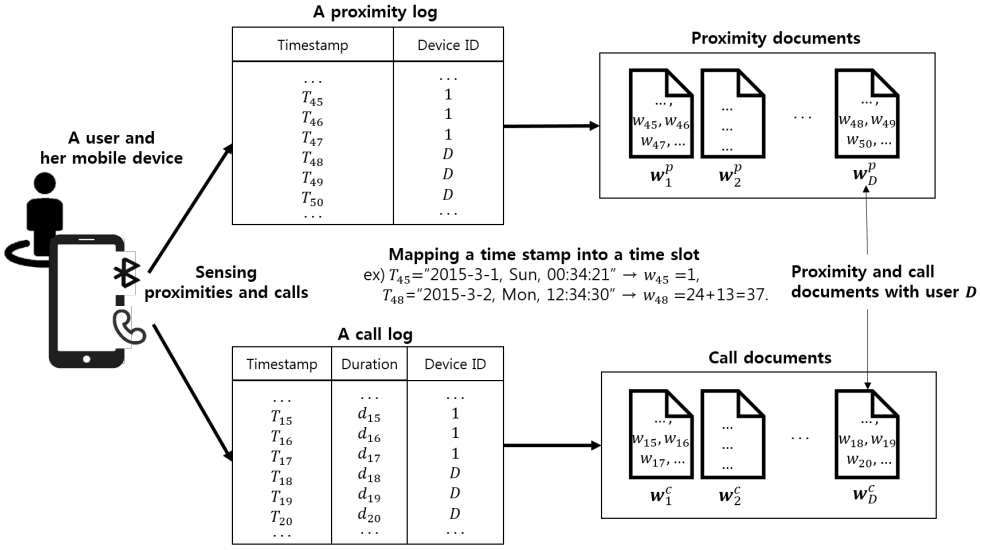


Figure 3.4: Generation of call and proximity documents from a user's call and proximity logs.

subsets according to interaction type. That is,

$$L^p = \{\mathbf{w}_1^p, \mathbf{w}_2^p, \ldots, \mathbf{w}_D^p\} \subset L, \tag{3.6}$$

$$L^c = \{\mathbf{w}_1^c, \mathbf{w}_2^c, \ldots, \mathbf{w}_D^c\} \subset L. \tag{3.7}$$

Figure 1 illustrates how interaction documents are generated from proximity and call logs of a specific user. A Bluetooth device embedded in the user's mobile phone senses nearby other users who interact with the target user. The timestamps of sensing moment and the device ID of the interacting user are recorded at a proximity log. The timestamps are then encoded into words which correspond to one of predefined time slots. A time slot is an hour of seven days. For instance, a timestamp, "2015-3-1, Sun, 00:34:21" is encoded as 1, since the timestamp falls into the first time slot (0 AM $\sim$ 1 AM of

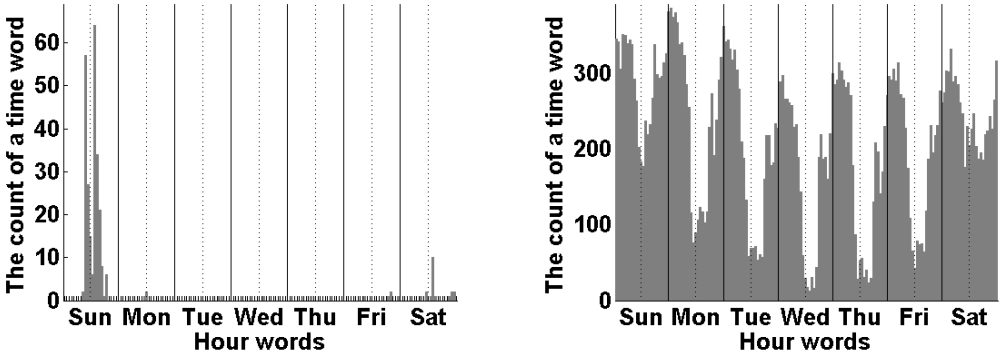
Sunday). In the same way, "2013-3-2, Mon, 12:34:30" becomes 37. Since D users interact with the target user, D proximity documents are generated by grouping words according to device ID. Call documents are also generated in a similar way except that a call is logged as a duration rather than a moment. Thus, a single duration is first discretized, and then is encoded into one or more words.

A time slot was defined as an hour of weekdays or weekends in the conference paper [31]. This is because daily routines are cyclic at weekdays, but are observed differently on weekends [24, 48]. That is, there are only 48 distinct words, where the first 24 words come from weekdays and the remaining 24 words come from weekends. However, in this dissertation, the days in a week are subdivided into hours. As a result, there are 24×7 distinct words, where 24 words are associated with each day of the week. Therefore, a time slot is expressed at a more specific level than our previous work.

3.1.4 Applying of a Topic Model

This dissertation utilizes a topic model for finding social interaction patterns. Especially, LDA is used as a basic model for this task. This chapter describes the results on applying LDA to finding social interaction patterns. *User112* is chosen from the Friends and Family data set [1] and only her proximity log is used for easy description.

The proximity log is first converted into a set of proximity documents as explained in the previous chapter. As a result, 13 documents are obtained from *user112*. Since a proximity document is a bag of unordered hour words,



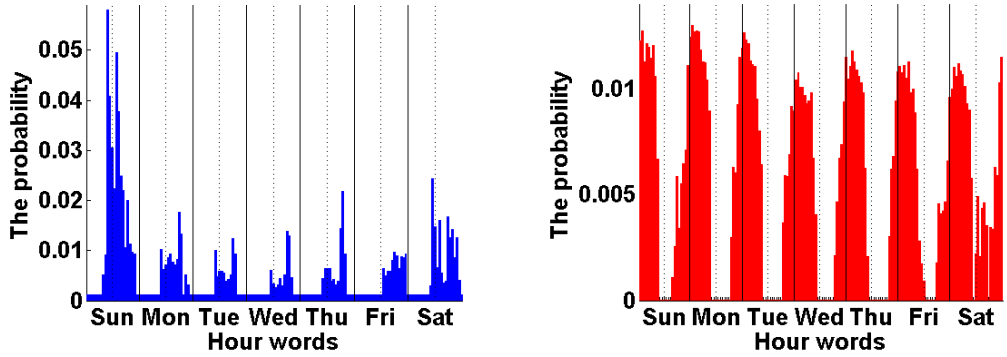
(a) A proximity document with *user1*

(b) A proximity document with *user10*

Figure 3.5: Visualization of two proximity documents from *user112* of the Friends & Family data set

it can be represented in a histogram plotting the count of its hour words. Figure 3.5 illustrates two proximity documents from *user112*. Figure 3.5(a) depicts the proximity document associated with *user1* and Figure 3.5(b) is that associated with *user10*. As shown in 3.5(a), the meetings between *user112* and *user1* were occurred mostly during the daytime on Sunday. These interactions are in general observed between friends. On the other hand, the meetings between *user112* and *user10* were occurred mainly from one evening to the next morning since they are the same family members.

What is expected from proximity documents of these characteristics is to find typical patterns between friends and also between family members. To verify the results, all the proximity documents are modeled using LDA with the topic number of 2 and topics are learned by the collapsed Gibbs sampling as explained in Chapter 3.1.2. Note that the topics draws probabilities

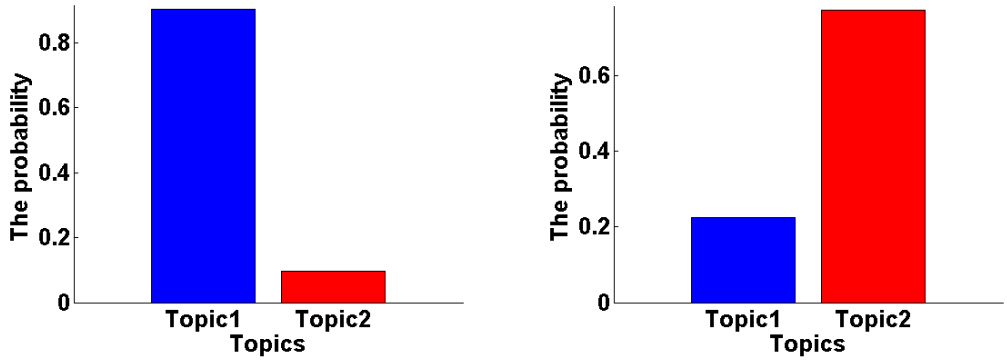


(a) Topic1 (b) Topic2

Figure 3.6: Topics from *user112* of the Friends & Family data set

of meetings over hour words. Hence topics are regarded as social interaction patterns. Figure 3.6 depicts two proximity topics for *user112*. Figure 3.6(a) shows interactions during the daytime. Especially, meetings on Friday and Sunday evening. In contrast, interactions from one morning to the next evening are shown in Figure 3.6(b). The results agree with the expectation for social interaction patterns.

The proximity topics govern the observed hour words in the proximity documents. A topic proportion from each document describes how much each topic influences the document. Figure 3.7 depicts two topic proportions for documents associated with *user1* and *user10*, respectively. Note that the proximity document with *user1* shows interactions mainly during the daytime on Sunday as shown in Figure 3.5(a). Hence the topic1 influences this document more than the topic2 as shown in Figure 3.7(a). In contrast, the proximity document with *user10* emphasizes interactions from one evening to the next morning as



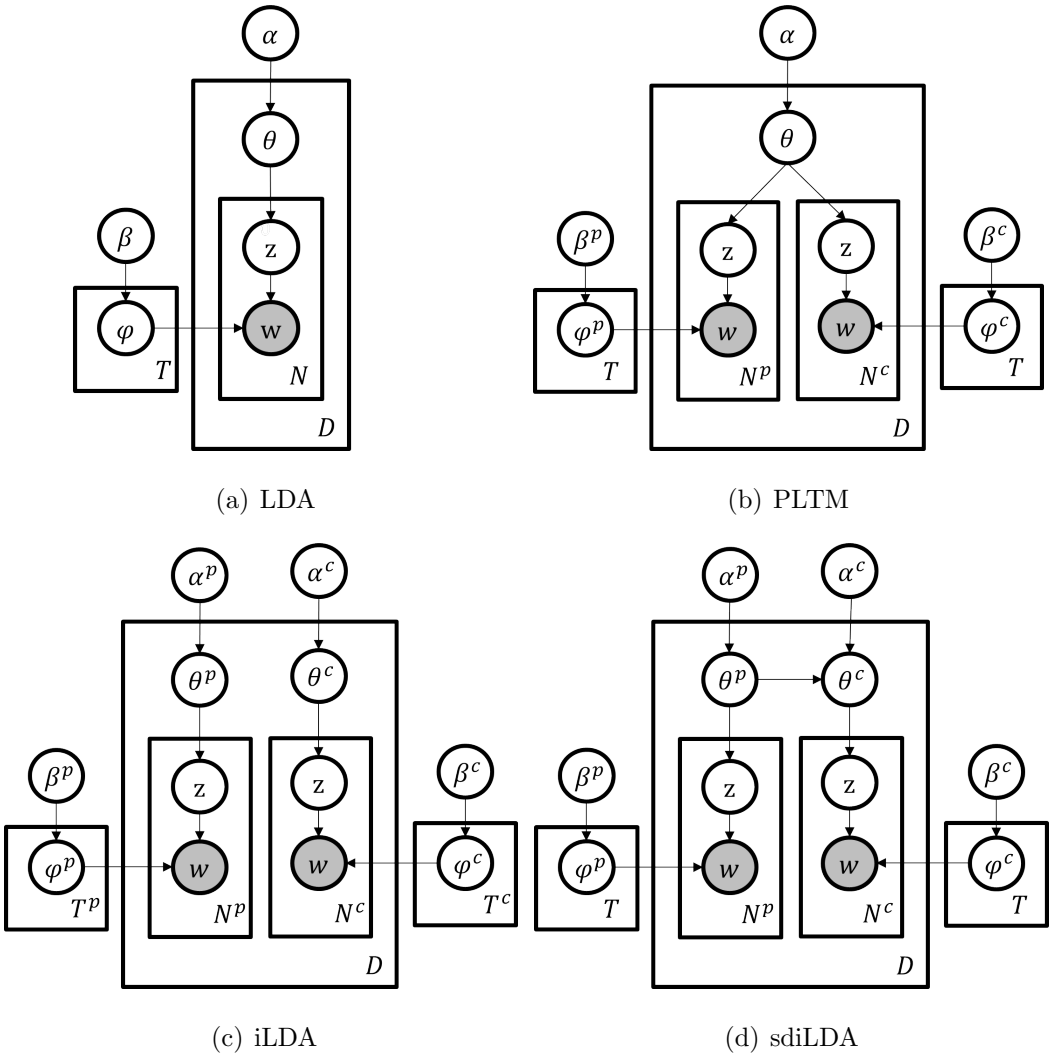
(a) The topic proportion with *user1* (b) The topic proportion with *user10*

Figure 3.7: Topic proportions from *user112* of the Friends & Family data set

shown in Figure 3.5(b). This fact results in the higher proportion of topic2 than topic1 as shown in Figure 3.7(b).

3.2 Topic Models Using Call and Proximity Logs Simultaneously

This chapter discusses topic models using call and proximity logs simultaneously. All topic models discussed in this dissertation are based on LDA [6]. Figure 3.8 depicts graphical representations of LDA and its three extensions for finding social interaction patterns. Figure 3.8(a) is LDA, Figure 3.8(b) is PLTM, and Figure 3.8(c) is independent LDA (iLDA). Figure 3.8(d) is single-directional influence LDA (sdiLDA) which is our final model.



(a) LDA (b) PLTM

(c) iLDA (d) sdiLDA

Figure 3.8: Graphical representation of four topic models for finding social interaction patterns.

3.2.1 Latent Dirichlet Allocation (LDA)

LDA employs two sets of latent variables to model documents as shown in Figure 3.8(a). One is a set of discrete-valued hidden variables z_i's. A hidden variable z_i assigns the i-th word of a document to one of T topics. The other is a set of latent variables θ_d's. A latent variable θ_d represents the influence proportion of the topics in the d-th document. The proportion of topics for each document is modeled as a Dirichlet distribution, while each topic is a multinomial distribution over words.

In order to employ proximities and calls simultaneously for identifying interaction patterns with LDA, proximities and calls should be regarded as the same information type. That is, they are forced to be governed by a common set of topics. This is performed by unifying proximity and call documents between two users into a single document. Then, the interaction patterns of a user are inferred as topics from the user's mobile log. Since this modeling does not distinguish calls and proximities, there is no explicit relationship between them.

3.2.2 Polylingual Topic Model (PLTM)

The calls between two office colleagues are often related to their work, and so are the proximities between them. Thus, their calls and proximities are understood to be related through their work, although the calls and proximities are not directly related. One method to model such a relationship between calls and proximities is to assume the correlations between topics of calls and those of proximities. This is a design principle behind PLTM [52]. PLTM extends

LDA to describe the joint distribution of loosely equivalent documents written in different languages. Even if most two words from different languages do not co-occur in a document, two documents written in different languages can describe almost the same content (topics) such as Wikipedia articles linked by different languages. Thus, PLTM takes the topic proportion θ as a common factor to correlate the topics of documents in different languages as shown in Figure 3.8(b).

The mobile log of a target user can be modeled by PLTM with ease. The modeling can be performed by regarding calls and proximities as two different languages. In this modeling, the same topic proportion governs both call and proximity documents observed between the target user and another user. Thus, it is assumed that the topics of a proximity document have a one-to-one mapping to those of a call document. Subsequently, interaction patterns are inferred as a set of topic pairs for calls and proximities.

PLTM is more flexible than LDA, because PLTM does not capture only correlations among topics, but also correlations of observed data through the topics. However, its assumption is too rigid to be used for modeling a mobile log. Proximities occur often during a long discussion, whereas most calls occur during a relatively short conversation. In most actual mobile logs, the number of proximities in a mobile log overwhelms that of calls, and proximities are observed more regularly than calls [24]. Thus, the topic proportion for a proximity document can get different from that of a call document. However, this is against the assumption that the topic proportion of both documents in different languages would be the same. That is, PLTM fails in managing the different topic proportions of call and proximity documents despite its flex-

ibility. In addition, when the number of topics is fixed, there is no explicit method in PLTM to control the strength of the correlations between calls and proximities.

3.2.3 Independent LDA (iLDA)

When identifying interaction patterns, it is also possible to ignore any dependency between calls and proximities. A new LDA that implements this assumption of independence between calls and proximities is given in Figure 3.8(c), and is referred to as independent LDA (iLDA). Whereas PLTM forces both proximity and call documents to share a single topic proportion θ, iLDA allows two different topic proportions θ^p and θ^c for proximity and call documents, respectively. That is, the topic proportions of call and proximity documents can be different in iLDA. One advantage of iLDA is that a different number of topics are allowed for call and proximity documents. However, iLDA loses the information obtainable when calls and proximities are analyzed simultaneously.

3.3 Modeling Single Directional Influences From Proximities to Calls

3.3.1 Single-directional Influence LDA (sdiLDA)

Single directional influence LDA (sdiLDA) shown in Figure 3.8(d) is our final model. It is almost same with iLDA except for the independent relationship between topic proportions for proximity and call documents. A topic

proportion for a call document depends on a topic proportion of a proximity document in sdiLDA, when the two documents are observed between two same users. Thus, the proximity document influences the call document, but not vice versa. The two topic proportions are actually topic distributions for the call and proximity documents. Then, the call topic distribution takes the proximity topic distribution as its prior by assuming that calls and proximities are a homogeneous information type. As a result, topic proportions for call documents are not trained only from the call documents, but also from topic proportions of proximity documents. Because the topics from a proximity topic proportion are mapped one-to-one onto those from a call topic proportion, the interaction patterns are inferred as a set of topic pairs of calls and proximities.

Let $\mathbf{w}_d^p = \{w_{1d}^p, w_{2d}^p, ..., w_{Nd}^p\}$ be a proximity document of N words expressing proximity interactions with a user d, and $\mathbf{w}_d^c = \{w_{1d}^c, w_{2d}^c, ..., w_{Md}^c\}$ be a call document of M words expressing call interactions with the same user d. Then, the generative process of sdiLDA for $\mathbf{w}_d^p$ and $\mathbf{w}_d^c$ is given as follows.

1. Sample $\theta_d^p|\alpha^p \sim Dir(\alpha^p)$.
2. For each hour word $w_{nd}^p \in \mathbf{w}_d^p$,
   (a) Sample $z \sim Multi(\theta_d^p)$.
   (b) Sample $w_{nd}^p \sim Multi(\varphi_z^p)$.
3. Sample

$$\theta_d^c|\alpha^c\theta_d^p \sim Dir(\alpha^c\theta_d^p). \tag{3.8}$$

4. For each hour word $w_{md}^c \in \mathbf{w}_d^c$,

(a) Sample $z \sim Multi(\theta_d^c)$.

(b) Sample $w_{md}^c \sim Multi(\varphi_z^c)$.

Here, $Dir(\cdot)$ and $Multi(\cdot)$ are a Dirichlet distribution and a multinomial distribution, respectively.

The proximity document $\mathbf{w}_d^p$ is generated first as in LDA. Then, the call document $\mathbf{w}_d^c$ is generated by taking the proximity topic distribution θ_d^p as a prior to the call topic distribution θ_d^c. This is actually performed by giving an asymmetric prior to θ^c, and regarding α^c and θ_d^p as a concentration parameter and a base measure of the asymmetric prior, respectively. Then, the prior of the Dirichlet distribution θ_d^c becomes a multiplication of α^c and θ_d^p as in Equation (3.8). θ_d^p influences θ_d^c with a one-to-one correspondence between their topics. Because α^c is a scalar, it adjusts the strength of the influences from θ_d^p to θ_d^c. The larger α^c is, the closer θ_d^c gets to θ_d^p. If α^c is small, θ_d^c is fitted to the observed data rather than θ_d^p.

3.3.2 Parameter Estimation

Only the documents L^p in Equation (3.6) and L^c in Equation (3.7) are observed explicitly, and they are given as a mobile log L in Equation (3.5). Since topics are regarded as interaction patterns, the latent topic variables φ^p and φ^c of sdiLDA should be estimated from L. Note from Figure 3.8(d) that only θ_d^p and φ^p affect the generation of proximity document $\mathbf{w}_d^p$. Thus, the probability of a proximity document $\mathbf{w}_d^p$ expressing interactions with a user d

is

$$p(\mathbf{w}_d^p|\varphi^p, \theta_d^p) = \prod_{n=1}^{N_d} \sum_{z=1}^{T} p(w_{id}n|\varphi_z^p)p(z|\theta_d^p),$$

where $p(\varphi_z^p)$ and $p(\theta_d^p)$ are $Dir(\beta^p)$ and $Dir(\alpha^p)$ distributions respectively. Then, the likelihood of a set of proximity documents L^p becomes

$$\mathcal{L}(\varphi^p, \theta_d^p|L^p) = \prod_{d=1}^{D} p(\mathbf{w}_d^p|\varphi^p, \theta_d^p p).$$

The optimal φ^p and θ^p are those that maximize the likelihood $\mathcal{L}(\varphi^p, \theta_d^p|L^p)$. However, direct maximization of $\mathcal{L}(\varphi^p, \theta_d^p|L^p)$ is computationally intractable due to coupling between φ^p and θ_d^p in the summation over latent topic z's. Griffiths and Steyvers [30] proposed the collapsed Gibbs sampling as an alternative method to find φ^p and θ^p. In the collapsed Gibbs sampling, only the latent variables denoted by z are sampled from L^p, a set of proximity documents. After the sampler burns out, θ_d and φ^p are estimated from z's.

Let $\mathbf{z}^p$ be topic assignments of all words in L^p and z_{nd}^p be a topic assignment of the n-th word w_{nd}^p in a proximity document $\mathbf{w}_d^p \in L^p$. According to Griffiths and Steyvers [30], the probability that w_{nd}^p is assigned to the j-th topic is given as

$$P(z_{nd}^p = j|\mathbf{z}_{-nd}^p, L^p) \propto \frac{n_j^{(w_{nd}^p)} + \frac{\beta^p}{W} - 1}{n_j^{(\cdot p)} + \beta^p - 1} \left(n_j^{(d^p)} + \frac{\alpha^p}{T} - 1 \right), \tag{3.9}$$

where $\mathbf{z}_{-nd}^p$ is the assignment of words to topics except w_{nd}^p, $n_j^{(w)}$ represents a count when the topic of a word w is j, and $n_j^{(\cdot p)}$ is a total number of words

to which topic j is assigned. Lastly, $n_j^{(d^p)}$ is the number of words in $\mathbf{w}_d^p$ whose topic is j. Then, $\mathbf{z}^p$ is computed by the collapsed Gibbs sampling.

Note that φ^p is a $W \times T$ matrix representing the probabilities of W words generated from T topics, and θ^p is a $D \times T$ matrix representing the probability of generating T topics in D proximity documents. After $\mathbf{z}^p$ is obtained, each element of φ^p and θ^p is estimated as follows.

$$\varphi_{w,j}^p = \frac{n_j^{(w^p)} + \frac{\beta^p}{W}}{n_j^{(\cdot^p)} + \beta^p},$$
$$\theta_{d,j}^p = \frac{n_j^{(d^p)} + \frac{\alpha^p}{T}}{n_\cdot^{(d^p)} + \alpha^p},$$

where $n_\cdot^{(d^p)}$ is the total number of words in $\mathbf{w}_d^p$. That is, φ^p and θ^p are estimated from $\mathbf{z}^p$.

Once θ^p is estimated, it is used as a prior for θ^c. Then, the estimation of φ^c can be calculated using the same method of estimating φ^p by the collapsed Gibbs sampling using L^c, a set of call documents. Thus, the probability that w_{nd}^c is assigned to the j-th topic is given as

$$P(z_{nd}^c = j | \mathbf{z}_{-nd}^c, L^c, \theta^p) \propto \frac{n_j^{(w_{nd}^c)} + \frac{\beta^c}{W} - 1}{n_j^{(\cdot^c)} + \beta^c - 1} \left(n_j^{(d^c)} + \alpha^c \theta_{d,j}^p - 1 \right).$$

Then, the latent variables φ^c and θ^c are estimated from $\mathbf{z}^c$ in the same manner to estimate φ^p and θ^p. That is, they are estimated by

$$\varphi_{w,j}^c = \frac{n_j^{(w^c)} + \frac{\beta^c}{W}}{n_j^{(\cdot^c)} + \beta^c},$$
$$\theta_{d,j}^c = \frac{n_j^{(d^c)} + \alpha^c \hat{\theta}_{d,j}^p}{n_\cdot^{(d^c)} + \alpha^c}. \tag{3.10}$$

In Equation (3.10), if $z_j^c \in \mathbf{z}^c$ is zero, then so is $n_j^{(d^c)}$. It indicates that the probability of generating the topic j becomes $\theta_{d,j}^p$. That is, $n_j^{(d^c)}$ is smoothed with a topic-specific quantity $\theta_{d,j}^p$. Consequently, different topics are more or less probable in all call documents in advance, and this prior is influenced by the topics used in proximity documents corresponding to those in call documents.

3.3.3 Hyperparameters

The performance of sdiLDA is affected by hyperparameters α^p, β^p, and β^c as well as α^c. Grid search is a simple way to find the optimal values of the hyperparameters. However, when each hyperarameter is descretized into K candidate values, K^4 parameter groups should be investigated by grid search, which is computationally infeasible. A simple and stable fixed-point iterative method introduced by Minka [53] is adopted to avoid this problem.

Since a fixed-point iteration provides an optimized hyperparameter to maximize likelihood of data regardless other hyperparameters, the four hyperparameters are learned independently. Thus, the time complexity becomes $O(4 \cdot L)$, where L is the number of iterations for a single hyperparameter. According to Minka [53], the update rule of the concentration parameter α^c is given as

$$\alpha^c \leftarrow \alpha^c \cdot \frac{\sum_{d=1}^{D} \sum_{j=1}^{T} \theta_{d,j}^p \Psi(n_j^{(d^c)} + \alpha^c \theta_{d,j}^p) - \theta_{d,j}^p \Psi(\alpha^c \theta_{d,j}^p)}{\sum_{d=1}^{D} \Psi(n_\cdot^{(d^c)} + \alpha^c) - \Psi(\alpha^c)},$$

where $\Psi(x)$ is the digamma function. $n_j^{d^c}$, $n_\cdot^{d^c}$, and $\theta_{d,j}^p$ are determined in advance as explained in Chapter 3.3.2. Thus, they are constants in this equation.

Note that α^c and θ_d^p compose an asymmetric prior for topic proportion of a call document. We regard α^p as an asymmetric prior of which base measure

is uniform over T topics. $\theta^p_{d,j}$'s are then replaced by $\frac{1}{T}$. Thus, the update rule for α^p becomes

$$\alpha^p \leftarrow \alpha^p \cdot \frac{\sum_{d=1}^{D}\sum_{j=1}^{T}\Psi(n_j^{(d^p)}+\frac{\alpha^p}{T})-\Psi(\frac{\alpha^p}{T})}{T\sum_{d=1}^{D}\Psi(n_{\cdot}^{(d^p)}+\alpha^p)-\Psi(\alpha^p)}, \tag{3.11}$$

where $n_j^{d^p}$ and $n_{\cdot}^{d}$ are constants computed in advance.

In order to determine the optimal values for β^p and β^c, a topic frequency in each document is replaced by a sum of word frequencies in each topic. Then, β^p and β^c can be estimated by fixed-point update rules similar to Equation (3.11). That is,

$$\beta^p \leftarrow \beta^p \cdot \frac{\sum_{j=1}^{T}\sum_{w=1}^{W}\Psi(n_j^{(w^p)}+\frac{\beta^p}{W})-\Psi(\frac{\beta^p}{W})}{W\sum_{j=1}^{T}\Psi(n_j^{(\cdot^p)}+\beta^p)-\Psi(\beta^p)},$$

$$\beta^c \leftarrow \beta^c \cdot \frac{\sum_{j=1}^{T}\sum_{w=1}^{W}\Psi(n_j^{(w^c)}+\frac{\beta^c}{W})-\Psi(\frac{\beta^c}{W})}{W\sum_{j=1}^{T}\Psi(n_j^{(\cdot^c)}+\beta^c)-\Psi(\beta^c)},$$

where $n_j^{(w^p)}$, $n_j^{(\cdot^p)}$, $n_j^{(w^c)}$, and $n_j^{(\cdot^c)}$ are constants.

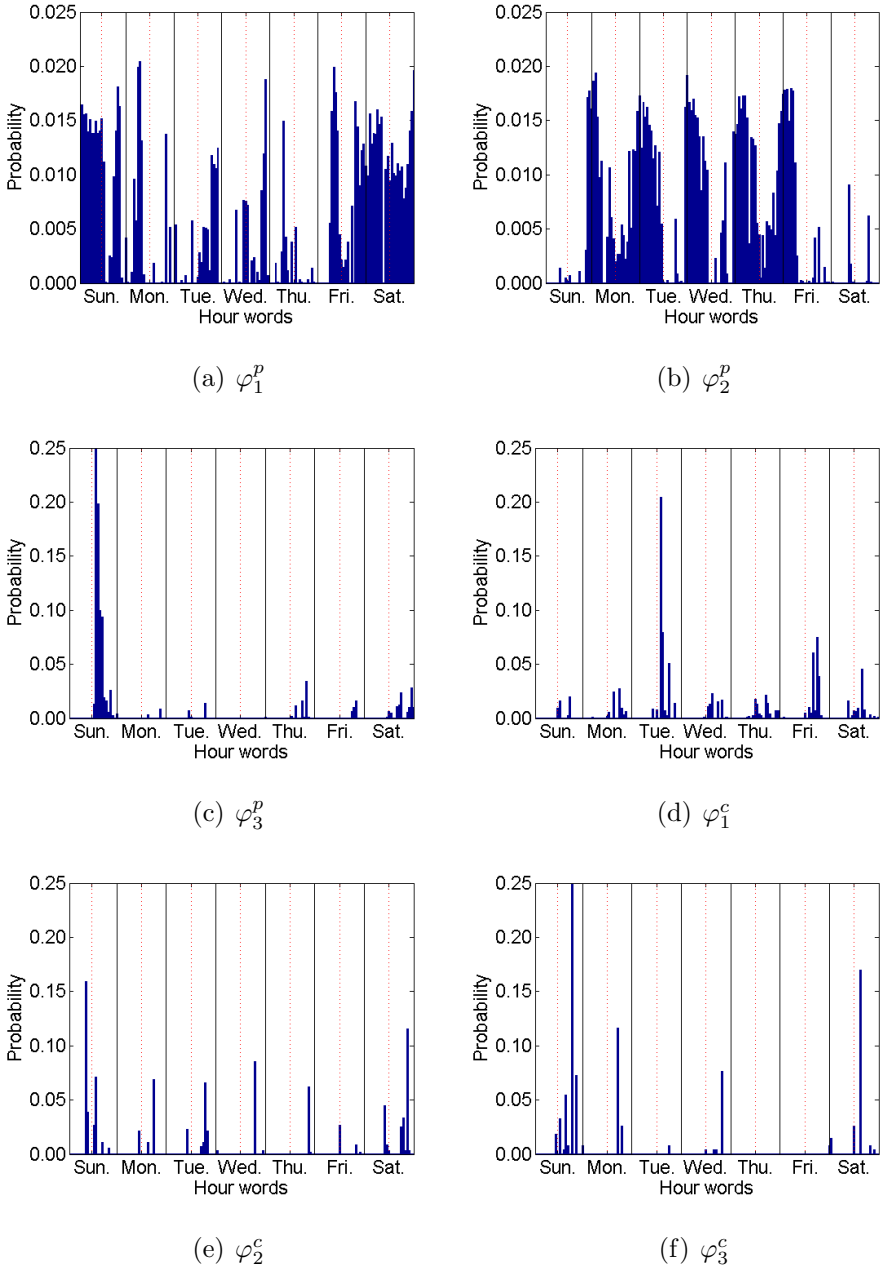
3.4 Examples of Finding Call and Proximity Patterns Simultaneously

Members of a family have, in general, a simple life cycle. For instance, they stay home together from evening to the next morning and spend most of their time together on the weekends. Calls usually occur during the daytime on weekdays since they are far away in that time. To verify social interaction patterns in this respect, *user*112 of Friends & Family data set [1] is chosen and topics by LDA, PLTM, iLDA, and sdiLDA are compared with each other.

For easy comparison among the topics, the number of topics, T is set to a small value, 3. Since only iLDA assumes no relationship between proximities and calls, topics by iLDA is discussed first. LDA infers topics by regarding proximities and calls as the same interaction type. The topics by LDA are then compared with those by iLDA. After that, topics by PLTM and sdiLDA are discussed in terms of modeling correlation between proximities and calls.

Figure 3.9 depicts six topics (three from proximities and another three from calls) inferred by iLDA. The X-axis of each topic graph represents hour words representing time slots, and Y-axis is probability of interactions during a time slot. Figure 3.9(a), Figure 3.9(b), and Figure 3.9(c) depicts three proximity topics, φ_1^p, φ_2^p, and φ_3^p respectively, while Figure 3.9(d), Figure 3.9(e), and Figure 3.9(f) are three call topics, φ_1^p, φ_2^p, and φ_3^p respectively. Figure 3.9(a) mainly draws a meeting distribution on weekends, and Figure 3.9(b) draws a meeting distribution from evening to the next morning during weekdays. Figure 3.9(c) shows a prominent meeting at Sunday noon. On the other hand, call patterns are sparsely drawn as seen in Figure 3.9(d), Figure 3.9(e), and Figure 3.9(f). Thus, calls at several specific hour words are prominent in each distribution. For instance, calls on Tuesday noon are prominent in Figure 3.9(d). These results reflect well our intuition on family members' life cycle.

Figure 3.10 depicts topics by LDA. Since LDA is trained by unifying proximity and call documents, there are only three topics in Figure 3.10. Figure 3.10(a) draws interactions on weekends and looks similar to the iLDA topic in Figure 3.9(a). The LDA topics in Figure 3.10(b) and Figure 3.10(c) are also similar to the iLDA topics in Figure 3.9(b) and Figure 3.9(c) respectively. This is because the number of proximities overwhelms that of calls. Thus LDA top-

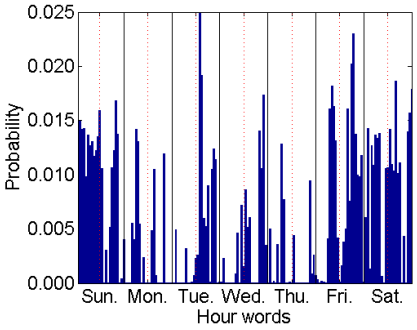


(a) φ_1^p (b) φ_2^p

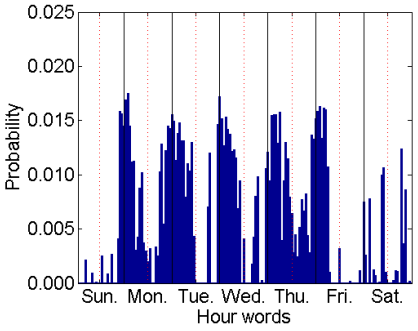
(c) φ_3^p (d) φ_1^c

(e) φ_2^c (f) φ_3^c

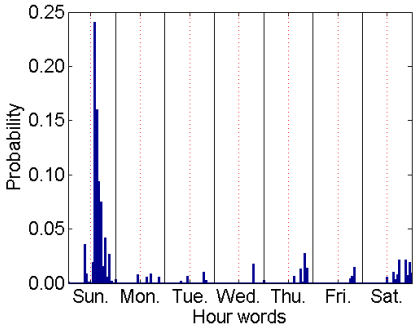
Figure 3.9: iLDA topics for *user112* with $T = 3$.



(a) φ_1



(b) φ_2

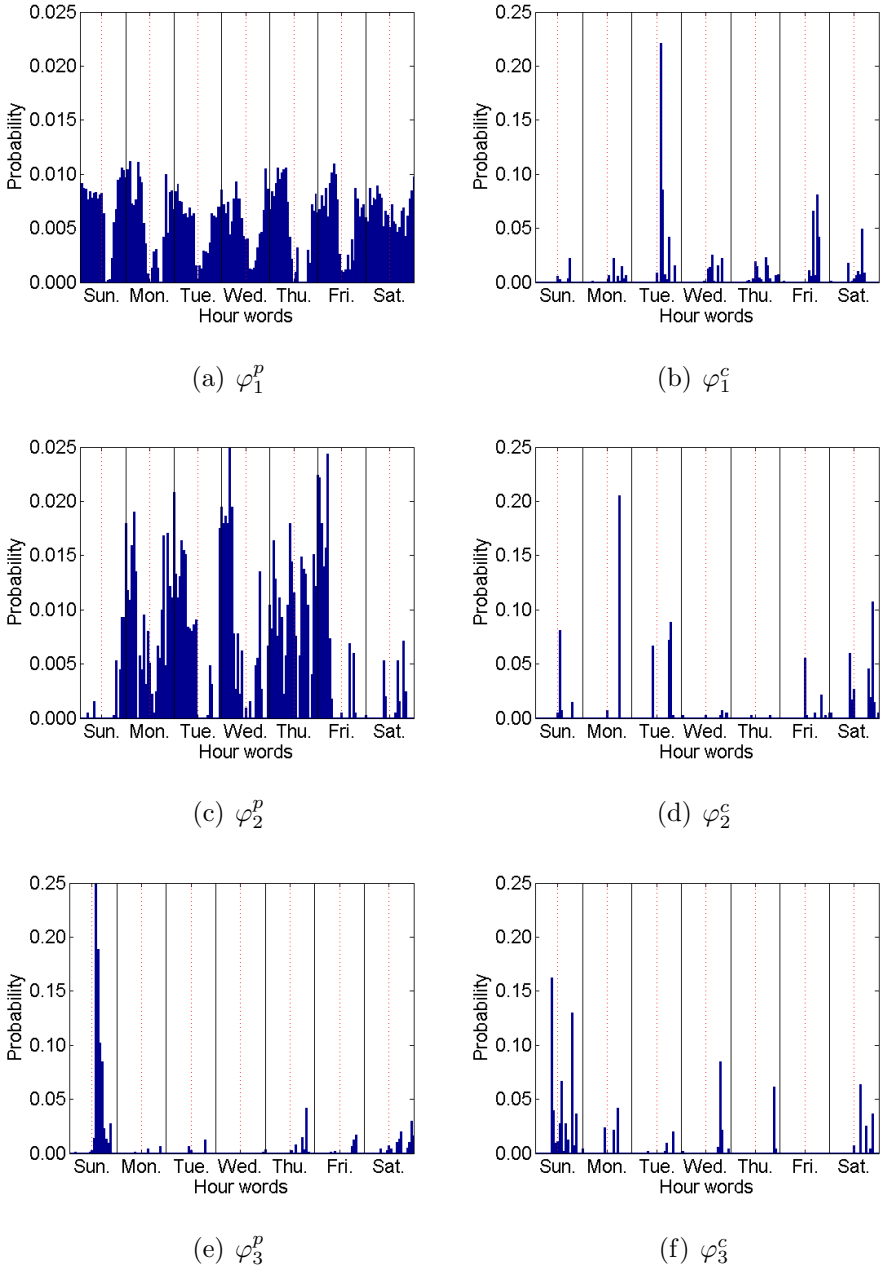


(c) φ_3

Figure 3.10: LDA topics for *user112* with $T = 3$.

ics are mainly governed by proximities. As a result, LDA fails to capturing call patterns, whereas iLDA provides proximity and call patterns separately. However, relationships between proximity and call topics are unknown in iLDA.

PLTM and sdiLDA provide pairs of proximity and call topics because they model explicitly the dependency between proximity and call topics as explained in Chapter 3.2.2 and Chapter 3.3.1. Figure 3.11 and Figure 3.12 depict topics inferred by PLTM and sdiLDA respectively. In these figures, topic pairs are aligned horizontally. For instance, Figure 3.11(a) and Figure 3.11(b) are a pair

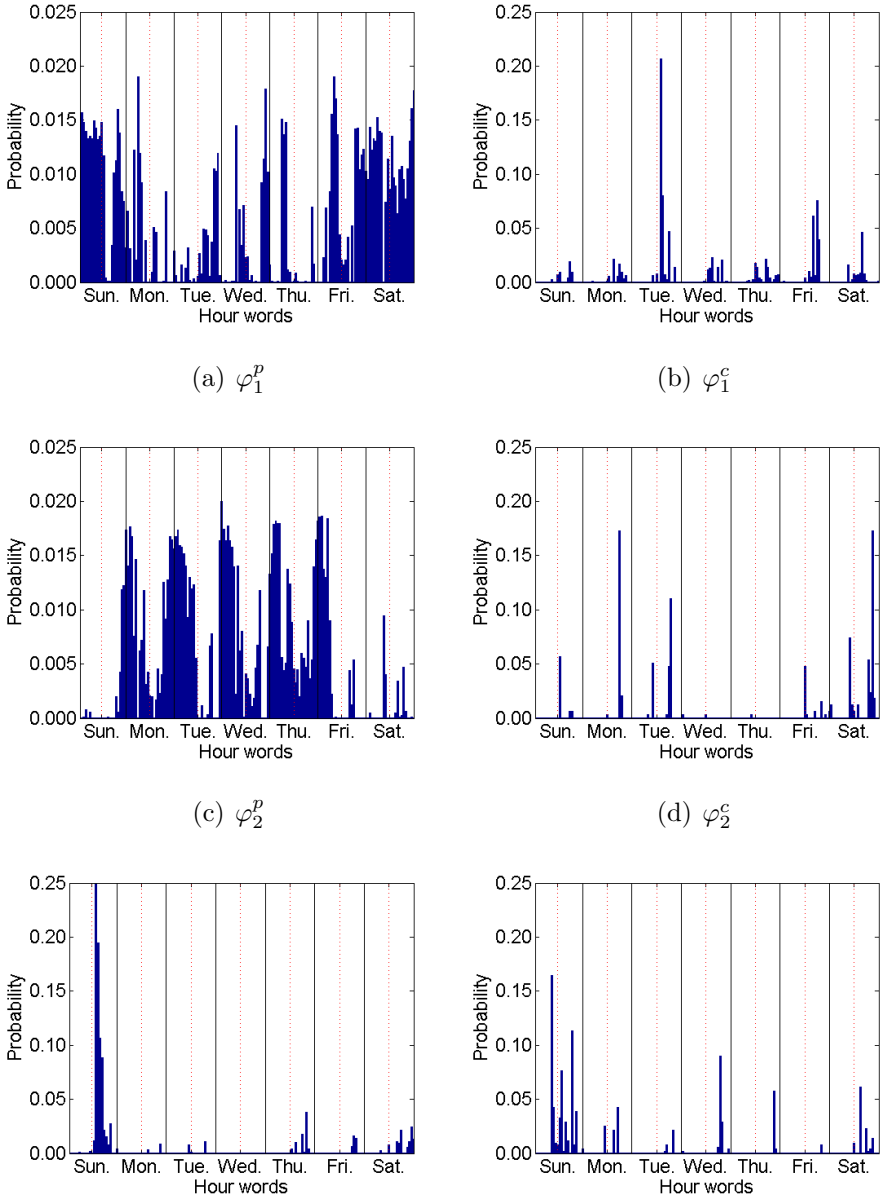


(a) φ_1^p (b) φ_1^c

(c) φ_2^p (d) φ_2^c

(e) φ_3^p (f) φ_3^c

Figure 3.11: PLTM topics for *user112* with $T = 3$.



(a) φ_1^p (b) φ_1^c

(c) φ_2^p (d) φ_2^c

(e) φ_3^p (f) φ_3^c

Figure 3.12: sdiLDA topics for *user112* with $T = 3$.

of proximity and call topics, φ_1^p and φ_1^c.

There are three things to note in these figures. First, PLTM and sdiLDA provides almost same call topics, and the call topics are different from those by iLDA. That is, Figure 3.11(b), Figure 3.11(d), and Figure 3.11(f) are almost same with Figure 3.12(b), Figure 3.12(d), and Figure 3.12(f) respectively. Though Figure 3.11(b) is almost same with Figure 3.9(d), the other call topics by PLTM and iLDA draws different distributions as seen in Figure 3.9(e), Figure 3.9(f), Figure 3.12(d), and Figure 3.12(f). This is because that call topics of sdiLDA and PLTM are affected by proximity topics while call topics of iLDA are inferred only from call documents.

Second, sdiLDA and iLDA provides almost same proximity topics, while proximity topics by PLTM and sdiLDA draw different interactions except φ_3^p. Figure 3.12(a), Figure 3.12(c), and Figure 3.12(e) are almost same with Figure 3.9(a), Figure 3.9(b), and Figure 3.9(c) respectively. In contrast, Figure 3.11(a) is completely different from Figure 3.12(a). Though both Figure 3.11(c) and Figure 3.12(c) draws meeting distributions on weekdays, Figure 3.12(c) is more cyclic than Figure 3.11(c). These results are natural since proximity topics are inferred by sdiLDA in the same way of iLDA. However, proximity topics by PLTM are affected by call topics, which results in drawing different interactions from sdiLDA.

Lastly, only PLTM provides completely different proximity topics from other models. As mentioned above, PLTM provides completely different proximity topics compared with sdiLDA, while proximity topics by sdiLDA and iLDA are almost same. LDA also draws topics similar to proximity topics of iLDA. These results imply that influences of a small number of calls to proximity topics are

stronger in PLTM than that in LDA, iLDA, and sdiLDA. The results prove that the assumption of PLTM is too rigid to be used for modeling a mobile log.

3.5 Perplexity for Topic Models

Perplexity is used as a measure of interaction pattern quality. It is widely used to measure the fitness of topic models for a test data set [6, 7, 60]. The perplexity decreases monotonically in the likelihood of a test data set. Thus, the lower the perplexity, the better the generalization performance of the model. According to Blei et al. [6], when L_{test}, a test set of M documents is given, the perplexity of a probabilistic topic model is measured by

$$Perplexity(L_{test}) = \exp\left(-\frac{\sum_{d=1}^{M}\log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right), \tag{3.12}$$

where N_d is the number of words in $\mathbf{w}_d$.

In LDA, the probability $p(\mathbf{w}_d)$ is given as

$$\begin{aligned} p(\mathbf{w}_d) &= \sum_{i=1}^{N_d}\sum_{j=1}^{T} p(w_{d_i}|j)p(j|d) \\ &= \sum_{i=1}^{N_d}\sum_{j=1}^{T} \varphi_{w_{d_i},j}\theta_{d,j}. \end{aligned}$$

In PLTM, θ is shared by $\mathbf{w}_d^p$ and $\mathbf{w}_d^c$. Thus, $p(\mathbf{w}_d^p)$ and $p(\mathbf{w}_d^c)$ become

$$\begin{aligned} p(\mathbf{w}_d^p) &= \sum_{i=1}^{N_d}\sum_{j=1}^{T} \varphi^p_{w_{d_i},j}\theta_{d,j}, \\ p(\mathbf{w}_d^c) &= \sum_{i=1}^{N_d}\sum_{j=1}^{T} \varphi^c_{w_{d_i},j}\theta_{d,j}. \end{aligned}$$

In iLDA and sdiLDA, φ^p and θ^p exist for proximity documents, and φ^c and θ^c exist for call documents. Thus, $p(\mathbf{w}_d^p)$ and $p(\mathbf{w}_d^c)$ are

$$\begin{aligned} p(\mathbf{w}_d^p) &= \sum_{i=1}^{N_d}\sum_{j=1}^{T} \varphi^p_{w_{d_i},j}\theta^p_{d,j}, \\ p(\mathbf{w}_d^c) &= \sum_{i=1}^{N_d}\sum_{j=1}^{T} \varphi^c_{w_{d_i},j}\theta^c_{d,j}. \end{aligned}$$

For all models, we use the collapsed Gibbs sampling to identify their own φ and θ [30]. φ's are estimated from a training set, and θ's are inferred from a test set using their own φ.

Note that the perplexity in Equation (3.12) is a per-word metric. Since a time word comes from calls or proximities, it is natural to measure perplexities separately according to the word type. Let L^p_{test} and L^c_{test} be the subsets of test proximity and call documents, respectively, where $L_{test} = L^p_{test} \cup L^c_{test}$. Then, the perplexity for L_{test} can be measured alternatively by

$$\begin{aligned} & Perplexity(L_{test}) \\ = \; & \exp\left(-\frac{\sum_{d=1}^{\frac{M}{2}} \log p(\mathbf{w}_d^p)}{\sum_{d=1}^{\frac{M}{2}} N_d^p} - \frac{\sum_{d=1}^{\frac{M}{2}} \log p(\mathbf{w}_d^c)}{\sum_{d=1}^{\frac{M}{2}} N_d^c} \right) \\ = \; & \exp\left(-\frac{\sum_{d=1}^{\frac{M}{2}} \log p(\mathbf{w}_d^p)}{\sum_{d=1}^{\frac{M}{2}} N_d^p} \right) \cdot \exp\left(-\frac{\sum_{d=1}^{\frac{M}{2}} \log p(\mathbf{w}_d^c)}{\sum_{d=1}^{\frac{M}{2}} N_d^c} \right) \\ = \; & Perplexity(L^p_{test}) \cdot Perplexity(L^c_{test}), \end{aligned}$$

where N_d^p and N_d^c are the number of words in $\mathbf{w}_d^p$ and $\mathbf{w}_d^c$ respectively.

For all experiments in this study, we set the symmetric Dirichlet parameters $\alpha = 50$ and $\beta = 0.01 \cdot W$, where W is the vocabulary size. This setting is common in most LDA-based studies [25, 30]. The concentration parameter α^c of sdiLDA is set to 50.

Chapter 4

Experiments

This chapter presents experiments in three sub-chapters. In Chapter 4.1, topic models introduced in Chapter 3 are evaluated in terms of perplexity. Especially, the newly proposed topic model, sdiLDA, is verified in three aspects, its assumption, flexibility, and effectiveness of parameter learning. Chapter 4.2 demonstrates effectiveness of the topic models in terms of distinguishing relationships among users. In Chapter 4.3, topic models are actually applied to classifying various social relationships between mobile users.

4.1 Evaluation of Topic Models

Topic models explained in Chapter 3 are evaluated in four aspects. First, it is verified that calls depend on proximities. For this, two sdiLDAs are compared with iLDA. One sdiLDA models influences from proximities to calls, which is the proposed model. The other sdiLDA is designed reversely. That is, it forces influences from calls to proximities. iLDA provides results that

has no influences from interactions of the other type. Second, patterns found from topic models are evaluated in quantitative. The perplexity is used as an evaluation measure. Third, effect of the concentration parameter α^c is investigated by varying α^c. Lastly, results by hyperparameter learning are compared. For this, hyperparameters are learned for not only sdiLDA but also the three topic models, LDA, iLDA, and PLTM. In the following chapter, data sets to be used are introduced and experimental settings are presented.

4.1.1 Experimental Settings

Three different data sets are adopted for the evaluation of the proposed model. The first is the Reality Mining data set [23]. This data set contains nine months of call and proximity logs for 96 academic users. The users are students or faculty members associated with the MIT Media Laboratory. For a single proximity, users' mobile phone used Bluetooth to periodically scan nearby devices at six minute intervals. The media access control (MAC) addresses of the detected devices were recorded along with the scanned time. For a call, the log recorded the time that the call started, its duration, and its direction (incoming or outgoing), along with the phone number involved in the call. However, the mobile logs of only 44 out of 96 total users were used in the experiments. The remaining 52 users have only proximity logs, even if both calls and proximities are required in our model.

The other two data sets used in the evaluation are the Social Evolution [26] and the Friends & Family [1] data sets. The two data sets were also built in the same manner as the Reality Mining data set; however, the user

Table 4.1: Basic statistics from the three data sets.

Data set	The number of users	The number of months (periods)	Average ratio of calls to proximities
Reality Mining	44	9 (2004.9∼2005.6)	0.186
Social Evolution	72	7 (2008.10∼2009.5)	0.015
Friends & Family	114	15 (2010.3∼2011.6)	0.090

groups and time periods are different. The Social Evolution data set contains calls and proximities of residents at an undergraduate dormitory in North America, while the Friends & Family data set was collected from the members of a young family residential community adjacent to a major North America research university.

Table 4.1 shows the basic statistics of the three data sets including data collection period. This table provides information regarding the numbers of users, the mobile log collection periods, and the average ratio of calls to proximities for each data set. The experiments utilize data from 72 and 114 users in the Social Evolution and Friends & Family data sets, respectively, whereas data from 44 users in the Reality Mining data set are used. All data in the three data sets was collected over different periods. The Reality Mining data was collected over a nine-month period, Social Evolution was collected over seven months, and Friends & Family over fifteen months. The average ratios

of calls to proximities are listed in the last column. These ratios are very small (less than 0.2), because the number of proximities usually overwhelms that of calls.

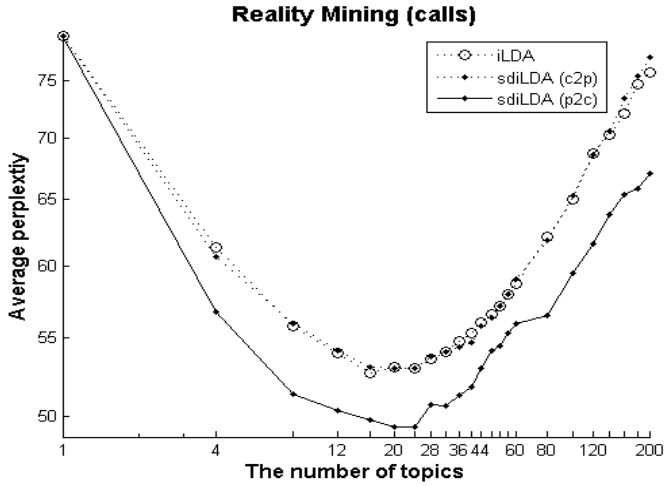
The mobile log of each user is used to evaluate topic models. Every log of a user is divided into five folds, where one fold is used as a test set and the remaining four folds are used as a training set. All data are converted into L in Eqaution (3.5), a set of interaction documents. As a result, there exist 220 ($= 5 \cdot 44$) different pairs of training and test data sets for the Reality Mining data set. Similarly, there are 360 ($= 5 \cdot 72$) and 570 ($= 5 \cdot 114$) pairs for the Social Evolution data set and Friends & Family data set, respectively.

4.1.2 Experimental Results

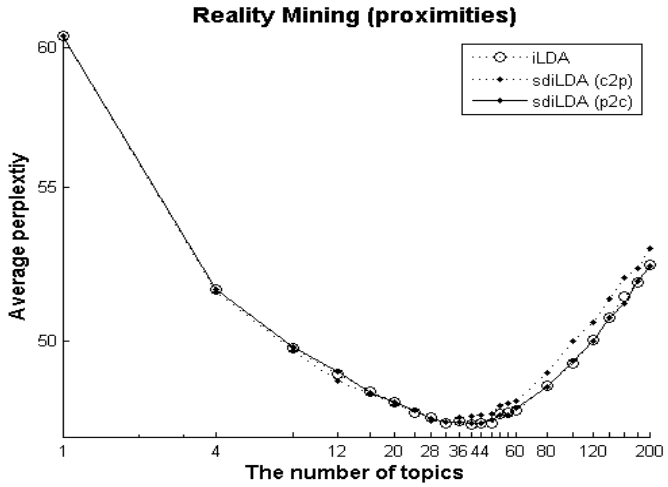
Verification of Single Directional Influences

In designing sdiLDA, it is assumed that the topics of call documents depend on those of proximity documents. In order to verify this assumption, we compare two sdiLDAs with iLDA. One is the proposed sdiLDA, denoted as sdiLDA(p2c), in which proximity topics have influence on call topics. The other is sdiLDA(c2p), in which call topics influence proximity topics. Because iLDA is configured to have the same number of topics for calls and proximities, the only difference between sdiLDAs and iLDA is that the sdiLDAs utilize single-directional influence between proximities and calls. For precise comparisons among the three models, the perplexities on calls and proximities are measured for each model.

Figure 4.1, 4.2, and 4.3 shows the comparison results for the three data



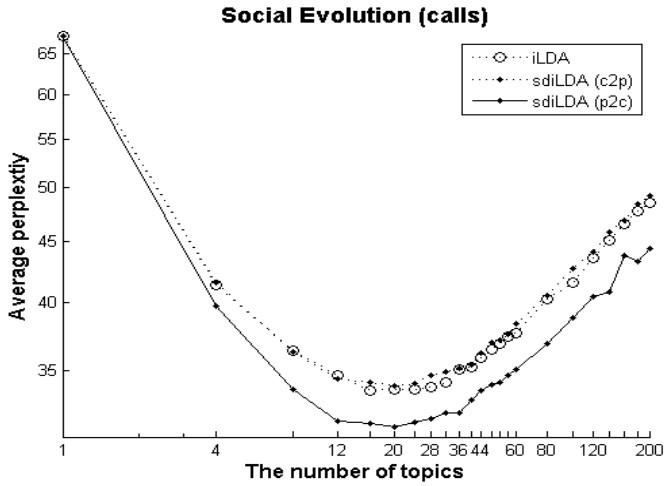
(a) Comparison on calls



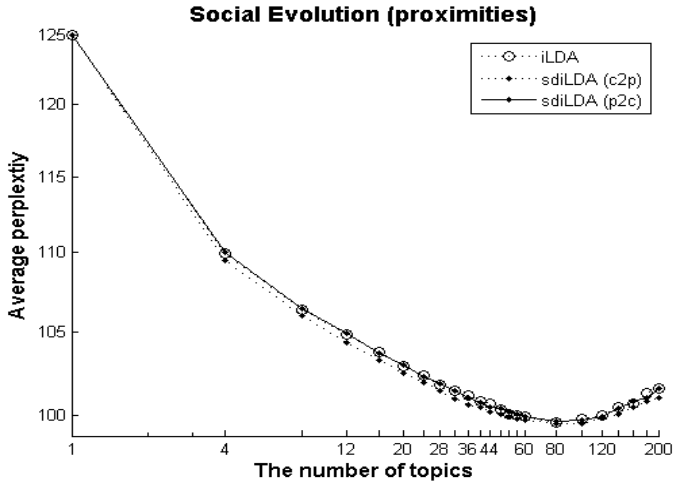
(b) Comparison on proximities

Figure 4.1: Comparisons of sdiLDA(p2c), sdiLDA(c2p), and iLDA from the Reality Mining data set.

sets. The X-axis of the figures represents the number of topics, and Y-axis



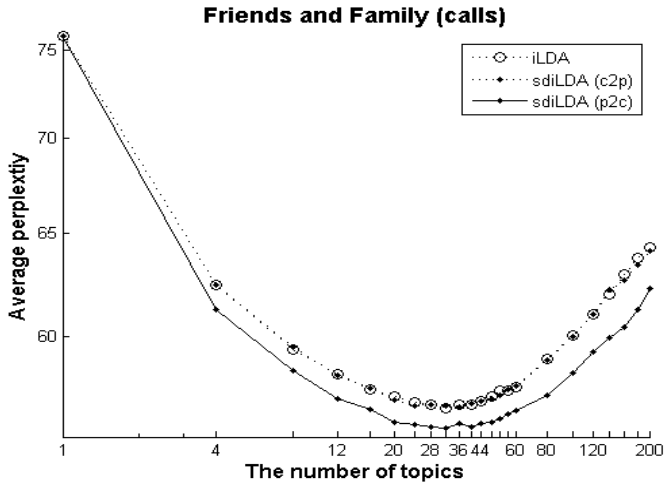
(a) Comparisons on calls



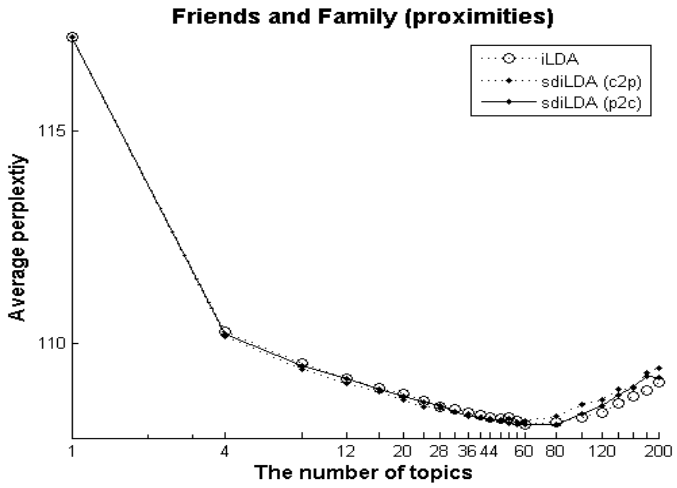
(b) Comparisons on proximities

Figure 4.2: Comparisons of sdiLDA(p2c), sdiLDA(c2p), and iLDA from the Social Evolution data set.

is the average perplexity of a test set from each data set. Both axes are at a log scale. The perplexities of all models tend to increase after a particu-



(a) Comparisons on calls



(b) Comparisons on proximities

Figure 4.3: Comparisons of sdiLDA(p2c), sdiLDA(c2p), and iLDA from the Friends & Family data set.

lar number of topics, which is overfitting of each model to its training data. However, the perplexity of sdiLDA(p2c) for calls is always lower than those

of sdiLDA(c2p) and iLDA, regardless of the number of topics for all three data sets. To determine statistical significance of the results, the perplexity differences from sdiLDA(p2c) to sdiLDA(c2p) and iLDA are evaluated with the paired t-test. The null hypothesis is that the difference mean is 0, while the alternative one-side hypothesis is that the mean is less than 0. According to our results, the differences are always statistically significant with all data sets at the 1% significance level. However, the perplexities of sdiLDA(c2p) on proximities are very similar to those of sdiLDA(p2c) and iLDA, as shown in Figures 4.1(b), 4.2(b), and 4.3(b). The perplexity difference gets statistically significant at the 5% significance level for the Reality Mining data set (Figure 4.1(b)), when the number of topics is larger than 160. We could not find any difference from the Social Evolution and the Friends & Family data sets shown in Figure 4.2(b) and 4.3(b). The fact that sdiLDA(p2c) shows similar performance to sdiLDA(c2p) and iLDA on proximities, but higher performance on calls proves that our assumption is correct.

Overall Quality of Interaction Patterns

To evaluate overall quality of interaction patterns identified by the sdiLDA, its average perplexities are compared with those of LDA, PLTM, iLDA, and a baseline. The baseline is a simple maximum likelihood model that predicts time words based on their frequency in training data. Thus, the baseline does not involve any latent variables. It is denoted as MLE. Figures 4.4, 4.5, and 4.6 shows the comparison results for the three data sets. As expected, sdiLDA outperforms the other models for most numbers of topics in all data sets. PLTM is better than sdiLDA for topics 160, 180, and 200 in the Reality

Mining data sets, and is comparable to sdiLDA for those topics in the Social Evolution data set. Though the difference between PLTM and iLDA is not conspicuous on the small numbers of topics in the Reality Mining and Social Evolution data sets, the lowest perpleixty of PLTM is still lower than that of iLDA. However, iLDA outperforms PLTM for most numbers of topics in the Friends and Family data set. In contrast, LDA shows the highest perplexities among the topic models. It achieves higher perplexities even than MLE for most numbers of topics in the three data sets. Note that LDA is the only model that does not distinguish calls and proximities. Thus, these results prove that some relationship exists between call and proximity topics. Therefore, it is important to model relationships between calls and proximities when identifying interaction patterns. The perpexity differences between sdiLDA and other models are statistically significant at 1% significance level. These results imply that sdiLDA is the best model for describing the relationships between calls and proximities.

Effect of Varying Concentration Parameter

The superiority of sdiLDA over PLTM comes from the fact that it has an independent α^c for calls, while calls and proximities share α, a Dirichlet prior of θ's, in PLTM. That is, α^c adjusts the strength of the influence from proximities to calls in sdiLDA. Note that the value of α^c is proportional to its influence. A small α^c implies a small influence, and a large α^c implies a large influence. Figures 4.7, 4.8, and 4.9 proves that controlling α^c is helpful. The graphs in this figure compare sdiLDAs with various α^c's and PLTM for the three data sets. The α^c's compared are 10^{-1}, 10^0, 10^1, 50, and 10^2. As seen in

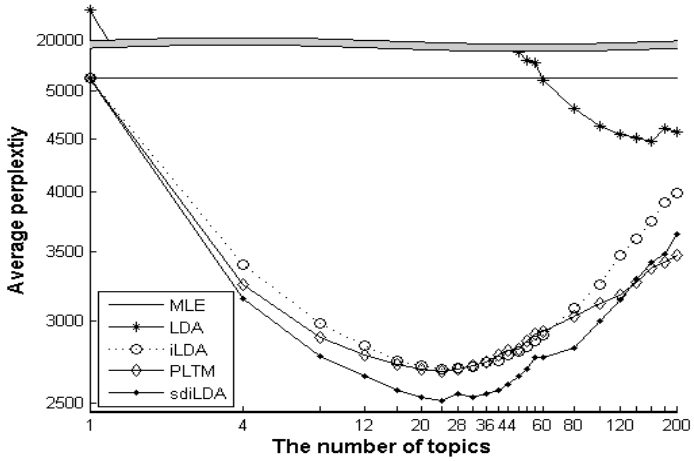


Figure 4.4: Comparisons of sdiLDA with MLE, LDA, iLDA, and PLTM using the Reality Mining data set.

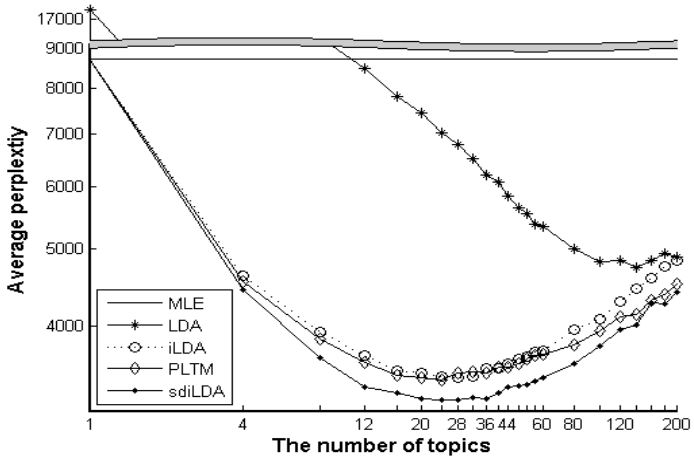


Figure 4.5: Comparisons of sdiLDA with MLE, LDA, iLDA, and PLTM using the Social Evolution data set.

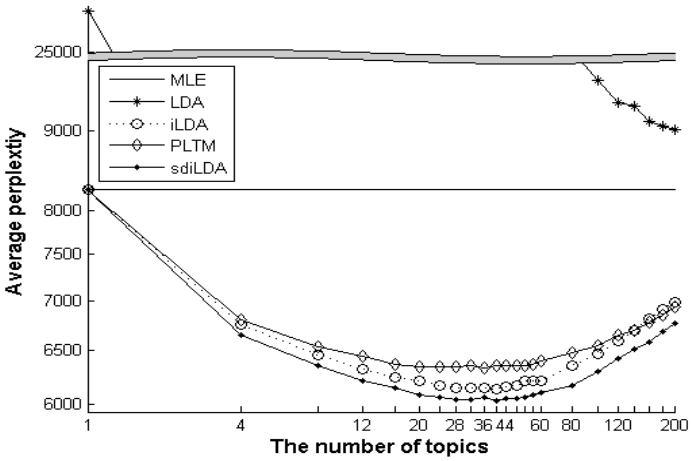


Figure 4.6: Comparisons of sdiLDA with MLE, LDA, iLDA, and PLTM using the Friends & Family data set.

these graphs, an sdiLDA with too small α^c overfits easily to training data as the number of topics increases. In particular, sdiLDAs with $\alpha^c = 10^{-1}$ show higher perplexities than PLTM for large numbers of topics in the three data sets. However, sdiLDAs outperform PLTM overall for the three data sets. In particularly, sdiLDA with $\alpha^c = 10^0$ in the Reality Mining data achieves the smallest perplexity of 2,151 at 16 topics, while PLTM shows the smallest perplexity of 2,691 at 24 topics. In the Social Evolution data set, an sdiLDA with $\alpha^c = 10^0$ achieves the smallest perplexity of 2587 at 12 topics, while PLTM shows the smallest perplexity of 3,410 at 24 topics. In the same manner, an sdiLDA with $\alpha^c = 10^0$ shows the best performance with a perplexity of 5,707 at 20 topics in the Friends & Family data set. PLTM shows the best perplexity of 6337 at 36 topics. The best sdiLDA in each data sets shows statistically significant differences against the other sdiLDAs and PLTM at 1% significance

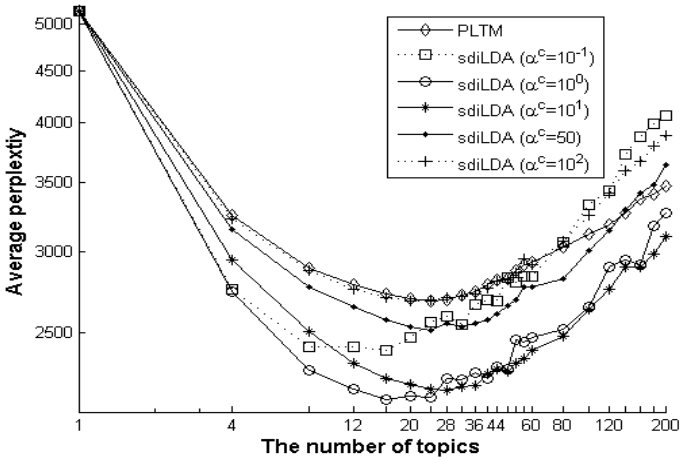


Figure 4.7: Perplexities of sdiLDAs with various α^c's and their comparison with PLTM using the Reality Mining data set.

level. Therefore, compared to PLTM, sdiLDA is more flexible and allows better modeling for identifying interaction patterns.

Effect of Hyperparameter Learning

Figures 4.10, 4.11, and 4.12 depicts hyperparameters learned from a training data set of *user112* discussed in Chapter 3.4 and Figure 4.13 shows performance of the learned models.

An iteration number is denoted as an index of a hyperparameter to represent its value at that iteration point. As shown in Figure 4.10, α^c's are converged to values less than 20 quickly. It tends to increase as the number of topics increases. Thus the influence from proximities to calls also increases with the increase of the topic number. In contrast, α^p's decreases monotonically as the number of topics increases as shown in Figure 4.11. At the same time, the

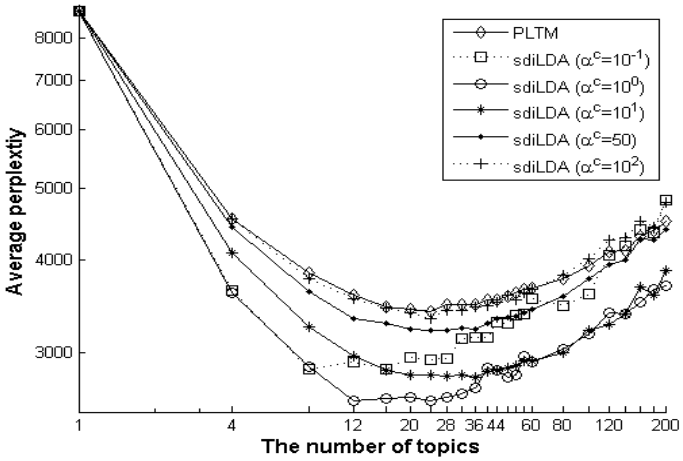


Figure 4.8: Perplexities of sdiLDAs with various α^c's and their comparison with PLTM using the Social Evolution data set.

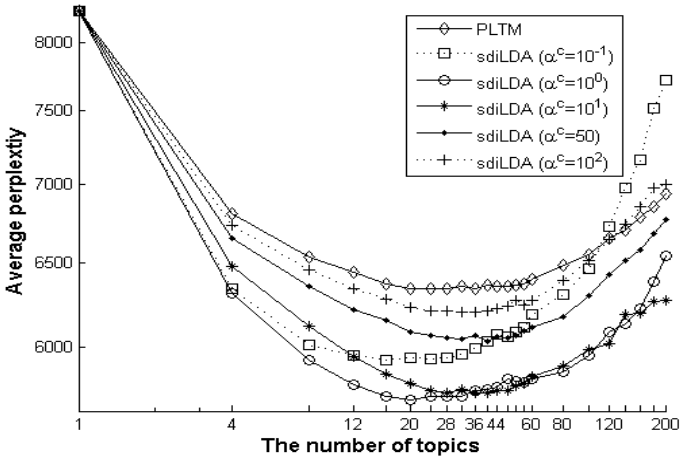


Figure 4.9: Perplexities of sdiLDAs with various α^c's and their comparison with PLTM using the Friends & Family data set.

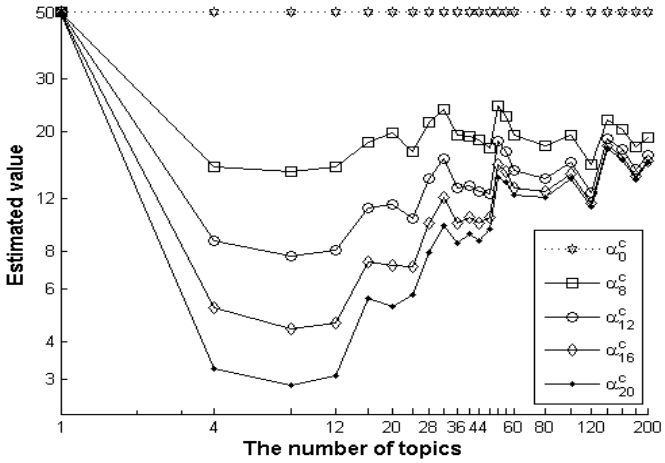


Figure 4.10: Comparisons of α^c's at iteration 8, 12, 18, and 20 for *user112* in the Friends and Family data set.

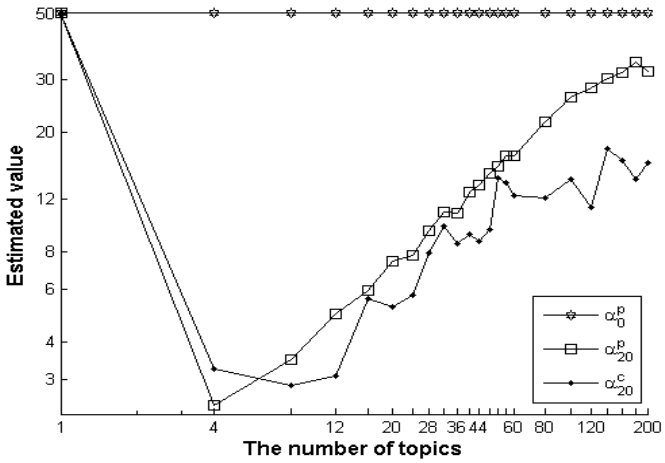


Figure 4.11: Comparisons of α^p's at initial time and after iteration 20 for *user112* in the Friends and Family data set.

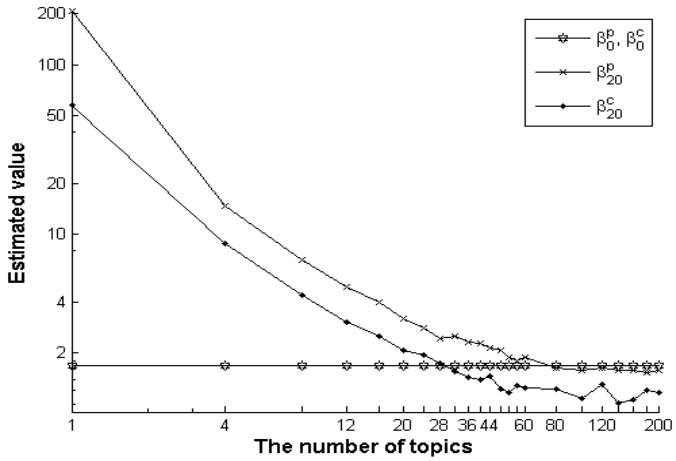


Figure 4.12: Comparisons of β^p's and β^c's at initial time and after iteration 20 for *user112* in the Friends and Family data set.

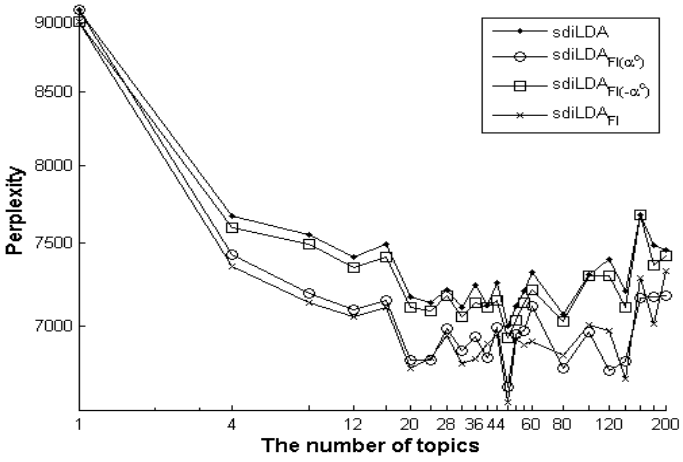


Figure 4.13: Comparisons of four sdiLDA with different fixed point iteration settings for *user112* in the Friends and Family data set.

gaps between α^p's at initial point and after 20 iterations decreases and closes to zero at the topic number 200. This tendency is also observed in fixed point iteration for β^p and β^c as shown in Figure 4.12. That is, α^p, β^p, and α^c tend to be similar to their own initial point. The results imply that fixed point iterations for the three hyperparameters affect to the performance of sdiLDA limitedly. In contrast, differences between α^c at initial point and after final iteration are always larger than 30, and thus its affect to the performance is expected to be large.

The effectiveness of the fixed point iterations is evaluated in Figure 4.13. In this figure, four sdiLDAs are learned from a training data set with different fixed point iteration settings and their perplexities on the corresponding test data set are depicted. One sdiLDA is learned without any fixed point iteration, while another one is learned with fixed point iterations for all hyperparameters. They are denoted by sdiLDA and sdiLDA$_{FI}$ respectively. The other two sdiLDAs are learned with and without only the fixed point iteration for α^c respectively. They are denoted by sdiLDA$_{FI(\alpha^c)}$ and sdiLDA$_{FI(-\alpha^c)}$ As shown in Figure 4.13, sdiLDA$_{FI(\alpha^c)}$ and sdiLDA$_{FI}$ outperform sdiLDA and sdiLDA$_{FI(-\alpha^c)}$ and their perplexities are similar on most numbers of topics. Though sdiLDA$_{FI(-\alpha^c)}$ shows improved results compared with sdiLDA, it is limited compared with sdiLDA$_{FI(\alpha^c)}$. As a result, the performance of sdiLDA$_{FI}$ totally depends on α^c. The results prove that the effect of α^c is larger than the other hyperparameters.

In order to verify the effectiveness of α^c against the other hyperparameters in general, the four sdiLDAs are evaluated on all test data sets in the three data sets. Figures 4.14, 4.15, and 4.16 shows the results. The tree

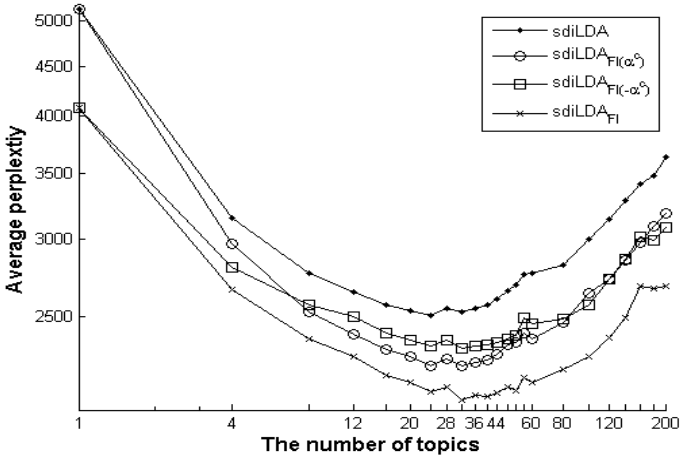


Figure 4.14: Comparisons of four sdiLDAs with different fixed point iteration settings using the Reality Mining data set.

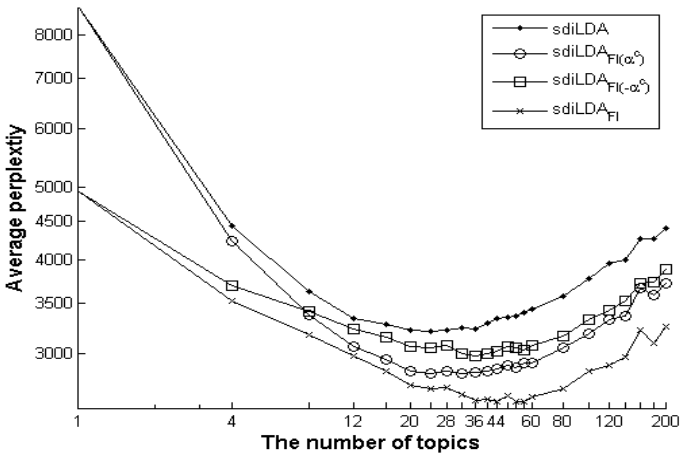


Figure 4.15: Comparisons of four sdiLDAs with different fixed point iteration settings using the Social Evolution data set.

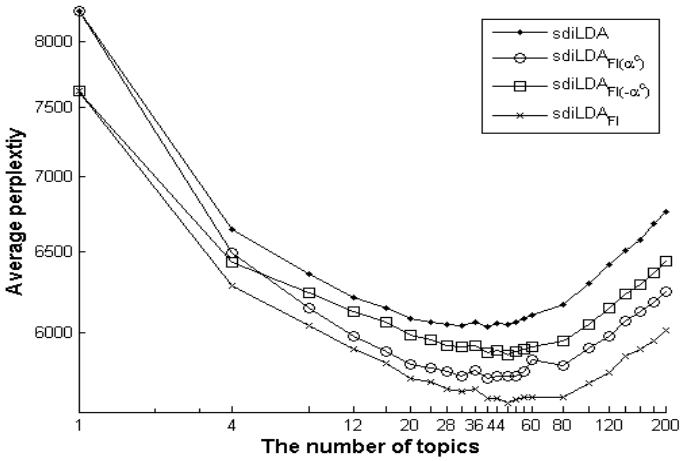


Figure 4.16: Comparisons of four sdiLDAs with different fixed point iteration settings using the Friends & Family data set.

models, sdiLDA$_{FI}$, sdiLDA$_{FI(\alpha^c)}$, and sdiLDA$_{FI(-\alpha^c)}$ outperform the ordinary sdiLDA. Thus fixed point iterations are benefitial to improve the performance of sdiLDA. In addition, sdiLDA$_{FI(\alpha^c)}$ outperforms sdiLDA$_{FI(-\alpha^c)}$ on most numbers of topics in the three data sets. The results proves that learning α^c is more effective than learning the other hyperparameters in improving sdiLDA.

The performances of LDA, iLDA, PLTM also can be improved by learning their own hyperparameters. For this, a fixed point iteration is used in the same way of sdiLDA. The results are depicted in Figures 4.17, 4.18, and 4.19. All the topic models are denoted with a subscript FI to distinguish their original versions. iLDA$_{FI}$ makes a remarkable improvement and is superior to PLTM$_{FI}$. However, sdiLDA$_{FI}$ outperforms the three models in the three data sets. The results demonstrate that the proposed single directional influence modeling is reasonable also in the advanced settings.

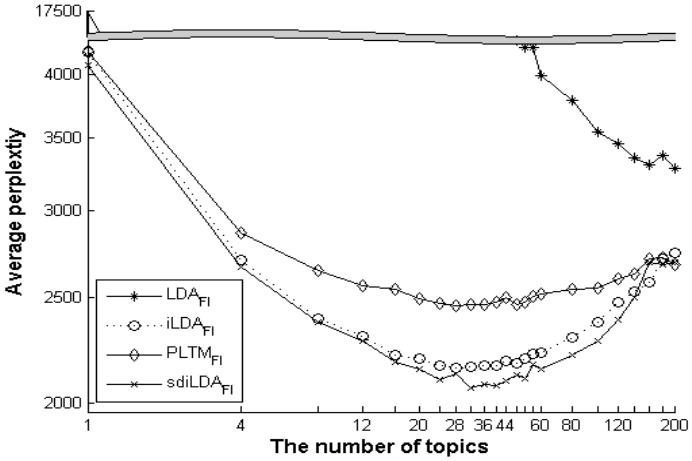


Figure 4.17: Comparisons of sdiLDA$_{FI}$ with LDA$_{FI}$, iLDA$_{FI}$, and PLTM$_{FI}$ using the Reality Mining data set.

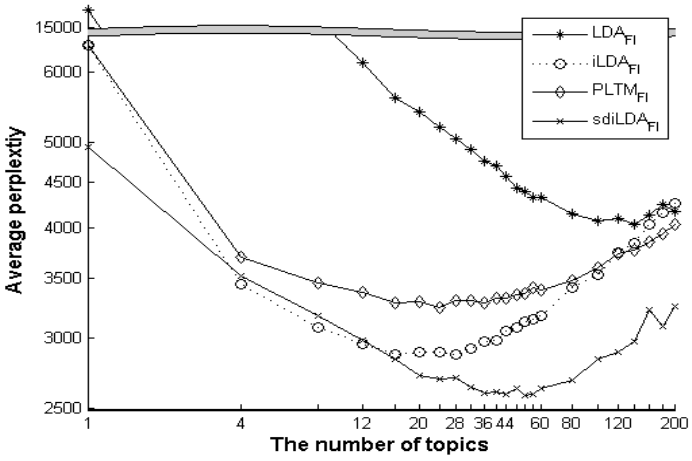


Figure 4.18: Comparisons of sdiLDA$_{FI}$ with LDA$_{FI}$, iLDA$_{FI}$, and PLTM$_{FI}$ using the Social Evolution data set.

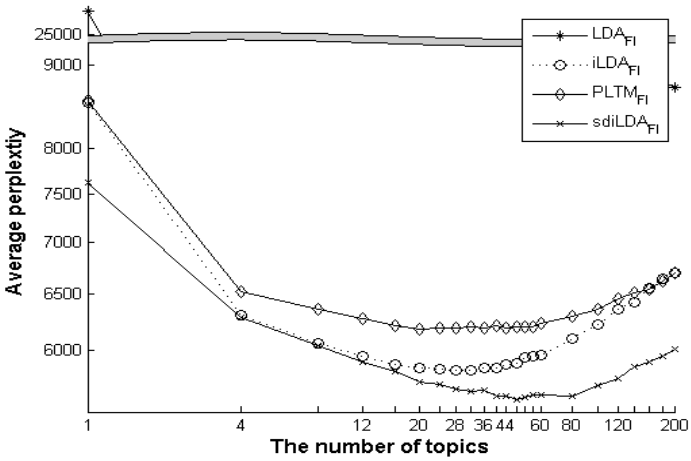


Figure 4.19: Comparisons of sdiLDA$_{FI}$ with LDA$_{FI}$, iLDA$_{FI}$, and PLTM$_{FI}$ using the Friends & Family data set.

4.2 Discriminant of Social Relationships

In the previous chapter, we have shown the quality and superiority of social interaction patterns learned by the proposed method. In order to see their actual effectiveness, the patterns are applied to distinguishing relationships among users. The patterns of PLTM and sdiLDA are compared, since they are the two best models in Figure 4.4.

4.2.1 Experimental Settings

The Reality Mining data set is used for the experiment, because it was collected to investigate human interactions according to relationship. The data set provides the survey results on exact relationships among users. There are three types of relationships between any two users in the survey.

- **symmetric relationship (sym. rel.)**: both users know each other.
- **asymmetric relationship (asym. rel.)**: only one user knows the other user.
- **no relationship (no rel.)**: both users do not know each other.

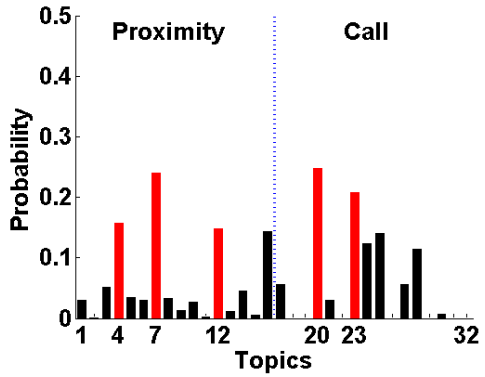
In the implementation of PLTM and sdiLDA ($\alpha^c = 1$), the number of topics is set to 24 and 16 respectively, because they show the best performance at those topics in Figure 4.7.

Note that interaction patterns of a target user are inferred as topics by both sdiLDA and PLTM. Because the topics govern social interactions regardless of other users who interact with the target user, they are not sufficient to distinguish users according to relationship type. Thus, we use topic distributions inferred from social interactions between the target user and other users. Because a topic distribution affects only social interactions of the target user and another user, it reflects relational characteristics of their interactions.

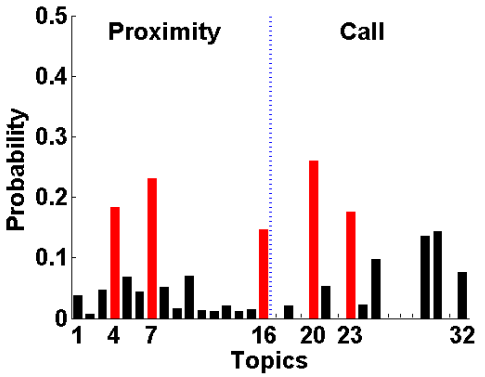
4.2.2 Experimental Results

Qualitative Evaluation

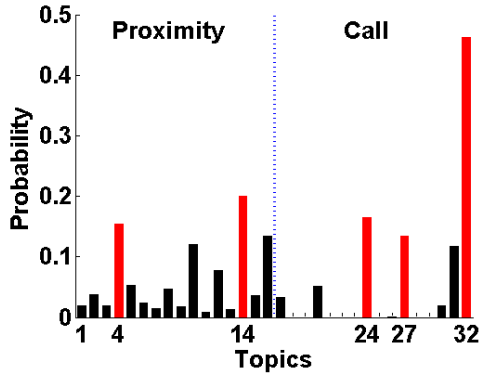
Figure 4.20 shows that the users who interact with a target user can be distinguished according to relationship type. The graphs in this figure are topic distributions of a randomly-selected target user named *user74* against other users. Because sdiLDA has two distinct topic distributions (θ^p and θ^c) between a target user and another user, the X-axis in this figure has 32 topics. Figure 4.20(a) is a topic distribution against *user81* who has a symmetric relationship



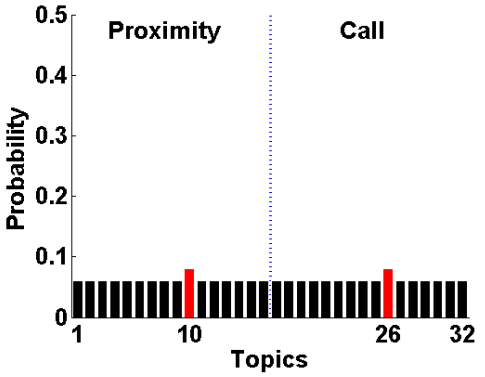
(a) Against *user81* with symmetric relationship



(b) Against *user93* with symmetric relationship



(c) Against *user76* with asymmetric relationship



(d) Against *user12* with no relationship

Figure 4.20: Topic distributions of *user74* against various users with various relationships.

with *user74*, and Figure 4.20(b) is a distribution against *user93* who also has a symmetric relationship with *user74*. In these figures, the top five topics are highlighted in red, and their indexes are also indicated. Because both *user81*

and *user93* have a symmetric relationship with *user74*, their distributions are similar. In addition, topics 4, 7, 20, and 23 are commonly prominent in both Figures 4.20(a) and 4.20(b).

Figure 4.20(c) shows a topic distribution of *user76* who has an asymmetric relationship with *user74*, and it also highlights the top five topics and their indexes in red. Because *user76* has a different relationship from *user81* and *user93*, its distribution is also different from Figure 4.20(a) and 4.20(b). Note that *user81* in Figure 4.20(a) shares only one topic with *user76* in Figure 4.20(c), while she shares four topics with *user93* in Figure 4.20(b). Figure 4.20(d) is a distribution against *user12* who has no relationship with *user74*. It highlights only two topics, because the probabilities of most topics are determined just by a prior. That is, no social interaction is observed between *user12* and *user74* except the two topics. As a result, Figure 4.20(d) is completely different from both Figure 4.20(a) and Figure 4.20(b).

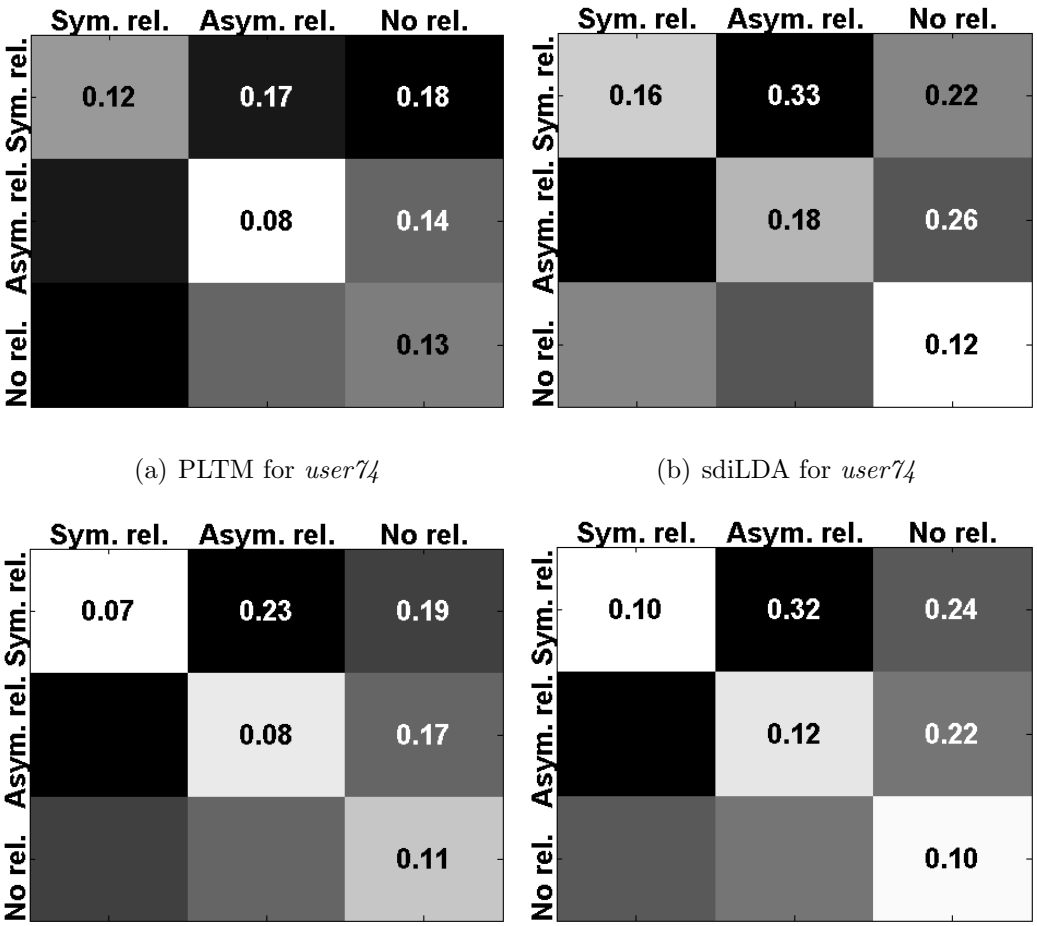
Quantitative Evaluation

The topic distributions of two users are similar if the users have the same relationship, and the distributions become dissimilar if the users have different relationships. For the numerical verification of this fact, the Jensen-Shannon divergence (JSD) [28] is adopted, which measures the similarity between two distributions. When two topic distributions, θ and θ', are given, the JSD between them is

$$JSD(\theta, \theta') = H\left(\frac{\theta + \theta'}{2}\right) - \frac{H(\theta) + H(\theta')}{2},$$

where $H(x)$ is the entropy of x. The smaller the JSD is, the more similar the two distributions are. The JSD between Figure 4.20(a) and Figure 4.20(b) is 0.136, while those between Figure 4.20(a) and Figure 4.20(c) and between Figure 4.20(a) and Figure 4.20(d) are 0.271 and 0.185 respectively. Thus, *user81* is numerically proved to be more similar to *user93* than *user76* and *user12*. This agrees with the statements discussed above.

Figure 4.21 shows similarity matrices of PLTM and sdiLDA. Both the X-axis and Y-axis in this figure represent the relationship type, and the brightness of an element indicates the average JSD of the element. The brighter an element is, the less distant the two relationship types are. Thus, a matrix in which only diagonal elements are bright is ideal. In addition, all matrices are symmetric. Figure 4.21(a) and 4.21(b) are the matrices for *user74*. According to these figures, sdiLDA is superior to PLTM, because the brightness difference between diagonal and non-diagonal elements is larger in sdiLDA than in PLTM. This can also be shown numerically. The value on each element is the average JSD for the relationship type expressed by the element. For instance, in sdiLDA, the average JSD of the users in a symmetric relationship with *user74* is 0.16, while the JSD between the users in an asymmetric relationship and the users in a symmetric relationship is 0.33 for *user74*. The average JSD of all diagonal elements is $(0.12+0.08+0.13)/3 = 0.11$ in PLTM, and that of all non-diagonal elements is $(0.17 + 0.18 + 0.14)/3 = 0.16$. As a result, the JSD difference between diagonal and non-diagonal elements is $0.16 - 0.11 = 0.05$ in PLTM. On the other hand, in sdiLDA, the average JSD of the diagonal elements is 0.15, and that of the non-diagonal elements is 0.27. Thus, the difference is $0.27 - 0.15 = 0.12$ in sdiLDA. Because the difference in sdiLDA is larger



(a) PLTM for *user74*

(b) sdiLDA for *user74*

(c) PLTM for all users

(d) sdiLDA for all users

Figure 4.21: Similarity matrices of sdiLDA and PLTM.

than that of PLTM, sdiLDA is superior to PLTM (at least for *user74*) in distinguishing users according to relationship.

In order to verify whether sdiLDA is superior to PLTM in general, JSDs are measured for all users. Figure 4.21(c) and 4.21(d) show the results. In PLTM, the average JSD of diagonal elements is 0.09, and that of non-diagonal elements

is 0.20. Thus, the JSD difference is 0.11. On the other hand, in sdiLDA, the average JSD of diagonal elements is 0.11, while that of non-diagonal elements is 0.26. Thus, the difference is 0.15. Therefore, even when we consider all users, the difference in sdiLDA is larger than that of PLTM. These results prove that sdiLDA is outstanding and superior to PLTM in discriminating users by relationship type.

4.3 Social Relationship Classification

The Chapter 4.2 verifies that topic distributions from sdiLDA are effective to distinguish acquaintance relationships compared with those from PLTM. In this chapter, the topic distributions are applied to classifying various social relationships between mobile users.

4.3.1 Experimental Settings

The Social Evolution data set is used for the experiment, because it provides the richest survey results on various social relationships among the three data sets. The data set provides monthly surveys on social relationships of five different types. The relationships used are as follows.

- **CloseFriend**: both users report close friendship.
- **Socializing**: both users report at least two common activities per week.
- **PoliticalDiscussant**: both users report to discuss on politics.
- **FacebookPhotos**: both users report to share tagged Facebook photos.

Table 4.2: Social relationships statistics from the Social Evolution data set

Relationship types	The number of positive rel.	The number of negative rel.	Imbalance ratio.
CloseFriend	232	3,549	15.38
Socializing	610	3,171	5.20
PoliticalDiscussant	340	3,441	10.10
FacebookPhotos	667	3,114	4.67
SharingBlogTwitter	651	3,130	4.81

- **SharingBlogTwitter**: both users report to share blog, live journal, and Twitter activities.

Social interactions between two users are regarded as different ones according to the period covered by each survey since relationships between two individuals can be changed over time. Table 4.2 shows statistics on social relationships from the survey results in the Social Evolution data set. This table provides information regarding the number of positive and negative examples for each relationship, and the imbalance ratio (IR), defined as the ratio of the number of instances in the majority class to the number of examples in the minority class. Though the number of positive and negative examples are different according to relationships, the numbers of total examples are the same for all the relationships, which is 3,781. The IRs are recorded in the last column of Table 4.2. The CloseFriend has the highest IR among the three relationships and this implies that the relationships has a more skewed class distribution than the others. However, class distributions of the other two

relationships also are highly skewed (see [47]). These imbalances cause that standard classification learning methods are often biased towards the majority class and therefore there is a higher misclassification rate for the minority class instances, so called the data imbalance problem [33].

Existing approaches are mainly concerned with given training and test instances. For example, data sampling methods make a balance data set from a given unbalance data set by using sampling techniques [3, 14]. A simple way is to create a superset of the original data set by replicating some instances or creating new instances from existing ones of the minority class. Cost sensitive methods assume higher misclassification costs for instances in the minority class and aim to minimize the high cost errors [20, 76]. These studies assume that a feature space is predefined and training and test instances expressed in the space are given in advance. However, the data imbalance problem is related to the data distribution over a feature space [27, 65]. Thus, it is a key factor to define a feature space for solving the data imbalance problem.

This dissertation utilizes topic models to define a feature space appropriate to the social relationship classification on the imbalance data set. For this, the topic models are used differently from the previous chapters in two points. First, a collection of interaction documents is built from all the users whereas, in the previous chapters, a collection is obtained from each user's interaction logs. This enables relationship instances to be expressed in the same feature space. In this dissertation, the feature space is defined as the topic space and thus a topic proportion from a user pair becomes a feature vector.

Second, topics are learned separately for positive and negative relationship classes respectively and the topics are used to express a feature space. Since

learning a topic model aims to find topics best fit to intrinsic properties of a given training data set, the model learned from all the training data is easily biased towards the majority class. By learning a feature space from positive and negative examples respectively, the feature space can reflect intrinsic properties of the minority class as well as of the majority class. This is a simple supervised approach since class labels are used in topic learning.

The proposed approach is novel in the sense that a feature space is divided in two areas, one area for the positive class and the other area for the negative class. In this dissertation, the positive area corresponds to the topic space obtained from instances of a positive relationship. This area is appropriate to describe interactions between users with the positive relationship, whereas the negative area obtained from instances of a negative relationship is fitted to interactions between users with the negative relationship. Therefore even a minor number of positive instances cluster together in at least the positive space and this cluster is distinguished with negative instances.

This dissertation adopts a support vector machine (SVM) with RBF (radial basis function) kernel as a basic classifier since the accuracy of the RBF kernel is better than the linear kernel in general [46]. All the topic models discussed in Chapter 3 are used with the basic classifier for the classification tasks. SVMs with LDA, iLDA, PLTM, and sdiLDA are denoted as SVM_{LDA}, SVM_{iLDA}, SVM_{PLTM}, and SVM_{sdiLDA}, respectively. The number of topics T is set to 3, 6, and 10. Thus each classifier has three versions according to the topic numbers. A cost-sensitive version of the basic classifier is used as a baseline with features proposed by Madan et al. [50]. The features are total phone calls, weekend/late-night calls, total proximity, and late-night/weekend proximity.

The baseline is denoted as $\text{SVM}_{Baseline}$.

Precision, recall, and F-score are used to evaluate performances and these metrics are defined as follows.

$$\begin{aligned} Precision &= \frac{\text{\# of examples labeled as positive correctly}}{\text{\# of of examples labeled as positive}} \\ Recall &= \frac{\text{\# of examples labeled as positive correctly}}{\text{\# of positive examples}} \\ F\text{-}score &= \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \end{aligned}$$

For the evaluation, 5-fold cross-validation is adopted.

4.3.2 Experimental Results

Table 4.3 shows comparison results for the CloseFriend classification. The results of topic-based methods are grouped according to the number topics and each group is expressed in the first column of this table. Each group has results from the four topic-based methods. The highest precision, recall, and F-score are highlighted with bold in each group and the highest ones in overall are denoted by asterisk (*). $\text{SVM}_{Baseline}$ has no concern with the topic numbers and thus its performance is recorded at the second row only.

There are several things to note in Table 4.3. First, all the topic-based methods outperform the baseline. The results implies that topics inferred automatically provide better representations than manually defined features.

Second, SVM_{iLDA} outperforms both SVM_{LDA} and SVM_{PLTM} over all the topic numbers. Since SVM_{LDA} simply combines call and proximity logs, the results imply that the combination of logs results in noise. On the other hand, the assumption of SVM_{PLTM} is too rigid in the classification task and thus it

shows lower performance than SVM_{iLDA}.

Lastly, though SVM_{iLDA} shows slightly better performance in the topic number 10, SVM_{sdiLDA} with the topic number 3 shows the best performance. The results demonstrate that sdiLDA provides the best representation for the CloseFriend classification.

Table 4.4 provides the results for the Socializing classification. There are two things to note. First, as shown in this table, the baseline shows better performance than that for the CloseFriend classification. This tendency is also

Table 4.3: Experimental results for the CloseFriend classification

Number of Topics	Mehtods	Recall	Precision	F-score
None	$\text{SVM}_{Baseline}$	0.0970	0.3348	0.1504
3	SVM_{LDA}	0.5644	0.3087	0.3715
	SVM_{iLDA}	0.8667	0.7174	0.7757
	SVM_{PLTM}	0.7609	0.4435	0.5490
	SVM_{sdiLDA}	**0.9586**	**0.9522***	**0.9546***
6	SVM_{LDA}	0.8621	0.4783	0.6144
	SVM_{iLDA}	**0.9800**	0.7739	0.8601
	SVM_{PLTM}	0.8872	0.6609	0.7552
	SVM_{sdiLDA}	0.9610	**0.8348**	**0.8909**
10	SVM_{LDA}	0.9463	0.5261	0.6742
	SVM_{iLDA}	**0.9938***	0.7391	**0.8467**
	SVM_{PLTM}	0.9677	0.6652	0.7851
	SVM_{sdiLDA}	0.9615	**0.7435**	0.8358

Table 4.4: Experimental results for the Socializing classification

Number of Topics	Mehtods	Recall	Precision	F-score
None	$\text{SVM}_{Baseline}$	0.1828	0.4984	0.2671
3	SVM_{LDA}	0.7835	0.5525	0.6446
	SVM_{iLDA}	0.8389	0.6787	0.7495
	SVM_{PLTM}	0.6761	0.4377	0.5174
	SVM_{sdiLDA}	**0.8861**	**0.7574**	**0.8136**
6	SVM_{LDA}	0.9141	0.5148	0.6574
	SVM_{iLDA}	0.8947	0.6918	0.7798
	SVM_{PLTM}	0.8782	0.6148	0.7223
	SVM_{sdiLDA}	**0.9272**	**0.7689**	**0.8354**
10	SVM_{LDA}	0.9296	0.5607	0.6981
	SVM_{iLDA}	0.9454	0.7066	0.8078
	SVM_{PLTM}	0.9453	0.6885	0.7964
	SVM_{sdiLDA}	**0.9668***	**0.8000***	**0.8741***

observed in the topic-based methods except SVM_{sdiLDA}. The results implies that the Socializing relationship is more related to proximities and calls than the CloseFriend relationship. Actually, the Socializing relationship is directly related to proximities and calls since the Socializing relationship is defined by common activities per week.

Second, though SVM_{sdiLDA} in the Socializing classification shows lower performance than that in the CloseFriend classification, it always outperform the other methods in recall, precision, and F-score in Table 4.4. This fact demon-

Table 4.5: Experimental results for the PoliticalDiscussant classification

Number of Topics	Mehtods	Recall	Precision	F-score
None	$\text{SVM}_{Baseline}$	0.1318	0.4118	0.1994
3	SVM_{LDA}	0.7287	0.4176	0.5262
	SVM_{iLDA}	**0.9257**	**0.8412**	**0.8784***
	SVM_{PLTM}	0.5935	0.3000	0.3625
	SVM_{sdiLDA}	0.8871	0.8000	0.8383
6	SVM_{LDA}	0.8623	0.4647	0.6011
	SVM_{iLDA}	**0.9411**	0.7471	0.8317
	SVM_{PLTM}	0.8489	0.5765	0.6780
	SVM_{sdiLDA}	0.7706	**0.8615***	**0.8354**
10	SVM_{LDA}	0.9196	0.4618	0.6118
	SVM_{iLDA}	0.9412	0.7059	0.8062
	SVM_{PLTM}	0.9428	0.6706	0.7828
	SVM_{sdiLDA}	**0.9579***	**0.7294**	**0.8258**

strates that modeling of single directional influences is helpful in the Socializing classification. It also proves that the proposed method to represent a feature space effectively works for relationships of different types.

Table 4.5 shows the results for the PoliticalDiscussant classification. In this table, SVM_{iLDA} shows the best performance with the topic number 3. The results implies that the PoliticalDiscussant relationship is related to proximities and calls independently. This is reasonable since, in general, a political discussion is carried out in face-to-face interactions. However, in spite of the results,

Table 4.6: Experimental results for the FacebookPhotos classification

Number of Topics	Mehtods	Recall	Precision	F-score
None	$\text{SVM}_{Baseline}$	0.1907	0.4977	0.2750
3	SVM_{LDA}	0.6304	0.4737	0.5387
	SVM_{iLDA}	0.8359	0.7534	0.7914
	SVM_{PLTM}	0.6619	0.5729	0.6101
	SVM_{sdiLDA}	**0.8906**	**0.8586***	**0.8738**
6	SVM_{LDA}	0.8546	0.5880	0.6957
	SVM_{iLDA}	0.8532	0.7293	0.7850
	SVM_{PLTM}	0.8178	0.6466	0.7201
	SVM_{sdiLDA}	**0.9125**	**0.8015**	**0.8520**
10	SVM_{LDA}	0.9227	0.5910	0.7188
	SVM_{iLDA}	0.9147	0.7519	0.8244
	SVM_{PLTM}	0.9304	0.6812	0.7853
	SVM_{sdiLDA}	**0.9368***	**0.8421**	**0.8866***

SVM_{sdiLDA} records the best recall and precision with the topic numbers 10 and 6 respectively. It also shows the best F-scores with the topic numbers. Thus the results by SVM_{sdiLDA} are still promising in the PolticalDiscussant classification.

Table 4.6 shows the results for the FacebookPhotos classification. One interesting thing is that the FacebookPhotos relationship is an on-line relationship and thus the relationship is not related directly to proximities and calls. As a result, the FacebookPhotos classification gives an answer whether on-line re-

Table 4.7: Experimental results for the SharingBlogTwitter classification

Number of Topics	Mehtods	Recall	Precision	F-score
None	$\text{SVM}_{Baseline}$	0.1884	0.4631	0.2675
3	SVM_{LDA}	0.6068	0.4631	0.5251
	SVM_{iLDA}	**0.8339**	**0.7888**	**0.8784**
	SVM_{PLTM}	0.7145	0.6215	0.6636
	SVM_{sdiLDA}	0.8047	0.7477	0.7734
6	SVM_{LDA}	0.8453	0.5477	0.6600
	SVM_{iLDA}	0.8412	0.6800	0.7494
	SVM_{PLTM}	0.8887	0.7108	0.7891
	SVM_{sdiLDA}	**0.9244**	**0.8292**	**0.8741**
10	SVM_{LDA}	0.9321	0.6292	0.7509
	SVM_{iLDA}	0.9225	0.7677	0.8378
	SVM_{PLTM}	0.9198	0.7200	0.8073
	SVM_{sdiLDA}	**0.9410***	**0.8446***	**0.8890***

lationship can be inferred from mobile logs or not. As shown in Table 4.6, all topic-based methods show comparable results to other tasks discussed in previous. Especially, SVM_{sdiLDA} always outperform the other methods and records the F-score of 0.8866 which is the largest one compared with the previous results. The results demonstrate that SVM_{sdiLDA} is effective in classifying the on-line relationship.

The SharingBlogTwitter relationship is another on-line relationship and the classification results of the relationship is shown in Table 4.7. As shown in

this table, all topic-based methods show higher performance than the baseline and SVM_{sdiLDA} records the best performance with the topic number 10. The results also proves that SVM_{sdiLDA} is useful in classifying the on-line relationship. In summary, SVM_{sdiLDA} shows superior performance to classify the four social relationships, CloseFriend, Socializing, FacebookPhotos, and SharingBlogTwitter. Though SVM_{iLDA} is the best one in classifying the PoliticalDiscussant relationship, SVM_{iLDA} also shows promising results. These results demonstrates that sdiLDA is effective in representing a feature space for the classification tasks.

Chapter 5

Conclusion

This dissertation has proposed a single-directional influenceLDA (sdiLDA), a novel topic model that can analyze calls and proximities simultaneously to identify interaction patterns. The main innovation in the proposed model comes from the explicit modeling of single-directional influences from proximities to calls. This modeling has two advantages. First, the proposed topic model infers call patterns related to proximity patterns, whereas this relatedness has been lost in most previous studies under the independent consideration of calls and proximities. Second, the proposed topic model provides an explicit method to adjust the strength of influences from proximities to calls, and thus it models the dependency between proximities and calls flexibly. Through the experiments with three datasets (RealityMining, SocialEvolution, and Friendsand Family), the proposed sdiLDA was shown to outperform LDA, iLDA,and PLTM in terms of perplexity. The experimental results prove that sdiLDA is the best model among them for identifying interaction patterns. In addition, the applicability of sdiLDA is also proven by demonstrating that the patterns

of sdiLDA are effective in distinguishing the relationships among users.

The findings of this dissertation offer new insights in identifying social interaction patterns. However, there are several limitations, which should be taken into account. First, there is a lack of ground truth for an object evaluation in the discovery of social interaction patterns. The dissertation uses perplexity as a quantitative measure like most of the previous studies. Perplexity can be understood to measure how well a given model explains a test data. However, given more ground truth, an objective evaluation of discovered patterns would be easier to perform.

Second, any topic models including sdiLDA cannot explain the actual causal relationship between calls and proximities. Although proximities are caused by a call that promises a meeting, this dependency is not observed explicitly. One possible way to model the dependency is to assume that a call and the proximity between two same users are related to each other if the call and the proximity occur within a small time interval. However, it is non-trivial work to determine an appropriate time interval since no actual dependency information is available.

Lastly, the users from the three data sets are related to the Reality Mining research group and thus, are likely to not be representative of the general public. However, despite this limitation, the three data sets, to the best of my knowledge, are the largest data sets that provide calls and proximities and mappings between device IDs and phone numbers. The user groups of the data sets are also different with from each other. Therefore, based on the results from the data sets, the proposed method is promising to find social interaction patterns from the general public.

The work done in the dissertation can be extended in two general directions. First, a promising future work is to apply the proposed idea, single directional influence modeling, to other topic models. For example, a neural network based topic model, so called replicated softmax [35], was proposed as an undirected counterpart of directed graphical models including LDA. Replicated softmax and its advanced version, document neural autoregressive distribution estimator (DocNADE) [42], showed superior results compared with LDA in information retrieval. Recently, DocNADE was utilized for multimodal data modeling in computer vision [78], and there are also some other neural network based deep learning methods [56, 67]. However, to the best of our knowledge, there is no prior work on undirected topic models or deep learning methods for modeling social interactions. Therefore, the proposed idea needs to be applied to these topic models in the context of social interaction modeling.

Second, new methods for applications using social interactions patterns are important. Development of a topic model to analyze the differences and similarities between communities is an interesting research topic. For example, it is important to understand the propagation of other types of opinions and habits in social networks. Most studies on this topic utilize an aggregate interaction that ignores specific interaction patterns. The basic idea of community comparison is to analyze communities in terms of patterns that are inferred as topics. It is also necessary to find patterns appropriate to specific tasks. In this work, the proposed method provides a basic principle utilizing different types of interactions simultaneously.

Bibliography

[1] N. Aharony, W. Pan, C. Ip, I. Khayal, and A. Pentland, "Social fmri: Investigating and shaping social mechanisms in the real world," *Pervasive and Mobile Computing*, Vol. 7, No. 6, pp. 643–659, 2011.

[2] L. Backstrom and J. Leskovec, "Supervised random walks: predicting and recommending links in social networks," In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 635–644, 2011.

[3] G. Batista, R. Prati, and M. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, Vol. 6, No. 1, pp. 20–29, 2004.

[4] B. Behmardi and R. Raich, "On confidence-constrained rank recovery in topic models," *IEEE Transactions on Signal Processing*, Vol. 60, No. 10, pp. 5146–5162, 2012.

[5] M. Black and R. Hickey, "Learning classification rules for telecom customer call data under concept drift," *Soft Computing*, Vol. 8, No. 2, pp. 102–108, 2003.

[6] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[7] D. Blei and M. Jordan, "Modeling annotated data," In *Proceedings of the 26th ACM Special Interest Group on Information Retrieval (SIGIR)*, pp. 127–134, 2003.

[8] D. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," *IEEE Signal Processing Magazine*, Vol. 27, No. 6, pp. 55–65, 2010.

[9] V. Blondel, A. Decuyper, and G. Krings, "A survey of results on mobile phone datasets analysis," *EPJ Data Science*, Vol. 4, No. 10.

[10] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertextual web search engine," *Computer Networks*, Vol. 56, No. 18, pp. 3825–3833, 2012.

[11] F. Calabrese, Z. Smoreda, V. Blondel, and C. Ratti, "Interplay between telecommunications and face-to-face interactions: A study using mobile phone data," *PLoS ONE*, Vol. 6, No. 7, p. e20814, 2011.

[12] J. Candia, M. González, P. Wang, T. Schoenharl, G. Madey, and A. Barabási, "Uncovering individual and collective human dynamics from mobile phone records," *Journal of Physics A: Mathematical and Theoretical*, Vol. 41, No. 22, p. 224015, 2008.

[13] A. Cardillo, M. Zanin, J. Gómez-Gardeñes, M. Romance, A. del Amo, and S. Boccaletti, "Modeling the multi-layer nature of the european air transport network: Resilience and passengers re-scheduling under random

failures," *The European Physical Journal Special Topics*, Vol. 215, No. 1, pp. 23–33, 2013.

[14] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321–357, 2002.

[15] T. Choudhury, "Sensing and modeling human networks," Ph.D. dissertation, Massachusetts Institute of Technology, 2003.

[16] T. Choudhury and A. Pentland, "Sensing and modeling human networks using the sociometer," In *Proceedings of the 7th IEEE International Symposium on Wearable Computers (ISWC)*, pp. 216–222, 2003.

[17] D. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg, "Inferring social ties from geographic coincidences," *National Academy of Sciences*, Vol. 107, No. 52, pp. 22 436–22 441, 2010.

[18] C. Diehl, G. Namata, and L. Getoor, "Relationship identification for social network discovery," In *Proceedings of the 22nd National Conference on Artificial Intelligence*, pp. 546–552, 2007.

[19] T. Do and D. Gatica-Perez, "Human interaction discovery in smartphone proximity networks," *Personal and Ubiquitous Computing*, Vol. 17, No. 3, pp. 413–431, 2013.

[20] P. Domingos, "Metacost: A general method for making classifiers cost-sensitive," In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 155–164, 1999.

[21] W. Dong, B. Lepri, and A. Pentland, "Modeling the co-evolution of behaviors and social relationships using mobile phone data," In *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia (MUM)*, pp. 134–143, 2011.

[22] J. Duesenberry, "Income, saving and the theory of consumer behavior," Vol. 87, pp. 369–386, 1967.

[23] N. Eagle and A. Pentland, "Reality mining: Sensing complex social systems," *Personal and Ubiquitous Computing*, Vol. 10, No. 4, pp. 255–268, 2006.

[24] N. Eagle, D. Lazer, and A. Pentland, "Inferring friendship network structure by using mobile phone data," *National Academy of Sciences*, Vol. 106, No. 36, pp. 15 274–15 278, 2009.

[25] K. Farrahi and D. Gatica-Perez, "Probabilistic mining of socio-geographic routines from mobile phone data," *IEEE Journal of Selected Topics in Signal Processing*, Vol. 4, No. 4, pp. 746–755, 2010.

[26] K. Farrahi, A. Madan, M. Cebrian, S. Moturu, and A. Pentland, "Sensing the "health state" of a community," *Pervasive Computing*, Vol. 11, No. 4, pp. 36–45, 2012.

[27] V. García, R. Mollineda, and J. Sánchez, "On the k-nn performance in a challenging scenario of imbalance and overlapping," *Pattern Analysis and Applications*, Vol. 11, No. 3-4, pp. 269–280, 2008.

[28] J. Gómez-Lopera, J. Martínez-Aroza, A. Robles-Pérez, and R. Román-Roldán, "An analysis of edge detection by using the jensen-shannon divergence," *Journal of Mathematical Imaging and Vision*, Vol. 13, No. 1, pp. 35–56, 2000.

[29] M. Granovetter, "The strength of weak ties," *American Journal of Sociology*, pp. 1360–1380, 1973.

[30] T. Griffiths and M. Steyvers, "Finding scientific topics," *National Academy of Sciences*, Vol. 101, No. 1, pp. 5228–5235, 2004.

[31] Y. Han, S. Cheng, S. Park, and S. Park, "Finding social interaction patterns using call and proximity logs simultaneously," In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Network Analysis and Mining (ASONAM)*, pp. 399 – 402, 2014.

[32] Y. Han, S.-Y. Park, and S.-B. Park, "A single-directional influence topic model using call and proximity logs simultaneously," *Soft Computing*, pp. 1–17, 2015.

[33] H. He and E. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 9, pp. 1263–1284, 2009.

[34] C. A. Hidalgo and C. Rodriguez-Sickert, "The dynamics of a mobile phone network," *Physica A: Statistical mechanics and its applications*, Vol. 387, No. 12, pp. 3017–3024, 2008.

[35] G. Hinton and R. Salakhutdinov, "Replicated softmax: an undirected topic model," In *Proceedings of Advances in Neural Information Processing systems (NIPS)*, pp. 1607–1614, 2009.

[36] T. Hofmann, "Probabilistic latent semantic indexing," In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 50–57, 1999.

[37] J. Hopcroft, T. Lou, and J. Tang, "Who will follow you back?: reciprocal relationship prediction," In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pp. 1137–1146, 2011.

[38] D. Hristova, M. Musolesi, and C. Mascolo, "Keep your friends close and your facebook friends closer: A multiplex network approach to the analysis of offline and online social ties," In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 206–215, 2014.

[39] T. Huynh, M. Fritz, and B. Schiele, "Discovery of activity patterns using topic models," In *Proceedings of the 2008 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pp. 10–19, 2008.

[40] J. Jung, "Contextualized mobile recommendation service based on interactive social network discovered from mobile users," *Expert System with Applications*, Vol. 36, No. 9, pp. 11 950–11 956, 2009.

[41] M. Kivelä, A. Arenas, M. Barthelemy, J. Gleeson, Y. Moreno, and M. Porter, "Multilayer networks," *Journal of Complex Networks*, Vol. 2, No. 3, pp. 203–271, 2014.

[42] H. Larochelle and S. Lauly, "A neural autoregressive topic model," In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pp. 2708–2716, 2012.

[43] D. Lazer, A. Pentland, L. A. Adamic, A. S. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Alstyne, "Computational social science," *Science*, pp. 721–723, 2009.

[44] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 7, pp. 1019–1031, 2007.

[45] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, "New perspectives and methods in link prediction," In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 243–252, 2010.

[46] H. Lin and C. Lin, "A study on sigmoid kernels for svm and the training of nonpsd kernels by smo-type methods," Technical report, Tech. Rep., 2003.

[47] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current

trends on using data intrinsic characteristics," *Information Sciences*, Vol. 250, pp. 113–141, 2013.

[48] A. Madan and A. Pentland, "Modeling social diffusion phenomena using reality mining," In *Proceedings of AAAI Spring Symposium on Human Behavior Modeling*, pp. 43–48, 2010.

[49] A. Madan, M. Cebrian, D. Lazer, and A. Pentland, "Social sensing for epidemiological behavior change," In *Proceedings of the 12th ACM international conference on Ubiquitous computing (Ubicomp)*, pp. 291–300, 2010.

[50] A. Madan, K. Farrahi, D. Gatica-Perez, and A. Pentland, "Pervasive sensing to model political opinions in face-to-face networks," In *Proceedings of Pervasive Computing*, pp. 214–231, 2011.

[51] D. Miller, *A theory of shopping*. John Wiley & Sons, 2013.

[52] D. Mimno, H. Wallach, J. Naradowsky, D. Smith, and A. McCallum, "Polylingual topic models," In *Proceedings of the 2014 Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pp. 880–889, 2009.

[53] T. Minka, "Estimating a dirichlet distribution," https://faculty.cs.byu.edu/~ringger/CS679/papers/Heinrich-GibbsLDA.pdf, 2000, accessed 15 July 2015.

[54] T. Minka and J. Lafferty, "Expectation-propagation for the generative

aspect model," In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 352–359, 2002.

[55] G. Mollenhorst, B. Völker, and F. H., "Social contexts and personal relationships: The effect of meeting opportunities on similarity for relationships of different strength," *Social Networks*, Vol. 30, No. 1, pp. 60 – 68, 2008.

[56] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 689–696, 2011.

[57] Ofcom, "The communications market 2012 (july)," http://stakeholders.ofcom.org.uk/market-data-research/market-data/communications-market-reports/cmr12/?a=0, 2012, accessed 25 October 2015.

[58] J. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. De Menezes, K. Kaski, A. Barabási, and J. Kertész, "Analysis of a large-scale weighted network of one-to-one human communication," *New Journal of Physics*, Vol. 9, No. 6, p. 179, 2007.

[59] W. Pan, N. Aharony, and A. Pentland, "Composite social network for predicting mobile apps installation," In *Proceedings of the 25th Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, pp. 821–827, 2011.

[60] D. Putthividhy, H. Attias, and S. Nagarajan, "Topic regression multimodal latent dirichlet allocation for image annotation," In *Proceedings*

of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3408–3415, 2010.

[61] C. Quadri, M. Zignani, L. Capra, S. Gaito, and G. Rossi, "Multidimensional human dynamics in mobile phone communications," *PLoS ONE*, Vol. 9, No. 7, p. e103183, 2014.

[62] M. Raento, A. Oulasvirta, and N. Eagle, "Smartphones an emerging tool for social scientists," *Sociological methods & research*, Vol. 37, No. 3, pp. 426–454, 2009.

[63] H. Raiffa and R. Schlaifer, *Applied statistical decision theory.* Cambridge, Mass.: Harvard Business School, 1961.

[64] S. Scellato, A. Noulas, and C. Mascolo, "Exploiting place features in link prediction on location-based social networks," In *Proceedings of the 17th International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 1046–1054, 2011.

[65] C. Seiffert, T. Khoshgoftaar, J. Van Hulse, and A. Folleco, "An empirical study of the classification performance of learners on imbalanced and noisy software quality data," *Information Sciences*, Vol. 259, pp. 571–595, 2014.

[66] V. Singh, L. Freeman, B. Lepri, and A. Pentland, "Classifying spending behavior using socio-mobile data," *HUMAN*, Vol. 2, No. 2, pp. 99–111, 2013.

[67] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep

boltzmann machines," In *Proceedings of Advances in Neural Information Processing systems (NIPS)*, pp. 2222–2230, 2012.

[68] M. Stumpf, C. Wiuf, and R. May, "Subnets of scale-free networks are not scale-free: Sampling properties of networks," *The National Academy of Sciences of the United States of America*, Vol. 102, No. 12, pp. 4221–4224, 2005.

[69] M. Szell, R. Lambiotte, and S. Thurner, "Multirelational organization of large-scale social networks in an online world," *The National Academy of Sciences*, Vol. 107, No. 31, pp. 13 636–13 641, 2010.

[70] J. Tang, T. Lou, and J. Kleinberg, "Inferring social ties across heterogenous networks," In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 743–752, 2012.

[71] S. Trautmann-Lengsfeld, J. Domínguez-Borràs, C. Escera, M. Herrmann, and T. Fehr, "The perception of dynamic and static facial expressions of happiness and disgust investigated by erps and fmri constrained source analysis," *PLoS ONE*, Vol. 8, No. 6, pp. 1–18, 06 2013.

[72] A. Vinciarelli, "Mobile phones and social signal processing for analysis and understanding of dyadic conversations," In *Proceedings of Mobile Social Signal Processing (MSSP)*, pp. 1–8, 2014.

[73] H. Wallach, D. Mimno, and A. McCallum, "Rethinking lda: Why priors matter," In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2009, pp. 1973–1981.

[74] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo, "Mining advisor-advisee relationships from research publication networks," In *Proceedings of the 16th International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 203–212, 2010.

[75] X. Wei and W. Croft, "LDA-based document models for ad-hoc retrieval," In *Proceedings of the 29th ACM Special Interest Group on Information Retrieval (SIGIR)*, pp. 178–185, 2006.

[76] B. Zadrozny, J. Langford, and N. Abe, "Cost-sensitive learning by cost-proportionate example weighting," In *Proceedings of the 3rd IEEE International Conference on Data Mining*, pp. 435–442, 2003.

[77] J. Zheng and M. Ni, "An unsupervised learning approach to social circles detection in ego bluetooth proximity network," In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pp. 721–724, 2013.

[78] Y. Zheng, Y. Zhang, and H. Larochelle, "Topic modeling of multimodal data: an autoregressive approach," In *Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1370–1377, 2014.

전화와 근접 로그를 이용하여 소통 패턴을 찾기 위한 토픽 모델

한 용 진

경북대학교 대학원 컴퓨터학부 컴퓨터공학전공
(지도교수 박세영)

(초 록)

사회적 소통에 대한 이해는 유비쿼터스 응용 개발의 핵심 요소 중 하나이다. 최근 사회적 소통을 이해하기 위한 방법으로 모바일 장치를 이용한 소통 감지 및 패턴을 찾는 방법이 연구되고 있다. 이들 연구는 주로 전화와 근접 로그을 이용하여 사회적 소통을 표현한다. 이러한 소통은 토픽으로 특징지을 수 있기 때문에 Latent Dirichlet Allocation (LDA)기반의 토픽 모델이 주요하게 연구되고 있다. 하지만 이러한 연구들은 전화와 근접 로그를 독립적인 것으로 간주하여 이들 두 가지 유형의 소통을 동시에 다룰 때 얻을 수 있는 정보를 손실하는 한계가 있다. 본 논문은 모바일 로그부터 얻은 전화와 근접 정보를 동시에 이용하는 소통 패턴 학습 방법을 제안한다. 먼저, 기존에 소개된 토픽 모델을 이용하여 전화와 근접 패턴을 동시에 찾는 방법들에 대해 논하고 이들 방법의 한계를 극복하기 위한 새로운 토픽 모델을 제안한다. 새롭게 제안하는 토픽 모델은 전화와 근접 정보를 동일한(homogeneous) 유형으로 간주한다. 즉, 전화와 근접 로그는 서로 다른 파라미터를 가지는 동일한 분포로부터 생성된다. 일반적으로 근접 로그가 전화 로그에 비해 압도적으로 많고 규칙적인 특징이 있다. 이러한 이유로 제안하는 방법은 근접 정보에서 전화로의 한 방향의 영향력을 모델링하고 각 전화와 근접 로그는 LDA로 모델링하였다. MIT(Massachusetts Institute of Technology)의 Reality Mining 프로젝트에서 제공하는 세 가지 데이터 셋을 이용한 실험에서 제안한 방법이 전화와 근접 로그를 독립적으로 간주하는 기존의 접근에 비해 월등히 높은 성능을 보였다. 또한 제안한 방법을 사회적 관계를 분류하는 응용에 적용하여 기존 접근에 비해 높은 성능을 보임으로써 제안한 방법의 유용성을 증명하였다.