Community-based Link Prediction with Text

David Mimno, Hanna Wallach, Andrew McCallum
Department of Computer Science
University of Massachusetts, Amherst
Amherst, MA
{mimno,wallach,mccallum}@cs.umass.edu

1. INTRODUCTION

There has been much recent interest in generative models for graphs. The intuition behind the study of such link prediction functions is that they provide a succinct description of the process by which networks grow and evolve: a model that accurately predicts small-scale actions such as coauthorships should help us understand the global properties of the network.

Previous work in social network analysis, such as Liben-Nowell and Kleinberg [5], has often focused on generative models that take into account only the graph structure of the network, without making any use of the individual properties of the nodes themselves. Frequently, however, much richer data is available than the link structure alone, such as text documents for coauthorship networks.

In this paper, we propose a generative model for documents that produces both text and authors based on a notion of communities, which each have a distribution over authors and over topics. We demonstrate this model on the proceedings of the NIPS conference, showing improved likelihood of held-out coauthorship data. Discovering latent structure can also be useful in analyzing long term trends, such as the growth and fragmentation of communities.

2. A COMMUNITY-BASED MODEL FOR COAUTHORSHIP

The proposed community-based generative model is in the same family as Latent Dirichlet Allocation (LDA), in which each word in each document is generated by one of set of topic distributions ϕ_t drawn from a Dirichlet prior $\mathcal{D}(\beta)$. This *Community-Author-Topic* (CAT) model generates documents by first selecting a community. A community consists of a distribution over topics θ_c and a distribution over authors ψ_c, drawn from Dirichlet distributions $\mathcal{D}(\alpha)$ and $\mathcal{D}(\zeta)$. Rather than specifying a fixed number of communities, we allow the number of communities to vary, according to a Chinese Restaurant Process prior [2]. Given the community c, a document is generated by first sampling authors from ψ_c and then, for each word, sampling a topic t from θ_c and a word from ϕ_t.

We use a blocked Gibbs sampler to train the model, alternating between sampling topic variables with communities fixed and sampling communities with topics fixed. The sampling distribution for a topic given all other topic and community assignments is

$$p(z_{di}|\boldsymbol{z}_{\setminus di}, \boldsymbol{w}, c_d) \quad \propto \quad (n_{ct} + \alpha_t)\frac{(n_{tw_i} + \beta_w)}{(n_{t\bullet} + \beta_\bullet)} \qquad (1)$$

where n_{ct} is the number of word tokens assigned to topic t in documents assigned to the current community. The sampling distribution over communities for a document given the authors and topic assignments in the document is

$$\begin{aligned} p(c|\boldsymbol{a}, \boldsymbol{z}, \boldsymbol{c}_{\setminus d}) \quad \propto \quad & p(c)\frac{\Gamma(\alpha_\bullet + n_{c\bullet})}{\prod_t \Gamma(\alpha_t + n_{ct})}\frac{\prod_t \Gamma(\alpha_t + n_{ct} + n_{dt})}{\Gamma(\alpha_\bullet + n_{c\bullet} + n_{d\bullet})} \qquad (2) \\ & \times \frac{\Gamma(\zeta_\bullet + n_{c\bullet})}{\prod_a \Gamma(\zeta_a + n_{ca})}\frac{\prod_a \Gamma(\zeta_a + n_{ca} + n_{da})}{\Gamma(\zeta_\bullet + n_{c\bullet} + n_{d\bullet})} \end{aligned}$$

Note that the authors are conditionally independent of the topics given the community. The prior over communities $p(c)$ is $\frac{n_c}{n_\bullet + \gamma}$ for an existing community and $\frac{\gamma}{n_\bullet + \gamma}$ for a new, empty community, where n_c is the number of documents assigned to community c and γ is a concentration parameter, which we set to 1. We sample the topics for 500 iterations before sampling cluster assignments for 200 iterations in each block. Plotting the log likelihood shows that this is sufficient for the model to come to a stable point.

Most previous topic models that have used author data [6, 8] have taken authors as observed, and have not explicitly specified a generative model for them. Newman et al. [7] present several models that generate named entities. The difference between these models and the CAT model is that our model explicitly includes a notion of clustering within the corpus, such that documents within a community share a topic distribution and an author distribution. Zhou et al. [9] present a model for generating communites that is equivalent to LDA in which each author's list of coauthors is treated as a document. Unlike CAT, this model does not generate text data, only a coauthorship network. There are also other community based link-only models, such as the models presented by Lescovec, Kleinberg and Faloutsos [4] and Goldenberg and Zheng [3].

3. EVALUATION

In addition to the full CAT model with both author and text data, we trained a similar community model using only author data and another model using only text data.

In order to compare link prediction models, we train each model on author and full text data from NIPS 1987 to 2003.[1] We then calculate the likelihood of the coauthorships from NIPS 2004–6 under each model. Given community assignments $\boldsymbol{c}$ and community-author counts n_{ca} we compute the

[1]This data was provided by Sam Roweis and Gal Chechik

Table 1: Topic distributions for the community with the largest number of papers by *Jordan_M* in three models, one trained only on authors, one only on topics, and one on both. The author-based model clusters Jordan and his coauthors, while the topic-based models distinguish between different areas of research.

authors only

tokens	topic words
4351	latent, variational, model, parameters
4221	number, algorithm, results, method
4137	theorem, proof, case, result
3827	mixture, density, gaussian, likelihood

Jordan_M (27), Gharamani_Z (18), Tenenbaum_J (10), Jaakkola_T (10), Griffiths_T_L (9)

topics only

tokens	topic words
15218	field, approximation, variational, distribution
13668	propagation, belief, inference, bp
7664	variables, inference, network, distribution
3737	number, algorithm, results, method

Jordan_M (11), Jaakkola_T (6), Saul_L (5), Kappen_H (5), Wainwright_M (4)

authors and topics

tokens	topic words
14903	propagation, belief, inference, bp
14282	field, approximation, variational, distribution
3244	theorem, proof, case, result
3109	bound, bounds, log, error

Jordan_M (11), Willsky_A (6), Jaakkola_T (6), Frey_B (5), Saul_L (5)

probability of any pair of authors as

$$p_c(x,y) = \sum_c \frac{n_{cx}+\zeta_x}{n_{c\bullet}+\zeta_\bullet} \frac{n_{cy}+\zeta_y}{n_{c\bullet}+\zeta_\bullet+1}. \quad (3)$$

This expression is equivalent to the probability of any community's author distribution emitting both x and y. For the Dirichlet prior we set $\zeta_a = 100 n_{\bullet a}/n_{\bullet\bullet}$.

We compare the community models to two baseline models. FIrst, one of the simplest generative models for social networks is preferential attachment [5]. Under this model, the probability of a link between two nodes x and y, $p_{pa}(x,y) \propto |\mathcal{C}_x||\mathcal{C}_y|$, where $\mathcal{C}_x$ is the set of coauthors of x.

Second, the Adamic-Adar heuristic [1] weights attributes of nodes by their frequency within the corpus. Commonly occurring attributes have lower weight than less frequent attributes. Following Liben-Nowell and Kleinberg, we define the attributes of a given pair of nodes as the intersection of the sets of coauthors for the two nodes.

$$AA(x,y) = \sum_{z \in \mathcal{C}_x \cap \mathcal{C}_y} \frac{1}{\log(|\mathcal{C}_z|)} \quad (4)$$

This function is zero when two authors share no coauthors, so we create a smoothed distribution by interpolating between the normalized Adamic-Adar score and the preferential attachment model. We find that this model achieves maximum likelihood at $\lambda \approx 0.64$.

4. RESULTS

Table 2 shows topic and author distributions for three communities trained with different data: one on authors, one on topics, and one on both. The first two methods correspond to eliminating either the second or third terms of the community sampling distribution in Equation 2. Communities trained solely on authors have relatively flat topic distributions but sharp author distributions. In communities trained with topic information, on the other hand, individual prolific authors may be spread through more communities reflecting the fact that they might write about different topics with different people. Log likelihood results for coauthorships in NIPS 2004–6 are shown in Figure 1. The two topic-based models outperform the author-based models, by a wide margin in the case of the Adamic-Adar and preferential attachment models. The model trained with just topic information has the best performance, but higher variance than the model trained with both author and topic information. This result suggests that taking node attributes such as text data into account can improve the quality of predictive models of network growth, and thus improve our understanding of the dynamics of social networks. In this case, topic information may help in the case where A writes with B and C, but on different topics. A model that looks only at common coauthors may predict a coauthorship between B and C, while a topic-based model would not.

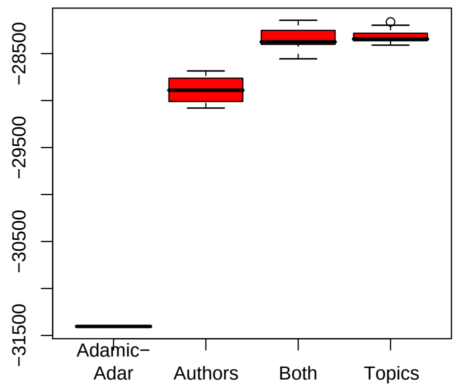

Figure 1: Log likelihood of NIPS 2004–6 coauthorships under each model. Preferential attachment (not shown) is much worse, at -40121.

5. REFERENCES

[1] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 23(3):211–230, July 2003.

[2] D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In *NIPS*, 2004.

[3] A. Goldenberg and A. X. Zheng. Exploratory study of a new model of evolving networks. In *ICML workshop on statistical network analysis*, 2006.

[4] J. Lescovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD*, 2005.

[5] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM*, 2003.

[6] A. McCallum, A. Corrada-Emmanuel, and X. Wang. Topic and role discovery in social networks. In *IJCAI*, 2005.

[7] D. Newman, C. Chemudugunta, P. Smyth, and M. Steyvers. Statistical entity-topic models. In *KDD*, 2006.

[8] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI*, 2004.

[9] D. Zhou, E. Manavoglu, J. Li, C. L. Giles, and H. Zha. Probabilistic models for discovering e-communities. In *WWW*, 2006.