sociological science

Finding Cultural Holes: How Structure and Culture Diverge in Networks of Scholarly Communication

Daril A. Vilhena,[a] Jacob G. Foster,[b] Martin Rosvall,[c] Jevin D. West,[a] James Evans,[d] Carl T. Bergstrom[a]

a) University of Washington; b) University of California–Los Angeles; c) Umeå University; d) University of Chicago

Abstract: Divergent interests, expertise, and language form cultural barriers to communication. No formalism has been available to characterize these "cultural holes." Here we use information theory to measure cultural holes and demonstrate our formalism in the context of scientific communication using papers from JSTOR. We extract scientific fields from the structure of citation flows and infer field-specific cultures by cataloging phrase frequencies in full text and measuring the relative efficiency of between-field communication. We then combine citation and cultural information in a novel topographic map of science, mapping citations to geographic distance and cultural holes to topography. By analyzing the full citation network, we find that communicative efficiency decays with citation distance in a field-specific way. These decay rates reveal hidden patterns of cohesion and fragmentation. For example, the ecological sciences are balkanized by jargon, whereas the social sciences are relatively integrated. Our results highlight the importance of enriching structural analyses with cultural data.

Keywords: cultural holes; jargon; scholarly communication; content analysis; complex networks; information theory

Editor(s): Jesper Sørensen, Delia Baldassarri; **Received:** December 20, 2013; **Accepted:** February 8, 2014; **Published:** June 9, 2014

Citation: Vilhena, Daril A., Jacob G. Foster, Martin Rosvall, Jevin D. West, James Evans, and Carl T. Bergstrom. 2014. "Finding Cultural Holes: How Structure and Culture Diverge in Networks of Scholarly Communication." Sociological Science 1: 221-238. **DOI:** 10.15195/v1.a15

Structural holes have provided a fruitful framework for analyzing the benefits that people and institutions reap from their location in a social network. Those whose networks span gaps in the social fabric obtain information, resources, and control through brokering social actors on either side. They also experience greater freedom of action (Burt, 1992). Those with no bordering holes experience greater competition for resources and increased constraint. Riffing on this generative term, Pachucki and Breiger coined the companion concept of "cultural holes" to label an emerging theme in cultural analysis (Pachucki and Breiger, 2010). This theme emphasizes that common culture—shared meanings, tastes, and interests—enables ties between individuals and institutions. When common culture is absent, the resulting "cultural hole" makes existing ties impoverished and new ties improbable or impossible. In other words, gaps in the *cultural* fabric may make it problematic or profitless to bridge coinciding gaps in the social fabric (Xiao and Tsui, 2007). Social actors who are structurally or physically proximate may be kept apart by deep divergences in matters of concern.

Despite widespread interest in the notion of cultural holes, no unifying measurement framework like Burt's (1992) calculus of structural constraint has emerged to locate or quantify them. This is largely to be expected, given their recent introduction. But formalization is made doubly difficult by the breadth of what sociologists label "culture," from tastes (Erickson, 1996; Lizardo, 2006; Vaisey and Lizardo, 2010) and values (Homans, 1961) to formal differences in artistic genres (Han, 2003; Sonnett, 2004) or rifts between institutional logics (Friedland, 2009). Nevertheless, underlying the apparent diversity of cultural objects is a common capacity to circulate. This suggests an analytical focus on human communication, construed here as the processes through which values, tastes, styles, and logics are transmitted. Indeed, cultural circulation is typically analyzed through its symbolic or linguis-

tic traces, and scholars of culture have explored the semiotics of everything from slang, gesture, and clothing to electronics and condoms (Tavory and Swidler, 2009). Even idiosyncratic cultural variations like the "important matters" probed by the General Social Survey are revealed through communication and comprise unusual, chatty topics like "eating less red meat" or "cloning headless frogs" (Bearman and Parigi, 2004). In other words, communication reflects a vast array of cultural objects and differences, ranging from matters of concern to forms of personal expression.

Insofar as culture is characterized by its shared, communicated quality—and differences in communicated content reflect underlying cultural differences in taste and interest—we define cultural holes in terms of the communicative burden placed on parties to an interaction. To put an ethnomethodological gloss on this definition, more *work* is required to sustain an interaction when underlying cultural differences lead one or both parties to transmit a stream of unfamiliar and unexpected symbols or behaviors (Garfinkel, 1991). More concretely, consider "epistemic cultures" in science (Knorr-Cetina, 1999) and their use of culture-specific language, or *jargon*. Jargon reflects a culture's matters of special concern, focus, and expertise, often through the coinage of compressed terms for frequently–used referents (arguably, this is exactly what Pachucki and Breiger have done with "cultural hole"). "Every science requires a special language," wrote the French philosopher Etienne Bonnot de Condillac, "because every science has its own ideas" (de Condillac, 1782). This maxim makes two points that are often overlooked in the quantitative analysis of scientific communication, although these points generalize to all forms of human communication. First, each (scientific) culture has an extensive mental catalog of concepts that reflect *distinct* concerns (de Condillac's "ideas"). Second, (scientific) cultures often develop specialized terms (jargon) to refer efficiently to the most common concepts. Discipline-specific jargon allows scientists to communicate with other members of their discipline more concisely and precisely than would be possible using everyday language. When an evolutionary biologist uses the term *fitness landscape*, this is a highly condensed shorthand for comparing expected relative reproductive success across multiple genotypes. An intertidal ecologist would require little further explanation to understand *fitness landscape*, but a financial economist might need to spend a long time on Wikipedia before she got a handle on it. Compared to the ecologist, the economist would need to put more interpretive work into sustaining the interaction. In other words, not only does every science have its own ideas, but these ideas, and their associated jargon, may or may not overlap with those dear to other disciplines.[1] Note that the same linguistic compressions occur in a regional dialect, a subcultural lingo, or a criminal argot, although the balance of intentions behind the compression—efficient communication with insiders or cultural distinction from and exclusion of outsiders—may vary from case to case.

These simple observations have profound implications for the study of culture and its communication. In science, information does not flow seamlessly from author to publication to reader. It is expressed in a particular language, reflecting particular concerns, with important consequences for its transmission. Jargon allows scientists to communicate new results quickly and effectively within the context of discipline-specific paradigms and interests, but it inhibits efficient knowledge transfer in many other situations (see Basil Bernstein's work on restricted and elaborated codes (Bernstein, 1964) for a similar concept). To list just a few examples, specialist language impedes communication when sharing medical information with a patient (Reach, 2009), publishing material intended for public outreach (Richardson, 2010), and presenting technical information to a multidisciplinary audience (Bischof and Eppler, 2010). Indeed, jargonistic compression can be used *intentionally* to obscure information (jargon as encryption) or reify cultural boundaries (jargon as shibboleth; (Sokal and Bricmont, 1998)). Technical language in patents provides an extreme case of deliberate obscurantism (Feldman, 2008). In other words, jargon accelerates the flow of information within disciplines by compressing language, but impedes communication between disciplines and makes knowledge transfer more difficult. As interests diverge and jargon becomes

[1] Consider "backward induction" versus "satisficing" as explanations for action; the underlying ideas and beliefs are not only distinct but mutually exclusive. We thank Gabriel Rossman for suggesting this example.

more central to scientific discourse in an area, it creates a chasm—a cultural hole—that impedes easy access from outside.

This trade–off between efficient communication within local cultures and inefficient communication between them merits closer attention. No general framework has been available for exploring how jargon and the underlying patterns of interest affect the structure of communication, and vice versa. In this article, we introduce an information-theoretic model of communication and develop a simple measure of cultural holes with a clear qualitative interpretation. To illustrate our method, we deploy it on a large collection of scientific papers in the JSTOR corpus. These articles have both full–text and citation information. This allows us to relate cultural information (captured by the full text) with structural traces of interaction (captured by the citation network). In citation networks, nodes represent papers and directed links represent citations among papers. We use the citation network to extract the macroscopic structure of scientific fields in JSTOR, building on extensive work that maps science in this way, e.g., (Rosvall and Bergstrom, 2008; Leydesdorff and Rafols, 2008; Small, 1999; Boyack et al., 2005); see "Data and Methods" for an extended technical description.[2] Note that our use of the structural information in citation networks to identify groups is entirely consistent with the parallel between structural and cultural holes set up by Pachucki and Breiger. As suggested by their account, we expect dense citations within a field to signal deep linkages of mutual agreement and awareness, or common culture. Conversely, we anticipate sparse citations or "structural holes" between fields to broadly coincide with cultural holes of varying depth. In the analysis that follows, we test these two contentions.

[2]Although we apply our cultural holes measure to scientific fields as identified using the map equation, this approach can be applied to any meaningful grouping of documents, authors, or corpora and can scale down to individual journals, authors, and articles. Indeed, a similar methodology is currently in use to screen submissions to the online preprint server ArXiv (P. Ginsparg, private communication, February 28, 2013). Articles with excessive jargon distance from existing fields are likely to be produced by authors outside the mainstream research community, on topics from perpetual motion machines to novel theories of everything.

Using Information Theory to Model Scholarly Culture and Communication

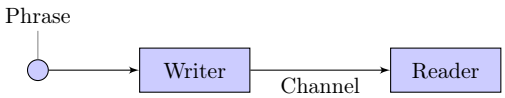
To quantify the communicative burden imposed by cultural holes, we construct a simple model of symbolic communication. To make the discussion concrete, we describe the model in the context of science, but it is very general in scope. We first consider a model for optimal communication within a given scientific culture/field; then we quantify the penalty for using these languages between fields. In the "Data and Methods" section, we show how to operationalize all building blocks of our model using a corpus that combines structural (citation) and cultural (text) traces.

Imagine a writer communicating with a reader through a channel, for example, a scientific article (Shannon, 1948). Let $\mathcal{X}$ denote the space of all phrases that the writer and the reader might use to communicate; these phrases broadly correspond to concepts. The writer is characterized by a codebook P_i that maps from phrases to codewords; the subscript i denotes the writer's field of science or scholarship. The codebook P_i has a corresponding probability distribution over a random variable X_i, with values $x \in \mathcal{X}$. This probability distribution tracks the importance of each phrase in field i; important phrases are used frequently, for example, *fitness landscape* in evolutionary biology. The writer generates a message by drawing phrases at random with probability $p_i(x)$. In other words, she chooses phrases in proportion to their importance in her particular scientific culture. Now imagine that she transcribes[3] the phrases into codewords from whatever "language" is appropriate for her reader's scholarly field, using that field's codebook P_j. We assume that the language of each scientific field is *optimized* based on how frequently a given phrase is used. This assumption is commonly used to explain the power–law distribution of English words (Zipf, 1935, 1949). The optimum codeword for a phrase x used in field i with probability $p_i(x)$ has length $-\log_2 p_i(x)$ in bits (Cover and Thomas,

[3]Of course, unless she is writing explicitly for an interdisciplinary audience, she will write in the language of her own field. The transcription process described here is a useful fiction, allowing us to estimate the effort required when readers from various fields decode the resulting text.

2006).[4] We use codeword length as a proxy for interpretive effort. Short codewords take less time and effort to "unpack" than long codewords. In essence, we turn the Zipfian principle of least effort around; assuming field-specific language is optimized for internal consumption, we can use phrase frequency and hence codeword length to infer interpretive effort.

First, consider a scientist from field i sending a message to a reader from the same field. Writer and reader have the same probability distribution and codebook, denoted by the blue boxes.

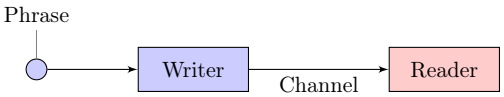


The writer selects phrase x with probability $p_i(x)$. Because she and her reader have the same codebook, she encodes x with a codeword of length $-\log_2 p_i(x)$. The expected message length per phrase is simply the Shannon entropy of X_i given probability distribution $p_i(x)$:

$$H(X_i) = -\sum_{x\in\mathcal{X}} p_i(x)\log_2 p_i(x). \qquad (1)$$

This result follows directly from the Shannon source coding theorem (Shannon, 1948). Indeed, this is the most efficient encoding of messages generated by sampling phrases $x \in \mathcal{X}$ with the writer's probability distribution $p_i(x)$. In cultures where many phrases are used with equal probability (i.e., probability mass is spread evenly across phrases), the entropy will be large, as will the average message length per phrase. If some phrases are used very frequently, the entropy will be smaller. These two situations model the *efficiency* of jargon for communication within fields.[5]

Now consider the more interesting case, in which the writer and the reader come from different fields and have different codebooks. This situation is represented below.



What is the average message length per phrase, given that the writer now encodes phrases optimally[6] for a reader in a different field? Denote the writer's probability distribution over phrases $p_i(x)$ and the reader's probability distribution $p_j(x)$. The writer selects phrases x with probability $p_i(x)$ and encodes them in codewords with length $-\log_2 p_j(x)$. The expected length of the writer's message per phrase is then the cross entropy of the two distributions (Cover and Thomas, 2006), that is, the entropy of X_i given $p_i(x)$ plus the Kullback–Leibler divergence between p_i and p_j (Kullback and Leibler, 1951):

$$Q(p_i||p_j) = -\sum_{x\in\mathcal{X}} p_i(x)\log_2 p_j(x). \qquad (2)$$

This quantity will always be larger than the Shannon entropy; a message sent to a reader in a different culture will, on average, be longer than the same message sent to a reader in the same one. Intuitively, a longer message will require more effort to read and understand than a shorter message, such that communication is more costly. Most of the extra cost incurred comes from phrases that are common in the author's field but rare in his or her reader's field. Nontechnical phrases ("here we show") will have comparable frequencies in both fields. If a particular phrase y is used very frequently in field i and very rarely in field j, the respective codewords will vary enormously in length, $-\log_2 p_i(y) \ll -\log_2 p_j(y)$. Insofar as the length of codewords corresponds to the time or effort required to decipher a message, an article written by someone in field i will require a reader from field j to expend much more energy to understand it than a reader from field i. The different message lengths thus reflect the inefficient communication between fields with distinct scientific cultures. In the worst case scenario, the two cultures concentrate their probability mass (i.e., interests) on completely disjoint subsets of phrases.[7] They have entirely distinct jargon, forc-

[4] Technically, codewords must have integer lengths $-\log_2 p_i(x)$ is the noninteger codeword length that gives the correct lower bound on the optimal average codeword length, $H(X_i) = -\sum_{x\in\mathcal{X}} p_i(x)\log_2 p_i(x)$.

[5] One could equivalently interpret $-\log_2 p_i(x)$ as the surprisal associated with the use of phrase x and $H(X_i)$ as the expected surprisal of a reader from field i reading a text generated according to the interests of field i, that is, her own interests.

[6] Again, keep in mind that this encoding is really a proxy for how much effort will be required from a reader in a different field

[7] As was famously described by C. P. Snow in a lecture rather appropriately titled "The Two Cultures" (Snow and Collini, 2012).

ing a reader from the one to look up *nearly every phrase* used by an author in the other.

With these results in hand, we can quantify the efficiency of communication from field i to field j as the ratio of the average message length *within* field i to the average message length *between* fields,

$$E_{ij} = \frac{H(X_i)}{Q(p_i||p_j)} = \frac{-\sum_{x\in\mathcal{X}} p_i(x)\log_2 p_i(x)}{-\sum_{x\in\mathcal{X}} p_i(x)\log_2 p_j(x)}, \quad (3)$$

and similarly define the cultural hole experienced by a reader from field j reading an article in field i as $C_{ij} = 1 - E_{ij}$. We denote the average cultural hole around field i as $C_i = \sum_j C_{ij}/N$, where we sum over the $N = 60$ potential "reading" fields.

This operationalization of cultural holes, though simple, has several advantages. First, unlike most attempts to quantify semantic information or measure cultural distance, ours is based on an explicit model of communication and has firm information-theoretic foundations (Cover and Thomas, 2006).[8] Second, because we are agnostic about what scholarly "phrases" are, we can incorporate many kinds of language, from mathematical formulae to canonical cases in legal scholarship. Moreover, we can incorporate other signals beyond those from language, like articles of clothing worn or gestures performed—anything for which we can define a probability distribution over types and hence a proxy for interpretive effort.[9] Finally, our framework is easily extensible to more complex models of phrase and signal generation, to hierarchical structure in the codebook, and so on, so long as the model assigns a probability to the appropriate semantic units.

Data and Methods

To illustrate our approach, we measure the cultural holes between scholarly cultures or fields in JSTOR. Recall that our method requires both the identification of "fields" and the assembly of "codebooks" that capture the culture of a field through its probability distribution over phrases. Fields are identified based on patterns of citation between articles using the map equation (Rosvall and Bergstrom, 2008). To assess culture, we assembled cataloges of phrases and their field-specific frequency distributions using full-text trigrams (distinct three word combinations) drawn from a representative subsample of articles in each field written between 1990 and 2010. These frequency distributions serve as the codebooks in our model of scholarly communication. In this section, we provide technical details on field identification and codebook creation. We also describe our procedures for measuring distance between fields and visualizing cultural holes as chasms on a citation map.

Field Identification

We studied 60 large scientific fields in the JSTOR citation network. This network includes more than 1.5 million interconnected articles. We identified fields using the map equation (Rosvall and Bergstrom, 2008), an algorithm for extracting the community structure of complex networks (Fortunato, 2010; Lancichinetti and Fortunato, 2009).

The map equation has been used extensively to extract scientific fields in citation networks, e.g., (Rosvall and Bergstrom, 2008). In our case, the map equation tracks scholarly citation flow in the JSTOR corpus. It partitions the articles into fields to minimize the description of an idealized researcher who navigates from article by article by following citations at random. The fields are the regularities that best compress the citation flow; in this sense, they are optimal. In practice, a field corresponds to a set of articles among which the idealized researcher would spend a long time before transitioning to another field.[10]

[8] To our knowledge, the closest alternative is (Livne et al., 2011), which uses the symmetrized Kullback–Leibler divergence to measure the semantic similarity of Twitter-using politicians. Although this shares our information-theoretic orientation, in our context, the cross entropy is easier to interpret

[9] To make this more plausible, consider the difficulty of correctly interpreting conventional gestures outside of their cultural context—Google "peace sign in Australia."

[10] Because citation networks are time directed, however, this idealized random walk approach can cause older articles to accumulate flow disproportionately. To resolve this, we used the *undirected* version of the citation network to infer the stationary distribution of the random walk (removing the time–directionality and the consequent problem of citation sinks). We then evaluated the quality of proposed partitions using the directed network.

Selecting the Sample and Naming Fields

To assure that conclusions about the culture of a given scientific field were well–founded, we first eliminated from the sample any field with fewer than 1,000 papers. This assures that the field in question is well–covered by JSTOR. We further restricted our sample to papers published between 1990 and 2010, to focus our analysis on contemporary science; any fields with fewer than 500 papers within this 20-year window were discarded. These sampling steps resulted in 78 fields. The logic behind our sampling procedure is straightforward: we wanted to make sure that we had enough papers from a field to construct reasonable proxies of its contemporary scientific culture. Note that scholarly fields vary in their representation in JSTOR. This heterogeneous coverage is responsible for the absence of many prominent scientific disciplines, such as physics, from our analysis.

Field names were then chosen manually, after identifying the phrases that best distinguish each cluster by measuring the mutual information between phrase i and cluster j (Manning et al., 2008). A list of the "most distinguishing phrases" for each cluster is available as supplementary material.

Because of the computational costs of calculating cultural holes, we further reduced the size of the sample from 78 to 60. These 60 fields were chosen to maintain balance across the major domains of scholarship in JSTOR. In particular, we retained every field in statistics and molecular biology. In ecology and evolution, we selected fields across subdomains to maximize the diversity of our sample.[11]

Codebooks

The phrase frequency distribution P_i for each scholarly field (its "codebook") was assembled using the empirical frequency of each triplet of consecutive words (trigram) in a random sample of 500 articles in that field, published between 1990 and 2010. We chose 500 because it was the largest number of articles that we could sample from every field in the study period; some fields had just over 500 articles from 1990 to 2010, making larger samples impossible. Samples of approximately 100 articles yielded a consistent field-level entropy, so we are confident that a larger 500–article sample is sufficient for our analysis.

In computational linguistics, a statistical language model assigns a probability to a sequence of m words $P(w_1, \ldots, w_m)$ by means of a probability distribution. Language models built on trigrams are especially reliable; they capture substantial complexity (syntactic structure and conditional probabilities between words) while being easy to implement. In many cases, more complicated language models reveal little additional information (Jelinek, 1991). We removed some function and stop words from the trigrams to decrease the overall number of terms in the database.[12] This procedure yields an augmented trigram language model, including bigrams and single words when flanked with stop words.[13]

To apply our information-theoretic model of scholarly communication (see "Using information theory to model scholarly culture and communication") to real data, we need the codebooks for field i and j to contain the same phrases (domain[P_i]=domain[P_j]) so that a concept from field i can always be expressed in the codebook of field j. To ensure this, we introduce a teleportation parameter α, which merges a discipline codebook P_i with the "corpus codebook" S generated from the articles of every field:

$$\begin{aligned} p_i^s(x) &= (1-\alpha)p_i(x) + \alpha s(x) \\ p_j^s(x) &= (1-\alpha)p_j(x) + \alpha s(x), \end{aligned}$$

where the lowercase p and s indicate the probability distribution function over phrases for the appropriate codebooks. Intuitively, this means that with some small probability α, the writer

[11] As discussed in footnote 16, it is unlikely that corpus representation of fields and their neighbors substantially affects our analysis. Missing fields would have to deviate substantially and systematically from what we observe to produce qualitative changes in our results. We are thus confident that our results are robust for scholarly domains and fields that are well–represented in JSTOR.

[12] Deleted words include *a*, *another*, *behind*, *no*, *none*, *something*, *such*, *than*, *that*, *wherever*, *will*. A complete list of these removed words, and more detailed methodology, are available in the supplementary material.

[13] Using an augmented trigram model will, in principle, improve our ability to detect language differences across fields when some fields use trigram phrases (e.g., *green fluorescent protein*) more frequently.

(or reader) consults not the discipline codebook P_i but the corpus codebook S. The value of this parameter does not change our results, so we chose the intuitive value of 1 percent.

Measuring Distance in the Citation Network

We measure distance in the citation network between fields i and j as follows. For randomly selected pairs of articles, one in field i and one in field j, we calculate the number of links in the shortest path between them. This topological distance is computed using the undirected version of the citation network. The average value of this quantity is an estimate of the average length of the shortest path between these two fields. This estimated average path length provides our distance measure.

Visualizing Cultural Holes

To construct the topographical map in Figure 3, we first embed the 60 fields in two dimensions using principal coordinates analysis (PCoA), that is, multidimensional scaling, as implemented in R (Borg and Groenen, 2005). We define the elements of the dissimilarity matrix using the average shortest path distance between two fields in the citation network. We refer to the xy coordinates of field i as F_x^i and F_y^i. Cultural canyons were calculated as the distance-weighted sum of pairwise, symmetrized cultural holes $\tilde{C}_{ij} = (C_{ij} + C_{ji})/2$.

The underlying logic is as follows. The height of each pixel in the map should be more affected by nearby fields. Thus we define a weight vector $\overrightarrow{w_{P_{xy}}}$ for pixel P_{xy}, at position x and position y. The components of this vector determine how much a given field i contributes to the pixel height at position x, y:

$$w^i_{P_{xy}} = \left(\sqrt{(x - F_x^i)^2 + (y - F_y^i)^2}\right)^{-\kappa},$$

where κ is a parameter controlling how rapidly the influence of a given field falls off with distance. For Figure 3, we used $\kappa = 1$, so the weight is just the inverse distance. Finally, we exclude the nearest field ($\max(\overrightarrow{w_{P_{xy}}})$) from calculations of the pixel height; inclusion of the nearest field leads to "defects" in the topographical map centered on each field (the sum is dominated by the self-hole $\tilde{C}_{ii} = 0$). The height of the pixel P_{xy} is then given by

$$P_{xy} = \sum_{i \neq \max(\overrightarrow{w_{P_{xy}}})} \sum_{j \neq \max(\overrightarrow{w_{P_{xy}}}), j \neq i} w^i_{P_{xy}} w^j_{P_{xy}} \tilde{C}_{ij}.$$

Results: Analyzing Scholarly Fields in JSTOR

We now turn to the analysis of our JSTOR corpus. We find systematic patterns in the distribution of cultural holes across the major domains of science cataloged in JSTOR. Biological sciences (including ecology, evolutionary, and molecular biology) are surrounded by deeper cultural holes on average (i.e., higher values of C_i) than behavioral and social sciences, such as psychology, economics, sociology, political science, business, religious studies, and education (T-test, $P < 10^{-12}$; see Figure 1A).

Furthermore, the social sciences are more accessible to the biological sciences than vice versa. If i is a social science field and j is a biological science field, then $C_{ij} < C_{ji}$ on average, that is, the cultural hole encountered by a biological scientist reading a social science paper tends to be shallower than for a social scientist reading a biological science paper ($C_{ij} < C_{ji}$ for 893 of 899 pairs; significant, see supplemental material for details of statistical test).

Note that this pattern does not follow trivially from the number of distinct phrases or technical terms (Fig. 1B). The number of distinct phrases is field- rather than domain-specific, with some fields from each domain having few terms and some having many. We find, however, that the *ratio* of distinct three-word phrases to distinct words does vary systematically by domain. Social science fields tend to use fewer words in more combinations, generating many distinct phrases from their stock of words. The ratio is small for biological science fields, suggesting that the constraints on word combination are stronger there and that fields with many distinct phrases have many distinct words (Fig. 1C). This domain-specific asymmetry may provide a partial explanation for the domain-level asymmetry in cultural holes. Distinct phrases from the biological sci-

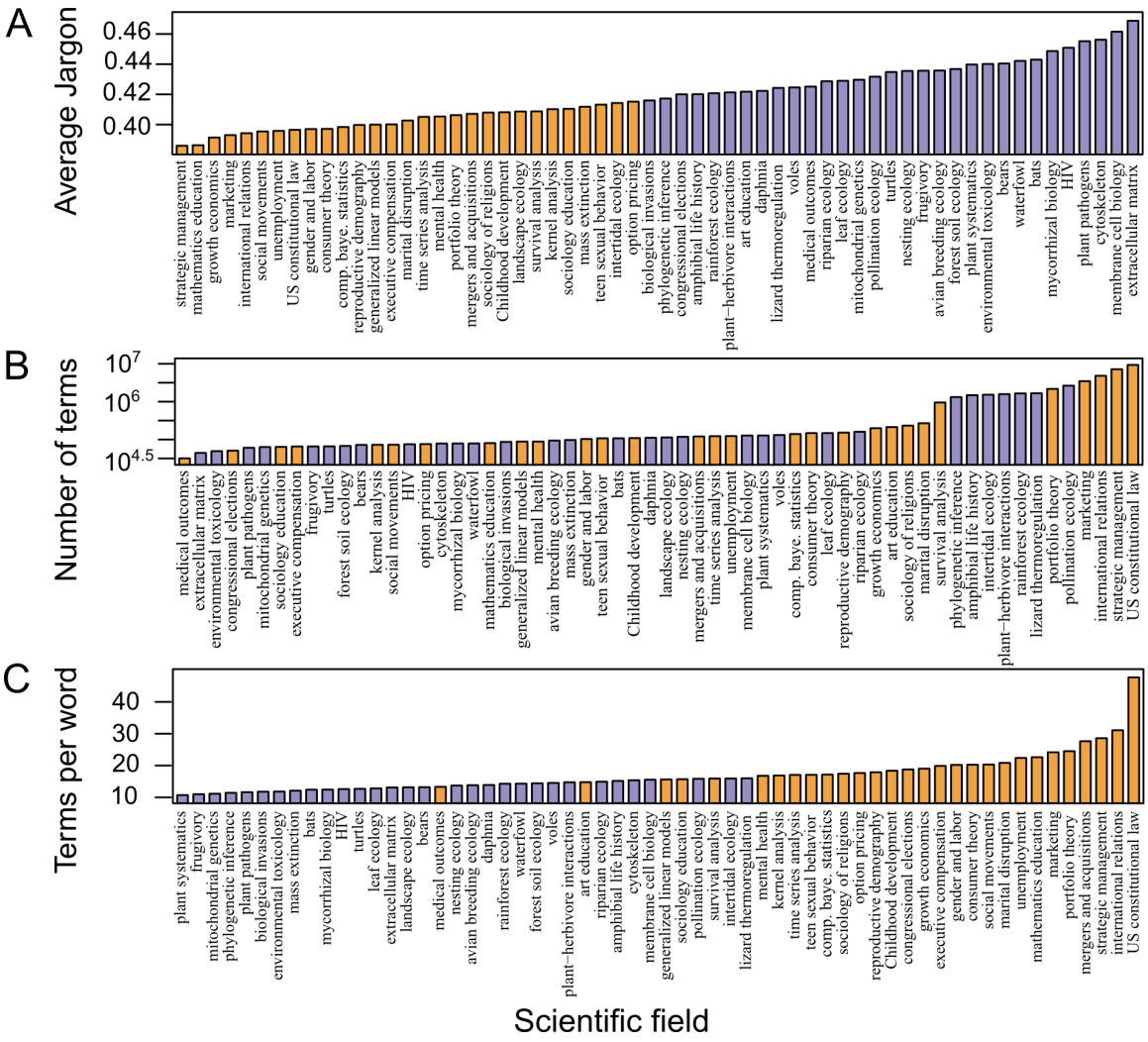


Figure 1: **(A)** Average cultural hole $C_i = \sum_j C_{ij}/N$ across fields. Note that biological sciences are surrounded by systematically larger cultural holes than social sciences. **(B)** Total number of phrases (distinct three word combinations/trigrams) in each codebook. Note that there is no pattern at the domain level, that is, biological sciences (blue) and behavioral/social sciences (orange) do not vary systematically in the number of phrases. **(C)** Number of phrases per distinct word in each field. Note the systematic difference between biological sciences (blue) and behavioral sciences (orange). This pattern suggests that social science fields use fewer words in more combinations, whereas word combination in the biological sciences is more constrained.

ences are likely to contain many words poorly represented in the social science codebooks.

Second, we find that the structure of scientific communication as traced by citations is *related to but distinct from* the structure traced by cultural or semantic difference. In other words, structural and cultural holes do not precisely align. Hierarchical clustering analysis (UPGMA) of the average shortest citation path distances between fields yields a dendrogram that splits the biological from the social sciences at the domain level (Fig. 2A) (Sokal, 1958); see "Data and Methods" for the estimation of citation path distance. When these fields are clustered by cultural hole, however, the dendrogram does not reflect these major domains (Fig.2B). We perform this clustering using a symmetrized version of the cultural hole measure: $\tilde{C}_{ij} = (C_{ij} + C_{ji})/2$. Now, the subfields broadly corresponding to ecology and evolution are separated from the molecular fields (Figs. 2A and 2B, molecular fields shown in red). This separation reveals a deep cultural hole within biology that cuts across the flow of the citation network. In economics, a smaller cultural hole is revealed: growth economics and consumer theory are clustered with portfolio theory, option pricing, and time series analysis by citation (Fig. 2A, shown in blue). When clustered by the cultural hole measure, however, they group with subfields having to do with labor (Fig. 2B, also blue).

These exploratory findings suggest the need to combine structural information (from citations) and cultural information (from full-text) systematically. To visualize these cultural holes as they cut across the traditional citation-based map of science, we embed scholarly fields in a two-dimensional topography defined by citation and jargon (Fig. 3). The relative location of fields is determined by the citation flows between them. Fields with substantial citation flow are placed close together, and those with little flow are placed far apart. The topographical features of this landscape, in turn, are determined by jargon.

More specifically, the x and y coordinates of each field are assigned via principal coordinate analysis of the citation distances. We measure dissimilarity between fields using the average shortest citation path between them (goodness of fit = 0.25) (Borg and Groenen, 2005). The depth of each pixel in the topographic overlay is a weighted average of the symmetrized cultural holes between nearby fields; see "Data and Methods" for details of map construction, including the dissimilarity matrix for PCoA and the weighting function. Deep chasms represent significant cultural holes. Researchers in fields separated by such holes must invest substantial resources to translate their neighbors' articles and incorporate them into work within their field. Wading into another literature with substantial jargon will literally take the reader under water.

This visualization exposes several key features of scientific communication. First, maps of science based purely on the structure of citations are missing a large part of the story—just like maps of society based purely on social ties. The numerous cultural holes crosscutting this landscape make it clear that the efficient flow of information assumed by classical citation analysis is often impeded by jargon. Second, the large-scale structure of the map makes sense. The social sciences cluster together on the left, the biological sciences on the right. At least in JSTOR, statistics sits between the social and the biological sciences, reflecting its role as a common resource. Interestingly, the cultural hole between statistics and the social sciences is shallower than between statistics and the biological sciences. Within the social sciences, there is a relatively clear path, with small cultural holes, from education to psychology to sociology to economics and business. This reflects the relative coherence of the social sciences in terms of jargon and, by extension, matters of common concern. In the biological sciences, by contrast, many modest cultural holes separate clusters of fields, with a more substantial chasm between the ecological sciences (bottom right) and genetics, phylogenetics, and systematics (middle right). Molecular biology fields (upper right) are far from the rest of biology in both citation distance and jargon, with massive cultural holes cutting molecular biology off from nearly every discipline in JSTOR.

Visual inspection of our topographical map suggests that the landscape is more rugged in the biological sciences than in the social sciences. The biological sciences are more balkanized by jargon and hence have more differentiated local cultures, which are reflected by the terms of interest from their articles. To make this intuition precise, we exploit the fact that the cultural hole between

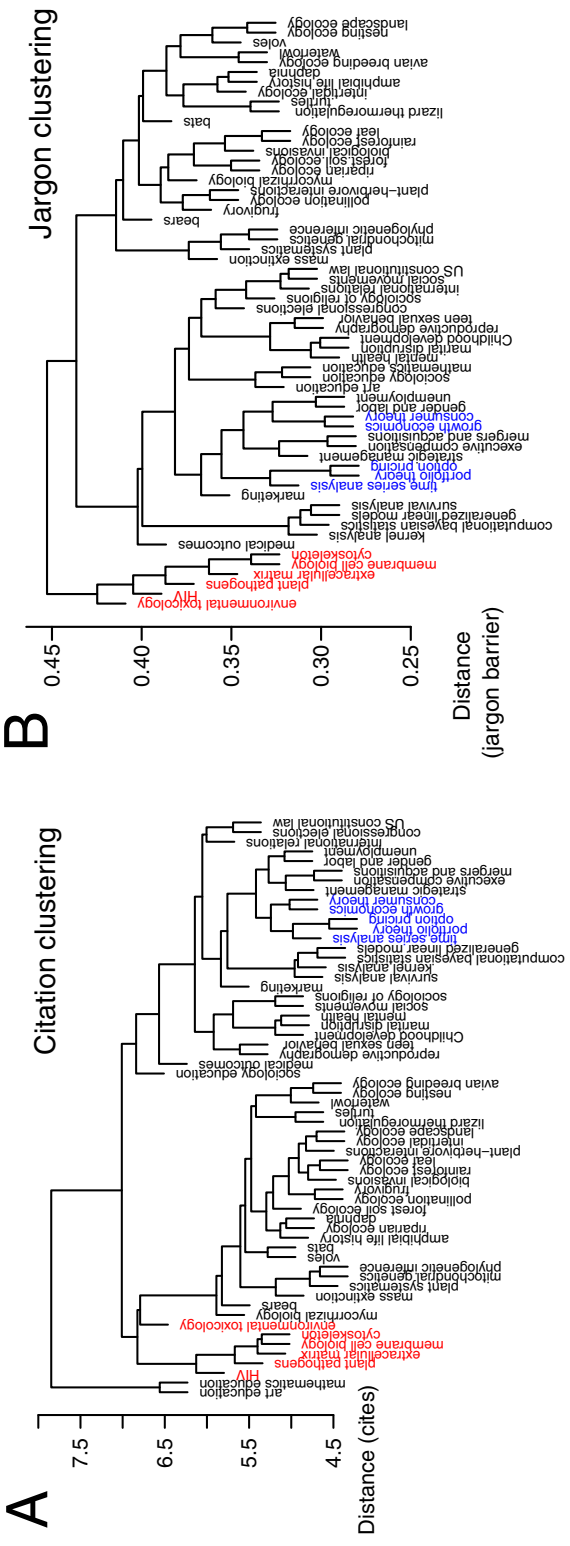


Figure 2: (A) Clustering the fields by shortest citation path. Note that the biological sciences cluster together on the left and the social and behavioral sciences cluster together on the right. (B) The biological sciences do not cluster together by symmetrized cultural hole $\tilde{C}_{ij}$. The molecular and cell biology subfields (red) are distant from the other biological science fields and the social science fields by jargon. Similarly, economic fields clustered by citation (blue) are grouped with other fields by jargon. Growth economics and consumer theory cluster with portfolio theory, option pricing, and time–series analysis by citation but with labor related fields (gender and labor; unemployment) by jargon. All clustering was performed using UPGMA, but the result holds for different clustering methods.

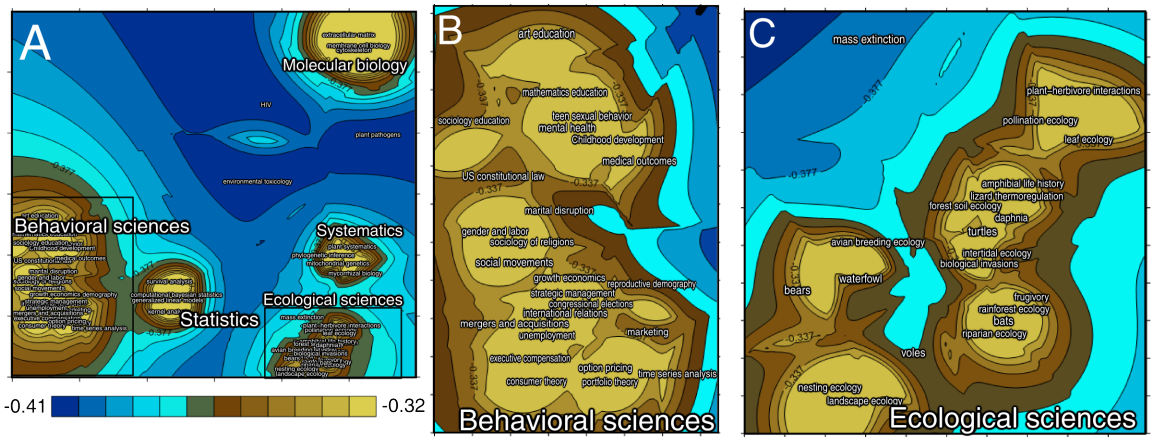


Figure 3: Topographical map of science combining textual and citation data. Fields that are close communicate frequently: positions in space are calculated by applying principal coordinate analysis to the matrix of shortest average citation paths and retaining the first and second principal coordinates. In this contour map, "oceans" (green shading to blue) represent the (negative of the) distance-weighted sum of symmetrized cultural holes between fields, $\tilde{C}_{ij}$; see "Data and Methods" for weighting function. Despite substantial citation flow between fields separated by these holes (e.g., survival analysis and medical outcomes), communication is inefficient. Note that social sciences cluster together on the left–hand side and biological sciences on the right–hand side, with statistics located between. Note also that molecular biology fields (upper right) are separated from other fields, including other biological sciences, by huge cultural holes, and that cultural holes sensibly separate the remaining fields (smaller panels). In the small panels, labels have been shifted slightly to reduce overlap and increase legibility.

field i and field j grows with the average shortest citation path between them. Equivalently, the efficiency $E_{ij} = 1 - C_{ij}$ decays with citation distance. To control for possible differences in citation practices between fields or domains, we normalize our measure of citation distance. Specifically, we divide the average shortest path from an article in field i to one in field j by the average shortest path between two articles in field i. We then model the decay in communication efficiency with citation distance as

$$E_{ij} = 1 - \beta(1 - e^{-\gamma d_{ij}}), \qquad (4)$$

where d_{ij} is the normalized citation distance, $1-\beta$ gives the asymptotic efficiency for fields at infinite distance, and γ controls the decay of E_{ij} with distance. Fits were obtained via nonlinear least squares in R. Figure 4A shows the field with the slowest decay (option pricing) and the field with the fastest decay (environmental toxicology). These examples suggest the conditions under which slow and fast decay occur. Fields with slow decay have substantial citation flow to several neighboring fields and relatively efficient communication with those fields, that is, small cultural holes. Fields with fast decay, by contrast, may have several fields at similar proximity but much less efficient communication with these fields, that is, deep semantic chasms. The decay rate γ varies across fields (Fig. 4B).

This analysis suggests that the topographical features observed in Figure 3 are not an artifact of embedding the high-dimensional citation network in two dimensions. We find that the decay rate is higher in the biological sciences than the social sciences (Fisher's exact test: top half vs. bottom half, $P < 0.0001$); in other words, the landscape of cultural holes in the biological sciences is indeed more rugged. There are a number of exceptions to this pattern, however; most surprisingly, several fields related to molecular biology have relatively small values of γ, so that efficiency of communication falls off slowly with citation distance. These counterexamples further illustrate that decay rate is not systematically related to the overall amount of jargon in a specific

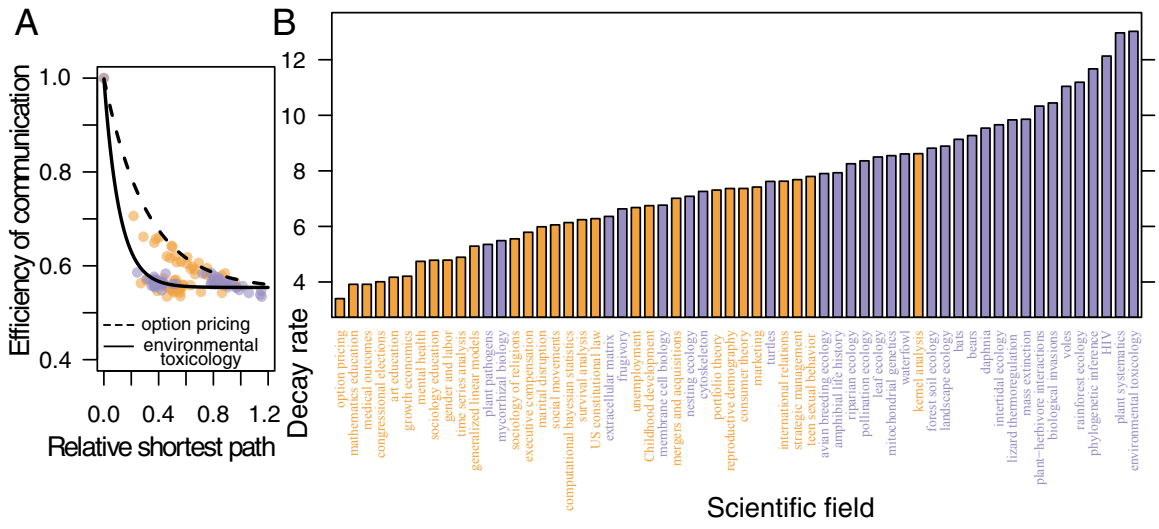


Figure 4: **(A)** Efficiency of communication from focal field i to target field j ($E_{ij} = 1 - C_{ij}$) decays with the distance to field j differently for different fields. Here we show the slowest decay (option pricing, dashed line) and the fastest decay (environmental toxicology, solid line) out of the 60 fields (see supporting information for others). **(B)** Decay rate γ plotted for behavioral science fields (orange) and biological science fields (blue). Focal fields with fast decay tend to have few fields nearby with which communication is efficient. Those with slow decay have several neighbors with relatively small cultural holes, that is, efficient communication. Distance is computed using the normalized shortest path: the average shortest path from a paper in field i to a paper in field j divided by the average shortest path from a paper in field i to another paper in field i. We subtract 1 from this value so that the normalized shortest path from a field to itself is 0. This normalization allows us to account for differences in citation norms that cause focal fields to be tightly or loosely connected, that is, to have a short or long average path distance within field.

field and hence the depth of the average cultural hole around it. Although the molecular biology fields have the most jargon in JSTOR (see Fig. 1A), decay rates for several of these fields are comparatively low, while two are very high (HIV and environmental toxicology).[14]

Discussion

Our results suggest that combining the structural analysis of citation flows with explicit models of communication processes (e.g., jargon-induced cultural holes) exposes important features of scholarly communication. We further suggest that the interaction of the structural and cultural dimensions of scholarly communication reveals previously neglected social processes. For example, we argue that the decay of communicative efficiency with citation distance reflects the relative insularity of scholarly cultures or fields. Fields with faster decay rates are less accessible to others close by. Scholars working in these fields make extensive use of jargon not shared with neighboring fields, creating cultural holes that make inter-field communication less efficient. Readers from neighboring fields are sufficiently aware of this *nearby knowledge* to reference it but can only understand it through potentially prohibitive study and decoding. By contrast, scholars in fields with slow decay rates likely use jargon that is shared with neighboring fields—their probability distribution over phrases is similar—making their work much more accessible through phrases of common concern. The absence of cultural holes

[14] Similarly, the decay rate is not systematically related to number of distinct terms (Fig. 1B) for example, HIV and plant pathogens have a similar number of distinct terms (and similar average cultural holes) but wildly different *decay rates*.

makes communication between these scientific cultures much easier.

This varying relationship between shortest citation paths and communicative efficiency helps us interpret between-field differences in the average depth of cultural holes, C_i. Molecular and cell biology fields have the largest cultural hole measures C_i in JSTOR (see Fig. 1A). The decay rate of some of these fields[15] is quite slow, however, in comparison with other fields in the social and biological sciences. This surprising combination suggests that molecular and cell biology does not have substantial jargon—and, consequently, deep cultural holes—because it is insular or exclusionary, per se. Slow decay rates exclude this interpretation at the field level. Indeed, the three cell biology fields cluster close together in citation distance and have relatively shallow cultural holes between them, reflecting a substantial overlap in matters of interest. Instead, molecular biology fields have high jargon simply because they are remote from most other fields in JSTOR, both in citation distance and semantic distance, that is, matters of concern. By contrast, vole research is close to *many* fields by citation (in the ecology cluster), but its communicative efficiency decays quickly as a function of citation distance. Voles are small, mouselike rodents, and while vole research has only a moderate amount of jargon overall, it overlaps little in matters of interest with its neighbors (e.g., bat and bear research). Vole research is thus surrounded by a moat of jargon and is much more insular than many molecular biology fields. These cases illustrate that decay rate can be used to assess the relative insularity of different fields, taking into account the distance between fields in the citation network as measured by shortest path.[16]

These examples suggest an interesting interpretation of the continuities and discontinuities between fields of science and scholarship in JSTOR. First, JSTOR fields fall into four great camps: the social sciences; the ecological and evolutionary branches of biology; the molecular and cellular branches of biology; and statistics. Concentrating on the substantive fields (social, ecological, and molecular), we note that each of these clusters, tied closely together by citation flow, is also united by concern with a particular scale (macroscopic and human; macroscopic and nonhuman; microscopic). The social sciences tend to have more shallow cultural holes, on average, and can communicate quite efficiently with neighboring social sciences. This suggests that these fields are not absorbed by particularities but instead share many matters of concern, remaining relatively integrated with one another. The ecological sciences, by contrast, have deeper cultural holes and communicate inefficiently with neighboring ecological fields. The example of vole research suggests a broader principle: these fields are absorbed in many particularities that they *do not share* with one another (I am concerned with voles; you with bears; she with bats).[17] Finally, the molecular biological sciences are surrounded by deep cultural holes, but many communicate *efficiently* with their immediate neighbors. This reflects an orientation toward many *shared* particularities: the molecules and processes that form the physical substrate of life. It is especially interesting that the ecological sciences—popularly associated with holistic, anti–reductionist thinking—are so balkanized in their matters of concern, while the molecular sciences, despite dividing life into so many distinct building blocks, nevertheless seem more integrated at the semantic level. A deeper explanation of these patterns is beyond the scope of this paper, but note that none of this would be apparent from a pure citation or semantic analysis alone. Note also that all aspects of our formalism are portable to any other type of data involving structural relationships and cultures revealed in language or signs.

Conclusion

Information theory provides a simple but powerful framework with which to model communi-

[15] Research on plant pathogens, the extracellular matrix, the cytoskeleton, and membrane cell biology.

[16] It is unlikely that these patterns in γ are an artifact of corpus representation of the focal field and its neighbors. Social science fields and ecological science fields both have many close neighbors in the corpus yet significantly different typical values of γ. It is possible that the relatively slow decay rates for molecular and cell biology fields could be affected by the undersampling of potential "middle-distance" neighbors, but the efficiency of communication to these missing neighbors would have to differ radically from sampled fields to shift the estimate of γ substantially.

[17] It is possible that these cultural holes may be somewhat attenuated by analogical, thesaurus-like mappings (e.g., voles and bats are both small mammals), but these mappings are likely to be limited in scope.

cation and measure cultural holes. We demonstrated its utility through the analysis of scientific communication, a central project in the science of science (de Solla Price, 1965). Qualitative studies have considered content of communication in conjunction with citations (Ceccarelli, 2001), but the vast majority of quantitative analyses rely exclusively on citations to map the structure and flow of scientific communication (Rosvall and Bergstrom, 2008). In cases where content is addressed directly, it is viewed as a substitute for citation analysis (Landauer et al., 2004; Gerrish and Blei, 2010), or citation patterns are used to construct a measure of semantic similarity (Moody and Light, 2006), or citations are treated as another type of content (Erosheva et al., 2004). Livne et al. (2011) is an important exception, although focused on a distinct literature, that is, social media and politics.

In this paper, we have demonstrated that scholarly semantic information is not a substitute for citation analysis. Rather, the two sources of information are complementary, together revealing patterns that otherwise remain invisible. We introduced an easily understood measure of cultural and semantic distance, grounded in a simple model of communication. We then showed that the social and biological sciences differ systematically in their use of jargon and the patterning of associated cultural holes. We demonstrated that science takes on a different shape when viewed through citation or content alone and introduced two procedures for combining such information: first, an attractive and easily interpretable "topographical map" of science, in which citation structure establishes location and jargon establishes topography; and second, a rigorous procedure for capturing how communication efficiency scales with citation distance. Using these two procedures, we made several surprising discoveries about the structure of the scholarly fields in JSTOR: the coherence of the social sciences; the balkanization of ecological sciences by jargon; and the surprising coherence of molecular and cell biology in the face of its isolation from the other fields in our sample. Moreover, though we pilot this approach by analyzing the structure and culture of scientific communication, we believe that it can be straightforwardly applied to any system involving a network of structural relationships and a distribution of cultural symbols, signals, or phrases.

Our analysis of scientific communication has marked limitations. Sampled trigrams (three-word phrases) map imperfectly to the phrases of most interest to scientists and scholars. Currently we do not distinguish different types of phrases or mark out those of special scientific importance. Note, however, that using a more sophisticated language model, for example, one taking syntactic information into account, will in general refine but not substantially alter our cultural hole measurements and subsequent conclusions. Under our core assumption that frequently–encountered linguistic units are easy to understand and infrequently–encountered ones are hard to understand, the primary contribution to cultural holes between fields will still come from phrases that are used frequently in one field and rarely in another (*capital asset pricing*, *Gibbs sampler*, *microtine cycles*, *vesicular stomatitis virus*). A more sophisticated language model might avoid splitting phrases that are longer than three words, for example, *Markov Chain Monte Carlo*, but this is essentially a difference in accounting. Such a phrase will still contribute under a trigram model, and because cultural holes are computed as ratios, the actual values are unlikely to change significantly. Likewise, a more sophisticated language model might infer deeper cultural holes between fields in which the same phrase is used robustly in different grammatical roles, but such situations are hard to imagine. When the same phrase is used in different semantic contexts, it might reasonably count as distinct and therefore deepen the cultural hole, but this argument suggests that our measure is in general a lower bound ripe for subsequent refinement.[18] Thus we are confident that our augmented trigram model provides a robust estimate of the cultural holes between two fields, given our general model of culture and communication.

More importantly, we cannot currently say why jargon is adopted: when it was introduced to maximize communicative efficiency within a

[18] In principle, we might underestimate the cultural holes between social science fields because we are failing to capture substantial differences in the context in which shared phrases are used. Such differences are unlikely in immediate neighbors, however, and thus unlikely to contribute to unmeasured balkanization in the social sciences.

field without regard to outside audiences (Kemp and Regier, 2012; Fawcett and Higginson, 2012; Zipf, 1935, 1949), and when it was used to distinguish the field and excavate a cultural hole that limits oversight, association and loss of status (Bourdieu and Thompson, 1991; Coleman, 1985; Holmes and Meyerhoff, 1999). Nevertheless, we believe that our investigation sheds substantial new light on jargon's distribution and its consequences for science. Future work that tags phrases and citations with temporality, authorship, institution, and semantic class (e.g., methods) may begin to disentangle the origins of jargon, the role of cultural holes in sustaining structural holes, and the circumstances in which actors can fill in cultural holes through intercultural communication. Similarly, research that identifies and interrogates the *emergence* of new symbols, signs, or phrases may be able to identify when efficiency or distinction is a primary motive for the creation of cultural holes.

More broadly, our findings point to a new research program in the analysis of culture and the science of science. On the empirical front, scholars now have a lean machinery to identify cultural holes and assess the efficiency of communication between communities as well as their relative insularity. On the methodological front, our framework can be extended to deal with complex syntactical rules and richer models of communication. Catalogs of technical terms can be expanded and structured to include term redundancy, hierarchy, and syntactic structure. Moreover, these lists can be broadened to include signals and symbols beyond written language. Communication rules can also be altered to capture the real social and cognitive processes through which scholars read and assess documents and people in other cultures evaluate messages of all types.[19]

Though our analysis highlights major patterns in JSTOR and science at large, we have barely scratched the surface of this landscape of possibility. Cultural holes constantly evolve as collaborative dynamics change. What happens to field-specific jargon when two cultures merge? Do changes in citation drive changes in jargon, or vice versa? Do jargon and other semiotic markers follow standard evolutionary birth–death dynamics, producing the culture-level patterns we observe—or are they subject to additional social processes? Our results underline both the exploding opportunities in the large-scale *structural* analysis of culture (Bail, 2014) and the importance of building and interconnecting new models and sources of information as we seek to quantify, understand, and shape behavior in science (Evans and Foster, 2011).

[19] On the theoretical front, we also note that our operationalization of Pachucki and Breiger's concept of cultural holes suggests intriguing connections between information theory, communication theory, and cultural sociology as well as the ethnomethodological perspective.

References

Bail, Christopher A. 2014. "The Cultural Environment: Measuring Culture with Big Data." *Theory and Society* 43. http://link.springer.com/article/10.1007/s11186-014-9216-5/fulltext.html.

Bearman, Peter and Paolo Parigi. 2004. "Cloning Headless Frogs and Other Important Matters: Conversation Topics and Network Structure." *Social Forces* 83:535–57. http://dx.doi.org/10.1353/sof.2005.0001.

Bernstein, Basil. 1964. "Elaborated and Restricted Codes: Their Social Origins and Some Consequences." *American Anthropologist* 66:55–69. http://dx.doi.org/10.1525/aa.1964.66.suppl_3.02a00030.

Bischof, Nicole and Martin J. Eppler. 2010. "Clarity in Knowledge Communication." In *Proceedings of the Tenth International Knowledge Management Conference IKnow*, volume 10, pp. 162–174. Verlag der Technischen Universität.

Borg, Ingwer and Patrick J. F. Groenen. 2005. *Modern Multidimensional Scaling: Theory and Applications*. New York: Springer.

Bourdieu, Pierre and John B. Thompson. 1991. *Language and Symbolic Power*. Cambridge MA: Harvard University Press.

Boyack, Kevin W., Richard Klavans, and Katy Börner. 2005. "Mapping the Backbone of Science." *Scientometrics* 64:351–74. http://dx.doi.org/10.1007/s11192--005--0255--6.

Burt, Ronald S. 1992. *Structural Holes: The Social Structure of Competition*. Cambridge MA: Harvard University Press.

Ceccarelli, Leah. 2001. *Shaping Science with Rhetoric: The Cases of Dobzhansky, Schrodinger, and Wilson*. Chicago: University of Chicago Press. http://dx.doi.org/10.7208/chicago/9780226099088.001.0001.

Coleman, Hywel. 1985. "Talking Shop: An Overview of Language and Work." *International Journal of the Sociology of Language* 1985:105–30. http://dx.doi.org/10.1515/ijsl.1985.51.105.

Cover, Thomas M. and Joy A. Thomas. 2006. *Elements of Information Theory*. Hoboken: Wiley-Interscience.

de Condillac, E. B. 1782. *Cours d'étude pour l'instruction du Prince de Parme*. Paris: Houel.

de Solla Price, D. J. 1965. "Networks of Scientific Papers." *Science* 149:510–15. http://dx.doi.org/10.1126/science.149.3683.510.

Erickson, Bonnie H. 1996. "Culture, Class, and Connections." *American Journal of Sociology* 102:217–51. http://dx.doi.org/10.1086/230912.

Erosheva, Elena, Stephen Fienberg, and John Lafferty. 2004. "Mixed-Membership Models of Scientific Publications." *Proceedings of the National Academy of Sciences* 101:5220–27. http://dx.doi.org/10.1073/pnas.0307760101.

Evans, James A. and Jacob G. Foster. 2011. "Metaknowledge." *Science* 331:721–25. http://dx.doi.org/10.1126/science.1201765.

Fawcett, Tim W. and Andrew D. Higginson. 2012. "Heavy Use of Equations Impedes Communication among Biologists." *Proceedings of the National Academy of Sciences* 109:11735–39. http://dx.doi.org/10.1073/pnas.1205259109.

Feldman, Robin. 2008. "Plain Language Patents." SSRN Scholarly Paper ID 1731651, Social Science Research Network, Rochester, NY.

Fortunato, Santo. 2010. "Community Detection in Graphs." *Physics Reports* 486:75–174. http://dx.doi.org/10.1016/j.physrep.2009.11.002.

Friedland, Roger. 2009. "The Endless Fields of Pierre Bourdieu." *Organization* 16:887–917.

Garfinkel, Harold. 1991. *Studies in Ethnomethodology*. Hoboken NJ: John Wiley.

Gerrish, Sean and David M. Blei. 2010. "A Language-Based Approach to Measuring Scholarly Impact." In *Proceedings of the 26th International Conference on Machine Learning, June 21–24*, pp. 375–382.

Han, Shin-Kap. 2003. "Unraveling the Brow: What and How of Choice in Musical Preference." *Sociological Perspectives* 46:435–459. http://dx.doi.org/10.1525/sop.2003.46.4.435.

Holmes, Janet and Miriam Meyerhoff. 1999. "The Community of Practice: Theories and Methodologies in Language and Gender Research." *Language in Society* 28:173–83. http://dx.doi.org/10.1017/S004740459900202X.

Homans, George Caspar. 1961. *Social Behavior: Its Elementary Forms*. New York: Harcourt, Brace and World, Inc.

Jelinek, Fred. 1991. "Up from Trigrams." In *Proceedings of Second European Conference on Speech Communication and Technology, EUROSPEECH*, volume 91, pp. 1037–40. Genova, Italy: September 24–26.

Kemp, Charles and Terry Regier. 2012. "Kinship Categories across Languages Reflect General Communicative Principles." *Science* 336:1049–54. http://dx.doi.org/10.1126/science.1218811.

Knorr-Cetina, K. 1999. *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, MA: Harvard University Press.

Kullback, Solomon and Richard A Leibler. 1951. "On Information and Sufficiency." *Annals of Mathematical Statistics* 22:79–86. http://dx.doi.org/10.1214/aoms/1177729694.

Lancichinetti, Andrea and Santo Fortunato. 2009. "Community Detection Algorithms: A Comparative Analysis." *Physical Review*

E 80:056117. http://dx.doi.org/10.1103/PhysRevE.80.056117.

Landauer, Thomas K., Darrell Laham, and Marcia Derr. 2004. "From Paragraph to Graph: Latent Semantic Analysis for Information Visualization." *Proceedings of the National Academy of Sciences of the United States of America* 101:5214–19. http://dx.doi.org/10.1073/pnas.0400341101.

Leydesdorff, Loet and Ismael Rafols. 2008. "A Global Map of Science Based on the ISI Subject Categories." *Journal of the American Society for Information Science and Technology* 60:348–62. http://dx.doi.org/10.1002/asi.20967.

Livne, Avishay, Matthew P. Simmons, Eytan Adar, and Lada A. Adamic. 2011. "The Party Is Over Here: Structure and Content in the 2010 Election." In *Proceedings of 5th International AAAI Conference on Weblogs and Social Media*. Barcelona, Spain.

Lizardo, Omar. 2006. "How Cultural Tastes Shape Personal Networks." *American Sociological Review* 71:778–807. http://dx.doi.org/10.1177/000312240607100504.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*, volume 1. Cambridge: Cambridge University Press.

Moody, James and Ryan Light. 2006. "A View from Above: The Evolving Sociological Landscape." *American Sociologist* 37:67–86. http://dx.doi.org/10.1007/s12108--006--1006--8.

Pachucki, Mark A. and Ronald L. Breiger. 2010. "Cultural Holes: Beyond Relationality in Social Networks and Culture." *Annual Review of Sociology* 36:205–24. http://dx.doi.org/10.1146/annurev.soc.012809.102615.

Reach, G. 2009. "Linguistic Barriers in Diabetes Care." *Diabetologia* 52:1461–63. http://dx.doi.org/10.1007/s00125--009--1404--x.

Richardson, Matthew L. 2010. "Publishing Scientific Outreach Materials in Educational and Social Science Journals." *American Entomologist* 56:11–13.

Rosvall, Martin and Carl T. Bergstrom. 2008. "Maps of Random Walks on Complex Networks Reveal Community Structure." *Proceedings of the National Academy of Sciences* 105:1118–23. http://dx.doi.org/10.1073/pnas.0706851105.

Shannon, Claude E. 1948. "The Mathematical Theory of Communication." *The Bell Systems Technical Journal* 27:379–423, 623–56. http://dx.doi.org/10.1002/j.1538--7305.1948.tb01338.x.

Small, Henry. 1999. "Visualizing Science by Citation Mapping." *Journal of the American Society for Information Science and Technology* 50:799–813. http://dx.doi.org/10.1002/(SICI)1097--4571(1999)50:9%3C799::AID--ASI9%3E3.0.CO;2--G.

Snow, C. P. and Stefan Collini. 2012. *The Two Cultures*. Cambridge MA: Cambridge University Press.

Sokal, Allan and Jean Bricmont. 1998. *Fashionable Nonsense: Postmodern Intellectuals' Abuse of Science*. London: Picador.

Sokal, Robert R. 1958. "A Statistical Method for Evaluating Systematic Relationships." *University of Kansas Scientific Bulletin* 38:1409–38.

Sonnett, John. 2004. "Musical Boundaries: Intersections of Form and Content." *Poetics* 32:247–64. http://dx.doi.org/10.1016/j.poetic.2004.05.007.

Tavory, Iddo and Ann Swidler. 2009. "Condom Semiotics: Meaning and Condom Use in Rural Malawi." *American Sociological Review* 74:171–89. http://dx.doi.org/10.1177/000312240907400201.

Vaisey, Stephen and Omar Lizardo. 2010. "Can Cultural Worldviews Influence Network Composition?" *Social Forces* 88:1595–1618. http://dx.doi.org/10.1353/sof.2010.0009.

Xiao, Zhixing and Anne S. Tsui. 2007. "When Brokers May Not Work: The Cultural Contingency of Social Capital in Chinese High-Tech Firms." *Administrative Science Quarterly* 52:1–31. http://dx/doi.org/10.2189/asqu.52.1.1.

Zipf, George K. 1935. *The Psycho-biology of Language*. Boston, MA: Houghton Mifflin.

Zipf, George K. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.

Acknowledgements: This work was supported in part by NSF grant SBE-0915005 to CTB, a WRF-Hall research fellowship to DAV, and Swedish Research Council grant 2012-3729 to MR. We thank JSTOR for a generous gift and for processing the initial data for this project. We thank Jake Fisher, Mark Mizruchi, Jim Moody, Gabriel Rossman, and Lynne Zucker for their comments on earlier drafts. Direct correspondence to Jacob G. Foster.

Daril A. Vilhena: Department of Biology, University of Washington. E-mail: daril@uw.edu

Jacob G. Foster: Department of Sociology, University of California—Los Angeles. E-mail: foster@soc.ucla.edu.

Martin Rosvall: Department of Physics, University of Umeå. E-mail: martin.rosvall@physics.umu.se

Jevin D. West: Information School, University of Washington. E-mail: jevinw@u.washington.edu

James Evans: Department of Sociology, University of Chicago. E-mail: jevans@uchicago.edu

Carl T. Bergstrom: Department of Biology, University of Washington. E-mail: cbergst@ u.washington.edu