Supplementary Materials

1. DATA COLLECTION AND PREPROCESSING

We obtained the NSF-OCE awards data directly from the NSF website (https://www.nsf.gov/) using the site's Advanced Search feature. Few records prior to 1985 contain award abstracts, so we focused on awards from 1985 to 2018. Of the 19,865 records for the period, 1,212 awards with missing abstracts and 4,755 duplicate records were removed from the data set. Collaborative Research awards are those in which investigators from two or more institutions collaborate on one research project. These awards have the same abstract but the different organizations receive separate awards. For this type of award we consider one abstract per project and sum the amounts awarded to each participating organization. Including multiple identical abstracts for the same project would bias the results, so an additional 2,660 repeating abstracts from Collaborative Research awards were removed from the data set, leaving a final total of 11,238 award abstracts. We do not expect the topic composition of the awards with missing abstracts (~6%) to be significantly different from that of the remaining awards. Therefore, their removal from the analysis should not impact the results significantly. In this study, we do not attempt to distinguish between the different types of NSF-OCE awards and use all the available award abstracts in the analysis.

Prior to applying the topic model, the collection of abstracts is converted to a *bag-of-words* sparse matrix in which each abstract is represented as a vector of word frequencies over the abstracts' vocabulary. The collection of award abstracts contains many similar documents and a high number of words that are very common across abstracts. Words that are too frequent or that only appear in a few abstracts are not very informative to the topic model and negatively impact computational speed and topic interpretability. Therefore, in the conversion to *bag-of-words*, we remove words with document frequency higher than 20% and only include words that appear in at least five abstracts. We also remove a list of stop words (i.e., articles, prepositions, different forms of the verb "to be" and other common words in the English language). To further reduce redundancy and noise in the vocabulary and improve model performance, we apply two forms of normalization to the collection of abstracts that extract normal forms of words by dropping common suffixes (stemming) and using a dictionary of known word forms (lemmatization).

2. TOPIC MODEL

We extract 20 topics by applying latent Dirichlet allocation (LDA) to the *bag-of-words* representation of our collection of abstracts. LDA is a Bayesian probabilistic model that identifies groups of words that tend to appear together frequently in the documents (Blei et al., 2003). It assumes that each document is composed of a mix of different topics, and that the topics are frequency distributions of the words in all the documents. The number of topics to be found by the model is specified by the researcher and is chosen based on interpretability, coherence, and analytic utility of the results (Blei and Lafferty, 2009). After several tests using different numbers, the LDA with 20 topics gave us the best results based on those criteria.

The LDA takes as input a m documents by n words *bag-of-words* sparse matrix:

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & x_{2,3} & x_{2,4} & \dots & x_{2,n} \\ x_{3,1} & x_{3,2} & x_{3,3} & x_{3,4} & \dots & x_{3,n} \\ x_{4,1} & x_{4,2} & x_{4,3} & x_{4,4} & \dots & x_{4,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{m,1} & x_{m,2} & x_{m,3} & x_{m,4} & \dots & x_{m,n} \end{bmatrix}; \quad X \in \mathbf{R}^{m \times n},$$

where $x_{i,j}$ is the frequency of word j in document i. The LDA output includes the k topics by n words latent topics (LDA components) matrix:

$$C = \begin{bmatrix} c_{1,1} & c_{1,2} & c_{1,3} & c_{1,4} & \dots & c_{1,n} \\ c_{2,1} & c_{2,2} & c_{2,3} & c_{2,4} & \dots & c_{2,n} \\ c_{3,1} & c_{3,2} & c_{3,3} & c_{3,4} & \dots & c_{3,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{k,1} & c_{k,2} & c_{k,3} & c_{k,4} & \dots & c_{k,n} \end{bmatrix}; \quad C \in \mathbf{R}^{k \times n},$$

where $c_{i,j}$ is the frequency of word j on topic k, and the m documents by k topics matrix of topic probabilities:

$$P = \begin{bmatrix} p_{1,1} & p_{1,2} & p_{1,3} & \dots & p_{1,k} \\ p_{2,1} & p_{2,2} & p_{2,3} & \dots & p_{2,k} \\ p_{3,1} & p_{3,2} & p_{3,3} & \dots & p_{3,k} \\ p_{4,1} & p_{4,2} & p_{4,3} & \dots & p_{4,k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{m,1} & p_{m,2} & p_{m,3} & \dots & p_{m,k} \end{bmatrix}; \quad P \in \mathbf{R}^{m \times k},$$

where $p_{i,j}$ is the probability of topic j for document i. Sorting the word frequencies in the rows of C in descending order gives us the most common words in each topic (Table 1). The rows of P contain the probability distribution of the different topics for each document $\left(\sum_{j=1}^{k} p_{i,j} = 1\right)$. The topic probabilities $p_{i,j}$ represent the proportions of the different topics in

each award. Adding these proportions across awards for each topic gives us the frequency of awards per topic:

$$W_j = \sum_{i=1}^{m} p_{i,j}. \quad (1)$$

We divide the award frequencies per topic by the total number of awards (m) to obtain the topic fractions of the number of awards (Equation 2):

$$L_j = \frac{1}{m}\sum_{i=1}^{m} p_{i,j}. \quad (2)$$

Multiplying the amount of money awarded for each project (a_i) by the award's topic probabilities ($p_{i,j}$) gives us the portion of the money that goes into each topic ($p_{i,j}\, a_i$). The sum of those portions across awards for each topic is the total amount of money awarded per topic:

$$A_j = \sum_{i=1}^{m} p_{i,j}\, a_i. \quad (3)$$

The topic fraction of the total amount awarded is computed by dividing the amount awarded per topic by the total amount of money awarded (all topics):

$$S_j = \frac{\sum_{i=1}^{m} p_{i,j} a_i}{\sum_{i=1}^{m} a_i}. \quad (4)$$

We obtain the mean amount awarded per project for each topic by dividing the amount awarded (A_j) by the award frequency (W_j) for each topic:

$$R_j = \frac{\sum_{i=1}^{m} p_{i,j} a_i}{\sum_{i=1}^{m} p_{i,j}}. \quad (5)$$

For each year and topic, we compute the fraction of the number of awards (L_j), the amount of money awarded (A_j), and the fractions of the total amount of money awarded (S_j) and the mean amount of money awarded per project (R_j) to examine trends in research and funding (Figures 3–5 and 7). Note that we are dividing the awards into components (one for each topic), partitioning the money awarded for each project between its different components, and using all components (topics) of the awards in the computations (Equations 1–5) so no topic information is lost.

The amount of money awarded for each project was adjusted for inflation to 2017 dollars using the US Bureau of Labor Statistics Consumer Price Index annual average. The years 1985, 1986, and 2018 contain only 2, 2, and 27 awards with valid abstracts, respectively. As a result, the annual statistics for these years are highly biased and not representative of all the awards in those years, and were removed from the time series presented in Figures 2–7.

3. VISUALIZATION OF AWARDS IN TOPIC SPACE

We visualize the distribution of awards in 20-dimensional topic space using t-Distributed Stochastic Neighbor Embedding (t-SNE; van der Maaten and Hinton, 2008). t-SNE is a state-of-the-art nonlinear dimensionality reduction algorithm used in the visualization of high-dimensional data. The algorithm finds a two- or three-dimensional representation of high-dimensional data that retains much of the local structure, while revealing important global structure such as clusters at multiple scales (van der Maaten and Hinton, 2008). These features make this algorithm ideal for visualizing the distribution of awards in topic space, as it makes it easier to identify the clusters of awards that form the different topics as well as groups of related topics (clusters) that form the different NSF-OCE programs (Figure 1a,b).

In Figure 1a, awards are assigned the topic for which the probability is the highest. For more interdisciplinary awards, the probability of the dominant topic might be only slightly higher than that of other topics, which can result in those awards being collocated with awards from a different topic but with similar topic composition.

4. AWARD INTERDISCIPLINARITY

The LDA model assumes that each award is composed of a mix of all 20 topics. The interdisciplinarity of each award is directly proportional to the evenness of the distribution of topic probabilities. Evenly distributed topic probabilities imply diverse topic composition and high interdisciplinarity. Conversely, nonuniform probabilities indicate dominant topics and lower interdisciplinarity. Here, we borrow from the ecological literature and use the Pielou's evenness index (Ludwig and Reynolds, 1988) to quantify award interdisciplinarity:

$$J_i = -\frac{1}{\log k}\sum_{j=1}^{k}(p_{i,j}\log p_{i,j}),$$

where $p_{i,j}$ is the probability of topic j for document i (see matrix P in Section 2). Pielou's index J_i varies between 0 and 1. Higher values mean a more even probability distribution and higher interdisciplinarity.

We examine temporal trends in interdisciplinary research by looking at how the annual mean award interdisciplinarity for each topic varies in time (Figure 2). To compute the annual mean award interdisciplinarity for the different topics we assign each award the topic for which it has the highest probability ($p_{i,j}$), group the awards by year and topic, and average the award interdisciplinarity for the awards within each group. The means for the years 1985, 1986, and 2018 were removed from the time series presented in Figure 2 due to the very low number of awards with valid abstracts in these years (see Section 2).