A Utility Model of Authors in the Scientific Community

Yanchuan Sim
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
ysim@cs.cmu.edu

Bryan R. Routledge
Tepper School of Business
Carnegie Mellon University
Pittsburgh, PA 15213, USA
routledge@cmu.edu

Noah A. Smith
Computer Science & Engineering
University of Washington
Seattle, WA 98195, USA
nasmith@cs.washington.edu

Abstract

Authoring a scientific paper is a complex process involving many decisions. We introduce a probabilistic model of some of the important aspects of that process: that authors have individual preferences, that writing a paper requires trading off among the preferences of authors as well as extrinsic rewards in the form of community response to their papers, that preferences (of individuals and the community) and tradeoffs vary over time. Variants of our model lead to improved predictive accuracy of citations given texts and texts given authors. Further, our model's posterior suggests an interesting relationship between seniority and author choices.

1 Introduction

Why do we write? As researchers, we write papers to report new scientific findings, but this is not the whole story. Authoring a paper involves a huge amount of decision-making that may be influenced by factors such as institutional incentives, attention-seeking, and pleasure derived from research on topics that excite us.

We propose that text collections and associated metadata can be analyzed to reveal optimizing behavior by authors. Specifically, we consider the ACL Anthology Network Corpus (Radev et al., 2013), along with author and citation metadata. Our main contribution is a method that infers two kinds of quantities about an author: her associations with interpretable research topics, which might correspond to relative expertise or merely to preferences among topics to write about; and a tradeoff coefficient that estimates the extent to which she writes papers that will be cited versus papers close to her preferences.

The method is based on a probabilistic model that incorporates assumptions about how authors decide what to write, how joint decisions work when papers are coauthored, and how individual and community preferences shift over time. Central to our model is a low-dimensional topic representation shared by authors (in defining preferences), papers (i.e., what they are "about"), and the community as a whole (in responding with citations). This method can be used to make predictions; empirically, we find that:

1. topics discovered by generative models outperform a strong text regression baseline (Yogatama et al., 2011) for citation count prediction;
2. such models do better at that task *without* modeling author utility as we propose; and
3. the author utility model leads to better predictive accuracy when answering the question, "given a set of authors, what are they likely to write?"

This method can also be used for exploration and to generate hypotheses. We provide an intriguing example relating author tradeoffs to age within the research community.

2 Notation and Representations

In the following, a document d will be represented by a vector $\boldsymbol{\theta}_d \in \mathbb{R}^K$. The dimensions of this vector might correspond to elements of a vocabulary, giving a "bag of words" encoding; in this work they correspond to latent topics.

Document d is assumed to elicit from the scientific community an observable response y_d, which might correspond to the number of citations (or downloads) of the paper.

Each author a is associated with a vector $\boldsymbol{\eta}_a \in \mathbb{R}^K$, with dimensions indexed the same as documents. Below, we will refer to this vector as a's "preferences," though it is important to remember that they could also capture an author's *expertise*,

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1510–1519,
Lisbon, Portugal, 17-21 September 2015.

and the model makes no attempt to distinguish between them. We use "preferences" because it is a weaker theoretical commitment.

3 Author Utility Model

We describe the components of our model—author utility (§3.1), coauthorship (§3.2), topics (§3.3), and temporal dynamics (§3.4)—then give the full form in §3.5.

3.1 Modeling Utility

Our main assumption about author a is that she is an optimizer: when writing document d she seeks to increase the response y_d while keeping the contents of d, $\boldsymbol{\theta}_d$, "close" to her preferences $\boldsymbol{\eta}_a$. We encode her objectives as a utility function to be maximized with respect to $\boldsymbol{\theta}_d$:

$$U(\boldsymbol{\theta}_d) = \kappa_a y_d - \frac{1}{2}\|\boldsymbol{\theta}_d - (\boldsymbol{\eta}_a + \boldsymbol{\epsilon}_{d,a})\|_2^2 \quad (1)$$

where $\boldsymbol{\epsilon}_{d,a}$ is an author-paper-specific idiosyncratic randomness that is unobserved to us but assumed known to the author. (This is a common assumption in discrete choice models. It is often called a "random utility model.")

Notice the tradeoff between maximizing the response y_d and staying close to one's preferences. We capture these competing objectives by formulating the latter as a squared Euclidean distance between $\boldsymbol{\eta}_a$ and $\boldsymbol{\theta}_d$, and encoding the tradeoff between extrinsic (citation-seeking) and intrinsic (preference-satisfying) objectives as the (positive) coefficient κ_a. If κ_a is large, a might be understood as a citation-maximizing agent; if κ_a is small, a might appear to care much more about certain kinds of papers ($\boldsymbol{\eta}_a$) than about citation.

This utility function considers only two particular facets of author writing behavior; it does not take into account other factors that may contribute to an author's objective. For this reason, some care is required in interpreting quantities like κ_a. For example, divergence between a particular $\boldsymbol{\eta}_a$ and $\boldsymbol{\theta}_d$ might suggest that a is open to new topics, not merely hungry for citations. Other motivations, such as reputation (notoriously difficult to measure), funding maintenance, and the preferences of peer referees are not captured in this model. Similarly for preferences $\boldsymbol{\eta}_a$, a large value in this vector might reflect a's skill or the preferences of a's sponsors rather than a's personal interest the topic.

Next, we model the response y_d. We assume that responses are driven largely by topics, with some noise, so that

$$y_d = \boldsymbol{\beta}^\top \boldsymbol{\theta}_d + \xi_d \quad (2)$$

where $\xi_d \sim \mathcal{N}(0, 1)$. Because the community's interest in different topics varies over time, $\boldsymbol{\beta}$ is given temporal dynamics, discussed in §3.4.

Under this assumption, the author's *expected* utility assuming she is aware of $\boldsymbol{\beta}$ (often called "rational expectations" in discrete choice models), is:

$$\mathbb{E}[U(\boldsymbol{\theta}_d)] = \kappa_a \boldsymbol{\beta}^\top \boldsymbol{\theta}_d - \frac{1}{2}\|\boldsymbol{\theta}_d - (\boldsymbol{\eta}_a + \boldsymbol{\epsilon}_{d,a})\|_2^2 \quad (3)$$

(This is obtained by plugging the expected value of y_d, from Eq. 2, into Eq. 1.)

An author's decision will therefore be

$$\hat{\boldsymbol{\theta}}_d = \arg\max_{\boldsymbol{\theta}} \kappa_a \boldsymbol{\beta}^\top \boldsymbol{\theta} - \frac{1}{2}\|\boldsymbol{\theta} - (\boldsymbol{\eta}_a + \boldsymbol{\epsilon}_{d,a})\|_2^2 \quad (4)$$

Optimality implies that $\hat{\boldsymbol{\theta}}_d$ solves the first-order equations

$$\kappa_a \beta_j - (\hat{\theta}_{d,j} - (\eta_{a,j} + \epsilon_{d,a,j})) = 0, \ \forall 1 \le j \le K \quad (5)$$

Eq. 5 highlights the tradeoff the author faces: when $\beta_j > 0$, the author will write more on $\theta_{d,j}$, while straying too far from $\eta_{a,j}$ incurs a penalty.

3.2 Modeling Coauthorship

Matters become more complicated when multiple authors write a paper together. Suppose the document d is authored by set of authors $\boldsymbol{a}_d$. We model the joint expected utility of $\boldsymbol{a}_d$ in writing $\boldsymbol{\theta}_d$ as the average of the group's utility.[1]

$$\mathbb{E}[U(\boldsymbol{\theta}_d)] = \frac{1}{|\boldsymbol{a}_d|}\sum_{a \in \boldsymbol{a}_d}\Big(\kappa_a \boldsymbol{\beta}^\top \boldsymbol{\theta}_d - \frac{1}{2} c_{d,a}\|\boldsymbol{\theta}_d - (\boldsymbol{\eta}_a + \boldsymbol{\epsilon}_{d,a})\|_2^2\Big) \quad (6)$$

where the "cost" term is scaled by $c_{d,a}$, denoting the fractional "contribution" of author a to document d. Thus, $\sum_{a \in \boldsymbol{a}_d} c_{d,a} = 1$, and we treat $\boldsymbol{c}_d$ as a latent categorical distribution to be inferred. The first-order equation becomes

$$\sum_{a \in \boldsymbol{a}_d} \kappa_a \boldsymbol{\beta} - c_{d,a}(\boldsymbol{\theta}_d - (\boldsymbol{\eta}_a + \boldsymbol{\epsilon}_{d,a})) = \mathbf{0} \quad (7)$$

[1] This assumption is a convenient starting place, but we can imagine revisiting it in future work. For example, an economist and a linguist with different expertise might derive "utility" from the collaboration that is non-linear in each one's individual preferences (Anderson, 2012). Further, contributions by complementary authors are not expected to be independent of each other.

3.3 Modeling Document Content

As noted before, there are many possible ways to represent and model document content $\boldsymbol{\theta}_d$. We treat $\boldsymbol{\theta}_d$ as (an encoding of) a mixture of topics. Following considerable past work, a "topic" is defined as a categorical distribution over observable tokens (Blei et al., 2003; Hofmann, 1999). Let $\boldsymbol{w}_d$ be the observed bag of tokens constituting document d. We assume each token is drawn from a mixture over topics:

$$p(\boldsymbol{w}_d \mid \boldsymbol{\theta}_d) = \sum_{\boldsymbol{z}_d} \prod_{i=1}^{N_d} p(z_{d,i} \mid \boldsymbol{\theta}_d) p(w_{d,i} \mid \boldsymbol{\phi}_{z_{d,i}})$$

where N_d is the number of tokens in document d, $z_{d,i}$ is the topic assignment for d's ith token $w_{d,i}$, and $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K$ are topic-term distributions. Note that $\boldsymbol{\theta}_d \in \mathbb{R}^K$; we define $p(z \mid \boldsymbol{\theta}_d)$ as a categorical draw from the softmax-transformed $\boldsymbol{\theta}_d$ (Blei and Lafferty, 2007).

Using topic mixtures instead of a bag of words provides us with a low-dimensional interpretable representation that is useful for analyzing authors' behaviors and preferences. Each dimension j of an author's preference is grounded in topic j. If we ignore document responses, this component of model closely resembles the author-topic model (Rosen-Zvi et al., 2004), except that we assume a different prior for the topic mixtures.

3.4 Modeling Temporal Dynamics

Individual preferences shift over time, as do those of the research community. We extend our model to allow variation at different timesteps. Let $t \in \langle 1, \ldots, T \rangle$ index timesteps (in our experiments, each t is a calendar year). We let $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{\eta}_a^{(t)}$, and $\kappa_a^{(t)}$ denote the community's response coefficients, author a's preferences, and author a's tradeoff coefficient at timestep t.

Again, we must take care in interpreting these quantities. Do changes in community interest drive authors to adjust their preferences or expertise? Or do changing author preferences aggregate into community-wide shifts? Or do changes in the economy or funding availability change authors' tradeoffs? Our model cannot differentiate among these different causal patterns. Our method is useful for tracking these changes, but it does not provide an explanation for *why* they take place.

Modeling the temporal dynamics of a vector-valued random variable can be accomplished using a multivariate Gaussian distribution. Following Yogatama et al. (2011), we assume the prior for $\boldsymbol{\beta}_j^{(\cdot)} = \langle \beta_j^{(1)}, \ldots, \beta_j^{(T)} \rangle$ has a tridiagonal precision matrix $\Lambda(\lambda, \alpha) \in \mathbb{R}^{T \times T}$:

$$\Lambda(\lambda, \alpha) = \lambda \begin{pmatrix} 1+\alpha & -\alpha & 0 & 0 & \ldots \\ -\alpha & 1+2\alpha & -\alpha & 0 & \ldots \\ 0 & -\alpha & 1+2\alpha & -\alpha & \ldots \\ 0 & 0 & -\alpha & 1+2\alpha & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

The two hyperparameters α and λ capture, respectively, autocorrelation (the tendency of $\beta_j^{(t+1)}$ to be similar to $\beta_j^{(t)}$) and overall variance. This approach to modeling time series allows us to capture temporal dynamics while sharing statistical strength of evidence across all time steps.

We use the notation $\mathcal{T}(\lambda, \alpha) \equiv \mathcal{N}(\mathbf{0}, \Lambda(\lambda, \alpha))$ for this multivariate Gaussian distribution, instances of which are used as priors over response coefficients $\boldsymbol{\beta}$, author preferences $\boldsymbol{\eta}_a$, and (transformed) author tradeoffs $\log \kappa_a$.

.

Observed evidence	
$w_{d,i}$	ith token in document d
V	vocabulary size
N_d	number of tokens in document d
y_d	response to document d
$\mathcal{A}$	the set of authors
$\boldsymbol{a}_d$	set of authors of document d ($\subseteq \mathcal{A}$)
T	number of timesteps
$\mathcal{D}_t$	the set of documents from timestep t
$\mathcal{D}$	the set of all documents ($= \bigcup_{t=1}^{T} \mathcal{D}_t$)
Latent variables	
$\boldsymbol{\beta}^{(t)}$	response coefficients at time t ($\in \mathbb{R}^K$)
$\boldsymbol{\eta}_a^{(t)}$	author a's topic preferences at time t ($\in \mathbb{R}^K$)
$\boldsymbol{\kappa}_a^{(t)}$	author a's tradeoff coefficient at time t ($\in \mathbb{R}_{\geq 0}$)
$\boldsymbol{\theta}_d$	document d topic associations ($\in \mathbb{R}^K$)
$c_{d,a}$	author a contrtibution to document d ($\sum_{a \in \boldsymbol{a}_d} c_{d,a} = 1$)
$\boldsymbol{\phi}_k$	distribution over terms for topic k
$z_{d,i}$	topic assignment of $w_{d,i}$
Constants and hyperparameters	
K	number of topics
ρ	symmetric Dirichlet hyperparameter for $\boldsymbol{\phi}_k$
σ_c^2	variance hyperparameter for author contributions $\boldsymbol{c}_d$
$\{\lambda^{(\beta)}, \alpha^{(\beta)}\}$, $\{\lambda^{(\eta)}, \alpha^{(\eta)}\}$, $\{\lambda^{(\kappa)}, \alpha^{(\kappa)}\}$	hyperparameters for priors of $\boldsymbol{\beta}, \boldsymbol{\eta}$, and $\log \kappa$ respectively

Table 1: Table of notation.

3.5 Full Model

Table 1 summarizes all of the notation. The log-likelihood of our model is:

$$\begin{aligned}\mathcal{L} &= \log p(\boldsymbol{\beta}) + \sum_{d \in \mathcal{D}} \log p(\boldsymbol{c}_d) \\ &+ \sum_{d \in \mathcal{D}} \log p(y_d \mid \boldsymbol{\theta}_d, \boldsymbol{\beta}) + \log p(\boldsymbol{w}_d \mid \boldsymbol{\theta}_d) \\ &+ \sum_{a \in \mathcal{A}} \log p(\boldsymbol{\eta}_a) + \log p(\kappa_a) \\ &+ \sum_{d \in \mathcal{D}} \sum_{a \in \boldsymbol{a}_d} \log p(\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \boldsymbol{\eta}_a, \kappa_a, c_{d,a}) \qquad (8)\end{aligned}$$

We adopt a Bayesian approach to parameter estimation. The generative story, including all priors, is as follows. Recall that $\mathcal{T}(\cdot, \cdot)$ denotes the time series prior discussed in §3.4. See also the plate diagram for the graphical model in Fig. 1.

1. For each topic $k \in \{1, \ldots, K\}$:
   (a) Draw response coefficients $\boldsymbol{\beta}_k^{(\cdot)} \sim \mathcal{T}(\lambda^{(\beta)}, \alpha^{(\beta)})$ and term distribution $\boldsymbol{\phi}_k \sim \text{Dirichlet}(\rho)$.
   (b) For each author $a \in \mathcal{A}$, draw preference strengths for topic k over time, $\langle \eta_{a,k}^{(1)}, \ldots, \eta_{a,k}^{(t)} \rangle \sim \mathcal{T}(\lambda^{(\eta)}, \alpha^{(\eta)})$.
2. For each author $a \in \mathcal{A}$, draw (transformed) tradeoff parameters $\langle \log \kappa_a^{(1)}, \ldots, \log \kappa_a^{(T)} \rangle \sim \mathcal{T}(\lambda^{(\kappa)}, \alpha^{(\kappa)})$.
3. For each timestep $t \in \{1, \ldots, T\}$, and each document $d \in \mathcal{D}_t$:
   (a) Draw author contributions $\boldsymbol{c}_d \sim \text{Softmax}(\mathcal{N}(\mathbf{0}, \sigma_c^2 \mathbf{I}))$. This is known as a logistic normal distribution (Aitchison, 1986).
   (b) Draw d's topic distributions (this distribution is discussed further below):

$$\boldsymbol{\theta}_d \sim \mathcal{N}\left(\sum_{a \in \boldsymbol{a}_d} \kappa_a^{(t)} \boldsymbol{\beta}^{(t)} + c_{d,a} \boldsymbol{\eta}_a^{(t)}, \|\boldsymbol{c}_d\|_2^2 \mathbf{I} \right) \qquad (9)$$

   (c) For each token $i \in \{1, \ldots, N_d\}$, draw topic $z_{d,i} \sim \text{Categorical}(\text{Softmax}(\boldsymbol{\theta}_d))$ and term $w_{d,i} \sim \text{Categorical}(\boldsymbol{\phi}_{z_{d,i}})$.
   (d) Draw response $y_d \sim \mathcal{N}\left(\boldsymbol{\beta}^{(t_d)\top} \boldsymbol{\theta}_d, 1\right)$; note that it collapses out ξ_d, which is drawn from a standard normal.

Eq. 9 captures the choice by authors $\boldsymbol{a}_d$ of a distribution over topics $\boldsymbol{\theta}_d$. Assuming that the $\boldsymbol{\epsilon}_{d,a}$s are i.i.d. and Gaussian, from Eq. 7, we get

$$\boldsymbol{\theta}_d = \sum_{a \in \boldsymbol{a}_d} \kappa_a \boldsymbol{\beta} + c_{d,a} \boldsymbol{\eta}_a + c_{d,a} \boldsymbol{\epsilon}_{d,a},$$

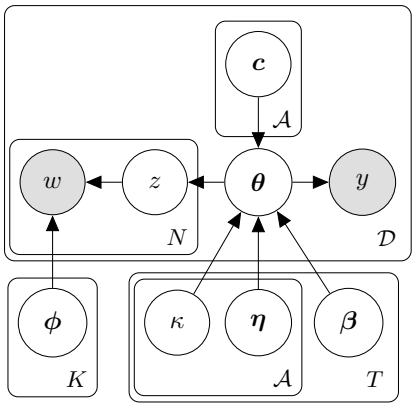

Figure 1: Plate diagram for author utility model. Hyperparameters and edges between consecutive time steps of $\boldsymbol{\beta}, \boldsymbol{\eta}$ and κ are omitted for clarity.

and the linear additive property of Gaussians gives us

$$\boldsymbol{\theta}_d \sim \mathcal{N}\left(\sum_{a \in \boldsymbol{a}_d} \kappa_a \boldsymbol{\beta} + c_{d,a} \boldsymbol{\eta}_a, \|\boldsymbol{c}_d\|_2^2 \mathbf{I} \right)$$

In §3.1 we described a utility function for each author. The model we are estimating is similar to those estimated in discrete choice econometrics (McFadden, 1974). We assumed that authors are utility maximizing (optimizing) and that their optimal topic distribution satisfies the first-order conditions (Eq. 7). However, we cannot see the idiosyncratic component, $\boldsymbol{\epsilon}_{d,a}$, which is assumed to be Gaussian; as noted, this is known as a random utility model. Together, these assumptions give the structure of the distribution over topics in terms of (estimated) utility, which allows us to naturally incorporate the utility function into our probabilistic model in a familiar way (Sim et al., 2015).

4 Learning and Inference

Exact inference in our model is intractable, so we resort to an approximate inference technique based on Monte Carlo EM (Wei and Tanner, 1990). During the E-step, we perform Bayesian inference over latent parameters $(\boldsymbol{\eta}, \boldsymbol{\kappa}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{\phi})$ using a Metropolis-Hastings within Gibbs algorithm (Tierney, 1994), and in the M-step, we compute maximum *a posteriori* estimates of $\boldsymbol{\beta}$ by directly optimizing the log-likelihood function. Since we are using conjugate priors for $\boldsymbol{\phi}$, we can integrate it out. We did not perform Bayesian posterior inference over $\boldsymbol{\beta}$ because the coupling of $\boldsymbol{\beta}$

would slow mixing of the MCMC chain.

E-step. We sample each $\boldsymbol{\eta}_a^{(t_d)}$, $\boldsymbol{c}_d$, $\log \boldsymbol{\kappa}_a^{(\cdot)}$, and $\boldsymbol{\theta}_d$ blockwise using the Metropolis-Hastings algorithm with a multivariate Gaussian proposal distribution, tuning the diagonal covariance matrix to a target acceptance rate of 15-45% (see appendix §A for sampling equations).

For $\boldsymbol{z}$, we integrate out $\boldsymbol{\phi}$ and sample each $z_{d,i}$ directly from

$$p(z_{d,i} = k \mid \boldsymbol{\theta}_d, \boldsymbol{\phi}_k) \propto \exp(\theta_{d,k}) \frac{C_{k,w_{d,i}}^{-d,i} + \rho}{C_{k,\cdot}^{-d,i} + V\rho}$$

where $C_{k,w}^{-d,i}$ and $C_{k,\cdot}^{-d,i}$ are the number of times w is associated with topic k, and the number of tokens associated with topic k respectively.

We run the E-step Gibbs sampler to collect 3,500 samples, discarding the first 500 samples for burn-in and only saving samples at every third iteration.

M-step. We approximate the expectations of our latent variables using the samples collected during the E-step, and directly optimize $\boldsymbol{\beta}^{(t)}$ using L-BFGS (Liu and Nocedal, 1989),[2] which requires a gradient. The gradient of the log-likelihood with respect to $\beta_j^{(t)}$ is

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \beta_j^{(t)}} &= -2\lambda^{(\beta)} \beta_j^{(t)} \\
&\quad - 2\lambda^{(\beta)} \alpha^{(\beta)} \mathbf{1}\{t > 1\} (\beta_j^{(t)} - \beta_j^{(t-1)}) \\
&\quad - 2\lambda^{(\beta)} \alpha^{(\beta)} \mathbf{1}\{t < T\} (\beta_j^{(t)} - \beta_j^{(t+1)}) \\
&\quad + 2 \sum_{d \in \mathcal{D}_t} \theta_{d,j} (y_d - \beta_j^{(t)} \theta_{d,j}) \\
&\quad + 2 \sum_{d \in \mathcal{D}_t} \kappa_d^{(t)} \left(\theta_{d,j} - \kappa_d^{(t)} \beta_j^{(t)} - \sum_{a \in \boldsymbol{a}_d} \frac{\eta_{a,j}^{(t)}}{|\boldsymbol{a}_d|} \right)
\end{aligned} \tag{10}$$

where $\kappa_d^{(t)} = \frac{1}{|\boldsymbol{a}_d|} \sum_{a \in \boldsymbol{a}_d} \kappa_a^{(t)}$.

We ran L-BFGS until convergence[3] and slice sampled the hyperparameters $\lambda^{(\eta)}, \alpha^{(\eta)}, \lambda^{(\kappa)}, \alpha^{(\kappa)}$ (with vague priors) at the end of the M-step. We fix the symmetric Dirichlet hyperparameter $\rho = 1/V$, and tuned $\lambda^{(\beta)}, \alpha^{(\beta)}$ on a held-out developement dataset by grid search over $\{0.01, 0.1, 1, 10\}$.

[2]We used libLBFGS, an open source C++ implementation (`https://github.com/chokkan/liblbfgs`).

[3]Relative tolerance of 10^{-4}.

During initialization, we randomly set the topic assignments, while the other latent parameters are set to 0. We ran the model for 10 EM iterations.

Inference. During inference, we fix the model parameters and only sample $(\boldsymbol{\theta}, \boldsymbol{z})$ for each document. As in the E-step, we discard the first 500 samples, and save samples at every third iteration, until we have 500 posterior samples. In our experiments, we found the posterior samples to be reasonably stable after the initial burn in.

5 Experiments

Data. The ACL Anthology Network Corpus contains 21,212 papers published in the field of computational linguistics between 1965 and 2013 and written by 17,792 authors. Additionally, the corpus provides metadata such as authors, venue and in-community citation networks. For our experiments, we focused on conference papers published between 1980 and 2010.[4] We tokenized the texts, tagged the tokens using the Stanford POS tagger (Toutanova et al., 2003), and extracted n-grams with tags that follow the simple (but effective) pattern of `(Adj|Noun)* Noun` (Justeson and Katz, 1995), representing the dth document as a *bag of phrases* ($\boldsymbol{w}_d$). Note that phrases can also be unigrams. We pruned phrases that appear in $< 1\%$ or $> 95\%$ of the documents, obtaining a vocabulary of V = 6,868 types. The pruned corpus contains 5,498 documents and 2,643,946 phrase tokens written by 5,575 authors. We let responses

$$y_d = \log(1 + \#\text{ of incoming citations in 3 years})$$

For our experiments, we used 3 different random splits of our data (70% train, 20% test, and 10% development) and averaged quantities of interest. Furthermore, we remove an author from a paper in the development or test set if we have not seen him before in the training data.

5.1 Examples of Authors and Topics

Table 2 illustrates ten manually selected topics (out of 64) learned by the author utility model. Each topic is labeled with the top 10 words most likely to be generated conditioned on the topic

[4]The conferences we included are: ACL, CoNLL, EACL, EMNLP, HLT, and NAACL. We ignored journal papers, as well as workshop papers, since they are characteristically different from conference papers.

(ϕ_k). For each topic, we compute an author's topic preference score:

$$\text{TPS}(a, k) = \eta_{a,k}^{(t_d)} \sum_{d \in D_a} [\text{Softmax}(\theta_d)]_k \times y_d$$

where $\text{Softmax}(\boldsymbol{x}) = \frac{\exp(\boldsymbol{x})}{\sum_i \exp(x_i)}$. The TPS scales the author's η preferences by the relative number of citations that the author received for the topic. This way, we can account for different ηs over time, and reduce variance due to authors who publish less frequently.[5] For each topic, the five authors with the highest TPS are displayed in the rightmost column of Table 2. These topics were among the roughly one third (out of 64) that seemed to coherently map to research topics within NLP. Some others corresponded to parts of a paper (e.g., explaining notation and formulae, experiments) or to stylistic groups (e.g., "rational words" including *rather*, *fact*, *clearly*, *argue*, *clear*, *perhaps*). Others were not interpretable to us.

5.2 Predicting Responses

We compare against two baselines for predicting in-community citations. Yogatama et al. (2011) is a strong baseline for predicting responses; they incorporated n-gram features and metadata features in a generalized linear model with the time series prior discussed in §3.4.[6] We also compare against a version of our model without the author utility component. This equates to replacing Yogatama et al.'s features with LDA topic mixtures, and performing joint learning of the topics and citations; we therefore call it "TimeLDA." Without the time series component, TimeLDA would instantiate supervised LDA (McAuliffe and Blei, 2008). Figure 2 shows the mean absolute error (MAE) for the three models.

With sufficiently many topics ($K \geq 16$), topic representations achieve lower error than surface features. Removing the author utility component from our model leads to better predictive performance. This is unsurprising, since our model forces β to explain both the responses (what is evaluated here) and the divergence between author preferences η_a and what is actually written. The utility model is nonetheless competitive with the Yogatama et al. baseline.

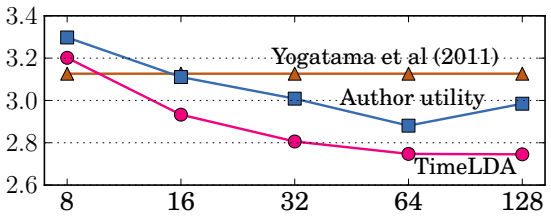

Figure 2: Mean absolute error (in citation counts) for predicted citation counts (y-axis) against the number of topics K (x-axis). Errors are in actual citation counts, while the models are trained with log counts. TimeLDA significantly outperforms Yogatama et al. (2011) for $K \geq 64$ (paired t-test, $p < 0.01$), while the differences between Yogatama et al. (2011) and author utility are not significant. The MAE is calculated over 3 random splits of the data with 809, 812, and 811 documents in the test set respectively.

5.3 Predicting Words

"Given a set of authors, what are they likely to write?" — we use perplexity as a proxy to measure the content predictive ability of our model. Perplexity on a test set is commonly used to quantify the generalization ability of probabilistic models and make comparisons among models over the same observation space. For a document $\boldsymbol{w}_d$ written by authors $\boldsymbol{a}_d$, perplexity is defined as

$$\text{perplexity}(\boldsymbol{w}_d \mid \boldsymbol{a}_d) = \exp\left(-\frac{\log p(\boldsymbol{w}_d \mid \boldsymbol{a}_d)}{N_d}\right)$$

and a lower perplexity indicates better generalization performance. Using S samples from the inference step, we can compute

$$p(\boldsymbol{w}_d \mid \boldsymbol{a}_d) = \frac{1}{S} \sum_{s=1}^{S} \prod_{i=1}^{N_d} \frac{1}{|\boldsymbol{a}_d|} \sum_{a \in \boldsymbol{a}_d, k} \theta_{d,k}^s \phi_{k,w_{di}}^s$$

where θ^s is the sth sample of θ, and ϕ^s is the topic-word distribution estimated from the sth sample of $\boldsymbol{z}$.

We compared the Author-Topic model of Rosen-Zvi et al. (2004). The AT model is similar to setting $\kappa_a = 0$ for all authors, $\boldsymbol{c}_d = \frac{1}{|\boldsymbol{a}_d|}$, and using a Dirichlet prior instead of logistic normal on η_a. Figure 3 present the perplexity of these

[5] The TPS is only a measure of an author's propensity to write papers in a specific topic area and is not meant to be a measure of an author's reputation in a particular research sub-field.

[6] For the ACL dataset, Yogatama et al. (2011)'s model predicts whether a paper will receive at least 1 citation within three years, while here, we train it to predict $\log(1 + \#\text{citations})$ instead.

Topic	Top words	Authors
"MT"	alignment, translation, align, decode, phrase, och, bleu, ney, bleu score, target language	Philipp Koehn, Chris Dyer, Qun Liu, Hermann Ney, David Chiang
"Empirical methods"	model, parameter, learn, iteration, maximize, prior, initialize, distribution, weight, crf	Noah Smith, Dan Klein, Percy Liang, John DeNero, Andrew McCallum
"Parsing"	parse, sentence, parser, accuracy, collins, dependency, tree, parse tree, head, charniak	Michael Collins, Joakim Nivre, Jens Nilsson, Dan Klein, Ryan McDonald
"Dialogue systems"	speak, speech, utterance, user, speaker, dialogue system, turn, act, recognition, transcription	Diane Litman, Marilyn Walker, Julia Hirschberg, Oliver Lemon, Amanda Stent
"NER"	name, entity, identify, person, location, list, organization, system, entity recognition, mention	Jenny Rose Finkel, Satoshi Sekine, Rion Snow, Christopher Manning, Abraham Ittycheriah
"Semantics"	argument, verb, predicate, syntactic, relation, semantic role, annotated, frame, assign	Martha Palmer, Alessandro Moschitti, Daniel Jurafsky, Sanda Harabagiu, Mirella Lapata
"Lexical semantics"	wordnet, noun, sense, concept, context, sens, relation, meaning, pair, disambiguate	Rion Snow, Rob Koeling, Eneko Agirre, Ido Dagan, Patrick Pantel
"Tagging & chunking"	method, sentence, propose, japanese, noun phrase, extract, table, analyze, precision, technology	Yuji Matsumoto, Hitoshi Isahara, Junichi Tsujii, Sadao Kurohashi, Kentaro Torisawa
"Coreference"	mention, instance, create, approach, report, due, text, pair, exist, system	Vincent Ng, Aria Haghighi, Xiaofeng Yang, Claire Cardie, Pascal Denis
"Sentiment classification"	classify, label, accuracy, positive, classification, annotated, annotator, classifier, review, negative	Janyce Wiebe, Soo Min Kim, Eduard Hovy, Carmen Banea, Ryan McDonald

Table 2: Top words from selected topics and authors with preferences in those topics. We manually labeled each of these topics.

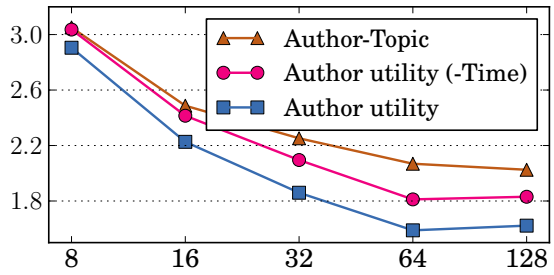

Figure 3: Held-out perplexity ($\times 10^3$, y-axis) with varying number of topics K (x-axis). The differences are significant between all models at $K \geq 64$ (paired t-test, $p < 0.01$). There are 523,381, 529,397, 533,792 phrase tokens in the random test sets.

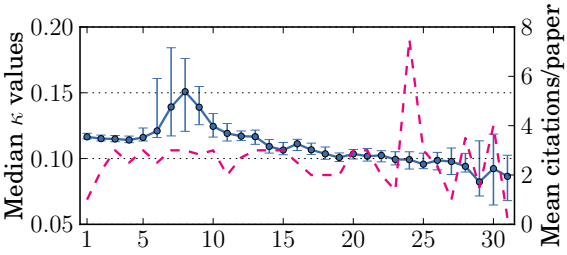

Figure 4: Plot of authors' median κ (blue, solid) and mean citation counts (magenta, dashed) against their academic age in this dataset (see text for explanation).

models at different values of K. We include a version of our author utility model that ignores temporal information ("–time"), i.e., setting $T = 1$ and collapsing all timesteps. We find that perplexity improves with the addition of the utility model as well as the temporal dynamics.

5.4 Exploration: Tradeoffs and Seniority

Recall that κ_a encodes author a's tradeoff between increasing citations (high κ_a) and writing papers on topics a prefers (low κ_a). We do not claim that individual κ_a values consistently represent authors' tradeoffs between citations and writing about preferred topics. We have noted a number of potentially confounding factors that affect authors' choices, for which our data do not allow us to control.

However, in aggregate, κ_a values can be explored in relation to other quantities. Given our model's posterior, one question we can ask is: do an author's tradeoffs tend to change over the course of her career? In Figure 4, we plot the median of κ (and 95% credible intervals) for authors at different "ages." Here, "age" is defined as the number of years since an author's first publication in this dataset.[7]

A general trend over the long term is observed: researchers appear to move from higher to lower κ_a. Statistically, there is significant dependence between κ of an author and her age; the Spearman's rank correlation coefficient is $\rho = -0.870$ with p-value $< 10^{-5}$. This finding is consis-

[7]This means that larger ages correspond to seniority, but smaller ages are a blend of junior researchers and researchers of any seniority new to this publication community.

tent with the idea that greater seniority brings increased and more stable resources and greater freedom to pursue idiosyncratic interests with less concern about extrinsic payoff. It is also consistent with decreased flexibility or openness to shifting topics over time.

To illustrate the importance of our model in making these observations, we also plot the mean number of citations per paper published (across all authors) against their academic age (magenta lines). There is no clear statistical trend between the two variables ($\rho = -0.017$). This suggests that through κ, our model is able to pick up evidence of author's optimizing behaviors, which is not possible using simple citation counts.

There is a noticeable effect during years 5–10, in which κ tends to rise by around 40% and then fall back. (Note that the model maintains considerable uncertainty—wider intervals—about this effect.) Recall that, for a researcher trained within the field and whose primary publication venue is in the ACL community, our measure of age corresponds roughly to academic age. Years 5–10 would correspond to the later part of a Ph.D. program and early postgraduate life, when many researchers begin faculty careers. Insofar as it reflects a true effect, this rise and fall suggests a stage during which a researcher focuses more on writing papers that will attract citations. However, more in-depth study based on data that is not merely observational is required to quantify this effect and, if it persists under scrutiny, determine its cause.

The effect in year 24 of mean citations per paper (magenta line) can be attributed to well cited papers co-authored by senior researchers in the field who published very few papers in their 24th year. Since there are relatively few authors in the dataset at that academic age, there is more variance in mean citations counts.

6 Related Work

Previous work on modeling author interests mostly focused on characterizing authors by their style (Holmes and Forsyth, 1995, *inter alia*),[8] through latent topic mixtures of documents they have co-authored (Rosen-Zvi et al., 2004) and their collaboration networks (Johri et al., 2011). Like our paper, the latter two are based on topic models, which have been popular for modeling the content of scientific articles. For instance, Gerrish and Blei (2010) measured scholarly impact using dynamic topic models, while Hall et al. (2008) analyzed the output of topic models to study the "history of ideas."

Predicting responses to scientific articles was explored in two shared tasks at KDD Cup 2003 (Brank and Leskovec, 2003; McGovern et al., 2003) and by Yogatama et al. (2011), which served as a baseline for our experiments and whose time-series prior we used in our model. Furthermore, there has been considerable research using topic models to predict (or recommend) citations (instead of aggregate counts), such as modeling link probabilities within the LDA framework (Cohn and Hofmann, 2000; Erosheva et al., 2004; Nallapati and Cohen, 2008; Kataria et al., 2010; Zhu et al., 2013) and augmenting topics with discriminative author features (Liu et al., 2009; Tanner and Charniak, 2015).

We modeled both interests of authors and responses to their articles jointly, by assuming authors' text production is an expected utility-maximizing decision. This approach is similar to our earlier work (Sim et al., 2015), where authors are rational agents writing texts to maximize the chance of a favorable decision by a judicial court. In that study, we did not consider the unique preferences of each decision making agent, nor the extrinsic-intrinsic reward tradeoffs that these agents face when authoring a document.

Our utility model can also be viewed as a form of natural language generator, where we take into account the context of an author (i.e., his preferences, the tradeoff coefficient, and what is popular) to generate his document. This is related to natural language pragmatics, where text is influenced by context.[9] Hovy (1990) approached the problem of generating text under pragmatic circumstances from a planning and goal-orientation perspective, while Vogel et al. (2013) used multi-agent decision-theoretic models to show cooperative pragmatic behavior. Vogel et al.'s models suggest an interesting extension of ours for future work: modeling cooperation among co-authors and, perhaps, in the larger scientific discourse.

[8] A closely related problem is that of authorship attribution. There has been extensive research on authorship attribution focusing mainly on learning "stylometric" features of authors; see Stamatatos (2009) for a detailed review.

[9] The $\boldsymbol{\beta}$ vectors can be seen as a naïve representation of world knowledge that motivates an author to select content that reflects his behavioral preferences and intentions.

7 Conclusions

We presented a model of scientific authorship in which authors trade off between seeking citation by others and staying true to their individual preferences among research topics. We find that topic modeling improves over state-of-the-art text regression models for predicting citation counts, and that the author utility model generalizes better than simpler models when predicting what a particular group of authors will write. Inspecting our model suggests interesting patterns in behavior across a researcher's career.

Acknowledgements

The authors thank the anonymous reviewers for their thoughtful feedback and members of the ARK group at CMU for their valuable comments. This research was supported in part by an A*STAR fellowship to Y. Sim, by a Google research award, and by computing resources from the Pittsburgh Supercomputing Center; it was completed while NAS was at CMU.

A Appendix: Sampling equations

We sample each $\boldsymbol{\eta}_{a,j}$, for $j = 1 \dots K$, and $\boldsymbol{\kappa}_a$ blockwise across time steps using Metropolis-Hastings algorithm with a multivariate Gaussian proposal distribution and likelihood:

$$p(\boldsymbol{\eta}_{a,j} \mid \boldsymbol{\eta}_{-(a,j)}, \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{\kappa}, \boldsymbol{\beta}, \Lambda^{(\eta)}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\eta}_{a,j}\Lambda^{(\eta)}\boldsymbol{\eta}_{a,j}^{\top} - \sum_{\substack{t \in T \\ d \in D_t}} \frac{\left(\theta_{d,j} - \sum_{a' \in \boldsymbol{a}_d} \kappa_{a'}^{(t)}\beta_j^{(t)} + c_{d,a'}\eta_{a',j}^{(t)}\right)^2}{2\|\boldsymbol{c}_d\|_2^2} \right)$$

$$p(\boldsymbol{\kappa}_a \mid \boldsymbol{\kappa}_{-(a)}, \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{\eta}, \boldsymbol{\beta}, \Lambda^{(\kappa)}) \propto \exp\left(-\frac{1}{2}\log(\boldsymbol{\kappa}_a)\Lambda^{(\kappa)}\log(\boldsymbol{\kappa}_a^{\top}) - \sum_{\substack{t \in T \\ d \in D_t}} \frac{\|\boldsymbol{\theta}_d - \sum_{a' \in \boldsymbol{a}_d} \kappa_{a'}^{(t)}\boldsymbol{\beta}^{(t)} + c_{d,a'}\boldsymbol{\eta}_{a'}^{(t)}\|_2^2}{2\|\boldsymbol{c}_d\|_2^2} \right)$$

$\Lambda^{(\eta)}$ and $\Lambda^{(\kappa)}$ are shorthands for the precision matrices $\Lambda(\lambda^{(\eta)}, \alpha^{(\eta)})$ and $\Lambda(\lambda^{(\kappa)}, \alpha^{(\kappa)})$ respectively. Likewise, $\boldsymbol{\theta}_d$ is sampled blockwise for each document with a multivariate Gaussian distribution and likelihood:

$$p(\boldsymbol{\theta}_d \mid \boldsymbol{c}_d, \boldsymbol{\eta}, \boldsymbol{\kappa}, \boldsymbol{\beta}) \propto \exp\left(-\frac{(y_d - \boldsymbol{\beta}^{(t_d)\top}\boldsymbol{\theta}_d)^2}{2} - \frac{\|\boldsymbol{\theta}_d - \sum_{a \in \boldsymbol{a}_d} \kappa_a^{(t_d)}\boldsymbol{\beta}^{(t_d)} + c_{d,a}\boldsymbol{\eta}_a^{(t_d)}\|_2^2}{2\|\boldsymbol{c}_d\|_2^2} \right)$$

For $\boldsymbol{c}_d$, we first sampled each $\boldsymbol{c}_d$ from a multivariate Gaussian distribution, and applied a logistic transformation to map it onto the simplex. The likelihood for $\boldsymbol{c}_d$ is:

$$p(\boldsymbol{c}_d \mid \boldsymbol{\theta}_d, \boldsymbol{\eta}, \boldsymbol{\kappa}, \boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2\sigma_c^2}\left\|\log\left(\frac{\boldsymbol{c}_d}{c_{d,|\boldsymbol{a}_d|}}\right)\right\|_2^2 - \frac{\|\boldsymbol{\theta}_d - \sum_{a \in \boldsymbol{a}_d} \kappa_a^{(t_d)}\boldsymbol{\beta}^{(t_d)} + c_{d,a}\boldsymbol{\eta}_a^{(t_d)}\|_2^2}{2\|\boldsymbol{c}_d\|_2^2} \right)$$

References

John Aitchison. 1986. *The Statistical Analysis of Compositional Data*. Chapman & Hall.

Katharine A. Anderson. 2012. Specialists and generalists: Equilibrium skill acquisition decisions in problem-solving populations. *Journal of Economic Behavior & Organization*, 84(1):463–473.

David M. Blei and John D. Lafferty. 2007. A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March.

Janez Brank and Jure Leskovec. 2003. The download estimation task on KDD Cup 2003. *SIGKDD Explorations Newsletter*, 5(2):160–162, December.

David A. Cohn and Thomas Hofmann. 2000. The missing link – a probabilistic model of document content and hypertext connectivity. In *NIPS*.

Elena Erosheva, Stephen Fienberg, and John Lafferty. 2004. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101(suppl. 1):5220–5227.

Sean Gerrish and David M. Blei. 2010. A language-based approach to measuring scholarly impact. In *Proc. of ICML*.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *Proc. of EMNLP*.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proc. of SIGIR*.

D. I. Holmes and R. S. Forsyth. 1995. The federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing*, 10(2):111–127.

Eduard H. Hovy. 1990. Pragmatics and natural language generation. *Artificial Intelligence*, 43(2):153–197, May.

Nikhil Johri, Daniel Ramage, Daniel A. McFarland, and Daniel Jurafsky. 2011. A study of academic collaboration in computational linguistics with latent mixtures of authors. In *Proc. of the Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27, March.

Saurabh Kataria, Prasenjit Mitra, and Sumit Bhatia. 2010. Utilizing context in generative Bayesian models for linked corpus. In *Proc. of AAAI*.

Dong C. Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528.

Yan Liu, Alexandru Niculescu-Mizil, and Wojciech Gryc. 2009. Topic-link LDA: Joint models of topic and author community. In *Proc. of ICML*.

Jon D. McAuliffe and David M. Blei. 2008. Supervised topic models. In *NIPS*.

Daniel McFadden. 1974. Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, pages 105–142. Academic Press.

Amy McGovern, Lisa Friedland, Michael Hay, Brian Gallagher, Andrew Fast, Jennifer Neville, and David Jensen. 2003. Exploiting relational structure to understand publication patterns in high-energy physics. *SIGKDD Exploration Newsletter*, 5(2):165–172, December.

Ramesh Nallapati and William W. Cohen. 2008. Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs. In *Proc. of ICWSM*.

Dragomir R. Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The ACL anthology network corpus. *Language Resources and Evaluation*, pages 1–26. Data available at `http://clair.eecs.umich.edu/aan/`.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proc. of UAI*.

Yanchuan Sim, Bryan Routledge, and Noah A. Smith. 2015. The utility of text: The case of amicus briefs and the Supreme Court. In *Proc. of AAAI*.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Chris Tanner and Eugene Charniak. 2015. A hybrid generative/discriminative approach to citation prediction. In *Proc. of NAACL*.

Luke Tierney. 1994. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):pp. 1701–1728.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of NAACL*.

Adam Vogel, Max Bodoia, Christopher Potts, and Daniel Jurafsky. 2013. Emergence of Gricean maxims from multi-agent decision theory. In *Proc. of NAACL*.

Greg C. G. Wei and Martin A. Tanner. 1990. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):pp. 699–704.

Dani Yogatama, Michael Heilman, Brendan O'Connor, Chris Dyer, Bryan R. Routledge, and Noah A. Smith. 2011. Predicting a scientific community's response to an article. In *Proc. of EMNLP*.

Yaojia Zhu, Xiaoran Yan, Lise Getoor, and Cristopher Moore. 2013. Scalable text and link analysis with mixed-topic link models. In *Proc. of KDD*.