Reconceptualizing the classification of PNAS articles

Edoardo M. Airoldi[a], Elena A. Erosheva[b], Stephen E. Fienberg[c,d,1], Cyrille Joutard[e], Tanzy Love[f], and Suyash Shringarpure[d]

[a]Department of Statistics and Faculty of Arts and Sciences Center for Systems Biology, Harvard University, Cambridge, MA 02138; [b]Department of Statistics, School of Social Work, and the Center for Statistics and the Social Sciences, University of Washington, Seattle, WA 98195; [c]Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213; [d]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213; [e]Département de Mathématiques, Université de Montpellier, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France; and [f]Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY 14642

Contributed by Stephen E. Fienberg, October 8, 2010 (sent for review August 6, 2009)

PNAS article classification is rooted in long-standing disciplinary divisions that do not necessarily reflect the structure of modern scientific research. We reevaluate that structure using latent pattern models from statistical machine learning, also known as mixed-membership models, that identify semantic structure in co-occurrence of words in the abstracts and references. Our findings suggest that the latent dimensionality of patterns underlying PNAS research articles in the Biological Sciences is only slightly larger than the number of categories currently in use, but it differs substantially in the content of the categories. Further, the number of articles that are listed under multiple categories is only a small fraction of what it should be. These findings together with the sensitivity analyses suggest ways to reconceptualize the organization of papers published in PNAS.

text analysis | hierarchical modeling | Monte Carlo Markov chain | variational inference | Dirichlet process

The *Proceedings of the National Academy of Sciences* (PNAS) is indexed by Physical, Biological, and Social Sciences categories, and, within these, by subclassifications that correspond to traditional disciplinary topics. When submitting a paper, authors classify it by selecting a major and a minor category. Although authors *may* opt to have dual or even triple indexing, only a small fraction of published PNAS papers do so. How well does the current classification scheme capture modern interdisciplinary research? Could some alternative structure better serve PNAS in fostering publication and visibility of the best interdisciplinary research? These questions may be thought of as falling under the broad umbrella of "knowledge mapping."

A special 2004 supplement of PNAS, based on the *Arthur M. Sackler Colloquium on Mapping Knowledge Domains*, presented a number of articles that applied various knowledge mapping techniques to the contents of PNAS itself (1). What was striking about the issue is that two articles by Erosheva, et al. (2, henceforth EFL) and Griffiths and Steyvers (3, henceforth GS), based on similar statistical machine learning models, made statements about the number of inferred categories needed to describe semantic patterns in PNAS articles that differed by more than an order of magnitude (10 versus 300). Here we revisit these earlier analyses in the light of a new one and attempt (*i*) to understand the differences between them and (*ii*) to estimate the minimal number of latent categories necessary to describe modern scientific research, often interdisciplinary, as reported in PNAS.

To set the stage, we provide a brief overview of the relevant models and summarize the similarities and differences between the two approaches and corresponding analyses presented in refs. 2 and 3. Using the same database as in EFL (2), we explore a wide range of analytic and modeling choices in our attempt to reconcile the differences in prior analyses. We approach the choice of the number of "latent categories," which are inferred from data, with multiple strategies including one similar to that used by GS (3). Our findings suggest that 20 to 40 latent categories suffice to describe PNAS Biological Sciences publications, 1997–2001. Thus a reconceptualization of the indexing for PNAS Biological Sciences articles would require at most doubling the 19 traditional disciplinary categories. Because the true number of underlying semantic patterns is unknown and unknowable, we also report on a simulation study that confirms that, were there as few as 20 topics, our methodology would come close to estimating this number in a reasonable way. We also suggest some implications of our reconceptualization for the multiple indexing of interdisciplinary research in PNAS and elsewhere.

Overview of the Earlier Analyses

EFL (2) and GS (3) both analyzed data extracted from PNAS articles from an overlapping time period using versions of mixed-membership models (4). A distinctive feature of mixed-membership models for documents is the assumption that articles may combine words (plus any other attributes such as references) from several latent categories according to proportions of the article's membership in each category. The latent categories are not observable. They are typically estimated from data together with the proportions. The latent categories need not correspond to existing PNAS disciplinary classifications. Rather, each category can be thought of as a probability distribution over document-specific attributes that specifies which set of, say, words and references, co-occur frequently. The latent categories are often a quantitative by-product of concepts and semantic patterns that are used in a specific disciplinary area more than in others.

A mixed-membership structure allows for a parsimonious representation of interdisciplinary research without the need to create separate categories to accommodate both existing disciplinary links and new forms of collaborative research. Mixed-membership models achieve this through specifying article-level membership parameter vectors. In general, formulating mixed-membership models requires a combination of assumptions at the population level (e.g., PNAS Biological Sciences), subject level (individual articles), latent variable level (article's membership vector), and the sampling scheme for generating subject's attributes (article's words and/or references). Variations of these assumptions can easily produce different mixed-membership models, and the models used by EFL and GS are special cases of the general mixed-membership model framework presented by EFL.

We summarize other aspects of analytic choices, model fitting, and model selection strategies by EFL and GS in Table 1. We believe that analytic decisions, such as working with the Biological Sciences articles* versus with all PNAS articles, including

Author contributions: E.M.A., E.A.E., S.E.F., C.J., and T.L. designed research; E.M.A., E.A.E., S.E.F., C.J., T.L., and S.S. performed research; E.M.A., E.A.E., S.E.F., C.J., T.L., and S.S. contributed new reagents/analytic tools; E.M.A., E.A.E., S.E.F., C.J., T.L., and S.S. analyzed data; and E.M.A., E.A.E., S.E.F., C.J., and T.L. wrote the paper.

The authors declare no conflict of interest.

*Of 13,008 research articles published during this five-year period, 12,036 or 92.53% were in the Biological Sciences.

[1]To whom correspondence should be addressed. E-mail: fienberg@stat.cmu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1013452107/-/DCSupplemental.

Table 1. Comparison of analytic choices in previous analyses

2004 analyses	Erosheva, et al.	Griffiths and Steyvers
PNAS database		
Years	1997–2001	1991–2001
Scope	Biological Sciences	All areas
Article type	Only research articles	All publications
Article data		
Data source	Words (abstract) and references	Words (abstract)
Types of words included	Frequent, rare, "stop"	Only frequent
Model structure		
Number of latent categories	K	K
Mixed membership	$\lambda \sim \text{Dirichlet}(\alpha_1,\ldots,\alpha_K)$	$\lambda \sim \text{Dirichlet}(\alpha_1,\ldots,\alpha_K), \alpha_k = \alpha\ \forall k$
Distribution for words	Multinomial $(\sum_{k=1}^{K} \lambda_k \theta_{1k}, R_1)$	Multinomial $(\sum_{k=1}^{K} \lambda_k \theta_{1k}, R_1)$
Distribution for references	Multinomial $(\sum_{k=1}^{K} \lambda_k \theta_{2k}, R_2)$	None
Estimation		
Strategy	Variational expectation–maximization	Gibbs sampler
Hyperparameters	Estimated $\alpha_1,\ldots,\alpha_K$	Set $\alpha = 50/K$
Dimensionality selection		
Main objective	Descriptive model	Predictive model
Dimensions considered	$K = 8$; 10	$50 \le K \le 1000$

commentaries and reviews in the database, or excluding rare words from the analysis, cannot account for the order of magnitude difference in the most likely number of latent categories inferred from the similar data. Given that the models were so similar, we questioned the discrepancy between 8 to 10 latent categories used by EFL and 300 likely latent categories reported by GS. Why was there such a large difference in this key feature around which all other results revolved? More importantly, in light of this issue, can this type of statistical model support a substantive reconceptualization of the classification scheme in use by PNAS?

Below, we report on new analyses and results for the PNAS data and offer evidence in support of the utility of mixed-membership analysis for grounding considerations about a useful reconceptualization of PNAS categories.

Main Analysis

Mixed-Membership Models. We attempted to reconcile the differences in the original analyses of EFL and GS as follows: First, we used a common database for all models considered in this paper. Second, we varied data sources and hyperparameter estimation strategies to closely match those of the original analyses. Third, we remedied the absence of dimensionality selection strategy in EFL by allowing the number of latent categories, K, to change between 2 and 1,000, and comparing goodness of fit for different values of K.

Table 2 summarizes the resulting four mixed-membership models in a 2×2 layout. Model 3 is the closest to EFL's model except that we now employ a symmetric Dirichlet distribution ($\alpha_k = \alpha$ for all k) that matches GS's assumption. Model 2 uses the same data source and hyperparameter estimation strategy as in GS. We include models 1 and 4 to complement the other two by balancing the choice of data and estimation strategies.

Let x_1 be the observed words in the article's abstract and x_2 be the observed references in the bibliography. We assume that words and references come from finite discrete sets (vocabularies) of sizes V_1 and V_2, respectively. For simplicity, we assume that the vocabulary sets are common to all articles, independent of the publication time. We assume that the distribution of words and references in an article is driven by an article's membership in each of K latent categories, $\lambda = (\lambda_1,\ldots,\lambda_K)$, representing proportions of attributes that arise from a given latent pattern; $\lambda_k \ge 0$ for $k = 1,2,\ldots,K$ and $\sum_{k=1}^{K} \lambda_k = 1$. We denote the probabilities of the V_1 words and the V_2 references in the kth pattern by θ_{k1} and θ_{k2}, for $k = 1,2,\ldots,K$. These vectors of probabilities define multinomial[†] distributions over the two vocabularies of words and references for each latent category. We assume that article-specific (latent) vectors of mixed-membership scores are realizations from a symmetric Dirichlet[‡] distribution. For an article with R_1 words in the abstract and R_2 references in the bibliography, the generative sampling process for the mixed-membership model is as follows:

Mixed-Membership Models: Generative Process.

1. Sample $\lambda \sim \text{Dirichlet}(\alpha_1,\alpha_2,\ldots,\alpha_K)$, where $\alpha_k = \alpha$, for all k.
2. Sample $x_1 \sim \text{Multinomial}(p_{1\lambda},R_1)$, where $p_{1\lambda} = \sum_{k=1}^{K} \lambda_k \theta_{k1}$.
3. Sample $x_2 \sim \text{Multinomial}(p_{2\lambda},R_2)$, where $p_{2\lambda} = \sum_{k=1}^{K} \lambda_k \theta_{k2}$.

This process corresponds to models 3 and 4 in Table 2. The process for models 1 and 2 relies on steps 1 and 2 where only words in abstracts, x_1, are sampled. The conditional probability of words and references in an article is then

$$\Pr(x_1^{1:R_1}, x_2^{1:R_2} | \theta,\alpha) = \int \prod_{j=1}^{2} \prod_{r=1}^{R_j} \sum_{k=1}^{K} \prod_{v=1}^{V_j} (\lambda_k \theta_{kjv})^{x_{jv}^r} dD_\alpha(\lambda).$$

Estimation and Posterior Inference. Given a collection of articles, we treat pattern-specific distributions of words and references, $\{\theta_{k1}\}$ and $\{\theta_{k2}\}$, as constant quantities to be estimated, and article-specific proportions of membership λ_k as incidental parameters whose posterior distributions we compute. We assume that the hyperparameter α is unknown and estimated from the data in models 1 and 3; we fix the value of α at $50/K$ following the GS's heuristic in models 2 and 4. We carry out estimation and

Table 2. Mixed-membership models in our analysis

Data source(s)	Hyperparameter α	
	Estimated	Set at $50/K$
Abstract	Model 1	Model 2
Abstract + bibliography	Model 3	Model 4

[†]A multinomial distribution quantifies the intuition that words (or references) occur at each position in an abstract (or a bibliography) with different probabilities. The data suggest which words and references are most popular in articles that express each latent category.

[‡]A symmetric Dirichlet distribution quantifies the intuition that an article tends to belong to a few latent categories, when $\alpha < 1$. As $\alpha > 1$, an article belongs to more and more latent categories. The data suggest that $\hat{\alpha} < 1$ for articles in the biological sciences, implying that each research article covers only a few scientific areas.

inference using the variational expectation-maximization algorithm (5, 6). Variational methods provide an approximation to a joint posterior distribution when the likelihood is intractable. When we fix the hyperparameter α, as in models 2 and 4, we can use a Gibbs sampler to obtain the exact joint posterior distribution as implemented by GS in their original analysis. When we estimate α, however, we rely on variational approximations for estimation and inference. Simulation studies for the Grade of Membership model have shown that results obtained from both estimation methods are similar (7). We give full details in *SI Text*.

Dimensionality Selection. Each time we fit a mixed-membership model to data, we must specify the number of latent categories, K, in the model. The goal of dimensionality selection is to identify a number of latent categories K^* that is *optimal* in some sense. We identify the number of latent categories that leads to an optimal model-based summary of the database of scientific articles in a predictive sense, by means of a battery of out-of-sample experiments involving a form of cross-validation. We use 5-fold cross-validation, common in the machine learning literature, e.g., ref. 8, and explain the rationale for this choice in *SI Text*. Each out-of-sample experiment consists of five model fits for a given value of K. First, we split the N articles into five batches. Then, in turn, we estimate the model parameters using the articles in four batches, and we compute the likelihood of the articles in the fifth held-out batch. This leads to mean and variability estimates of quantities that summarize the goodness of fit of the model for a given K, on a batch of articles not included in the estimation. We consider a grid of values for K that range from a small to a large number of latent categories; namely, $K = 2,\ldots,5,10,\ldots,45,50,75,100,200,\ldots,900,1{,}000$.

Sensitivity Analyses

Fitting the four mixed-membership models from Table 3 to the PNAS dataset allows us to examine the impact of using references and estimating the hyperparameter α in a 2×2 design. We examine the sensitivity of empirical PNAS results obtained with mixed-membership models by considering the impact on model fit and selection of (*i*) our key assumption of mixed membership and our simple bag-of-references model. In addition, (*ii*) we use a simulation study to investigate the methodological issue of the potential impact on dimensionality selection due to fixing hyperparameter α, following the strategy of GS. Finally, to address interpretation issues, we study (*iii*) the distributions of shared memberships for different values of K and investigate (*iv*) whether increases in model dimension K beyond some optimal value change the macrostructure of the latent categories.

(i) Alternative Models. To study sensitivity of our latent dimensionality results to the key assumption of mixed membership, we implement another mixture model assuming that research reports belong to only one of the latent categories. This full-membership model can be thought of as a special case of the mixed-membership model where, for each article, all but one of the membership scores are restricted to be zero. As opposed to traditional finite mixture models that are formulated conditional on the number of latent categories K, this model variant allows the joint estimation of the latent categories, θ, and of the model dimension K.

We assume an infinite number of categories and implement this assumption through a Dirichlet process prior, D_α; for λ, e.g., see refs. 9 and 10. The distribution D_α models the prior probabilities of latent pattern assignment for the collection of documents. In particular, for the nth article, given the set of assignments for the remaining articles, λ_{-n}, this prior puts a probability mass on the kth pattern (out of K_{-n} distinct patterns observed in the collection of documents excluding the nth one), which is proportional to the number of documents associated with it. The prior distribution also puts a probability mass on a new, $(K_{-n}+1)$th latent semantic pattern, which is distinct from the patterns $(1,\ldots,K_{-n})$ observed in λ_{-n}. That is, D_α entails prior probabilities for each component of λ_n as follows:

$$p(\lambda_{n[k]} = 1 | \lambda_{-n}) = \begin{cases} \frac{m(-n,k)}{N-1+\alpha} & \text{if } m(-n,k) > 0 \\ \frac{\alpha}{N-1+\alpha} & \text{if } k = K_{-n}+1 \\ 0 & \text{otherwise,} \end{cases}$$

where $m(-n,k)$ is the number of documents that are associated with the kth latent pattern, excluding the nth document, i.e., $m(-n,k) = \sum_{m=1}^{N} I(\lambda_{m[k]} = 1, m \neq n)$.

The generative sampling process for this full-membership model is as follows:

1. Sample $\lambda \sim$ Dirichlet Process(α)
2. Sample $x_1 \sim$ Multinomial(θ_{k1}, R_1), where $\lambda_{n[k]} = 1$.
3. Sample $x_2 \sim$ Multinomial(θ_{k2}, R_2), where $\lambda_{n[k]} = 1$.

As with mixed-membership models, we considered two versions of the data: words from the abstract and references from the bibliography of the collection of articles. Model 5 corresponds to this process with steps 1 and 2, where we sample only words, x_1. Model 6 corresponds to this process with steps 1–3, where we sample words and references, x_1 and x_2. We provide full details about estimation and inference via Markov chain Monte Carlo methods in *SI Text*.

Additionally, we fit a mixed-membership model with a time-dependent bag of references. This confirmed that giving up the time resolution of the articles in our database has a negligible impact on model selection results.

(ii) Simulation Study. We simulated data from a mixed-membership model with $K^* = 20$ latent categories to obtain a corpus of documents we could use as ground truth. We used a vocabulary of 1,000 words and simulated 5,000 documents. We sampled the length of each document from a Poisson distribution with a mean of 100 words. We set the hyperparameter controlling mixed membership equal to $\alpha = 0.01$, whereas we sampled the banks of Bernoulli parameters corresponding to latent patterns from a symmetric Dirichlet with hyperparameter 0.01. We then treated K^* as unknown and approached model estimation in two ways: (*i*) by estimating the hyperparameter α as in our main analysis and (*ii*) by setting the hyperparameter $\alpha = 50/K$, for a given K, according to the ad hoc strategy implemented by GS.

For model selection purposes, we considered a grid for K as follows: increments of 4 for $10 \leq K \leq 50$, increments of 10 for $60 \leq K \leq 100$, and increments of 50 for $150 \leq K \leq 500$. Thus, we fit the model 25 times for each of 24 values of K.

(iii) Shared Memberships. Assume that a document is associated with latent category $k \in \{1,\ldots,K\}$ if and only if its membership score for this category is greater than $sd + 1/K$, where sd is the posterior standard deviation of the membership scores. For each value of K in our grid, we computed the number of documents associated with exactly $k \in \{1,\ldots,K\}$ latent categories. We then examined these distributions of the shared membership for a range of models with up to $K = 300$.

(iv) Macrostructure. Attempting to interpret the latent categories manually for all values of K in our grid is unreasonable. Hence, we analyzed computationally whether increases in model dimension K destroy the macrostructure and reorganize the latent categories by comparing multinomial probabilities for the latent patterns from the model with a smaller dimension K^* with the closest-matching ones of the model with a larger dimension.

We provide details for sensitivity analyses (iii) and (iv) in *SI Text*.

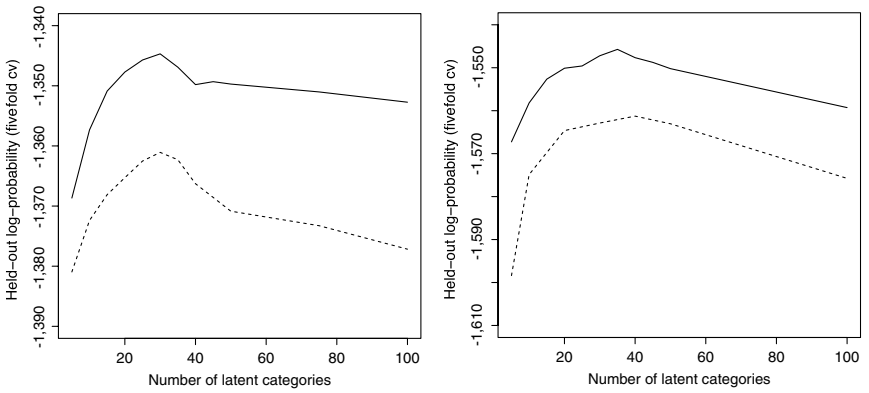

Fig. 1. Average held-out log probability corresponding to four mixed-membership models we fit (Table 3) to the PNAS Biological Sciences articles, 1997–2001, using words from article abstracts (*Left*) and words and references (*Right*). Solid lines correspond to models fitted by estimating the hyperparameter α; dashed lines correspond to models fitted by setting the hyperparameter equal to $\alpha = 50/K$.

Main Results

Dimensionality. Our primarily goal is to assess qualitatively and quantitatively a reasonable range for the number of latent categories underlying the PNAS database. Our analysis offers some insights into the impact on the results from differences in the models and the inference strategies. The simulation study also investigates the impact of such differences on model fit and dimension selection in a controlled setting.

Dimensionality for Mixed-Membership Models. To provide a quantitative assessment of model fit in terms of the number of latent categories K, we relied on their predictive performance with out-of-sample experiments, as described above. Recall that for mixed-membership analysis with models 1–4, we assume that K is an unknown constant. We split the articles into five batches to be used for all values of K. We considered values of K on a grid, spanning a range between 2 and 1,000. To summarize goodness of fit of the model in a predictive sense, we examine the held-out probability, that is, the probability computed on the held-out batch of articles.[§]

For each value of K on the grid, we computed the average held-out log-probability value over the five model fits. Fig. 1 summarizes predictive performance of the mixed-membership models 1–4, for values of $K = 2,\ldots,100$ (the average log-probability values continued to decline gradually for K greater than 100). The goodness of fit improves when we estimate the hyperparameter α (solid lines); however, all plots suggest an optimal choice of K falls in the range of 20–40, independent of the estimation strategy for α and of references inclusion. Values of K that maximize the held-out log probability are somewhat greater when the database includes references. We obtained similar dimensionality results using the Bayesian information criterion (11).

Dimensionality for Full-Membership Models. Although we base the choice of K for mixed-membership models 1–4 on their predictive performance, semiparametric full-membership models 5 and 6 allow us to examine posterior distribution of K.

Fig. 2 shows the posterior distributions on K—density on the Y axis versus values of K on the X axis—obtained by fitting data to semiparametric models with words only (model 5, solid line) and words and references (model 6, dashed line). The maximum a posteriori estimate of K is smaller for the model including references compared to the model with words only. Further, the posterior range of K is smaller for the model including references. Thus adding references to the models reduces the posterior uncertainty about K.

Dimensionality: Overall. Our simulation showed that setting the hyperparameter α as a function of K in the same way as GS did had the greatest impact on estimates of the document-specific mixed-membership vectors, leading to a modest upward bias in the choice of an optimal K, but did not result in an order of magnitudes difference. We provide more detailed results on our simulation study in *SI Text*.

Overall, for all six models, values of K in the range of 20–40 are plausible choices for the number of latent categories in PNAS Biological Sciences research reports, 1997–2001.

Qualitative and Quantitative Analysis of Inferred Categories. For illustrative purposes, we consider $K^* = 20$ for the mixed-membership model with words and references. We obtain qualitative descriptions of the latent categories using two approaches: via examining high probability words and references in each category and via comparing the model-based inferred article categories with the original PNAS classifications.

Studying the lists of words and references that are most likely to occur according to the distribution of each latent category, we see some interesting patterns that are distinct from current PNAS

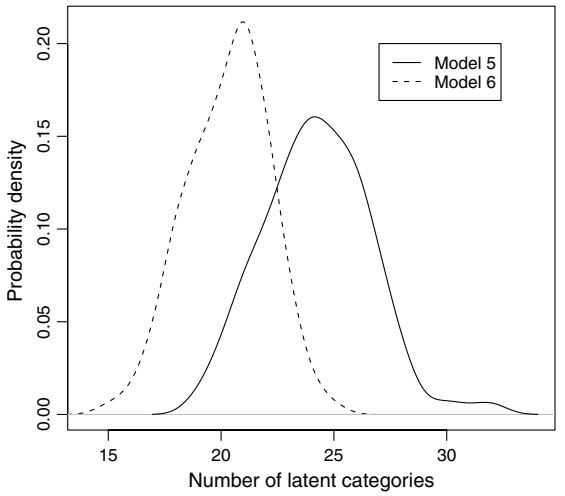

Fig. 2. Posterior distribution of the number of mixture components K for full-membership models for the PNAS Biological Sciences articles, 1997–2001, using words from article abstracts (solid line) and words and references (dashed line).

[§]Technically, the held-out probability is a variational lower bound on the likelihood of the held-out documents, as we detail in *SI Text*.

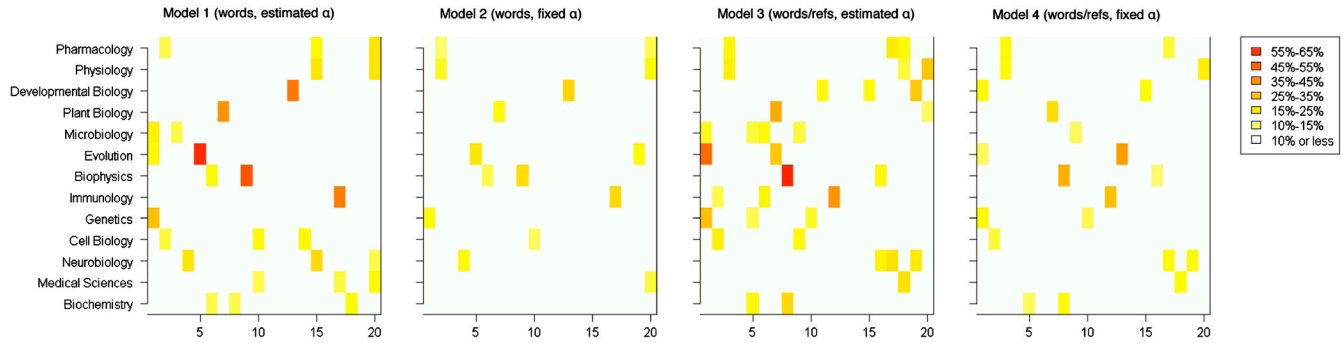

Fig. 3. Estimated average membership of articles in the 20 latent categories by PNAS classifications for mixed-membership models 1–4 in Table 3, *Left* to *Right*.

classifications. For example, category 5 focuses on the process of apoptosis and genetic nuclear activity in general. Category 12 concerns peptides. Several categories relate to protein studies including pattern 8 that deals with protein structure and binding. We offer an interpretation of *all* the topics in *SI Text* in an effort to demonstrate what a reasonable model fit should look like.

To examine the relationship between the 20 inferred categories and the 19 original PNAS categories in the Biological Sciences, we plot in Fig. 3 the average membership of the set of documents in the *i*th PNAS class (row) in the *k*th latent category (column).[¶] We threshold the average membership scores so that small values (less than 10%) would not distract from the visual pattern.

Fig. 3 (*Left* to *Right*) details results for models 1 and 2 (words only) and models 3 and 4 (words and references). The results reveal the impact of expanding the database to include references and of setting the hyperparameter at $\alpha = 50/K$ (models 2 and 4). When we include the references, the relationship of estimated latent categories with designated PNAS classifications becomes more composite for each estimation method. When we estimate the hyperparameter α, we observe a better agreement between estimated latent categories and the original PNAS classifications. A greater number of darker color blocks point to more articles with estimated substantial membership in just a few latent categories for the α-estimated models. Lighter blocks for the constrained-α models may be due to more spread-out membership (due to small membership values of all articles) or to an apparent disagreement of estimated membership vectors among articles from original PNAS classifications. Either explanation leads us to conclude that estimating hyperparameters gives us a model that has a better connection to the original PNAS classification.

From an inspection of the estimated categories, we see that small subclassifications such as Anthropology do not result in separate categories and broad ones such as Microbiology and Pharmacology have distinct subpatterns within them. Nearly all of the PNAS classifications are represented by several word-and-reference co-occurrence patterns, consistently across models.

Fig. 4 shows the distributions of shared memberships for varying values of K based on model 3. Overall, no matter what the dimensionality of the model, most articles tend to be associated with about five or fewer latent categories. For $K^* = 20$, 37% of articles are associated with two latent categories and 2% with three categories, the theoretical upper bound on the number of associations in this case. *SI Text* provides further details.

When we investigated the impact of increases in dimensionality K on interpretation, we found substantial reorganization among distributions of words and references in the latent categories. We compared estimated multinomial distributions for words and references between categories from pairs of models of dimensions K_1 and K_2, where $K_1 < K_2$, by computing correlations between all pairs of vectors. We found that correlations between the K_1 vectors in the smaller model and K_1 best-matching vectors from the larger model tend to diminish as K_2 increases, indicating that the macrostructure is not preserved. As expected, we also found that correlations between the K_1 vectors in the smaller model and additional vectors from the larger model were small.

Predictions. Recall that our database includes 11,988 articles, classified by the authors into 19 subcategories of the Biological Sciences section. Of these, 181 were identified by their authors as having dual classifications. Here, we identify publications that have similar membership vectors to dual-classified articles, i.e., single-classified articles that may have been also cross-submitted. Table 3 summarizes these results. The parametric models 1–4 predict that respectively 554, 114, 1,008, and 538 additional articles were *similar* to the author identified dual-classified articles. By similar, we mean that their mixed-membership vectors in the 20 latent semantic patterns match a membership vector of a dual-classified article to the first significant digit. Of particular interest is the large proportion of Biochemistry, Neurobiology, Biophysics, and Evolution articles that our analyses suggest as potential dual-classified articles.

Discussion

We have focused on alternative specifications for mixed-membership models to explore ways to classify papers published in PNAS that capture, in a more salient fashion, the interdisciplinary nature of modern science. Through the data analysis of 5 y of PNAS Biological Science articles, we have demonstrated that a small number of classification topics do an adequate job of cap-

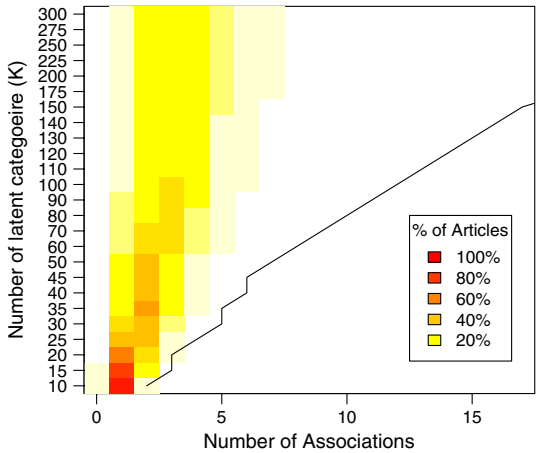

Fig. 4. Distribution of shared membership across the latent categories for different values of K using model with words and references. Black solid line indicates the upper bound on the number of associations for each K.

[¶]We show only the 13 most frequently used disciplinary categories here, but we provide the complete figure in *SI Text*.

Table 3. Summary statistics for dual-classified articles and predictions based on mixed-membership model applied to 5 years of PNAS data

	Articles in database		Prediction with model:			
Published Category	Primary	Sec.	1	2	3	4
Biochemistry	2,580	33	51	8	230	109
Medical Sciences	1,547	13	13	2	39	18
Neurobiology	1,343	10	65	29	104	10
Cell Biology	1,230	10	5	3	10	3
Genetics	980	14	20	2	55	15
Immunology	865	9	43	1	80	45
Biophysics	637	40	139	37	231	131
Evolution	510	12	101	6	103	133
Microbiology	498	11	8	3	13	9
Plant Biology	488	4	2	0	8	1
Developmental Bio	367	2	2	0	3	1
Physiology	341	2	0	2	17	3
Pharmacology	189	2	0	1	9	2
Ecology	133	5	49	3	34	42
Applied Bio Sci	95	6	5	0	2	1
Psychology	88	1	34	14	52	1
Agricultural Sci	43	2	2	0	4	0
Population Biology	43	5	10	3	12	13
Anthropology	10	0	5	0	2	1
Total	11,988	181	554	114	1,008	538

turing the semantic structure of the published articles. They also provide us with a reasonable correspondence to the current PNAS classification structure.

The machine learning literature contains many variants of mixed-membership models for classification and clustering problems. For example, Blei and Lafferty (12) describe a dynamic topic model and apply it to data from 125 y of *Science*. A different approach to references might exploit the network structure of authors with the mixed-membership stochastic block model of ref. 13 or the author–topic model of ref. 14; see also a review of such models in the psychological literature (15). The selection of appropriate dimension for number of latent categories, K, is often hidden behind the scene in applications, with some exceptions such as those involving a probability distribution over the number of dimensions such as models with the Dirichlet process (16) and its many variants (17–20).

Here we provide an extended analysis of dimensionality in a database of PNAS publications, contrasting our findings with earlier published ones (2, 3). The consistency of our results across multiple variants of mixed-membership models indicates that this type of statistical analysis, when done carefully, could support a substantive reconceptualization of the classification scheme used by PNAS. A more in-depth study of semantic patterns, inferred from actual data extracted from papers published in PNAS using tools such as those described in this paper, would also assist in the review process and the indexing of published papers, to reflect modern, overlapping, and interdisciplinary scientific publications. Finally, instead of relying solely on citations, researchers could be suggested related work via articles with the "most similar" semantic patterns, in an automated manner.

1. Shiffrin RM, Börner K (2004) Mapping knowledge domains. *Proc Natl Acad Sci USA* 101:5183–5185.
2. Erosheva EA, Fienberg SE, Lafferty J (2004) Mixed-membership models of scientific publications. *Proc Natl Acad Sci USA* 101:5220–5227.
3. Griffiths TL, Steyvers M (2004) Finding scientific topics. *Proc Natl Acad Sci USA* 101:5228–5235.
4. Erosheva EA, Fienberg SE (2005) *Classification—The Ubiquitous Challenge*, eds C Weihs and W Gaul (Springer, Berlin), pp 11–26.
5. Jordan MI, Ghahramani Z, Jaakkola T, Saul L (1999) Introduction to variational methods for graphical models. *Mach Learn* 37:183–233.
6. Airoldi EM (2007) Getting started in probabilistic graphical models. *PLoS Comput Biol* 3(12):e252.
7. Erosheva EA, Fienberg SE, Joutard C (2007) Describing disability through individual-level mixture models for multivariate binary data. *Ann Appl Stat* 1:502–537.
8. Hastie T, Tibshirani R, Friedman JH (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York).
9. Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. *Ann Stat* 1:209–230.
10. Neal R (2000) Markov chain sampling methods for Dirichlet process mixture models. *J Comput Graph Stat* 9:249–265.
11. Schwarz GE (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464.
12. Blei DM, Lafferty JD (2006) *Proceedings of the Twenty-Third International Conference*, eds WW Cohen and A Moore (ACM, Pittsburgh), pp 113–120.
13. Airoldi EM, Blei DM, Fienberg SE, Xing EP (2008) Mixed membership stochastic blockmodels. *J Mach Learn Res* 9:1981–2014.
14. Rosen-Zvi M, Chemudugunta C, Griffiths TL, Smyth P, Steyvers M (2010) Learning author-topic models from text corpora. *ACM T Inform Syst* 28:1–38.
15. Griffiths TL, Steyvers M, Tenenbaum J (2007) Topics in semantic representation. *Psychol Rev* 114:211–244.
16. Escobar M, West M (1995) Bayesian density estimation and inference using mixtures. *J Am Stat Assoc* 90:577–588.
17. Griffiths TL, Ghahramani Z (2005) Infinite latent feature models and the Indian buffet process. (University College, London), Technical Report GCNU-TR 2005–001.
18. Griffin J, Steele M (2006) Order-based dependent Dirichlet processes. *J Am Stat Assoc* 101:179–194.
19. Teh YW, Jordan MI, Beal MJ, Blei DM (2006) Hierarchical Dirichlet processes. *J Am Stat Assoc* 101:1566–1581.
20. Duan JA, Guindani M, Gelfand AE (2007) Generalized spatial Dirichlet process models. *Biometrika* 94:809–825.

Supporting Information

Airoldi et al. 10.1073/pnas.1013452107

SI Text

In the main paper, we summarized our main analyses and the sensitivity analysis using models 1–4. Here we present further details and supplementary results that support the claims and conclusions in the paper.

We begin with a detailed description of the estimation and inference associated with the mixed-membership models and some technical justification regarding the cross-validation approach for model assessment. We then provide the details of estimation and inference for the full-membership sensitivity analysis and follow up with the results of the rest of our sensitivity analyses concerning the distribution of shared membership and the effects of model size on the macrostructure of latent categories. Next, we present results from the simulation study described in the main paper. Finally, we provide a more complete discussion of results and interpretation of all of the latent semantic patterns in the fit of the mixed-membership model presented in the main paper.

Estimation for Mixed-Membership Models. In the mixed-membership case, we assume the number of topics ($K < \infty$) is fixed during inference. The probability of a document according to this model is

$$p(x_1^{1:R_1}, x_2^{1:R_2}|\alpha,\theta) = \int \left(\prod_{r=1}^{R_1}\sum_{k=1}^{K}\prod_{v=1}^{V_1}(\lambda_k\theta_{k1[v]})^{x^r_{1[v]}}\right) \times \left(\prod_{r=1}^{R_2}\sum_{k=1}^{K}\prod_{v=1}^{V_2}(\lambda_k\theta_{k2[v]})^{x^r_{2[v]}}\right) D_\alpha(d\lambda), \quad \text{[S1]}$$

and the integral does not have a closed form solution. We need the probability to compute the joint posterior distribution of the latent variables encoding mixed membership, and latent category indicators for words and references to compute the denominator of

$$p(\lambda, z_1^{1:R_1}, z_2^{1:R_2}|x_1^{1:R_1}, x_2^{1:R_2}, \alpha,\theta) = \frac{p(\lambda, z_1^{1:R_1}, z_2^{1:R_2}, x_1^{1:R_1}, x_2^{1:R_2}|\alpha,\theta)}{p(x_1^{1:R_1}, x_2^{1:R_2}|\alpha,\theta).}, \quad \text{[S2]}$$

The variational method prescribes the use of a mean-field approximation to the posterior distribution in Eq. **S2**, described below. Such an approximation leads to a lower bound for the probability of a document, which depends upon a set of free parameters ($\gamma, \phi_1^{1:R_1}, \phi_2^{1:R_2}$). These free parameters are introduced in the mean-field approximation and are chosen to minimize the Kullback–Leibler (KL) divergence between true and approximate posteriors.

Variational Expectation–Maximization (EM) ($\{x_{1n}^{1:R_1}, x_{2n}^{1:R_2}\}_{n=1}^N$).

1. initialize $\alpha_{[k]} := 1/K \ \forall k$
2. initialize $\theta_{k1[v]} := 1/V_1 \ \forall v, k$
3. initialize $\theta_{k2[v]} := 1/V_2 \ \forall v, k$

   repeat

   for $n \leftarrow 1$ **to** N **do**

4. $(\gamma_n, \phi_{1n}^{1:R_1}, \phi_{2n}^{1:R_2}) \mapsto$ **MFLB**$(x_{1n}^{1:R_1}, x_{2n}^{1:R_2}) \ \forall n$

   end

5. $\theta_{1[vk]} \mapsto \sum_{n=1}^{N}\sum_{r=1}^{R_1}\phi^r_{1n[k]}x^r_{1n[v]} \ \forall v, k$
6. $\theta_{2[vk]} \mapsto \sum_{n=1}^{N}\sum_{r=1}^{R_2}\phi^r_{2n[v]}x^r_{2n[k]} \ \forall v, k$
7. normalize the vectors of θ to sum to 1
8. find pseudo MLE for α with Newton-Raphson (Eq. **S5–S6**)

   until convergence ;
9. **return** (α, θ)

Algorithm 1: The variational EM algorithm to solve the Bayes problem in finite mixture model of text and references. The **MFLB** algorithm called in step 4 is the Mean-Field Lower-Bound procedure detailed in Algorithm 2. In our implementation, the M step updates (steps 5–6) are performed incrementally, within step 4 of the algorithm outlined above, thus speeding up the overall run time. The variational EM algorithm we develop for performing posterior inference, see Algorithm 1, is an approximate EM-like algorithm, with estimation and maximization steps. During the M step, we maximize the lower bound for the probability over the hyperparameters of the model, (α, θ), to obtain to (pseudo) maximum-probability estimates. During the E step, we tighten the lower bound for the probability by minimizing the KL divergence between the true and the approximate posteriors over the free parameters, $(\gamma, \phi_1^{1:R_1}, \phi_2^{1:R_2})$, given the most recent estimates for the hyperparameters.

In the M step, we update the hyperparameters of the model, $(\alpha, \theta_1, \theta_2)$, by maximizing the tight lower bound for the probability over such hyperparameters. Given the most recent updates of the free parameters the bound depends on, $(\gamma, \phi_1^{1:R_1}, \phi_2^{1:R_2})$. This leads to the following (pseudo) maximum-probability estimates for the parameters:

$$\theta_{k1[v]} \propto \sum_{n=1}^{N}\sum_{r=1}^{R_1}\phi^r_{1n[k]}x^r_{1n[v]}, \quad \text{[S3]}$$

$$\theta_{k2[v]} \propto \sum_{n=1}^{N}\sum_{r=1}^{R_2}\phi^r_{2n[k]}x^r_{2n[v]}, \quad \text{[S4]}$$

where n is the document index, introduced above. The document index is necessary as all documents are used to estimate the latent semantic patterns. A closed form solution for the (pseudo) maximum-probability estimates of α does not exist. We can produce a method that is linear in time by using a Newton–Raphson algorithm, with the following gradient and Hessian for the log probability:

$$\frac{\partial L}{\partial \alpha_{[k_1]}} = N\left(\Psi\left(\sum_{k=1}^{K}\alpha_{[k]}\right) - \Psi(\alpha_{[k_1]})\right) + \sum_{n=1}^{N}\left(\Psi(\gamma_{n[k_1]}) - \Psi\left(\sum_{k=1}^{K}\gamma_{n[k]}\right)\right), \quad \text{[S5]}$$

$$\frac{\partial L}{\partial \alpha_{[k_1]}\alpha_{[k_2]}} = N\left(I_{k_1=k_2}\cdot\Psi'(\alpha_{[k_1]}) - \Psi'\left(\sum_{k=1}^{K}\alpha_{[k]}\right)\right). \quad \text{[S6]}$$

In the approximate E step we update the free parameters for the mean-field approximation of the posterior distribution in Eq. **S2**, $(\gamma, \phi_1^{1:R_1}, \phi_2^{1:R_2})$, given the most recent estimates of the hyperparameters of the model, (α, θ), as follows:

$$\phi^r_{1[k]} \propto \prod_{v=1}^{V_1}\left[\theta_{k1[v]}\cdot\exp\left(\Psi(\gamma_{[k]}) - \Psi\left(\sum_{k=1}^{K}\gamma_{[k]}\right)\right)\right]^{x^r_{1[v]}}, \quad \text{[S7]}$$

$$\phi^r_{2[k]} \propto \prod_{v=1}^{V_2}\left[\theta_{k2[v]}\cdot\exp\left(\Psi(\gamma_{[k]}) - \Psi\left(\sum_{k=1}^{K}\gamma_{[k]}\right)\right)\right]^{x^r_{2[v]}}, \quad \text{[S8]}$$

$$\gamma_{[k]} = \alpha_{[k]} + \sum_{r=1}^{R_1} \phi^{r}_{1[k]} + \sum_{r=1}^{R_2} \phi^{r}_{2[k]}. \quad \text{[S9]}$$

This minimizes the posterior KL divergence between true and approximate posteriors, at the document level, and leads to a new lower bound for the probability of the collection of documents. Note that the products over words and references in formulas **S7** and **S8** serve the purpose of selecting the correct probabilities of occurrence in the respective vocabularies, at a specific position, (r_1, r_2), in the document. That is, the updates of the free parameters $(\phi^{r_1}_{1[k]}, \phi^{r_2}_{2[k]})$ depend only on the probabilities $(\theta_{k1[v_1]}, \theta_{k2[v_2]})$, where $v_1 := \{v \in [1,V_1] \text{ s.t.} x^{r_1}_{1[v]} = 1\}$ and $v_2 := \{v \in [1,V_2] \text{ s.t.} x^{r_2}_{2[v]} = 1\}$. Using this notation, the updates simplify to

$$\phi^{r_1}_{1[k]} \propto \theta_{k1[v_1]} \cdot \exp\left(\Psi(\gamma_{[k]}) - \Psi\left(\sum_{k=1}^{K} \gamma_{[k]}\right)\right), \quad \text{[S10]}$$

$$\phi^{r_2}_{2[k]} \propto \theta_{k2[v_2]} \cdot \exp\left(\Psi(\gamma_{[k]}) - \Psi\left(\sum_{k=1}^{K} \gamma_{[k]}\right)\right). \quad \text{[S11]}$$

The mean-field approximation to the likelihood we described above is summarized in Algorithm 2.

Mean-Field Lower-Bound ($x_1^{1:R_1}$, $x_2^{1:R_2}$).

1. initialize $\phi^{r}_{1[k]} := 1/K \ \forall r, k$
2. initialize $\phi^{r}_{2[k]} := 1/K \ \forall r, k$
3. initialize $\gamma_{[k]} := \alpha_{[k]} + R_1/K + R_2/K \ \forall k$
   repeat
   for $r \leftarrow 1$ **to** R_1 **do**
   for $k \leftarrow 1$ **to** K **do**
4. $\phi^{r}_{1[k]} \mapsto \theta_{k1[v]} \times \exp(\Psi(\gamma_{[k]}) - \Psi(\sum_{k=1}^{K} \gamma_{[k]}))$
   end
5. normalize ϕ^{r}_{1} to sum to 1
   end
   for $r \leftarrow 1$ **to** R_2 **do**
   for $k \leftarrow 1$ **to** K **do**
6. $\phi^{r}_{2[k]} \mapsto \theta_{k2[v]} \times \exp(\Psi(\gamma_{[k]}) - \Psi(\sum_{k=1}^{K} \gamma_{[k]}))$
   end
7. normalize ϕ^{r}_{2} to sum to 1
   end
8. $\gamma = \alpha + \sum_{r=1}^{R_1} \phi^{r}_{1} + \sum_{r=1}^{R_2} \phi^{r}_{2}$
   until convergence;
9. **return** $(\gamma, \phi_1^{1:R_1}, \phi_2^{1:R_2})$

Algorithm 2: The mean-field approximation to the likelihood for the finite mixture model of text and references. In order to develop the mean-field approximation for the posterior distribution in Eq. **S2** we use in the E step above. We posit N independent fully factorized joint distributions over the latent variables, one for each document,

$$q(\lambda, z_1^{1:R_1}, z_2^{1:R_2} | \gamma, \phi_1^{1:R_1}, \phi_2^{1:R_2}) = q(\lambda|\gamma)\left(\prod_{r_1=1}^{R_1} q(z_1^{(r_1)}|\phi_1^{(r_1)}) \prod_{r_2=1}^{R_2} q(z_2^{(r_2)}|\phi_2^{(r_2)})\right),$$

which depends on the set of previously mentioned free parameters, $(\gamma, \phi_1^{1:R_1}, \phi_2^{1:R_2})$. The mean-field approximation consists in finding an approximate posterior distribution,

$$\tilde{p}(\lambda, z_1^{1:R_1}, z_2^{1:R_2} | \tilde{\gamma}, \tilde{\phi}_1^{1:R_1}, \tilde{\phi}_2^{1:R_2}, \alpha, \theta),$$

where the conditioning on the data is now obtained indirectly, through the free parameters,

$$\tilde{\gamma} = \tilde{\gamma}(x_1^{1:R_1} x_2^{1:R_2}), \quad \tilde{\phi}_2^{1:R_1} = \tilde{\phi}_1^{1:R_1}(x_1^{1:R_1} x_2^{1:R_2}), \quad \tilde{\phi}_2^{1:R_2} = \tilde{\phi}_2^{1:R_1}(x_1^{1:R_1} x_2^{1:R_2}).$$

The factorized distribution leads to a lower bound for the probability; in fact, it is possible to find a closed form solution to the integral in Eq. **S1** by integrating the latent variables out with respect to the factorized distribution. An approximate posterior, $\tilde{p}$, is computed by substituting the lower bound for the probability at the denominator of Eq. **S2**. The mean-field approximation is then obtained by minimizing the Kullback–Leibler divergence between the true and the approximate posteriors, over the free parameters.

The mean-field approximation has been used in many applications over the years (1–4). Intuitively, the approximation aims at reducing a complex problem into a simpler one by "decoupling the degrees of freedom in the original problem." Such decoupling is typically obtained via an expansion that involves additional, free parameters that are problem dependent, e.g., $\{\gamma, \phi_{1n}^{1:R_1}, \phi_{2n}^{1:R_2}\}_{n=1}^{N}$ in our model above. For a thorough treatment of such methods focused on applications to statistics and machine learning, see refs. 5–7).

Cross-validation methodology. In the article, we use several methods to arrive at choices for K, the number of latent semantic patterns including 5-fold cross-validation for the mixed-membership models, following the approach described in Hastie et al. (8) and widely used in other machine learning applications. There are many ways to approach the cross-validation methodology depending on the fraction of data used for estimation, and the complementary fraction used for cross-validation.

Results in Breiman et al. (9) and Kohavi (10) suggest that leave-one-out cross-validation typically leads to estimates with low bias and high variance. Using a larger portion of the database as the test set, as in k-fold cross-validation, can reduce the variance of the estimates at a cost of little-to-modest increase in bias. For large databases, such as the one we consider in our article, many authors including Mahmood and Khan (11) suggest the use of 10-fold or 5-fold cross-validation, perhaps coupled to other methods (such as the bootstrap) to increase the diversity of the otherwise few folds, but seldom 3-fold and 2-fold. The decision on the number of "folds" to use is somewhat arbitrary, depending on the objective of the analysis and any resolution of differential performance is typically settled empirically. When cross-validation is combined with the bootstrap or other sampling schemes, "shared wisdom" suggests leaving enough data in the test set to avoid reporting overconfident results [see, e.g., Hastie, et al. (8)]. Overconfidence may still happen with 10-fold hybrid cross-validation/sampling schemes, in our experience, but seldom with 5-fold. This is the reason why we chose the 5-fold cross-validation scheme.

There are other related suggestions in the literature that appear to relate to both cross-validation and some of the other strategies for model selection emerging in the literature that we have chosen not to pursue at this time such as the switch-back scheme in Grunwald, et al. (12).

Estimation for Full-Membership Models. In the infinite mixture case, we assume the total number of latent semantic patterns, K, to be unknown and possibly infinite. The posterior distribution of λ, which is the goal of the posterior inference in this model, cannot be derived in closed form. However, the component-specific full conditional distributions, i.e., $\Pr(\lambda_n | \lambda_{-n})$ for $n = 1,\ldots,N$, are known up to a normalizing constant. Therefore we can explore

the desired posterior distribution of the vector λ through Markov chain Monte Carlo (MCMC) sampling methods.

Following Algorithm 3 in ref. 13, we derive the full conditional distribution of the pattern assignment vector. The full conditional probability that document (x_{1n}, x_{2n}) belongs in an existing topic k, given all documents, $\{x_{1n}, x_{2n}\}_{n=1}^{N}$, and the full-membership vectors of all other documents, λ_{-n}, is given by

$$\begin{aligned}
&\Pr(\lambda_{n[k]}=1|\lambda_{-n},\{x_{1n},x_{2n}\}_{n=1}^{N})\\
&\propto \frac{m(-n,k)}{N-1+\alpha}\times\binom{R_{1n}}{x_{1n}}\frac{\Gamma(\eta_1+\sum_v\sum_{i\neq n:\ \lambda_{i[k]}=1}x_{1i[v]})}{\prod_v\Gamma(\eta_1/V_1+\sum_{i\neq n:\ \lambda_{i[k]}=1}x_{1i[v]})}\\
&\times\frac{\prod_v\Gamma(x_{1n[v]}+\eta_1/V_1+\sum_{i\neq n:\ \lambda_{i[k]}=1}x_{1i[v]})}{\Gamma(\sum_v x_{1n[v]}+\eta_1+\sum_v\sum_{i\neq n:\ \lambda_{i[k]}=1}x_{1i[v]})}\\
&\times\binom{R_{2n}}{x_{2n}}\frac{\Gamma(\eta_2+\sum_v\sum_{i\neq n:\ \lambda_{i[k]}=1}x_{2i[v]})}{\prod_v\Gamma(\eta_2/V_2+\sum_{i\neq n:\ \lambda_{i[k]}=1}x_{2i[v]})}\\
&\times\frac{\prod_v\Gamma(x_{2n[v]}+\eta_2/V_2+\sum_{i\neq n:\ \lambda_{i[k]}=1}x_{2i[v]})}{\Gamma(\sum_v x_{2n[v]}+\eta_2+\sum_v\sum_{i\neq n:\ \lambda_{i[k]}=1}x_{2i[v]})},
\end{aligned}\tag{S12}$$

if $m(-n,k)>0$, where λ_{-n} is the topic assignment vector for all documents other than (x_{1n}, x_{2n}). The full conditional probability that document (x_{1n}, x_{2n}) belongs to a semantic pattern that no other document in the collection belongs to (number $k = K_{-n}+1$) is the following:

$$\begin{aligned}
&\Pr(\lambda_{n[k]}=1|\lambda_{-n},\{x_{1n},x_{2n}\}_{n=1}^{N})\\
&\propto\frac{\alpha}{N-1+\alpha}\times\binom{R_{1n}}{x_{1n}}\frac{\Gamma(\eta_1)\prod_v\Gamma(x_{1n[v]}+\eta_1/V_1)}{\Gamma(\eta_1/V_1)^{V_1}\Gamma(\sum_v x_{1n[v]}+\eta_1)}\\
&\times\binom{R_{2n}}{x_{2n}}\frac{\Gamma(\eta_2)\prod_v\Gamma(x_{2n[v]}+\eta_2/V_2)}{\Gamma(\eta_2/V_2)^{V_2}\Gamma(\sum_v x_{2n[v]}+\eta_2)}.
\end{aligned}\tag{S13}$$

The sparseness of $\{x_{1n}, x_{2n}\}_{n=1}^{N}$ and symmetry of the Dirichlet prior leads to a form of **S12** and **S13** that are more quickly computed. Recall that the dimensions of the vectors x_{1n} and x_{2n} are $V_1 = 30,179$ and $V_2 = 73,321$, respectively. Although the dimensions are large, many elements of each vector are equal to zero, corresponding to words that were not used in that abstract and references that were not cited in that bibliography. Only the nonzero values and their coordinates are needed for computation; the full vectors need not be held in memory. Only the nonzero elements contribute to the sums in **S12** and **S13**. Define $\mathscr{V}_i = \{v: x_{in[v]} > 0\}$, for $i = 1, 2$. The product terms associated with the zero elements of x_{1n} and x_{2n} cancel yielding the following equations:

$$\begin{aligned}
&\Pr(\lambda_{n[k]}=1|\lambda_{-n},\{x_{1n},x_{2n}\}_{n=1}^{N})\\
&\propto\frac{m(-n,k)}{N-1+\alpha}\times\frac{R_{1n}!\Gamma(\eta_1+\sum_v\sum_{i\neq n:\ \lambda_{i[k]}=1}x_{1i[v]})}{\prod_{\mathscr{V}_1}x_{1n[v]}!\Gamma(\sum_{\mathscr{V}_1}x_{1n[v]}+\eta_1+\sum_v\sum_{i\neq n:\ \lambda_{i[k]}=1}x_{1i[v]})}\\
&\times\frac{\prod_{\mathscr{V}_1}\Gamma(x_{1n[v]}+\eta_1/V_1+\sum_{i\neq n:\ \lambda_{i[k]}=1}x_{1i[v]})}{\prod_{\mathscr{V}_1}\Gamma(\eta_1/V_1+\sum_{i\neq n:\ \lambda_{i[k]}=1}x_{1i[v]})}\\
&\times\frac{R_{2n}!\Gamma(\eta_2+\sum_v\sum_{i\neq n:\ \lambda_{i[k]}=1}x_{2i[v]})}{\prod_{\mathscr{V}_2}x_{2n[v]}!\Gamma(\sum_{\mathscr{V}_2}x_{2n[v]}+\eta_2+\sum_v\sum_{i\neq n:\ \lambda_{i[k]}=1}x_{2i[v]})}\\
&\times\frac{\prod_{\mathscr{V}_2}\Gamma(x_{2n[v]}+\eta_2/V_2+\sum_{i\neq n:\ \lambda_{i[k]}=1}x_{2i[v]})}{\prod_{\mathscr{V}_2}\Gamma(\eta_2/V_2+\sum_{i\neq n:\ \lambda_{i[k]}=1}x_{2i[v]})},\\
&\text{if } m(-n,k)>0,
\end{aligned}\tag{S14}$$

$$\begin{aligned}
&\Pr(\lambda_{n[k]}=1|\lambda_{-n},\{x_{1n},x_{2n}\}_{n=1}^{N})\\
&\propto\frac{\alpha}{N-1+\alpha}\times\frac{R_{1n}!\Gamma(\eta_1)\prod_{\mathscr{V}_1}\Gamma(x_{1n[v]}+\eta_1/V_1)}{\prod_{\mathscr{V}_1}x_{1n[v]}!\Gamma(\sum_{\mathscr{V}_1}x_{1n[v]}+\eta_1)\Gamma(\eta_1/V_1)^{\sum^{1}_{\mathscr{V}_1}}}\\
&\times\frac{R_{2n}!\Gamma(\eta_2)\prod_{\mathscr{V}_2}\Gamma(x_{2n[v]}+\eta_2/V_2)}{\prod_{\mathscr{V}_2}x_{2n[v]}!\Gamma(\sum_{\mathscr{V}_2}x_{2n[v]}+\eta_2)\Gamma(\eta_2/V_2)^{\sum^{1}_{\mathscr{V}_2}}},\quad\text{otherwise.}
\end{aligned}\tag{S15}$$

The parameters of the model estimated in this way are the vector λ of pattern assignments and the total number of latent semantic patterns, K. The posterior distributions of λ and K can be found using a Gibbs sampler with these full conditional distribution as shown in Algorithm 3.

MCMC $(\{x_{1n}, x_{2n}\}_{n=1}^{N})$.

1. initialize K between 1 and N
   for $k \leftarrow 1$ **to** $K-1$ **do**
2. set $\lambda_{n[k]} := 1$ for $n = (k-1)\lfloor N/K \rfloor + 1$ to $k\lfloor N/K \rfloor$
3. set $\lambda_{n[K]} := 1$ for $n = (K-1)\lfloor N/K \rfloor + 1$ to N
   end
   repeat
   for $n \leftarrow 1$ **to** N **do**
4. sample λ_n from Multinomial (Eq. **S12**–**S13**)
5. update $K := \dim(\lambda)$
   for $k \leftarrow 1$ **to** K **do**
6. $\phi^{r}_{2[k]} \mapsto \theta_{k2[v]} \times \exp(\Psi(\gamma_{[k]}) - \Psi(\sum_{k=1}^{K}\gamma_{[k]}))$
   end
7. normalize ϕ^{r}_{2} to sum to 1
   end
8. $\gamma = \alpha + \sum_{r=1}^{R_1}\phi^{r}_{1} + \sum_{r=1}^{R_2}\phi^{r}_{2}$
   until 50 iterations after convergence (see *discussion*);
9. **return** posterior distribution of (λ, K)

Algorithm 3: The MCMC algorithm to find the posterior distribution of classification in the infinite mixture model of text and references, described in the main article.

In order to assess convergence of the Markov chain, we examine the total number of latent semantic patterns, K (which varies by Gibbs sample) and consider the Markov chain converged when the distribution of K has converged. We started chains with 10, 25, 40, and 11,988 semantic patterns, and they converged after approximately 30 iterations. Thus we are reasonably confident of convergence despite the small number of iterations because of the diversity of chain starting values.

In the estimation of the posterior distribution of λ and K, there are two hyperparameters that must be chosen. The prior distribution on λ depends on the value of α; values of α greater than one encourage more groups, whereas values of α smaller than one discourage new groups. We interpret α as the number of documents that we a priori believe belong in a new topic started by one document. However, once a document has started a new group, other documents will be less likely to join that group based on its small size. Therefore we use $\alpha = 1$ as the standard value.

The posterior distribution of λ also depends, through θ, on the η parameters. This is the Dirichlet prior on the probability vector over words or references for each semantic pattern. A value of η smaller than V, the vocabulary size, implies a prior belief that the word distributions will be highly skewed (a few likely words in every pattern and many words with almost no probability of use in any pattern). These values of η cause all documents to appear in one large group, $K = 1$. A value of η larger than V implies a prior belief that all words are equally likely to be used in a given pattern. Here, we take $\eta_1 = 1{,}000 \times V_1$ and $\eta_2 = 1{,}000 \times V_2$ as values that encourage a range of values of K.

Distributions of Shared Memberships. One advantage of using mixed-membership models is that they can model articles that do not fit into a crisp classification scheme. Discussing the details and the interpretations of the shared memberships addresses a fundamental aspect of the models being used. According to our analyses, the model of text and references with $K = 20$ is interpretable and fits the data well. In the main article, Table 3 and the discussion of the predictions address the issue of shared membership predicted by our model. Here, we look into the shared memberships for this model in more detail.

To address interpretation issues, we begin by studying the distributions of shared memberships for different values of K. We identified a threshold to decide what magnitude of posterior membership may be counted as association of an article to a latent semantic pattern, and we computed the empirical distribution of how many patterns the articles in our database are associated with, for each value of K. We consider two thresholds to establish associations of documents to latent semantic patterns: namely, a simple threshold $T_1 = 1/K$, and a threshold $T_2 = (1/K +$ standard deviation of the posteriordistribution of the memberships) that takes into account the posterior variability of the memberships. For instance, with an estimated $\alpha = 0.01$ and a model with $K = 20$ latent categories, the standard deviation of the membership of document n to latent category k, $\sigma(\lambda_{n[k]})$, is 0.209. The simple threshold T_1 would designate as associations all memberships larger than $1/20 = 0.05$. Using threshold T_2, we would call associations only those memberships of magnitude larger than $0.209 + 1/20 = 0.259$. Note that at most three such memberships can happen for each document with $K = 20$. Each threshold leads to a theoretical maximum number of associations per document, depending on the estimated value of α and the prespecified number of latent categories K.

Using these thresholds, we computed the number of documents associated with exactly zero, one, two, three, etc., latent categories corresponding to the best models of text and text and references with $K = 20$ described in the main text. The percentage of documents associated with X topics is given in Table S1. (There are 11,988 documents in total.)

This analysis suggests that most of the documents are associated with a few latent categories, much less than the theoretical limit. The words in the article abstracts are informative for the 20 latent semantic patterns; they assign most articles to three or fewer patterns. Adding the articles' references to the model helps to further focus the documents' memberships over fewer patterns.

Next, we pursued this analysis on a much larger scale. Although the new analysis does not involve careful interpretation, it aims at quantifying the extent to which our approach leads to potentially interpretable latent categories. We evaluate the potential for interpretability, by looking at the number of documents associated with exactly zero, one, two, etc., latent categories for a battery of models of text and of text and references with a prespecified number of latent categories ranging from $K = 10$ to $K = 300$. We use both thresholds described above to identify substantial associations of documents to latent categories, and we compare the number of associations to their corresponding theoretical maxima. The results are summarized in Fig. S1. We plot models with K latent categories on the Y axis, and the proportion of documents with exactly zero, one, two, etc., latent category associations on the X axis. The intensity of the color indicates the frequency of articles with exactly that many associations to latent categories for a given model and threshold. The line on each figure is the theoretical maximum number of associations for a given threshold.

This large-scale analysis confirms the findings reported above for the models of words and words and references with $K = 20$. Independent of the model and the threshold used, most documents are associated with a few latent categories, much less than the theoretical limit would allow. Overall, most documents are associated with around five latent categories in the models that consider words only as data, even as K gets large. Models that consider both words and references as data lead to a smaller number of latent categories associations per document; this result is independent of the threshold and of K. Using threshold T_2, which takes into account variability, to designate pattern association also leads to memberships that associate documents with fewer semantic patterns; this result is independent of K and of whether we consider text or text and references as data.

Macrostructure as the Model Size Increases. We sought to quantify the extent to which a model with a large number of latent categories (say $K = 300$) retains, in some sense, the macrostructure seen for the model with $K = 20$ described in the main article. Manually interpreting the results of the models with $K = 300$ latent categories is not feasible; thus we choose not to attempt a qualitative interpretation of these models. Instead, we carried out a battery of experiments aimed at exploring the impact of an increase in the number of latent categories, K, on the macrostructure and organization of the latent categories, quantitatively.

The main obstacle to this analysis is that the log-probability optimization problem is not convex for the class of membership models we consider. We need to be able to distinguish the effect of nonidentifiability present in membership models (i.e., the label switching problem described by ref. 14) on macrostructure and reorganization of the latent categories from the effect of increases in K, which is the quantity of interest. We devised a simple strategy to do this. We performed several local analyses to quantify how the K estimated distributions change as we increase K. Each of these analyses is local to a mode in the parameter space —in the sense that each analysis corresponds to a parameter set that evolves from the mode identified by the first model fit with $K = 10$. Each analysis includes fitting a battery of models with different values of K. In detail, we repeat B times the following three steps, which instantiate a local analysis:

1. We set a grid for K: 10–50, every 5, 60–150 every 10, and 175–300 every 25.

2. We fit a model of text with $K = 10$. As we repeat this step, we require that the model of text with $K = 10$ is different (modulo a Procrustes transform; e.g., see ref. 15) from other previously fit models with $K = 10$, to make sure our B analyses explore B different modes in the parameter space.
3. We fit each subsequent, larger model with K on the grid by initializing as many of the latent distributions θ as possible using the preceding model estimates for θ, and by sampling the remaining θs from a symmetric Dirichlet distribution with parameter α estimated in the previous model.

We also tried an alternative initialization scheme, where we set all the remaining θ equal to $1/V$, where V is the number of words in the vocabulary. This scheme attempts to avoid sampling distributions close to one of the existing ones, thus possibly creating conflicts that may not be trivially resolved using additional Procrustes transforms.

According to the procedure above, we performed $B = 5$ analyses from five different modes identified by the model fit with $K = 10$. This resulted in five examples of evolution of θs as K increased. Using these results, we can analyze to what extent increases in K lead to a different macrostructure and organization of the inferred classes. In each of the five analyses we observe substantial reorganization, as well as addition of θ vectors. We computed the correlation between pairs of θ vectors in order to discover for how many of them the corresponding θ vector in the larger model would still be the best match. To test whether there is stability of the inferred θ vectors to multiple evolutionary histories, this was done for each of the $B = 5$ experiments corresponding to five distinct modes in the parameter space.

In Fig. S2, we plot the evolution of θ vectors as a function of K that corresponds to one starting mode (out of five). The top-left panel, panel (1,1) in the grid, quantifies the effects on the θ vectors of increasing K from 10 to 15. In this figure, horizontally the larger value of K increases and vertically the smaller value of K in the comparison increases. In each panel, we plot two empirical cumulative distribution function (CDFs), with error bands: in black, the CDF of the correlations between corresponding θ vectors, across models, in red the CDF of the maximum correlations between additional θ vectors in the larger model and all the θ vectors in the smaller model. For instance, panel (3,4) compares the θ vectors in models with $K = 20$ and $K = 30$. In this panel, the black line is the empirical CDF of the 20 correlations between matching vector pairs (θ_1, θ_1) up to $(\theta_{20}, \theta_{20})$. The red line is the empirical CDF of the 10 maximum correlations between vector pairs $\{(\theta_i,\theta_j), i = 1,\ldots,20, j = 21,\ldots,30\}$, each corresponding to an additional θ vector in the larger model, but not in the smaller model).

In Fig. S2 we see a clear pattern, one that we consistently found in the results corresponding to the other four modes (not presented here). The red CDFs rise very rapidly because nearly all of the correlations are near 0. The black CDFs rise more slowly because many of the correlations between matching θ vectors are large.

Consider, for instance, the third row of the panels, corresponding to the evolution of the model of text with $K = 20$, discussed in the main article. First, each time we compare the model with $K = 20$ to a larger model, there is little correlation between the additional θ vectors in the larger model and those in the $K = 20$ model. The newly added θ vectors tend to have maximum correlation with the θ vectors carried over from the previous (smaller) model that are negligible as K increases—red empirical CDFs rise rapidly. We expected this because the θ vectors live in a simplex of about 30,000 dimensions. Even a model with $K = 300$ is far from spanning the whole space. Second, there is a substantial amount of reorganization among the θ vectors as K grows as suggested by the black empirical CDFs: The empirical mass shifts to lower correlations as the difference in K increases.

In order to further explore this issue, we performed a related analysis. For each model pair in the 24×24 grid in Fig. S2, we counted the number of θ vectors in the smaller model for which the correlation with the θ vector with the same index in the larger model was lower than the correlation with a θ vector with a different index. The idea is that each time such a correlation is lower, there exists a permutation of the θ vectors that leads to a better matching between the smaller and larger models. The amount of nonmatching θ vectors is an indicator of the amount of reorganization that is happening as we consider a larger model. The results in Fig. S1 are the average percent of nonmatching θs over the $B = 5$ local analyses performed.

Taken together with the results of Fig. S2, the numbers in Fig. S1 suggest that models with $K = 40$ or less have θ vectors that are the most stable. For instance, the structure of the model with $K = 20$ discussed in the main text starts breaking down when compared to the model with $K = 70$. More than 25% of the θ vectors are no longer matching the initial θ vectors as we consider the model with $K = 300$.

Simulation Study. We quantify possible bias in the optimal number of latent categories K induced by the strategy adopted by Griffiths and Steyvers (16, henceforth GS). The issue here is that by setting the constant α to $50/K$, rather than estimating it, say, via empirical Bayes, strategies for model choice based on inference mechanism would not be able to recover the number of true underlying latent categories from data. We explored this question with a simulation, where data were generated from a model with $K = 20$.

We simulated a corpus of 5,000 documents using model 1 in Table 2 of the main manuscript, from a set of patterns of co-occurring words corresponding to $K = 20$ latent categories. The documents contained 100 word occurrences, on average, from a vocabulary of 1,000 unique words. The 20 patterns of co-occuring words, defined as multinomial distributions over the vocabulary, were sampled independently from a symmetric Dirichlet distribution with a constant set to 0.01, whereas the vectors that encoded the mixed membership of documents to patterns were sampled independently from a symmetric Dirichlet distribution with a constant α set to 0.1.

Results on Dimensionality. We fit a mixed-membership model using the two strategies to assign a value to α used by Erosheva, et al. (17) and GS. We generated 25 subsamples of size 4,000, from the original 5,000 documents, fit the two models on each of them, and computed the average held-out probability on the remaining 1,000 documents. We set a grid of values for K between 10 and 1,000, and a grid of α values between 0.01 and 0.50. We both estimated α using empirical Bayes (model 1) and we set it to $\alpha = 5/K$ (model 2). Two alternative model selection strategies, Bayesian information criterion (BIC) and held-out log-probability, lead to similar number of topics for the respective best models, as seen in Fig. S3. The two alternative estimation strategies for the constant α do not impact model fit, significantly. This result confirms our findings on the real data. Rather than a strong bias toward a higher number of latent patterns, the impact of the strategy adopted by GS takes the form of a comparatively poor model fit. The poor model fit can be attributed to a suboptimal estimation—in the probability sense—of the variability of the memberships of documents to latent patterns. A possible source of variability is given by the strategy of fixing $\alpha(K) = 5/K$ in model 2, the strategy used by ref. 16, where the constant is now 5 because we will consider fitting the model with as little as $K = 5$ latent categories. In Fig. S5 we contrast $\alpha(K)$ to $\hat{\alpha}$ obtained via empirical Bayes. The value of $\hat{\alpha}$ is always lower than $\alpha(K)$, thus inducing more variability in the memberships of documents to latent patterns, by increasing the sparsity. Although this difference is noticeable, the behavior of a symmetric Dirichlet

distribution changes dramatically for α values above and below one, whereas both $\alpha(K)$ and $\hat{\alpha}$ are well below one in this case. In light of this, we expect the impact of the two strategies to be minimal on the goodness of fit, as measured by the probability and functions of it. The difference on model fit, however, is more subtle and it can be significant. For instance, see the discussion of Fig. 3, in the main manuscript, and of Fig. S6. In Fig. S3 we measure the actual impact of the two strategies on model fit for a value of $\alpha = 0.1$ in the simulated data. The two panels show that similar values of K are chosen under both models with either model selection technique. This confirms that the effect of the choice of α on the estimation of K is minimal.

A Tale of Two Maxima. In our simulation, the maxima for BIC and held-out log probability are at about the correct value, $K \approx 20$. We consistently observe a local maximum at high K using the strategy that fixes $\alpha(K)$, for both model selection criteria, in Fig. S4.

A local maximum exists for K large (e.g., $K > 300$) as we fix α and we consider the smoothed probability surface as a function of K, in Fig. S4. This local maximum is dominated by a global maximum of the smoothed probability surface for K small (e.g., $K < 50$). This finding is consistent with the empirical findings, presented in the main article.

Running Time and Complexity. In all experiments, the distributions over the vocabulary corresponding the K latent patterns were initialized using distributions of words in randomly chosen documents. For models with $K \leq 200$ there is little practical difference in the running time of fitting models 1 and 2. In particular, in the range of values that fit the data well, the two models run in approximately the same amount of time.

For N documents, K latent patterns on a vocabulary of V words, let us denote the number of word occurrences with I, and the number of iterations till convergence of the posterior inference algorithm employed by T. Then, the complexity of fitting a model of multivariate attributes that follows the general specifications in ref. 17 is

$$O(I + NVKT + K^2T),$$

as derived in ref. 18.

Description of the Latent Categories. The output from the model using abstract text and references with α estimated and $K = 20$ was used as an illustration in the main paper. This model fits the data reasonably well, and we can interpret all 20 of the latent categories. Amplifying the results presented in the main paper, here we give further qualitative and quantitative analysis of the fit of this model and of the predictions for dual-classified articles.

Qualitative analysis of the latent categories. Table S4 gives the top 20 most probable words and references in each of the 20 latent categories. Using the likely words and references, we found the following interpretations of the 20 estimated latent categories: latent category 1 is about population genetics. Category 2 concerns the activation of enzymes by protein kinases. Category 3 focuses on problems of hormone levels, and categories 4 and 5 on nuclear activity (production of cdna and mrna) and (catalysts for DNA copying). We observe two categories associated with HIV, category 6 (HIV and immune response) and category 12 (T-cell response to HIV infection). Category 7 is related to plant evolution and phylogenetic relationships. Several categories are related to protein studies, for instance, category 8 (protein structure and folding) and category 11 (protein promotion by transcription binding factors). Category 14 deals with cancer markers. We can identify two categories related to tumor experiments with mice: category 13 (mutant mice and tumor suppression), category 18 (tumor treatment for mice and humans), whereas category 15 has to do with bone marrow stem cells. Four topics relate to the brain and neurons: category 16 (functional and visual responses to changes in the brain), category 17 (neurons and neurotransmitters), category 19 (nervous system development), and category 20 (electrical excitability of cell membranes). Categories 9 and 10 are the most ambiguous, relating to procedural explanations and genetic mutation broadly.

Quantitative analysis of the latent categories. Fig. S6 is an extended version of Fig. 3 in the main article, including all 19 PNAS classifications. These extended plots suggest a pattern similar to that observed in Fig. 3. When the hyperparameter α is estimated, we observe a better agreement between estimated latent categories and the original PNAS classifications. A greater number of darker color blocks point to more articles with estimated substantial membership in just a few latent categories for the α-estimated models. Lighter blocks for the constrained-α models may be due to more spread out membership (due to small membership values of all articles) or to an apparent disagreement among estimated membership vectors among articles from original PNAS classifications (due to dissimilar membership vectors). Either explanation leads us to conclude that estimating hyperparameters leads to a set of estimated latent categories that is better correlated to the original PNAS classification.

Table S3 helps us examine the composition of latent categories in terms of nouns, verbs, and adjectives. We observe that the proportions of nouns range from 26 to 49%, the proportion of verbs—from 10 to 20%, and stop words—from 24 to 42%. None of the inferred latent categories appears to be a technical service category that includes disproportional amounts of words of any one particular type.

Predictions of dual classification. In the main paper, Table 3 describes articles that are similar to 181 articles that the authors classified into more than one of the 19 subcategories of the Biological Sciences section. A total of 1,489 articles were predicted to be similar to the author dual-classified articles. Most of these were predicted as dual classified by only one model. We further examined 13 articles that are predicted to be similar to dual-classified articles. For all of them, a case could be made for dual classification. For 3 of them, however, such an argument is not straightforward.

For 10 of the 13 articles further examined, a dual classification appears to be justified because of the interdisciplinary scientific approach. Four of the papers were biochemistry and biophysics papers regarding protein structures (similar to papers dual classified in the Physical Sciences by their authors). Three papers were in the Plant Biology PNAS classification and had significant genetic and population biology content. Three more papers from the Neurobiology and Psychology classifications blended sophisticated measurement of the brain with hypotheses and conclusions regarding emotion and learning.

For example, article number 238, entitled "Emotion-induced changes in human medial prefrontal cortex: I. During cognitive task performance," is predicted to have dual classification by three of the four mixed-membership models. The paper's hypothesis involves the psychological quantity of emotional content. Changes in cerebral blood flow were identified in regions of the medial prefrontal cortex using neurobiology techniques. Although the authors submitted the paper to the Neurobiology classification, we conclude that it could also have been classified in Psychology, as the models suggest.

1. Rustagi J (1976). *Variational Methods in Statistics* (Academic Press, New York).
2. Sakurai J (1985). *Modern Quantum Mechanics* (Addison-Wesley, Redwood City, CA).
3. Parisi G (1988). *Statistical Field Theory* (Addison-Wesley, Redwood City, CA).
4. Bathe KJ (1996). *Finite Element Procedures* (Prentice Hall, Englewood Cliffs, NJ).
5. Jordan MI, Ghahramani Z, Jaakkola T, Saul L (1999). Introduction to variational methods for graphical models. *Mach Learn* 37:183–233.
6. Xing EP, Jordan MI, Russell S (2003). *Uncertainty in Artificial Intelligence* (Morgan Kaufmann, San Francisco), pp 583–591.
7. Wainwright MJ, Jordan MI (2003). Graphical models, exponential families and variational inference. *Found Trends Mach Learn* 1:1–305.
8. Hastie T, Tibshirani R, Friedman JH (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York), 2nd Ed.
9. Breiman L, Freidman JH, Olshen RA, Stone CJ (1984) *Classification and Regression Trees* (Wadsworth, Belmont, CA).
10. Kohavi, R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)* (Morgan Kaufmann, San Francisco), pp 1137–1143.
11. Mahmood Z, Khan, S (2009) On the use of K-fold cross-validation to choose cutoff values and assess the performance of predictive models in stepwise regression. *Int J Biostatistics* 5(1):Article 25.
12. Grünwald, PD (2007) *The Minimum Description Length Principle* (MIT Press, Cambridge, MA).
13. Neal R (2000). Markov chain sampling methods for Dirichlet process mixture models. *J Comp Graph Stat* 9:249–265.
14. Stephens M (2000). Dealing with label switching in mixture models. *J R Stat Soc B Met* 62:795–809.
15. Dryden IL, Mardia KV (1998). *Statistical Shape Analysis*. (Wiley, New York).
16. Griffiths TL, Steyvers M (2004). Finding scientific topics. *Proc Natl Acad Sci USA* 101:5228–5235.
17. Erosheva EA, Fienberg SE, Lafferty J (2004) Mixed-membership models of scientific publications. *Proc Natl Acad Sci USA* 101:5220–5227.
19. Joutard CJ, Airoldi EM, Fienberg SE, Love TM (2008). *Data Mining Patterns: New Methods and Applications*, eds Poncelet P, Masseglia F, Teisseire M (IGI Global, Hershey, PA), pp 240–275.
20. Airoldi EM (2006). Bayesian mixed membership models of complex and evolving networks. Doctoral dissertation (School of Computer Science, Carnegie Mellon University, Pittsburgh).

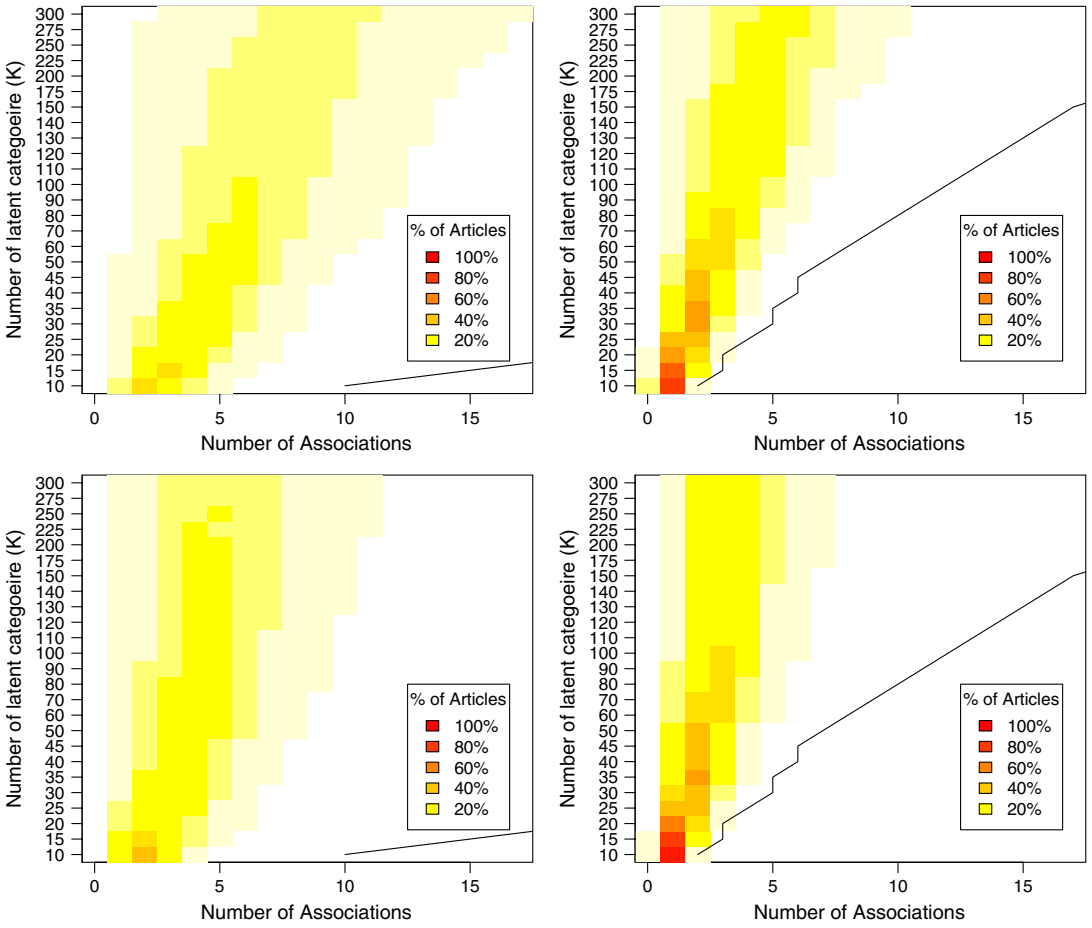

Fig. S1. Distributions of the number of article associations obtained with simple threshold T_1 (*Left*) and threshold T_2 (*Right*) for models fitted with words only (*top*) and words and references (*bottom*). The black line indicates the theoretical maximum number of associations.

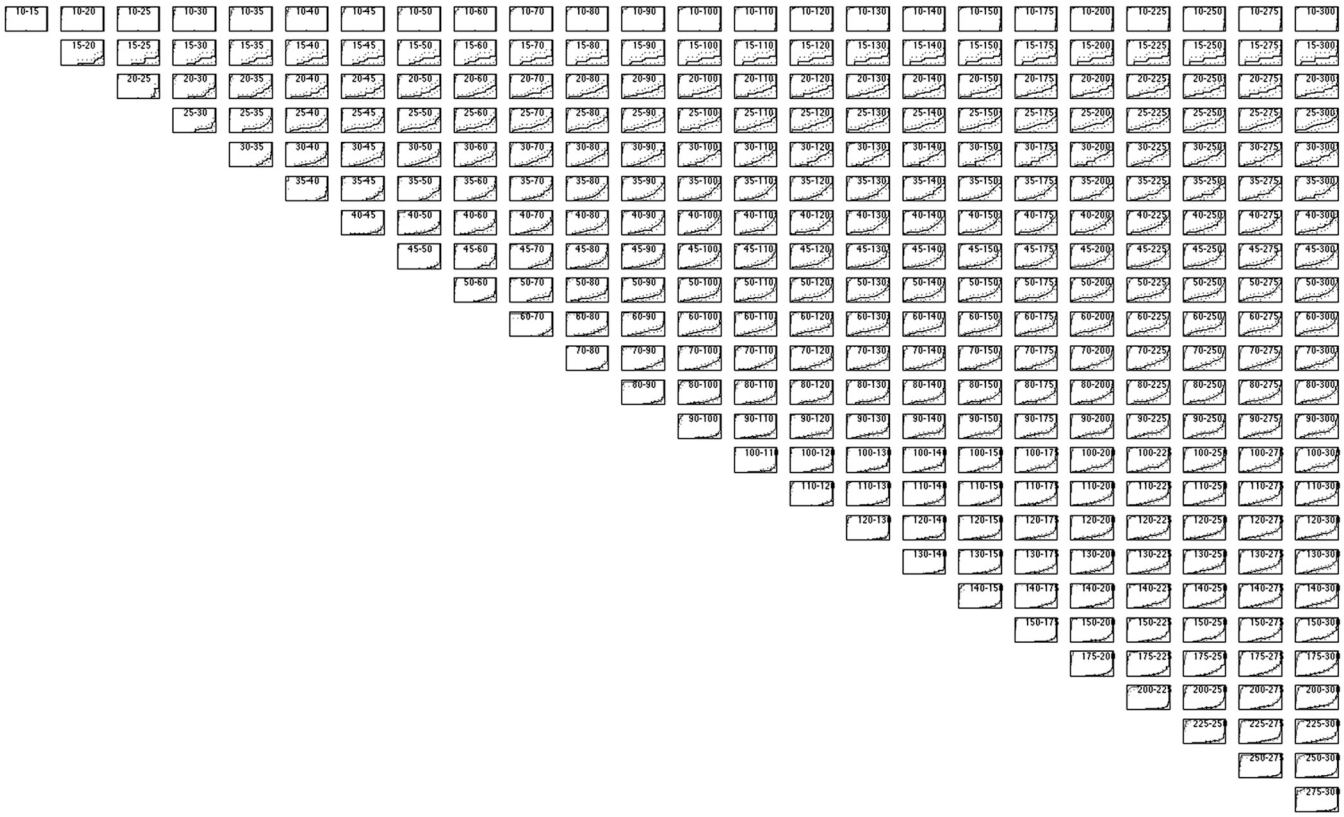

Fig. S2. The complete evolution of θ vectors as a function of *K*. See text for details.

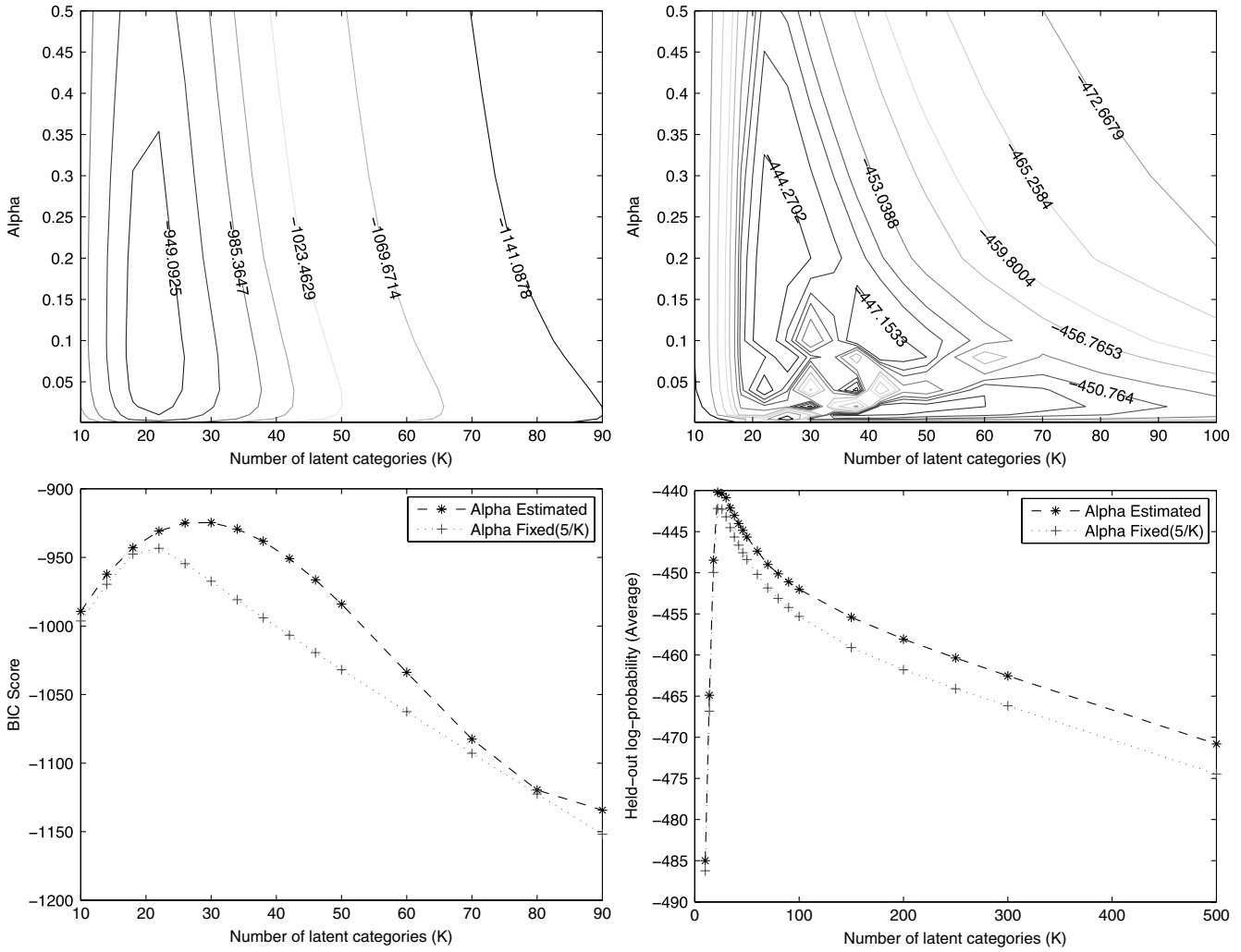

Fig. S3. The contour plots of average BIC (*Top Left*) and average held-out log probability (*Top Right*) on a portion of the (α; K) grid that includes the peaks—the two model selection strategies agree. The average BIC (*Bottom Left*) and held-out log probability (*Bottom Right*) vs $\alpha(K)$ and $\hat{\alpha}$.

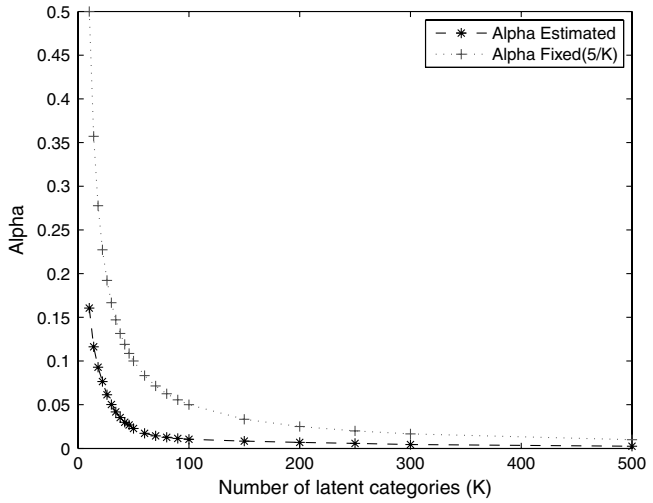

Fig. S4. Smoothed probability surface estimated by regression from 25 held-out log-probability estimates on covariates (α; K).

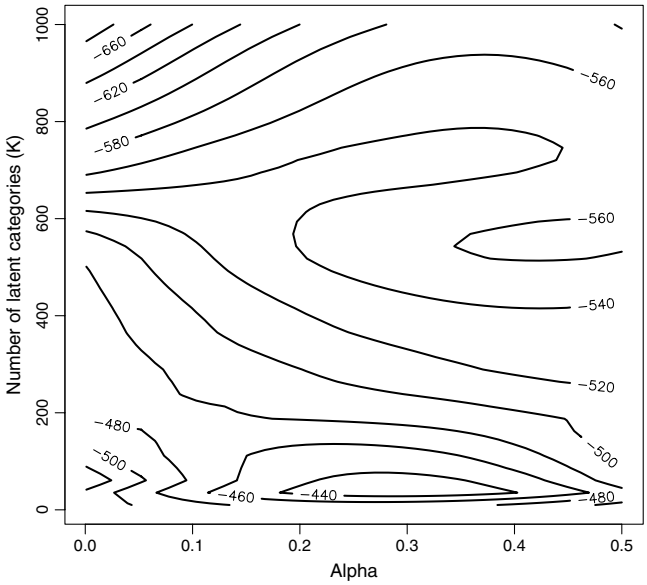

Fig. S5. The behavior of $\alpha(K)$ is contrasted to that of $\hat{\alpha}$.

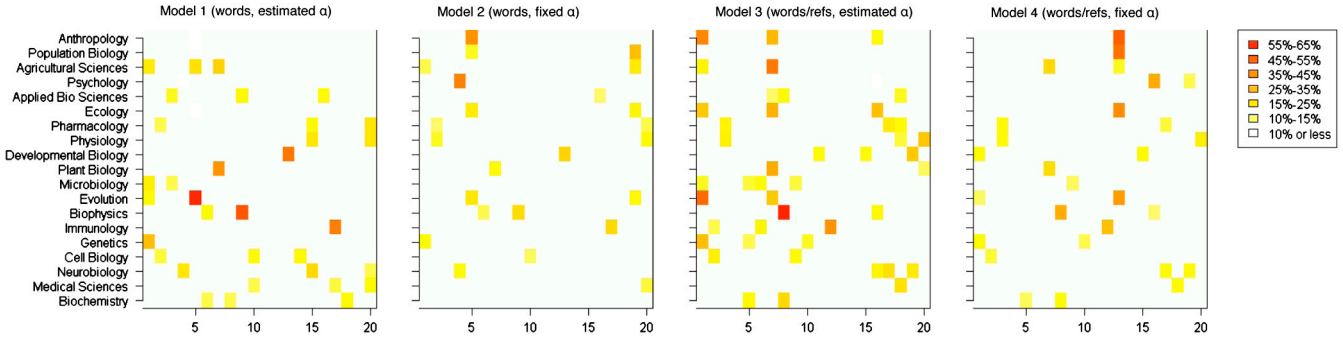

Fig. S6. Estimated membership of articles in the 20 latent categories by PNAS classifications for mixed-membership models 1–4 in Table 2 of the main article, left to right.

Table S1. Empirical distribution of mixed memberships for the models with $K = 20$

			Article associated with exactly this many categories (in % points)									
Model	Threshold	Max Assoc	0	1	2	3	4	5	6	7	8	9
Text	0.050	20	0.00	4.63	18.51	30.08	27.14	14.31	4.15	1.04	0.13	0.01
Text and refs.	0.050	20	0.00	12.95	30.08	31.77	17.43	6.09	1.49	0.17	0.03	0.01
Text	0.259	3	1.66	57.33	39.11	1.90	0	0	0	0	0	0
Text and refs.	0.259	3	0.55	60.21	37.44	1.80	0	0	0	0	0	0

Table S2. Percentage of nonmatching θ vectors between pairs of models of different sizes

to → / from ↓	15	20	25	30	35	40	45	50	60	70	80	90	100	110	120	130	140	150	175	200	225	250	275
10	4	6	6	8	9	9	11	12	12	12	16	16	16	16	16	16	16	16	16	16	16	16	16
15		0	1	2	5	7	7	7	8	12	14	17	18	18	19	19	19	19	19	20	21	21	21
20			0	1	1	3	3	7	7	10	15	16	19	21	23	24	24	24	24	25	26	27	27
25				0	1	2	3	5	7	8	12	14	14	16	20	20	21	21	21	22	23	24	24
30					0	1	1	2	2	4	6	10	10	10	12	14	16	17	17	18	19	19	20
35						0	0	1	2	2	4	7	8	9	9	11	12	13	14	16	18	18	18
40							0	0	0	1	4	6	7	7	8	10	12	12	12	13	14	15	15
45								1	1	2	3	7	7	8	9	11	13	14	14	15	17	18	19
50									0	1	4	6	6	9	11	12	14	15	15	15	17	18	20
60										0	0	2	3	5	7	8	10	12	13	14	16	17	17
70											0	0	2	2	4	5	5	6	8	11	12	12	12
80												0	0	0	2	3	4	5	6	7	9	12	13
90													0	0	1	3	3	3	4	6	7	9	10
100														0	0	1	3	3	5	6	7	7	8
110															0	0	0	0	5	6	7	7	8
120																0	0	1	3	6	8	8	9
130																	0	0	3	3	4	5	5
140																		0	1	1	1	4	5
150																			0	0	0	2	2
175																				0	0	1	2
200																					0	0	0
225																						14	15
250																							29
275																							

Table S3. Word type distributions of the patterns

Pattern	Nouns	Verbs	Stop words
1	31%	19%	35%
2	40%	18%	30%
3	38%	23%	24%
4	30%	14%	32%
5	26%	15%	42%
6	30%	12%	33%
7	42%	14%	30%
8	38%	17%	31%
9	31%	15%	34%
10	49%	16%	29%
11	42%	13%	30%
12	33%	15%	34%
13	33%	16%	34%
14	42%	14%	30%
15	37%	10%	28%
16	43%	15%	27%
17	38%	13%	30%
18	36%	15%	31%
19	35%	19%	28%
20	42%	13%	27%

Table S4. Word usage patterns corresponding to the model of text and references, with $K = 20$ topics

Latent Category 1: Population genetics

Words:

genes, gene, genetic, chromosome, number, species, sequence, analysis, genome, sequences, expression, human, dna, selection, region, different, population, evolution, data, loci

References:

Molecular Cloning: A Laboratory Manual (2nd ed.)
Basic local alignment search tool
Cluster analysis and display of genome-wide expression patterns
Gapped BLAST and PSI-BLAST: A new generation of protein database search programs
CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-spe
Quantitative monitoring of gene expression patterns with a complementary DNA microarray
The neighbor-joining method: A new method for reconstructing phylogenetic trees
Genetic transformation of Drosophila with transposable element vectors
Expression monitoring by hybridization to high-density oligonucleotide arrays
Exploring the metabolic and genetic control of gene expression on a genomic scale
The Genome of Drosophila melanogaster
Drosophila a Laboratory Manual
The transcriptional program in the response of human fibroblasts to serum
Statistical method for testing the neutral mutation hypothesis by DNA polymorphism
Serial analysis of gene expression
The principles and practice of statistics in biological research
The neutral theory of molecular evolution
Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction
Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentia
The future of genetic studies of complex human diseases

Latent Category 2: Activation of enzymes by protein kinases

Words:

kinase, cells, cell, activation, protein, apoptosis, induced, phosphorylation, signaling, activity, tyrosine, activated, expression, receptor, factor, growth, death, kappa, domain, nf

References:

Apoptosis in the pathogenesis and treatment of disease
Opposing effects of ERK and JNK-p38 MAP kinases on apoptosis
Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins
Apaf-1, a human protein homologous to C. elegans CED-4, participates in cytochrome c-dependent activation of caspase-3
FLICE, a novel FADD-homologous ICE/CED-3-like protease, is recruited to the CD95 (Fas/APO-1) death-inducing signaling c
Cytochrome c and dATP-dependent formation of Apaf-1/caspase-9 complex initiates an apoptotic protease cascade
Involvement of MACH, a novel MORT1/FADD-interacting protease, in Fas/APO-1 and TNF receptor-induced cell death
Production of high-titer helper-free retroviruses by transient transfection
JNK1: A protein kinase stimulated by UV light and Ha-Ras that binds and phosphorylates the c-Jun activation domain
STATs and gene regulation
Akt phosphorylation of BAD couples survival signals to the cell-intrinsic death machinery
The NF-kappa B and I kappa B proteins: New discoveries and insights
Dissection of TNF receptor 1 effector functions: JNK activation is not linked to apoptosis while NF-kappaB activation pr
Identification and inhibition of the ICE/CED-3 protease necessary for mammalian apoptosis

Induction of apoptotic program in cell-free extracts: Requirement for dATP and cytochrome c
The Bcl-2 protein family: Arbiters of cell survival
Apoptosis by death factor
Caspases: Enemies within
Specificity of receptor tyrosine kinase signaling: transient versus sustained extracellular signal-regulated kinase acti
FADD, a novel death domain-containing protein, interacts with the death domain of Fas and initiates apoptosis

Latent Category 3: Hormone levels

Words:

alpha, beta, cells, receptor, induced, expression, increased, activity, levels, hormone, estrogen, cox, effects, cardiac, effect, mrna, gene, human, fold, insulin

References:

Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction
A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding
Cloning of a novel receptor expressed in rat prostate and ovary
Nitric oxide: Physiology, pathophysiology, and pharmacology
Sequence and characterization of a coactivator for the steroid hormone receptor super-family
Differential expression of estrogen receptors alpha and beta mRNA during differentiation of human osteoblast SV-HFO cell
The SREBP pathway: Regulation of cholesterol metabolism by proteolysis of a membrane-bound transcription factor
Protein measurement with the folin phenol reagent
The RXR heterodimers and orphan receptors
Alteration of reproductive function but not prenatal sexual development after insertional disruption of the mouse estrog
Molecular mechanisms of action of steroid/thyroid receptor superfamily members
Positional cloning of the Werner s syndrome gene
Apparent hydroxyl radical production by peroxynitrite: implications for endothelial injury from nitric oxide and superox
Oxidants, antioxidants, and the degenerative diseases of aging
Molecular basis of agonism and antagonism in the oestrogen receptor
Reactions between nitric oxide and haemoglobin under physiological conditions
The steroid and thyroid hormone receptor superfamily
Differential ligand activation of estrogen receptors ER and ER at AP1 sites
SREBP-1, a membrane-bound transcription factor released by sterol-regulated proteolysis
The Bloom s syndrome gene product is homologous to RecQ helicases

Latent Category 4: Production of cdna and mrna

Words:

protein, mrna, rna, nuclear, cells, proteins, splicing, sequence, activity, cdna, gene, binding, expression, human, telomerase, cell, export, specific, mrnas, exon

References:

Molecular Cloning: A Laboratory Manual (2nd ed.)
Specific association of human telomerase activity with immortal cells and cancer
Extension of life-span by introduction of telomerase into normal human cells
A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding
The HIV-1 Rev activation domain is a nuclear export signal that accesses an export pathway used by specific cellular RNA
Cleavage of structural proteins during the assembly of the head of bacteriophage T4
Telomeres shorten during ageing of human fibroblasts
CRM1 is an export receptor for leucine-rich nuclear export signals
Nucleocytoplasmic transport
Telomerase catalytic subunit homologs from fission yeast and human
Identification of a signal for rapid export of proteins from the nucleus
The RNA component of human telomerase
Nucleocytoplasmic transport: Signals, mechanisms and regulation
Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei
Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction
Nucleocytoplasmic transport: the soluble phase
Exportin 1 (Crm1p) is an essential nuclear export factor
Splicing of precursors to mRNA by the spliceosome
Telomere shortening associated with chromosome instability is arrested in immortal cells which express telomerase activi
CRM1 is responsible for intracellular transport mediated by the nuclear export signal

Latent Category 5: Catalysts for dna copying

Words:

RNA, protein, binding, transcription, sequence, gene, site, promoter, strand, specific, complex, polymerase, region, activity, single, base, sequences, sites, histone, proteins

References:

Molecular Cloning: A Laboratory Manual (2nd ed.)
Experiments in molecular genetics
A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p
Cleavage of structural proteins during the assembly of the head of bacteriophage T4
Histone acetylation in chromatin structure and transcription
Crystal structure of the nucleosome core particle at 2.8 A resolution
DNA sequencing with chain-terminating inhibitors
Histone acetylation and transcription regulatory mechanisms
System of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae
A p300/CBP-associated factor that competes with the adenoviral oncoprotein E1A

Biochemistry of homologous recombination in E. coli
Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii
The transcriptional coactivators p300 and CBP are histone acetyltransferases
A complex containing N-CoR, mSin3 and histone deacetylase mediates transcriptional repression
The minimal gene complement of Mycoplasma genitalium
Role for N-CoR and histone deacetylase in Sin3-mediated transcriptional repression
Communication modules in bacterial signaling proteins
Tetrahymena histone acetyltransferase A: A homolog to yeast Gcn5p linking histone acetylation to gene activation
Gapped BLAST and PSI-BLAST: A new generation of protein database search programs
Transcriptional silencing in yeast is associated with reduced nucleosome acetylation

Latent Category 6: HIV and immune response
Words:
virus, cells, hiv, infection, beta, viral, infected, alpha, cell, gamma, immune, response, specific, disease, replication, reproduction, mice, human, host, ifn
References:
Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease
Rapid turnover of plasma virion and CD4 lymphocytes in HIV-1 infection.
Secreted amyloid beta-protein similar to that in the senile plaques of Alzheimer's disease is increased in vivo by the p
Endoproteolysis of presenilin 1 and accumulation of processed derivatives in vivo
Viral dynamics in human immunodeficiency virus type 1 infection
Vigorous HIV-1-specific CD4+ T cell responses associated with control of viremia
Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection
Deficiency of presenilin-1 inhibits the normal cleavage of amyloid precursor protein
Recovery of replication-competent HIV despite prolonged suppression of plasma viremia
Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy
HIV infection is active and progressive in lymphoid tissue during the clinically latent stage of disease
Skeletal and CNS defects in presenilin-1-deficient Mice
Correlative memory deficits, Abeta elevation, and amyloid plaques in transgenic mice
Temporal association of cellular immune responses with the initial control of viremia in primary HIV-1 syndrome
Alzheimer-type neuropathology in transgenic mice overexpressing V717F beta-amyloid precursor protein
Increased amyloid-beta42(43) in brains of mice expressing mutant presenilin
Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer s disease
A presenilin-1 dependent g-secretase-like protease mediates release of Notch intracellular domain
HIV population dynamics in vivo: Implications for genetic variation, pathogenesis, and therapy
Facilitation of lin-12-mediated signalling by sel-12, a Caenorhabditis elegans S182 Alzheimer s disease gene

Latent Category 7: Plant evolution and phylogenetic relationships
Words:
plants, plant, gene, genes, species, acid, arabidopsis, mitochondrial, resistance, sequence, biosynthesis, phylogenetic, expression, data, analysis, leaves, mutant, transgenic, results, sequences
References:
Molecular Cloning: A Laboratory Manual (2nd ed.)
CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight
Basic local alignment search tool
The neighbor-joining method: A new method for reconstructing phylogenetic trees
BLAST and PSI-BLAST: A new generation of protein database search programs
Confidence limits on phylogenies: An approach using the bootstrap
A rapid and sensitive method for the quantification of microgram quantities of protein utilizing the principle of protein-dye binding
Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome
Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branch
A revised medium for rapid growth and bioassays with tobacco tissue cultures
Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies
Cleavage of structural proteins during the assembly of the head of bacteriophage T4
In planta Agrobacterium mediated gene transfer by infiltration of adult Arabidopsis thaliana plants
Systemic acquired resistance
A comprehensive set of sequence analysis programs for the VAX
A procedure for mapping Arabidopsis mutations using co-dominant ecotype-specific PCR-based markers
Dating of the human-ape splitting by a molecular clock of mitochondrial DNA
Resistance gene-dependent plant defense responses
Genes galore: A summary of methods for accessing results from large-scale partial sequencing of anonymous Arabidopsis cD
Assignment of 30 microsatellite loci to the linkage map of Arabidopsis

Latent Category 8: Protein structure and folding
Words:
protein, structure, binding, residues, site, proteins, enzyme, domain, state, active, folding, beta, structural, amino, structures, helix, reaction, substrate, energy, complex
References:
MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures
Improved methods for building protein models in electron density maps and the location of errors in these models
Processing of X-ray diffraction data collected in oscillation mode
Crystallography and NMR system: A new software suite for macromolecular structure determination

PROCHECK—A program to check the stereochemical quality of protein structures
Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons
AMoRe: An automated package for molecular replacement
Raster3D version 2.0. A program for photorealistic molecular graphics
A novel statistical quantity for assessing the accuracy of crystal structures
Raster3D: Photorealistic molecular graphics
NMR of proteins and nucleic acids
Cleavage of structural proteins during the assembly of the head of bacteriophage T4
Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features
Protein structure comparison by alignment of distance matrices
CHARMM: A program for macromolecular energy, minimization, and dynamics calculations
Funnels, pathways, and the energy landscape of protein folding: A synthesis
MOLMOL: A program for display and analysis of macromolecular structures
Oscillation data reduction program
The CCP4 suite: Programs for protein crystallography
Accurate bond and angle parameters for X-ray protein structure refinement

Latent Category 9: Procedural explanations
Words:
proteins, membrane, domain, binding, terminal, cells, fusion, cell, complex, domains, kda, surface, receptor, interaction, membranes, amino, sequence, transport, golgi, cytoplasmic
References:
Cleavage of structural proteins during the assembly of the head of bacteriophage T4
Molecular Cloning: A Laboratory Manual (2nd ed.)
Basic local alignment search tool
Antibodies: A laboratory manual
Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: Procedure and some applications
A novel genetic system to detect protein-protein interactions
A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding
DNA sequencing with chain-terminating inhibitors
Functional rafts in cell membranes
Mechanisms of intracellular protein transport
A simple method for displaying the hydropathic character of a protein
SNAP receptors implicated in vesicle targeting and fusion
Protein sorting by transport vesicles
Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites
The t(4;11) chromosome translocation of human acute leukemias fuses the *ALL-1* gene, related to Drosophila *trithorax*, to the *AF-4* gene
Predicting coiled coils from protein sequences
A new method for predicting signal sequence cleavage sites
The structure of influenza haemagglutinin at the pH of membrane fusion
Protein measurement with the folin phenol reagent
Involvement of a homolog of Drosophila trithorax by 11q23 chromosomal translocations in acute leukemias

Latent Category 10: Genetic mutation
Words:
DNA, protein, cell, yeast, cells, gene, proteins, activity, mutations, complex, human, replication, mutant, mutation, results, delta, binding, function, nuclear, mutants
References:
Molecular Cloning: A Laboratory Manual (2nd ed.)
A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisia
Experiments in molecular genetics
Lessons from hereditary colorectal cancer
The RXR heterodimers and orphan receptors
DNA repair and mutagenesis
Ubiquitin-dependent protein degradation
The ubiquitin system
The nuclear receptor superfamily: The second decade
K. Ligand-independent repression by the thyroid hormone receptor mediated by a nuclear receptor co-repressor
A transcriptional co-repressor that interacts with nuclear hormone receptors
Inhibition of proteasome activities and subunit-specific amino-terminal threonine modification by lactacystin
ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex
Transformation of intact yeast cells treated with alkali cations
Structure and functions of the 20S and 26S proteasomes
A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1
Methods in Yeast Genetics: A Laboratory Course Manual
Clues to the pathogenesis of familial colorectal cancer
How proteolysis drives the cell cycle
Inactivation of the mouse Msh2 gene results in mismatch repair deficiency, methylation tolerance, hyperecombination and

Latent Category 11: Protein promotion by transcription binding factors
Words:
transcription, protein, gene, expression, transcriptional, beta, binding, factor, promoter, activation, human, proteins, cells, domain, activity, genes, cell, tgf, nuclear, specific
References:

p53, the cellular gatekeeper for growth and division
Molecular Cloning: A Laboratory Manual (2nd ed.)
Mdm2 promotes the rapid degradation of p53
TGF-beta signalling from cell membrane to nucleus through SMAD proteins
A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polym
p53: Puzzle and paradigm
Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei
Partnership between DPC4 and SMAD proteins in TGF-beta signalling pathways
Regulation of p53 stability by Mdm2
A CBP Integrator Complex Mediates Transcriptional Activation and AP-1 Inhibition by Nuclear Receptors
Transcriptional activation by recruitment
p53 mutations in human cancers
The CBP co-activator is a histone acetyltransferase
Activation of p53 sequence-specific DNA binding by acetylation of the p53 C-terminal domain
The mdm-2 oncogene product forms a complex with the p53 protein and inhibits p53-mediated transactivation
Receptor-associated Mad homologues synergize as effectors of the TGF-beta response
The general transcription factors of RNA polymerase II
MADR2 is a substrate of the TGFbeta receptor and its phosphorylation is required for nuclear accumulation and signaling
The role of general initiation factors in transcription by RNA polymerase II
The p53 mdm-2 autoregulatory feedback loop

Latent Category 12: T-cell response to HIV infection
Words:
cells, cell, class, receptor, antigen, peptide, mhc, specific, cd4, tcr, alpha, mice, molecules, beta, human, hla, peptides, tumor, cd8, immune
References:
Identification of a major co-receptor for primary isolates of HIV-1
HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5
A Dual-Tropic Primary HIV-1 Isolate That Uses Fusin and the beta-Chemokine Receptors CKR-5, CKR-3, and CKR-2b as Fusion
The beta-Chemokine Receptors CCR3 and CCR5 Facilitate Infection by Primary HIV-1 Isolates
CC CKR5: a RANTES, MIP-1alpha, MIP-1beta receptor as a fusion cofactor for macrophage-tropic HIV-1
HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor
Identification of Rantes, MIP-1alpha, and MIP-1beta as the major HIV-suppressive factors produced by CD8+ T cells
Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection
Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene
Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene
CD4-induced interaction of primary HIV-1 gp120 glycoproteins with the chemokine receptor CCR-5
Structure of the complex between human T-cell receptor, viral peptide and HLA-A2
The dendritic cell system and its role in immunogenicity
CD4-dependent, antibody-sensitive interactions between HIV-1 and its co-receptor CCR
Phenotypic analysis of antigen-specific T lymphocytes
Efficient presentation of soluble antigen by cultured human dendritic cells is maintained by granulocyte/macrophage-colony-stimulating factor plus interleukin 4 and downregulated by tumor necrosis factor alpha
T cell receptor antagonist peptides induce positive selection
MHC ligands and peptide motifs: First listing
Dendritic cells and the control of immunity
An alpha-beta T Cell Receptor structure at 2.5 Å and its orientation in the TCR-MHC complex

Latent Category 13: Mutant mice and tumor suppression
Words:
mice, type, wild, p53, mutant, deficient, mutation, mutations, normal, null, gene, phenotype, dna, mouse, knockout, homozygous, role, transgenic, radiation, mutant
References:
WAF1, a potential mediator of p53 tumor suppression
Molecular Cloning: A Laboratory Manual (2nd ed.)
A mammalian cell cycle checkpoint pathway utilizing p53 and GADD45 is defective in ataxia-telangiectasia
The p21 Cdk-interacting protein Cip1 is a potent inhibitor of G1 cyclin-dependent kinases
Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours
Mice Lacking p21 CIP1/WAF1 undergo normal development, but are defective in G1 checkpoint control
Enhanced Phosphorylation of p53 by ATM in Response to DNA Damage
Activation of the ATM Kinase by Ionizing Radiation and Phosphorylation of p53
A single ataxia telangiectasia gene with a product similar to PI-3 kinase
Crystal structure of a p53 tumor suppressor-DNA complex: Understanding tumorigenic mutations
p21 is a universal inhibitor of cyclin kinases
Tumor spectrum analysis in p53-mutant mice
p53, guardian of the genome
Defective DNA-dependent protein kinase activity is linked to V(D)J recombination and DNA repair defects associated with
Transgenetic studies implicate interactions between homologous PrP isoforms in scrapie prion replication
Role of the INK4a Locus in Tumor Suppression and Cell Mortality
ARF Promotes MDM2 Degradation and Stabilizes p53: ARF-INK4a Locus Deletion Impairs Both the Rb and p53 Tumor Suppression
A model for p53-induced apoptosis
Prion propagation in mice expressing human and chimeric PrP transgenes implicates the interaction of cellular PrP with a
Wild-type p53 is a cell cycle checkpoint determinant following irradiation

Latent Category 14: Cancer markers

Words:

beta, cell, factor, cells, protein, human, vegf, binding, expression, endothelial, activity, catenin, expressed, hypoxia, growth, alpha, thrombin, vascular, tf, cadherin

References:

Activation of Beta-Catenin-Tcf Signaling in Colon Cancer by Mutations in Beta-Catenin or APC
Constitutive Transcriptional Activation by a Beta-Catenin-Tcf Complex in APC/Colon Carcinoma
Abnormal blood vessel development and lethality in embryos lacking a single VEGF allele
Functional interaction of Beta-catenin with the transcription factor LEF-1
Hypoxia-inducible factor 1 is a basic-helix-loop-helix-PAS heterodimer regulated by cellular O2 tension
Regulation of intracellular beta-catenin levels by the adenomatous polyposis coli (APC) tumor-suppressor protein
Heterozygous embryonic lethality induced by targeted inactivation of the VEGF gene
Stabilization of Beta-Catenin by Genetic Defects in Melanoma Cell Lines
XTcf-3 Transcription Factor Mediates Beta-Catenin-Induced Axis Formation in Xenopus Embryos
Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction
Binding of the von Hippel-Lindau tumor suppressor protein to Elongin B and C
Mechanisms of angiogenesis
Inhibition of transcription elongation by the VHL tumor suppressor protein
Binding of GSK3Beta to the APC-Beta-Catenin Complex and Regulation of Complex Assembly
Cleavage of structural proteins during the assembly of the head of bacteriophage T4
Molecular cloning of a functional thrombin receptor reveals a novel proteolytic mechanism of receptor activation
Cell Adhesion: The Molecular Basis of Tissue Architecture and Morphogenesis
Negative regulation of hypoxia-inducible genes by the von Hippel-Lindau protein
Endothelial PAS domain protein 1 (EPAS1), a transcription factor selectively expressed in endothelial cells
Failure of blood-island formation and vasculogenesis in Flk-1-deficient mice

Latent Category 15: Bone marrow stem cells

Words:

cells, cell, stem, bone, differentiation, expression, development, tRNA, marrow, hematopoietic, gene, human, culture, normal, embryonic, vitro, mouse, proliferation, growth, epithelial

References:

Manipulating the mouse embryo: A laboratory manual
RAG-2-deficient mice lack mature lymphocytes owing to inability to initiate V(D)J rearrangement
Purification and characterization of mouse hematopoietic stem cells
Resolution and characterization of pro-B and pre-pro-B cell stages in normal mouse bone marrow
Clonal selection and learning in the antibody system
Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction
Promoter traps in embryonic stem cells: A genetic screen to identify and mutate developmental genes in mice
Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs
Long-Term Lymphohematopoietic Reconstitution by a Single CD34-Low/Negative Hematopoietic Stem Cell
Derivation of completely cell culture-derived mice from early-passage embryonic stem cells
The active cochlea
Hematopoietic commitment during embryonic stem cell differentiation in culture
The long-term repopulating subset of hematopoietic stem cells is deterministic and isolatable by phenotype
Isolation of a candidate human hematopoietic stem-cell population
Stem Cells in the Central Nervous System
Mechanical amplification of stimuli by hair cells
RAG-1-deficient mice have no mature B and T lymphocytes
Disruption of overlapping transcripts in the ROSA Beta-geo 26 gene trap strain leads to widespread expression of Beta-ga
Somatic generation of antibody diversity
In Vivo Gene Delivery and Stable Transduction of Nondividing Cells by a Lentiviral Vector

Latent Category 16: Functional and visual responses to changes in the brain

Words:

visual, actin, time, cortex, model, myosin, light, changes, cdata, results, activity, response, cortical, single, functional, different, brain, spatial, temporal, stimulus

References:

Coplanar stereotaxic atlas of the human brain
Three-dimensional structure of myosin subfragment-1: A molecular motor
Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation
Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging
Structure of the actin-myosin complex and its implications for muscle contraction
Single myosin molecule mechanics: piconewton forces and nanometre steps
The regulation of rabbit skeletal muscle contraction. I. Biochemical studies of the interaction of the tropomyosin-troponin complex with actin and the proteolytic fragments of myosin
On/off blinking and switching behaviour of single molecules of green fluorescent protein
The Green Fluorescent Protein
Functional and effective connectivity in neuroimaging: A synthesis
Numerical Recipes in C, the art of scientific computing
Fluorescent indicators for Ca2+ based on green fluorescent proteins and calmodulin
Brain magnetic resonance imaging with contrast dependent on blood oxygenation
The hippocampus as a cognitive map
Imaging of single fluorescent molecules and individual ATP turnovers by single myosin molecules in aqueous solution
Direct observation of kinesin stepping by optical trapping interferometry

Actin-Based Cell Motility and Cell Locomotion
Visual feature integration and the temporal correlation hypothesis
Sorting single molecules: application to diagnostics and evolutionary biotechnology
Atomic model of plant light-harvesting complex by electron crystallography

Latent Category 17: Neurons and neurotransmitters
Words:
alpha, receptor, receptors, neurons, protein, beta, gamma, synaptic, cells, activity, activation, glutamate, hippocampal, dependent, gaba, induced, mediated, acid, neuronal, brain
References:
A synaptic model of memory—long-term potentiation in the hippocampus
GAIP and RGS4 are GTPase-activating proteins for the Gi subfamily of G protein alpha subunits
G proteins: Transducers of receptor-generated signals
Cloned glutamate receptors
Contrasting properties of two forms of long-term potentiation in the hippocampus
Heterotrimeric C proteins: Organizers of transmembrane signals
EGL-10 Regulates G Protein Signaling in the C. elegans Nervous System and Shares a Conserved Domain with Many Mammalian
RGS family members: GTPase-activating proteins for heterotrimeric G-protein alpha-subunits
Modulation of Ca2+ channels G-protein py subunits
The 2.0 Å crystal structure of a heterotrimeric G protein
Short-term synaptic plasticity
Improved patch-clamp techniques for high-resolution current recordings from cells and cell free membrane patches
GAIP, a protein that specifically interacts with the trimeric G protein $G\alpha_{i3}$, is a member of a protein family with a highly conserved core domain
Voltage-dependent modulation of N-type calcium channels by G-protein subunits
G1 phase progression: Cycling on cue
Activation of postsynaptically silent synapses during pairing-induced LTP in CA1 region of hippocampal slice
RGS Proteins and Signaling by Heterotrimeric G Proteins
New functional activities for the p21 family of CDK inhibitors
D-type cyclin-dependent kinase activity in mammalian cells
G protein-coupled receptors. III. New roles for receptor kinases and b-arrestins in receptor signaling and desensitizati

Latent Category 18: Tumor treatment for mice and humans
Words:
tumor, mice, expression, gene, human, cells, cancer, tumors, levels, patients, growth, treatment, normal, increased, vivo, disease, tissue, liver, leptin, cell
References:
Receptor-associated Mad homologs synergize as effectors of the TGF-b response
Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction
Patterns and Emerging Mechanisms of the Angiogenic Switch during Tumorigenesis
Recombinant mouse ob protein: Evidence for a peripheral signal linking adiposity and central neural networks
Effects of the obese gene product on body weight regulation in ob/ob mice
Weight-reducing effects of the plasma protein encoded by the obese gene
Spontaneous hypercholesterolemia and arterial lesions in mice lacking apolipoprotein E
Angiostatin: A novel angiogenesis inhibitor that mediates the suppression of metastases by a lewis lung carcinoma
Severe hypercholesterolemia and atherosclerosis in apolipoprotein E-deficient mice created by homologous recombination i
Endostatin: An Endogenous Inhibitor of Angiogenesis and Tumor Growth
Protein measurement with the Folin-Phenol reagents
Cellular immunity to viral antigens limits E1-deleted adenoviruses for gene therapy
Abnormal splicing of the leptin receptor in diabetic mice
Identification and expression cloning of a leptin receptor
Angiogenesis in cancer, vascular, rheumatoid and other disease
Cleavage of structural proteins during the assembly of the head of bacteriophage T4
Gene delivery to skeletal muscle results in sustained expression and systemic delivery of a therapeutic protein
A rapid and sensitive method for the quantification of microgram quantities of protein utilizing the principle of protei
Evidence That the Diabetes Gene Encodes the Leptin Receptor: Identification of a Mutation in the Leptin Receptor Gene in
Efficient long-term gene transfer into muscle tissue of immunocompetent mice by adeno-associated virus vector

Latent Category 19: Nervous system development
Words:
expression, gene, neurons, development, brain, expressed, cells, system, neuronal, role, cell, mouse, nervous, adult, mice, receptor, protein, neural, genes, central
References:
Targeted gene expression as a means of altering cell fates and generating dominant phenotypes
Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction
The Rat Brain in Stereotaxic Coordinates
Neurotrophins and Neuronal Plasticity
Aggregation of Huntingtin in Neuronal Intranuclear Inclusions and Dystrophic Neurites in Brain
Manipulating the Mouse Embryo: A Laboratory Manual
Formation of Neuronal Intranuclear Inclusions Underlies the Neurological Dysfunction in Mice Transgenic for the HD Mutat
Physiology of the neurotrophins
A novel multigene family may encode odorant receptors: A molecular basis for odor recognition
Manipulating the Mouse Embryo
The Mouse Brain: In Stereotaxic Coordinates

A single protocol to detect transcripts of various types and expression levels in neural tissue and cultured cells: In situ hybridization using digoxigenin-labelled cRNA probes
A non-radioactive in situ hybridization method for the localization of specific RNAs in *Drosophila* embryos reveals translational control of the segmentation gene *hunchback*
A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes
Molecular bases for circadian clocks
A Derivation of Completely Cell Culture-Derived Mice from Early-Passage Embryonic Stem Cells
Disruption of the proto-oncogene int-2 in mouse embryo-derived stem cells: A general strategy for targeting mutations to
Huntingtin-Encoded Polyglutamine Expansions Form Amyloid-like Protein Aggregates In Vitro and In Vivo
The nerve growth factor 35 years later
Functions of the neurotrophins during nervous system development: What the knockouts are teaching us

Latent Category 20: Electrical excitability of cell membranes
Words:
ca2+, channel, channels, alpha, cells, atp, membrane, calcium, k+, voltage, binding, protein, dependent, cell, beta, transport, activity, release, na+, current
References:
Improved patch-clamp techniques for high-resolution current recordings from cells and cell free membrane patches
Ionic Channels of Excitable Membranes
Molecular Cloning: A Laboratory Manual (2nd ed.)
Cleavage of structural proteins during the assembly of the head of bacteriophage
A new generation of Ca2+ indicators with greatly improved fluorescence properties
Inositol trisphosphate and calcium signalling
Amiloride-sensitive epithelial Na+ channel is made of three homologous subunits
Reconstitution of IKATP: An Inward Rectifier Subunit Plus the Sulfonylurea Receptor
The Structure of the Potassium Channel: Molecular Basis of K+ Conduction and Selectivity
Structure and function of voltage-gated ion channels
A family of sulfonylurea receptors determines the pharmacological properties of ATP-sensitive K+ channels
Cloning of the beta cell high-affinity sulfonylurea receptor: a regulator of insulin secretion
Primary structure and functional expression from complementary DNA of a brain calcium channel
Functional expression of a probable Arabidopsis thaliana potassium channel in Saccharomyces cerevisiae
Truncation of Kir6.2 produces ATP-sensitive K+ channels in the absence of the sulphonylurea receptor
Calcium signaling
The synaptic vesicle cycle: A cascade of protein-protein interactions
Identification of the cystic fibrosis gene: Cloning and characterization of complementary DNA
Calcium sparks: Elementary events underlying excitation-contraction coupling in heart muscle
Cloning and expression of an inwardly rectifying ATP-regulated potassium channel