MINING CHAT LOGS TO EXTRACT INFORMATION ABOUT AUTHORS AND TOPICS FOR CRIME INVESTIGATION

ABDUR RAHMAN M. A. BASHER

A THESIS

IN

THE CONCORDIA INSTITUTE FOR INFORMATION SYSTEMS ENGINEERING

PRESENTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF APPLIED SCIENCE IN INFORMATION SYSTEMS SECURITY

CONCORDIA UNIVERSITY

MONTRÉAL, QUÉBEC, CANADA

NOVEMBER 2011

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Abdur Rahman M. A. Basher**

Entitled: **Mining Chat Logs to Extract Information about Authors and Topics for Crime Investigation**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science in Information Systems Security

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

	Dr. Lingyu Wang	Chair
	Dr. Yong Zeng	Examiner
	Dr. Juergen Rilling	Examiner
	Dr. Benjamin Fung	Supervisor
Approved by	**Dr. Chadi M. Assi**	Graduate Program Director

20

Dr. Robin A. L. Drew, Dean

Faculty of Engineering and Computer Science

Abstract

Mining Chat Logs to Extract Information about Authors and Topics for Crime Investigation

Abdur Rahman M. A. Basher

Cybercriminals have been using the Internet to accomplish illegitimate activities and to execute catastrophic attacks. Computer Mediated Communication, such as online chat, provides an anonymous channel for predators to exploit victims. In order to prosecute criminals in a court of law, an investigator often needs to extract evidence from a large volume of chat messages. Most of the existing search tools are keyword-based, and the search terms are provided by an investigator. The quality of the retrieved results depends on the search terms provided. Due to the large volume of chat messages and the large number of participants in public chat rooms, the process is usually time-consuming and error-prone. This thesis presents a topic search model to analyze archives of chat logs for segregating crime-relevant logs from others. Specifically, we propose an extension of the Latent Dirichlet Allocation (LDA)-based model to extract topics, compute the contribution of authors in these topics, and study the transitions of these topics over time. In addition, we present another unique model for characterizing authors-topics over time. This is crucial for investigation because it provides a view of the activity in which authors are involved

in certain topics. Experiments on two real-life datasets suggest that the proposed approach can discover hidden criminal topics and the distribution of authors to these topics.

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Dr. Benjamin Fung, for his many suggestions, enthusiasm, and constant support over the last year. I am thankful for all the assistance he gave me, especially at times when I found it is very difficult to continue.

My gratitude also goes to the thesis reviewers for the time they spent patiently reading through my thesis, and providing valuable feedback that has served to improve it. This thesis would not have been possible without their strongest support.

I am indebted to my many of my friends for the help and knowledge they shared with me on several aspects of life, and I am grateful for their companionship during the writing of my thesis.

Last, but definitely not least, I am endlessly grateful to my dear parents whose dedication, love and persistent guidance, has taken the load off my shoulder. I also like to thank my Siblings for their unwavering support and motivation throughout this entire process. This thesis would not have been possible without the continuous assistance from my family who gave me the strength and will to succeed.

"Read! In the Name of your Lord, Who has created (all that exists), has created man from a clot of congealed blood. Read! And your Lord is the Most Generous, Who has taught (the writing) by the pen, has taught man that which he knew not. Nay! Verily, man does transgress all bounds (in disbelief and evil deed, etc.). Because he believes himself self-sufficient. Surely to your Lord is your return." - Chapter Al-Àlaq (The Clot) [96:1-8], The Holy Qurán.

To my Parents and Family

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Demand for Computer Mediated Communication, such as online chat, instant messaging, blogs, and twitts, are growing tremendously due to its efficiency both in delivering messages on time and its costs effectiveness. Many software applications have been developed to serve this demand. Instant messaging seems to be the preferred type of communication, especially chatting, because it provides one-to-one or one-to-many instant communication, and it can also handle video and audio calls as well. They provide effectiveness not for only personal uses, but also for business, advertising, and e-commerce.

Cyber chat is becoming a global concern since it has become a venue for conducting illegitimate activities. Illegitimate activities include *cyber stalking*, *online contact*, *online harassment*, and *degradation* [Han08]. Cyber stalking is the spreading offensive words and statements against the another person online within the same channel or other channels as long as the predator knows the real identity of the selected victim. The aim of the perpetrator's in conducting this crime is the desire for control and power. Cyber stalking has the potential to very quickly move from cyberspace to real life. Online contact can lead to offline harm when the predator gains the trust of the victim in order to abuse them in real life, either physically, sexually, or financially. It is certainly intended for conducting crimes. Online harassment includes the use of words or actions that abuse others through instant

messaging and especially live video streaming. This may also include threats, rumors, mocking, disclosure of unauthorized sensitive information, defamation of character, coarse language, name calling, personal attacks, child harassment, sexual intimidation, sexual harassment, and so on. Degradation refers to insulting individuals and groups through disrespectful images or words that may cause harm to them. This is mostly used in the sexual arena but could also be extended to racial, religious, and political insults. [Han08].

These traditional crimes, which are conducted through the internet as a medium, poses new challenges for law enforcement agencies to prevent, detect, investigate, and prosecute perpetrators. Unfortunately, the capability of current crime-investigation software tools does not fully meet the actual needs of real-life investigation.

In this thesis, we introduce a method for forensics investigators to utilize when performing a search analysis, given a collection of chat logs. The work is divided into four core stages: searching crime-relevant logs, discovering crime-relevant topics from identified criminal logs, estimating the contribution of authors in the discovered topics, and representing transitions of the crime-related topics over time. We first identify whether a given chat log is crime-relevant or not, based on the predefined criminal topics. Once the crime-related chat log is determined, we deploy a probabilistic topic model to extract the hidden semantic structure of the logs. Next, the authors' contributions within the discovered topics are estimated. Finally, an evolution of topics under some specific time intervals is generated. In certain cases, investigators are required to distinguish certain authors from others within some interval of time. This is obtained by including another stage to compute the bond composed of authors-topics trends over time.

1.1 Motivation

Suppose an investigator seizes a suspect's computer that has an enormous amount of chat logs from, e.g., Windows Live Messenger or IRC chat rooms. The chat logs sometimes

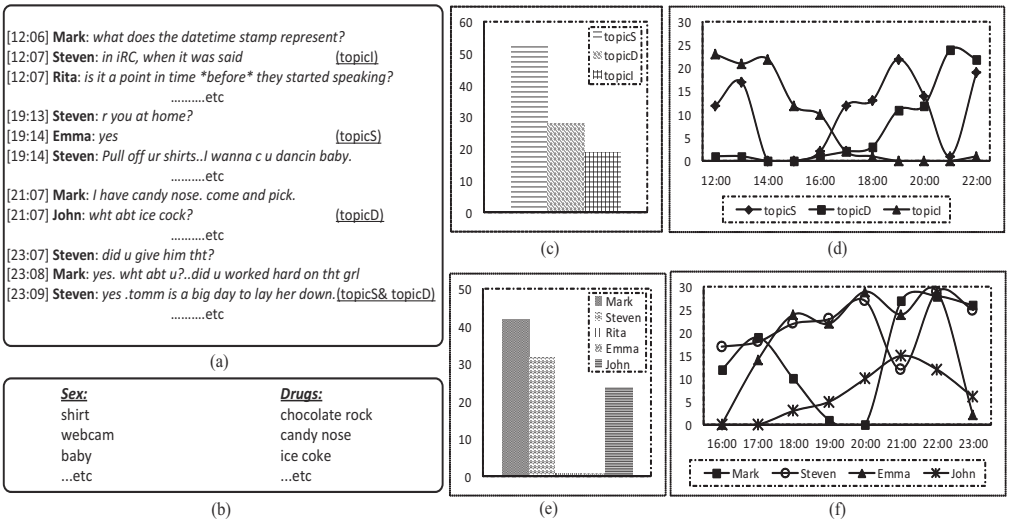


Figure 1: (a)- A detained chat log d. (b)- Criminal topics (Sex and Drugs) with their associated terms. (c)- Topics distribution in the chat log d. (d)- Topics over time in the chat log d. (e)- Authors distribution over topic topicD in the chat log d. (f)- Authors-Topics over time in the chat log d for topic topicS.

contain important information that is directly or indirectly related to the criminal activities under investigation. In Figure 1(a), it presents a general form of chat log that contains information about criminal activities, such as Sex and Drugs.

The challenge is how to effectively and efficiently extract the relevant information and evidence from a large volume of chat messages. In this thesis, we propose a discovery method, in a context of chat log-topic and topic-author relations, to answer the following questions that are frequently raised by investigators:

Q 1. *How can an investigator determine which logs are crime-relevant? In identifying a crime-relevant log, what are the contributed topics in the log file? How have they evolved over time? Moreover, how can an investigator extract the topics that are crime-related from the identified crime-relevant log files?*

Q 2. *Who are the contributors to a topic in a given chat log? How can an investigator track the activity of authors in a log file?*

In general, we are concerned with generating Figures 1 (c)(d)(e) and (f), as results from our research questions.

However, the existing topics discovery methods [BNJ03,RZGSS04,WMM05,RHNM09] cannot be directly manipulated to address the problem illustrated in the context of crime investigation, because of the differences in characteristics of chat messages from traditional documents [HD10], such as historical or scientific articles:

- Chat is informal and its content is not well structured. Chat often contains spoken languages with a lot of grammatical and spelling mistakes.
- Transliteration is often used and refers to writing, or spelling words, or letters in a language written using a different alphabet or script.
- The contents (topics) in chat logs change frequently and implicitly over time as consequences of incoherence of message sequences.
- Messages on chat logs are often shorter per author, ranging from a few words to a few lines.
- Authors within these short messages use many deceptive techniques for covert communication. For example, they use emoticons to express human facial behavior that complements a text message. Moreover, the actual meaning of the words used within a chat log is different from their apparent meaning; street terms are more frequently used in the context of illegitimate activities. For example, the word 'snow' used in drugs trafficking means cocaine.

As a result, criminal topics extraction from log files requires special handling, and the analytical techniques used widely for mining texts of literary and historic documents may

not achieve the same accuracy when applied to online documents. Furthermore, these techniques do not collect information about authors composing criminal topics.

Therefore, it is essential to present a method that precisely captures the various characteristics of chat logs. This method includes the ability to discover crime-related topics and to predict the authors of these topics. The topics discovered by this method would also expose characteristics of different topics for further investigation, such as the percentage of the topics, highly used terms, and the evolution of topics as a distribution over a given time. All of the aftermentioned problems motivate us to build a tool that includes all these tasks. We will present a detailed explanation on this tool later on in this thesis.

1.2 Problem Definition

In this thesis, we assume the user of our method is a crime investigator who has access to a collection of chat log documents, and who would like to analyze the relationship between the topics discussed and the participating authors. We formally define an abstract representation of chat log documents, user-specified criminal topics, and some basic notions of topics and authors, followed by a problem statement.

Definition 1 (**Chat log document**). A *chat message* is a triplet (a, μ, τ), representing an author a writing a piece of text μ at time τ. A *chat log document*, denoted by d, is a sequence of chat messages ordered by τ. ■

***Example* 1.** In Figure 1(a), *Mark* wrote the text message *"I have candy nose. come and pick."* at time *[21:07]*. This chat message is represented by a triplet *(Mark, "I have candy nose. come and pick.", [21:07])*. The chat log document is a sequence of chat messages ordered by time. ■

An investigator wants to identify the crime-relevant topics discussed in a chat log document and the authors participated in the discussion of the topics. The following definitions

define such notions.

Definition 2 (**Crime-relevant topic**). Let c be a criminal topic from a set of investigator-specified criminal topics C. Let t be a topic, discussed in a chat log d, from a set of topics K_d discussed in d. Let $distance(t_1, t_2)$ be a function that describes the dissimilarity of the two topics t_1 and t_2. A topic t is *relevant* to a criminal topic c, if $distance(t, c) \leq \gamma$, where γ is an investigator-specified relevance threshold. ■

***Example* 2.** The chat log document d in Figure 1(a) contains three topics $K_d = \{topicI, topicS, topicD\}$. Figure 1(b) illustrates two investigator-specified criminal topics $C = \{Sex, Drugs\}$. Suppose $\gamma = 0.2$ and $support(topicD, Drugs) = 0.56$. The $topicD$, discussed in d, is relevant to the criminal topic $Drugs$, if $distance(topicD, Drugs) \leq \gamma = 0.2$. Chapter 4 will define the $distance$ function. ■

To identify relevant criminal information from a large collection of chat log documents, an investigator first has to identify the crime-relevant documents, and then the topics' distribution with respect to authors over time. The following definitions formally capture these notions.

Definition 3 (**Crime-relevant document**). A chat log document d is *crime-relevant*, if d contains at least one crime-relevant topic. ■

Definition 4 (**Active topic**). Let $[\tau_t^s, \tau_t^f]$ be a time interval of topic t discussed in a chat log document. The *active* level of t over the time interval $[\tau_t^s, \tau_t^f]$ is described by function $F(t)_{\tau^s}^{\tau^f}$. ■

Definition 5 (**Active author**). Let Λ_d be a set of authors participating in chat log document d. Let Λ_d^t be a set of authors participating in a topic t in d, where $\Lambda_d^t \subseteq \Lambda_d$. The *active* level of an author $a^t \in \Lambda_d^t$ is defined by $F(a_d^t)_{\tau^s}^{\tau^f}$ provided t is active during $[\tau_t^s, \tau_t^f]$. ■

***Example* 3.** Figure 1(d) depicts the active levels of $topicI$, $topicS$, and $topicD$ between *12:00* and *22:00*. For example, $topicD$ is actively discussed between *20:00* and *22:00*, but

is relatively inactive between *12:00* and *13:00*. Figure 1(f) defines the evolution (active level) of authors over the previous time intervals. ■

The problem studied in this thesis is formally defined as follows:

Definition 6 (**Authors-Criminal topics activity over time in a chat log).** Given a collection of chat log documents L, a set of criminal topics C, and a relevance threshold γ, the problem is:

1. to identify all crime-relevant documents from L,
2. to identify all crime-relevant topics in each document $d \in L$ with respect to C and γ, and
3. to identify the active level of crime-relevant topics, and all their associated active authors over a given time interval $[\tau_t^s, \tau_t^f]$ for each identified crime-relevant document. ■

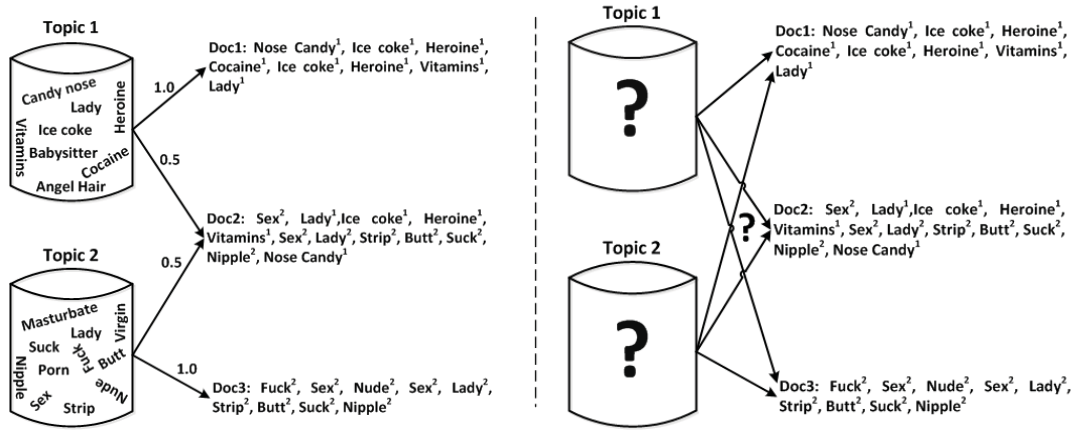
Before moving forward, we first introduce the terminology and notations provided in Table 1 that will be used throughout the rest of this thesis.

1.3 Probabilistic topic model

If we augment that a document is composed of words and a subset of these words describe a topic, then a document is considered to be mixtures of topics. This is the intuition behind *topic modeling* [BL09]. As from the definition, a topic model is a generative model, which describes how words are generated from the latent random variables (topics) through some probabilistic procedure. In this model, the words are observed while the topics are hidden (latent), and a topic is a probability distribution over words. Therefore, the primary goal for generative model is to explore the best set of topics that can explain the observed words, under the assumption that the model actually generated the documents.

Table 1: Notations used in this thesis

SYMBOL	DESCRIPTION
α	Dirichlet parameters for topics (Dirichlet prior)
$\bar{\alpha}$	Dirichlet parameters for authors (Dirichlet prior)
β	Topic-dependent Dirichlet parameters for word index (Dirichlet prior)
λ	Topic-dependent Dirichlet parameters for time slots (Dirichlet prior)
θ	Multinomial distribution of topics given the documents in the corpus
ϑ	Multinomial distribution of topics given the authors for the documents in the corpus
φ	Multinomial distribution of words to topics
η	Multinomial distribution of time intervals to topics for the documents in the corpus
D	Number of documents
d_c	Crime-relevant document (chat log)
T	Number of topics
c	Criminal topic
A	Number of authors
V	Number of unique words in the vocabulary
N_d	Number of word tokens specific to the document d
z	Topic indices
Λ_d	Set of authors in the dth document
x	Author assignments



(a) Probabilistic generative process (b) Statistical inference

Figure 2: Illustration of (a)- the generative process and (b)- the problem of statistical inference underlying topic models. The superscript numbers associated with the words in documents represents the topic that words are sampled from.

We describe the the topic modeling procedure by Figure 2 (from the paper [SG07]). In this figure, topic modeling method is described in two separate ways: as a generative model, and as a problem of statistical inference. From Figure 2(a), topics 1 and 2 are related to drugs and sex, respectively, and they contain different distribution over their relative words. Documents are generated by choosing words that correspond to topics, depending on the weights (with the arrow) provided to the topics. For example, documents 1 and 3 are generated by picking words only from topic 1 and 2, respectively, while document 2 is generated from the two topics with equal distributions. Therefore, documents with different content are generated by choosing different distributions over topics; there is no notion of mutual exclusivity that restricts words to be drawn only from a single topic. In addition, the same word, such as "Lady", can appear in both topics with different probabilities; this is known by *polysemy*. Furthermore, this process does not address the order of words as they appear in document. Thus, the *"bag of words"* assumption in this model is applied [SG07].

On the other hand, if the task is to search for topics that compromise a given document

in a reverse process, then the statistical inference is applied. Figure 2(b) illustrates this assumption. This involves inferring the two distributions: multinomial distribution over words associated to each topic, and multinomial distribution over topics for each document. That is what the probabilistic topic modeling is about. Probabilistic topic modeling algorithms are statistical methods that analyze the words of the original texts to discover the topics of documents, how those topics are connected to each other, and how have they changed over time. In this thesis, we use probabilistic topic model to infer the set of topics that are responsible for generating a collection of chat logs. Afterward, crime-related topics are explored, if existed in chat logs, from the discovered topics [SG07].

In this thesis, our approach is based on probabilistic topic models, because these models posit in general several advantages [BL09]. First, they rely on the semantic information derived from a word-document co-occurrence matrix. Second, they reduce the feature dimensions. Third, generative models are easily applied to new data, especially for information retrieval or classification. Fourth, they can be used easily as a component in far more complicated topic models. Finally, generative models are *general*; it could be other data instead of words, *LDA* has been extended to encounter other research fields, such as: object recognition [CFF07], natural language processing [GSBT05], video analysis [WMG07], collaborative filtering [Mar03], spam filtering [BSSB09], web-mining [MLSZ06], authorship disambiguation [RZGSS04, RZCG+10], and dialogue segmentation [PGKT06].

1.4 Contributions of the Thesis

The major contributions of this thesis can be summarized as follows:

Criminal Topic model. We present a *Criminal Topic (CT)* model that is an extended version from the *mixture of unigrams* model. In this model, we assume two observed variables: a single topic, and its associated words. For estimating the distribution

of words, we train the model with several hundreds of crime-related logs. From the learning phase, each term in a single criminal topic c is associated to a weight that describes the appearance of this term in c. Therefore, when chat logs are collected, CT model is applied to infer whether these detained logs are crime-relevant or not.

Identifying crime-related chat logs. We provide a mechanism in identifying crime-relevant logs. At the beginning, we transfer each chat log d to a language model M_d, and then we compare M_d with the Criminal Topic model. Based on the results, we identify crime-related logs. The algorithm is proven efficient in searching the crime-related logs, in a huge collection, whenever adequate terms are provided in the CT model.

Identifying topics activity in chat logs. We develop a *LDA-TOT* model that discovers the distribution of topics over passage of time intervals. In order to capture the activity of topics, we define the transition function $F(t)_{\tau^s}^{\tau^f}$. This contribution provides the investigators enough data to help them deducing the active crime-related topics in logs. Moreover, LDA-TOT can predict topics expressed in logs, if the distributions over time is determined. In addition, the model can identify topics activity in chat logs without knowing the timestamps by defining the distribution of topics. For the inference purpose, we use the Gibbs sampling algorithm.

Identifying authors activity in chat logs. We present a *A-TOT* model to describe authors' distribution over topics given time. In addition, we define the transition function $F(a_d^t)_{\tau^s}^{\tau^f}$ to capture the authors' activity over topics during time intervals. This is the major contribution in our work. To the best of our knowledge, A-TOT is the first model that explains authors' distributions in topics together with the topics distributions over time.

This model helps investigators predict authors given a document with unknown writers either, by the distribution of topics over document (θ), or by the distribution of

topics over time (η). This thesis adds a new dimension into the pre-existing models by introducing the evolution of authors over timeslots. For the inference purpose, we employ the Gibbs sampling, as we do for LDA-TOT.

In addition to the above contributions, we study the two research questions on real data, where efficiency and scalability is achieved using both new models, LDA-TOT and A-TOT.

1.5 Thesis Organization

The rest of the thesis is organized as follows:

In Chapter 2, we provide a comprehensive literature review related to topics discovery and the extended models from LDA.

In Chapter 3, we describe the background information relevant to our proposed models specifically LDA, AT, and TOT models.

In Chapter 4, we explain the n-grams model in general and elaborate on Criminal Topic model, as an extension from the mixture of unigrams model. In addition, we explore some of the measurements used to evaluate our proposed method.

In Chapter 5, we present the LDA-Topics over time model that explicitly models time jointly with words co-occurrence patterns. This model is an extension from both models, LDA and TOT. It uses discretization of time in describing the evolution of topics over time. We also describe the algorithm for searching crime-related logs and show the experimental results using LDA-TOT model on real-life datasets.

In Chapter 6, we present a new Author-Topics over Time model. We explain the ultimate objectives of proposing this model and describe the results obtained using this model.

Finally, Chapter 7 summarizes our research and describes the limitations and possible future research directions.

Chapter 2

Literature Review

We summarize the state of the art in the literature of topics discovery and modeling. Blei et al. [BNJ03] proposed the *Latent Dirichlet Allocation* (*LDA*) model to extract topics and summarize a document corpus. The general idea of LDA is to generate a discrete distribution of words per topic and a discrete distribution over topics per document. In the following sections, we describe extensions of LDA to topics labeling, authors' distributions in topics and their relationships, predicting authors using writeprints, topics' progressive information over time, and modeling topics in microblogging environment.

2.1 Labeling topics

LDA is expressive enough to reveal topics in a document, but does not provide a way of including labels in its learning procedure. Hence, LDA has been adapted in applications for topic labeling, as in [BM08, LJSJ08, RHMGM09, RHNM09]. Blei et al. [BM08] proposed *Supervised LDA* (*sLDA*), where a label is generated from each document's empirical topic mixture distribution. Lacoste et al. [LJSJ08] proposed *Discriminate variation on Latent Dirichlet Allocation (DiscLDA)*, where a document is related to a categorical variable or class label, and a topic mixture distribution is associated with each label. However,

these models use single labeling to a document and do not provide multiple labels to each document. *Multi-Multinomial LDA* (*MM-LDA*) [RHMGM09] assigns multiple labels for each document. Unfortunately, MM-LDA's learned topics do not link directly with the label. Therefore, Ramage et al. [RHNM09] proposed the *Labeled-LDA* (*L-LDA*) model to directly associate each observed document's label set with one topic.

Our way to solve topics labeling is by introducing a *Criminal Topic* model that includes predefined terms, with their distributions associated to each criminal topic. The discovered topics are labeled as crime-relevant whenever the distributions of these topics and topics from the Criminal Topic model are assumed to be relevant through some distance measurement.

2.2 Modeling authors with topics and their relationships among each other

Several extensions of LDA models have been proposed to identify authors and the proportion of each author in a document. For example, Rosen-Zvi et al. [RZGSS04] introduced an *Author-Topic* (*AT*) model, a generative model for authors and their corresponding topic distributions. In their experiments, AT seems to outperform LDA when the test documents contain few observed words.

Other works have been extended further to deduce the social networks between entities in different types of documents [CBGB09, MWM07, ZGFY07, LNMG09, NAXC08, SLTS05]. Chang et al. [CBGB09] presented a probabilistic topic model to describe the relationships between pairs of entities encoded in a collection of texts. First, the entities are extracted, and then document is divided into two different class of bag of words: entity context relates to an entity, and pair context relates to the pair of entities. From this assumption, the topic is modeled and the relationships between entities are inferred. McCallum

et al. [WMM05] proposed a *Group-Topic (GT)* model to cluster entities into groups with relations between them. In their model, the discovery of groups is guided by the emerging topics and the discovery of topics is guided by emerging groups. In addition, the model is able to capture the language attribute being used within entities, and this helps to assign group memberships. Their experimental results suggest that the inference of joint probability improves both the performance of both groups and topics discovery. Zhang et al. [ZGFY07] introduced a *Generic weighted network-Latent Dirichlet Allocation (GWN-LDA)* model for discovering probabilistic community profiles as distributions on the entire social actor space. Therefore, each social actor belongs to every community with different probability and contributes a part, big or small, to every community in the society. Similar work can be found in Liu et al. [LNMG09]. Nallapati et al. [NAXC08] introduced a model called *Pairwise-Link-LDA*, which models the activity in term of absence or presence of a link between every pair of documents. Their work mainly addresses the problem of joint modeling of text and citations in the topic modeling framework. However, they perform modeling, based on absence or existence of a link in every pair of documents, and this does not fit with large scale authors' networks. *CommunityNet*, a personal profile, is developed by Song et al. [SLTS05], which went a further step in predicting authors behavior in receiving and sending information, by analyzing the contact and content of personal communications.

In our approach, we modify the AT model to accommodate the evolution of topics discovered and the proportion of authors to these topics over time.

2.3 Predicting authors using writeprints

A somewhat different approach to topic modeling in predicting authors is to extract a set of features, writeprints, from collections of online documents, where values in this set differ from each author. Based on the features, authors profile is built as in [KCAC08,

ZQHC03, dVACM01, AC08, IHFD08, IBFDss, IBFD10]. To predict the plausible authors in unknown authors of documents, the same set of features is applied in these documents. Next, similarity is used to infer authors, based on the extracted features from unknown documents and authors profile. While this approach can provide useful broad information about authors, associating authors to topics is not studied in their works.

2.4 Modeling temporal information in topics

Studying the evolution of topics over time is valuable, because it reveals different characteristics of topics and their authors. Wang et al. [WM06] proposed the *Topics Over Time* (*TOT*) model, a non-Markov continuous time model of topical transitions. TOT models timestamps by parameterizing a continuous beta distribution over time with each topic. They assume that the meaning of a particular topic can be relied upon as constant, but its occurrence and correlations change significantly over time. *Dynamic Topic Model (DTM)* by Blei et al. [BL06] takes a slightly different approach. It explicitly models the evolution of topics with time by estimating the topic distribution at various time stamps. Therefore, it is easier to predict the words in a particular topic at different points in time. However, DTM does not yield a simple solution to the problems of inference and estimation, and it ignores the time dependency of individual documents inside a collection/period. The *Continuous Time Dynamic Topic Model* (*cDTM*) [WBH08] replaces the discrete state space model of the DTM [BL06] with its continuous form, called *Brownian motion*. The topics are modeled through a sequential collection of documents, where a topic is a pattern of word use that is expected to change over the course of the collection. Significantly, *cDTM* generalizes the *DTM* in that the only discretization it models is the resolution at which the timestamps of the documents are measured. Nallapati et al. [NDLU07] *Multiscale Topic Tomography Model (MTTM)*, which employs conjugate priors using non-homogeneous Poisson processes to model generation of word-counts. In addition, the evolution of topics could be

modeled at various time-scales of resolution using *Haar wavelets*. AlSumait et al. [ABD08] used an online version of LDA model (*OLDA*), where topics are evolved through incremental updates for new data based on the current position.

To collect the distribution of topics over time, we employ an extended combination of three models, LDA, AT, and TOT, where discretization of timeslots is used, because the time intervals in a chat log are relatively short, from a few minutes to a few hours.

2.5 Modeling topics in microblogging environment

The topic models discussed in most of the current literature are applied to structured documents, which are quite different from chat logs. As a result, it becomes very difficult to obtain an accurate model from logs. Hong et al. [HD10] focused on online messages, particularly Twitter. They conducted an empirical study of different strategies to aggregate tweets, based on the existing models. Li et al. [LJW10] presented an approach to resolve the sparsity of data in short texts environment, such as chat logs, by assigning a single topic for a whole sentence. This is done by clustering semantically related sentences patterns that are likely about the same aspect, and then frequent subtree pattern mining is applied to generate sentence patterns that can represent the aspects. Similar aspects can be found in [SSRZG04].

In contrast, our work focuses on four major aspects: criminal topics discovery, authors' proportions with respect to topics, evolution of topics with respect to time, and evolution of authors-topics over time.

Chapter 3

Preliminaries

Due to the massive amount of electronic documents available today, it is necessary to have some efficient methods for summarization, organization, management, and information retrieval. For example, it is always required to know the summary of the documents, their relationships among them in a corpus, tracking authors of these documents, and their trends over time. Therefore, it becomes increasingly important for searching and indexing a large collection of text data. Topics modeling of text collections is a widely growing field of study that serves such those needs. *Latent Dirichlet Allocation (LDA)* is the most widely used for modeling purposes. In this chapter, we briefly describe the statistical topic models, *LDA*, *AT*, and *TOT* to provide a sound theoretical foundation to our research problem in Chapter 1.

3.1 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (*LDA*) [BNJ03] is an unsupervised generative probabilistic model that discovers latent semantic topics in a corpus with large collections of discrete data, such as the words in a set of documents. It is based on a *"bag of words"* assumption, which treats each document as a frequency of word counts, ignoring the order of

appearance. In the language of probability theory, this is an assumption of *"exchangeability"*: words are *independent* and *identically* distributed over the topics, and the topics are *infinitely* exchangeable throughout the document, based on some conditional parameters [BNJ03, ABD08]. This conditionally independence allows us to build a hierarchical Bayesian model for a corpus of documents and words.

The intuition behind LDA is the assumption that words carry strong semantic information about a document. Therefore, it is reasonable to assume that documents with similar topics will use the same collection of words. These similar topics, latent topics, are discovered by identifying groups of words in the corpus that frequently occur together within documents. The mixture of (latent) topics in document collection summarizes the content and the underline thematic structure of documents quantitatively. Moreover, this distribution of topics assists in searching and indexing a large collection of text data, by comparing how similar one document is to another through measuring the similarity on the corresponding topic mixtures.

3.1.1 Generative Process

In LDA, a document can be viewed as a random mixture of hidden variables (i.e., topics) and observed data (i.e., words). Words in a document are generated from the hidden topics and are not linked to the documents directly, but are linked via latent variables (topics) that are responsible for using a particular word in the document drawn from a specific topic distribution that the document focuses on. Therefore, LDA is considered to be a three-level hierarchical Bayesian network.

In general, the graphical model of LDA is represented by plate notation in Figure 3. For readers not familiar with plate notation, shaded and unshaded variables indicate observed and latent variables, respectively. An arrow between variables indicates a conditional dependency, and the boxes (plates) in the figure indicate replication with the number

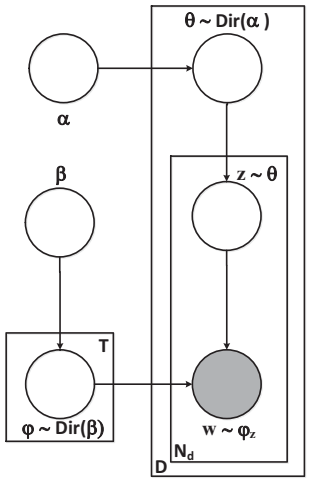


Figure 3: The graphical model representation (plate notation) of Latent Dirichlet Allocation (LDA)

of repetitions given by the variable in the bottom [Bun94]. The generative process can be described as follows:

1. For each document d, choose D multinomials $\theta_d \sim$ Dirichlet prior α;

2. For each topic t, choose T multinomials $\varphi_t \sim$ Dirichlet prior β;

3. For each word w_{di} per document d, in the corpus:

   - choose a topic $z_i \sim$ multinomial θ_d; $(P(z_i \mid \alpha))$
   - choose a word $w_i \sim$ multinomial φ_z; $(P(w_i \mid z_i, \beta))$

More specifically, for each document d in the corpus, the LDA first picks a multinomial distribution θ_d, from the Dirichlet prior α, and then the a topic $z_i = t$ is assigned to the *i*th word in the document, according to θ_d that determines which topics are most likely to appear in a document. Based on $z_i = t$, the model then chooses a word w_i, from the vocabulary of V words, according to the multinomial distribution φ_z that is generated from the Dirichlet prior β for each topic t. From this procedure, we observe that each word in

a document is generated by a different topic at random. As a result, documents in LDA exhibit mixture of topics distributions unlike the *mixture of unigrams* model (discussed in the Chapter 4). In addition, LDA does not attempt to model the order of words within a document. Thus, the "bag of words" concept is assumed in this model.

The topic weight vector θ is a $D \times T$ matrix that is estimated from data. It describes the distribution of each of the T topics over D documents, where $\sum_t^T \theta_{d,t}$=1. The word weight vector φ is a $V \times T$ matrix that is also estimated from data, and it defines the distribution of each of the T topics over V words, where $\sum_v^V \theta_{v,t}$=1.

3.1.2 Inference using Gibbs sampling

The previous generative procedure defines the following joint distribution of all variables:

$$\begin{aligned} P(z, w, \theta, \varphi \mid \alpha, \beta) &= P(\theta \mid \alpha) P(z \mid \theta) P(\varphi \mid \beta) P(w \mid z, \varphi) \\ &= (\prod_{d=1}^{D} \frac{\Gamma(\alpha)}{\prod_{t=1}^{T} \Gamma(\alpha_t)} \prod_{t=1}^{T} \theta_{t,d}^{\alpha_t + n_{td} - 1}) \\ &\quad \times (\prod_{t=1}^{T} \frac{\Gamma(\beta_t)}{\prod_{v=1}^{V} \Gamma(\beta_{v,t})} \Pi_{v=1}^{V} \varphi_{v,t}^{\beta_{b,t} + n_{v,t} - 1}) \end{aligned} \tag{1}$$

The detail of the joint probability is outlined in paper [Hei04]. Now, the distribution of the latent topic variables conditioned on the words is computed as:

$$P(z, \theta \mid w, \alpha, \beta) = \frac{P(z, \theta, w \mid \alpha, \beta)}{P(w \mid \alpha, \beta)} \tag{2}$$

$P(w \mid \alpha, \beta)$ represents the marginal distribution, likelihood, of a document. We normalize $P(w \mid \alpha, \beta)$ by marginalize over the hidden variables, and the resulting margin probability is expressed by:

$$\begin{aligned} P(w \mid \alpha, \beta) &= \int_{\varphi} \int_{\theta} \Sigma_z (\prod_{d=1}^{D} \frac{\Gamma(\alpha)}{\prod_{t=1}^{T} \Gamma(\alpha_t)} \prod_{t=1}^{T} \theta_{t,d}^{\alpha_t + n_{t,d} - 1}) \\ &\quad \times (\prod_{t=1}^{T} \frac{\Gamma(\beta_t)}{\prod_{v=1}^{V} \Gamma(\beta_{v,t})} \prod_{v=1}^{V} \varphi_{v,t}^{\beta_{b,t} + n_{v,t} - 1}) \end{aligned} \tag{3}$$

Estimating θ and φ, which provides the topics' proportions in each document and words' proportions to these topics, respectively, from the above function are intractable due to the coupling between θ and φ in the summation over latent topics z. Therefore, different complex algorithms have been proposed, including *variational inference* [BNJ03], *expectation propagation* [ML02], and *Gibbs sampling* [GS04]. Gibbs sampling is a form of *Markov Chain Monte Carlo*, which is used for obtaining an approximate inference about parameters in an iterative process. Throughout this thesis, we apply Gibbs sampling for inference purpose. In this model, the posterior distribution of topics over words is calculated as follows:

$$P(z_i = t \mid z_{-i}, w_i, \alpha, \beta) \propto \frac{n_{w_i}^{V,T} + \beta}{\sum_v^V n_{w_{-i}t}^{V,T} + V\beta} \times \frac{n_{d_i,t}^{D,T} + \alpha}{\sum_t^T n_{d_{-i},t}^{D,T} + T\alpha} \tag{4}$$

where $n_{w_{-i},t}^{V,T}$ is the vector count of the word w being assigned to the topic t, not including current word i. $n_{d_{-i},t}^{D,T}$ is the vector count of topic t being assigned to some words, not including the current word i, in a document d. After several iterations specified by the user, the multinomial distribution of documents over topics θ and the multinomial distribution of topics over words φ are obtained from the posterior distribution of topics. The details for the Gibbs sampling and LDA can be found in [Hei04, BNJ03], respectively.

3.2 Author-Topic (AT)

LDA discloses the underlying topics in the documents in a corpus. However, LDA does not identify a document's authors nor authors' association to each topic in a document's topics.

In this section, we discuss an algorithm that extracts both: the topics expressed in document collections and the authors' distributions over these topics. The algorithm follows

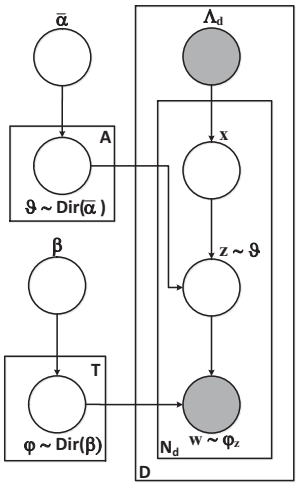


Figure 4: The graphical model representation (plate notation) of Author-Topic (AT) model

the probabilistic modeling, where documents represent as a mixture of multiple topics associated to authors and topics are probability distributions over words. This model is called *Author-Topic (AT)*, and it is proposed by Rosen-Zvi et al. [RZGSS04]. The model combines the strength of the two models, LDA and *Author* (also termed a *Multilabel Mixture Model*) [McC99]. AT model assumes that the authorship information in each document, in a corpus, is provided.

AT model provides support to a variety of communicating and exploratory queries in a set of documents with authors, including finding the authors who are most likely to write a given topic, and finding the most unusual paper written by a given author [SSRZG04].

3.2.1 Generative Process

AT model is an extension of LDA model and it does not only discover what topics are expressed in a document, but also which authors are associated with each topic. The model is based on the "bag of words" assumption as LDA. A document, in a collection, exhibit multiple topics that are a mixture of distributions associated with the authors.

The generative process for this model is shown in Figure 4:

1. For each author a, choose A multinomials $\vartheta_a \sim$ Dirichlet prior $\bar{\alpha}$;

2. For each topic t, choose T multinomials $\varphi_t \sim$ Dirichlet prior β;

3. For each word w_{di} in each document d, in the corpus:

   - choose an author $x_i \sim$ uniform Λ_d; $(P(x_i \mid \Lambda_d))$
   - choose a topic $z_i \sim$ multinomial ϑ_a; $(P(z_i \mid x_i, \bar{\alpha}))$
   - choose a word $w_i \sim$ multinomial φ_z; $(P(w_i \mid z_i, \beta))$

Formally, the procedure for generating a document starts by choosing an author x, uniformly at random, from the set of authors Λ_d for each word w_i specific to the document d, and then a topic is sampled from the distribution of topics specific to that author x. Finally, the words are sampled from the distribution of topics over words [RZGSS04]. This process is continued for all words in the document. However, it is important to note that there is no topic mixture for an individual document [HD10]. In other words, the multinomial distribution θ_d of topics, given documents, is not sampled in AT model unlike the LDA model.

3.2.2 Inference using Gibbs sampling

In the AT model, observed variables are not only include the words w in a document, but also the set of authors Λ_d in each document d. In addition, each word w in a document is consists of two latent variables: an author x and a topic z.

An analogy to LDA, the Gibbs sampler for the posterior distribution of topics is:

$$P(z_i = t, x_i = a \mid w_i = w, z_{-i}, w_{-i}, x_{-i}, A, \bar{\alpha}, \beta) \propto \frac{n_{w_i}^{V,T} + \beta}{\sum_v^V n_{w_{-i},t}^{V,T} + V\beta} \times \frac{n_{x_i,t}^{A,T} + \bar{\alpha}}{\sum_t^T n_{x_{-i},t}^{A,T} + T\bar{\alpha}} \quad (5)$$

where $n^{V,T}_{w_{-i},t}$ is the vector counts of the word w being assigned to the topic t, not including current word i, and $n^{A,T}_{x_{-i},t}$ is the vector count of words being assigned to topic t for author a to some words, not including the current word i.

Comparing to the LDA, AT seems to outperform LDA when relatively little is known about a new document, but the LDA model produces better distribution over topics of the content of individual documents when the observed words are outnumbered.

In summary, the AT model is a relatively simple probabilistic model for discovering the relationships between authors, documents, topics, and words. The important of this model can be explained in terms of providing a general framework for queries that explore authors together with documents. Furthermore, this model could be incorporated in identifications of authors in document collection, not only on the basis of stylistic features, but also the topics distributions in the collection. In addition, the set of authors could be redefined with a set of other interested information, as citation, journal source, and the publication year,etc., to explore topics conditioned on these sets. These extensions do not require changes to the generative model.

For more details on AT model, refer to the following papers [RZGSS04, RZCG$^+$10].

3.3 Topics over Time (TOT)

In this section, we provide the details of the *TOT* model that discovers topics dynamically. We argue that topics generally spot different patterns throughout time, as they fall and rise; split apart; merge to form new topics. The previous LDA and AT models do not consider timestamps in documents; therefore, resulting in misleading of topics occurrence within time. To collect topics' temporal information, *Topics over Time (TOT)* [WM06] is introduced, which is a simple model to integrate progressive information in extracting topics.

TOT jointly models both: word co-occurrence in a document and localization of information in estimating topics. TOT parameterizes a continuous beta distribution over time to each topic, rather than taking the Markov assumptions over state transition in time. Specifically, each discovered topic is associated with a continuous distribution over time, and it is responsible for generating two observed variables: timestamps and words.

Generally, topics pose a narrow time distribution when strong word co-occurrence pattern are observed within time intervals, and have a broad time distribution when frequent of words pattern remains consistent across long time span [WM06]. Therefore, this continues beta distribution over time span produces various fluctuations, and shapes of rising and falling of topics over passage of time, and provides interesting results in collections of documents.

3.3.1 Generative Process

There are two ways of describing *TOT* generative process [WM06], and the one corresponds to the Gibbs sampling process of variables estimation is illustrated here. As mentioned previously, in addition to words, the timestamps are considered observed variables and associate to the latent topics. Thus, parameter estimation is driven to discover topics that simultaneously capture word co-occurrences and locality of those patterns in time.

The generative process corresponding to Figure 5 is:

1. For each document d, choose D multinomials $\theta_d \sim$ Dirichlet prior α ;

2. For each topic t, choose T multinomials $\varphi_t \sim$ Dirichlet prior β ;

3. For each word $w_{d,i}$ per document d, in the corpus:

   - choose a topic $z_i \sim$ multinomial θ_d; $(P(z_i \mid \alpha))$
   - choose a word $w_i \sim$ multinomial φ_z; $(P(w_i \mid z_i, \beta))$

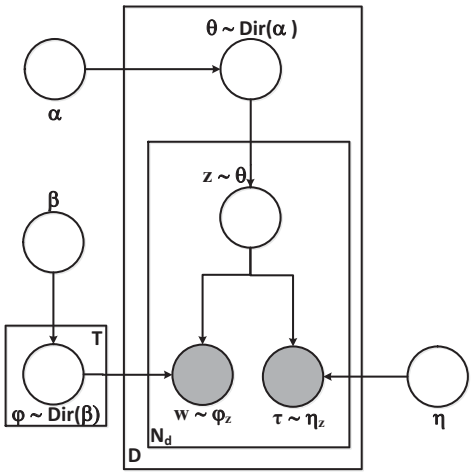


Figure 5: The graphical model representation (plate notation) of Topics over Time (TOT) model

- choose a timestamp $t_i \sim$ *Beta* η_z; $(P(t_i \mid z_i))$

Each document d is represented as a mixture of topics θ. Each topic z is a multinomial distribution over a word vocabulary φ, and z is also a beta distribution over timestamp η.

3.3.2 Inference using Gibbs sampling

Again, we employ Gibbs sampling to conduct approximate inference for TOT. Starting from the joint distribution $P(w, \tau, z, \theta, \vartheta \mid \alpha, \beta, \eta)$, the conditional probability distribution of topics over other variables for this model is derived as follows:

$$P(z_i = t \mid w_i = w, \tau, z_{-i}, \alpha, \beta, \eta) \propto (n_{d_i,t}^{D,T} + \alpha - 1) \times \frac{n_{w_i}^{V,T} + \beta - 1}{\sum_v^V n_{w_{-i},t}^{V,T} + V\beta - 1} \times \frac{(1-\tau_i)^{\eta_{t_i,1}-1} \tau_i^{\eta_{t_i,2}-1}}{B(\eta_{t_i,1}, \eta_{t_i,2})} \quad (6)$$

where $n_{w_{-i},t}^{V,T}$ is the number of words w being assigned to the topic t, not including current word i, and $n_{d_{-i},t}^{D,T}$ is the vector count of topic t being assigned to some words, not including current word i, in a document d. B is a beta function and η_t is the beta distribution for topic

t. For simplicity, η is updated after each sampling by the method of moments as follows:

$$\eta_{t,1} = \bar{\tau}_t(\frac{\tau_t(1-\bar{\tau}_t)}{s_t^2} - 1) \tag{7}$$

$$\eta_{t,2} = (1-\bar{\tau}_t)(\frac{\tau_t(1-\bar{\tau}_t)}{s_t^2} - 1) \tag{8}$$

$\bar{\tau}_t$ and s_t^2 represent the sample mean and the biased sample variance of the timestamps belonging to topic t, respectively.

From the generative process, we observe that the discovered topics are *constant*, and the time information is used to better discover topics [WBH08]. In general, TOT can predict absolute time values given an unknown timestamps of documents, by extracting topics' distribution in documents, and in other way it helps predict topics' distribution in documents given a timestamp.

TOT has been extended to perform topics and group membership over time, with a *Group-Topic (GT)* model [MWM07]. For more extensive details on TOT model, including generative process and experimental results, readers can refer to [WM06].

3.4 Summary

In this chapter, we described a simple probabilistic topic model, *Latent Dirichlet allocation (LDA)*, and its two extensions: *Author-Topic (AT)* and *Topics over Time (TOT)* models.

These exploratory models provide an automatic procedure in summarizing and extracting information about topics, the relationships between authors, and topic time-trends from large text corpora, which is hard to obtain manually. In other words, these models uncover the underlying structure of documents, by extracting the mixture of topics per document, and expose the connections between authors and topics, by extracting the mixture of authors that might be useful in predicting authors in documents. In addition, TOT model demonstrates various localization of topics over time, as evolution of topics over time.

In the subsequent chapters, we describe our algorithms and propose models that are based on these three models.

Chapter 4

Measurements and Criminal Topic (CT) model

This chapter covers the distance and evolution measurements used in language modeling. We note that the measurements are discussed in the context of our research area. Later on, we describe the *Criminal Topic (CT)* model and its usages in details.

4.1 Kullback-Leibler divergence

In 1951, Kullback and Leibler [KL51] studied the scientific meaning related to Fisher's concepts of a sufficient statistic [BA01]. Their work is now known as *Kullback-Leibler divergence (KL)* or *Relative Entropy*. It has been studied in Information Retrieval as a measurement on how different two probability distributions are [XC99, MH04].

The *KL divergence* is considered a distance measurement between the two probability densities, from a true probability distribution to a target probability distribution.

Let c (a criminal topic) be a true distribution having probability function M_c, and let a second or targeted distribution d (a chat log) have probability function M_d. Then the KL distance is defined by:

$$KL(M_c \parallel M_d) = \Sigma_{w \in V} P(w \mid M_c) \log \frac{P(w \mid M_c)}{P(w \mid M_d)} \tag{9}$$

In this thesis, we denote M_d and M_c as two language models, and they consist of distribution of words. In language models, KL is often used in clustering, as a measure of (dis)similarity of some given language models. Therefore, we employ KL to measure the dissimilarity between M_d and M_c. When using a code based on d, KL measures the expected number of additional bits required to code samples from c [ABD08]. In other words, it measures how bad the probability distribution M_d is at modeling M_c. Although it is often intuited for distance metric, KL divergence is not symmetric. Therefore, in our work, we simply compute the average of KL($M_c \parallel M_d$) and KL($M_d \parallel M_c$).

4.2 Normalized Mutual Information

In addition to KL divergence, *Normalized Mutual Information (NMI)* is applied as a distance function. Before proceeding on NMI, we present a brief detail on *Mutual Information (MI)*, which is the basis for the NMI measure. We describe both of them only in context of our approach. MI measures the contribution of the presence/absence of a term for making the correct classification decision on c. In our application, it measures the mutual dependence of t and the given c.

$$I(\Omega_t, C_c) = \Sigma_{e_w \in 0,1} \Sigma_{e_c \in 0,1} P(\Omega = e_w, C = e_c) \log_2 \frac{P(\Omega = e_w, C = e_c)}{P(\Omega = e_w) P(C = e_c)} \tag{10}$$

Ω_t is a random topic t that takes values: e_w=1 and e_w=0 (1 means the topic contains term w and 0 is not), and *C* is a random variable that takes values: e_c=1 and e_c=0 (1 means the topic t is in class c and 0 is not). $I(\Omega_t, C_c)$=0, only if a term's distribution is same in the class c and topic t. Therefore, MI is a perfect indicator for class membership of topic t. For the calculation proposes in this thesis, we apply the following:

$$I(\Omega_t, C_c) = \Sigma_t \Sigma_c \frac{\mid w_t \cap w_c \mid}{N} \log \frac{N \mid w_t \cap w_c \mid}{\mid w_t \mid\mid w_c \mid} \tag{11}$$

where w_t is the number of terms in topic t, and w_c is the number of terms in class c.

Now, the Normalized Mutual Information is defined by:

$$NMI(\Omega_t, C_c) = \frac{I(\Omega_t, C_c)}{(H(\Omega_t) + H(C_c))/2} \tag{12}$$

where $I(\Omega_t, C_c)$ referred to a mutual information between the relevant topic Ω_t and a given criminal topic c. H stands for the entropy [MRS08]:

$$H(\Omega_t) = -\Sigma_t \frac{\mid w_t \mid}{N} \log \frac{\mid w_t \mid}{N} \tag{13}$$

NMI is always a number between 0, implies two topics are independent, and 1, implies complete match, and in other way it can be assumed as reduction in uncertainty about one random variable given the knowledge of another. High mutual information shows a large reduction in uncertainty; low mutual information shows a small reduction; and zero mutual information means the variables are completely independent. We emphasize that the MI is intimately related to the KL divergence [CT91].

4.3 Evaluation Measures

In this section, we provide some of the measurements used in the information retrieval to evaluate the approach discussed in Chapters 5 and 6. The evaluation measures are *Precision*, *Recall*, and *F-Measure*.

The *Precision* of a model describes the number of the discovered chat logs d_c that are correct from overall retrieved logs that seem to be relevant; d_c is the crime-related chat log.

$$Precision = \frac{Number\ of\ the\ truth\ d_c\ is\ discovered}{Retrieved\ Documents} = \frac{t\_pos}{t\_pos + f\_pos} \tag{14}$$

The *Recall* of a model describes the number of the relevant (truth) chat logs d_c are

Table 2: Contingency table between Relevant and Retrieved values

	Relevant	Non-Relevant
Retrieved	true positives (t_pos)	false positives (f_pos)
Not-Retrieved	false negatives (f_neg)	true negatives (t_neg)

successfully discovered by the model.

$$Recall = \frac{Number\ of\ the\ truth\ d_c\ are\ discovered}{Number\ of\ d_c\ are\ correct} = \frac{t\_pos}{t\_pos + f\_neg} \tag{15}$$

Both, precision and recall, measures can be made clear using the notation in Table 2. True positives refer to the relevant chat logs that are correctly retrieved, while true negatives are the non-relevant chat logs that are not retrieved. False positives are the non-relevant chat logs that are retrieved, while false negatives are the relevant chat logs that are not retrieved.

The *F-Measure* computes the weighted harmonic mean of precision and recall for a model.

$$F_\pi = (\pi^2 + 1) \cdot \frac{Precision \cdot Recall}{\pi^2 \cdot Precision + Recall} \tag{16}$$

where $\pi \in [0, \infty]$. In this thesis, we use $\pi = 2$, which weighs recall higher than precision, and $\pi = 1$, which gives an equal weight for both measures, recall and precision.

4.4 Criminal Topic (CT) model

n-grams are the most commonly used natural language model [Cha94]. It is a probabilistic model that takes the assumption that only the previous *n*-1 words, in a sequence, have any effect on the probabilities for the next word. In other word, the probability of a current word depends on the previous *n*-words.

A *n*-gram model of size 1 is called a *Unigram* model. Figure 6 represents the Unigram model in plate notation. In this model, the words for each document are drawn from a single multinomial distribution, independently. This can be represented by:

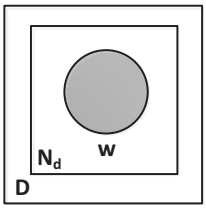


Figure 6: The graphical model representation (plate notation) of Unigram model

$$P(w) = \prod_{i=1}^{N} p(w_i) \tag{17}$$

If we extend the unigram model by adding a discrete random topic $(c = z)$, the *mixture of unigrams* model is obtained [NMTM00, BNJ03]. In this model, each document is generated by, first picking a random topic c, and then generating N words, independently, from the conditional multinomial $p(w|c)$. Therefore, the probability of a document is [BNJ03]:

$$P(w) = \sum_{c} p(c) \prod_{i=1}^{N} p(w_i|c) \tag{18}$$

In this thesis, we apply the mixture of unigrams model to explore a chat log and its relation to criminal activities. Throughout this thesis we use *Criminal Topic (CT)* to refer to the mixture of unigrams model. Under this model, a single topic c generates N words. We assume that the topic c and the words w are observable in CT model. The key point of developing this model is that the assumption for any detained chat logs, it might exhibit several criminal topics, and each of these topics is composed of its own distribution of words. Therefore, comparing the topics distributions in d with c indicates the relevance of d to crime.

The words are drawn from a single topic distribution:

$$P(w|c, \varphi) = \prod_{n=1}^{N} \varphi_{w_n,c} \tag{19}$$

where φ is the distribution of words under c. It describes the probability of each word w

conditioned on c. φ_c is calculated as follows:

$$\varphi_c = \frac{n_{w_i,c}^{W_c,C} + \beta}{\sum_n^{W_c} n_{w_i,c}^{W_c,C} + W_c\beta}, \tag{20}$$

where W_c is the number of words in criminal topic c, and $(C = T)$ in the Figure 7. Using CT model, KL is applied to estimate the distance between d and c in order to distinguish crime-related logs from others. In addition, it is also applied to compute the distance between discovered crime-related topics and c after the two extended models, LDA-TOT and A-TOT (described in Chapters 5 and 6), have generated topics from d.

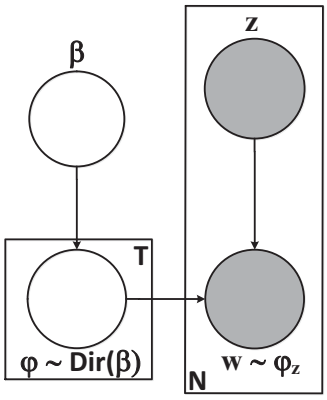


Figure 7: The graphical model representation (plate notation) of Criminal Topic (CT) model

4.5 Summary

In this chapter, we introduced the nonsymmetric distance measurement *Kullback-Leibler divergence*, and Symmetric *Normalized Mutual Information*, and the evaluation measures for our approach that is discussed in the next two chapters. We also presented a *Criminal Topic* model consisting of a single topic that generates distribution of words. For references to the comprehensive literature on the subjects discussed in this chapter, readers may refer to [MRS08, HK06, Cha94, KKK80].

Chapter 5

LDA-Topics over Time (LDA-TOT)

Probabilistic topic models, such as LDA, AT, and TOT described in Chapter 3, model the hidden semantic structure of a document collection without pre-specifying whether a document contains a specific topic or not. In this chapter, we will introduce a methodology of assigning a particular document or chat log to crime-related, based on some distance measurements. We begin by outlining the new extended model from LDA, describing the algorithm in details, and then analyzing the results in an extensive manner.

5.1 Overview

LDA is usually performed on large size documents, and it is inappropriate for small size documents, such as chat logs. Moreover, chat logs classification through LDA is error-prone. A chat log that is biology relevant might be misclassified to be crime-relevant. We define two main reasons for misclassification. First, LDA is based on the "bag of words" assumption, which treats a document as a frequent of words count; therefore, the weight of the words depends on the number of words occurrences in a collection. Second, the size of chat logs are small comparing to the traditional documents, such as scientific articles; thus, it is hard to obtain mixture of topics in chat logs. As a result, preprocessing is required

before employing language modeling techniques. In particularly, we propose an algorithm that determines crime-related logs through measuring the difference between a chat log d and a provided criminal topic c. The key point is to compute the probability of a language model M_c generating the document d and to determine the topics in interest given d.

Topic discovery is influenced not only by the occurrence of words and their frequencies, but also by the timestamp associated with each word in a chat log. The transition of topics over time in a given chat log can be estimated, by introducing an observable variable t into the standard LDA model. Various models have been proposed to illustrate the transition of topics over time, such as the *TOT* model [WM06]. Nonetheless, we depict *LDA-Topics over Time (LDA-TOT)* model, in Figure 8, that identifies topics and their evolution over time.

The primary difference between this model and TOT is the use of discrete intervals of time instead of continuous time, as in TOT (see Chapter 3). Time intervals in a chat log are relatively short, ranging from a few minutes to a few hours. Therefore, we employ discrete time intervals in this model. Moreover, it is easy to include discretization of time, for the learning and computation purposes, in order to generate the topics' distribution over time (η), rather then using continuous beta distribution.

5.2 Generative Process for LDA-TOT

The generative process, in Figure 8, for LDA-TOT uses Gibbs sampling, for estimating the parameters, and it is as follows:

1. For each document d, choose D multinomials $\theta_d \sim$ Dirichlet prior α ;
2. For each topic t, choose T multinomials $\varphi_t \sim$ Dirichlet prior β ;
3. For each word w_{di} in each document d, in the corpus:

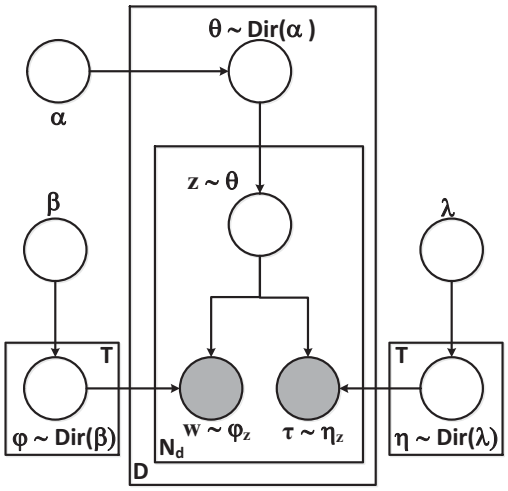


Figure 8: The graphical model representation (plate notation) of LDA-Topics over Time (LDA-TOT) model

- choose a topic $z_i \sim$ multinomial θ_d; $(P(z_i \mid x_i, \alpha))$
- choose a word $w_i \sim$ multinomial φ_z; $(P(w_i \mid z_i, \beta))$
- choose a timeinterval $\tau_i \sim$ multinomial η_z; $(P(\tau_i \mid z_i, \lambda))$

In general, LDA-TOT model starts by picking a multinomial distribution θ_d, from the Dirichlet distribution α, that determines which topics are most likely to appear in a chat log d. Next, the model chooses a single topic $z_i = t$ and assigns the ith word (w_i) in the chat log to $z_i = t$, based on the multinomial distribution θ_d. To generate a word, the model picks a word w_i, from the vocabulary of V words, according to the multinomial distribution φ_z, which is generated from the Dirichlet distribution β for each topic t, and assigns a time stamp τ_i to w_i from η_z. The η_z defined in this model is a multinomial distribution for each word token w_i over time stamp τ_i, under a topic $z = t$. From the procedure, we notice that each word in a chat log is generated by different topics at random.

5.3 Inference using Gibbs sampling

The posterior distribution of topics in LDA-TOT depends on both word and time. The Gibbs sampling algorithm, as done for the other models in Chapter 3, reduces the parameter estimation problem to a simple counting and sampling process. We begin deriving with the joint distribution $P(w, \tau, z|\lambda, \alpha, \beta)$.

$$
\begin{aligned}
P(w,\tau,z|\lambda,\alpha,\beta) &= P(w|z,\beta)P(\tau|z,\lambda)P(z|\alpha) \\
&= \int P(w,\varphi|z,\beta)\,d\varphi \int P(\tau,\eta|z,\lambda)\,d\eta \int P(z,\theta|\alpha)\,d\theta \\
&= \int P(w|z,\varphi)P(\varphi|\beta)\,d\varphi \int P(\tau|z,\eta)P(\eta|\lambda)\,d\eta \int P(z|\theta)P(\theta|\alpha)\,d\theta \\
&= \int \prod_{t=1}^{T}\left(\prod_{v=1}^{V}\varphi_{v,t}^{n_{v,t}}\right)\frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\left(\prod_{v=1}^{V}\varphi_{v,t}^{\beta_v-1}\right)d\varphi \\
&\quad \times \int \prod_{t=1}^{T}\left(\prod_{v=1}^{V}\varphi_{v,t}^{n_{v,t}}\right)\frac{\Gamma(\sum_{v=1}^{V}\lambda_v)}{\prod_{v=1}^{V}\Gamma(\lambda_v)}\left(\prod_{v=1}^{V}\eta_{v,t}^{\lambda_v-1}\right)d\eta \\
&\quad \times \int \prod_{d=1}^{D}\left(\prod_{t=1}^{T}\theta_{d,t}^{n_{d,t}}\right)\frac{\Gamma(\sum_{t=1}^{T}\alpha_t)}{\prod_{t=1}^{T}\Gamma(\alpha_t)}\left(\prod_{t=1}^{T}\theta_{d,t}^{\alpha_t-1}\right)d\theta \\
&= \int \prod_{t=1}^{T}\frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\prod_{v=1}^{V}\varphi_{v,t}^{n_{v,t}+\beta_v-1}\,d\varphi \\
&\quad \times \int \prod_{t=1}^{T}\frac{\Gamma(\sum_{v=1}^{V}\lambda_v)}{\prod_{v=1}^{V}\Gamma(\lambda_v)}\prod_{v=1}^{V}\eta_{v,t}^{n_{v,t}+\lambda_v-1}\,d\eta \\
&\quad \times \int \prod_{d=1}^{D}\frac{\Gamma(\sum_{t=1}^{T}\alpha_t)}{\prod_{t=1}^{T}\Gamma(\alpha_t)}\prod_{t=1}^{T}\theta_{d,t}^{n_{d,t}+\alpha_t-1}\,d\theta \\
&= \prod_{t=1}^{T}\frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\int \prod_{v=1}^{V}\varphi_{v,t}^{n_{v,t}+\beta_v-1}\,d\varphi \\
&\quad \times \prod_{t=1}^{T}\frac{\Gamma(\sum_{v=1}^{V}\lambda_v)}{\prod_{v=1}^{V}\Gamma(\lambda_v)}\int \prod_{v=1}^{V}\eta_{v,t}^{n_{v,t}+\lambda_v-1}\,d\eta \\
&\quad \times \prod_{d=1}^{D}\frac{\Gamma(\sum_{t=1}^{T}\alpha_t)}{\prod_{t=1}^{T}\Gamma(\alpha_t)}\int \prod_{t=1}^{T}\theta_{d,t}^{n_{d,t}+\alpha_t-1}\,d\theta
\end{aligned}
$$

$$
\begin{aligned}
P(w,\tau,z|\lambda,\alpha,\beta) &= \prod_{t=1}^{T}\frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\frac{\prod_{v=1}^{V}\Gamma(\beta_v+n_{v,t})}{\Gamma(\sum_{v=1}^{V}\beta_v+n_{v,t})} \\
&\times\prod_{t=1}^{T}\frac{\Gamma(\sum_{v=1}^{V}\lambda_v)}{\prod_{v=1}^{V}\Gamma(\lambda_v)}\frac{\prod_{v=1}^{V}\Gamma(\lambda_v+n_{v,t})}{\Gamma(\sum_{v=1}^{V}\lambda_v+n_{v,t})} \\
&\times\prod_{d=1}^{D}\frac{\Gamma(\sum_{t=1}^{T}\alpha_t)}{\prod_{t=1}^{T}\Gamma(\alpha_t)}\frac{\prod_{t=1}^{T}\Gamma(\alpha_t+n_{d,t})}{\Gamma(\sum_{t=1}^{T}\alpha_t+n_{d,t})} \qquad (21)
\end{aligned}
$$

Using the chain rule, we obtain the conditional distribution of topics:

$$
\begin{aligned}
P(z_i=t|w_i=w,z_{-i},w_{-i},\tau_{-i},\lambda,\alpha,\beta) &= \frac{P(w,z,\tau|\lambda,\alpha,\beta)}{P(w,z_{-i},\tau|\lambda,\alpha,\beta)} \\
&= \frac{P(w|z,\beta)P(z|\alpha)P(\tau|z,\lambda)}{P(z_{-i},w_{-i},\tau_{-i}|\lambda,\alpha,\beta)P(w_i|\lambda,\alpha,\beta)} \\
&\times\frac{1}{P(\tau_i|\lambda,\alpha,\beta)} \\
&\propto \frac{P(w|z,\beta)P(z|\alpha)P(\tau|z,\lambda)}{P(w_{-i}|z_{-i},\beta)P(z_{-i}|\alpha)P(\tau_{-i}|z_{-i},\lambda)} \\
&\propto \frac{n_{w_i}^{V,T}+\beta}{\sum_v^V n_{w_{-i},t}^{V,T}+V\beta}\times\frac{n_{d_i,t}^{D,T}+\alpha}{\sum_t^T n_{d_{-i},t}^{D,T}+T\alpha} \\
&\times\frac{n_{\tau_i}^{V,T}+\lambda}{\sum_v^V n_{\tau_{-i},t}^{V,T}+V\lambda} \qquad (22)
\end{aligned}
$$

where $n_{w_{-i},t}^{V,T}$ is the vector counts of the word *w* being assigned to the topic t, not including current word i. $n_{d_{-i},t}^{D,T}$ is the vector count of topic t being assigned to some words, not including the current word i, in a document d. $n_{\tau_{-i},t}^{V,T}$ is the vector counts of the word *w* being assigned to the topic t under timeinterval τ, not including current word i. The equation 22 is the conditional probability derived by marginalizing out the random variables, θ, φ, and η.

5.4 Mining for crime-relevant chat logs, topics, and topics over time using LDA-TOT

In this section, we provide an algorithm to classify crime-related chat logs and to extract the underlying crime-related topics in these logs. We emphasize that this algorithm searches for a particular criminal topic in a chat log. An overview of the algorithm is shown in Algorithm 1. The process starts by employing the CT model to estimate φ for a single topic c (Line 2). This is the learning process for the CT model. It is common for a criminal topic c to contain a set of words W_c that might not be included in the pre-existing vocabulary set $\bar{V}$. Therefore, we combine the words W_c in a criminal topic with the existing $\bar{V}$ words (Line 3). Next, the distance between a chat log d and the criminal topic c is calculated using KL divergence (Line 4), under the same vocabulary used for both c and d. The results obtained from KL might or might not pass the user-specified threshold ϵ. In case the distance measurement KL is lower than or equal to ϵ (Line 5), the algorithm proceeds to the subsequent steps (Line 6-13); otherwise it terminates (Line 14). Then, LDA-TOT is applied to extract crime-relevant topics in a chat log, where all words in d are randomly assigned to topics (Line 6). The iteration process starts by executing Gibbs sampling and computing KL distance between each topic t and the provided c (Line 8-11). Finally, the algorithm terminates when a topic t (Line 12) satisfies the threshold γ. The outputs are the three distributions (θ, φ, η); these are further analyzed in the next experimental section, using other evaluation measures to evaluate the performance of the proposed procedure.

5.5 Experiments

In this section, we perform an empirical study on the first research question, presented in Chapter 1, and provide the results with extensive details. We emphasize that the thesis is concerned on the two research questions only, in Chapter 1; therefore, the comparison of

Algorithm 1 Mining for crime-relevant chat logs, topics, and topics over time using LDA-TOT

Input: $\alpha, \beta, \lambda, \epsilon, \gamma$
φ=Calculate criminal topic-word distribution($D, \alpha, \beta, \lambda$)
$V = \bar{V} \bigcup W$
Δ=KL(d_i, c)
if $\Delta \leq \epsilon$ **then**
    Initialize randomly for all words w_i^N in a chat log d to topics z_t^T
    repeat
        [$\theta_d, \varphi, \eta, z_t$] = GibbsSampling($d, \tau_d, \alpha, \beta, \lambda$)
        for t=1 to T **do**
            σ_t^T=KL($\theta_{d,t}^T, c$)
        end for
        L = GetLowest(σ_t^T)
    until $L \leq \gamma$
end if

the proposed model with other existing models is not addressed in this thesis.

5.5.1 Datasets

The chat logs used in the experiment are obtained from a website called *perverted-justice.com* and *IRC* logs.

Perverted-Justice. This dataset consists of chat logs from various instant messages, e.g., Yahoo! and AOL, containing information about adults who seek online sexual conversations with others who are posing as children or underage teenagers (pseudo-victims). It contains over 546 log files, as of July 11, 2011, and over 1000 authors. For simplicity, we use only the time intervals associated with messages in these chat logs, without considering the date.

IRC. This dataset is collected from various IRC channels by running a mIRC application for about 10 days. The dataset contains 160 authors and 50 log files with a total of 4086 word tokens. There are 5 categories classified in multiple topics. Each message in the chat logs has a timestamp that is determined by the date and time intervals. As in the

Table 3: Summary of the datasets used in this paper

Dataset	Documents	Words	Unique Words
Perverted-Justice	250	27866	1455
IRC	50	4086	276

previous dataset, we use only the time intervals and ignore the date.

For both datasets, we first remove all the links from the messages, stop words, numbers, and non-English letters. The words are downcased and stemmed to their root source, using porter stemmer. However, words that rarely appear in a chat log are not removed, because the chat log differs from the structured documents and the words might be of value to the results. We furthermore prune the corpus, by including only the log files with more than 500 words, and we use these for further processing. The results from the pre-processing step for the both datasets consist of 300 logs, with 670 authors, and a total of 31952 word tokens. Some statistics of the two datasets after pre-processing are summarized in Table 3.

As for the other settings, we do not estimate the hyper parameters α, β, and λ; instead, they are fixed at α=1, β=0.01, and λ=0.01, respectively. The number of topics T is also fixed at T=5 for both models. Two sets of c are used, one contains 30 words and other 50 words, to capture the characteristics of the discovered topics and their transitions.

We train the CT model with 200 chat logs, from *perverted-justice*, to compute the φ distribution of topic c, where c is only sex related, and keep 100 logs for testing the outcomes from LDA-TOT. Note, c could be any criminal topic. The chat logs are renamed to $d_1, d_2, d_3, \ldots$ and authors are renamed to $a_1, a_2, a_3, \ldots$, instead of using their true names due to privacy concerns. The experiments are executed on a PC running Windows 7 (32-bit) with Intel 2.13GHz (2 CPUs) and 2GB memory. We run the application several times at a fixed number of 2000 iterations, and we record the outcomes each time in terms of KL$(t_{i,d}, c)$, NMI$(t_{i,d}, c)$, and θ_d.

Table 4: KL divergence between documents (d_1, d_2, d_3, d_4) and c when $\mid c \mid$=30 and $\mid c \mid$=50 using LDA-TOT and A-TOT

Documents	Size	KL(d_i, c)
d_1	$\mid c \mid$=30	0.7162
	$\mid c \mid$=50	0.5906
d_2	$\mid c \mid$=30	1.1261
	$\mid c \mid$=50	0.9770
d_3	$\mid c \mid$=30	0.8526
	$\mid c \mid$=50	0.6327
d_4	$\mid c \mid$=30	0.6953
	$\mid c \mid$=50	0.5361

5.5.2 Case study

In this subsection, we evaluate the effectiveness of our algorithm, by performing an in-depth case study, on the two aforementioned datasets, to answer the first research question. The two research questions elaborate on problem 6, in Chapter 1. The resulting distributions (θ, φ, η) from LDA-TOT model are further analyzed to capture various characteristics of topics and their evolutions over time. Before moving forward, the following Table 6 contains some sexual terminologies, which is inappropriate to some readers of this thesis; therefore, readers' discretion is advised.

Q 1. *How can an investigator determine which logs are crime-relevant? In identifying a crime-relevant log, what are the contributed topics in the log file? How have they evolved over time? Moreover, how can an investigator extract the topics that are crime-related from the identified crime-relevant log files?*

To answer this question, we apply the mining algorithm, using LDA-TOT to extract the crime-related topics, one chat log at a time. We select several logs randomly and record the similarities among these logs. Next, we adopt two expected cases, based on the results from KL(d, c) : 1- KL$(d, c) \leq \epsilon$ when d is crime-relevant. 2- KL$(d, c) > \epsilon$ when d is not

Table 5: KL divergence and NMI between crime-related topics from documents (d_1, d_2, d_3, d_4) and c when $\mid c \mid$=30 and $\mid c \mid$=50 using LDA-TOT

Documents	Size	KL$(t_{i,d}, c)$	NMI$(t_{i,d}, c)$
d_1	$t_2(\mid c \mid$=30)	1.2250	0.2189
	$t_4(\mid c \mid$=50)	1.2228	0.3700
d_2	$t_4(\mid c \mid$=30)	0.9608	0.0977
	$t_0(\mid c \mid$=50)	1.0684	0.0600
d_3	$t_0(\mid c \mid$=30)	1.1768	0.1156
	$t_3(\mid c \mid$=50)	0.6214	0.2429
d_4	$t_0(\mid c \mid$=30)	1.2460	0.2181
	$t_1(\mid c \mid$=50)	1.0986	0.6761

crime-relevant. We set the users' threshold ϵ to 1.25 and γ to 0.86. Below is the description of the two cases on 4 selected chat logs:

Case 1 (KL$(d, c) \leq \epsilon$)**:** From Definition 3, the document d is crime-related under this case. Based on the results from KL between d and c, as shown in Table 4, it is clear that 3 chat logs $\{d_1, d_3, d_4\}$ follow this case, and they are related to crime. We remind the reader again that topic c is sex related. LDA-TOT generates 5 topics from each of these 3 chat logs, and the crime-relevant topics are shown in Table 6.

Not surprisingly, the top 10 relevant words, with high probabilities, provide sufficient information to classify these topics as crime-related, and the measurements from KL and NMI support our prospects as well. The θ_d^t distributions (between the round brackets) for these topics are above 0.2, which represents about one-fifth of the logs. This computation is far more essential, because it distinguishes the crime-related chat logs from others, and the importance of θ_d is well demonstrated in case 2.

By observing KL(d, c) and KL$(t_{i,d}, c)$, from Tables 4 and 6, we notice that the results are not always monotonic. For example, KL(d_1, c)=0.7162 and KL(d_4, c)=0.6953 when the size of $\mid c \mid$=30. However, KL(t_{2,d_1}, c)=1.2250 is more relevant to c than KL(t_{0,d_4}, c)=1.2460. In addition, NMI seems to behave the same for both of these topics t_{2,d_1}=0.2189 and

Table 6: Top 10 relevant words extracted for crime-related topics from documents (d_1, d_2, d_3, d_4) and their distribution over documents using LDA-TOT

d_1				**d_2**			
|c| = 30		**|c| = 50**		**|c|=30**		**|c| = 50**	
t_2 (0.2149)	**Prob**	**t_4 (0.2209)**	**Prob**	**t_4 (0.0632)**	**Prob**	**t_0 (0.1944)**	**Prob**
girl	0.0399	girl	0.0389	spend	0.0109	wed	0.0302
sex	0.0099	fuck	0.0197	maevlyn	0.0109	minut	0.0151
pretti	0.0092	happi	0.0197	love	0.0037	thing	0.0121
kiss	0.0089	pic	0.0155	blow	0.0037	actual	0.0091
vid	0.0081	suck	0.0125	puta	0.0037	anniversari	0.0091
pussi	0.0070	bed	0.0079	misskriss	0.0037	morn	0.0091
young	0.0055	sleepi	0.0049	yesterdai	0.0037	julianu	0.0061
cute	0.0048	dick	0.0049	pleas	0.0037	that	0.0061
bodi	0.0044	bodi	0.0044	biggi	0.0037	state	0.0061
virgin	0.0025	sexi	0.0039	develop	0.0037	grei	0.0061

d_3				**d_4**			
|c| = 30		**|c| = 50**		**|c|=30**		**|c| = 50**	
t_0 (0.2353)	**Prob**	**t_3 (0.2010)**	**Prob**	**t_0 (0.2174)**	**Prob**	**t_1 (0.1744)**	**Prob**
luv	0.0443	love	0.0212	girl	0.0237	nite	0.0258
miss	0.0365	look	0.0114	nite	0.0213	nude	0.0199
peni	0.0348	peni	0.0110	nude	0.0164	sweet	0.0187
kiss	0.0155	luv	0.0098	pussi	0.0155	pussi	0.0187
stuff	0.0148	suck	0.0094	sweet	0.0155	babi	0.0146
suck	0.0128	bed	0.0063	bf	0.0073	gf	0.0141
touch	0.0125	kiss	0.0063	sex	0.0068	kiss	0.0117
sleep	0.0104	feel	0.0063	ass	0.0068	bodi	0.0105
pretti	0.0095	nake	0.0051	nake	0.0063	figur	0.0105
big	0.0091	butt	0.0035	leg	0.0044	sexxi	0.0094

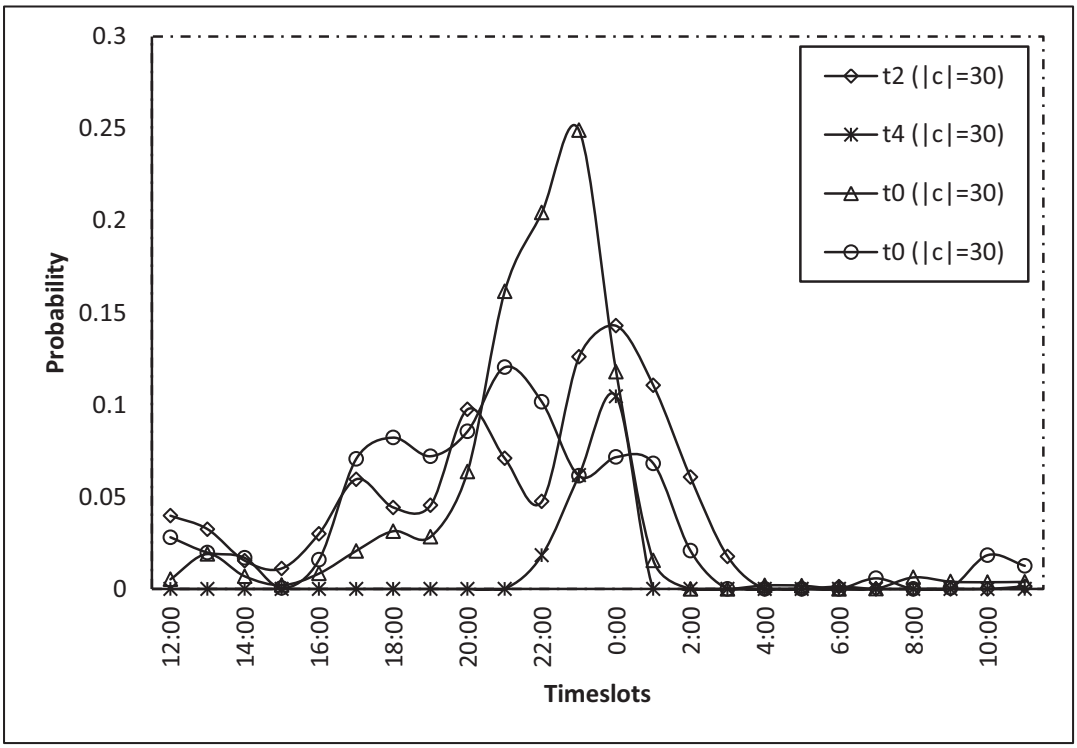


Figure 9: Evolution of crime-related topics using LDA-TOT when $\mid c \mid$=30

t_{0,d_4}=0.2181.

Probabilistic topic models, such as LDA-TOT, are based on the concept of generating topics randomly; each time it extracts topics with different probability distributions. Therefore, the results obtained from KL and NMI between discovered topics and c are not necessarily monotonic. Nevertheless, the algorithm discloses crime-related chat logs, if they exist in a collection of data texts.

Furthermore, KL(t_{0,d_4}, c)=1.2460 and the NMI of topic t_{0,d_4} should obtain better results. This is because KL(d_4, c)=0.6953 clearly indicates that d_4 is more proximate to be classified as a crime-related log than the other chat logs shown in Table 4. After 4000 iterations, we found KL(t_{2,d_4}, c)=0.9334 and NMI of this topic is t_{2,d_4}=0.4672.

Case 2 (KL$(d, c) > \epsilon$)**:** This case occurs whenever a chat log does not satisfy the user's threshold ϵ. From Table 4, the only chat log that falls under this category is d_2. Obviously,

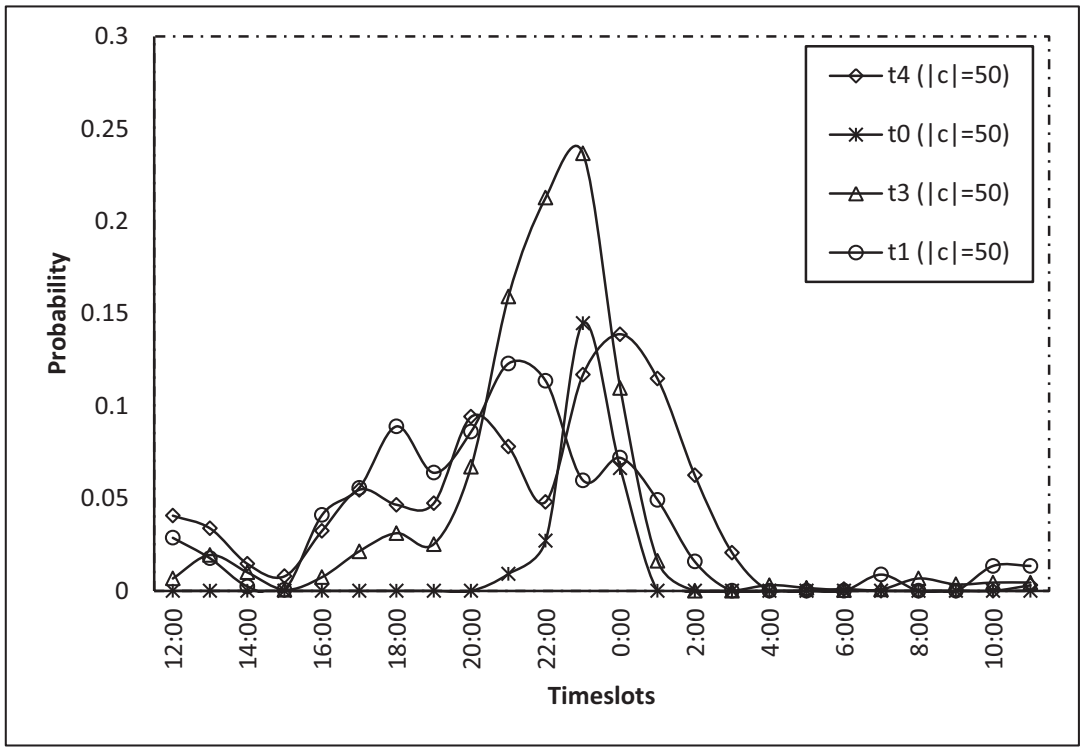


Figure 10: Evolution of crime-related topics using LDA-TOT when $\mid c \mid$=50

t_{4,d_2} (c=30) for this chat log does not contain the expected words to be classified as crime-relevant.

We observe an interesting result from KL(t_{4,d_2}, c), and it satisfies the user's threshold γ. In general, KL measures the distance between the two models (t and c). This is achieved by comparing the probability of the shared words in both topics c and t_{4,d_2}. We do not consider fixed vocabulary in the comparison, rather we depend on the mutual words. Suppose the unique words for both $\mid c \mid$=30 and $\mid t_1 \mid$=500. If the two models have joint words, with similar probability, then the KL distance for both models is similar. Consequently, the result from KL(t_{4,d_2}, c) fits with the threshold γ.

The $\theta_d^{t_1}$ distribution shows that approximately 0.0632 of d_2 is about criminal subjects. In conclusion, d_2 is not crime-relevant, based on the results from KL(d_2, c) and $\theta_{d_2}^{t_1}$.

One might ask whether the condition KL$(t_{i,d}, c)$=0 applies for the both cases. This

might occur, but it does not necessarily mean that a topic is crime-related and case 2 sheds some light on it. A topic t is considered to be crime-related whenever the two conditions hold: KL$(d, c) \leq \epsilon$ and KL$(t_{i,d}, c) \leq \gamma$.

When we alter the size of c by increasing the number of criminal terms to 50, the results from KL$(t_{i,d}, c)$ and NMI are improved, as observed in Table 5. The top 10 words in Table 6 include new crime-relevant terms that were not observed when $\mid c \mid$=30. This is not a coincidence, since the words used in c are drawn from the two datasets. In general, increasing the size of c gives better predictions about the distance between discovered topics and c.

In addition to topics extraction, the LDA-TOT is able to predict the time associated with each message in a chat log. Figures 9 and 10 include the fluctuations of relevant topics from 4 chat logs when $\mid c \mid = 30$ and $\mid c \mid = 50$. The characteristics of the transitions can be classified through the transition function $F(t)_{\tau^s}^{\tau^f}$, as *active* and *not-active*. In many cases, topics' activity is provided by investigators to assist them in analyzing different rise and falls of topics. Therefore, we define the transition function as:

$$F(t)_{\tau^s}^{\tau^f} = \begin{cases} active & \text{if } \sum_{\tau^s}^{\tau^f} p(t)^{\tau^s} \geq \text{users' threshold} \\ not-active & \text{if } \sum_{\tau^s}^{\tau^f} p(t)^{\tau^s} < \text{users' threshold} \end{cases}$$

$\sum_{\tau^s}^{\tau^f} p(t)^{\tau^s}$ sums the probability of a topic t during interval $[\tau_s, \tau_f]$. $F(t)_{\tau^s}^{\tau^f}$ indicates the activity of t. We found the best results are obtained when an average of θ_d over the three highest topics is considered for estimating the users' threshold. For instance, when setting the users' threshold to 0.2143, as an average of θ_{d_1} over 3 topics, the topic t_{4,d_1} ($\mid c \mid$=50) is active during [22:00, 1:00] and not active elsewhere.

In general, the topics $t_{4,d_1}, t_{0,d_2}, t_{3,d_3}, t_{1,d_4}$ ($\mid c \mid$=50) are widely active during time intervals [15:00,3:30] when $p(t)_{\tau^s}^{\tau^f} \geq$0.2143, with a peak on [21:00,1:00]. Investigators collect information, within certain intervals, that indicate the activity of crime-related topics, and

thus provide the start point for the investigations process.

We conclude that a crime-relevant chat log d can be recognized through KL(d, c), and the crime-related topics are determined by three factors: θ_d^t, KL(d, c), and KL$(t_{i,d}, c)$. The characteristics of the relevant topics are studied through NMI, whereas high probability means obtaining a better quality of discovered topics. In addition, the evolution of topics is demonstrated through the transition function $\digamma(t)_{\tau^s}^{\tau^f}$, in terms of active or not active, in the given time intervals associated with each message in logs.

5.6 Summary

Given a corpus with large collection of chat logs, it is trivial to segregate these logs manually to crime-relevant or not. Furthermore, if the crime-related logs are identified, the underline hidden structure of these logs are required to expose. Therefore, in this chapter, we proposed an algorithm to accomplish these tasks, automatically. We described *LDA-TOT* model that combines LDA and TOT models. This model not only discovers topics, but also detects the flaws of these topics over discrete time intervals.

Using this model, we studied the first research question on two datasets. Experimental results highlighted on two cases, depending on how logs are criminally associated. Through our approach, investigators now can distinguish logs, discover related topics, unveil the distribution of words in logs, and track the progress of these topics over timeslots.

One interesting subject, that remains to discuss, is to identify authors in topics. LDA-TOT model cannot achieve authors' distributions over topics, and this will be our major concern in the next Chapter 6.

Chapter 6

Author-Topics over Time (A-TOT)

Chapter 5 discussed on how to identify crime-relevant chat logs, and spent much time to explain topics discovery and their evolutions over time. The purpose of cybercrime investigators is to extract related topics and to identify the plausible author(s) in logs. In this chapter, we concentrate on authors' contribution in topics: how do the authors flow with topics over time and the plausible authors over passage of time? We discuss empirically the interpretation of the results, based on the newly proposed model.

6.1 Overview

The primary purpose from detaining criminal chat logs is to explore authors within discovered topics in these logs. Unfortunately, the model dictated in Chapter 5 does not provide this interesting part. Furthermore, joint author-topics over time modeling has received little or no attention as far as we are aware. In order to address the authors' distribution over topics, *A-TOT* is developed. Our motivation is to model authors-topics and detect the authors' movement throughout the time intervals, which assists investigators to trace their activities within topics in logs.

A-TOT model can be thought as an extension combined from both models, AT and TOT.

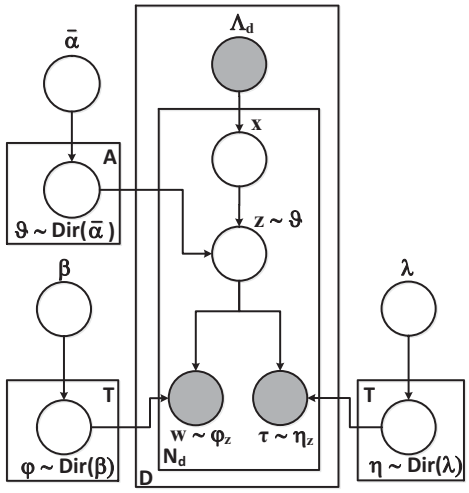


Figure 11: The graphical model representation (plate notation) of *Author-Topics over Time (A-TOT)* model

The aim of this unsupervised learning model is to achieve topics extraction, authors-topics distribution, and authors-topics distribution over time.

There are two graphical representations of A-TOT model. In first model, the procedure starts by choosing an author a timestamp, and then picks a topic where terms are formulated from this topic. In this model, a single timestamp is assigned to a topic. As for the other alternative, an author first chooses a topic, and then the topic generates terms associated to it and draws timestamps to each of these terms. This one corresponds to the Gibbs sampling in generative procedure, and we elaborate it in the next section. Figure 11 illustrates this model in plate notation. We note that the algorithm for searching a crime-related chat logs is similar to the previous one outlined in Chapter 5, except A-TOT is used instead of LDA-TOT.

6.2 Generative Process for A-TOT

The generative process for A-TOT, that corresponds to the Gibbs sampling for estimating the parameters, is as follows:

1. For each author a, choose A multinomials $\vartheta_a \sim$ Dirichlet prior $\bar{\alpha}$;
2. For each topic t, choose T multinomials $\varphi_t \sim$ Dirichlet prior β ;
3. For each word w_{di} in each document d, in the corpus:
   - choose an author $x_i \sim$ uniform Λ_d; $(P(x_i \mid \Lambda_d))$
   - choose a topic $z_i \sim$ multinomial ϑ_a; $(P(z_i \mid x_i, \bar{\alpha}))$
   - choose a word $w_i \sim$ multinomial φ_z; $(P(w_i \mid z_i, \beta))$
   - choose a timeinterval $\tau_i \sim$ multinomial η_z; $(P(\tau_i \mid z_i, \lambda))$

Formally, the set of authors Λ_d in a chat log d is observed. The procedure begins by choosing an author x, randomly at uniform, from the set of authors Λ_d. Afterward, the multinomial distribution ϑ_a, from the Dirichlet distribution $\bar{\alpha}$, is picked, and this distribution determines which topics are most likely to be assigned to the author x in a chat log d. Next, a single topic $z_i = t$ is sampled for each ith word (w_i) in d, from the multinomial distribution ϑ_a associated with the author x for that word. In general, we assume the ith word (w_i) in d is written by x for the topic $z_i = t$. Finally, in order to generate a word, the model chooses a word w_i, from the vocabulary of V words, based on the multinomial distribution φ_z, and assigns a single timestamp τ_i from η_z to w_i. φ_z is generated from the Dirichlet distribution β for each topic t.

From the procedure, A-TOT depends on both word and time for generating topics. A topic in this model is sampled from the distribution of topics specific to author x, and the

words are sampled from the distribution of words over topics. The distribution of words over topics $\Sigma_{v=1}^{V}\varphi_{v,t}$=1 is the same for both models, LDA-TOT and A-TOT. As for A-TOT, the distribution of topics over authors $\Sigma_{t=1}^{T}\vartheta_{a,t}$=1. Like LDA-TOT, η_z is a multinomial distribution for each word token w_i over time stamp τ_i, under a topic z.

6.3 Inference using Gibbs sampling

We begin deriving with the joint distribution $P(w, \tau, x, z|A, \lambda, \alpha, \beta)$.

$$\begin{aligned} P(w,\tau,x,z|A,\lambda,\alpha,\beta) &= P(w|z,\beta)P(\tau|z,\lambda)P(z|\alpha)P(x|A) \\ &= \prod_{t=1}^{T}\frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\frac{\prod_{v=1}^{V}\Gamma(\beta_v+n_{v,t})}{\Gamma(\sum_{v=1}^{V}\beta_v+n_{v,t})} \\ &\quad\times\prod_{t=1}^{T}\frac{\Gamma(\sum_{v=1}^{V}\lambda_v)}{\prod_{v=1}^{V}\Gamma(\lambda_v)}\frac{\prod_{v=1}^{V}\Gamma(\lambda_v+n_{v,t})}{\Gamma(\sum_{v=1}^{V}\lambda_v+n_{v,t})} \\ &\quad\times\prod_{d=1}^{D}\frac{\Gamma(\sum_{t=1}^{T}\alpha_t)}{\prod_{t=1}^{T}\Gamma(\alpha_t)}\frac{\prod_{t=1}^{T}\Gamma(\alpha_t+n_{d,t})}{\Gamma(\sum_{t=1}^{T}\alpha_t+n_{d,t})}\times\prod_{d=1}^{D}\frac{1}{A_d^{N_d}} \end{aligned} \quad (23)$$

Using the chain rule, the conditional distribution $P(z_i = t, x_i = a \mid w_i = w, z_{-i}, w_{-i}, x_{-i}, \tau_{-i}, A, \lambda, \bar{\alpha}, \beta)$ uses the Gibbs sampling and is obtained by:

$$\begin{aligned} P(z_i=t,x_i=a \mid w_i=w,z_{-i},w_{-i},x_{-i},\tau_{-i},A,\lambda,\bar{\alpha},\beta) &\propto \frac{n_{w_i}^{V,T}+\beta}{\sum_v^V n_{w_{-i},t}^{V,T}+V\beta} \\ &\quad\times\frac{n_{x_i,t}^{A,T}+\bar{\alpha}}{\sum_t^T n_{x_{-i},t}^{A,T}+T\bar{\alpha}} \\ &\quad\times\frac{n_{\tau_i}^{V,T}+\lambda}{\sum_v^V n_{\tau_{-i},t}^{V,T}+V\lambda} \end{aligned} \quad (24)$$

where $n_{w_{-i},t}^{V,T}$ is number of the word w being assigned to topic t, not including current word token i. $n_{x_{-i},t}^{A,T}$ is number of words being assigned to topic t for author a to some words, not including the current word i. $n_{\tau_{-i},t}^{V,T}$ is number of the word w being assigned to the topic t under timeinterval τ, not including current word i. The equation 24 represents the conditional probability derived by marginalizing out the random variables, ϑ, φ, and η.

6.4 Mining for crime-relevant chat logs, topics, authors, and authors-topics over time using A-TOT

Before proceeding with the algorithm, we remind the readers that the purpose of developing A-TOT model is to study the authors-topics distribution and to see the impacts of authors over time within these topics. We employ the same algorithm, in Chapter 5, except A-TOT model is used instead of LDA-TOT. Algorithm 2 searches for the crime-related logs, extracts topics with authors, and collects the authors' distributions over time intervals within the discovered topics using A-TOT. The outputs are in the form of four distributions, θ,ϑ,φ, and η. Though θ is not applied in this model, but, in our algorithm, we employ a slightly different implementation of A-TOT model.

The distribution ϑ represents authors-topics probabilities, whereas η outlines the topics-time intervals distribution, and from ϑ and η, the author-topics over time is computed. The detail explanation of the algorithm is illustrated in Chapter 5.

Algorithm 2 Mining for crime-relevant chat logs, topics, authors, and authors-topics over time using A-TOT

Input: $\alpha, \beta, \lambda, \epsilon, \gamma$
φ=Calculate criminal topic-word distribution($D, \alpha, \beta, \lambda$)
V=$\bar{V} \bigcup W$
Δ=KL(d_i,c)
if $\Delta \leq \epsilon$ **then**
    Initialize randomly for all words w_n^N in a chat log d to topics z_t^T
    repeat
        [$\theta_d, \vartheta_a, \varphi, \eta, z_t$] = GibbsSampling($d, a_d, \tau_d, \alpha, \beta, \lambda$)
        for t=1 to T **do**
            σ_t^T=KL($\theta_{d,t}^T, c$)
        end for
        L = GetLowest(σ_t^T)
    until $L \leq \gamma$
end if

6.5 Experiments

Although A-TOT can discover topics, we report that our experiments, in this section, is focused on the second research question, in Chapter 1.

Similar to LDA-TOT, we train the CT model with 200 chat logs, from *perverted-justice*, to compute the distribution of topic c, where c is only sex related, and keep 100 logs for testing the outcomes from A-TOT.

6.5.1 Datasets

The same two datasets, as discussed in Chapter 5, are applied in the experiment. We proceed to the next subsection, and refer readers to the experiment section in Chapter 5 for more details on the two datasets, *perverted-justice* and *IRC*.

6.5.2 Case study

In this subsection, we study the second research question:

Q 2. *Who are the contributors to a topic in a given chat log? How can an investigator track the activity of authors in a log file?*

We divide this question into two parts. First, we determine the proportions of each author contributing in each of the extracted topics. Second, we explore the impacts of the authors throughout the time intervals on the extracted topics. This time, we employ the mining algorithm, using a A-TOT model to study the two parts of the question. We remind that the two users' threshold ϵ and γ are set to 1.25 and 0.86, respectively.

As mentioned previously, A-TOT implementation is slightly different from the proposed one, because we are concerned with collecting information related to θ_d and ϑ_a distributions. We apply the same 4 chat logs, used in Chapter 5, that explore the first research question. From each of these chat logs, A-TOT generates 5 topics with authors associated

Table 7: KL divergence and NMI between crime-related topics from documents (d_1, d_2, d_3, d_4) and c when $\mid c \mid$=30 and $\mid c \mid$=50 using A-TOT

Documents	Size	KL$(t_{i,d}, c)$	NMI$(t_{i,d}, c)$
d_1	$t_4(\mid c \mid$=30)	1.0390	0.1968
	$t_4(\mid c \mid$=50)	1.1812	0.3732
d_2	$t_2(\mid c \mid$=30)	0.7942	0.0421
	$t_4(\mid c \mid$=50)	1.0841	0.0556
d_3	$t_2(\mid c \mid$=30)	1.1768	0.1156
	$t_1(\mid c \mid$=50)	0.6214	0.2429
d_4	$t_4(\mid c \mid$=30)	1.2218	0.2038
	$t_1(\mid c \mid$=50)	0.7988	0.4135

to each. The θ_d distribution for the crime-related topics, from the 4 chat logs, is displayed (between the round brackets) in Table 8. We observe similar results when comparing the distribution θ_d, from Tables 6 and 8, for both models, LDA-TOT and A-TOT. Additionally, the transitions of topics are also similar, as shown in Figures 9,10,12 and 13. However, the comparison between A-TOT and LDA-TOT models is not addressed in this thesis.

The generated ϑ_a^t distribution, using A-TOT is shown in Table 8. The top 3 authors, with the highest probabilities, for each of the crime-relevant topics in each of the 4 chat logs are displayed. For example, author a_1 in d_3 has a probability of 0.2353 for topic t_2, which outlines the contribution of a_1 out of all authors to the crime-relevant topic t_2 when $\mid c \mid$=30.

Though ϑ_a^t distribution assists investigators to identify the plausible authors in the crime-related topics, it does not provide the contributions and activity of each author during specific time intervals within topics. From Definition 1, time τ_d is associated with both message μ_d and author a_d. Hence, for the second part of the question, we keep tracking the times since the messages were composed. Following up, we characterize authors' contributions during time interval $[\tau^s, \tau^f]$ by:

Table 8: Top 10 relevant words extracted for crime-related topics from documents (d_1, d_2, d_3, d_4), their distribution over documents and their distribution over top 3 authors using A-TOT

d_1				d_2			
|c|=30		|c|=50		|c|=30		|c|=50	
t_4 (0.2617)	**Prob**	**t_4 (0.1940)**	**Prob**	**t_2 (0.2248)**	**Prob**	**t_4 (0.2084)**	**Prob**
girl	0.0331	girl	0.0440	wed	0.0290	good	0.0208
happi	0.0168	feel	0.0259	go	0.0174	time	0.0119
fuck	0.0168	show	0.0195	earlier	0.0145	thing	0.0119
show	0.0160	fuck	0.0183	talk	0.0145	care	0.0119
nite	0.0135	cool	0.0167	care	0.0116	wb	0.0089
suck	0.0106	hot	0.0064	hour	0.0087	went	0.0089
sleep	0.0105	sex	0.0064	night	0.0087	wine	0.0060
busi	0.0100	babe	0.0060	maevlyn	0.0087	pc	0.0060
night	0.0090	sleep	0.0048	thank	0.0058	gnite	0.0060
hurt	0.0084	naughti	0.0040	pc	0.0058	old	0.0060
Authors	**Prob**	**Authors**	**Prob**	**Authors**	**Prob**	**Authors**	**Prob**
a_1	0.4286	a_5	0.2143	a_1	0.3077	a_5	0.2667
a_2	0.2819	a_3	0.2005	a_2	0.2353	a_4	0.2523
a_3	0.2629	a_4	0.1904	a_3	0.2329	a_9	0.2143

d_3				d_4			
|c|=30		|c|=50		|c|=30		|c|=50	
t_2 (0.1870)	**Prob**	**t_1 (0.1636)**	**Prob**	**t_4 (0.2383)**	**Prob**	**t_4 (0.1891)**	**Prob**
love	0.0170	peni	0.0491	wear	0.0201	sweet	0.0104
luv	0.0166	suck	0.0181	nite	0.0196	nite	0.0093
feel	0.0085	cam	0.0176	nude	0.0152	pussi	0.0077
babi	0.0085	touch	0.0176	pussi	0.0143	tit	0.0066
bed	0.0050	girl	0.0102	fuck	0.0138	feel	0.0066
vagina	0.0046	panti	0.0094	babi	0.0112	sleep	0.0060
finger	0.0042	hair	0.0087	gf	0.0107	girl	0.0060
peni	0.0037	luv	0.0047	bf	0.0067	drink	0.0055
kiss	0.0033	bed	0.0037	swallow	0.0049	romant	0.0055
big	0.0028	stare	0.0035	pretti	0.0040	suck	0.0049
Authors	**Prob**	**Authors**	**Prob**	**Authors**	**Prob**	**Authors**	**Prob**
a_1	0.2353	a_2	0.2000	a_1	0.2564	a_2	0.1995
a_2	0.2000	a_4	0.1731	a_2	0.2396	a_5	0.1925
a_3	0.1930	a_3	0.1477	a_3	0.2292	a_4	0.1916

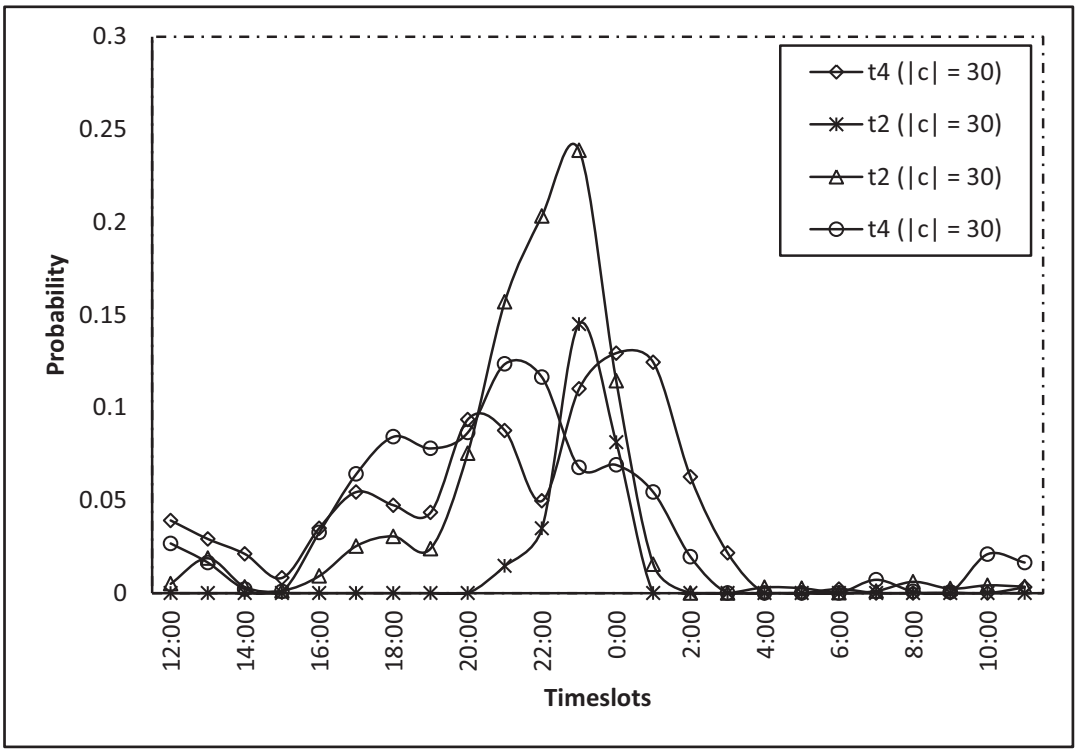


Figure 12: Evolution of crime-related topics using A-TOT when $\mid c \mid$=30

$$F(a_d^t)_{\tau^s}^{\tau^f} = \begin{cases} active & \text{if } p(a_d^t)_{\tau^s}^{\tau^f} \geq \text{users' threshold, } F(t)_{\tau^s}^{\tau^f} \text{ is active} \\ not-active & \text{otherwise} \end{cases}$$

An author is said to be *active* during the interval $[\tau^s, \tau^f]$ for topic t, if the probability of an author participating in t, during that interval, exceeds the users' threshold, and $F(t)_{\tau^s}^{\tau^f}$ is *active* within that period. The users' threshold is calculated, by taking an average of ϑ_a^t over authors for t. To compute $p(a_{i,d}^t)_{\tau^s}^{\tau^f}$, we first map the contribution of an author $a_{i,d}^t$, within $[\tau^s, \tau^f]$, using $P(a^{\tau^s}|t)$=$\frac{p(a^{\tau^s}|d^{\tau^s})\cdot p(t^{\tau^s}|d^{\tau^s})}{p(d^{\tau^s})}$ per time instance s. Next, we calculate $\sum_{\tau^s}^{\tau^f} P(a^{\tau^s}|t)$, as a total probability for author a^t during $[\tau^s, \tau^f]$.

The transitions of the crime-related topics when $\mid c \mid$=30 and $\mid c \mid$=50, using A-TOT are shown in Figures 12 and 13. From these figures and the mapping function, we determine authors' activity over time. For example, let us analyze authors' activity in topic t_{4,d_4} during

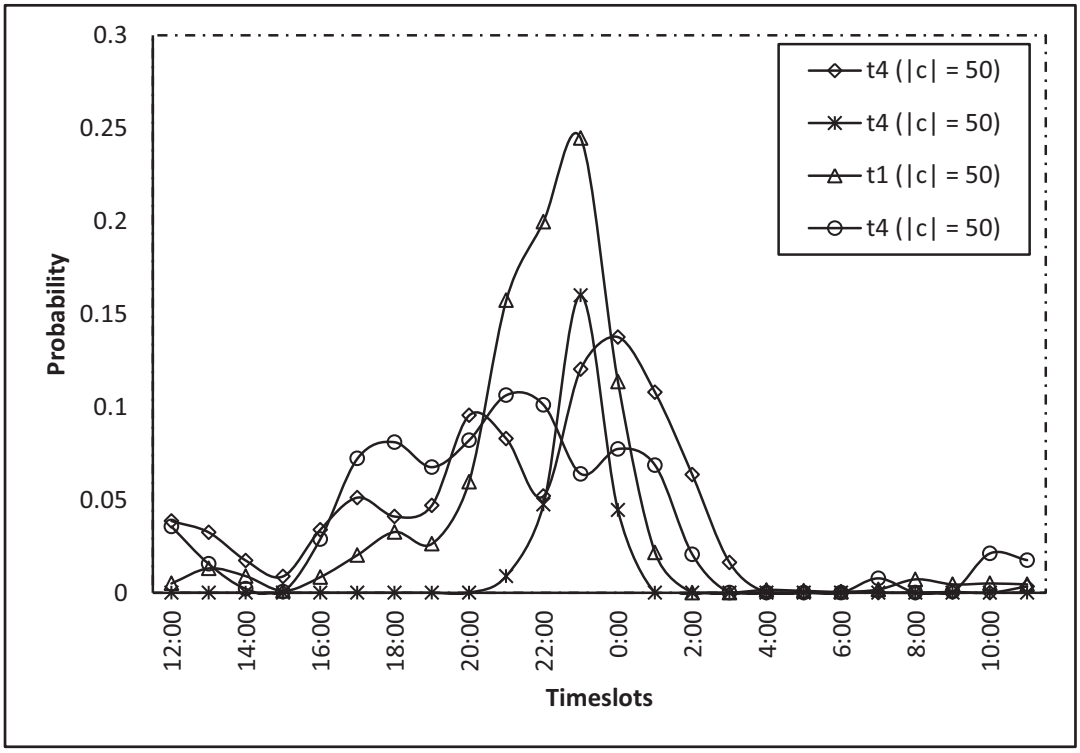


Figure 13: Evolution of crime-related topics using A-TOT when $\mid c \mid$=50

[16:00,19:00]. First, we determine the users' threshold, which is 0.1862 as an average of ϑ^{t_4}. Next, the mapping function is calculated for all authors. For simplicity, let us pick an author a_4 and time instance s=16:00. Then, we compute the mapping function, which is $P(a_{4,\tau^{16:00}}|t_4)$= 0.0467. Afterwards, the total probability of a_2 is estimated, by computing $\sum_{\tau^{16:00}}^{\tau^{19:00}} p(a_{4,\tau^s}|t_4)$=0.2660. Consequently, we say the authors (a_1, a_4) for topic t_{4,d_1} are *active* for satisfying the two conditions when applying the transition function $F(a_d^t)_{\tau^s}^{\tau^f}$, while the authors (a_2, a_3, a_5) are not.

Figure 14 summarizes the activity of authors for the crime-related topic t_{4,d_1}. It can be observed that the most active time for authors occurred during [0:00,7:00] and [15:00,23:00]. This helps the investigators determine the initiator of a topic and to capture the plausible authors within intervals. If the given time period [15:00,19:00] is an important interval for an investigator, then the suspected authors are (a_1, a_4), since they are active during that

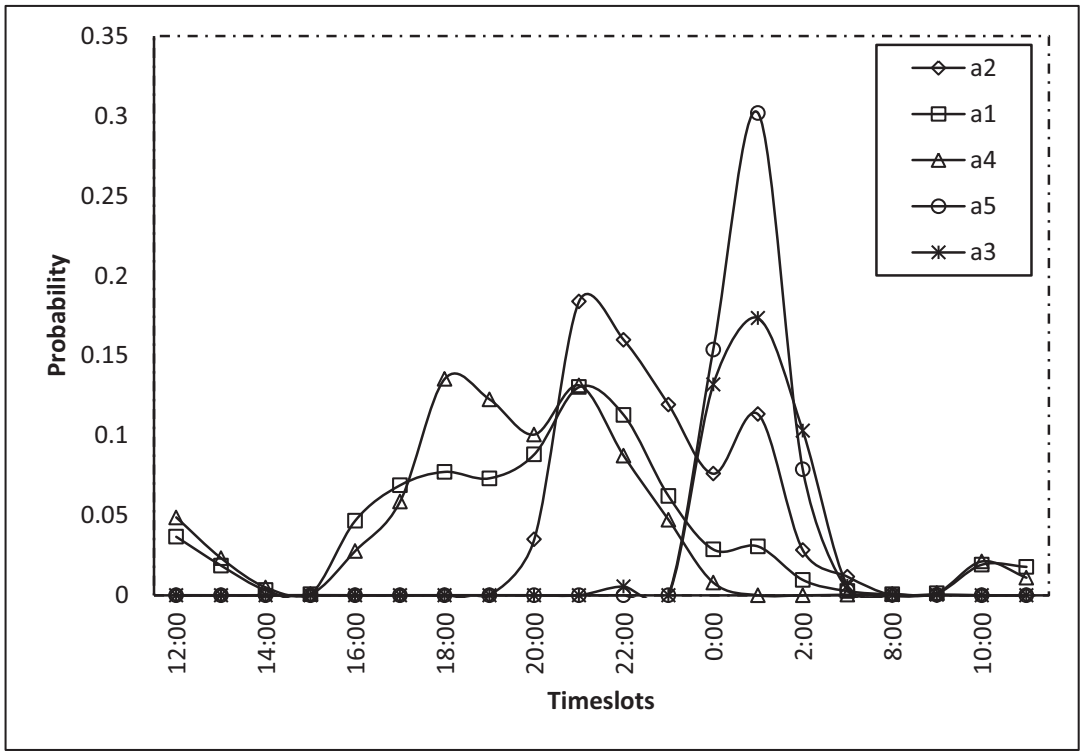


Figure 14: Authors activity for crime-related topic t_{4,d_4} using A-TOT when $\mid c \mid$=50

phase of time, while (a_2, a_3, a_5) are not active.

Analogy to LDA-TOT, when we increase the size of c, the probability of authors-topics are different in the context of crime-relevant topics. For example, from Table 8, the probability of author a_3 in t_{4,d_1} when $|c|$=30 is 0.2005, unlike 0.2629 when $|c|$=50. The NMI for the discovered crime-relevant topics, in Table 7, are improved and new words are obtained, as explored in Table 8. Hence, we determine that the NMI value of topics quantifies the best obtained results. Note, the criminal words used in c are collected from the two datasets.

Table 9: Precision, Recall, F_1, and F_2 using LDA-TOT and A-TOT

	Precision	Recall	F_1	F_2
$\mid c \mid$=30	0.72	1	0.84	0.93
$\mid c \mid$=50	0.78	1	0.88	0.95

In Table 9, we list the precision, recall, F_1, and F_2 measures for the two models previously described, LDA-TOT and A-TOT. Both models found all the truth-relevant chat logs, achieving recall values of 1.0 for the two conditions ($|c|$=30 and $|c|$=50). For precision, there are 19 incorrect logs being retrieved for $|c|$=30 and 14 for $|c|$=50; therefore, the values are 0.72 and 0.78, respectively. The different precision values with the two different sizes of c can be explained through KL(d, c). Using fewer terms in c increases the KL(d, c) value, and thus decreases the precision, and vice versa is also true. The calculated results seem to be subjective. This is because the datasets are not large enough, and we expect precision to be low whenever the size of terms provided in c is small in a huge collection of data.

We conclude that ϑ_a^t, which describes the authors-topics distribution, defines authors contributions in each topic. The characteristics of the authors during several intervals is studied through the transition function $F(a_d^t)_{\tau^s}^{\tau^f}$. In addition, integrating the two distributions, ϑ_a and η_t, into the A-TOT model assists investigators in searching for authors-topics and topics over time, instead of relying on separate time-consuming computation.

6.6 Summary

In this chapter, we presented a Bayesian network *Author-Topics over Time (A-TOT)* model for discovering topics with authors and exploring the movement of authors over time in these topics given a corpus of text messages.

We employed A-TOT model, in the mining algorithm, to address the second research question, and the obtained results provide the measurement for tracing authors over time within discovered topics.

In conclusion, we describe that the probabilistic language model A-TOT would form a useful component in systems for expert-finding of authors, topics recommendation and

prioritization, and understanding the flow of the topics in a relation with authors, in order to make decision on the most plausible authors given a chat log efficiently.

Chapter 7

Conclusion and Future Work

We propose an effective method, using *LDA-TOT* and *A-TOT* models, to extract information from collections of documents. The collected information includes authors, topics, topics time-trends, and authors-topics over time. The algorithm helps investigators to analyze criminal logs in a corpus of detained chat logs. We sought to study chat logs, in different granularity, to identify and segregate crime-related logs and topics associated with these logs. Next, we studied the concept of evolution of topics over time in order to explore the temporal information in these topics. We went a step further by exploring authors' activity within these topics, which represents the evolution of authors-topics over time. In an attempt to build our proposed method, we developed two models with multiple modality attributes influenced by three past models, *LDA*, *AT*, and *TOT*. As for evolution, we used discretization of time to capture different fluctuations of topics over discrete time stamps, instead of using continuous time as does the TOT model. Although our proposed *LDA-TOT* and *A-TOT* models are intended for crime investigation in chat logs, these models could be also applied to other dataset for different purposes, such as twitter and ebay.

We conducted extensive empirical study on the proposed models, by applying results to two datasets, *perverted-justice* and *IRC*. Through our experiments, we demonstrated that our approach can be very useful for an investigator, because it helps identify crime-related

topics. Furthermore, the system is capable of determining the most plausible authors, based on topics expressed in a log and the activity of the authors.

Despite the advantages, probabilistic models, ours and in general, suffer from several shortcomings. These limitations motivate us to consider additional future research directions to supplement the limitations in the area of topic extraction in microblogging environment. For now, we list the limitations when applying to chat logs:

Document size. Due to the short size of chat logs in general, it was hard to obtain the best mixture of topics θ_d and the authors-topics distribution ϑ_a, during conducting experiments, and we applied the two algorithms several times until we obtained the best results. Therefore, we deduce that the accuracy of the extracted topics depends on the size of the chat logs. Although several works, as [HD10, LJW10], deal with short text environments (microblogging), such as Twitter, none of them define a proper method for dealing with texts in chat logs.

Input processing. In a probabilistic model, the "bag of words" assumption is used for modeling topics. We observe that depending on this assumption, in many cases, might not infer a true topic in a chat log. For example, if a chat log is related to a drug topic and drug-related terms occur a few times, the model might generate topics not related to drug. Additionally, none of these models care much about the words processing. These words might contain a lot of noise, ambiguity, and even imprecision. Moreover, as TOT model, we assume that the meaning of topics, generated from the LDA-TOT and A-TOT, are constant though their occurrences and correlations change significantly over time. However, drawing timestamps from a single distribution does not provide a good mechanism for dealing with bursty data, which is common in data streams. Consequently, a generative model that deals with the inputs is one of our future research directions.

Users' threshold. We used several thresholds in our experiments, such as the number of topics (T). Though Teh et al. [TJBB05] proposed a *Hierarchical Dirichlet Processes model* that automatically infers the number of topics among the documents, other thresholds as $F(a_d^t)_{\tau^s}^{\tau^f}$, which lies outside the A-TOT model, are not defined, automatically. These thresholds are synthetic and do not explicitly relate to the prior knowledge of an investigator. Nonetheless, our choices are somewhat subjective, as there is no standard way to determine the optimal values.

Topic correlation. A chat log contains topics that overlap each other, and the proposed models do not capture topic correlation. For example, a document about genetic might also be about disease. Several works on correlation of topics are adopted in many variants of LDA, as in ([BL05,LM06,SM06]). In future work, we will consider this issue in a great detail.

Criminal Topic model. Although CT model is used to segregate crime-relevant logs and to discover the crime-related topics, it reduces the required calculations and improves the quality of the discovered topics, if the CT model was integrated with the existing proposed models. From the experiments, we observe that the quality and the accuracy of the discovered criminal logs depend on the provided terms in the CT model. This is because the CT model is based on the previously outlined "bag of words" assumption. Therefore, as a part of future research, we consider integrating this model to the proposed models, and depend less on the terms availability, by introducing some other measurements.

As time will progress, new research area will possibly emerges to extend or improve the proposed framework, which efficiently serves the forensics investigators' requirements, and to present a more robust model that integrates all the shortcomings of the current models.

Bibliography

[ABD08] Loulwah AlSumait, Daniel Barbará, and Carlotta Domeniconi. On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *Proceedings of the 8th International Conference on Data Mining*, pages 3–12, 2008.

[AC08] Ahmed Abbasi and Hsinchun Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Trans. Inf. Syst.*, 26:7:1–7:29, 2008.

[BA01] Kenneth P. Burnham and David R. Anderson. Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research*, 28:111–119, 2001.

[BL05] David M. Blei and John D. Lafferty. Correlated topic models. In *Proceedings of NIPS*, 2005.

[BL06] David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120. ACM, 2006.

[BL09] David M. Blei and John D. Lafferty. *Topic models*. Chapman & Hall/CRC, 1st edition, 2009.

[BM08] David M. Blei and Jon McAuliffe. Supervised topic models. In *Proceedings of the Advances in Neural Information Processing Systems 20*, pages 121–128. 2008.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[BSSB09] István Bíró, Dávid Siklósi, Jácint Szabó, and András A. Benczúr. Linked latent Dirichlet allocation in web spam filtering. In *Proceedings of the 5th ACM International Workshop on Adversarial Information Retrieval on the Web*, pages 37–40, 2009.

[Bun94] Wray L. Buntine. Operations for learning with graphical models. *J. Artif. Int. Res.*, 2:159–225, 1994.

[CBGB09] Jonathan Chang, Jordan Boyd-Graber, and David M. Blei. Connections between the lines: augmenting social networks with text. In *Proceedings of the 15th International Conference on Knowledge Discovery and Data Mining*, pages 169–178, 2009.

[CFF07] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent object segmentation and classification. In *Proceedings of International Conference in Computer Vision (ICCV)*, 2007.

[Cha94] Eugene Charniak. *Statistical Language Learning*. MIT Press, Cambridge, MA, USA, 1994.

[CT91] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, NY, USA, 1991.

[dVACM01] O. de Vel, A. Anderson, M. Corney, and G. Mohay. Mining e-mail content for author identification forensics. *SIGMOD Rec.*, 30:55–64, 2001.

[GS04] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *In Proceedings of the National Academy of Sciences*, 101:5228–5235, 2004.

[GSBT05] T. Griffiths, M. Steyvers, D. Blei, and J. Tenenbaum. Integrating Topics and Syntax. In *Proceedings of the Advances in Neural Information Processing Systems 17*, 2005.

[Han08] Ria Hanewald. Confronting the pedagogical challenge of cyber safety. *Australian Journal of Teacher Education*, 33(3):1–16+, 2008.

[HD10] Liangjie Hong and Brian D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the 1st Workshop on Social Media Analytics*, pages 80–88, 2010.

[Hei04] Gregor Heinrich. Parameter estimation for text analysis. Technical report, 2004.

[HK06] J. Han and M. Kamber. *Data mining: concepts and techniques*. Elsevier, San Francisco, CA, USA, 2006.

[IBFD10] F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi. Mining writeprints from anonymous e-mails for forensic investigation. *Digital Investigation*, 7:56–64, 2010.

[IBFDss] F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi. A unified data mining solution for authorship analysis in anonymous textual communications. *Information Sciences: Special Issue on Data Mining for Information Security*, in press.

[IHFD08] F. Iqbal, R. Hadjidj, B. C. M. Fung, and M. Debbabi. A novel approach of mining write-prints for authorship attribution in e-mail forensics. *Digital Investigation*, 5:S42–S51, 2008.

[KCAC08] Tayfun Kucukyilmaz, B. Barla Cambazoglu, Cevdet Aykanat, and Fazli Can. Chat mining: Predicting user and message attributes in computer-mediated communication. *Inf. Process. Manage.*, 44:1448–1466, 2008.

[KKK80] S. Kullback, J. C. Keegel, and J. H. Kullback. *Topics in Statistical Information Theory*. Springer-Verlag, Berlin, 1980.

[KL51] Solomon Kullback and Richard A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951.

[LJSJ08] Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification, 2008.

[LJW10] Peng Li, Jing Jiang, and Yinglin Wang. Generating templates of entity summaries with an entity-aspect model and pattern mining. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 640–649, 2010.

[LM06] Wei Li and Andrew McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 577–584. ACM, 2006.

[LNMG09] Yan Liu, Alexandru Niculescu-Mizil, and Wojciech Gryc. Topic-link LDA: joint models of topic and author community. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 665–672. ACM, 2009.

[Mar03] Benjamin Marlin. Modeling user rating profiles for collaborative filtering. In *NIPS*17*. MIT Press, 2003.

[McC99] Andrew McCallum. Multi-label text classification with a mixture model trained by EM. In *AAAI 99 Workshop on Text Learning*, 1999.

[MH04] Gheorghe Muresan and David J. Harper. Topic modeling for mediated access to very large document collections. *J. Am. Soc. Inf. Sci. Technol.*, 55:892–910, 2004.

[ML02] Thomas Minka and John Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 352–359. Morgan Kaufmann, 2002.

[MLSZ06] Qiaozhu Mei, Chao Liu, Hang Su, and ChengXiang Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th International Conference on World Wide Web*, pages 533–542. ACM, 2006.

[MRS08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[MWM07] Andrew McCallum, Xuerui Wang, and Natasha Mohanty. Joint group and topic discovery from relations and text. In *Proceedings of the 2006 conference on Statistical Network Analysis*, pages 28–44. Springer-Verlag, 2007.

[NAXC08] Ramesh M. Nallapati, Amr Ahmed, Eric P. Xing, and William W. Cohen. Joint latent topic models for text and citations. In *Proceeding of the 14th International Conference on Knowledge Discovery and Data Mining*, pages 542–550, 2008.

[NDLU07] Ramesh M. Nallapati, Susan Ditmore, John D. Lafferty, and Kin Ung. Multiscale topic tomography. In *Proceedings of the 13th International Conference on Knowledge Discovery and Data Mining*, pages 520–529, 2007.

[NMTM00] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39:103–134, 2000.

[PGKT06] Matthew Purver, Thomas L. Griffiths, Konrad P. Körding, and Joshua B. Tenenbaum. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 17–24, 2006.

[RHMGM09] Daniel Ramage, Paul Heymann, Christopher D. Manning, and Hector Garcia-Molina. Clustering the tagged web. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*, pages 54–63, 2009.

[RHNM09] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: Volume 1*, pages 248–256, 2009.

[RZCG+10] Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers. Learning author-topic models from text corpora. *ACM Trans. Inf. Syst.*, 28:4:1–4:38, 2010.

[RZGSS04] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494. AUAI Press, 2004.

[SG07] Mark Steyvers and Tom Griffiths. *Probabilistic Topic Models*. Lawrence Erlbaum Associates, 2007.

[SLTS05] Xiaodan Song, Ching-Yung Lin, Belle L. Tseng, and Ming-Ting Sun. Modeling and predicting personal information dissemination behavior. In *Proceedings of the 11th International Conference on Knowledge Discovery and Data Mining*, pages 479–488, 2005.

[SM06] M. Mahdi Shafiei and Evangelos E. Milios. Latent Dirichlet co-clustering. In *Proceedings of the 6th International Conference on Data Mining*, pages 542–551, 2006.

[SSRZG04] Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining*, pages 306–315, 2004.

[TJBB05] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems*, pages 1385–1392. MIT Press, 2005.

[WBH08] Chong Wang, David M. Blei, and David Heckerman. Continuous time dynamic topic models. In *UAI'08*, pages 579–586, 2008.

[WM06] Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining*, pages 424–433, 2006.

[WMG07] Xiaogang Wang, Xiaoxu Ma, and Eric Grimson. Unsupervised activity perception by hierarchical bayesian models. In *Proc. CVPR*, 2007.

[WMM05] Xuerui Wang, Natasha Mohanty, and Andrew McCallum. Group and topic discovery from relations and text. In *Proceedings of the 3rd International Workshop on Link Discovery*, LinkKDD '05, pages 28–35, 2005.

[XC99] Jinxi Xu and W. Bruce Croft. Cluster-based language models for distributed retrieval. In *Proceedings of the 22nd Annual International ACM Conference on Research and Development in Information Retrieval*, pages 254–261, 1999.

[ZGFY07] Haizheng Zhang, C. Lee Giles, Henry C. Foley, and John Yen. Probabilistic community discovery using hierarchical latent gaussian mixture model. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 1*, pages 663–668. AAAI Press, 2007.

[ZQHC03] Rong Zheng, Yi Qin, Zan Huang, and Hsinchun Chen. Authorship analysis in cybercrime investigation. In *Proceedings of the 1st NSF/NIJ Conference on Intelligence and Security Informatics*, pages 59–73. Springer-Verlag, 2003.