Predicting Social Networks in Weblogs

Patrick Jähnichen

Natural Language Processing Group
Department of Computer Science
University of Leipzig
jaehnichen@informatik.uni-leipzig.de

Abstract: Weblogs and other platforms used to organize a social life online have achieved an enormous success over the last few years. Opposed to applications directly designed for building up and visualizing social networks, weblogs are comprised of mostly unstructured text data, that comes with some meta data, such as the author of the text, its publication date or the URL it is available under. In this paper, we propose a way, how these networks may be inferred not by the available meta data, but by pure natural language analysis of the text content, allowing inference of these networks without any meta data at hand. We discuss results of first experiments and outline possible enhancements as well as other ways to improve prediction of social networks based solely on content analysis.

1 Introduction

In the advent of Web 2.0, where technologies are centered on user generated content and the presentation of this content, applications such as networking websites or weblogs achieved an enormous popularity among internet users. In weblogs (or blogs), an author gathers information about a certain field of interest[1] in articles (so called blog posts), that are (mostly) ordered in a descending chronological way. In these articles, authors tend to link to other blog posts, which indicates a thematic similarity and pointer to further information about a topic. We assume that a network, built according to these hyperlinks between authors' posts, suggests either strong thematically motivated reference or personal acquaintance between authors (that is, neighboring nodes in a social network) or both, where authors are nodes and hyperlinks between two authors' blog posts are edges between the corresponding nodes. It is also possible to analyze the blogs' text content itself by natural language processing techniques to identify topics that individual blog posts comprise. We presume that the statistic similarity between two authors' specific topic profiles, computed by analyzing their personal blog posts, correlates with the path distance between the same two authors in the aforementioned network, in other words, that these two authors are more likely to be personally acquainted with each other. The contribution of this paper is in showing that the hyperlink structure immanent to weblogs exhibits what is known as the Small-World phenomenon. Further, we apply two different topic models

[1] this may be very divers; authors write about technology, art, music, their private life etc.

to the data and compare author-to-author distances in terms of author specific probability distributions over topic concepts to their respective path distance in the graph formed by weblogs' hyperlink structure and show, that the above mentioned proposition holds true.

2 Related Work

Topic modeling has experienced a lot of attention over the last decade. Based on the ground-breaking work of Blei et. al. in [BNJ03] on Latent Dirichlet Allocation (LDA), many different probabilistic models for a rich variety of applications have been developed (e.g. [PGKT06], [BGBZ07], [MCEW04], [WM06]). The LDA model (cf. Fig. 1 (a)) describes documents as a mixture over T topics and each topic as a multinomial distribution over a vocabulary of V words. That is, each word in each document is assigned a latent variable, representing the virtual concept of a topic this word belongs to. Following the "bag-of-words" assumption[2] and de Finetti's theorem[3] and using machine learning techniques (cf. [GS04]), the model infers the distribution of the latent variables,

$$p\left(z_i = j|\mathbf{z}_{\setminus i}, \mathbf{w}\right) \propto \frac{n^{(w_i)}_{\setminus i,j} + \beta}{n^{(\cdot)}_{\setminus i,j} + V\beta} \frac{n^{(d_i)}_{\setminus i,j} + \alpha}{n^{(d_i)}_{\setminus i,\cdot} + T\alpha},$$

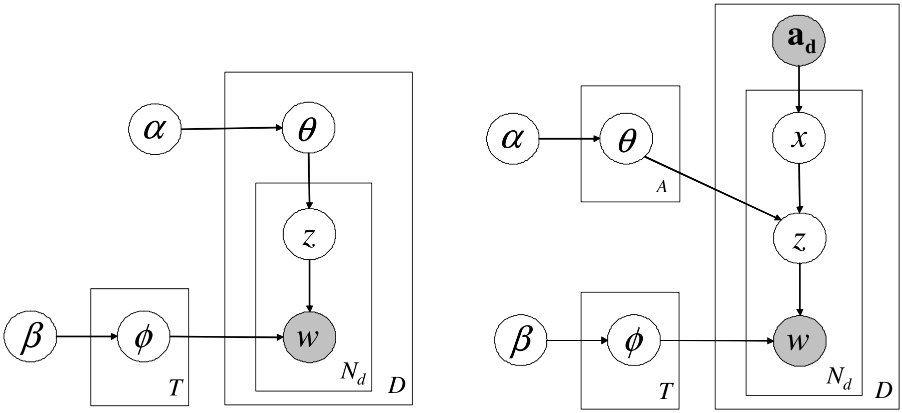
from which the document-specific distributions over topics as well as the topic-specific distributions over words can be derived. Here, $n^{(w_i)}_{\setminus i,j}$ is the number of times, word w_i has been assigned to topic j, $n^{(d_i)}_{\setminus i,j}$ is the number of times, a word from document d_i has been assigned to topic j, both excluding the current assignment of z_i. $(\cdot)$ indicates iteration over the whole parameter space of a variable. The first term of the above equation corresponds to the probability of word w_i in topic j and the second term is the probability of topic j in document d_i. Based on the LDA model, we review the Author-Topic model (AT) introduced by [RZGSS04], that explicitly assigns a probability distribution over topics to each author instead of each document. This is done by decoupling the topic distribution from documents and instead infer a topic distribution for every author. The decision on a topic for a word in a document is then not only based on the document specific topic probability distribution, but on an author, chosen from possible authors of a document, and her specific topic probability distribution (cf. Fig. 1 (b)) In this model, the latent variable distribution (and with it the authors' distributions over topics and the topics' distributions over words) can be estimated by

$$p\left(z_i = j|\mathbf{z}_{\setminus i}, \mathbf{x}_{\setminus i}, \mathbf{w}\right) \propto \frac{C^{WT}_{mj} + \beta}{\sum_{m'} C^{WT}_{m'j} + V\beta} \frac{C^{AT}_{kj} + \alpha}{\sum_{j'} C^{AT}_{kj'} + T\alpha},$$

where C^{WT}_{mj} and C^{AT}_{kj} are the number of times a word m has been assigned to topic j and the number of times an author k is assigned to topic j respectively.

[2] the "bag-of-word" approach assumes that the order of events does not influence the joint probability of all events (cf. [AIJ85])

[3] the de Finetti theorem states that any collection of exchangeable random variables follows a mixture distribution, in general an infinite one (cf. [dFMS75])



(a) LDA model

(b) AT model

Figure 1: Topic models in plate notation, as in [RZGSS04]

Several different techniques for analyzing weblogs and hyperlink structure therein have been proposed throughout the literature that model either the source ([Kle99]) or the flow ([GGLNT04], [AZAL04]) of information in blogspace (i.e. the entirety of web log documents). Concerning the connectivity of blog posts, [KSV06] propose a model in which blog documents may be implicitly linked through sharing a common topic, even if there is no explicit hyperlink between them. This is motivated by the fact, that the authors observe a low per node edge count of only 0.27 (cf. [KSV06], section 3). They create links based on the collected metadata, such as common tags or common authors of blog posts to increase this edge count. As seen in the next section, the data used for our experiments exhibits a much higher node-edge ratio. Additionally, in [KL06] the authors propose a method to link documents by means of content similarity based on the Maximum-Likelihood estimate, also to enrich the connectivity between blog posts. In [GRZW08] the authors go a step further and derive a model that probabilistically introduces links between documents that are not at all based on the available hypertext structure. Here, a link may exist between a word in a document and any other document in the corpus and each link is assigned a topic (in the topic modeling sense). In fact, all previous models consider a document-to-document linking structure, whereas our approach works on a higher level of abstraction, i.e. author-to-author linking structure based on the authors' documents.

3 Finding Social Networks in Weblogs

Given a corpus comprised of blog posts that come with minimal meta data, i.e. the author of the post and the original URI it was available at, the aim is, to find out if the different authors of blog posts are in some way interconnected. If so, we want to examine, if these connections form a network, that follows the same principles as a social network does.

3.1 The spinn3r.com ICSWM09 Data Set

The spinn3r.com ICSWM09 data set [BJS09] has been provided by spinn3r.com, a company focussing on indexing blog websites and providing an interface to access them. The data set consists of a two months crawl (Aug 1 - Oct 1 2008) of randomly selected weblogs and comprises 127 GB of uncompressed data. The whole timespan has been split into nine weeks to reduce computational complexity. The blogs are divided into different tier groups that are determined using the *tailrank* algorithm[4] and only tier group one has been used, still consisting of about 36 GB of data. As the majority of the data is in English, we restricted ourselves to that language, resulting in a total of 5.3 million blog posts. After that, all hyperlinks in the content were extracted and stored, then all HTML tags were removed from the text content, as well as stop words and punctuation. Each post has been saved together with its meta data, i.e. the author, timestamp, original URL and the links contained in the content.

3.2 Hyperlink Graph in Weblogs

One characteristic used to determine social network behavior of graphs is the characteristic path length L of a graph. It can be determined by building the median of the means of the shortest path lengths between all vertex pairings. The second characteristic property of social network graphs is the cluster coefficient γ of a graph. It is the average over all nodes' cluster coefficients, which are defined as $\gamma_v = \frac{|E(G)|}{\binom{k_v}{2}}$, where $E(G)$ is the edge list of the graph G and k_v is the degree[5] of vertex v. Thus, $\binom{k_v}{2}$ is the maximum number of edges between v and its neighboring nodes, which makes the cluster coefficient a fraction of actually existing over all possible edges. In a social network context, this is often described as a measure on how probable it is, that neighbors of v also know each other. As Watts states in [Wat99], social networks exhibit a similar characteristic path length as random graphs of the same size and order, but a significantly higher cluster coefficient. These are also known as a small-world graphs. All documents have been stored together with the URLs contained in their text content and the URL under which they were originally available. Now, if a text- contained URL matches the URL of another document, this means that the author of the first (taken to be author A) has linked in one of her blog posts to another document, of which the author is known (taken to be author B). As the overall goal is, to predict social networks by analyzing text content similarity,

[4] in tailrank, older posts gain a lower score than newer ones, same as less popular gain a lower score than more popular ones (popularity in terms of being cited by others); the lower the tier group, the more successful a blog is in placing its posts on the list of top stories with high tailrank score

[5] the degree of a vertex is equal to the size of its neighborhood Γ_v, which is the set of vertices, v is connected to via an edge

consider that, if A links to another document in one of her posts, it is highly likely that the other document's content is statistically similar to the content of her own blog post. Additionally, A has to know the blog post, she has linked to. This is only possible, if (a) A knows B and regularly reads Bs blog posts or (b) another author (author C) that A is acquainted[6] to, is also acquainted to B, giving A the possibility to come across B's post by reading C's posts and following the link there. The second possibility might also be extended to a chain of arbitrary length, although the longer this chain, the lesser the probability of its existence. To build up the network graph, we applied the following steps to each document in a week segment:

1. determine author and hyperlinks contained in the text content of the document,
2. compare the hyperlinks to a list of links to other documents,
3. if a text-contained link in a document matches the unique link of another document and given that the matched document belongs to the same week segment,
   (a) add both documents' authors (A and B) to $V(G)$, such that $V(G) = V(G) \cup \{A\} \Leftrightarrow A \notin V(G)$ and $V(G) = V(G) \cup \{B\} \Leftrightarrow B \notin V(G)$,
   (b) add an edge (A, B) to $E(G)$, such that $E(G) = E(G) \cup \{(A, B)\} \Leftrightarrow (A, B) \notin E(G) \wedge (B, A) \notin E(G)$,

where $V(G)$ is the list of vertices of a graph G.

The resulting networks are described in Tab. 1. Here, the largest possible graph (maximal values) and the largest connected component in the found graph are characterized in terms of their size, order and their fraction of the maximum graph. Additionally, a visualization of the fifth week's hyperlink graph, centered around the node with maximum degree, is shown in Fig. 2.

	maximal values		largest connected component				
week	order	size	order	%	size	$\%\cdot 10^{-4}$	edges per node
1	87831	$3.9 \cdot 10^9$	3830	4.36	5368	1.37	1.4
2	104440	$5.45 \cdot 10^9$	5390	5.16	7785	1.42	1.44
3	102027	$5.2 \cdot 10^9$	5129	5.03	7371	1.41	1.44
4	101315	$5.13 \cdot 10^9$	5361	5.29	7684	1.49	1.43
5	99786	$4.97 \cdot 10^9$	6383	6.4	9554	1.92	1.5
6	109155	$5.95 \cdot 10^9$	6041	5.53	8945	1.5	1.48
7	107841	$5.81 \cdot 10^9$	5851	5.43	8632	1.48	1.48
8	112153	$6.28 \cdot 10^9$	5965	5.32	8896	1.42	1.49
9	82846	$3.43 \cdot 10^9$	4080	4.92	5533	1.61	1.36

Table 1: Comparison of maximal possible and largest connected component of found networks in the data

[6] being acquainted or to know each other is used interchangeably to represent the fact that an author links to a document of another author in one of her blog posts

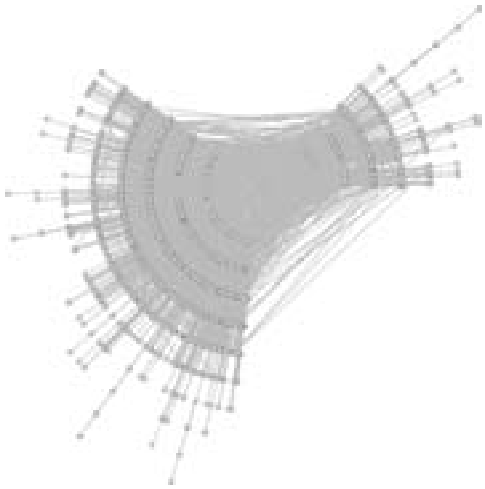


Figure 2: A visualization of the extracted hyperlink graph

As has been suggested in section 1, we observe a high node-edge ratio in the largest connected component of each data segment's hyperlink graph (compared to the findings of [KSV06]. Looking at the characteristics in Tab. 2, we can show that the found network graphs exhibit the same properties as small-world graphs (i.e. their characteristic path lengths are similar, but the cluster coefficient of the found networks is considerably higher than in random graphs of same size and order) and can thus be treated as representations of actual social networks latent to the underlying weblog data. This also gives an excellent evaluation measure for comparing distances in a probabilistically derived author network to their counterparts in the inherent hyperlink structure of weblogs.

	network graph			random graph		
week	L	γ	D	L	γ	D
1	6.3	0.092	17	7.8	$3.2 \cdot 10^{-4}$	26
2	6.2	0.11	7.9	18	$4.7 \cdot 10^{-4}$	30
3	6.15	0.099	21	7.99	$5.9 \cdot 10^{-4}$	30
4	6.15	0.115	22	8.1	$5.3 \cdot 10^{-4}$	30
5	5.35	0.113	18	7.9	$3.1 \cdot 10^{-4}$	23
6	5.6	0.107	18	7.94	$3.2 \cdot 10^{-4}$	23
7	5.84	0.099	20	7.94	$3.5 \cdot 10^{-4}$	26
8	5.76	0.098	20	7.86	$3.2 \cdot 10^{-4}$	21
9	6.29	0.104	19	8.02	$3.2 \cdot 10^{-4}$	25

Table 2: Average path lengths, cluster coefficients and diameters of networks extracted from the data set and corresponding random networks.

Opposed to [KSV06], [KL06] and [GRZW08], we do not focus on predicting links between documents directly. Instead, we analyze author-specific content with probabilistically driven topic models and arrive at a topic probability distribution for each author. The distances between authors are then compared to the corresponding network path distances in the network graph described above. That is, we determine correlations between author distances in probabilistic models and network distances in the hyperlink graph.

4 Experiments

We trained both the LDA and AT model on the data of the fifth week segment with a Gibbs sampler run for 2000 iterations. The fifth segment has been chosen, because its largest connected component contains the largest fraction of authors and edges (cf. Tab. 1) as well as the highest node-edge ratio of all data segments. Since running a Gibbs sampler[7] with the data of one week already takes almost four days of computing time, we restricted ourselves to the data promising the best results (i.e. the fifth data segment).

By applying the LDA model, we determined a topic probability distribution for each document. To generate a topic probability distribution for authors, the probability distributions of all documents of an author are averaged. As the AT model arrives directly at a specific topic probability distribution for each author, no averaging has to be done.

After that, we computed the distances between all author specific probability distributions and averaged found path distances corresponding to author pairings having a similar probability distribution distance. Finally, we compared both the similarity between the generated topic probability distributions of authors and the actual path length in the social network. Following [Lee01], we used the skew divergence (with $\alpha = 0.99$), a Kullback-Leibler (KL) based distance metric for probability distributions to measure the distance between two probability distributions. It is defined as

$$s_\alpha(P, Q) = D_{KL}(Q \| \alpha P + (1 - \alpha) Q)$$

where

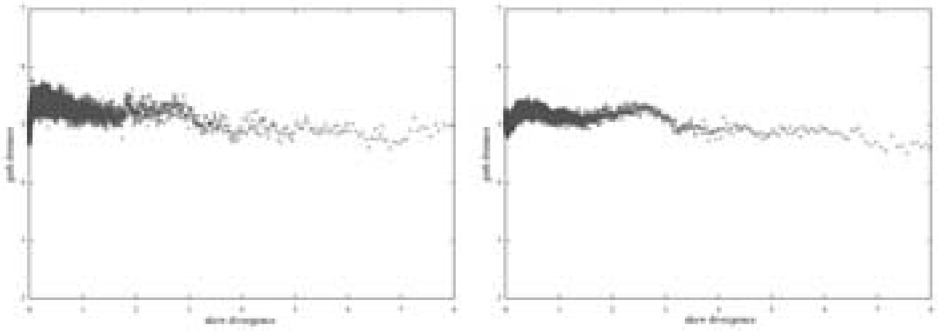
$$D_{KL}(P \| Q) = \sum_i P(i) \, log \left(\frac{P(i)}{Q(i)} \right)$$

is the Kullback-Leibler distance.

Using the LDA model, we encounter an increasing average path length between authors in the graph as the averaged author specific topic probability distribution distance of such two authors rises, i.e. the topics appearing in their blog posts are less similar. Also, the path distance stagnates at around average path length of the underlying network.

Looking at the correlation between KL-divergence based measures and the path distance in the network in the AT model it can be seen, that for low divergences path distance increases and also stagnates at around average path length. The fact that with increasing similarity measure values (and hence with less similarity), path distances in the social network grow is shown even better than in the LDA model. Interestingly, the path distance stabilizes at a higher value for the LDA model than for the AT model (cf. Fig. 4), which might be caused

[7] we used a reimplementation of the parallelized Gibbs sampler described in [WBS$^+$09]



(a) LDA model

(b) AT model

Figure 3: Skew divergence against network path length

by the process of simply averaging over all documents of an author instead of directly using author topic probabilities provided by the AT model.

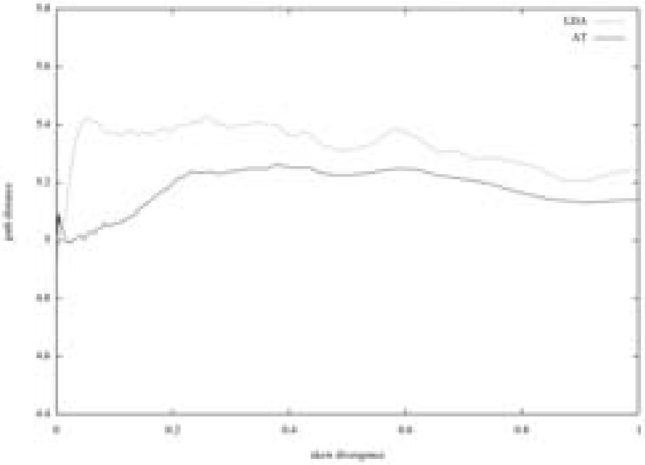


Figure 4: Skew diversion against network path length - all models

5 Conclusion

In this paper we have shown, that hyperlink structures in weblogs indeed exhibit the Small-World phenomenon and might thus be treated as social networks formed between blog posts' authors. We have introduced two methods to infer latent topics from analyzing natural language and applied these to arbitrarily chosen weblog documents. Further, we described some methods to analyze weblogs and how they differ from our approach.

We have also shown, that the ”difference” between authors in the context of topic models correlates with network path distance between authors and could possibly be used to infer, if not identical, at least similar networks that also exhibit characteristics typical to social networks. Distance measures show an expected correlation between their values and path distances, where a lower author similarity (and thus fewer shared interests) result in a higher distance in the social network. As the improvement from the LDA to AT model suggests, further enhancements might be given by more sophisticated generative probabilistic models, one of which, [Jäh09], adds one extra level of abstraction and tries to infer community specific probability distributions over authors, i.e. local author clusters in a social network, directly. Additionally, the used topic models have a fixed number of topics as a parameter to be chosen by hand. In [TJBB06], the authors describe a nonparametric bayesian approach to arrive at an optimal number of topics, fitting the data best. This might also be a valuable enhancement to the proposed approach.

Bibliography

[AIJ85] David Aldous, Illdar Ibragimov, and Jean Jacod. Exchangeability and related topics. volume 1117, pages 1–198. 1985.

[AZAL04] Eytan Adar, Li Zhang, Lada A Adamic, and RM Lukose. Implicit structure and the dynamics of blogspace. *Workshop on the Weblogging Ecosystem*, 2004.

[BGBZ07] Jordan Boyd-Graber, David M Blei, and Xiaojin Zhu. A topic model for word sense disambiguation. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1024–1033, 2007.

[BJS09] Kevin Burton, Akshay Java, and Ian Soboroff. The ICWSM 2009 Spinn3r Dataset, May 2009.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, Mar 2003.

[dFMS75] Bruno de Finetti, Antonio Machí, and Adrian Smith. *Theory of probability: a critical introductory treatment*. 1975.

[GGLNT04] Daniel Gruhl, Ratan Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. *Proceedings of the 13th international conference on World Wide Web*, pages 491–501, 2004.

[GRZW08] Amit Gruber, Michal Rosen-Zvi, and Yair Weiss. Latent topic models for hypertext. *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence*, Jan 2008.

[GS04] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl 1):5228–5235, 2004.

[Jäh09] Patrick Jähnichen. Finding and Analyzing Social Networks in unstructured web log data using probabilistic topic modeling. Master’s thesis, 2009.

[KL06] Oren Kurland and Lillian Lee. PageRank without hyperlinks: Structural re-ranking using links induced by language models. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05:306—313, Jan 2006.

[Kle99] Jon Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, Jan 1999.

[KSV06] Apostolos Kritikopoulos, Martha Sideri, and Iraklis Varlamis. BlogRank: ranking weblogs based on connectivity and similarity features. *Proceedings of the 2nd international workshop on Advanced architectures and algorithms for internet delivery and applications*, AAA-IDEA '06, 2006.

[Lee01] Lillian Lee. On the effectiveness of the skew divergence for statistical language analysis. *Artificial Intelligence and Statistics*, pages 65–72, 2001.

[MCEW04] Andrew McCallum, Andrés Corrada-Emmanuel, and Xuerui Wang. The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email. *NIPS'04 Workshop on Structured Data and Representations in Probabilistic Models for Categorization*, 2004.

[PGKT06] Matthew Purver, Thomas L Griffiths, KP Körding, and Joshua B Tenenbaum. Unsupervised topic modelling for multi-party spoken discourse. pages 17–24, 2006.

[RZGSS04] Michal Rosen-Zvi, Thomas L Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. *Proceedings of the 20th conference on Uncertainty in Artificial Intelligence*, Jan 2004.

[TJBB06] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, Jan 2006.

[Wat99] Duncan J Watts. *Small worlds: The Dynamics of Networks between Order and Randomness*. 1999.

[WBS$^+$09] Yi Wang, H Bai, M Stanton, W.Y Chen, and E Chang. Plda: Parallel latent dirichlet allocation for large-scale applications. *Algorithmic Aspects in Information and Management*, pages 301–314, 2009.

[WM06] Xuerui Wang and Andrew McCallum. Topics over time: a non-Markov continuous-time model of topical trends. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433, 2006.