Bridging "The Two Cultures": A View into the Sciences and Humanities through Textual Analysis

Kim Ruelle
Research Experience for Undergraduates 2008, Santa Fe Institute. Mentor: David Krakauer.

Abstract

In a 1959 lecture, novelist and scientist C.P. Snow posited a chasm between the sciences and the humanities. Many have since discussed and debated his claim, but the idea of such a break continues to persist. We are utilizing textual analysis to study representative texts in each of these areas, novels and scientific papers. We have used a multi-pronged approach in our work, drawing on information theory, network analysis, differential equation modeling, and principal component analysis. We aim to learn more about the texts themselves, their related fields, and the bridges that may connect the "Two Cultures." This study is ongoing, and this paper is a summary of some of our current findings.

Introduction

In 1959, British novelist and scientist C. P. Snow gave a Rede Lecture at the University of Cambridge on the relationship between the sciences and the humanities. Having a background in both of these areas, and frequently socializing with scholars in both groups, Snow felt he had a unique insight into problems plaguing these disciplines. He dubbed this "The Two Cultures": that "intellectual life ... [was] increasingly being split into two polar groups. ... Literary intellectuals at one pole—at the other scientists.... Between the two a gulf of mutual incomprehension—sometimes ... hostility and dislike, but most of all lack of understanding." ([1])

Snow's speech centered on the deep divide between these disciplines, particularly between the physical scientists and literary scholars ([1]). The idea was quickly seized upon by others, and the idea

of "Two Cultures" became popular. It is obvious some kind of divide still exists between the sciences and humanities. Traditionally, this problem has been attacked from a perspective most firmly rooted in the humanities: in a narrative or essay form. We have decided to approach the issue from the other direction, using quantitative data and analyses. We chose to access these "Two Cultures" by studying texts representative of each culture: novels for the humanities, and scientific papers for the sciences. Through textual analyses of the titles (and abstracts, for papers) of these works, we aim to gain insights into both fields individually, and to find differences and similarities between the two.

Our research was initially inspired by the work of literary scholar Franco Moretti. He conducted analyses on British novel titles published from 1740-1850. In his results, he found that the mean length of novel titles was greatly decreasing over time, while the number of novels produced each year was increasing. He proposed several potential explanations for this trend. As the number of novels on the market grew, shorter novels would be better at catching the eye of a browsing reader. Circulating libraries were also becoming more popular, and tended to shorten the titles of books to fit in their library catalogues, and to fit on book spines. New novels with shorter titles would not risk titles being arbitrarily shortened in libraries. People were also more familiar with the novel as a form as time went on, and no longer needed a long summary-like title to induce them to read a book. Our initial choice of novel titles as a subject of study was through dialogues with Moretti, and we hoped to expand on his work. ([2])

We identified several differences and similarities between our chosen representative texts. Novels are typically single author, long, and either self- or editor-reviewed. Papers are typically multi-author, short, and peer-reviewed. However, old novel titles were used in a form very similar to paper abstracts. Additionally, the titles (and abstracts) of both are used to summarize the ideas contained within through the English language (since all our records were in English). Titles also provide a window into the larger text, and the texts can be said to provide a window into the larger culture. Finally, since both our novels and papers were drawn from many years, we can treat these records as

cultural or conceptual time series.

Data

We drew novel title data from three databases, all supplied us by Franco Moretti. All novel titles were English-language titles of British novels. The time periods covered by the three databases were 1600-1660, 1660-1715, and 1740-1850. There were 203 titles from 1600-1660, 488 titles from 1660-1715, and 7304 titles from 1740-1850. We were most interested in the 1740-1850 time period, since Moretti had suggested that was the area with the best data and most worthy of study. However, for some of our investigations, we did look at data from all three time periods.

For the scientific papers, we used working papers from the Santa Fe Institute (SFI). We were given data for the titles and abstracts of 907 working papers from the years of 1989-2001.

Methods

We decided to approach these two data sets from four main directions: Dynamics, Information, Networks, and Dimensions. For the novels, all four of these methods of investigation proved useful. For the working papers, we were only interested in the latter two.

Novels: Dynamics

Our first thought was to attempt to create a crude model to explain the trends in title length and numbers of novels published from 1740-1850, as previously observed by Franco Moretti. We decided to build a model using ordinary differential equations. The equations of the model are shown below (figure 1). In this model: λ_i is the number of books of title length i produced per year; N is the population of free (unoccupied) readers; r_i is a function representing the preference of readers for books of title length i as a function of i; x_i is the number of books of title length i published; y_i are the number

Figure 1: Novel Production Model Equations

$$\lambda_i = \frac{N}{1 + x_i^p}$$

Production function

$$r_i = \frac{1}{i^p}$$

Preference function

$$\frac{dx_i}{dt} = \lambda_i - r_i x_i N$$

Books published

$$\frac{dy_i}{dt} = r_i x_i N - s y_i$$

Books being read

$$\frac{dN}{dt} = -N \sum_j^L r_j x_j + s \sum_j^L y_j + gN$$

Population of free readers

$$\lambda_T = \sum_j^L \lambda_j$$

Total # books / year

$$\langle \lambda \rangle = \frac{\sum_j^L j \lambda_j}{\lambda_T}$$

Mean title length / year

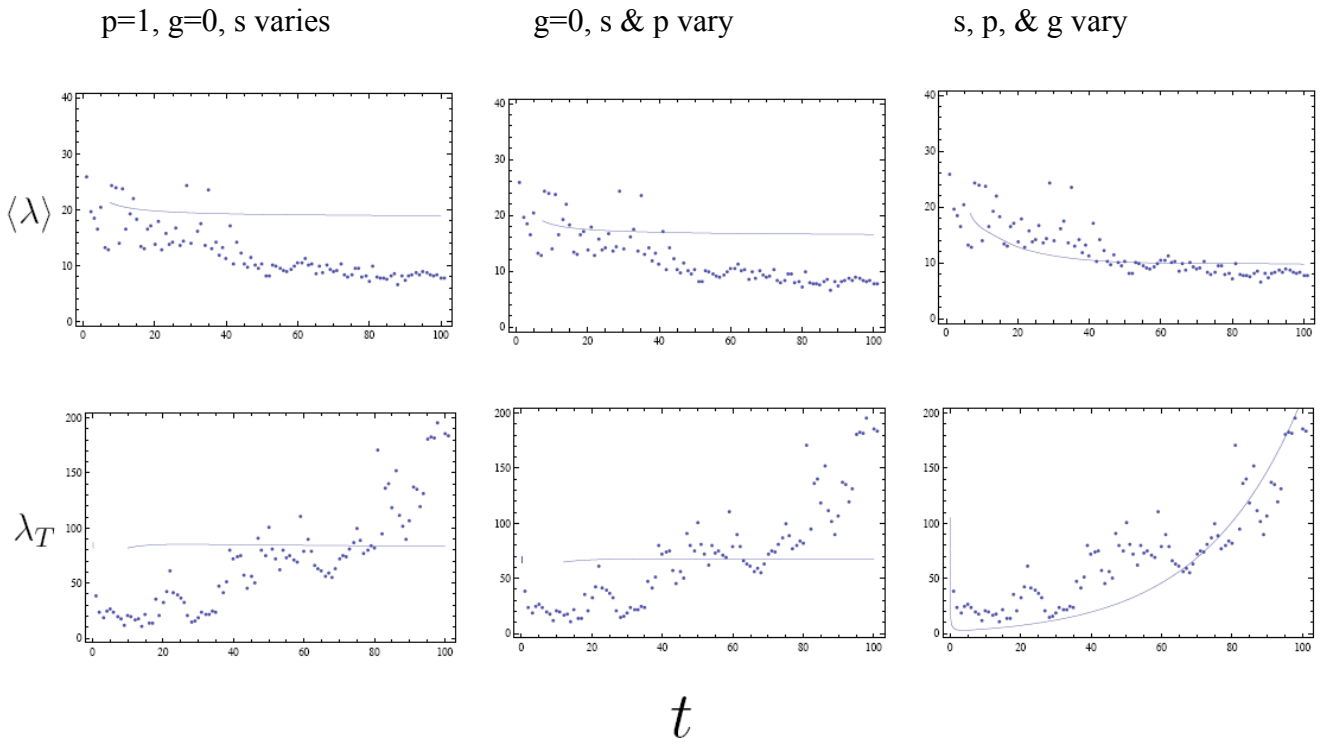
---

of books currently being read. The parameters p, s, and g control the preference function, reading speed, and population growth, respectively. The range of i is (1, L), where L is some value for the maximum title length. There were a set of initial conditions imposed on the model: $N(0)=100$; $y_i(0)=0$; $x_L(0)=0$; $x_{i \neq L}(0)=C$ (where C is a small constant). Due to these conditions, at time 0, there are N books of title length L being produced, while a small (and equal) amount of books of all other title lengths are being produced.

After we had developed this model, we attempted to fit it to the data. Initially, we set p=1, g=0, and allowed s to vary. This did not produce a very good fit. Then we set g=0 and allowed both s and p

to vary. This produced a slightly better fit, but not by much. Finally, we allowed all three parameters s, p, and g to vary. This produced a fairly good fit to our data. Figure 2 below displays these three instances.

Figure 2: Novel Production Model Fit Graphs



Since population growth and the preference function were both crucial in fitting this model, we conclude that these factors may explain a lot of the trend in novel title lengths and the number of novels published. Additionally, we hope that in the future we can get accurate figures for the reading population size and rates of growth. From this, we would be able to perhaps extract the parameters for the preference function or the speed of reading function.

Novels: Information

Before we could study information measures, we needed to decide on a way to extract the data

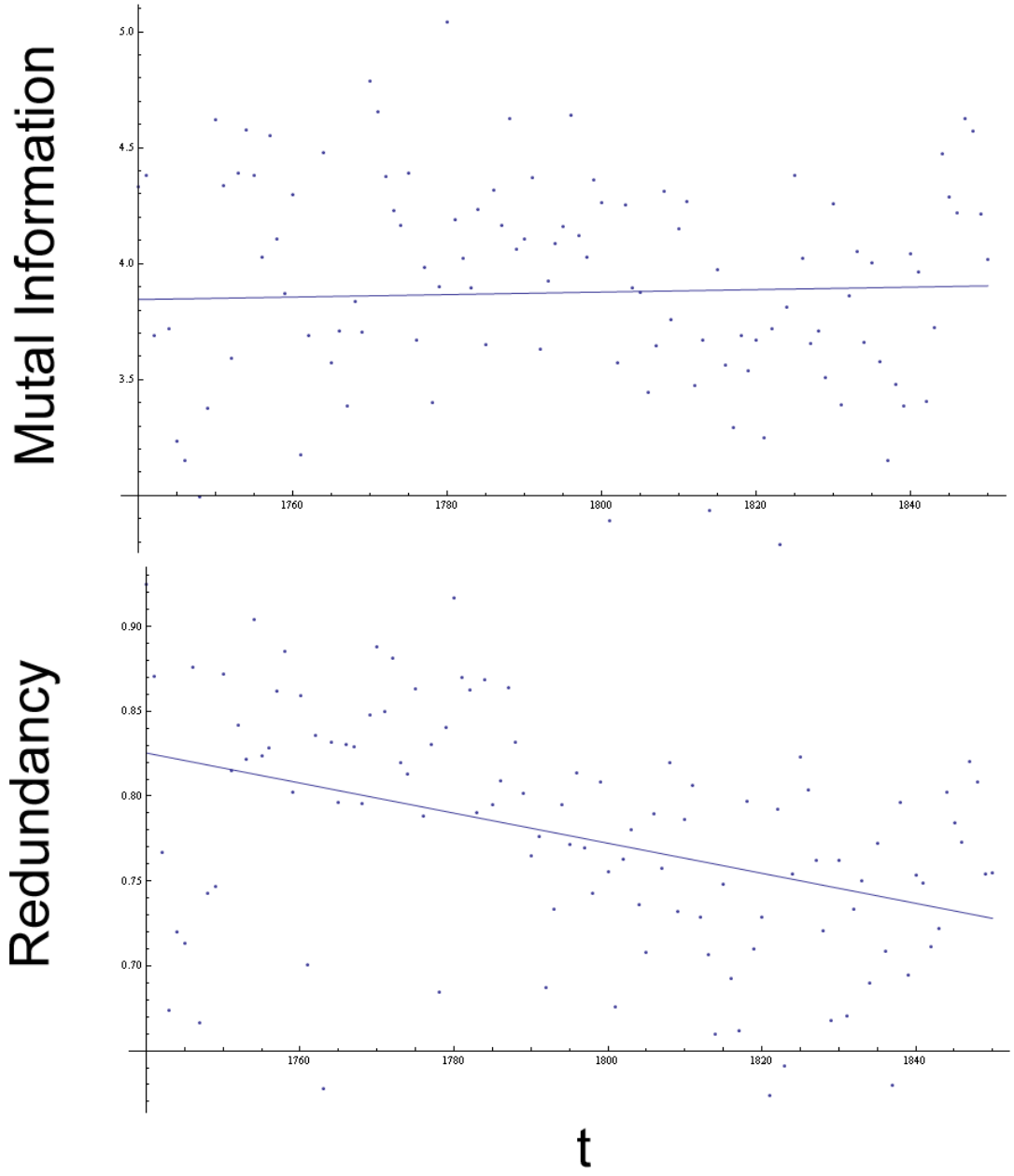
we needed from our raw data. We decided to study keyword co-occurrence. Thus, we would pick a list of n keywords for a certain period of time, and then we would count the number of times each pair of words co-occurred in the same record. However, we needed a way of extracting these keywords. Originally we tried manual selection, but were worried about the obvious bias this incorporated. Then, we attempted to create an automatic keyword selection method. We used a filter list of function words to eliminate words without useful information (e.g., articles, pronouns, prepositions). After applying this filter, we found that in some cases certain keywords were still being selected that we felt did not carry significant meaning. Therefore, we played around with making a combination automatic and manual word filter. However, for most of our measures, we just used the automatic filter, and only used the combination filter in order to compare the two.

For the novels, there were not enough words per year in some early years to do keyword selection by year, so we did keyword selection by decade. We set our number of keywords to 20. All keyword co-occurrence matrices were thus symmetric 20x20 matrices with integer entries, and 0's along the diagonal.

Once we began looking at information, our main question was: What happens to information as novel titles get shorter? As the titles get shorter, fewer and fewer words are available to the author to convey the same amount of information to the reader. We theorized that this might be dealt with by getting rid of words with high mutual information. Thus, similar words such as "king" and "monarch" would not appear in the same title together, since one already tells a lot of the information contained in the other one. Shorter titles would eliminate the redundancy present in longer titles. However, when we analyzed the mutual information in the data, we found the following (figure 3).

In our analysis, we found that mutual information actually remained almost constant, with a slight increase over time. This was not what we expected, and we theorized that maybe this was happening because of the greatly increasing number of books published over time. Therefore we also

Figure 3: Mutual Information in Novel Titles (Keyword Co-Occurrence)



calculated the redundancy for the data (figure 3), a normalized variant of mutual information. The redundancy was indeed shown to be decreasing over time, suggesting a pattern more in line with our theory. However, because of the inconsistency of the mutual information data, we feel neither of these results are conclusive yet, and require further study.

Novels: Networks

We also used the keyword co-occurrence matrices as adjacency matrices for generating networks. Thus, each keyword was a node, and if two keywords co-occurred four times, there would be four edges connecting those nodes. From these we generated networks to which we could apply various algorithms in an effort to learn more about how specific keywords were related to each other. With the novels, we were hypothesizing that strongly related groups of keywords within the network might relate to genres of novels.

One issue we have not dealt with yet is the generation of null distribution networks to compare to our data-generated networks. We realize this is an important issue for assessing the significance of our results, and plan to do this, but have not had time to do it yet now.

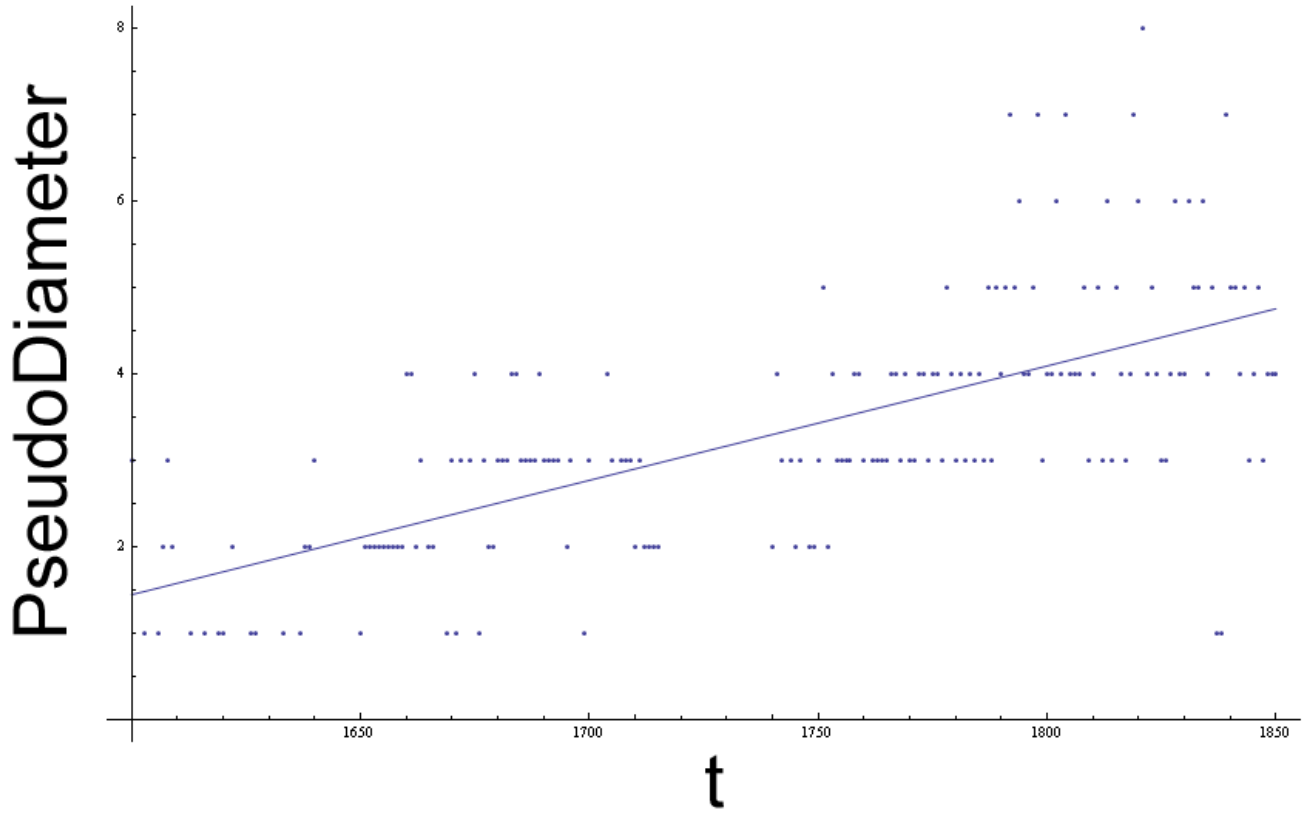
Another issue with our novel networks was that particularly in years with very few records or with very short titles, we had some graphs that were unconnected (thus, those keywords did not co-occur with any of the other keywords in any of the records for that year). We believe these unconnected graphs could cause issues for some of our analyses, but we have not yet been able to really delve into what those specific problems might be.

The first analysis we used to assess these networks was closeness centrality. Closeness is the mean shortest path between a node and all other nodes reachable from it, and is a measure of centrality ([3]). We calculated the closeness for each node in the network, and then ranked the nodes by closeness to find the most central nodes. For example, in 1847, "love" and "family" were the top two most central nodes. We think that the most central nodes were certain years were perhaps indicative of themes that were overpowering the books published in that year, and thus might represent genres that were most central.

Another measure we used was pseudo diameter, an algorithm which approximates the diameter of a network. The diameter is the longest of all the shortest paths between any two pairings of nodes in the network ([3]). Here, of course, the unconnected graphs might cause us problems (as they have

infinite diameter, but not pseudo-diameter). We took the pseudo diameter as a rough measure of how spread out the two most distantly related concepts (among the 20 keywords we selected) were for each year. For example, in 1819, "romance" and "history" were spread apart with a pseudo diameter of 7. When we graphed the pseudo diameters over time, we found that they appeared to be increasing (figure 4). We hypothesize that this might be indicative of words becoming more and more spread out over time. We also think this might relate back to our previous redundancy measures, as increasing distance between words would decrease redundancy.

Figure 4: Pseudo Diameters of Novel Co-Occurrence Networks



Novels: Dimensions

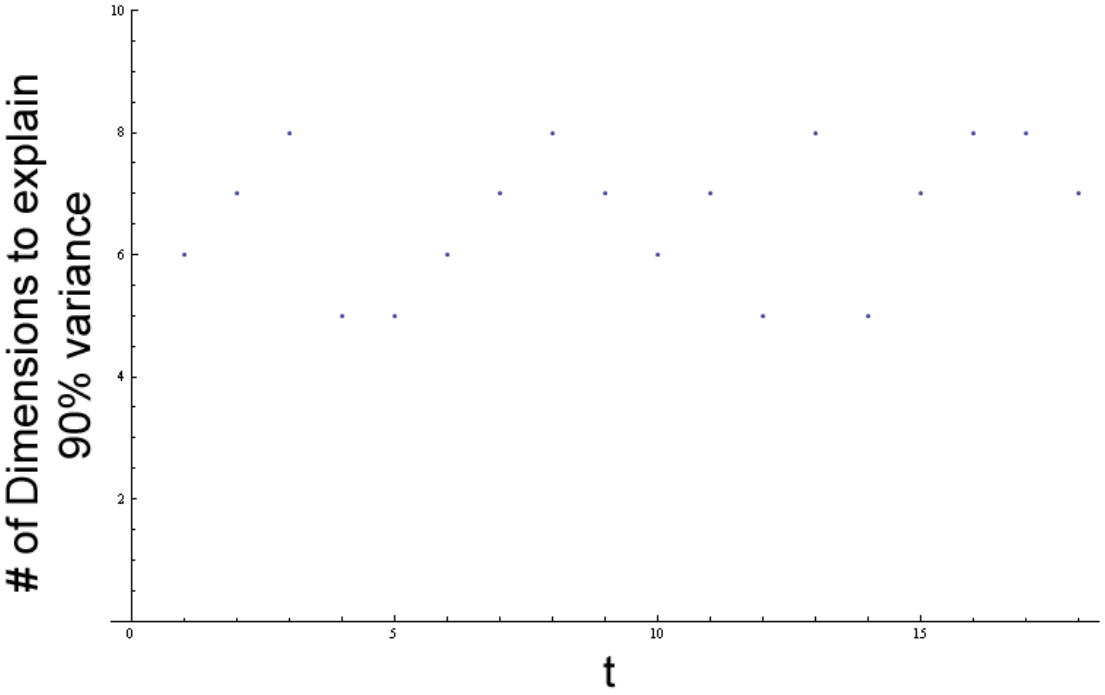
Our last question of interest for the novel titles was: What are the principal conceptual dimensions in the data over time? In order to assess this, we first calculated keyword occurrence (not co-occurrence) per year per decade. Then, we used principal components analysis (PCA), a method for finding patterns in high-dimensional data, to analyze these resulting matrices. We set each keyword as a dimension and calculated the covariance matrix for our data. Then we calculated the eigenvectors and eigenvalues for these covariance matrices. In PCA, these eigenvectors (which are of course all orthogonal) each correspond to a dimension in the data. Additionally, the eigenvectors with the highest eigenvalues correspond for most of the variance. ([4]) From these ranked eigenvectors, then, it can be determined how many eigenvectors (i.e., principle components or dimensions) were necessary to account for 90% of the variance. The principle components could also be analyzed to see, for example, which keywords accounted for most of the direction of that vector. In the tenth decade of our data, the top five keywords in the first principle component were "France," "containing," "king," "French," and "history." Also in this decade, the top two dimensions together accounted for 75% of the variance.

When we graphed the number of dimensions over time for the novel data (figure 5), we saw that they appeared to be remaining fairly constant over time, with an upper and lower bound. (The mean was 6.7 and the standard deviation was 1.1, with a minimum of 5 and a maximum of 8.) There appeared to possibly some kind of periodic behavior going on. When we plotted a line to this data, we did find a slight upward trend, however, the strong bounding of the data over time suggested to us that the number of dimensions in the novel data was remaining roughly constant, varying around 6.7. This is a very interesting result because some have suggested that the number of genres (and here our dimensions might correspond to genres) in novels has always been constant, ranging between 5 and 7. This is very close to matching the data we found. Another interesting interpretation is that since the number of dimensions is remaining roughly constant over time, this might mean that no really new ideas are being incorporated into novels, that all that can be said has already been said, and the same

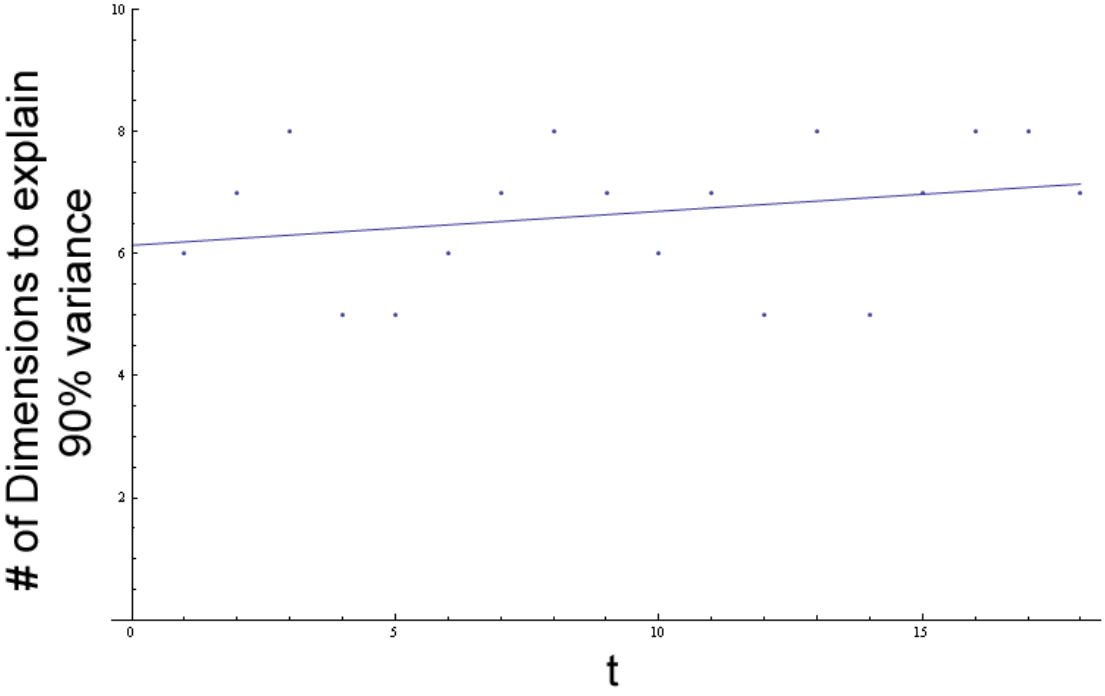
themes are simply re-explored over and over.

Figure 5: Dimensions in Novels

Without fit:



With linear fit:



Working Papers: Networks

In studying the working papers of SFI, we hoped to gain some insights into science in this specific culture. First, we used the same keyword co-occurrence method used for the novels, except here we had enough data to extract keywords for each year. We thought of these word co-occurrences as shared concepts in this space, and were interested in exploring the space of those concepts and how that landscape has changed historically at SFI.

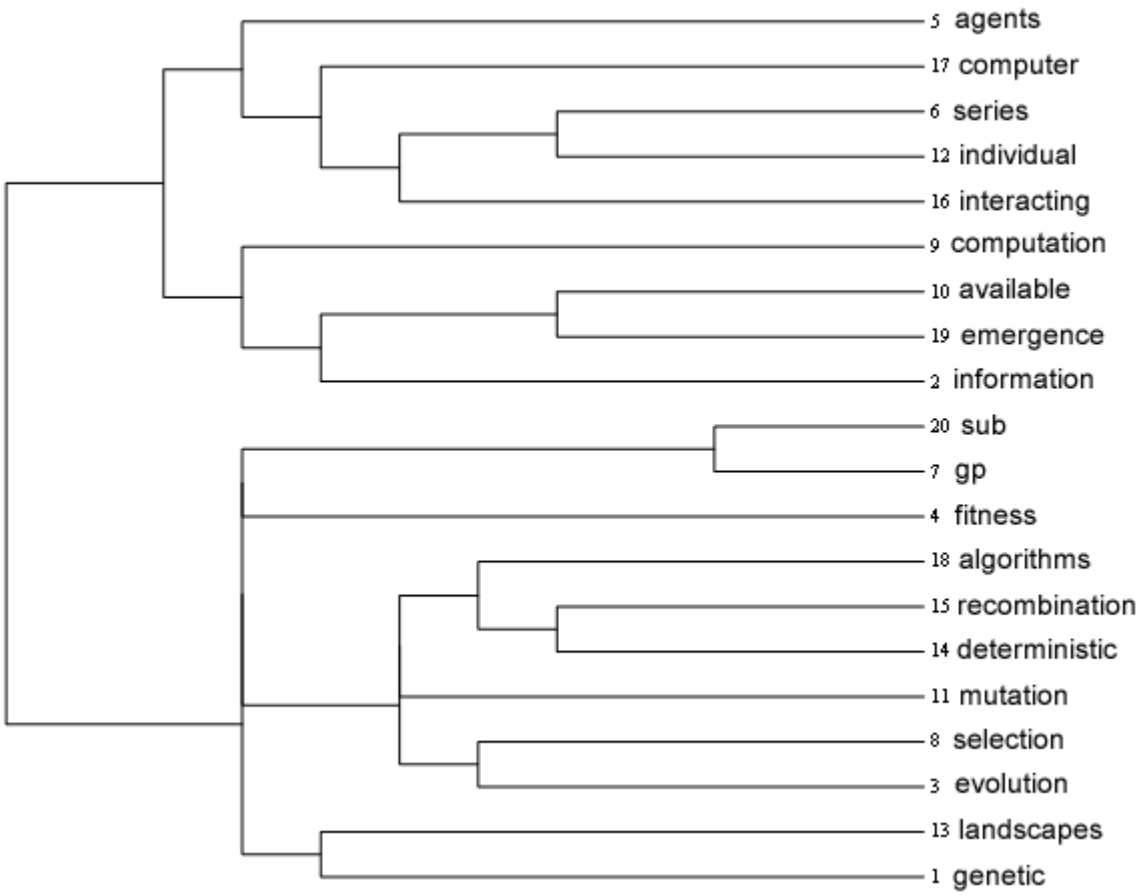
Using the co-occurrence matrix as our adjacency matrix again, we constructed our networks. (And once more, we have not yet calculated the null models to compare these to, but are planning to do so.) In this case, almost all of our networks were connected, regardless of whether we used the automatic filter or the combined filter that also incorporated manually excluded words. Only one or two years, through either filter, had single points that were unconnected. Using these networks, we wanted to study the robustness of the concept space, and which concepts were central.

Setting n equal to some value between 2 and 20 (the number of keywords we had selected), we used an algorithm to "find a partition of the vertices into n parts so that each part has roughly the same number of vertices, and so that the number of edges between these parts … is minimized" ([3]). That is, we cut our graph into n parts, while cutting as few edges as possible. The resulting n sub-graphs would be those more densely connected, and correspondingly, we hypothesized, more tightly conceptually related. For example, for the year 1994 and n=5, we got the following groupings: (1) "information," "computation," "available," "emergence"; (2) "agents," "series," "individual," "interacting"; (3) "genetic," "GP," "computer," "sub"; (4) "evolution," "fitness," "deterministic," "recombination"; (5) "selection," "mutation," "landscapes," "algorithms." In addition, using this algorithm, we can count the percentage of edges that needed to be cut for each year in order to partition the graph. We believe that this measure could function as a representation of robustness, but this will require further study.

Next, we took the above "min cut" algorithm and created an iterated form of it. We ran the algorithm for each year from n=2 to n=20, and counted the number of times each pair of keywords

were grouped together in the resulting sub-networks. We used these counts as inverse indicators of distance: words that were grouped together many times were more closely related. From this, then, we were able to construct a hierarchical graph relating the relative distances between concepts. We feel this is a useful conceptual visualization, because it shows which groups of words appear to be more closely related, and which are very easily broken apart. An example hierarchical tree for the year 1994 can be seen in figure 6 below:

Figure 6: Iterated Min Cut Hierarchical Tree for 1994



Working Papers: Dimensions

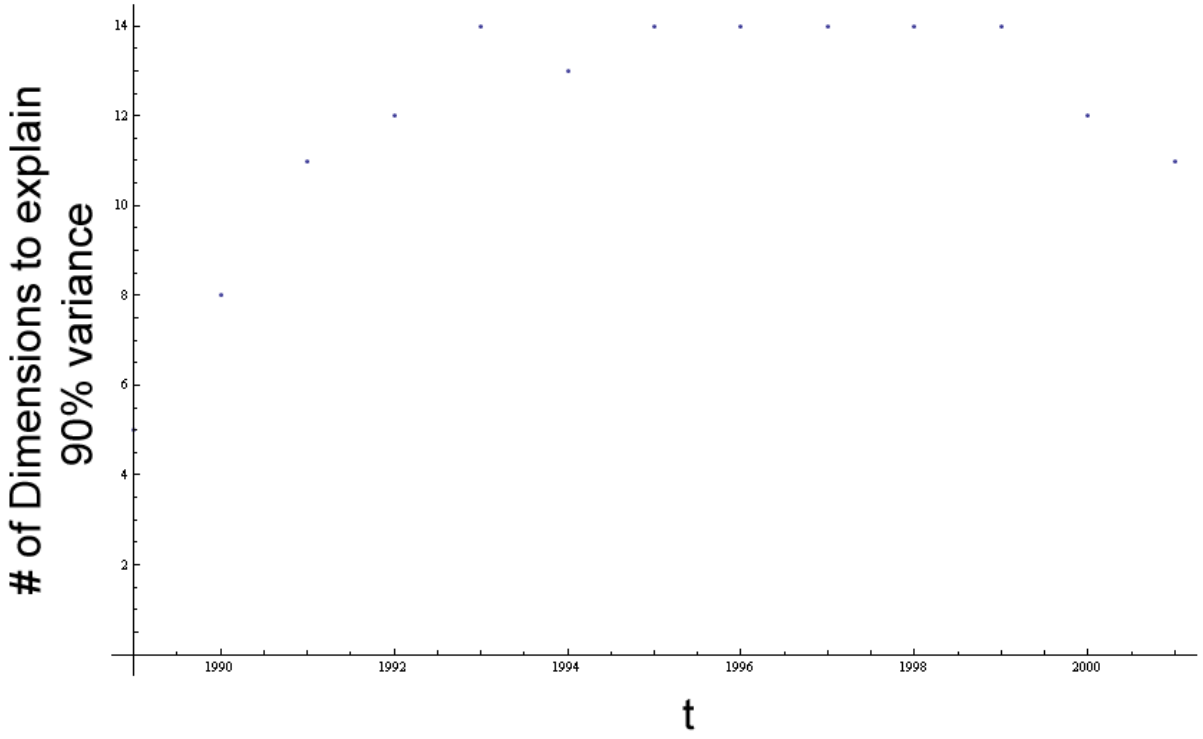
We studied working paper dimensions using the same PCA method used for the novels, but using the number of keyword occurrences per record per year. For the year 1994, for example, the top five keywords making up the first principle component were "selection," "time," "GP," "fitness," and "genetic." The top two principle components together accounted for only 31% of the variance.

When we plotted the number of dimensions necessary to account for 90% of the variance (figure 7), we saw a general upward trend, with a mean of 12 and standard deviation of 2.8. Fitting a linear regression to the data confirmed this. This suggests that originally, SFI was operating in a smaller conceptually dimensional space where certain concepts dominated, and that as time went on, the space became more complicated with higher dimensions as new concepts and ideas were realized. This implies that early in, sub-spaces in this high-dimensional space were not occupied. This is the opposite of redundancy, with more and more concepts not co-occurring as time went on, and realizing new ideas is really what science is supposed to be all about.

However, we also noticed what appeared to be a dip in the last few years of the data. Fitting a quadratic regression to the data also showed this downward dip (figure 8), with a maximum number of dimensions reached in the mid- to late-1990s. This beginning of a downward trend is worrying to us. Does this mean that the number of dimensions is beginning to decrease? Are fewer new ideas being realized? One idea is that a particularly attractive or unifying new conceptual area arose, and a lot of people started doing flocking to and doing research in that area. Ideally, we would like to obtain data for the years beyond 2001 to see whether this downward trend continued, leveled off, or reversed itself. Also, when comparing the dimensions of the novels and working papers, another interesting potential explanation presented itself, which will be described in the next section.

Figure 7: Dimensions in Working Papers (1)

Without fit:



With linear fit:

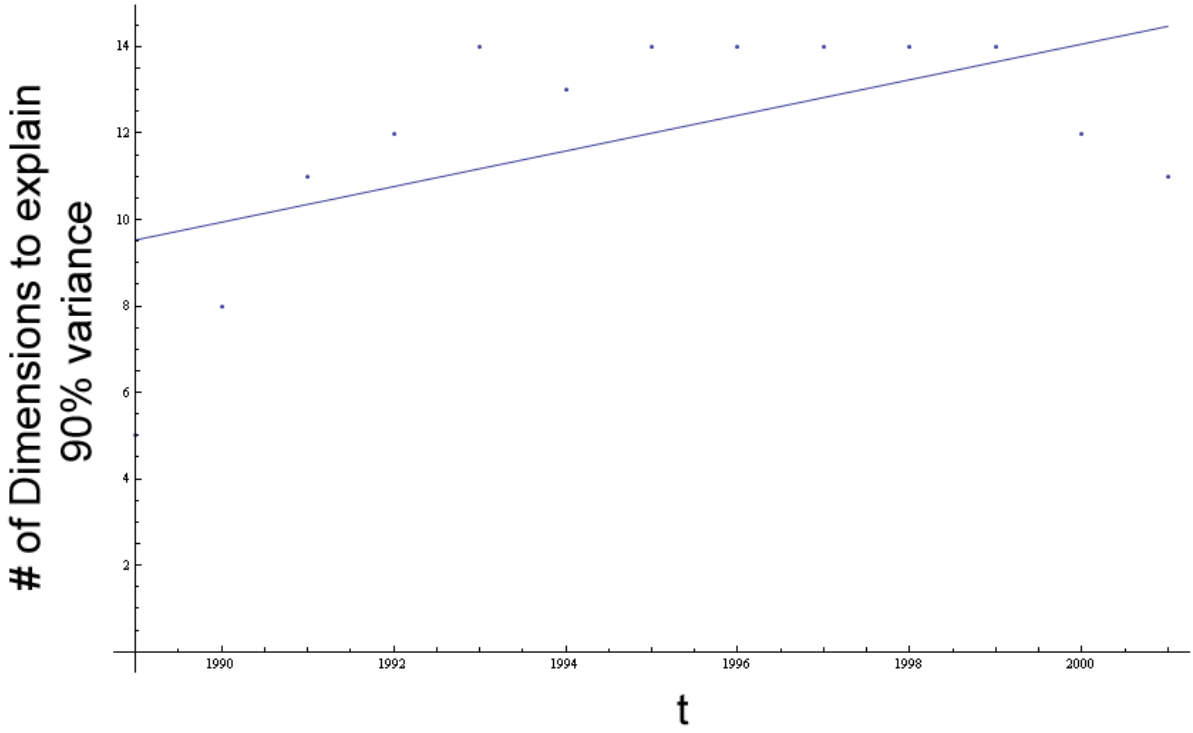
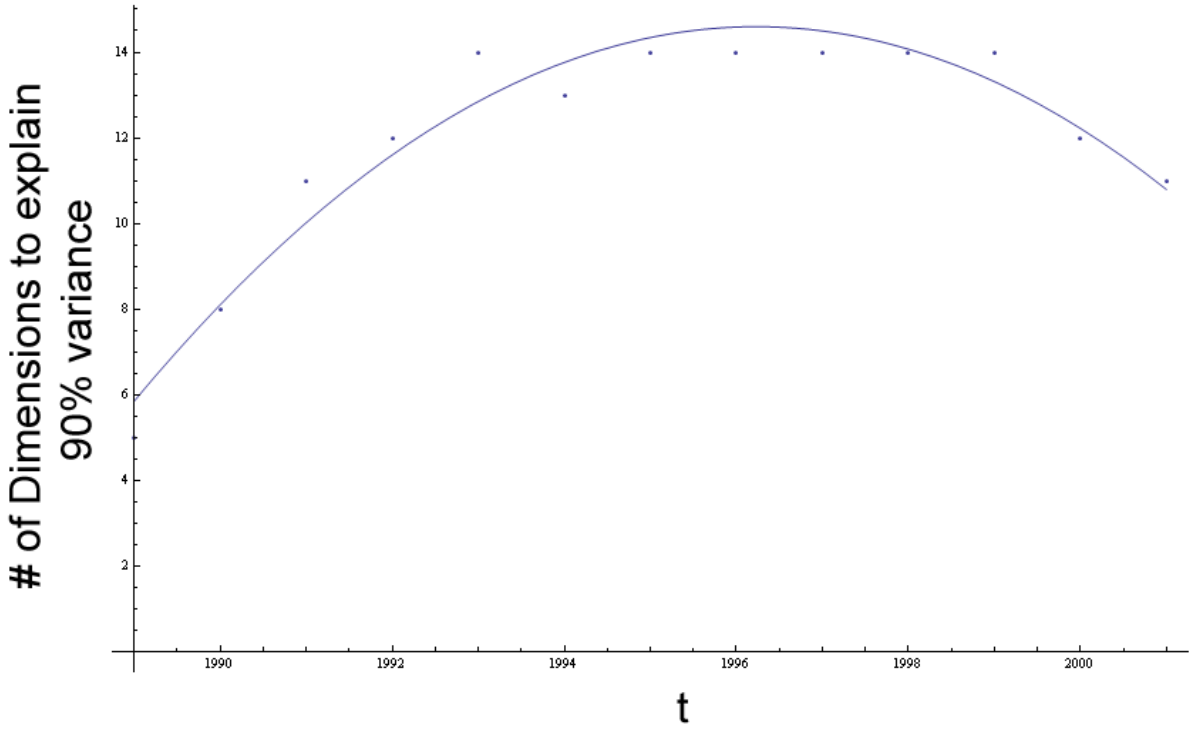


Figure 8: Dimensions in Working Papers (2)

With quadratic fit:

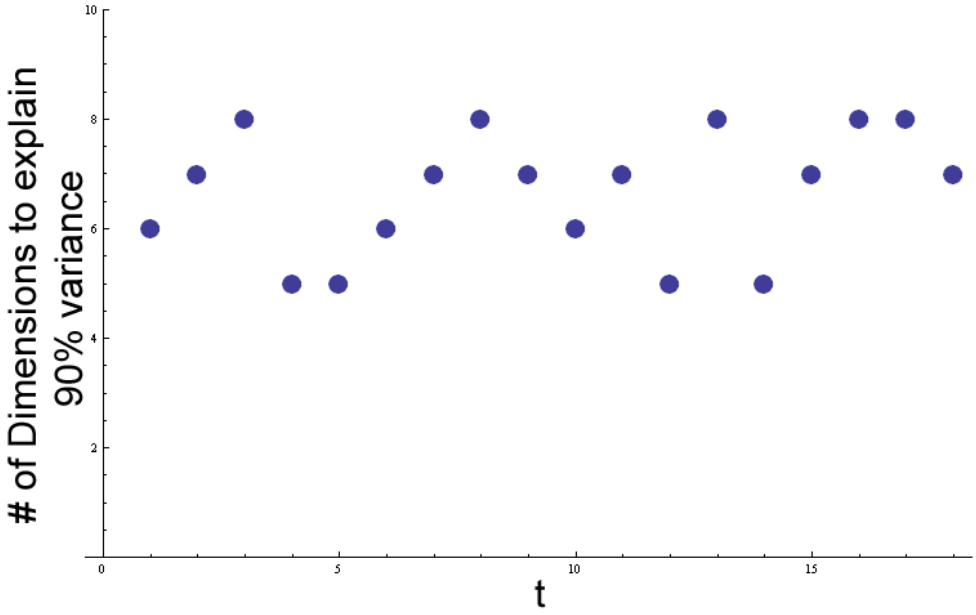


Bridging the Cultures: Dimensions

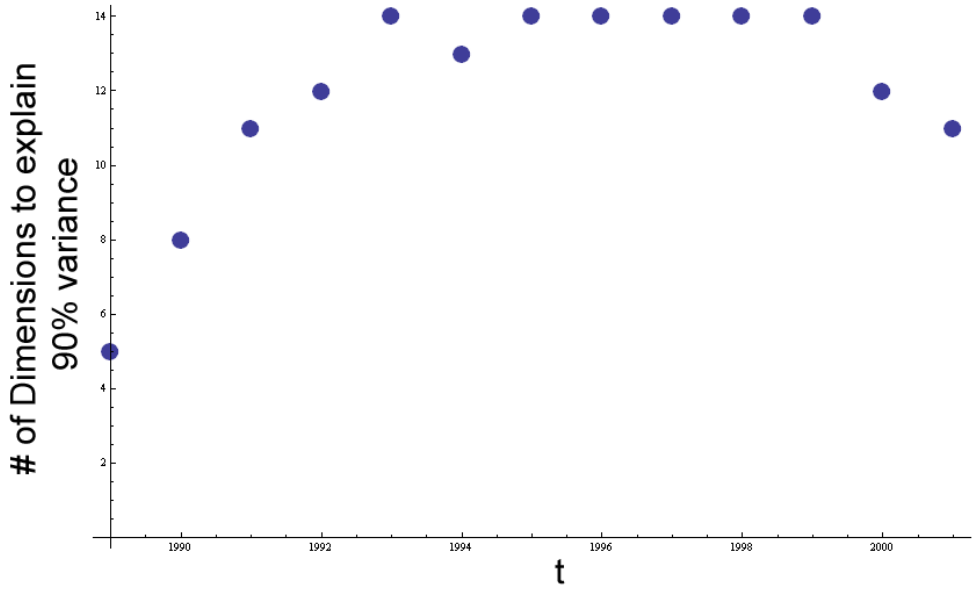
When comparing the dimensions for the novels and the working papers, we saw that the working papers were a much higher conceptually dimensional space. The novels had a mean of 6.7 ± a standard deviation of 1.1, while the working papers had a mean of 12 ± a standard deviation of 2.8. Of course, we must take into account the different number and length of records for each of these data sets, so it is uncertain whether the number of dimensions can be directly compared and yield useful results. However, we feel we can compare the trends within each of these dimension time series (figure 9). The novel data seems largely invariant and bounded, whereas the working paper dimensions seem to be increasing over time (discounting the slight decrease at the end). This might suggest that for the novels, everything has "already been said" and the number of genres is constant, whereas for the working papers, new concepts spaces are introduced as time goes on.

Figure 9: Visual Comparison of Dimensions in Novels and Working Papers

Novels:



Working Papers:



However, another interesting thing to consider is what would happen to the working papers over a longer time period. We had data for the novels for almost 250 years, while we only had working paper data for 13 years. One proposed explanation for this dip at the end of the working paper data, then, is the beginning of periodic behavior. Perhaps, if we had scientific paper data for a longer period

of time, we would see the kind of possibly periodic behavior we see in the novel data, with an upper bound and a lower bound, and periods of higher and lower dimensions. One conceptual explanation for this is that science is undergoing periods of "normal" and "revolutionary" science, as suggested by Thomas Khun in his book *The Structure of Scientific Explanation*. Therefore, the dip at the end of our time period for the working papers might be the beginning of a normal period in science. ([5])

Bridging the Cultures: Networks

Finally, we compare one of our network measures between the novel and working papers. For pseudo diameter, you will recall, we found an upward trend (increasing diameter) for the working papers, ranging from 1 to 8 (figure 4). In contrast, for the novels, every year had a pseudo diameter of 2. It is possible that this is just noise, maybe caused by the greater number of words in the working paper records, and the larger amount of data per year for many of those years However, another possibility is that this is indeed signal. Maybe in the working papers, we have to travel through fewer concepts to reach new conceptual areas, whereas in the novels, the concepts are much more spread out. This would indicate that connections can be found between many scientific areas of thought through few intermediary areas. To determine whether this is noise or signal, we feel this bears further study.

In Conclusion

We have shown that it is possible to analyze both of these "Two Cultures," the sciences and the humanities, using quantitative analytic methods, and that interesting insights may be gained as a result. We have also shown that it may be possible to compare these two areas using these methods. This scientific and quantitative approach, is vastly different from the usual approach which has been taken in attacking this problem, which is usually in the forms of discourse or essays and thus rooted more firmly in the humanities.

Future Directions

First, this research is still in progress and so obviously we would like to continue it in the future. We need to generate null distribution networks to compare to our data-generated networks, and we would like to both refine our other analyses and explore some new avenues of analysis. Eventually we hope this will lead to a publishable paper on the novel in collaboration with Franco Moretti, as well as an SFI working paper on the "Two Cultures."

Additionally, we have some ideas for future directions we could take to expand this research. We would like to try comparing works from the same time periods, as well as getting data for scientific papers over a longer time period comparable to what we had for the novels here. We would like to explore papers from different scientific communities, and perhaps within disciplines. Another idea is that instead of using novels, we could study scholarly literary papers, a form perhaps more easily compared with scientific papers. Finally, we are curious about the idea of exploring different forms outside of texts, such as music, and relating those to the sciences and the humanities.

Acknowledgments

I would like to thank Franco Moretti for providing inspiration and data for this project and for sharing his work and ideas with us. I would like to thank the Santa Fe Institute and all the people there involved with the REU program for setting up such a great opportunity for undergraduates to do research. And especially, I would like to thank my mentor David Krakauer for working with me on this project and sharing his insights and ideas with me.

References

[1] Snow, C. P. (1959). "The Two Cultures." Rede Lecture.

[2] Moretti, Franco. (N/A). "Quantitative data, formal analysis. Reflections on 7,000 titles [British novels, 1740-1850]." (Pre-published)

[3] Wolfram Mathematica 6 for Students. Help Articles: "ClosenessCentrality," "PseudoDiameter," "MinCut."

[4] Smith, Lindsay I. (2002). "A tutorial on Principal Components Analysis." (February 26, 2002).

[5] Khun, Thomas. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.