

Chapter 1

APPLYING TOPIC MODELLING ON FORENSIC DATA: A CASE STUDY

Alta de Waal, Jacobus P. Venter and Etienne Barnard

Abstract Most actionable evidence for investigation purposes is identified during the analysis phase of a digital investigation process. The objective of the analysis phase (digital analysis) is to reduce the quantity and enhance the intelligibility of data that must be reviewed by a human analyst. Currently, this is done through expression based searching, which assumes a good understanding of the evidence prior to the search. Therefore, latent evidence will not be found with such methods. This suggests a clear role for knowledge discovery and data mining (KDD) techniques to enhance the digital analysis process. The research described in this article investigates the application of topic modelling as a KDD technique on forensic data and its ability to contribute to digital analysis. Topic models infer the underlying semantic context of a text collection and summarises it as topics described by words. The data used for this application was extracted from a real digital investigation case. This novel application highlights several challenges that forensic data poses to topic modelling algorithms and we report on lessons learned from the case study.

Keywords: Evidence Mining, Digital Investigation, Data Mining, Topic Modelling.

1. Introduction

The use and value of information obtained from digital sources in various investigations has been argued widely [1], [11]. This has led to the establishment of the term digital forensics and the growth of this area in practical application and scientific research. The four major phases in digital investigation are acquisition, examination, analysis and reporting [14]. It has been argued that support for the analysis phase (called digital analysis from here on), where most of the actionable evidence is being gathered, lacks sufficient definition and support in terms of prin-

ciples, methods, tools, etc. [14], [17]. The use of knowledge discovery and data mining (KDD) to enhance digital analysis has received some recent attention [14], [17] and the use of KDD principles and tools in digital investigations was defined as evidence mining in [17].

Textual artifacts are very important to many digital investigations [1], [11] and include e-mails, reports, letters, notes, text messages, etc. - collectively referred to as documents. In a typical forensic evidence set thousands of documents are found. Herein lies one of the problems of digital analysis. Of the thousands of documents in an evidence set only a small proportion would be relevant and of the relevant documents only a small proportion may contain actionable evidence. Processing the thousands of text documents manually, in order to find the relevant evidence is a difficult and time consuming task.

Digital analysis is mostly done through expression based searching. This implies that a good understanding of the evidence that is looked for must exist before a search can commence. Information retrieved is not ranked (e.g. based on relevance to the case) in any manner. This situation means that latent evidence will not be found (In this context we use the phrase "latent evidence" to refer to evidence that exists but is not directly accessible to the investigator). The latent evidence must first become visible before it can be considered in the investigation. Evidence mining aims to uncover, through the application of KDD principles and techniques, electronic artifacts that can form part of the evidence set to assist in the development of crime scenarios [17]. This includes known and latent evidence.

A process to support evidence mining was defined in [17] as CRISP-EM (a specialisation of the well-known CRISP-DM process [5]). The work described in this article falls within the scope of the data-preparation tasks of CRISP-EM (see Figure 1). The data preparation phase covers all activities to construct a dataset to be used in the next phase to do event reconstruction and modelling. The construction of this dataset is obviously a challenging task and is a trade-off between choosing relevant data and losing vital information necessary for event reconstruction. A summary of the data could be extremely helpful for the investigator in order to facilitate better understanding of the data content and focusing the data preparation task on relevant data.

Topic modelling as a latent variable analysis technique can assist in associating relevant documents by modelling the underlying (latent) topics in the text collection. Additionally, it suggests prevalent themes within the text which provides a summary of the document collection. As a KDD technique, it has the potential to discover latent evidence often missed with expression-based searching. However, digital evidence

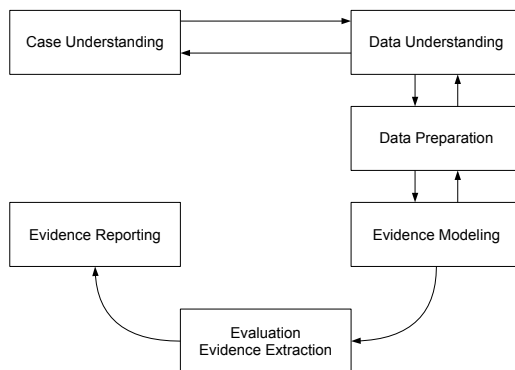


Figure 1. The Main Phases of the CRISP-EM process.

is inhomogeneous in terms of format and content, which poses unique challenges to KDD techniques. In the current research, we investigate the matters that need to be addressed in order to apply topic modelling to forensic data, and we study how useful such models are when applied appropriately.

The article is organised as follows. In section 2 a brief overview of topic modelling is given, along with an example. In section 3 topic modelling is applied to forensic data obtained from a real case and the results are presented. Section 4 addresses the important aspect of the interpretation of the results and how it can assist the forensic investigator. Merging the research fields of digital forensics and topic modelling is not a straightforward exercise and section 5 lists a number of valuable lessons learned from this case study when applying topic modelling as a KDD technique on forensic data. Future work is covered in section 6.

2. Topic Modelling

Today, large collections of digital data are widely available and continue to grow in size at an increasing pace. Trying to understand the meaning of such data is a difficult task and in general the first option is to perform keyword searches. The results of keyword searches do not always describe the meaning of the data collection in a satisfactory way, especially if the user has limited insight into the collection. A summary of the data would be very useful and would ideally encapsulate the main topics within the data [12]. One example of a data collection is a text corpus where documents represent entities in the text corpus. Examples include news articles, conference proceedings or minutes of meetings. In the case of a text corpus of news articles, a summary of topics could include politics, sport, finance, culture and local news.

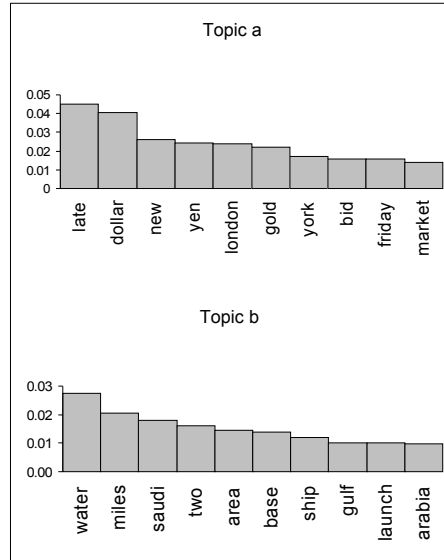


Figure 2. Word probability distributions over two topics (top-10 words)

When one thinks of a text corpus as a collection of documents, it makes sense that each document has an underlying semantic context. This semantic context develops as the document is generated and refers to the intended meaning of the document. For example, a newspaper article has the purpose of reporting on a news event and as we read the article, we become aware of the intended message the author(s) is hoping to communicate. Although the semantic context is hidden, it is represented in the words of a document. Topic modelling addresses the retrieval of semantic context from a text corpus and can be described as a problem of statistical inference. When given the data (words), the latent semantic context from which it was generated can be inferred [7]. A topic is defined as a probability distribution over words. In statistical terms, a topic model is a latent variable model where the latent variables describe the topics [2]. Figure 2 is an example of two topics. The topics resulted from a subset of the TREC AP [8] corpus and the top-10 words (10 words with highest probabilities) are illustrated in the graph with their respective probabilities indicated on the y-axis. The top-10 words describe the particular topic. Topic 4 clearly has to do with financial markets whereas topic 14 deals with a naval event in Saudi Arabia.

The fundamental assumption in topic modelling is that the semantic context of a document is a mixture of topics [7]. A bag-of-words approach is commonly adopted for topic modelling, which means that

each document is treated as a collection of words, ignoring the structure of the document. The output of the bag-of-words approach is a word \times document frequency matrix where $cell_{ij}$ represents the frequency of $word_i$ in $document_j$.

3. Topic Modelling applied on Forensic Data

When applied to text data, topic modelling provides a summary of the documents by describing the latent topics in the data as illustrated in Figure 2. This leads to two useful outputs: the first output is a visual summary of the topics and the second output is a visual representation of the document space.

3.1 Topic Modelling Process

Figure 3 indicates the process followed in order to apply topic modelling to the analysis of the original forensic data. Each level represents a data set of a different nature and size. Level 1 in the figure represents the original forensic data set. Levels 2 to 4 describe the data filtering process and level 5 involves the data pre-processing step which results in the word \times document input matrix for topic modelling. The topic modelling results define level 6.

3.2 Data Set

The data set can be described in parallel with the levels in the process graph:

- 1 The data used as a text corpus in the experiment was taken from a real digital investigation (level 1 of figure 3) and includes various entities (more than 100,000) such as documents, operating system files, deleted entities, page files, etc.
- 2 The data set and data type was selected (CRISP-EM tasks 3.1-A: Select sites/equipment/device and 3.1-B: Select types of data to be included). All the documents type files (.doc, .txt, .pdf, .html and .rtf) in the evidence set were extracted using the Forensics Toolkit from AccessData (*FTKTM*). This includes only allocated, or logical type files. The data was extracted from three devices from one case. This data set of document type files is on level 2 of figure 3 and contained 12,483 documents.
- 3 The data set was reduced to documents with natural language content (CRISP-EM task 3.2-A: Data Reduction). After converting the documents to text files (CRISP-EM task 3.5-A: Convert

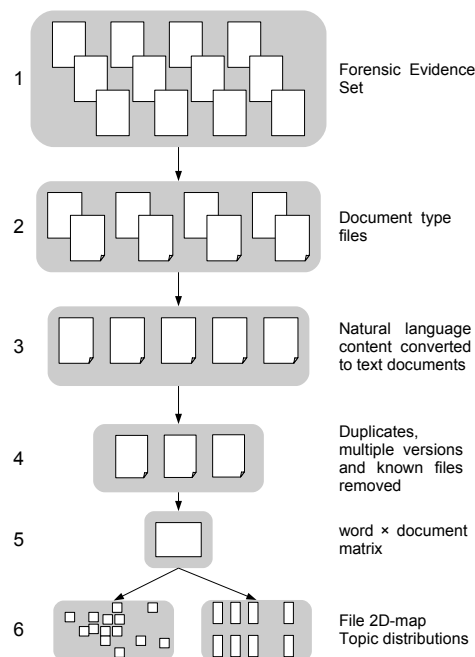


Figure 3. Topic modelling output and interpretation scheme for forensic data

Data Formats), the data set at level 3 of figure 3 contained 1,661 documents.

4 Removing files such as keystroke log files, software documentation, multiple versions of the same document and files with no text (CRISP-EM task 3.2-A: Data Reduction), provided the data set of 837 files at level 4 of figure 3.

The rest of this section indicates the further processing done on the data. This relates to CRISP-EM task 3.3-D: Perform Text Processing.

3.3 Data Pre-processing

As a data pre-processing task, stop words (common words that appear frequently in text), words occurring only once in the corpus, numbers, special characters and words with two characters or less were removed from the data. The data pre-processing was done in the open source programming language *Python*. Data pre-processing results in a word \times document matrix and for this case study the matrix size was approximately 11,000 words \times 837 documents (depending on stemming or no stemming included in the pre-processing task). This matrix defines the

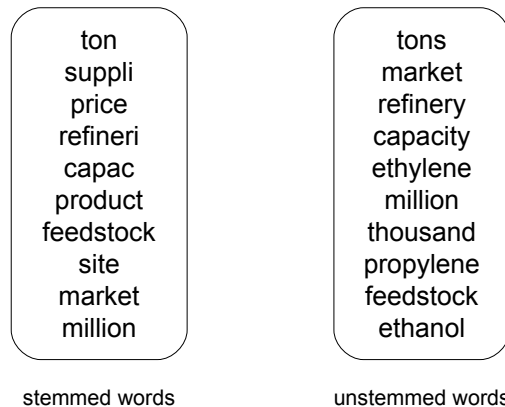


Figure 4. Comparison of topic: with and without stemming

data set of level 5 in figure 3. The word \times document matrix is the input to topic modelling described in the next section.

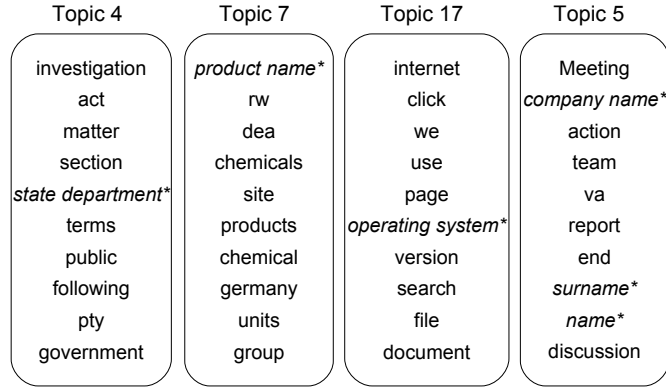
3.4 Experimental Setup

Early into experimentation it became clear that the forensic data poses unique challenges to the topic modelling approach, which are discussed in detail in section 5. One challenge that deserves mentioning in this section is the use of stemming as a standard data pre-processing technique. Stemming is a process for reducing derived words to their stem. For example, ‘wait’ is the stem for ‘waiting’, ‘waits’ and ‘waited’. The Porter stemming algorithm [15] in the Natural Language Toolkit for *Python* was used to perform stemming. Stemming was planned as a standard pre-processing task, but the stemmed words hampered the intelligibility and interpretation of topic distributions. We ran two experiments:

- 1 Apply stemming to words in the pre-processing task,
- 2 Use inflections and derived versions of words, without stemming.

Figure 4 illustrates the comparison of a topic with and without stemming. It is important to understand the influence that stemming has on the interpretation of the results: if stemming may hamper an investigator in grasping the gist of a topic due to not being able to see the original unstemmed word, then it would be more appropriate to develop topics without the use of stemming (even though this will increase the dimensionality of the problem).

A number of topic models exist with different assumptions about the origin of the distribution over topics [7]. The Latent Dirichlet Alloca-



**Information changed due to sensitive nature of original data*

Figure 5. Illustration of topic modelled from forensic data

tion (LDA) model makes the assumption that the set of topics originated from a Dirichlet distribution. In practice, this results in more reasonable mixtures of topics in a document, compared to previous approaches where an explicit model had not been employed [2]. For the purpose of this experiment, LDA as the topic model was chosen. For simplification, the number of topics was fixed to 20. In future, the LDA model will be extended by defining the number of topics as a random variable, which allows the topic model to infer the natural number of topics inherent in the text corpus. The *Matlab*[®] Topic Modelling Toolbox [6] was used to perform the LDA topic modelling.

3.5 Results

The output of topic modelling can be represented in two ways, a word \times topic matrix and topic \times document matrix, which define the data set of level 6 in figure 3. These two representations are described below.

- Word \times topic matrix: Each column represents a topic as a probability distribution over words. The top-10 words (words with the highest probabilities) are a good description of what the topic is about. Listing the top-10 words of each topic provides a summary of the document collection. Examples of the topics found are shown in Figure 5. Topic 17 indicates computer usage and internet access/searching. Topic 5 indicates company meetings with the frequent involvement of a specific individual.
- Topic \times document matrix: Each column represents a mixture of topics for a document. The mixture of topics describes the seman-

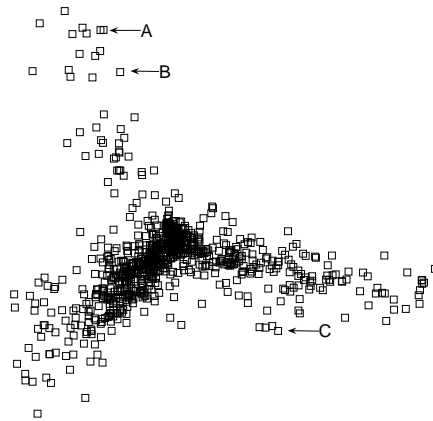


Figure 6. Visualisation of documents in 2D map

tic context, or gist of the document [7]. Documents with a similar mixture (topic distribution) are closely related in terms of semantic context. This ‘relatedness’ of documents can be visualized in a 2D map: For each document pair, the symmetrised Kullback-Leibler divergence between topic distributions is calculated. (The Kullback-Leibler divergence is a measure of difference between two probability distributions [10].) Classical multidimensional scaling is used to visualize all pairwise document distances in the 2D map. Figure 6 illustrates the 2D visualisation of the forensic data documents where each block represents a document. The graph can be interpreted as follows: document A (indicated by symbol A in Figure 6) are closely related to document B in terms of their respective mixture of topics (semantic context). Document A and C differ significantly in terms of semantic context. For the purpose of forensic analysis it means that if document A is identified as a relevant document to the case, one would rather focus investigation efforts on document B than document C. A similar 2D map can be generated for topics, in which case it will indicate the relatedness between topics. The implication of such a map for digital forensics means that if a topic is identified as being relevant to the case, then neighbouring topics on the 2D map can be prioritised in the investigation.

4. Forensic benefit of the results

Topic modelling can assist cyber forensic analysts and investigators in different ways. Firstly, in a large case, with multiple datasets from multiple sites, performing topic modeling on natural language data will provide the analyst/investigator with an overview of the data in terms of the broad semantic contexts. The benefit of this is a summary of the natural language data which could assist the investigator in prioritizing the data to be analysed. Conventional techniques use keyword searches to identify relevant documents. The document 2D map (Figure 6) can then be used to identify closely related documents that would typically not be found from the keywords. This assists in finding other relevant documents and therefore expanding the set of relevant documents. Words associated with topics can be used to expand existing keyword sets. Where an existing keyword occurs within the top-10 words defining a topic, the other words defining that topic can then be included in the keyword set. This expands the set of keywords based on the actual characteristics of the forensic data and not prior case knowledge.

5. Lessons Learned

This research has shown that topic modeling can be a valuable component for digital analysis, both for the purposes of reducing the quantity of data that must be reviewed by a human analyst and for suggesting themes that are prevalent within a set of documents to be analyzed. Although much research remains to be done on the development and application of algorithms that perform well in this context, our work has shown that even off-the-shelf algorithms can function usefully. One issue that deserves urgent attention is the design of performance metrics that reflect the goals of forensic modeling during the development of topic models. For standard applications of topic models this is a significant challenge [16], and for this novel application even more so. Such metrics should reflect the particular requirements of the forensic environment (e.g. intelligibility to a human analyst, salience of topics detected); they will serve as a crucial guide for the development of more sophisticated algorithms.

In addition, several practical matters were found to be important during our investigation. These include the following:

- Many documents are represented by a variety of versions within the extracted set. Treating these versions as independent documents skews the topics extracted and increases the computational complexity of the modeling problem unnecessarily; on the other hand, detecting different versions is a computational challenge (for exam-

ple, one version may contain only a portion of another; hence, the percentage overlap may be low). It is also a non-trivial challenge to merge the different versions without risking the loss of relevant information.

- Named entities (person names, locations and organizations) are important in digital investigations and have high evidence potential. It needs to be treated with some care. We recommend that named entities be recognized (using an algorithm such as that described by Louis et al. [9]) and removed from documents temporarily (to exclude them from data pre-processing tasks such as stemming and removal of stop words). Once the pre-processing is complete, they can be returned to the document as concepts and not as individual words. Newman et al. [13] combined topic models and named entity recognizers to jointly analyse named entities and topics. This way topics relate entities to one another which provides a wealth of information on people, organisations and location mentioned in the text corpus.
- In many cases, documents from several languages may be present in the same corpus. Documents from different languages should be treated separately from one another, for several reasons (investigators may be proficient in only a subset of the languages, data pre-processing tasks such as stemming and spell checking are language dependent, current techniques are not suitable for topic modeling across languages, etc.) It is therefore advisable to separate documents using an automatic system such as that developed by Botha et al. [3].
- Stemming successfully reduces the number of parameters in the corpus and consolidates semantically related words; it also increases the number of occurrences of individual words in the corpus, which leads to better modeling. However, as mentioned in section 3.4, the use of stemming for forensic data may hamper the understanding of topic distributions; it may therefore be advisable to revert to the original words when presenting the topics to an investigator.
- ‘Known files’ include help files of purchased software, license agreements, etc. They need to be removed from the corpus before analysis, to reduce the amount of spurious data presented to the analyst. Fortunately, this can be done efficiently (e.g. through the use of hash values for lists of ‘known’ documents such as ‘readme.txt’ files of known programs).

- Spelling mistakes add parameters to the model and result in incorrect word statistics - the count of one word is split into more than one spelling variant. However, it is difficult to automate spell-checking reliably in an informal context: important neologisms and jargon related to the investigation could be transcribed wrongly. It is probably preferable to have a low precision rather than to wrongly correct spelling mistakes, but this matter deserves further investigation.
- It is standard practice in topic modeling to remove words occurring only once in the corpus. This usually leads to the removal of approximately 5% of the vocabulary in the corpus. When this was done for the forensics data set, approximately 50% of the vocabulary was removed, suggesting that much useful information may be discarded in the process. A more inventive way of dealing with unique occurrences is therefore likely to be useful in this context.
- Text corpora used for topic modeling are typically homogeneous (e.g. news articles, conference proceedings or book chapters). Forensic data, on the other hand, are generally a mixture of documents, reports, letters, email bodies and faxes. It may therefore be beneficial to modify topic modeling approaches to cope with such inhomogeneous data more successfully (e.g. to avoid the bias towards longer documents that is inherent in the statistics used by current approaches).

6. Future Work

This article reports on one case study of topic modelling applied to forensic data very early into an ongoing investigation. No evidence has been discovered on this investigation as yet. Future work includes reporting on the benefits of including topic modelling as a digital analysis technique in the ongoing investigation. More studies are needed on different case types, crimes and evidence sets in order to validate the lessons learned and possibly expanding the list. Topic model algorithms can be expanded to take into account evolution of topics and the temporal component of the data [12]. Such topic models captures how the structure of the data changes over time [18] which could indicate the rise and fall in prominence of topics over time.

One useful characteristic of forensic data is the abundance of metadata. Metadata could be use to add the temporal component as well as other information (e.g. author indication, machine origin, external links, etc.) to KDD techniques in order to generate more informative results.

References

- [1] Beebe, N. and J. Clark, Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results, *Digital Investigation* Vol. 4S, 2007
- [2] Blei, D., A. Ng and M. Jordan, Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993-1022, 2003.
- [3] Botha, G., V. Zimu and E. Barnard, Text-based Language Identification for the South African Languages, *Proceedings of the 17th Annual Symposium of Pattern Recognition Association of South Africa*, Parys, South Africa, November, 2006.
- [4] Case, E., *Digital Evidence and Computer Crime*, Academic Press, 2000.
- [5] Chapman, P., et al, *CRISP-DM 1.0 Step-by-step data mining guide*, SPSS, 1999.
- [6] Griffiths, T. and M. Steyvers, Finding Scientific Topics. *Proceedings of the National Academy of Sciences*, 101 (suppl. 1), pp 5228-5235, 2004.
- [7] Griffiths, T., M. Steyvers and J. Tenenbaum, J, Topics in Semantic Representation. *Psychological Review*, 114(2), 211-244, 2007.
- [8] Harman, D., Overview of the first text retrieval conference (TREC-1). In *Proceedings of the First Text Retrieval Conference (TREC-1)*, pp 1-20 1992.
- [9] Louis, A., A. De Waal and J. Venter, Named entity recognition in a South African context. *Proceedings of the 2006 SAICSIT conference*, pp 170-179, South Africa, 2006.
- [10] Mackay, D., *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
- [11] McCue, C., *Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis*, Elsevier, 2007
- [12] Mei, Q. and C. Zhai, Discovering Evolutionary Theme Patterns from Text - An Exploration of Temporal Text Mining, *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD' 05)*, pp 198-207, 2005
- [13] Newman, D., C. Chemudugunta, P. Smyth and M. Steyvers, Analyzing Entities and Topics in News Articles Using Statistical Topic Models, S. Mehrotra et al (Eds): *ISI 2006, LNCS 3975*, pp.93-104, Springer-Verlag Berlin Heidelberg, 2006.

- [14] Pollitt, M. and A. Whitledge, Exploring Big Haystacks Data Mining and Knowledge Management, International Federation for Information Processing, Volume 222, Advances in Digital Forensics II, eds. Olivier, M., Sheno, S., Boston: Springer, pp. 67-76, 2006.
- [15] Porter, M., An algorithm for suffix stripping, Program, 14(3) pp 130:137, 1980.
- [16] Rigouste, L., O. Capp and F. Yvon, Inference and evaluation of the multinomial mixture model for text clustering. Inf. Process. Manage. 43(5) pp 1260-1280, 2007.
- [17] Venter, J., A. De Waal and N. Willers, Specialising CRISP-DM for Evidence Mining, International Federation for Information Processing, Advances in Digital Forensics III, eds. Sheno, S., Boston: Springer, 2007.
- [18] Wang, X. and A. McCallum, Topic over Time: A Non-Markov Continuous-Time Model of Topical Trends, KDD'06 Philadelphia, USA, August 20-23, 2006.