Mixed Membership Models of Scientific Publications

ELENA EROSHEVA*, STEPHEN FIENBERG†§, AND JOHN LAFFERTY‡§

*Department of Statistics, School of Social Work, and the Center for Statistics and the Social Sciences, University of Washington, Seattle, WA 98195
†Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213
‡Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213
§Center for Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh, PA 15213

The *Proceedings of the National Academy of Sciences* (PNAS) is one of world's most cited multidisciplinary scientific journals. The PNAS official classification structure of subjects is reflected in topic labels submitted by the authors of manuscripts, largely related to traditionally established disciplines. These include broad field classifications into Physical Sciences, Biological Sciences, Social Sciences, and further subtopic classifications within the fields. Focusing on Biological Sciences, we explore an internal soft classification structure of articles based only on semantic decompositions of abstracts and bibliographies, and compare it with the formal discipline classifications.

Our model assumes that there is a fixed number of internal categories, each characterized by multinomial distributions over words (in abstracts) and references (in bibliographies). Soft classification for each article is based on proportions of the article's content coming from each category. We discuss the appropriateness of the model for the PNAS database as well as other features of the data relevant to soft classification.

"The Proceedings is there to help bring new ideas promptly into play. New ideas may not always be right, but their prominent presence can lead to correction. We must be careful not to censor those ideas which seem to be off beat." Saunders MacLane, PNAS, Vol. 94, pp. 5983-5985, June 1997

Are there internal categories of papers in the *Proceedings of the National Academy of Sciences* that we can obtain empirically with statistical data mining tools based only on semantic decompositions of words and references used? Can we identify MacLane's "off-beat" but potentially path-breaking PNAS papers using these internal categories? Do these empirically defined categories correspond in some natural way to the classification by field used to organize the papers for publication, or does PNAS publish substantial numbers of interdisciplinary papers that transcend these disciplinary boundaries? These are examples of questions that our contribution to the mapping of knowledge domains represented by PNAS attempts to explore.

A number of mathematical and statistical techniques have been developed for analyzing complex data in ways that could reveal underlying data patterns through some form of classification. Computational advances have made some of these techniques extremely popular in recent years. For example, two of the ten most cited articles from 1997-2001 PNAS publications are on applications of clustering for gene expression patterns [1, 2]. The traditional assumption in most methods that aim to discover knowledge in underlying data patterns has been that each subject (object or individual) from the population of interest inherently belongs to only one of the underlying subpopulations (clusters, classes, aspects, or pure type categories). This implies that a subject shares all of its attributes, usually with some degree of uncertainty, with the subpopulation to which it belongs. Given that a relatively small number of subpopulations is often necessary for a meaningful interpretation of the underlying patterns, many data collections do not conform with the traditional assumption. Subjects in such populations may combine attributes from several subpopulations simultaneously. In other words, they may have a mixed collection of attributes originating from more than one subpopulation.

Several different disciplines have developed approaches that have a common statistical structure which we refer to as mixed membership. In genetics, mixed membership models can account for the fact that individual genotypes may come from different subpopulations according to (unknown) proportions of an individual's ancestry. Rosenberg et al. [3] use such a model to analyze genetic samples from 52 human populations around the globe,

identifying major genetic clusters without using the geographic information about the origins of individuals. In the Social Sciences, such models are natural since members of a society can exhibit mixed membership with respect to the underlying social or health groups for a particular problem being studied. Hence, individual responses to a series of questions may have mixed origins. Woodbury et al. [4] employ this idea to develop medical classification. In text analysis and information retrieval, mixed membership models have been used to account for different topical aspects of individual documents.

In the following section we describe a class of mixed membership models that unifies existing special cases [5]. We then explain how this class of models can be adapted to analyze both the semantic content of a document and its citations of other publications. We fit this document-oriented mixed membership model to a subcollection of the PNAS database supplied to the participants in the Arthur M. Sackler Colloquium on "Mapping Knowledge Domains." We focus in our analysis on a high level description of the fields in Biological Sciences in terms of a small number of extreme or basis categories. Griffiths and Steyvers [6] use a related version of the model for abstracts only and attempt a finer level of description.

Mixed Membership Models

The general mixed membership model we work with relies on four levels of assumptions: population, subject, latent variable, and sampling scheme. Population level assumptions describe the general structure of the population that is common to all subjects. Subject level assumptions specify the distribution of observable responses given individual membership scores. Membership scores are usually unknown and hence can also be viewed as latent variables. The next assumption is whether the membership scores are treated as fixed or random in the model. Finally, the last level of assumption specifies the number of distinct observed characteristics (attributes) and the number of replications for each characteristic. We describe each set of assumptions formally in turn.

Population level. Assume there are K original or basis subpopulations in the populations of interest. For each subpopulation k, denote by $f(x_j|\theta_{kj})$ the probability distribution for response variable j, where θ_{kj} is a vector of parameters. Assume that within a subpopulation, responses to observed variables are independent.

Subject level. For each subject, membership vector $\lambda = (\lambda_1, \ldots, \lambda_K)$ provides the degrees of a subject's membership in each of the subpopulations. The probability distribution of observed responses x_j for each subject is fully defined by the conditional probability $Pr(x_j|\lambda) = \sum_k \lambda_k f(x_j|\theta_{kj})$, and the assumption that response variables x_j are independent, conditional on membership scores. In addition, given the membership scores, observed responses from different subjects are independent.

Latent variable level. With respect to the latent variables, one could either assume that they are fixed unknown constants or that they are random realizations from some underlying distribution.

1. If the membership scores λ are fixed but unknown, the conditional probability of observing x_j, given the parameters $\boldsymbol{\theta}$ and membership scores, is

$$Pr(x_j|\lambda;\boldsymbol{\theta}) \quad = \quad \sum_{k=1}^{K} \lambda_k f(x_j|\theta_{kj}). \tag{1}$$

2. If membership scores λ are realizations of latent variables from some distribution D_α, parameterized by vector α, then the probability of observing x_j, given the parameters, is:

$$Pr(x_j|\alpha,\boldsymbol{\theta}) \quad = \quad \int \left(\sum_{k=1}^{K} \lambda_k f(x_j|\theta_{kj}) \right) dD_\alpha(\lambda). \tag{2}$$

Sampling scheme. Suppose R independent replications of J distinct characteristics are observed for one subject, $\{x_1^{(r)}, \ldots, x_J^{(r)}\}_{r=1}^R$. Then, if the membership scores are treated as realizations from distribution D_α, the conditional probability is

$$Pr\left(\{x_1^{(r)}, \ldots, x_J^{(r)}\}_{r=1}^R | \alpha, \boldsymbol{\theta}\right) = \int \left(\prod_{j=1}^{J} \prod_{r=1}^{R} \sum_{k=1}^{K} \lambda_k f(x_j^{(r)} | \theta_{kj}) \right) dD_\alpha(\lambda). \quad (3)$$

When the latent variables are treated as unknown constants, the conditional probability for observing R replications of J variables can be derived analogously. In general, the number of observed characteristics J need not be the same across subjects, and the number of replications R need not be the same across observed characteristics.

One can derive examples of mixed membership models from this general set up by specifying different choices of J and R, and different latent variable assumptions. Thus, the *Grade of Membership* model of Manton et al. [7] assumes polytomous responses are observed to J survey questions without replications and uses the fixed-effects assumption for the membership scores. Potthoff et al. [8] employ a variation of the Grade of Membership model by treating the membership scores as Dirichlet random variables; the authors refer to the resulting model as *Dirichlet generalization of latent class models*. Erosheva [5] provides a formal latent class representation for the Grade of Membership model approach. In genetics, Pritchard et al. [9] use a *clustering model with admixture*. For diploid individuals the clustering model assumes that $R = 2$ replications (genotypes) are observed at J distinct locations (loci), treating the proportions of a subject's genome that originated from each of the basis subpopulations as random Dirichlet realizations. Variations of mixed membership models for text documents called *probabilistic latent semantic analysis* [10] and *latent Dirichlet allocation* [11] both assume that a single characteristic (word) is observed a number of times for each document, but the former model considers the membership scores as fixed unknown constants, whereas the latter treats them as random Dirichlet realizations.

The mixed membership model framework presented above unifies several specialized models that have been developed independently in the social sciences, genetics, and text mining applications. In the text mining area, initial work by Hofmann on *probabilistic latent semantic analysis* [10] was followed by the work of Blei et al. [11], who proposed a Dirichlet generating distribution for the membership scores and the use of variational methods to estimate the *latent Dirichlet allocation* model parameters. Minka and Lafferty [12] develop a more accurate approximation method for this model.

A natural extension of the original analyses in the text mining area that have been based on a single source is to combine information from multiple sources. Cohn and Hofmann [13] propose a probabilistic model of document content and hypertext connectivity for text documents by considering links (or references) in addition to words, thus, essentially combining two distinct characteristics; they treat the membership scores as fixed. Following Cohn and Hofmann, we adopt a mixed membership model for words and references in journal publications but treat the membership scores as random Dirichlet realizations. Barnard et al. [14] develop similar and alternative approaches for combining different sources of information.

Mixed Membership Models for Documents

We can use the general model framework for documents consisting of abstracts and references by representing a document as $d = (\{x_1^{(r_1)}\}, \{x_2^{(r_2)}\})$ where $x_1^{(r_1)}$ is a word (w) in the abstract and $x_2^{(r_2)}$ is a reference (r) in the bibliography, $r_j = 1, \ldots, R_j$. By adopting the "bag of words" assumption, we treat the words in each abstract as independent replications of the first observed characteristic (word). Similarly, under the assumption of a "bag of references" we treat references as independent replications of the second observed characteristic (reference). Thus, the representation of a document consists of word counts $n(w, d)$ (the number of times word w appears in document d) and reference counts $n(r, d)$ (1 if the bibliography of d contains a reference to r, and 0 otherwise). In this context, subpopulations refer to topical aspects.

The parameters $\boldsymbol{\theta}$ of our model are:

$$\text{Dirichlet parameters:} \quad \alpha_1, \ldots, \alpha_K, \tag{4}$$

$$\text{word aspect parameters:} \quad \theta_{1k}(w) = p(w \mid k),\ k = 1, 2, \ldots, K, \tag{5}$$

$$\text{reference aspect parameters:} \quad \theta_{2k}(r) = q(r \mid k),\ k = 1, 2, \ldots, K. \tag{6}$$

In the generative model, documents $d = (\{x_1^{(r_1)}\}, \{x_2^{(r_2)}\})$ are sampled according to the following sequence:

$$\lambda \sim \text{Dirichlet}(\alpha), \tag{7}$$

$$x_1^{(r_1)} \sim \text{Multinomial}(p_\lambda),\ \text{where } p_\lambda = \sum_{k=1}^{K} \lambda_k\, \theta_{1k}, \tag{8}$$

$$x_2^{(r_2)} \sim \text{Multinomial}(q_\lambda),\ \text{where } q_\lambda = \sum_{k=1}^{K} \lambda_k\, \theta_{2k}, \tag{9}$$

where $\sum_w \theta_{1k}(w) = 1$ and $\sum_r \theta_{2k}(r) = 1$, $k = 1, \ldots, K$. Since distributions of words and references in a document are convex combinations of the aspects' distributions, the aspects can be thought of as extreme or basis categories for a collection of documents. The sampling of words and references in the model can also be interpreted as a latent classification process where an aspect of origin is drawn first for each word and for each reference in a document, according to a multinomial distribution parameterized by the document-specific membership scores λ, and words and references are then generated from corresponding distributions of the aspects of origin [5]. Rather than a mixture of K latent classes, the model can be thought of as a "simplicial mixture" [12] since the word and reference probabilities range over a simplex with corners θ_{1k} and θ_{2k}, respectively.

The likelihood function is thus

$$p(\boldsymbol{\theta} \mid d) = \int_{\Delta_{K-1}} \text{Dir}(\lambda \mid \alpha) \prod_w p_\lambda(w)^{n(w,d)} \prod_r q_\lambda(r)^{n(r,d)}\, d\lambda \tag{10}$$

$$= \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int_{\Delta_{K-1}} \prod_{i=1}^{k} \lambda_i^{\alpha_i - 1} \prod_w p_\lambda(w)^{n(w,d)} \prod_r q_\lambda(r)^{n(r,d)}\, d\lambda \tag{11}$$

where Δ_{K-1} is the $(K-1)$-simplex.

It is important to note that the assumption of exchangeability among words and references (conditional independence given the membership scores) does not imply joint independence among the observed characteristics. Instead, the assumption of exchangeability means that dependencies among words and references can be fully explained by the documents' membership scores. For an extended discussion on exchangeability in this context, see [15].

Alternative Models for References

For the analysis of PNAS publications in the next section, we assume multinomial sampling of words and references. While multinomial sampling is computationally convenient, it is not a realistic model of the way in which authors select references for the bibliography of a paper. We briefly describe examples of more realistic generative assumptions for references.

Suppose an article focuses on a sufficiently narrow scientific area. In this case the authors may have essentially perfect knowledge of the literature and thus they would pay separate attention to each paper in their pool of references as they consider whether to include it in the bibliography. Under these circumstances, given that the pool of references contains R papers, we assume that a document is represented as $d = (\{x_1^{(r_1)}\}, x_2, x_3, \ldots, x_{R-1})$ where $x_1^{(r_1)}$ is a word (w) in the abstract, R is the number of references, and $x_2, \ldots, x_{R-1}$ are all references in the pool. Reference counts do not change: they are given by $n(r, d) = 1$, if the bibliography of d contains a reference to r, and by $n(r, d) = 0$ otherwise.

Then our model for generating documents would be to sample $d = (\{x_1^{(r_1)}\}, x_2, x_3, \ldots, x_{R-1})$ according to:

$$\lambda \sim \text{Dirichlet}(\alpha) \tag{12}$$

$$x_1^{(r_1)} \sim \text{Multinomial}(p_\lambda), \text{ where } p_\lambda = \sum_{k=1}^{K} \lambda_k \, \theta_{1k} \tag{13}$$

$$x_j \sim \text{Bernoulli}(q_\lambda(x_j)), \text{ where } q_\lambda(x_j) = \sum_{k=1}^{K} \lambda_k \, \theta_{jk}, \; j = 2, 3, \ldots, R-1, \tag{14}$$

where $\sum_w \theta_{1k}(w) = 1, \; k = 1, \ldots, K$. The likelihood function based on this alternative model would not only take into account which documents contain which references, but it would also incorporate the information about which references documents do not contain.

Alternatively we could consider a reference list as being generated by a two-step combination of multinomial and Bernoulli draws where authors first select a pool of references determined by the compositional structure of the research (multinomial sampling) and then decide which references are most relevant for the current document and which are not (Bernoulli sampling). Yet another alternative is to assume that the authors have knowledge of importance only regarding those papers that are of similar decomposition. The probabilities of including a reference to a paper that is within an area of expertise may then depend on the contextual decomposition of citing that paper (multinomial or Bernoulli sampling). The probabilities of giving a reference to a paper that is outside of the area of expertise may then be considered a constant, which is equivalent to saying that references to papers that are outside of the area of expertise occur by chance.

Both the basic model for references and any alternatives would still need to reflect the time ordering on publications and include in the pool of possible references only those that have already been published, perhaps even with a short time lag. But even such changes are unlikely to produce a "correct" model for citation practices. As Box [16] reminds us, no model can be right but some models are more sensible and useful than others.

Estimating the Model

The primary complication in using a mixed membership model such as (7)–(9) where the membership probabilities are random rather than fixed, is that the integral in (10) cannot be computed explicitly and therefore must be approximated. Two approximation schemes have been recently investigated for this problem, and the associated problem of fitting the model. In the variational approach [11], the mixture terms $p_\lambda(w) = \sum_{k=1}^{K} \lambda_k \, \theta_{1k}(w)$ are bounded from below in a product form that leads to a tractable integral; the lower bound is then maximized. A related approach, called Expectation-Propagation (EP) [12], also approximates each mixture term in a product form, but chooses the parameters of the factors by matching first and second moments. Either of these approximations to the integral (10) can be used in an approximate EM algorithm to estimate the parameters of the models. It is shown in [12] that EP in general leads to better approximations than the simple variational method for mixed membership models, although we have obtained comparable results with both approaches on the PNAS collection. The results reported below use the variational approximation.

The PNAS Database

The *National Academy of Sciences* provided the database for the participants of the Arthur M. Sackler Colloquium "Mapping Knowledge Domains." We have focused on a subset which contains all Biological Sciences articles in volumes 94–98 of the Proceedings which correspond to Julian years 1997-2001, thereby ignoring articles published in the Social and Physical Sciences unless they have official dual classifications with one classification in the Biological Sciences. The reason for this narrowing of focus is twofold. First, the major share of Proceedings publications in recent years represents research developments in the Biological Sciences. Thus, out of 13,008 articles published in 94–98 volumes, 12,036 or 92.53% are in the Biological Sciences. The share of Social and Physical Sciences articles in volumes 94–98 is a much more modest 7.47%. Second, we assume a collection of papers is characterized by mixed membership in a number of internal categories, and Social and Physical Sciences

papers are unlikely to share the same internal categories with papers from the Biological Sciences. We also automatically ignore other types of PNAS publications such as corrections, commentaries, letters, and reviews because these are not traditional research reports. Among the Biological Sciences papers in our database, eleven articles were not processed because they did not have an abstract, and one article was not processed because it did not contain any references.

PNAS is one of world's most cited multidisciplinary scientific journals. Historically, when submitting a research paper to the Proceedings, authors have to select a major category from Physical, Biological, or Social Sciences, and a minor category from the list of topics. The Proceedings permits dual classifications between major categories and, in exceptional cases, within a major category. The lists of topics change over time to reflect changes in the Academy sections. As stated in the PNAS information for authors revised in June 2002, official classification of PNAS publications in Biological Sciences contains 19 topics: Biochemistry, Medical Sciences, Neurobiology, Cell Biology, Genetics, Immunology, Biophysics, Evolution, Microbiology, Plant Biology, Developmental Biology, Physiology, Pharmacology, Ecology, Applied Biology, Psychology, Population Biology, Agricultural Sciences, and Anthropology. The percentages of published papers and numbers of dual classification papers in each topic are shown in Table 1.

	Topic	Number		Percent
1	Biochemistry	2578	(33)	21.517
2	Medical Sciences	1547	(13)	12.912
3	Neurobiology	1343	(9)	11.209
4	Cell Biology	1231	(10)	10.275
5	Genetics	980	(14)	8.180
6	Immunology	865	(9)	7.220
7	Biophysics	636	(40)	5.308
8	Evolution	510	(12)	4.257
9	Microbiology	498	(11)	4.157
10	Plant Biology	488	(4)	4.073
11	Developmental Biology	366	(2)	3.055
12	Physiology	340	(1)	2.838
13	Pharmacology	188	(2)	1.569
14	Ecology	133	(5)	1.110
15	Applied Biological Sciences	94	(6)	0.785
16	Psychology	88	(1)	0.734
17	Agricultural Sciences	43	(2)	0.359
18	Population Biology	43	(5)	0.359
19	Anthropology	10	(0)	0.083
	Total	11981	(179)	100

Table 1: Biological Sciences publications in PNAS volumes 94–98, by subtopic. Numbers of papers with dual classifications are given in parentheses.

The topic labels provide a classification structure for published materials and most of the articles are members of only a single topic. For our mixed membership model, we assume that there is a fixed number of extreme internal categories or aspects, each of which is characterized by multinomial distributions over words (in abstracts) and references (in bibliographies). Aspects are determined from contextual decompositions in such a way that a multinomial distribution of words and references in each document is a convex combination of the corresponding distributions from the aspects. The convex combination for each article is based on proportions of the article's content coming from each category. These proportions, or membership scores, determine soft classifications of articles with respect to internal categories.

Results

Choosing a suitable value for the number of internal categories or aspects, K, in this type of setting is difficult. In our analyses, we focused largely on two versions of the model, one with eight aspects and the other with ten. The set of parameters in our model is given by multinomial word and reference probabilities for each aspect, and by the parameters of Dirichlet distribution, which is a generating distribution for membership scores. There are $39{,}616$ unique words and $77{,}115$ unique references in our data, hence adding an aspect corresponds to having $39{,}615 + 77{,}114 + 1 = 116{,}730$ additional parameters. Because of the large numbers of parameters involved, it is difficult to assess the extent to which the added pair of aspects actually improve the fit of the model to the data. On the basis of a set of preliminary comparisons we found little to choose between them in fit and greater ease of interpretation for the eight aspect model. Therefore we report only the results of the eight aspect model here.

To see whether there are certain contexts that correspond to the aspects, we examine the most common words in the estimated multinomial distributions. In Table 2 we report the first 15 of the high probability words for each aspect, filtering out so called "stop words", words that are generally common in English. An alternative way would be to discard the words from the "stop list" before fitting the model. If the distribution of "stop words" is not uniform across the internal categories, this alternative approach may potentially produce different results.

The following interpretations are based on examination of 50 high probability words for each aspect. Note that enumeration of the aspects is arbitrary. The first aspect includes a number of words such as Ca2+, kinase, phosphorylation, receptor, g (protein) channel, that pertain to cell signaling and intracellular signal transduction. It is likely that in this aspect signal transduction is considered as applied to neuron signaling as indicated by the words synaptic, neurons, voltage. It is interesting that Ca2+ in the first aspect is the highest probability contextual word over all the aspects. Frequent words for the second aspect indicate that its context is related to molecular evolution that deals with natural selection on population and intraspecies level and mechanisms of acquiring genetic traits. Words in aspect 3 pertain mostly to plant molecular biology area. High probability words in aspect 4 relate to studies of neuronal responses in mice and humans, which identify this aspect as related to developmental biology and neurobiology. Aspect 5 contains words that can be associated with biochemistry and molecular biology. Words in aspect 6 point to genetics and molecular biology. Frequent words for aspect 7 contain such terms as immune, IL or interleukin, antigen, (interferon) gamma, and MHC class II, that point to a relatively new area in immunology, namely, tumor immunology. Presence of such words as HIV and virus in aspect 7 indicates a more general immunology content. For aspect 8, words such as increase or reduced, treatment, effect, fold, and p (assuming it stands for p-value) correspond to general reporting of experimental results, likely in the area of endocrinology.

As for words, multinomial distributions are estimated for the references that are present in our collection. For estimation we only need unique indicators for each referenced paper. After the model is fitted, attributes of high probability references for each aspect provide additional information about its contextual interpretation. Table 3 provides attributes of 15 high probability references for each aspect that were available in the database, together with PNAS citation counts (number of times cited by PNAS papers in the database). Notice that because the model draws from the contextual decomposition, high reference count is not necessary for a reference to have high aspect probability. In Table 3 high probability references for aspect 1 are dominated by publications in *Nature*, and, similarly, references in aspect 7 are mostly *Nature*, *Cell*, and *Science* publications from the mid-1990s.

Examining titles of the references (not shown), we see that manuals, textbooks, and references to articles that describe particular methodology appear to be prominent for many aspects. Thus, among the first 15 high probability references, all 15 from aspect 3, and more than half from aspect 4 are of this methodological type. In contrast, most high probability references for aspect 7 are those that report new findings. Titles of the references indicate neurobiology content for aspect 1, molecular evolution for aspect 2, and plant molecular biology for aspect 3, which is in agreement with our conclusions based on high probability words. For other aspects, titles of hight probability references help us to refine the aspects. Thus, aspect 4 mostly pertains to study of brain development, in particular, via genetic manipulation of mouse embryo. Aspect 5, identified as biochemistry and molecular biology by the words, can be described as protein structural biology by the references. Aspect 6 may be labelled more detailed as "DNA repair, mutagenesis, and cell cycle." The references for aspects 7 and 8 shift their focuses more

Aspect 1		Aspect 2		Aspect 3		Aspect 4	
ca2+	0.0062	species	0.0040	sequence	0.0024	development	0.0034
channel	0.0047	sequence	0.0026	acid	0.0020	neurons	0.0034
membrane	0.0047	sequences	0.0024	plants	0.0018	brain	0.0029
channels	0.0040	genetic	0.0024	cdna	0.0017	mouse	0.0025
receptors	0.0028	genome	0.0022	mutant	0.0015	normal	0.0024
synaptic	0.0026	evolution	0.0020	single	0.0015	expressed	0.0021
neurons	0.0022	among	0.0017	enzyme	0.0015	cortex	0.0019
g	0.0021	population	0.0016	plant	0.0014	embryonic	0.0017
calcium	0.0021	most	0.0016	identified	0.0013	adult	0.0017
activation	0.0020	chromosome	0.0015	amino	0.0013	neuronal	0.0016
release	0.0020	selection	0.0015	expressed	0.0013	function	0.0016
kinase	0.0019	populations	0.0014	mutants	0.0013	neural	0.0015
subunit	0.0019	three	0.0014	molecules	0.0012	early	0.0014
intracellular	0.0017	based	0.0013	based	0.0012	patients	0.0014
acid	0.0016	variation	0.0013	kda	0.0011	functional	0.0013

Aspect 5		Aspect 6		Aspect 7		Aspect 8	
residues	0.0028	transcription	0.0060	il	0.0046	increased	0.0027
enzyme	0.0023	nuclear	0.0036	tumor	0.0040	receptors	0.0023
active	0.0020	promoter	0.0031	activation	0.0036	g	0.0022
terminal	0.0019	transcriptional	0.0030	hiv	0.0032	p	0.0022
amino	0.0019	p53	0.0029	apoptosis	0.0031	insulin	0.0018
rna	0.0018	rna	0.0027	kinase	0.0028	effects	0.0018
structural	0.0018	kinase	0.0024	antigen	0.0026	increase	0.0018
state	0.0018	yeast	0.0024	virus	0.0025	acid	0.0018
folding	0.0017	function	0.0022	gamma	0.0021	effect	0.0016
sequence	0.0017	activation	0.0020	infection	0.0021	fold	0.0016
form	0.0016	sequence	0.0018	immune	0.0020	reduced	0.0016
peptide	0.0016	terminal	0.0018	signaling	0.0018	treatment	0.0016
atp	0.0015	cycle	0.0018	death	0.0017	glucose	0.0016
helix	0.0015	mutations	0.0017	activated	0.0017	mrna	0.0015
substrate	0.0015	factors	0.0017	vivo	0.0017	rats	0.0015

Table 2: High probability words for each aspect.

towards HIV infection and studies of molecular mechanisms of obesity.

Among frequent references for the eight aspects, there are seven PNAS papers, two of which appear in two aspects. These seven papers share a special feature: they were all either co-authored or contributed by a distinguished member of the National Academy of Sciences. In fact, one paper was co-authored by a Nobel prize winner and two were contributed by other Nobelists. Although these papers do not have the highest counts in the database, they are notable for various reasons; e.g., one is on clustering and gene expression [1], and it is also one of the two highly cited PNAS papers on clustering which we mentioned in the introduction. These seven papers may not necessarily be "off-beat," but they may be among those that fulfill Saunders MacLane's petition regarding the special nature of PNAS.

From our analysis of high probability words, it is difficult to see whether the majority of aspects correspond to a single topic from the official classifications in PNAS Biological Science publications. To investigate whether there is a correspondence between the estimated aspects and the given topics, we examine aspect "loadings" for each paper. Given estimated parameters of the model, the distribution of each article's "loadings" can be obtained via Bayes' theorem. The variational and EP procedures provide Dirichlet approximations to the posterior distribution $p(\boldsymbol{\lambda} \mid d, \boldsymbol{\theta})$ for each document d. We employ the mean of this Dirichlet as an estimate of the weight of the document on each aspect. Histograms of these loadings are provided in Figure 4 for four topics: Biophysics, Biochemistry,

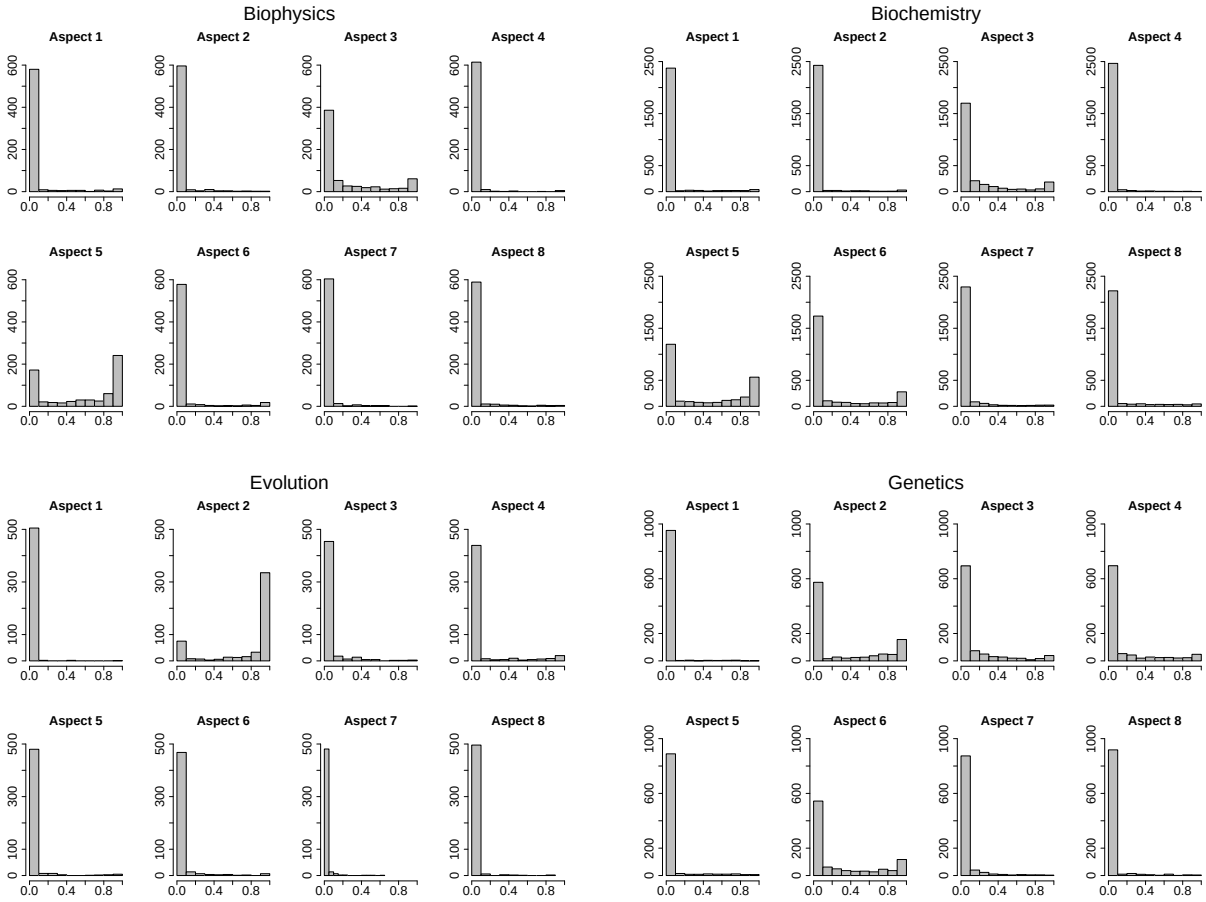


Figure 1: Distributions by aspect of the posterior means of membership scores for articles published in Biophysics, Biochemistry, Evolution, and Genetics.

Evolution, and Genetics. Relatively high histogram bars near zero correspond to the majority of articles having small posterior membership scores for the given aspect. About half of the articles in Biophysics can be considered as full members in aspect 5, but there are also articles from this topic that are full members in aspects 3, 4 and 6. The rest have mixed membership mostly in aspects 3 and 5. Distribution of the aspect loadings for articles in Biochemistry and Genetics similarly indicates that articles are coming from different aspects, and a substantial part of them is of mixed membership. Papers published in Evolution, on the other hand, show a somewhat different behavior—the majority of these papers comes fully from aspect 2.

The sparsity of the loadings can also be gauged by the parameters of the Dirichlet distribution, which are estimated as $\alpha_1 = 0.0195$, $\alpha_2 = 0.0203$, $\alpha_3 = 0.0569$, $\alpha_4 = 0.0346$, $\alpha_5 = 0.0317$, $\alpha_6 = 0.0363$, $\alpha_7 = 0.0411$, $\alpha_8 = 0.0255$. The estimated Dirichlet, which is the generative distribution of membership scores, is "bathtub shaped" on the simplex; as a result, articles will tend to have relatively high membership scores in only a few aspects.

To summarize the aspect distributions for each topic, we provide mean loadings and the graphical representation of these values in Figure 4. Larger values correspond to darker colors, and the values below some threshold are not shown (white) for clarity. As an example, the mean loading of 0.2883 for Pharmacology in the first aspect is the average of the posterior means of the membership scores for this aspect over all Pharmacology publications in the database. Note that this percentage is based on the assumption of mixed membership, and can be interpreted

as indicating that 29% of the words in Pharmacology papers originate from aspect 1, according to our model.

Examining the rows of Figure 4, we see that most subtopics in Biological Sciences have major components from more than one aspect (extreme or basis category). Examining the columns, we can gain further insights in interpretation of the extreme categories. Aspect 8, for example, is the aspect of origin for a combined 37% of Physiology, 30% of Pharmacology, and 25% of Medical Sciences papers, according to the mixed membership model. The most prominent subtopic is Evolution; it has the greatest influence in defining an extremal category, aspect 2. This is consistent with a special place Evolution holds among the Biological Sciences by standing apart both conceptually and methodologically.

Concluding Remarks

We have presented results from fitting a mixed membership model to a collection of Biological Sciences publications in the *Proceedings of the National Academy of Sciences*, 1997–2001, resulting in an implicit semantic decomposition of words and references in the papers. The model allows us to identify extreme internal categories of publications and to provide soft classifications of papers into these categories. Our results show that the traditional discipline classifications correspond to a mixed distribution over the internal categories. Our analyses and modeling were intended to capture a high level description of a subset of PNAS papers. By contrast, Griffiths and Steyvers [6] attempt a more fine-grained description by applying a restricted version of the mixed membership model using only words in the abstracts for a larger subset of PNAS papers. We plan to present a careful comparison between their approach and ours in the near future.

As noted in a famous statement by George Box [16], "all models are wrong." In our case, the assumption of a "bag of words and references" in the mixed membership model clearly oversimplifies reality; the model does not account for the general structure of the language nor does it capture the compositional structure of bibliographies. Many interesting extensions of the basic model we have explored are possible, from hierarchical models of topics, to more detailed models of citations and dynamic models of the evolution of scientific fields over time. Nevertheless, as Box notes, even wrong models may be useful. Our results indicate that mixed membership models can be useful for analyzing the implicit semantic structure of scientific publications.

Acknowledgments

The authors thank Dr. Anna Lokshin (University of Pittsburgh, Pittsburgh, PA) for her help in interpreting model results from a biologists' perspective. Elena Erosheva's work was supported in part by NIH grant R01 CA94212-01, and John Lafferty's work was supported by NSF grant CCR-0122581 and ARDA contract MDA904-00-C-2106.

1. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D., (1998) *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868.
2. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander E. S., and Golub, T. R., (1999) *Proc. Natl. Acad. Sci. USA* **96**, 2907–2912.
3. Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W., (2002) *Science* **298**, 2381–2385.
4. Woodbury, M. A., Clive, J., and Garson, A., (1978) *Computers and Biomedical Research* **11**, 277–298.
5. Erosheva, E. A., (2002) *PhD thesis, Carnegie Mellon University*.
6. Griffiths, T. L., and Steyvers, M., (2003) *Arthur M. Sackler Colloquium "Mapping Knowledge Domains"*.
7. Manton, K. G., Woodbury, M. A., and Tolley, H. D., (1994) *Wiley-Interscience*, 312.
8. Potthoff, R. G., Manton, K. G., Woodbury, M. A., and Tolley, H. D., (2000) *Journal of Classification* **17**, 315–353.
9. Pritchard, J. K., Stephens, M., and Donnelly, P., (2000) *Genetics* **155**, 945–959.
10. Hofmann, T., (2001) *Machine Learning* **42**, 177–196.

11. Blei, D. M., Ng, A. Y., and Jordan, M. I., (2003) *Journal of Machine Learning Research*, 3:993–1002.

12. Minka, T. P., and Lafferty, J., (2002) *Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI-2002)*, 352–359.

13. Cohn, D., and Hofmann, T., (2001) *Neural Information Processing Systems 13*.

14. Barn ard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., and Jordan, M. I. (2003) *Journal of Machine Learning Research* 3:1107–1135.

15. Blei, D. M., Jordan, M. I., and Ng, A. Y., (2003) *Bayesian Statistics 7. Proceedings of the Seventh Valencia International Meeting*, 25-44.

16. Box, G., (1979) *Robustness in Statistics*, 202.

Aspect 1

Author	Journal, Year	C
HAMILL OP	PFLUG ARCH EUR J PHY, 1981	72
LAEMMLI UK	Nature, 1970	322
HILLE B	IONIC CHANNELS EXCIT, 1992	58
BLISS TVP	NATURE, 1993	54
SUDHOF TC	NATURE, 1995	33
GRYNKIEWICZ G	J BIOL CHEM, 1985	31
SAMBROOK J	MOL CLONING LAB MANU, 1989	764
SHERRINGTON R	NATURE, 1995	33
ROTHMAN JE	NATURE, 1994	27
SIMONS K	NATURE, 1997	35
SOLLNER T	NATURE, 1993	25
ROTHMAN JE	SCIENCE, 1996	24
THINAKARAN G	NEURON, 1996	23
TOWBIN H	P NATL ACAD SCI USA, 1979	86
BERMAN DM	CELL, 1996	21

Aspect 2

Author	Journal, Year	C
SAITOU N	MOL BIOL EVOL, 1987	96
THOMPSON JD	NUCLEIC ACIDS RES, 1994	147
ALTSCHUL SF	NUCLEIC ACIDS RES, 1997	160
SAMBROOK J	MOL CLONING LAB MANU, 1989	764
ALTSCHUL SF	J MOL BIOL, 1990	253
FELSENSTEIN J	EVOLUTION, 1985	51
KISHINO H	J MOL EVOL, 1989	31
STRIMMER K	MOL BIOL EVOL, 1996	31
KIMURA M	J MOL EVOL, 1980	34
EISEN MB	P NATL ACAD SCI USA, 1998	60
SWOFFORD DL	PAUP PHYLOGENETIC AN, 1993	25
KIMURA M	NEUTRAL THEORY MOL E, 1983	28
KUMAR S	MEGA MOL EVOLUTIONAR, 1993	26
HASEGAWA M	J MOL EVOL, 1985	24
NEI M	MOL EVOLUTIONARY GEN, 1987	28

Aspect 3

Author	Journal, Year	C
SAMBROOK J	MOL CLONING LAB MANU, 1989	764
LAEMMLI UK	NATURE, 1970	322
ALTSCHUL SF	J MOL BIOL, 1990	253
BRADFORD MM	ANAL BIOCHEM, 1976	209
SANGER F	P NATL ACAD SCI USA, 1977	140
MILLER JH	EXPT MOL GENETICS, 1972	102
ALTSCHUL SF	NUCLEIC ACIDS RES, 1997	160
THOMPSON JD	NUCLEIC ACIDS RES, 1994	147
CHOMCZYNSKI P	ANAL BIOCHEM, 1987	206
HARLOW E	ANTIBODIES LAB MANUA, 1988	129
BLATTNER FR	SCIENCE, 1997	56
SCHENA M	SCIENCE, 1995	40
KYTE J	J MOL BIOL, 1982	51
MURASHIGE T	PHYSL PLANTARUM , 1962	33
TOWBIN H	P NATL ACAD SCI USA, 1979	86

Aspect 4

Author	Journal, Year	C
HOGAN B	MANIPULATING MOUSE E, 1994	68
CHOMCZYNSKI P	ANAL BIOCHEM, 1987	206
TALAIRACH J	COPLANAR STEREOTAXIC, 1988	60
PAXINOS G	RAT BRAIN STEREOTAXI, 1986	38
SAMBROOK J	MOL CLONING LAB MANU, 1989	764
NAGY A	P NATL ACAD SCI USA, 1993	39
MANSOUR SL	NATURE, 1988	37
BRAND AH	DEVELOPMENT, 1993	46
HOGAN B	MANIPULATING MOUSE E, 1986	32
TYBULEWICZ VLJ	CELL, 1991	46
KWONG KK	P NATL ACAD SCI USA, 1992	24
DUNLAP JC	CELL, 1999	19
LI E	CELL, 1992	35
ALTSCHUL SF	J MOL BIOL, 1990	253
EISEN MB	P NATL ACAD SCI USA, 1998	60

Aspect 5

Author	Journal, Year	C
KRAULIS PJ	J APPL CRYSTALLOGR, 1991	202
JONES TA	ACTA CRYSTALLOGR A, 1991	174
OTWINOWSKI Z	METHOD ENZYMOL, 1997	140
BRUNGER AT	ACTA CRYSTALLOGR D 5, 1998	118
LASKOWSKI RA	J APPL CRYSTALLOGR, 1993	96
NICHOLLS A	PROTEINS, 1991	85
NAVAZA J	ACTA CRYSTALLOGR A, 1994	81
SAMBROOK J	MOL CLONING LAB MANU, 1989	764
LAEMMLI UK	NATURE, 1970	322
MERRITT EA	ACTA CRYSTALLOGR D, 1994	66
BRUNGER AT	NATURE, 1992	48
BRADFORD MM	ANAL BIOCHEM, 1976	209
MERRITT EA	METHOD ENZYMOL, 1997	41
WUTHRICH K	NMR PROTEINS NUCL AC, 1986	40
KABSCH W	BIOPOLYMERS, 1983	39

Aspect 6

Author	Journal, Year	C
SAMBROOK J	MOL CLONING LAB MANU, 1989	764
SIKORSKI RS	GENETICS, 1989	102
DIGNAM JD	NUCLEIC ACIDS RES, 1983	68
LEVINE AJ	CELL, 1997	57
ELDEIRY WS	CELL, 1993	54
HARLOW E	ANTIBODIES LAB MANUA, 1988	129
HARPER JW	CELL, 1993	50
FRIEDBERG EC	DNA REPAIR MUTAGENES, 1995	58
ALTSCHUL SF	J MOL BIOL 1990	253
OGRYZKO VV	CELL, 1996	41
WEINBERG RA	CELL, 1995	40
KAMEI Y	CELL, 1996	39
HOLLSTEIN M	SCIENCE, 1991	41
FIELDS S	NATURE, 1989	67
YANG XJ	NATURE, 1996	37

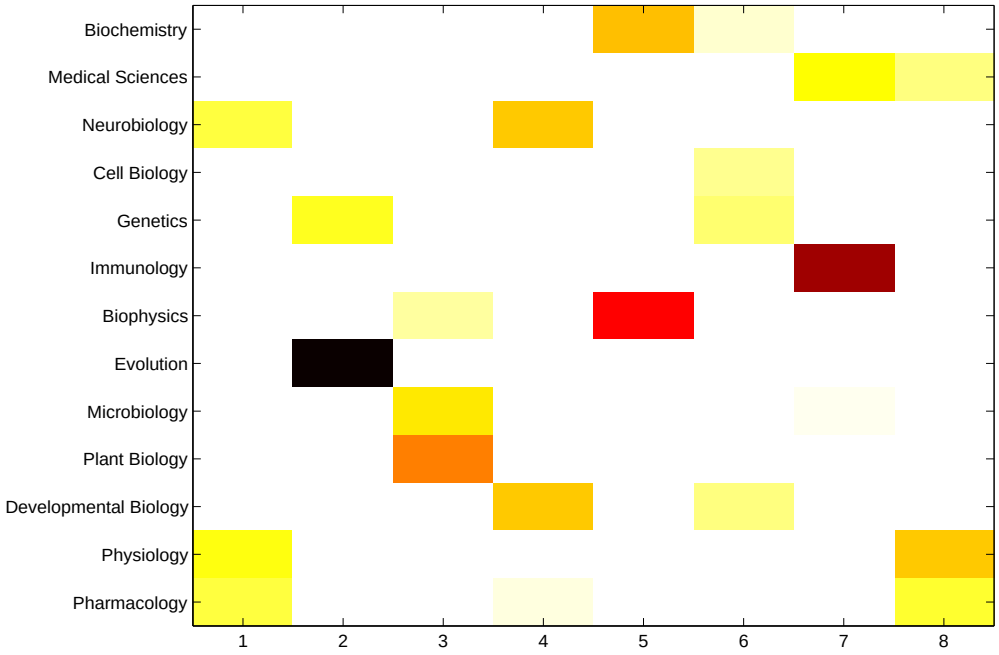
Aspect 7

Author	Journal, Year	C
DENG HK	NATURE, 1996	46
DRAGIC T	NATURE, 1996	45
DORANZ BJ	CELL, 1996	45
FENG Y	SCIENCE, 1996	43
ALKHATIB G	SCIENCE, 1996	43
COCCHI F	SCIENCE, 1995	41
CHOE H	CELL, 1996	41
THOMPSON CB	SCIENCE, 1995	38
ZOU H	CELL, 1997	40
DARNELL JE	SCIENCE, 1994	40
MUZIO M	CELL, 1996	35
LI P	CELL, 1997	36
XIA ZG	SCIENCE, 1995	38
BOLDIN MP	CELL, 1996	34
PEAR WS	P NATL ACAD SCI USA 1993	57

Aspect 8

Author	Journal, Year	C
CHOMCZYNSKI P	ANAL BIOCHEM, 1987	206
BRADFORD MM	ANAL BIOCHEM, 1976	209
LAEMMLI UK	NATURE, 1970	322
LOWRY OH	J BIOL CHEM, 1951	73
ZHANG Y	NATURE, 1994	31
KUIPER GGJM	P NATL ACAD SCI USA, 1996	27
SAMBROOK J	MOL CLON LAB MANU, 1989	764
MONCADA S	PHARMACOL REV, 1991	25
PELLEYMOUNTER MA	SCIENCE, 1995	23
CAMPFIELD LA	SCIENCE, 1995	23
KUIPER GGJM	ENDOCRINOLOGY, 1997	22
HALAAS JL	SCIENCE, 1995	21
BLIGH EG	CAN J BIOCH PHYSL, 1959	45
BROWN MS	CELL, 1997	28
ZHANG SH	SCIENCE 1992	18

Table 3: High probability references by aspect. For each aspect, the top references are shown in order of decreasing probability, according to the model. The count of each reference in the PNAS collection is shown in the rightmost column (C).



Topic	1	2	3	4	5	6	7	8
Biochemistry	0.0469	0.0347	0.1810	0.0178	0.3838	0.2057	0.0477	0.0823
Medical Sciences	0.0244	0.0502	0.0938	0.1274	0.0181	0.1075	0.3286	0.2500
Neurobiology	0.2875	0.0398	0.0722	0.3768	0.0196	0.0296	0.0441	0.1304
Cell Biology	0.1691	0.0165	0.1420	0.0684	0.1097	0.2423	0.1637	0.0884
Genetics	0.0141	0.3056	0.1422	0.1532	0.0487	0.2621	0.0395	0.0347
Immunology	0.0127	0.0593	0.1003	0.0413	0.0422	0.0915	0.6244	0.0283
Biophysics	0.0507	0.0295	0.2398	0.0162	0.5496	0.0542	0.0176	0.0423
Evolution	0.0042	0.7679	0.0465	0.0913	0.0289	0.0378	0.0101	0.0133
Microbiology	0.0158	0.1725	0.3431	0.0335	0.0647	0.1174	0.1870	0.0661
Plant Biology	0.1333	0.0983	0.4400	0.0360	0.0462	0.0954	0.0166	0.1344
Developmental Biology	0.0475	0.0288	0.1071	0.3729	0.0274	0.2558	0.0974	0.0631
Physiology	0.3179	0.0275	0.0712	0.1123	0.0258	0.0116	0.0595	0.3743
Pharmacology	0.2883	0.0161	0.0772	0.1965	0.0299	0.0349	0.0537	0.3033

Table 4: Mean decompositions of aspect membership scores (bottom), together with a graphical representation of this table (top). For clarity, the six lowest frequency topics, which make up 3.4% of the Biological Sciences articles, are not shown.