Susan Biancani
CS224W Project Proposal
October 20, 2010

The Problem:

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. 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 (Biancani et al. 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.

This is a complicated question, because departments both contain people with similar training and research interests, and bring these people into fairly regular contact with each other. Two competing hypotheses are possible:

> Hypothesis 1: Departments collect people with similar research interests. These people would collaborate in the absence of departmental co-affiliation. *The homophily hypothesis.*
>
> Hypothesis 2: Departments add value by bringing people into more contact with each other than they would otherwise have. *The propinquity hypothesis.*

In order to test the relative merits of these two hypotheses, we need to disentangle the effects of similar research interests, and shared membership in administrative units. One useful resource in this task is each scholar's body of written work. Published articles provide a "paper trail" of their author's research interests. Comparing each scholar's body of written work to that of her colleagues allows us to infer similarities (and differences) among scholars—both those who share a department, and those who don't. I propose to compute a suitable similarity measure, based on written work, for each pair of faculty members at Stanford, and to use this to help distinguish between the homophily hypothesis and the propinquity hypothesis.

The Data:

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:

[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.

- 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 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)

For Part 1 of this project, I will use a corpus of 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.

The corpus contains 66,000 abstracts. The vast majority (50,000) have a single author within the Stanford faculty network (co-authors from outside Stanford, and those who are not faculty members, have been excluded from the analysis). A further 11,000 have two Stanford authors, and 3,000 have three, leaving 2,000 with more than three authors. This distribution is summarized in Figure 1.

Co-authorships present an interesting conundrum. In the year that two people are co-authors on the same paper, they will necessarily be very similar—in part because they are likely to share research

[2] Publication and citation information are derived from ISI Web of Knowledge. Patent information is collected from the United States Patent and Trademark Office (USPTO). Grant information was collected from the Sponsored Research Office at Stanford. And dissertation information was collected from UMI's dissertation database as well as Stanford Library's Collections. ISI and the USPTO are widely used and highly esteemed data sources for studies of knowledge creation and innovation (Stuart and Ding 2006; Leahey 2007; Jones, Wuchty, and Uzzi 2008)

interests, and in part as an artifact of attributing the same abstract to both of them. For this reason, it's important to allow similarity to vary over time. This was, I can ask questions like, How does similarity at time t predict co-authorship at time t+1?

Doing the Project, Part (a): Deriving a Similarity Measure

In broad strokes, I would like to characterize each author as a point in multidimensional space, on the basis of their written work. I want to do so in a way that allows an author's location to vary over time. Then, for each time-point in my study, I will calculate the person-by-person similarity for every faculty member in my dataset. These similarity measures can then be used as a control variable in a regression predicting the presence of collaborative ties. I plan to try two different approaches to place authors in n-space, which I can then compare: term frequency-inverse document frequency (tf-idf), and latent Dirichlet allocation (LDA).

Tf-idf is described by Manning, Raghavan and Schütze (2009). The term frequency of a term t in a document d is computed as the count of the number of times the term occurs in the document and is written as $tf_{t,d}$. Two documents that contain many terms in common will have a high tf. However, tf does not account for words that occur frequently in all documents, and thus are less meaningful. For this reason, document frequency is a popular correction factor. The document frequency df_t of a term t in a collection is the count of the number of documents in the collection that contain the term. The term frequency and the inverse of the document frequency are combined to create a composite weight for each term t in each document d:

$$tf\text{-}idf_{t,d} = \frac{tf_{t,d}}{df_t} = tf_{t,d} \times idf_t$$

Because $tf_{t,d}$ is computed as a count over a document, it's easy to adapt this to calculate $tf_{t,a}$, the frequency of term t over all documents written by author a, by defining each author to be the set of all documents he has written. The tf-idf of a term t in the writing of an author a is: [3]

$$tf\text{-}idf_{t,a} = \frac{tf_{t,a}}{df_t} = tf_{t,a} \times idf_t, \;\; tf_{t,a} = \sum_{d \in a} tf_{t,d}$$

Similarly, I can compute an author's tf-idf for a subset of years in my study by summing tf-idf only over those years. In order to allow an author's tf-idf vector to vary over time, one possibility would be to generate a different vector in each year for each person. However, many faculty members publish 0 or 1 papers in many years, so slicing this finely may introduce a lot of noise into my distance measures. Another possibility would be to treat similarity as a cumulative measure. To do so, I would first slice the data by person and into two time chunks: 1993-1998, and 1999-2007[4]. The pre-1999-per-person vectors would be used to calculate a person-by-person starting similarity. For each subsequent year y, I

[3] By convention, tf-idf is usually written as tf x idf, rather than as tf/df.

[4] The selection of 1998 is somewhat arbitrary, and subsequently, I could investigate whether choosing a different starting year affects my results.

would add y to the starting vector, and recalculate similarity. So, an author's vector in 1999 would be based on all papers written 1993-1999; the vector for 2000 would include all papers written 1993-2000, and so on. This measure conceives of each person's location in n-space as reflecting the total of their past work. It can allow for change over time, but without forcing a person to be defined by a single paper.[5]

Computing the tf-idf for all terms in the corpus, for all authors yields a term-by-author matrix for each time-point, in which each author is represented by a vector of tf-idf values of length n = (number of terms in the corpus). I can compute the distance between two authors as the distance between their vectors. To do so, I will use cosine similarity, to account for the fact that my documents (and each author's set of documents) can vary in length. Thus instead of computing the absolute (Euclidean) distance between vectors, I will compute the cosine of the angle between them. The cosine also has the convenient property of varying from 0 (orthogonal vectors) to 1 (identical vectors); that is, the more similar two vectors are, the greater their cosine. In contrast, the more similar two vectors are, the smaller their Euclidean distance will be, making it less intuitive to use as the weight of an edge between two nodes in a network.

This is one reasonable approach I can use to define the similarity of two scholar's research interests (i.e. their research-interest homophily): the cosine similarity of the tf-idf vector based on their written work. There are two disadvantages to using tf-idf to define my similarity measure. First, tf-idf does not provide a great deal of dimension-reduction: each author has a vector of length n = (number of terms in the corpus). In a corpus of 66,000 abstracts covering a wide range of disciplines, this can be a lot of dimensions. Second, tf-idf does not account for words that have multiple meanings. All homonyms are collapsed into one term. This may present a problem in my multi-discipline corpus, where words may have very divergent meanings in different disciplines.

One approach that addresses both of these concerns is LDA, which reduces the number of dimensions to the number of topics in the corpus (a value determined by me at run-time, and always significantly less than the number of terms). While LDA doesn't take different word senses as an input, it does allow a single term to belong to multiple topics. In this way, identical terms in two different documents may load on different topics. Because the topic loading determines document- or author-similarity, LDA decreases the chance that homonyms will falsely inflate the similarity of two documents or authors.

LDA is a generative probabilistic model that characterizes each document in a corpus of documents as a mixture of topics, and thus can be used to identify themes and trends in the corpus (Blei, Ng and Jordan 2003). Briefly, LDA uses a "bag of words" model, in which a document is treated as a collection of word frequencies, ignoring information about the word's position in a document, grammar and syntax. The procedure inductively identifies words that tend to co-occur, and groups these together into topics. The number of topics is defined by the user at the outset. LDA imagines each document in the corpus as a

[5] A slight variant of this would conceive each person's topical identity as being based on her work in the last k years (perhaps k=5). So a person's topic vector in 2000 would use their papers from 1996-2000. This approach assumes work greater than k years old is no longer relevant.

distribution over topics, and each topic as a distribution over words. To generate a document, the model first chooses a topic from the distribution of topics, and then chooses a word from the distribution of words in that topic. Topics are selected through an iterative algorithm that seeks to maximize the probability of the observed corpus.

The Stanford NLP group has published a Topic Modeling Toolkit (TMT) to facilitate the use of LDA (Ramage et al. 2009; http://nlp.stanford.edu/software/tmt/tmt-0.3/). I plan to use this toolkit on my corpus of abstracts to assign to each document a distribution over the set of topics in the corpus. The topic distribution for document i can be expressed as:

$$d_i = (z_{i1}, z_{i2}, \ldots z_{in}), \quad \sum_{k=1}^{n} z_{ik} = 1$$

where z_k is the probability that a word w in document i was drawn from topic k, and n is the total number of topics in the model.

The use of LDA on academic abstracts is well established (Rosen-Zvi et al. 2004, Chang and Blei 2009). The toolkit allows users to group documents into "slices" after training, and sums the distribution of topics in each slice. I propose to slice the corpus by author, thus computing a vector of topic loadings per author. The distribution for an author a can be expressed with identical notation to the distribution for a document:

$$d_a = (z_{a1}, z_{a2}, \ldots z_{an}), \quad \sum_{k=1}^{n} z_{ak} = 1$$

Again, I plan to treat an author's work cumulatively. I will first slice the data by person and into two time chunks: 1993-1998, and 1999-2007, and run the TMT slicing procedure[6]. The pre-1999-per-person vectors would be used to calculate a person-by-person starting similarity. I'll then rerun the TMT slicing procedure for each year y, slicing at year y. An author's vector in 1999 would be based on all papers written 1993-1999; the vector for 2000 would include all papers written 1993-2000, and so on. Finally, I'll recalculate similarity among all author-author pairs in each year.

Because I will recalculate the similarity between every pair of nodes 9 times (for years 1999-2007), the lower dimensionality of the LDA vectors will be an advantage over the higher-dimension tf-idf vectors. On the other hand, generating the LDA vectors in the first place will likely take longer than generating the tf-idf vectors.

Doing the Project, Part (b): Testing Homophily vs. Propinquity

[6] The selection of 1998 is somewhat arbitrary, and subsequently, I could investigate whether choosing a different starting year affects my results.

Now that I have identified a measure of the similarity between each pair of authors in the network, I can use this to investigate how much departmental co-membership promotes new tie formation, net of similarity.

I propose to use a logistic regression model on the dyad-level, to predict the presence of a collaborative tie between two individuals. I will estimate the model three times, once for each type of collaborative tie: co-authoring a publication, co-application for a grant, and co-service on a dissertation committee. I'll also repeat it using each of my two similarity measures.[7] Each of these three activities is a substantive part of the work of academics. Moreover, departments may promote them through different mechanisms; for example, homophily in research interest may matter more for co-authoring of publications, while propinquity may matter more for dissertation co-service.

I am primarily interested in how departments promote the formation of new ties, rather than promoting repeated ties. Thus, it makes sense to restrict my study to new ties. To do so, I can start my analysis in 1999, but use the years 1993-1998 to look for prior ties between i and j. Ties will be "at risk" in year t if they haven't existed in any year prior to t. If a tie between two individuals exists in the 1993-1998 span, or if it later forms, it will be dropped from the risk pool for subsequent years. Such a restriction also prevents against circular causation, in which an early tie both encourages people to write about similar topics, and causes them to collaborate again.

My dependent variable will be whether individuals i and j form a new tie of the given type in year t. The model I propose is:

$$Tie_{ijt} = \beta_0 + \beta_1 D_{ijt} + \beta_2 s_{ijt} + controls$$

where D_{ijt} is a binary variable indicating whether i and j were (non-courtesy) members of the same department in year t, s_{ijt} is the similarity between i and j in year t (based on tf-idf of LDA), and "controls" includes other relevant dyad-level control variables. Variables, including potential controls, are summarized in Table 1.

I can compare this to a baseline model that does not include similarity, and to a model that includes an interaction term (to see whether similarity matters more or less for two people in the same department):

$$Tie_{ijt} = \beta_0 + \beta_1 D_{ijt} + controls$$

$$Tie_{ijt} = \beta_0 + \beta_1 D_{ijt} + \beta_2 s_{ijt} + \beta_3 D_{ijt} \times s_{ijt} + controls$$

By comparing across these nested model, I can determine whether the previously observed effect of departments is altered by the inclusion of similarity. I contend that the observed power of academic

[7] I have not yet identified a metric for evaluating which similarity measure is better. At minimum, I can see if they're very different, or if changing between them affects the results of my regression. One possibility for evaluating the two measures is to see which better predicts co-publication ties. However, I hesitate to use that here, as this is one of the outcomes I'm trying to test for.

departments in promoting collaborative ties stems from two sources: homophily (the fact that two people in the same department have more similar research interests than two random people in the university), and propinquity (the fact that two people in the same department come into contact more than two random people in the university). By using topic-based-similarity as a proxy for research-interest-homophily, I hope to disentangle the relative importance of these two influences.

References

Biancani, Susan M., Daniel A. McFarland, Linus Dahlander and Lindsay Owens. "Interdisciplinary Super-Centers and the Transformation of the American Research University." Submitted August, 2010 to the *American Journal of Sociology.*

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet allocation." *Journal of Machine Learning Research*, 3: 993-1022.

Celikyilmaz, Asli, Dilek Hakkani-Tur and Gokhan Tur. 2010. "LDA Based Similarity Modeling for Question Answering." Procedings of the Workshop on Semantic Search, North American Chapter of the Association for Computational Linguistics - Human Language Technologies. June 5, 2010, Los Angeles, California.

Chang, Jonathan and David M. Blei. 2009. "Relational Topic Models for Document Networks." Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5.

Daniel Ramage, Evan Rosen, Jason Chuang, Christopher D. Manning, and Daniel A. McFarland. 2009. "Topic Modeling for the Social Sciences." In *NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond*.

Jones, Benjamin F., Stefan Wuchty, and Brian Uzzi. 2008. "Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science." *Science* 322: 1259-62.

Leahey, Erin. 2007. "Not by Productivity Alone: How Visibility and Specialization Contribute to Academic Earnings." *American Sociological Review* 72 (4): 533-61.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. Banff, Alberta, Canada.

Stuart, Toby and Waverly Ding. 2006. "When Do Scientists Become Entrepreneurs? The Social Structural Antecedents of Commercial Activity in the Academic Life Sciences." *American Journal of Sociology* 112 (1): 97-144.

Figure 1. Count of publications in the Mimir Corpus with a given number of authors. The vast majoring (~50,000) have a single author at Stanford.

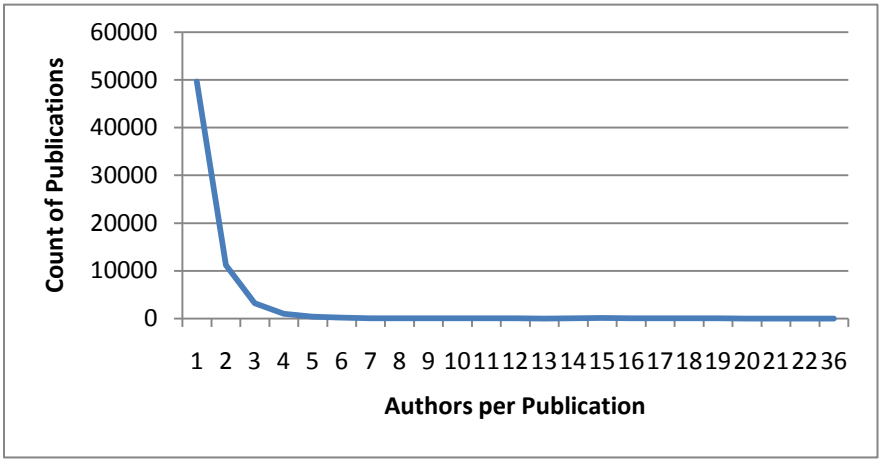

TABLE A-1. Dyad-level Variables: Definition and Description

Variable	Definition
Dependent variables (new ties)	
Dissertation Co-Advising	Tie formed between i and j through a dissertation committee
Grant Co-Authoring	Tie formed between i and j through a jointly proposed grant
Publication Co-Authoring	Tie formed between i and j through a co-authored publication
Shared Referencing	Tie formed between i and j through citing the exact same reference
Independent variables	
Same department	i and j work in the same department
Topic similarity	Weighted value derived from LDA analysis
Shared membership in Bio-X	i and j are involved in Bio-X
Shared membership in Woods	i and j are involved in Woods
Controls	
Courtesy department	i and j are linked through a courtesy appointment
Proportion shared references	Count of shared cites between i and j, divided by row minimum of i and j's total number of cites
Same gender	i and j have the same gender
Same ethnicity	i and j have the same ethnicity
Same tenure status	i and j have the same tenure status
Status difference	Absolute difference in appointment year between i and j
Degree conditioning - dissertations	i and j's combined degree centrality in the dissertation network at time t-1
Degree conditioning - grants	i and j's combined degree centrality in the grant network at time t-1
Degree conditioning - publications	i and j's combined degree centrality in the publication network at time t-1
Total grant awards prior 3 years	Sum of i's and j's funding through grants in the prior 3 years