Supporting Information
for
Measuring Discursive Influence Across Scholarship

Aaron Gerow Yuening Hu Jordan Boyd-Graber
David M. Blei James A. Evans

Contents

1 The Regression-based Document Influence Model

The regression-based document influence model (RDIM) developed for this work is a dynamic topic model (DTM) that extends the document influence model (DIM) [6, 16]. DTMs are themselves an extension of the static, latent Dirichlet allocation topic model (LDA; [9]). These models share a statistical pedigree in probabilistic graphical models, which specify dependence relationships among observed and latent variables. All topic models assume a set of latent term-distributions, referred to as "topics", and are responsible for observed term co-occurrences across documents [5]. In RDIM, there are only two sets of observed variables: the occurrence of terms in documents, and a vector of covariates for each document. Latent variables are fit by expectation maximization (EM), which iterates between updating model parameters given the data, and computing the expectation of the data give the desired parameters [23]. We define RDIM in more detail in Section 1.1, explain the approximate inference in Section 1.2, and derive the E- and M-step updates for variables that differ from DIM in Section 1.3.

1.1 Definition

RDIM learns the influence of documents in a time-ordered collection of text. Here, we adopt a technical definition of influence: how much a document changes future discourse. This discourse is represented as a set of topics $\{1, \cdots, K\}$, which constitute probability distributions over terms [9, 13]. Dynamic topic models derive a set of such topics from documents binned into time-steps $\{1, \cdots, T\}$, that drift with Gaussian noise but which are also guided by variation in the preceding documents. This variation consists of words $w_d \in N$, their topic-assignments z_d, and the influence, $\ell_{d,k}$, of previous documents.[1]

While DTMs allow topics to change, they do not provide a measure of how the change takes place. The dynamic influence model [16, DIM] learns a document's influence based on how its linguistic variation is reflected in future topics. This is carried out using a latent vector over topics for influence, ℓ_d, which, along with w_d and z_d, is used to learn topics in the next time-slice, $\beta_{k,t+1}$. Our model, RDIM, is

[1] The subscripts d, t, and k pick out a document, time-stamp and topic respectively. However, d and t are redundant because each document d has a single time-stamp t. We omit t where time is not relevant to the interpretation of a variable. In all cases $\ell_{t,d,k} = \ell_{d,k}$ and $\tau_{d,k} = \tau_{t,d,k}$.

specifically designed to measure exogenous factors that shape discursive influence such as authorship, publication venue and other document metadata. RDIM operationalizes a robust notion of topic-specific influence as the *mixture* of content (as in DIM) and the results of a latent regression on document covariates, which are specific to each topic.

Denote the influence of document d as $\ell_{t,d}$, which we take to be informed by a regression on a vector of S observed document-level covariates, denoted by $\tau_{t,d}$. To incorporate a latent regression on these covariates, we assume influence is drawn from a Gaussian, the mean of which is given by a projection on $\tau_{t,d}$:

$$\ell_{t,d} \sim \mathcal{N}(\mu\tau_{t,d}, \sigma_\ell^2 \mathbb{I}) \tag{1}$$

where μ is a $K \times S$ matrix of topic-specific coefficients. Furthermore, we assume μ is drawn from a Gaussian of mean 0 and specified variance:

$$\mu_k \sim \mathcal{N}(0, \sigma_\mu^2 \mathbb{I}) \tag{2}$$

Figure 1 illustrates the plate diagram for RDIM. In RDIM, the generative process for each time-slice t, is assumed to be

1. Draw topics $\beta_{t+1}|\beta_t, (w, \ell, z)_{t,1:D_t} \sim \mathcal{N}(\beta_t + \exp(-\beta_t) \sum_d \ell_{t,d} \sum_n w_{t,d,n} z_{t,d,n}, \sigma^2 \mathbb{I})$.
2. Draw coefficients $\mu_k \sim \mathcal{N}(0, \sigma_\mu^2 \mathbb{I})$.
3. For each document d at time t:
    - (a) Draw $\theta_{t,d} \sim \text{Dir}(\alpha)$.
    - (b) For each word $w_{t,d,n}$
        - i. Draw $z_{t,d,n} \sim \text{Mult}(\theta_{t,d})$.
        - ii. Draw $w_{t,d,n} \sim \text{Mult}(\pi(\beta_{t,z_{t,d,n}}))$.
    - (c) Draw $\ell_{t,d} \sim \mathcal{N}(\mu\tau_{t,d}, \sigma_\ell^2 \mathbb{I})$

where the function $\pi(x)$, used in the word draws, maps the multinomial parameters to their mean: $\pi(x)_w = \frac{\exp(x_w)}{\sum_w \exp(x_w)}$.

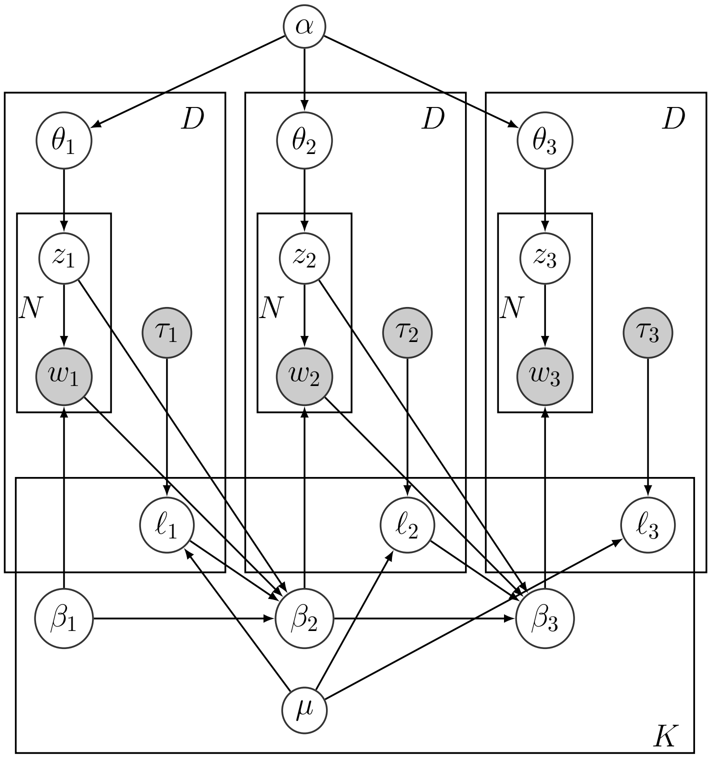

Figure 1: Plate diagram for RDIM, shown with three time-slices. Note that document influence, $\ell_{t,d}$, is a product of the document metadata in $\tau_{t,d}$ and topic-specific coefficients, μ_k. Influence is then learned by observing its effect on topics of subsequent time-steps, $\beta_{\{t+1,\cdots,T\},k}$.

1.2 Approximate Inference

Similar to DTM and DIM, RDIM aligns topics using draws from a Gaussian:

$$\beta_{t+1,k}|\beta_{t,k},(w,\ell,z)_{t,1:D_t} \sim \mathcal{N}(\beta_{t,k} + exp(-\beta_{t,k})\sum_{d=1}^{D_t}\ell_{t,d,k}\sum_{n=1}^{N_{t,d}} w_{t,d,n}z_{t,d,n}, \sigma^2\mathbb{I}) \tag{3}$$

This introduces non-conjugacy between the log-normal prior topic mixtures, β, and the multinomial word observations, w, however, precluding collapsed Gibbs sampling or traditional EM. RDIM uses variational inference based on Kalman filtering [19], which estimates the variational parameters to minimize the KL divergence to the true posterior. In variational inference, the topic alignment parameters are a Gaussian chain, $\{\beta_{1,k}, \cdots, \beta_{T,k}\}$, governed by the variational parameters $\{\hat{\beta}_{1,k}, \cdots, \hat{\beta}_{T,k}\}$ that describe the mean of each distribution.

The addition of τ does not change this distribution, but μ does. RDIM learns μ by variational estimation to the approximate $\hat{\mu}$. This means topics fit with DTM, DIM, and RDIM will be different. Intuitively, this is because future topics have been "influenced" by the covariates apparent in τ. We assume the variational distribution for influence is given by the Gaussian of the mean influence and a fixed variance:

$$\ell_{t,d} \sim \mathcal{N}(\hat{\ell}_{t,d}, \nu_\ell^2 \mathbb{I}) \tag{4}$$

We also assume the variational distribution for μ is drawn from a Gaussian with mean $\hat{\mu}_k$ and fixed variance:

$$\mu_k \sim \mathcal{N}(\hat{\mu}_k, \nu_\mu^2 \mathbb{I}) \tag{5}$$

For the variational distributions of other parameters, we follow the assumptions in DIM [16]:

$$\beta_t \sim \mathcal{N}(\hat{\beta}_{1:t}, \hat{\sigma}^2 \mathbb{I}) \tag{6}$$

$$\theta_{t,d} \sim \text{Dir}(\gamma_{t,d}) \tag{7}$$

$$z_{t,d,n} \sim \text{Mult}(\phi_{t,d,n}) \tag{8}$$

This defines the variational distribution as

$$q(\beta, \ell, \theta, \mathbf{z}, \mu | \hat{\beta}, \hat{\ell}, \gamma, \phi, \hat{\mu}) = \prod_k q(\beta_{1,k}, \cdots, \beta_{T,k} | \hat{\beta}_{1,k}, \cdots, \hat{\beta}_{T,k}) \tag{9}$$

$$\times \prod_k q(\mu_k | \hat{\mu}_k) \times \prod_{t=1}^{T} \Big(\prod_{d=1}^{D_t} q(\theta_{t,d} | \gamma_{t,d}) q(\ell_{t,d} | \hat{\ell}_{t,d}) \prod_{n=1}^{N_{t,d}} q(z_{t,d,n} | \phi_{t,d,n}) \Big)$$

RDIM is fit by variational expectation maximization (EM) described in Algorithm 1, and here we focus on the updates of parameter $\hat{\ell}$ and $\hat{\mu}$. We define $X = \text{Diag}(\exp(-\beta_{t,k}))(\mathbf{w}_t \circ \mathbf{z}_{t,k})$, and $\Delta\beta_{t,k} = \beta_{t+1,k} - \beta_{t,k}$. With S covariates, $\tau_{t,d}$ is an S-length vector and τ_t is an $S \times D$ matrix coding the observed covariates of D documents. μ is a $K \times S$ matrix, containing the coefficients for each each covariate. The lower bound of $\hat{\ell}_{t,k}$ is given by:

$$\mathcal{L}_{\hat{\ell}_{t,k}} = \frac{1}{\sigma^2} \mathbb{E}_q[X^\mathsf{T} \Delta\beta_{t,k}] \hat{\ell}_{t,k} - \frac{1}{2\sigma^2} \mathbb{E}_q[X^\mathsf{T} X] \hat{\ell}_{t,k}^2 - \frac{1}{2\sigma_d^2} (\hat{\ell}_{t,k} - \hat{\mu}_k \tau_t)^2 \tag{10}$$

This provides the E-step update for $\hat{\ell}_{t,k}$:

$$\hat{\ell}_{t,k} \leftarrow (\mathbb{E}_q[X^\mathsf{T} X] + \frac{\sigma^2}{\sigma_d^2} \mathbb{I})^{-1} (\mathbb{E}_q[X^\mathsf{T} \Delta\beta_{t,k}] + \hat{\mu}_k \tau_t) \tag{11}$$

Algorithm 1 EM inference in RDIM

```
for $i$ iterations do
    for Time $t$ do
        for Document $d \in t$ do
            for Token $n \in d$ do
                Update $\phi_{t,d,n,k}$
            end for
            Update $\gamma_{t,d,k}$
        end for
        Update $\hat{\ell}_{t,k}$ by Eq. 11
    end for
    for Time $t$ do
        for Topic $k$ do
            Update $\hat{\beta}_{t,k}$
        end for
    end for
    Update $\hat{\mu}$ by Eq. 13
end for
```

Similarly, the lower-bound of μ is given by

$$\mathcal{L}_{\mu_k} = \sum_t \sum_d \sum_k -\frac{1}{2\sigma_\ell^2}(\hat{\ell}_{t,d,k} - \hat{\mu}_k \tau_{t,d})^2 - \sum_k \frac{\hat{\mu}_k^2}{2\sigma_\mu^2} \tag{12}$$

This yields the M-step update for μ_k:

$$\hat{\mu}_k^\mathsf{T} \leftarrow \left(\sum_t \sum_d \tau_{t,d} \tau_{t,d}^\mathsf{T} + \frac{\sigma_\ell^2}{\sigma_\mu^2} \mathbb{I} \right)^{-1} \sum_t \sum_d \tau_{t,d} \hat{\ell}_{t,d,k} \tag{13}$$

The update for $\hat{\ell}_{t,k}$ by Eq. 11 is similar to DIM and relatively straightforward. An extra term $\hat{\mu}_k \tau_t$ is added to the second term of Eq. 11 to model the influence from the covariates. However, updating $\hat{\mu}$ by Eq. 13 can be difficult because, although it is entirely observed, the first term is a large, dense and potentially degenerate $S \times S$ matrix. Numerical and memory considerations rule out direct inversion. Instead, we use a high-performance solver that accommodates general matrices and provides bounds on any numerical error in the solutions.[2] The model's complexity, primarily due to the K-updates for $\hat{\mu}$ and $\hat{\ell}$, is overcome by reducing the number of ϕ-updates in early iterations (prior to which the topic-chains remain unaligned)

[2] PARDISO; `www.pardiso-project.org`

and with local multi-threading, pipelining and a distributed memory solver. Our implementation of RDIM is available online.[3]

1.3 Derivation

This section derives the E- and M-step updates for influence ℓ (as it differs from DIM) and the coefficients, μ, for the document-level covariates $\tau_{t,d}$.

Following the model assumptions in Sections 1.1 and 1.2, we obtain the variational distribution (Eq. 9). Now we can write out the evidence lower bound $\mathcal{L}(\hat{\beta}, \hat{\ell}, \gamma, \phi, \hat{\mu}; \alpha)$ and factorize it as

$$
\begin{aligned}
&\mathcal{L}(\hat{\beta},\hat{\ell}, \gamma, \phi, \hat{\mu}; \alpha) \\
&\quad = \mathbb{E}_q[\log p(\mathbf{w}, \mathbf{z}, \theta, \ell, \mu, \beta; \alpha)] - \mathbb{E}_q[\log q(\beta, \ell, \mu, \theta, \mathbf{z}|\hat{\beta}, \hat{\ell}, \hat{\mu}, \gamma, \phi)] && (14)\\
&\quad = \sum_k \sum_t \mathbb{E}_q[\log p(\beta_{t+1,k}|\beta_{t,k}, (w, \ell, z)_{t,1:D_t})] && (15)\\
&\qquad + \sum_t \sum_d \mathbb{E}_q[\log p(\theta_{t,d}|\alpha_t)] && (16)\\
&\qquad + \sum_t \sum_d \mathbb{E}_q[\log p(\ell_{t,d}|\mu)] && (17)\\
&\qquad + \sum_t \sum_d \sum_n \mathbb{E}_q[\log p(z_{t,d,n}|\theta_{t,d})] && (18)\\
&\qquad + \sum_t \sum_d \sum_n \mathbb{E}_q[\log p(w_{t,d,n}|\beta_{t,z_{t,d,n}})] && (19)\\
&\qquad - \sum_t \sum_k \mathbb{E}_q[\log(\beta_{k,t}|\hat{\beta}_{k,1:T})] && (20)\\
&\qquad - \sum_t \sum_d \mathbb{E}_q[\log q(\theta_{t,d}|\gamma_{t,d})] && (21)\\
&\qquad - \sum_t \sum_d \mathbb{E}_q[\log q(\ell_{t,d}|\hat{\ell}_{t,d})] && (22)\\
&\qquad - \sum_t \sum_d \sum_n \mathbb{E}_q[\log q(z_{t,d,n}|\phi_{t,d,n})] && (23)\\
&\qquad + \sum_k \mathbb{E}_q[\log p(\mu_k)] && (24)\\
&\qquad - \sum_k \mathbb{E}_q[\log q(\mu_k|\hat{\mu}_k)] && (25)
\end{aligned}
$$

[3] https://github.com/gerowam/influence.git

Updating Influence ℓ Three terms (Eqs. 15, 17 and 22) are related to the influence parameter ℓ. Eq. 15 and can be expanded as

$$\begin{aligned}
&\sum_t \sum_k \mathbb{E}_q[\log p(\beta_{t+1,k}|\beta_{t,k},(w,\ell,z)_{t,1:D_t})] \qquad (26)\\
&= -\frac{VTK}{2}(\log 2\pi + \log\sigma^2)\\
&\quad - \frac{1}{2\sigma^2}\sum_t\sum_k\sum_v \mathbb{E}_q[(\beta_{t+1,k,v} - \beta_{t,k,v} - \exp(-\beta_{t,k,v})(\mathbf{w}_{t,v}\circ \mathbf{z}_{t,v,k})\ell_{t,k})^2]\\
&= -\frac{VTK}{2}(\log 2\pi + \log\sigma^2)\\
&\quad - \frac{1}{2\sigma^2}\sum_t\sum_k\sum_v \mathbb{E}_q[(\beta_{t+1,k,v} - \beta_{t,k,v})^2]\\
&\quad + \frac{1}{\sigma^2}\sum_t\sum_k\sum_v \mathbb{E}_q[(\beta_{t+1,k,v} - \beta_{t,k,v})\exp(-\beta_{t,k,v})(\mathbf{w}_{t,v}\circ \mathbf{z}_{t,v,k})\ell_{t,k}]\\
&\quad - \frac{1}{2\sigma^2}\sum_t\sum_k\sum_v \mathbb{E}_q[\exp(-2\beta_{t,k,v})((\mathbf{w}_{t,v}\circ \mathbf{z}_{t,v,k})\ell_{t,k})^2]
\end{aligned}$$

where $\mathbf{w}_{t,v}$ is a D_t-dimensional vector with each entry as the frequency of word type v in document $d.D_t$ is the number of documents at time t; $\mathbf{z}_{t,v,k}$ is a D_t-dimensional vector, where each entry is 1 if the topic is k otherwise 0; $\ell_{t,k}$ is also a D_t-dimensional vector, where each entry is the influence of document d with time-stamp t on topic k. V is the vocabulary size; T is the number of time intervals; and K is the number of topics.

Similarly, we know $\tau_{t,d}$ is an S-dimensional vector, and we use τ_t to denote an $S \times D_t$ matrix; $\hat{\mu}$ is a $K \times S$ matrix and we use $\hat{\mu}_k$ to denote the S-dimensional vector for topic k. Eq. 17 can be expanded as

$$\begin{aligned}
&\sum_t \sum_d \mathbb{E}_q[\log p(\ell_{t,d}|\mu)] \qquad (27)\\
&\qquad = \sum_t\sum_d\sum_k \mathbb{E}_q[-\frac{1}{2\sigma_\ell^2}(\ell_{t,d,k} - \mu_k\tau_{t,d})^2 - \frac{1}{2}(\log 2\pi + \log\sigma_\ell^2)]\\
&\qquad = \sum_t\sum_d\sum_k -\frac{1}{2\sigma_\ell^2}((\hat{\ell}_{t,d,k} - \hat{\mu}_k\tau_{t,d})^2 + \nu_\ell^2) - \frac{1}{2}(\log 2\pi + \log\sigma_\ell^2)\\
&\qquad = \sum_t\sum_k -\frac{1}{2\sigma_\ell^2}((\hat{\ell}_{t,k} - \hat{\mu}_k\tau_t)^2 + \nu_\ell^2) - \frac{1}{2}(\log 2\pi + \log\sigma_\ell^2)
\end{aligned}$$

Eq. 22 can be expanded as

$$
\begin{aligned}
\mathbb{E}_q[\log q(\ell_{t,d,k}|\hat{\ell}_{t,d,k})] &= \mathbb{E}_q[-\frac{1}{2}\log 2\pi - \log \sigma_l - \frac{(\ell_{t,d,k} - \hat{\ell}_{t,d,k})^2}{2\sigma_l^2}] \\
&= -\frac{1}{2}\log 2\pi - \log \sigma_l - \frac{\nu_\ell^2}{2\sigma_l^2}
\end{aligned}
\tag{28}
$$

Combining the three expanded terms, we obtain the evidence lower bound for influence $\hat{\ell}_{t,k}$:

$$
\begin{aligned}
\mathcal{L}_{\hat{\ell}_{t,k}} =& \frac{1}{\sigma^2}\mathbb{E}_q[(\beta_{t+1,k} - \beta_{t,k}) \exp(-\beta_{t,k})(\mathbf{w}_t \circ \mathbf{z}_{t,k})\ell_{t,k}] \\
& - \frac{1}{2\sigma^2}\mathbb{E}_q[\exp(-2\beta_{t,k})((\mathbf{w}_t \circ \mathbf{z}_{t,k})\ell_{t,k})^2] - \frac{1}{2\sigma_\ell^2}(\hat{\ell}_{t,k} - \hat{\mu}_k\tau_t)^2
\end{aligned}
\tag{29}
$$

We define $X = \mathrm{Diag}(\exp(-\beta_{t,k}))(\mathbf{w}_t \circ \mathbf{z}_{t,k})$, and $\Delta\beta_{t,k} = \beta_{t+1,k} - \beta_{t,k}$. As a result, we have

$$
\mathcal{L}_{\hat{\ell}_{t,k}} = \frac{1}{\sigma^2}\mathbb{E}_q[X^\top \Delta\beta_{t,k}]\ell_{t,k} - \frac{1}{2\sigma^2}\mathbb{E}_q[X^\top X]\hat{\ell}_{t,k}^2 - \frac{1}{2\sigma_\ell^2}(\hat{\ell}_{t,k} - \hat{\mu}_k\tau_t)^2 \tag{30}
$$

We take the derivative of the evidence lower bound with respect to $\hat{\ell}_{t,k}$, and generate

$$
\frac{\partial \mathcal{L}_{\hat{\ell}_{t,k}}}{\partial \hat{\ell}_{t,k}} = \frac{1}{\sigma^2}\mathbb{E}_q[X^\top \Delta\beta_{t,k}] - \frac{1}{\sigma^2}\mathbb{E}_q[X^\top X]\hat{\ell}_{t,k} - \frac{1}{\sigma_\ell^2}\hat{\ell}_{t,k} + \frac{1}{\sigma_\ell^2}\hat{\mu}_k\tau_t \tag{31}
$$

Setting the derivative to zero, we obtain

$$
\frac{1}{\sigma^2}\mathbb{E}_q[X^\top \Delta\beta_{t,k}] - \frac{1}{\sigma^2}\mathbb{E}_q[X^\top X]\hat{\ell}_{t,k} - \frac{1}{\sigma_\ell^2}\hat{\ell}_{t,k} + \frac{1}{\sigma_\ell^2}\hat{\mu}_k\tau_t = 0 \tag{32}
$$

$$
\rightarrow \quad (\mathbb{E}_q[X^\top X] + \frac{\sigma^2}{\sigma_\ell^2}\mathbb{I})\hat{\ell}_{t,k} = \mathbb{E}_q[X^\top \Delta\beta_{t,k}] + \hat{\mu}_k\tau_t \tag{33}
$$

This provides the E-step update for influence:

$$
\hat{\ell}_{t,k} \leftarrow (\mathbb{E}_q[X^\top X] + \frac{\sigma^2}{\sigma_\ell^2}\mathbb{I})^{-1}(\mathbb{E}_q[X^\top \Delta\beta_{t,k}] + \hat{\mu}_k\tau_t) \tag{34}
$$

where $\mathbb{E}_q[X^\top \Delta\beta_{t,k}]$ and $\mathbb{E}_q[X^\top X]$ can be computed the same as in DIM [16]. $\mathbb{E}_q[X^\top \Delta\beta_{t,k}]$ is a D_t dimensional vector with each element as:

$$
\sum_n (\mathbf{w}_{t,d,n}\phi_{t,d,n,k})(\hat{m}_{t+1,k,n} - \hat{m}_{t,k,n} + \hat{\Sigma}_{t,k,n}/2)\exp(-\hat{m}_{t,k,n} + \hat{\Sigma}_{t,k,n}/2) \tag{35}
$$

where $\hat{m}$ and $\hat{\Sigma}$ are the mean and variance of the variational posterior for $\hat{\beta}$. These two parameters are estimated using standard Kalman filter calculations (see [6]). $\mathbb{E}_q[X^\mathsf{T} X]$ is a $D_t \times D_t$ matrix with each element as

$$\sum_n \exp(-2\hat{m}_{t,k,n} + 2\hat{\Sigma}_{t,k,n})(\mathbf{w}_{t,d,n}\mathbf{w}_{t,d',n}\phi_{t,d,n,k}\phi_{t,d',n,k}) \tag{36}$$

when $d \neq d'$. For $d = d'$, the element is

$$\mathbb{E}_q[X^\mathsf{T} X]_{d,d} = \sum_n \exp(-2\hat{m}_{t,k,n} + 2\hat{\Sigma}_{t,k,n})(\mathbf{w}^2_{t,d,n}\phi_{t,d,n,k}) \tag{37}$$

Updating Coefficients μ Three terms (Eqs. 17, 24 and 25) in the model are related to the evidence lower bound for μ, and these three terms can be expanded as follows, respectively:

Eq. 17:

$$\begin{aligned}
&\sum_t \sum_d \mathbb{E}_q[\log p(\ell_{t,d})] \\
&\quad = \sum_t \sum_d \sum_k \mathbb{E}_q[-\frac{1}{2\sigma_\ell^2}(\ell_{t,d,k} - \mu_k \tau_{t,d})^2 - \frac{1}{2}(\log 2\pi + \log \sigma_\ell^2)] \\
&\quad = \sum_t \sum_d \sum_k \Big(- \frac{1}{2\sigma_\ell^2}(\hat{\ell}_{t,d,k} - \hat{\mu}_k \tau_{t,d})^2 + \nu_\ell^2 - \frac{1}{2}(\log 2\pi + \log \sigma_\ell^2)\Big) \qquad (38)
\end{aligned}$$

Eq. 24:

$$\begin{aligned}
\sum_k \mathbb{E}_q[\log p(\mu_k)] &= \sum_k \sum_s \mathbb{E}_q[-\frac{1}{2\sigma_\mu^2}\mu_{k,s}^2 - \frac{1}{2}(\log 2\pi + \log \sigma_\mu^2)] \\
&= \sum_k \sum_s \Big(- \frac{1}{2\sigma_\mu^2}(\hat{\mu}_{k,s}^2 + \nu_\mu^2) - \frac{1}{2}(\log 2\pi + \log \sigma_\mu^2)\Big) \qquad (39)
\end{aligned}$$

Eq. 25:

$$\begin{aligned}
-\sum_k \mathbb{E}_q[\log q(\mu_k|\hat{\mu}_k)] &= -\sum_k \sum_s \mathbb{E}_q[\frac{1}{2}\log 2\pi - \log \sigma_\mu - \frac{(\mu_{k,s} - \hat{\mu}_{k,s})^2}{2\sigma_\mu^2}] \\
&= \sum_k \sum_s \Big(- \frac{1}{2}\log 2\pi - \log \sigma_\mu - \frac{\nu_\mu^2}{2\sigma_\mu^2}\Big) \qquad (40)
\end{aligned}$$

which, together, give us the lower-bound of μ_k:

$$\begin{aligned}\mathcal{L}_{\mu_k} &= \sum_t \sum_d \sum_k -\frac{1}{2\sigma_\ell^2}(\hat{\ell}_{t,d,k} - \hat{\mu}_k \tau_{t,d})^2 - \sum_k \sum_s \frac{\hat{\mu}_{k,s}^2}{2\sigma_\mu^2} \\ &= \sum_t \sum_d \sum_k -\frac{1}{2\sigma_\ell^2}(\hat{\ell}_{t,d,k} - \hat{\mu}_k \tau_{t,d})^2 - \sum_k \frac{\hat{\mu}_k^2}{2\sigma_\mu^2} \qquad (41)\end{aligned}$$

Taking the derivative with respect to μ_k, we have

$$\begin{aligned}\frac{\partial \mathcal{L}_{\mu_k}}{\partial \mu_k} &= \sum_t \sum_d -\frac{1}{\sigma_\ell^2}(\hat{\ell}_{t,d,k} - \hat{\mu}_k \tau_{t,d})(-\tau_{t,d}) - \frac{\hat{\mu}_k}{\sigma_\mu^2} \qquad (42) \\ &= \sum_t \sum_d \frac{1}{\sigma_\ell^2} \tau_{t,d}(\hat{\ell}_{t,d,k} - \hat{\mu}_k \tau_{t,d}) - \frac{\hat{\mu}_k}{\sigma_\mu^2} \qquad (43)\end{aligned}$$

which, when set to zero yields the update for $\hat{\mu}_k$:

$$\hat{\mu}_k^{\mathsf{T}} \leftarrow \left(\sum_t \sum_d \tau_{t,d} \tau_{t,d}^{\mathsf{T}} + \frac{\sigma_\ell^2}{\sigma_\mu^2} \mathbb{I} \right)^{-1} \sum_t \sum_d \tau_{t,d} \hat{\ell}_{t,d,k} \qquad (44)$$

1.4 Modeling Framework: Assumptions & Limitations

Like all topic models, RDIM is generative in that it treats observations as the outcome of an underlying process. This process is defined with a graph of dependence relationships between observed and latent variables (Figure 1). The generative assumption allows for the inference process, but it also imposes constraints on the data. The primary assumption in topic models is that documents, which consist of tokens, tend to be produced by topics, responsible for the constituent term co-occurrence patterns. The intuition is straight-forward: documents tend to be *about* something, which dictates what words are used. A second assumption, specific to DTMs, is that a time-ordered "diachronic" collection has a consistent set of topics throughout. This is no small assumption, and some work has sought to relieve it [2, 12]. As we will see, however, in growing collections of scientific research (ACL-ARC, APS and JSTOR datasets) the emergence of topics can indeed be seen over the course of the data. This is typically observed as early years which have vague or "background" topics, which individuate over time, becoming increasingly coherent and specific. In some ways, the constant-K assumption is helpful because it forces the model to establish the genesis of a topic. This assumption further highlights the importance of choosing an optimal K for a given dataset. We address this by measuring topic usage in static models fit with many topics and sparse hyperpriors, which approximates a Bayesian nonparametric search over possible values of K. A third assumption, made by DIM and RDIM in particular, is that documents contribute change to future topics. In RDIM this influence is adjusted by extrinsic, document-level features coded in the metadata, τ_d. We implement this as a projection of the covariates on a linear space, enabling estimation of their *marginal* effects. A final assumption made by nearly all statistical analyses is the ability for a sample to represent a population. Some of our claims are generalizable to research as a whole, and where they are not, it has more to do with limitations of data than the modeling framework. Our two primary datasets, the APS and JSTOR corpora, offer some of the largest digitized collections of published research, but they are by no means exhaustive. Neither give adequate coverage to orally presented research. JSTOR contains books and monograms, which were more popular earlier in the collection. However the APS dataset contains only peer-reviewed journal publications. These aspects of the data mediate our conclusions, but do not violate assumptions of the modeling framework: documents still exhibit systematic variation in term co-occurrence, the force of influence over time and a significant role of covariates. As such, attributing discursive influence to documents in different temporal regions of the corpus does not change their interpretation.

2 Evaluations

2.1 Quantitative Evaluation

Convergence Results of topic models can be difficult to qualify [13]. Here, we assess the stability of RDIM and compare it to its simpler parent model, DIM, using the Association for Computational Linguistics Anthology Reference Corpus[4] (ACL-ARC) [4]. The corpus contained 10,331 full-text articles, 43M tokens (51,142 unique) over 24 years. The top 2,000 strongest bigram collocations, extracted using normalized PMI [30] were also included, and only terms occurring in five or more documents, with a mean TF*IDF of 1.0E-5 were kept. This produced a vocabulary of 23,104 unique tokens. Publication venue, authorship, length (in pages) and number of authors were coded in τ.

In our analysis of model stability, we fit 10- 20- and 50-topic models to random folds of the ACL-ARC. The folding procedure was performed over documents in each time-slice D_t, which induces folds across N and τ. Following [6] and [16], the model was initialized with a static LDA model fit to the entire corpus. We set σ_μ=0.0001 and all other parameters to published defaults: σ_d=0.0001, σ_ℓ=0.0001, α=0.01, and the chain variance for $\{\beta_{k,1}, \cdots, \beta_{k,T}\}$, $\hat{\sigma}^2 = 0.005$. The convergence criterion was met when the model lower bound changed less than 0.01%

[4] acl-arc.comp.nus.edu.sg

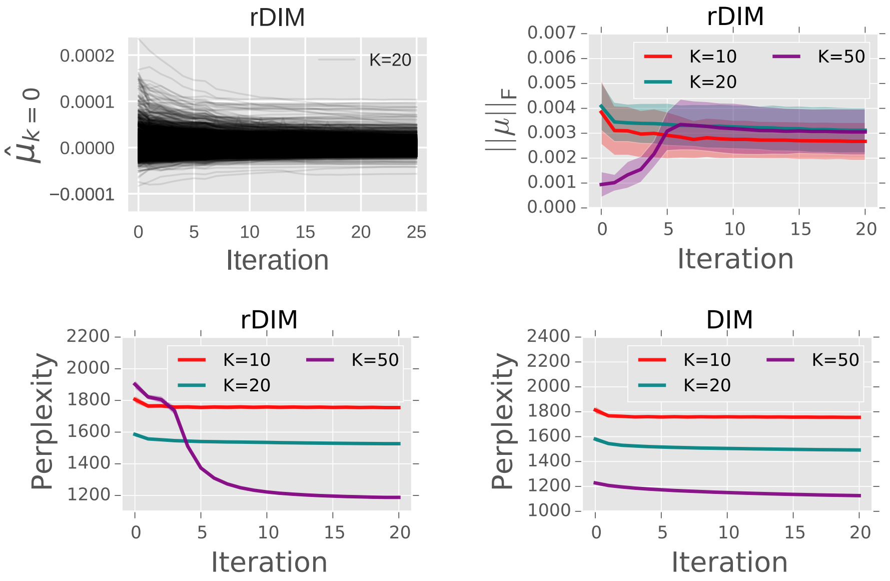

Figure 2: A sample of coefficients in $\hat{\mu}$ during training (a; top left), the convergence of $\|\mu\|_{\mathsf{F}}$ (b; top right), held-out perplexity on 10, 70 / 30 folds of the ACL-ARC using RDIM (c; bottom left) and DIM (d; bottom right). Error-bands are ±1 s.d. of the mean.

from the previous iteration.

We compared our model to DIM, which does not estimate how document covariates contribute to influence. The held-out log-likelihood and perplexity were computed for 10 random folds. We also assessed $\|\hat{\mu}\|_{\mathsf{F}}$, the stabilization of which signals convergence in the estimated coefficients. Our model exhibits similar convergence behavior to its less complex counter-part (Figure 2c-d), where the 50-topic models performed best. After 25 iterations, the final values for RDIM were all within 1 s.d. of the mean compared to DIM. Looking at the coefficient matrix, which is unique to RDIM, we demonstrate that $\|\hat{\mu}\|_{\mathsf{F}}$ tends to converge before perplexity. Convergence in the norm of $\hat{\mu}$, however, does not necessarily imply all coefficients converged. Although no coefficients exhibited super-linear trends, a small number were numerically unstable (observed as oscillation): about 2 of 8,804 in each topic on average.

***K* and Influence** Applications of topic models and DTMs have been described, verified and critiqued elsewhere in the literature [11]. However, our work relies centrally on the interpretation of the latent influence variable, $\hat{\ell}$. Posterior estimates for influence are affected by model specification and the choice for K. Table 1 shows the correlations of influence for models fit to the full ACL-ARC data with varying number of topics. The greatest discrepancy is seen comparing a $K = 1$ model to a $K = 100$ model, where influence scores correlated at $r = 0.26$. Although all model pairs are positively correlated and models closer in specification are considerably more correlated, there continues to be significant variation. This underscores the importance of choosing the number of topics in a principled, data-driven fashion, as discussed in the following section.

K	**1**	**2**	**5**	**10**	**20**	**50**	**75**	**100**
1	1							
2	.84	1						
5	.72	.80	1					
10	.56	.69	.80	1				
20	.51	.63	.74	.76	1			
50	.32	.42	.56	.60	.61	1		
75	.30	.39	.51	.56	.56	.69	1	
100	.26	.38	.47	.48	.51	.54	.58	1

Table 1: Pearson correlation between document influence (Eq. 8; main text) for models with different numbers of topics. In all cases $p < 0.01$ and all differences are significant ($p < 0.01$; Fisher's z-transformation).

Influence and Missing Data Influence scores are also stable with respect to missing data and random starting conditions. On ten folds of the ACL-ARC data, randomly removing 20% of documents, the average correlation for influence (over all pairs of folds) was $r = .84$ ($K = 10$), $r = .83$ ($K = 20$), $r = .83$ ($K = 50$) and $r = .83$ (K=75) (all $p < 0.01$). On the same data, random initialization produced approximately 4% variation in influence scores.

2.2 Influence in Computational Linguistics Research

Here, we justify the complexity of our model and demonstrate its usefulness, surveying results from a 20-topic model fit to the ACL-ARC dataset described above. We use the ACL-ARC to exemplify the forms of analysis enabled by RDIM, and to compare simple uses of our model to its predecessors. The ACL-ARC data has been analyzed in a number of places (see [1] for a collection of analyses), often with models fitted with more than 20 topics. For example, [3] fit 100 topics, labeled 73, grouped those into three-year temporal bins which were then clustered before analyzing the publishing trajectories of academics. We are primarily interested in demonstrating the viability of influence as we conceive it and showing that contextual features, uniquely captured by our model, are both sensible and useful. Choosing the number of topics for our primary datasets, APS and JSTOR, is discussed in Section 3.

Both DIM and RDIM produce topics and influence scores. The composition of these topics is itself guided, in part, by influence. In DIM, because the conception of influence is a simple scalar value, whereas in RDIM it is regressed on document covariates, the topics themselves will be different. This relationship is grounded in the generative assumptions: covariates affect how a document is created and perceived, and the derived topics should reflect this. Influence, then, serves as an explanatory trace of a document's lasting impact. Expanding this explanation to include the marginal effects of covariates enables RDIM not only to better explain lasting influence, but also to provide better topics.

We examined estimated values in $\hat{\mu}$ to explore how document covariates effect influence in the ACL-ARC. Recall that $\hat{\mu}_{k,s}$ is the *marginal* effect of parameter s on $\hat{\ell}_{d,k}$ (where s is a feature of d coded in τ). In the ACL-ARC, authors tend to be the largest contributors of extrinsic influence. Tables 2a-b show two topics, Grammar & Parsing and Information Retrieval, and the covariates with the largest marginal effect. In Parsing & Grammar, the top coefficients were for two linguists, Gerald Gazdar and Robert Berwick, who developed theories of natural language syntax influenced by computability constraints. Note that coefficients are topic-specific: neither Gazdar nor Berwick were in the top coefficients for other topics. In Information Retrieval, the strongest covariates include Donna Harman, who co-authored authoritative datasets for TIPSTER conferences, and Gerard Salton, the namesake of the "Salton Vector Space Model".[5] Despite comprising the top coefficients in all topics, the average author detracts from influence. Although the median author coefficient was below zero in all topics, the distributions were still skewed to the negative ($\gamma_1(\hat{\mu}_k) > 0$, $p < 0.01$ for all k; Pearson's moment coefficient). That is, there are more positive outliers than negative, suggesting that there is more room to be an outstanding author, with a positive reputation boosting influence for your opus, than a terrible author with a negative reputation that detracts attention from it.

[5] Interestingly, Salton did not write an often-cited 1975 paper *A Vector Space Model for Information Retrieval* [15]. Nonetheless, he is widely credited with pioneering research in vector space models that emerged in the 1970s and were used in various information retrieval tasks.

Parsing & Grammar	**Information Retrieval**
tree	document
grammar	term
node	query
structure	text
feature	topic
language	retrieval
figure	collection
form	result
constituent	information
set	sentence

Parsing & Grammar		**Information Retrieval**	
Parameter	**Value in $\hat{\mu}_k$**	**Parameter**	**Value in $\hat{\mu}_k$**
Gerald Gazdar	0.00013	Donna Harman	0.00017
Robert C. Berwick	0.00012	G. Vladutz	0.00014
Monique Rolbert	0.00012	Gerard Salton	0.00013
J. N. Verastegui-Carvajal	0.00011	Jade Goldstein	0.00012
C. Raymond Perrault	0.00001	Chris Buckley	0.00011
Jan Landsbergen	0.00009	David D. Lewis	0.00011
James Kilbury	0.00008	Tomek Strzalkowski	0.00010
Luis Damas	0.00008	K. L. Kwok	0.00009
R. C. Bainbridge	0.00008	John Broglio	0.00009
C. S. Mellish	0.00008	Andy Lauriston	0.00007

Table 2: Top words from the final time-step of two topics in the ACL-ARC (a; top) and estimated covariate coefficients in $\hat{\mu}_k$ with the largest positive effect on influence in these topics (b; bottom). In all 20 topics, authors comprised the ten strongest effects on influence.

The venue in which a paper is published affects its reception. Nevertheless, such an effect may not be uniform: certain venues may promote papers' influence in particular topics while inhibiting influence in others. Figure 3 shows the coefficients for publication venues in the ACL-ARC model. Topic #4, Information Retrieval, shows a disproportionately positive effect for papers published in the (now defunct) TIPSTER conferences. TIPSTER was a competition-based event sponsored by the U.S. NIST, alternatively referred to as the Text REtrieval Conference (TREC). The TIPSTER / TREC conferences focused on document retrieval in different settings (libraries, email, web, newswire, etc.), it is logical that documents about information retrieval would receive a positive boost for being published here, as opposed to elsewhere. Another positive outlier is seen in topic 16, Entity Extraction, for papers published in Message Understanding conferences,

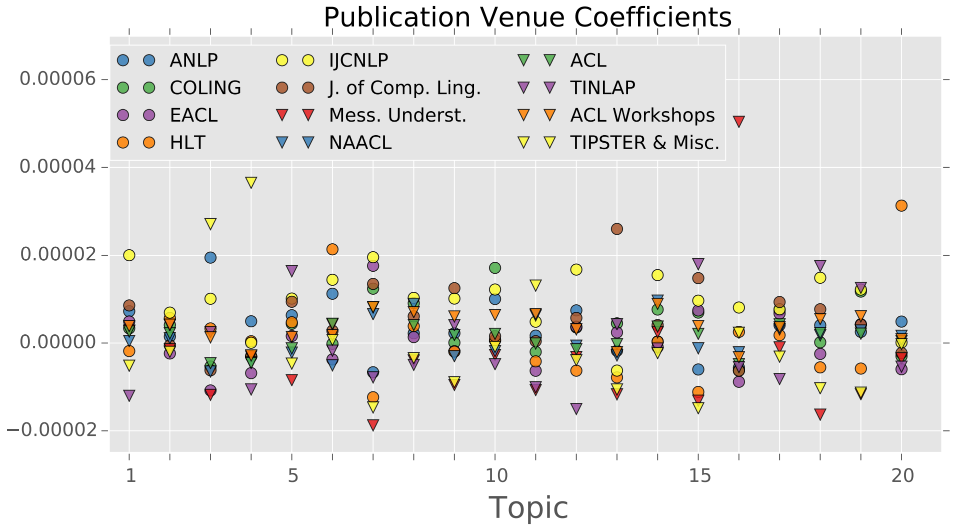

Figure 3: Estimated coefficients in $\hat{\mu}_k$ for publication venue in the ACL-ARC data. In most topics, values cluster near zero. Like authors, there are more positive outliers than negative.

sponsored by the U.S. DARPA from 1987 to 1997. This conference focused on extracting structured information from formal communications such as military reports, terrorist communications and newswires. Publishing in these conferences would increase the influence of papers on named entity, event extraction and other entity recognition tasks. A third outlier is the effect of publishing in Human Language Technology (HLT) conferences for topic 20, Speech Processing. Although speech is important to many aspects of computational linguistics, HLT focuses primarily on enabling language-based interaction between people and computers, where speech is a central component.

2.2.1 Comparison to DIM

Our model assigned the highest posterior influence to the 1982 paper "From English to Logic: Context-Free Computation of 'Conventional' Logical Translation", which was published in the American Journal of Computational Linguistics. This paper, by Lenhart Schubert and Francis Pelletier, describes a class of context free grammars that account for grammatical and logical structure in English. The paper introduced concepts that are now main-stays of the wider field, such as induction, learning, inference, parsing and translation. Examining the paper's metadata, we found that Pelletier contributed a substantial boost to its influence, though it would still have been highly influential (by content alone) without his by-line. The second most influential publication, "The Text REtrieval Conferences (TRECs)", which summarizes the TREC conference datasets, received a sizable boost from its venue and authors. Compare this to the highest score given by DIM to the 1986 "Q&A: Already a Success?" a miniature precis on automated question answering by Gary Hendrix.[6] This paper offered the ACL community important vocabulary on an emerging topic (mostly notably the bigram *question answering*). The second most influential paper from the DIM model was a 1986 opinion piece by Harry Tennant on the "The Commercial Application of Natural Language Interfaces", which introduced a menu-based linguistic interface developed by Texas Instruments. The influence of both these papers comes from having introduced lexical variation that later became common in particular topics. The primary difference in how RDIM and DIM attribute this influence lies in the combination of internal / textual and external / contextual contributions: whereas the top papers from DIM contribute significant lexical variation, the CFGs and TREC papers given by RDIM brought both formative content *and* had external boosts from eminent authors and well-suited publication venues.

Explicitly modeling document metadata makes for a more complex model, but it also provides a more accurate account of citations. In DIM, the maximum posterior influence scores were found to correlate with citation counts [16]. Document Influence in RDIM (I_d; Eq. 8; main text) is the product of both content and metadata, and should be more strongly correlated to citations than in DIM. In a 10-topic model on the ACL-ARC data, our I_d correlated with citations at $\rho = 0.28$ and DIM produced $\rho = 0.22$. For $K = 20$, the RDIM's scores correlated at $\rho = 0.27$ and 0.21 for DIM. At $K = 50$, our model correlated with citations at 0.23 and DIM again at 0.21. Finally, at $K = 75$ our model correlated at $\rho = 0.22$ and DIM at $\rho = 0.18$ (all $p < 0.01$). This confirms that the contextual features captured by RDIM have a significant, observable effect on predicting variation in citations.

3 Data & Model Specification

3.1 APS Collection

The APS collection contained 509,007 abstracts dating from 1913 to 2015. Documents were coded with their type (comment, essay, rapid, erratum, brief, miscellaneous, letter, article, reply or Nobel), venue (*Physical Review, Physical Review A, Physical Review B, Physical Review C, Physical Review D, Physical Review E, Reviews of Modern Physics, Physical Review Special Topics, Physical Review Let-*

[6] The paper, about a system called Q&A, is humorously presented as a Q&A.

ters, or *Physical Review X*), authorship and author affiliation. Categorical covariates were coded as indicators and centered to mean 0 and unit variance. Affiliations can be a sensitive subject: the institutional resolution at which authors choose to affiliate may be motivated by political, financial, or by expected perception. Authors can choose to affiliate with their lab, department, school, division, institute, university, etc.. We did not collapse affiliations into canonical institutions of any level. Instead, to maintain a principled sample and avoid spurious metadata (redundancies, misspellings, ambiguities), papers were removed if they did not have an author and affiliation that occurred twice in the corpus. This filtering process helps avoid certain issues with the data, particularly when analyzing author and affiliation effects. The process resulted in a similar distribution across publication venues in the sample compared to the full collection. The resulting set contained 251,382 documents, dating from 1918 to 2015 with 74,459 covariates coded in τ. In those documents, open-class words[7] were kept if they occurred in five or more documents. We also included the 5,000 strongest bigram collocations (also occurring in five or more documents) [30]. Unigrams and bigrams with a mean TF*IDF less than or equal to 1.0E-5 were removed. The final vocabulary consisted of 15,312 tokens (13,833 unigrams and 1,479 bigrams). This vocabulary size was data-driven, filtered by frequency, collocation strength and TF*IDF scores, but it was also affirmed by three professional researchers with doctoral degrees in physics. We fit a model with $K = 37$ topics; other parameters were the same as the ACL-ARC models. The model was run on a tightly-coupled computing cluster and converged after 26 iterations. Across all topics, we identified 8,103 (0.3%) coefficients were oscillating beyond a tolerance of ±0.00001 per iteration. These coefficients were discarded without further analysis. The derived dataset is available by contacting the authors.

We chose to fit the APS data with 37 topics. This specification was selected by fitting a static model which is less intensive than RDIM. A 500-topic LDA [9] was fit with sparse hyper-parameters ($\alpha_k = 1/K$ and $\beta_w = 10/N$).[8] This configuration estimates the Bayesian non-parametric solution [25, 26] and induces sparsity in words' topic assignments, $z_{n,k}$, yielding a measure of topic use. This usage is observed by a low proportion of documents having more than a certain number of tokens. Specifically, for each of the 500 topics, we calculated the proportion of documents that had 10 or more assigned tokens. To select a threshold above which a topic is considered sufficiently "used", we minimize the density estimate of the topic usage distribution for scores above the mean (Figure 4). For APS, the

[7] Open-class refers to words that have a part-of-speech class to which new words can be added. For example, a new noun can be added to a lexicon because nouns are an open class. The lexicon tends not to accommodate new function words like determiners, articles, pronouns, etc..

[8] N being the number of tokens in the vocabulary.

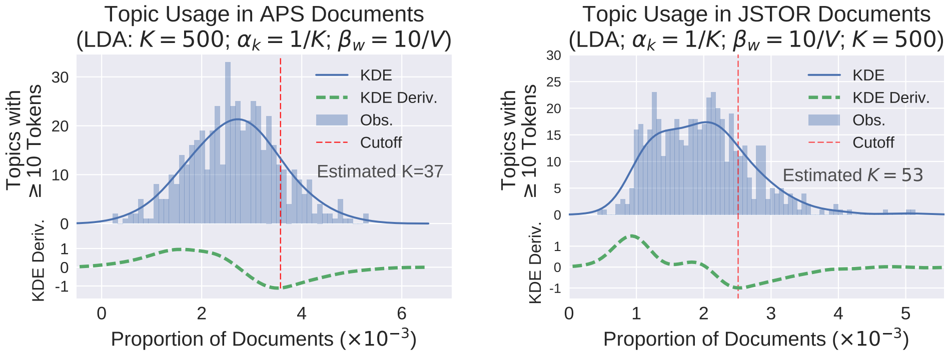

Figure 4: Topic usage distribution for APS (left) and JSTOR (right) static models. The topic usage distribution in static 500-topic models of each dataset, was used to calculated a threshold above which topics were sufficiently "used". This was done by minimizing the derivative (green dashed line) of the KDE (blue line) above the mean. The thresholds are denoted with a vertical dashed red line.

threshold was the 463rd most used topic, yielding an estimate of $K = 37$ topics. We also computed a number of post-hoc topic diagnostics[9] on static models fit to the data with different numbers of topics. The results of these—available by request from the authors—showed that $K = 37$ for the APS data constituted a compromise between coherent, consistent and individuated topics.

Using $K = 37$ for the APS also provides a cognitively manageable set of topics, offering space for canonical physics subjects. 37 topics were few enough that the topics could be labeled by physicists with experience in a variety of sub-disciplines. Some research has used models with many topics to provide detailed semantic analyses [17, 18], assist machine translation [10, 22], or to examine variation across collections [27, 36]. Our goal was not to uncover new topics in physics, nor to provide detailed explanations of sub-fields, but to offer a measure of scholarly influence that is sensitive to the overarching organization of physics. In our model, there were two topics that did not directly relate to sub-disciplines. The first was a generalist topic, Academic Reporting, with prominent words like *find*, *show* and *calculate*. Such a topic is not uncommon in models of academic text [7, 31, 32]. Nonetheless, we exclude this topic from consideration when computing document influence and other metrics. The second generalist topic was more specific to physics, labeled Experimentation, with prominent words such as *experiment*, *construct* and *setup*. This topic was retained because, though not a sub-discipline, per se, experimentation is a unique aspect to physics research.

[9] See mallet.cs.umass.edu/diagnostics.php.

APS topics were mainly related to sub-disciplines of physics, many of which emerged relatively late in the dataset. This data spans nearly 100 years, during which new concepts emerged in the literature, some of which go on to be important in various topics. Each topic chain can be thought of as tracing the origin of the topic's eventual composition—much like how evolution lacks a destination, but has trajectory all the same. Figure 5 illustrates the number of words assigned to each topic, $|z_k|$ over time. This is effectively the prevalence of topics in each year of the data. We omit the omnipresent, generalist topic Academic Reporting (topic #36). Note how most topics emerge between 1950 and 1980. We also assessed year-to-year divergence for each topic and found it correlated with growth. Because such yearly divergence is a macroscopic, distributional effect, which means that attributing *periods of change* to specific documents would be an over-reach. Instead, the topic contribution metric described in the main text (Eq. 10) provides an egocentric view of what future variation is attributable to a given document.

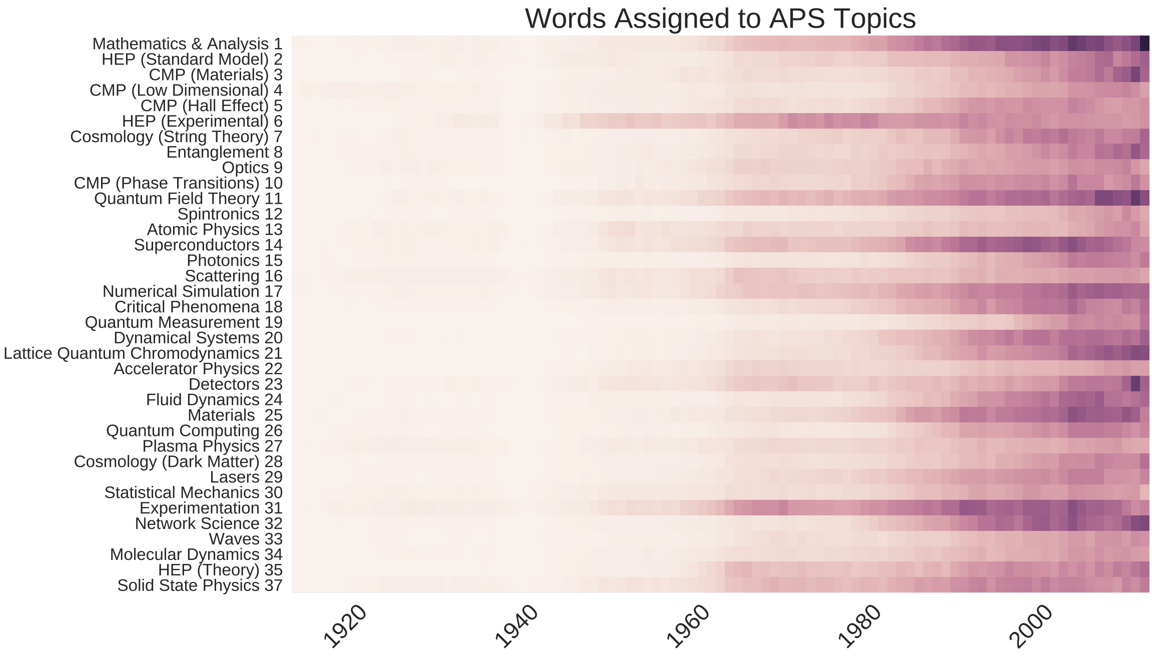

Figure 5: Heatmap of the number of words assigned to each topic over the course of the APS collection. Darker cells indicate more words.

3.2 JSTOR Collection

The JSTOR collection initially contained 2,205,970 full-text documents in 2,549 journals across 66 different disciplines from 1894 to 2014; more recent research is under-represented in JSTOR. Disciplines in JSTOR are attributed by curators, specified for each journal and organized into nine "domains" (Area Studies, Arts, Business & Economics, History, Humanities, Law, Medicine & Applied Health, Science & Mathematics, and Social Science). Documents were excluded if they did not contain at least one author who wrote at least three documents, or they were classified in a subject that had a gap exceeding 20 years (e.g. Railroad Science). Of the remaining documents, we were able to use a random, 50% sample. This document-wise sampling was performed per-year, not over the whole collection, resulting in a year-document distribution similar to the full collection. The vocabulary was extracted similarly to the APS collection except that only unigram tokens were included, non-English words were excluded and fewer tokens were discarded by lowering the TF*IDF threshold. The resulting vocabulary consisted of 20,155 tokens. The final dataset included 428,034 full text articles with metadata for 28,861 variables coded in τ, representing authorship, publication venue, publisher and discipline.

The sampled collection had a similar composition to the full collection in terms of domain (Figure 6). The biggest relative discrepancy was in Medicine and Allied Health, which comprises 5.9% in the full collection but only 1.2% in the sample. This is because many disciplines within this domain are recent additions to JSTOR and have not yet been contiguously indexed for 20 or more years. Pruning terms in JSTOR, as with APS, was carried out in a data-driven manner to reduce semantically uninteresting words by frequency and TF*IDF. While our approach is typical in topic modeling and other NLP tasks, we expanded the vocabulary to over 20k—near the advised upper limit for topic models [13, 8]—because JSTOR is such a diverse corpus. Term pruning, can result in removing documents if they contain no words after the filtering, but this only occurred 11 times in JSTOR.

A 53-topic model was fit with parameters similar to the APS data. The choice of 53 topics was made using the same process as APS (See Figure 4). Nearly all of the topics exhibited a clear subject, and we used Google Scholar searches of high loading terms to confirm that the fields of the journals in which they were common were stable. Three of the resulting topics were indiscernible, generalist topics, each labeled Academic Verbiage. These topics were relatively static throughout the corpus, but were not excluded because influence scores were so low (owing to topic stasis) that they contributed little to calculations of document influence and other model-based metrics. The JSTOR data was also submitted to the same diagnostic regime as APS: static models were fit using different numbers of topics.

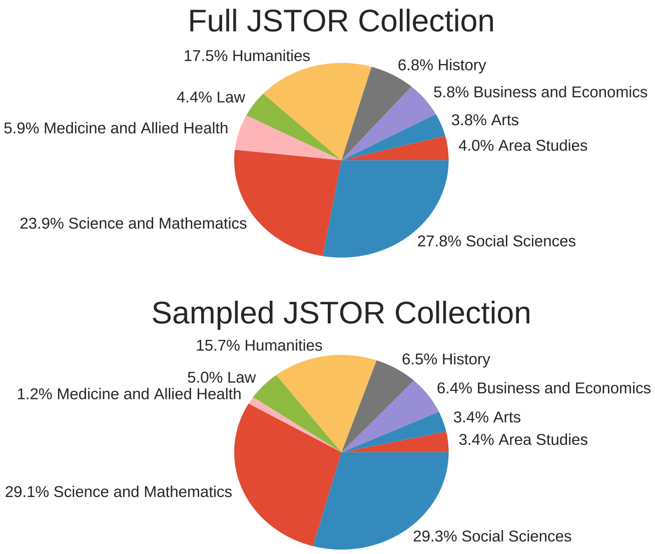

Figure 6: Composition of the full JSTOR collection (top) and the sample used in our dataset (bottom).

The diagnostics, as well as the three generalist topics, suggest that 53 is close to the upper limit while remaining labelable and individuated. Results of these diagnostics are available upon request.

Topic emergence in JSTOR is perhaps more interesting than in physics. JSTOR represents a wider range of academic research and while some areas were permanent throughout, others emerged over the course of the data. The number of words assigned to each topic for each year of the JSTOR collection is shown in Figure 7. Many topics are well-represented throughout, such as Literary Theory and, Plant Biology. Dips in the word-topic assignments (for all topics) correlate with the two world wars. While some topics exhibit a permanence, some do not. One example, explored more in Section 4.2, is the Environmental Science topic, which swells in the 1960s.

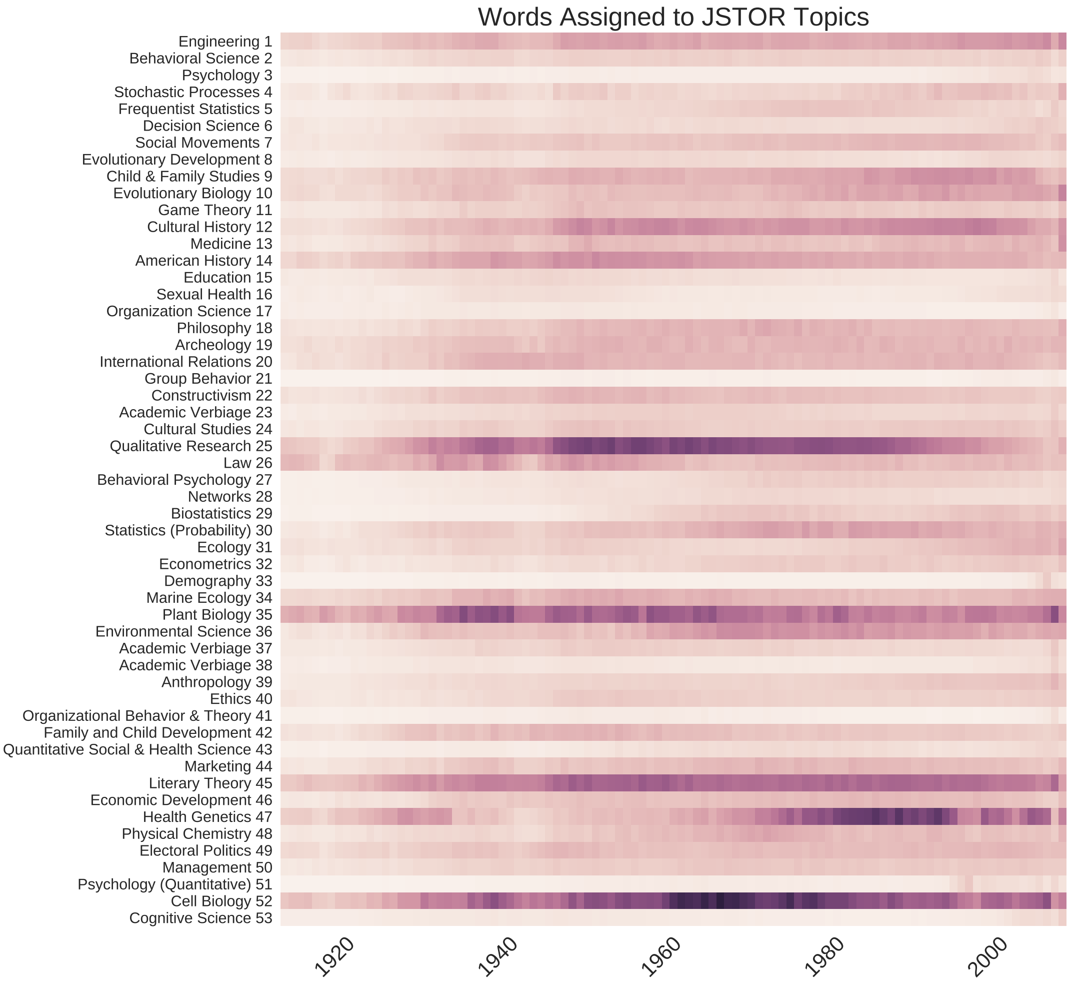

Figure 7: Heatmap of the number of words assigned to each topic over the course of the JSTOR collection. Darker cells indicate more words.

4 Supplemental Results

4.1 APS Topics

Like DTMs, the results of these models provide insight into topic and term dynamics over time. A good example in the APS data is the HEP (Theory) topic, which, like other topics, emerged from a relatively general composition in the early years, into a more discernible, specific topic. This emergence can be tracked by looking at various terms' likelihood in yearly topics (Figure 8a). Top words in HEP (Theory) over time are shown in Table 3. Indeed, *standard model*, a concept that has dominated HEP theory in recent years, did not approach the top of the distribution until the late 80s. With regards to topics, individual terms can be used as probes. For example, the concept of a *quark* emerged primarily in HEP (Theory) and later in HEP (Standard Model) and the experimental side HEP, Accelerator Physics. Together, these three topics dominate assignments of *quark* to all others (Figure 8b). These term and topic dynamics are a result of DTMs, but it is important to remember that topic composition is guided by influence, itself a combination of metadata, words, and topic assignments. Therefore, topics represent a model-aware source of influence, uniquely captured by RDIM.

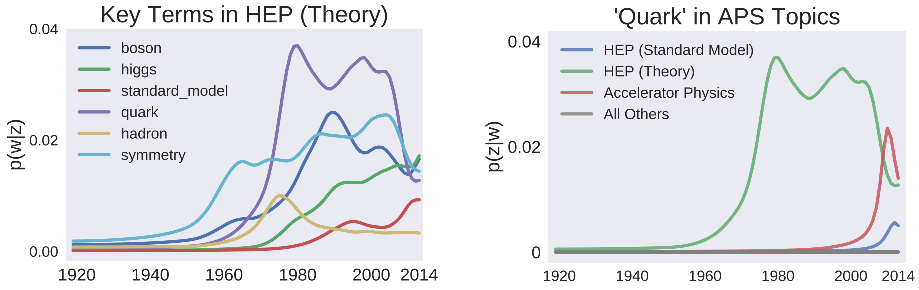

Figure 8: Likelihood of six key terms in the HEP (Theory) topic over the course of the APS corpus (a; left). Likelihood of topic-assignment given the term *quark* throughout the APS data (b; right). Here, the term *quark* is likely to be drawn from one of only three topics.

As a process, science is often characterized by bifurcation and consolidation [21]. Given the expanding volume and breadth of science, it is difficult to exemplify consolidation processes aside from certain specific examples.[10] Bifurcation, on the other hand, is not only common, but observable at larger scales.

[10] See `plato.stanford.edu/entries/scientific-reduction` for a philosophical discussion, particularly on *supervenience* in Section 4.5.3.

1920	1930	1940	1950	1960	1970	1975	1980	1985
mass	mass	mass	meson	mass	mass	model	mass	mass
doublet	rule	meson	mass	meson	model	mass	model	model
rule	coupling	coupling	coupling	coupling	current	coupling	quark	quark
coupling	doublet	rule	nucleon	nucleon	coupling	rule	coupling	coupling
sum	meson	heavy	interaction	interaction	rule	current	meson	symmetry
heavy	heavy	constant	rule	constant	meson	quark	symmetry	boson
neutral	sum	interaction	constant	coupling_constant	sum	meson	rule	meson
meson	neutral	nucleon	coupling_constant	rule	trajectory	sum	weak	rule
constant	constant	sum	sum	vector	sum_rule	symmetry	current	constant
exception	interaction	neutral	neutral	sum	regge	weak	sum	coupling_constant
interaction	exception	doublet	scalar	relation	symmetry	sum_rule	boson	breaking
pole	diagram	exception	vector	symmetry	relation	constant	decay	sum
diagram	nucleon	intermediate	heavy	model	pole	coupling_constant	constant	baryon
color	pole	coupling_constant	relation	weak	amplitude	decay	coupling_constant	chiral
intermediate	intermediate	vector	weak	decay	constant	vector	gauge	decay
nucleon	sum_rule	sum_rule	strong	strong	coupling_constant	interaction	sum_rule	gauge
sum_rule	color	particle	intermediate	scalar	vector	relation	interaction	sum_rule
particle	relation	relation	discussed	pole	decay	hadron	hadron	current
relation	particle	scalar	assumption	current	baryon	regge	breaking	nucleon
vector	vector	pole	one	also	interaction	also	also	weak

1990	1995	2000	2005	2010	2011	2012	2013	2014
mass	mass	mass	mass	mass	mass	mass	mass	mass
model	model	model	quark	gauge	gauge	gauge	gauge	gauge
coupling	quark	quark	symmetry	symmetry	symmetry	standard	higgs	higgs
quark	coupling	coupling	coupling	quark	higgs	symmetry	boson	boson
boson	boson	symmetry	model	coupling	coupling	higgs	standard	symmetry
symmetry	symmetry	boson	chiral	higgs	quark	boson	symmetry	standard
meson	meson	chiral	boson	chiral	standard	coupling	coupling	coupling
rule	chiral	breaking	breaking	boson	boson	quark	quark	quark
breaking	higgs	meson	higgs	breaking	chiral	chiral	chiral	chiral
higgs	heavy	heavy	gauge	standard	breaking	breaking	breaking	breaking
coupling_constant	breaking	higgs	meson	value	value	value	value	value
constant	rule	gauge	heavy	scale	scale	scale	scale	scale
chiral	gauge	rule	scale	model	model	standard_model	standard_model	standard_model
sum	sum	baryon	baryon	meson	standard_model	model	model	model
gauge	constant	find	rule	theory	decay	decay	decay	decay
baryon	coupling_constant	scale	value	decay	theory	flavor	flavor	lattice
find	correction	value	sum	flavor	flavor	theory	lattice	flavor
sum_rule	find	sum	light	heavy	meson	heavy	higgs_boson	higgs_boson
scalar	baryon	parameter	find	standard_model	heavy	lattice	theory	heavy
parameter	parameter	constant	symmetry_breaking	baryon	scalar	meson	heavy	effective

Table 3: Top words in HEP (Theory) throughout the APS collection.

Even though the number of topics is held constant over time in our models, some topics exhibit specialization as they shift from their initial general state. The Quantum Computing topic exemplifies this process (Table 4). Early in the corpus, its broad focus on semiconductors and electricity is clear, but beginning in the 1960s, terms like *tunneling*, *transport* and *barrier* make their way to top. By the turn of the century, concepts like edging effects, and the role of surfaces and interfaces emerge. In the final years, *quantum*, *dot* and even *device* are among the most likely words. *Device* is particularly important because it signals that research is beginning to look at *engineering* in addition to explaining natural phenomena. Comparing Quantum Computing to other quantum-related topics such as Superconductors or Lattice Quantum Chromodynamics, it begins to individuate in the 1960s. Both of these other topics began in similar ways to Quantum Computing, but diverged midway through the dataset.

1920	1930	1940	1950	1960	1970	1980	1990	1995	2000
current	current	surface	current	surface	surface	layer	interface	tunneling	tunneling
potential	surface	current	surface	current	metal	current	layer	interface	interface
voltage	potential	potential	charge	metal	current	interface	tunneling	layer	layer
surface	voltage	voltage	potential	charge	tunneling	tunneling	barrier	barrier	current
charge	metal	layer	voltage	electron	electron	barrier	electron	electron	barrier
metal	electrode	metal	metal	voltage	layer	surface	current	conductance	electron
electrode	charge	charge	layer	layer	junction	electron	conductance	current	conductance
wire	layer	wire	electron	potential	barrier	junction	voltage	step	transport
layer	wire	electrode	wire	barrier	voltage	metal	step	voltage	step
electron	electron	electron	region	region	effect	step	transport	transport	microscopy
contact	contact	function	contact	contact	bulk	transport	junction	wire	wire
cobalt	function	contact	function	effect	charge	effect	structure	microscopy	scanning
function	cobalt	difference	electrode	normal	contact	voltage	effect	edge	voltage
difference	difference	cobalt	barrier	junction	normal	contact	image	scanning	junction
connected	height	height	height	wire	region	height	resistance	junction	edge
height	observed	observed	normal	tunneling	interface	region	contact	structure	image
vacuum	collector	region	effect	density	density	observed	observed	image	contact
collector	normal	barrier	carrier	carrier	observed	normal	oscillation	effect	scanning_tunneling
applied	vacuum	normal	cobalt	electrode	potential	tunnel	wire	contact	bias
observed	applied	carrier	observed	height	height	thickness	thickness	oscillation	effect

2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
current	current	current	current	current	current	current	current	current	current
tunneling	tunneling	tunneling	tunneling	tunneling	tunneling	tunneling	tunneling	tunneling	tunneling
interface	transport	transport	transport	transport	transport	transport	transport	transport	transport
transport	interface	interface	interface	conductance	conductance	interface	interface	interface	interface
conductance	conductance	conductance	conductance	interface	interface	conductance	conductance	conductance	conductance
electron	electron	electron	voltage	voltage	voltage	voltage	scanning	layer	layer
barrier	junction	junction	junction	junction	junction	scanning	dot	quantum	quantum
junction	voltage	voltage	electron	electron	electron	junction	device	scanning	scanning
voltage	barrier	wire	wire	wire	scanning	electron	layer	electron	electron
wire	wire	barrier	barrier	barrier	wire	wire	electron	wire	edge
contact	bias	bias	bias	bias	barrier	device	wire	dot	wire
bias	contact	contact	scanning	scanning	bias	barrier	voltage	effect	effect
microscopy	scanning	scanning	contact	contact	contact	contact	junction	voltage	contact
scanning	microscopy	microscopy	microscopy	device	device	dot	contact	edge	voltage
layer	layer	device	device	microscopy	microscopy	bias	barrier	junction	junction
effect	effect	layer	layer	layer	layer	layer	bias	contact	device
scanning_tunneling	device	effect	effect	effect	effect	microscopy	quantum	bias	dot
device	charge	charge	charge	charge	charge	effect	edge	charge	charge
charge	scanning_tunneling	scanning_tunneling	scanning_tunneling	scanning_tunneling	scanning_tunneling	charge	microscopy	barrier	barrier
image	image	image	image	edge	dot	quantum	charge	device	bias

Table 4: Top words in Quantum Computing throughout the APS collection.

4.2 JSTOR Topics

Our JSTOR collection contains a wide range of academic writing. Among the 53 topics were three that consisted of terms relating primarily to academic writing. Throughout the collection, these Academic Verbiage topics were typified by terms like *example*, *show*, *question* and *work*. Topics of this kind are common in text collections from a given domain [13]. The lexical first-mover effect discussed in the main text is observable in these topics, despite their stability over time. Figure 9 plots the yearly average influence in each topic and the mean over all topics. Note the period of generally high influence in the first ten years. The first documents in our sample appear to shape the future more than others because they are simply the first ones to use *most* terms. This “burn-in” period is unavoidable without access to the entirety of written work. As such, the first twelve years of JSTOR (1913 to 1930) were discarded from document-specific analyses.

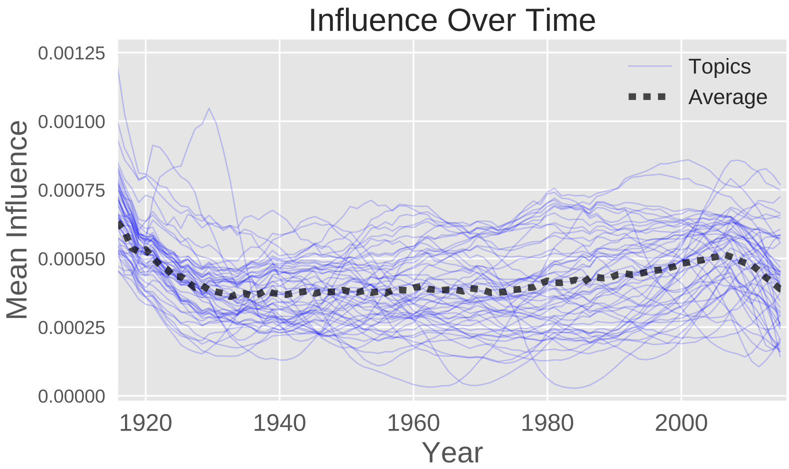

Figure 9: Mean estimated influence over documents in each year of the JSTOR corpus: $\frac{1}{|D_t|}\sum_t \hat{\ell}_{d,t}$.

Table 5 shows the top words for the Environmental Science topic in JSTOR. This topic is typical of topics that emerged part way through the corpus. In the early years, the topic is concerned primarily with natural resources, specifically water, oil and gas. In the 1960s, the terms *temperature* and *treatment* rise to the top. This is the period during which the topic became considerably more prominent in documents (Figure 7). While some topics were used consistently throughout the corpus, many are like Environmental Science in that their rise in usage is matched by coincidental rises in specific terms—often marking a conceptual shift, in this case a focus on climate change.

1920	**1930**	**1940**	**1950**	**1960**	**1965**	**1970**	**1975**
used	water	water	plant	plant	water	plant	plant
water	used	used	water	water	plant	water	water
use	supply	plant	used	used	temperature	temperature	temperature
supply	foot	use	per	temperature	used	used	rate
air	use	air	air	condition	soil	surface	leaf
work	inch	supply	use	per	surface	rate	soil
gas	plant	hour	condition	weight	condition	soil	used
made	pipe	flow	supply	treatment	weight	condition	weight
pressure	flow	material	material	content	day	weight	surface
inch	capacity	made	made	surface	rate	table	condition
plant	gallon	gas	amount	made	treatment	day	value
pipe	tank	work	flow	day	content	treatment	treatment
well	made	time	time	amount	table	leaf	table
iron	thompson	condition	high	air	air	content	content
capacity	work	tank	content	soil	leaf	air	high
main	air	capacity	temperature	period	high	value	dry
flow	filter	pressure	period	high	low	high	low
foot	gas	operation	treatment	leaf	study	growth	level
condition	pressure	per	method	material	growth	dry	day
per	condition	amount	ing	rate	seed	low	growth

1980	**1985**	**1990**	**1995**	**2000**	**2005**	**2010**	**2014**
water	water	rate	rate	rate	rate	rate	rate
plant	rate	water	water	fig	per	per	per
temperature	temperature	temperature	table	per	fig	increase	increase
rate	plant	treatment	temperature	water	increase	ratio	low
soil	treatment	table	per	increase	low	low	use
treatment	soil	soil	fig	low	ratio	fig	ratio
energy	weight	low	treatment	table	density	mass	mass
content	low	high	low	temperature	value	value	value
level	table	weight	increase	density	higher	higher	higher
weight	high	area	high	value	condition	condition	fig
high	content	increased	increased	ratio	measured	density	condition
low	level	per	value	higher	increased	measured	measured
used	dry	content	soil	increased	water	use	nutrient
value	plot	level	higher	measured	table	increased	potential
condition	area	increase	measured	condition	mass	nutrient	density
table	used	dry	content	high	nutrient	system	increased
surface	condition	value	condition	treatment	maximum	maximum	system
dry	value	condition	density	depth	decrease	data	loss
range	production	higher	area	composition	depth	loss	data
study	range	used	ratio	lower	system	respectively	using

Table 5: Top words, p$(w|z)$, in Environmental Science throughout the JSTOR collection. Like many topics in JSTOR, it exhibits a conceptual transition part way through the corpus: here from natural resources to a more specific focus on the environment.

4.3 Relative Effect of Covariates

By explicitly modeling covariates' effect on influence, RDIM provides insight into the *composition* of influence. The proportion of intrinsic influence (influence without the covariate effects; $\hat{\ell}_{d,k} - \sum_{s \in \tau_d} \hat{\mu}_{k,s}$) and the extrinsic influence (total effect of covariates; $\sum_{s \in \tau_d} \hat{\mu}_{k,s}$) can be used to assess the proportion of influence attributable to metadata. Figure 10 show the proportional effect of covariates on influence for APS and JSTOR. Positive values indicate covariates' net effect *added* influence to documents in a topic, whereas negative values mean covariates decreased influence. In both collections there is a range of proportional effects across topics, though all means were within $\pm 15\%$ of the total influence.

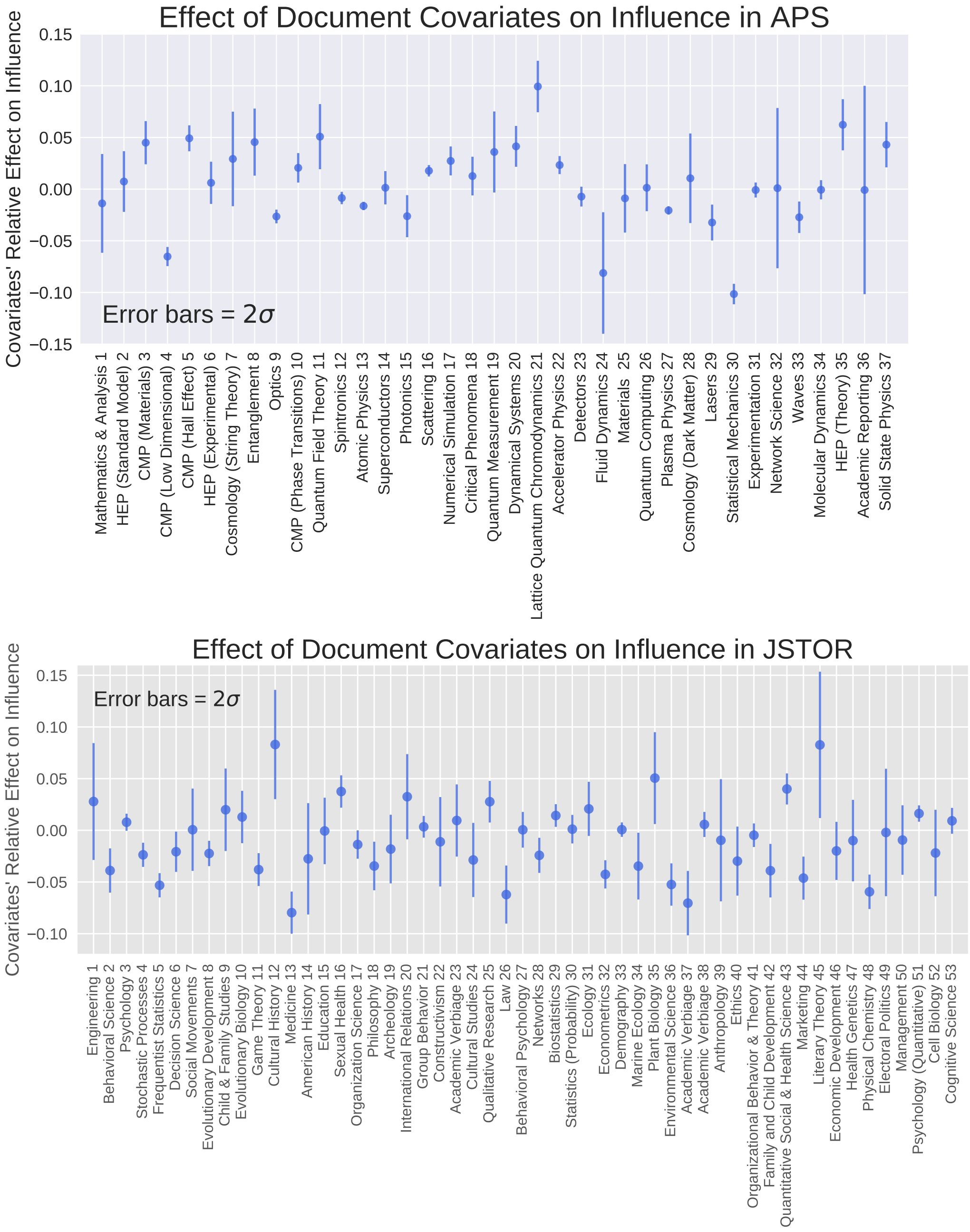

Figure 10: Proportion of influence from covariates in each topic of the APS corpus (top) and the JSTOR corpus (bottom).

4.4 Author Affiliations

The APS collection contained 26,948 distinct institutions. Of these, the institution with the greatest marginal effect on influence was the *The Physical Laboratory* at Cornell University, which was in the top percentile of six topics. The institution with the second highest marginal effect over all topics was *Bell Telephone Labs*, which dates back to 1925, when it pioneered work on information theory and computation, publishing much of Claude Shannon's work, inventing the transistor as well as developing UNIX and the C programming language. The third and fourth strongest institutional name-sakes were Nankai University—a top-ranked Chinese university in Tianjin—and the *Institut fur Theoretische Physik* at the University of Vienna, once chaired by Erwin Schrödinger.

4.5 Authors

47,482 authors were coded in the APS data. The most influential authors, measured as their sum influence over topics, $\sum_k \hat{\mu}_{k,s=author}$ are all authors of early papers that introduce key terms which later came to define topics. The authors include William Duane—a Harvard physicist, student of Max Planck and colleague of Marie Curie—who was awarded the Comstock Prize in Physics (1923) for the production of radon in the lab. Duane tops other authors in influence, in part, because he used the term *absorption* four times in the abstract of a 1919 paper, a term later prominent in Optics and other topics. The next most influential name-sake was William Swann—who used the term *vector potential* in 1920. Other notable authors among those with exceptionally influential documents are Edwin Hall of the Hall Effect and Raymond Birge who lead Berkeley's Physics Department and hired Robert Oppenheimer.

4.6 Publication Venue

There were 11 venues coded in the APS data, 8 of which have been published for 20 or more years. To explore the effect of publication venue on document influence, venue coefficients were examined. There is considerable clustering in most topics, suggesting that often, there is no particular journal that boosts influence. However, there were certain topical trends consistent with the publishing mandate of APS journals (Figure 11a). For example, Physical Review D, which publishes work in HEP, cosmology, field theory, and particle physics, stands above other outlets in topics related to high energy and cosmological physics. This implies that "correct placement" of a paper in Physical Review D provides the document with more influence than had it been published elsewhere. Across all topics, the highest extrinsic boost to influence comes in the Academic Reporting topic—which is a broad topic regarding scholarly reporting. The venues that provided the most additional influence here were Physical Review (the original APS journal) and Reviews of Modern Physics. Neither venue has a subject-specific mandate and both contain research relevant to the greater physics community.

Physical Review E, which has the broadest remit in terms of subject matter, gives a sizable boost to Network Science, Fluid Dynamics and Dynamical Systems. Papers here range from complex systems, to chemical and biological physics, materials science, plasma and information theory. Papers in P.R.E. tend to be more influential than those in other subject-mandated journals (P.R. A-D; Figure 11b). This may be due to the journal's breadth: whereas venues like Reviews of Modern Physics provide a sizable extrinsic boost to many topics, and specific journals like P.R.D do well in topics related to their mandate, papers in P.R.E have more *available* areas to influence, which increases their sum influence across all topics.

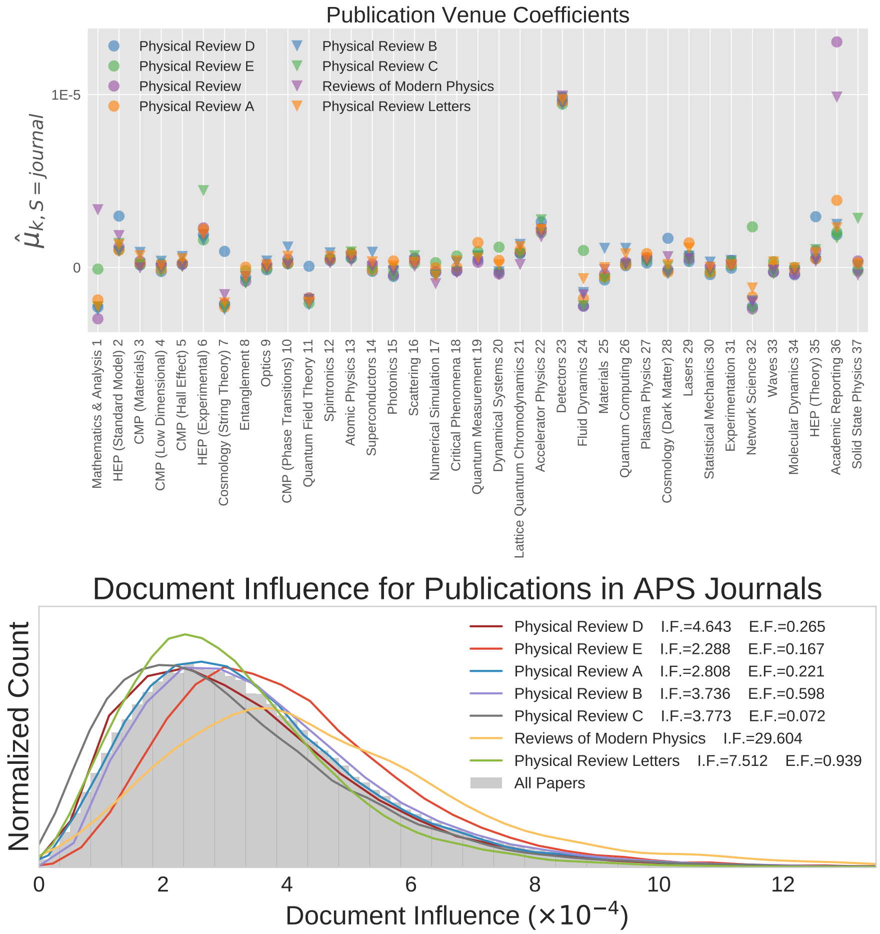

Figure 11: Coefficients for APS publication venues for each topic (a; top) and the distribution of influence for documents published in each journal (b; bottom). Also shown in (b) are journals' Impact Factor (I.F.) and Eigenfactor (E.F.), two measures of journal quality. Only journals published 20 or more years are shown.

4.7 Citation Patterns & Influence

Citations are very much the currency of scientific impact as they act like a proxy for what communities find important. Citation patterns have been analyzed in a number of places [24, 33, 34], but typical analyses treat them as observed outcomes. The increased visibility of citation metrics (for institutions, journals, authors and individual papers) has led to a bias for citing highly cited papers [35]. Document influence provides a content-driven measure of *lasting* impact compared to the *perceived* impact of citation counts. Recall that citations were not coded in τ for the APS or JSTOR models, allowing comparison to document influence. An important question is whether citations exhibit similar behavior as influence. Figure 12 illustrates a three-dimensional plot of a paper's document influence vs. the mean influence of the documents the paper cites and that of the documents that cite it. First, note the positive skew for cited papers' influence. This implies that, overall, people tend to cite papers of higher-than-average influence. Second, papers with high influence tend to land in quadrant A, papers that cite are cited influential papers. This confirms the guiding intuition behind our analysis: that document influence has a discernible effect on how people write (changes in topics) as well as what they choose to cite.

Figure 13 depicts a similar picture using citation counts. Points represent the citation count of a paper as of 2015 (internal to our sample), and the x- and y-axes represent mean citation counts for papers cited by and citing each document. The high-citing-high dynamic is observed in citation counts as it was with document influence, but to a greater extent. Generally, there is some skew toward citing higher cited papers, but this effect is quite pronounced for highly-cited papers themselves: a highly cited paper is very likely to be in quadrant A where it cites and is cited by other highly cited papers. The high-citing / highly-cited skew is likely due to the positive feedback whereby highly cited papers are that much more likely to be cited. Unlike influence, which is a discursive measure of impact, *citations are more self-reinforcing*.

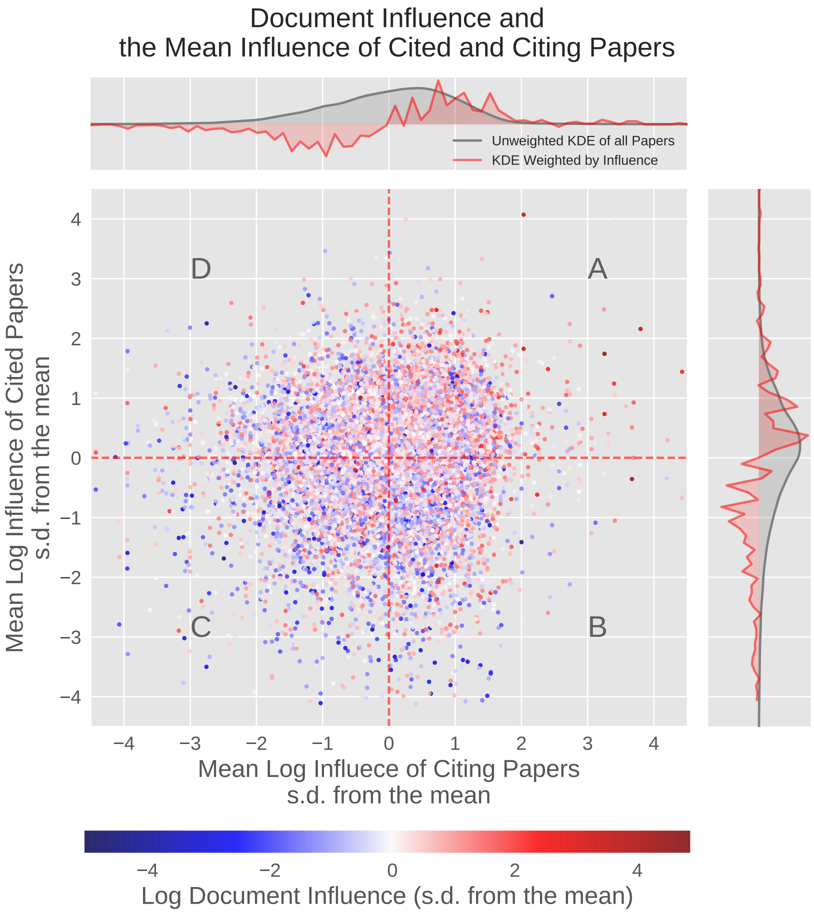

Figure 12: Influence of papers cited at least once in the APS (color points) and the mean log influence of the papers that cite them (x-axis) and mean log influence of the papers they cite (y-axis). Also shown is the kernel density estimate (KDE) in both spatial dimensions for unweighted (gray) and influence-weighted (red) distributions. In all dimensions, points represent s.d. from the mean influence. For presentations, outliers beyond ±20 s.d. of the mean are omitted. Papers in quadrant A cite and are cited by highly influential papers. Papers in quadrant B are cited by influential papers but cite uninfluential papers. In quadrant C, papers cite and are cited by uninfluential papers. And in D, papers cite influential papers, but are cited by uninfluential papers. The weighted KDEs (red distributions in plots above and to the right of the scatter plot) demonstrate that influential papers cite and are cited by papers of greater-than-average influence.

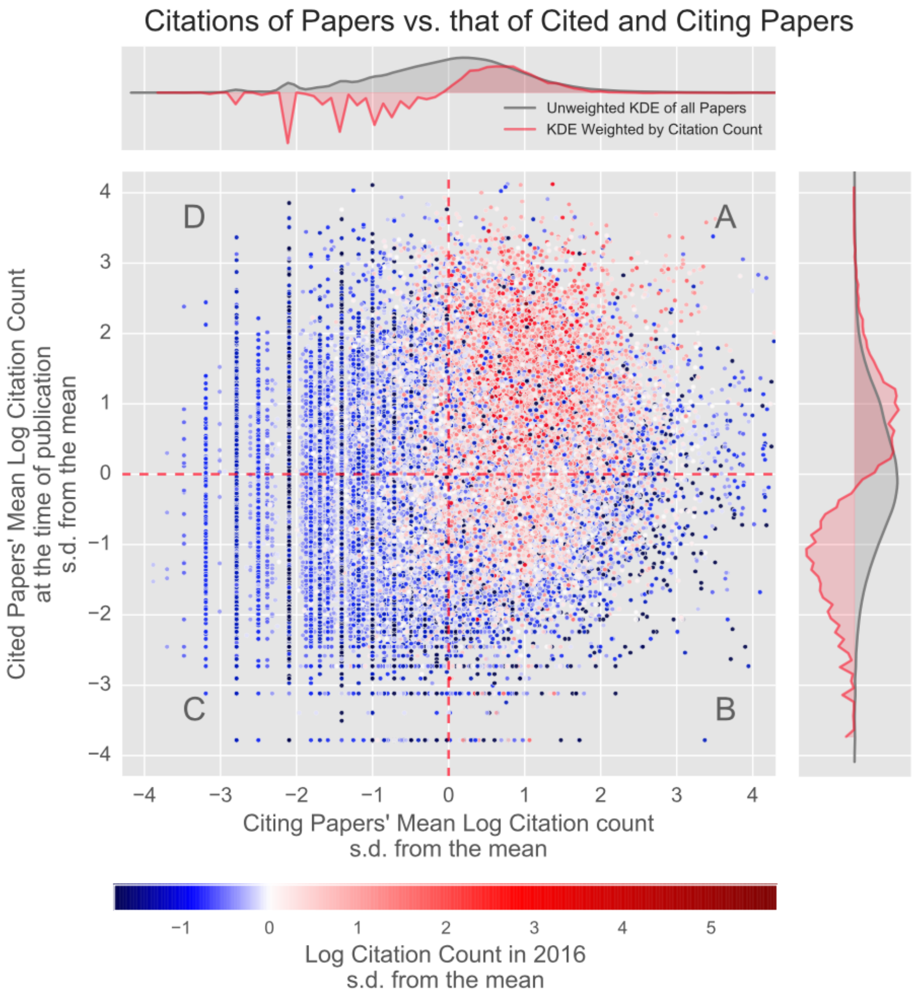

Figure 13: Log citation count of papers cited at least once in the APS (color points) and the mean log citation count of the papers that cite them (x-axis) and mean log citation count of the papers they cite (at the time of publication; y-axis). KDEs are depicted for unweighted (gray) and citation-weighted (red) distributions. In all dimensions, points represent s.d. from the mean and outliers beyond ±20 s.d. are omitted. Papers in quadrant A cite and are cited by highly-cited papers. Papers in quadrant B are cited by highly cited papers but cite under-cited papers. In quadrant C, papers cite and are cited by under-cited papers. In D, papers cite highly cited papers, but are cited by under-cited ones. The weighted KDEs show that highly cited papers cite and are cited by papers with greater-than-average citation counts.

4.8 Document Influence vs. Persistence & Sleeping Beauty Scores

We found a complex relationship between discursive influence, citation counts and other aspects of a publication. In particular, the topic uniformity of authors' bibliographies (Persistence; Eq. 11 in the main text) and the convexity of papers' citations over time (sleeping beauty scores [20]) were correlated to influence and citations. Figure 14 illustrates the three-dimensional relationship between influence, citation counts and author persistence. Papers written by persistent authors were cited more than those by less persistent authors. However, papers written by persistent authors tended to be less influential. This is primarily due to the fact that influence is summed over topics—the same topics across which, if an author is spread thin, they are less persistent. In other words, the "easiest" way for a document to be highly influential is to contribute discursive changes to *many* topics. This diversity, then, diminishes the persistence of an author. Also, persistence is unbounded and it is fairly difficult to be exceptionally persistent: one must have written many papers in the same topic(s) over a long period of time. It might be that overcoming this difficulty is related, in a professional sense, to overcoming the difficulty of having a highly cited paper.

Papers with relatively high influence but few citations (upper left quadrant in Figure 1 in the main text) were those that made discursive contributions that went un-credited with citations. One hypothesis was that these papers tended to introduce new concepts, the value of which were not taken up by the typical news cycle of science. Figure 15 shows the relationship between papers' influence, citation counts and sleeping beauty score. In this case, the trend is different than for author persistence. Papers with above-average influence tended to score higher on the sleeping beauty index. The inflection point for this effect—where papers of above-average influence tend to have an above-average sleeping beauty score—is approximately $+.5$ s.d. of the mean. In terms of citations, the sleeping beauty score is somewhat circular because it uses the maximum citations in certain years after publication. Highly cited papers do tend to have slightly above-average sleeping beauty scores. This three-way relationship reflects the fact that sleeping beauties are not randomly selected for an over-due spike in citations, but instead, that a community was late in acknowledging their contribution to discourse.

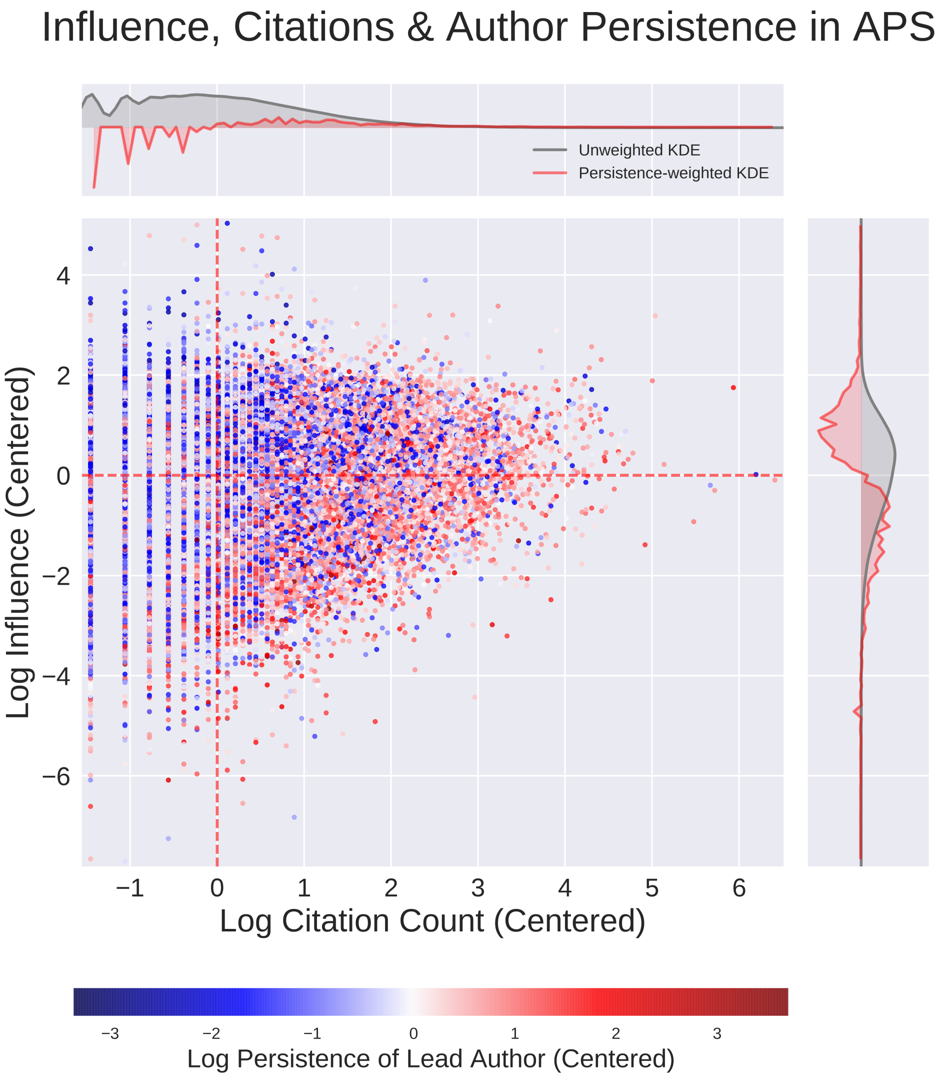

Figure 14: Influence vs. citations vs. author persistence in the APS data. All dimensions are logged and centered for visualization. Unweighted (gray) and persistence-weighted KDEs are displayed for each spatial axis. Higher influence papers are less likely to be authored by persistent authors, whereas highly cited papers are more likely to have a more persistent author.

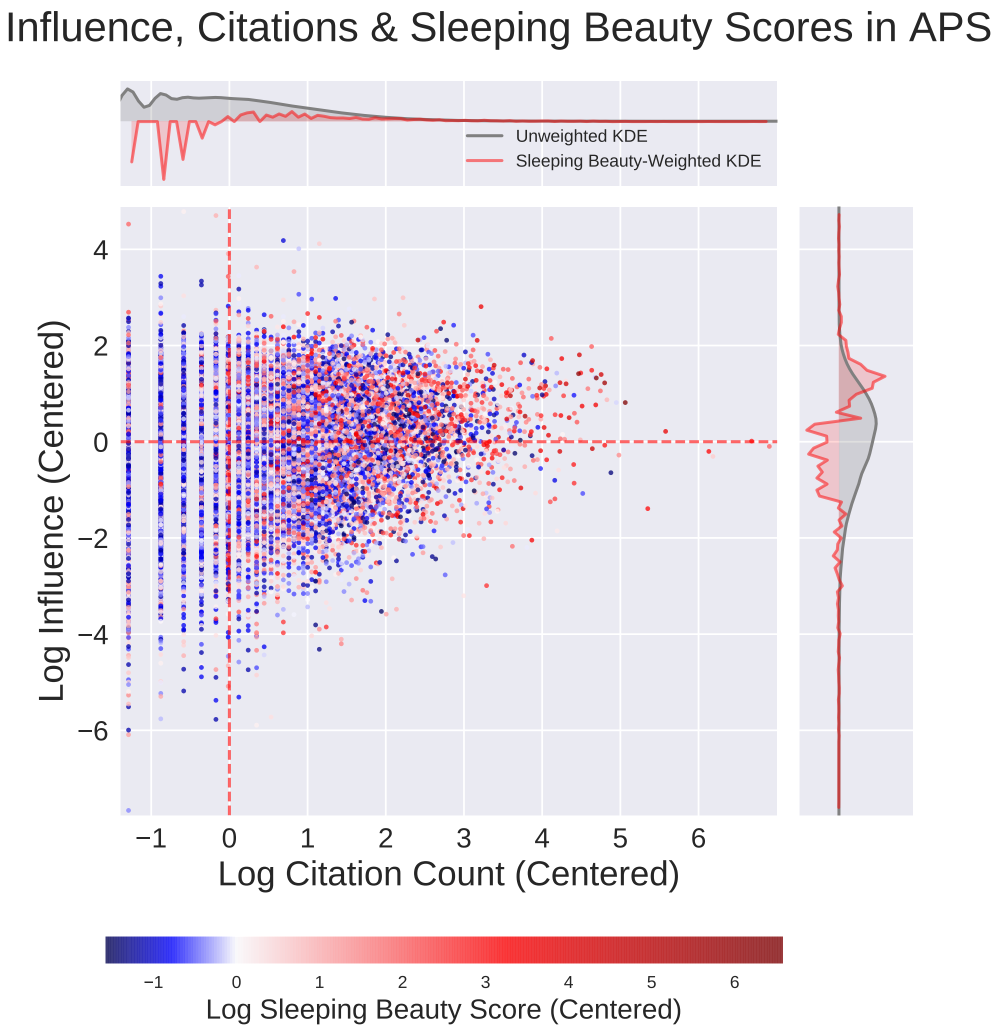

Figure 15: Influence vs. citations vs. sleeping beauty scores in the APS data. All dimensions are logged and centered. Unweighted (gray) and SB-weighted KDEs are shown for each spatial axis. Higher influence papers tend to have higher SB scores, as are highly cited papers.

4.9 Influence, Citations and Author Influence in JSTOR

The relationship between influence and citations is reviewed in the main text (see Figure 1). In the JSTOR collection, we determined that this relationship is specific to topics. For each topic, the correlation between influence and citation counts was computed ($\hat{\ell}_d \times C_d$; Table 6). These correlations were mostly significant, but fairly weak, ranging from -0.05 to 0.06. Given the design of RDIM, influence is topic-specific but citations are not. To explore how citations relate to each topic specifically, we scaled citation counts by the document-topic mixture and computed the same correlations ($\hat{\ell}_d \times \theta_d C_d$). With topic-relative citation counts, the correlations increased for all topics compared to raw counts. The correlation was positive for all but two topics and ranged from -0.02 to 0.13. While the strength of these correlations make it hard to use discursive influence as the sole or even primary predictor of citation influence, our finding highlights the model's ability to pick up topics sensitive to different citation habits.

RDIM is primarily a model of documents, but it affords us with estimates of how contextual features, like authorship, can change influence. The relationship between authors extrinsic influence and their own citation counts can be compared by looking at an author's bibliography and their coefficients in $\hat{\mu}$. Table 7 shows the correlation between authors' coefficients and their total citations, scaled by the document-topic mixture for each document. When topics are ordered by correlation, an interesting picture emerges: topics related to math and the natural sciences are found near the bottom (shown in red) and topics about arts and humanities are close to the top. Social sciences span most of the range, but are roundly displaced by natural sciences on the bottom. This is highly suggestive of the nature of *how* authors impact different areas of research. In the natural sciences, authors' extrinsic impact on their articles' discursive influence is not predictive of those articles' citation counts. However, in the humanities and social sciences, where narrative may be more central, authors who's name-sake boost discursive influence are more highly cited. Authors' ability to change the mix of discourse appears to be valued more in the humanities and social sciences than in math and natural sciences. In fields focused on natural and inflexible or formal objects of inquiry (e.g. Cell Biology, Physical Chemistry, and various statistics topics), authors that change the narrative are less rewarded than in fields where inquiry is focused on artistic, humanistic or social objects (e.g. Philosophy, Literary Theory, Education).

Topic	$\hat{\ell}_d \times C_d$	$\hat{\ell}_d \times \theta_d C_d$
Academic Verbiage	0.0507	0.0954
Academic Verbiage	0.0410	0.0982
Academic Verbiage	0.0534	0.0910
American History	(0.0069)	0.0455
Anthropology	0.0152	0.0839
Archaeology	0.0160	0.0737
Behavioral Psychology	(-0.0028)	0.0580
Behavioral Science	0.0588	0.1140
Biostatistics	-0.0007	0.0594
Cell Biology	0.0138	0.0693
Child & Family Studies	-0.0294	0.0329
Cognitive Science	0.0656	0.1191
Constructivism	0.0295	0.0926
Cultural History	-0.0486	-0.0141
Cultural Studies	(0.0022)	0.0647
Decision Science	0.0422	0.1053
Demography	0.0659	0.1152
Ecology	0.0469	0.1050
Econometrics	0.0306	0.0984
Economic Development	(-0.0030)	0.0684
Education	0.0414	0.1058
Electoral Politics	(0.0037)	0.0777
Engineering	0.0271	0.0752
Environmental Science	(0.0089)	0.0711
Ethics	0.0171	0.0890
Evolutionary Biology	0.0268	0.0882
Evolutionary Development	0.0561	0.1063
Family and Child Development	0.0128	0.0753
Frequentist Statistics	(-0.0005)	0.0569
Game Theory	0.0255	0.0919
Group Behavior	0.0257	0.0813
Health Genetics	(0.0066)	0.0625
International Relations	-0.0356	0.0357
Law	(0.0025)	0.0738
Literary Theory	-0.0448	-0.0199
Management	0.0337	0.0961
Marine Ecology	0.0522	0.1111
Marketing	(0.0076)	0.0816
Medicine	0.0460	0.1096
Networks	0.0184	0.0816
Organization Science	0.0516	0.1121
Organizational Behavior & Theory	0.0622	0.1252
Philosophy	0.0216	0.0879
Physical Chemistry	0.0163	0.0784
Plant Biology	(0.0003)	0.0519
Psychology	0.0262	0.0876
Psychology (Quantitative)	0.0391	0.0858
Qualitative Research	0.0129	0.0203
Quantitative Social & Health Science	0.0276	0.0949
Sexual Health	0.0444	0.0957
Social Movements	-0.0256	0.0420
Statistics (Probability)	0.0186	0.0800
Stochastic Processes	0.0292	0.0887

Table 6: For each JSTOR topic, Spearman correlations between influence, $\hat{\ell}_d$ and unscaled documents citation counts, C_d, as well as between influence and citations scaled by document-topic mixtures, $\theta_d C_d$. Correlations significant at $p < 0.01$ are shown in green (positive) and red (negative).

Topic	$\hat{\mu}_A \times \sum_{d \in A} C_d \theta_d$
Anthropology	0.3728
Child & Family Studies	0.3549
Education	0.3459
Electoral Politics	0.3299
Organization Science	0.3252
Philosophy	0.3199
Academic Verbiage	0.3099
Law	0.3022
Ethics	0.2982
Organizational Behavior & Theory	0.2978
Cognitive Science	0.2950
Cultural Studies	0.2939
Constructivism	0.2917
Group Behavior	0.2783
Social Movements	0.2662
International Relations	0.2617
Cultural History	0.2521
Literary Theory	0.2504
Academic Verbiage	0.2499
Sexual Health	0.2333
Demography	0.2317
American History	0.2290
Economic Development	0.2203
Decision Science	0.2157
Family and Child Development	0.2147
Game Theory	0.2130
Marketing	0.2102
Quantitative Social & Health Science	0.2046
Academic Verbiage	0.1923
Behavioral Psychology	0.1817
Management	0.1765
Psychology (Quantitative)	0.1711
Networks	0.1705
Psychology	0.1360
Qualitative Research	0.1015
Econometrics	0.0892
Stochastic Processes	0.0841
Behavioral Science	0.0810
Engineering	0.0776
Archaeology	0.0741
Evolutionary Development	0.0706
Ecology	0.0595
Medicine	0.0521
Marine Ecology	0.0367
Statistics (Probability)	0.0331
Plant Biology	0.0173
Biostatistics	(-0.0037)
Frequentist Statistics	(-0.0039)
Physical Chemistry	-0.0230
Evolutionary Biology	-0.0295
Environmental Science	-0.0773
Health Genetics	-0.0885
Cell Biology	-0.1068

Table 7: Spearman correlations between author coefficients and authors' total citations, scaled by document mixture. Arts & Humanities-related topics are shown in blue, Social Sciences in green and Math & Natural Sciences in red. Correlations not significant at $p < 0.01$ are given in parentheses.

5 De-biasing Citations: Half-life of Discourse

Citation counts exhibit preferential attachment: highly cited papers continue to be cited at a higher rate than less cited papers [28, 29]. Such reinforcement can be observed as a long tail in the distribution of citation counts over papers [14]. In both datasets, we found the distribution for document influence, I_d, had a shorter tail than citations counts, C_d. For individual papers, citations follow a log-normal decay over time [34], but the distribution of citation counts across papers is scale-free [29]. To assess the size of the tail in both distributions, type 1 power-laws of the un-scaled form $y = x^{-\gamma}$ were fit to C_d and I_d with a cutoff [14]. Smaller values of γ denote longer-tailed distributions. In APS, $\gamma_C = .20$ and $\gamma_I = .29$ and in JSTOR $\gamma_C = .21$ and $\gamma_I = .48$: citations have a longer tail in both sets. While citation counts have been shown to follow such a curve, document influence is observed. We also fit negative binomial PMFs to each distribution and found that citations have significantly higher dispersion than document influence (APS $Disp(C) = 1.10$, $Disp(I) = 1.00$; JSTOR $Disp(C) = 1.04$, $Disp(I) = 1.00$), affirming their longer tail.

While citation counts are bursty and decay, topic contribution is more stable. Much of the variance in topic contributions is because documents tend not to contribute much to most topics, but instead, they make sizable contributions to a few topics. These contributions decay less quickly. With a 10,000-document sample from JSTOR, the average topic contribution decayed at a near-linear, normalized rate of 0.02 per year (about 2% / year). This confirms that our model's topic contribution metric helps de-bias the credit assigned by citations. It also suggests that, while some genuine citation-traced impact may not be present in text, there is a sizable portion of influence simply not represented in citations.

To quantify the stability of topic contribution (Eq. 10; main text) compared to citation counts, we computed the half-life of each for a 1,000-document sample from every percentile of the influence distribution. Half-life, $T_{1/2}$, is the number of time-steps after which the given score is half of what it was at the beginning. $T_{1/2}$ was calculated for both citations and topic contribution. Documents in higher percentiles had longer half-lives for each metric. But in every case, citations had shorter half-lives and smaller variance than topic contribution (Table 8). This confirms that topic contribution is more stable over all, but it also shows that higher influence papers tend to make longer-lasting discursive contributions.

Influence	**Citation Counts**		**Topic Contribution**	
Percentile	**Variance**	$\boldsymbol{T_{1/2}}$	**Variance**	$\boldsymbol{T_{1/2}}$
0-10	0.86	2.74	2.56	5.50
10-20	0.51	2.54	2.24	6.65
20-30	0.64	2.30	2.08	6.53
30-40	0.76	1.34	2.64	6.38
40-50	0.62	1.47	2.01	6.96
50-60	0.83	2.16	1.77	7.34
60-70	1.57	1.65	2.03	7.81
70-80	1.43	1.75	2.08	7.36
80-90	0.73	2.03	1.82	8.54
90-100	1.90	2.95	1.96	10.07

Table 8: Variance and half-life, $T_{1/2}$, for citation counts and topic contribution (Eq. 10; main text). 1,000 documents were randomly sampled from each percentile of the document influence distribution. Variance was calculated on observed data, citation half-life was calculated using continuous log-normal PDFs fit to observed data, and contribution half-lives were observed. For variance, scales were normalized to make them comparable and half-life is presented in years.

6 Models of Citations

Within the JSTOR collection, only 29% of documents were cited by another document. Many of our analyses of APS data relied on the citation-rich environment of physics research. JSTOR offers a broader challenge: predicting what documents will be cited. Here, we explore how discursive influence can help predict citedness and citations counts in JSTOR. As citation habits are time-dependent we used both document influence, I_d, and date, t_d, as predictors for whether or not a document would be cited at least once. A stack of increasingly specified logit models were fit to citedness, $C_d > 0$ (Table 9). The fully specified model minimized AIC and estimated a negative constant effect, and positive effects for influence and date. These models were not designed to be accurate predictors of citedness, but to show that influence is time-agnostic and *statistically helpful* in situations where there is a temporal effect.

A second stack of models was used to predict citations counts, C_d. These models consisted of increasingly specified logistic-link negative binomial regressions on C_d (Table 10). The fully specified model had the lowest AIC and estimated a negative constant effect, positive singular effects for influence and date, and a small negative effect for the interaction between document influence and date. This second order effect was too weak to warrant further characterization.

Model	AIC	Parameter	Coef.	s.e.	95% C.I.	P>$\lvert z\rvert$
$C_d > 0 \sim 1$	16,573.5					
		Intercept	-0.7948	0.019	[-0.831, -0.758]	0.000
$C_d > 0 \sim I_d + 1$	16,031.4					
		Intercept	-0.8286	0.019	[-0.866, -0.791]	0.000
		I_d	0.4388	0.019	[0.401, 0.476]	0.000
$C_d > 0 \sim t_d + 1$	16,504.5					
		Intercept	-0.7994	0.019	[-0.836, -0.763]	0.000
		t_d	0.1583	0.019	[0.121, 0.195]	0.000
$C_d > 0 \sim I_d + t_d + 1$	16,003.0					
		Intercept	-0.8308	0.019	[-0.869, -0.793]	0.000
		I_d	0.4268	0.019	[0.389, 0.465]	0.000
		t_d	0.1100	0.019	[0.072, 0.148]	0.000
$C_d > 0 \sim I_d * t_d + 1$	16,002.5					
		Intercept	-0.8310	0.019	[-0.869, -0.793]	0.000
		I_d	0.4268	0.019	[0.389, 0.465]	0.000
		t_d	0.1097	0.020	[0.071, 0.148]	0.000
		$I_d * t_d$	-0.0014	0.019	[-0.036, 0.039]	0.943

Table 9: Increasingly specified logit models predicting citedness, $C_d > 0$, of documents in the JSTOR collection.

Acknowledgments

Thanks to Sean Gerrish for clarifications on the DIM and to John Lafferty and Sean for making the DTM and DIM code freely available. Thanks also Yadu Nand, the UChicago *Research Computing Center*, *Open Science Grid* and *SWIFT* teams for generous computing resources and support.

Model	AIC	Parameter	Coef.	s.e.	95% C.I.	P>$\|z\|$
$C_d \sim 1$	768,344.3					
		Intercept	-0.7787	0.003	[-0.784, -0.773]	0.000
$C_d \sim I_d + 1$	754,381.2					
		Intercept	-0.8307	0.003	[-0.836, -0.825]	0.000
		I_d	0.3262	0.003	[0.321, 0.332]	0.000
$C_d \sim t_d + 1$	767,628.1					
		Intercept	-0.7814	0.003	[-0.787, -0.776]	0.000
		t_d	0.0766	0.003	[0.071, 0.082]	0.000
$C_d \sim I_d + t_d + 1$	754,108.4					
		Intercept	-0.8320	0.003	[-0.838, -0.826]	0.000
		I_d	0.3226	0.003	[0.317, 0.328]	0.000
		t_d	0.0479	0.003	[0.042, 0.053]	0.000
$C_d \sim I_d * t_d + 1$	753,927.2					
		Intercept	-0.8356	0.003	[-0.841, -0.830]	0.000
		I_d	0.3267	0.003	[0.321, 0.332]	0.000
		t_d	0.0369	0.003	[0.031, 0.043]	0.000
		$I_d * t_d$	-0.0377	0.003	[-0.032, -0.043]	0.000

Table 10: Increasingly specific negative binomial regression models predicting citation counts, C_d, of papers in the JSTOR collection.

References

[1] *ACL '12: Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries* (Stroudsburg, PA, USA, 2012), Association for Computational Linguistics.

[2] AHMED, A., AND XING, E. P. Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI 2010)* (2010), vol. 20, pp. 29–39.

[3] ANDERSON, A., MCFARLAND, D., AND JURAFSKY, D. Towards a computational history of the acl: 1980-2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries* (2012), pp. 13–21.

[4] BIRD, S., DALE, R., DORR, B. J., GIBSON, B., JOSEPH, M. T., KAN, M.-Y., LEE, D., POWLEY, B., RADEV, D. R., AND TAN, Y. F. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proceedings of the 6th International*

Conference on Language Resources and Evaluation (LREC 2008) (2008), pp. 1755–1759.

[5] BLEI, D. M. Probabilistic topic models. *Communications of the ACM 55*, 4 (2012), 77–84.

[6] BLEI, D. M., AND LAFFERTY, J. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning (ICML 2006)* (2006), pp. 113–120.

[7] BLEI, D. M., AND LAFFERTY, J. D. A correlated topic model of science. *The Annals of Applied Statistics*, 1 (2007), 17–35.

[8] BLEI, D. M., AND LAFFERTY, J. D. Topic models. *Text mining: classification, clustering, and applications 10*, 71 (2009), 34.

[9] BLEI, D. M., NG, A. Y., AND JORDAN, M. I. Latent Dirichlet allocation. *Journal of Machine Learning Research 3* (2003), 993–1022.

[10] BOYD-GRABER, J., AND BLEI, D. M. Multilingual topic models for unaligned text. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)* (2009), pp. 75–82.

[11] BOYD-GRABER, J., HU, Y., AND MIMNO, D. Applications of topic models. *Foundations and Trends in Information Retrieval 11*, 2-3 (2017), 143–296.

[12] CARON, F., DAVY, M., AND DOUCET, A. Generalized Polya urn for time-varying Dirichlet process mixtures. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (UAI 2007)* (2007), pp. 33–40.

[13] CHANG, J., GERRISH, S., WANG, C., BOYD-GRABER, J. L., AND BLEI, D. M. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, Eds. 2009, pp. 288–296.

[14] CLAUSET, A., SHALIZI, C. R., AND NEWMAN, M. E. Power-law distributions in empirical data. *SIAM Review 51*, 4 (2009), 661–703.

[15] DUBIN, D. The most influential paper Gerard Salton never wrote. *Library Trends 52*, 4 (2004), 748–764.

[16] GERRISH, S., AND BLEI, D. M. A language-based approach to measuring scholarly impact. In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)* (2010), pp. 375–382.

[17] GRIFFITHS, T. L., AND STEYVERS, M. Finding scientific topics. *Proceedings of the National Academy of Sciences 101*, 1 (2004), 5228–5235.

[18] HALL, D., JURAFSKY, D., AND MANNING, C. D. Studying the history of ideas using topic models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)* (2008), pp. 363–371.

[19] KALMAN, R. E. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering 82*, 1 (1960), 35–45.

[20] KE, Q., FERRARA, E., RADICCHI, F., AND FLAMMINI, A. Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences 112*, 24 (2015), 7426–7431.

[21] KUHN, T. S. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.

[22] MIMNO, D., WALLACH, H. M., NARADOWSKY, J., SMITH, D. A., AND MCCALLUM, A. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)* (2009), pp. 880–889.

[23] MOON, T. K. The expectation-maximization algorithm. *IEEE Signal Processing Magazine 13*, 6 (1996), 47–60.

[24] NEWMAN, M. E. J. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences 101*, 1 (2004), 5200–5205.

[25] ORBANZ, P., AND TEH, Y. W. Bayesian nonparametric models. In *Encyclopedia of Machine Learning*. Springer, 2011, pp. 81–89.

[26] ORBANZ, P., AND TEH, Y. W. Modern bayesian nonparametrics. In *Neural Information Processing Systems* (2011).

[27] PAUL, M., AND GIRJU, R. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)* (2009), pp. 1408–1417.

[28] PRICE, D. D. S. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science 27*, 5 (1976), 292–306.

[29] REDNER, S. How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B – Condensed Matter and Complex Systems 4*, 2 (1998), 131–134.

[30] SMADJA, F. Retrieving collocations from text: Xtract. *Computational Linguistics 19*, 1 (1993), 143–177.

[31] STEYVERS, M., AND GRIFFITHS, T. Probabilistic topic models. *Handbook of Latent Semantic Analysis 427*, 7 (2007), 424–440.

[32] STEYVERS, M., SMYTH, P., ROSEN-ZVI, M., AND GRIFFITHS, T. Probabilistic author-topic models for information discovery. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2004)* (2004), pp. 306–315.

[33] UZZI, B., MUKHERJEE, S., STRINGER, M., AND JONES, B. Atypical combinations and scientific impact. *Science 342*, 6157 (2013), 468–472.

[34] WANG, D., SONG, C., AND BARABÁSI, A.-L. Quantifying long-term scientific impact. *Science 342*, 6154 (2013), 127–132.

[35] WANG, M., YU, G., AND YU, D. Measuring the preferential attachment mechanism in citation networks. *Physica A: Statistical Mechanics and its Applications 387*, 18 (2008), 4692–4698.

[36] ZHANG, J., GEROW, A., ALTOSAAR, J., EVANS, J., AND SO, R. J. Fast, flexible models for discovering topic correlation across weakly-related collections. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)* (2015), pp. 1554–1564.