

words and documents.

A somewhat different approach is to cluster the documents into groups containing similar semantic content, using any of a variety of well-known document clustering techniques (e.g., Cutting et al., 1992; McCallum, Nigam, and Ungar, 2000; Popescul et al., 2000). Each cluster of documents can then be associated with a latent topic (e.g., as represented by the mean term vector for documents in the cluster). While clustering can provide useful broad information about topics, clusters are inherently limited by the fact that each document is (typically) only associated with one cluster. This is often at odds with the multi-topic nature of text documents in many contexts. In particular, combinations of diverse topics within a single document are difficult to represent. For example, this present paper contains at least two significantly different topics: document modeling and Bayesian estimation. For this reason, other representations (such as those discussed below) that allow documents to be composed of multiple topics generally provide better models for sets of documents (e.g., better out of sample predictions, Blei, Ng, and Jordan (2003)).

Hofmann (1999) introduced the aspect model (also referred to as probabilistic LSI, or pLSI) as a probabilistic alternative to projection and clustering methods. In pLSI, topics are modeled as multinomial probability distributions over words, and documents are assumed to be generated by the activation of multiple topics. While the pLSI model produced impressive results on a number of text document problems such as information retrieval, the parameterization of the model was susceptible to overfitting and did not provide a straightforward way to make inferences about new documents not seen in the training data. Blei, Ng, and Jordan (2003) addressed these limitations by proposing a more general Bayesian probabilistic topic model called latent Dirichlet allocation (LDA). The parameters of the LDA model (the topic-word and document-topic distributions) are estimated using an approximation technique known as variational EM, since standard estimation methods are intractable. Griffiths and Steyvers (2004) showed how Gibbs sampling, a Markov chain Monte Carlo technique, could be applied in this model, and illustrated this approach using 11 years of abstract data from the *Proceedings of the National Academy of Sciences*.

Our focus here is to extend the probabilistic topic models to include authorship information. Joint author-topic modeling has received little or no attention as far as we are aware. The areas of stylometry, authorship attribution, and forensic linguistics focus on the problem of identifying what author wrote a given piece of text. For example, Mosteller and Wallace (1964) used Bayesian techniques to infer whether Hamilton or Madison was the more likely author of disputed Federalist papers. More recent work of a similar nature includes authorship analysis of a purported poem by Shakespeare (Thisted and Efron, 1987), identifying authors of software programs (Gray, Sallis, and MacDonell, 1997), and the use of techniques such as support vector machines (Diederich et al., 2003) for author identification.

These author identification methods emphasize the use of distinctive stylistic features (such as sentence length) that characterize a specific author. In contrast, the models we present here focus on extracting the general semantic content of a document, rather than the stylistic details of how it was written. For example, in our model we omit common

“stop” words since they are generally irrelevant to the topic of the document—however, the distributions of stop words can be quite useful in stylometry. While “topic” information could be usefully combined with stylistic features for author classification we do not pursue this idea in this particular paper.

Graph-based and network-based models are also frequently used as a basis for representation and analysis of relations among scientific authors. For example, Newman (2001), Mutschke (2003) and Erten et al. (2003) use methods from bibliometrics, social networks, and graph theory to analyze and visualize co-author and citation relations in the scientific literature. Kautz, Selman, and Shah (1997) developed the interactive ReferralWeb system for exploring networks of computer scientists working in artificial intelligence and information retrieval, and White and Smyth (2003) used PageRank-style ranking algorithms to analyze co-author graphs. In all of this work only the network connectivity information is used—the text information from the underlying documents is not used in modeling. Thus, while the grouping of authors via these network models can implicitly provide indications of latent topics, there is no explicit representation of the topics in terms of the text content (the words) of the documents.

The novelty of the work described in this paper lies in the proposal of a probabilistic model that represents both authors and topics, and the application of this model to a large well-known document corpus in computer science. As we will show later in the paper, the model provides a general framework for exploration, discovery, and query-answering in the context of the relationships of author and topics for large document collections.

The outline of the paper is as follows: in Section 2 we describe the author-topic model and outline how the parameters of the model (the topic-word distributions and author-topic distributions) can be learned from training data consisting of documents with known authors. Section 3 illustrates the application of the model to a large collection of abstracts from the CiteSeer system, with examples of specific topics and specific author models that are learned by the algorithm. In Section 4 we illustrate a number of applications of the model, including the characterization of topic trends over time (which provides some interesting insights on the direction of research in computer science), and the characterization of which papers are most typical and least typical for a given author. An online query interface to the system is described in Section 5, allowing users to query the model over the Web—an interesting feature of the model is the coupling of Bayesian sampling and relational database technology to answer queries in real-time. Section 6 contains a brief discussion of future directions and concluding comments.

2. AN OVERVIEW OF THE AUTHOR-TOPIC MODEL

2.1 The Probabilistic Generative Model

The author-topic model reduces the process of writing a scientific document to a simple series of probabilistic steps. The model not only discovers what topics are expressed in a document, but also which authors are associated with each topic. To simplify the representation of documents, we use

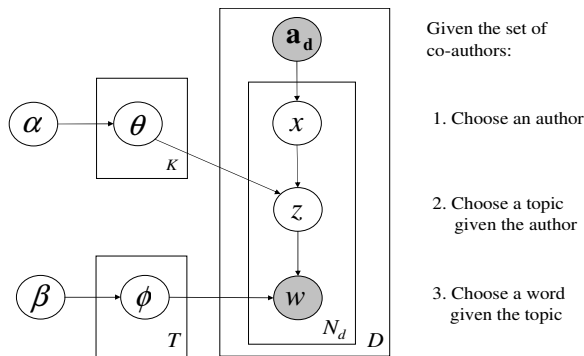


Figure 1: The graphical model for the author-topic model using plate notation.

a bag of words assumption that reduces each document to a vector of counts, where each vector element corresponds to the number of times a term appears in the document.

Each author is associated with a multinomial distribution over topics. A document with multiple authors has a distribution over topics that is a mixture of the distributions associated with the authors. When generating a document, an author is chosen at random for each individual word in the document. This author picks a topic from his or her multinomial distribution over topics, and then samples a word from the multinomial distribution over words associated with that topic. This process is repeated for all words in the document.

In the model, the authors produce words from a set of T topics. When T is kept relatively small relative to the number of authors and vocabulary size, the author-topic model applies a form of dimensionality reduction to documents; topics are learned which capture the variability in word choice across a large set of documents and authors. In our simulations, we use 300 topics (see Rosen-Zvi et al. (2004) for an exploration of different numbers of topics).

Figure 1 illustrates the generative process with a graphical model using plate notation. For readers not familiar with plate notation, shaded and unshaded variables indicate observed and latent variables respectively. An arrow indicates a conditional dependency between variables and plates (the boxes in the figure) indicate repeated sampling with the number of repetitions given by the variable in the bottom (see Buntine (1994) for an introduction). In the author-topic model, observed variables not only include the words w in a document but also the set of coauthors A_d on each document d . Currently, the model does not specify the generative process of how authors choose to collaborate. Instead, we assume the model is provided with the authorship information on every document in the collection.

Each author (from a set of K authors) is associated with a multinomial distribution over topics, represented by θ . Each topic is associated with a multinomial distribution over words, represented by ϕ . The multinomial distributions θ and ϕ have a symmetric Dirichlet prior with hyperparameters α and β (see Rosen-Zvi et al. (2004) for details). For each word in the document, we sample an author x uniformly from A_d , then sample a topic z from the multinomial distribution θ associated with author x and sample a word w from a multinomial topic distribution ϕ associated with topic z . This sampling process is repeated N times to form

document d .

2.2 Bayesian Estimation of the Model Parameters

The author-topic model includes two sets of unknown parameters—the K author-topic distributions θ , and the T topic distributions ϕ —as well as the latent variables corresponding to the assignments of individual words to topics z and authors x . The Expectation-Maximization (EM) algorithm is a standard technique for estimating parameters in models with latent variables, finding a mode of the posterior distribution over parameters. However, when applied to probabilistic topic models (Hofmann, 1999), this approach is susceptible to local maxima and computationally inefficient (see Blei, Ng, and Jordan, 2003). We pursue an alternative parameter estimation strategy, outlined by Griffiths and Steyvers (2004), using Gibbs sampling, a Markov chain Monte Carlo algorithm to sample from the posterior distribution over parameters. Instead of estimating the model parameters directly, we evaluate the posterior distribution on just x and z and then use the results to infer θ and ϕ .

For each word, the topic and author assignment are sampled from:

$$P(z_i = j, x_i = k | w_i = m, \mathbf{z}_{-i}, \mathbf{x}_{-i}) \propto \frac{C_{m_j}^{WT} + \beta}{\sum_{m'} C_{m'_j}^{WT} + V\beta} \frac{C_{kj}^{AT} + \alpha}{\sum_{j'} C_{kj'}^{AT} + T\alpha} \quad (1)$$

where $z_i = j$ and $x_i = k$ represent the assignments of the i th word in a document to topic j and author k respectively, $w_i = m$ represents the observation that the i th word is the m th word in the lexicon, and $\mathbf{z}_{-i}, \mathbf{x}_{-i}$ represent all topic and author assignments not including the i th word. Furthermore, $C_{m_j}^{WT}$ is the number of times word m is assigned to topic j , not including the current instance, and C_{kj}^{AT} is the number of times author k is assigned to topic j , not including the current instance, and V is the size of the lexicon.

During parameter estimation, the algorithm only needs to keep track of a $V \times T$ (word by topic) count matrix, and a $K \times T$ (author by topic) count matrix, both of which can be represented efficiently in sparse format. From these count matrices, we can easily estimate the topic-word distributions ϕ and author-topic distributions θ by:

$$\phi_{mj} = \frac{C_{m_j}^{WT} + \beta}{\sum_{m'} C_{m'_j}^{WT} + V\beta} \quad (2)$$

$$\theta_{kj} = \frac{C_{kj}^{AT} + \alpha}{\sum_{j'} C_{kj'}^{AT} + T\alpha} \quad (3)$$

where ϕ_{mj} is the probability of using word m in topic j , and θ_{kj} is the probability of using topic j by author k . These values correspond to the predictive distributions over new words w and new topics z conditioned on w and z .

We start the algorithm by assigning words to random topics and authors (from the set of authors on the document). Each Gibbs sample then constitutes applying Equation (1) to every word token in the document collection. This sampling process is repeated for I iterations. In this paper we primarily focus on results based on a single sample so that specific topics can be identified and interpreted—in tasks involving prediction of words and authors one can average over topics and use multiple samples when doing so (Rosen-Zvi

TOPIC 95		TOPIC 293		TOPIC 29		TOPIC 58	
WORD	PROB.	WORD	PROB.	WORD	PROB.	WORD	PROB.
PATTERNS	0.1965	USER	0.3290	MAGNETIC	0.0155	METHODS	0.5319
PATTERN	0.1821	INTERFACE	0.1378	STARS	0.0145	METHOD	0.1403
MATCHING	0.1375	USERS	0.1060	SOLAR	0.0135	TECHNIQUES	0.0442
MATCH	0.0337	INTERFACES	0.0498	EMISSION	0.0127	DEVELOPED	0.0216
TEXT	0.0242	SYSTEM	0.0434	MASS	0.0125	APPLIED	0.0162
PRESENT	0.0207	INTERACTION	0.0296	OBSERVATIONS	0.0120	BASED	0.0153
MATCHES	0.0167	INTERACTIVE	0.0214	STAR	0.0118	APPROACHES	0.0133
PAPER	0.0126	USABILITY	0.0132	RAY	0.0112	COMPARE	0.0113
SHOW	0.0124	GRAPHICAL	0.0092	GALAXIES	0.0105	PRACTICAL	0.0112
APPROACH	0.0099	PROTOTYPE	0.0086	OBSERVED	0.0098	STANDARD	0.0102
AUTHOR	PROB.	AUTHOR	PROB.	AUTHOR	PROB.	AUTHOR	PROB.
Navarro_G	0.0133	Shneiderman_B	0.0051	Falcke_H	0.0140	Srinivasan_A	0.0018
Amir_A	0.0099	Rauterberg_M	0.0046	Linsky_J	0.0082	Mooney_R	0.0018
Gasieniec_L	0.0062	Harrison_M	0.0025	Butler_R	0.0077	Owren_B	0.0018
Baaza-Yates_R	0.0048	Winiwarter_W	0.0024	Knapp_G	0.0067	Warrow_T	0.0016
Baker_B	0.0042	Ardissano_L	0.0021	Bjorkman_K	0.0065	Fensel_D	0.0016
Arikawa_S	0.0041	Billsus_D	0.0019	Kundu_M	0.0060	Godsill_S	0.0014
Crochemore_M	0.0037	Catardi_T	0.0017	Christensen-D_J	0.0057	Saad_Y	0.0014
Rytter_W	0.0034	St_R	0.0017	Mursula_K	0.0054	Hansen_J	0.0013
Raffinot_M	0.0032	Picard_R	0.0016	Cranmer_S	0.0051	Zhang_Y	0.0013
Ukkonen_E	0.0032	Zukerman_I	0.0016	Nagar_N	0.0050	Dieterich_T	0.0013

TOPIC 52		TOPIC 68		TOPIC 298		TOPIC 139	
WORD	PROB.	WORD	PROB.	WORD	PROB.	WORD	PROB.
DATA	0.1822	PROBABILISTIC	0.0869	RETRIEVAL	0.1208	QUERY	0.1406
MINING	0.0657	BAYESIAN	0.0791	INFORMATION	0.0613	QUERIES	0.0947
DISCOVERY	0.0408	PROBABILITY	0.0740	TEXT	0.0461	DATABASE	0.0932
ATTRIBUTES	0.0343	MODEL	0.0533	DOCUMENTS	0.0385	DATABASES	0.0468
ASSOCIATION	0.0328	MODELS	0.0466	INDEXING	0.0369	DATA	0.0426
LARGE	0.0279	PROBABILITIES	0.0308	DOCUMENT	0.0316	RELATIONAL	0.0384
DATABASES	0.0257	INFERENCE	0.0306	QUERY	0.0261	JOIN	0.0188
KNOWLEDGE	0.0175	CONDITIONAL	0.0274	CONTENT	0.0256	PROCESSING	0.0165
PATTERNS	0.0174	PRIOR	0.0273	SEARCH	0.0174	SOURCES	0.0114
ITEMS	0.0173	POSTERIOR	0.0228	RELEVANCE	0.0171	OPTIMIZATION	0.0110
AUTHOR	PROB.	AUTHOR	PROB.	AUTHOR	PROB.	AUTHOR	PROB.
Han_J	0.0164	Koller_D	0.0104	Oard_D	0.0097	Levy_A	0.0092
Zaki_M	0.0089	Heckerman_D	0.0079	Hawking_D	0.0065	Naughton_J	0.0078
Liu_B	0.0071	Ghahramani_Z	0.0060	Croft_W	0.0057	Suciu_D	0.0075
Cheung_D	0.0066	Friedman_N	0.0060	Jones_K	0.0053	Raschid_L	0.0075
Shim_K	0.0051	Mylymaki_P	0.0057	Schauble_P	0.0052	DeWitt_D	0.0062
Mannila_H	0.0049	Lukasiewicz_T	0.0054	Voorhees_E	0.0050	Widom_J	0.0058
Rastogi_R	0.0049	Geiger_D	0.0045	Callan_J	0.0046	Abiteboul_S	0.0057
Ganti_V	0.0048	Muller_P	0.0044	Fuhr_N	0.0042	Chu_W	0.0055
Toivonen_H	0.0043	Berger_J	0.0044	Smeaton_A	0.0042	Libkin_L	0.0054
Liu_H	0.0043	Xiang_Y	0.0042	Sanderson_M	0.0041	Kriegel_H	0.0054

Figure 2: Eight example topics extracted from the CiteSeer database. Each is illustrated with the 10 most likely words and authors with corresponding probabilities.

et al., 2004).

3. AUTHOR-TOPICS FOR CITESEER

3.1 Learning the Model

Our collection of CiteSeer abstracts contains $D = 162,489$ abstracts with $K = 85,465$ authors. We preprocessed the text by removing all punctuation and common stop words. This led to a vocabulary size of $V = 30,799$, and a total of 11,685,514 word tokens.

There is inevitably some noise in data of this form given that many of the fields (paper title, author names, year, abstract) were extracted automatically by CiteSeer from PDF or postscript or other document formats. We chose the simple convention of identifying authors by their first initial and second name, e.g., A_Einstein, given that multiple first initials or fully spelled first names were only available for a relatively small fraction of papers. This means of course that for some very common names (e.g., J_Wang or J_Smith) there will be multiple actual individuals represented by a single name in the model. This is a known limitation of working with this type of data (e.g., see Newman (2001) for further discussion). There are algorithmic techniques that could be used to automatically resolve these identity problems—

TOPIC 276		TOPIC 158		TOPIC 213		TOPIC 15	
WORD	PROB.	WORD	PROB.	WORD	PROB.	WORD	PROB.
DATA	0.1468	PROBABILISTIC	0.0826	RETRIEVAL	0.1381	QUERY	0.1699
MINING	0.0631	BAYESIAN	0.0751	INFORMATION	0.0600	QUERIES	0.1209
DISCOVERY	0.0396	PROBABILITY	0.0628	INDEX	0.0529	JOIN	0.0258
ATTRIBUTES	0.0392	MODEL	0.0364	INDEXING	0.0469	DATA	0.0212
ASSOCIATION	0.0316	PROBABILITIES	0.0313	QUERY	0.0319	OPTIMIZATION	0.0171
RULES	0.0252	INFERENCE	0.0294	CONTENT	0.0299	PROCESSING	0.0162
PATTERNS	0.0210	MODELS	0.0273	BASED	0.0224	RELATIONAL	0.0131
LARGE	0.0207	CONDITIONAL	0.0262	SEARCH	0.0219	DATABASE	0.0128
ATTRIBUTE	0.0183	DISTRIBUTION	0.0261	RELEVANCE	0.0212	AGGREGATION	0.0117
DATABASES	0.0179	PRIOR	0.0259	SIMILARITY	0.0178	RESULT	0.0106
AUTHOR	PROB.	AUTHOR	PROB.	AUTHOR	PROB.	AUTHOR	PROB.
Han_J	0.0157	Koller_D	0.0109	Oard_D	0.0080	Naughton_J	0.0103
Zaki_M	0.0104	Heckerman_D	0.0079	Voorhees_E	0.0053	Suciu_D	0.0091
Liu_B	0.0080	Friedman_N	0.0076	Hawking_D	0.0053	Levy_A	0.0080
Cheung_D	0.0075	Ghahramani_Z	0.0060	Schauble_P	0.0051	DeWitt_D	0.0077
Hamilton_H	0.0058	Lukasiewicz_T	0.0053	Croft_W	0.0051	Wong_L	0.0071
Mannila_H	0.0056	Mylymaki_P	0.0053	Jones_K	0.0041	Ross_K	0.0067
Brin_S	0.0055	Poole_D	0.0050	Bruza_P	0.0041	Kriegel_H	0.0055
Ganti_V	0.0050	Xiang_Y	0.0048	Lee_D	0.0040	Mumick_I	0.0054
Liu_H	0.0050	vanderGaag_L	0.0047	Smeaton_A	0.0040	Raschid_L	0.0053
Toivonen_H	0.0049	Berger_J	0.0040	Callan_J	0.0039	Kossmann_D	0.0053

Figure 3: The four most similar topics to the topics in the bottom row of Figure 2, obtained from a different Markov chain run.

however, in this paper, we don't pursue these options and instead for simplicity work with the first-initial/last-name representation of individual authors.

In our simulations, the number of topics T was fixed at 300 and the smoothing parameters α and β (Figure 1) were set at 0.16 and 0.01 respectively. We ran 5 independent Gibbs sampling chains for 2000 iterations each. On a 2GHz PC workstation, each iteration took 400 seconds, leading to a total run time on the order of several days per chain.

3.2 Author-Topic and Topic-Word Models for the CiteSeer Database

We now discuss the author-topic and topic-word distributions learned from the CiteSeer data. Figure 2 illustrates eight different topics (out of 300), obtained at the 2000th iteration of a particular Gibbs sampler run.

Each table in Figure 2 shows the 10 words that are most likely to be produced if that topic is activated, and the 10 authors who are most likely to have produced a word if it is known to have come from that topic. The words associated with each topic are quite intuitive and, indeed, quite precise in the sense of conveying a semantic summary of a particular field of research. The authors associated with each topic are also quite representative—note that the top 10 authors associated with a topic by the model are not necessarily the most well-known authors in that area, but rather are the authors who tend to produce the most words for that topic (in the CiteSeer abstracts).

The first 3 topics at the top of Figure 2, topics #163, #87 and #20 show examples of 3 quite specific and precise topics on string matching, human-computer interaction, and astronomy respectively. The bottom four topics (#205, #209, #289, and #10) are examples of topics with direct relevance to data mining—namely data mining itself, probabilistic learning, information retrieval, and database querying and indexing. The model includes several other topics related to data mining, such as predictive modeling and neural networks, as well as topics that span the full range of research areas encompassed by documents in CiteSeer. The full list is available at <http://www.datalab.uci.edu/author-topic>.

Topic #273 (top right Figure 2) provides an example of a topic that is not directly related to a specific research area.

A fraction of topics, perhaps 10 to 20%, are devoted to “non-research-specific” topics, the “glue” that makes up our research papers, including general terminology for describing methods and experiments, funding acknowledgments and parts of addresses (which inadvertently crept in to the abstracts), and so forth.

We found that the topics obtained from different Gibbs sampling runs were quite stable. For example, Figure 3 shows the 4 most similar topics to the topics in the bottom row of Figure 2, but from a different run. There is some variability in terms of ranking of specific words and authors for each topic, and in the exact values of the associated probabilities, but overall the topics match very closely.

4. APPLICATIONS OF THE AUTHOR-TOPIC MODEL TO CITESEER

4.1 Topic Trends over Time

Of the original 162,489 abstracts in our data set, estimated years of publication were provided by CiteSeer for 130,545 of these abstracts. There is a steady (and well-known) increase year by year in the number of online documents through the 1990’s. From 1999 through 2002, however, the number of documents for which the year is known drops off sharply—the years 2001 and 2002 in particular are under-represented in this set. This is due to fact that it is easier for CiteSeer to determine the date of publication of older documents, e.g., by using citations to these documents.

We used the yearly data to analyze trends in topics over time. Using the same 300 topic model described earlier, the documents were partitioned by year, and for each year all of the words were assigned to their most likely topic using the model. The fraction of words assigned to each topic for a given year was then calculated for each of the 300 topics and for each year from 1990 to 2002.

These fractions provide interesting and useful indicators of relative topic popularity in the research literature in recent years. Figure 4 shows the results of plotting several different topics. Each topic is indicated in the legend by the five most probable words in the topic. The top left plot shows a steady increase (roughly three-fold) in machine learning and data mining topics. The top right plot shows a “tale of two topics”: an increase in information-retrieval coupled to an apparent decrease in natural language processing.

On the second row, on the left we see a steady decrease in two “classical” computer science topics, operating systems and programming languages. On the right, however, we see the reverse behavior, namely a corresponding substantial growth in Web-related topics.

In the third row, the left plot illustrates trends within database research: a decrease in the transaction and concurrency-related topic, query-related research holding steady over time, and a slow but steady increase in integration-related database research. The plot on the right in the third row illustrates the changing fortunes of security-related research—a decline in the early 90’s but then a seemingly dramatic upward trend starting around 1995.

The lower left plot on the bottom row illustrates the somewhat noisy trends of three topics that were “hot” in the 1990’s: neural networks exhibits a steady decline since the early 1990’s (as machine learning has moved on to areas such as support vector machines), genetic algorithms appears to

be relatively stable, and wavelets may have peaked in the 1994–98 time period.

Finally, as with any large data set there are always some surprises in store. The final figure on the bottom right shows two somewhat unexpected “topics”. The first topic consists entirely of French words (in fact the model discovered 3 such French language topics). The apparent peaking of French words in the mid-1990s is likely to be an artifact of how CiteSeer preprocesses data rather than any indication of French research productivity. The lower curve corresponds to a topic consisting of largely Greek letters, presumably from more theoretically oriented papers—fans of theory may be somewhat dismayed to see that there is an apparent steady decline in the relative frequency of Greek letters in abstracts since the mid-1990s!

The time-trend results above should be interpreted with some caution. As mentioned earlier, the data for 2001 and 2002 are relatively sparse compared to earlier years. In addition, the numbers are based on a rather skewed sample (online documents obtained by the CiteSeer system for which years are known). Furthermore, the fractions per year only indicate the relative number of words assigned to a topic by the model and make no direct assessment of the quality or importance of a particular sub-area of computer science. Nonetheless, despite these caveats, the results are quite informative and indicate substantial shifts in research topics within the field of computer science.

In terms of related work, Popescul et al. (2000) investigated time trends in CiteSeer documents using a document clustering approach. 31K documents were clustered into 15 clusters based on co-citation information while the text information in the documents was not used. Our author-topic model uses the opposite approach. In effect we use the text information directly to discover topics and do not explicitly model the “author network” (although implicitly the co-author connections are used by the model). A direct quantitative comparison is difficult, but we can say that our model with 300 topics appears to produce much more noticeable and precise time-trends than the 15-cluster model.

4.2 Topics and Authors for New Documents

In many applications, we would like to quickly assess the topic and author assignments for new documents not contained in our subset of the CiteSeer collection. Because our Monte Carlo algorithm requires significant processing time for 160K documents, it would be computationally inefficient to rerun the algorithm for every new document added to the collection (even though from a Bayesian inference viewpoint this is the optimal approach). Our strategy instead is to apply an efficient Monte Carlo algorithm that runs only on the word tokens in the new document, leading quickly to likely assignments of words to authors and topics. We start by assigning words randomly to co-authors and topics. We then sample new assignments of words to topics and authors by applying Equation 1 only to the word tokens in the new document each time temporarily updating the count matrices C^{WT} and C^{AT} . The resulting assignments of words to authors and topics can be saved after a few iterations (10 iterations in our simulations).

Figure 5 shows an example of this type of inference. Abstracts from two authors, B.Scholkopf and A.Darwiche were combined together into 1 “pseudo-abstract” and the document treated as if they had both written it. These two au-

