Using Research-Interest Similarity and Departmental Co-membership to Predict Collaborative Ties

Susan Biancani
Stanford University School of Education
Stanford, CA USA
biancani@stanford.edu

Abstract

I investigate the complimentary roles of similarity on research interests, and of departmental co-membership on promoting new collaborative ties among professors at an elite American research university. Similarity is measured as cosine similarity of tf-idf calculations of their abstracts for published work, and is recalculated for each year in the study period. I find that similarity is a strong predictor of collaboration, but that department co-membership is even stronger.

1 Introduction

Among sociologists who study universities, one of the central questions surrounding the worth of disciplinary departments is their role in promoting knowledge creation among scholars (Abbott, 2001; Gumport and Snydman, 2002; Hearn, 2007; Peterson, 2007; Pfeffer and Moore, 1980;). In prior work, colleagues and I have demonstrated that both departments and interdisciplinary centers promote new scholarly ties: co-membership in a department or center predicts new collaborations in the form of co-authorship of publications, co-application and co-award of grants, and co-service on dissertation committees (under review)[1]. While it is clear that sharing a membership in these administrative units makes two faculty members more likely to work together than baseline, it is not clear through what mechanism departments and centers operate.

One obvious possibility is that departments collect scholars with similar research interests, and it is this similarity, rather than shared department membership per se, that promotes collaboration. To investigate this alternative, I have developed a text-based similarity metric, which can account for similarity of research interests among scholars. Including this measure in a regression to predict new collaborations, I can assess the effect shared department membership has, over and above the effect of research-interest similarity.

While there has been significant debate around the importance of departments as organizing structures, little empirical work directly addresses this question (but see: Blau, 1973; Friedkin, 1978).

2 The Corpus

This study is based on a rich dataset from Stanford University, covering the years 1993-2007. From the university's budget office, I have an accounting of all money spent during this time and the funding source whence it derived. In addition, I have detailed longitudinal information on all 3057 Academic Council members at Stanford during this period, including:

- Employment: departmental affiliation(s), year of hire, rank in each year, and job classification
- Highest degree obtained, including subject and granting university
- Personal attributes: age, gender, ethnicity
- Research and advising: dissertation committees, grants applied for, grants

[1] This earlier paper primarily investigated the role of interdisciplinary centers in tie formation at Stanford. As part of that research, we showed that departments are strong predictors of new collaborative ties. Now, I intend to investigate how departments achieve this effect.

received, publications, patents, citations referenced[2]

These data allow me to derive networks from over 400,000 collaborations in distinct types of core intellectual activity. For each type of network, I can note the presence or absence of a tie between two faculty members in a given year. Types of ties include:

- Co-authorship: two faculty members are both listed as authors on an ISI Web of Science publication
- Co-grantee: two faculty members apply for a grant together
- Co-mentoring: two faculty members serve together on a dissertation committee
- Shared references: two faculty members independently cite the exact same publication in work they publish (note: shared references do not include references in papers co-authored by the two faculty members in question)

I will base the similarity measure on a corpus of 66,000 abstracts of all papers published by Stanford faculty members between 1993 and 2007. I have significant meta-data about each publication, including:

- Year of publication
- Journal name and ISSN number
- Number of pages
- Number of references
- Publication type and article type
- Number of authors and all author names

For authors who are Stanford faculty members, the publication is linked into the database, and can be joined to personal information about authors.

3 Measuring Similarity

I conceptualized each author in year Y as the sum of their abstracts from 1993 (the start of the observation) to year Y. I used tf-idf to generate a vector for each author in each year, based on the cumulative total of their writing up to that year. I then used cosine similarity to measure the similarity of each pair of authors in each year from 1998 to 2007.

[2] Publication and citation information are derived from ISI Web of Knowledge. Grant information was collected from the Sponsored Research Office at Stanford. Dissertation information was collected from UMI's dissertation database as well as Stanford Library's Collections.

I treated abstracts cumulatively because in many disciplines in the social sciences, scholars may publish one paper per year or per two years, and a non-cumulative measure would have suffered from significant noise. I began the measure in 1998 both in order to reduce noise in the early years, and because this allowed me to filter out repeated collaborative ties (see below). Descriptive statistics are as follows:

Observations	Min	Max	Mean	Standard Deviation
9,559,714	0	0.741	0.000682	0.00480

Table 1: Descriptive statistics for similarity metric

3.1 Validating the Similarity Measure

I conducted three assessments, via regression, of the measure to assess its validity. First, I predicted that two papers by the same author would be more similar, and used "same author" to predict similarity. Second, I used the number of keywords in common between two papers to predict similarity, predicting that papers with more keywords in common should be more similar. Finally, I predicted that two papers from the same department should be more similar than two from different departments. All regressions yielded expected results. Moreover, it is encouraging that the coefficient for shared department is smaller than that for shared author, which is as expected.

	Estimated Coefficient	Significance
Same author	2.588×10^{-2}	$p < 2 \times 10^{-16}$
Count of shared keywords	1.338×10^{-2}	$p < 2 \times 10^{-16}$
Same department	1.630×10^{-3}	$p < 2 \times 10^{-16}$

Table 2: Validation results - Estimated coefficients in separate OLS regressions predicting similarity

4 The Model

I introduced the similarity as a control variable in a logistic regression to predict new tie formation at the dyad level. I considered three different types of ties: co-authoring a publication, co-service on dissertation committees, and co-application for grants. All models described here were logistic regressions predicting the formation of a new tie in a given year. I focused on new tie formation because the decision to repeat a tie may take into account different factors (such as the pleasantness or success of a prior collaboration) than would a new tie. Additionally, I'm interested in the role of research-interest similarity, which may be influenced by prior collaborations; thus a study of repeated ties runs the risk

of reverse causality. In all models, I clustered the standard errors on person i of the ij tie, to account for autocorrelation in the yearly measurements. I used the first six years as the baseline, beginning my analysis in 1998. Doing so allowed me to identify prior ties (in the first five years), and thus to distinguish between new and repeated ties in my study period, 1998-2007.

4.1 Variables

Dependent variables: I used three dependent variables that each reflect substantive intellectual collaboration: a new tie formed between i and j in year t, for the activities of co-authoring publications, co-advising dissertations, and co-authoring grants.

Independent variables: My primary interest is the role of research-interest similarity in predicting new ties. I calculated similarity through tfidf, as described above. Plotting a histogram of the similarity distribution showed it to be either exponentially or power-law distributed. Thus, I took the natural log of similarity, and then standardized. In addition, I assess a measure of similarity based on i's and j's proportion of shared references in prior years. Shared referencing is a commonly used measure of intellectual similarity. Further, I assess the role of departmental co-membership and shared courtesy appointments; I am interested in assessing whether the inclusion of tfidf similarity markedly decreases the effect of shared departmental memberships.

Control variables: I included several variables to control for homophily on personal traits: same gender, difference in year of receiving highest degree, age difference, same ethnicity, same tenure status, and difference in year of appointment (McPherson et al. 2001). I also included a measure of degree centrality, to control for preferential attachment processes. This variable, degree conditioning, was measured as the sum of i's and j's degree centrality at year Y-1. This is similar to Ingram and Morris's (2007) strategy of including a fixed effect for each individual in a dyad. I included a measure of prior funding, calculated as the sum of i's and j's total funding through grants in the prior three years. Finally, I included yearly dummy variables to control for secular time trends in the data (these are not reported, in order to streamline the tables, but are available upon request).

5 Results

Results for each of the three dependent variables (publication, grant, and dissertation ties) are reported separately. Tables follow.

5.1 Publications

See Table 3. I began by estimating a baseline model, using only the control variables. In Model 2 I added same department, and in Model 3 added same department and courtesy department. Models 4 and 5 considered similarity and shared references respectively, adding one or the other to the control variables. In Model 6 I considered shared references and similarity together, and I find that the combination dampens the effects of each, but the difference is more pronounced for shared references than for similarity. In Model 7, I considered shared references together with shared department and courtesy. I find that the inclusion of shared references decreases the estimated coefficient for department and courtesy only very slightly. In Model 8, I include similarity as well, and find that the estimated coefficient for department decreases more substantially, to about 71% of its value in model 7. At the same time, in the model with department, the coefficient for similarity decreases relative to that estimated in Model 6, to about 83% of the prior estimate.

5.2 Grants

See Table 4. I estimated the same series of models for grant ties and dissertation ties. For grants, I found very similar results to those for publications. The estimate for department in Model 8 is 68% of that estimated in Model 7. The estimate for similarity in Model 8 is 81% that estimated in Model 6.

5.3 Dissertations

See Table 5. The results for dissertation were similar to those for publications and grants. The estimate for department in Model 8 was 75% of that in Model 7. The estimate for similarity in Model 8 was 75% of that in Model 6. Thus, for dissertations, the departmental effect seems somewhat more robust than for the other type of ties.

6 Discussion

In all three cases, I found that similarity was a positive predictor of new tie formation. Further, in all three cases, I found that including similari-

ty in the model had a dampening effect on the role of department in mediating new tie formation. However, even given this dampening, the coefficient on shared department ranged from 1.8 (for publications) to 2.7 (for dissertations). Because these coefficients report increase in log-odds, these translate to an increase of 6.0 times the baseline odds of co-publishing, and 14.5 times the baseline odds of serving on a dissertation committee together. Thus, even taking the similarity in research interests between department co-members into account, we can see that department retain an important role in promoting new ties.

By comparing the change in department from Model 7 to 8 (adding similarity) to the change from Model 3 to Model 7 (adding shared references), we can get a sense of scale. As described above, adding similarity to the model decreased the estimated effect of department on tie formation by about 30% across tie types. In contrast, adding shared references (comparing Model 3 to Model 7) decreases the estimated coefficient for department by 4% for publications, by 2% for grants, and by 2% for dissertations.

One possible reason for this difference in the size of changes between models is that being in the same department constrains what professors write about (and thus their tfidf similarity) more than it influences which references they cite. In other words, your choice of the references you cite may depend on your departmental affiliation less than your choice of research subject does. Looking at the correlations between these variables may help:

	Similarity	Shared References	Same Department
Similarity	1.0		
Shared References	0.141	1.0	
Same Department	0.139	0.106	1.0

Table 6: Correlations between independent variables

The correlations are all fairly similar, though indeed the correlation between same department and shared references is less than that with similarity. At the same time, on a theoretical level it seems unlikely that departmental membership constrains word use in abstracts more than it does references. Professors often don't cite articles that are closely related to their work if those articles are published in journals with which they're unfamiliar—for example, journals from outside their field. Moreover, the shared reference variable counts occurrences in which two people cite the exact same article; this makes it a fairly coarse measure. Two professors who cite closely related articles, or two different articles by the same author, will not show up as having a shared reference. In contrast, tfidf similarity is a much more fine-grained measure, and can capture relationships that shared references misses.

7 Extension 1: Investigating the Similarity Sweet-spot

I hypothesize that there may be a sweet-spot for similarity in tie formation. That is, it's possible that people choose to collaborate with those who are similar to them, but not too similar. I can assess this by adding a similarity-squared term to my analysis. If similarity-squared is negative, while similarity remains positive, the net effect would be a downward-facing parabola. In this case, similarity would be associated with increasing likelihood of collaboration up to a point, beyond which, greater levels of similarity decrease a dyad's chance of collaborating.

I addressed this question by repeating my above model (without degree conditioning and prior award), adding a variable similarity squared. I generated this variable by squaring the log-normalized similarity value. Results are presented in Table 6. Similarity squared is not significant for any of the three types of ties studied here: the data fails to support the sweet-spot hypothesis.

8 Conclusion

It is clear from the data presented that, within this dataset, academic departments play a substantial role in promoting new tie formation on publications, dissertations, and grants. This effect is largest for dissertations, which is not surprising, as dissertation committees are often governed by rules stipulating a minimum number of department members who must be involved. The effect remains large for both publications and grants. This is particularly interesting as Stanford, the site of our research, has a reputation for its "low walls" between departments and schools, facilitating interdepartmental collaboration.

In future work, it would be interesting to investigate further the mechanisms through which departments promote new ties. These may include: face-to-face meetings, such as regular faculty meetings or faculty lunches; presentations of research, such as lab meetings or informal col-

loquia; physical co-location, such as having offices near each other. The first two of these may be difficult to measure, though a qualitative investigation (for example, interviewing professors about how they find collaborators) may shed light.

It may be more feasible to investigate the role of office proximity. Using old directories, I may be able to locate professors' offices for the years in my study. Through GIS, I can measure the geographic distance between each pair of scholars. In similar research, conducted in an architectural firm, Owen-Smith found that desk location has significant effects on social ties and productivity (2009).

This would allow me to ask several interesting questions:

- Does having offices near each other make professors more likely to collaborate?
- If so, how much of the observed departmental effect can be attributed to office proximity?
- How does office proximity relate to research interest similarity? Are these measures correlated?
- How do office proximity and research similarity co-vary over time? Does moving your office closer to someone make you likely to develop research interests more similar to theirs?

In recent years, Stanford has built several large new buildings housing offices and labs for groups of professors doing related work. In addition, due to infrastructural decay in the Terman Engineering Building, several whole departments have been forced to move into these new spaces (in other cases, only some members of a department might choose to move into the new space). This movement, and variation in movement, provides the opportunity for interesting "natural experiments" on the effect of co-location on research focus, and on similarity.

Reference

Abbott, Andrew. 2001. *Chaos of Disciplines*. Chicago: University of Chicago Press.

Blau, Peter M. 1973. *The Organization of Academic Work*. New York: Wiley.

Friedkin, Noah E. 1978. "University social structure and social networks among scientists." *American Journal of Sociology* 83 (6): 1444-65.

Gumport, Patricia J. and Stuart K. Snydman. 2002. "The Formal Organization of Knowledge: An Analysis of Academic Structure." *The Journal of Higher Education* 73 (3): 375-408.

Hearn, James C. 2007. "Sociological Studies of Academic Departments." Pp. 222-65 in *Sociology of Higher Education: Contributions and their Contexts*, edited by Patricia J. Gumport. Baltimore, MD: The Johns Hopkins University Press

Ingram, P., M.W. Morris. 2007. "Do people mix at mixers? Structure, homophily, and the 'life of the party.'" *Administrative Science Quarterly* 52: 558-585

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27: 415-44.

Owen-Smith, Jason. 2009. "Toward a Behavioral Network Theory: Networks, Institutions, and Space." Presentation to the American Sociological Association, Aug 8-11, San Francisco, CA.

Peterson, Marvin W. 2007. "The Study of Colleges and Universities as Organizations." Pp. 147-86 in *Sociology of Higher Education: Contributions and their Contexts*, edited by Patricia J. Gumport. Baltimore, MD: The Johns Hopkins University Press.

Pfeffer, Jeffrey, and William L. Moore. 1980. "Average Tenure of Academic Department Heads: The Effects of Paradigm, Size, and Departmental Demography." *Administrative Science Quarterly*. 25(3): 387-406.

Research-Interest Similarity as a Measure of Scientific Paradigm Development: A Preliminary Investigation

Susan Biancani
Stanford University School of Education
Stanford, CA USA
biancani@stanford.edu

Abstract

The notion of a "hierarchy of sciences", in which academic fields can be ordered from "hard" to "soft", is an old one. However, efforts to develop a measure of hardness of a field, or of the field's level of paradigm development, to date have not been highly successful. Here I explore the possibility of using a text-based similarity measure to quantify the extent of consensus in a field, which is theorized to correlate with hardness.

9 Introduction

Thomas Kuhn's 1962 *The Structure of Scientific Revolutions* advances the notion that a scientific fields is characterized by a paradigm— a guiding set of assumptions, methods and values that shape how research in the field is conducted and evaluated, as well as what constitutes appropriate objects of study. Moreover, Kuhn argues that different fields are characterized by different levels of paradigm development. In low-paradigm fields, little consensus exists on the important questions in the field or the best methods with which to investigate them; research proceeds in fits and starts, and new findings may not build directly on prior findings. This description tends to fit fields in the social sciences. By contrast, high-paradigm fields—often those in the natural sciences—show much greater agreement on methods and research questions; there is often a race to publish important results, out of fear of getting "scooped." New findings build directly on—or challenge—prior findings, allowing knowledge to accumulate rapidly. High-paradigm fields have elsewhere been described as "high-consensus", "rapid-discovery", and "progressive" (Collins, 1994).

Sociologists of science have attempted to characterize fields according to their level of paradigm development, but many efforts to date have not been satisfying. Here I explore the use of a text-based similarity metric as a measure of the level of cohesion—and thus paradigm development—in a field.

10 Prior Work

Auguste Comte first advanced the notion of a "hierarchy of sciences" in the nineteenth century. Since then, much work on this question has come from the field of bibliometrics. Derek de Solla Price developed an "Immediacy Index", which showed faster rates of obsolescence of findings in the natural sciences than in the social sciences (1970). However, this metric was later shown to be an artifact of the differing volumes of work produced in a given time interval in different fields (Cole et al. 1978). Cole (1983) summarized findings from seven different approaches seeking to find a variable that reliably correlated with widespread perceptions of paradigm development, but found none that did; he concluded, "there are no systematic differences between sciences at the top and at the bottom of the hierarchy in either cognitive consensus or the rate at which new ideas are incorporated" (111).

Promising work comes from Susan Cozzens (1985), who demonstrated an intriguing pattern in a detailed qualitative study. Cozzens compared citations of two highly influential papers: one in neuropharmacology, and one in sociology of science. Cozzens finds that citations to the neuropharmacology paper varied over time: early on, citations mentioned either the main finding or other peripheral findings, and frequently commented on experimental technique. Later papers developed a formulaic citation of the main finding, suggesting that this finding had been vetted, and was now taken as fact. In contrast, no such shift was observed in citations of the sociology of science paper. Cozzens attributes this to the fact that very few citing papers at any time referred to the main finding; more often, the paper was abstracted, and was mentioned as an exam-

ple of a larger trend. Cozzens findings are enlightening, but as a close, qualitative study cannot easily be extended to new fields.

One technique that has successfully distinguished low-paradigm from high-paradigm fields, and which has the potential to be replicated automatically, is the measure of "fractional graph area" (FGA). FGA is a measure of the total fraction of page space in a given article that is taken up by graphs. Smith et al. (2000) hypothesized that papers from higher-paradigm fields would be characterized by higher FGA. In doing so, they drew on Latour's assertion that graphs distinguish science from non-science (1990). Graphs are a highly encoded means of communication; they can present a large amount of information in a compact form, because they build on a vast quantity of shared knowledge between the writer and the reader. Much information is embedded in a graph without elaborate explanation; it is assumed that the reader has sufficient prior familiarity with the form of a graph to be able to extract the new finding quickly. Thus, the use of graphs captures much of the nature of a high-paradigm field. In a random sample of 50 articles from each of 30 journals, Smith et al. found that FGA does indeed correlate with scientists' perceptions of the level of paradigm development of seven fields.

Smith et al. relied on prior coding by William Cleveland (1984) who measured the FGA of the papers used in the sample. Cleveland describes the process as "detailed and intensive" (261). Clearly, it would be useful to develop an automated measure of paradigm development.

Cole offers an interesting insight into the difference between high- and low-paradigm fields (1994). He argues that knowledge in each discipline can be divided into the core and the frontier. The core "consists of a small group of theories, methods, and exemplars that are almost universally accepted by the relevant scientific community as being both true and important", while the frontier consists of new all new knowledge (133). Some of the frontier will eventually move into the core, but most will not. "The problem with fields like sociology," Cole (a sociologist) argues, "is that they have virtually no core knowledge…There seems to be no sociological work that the great majority of the community will regard as both true and important" (134). This notion of core knowledge captures the essence of paradigm: it is the body of knowledge all members of the discipline are presumed to share; it is the knowledge taken for granted, and treated as background, by cutting-edge publications in the field.

Here, I explore the potential of using person-person similarity, based on the text of published work. Comparing texts directly in this way has the potential to capture the extent of shared, "core" knowledge in a field. Papers will vary in their use of terms related to frontier knowledge, but they should share a certain amount of core knowledge, and this should be reflected in overlapping term use.

The similarity metric I use is the cosine similarity of the tf-idf vectors representing the abstracts from professors' published work (see prior paper for details). In this investigation, I summed each professor's work over the entire time period, rather than using a yearly similarity measure. Because high-paradigm fields are characterized by higher levels of consensus on research interests and methods than low-paradigm fields—because they share a larger body of core knowledge—I hypothesize that they will show greater person-person similarity, on average, than low-paradigm fields.

11 Data

The data for this paper are drawn from the same corpus as the prior paper. Here, I restrict the study to four departments: physics, biology, psychology, and sociology. Smith et al. and other researchers have found that these fields are widely perceived to be ranked for paradigm development in the above order, with physics showing the highest level of development, and sociology the lowest (Lodahl and Gordon 1973; Ashar & Shapiro 1990; Biglan 1973).

Although the publications in our database have undergone extensive efforts at author disambiguation, I did identify some errors. For example, Professor Xeuguang Zhou of sociology was matched to another X. Zhou in physics, leading to a number of misattributed articles. While it would have been too laborious to hand-correct such errors throughout the entire corpus of 66,000 abstracts used in the prior study, for this smaller corpus, I did attempt to filter out incorrect matches.

First, I filtered according to the round of name-matching on which an author was assigned to a paper. The matching algorithm completed twenty rounds of matching, and earlier rounds are more reliable than later rounds. Thus, I took only papers matched on rounds 1-15. Then, I ex-

amined the complete list of author-keyword pairs for each department. Any author-keyword pair that arose in two or fewer articles, I checked by hand. I looked for another author in the Stanford database with the same last name, who might be in a more plausible department. If I didn't find any, I searched for the professor's CV or other information pertaining to research interests on the web. If a CV was available, I searched for the publication in question, or any others in the same journal, or on a closely related topic. If the paper was not listed on the CV, and no closely related papers were, I removed it from the given department's sample. In general, this was a fairly straightforward task; there was little "gray area" in these decisions.

This resulted in a corpus of 191 professors and nearly 6,000 abstracts.

Department	People	Publications	Keywords
Physics	43	1653	63
Biology	67	2912	134
Psychology	47	1192	109
Sociology	34	235	46
Total	191	5992	352

Department	Publications per Person	Keywords per Person	Keywords per Publication
Physics	38.44186	1.465116	0.038113
Biology	43.46269	2	0.046016
Psychology	25.3617	2.319149	0.091443
Sociology	6.911765	1.352941	0.195745

Table 1: Descriptive Statistics for the Corpus[3]

A few facts stand out about the distribution of papers in the corpus. First, sociology produced far fewer papers in the study interval than the other departments, even relative to its smaller size. Additionally, sociology has by far the greatest diversity of keywords relative to the number of papers produced. In general, it seems that the high-paradigm fields (physics, biology) publish more papers per person than the lower-paradigm fields (psychology, sociology). Additionally, the higher-paradigm fields use a smaller set of keywords, relative to the volume of output, than the lower-paradigm fields.

[3] Note: the ratios reported are the total number of publications or keywords for the whole department, divided by the total number of people or publications. In this respect, they give us a sense of the diversity or dispersion of the department, rather than indicating how many keywords are typically associated with an article in a given discipline.

11.1 Average Similarity by Department

I used three approaches to assess whether the higher paradigm fields were characterized by greater consensus, as reflected by greater similarity. First, I calculated the average person-person similarity in each department. It is possible that the dispersion of a department would increase as a function of the departmental size. That is, as more members join a department, the average distance between members increases. To account for this possibility, I took a random sample of 25 people from each of the larger three departments; I chose 25 because this is the size of the smallest department, sociology. I recalculated the average person-person similarity of the department using this sample of 25 department members. Finally, in another attempt to account for the larger size and greater productivity of the natural science departments, I chose a random sample of 46 keywords (again, because this is the smallest value, from sociology) from each department's list of keywords. I then created a subset of papers published in each department, including only papers using one or more of these keywords. I repeated this sampling procedure four times. I then calculated the average person-person distance on the basis of this subset of papers. Here are the results:

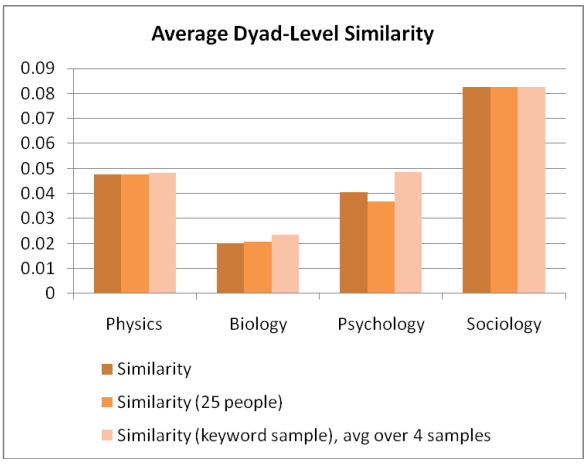


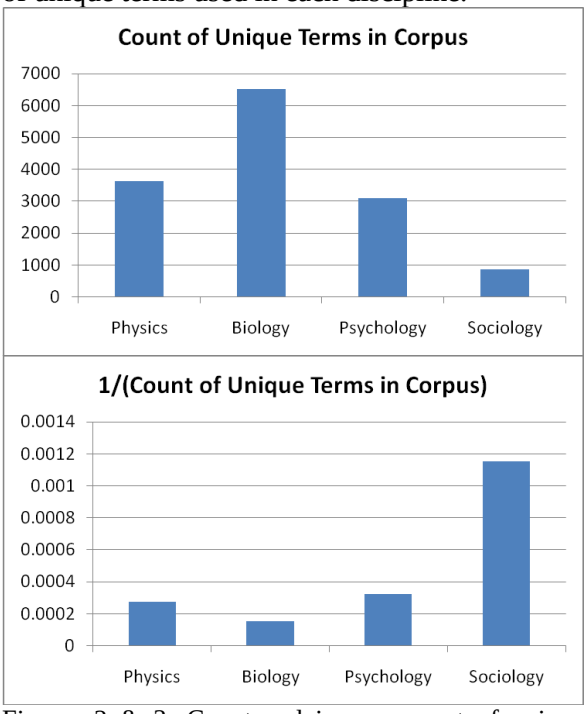
Figure 1: Average similarity between each pair of professors in a department

First, it is worth noting that the three different sampling techniques yielded broadly similar results. Thus, it appears the simply using person-person similarity from the whole department is a reasonable approach.

On the whole, these results are surprising. I had predicted that physics, as the highest-paradigm field, would have the greatest similarity among professors, and that sociology, as a low-paradigm field, would have the least. In fact,

these results do not accord with the pattern seen for keywords. Sociology had the greatest range of keywords; this would seem to predict that it should have lower levels of overlap between professors' abstracts.

One possibility is that these results reflect a quirk of tf-idf. This approach gives the greatest weight to rare words. Because physics and biology are technical fields, they may have more rare (i.e. specialized and technical) words in their lexicon than the social sciences. I can investigate this possibility with a count of the total number of unique terms used in each discipline.



Figures 2 & 3: Count and inverse count of unique terms in each corpus

The values obtained for the similarity measures above are very similar to the inverse of the count of unique terms used in each discipline's corpus. Thus, it appears that the tf-idf similarity measure is primarily reflecting the diversity of terms used in each discipline. We can confirm this observation in a slightly different way: by examining how many new terms are added to the corpus with additional papers. To investigate this, I counted the number of new terms added in each consecutive set of 50 papers.

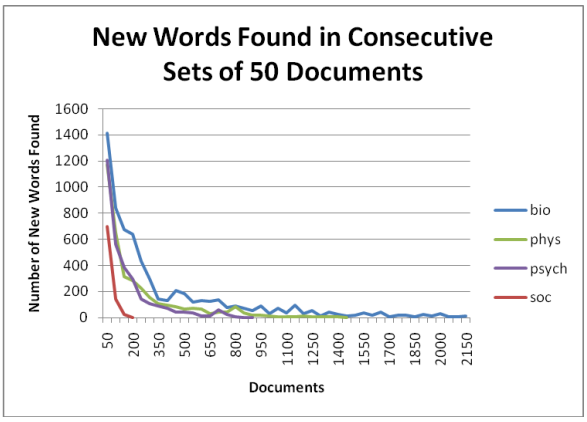


Figure 4: Count of new words found in consecutive sets of 50 papers

Here, it is clear that although sociology has 235 papers, almost all of these are captured in the first 150 papers. The slope of the curve for sociology is similar to that for the other departments, but even the first 50 papers examined contain fewer terms than 50 papers from another field. Meanwhile, biology looks qualitatively different from other departments; first, it has many more papers than other departments (75% more than physics, the second-most productive department). Additionally, biology keeps adding new terms at a rate of more than 100 per 50 papers until after 700 papers have already been added. Physics and psychology look very similar.

As noted, tf-idf measures are very sensitive to rare terms. The more rare terms in a field—and the greater the total size of the lexicon of that field—the less overlap there will be in term use between field members, all other things being equal. I can attempt to correct for this sensitivity by multiplying similarity by the number of terms used by the field in total. Doing so will give me a measure of person-person similarity, relative to the size of the field's lexicon.

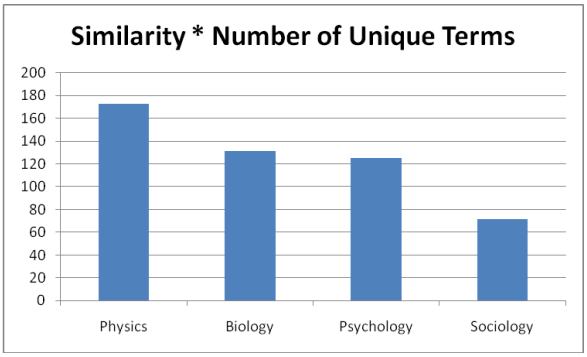


Figure 5: Dyadic similarity multiplied by number of terms in a department's corpus

Here we see the expected ordering of disciplines, although it is surprising that biology and psychology are so similar.

12 Conclusion

Text-based tf-idf similarity appears to be a promising option with which to measure the level of paradigm development of a scientific field. In order for this to be a useful metric for a wide array of applications, it needs to be easily repeatable on other datasets. Certainly, when abstracts are digitized, tokenizing them and calculating tf-idf-based similarity is feasible. The most significant challenge in repeating this analysis on other datasets may be handling author disambiguation. It would be interesting to reconstruct this dataset using more rudimentary name-matching techniques, and repeat the analysis to see how much author disambiguation affects the results.

If the measure is robust to some noise in author disambiguation, I would like to repeat the study on a larger dataset, for example all of the ISI database for the fields in question (rather than the Stanford authors only), to determine whether the findings hold.

Reference

Ashar, Hanna and Jonathan Z. Shapiro. 1990. "Are Retrenchment Decisions Rational? The Role of Information in Times of Budgetary Stress." *Journal of Higher Education*. 61: 123-41.

Biglan, Anthony . 1973. "Relationships Between Subject Matter Characteristics and the Structure and Output of University Departments." *Journal of Applied Psychology*. 57: 204-13.

Cleveland, William S. 1984. "Graphs in Scientific Publications." *American Statistician.* 38: 261-69.

Stephen Cole. 1983. "The Hierarchy of the Sciences?" *American Journal of Sociology*. 89: 111-39.

----------. 1994. "Why Sociology Doesn't Make Progress Like the Natural Sciences." *Sociological Forum.* 9 (2): 133-154.

Cole, Stephen, Jonathan R. Cole and Lorraine Dietrich. 1978. "Measuring the Cognitive State of Scientific Disciplines." In Yehuda Elkana, Joshua Lederberg, Robert K. Merton, Arnold Thackray and Harriet Zuckerman (eds), *Toward a Metric of Science*. New York: John Wiley & Sons, 209-51.

Collins, Randall. 1994. "Why the social sciences won't become high-consensus, rapid-discovery science." *Sociological Forum*. 9 (2): 155-177.

Susan E. Cozzens. 1985. "Comparing the Sciences: Citation Context Analysis of Papers from Neuropharmacology and the Sociology of Science." *Social Studies of Science*. 15 (1): 127-53.

de Solla Price, Derek J. 1970. "Citation Measures of Hard Science, Soft Science, Technology, and Non Science", in Carnot E. Nelson and Donald K. Pollack (eds), *Communication Among Scientists and Engineers* Lexington, MA: D.C. Heath, 3-22.

Kuhn, Thomas. 1962. *The Structure of Scientific Revolutions*. Chicago: The University of Chicago Press.

Latour, Bruno. 1990. "Drawing Things Together", in Michael Lynch and Steve Woolgar (eds), *Representation in Scientific Practice* Cambridge, MA: MIT Press, 1990: 19-68.

Lodahl, Janice Beyer and Gerald Gordon. 1972. "The Structure of Scientific Fields and the Functioning of University Graduate Departments." *American Sociological Review*. 37: 57-72.

Smith, Laurence D., Lisa A. Best, D. Alan Stubbs, John Johnston, Andrea Bastiani Archibald. 2000. "Scientific Graphs and the Hierarchy of the Sciences: A Latourian Survey of Inscription Practices." *Social Studies of Science*, 30 (1): 73-94.