Discovering Factions in the Computational Linguistics Community

Yanchuan Sim Noah A. Smith
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{ysim,nasmith}@cs.cmu.edu

David A. Smith
Department of Computer Science
University of Massachusetts
Amherst, MA 01003, USA
dasmith@cs.umass.edu

Abstract

We present a joint probabilistic model of who cites whom in computational linguistics, and also of the words they use to do the citing. The model reveals latent *factions*, or groups of individuals whom we expect to collaborate more closely within their faction, cite within the faction using language distinct from citation outside the faction, and be largely understandable through the language used when cited from without. We conduct an exploratory data analysis on the ACL Anthology. We extend the model to reveal changes in some authors' faction memberships over time.

1 Introduction

The ACL Anthology presents an excellent dataset for studying both the language and the social connections in our evolving research field. Extensive studies using techniques from the field of bibliometrics have been applied to this dataset (Radev et al., 2009a), quantifying the importance and impact factor of both authors and articles in the community. Moreover, recent work has leveraged the availability of digitized publications to study trends and influences within the ACL community (Hall et al., 2008; Gerrish and Blei, 2010; Yogatama et al., 2011) and to analyze academic collaborations (Johri et al., 2011).

To the best of our knowledge, however, existing work has mainly pursued "macroscopic" investigations of the interaction of authors in collaboration, citation networks, or the textual content of whole papers. We seek to complement these results with a "microscopic" investigation of authors' interactions by considering the individual sentences authors use to cite each other.

In this paper, we present a joint model of who cites whom in computational linguistics, and also of *how* they do the citing. Central to this model is the idea of *factions*, or groups of individuals whom we expect to (i) collaborate more closely within their faction, (ii) cite within the faction using language distinct from citation outside the faction, (iii) be largely understandable through the language used when cited from without, and (iv) evolve over time.[1] Factions can be thought of as "communities," which are loosely defined in the literature on networks as subgraphs where internal connections are denser than external ones (Radicchi et al., 2004). The distinction here is that the strength of connections depends on a latent language model estimated from citation contexts.

This paper is an exploratory data analysis using a Bayesian generative model. We aim both to discover meaningful factions in the ACL community and also to illustrate the use of a probabilistic model for such discovery. As such, we do not present any objective evaluation of the model or make any claims that the factions optimally explain the research community. Indeed, we suspect that reaching a broad consensus among community members about factions (i.e., a "gold standard") would be quite difficult, as any social community's factions are likely perceived very

[1] Our factions are computational abstractions—clusters of authors—discovered entirely from the corpus. We do not claim that factions are especially contentious, any more than "sub-communities" in social networks are especially collegial.

Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, pages 22–32,
Jeju, Republic of Korea, 10 July 2012. ©2012 Association for Computational Linguistics

subjectively. It is for this reason that a probabilistic generative model, in which all assumptions are made plain, is appropriate for the task. We hope this analysis will prove useful in future empirical research on social communities (including scientific ones) and their use of language.

2 Model

In this paper, our approach is a probabilistic model over (i) coauthorship relations and (ii) the words in sentences containing citations. The words are assumed to be generated by a distribution that depends on the (latent) faction memberships of the citing authors, the cited authors, and whether the authors have coauthored before. To model these different effects on language, we use a sparse additive generative (SAGE) model (Eisenstein et al., 2011). In contrast to the popular Dirichlet-multinomial for topic modeling, which directly models lexical probabilities associated with each (latent) topic, SAGE models the deviation in log frequencies from a background lexical distribution. Imposing a sparsity-inducing prior on the deviation vectors limits the number of terms whose probabilities diverge from the background lexical frequencies, thereby increasing robustness to limited training data. SAGE can be used with or without latent topics; our model does not include topics. Figure 1 shows the plate diagram for our model.

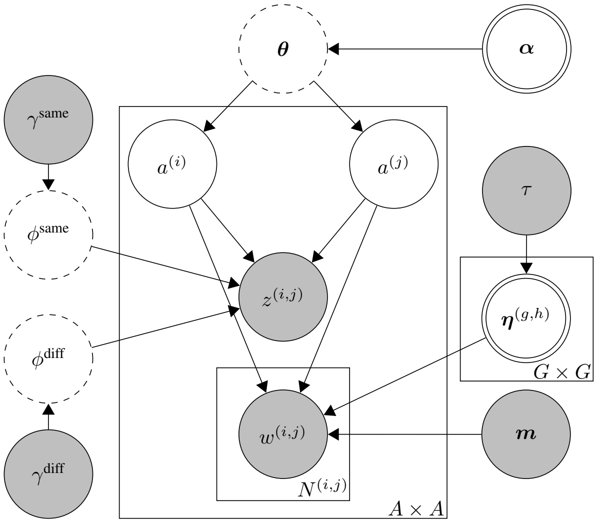

Figure 1: Plate diagram for our graphical model. A and G are the fixed numbers of authors and factions, respectively. $\boldsymbol{m}$ is the background word distribution, $\boldsymbol{\alpha}$, τ, γ are hyperparameters, $\boldsymbol{a}$ are latent author factions, $\boldsymbol{z}$ and $\boldsymbol{w}$ are the observed coauthorship relations and observed words in citation sentences between authors, respectively. Each of the $a^{(i)}$, denoting author i's faction alignment, are sampled once every iteration conditioned on all the other $a^{(j)}$. If i and j are coauthors or i cited j in some publication, $a^{(i)}$ and $a^{(j)}$ will not be conditionally independent due to the v-structure. ϕ^{same} and ϕ^{diff} are binomial distributions over whether two authors have collaborated together before, given that they are assigned to the same/different factions. Dashed variables are collapsed out in the Gibbs sampler, while double bordered variables are optimized in the M-step.

We describe the generative process:

- Generate the multinomial distribution over faction memberships from a Dirichlet distribution: $\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\alpha})$.

- Generate the binomial distribution for whether two authors coauthor, given that they are in the same faction, from a Beta distribution: $\phi^{\text{same}} \sim \text{Beta}(\gamma_0^{\text{same}}, \gamma_1^{\text{same}})$. Generate the analogous binomial, given that they are in different factions: $\phi^{\text{diff}} \sim \text{Beta}(\gamma_0^{\text{diff}}, \gamma_1^{\text{diff}})$.

- For each author i, draw a faction indicator $a_i \sim$ Multinomial($\boldsymbol{\theta}$).

- For all ordered pairs of factions (g, h), draw a deviation vector $\boldsymbol{\eta}^{(g,h)} \sim$ Laplace$(0, \tau)$. This vector, which will be sparse, corresponds to the deviations in word log-frequencies when faction g is citing faction h.

- For each word v in the vocabulary, let the unigram probability that an author in faction g uses to cite an author in faction h be

$$\beta_v^{(g,h)} = \frac{\exp(\eta_v^{(g,h)} + m_v)}{\sum_{v'} \exp(\eta_{v'}^{(g,h)} + m_{v'})}$$

.

- For each ordered pair of authors (i, j),
  - For each word that i uses to cite j, draw $w_k^{(i,j)} \sim$ Multinomial($\boldsymbol{\beta}^{(a^{(i)}, a^{(j)})}$).
  - If the authors are from the same faction, i.e., $a^{(i)} = a^{(j)}$, draw coauthorship indi-

cator $z^{(i,j)} \sim \text{Binomial}(\phi^{\text{same}})$; else, draw $z^{(i,j)} \sim \text{Binomial}(\phi^{\text{diff}})$.

Thus, our goal is to maximize the conditional likelihood of the observed data

$$p(\boldsymbol{w}, \boldsymbol{z} \mid \boldsymbol{\alpha}, \boldsymbol{\eta}, \tau, \boldsymbol{m}, \boldsymbol{\gamma}) = \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\phi}} \int_{\boldsymbol{a}} p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{a} \mid \boldsymbol{\alpha}, \boldsymbol{\eta}, \tau, \boldsymbol{m}, \boldsymbol{\gamma})$$

with respect to $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$. We fix τ and γ, which are hyperparameters that encode our prior beliefs, and $\boldsymbol{m}$, which we assume to be a fixed background word distribution.

Exact inference in this model is intractable, so we resort to an approximate inference technique based on Markov Chain Monte Carlo simulation. We perform Bayesian inference over the latent author factions while using maximum *a posteriori* estimates of $\boldsymbol{\eta}$ because Bayesian inference of $\boldsymbol{\eta}$ is problematic due to the logistic transformation. We refer the interested reader to Eisenstein et al. (2011). We take an empirical Bayes approach to setting the hyperparameter $\boldsymbol{\alpha}$. Our overall learning procedure is a Monte Carlo Expectation Maximization algorithm (Wei and Tanner, 1990).

3 Learning and Inference

Our learning algorithm is a two-step iterative procedure. During the E-step, we perform collapsed Gibbs sampling to obtain distributions over factions for each author, given the current setting of the hyperparameters. In the M-step, we obtain point estimates for the hyperparameters $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$ given the current posterior distributions for the author factions.

3.1 E-step

As the Dirichlet and Beta distributions are conjugate priors to the multinomial and binomial respectively, we can integrate out the latent variables $\boldsymbol{\theta}, \phi^{(\text{same})}$ and $\phi^{(\text{diff})}$. For an author i, we sample his faction alignment $a^{(i)}$ conditioned on faction assignments to all other authors and citation words between i and other authors (in both directions). Denoting $\boldsymbol{a}^{-i}$ as the current faction assignments for all the authors except i,

$$\begin{aligned} & p(a^{(i)} = g \mid \boldsymbol{a}^{(-i)}, \boldsymbol{w}, \boldsymbol{\eta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \\ & \quad \propto p(a^{(i)} = g, \boldsymbol{a}^{(-i)}, \boldsymbol{w} \mid \boldsymbol{\eta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \\ & \quad \propto (N_g + \alpha_g) \prod_{j}^{A} \frac{\gamma_z^\epsilon + N_z^\epsilon}{\gamma_0^\epsilon + \gamma_1^\epsilon + N_0^\epsilon + N_1^\epsilon} p(\boldsymbol{w}^{(i)} \mid \eta) \end{aligned}$$

where N_g is the number of authors (except i) who are assigned to faction g, ϵ_{ij} = "same" if $g = a^{(j)}$ and ϵ_{ij} = "diff" otherwise, and $N_1^\epsilon, N_0^\epsilon$ denotes the number of author pairs that have/have not coauthored before respectively, given the status of their factions ϵ. We elide the subscripts of ϵ and superscript of z for notational simplicity and abuse notation to let $\boldsymbol{w}^{(i)}$ refer to all author i's citation words, both incoming and outgoing. Using SAGE, the factor for an author's words is

$$p(\boldsymbol{w}^{(i)} \mid \eta) = \prod_j \prod_v \left(\beta_v^{(g,a^{(j)})}\right)^{w_v^{(i,j)}} \left(\beta_v^{(a^{(j)},g)}\right)^{w_v^{(j,i)}}$$

where $w_v^{(i,j)}$ is the observed count of the number of times word v has been used when author i cites j; j ranges over the A authors.

We sample each author's faction in turn and do so several times during the E-step, collecting samples to estimate our posterior distribution over $\boldsymbol{a}$.

3.2 M-step

In the M-step, we optimize all $\boldsymbol{\eta}^{(g,h)}$ and $\boldsymbol{\alpha}$ given the posterior distribution over author factions.

Optimizing $\boldsymbol{\eta}$. Eisenstein et al. (2011) postulated that the components of $\boldsymbol{\eta}$ are drawn from a compound model $\int \mathcal{N}(\eta; \mu, \sigma)\mathcal{E}(\sigma; \tau)d\sigma$, where $\mathcal{E}(\sigma; \tau)$ indicates the Exponential distribution. They fit a variational distribution $Q(\boldsymbol{\sigma})$ and optimized the log-likelihood of the data by iteratively fitting the parameters $\boldsymbol{\eta}$ using a Newton optimization step and maximizing the variational bound.

The compound model described is equivalent to the Laplace distribution $\mathcal{L}(\eta; \mu, \tau)$ (Lange and Sinsheimer, 1993; Figueiredo, 2003). Moreover, a zero mean Laplace prior has the same effect as placing an L_1 regularizer on $\boldsymbol{\eta}$. Therefore, we can equivalently maximize the regularized likelihood

$$\langle \boldsymbol{c}^{(g,h)} \rangle^T \boldsymbol{\eta}^{(g,h)} - \langle C^{(g,h)} \rangle \log \sum_v \exp(\eta_v^{(g,h)} + m_v)$$
$$- \lambda \left\| \boldsymbol{\eta}^{(g,h)} \right\|_1$$

with respect to $\eta^{(g,h)}$. $\langle \boldsymbol{c}^{(g,h)} \rangle$ is a vector of expected count of the words that faction g used when citing faction h, $\langle \boldsymbol{c}^{(g,h)} \rangle = \sum_v \langle c_v^{(g,h)} \rangle$ and λ is the regularization constant. The regularization constant and Laplace variance are related by $\lambda = \tau^{-1}$ (Tibshirani, 1996).

We use the gradient-based optimization routine OWL-QN (Andrew and Gao, 2007) to maximize the above objective function with respect to $\boldsymbol{\eta}^{(g,h)}$ for each pair of factions g and h.

Optimizing $\boldsymbol{\alpha}$. As in the empirical Bayes approach, we learn the hyperparameter setting of $\boldsymbol{\alpha}$ from the data by maximizing the log likelihood with respect to $\boldsymbol{\alpha}$. By treating $\boldsymbol{\alpha}$ as the parameter of a Dirichlet-multinomial compound distribution, we can directly use the samples of author factions produced by our Gibbs sampler to estimate $\boldsymbol{\alpha}$. Minka (2009) describes in detail several iterative approaches to estimate $\boldsymbol{\alpha}$; we use the linear-time Newton-Raphson iterative update to estimate the components of $\boldsymbol{\alpha}$.

4 Data Analysis

4.1 Dataset

We used the ACL Anthology Network Corpus (Radev et al., 2009b), which currently contains 18,041 papers written by 12,777 authors. These papers are published in the field of computational linguistics between 1965 and 2011.[2] Furthermore, the corpus provides bibliographic data such as authors of the papers and bibliographic references between each paper in the corpus. We extracted sentences containing citations using regular expressions and linked them between authors with the help of metadata provided in the corpus.

We tokenized the extracted sentences and downcased them. Words that are numeric, appear less than 20 times, or are in a stop word list are discarded. For papers with multiple authors, we divided the word counts by the number of pairings between authors in both papers, assigning each word to each author-pair (i.e., a count of $\frac{1}{nn'}$ if a paper with n authors cites a paper with n' authors).

Due to the large number of authors, we only used the 500 most cited authors (within the corpus) who have published at least 5 papers. Papers with no authors left are removed from the dataset. As a result, we have 8,144 papers containing 80,776 citation sentences (31,659 citation pairs). After text processing, there are 391,711 tokens and 3,037 word types.

In each iteration of the EM algorithm, we run the E-step Gibbs sampler for 300 iterations, discarding the first 100 samples for burn-in and collecting samples at every 3rd iteration to avoid autocorrelation. At the M-step, we update our $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$ using the samples collected. We run the model for 100 EM iterations.

We fixed $\lambda = 5$, $\boldsymbol{\gamma}^{\text{same}} = (0.5, 1)$ and $\boldsymbol{\gamma}^{\text{diff}} = (1, 0.5)$. Our setting of $\boldsymbol{\gamma}$ reflects our prior beliefs that coauthors tend to be from the same faction.

4.2 Factions in ACL (1965–2011)

We ran the model with $G = 30$ factions and selected the most probable faction for each author from the posterior distribution of the author-faction alignment obtained in the final E step. Only 26 factions were selected as most probable for some author.[3] Table 1 presents members of selected factions, along with citation words that have the largest positive log frequency deviation from the background distribution.[4] Table 2 shows a list of the top three authors associated with factions not shown in Table 1. Incoming (outgoing) citation words are found by summing the log deviation vectors $\boldsymbol{\eta}$ across citing (cited) factions. The author factions are manually labeled.

We see from Table 1, the model has selected keywords that are arguably significant in certain subfields in computational linguistics. Incoming citations are generally indicative of the subject areas in

[2] For a list of the journals, conferences and workshops archived by the ACL anthology, please visit `http://aclweb.org/anthology-new`.

[3] In future work, nonparametric priors might be employed to automate the selection of G.

[4] We found it quite difficult to make sense of terms with *negative* log frequency deviations. This suggests exploring a model allowing only positive deviations; we leave that for future work.

Formalisms (31)	*Fernando Pereira, Jason M. Eisner, Stuart M. Shieber, Walter Daelemans, Hitoshi Isahara*
Self cites:	parsing
In cites:	parsing, semiring, grammars, tags, grammar, tag, lexicalized, dependency
Out cites:	tagger, regular, dependency, transformationbased, tagging, stochastic, grammars, sense
Evaluation (17)	*Salim Roukos, Eduard Hovy, Marti A. Hearst, Chin-Yew Lin, Dekang Lin*
Self cites:	automatic, bleu, linguistics, evaluation, computational, text, proceedings
In cites:	automatic, bleu, segmentation, method, proceedings, dependency, parses, text
Out cites:	paraphrases, cohesion, agreement, hierarchical, entropy, phrasebased, evaluation, treebank
Semantics (26)	*Martha Palmer, Daniel Jurafsky, Mihai Surdeanu, David Weir, German Rigau*
Self cites:	sense, semantic, wordnet
In cites:	framenet, sense, semantic, task, wordnet, word, project, question
Out cites:	sense, wordnet, moses, preferences, distributional, semantic, focus, supersense
Machine Translation (MT1) (9)	*Kevin Knight, Michel Galley, Jonathan Graehl, Wei Wang, Sanjeev P. Khudanpur*
Self cites:	inference, scalable, model
In cites:	scalable, inference, machine, training, generation, translation, model, syntaxbased
Out cites:	phrasebased, hierarchical, inversion, forest, transduction, translation, ibm, discourse
Word Sense Disambiguation (WSD) (42)	*David Yarowsky, Rada Mihalcea, Eneko Agirre, Ted Pedersen, Yorick Wilks*
Self cites:	sense, word
In cites:	sense, preferences, wordnet, acquired, semcor, word, semantic, calle
Out cites:	sense, subcategorization, acquisition, automatic, corpora, lexical, processing, wordnet
Parsing (20)	*Michael John Collins, Eugene Charniak, Mark Johnson, Stephen Clark, Massimiliano Ciaramita*
Self cites:	parser, parsing, model, perceptron, parsers, dependency
In cites:	parser, perceptron, supersense, parsing, dependency, results, hmm, models
Out cites:	parsing, forest, treebank, model, coreference, stochastic, grammar, task
Discourse (29)	*Daniel Marcu, Aravind K. Joshi, Barbara J. Grosz, Marilyn A. Walker, Bonnie Lynn Webber*
Self cites:	discourse, structure, centering
In cites:	discourse, phrasebased, centering, tag, focus, rhetorical, tags, lexicalized
Out cites:	discourse, rhetorical, framenet, realizer, tags, resolution, grammars, synonyms
Machine Translation (MT2) (9)	*Franz Josef Och, Hermann Ney, Mitchell P. Marcus, David Chiang, Dekai Wu*
Self cites:	training, error
In cites:	error, giza, rate, alignment, training, minimum, translation, phrasebased
Out cites:	forest, subcategorization, arabic, model, translation, machine, models, heuristic

Table 1: Key authors and citation words associated with some factions. For each faction, we show the 5 authors with highest expected incoming citations (i.e p(faction | author) × citations). Factions are labeled manually, referring to key sub-fields in computational linguistics. Faction sizes are in parenthesis following the labels. The citation words with the strongest positive weights in the deviation vectors are shown.

which the faction holds recognized expertise. For instance, the faction labeled "semantics" has citation terms commonly associated with propositional semantics: *sense*, *framenet*, *wordnet*. On the other hand, outgoing citations hint at the related work that a faction builds on; discourse might require building on components involving *framenet*, *grammars*, *synonyms*, while word sense disambiguation involves solving problems like *acquisition* and modeling *subcategorization*.

4.3 Sensitivity

Given the same initial parameters, we found our model to be fairly stable across iterations of Monte

Adam Lopez, Paul S. Jacobs (2)
Regina Barzilay, Judith L. Klavans, Robert T. Kasper (3)
Lauri Karttunen, Kemal Oflazer, Kimmo Koskenniemi (3)
John Carroll, Ted Briscoe, Scott Miller (7)
Vincent J. Della Pietra, Stephen A. Della Pietra, Robert L. Mercer (25)
Thorsten Brants, Liang Huang, Anoop Sarkar (9)
Christoph Tillmann, Kenji Yamada, Sharon Goldwater (7)
Alex Waibel, Keh-Jiann Chen, Katrin Kirchhoff (3)
Lynette Hirschman, Claire Cardie, Vincent Ng (26)
Erik F. Tjong Kim Sang, Ido Dagan, Marius Paşca (21)
Yuji Matsumoto, Dragomir R. Radev, Chew Lim Tan (18)
Christopher D. Manning, Owen Rambow, Ellen Riloff (19)
Richard Zens, Hieu Hoang, Nicola Bertoldi (9)
Dan Klein, Jun'ichi Tsujii, Yusuke Miyao (6)
Janyce Wiebe, Mirella Lapata, Kathleen R. McKeown (50)
I. Dan Melamed, Ryan McDonald, Joakim Nivre (10)
Philipp Koehn, Lillian Lee, Chris Callison-Burch (80)
Kenneth Ward Church, Eric Brill, Richard M. Schwartz (19)

Table 2: Top 3 authors of the remaining 18 factions not displayed in Table 1.

Carlo EM. We found that when G was too small (e.g., 10), groups were more mixed and the $\boldsymbol{\eta}$ vectors could not capture variation among them well. When G was larger, the factions were subjectively cleaner, but fields like translation split into many factions (as is visible in the $G = 30$ case illustrated in Tables 1 and 2. Strengthening the L_1 penalty made $\boldsymbol{\eta}$ more sparse, of course, but gave less freedom in fitting the data and therefore more grouping of authors into a fewer effective factions.

4.4 Inter-Faction Relationships

By using the most probable *a posteriori* faction for each author, we can compute the number of citations between factions. We define the average inter-faction citations by:

$$\text{IFC}(g, h) = \frac{\Psi(g \rightarrow h) + \Psi(h \rightarrow g)}{N_g + N_h} \qquad (1)$$

where $\Psi(g \rightarrow h)$ is the total number of papers written by authors in g that cite papers written by authors in h.

Figure 2 presents a graph of selected factions and how these factions talk about each other. As we would expect, the machine translation faction is quite strongly connected to formalisms and parsing factions, reflecting the heavy use of grammars and parsing algorithms in translation. Moreover, we can observe that "deeper" linguistics research, such as semantics and discourse, are less likely to be cited by the other factions. This is reflected in Figure 3, where the statistical MT and parsing factions in the bottom left exhibit higher citation activity amongst each other. In addition, we note that factions tend to self-cite more often than out of their own factions; this is unsurprising given the prior we selected.

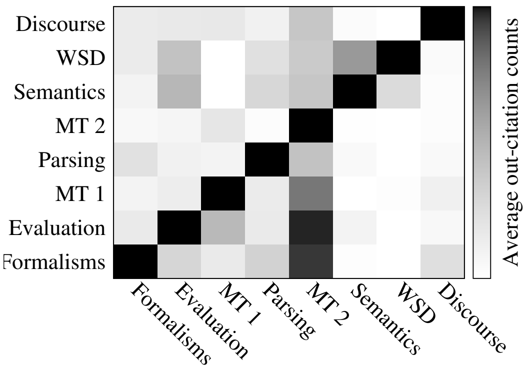

Figure 3: Heat map showing citation rates across selected factions. Factions on the horizontal axis are being cited; factions on the vertical axis are citing. Darker shades denote higher average $\frac{\Psi(g \rightarrow h)}{N_g}$.

The IFC between discourse and MT2 (as shown by the edge thickness in figure 2) is higher than expected, given our prior knowledge of the computational linguistics community. Further investigation revealed that, Daniel Marcu, posited by our model to be a member of the discourse faction, has coauthored numerous highly cited papers in MT in recent years (Marcu and Wong, 2002). However, the model split the translation field, which fragmented the counts of MT related citation words. Thus, assigning Daniel Marcu to the discourse faction, which also has a less diverse citation vocabulary, is more probable than assigning him to one of the MT factions. In §4.6, we consider a model of factions over time to mitigate this problem.

4.5 Comparison to Graph Clustering

Work in the field of bibliometrics has largely focused on using the link structure of citation networks to study higher level structures. See Osareh (1996) for a review. Popular methods include bibliographic coupling (Kessler, 1963), and co-citation

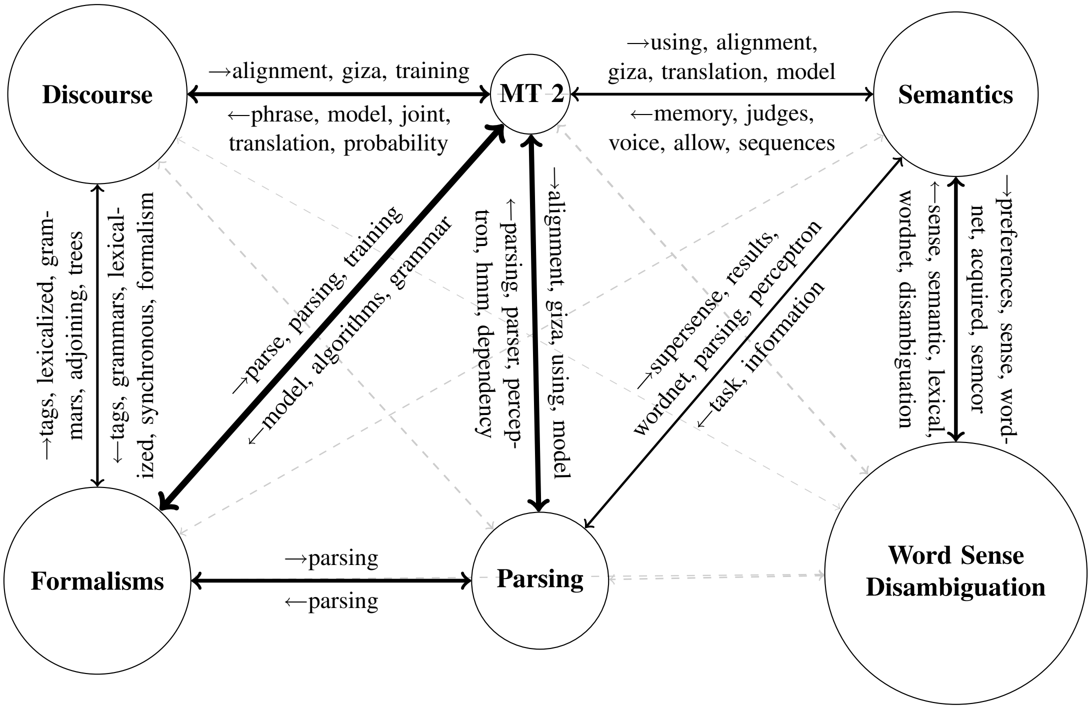
Discourse
MT 2
Semantics
Formalisms
Parsing
Word Sense Disambiguation
→alignment, giza, training
←phrase, model, joint, translation, probability
→using, alignment, giza, translation, model
←memory, judges, voice, allow, sequences
→tags, lexicalized, grammars, adjoining, trees
←tags, grammars, lexicalized, synchronous, formalism
→parse, parsing, training
←model, algorithms, grammar
←parsing, parser, perceptron, hmm, dependency
→alignment, giza, using, model
→supersense, results, wordnet, parsing, perceptron
←task, information
←sense, semantic, lexical, wordnet, disambiguation
→preferences, sense, wordnet, acquired, semcor
→parsing
←parsing

Figure 2: Citations among some factions. The size of a node is relative to the faction size and edge thickness is relative to the average number of inter-faction citations (equation 1). The words on the edges are the highest weighted words from the deviation vectors $\boldsymbol{\eta}$, with the arrow denoting the direction of the citation. Edges with below average IFC scores are represented as dashed lines, and their citations words are not shown to preserve readability.

analysis (Small, 1973). By using authors as an unit of analysis in co-citation pairs, author co-citations have been presented as a technique to analyze their subject specialties (White and Griffith, 1981). Using standard graph clustering algorithms on these author co-citation networks, one can obtain a semblance of author factions. Hence, we performed graph clustering on both collaboration and citation graphs[5] of authors in our dataset using Graclus[6], a graph clustering implementation based on normalized cuts and ratio associations (Dhillon et al., 2004).

In Table 3, we compare, for selected authors, how their faction-mates obtained by our model and graph clustering differ. When clustering on the author collaboration network, we obtained some clusters easily identified with research labs (e.g., Daniel Marcu at the Information Sciences Institute). The co-citation graph leads to groupings dominated by heavily co-cited papers in major research areas. While we do not have an objective measurement of quality or usefulness, we believe that the factions identified by our model align somewhat better with familiar technical themes around which sub-communities naturally form than major research problems or institutions.

4.6 Factions over Time

Faction alignments may be dynamic; we expect that, over time, individual researchers may move from one faction to another as their interests evolve. We consider a slightly modified model whereby authors are split into different copies of themselves during a non-overlapping set of discrete time periods. Given a set of disjoint time periods T, we denote each author-faction node by $\{a^{(i,t)} \mid (i,t) \in A \times T\}$. As we treat each "incarnation" of an author as a distinct individual, we can simply use the same inference algorithm described in §2. (In future work we might impose an expectation of gradual changes along a more continuous representation of time.)

[5] We converted the directed citation graph into a symmetric graph by performing bibliometric symmetrization described in Satuluri and Parthasarathy (2011, section 3.3).

[6] http://www.cs.utexas.edu/users/dml/Software/graclus.html

Our Model	Collaboration Network	Co-citation Network
Franz Josef Och		
Franz Josef Och, Hermann Ney, Mitchell P. Marcus, David Chiang, Dekai Wu error, giza, rate, alignment, training	Franz Josef Och, Hermann Ney, Richard Zens, Stephan Vogel, Nicola Ueffing giza, mert, popovic, moses, alignments	Franz Josef Och, Hermann Ney, Vincent J. Della Pietra, Daniel Marcu, Robert L. Mercer giza, bleu, phrasebased, alignment, mert
Daniel Marcu		
Daniel Marcu, Aravind K. Joshi, Barbara J. Grosz, Marilyn A. Walker, Bonnie Lynn Webber discourse, phrasebased, centering, tag, focus	Daniel Marcu, Kevin Knight, Daniel Gildea, David Chiang, Liang Huang phrasebased, forest, cube, spmt, hiero	Franz Josef Och, Hermann Ney, Vincent J. Della Pietra, Daniel Marcu, Robert L. Mercer giza, bleu, phrasebased, alignment, mert
Michael John Collins		
Eugene Charniak, Michael John Collins, Mark Johnson, Stephen Clark, Massimiliano Ciaramita parser, perceptron, supersense, parsing, dependency	Michael John Collins, Joakim Nivre, Lluís Márquez, Xavier Carreras, Jan Hajič pseudoprojective, maltparser, perceptron, malt, averaged	Michael John Collins, Christopher D. Manning, Dan Klein, Eugene Charniak, Mark Johnson tnt, prototypedriven, perceptron, coarsetofine, pcfg
Kathleen R. McKeown		
Mirella Lapata, Janyce Wiebe, Kathleen R. McKeown, Dan Roth, Ralph Grishman semantic, work, learning, corpus, model	Kathleen R. McKeown, Regina Barzilay, Owen Rambow, Marilyn A. Walker, Srinivas Bangalore centering, arabic, pyramid, realpro, cue	Kenneth Ward Church, David Yarowsky, Eduard Hovy, Kathleen R. McKeown, Lillian Lee rouge, minipar, nltk, alignment, montreal

Table 3: Comparing selected factions between our model and graph clustering algorithms. Authors with highest incoming citations are shown. For our model, we show the largest weighted words in the SAGE vector of incoming citations for the faction, while for graph clustering, we show words with the highest tf-idf weight.

We split the same data as the earlier sections into four disjoint time periods, 1965–1989, 1990–1999, 2000–2005 and 2006–2011. The split across time is unequal due to the number of papers published in each period: these four periods include 1,917, 3,874, 3,786, and 8,105 papers, respectively. Here we used $G = 20$ factions for faster runtime, leading to diminished interpretability, though the sparsity of the deviation vectors mitigates this problem somewhat. Figure 4 shows graphical plots of selected authors and their faction membership posteriors over time (drawn from the final E-step).

With a simple extension of the original model, we can learn shifts in the subject area the author is publishing about. Consider Eugene Charniak: the model observed a major change in faction alignment around 2000, when one of the popular Charniak parsers (Charniak, 2000) was released; this is somewhat later than Charniak's interests shifted, and the earlier faction's words are not clearly an accurate description of his work at that time. More fine-grained modeling of time and also accounting for the death and birth of factions might ameliorate these inconsistencies with our background knowledge about Charniak. The model finds that Aravind Joshi was associated with the tagging/parsing faction in the 1990s and in recent years moved back towards discourse (Prasad et al., 2008). David Yarowsky, known for his early work on word sense disambiguation, has since focused on applying word sense disambiguation techniques in a multilingual context (Garera et al., 2009; Bergsma et al., 2011). As mentioned in the previous section, we observe that the extended model is able to capture Daniel Marcu's shift from discourse-related work to MT with his work in phrase-based statistical MT (Marcu and Wong, 2002).

5 Related Work

A number of algorithms use topic modeling to analyze the text in the articles. Topic models such as latent Dirichlet allocation (Blei et al., 2003) and its variations have been increasingly used to study trends in scientific literature (McCallum et al., 2006; Dietz et al., 2007; Hall et al., 2008; Gerrish and Blei, 2010), predict citation information (McNee et al.,

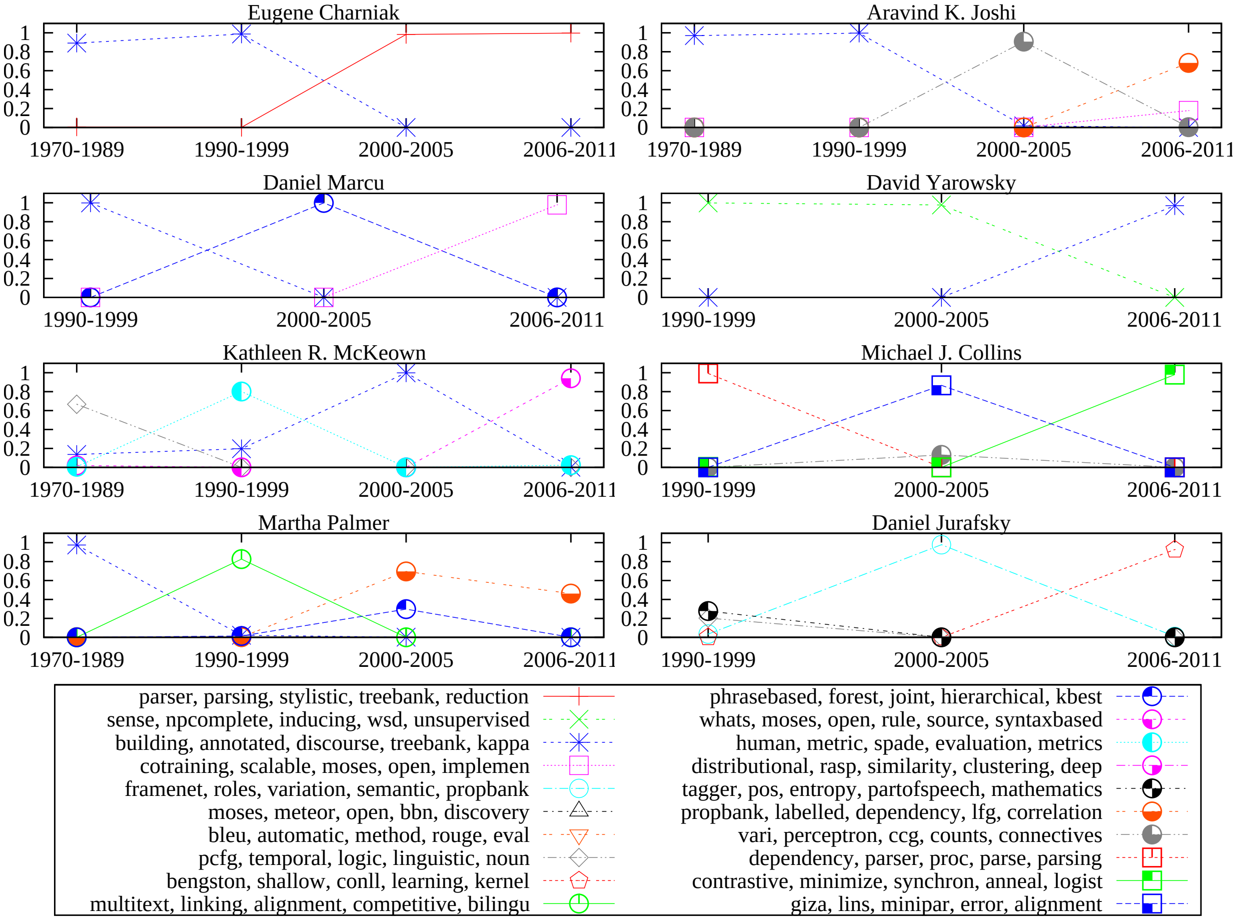

Figure 4: Posterior probability of faction alignment over time periods for eight researchers with significant publication records in at least three periods. The key for each entry contains the five highest weighted words in the deviation vectors for the faction's incoming citations. For each author, we show factions with which he or she is associated with probability > 0.1 in at least one time period.

2002; Ibáñez et al., 2009; Nallapati et al., 2008) and analyze authorship (Rosen-Zvi et al., 2004; Johri et al., 2011).

Assigning author factions can be seen as network classification problem, where the goal is to label nodes in a network such that there is (i) a correlation between a node's label and its observed attributes and (ii) a correlation between labels of interconnected nodes (Sen et al., 2008). Such collective network-based approaches have been used on scientific literature to classify papers/web pages into its subject categories (Kubica et al., 2002; Getoor, 2005; Angelova and Weikum, 2006). If we knew the word distributions between factions beforehand, learning the author factions in our model would be equivalent to the network classification task, where our edge weights are proportional to the probability of coauthorship multiplied by the probability of observing the citation words given the author's faction labels.

6 Conclusion

In this work, we have defined factions in terms of how authors talk about each other's work, going beyond co-authorship and citation graph representations of a research community. We take a first step toward computationally modeling faction formation by using a latent author faction model and applied it to the ACL community, revealing both factions and how they cite each other. We also extended the model to capture authors' faction changes over time.

Acknowledgments

The authors thank members of the ARK group and the anonymous reviewers for helpful feedback. We gratefully acknowledge technical assistance from Matthew Fiorillo. This research was supported in part by an A*STAR fellowship to Y. Sim, NSF grant IIS-0915187 to N. Smith, and the Center for Intelligent Information Retrieval and NSF grant IIS-0910884 for D. Smith.

References

G. Andrew and J. Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proc. of ICML*.

R. Angelova and G. Weikum. 2006. Graph-based text classification: learn from your neighbors. In *Proc. of SIGIR*.

S. Bergsma, D. Yarowsky, and K. Church. 2011. Using large monolingual and bilingual corpora to improve coordination disambiguation. In *Proc. of ACL*.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

E. Charniak. 2000. A maximum-entropy-inspired parser. In *Proc. of NAACL*.

I. S. Dhillon, Y. Guan, and B. Kulis. 2004. Kernel k-means: spectral clustering and normalized cuts. In *Proc. of KDD*.

L. Dietz, S. Bickel, and T. Scheffer. 2007. Unsupervised prediction of citation influences. In *Proc. of ICML*.

J. Eisenstein, A. Ahmed, and E. P. Xing. 2011. Sparse additive generative models of text. In *Proc. of ICML*.

M. A. T. Figueiredo. 2003. Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1150–1159.

N. Garera, C. Callison-Burch, and D. Yarowsky. 2009. Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In *Proc. of CoNLL*.

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

L. Getoor. 2005. Link-based classification. In *Advanced Methods for Knowledge Discovery from Complex Data*, pages 189–207. Springer.

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

A. Ibáñez, P. Larrañaga, and C. Bielza. 2009. Predicting citation count of bioinformatics papers within four years of publication. *Bioinformatics*, 25(24):3303–3309.

N. Johri, D. Ramage, D. A. McFarland, and D. Jurafsky. 2011. A study of academic collaborations in computational linguistics using a latent mixture of authors model. In *Proc. of the ACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*.

M. M. Kessler. 1963. Bibliographic coupling between scientific papers. *American documentation*, 14(1):10–25.

J. Kubica, A. Moore, J. Schneider, and Y. Yang. 2002. Stochastic link and group detection. In *Proc. of AAAI*.

K. Lange and J. S. Sinsheimer. 1993. Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics*, 2(2):175–198.

D. Marcu and W. Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proc. of EMNLP*.

A. McCallum, G. S. Mann, and D. Mimno. 2006. Bibliometric impact measures leveraging topic analysis. In *Proc. of JCDL*.

S. M. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S. K. Lam, A. M. Rashid, J. A. Konstan, and J. Riedl. 2002. On the recommending of citations for research papers. In *Proc. of CSCW*.

T. P. Minka. 2009. Estimating a Dirichlet distribution. Available online at `http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/minka-dirichlet.pdf`.

R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen. 2008. Joint latent topic models for text and citations. In *Proc. of KDD*.

F. Osareh. 1996. Bibliometrics, citation analysis and co-citation analysis: A review of literature I. *Libri*, 46(3):149–158.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008. The Penn discourse treebank 2.0. In *Proc. of LREC*.

D. R. Radev, M. T. Joseph, B. Gibson, and P. Muthukrishnan. 2009a. A bibliometric and network analysis of the field of computational linguistics. *Journal of the American Society for Information Science and Technology*.

D. R. Radev, P. Muthukrishnan, and V. Qazvinian. 2009b. The ACL Anthology Network corpus. In *Proceedings of the Workshop on Text and Citation Analysis for Scholarly Digital Libraries*.

F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, and G. Parisi. 2004. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2658–2663.

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

V. Satuluri and S. Parthasarathy. 2011. Symmetrizations for clustering directed graphs. In *Proc. of International Conference on Extending Database Technology*.

P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad. 2008. Collective classification in network data. *AI magazine*, 29(3):93.

H. Small. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, 24(4):265–269.

R. Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

G. C. G. Wei and M. 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):699–704.

H. D. White and B. C. Griffith. 1981. Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32(3):163–171.

D. Yogatama, M. Heilman, B. O'Connor, C.Dyer, B. R. Routledge, and N. A. Smith. 2011. Predicting a scientific community's response to an article. In *Proc. of EMNLP*.