Scientific Authorship, Collaboration, Interdisciplinarity, and Productivity

Jonathan R. Karr, Jake J. Hughey, Tim K. Lee
Department of Bioengineering, Stanford University, Stanford CA, 94305

December 12, 2008

Traditionally the path to scientific success has been a road of intense focus and specialization. Over the last few decades, an alternative model of scientific success has emerged – that of collaboration and interdisciplinary exchange. It has remained unclear, however, which strategy is optimal, and at what times during a scientist's lifetime. Separately, calls for a comprehensive digital author identifier system have become increasingly frequent. Here we propose a solution to the digital author identifier problem, and computationally construct a dataset consisting of scientists, relationships among them, and their publications and grants. Next we use this dataset to evaluate interdisciplinary as a career strategy. Additionally we build a web-based interface to our dataset. We show that scientific productivity increases exponentially with a scientist's interdisciplinarity and that of their collaborators. We conclude that high interdisciplinarity is an advantageous strategy for senior scientists, whereas intense focus is optimal for young scientists.

Availability: http://covertlab.stanford.edu/projects/ScienceGenealogy.

1 Introduction

Traditionally the path to scientific success has been a road of intense focus and specialization. Over the last few decades, an alternative model of scientific success has emerged – that of collaboration and interdisciplinary exchange. It has remained unclear, however, which strategy is optimal, and at what times during a scientist's lifetime.

Separately, over the last several years, the demand for unique digital author identifiers which accurately link scientists to their publications, has exploded. In the last year, six articles published in *Nature* advocated the creation of a comprehensive digital author identifier system which transcends non-unique scientist names, misspellings, and inconsistent transliteration. The demand for digital author identifiers is strongest in the Chinese and Indian scientific communities which suffer most from inconsistent transliteration and non-unique names.

First, we propose a solution to the digital author identifier problem, and create a dataset of scientists, the professional relationships among them, and their publications and grants. Second we use our scientist dataset to investigate how interdisciplinary research impacts a scientist's productivity, defined as the numbers of publications a scientist has authored and grants they have won.

We solve the digital author identifier problem by 1) learning scientists from the NIH publication and grant records, and 2) assigning unique and stable digital identifiers to each learned scientist. Second we learn adviser-advisee and collaboration relationships between scientists. Third we compute the distribution of pairs of MeSH headings over publications and scientists. Next we calculate each scientist's interdisciplinarity, or the negative average mutual information of pairs of MeSH headings associated with each scientist, and the average of that of their advisees and collaborators. Fifth we investigate scientific productivity as a function of interdisciplinarity. Finally we build a web-based interface to our dataset. The web-based interface displays the unique identifier computed for each scientist, and provides permanent hyperlinks using this identifier.

We show that a scientist's productivity increases exponentially with their interdisciplinarity and that of their collaborators, but in contrast correlates poorly with that of their advisees. We conclude that high interdisciplinarity is an advantageous strategy for senior scientists, whereas low interdisciplinarity is optimal for young scientists.

2 Materials and Methods

Below we discuss the eight steps of the digital author identifier assignment-scientist dataset construction algorithm illustrated in Figure (1). We use the algorithm to identify, among the 23,469 author names in a corpus of 7,614 publications and 1,050 grants, the scientists responsible for each document. The algorithm was implemented using a combination of MySQL, PHP, and MATLAB.

2.1 Primary Data

First we obtained a corpus of publications and grants, and the names of their authors from PubMed and CRISP, the NIH publication and grant databases. Specifically, we obtained 7,614 publications and 1,050 grants with the keyword "computational biology". We also obtained the NIH Journal database, and the PubMed and CRISP keyword ontologies, MeSH and

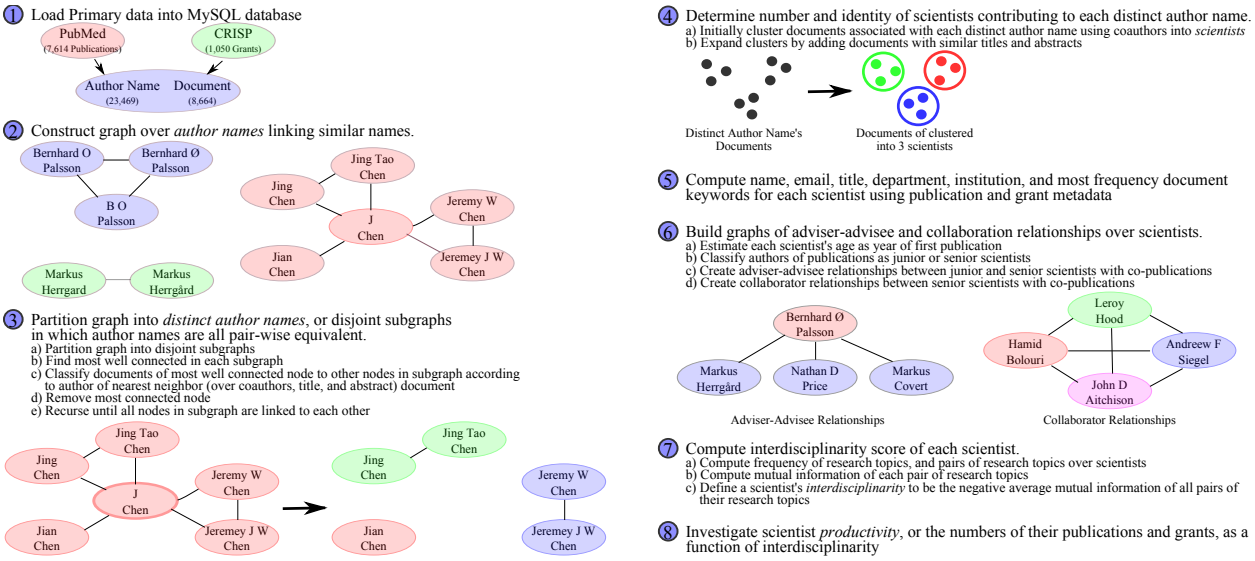


Figure 1 – Digital author identifier assignment-scientist dataset construction algorithm.

CRISP Thesaurus.

2.2 Scientist Inference

2.2.1 Distinct Author Names

To infer the set of scientists, as distinct from author names, which authored each of the 7,614 publications and 1,050 grants, we first computed the set of distinct author names and the publications and grants associated with each such name. First we compared author names pairwise and created a graph over author names where each author name was a node, and author names differing only by the presence/absence of full names versus initials, insertion of space characters in the middle of names, and/or the presence/absence of accent characters were connected by edges. Next we recursively partitioned each subgraph of connected author names into smaller subgraphs by 1) finding the most connected author name, and 2) classifying each of the publications and grants associated with this author name by their nearest neighbor among the bag-of-stemmed-words of publication and grant titles and abstracts of the publications and grants of the remaining author names. Stemming was performed using the Porter algorithm.

We investigated several methods for performing the latter including k-nearest neighbor (kNN) and multinomial naïve bayes (MNB). We did not consider SVM classification because it would be prohibitively slow on the full PubMed and CRISP datasets. We used leave-one-out cross validation to create training and tests sets of publications contained publications from multiple authors, where we knew the true authors of each publication. We compared the accuracy and runtime of the methods, and for kNN for various k. We found kNN accuracy highest with k=1. Additionally, we found similar accuracy between kNN and MNB on groups of publications with two (79% kNN, 66% MNB) and three authors (70% kNN, 67% MNB). We choose neighbor neighbor classification because the method is faster, and does not depend on our current labels of the publications.

2.2.2 Publication/Grant Clustering

Next we clustered the publications and grants corresponding to each distinct author name. This provided the number and identity of the scientists contributing to each distinct author name. We created initial clusters by grouping together documents with overlapping coauthors. Second we expanded clusters by adding documents with similar bag-of-stemmed-words of the title and abstract. Finally we computed each scientist's name from their associated author names.

2.3 Professional Relationships

2.3.1 Adviser-advisee Relationships

To infer adviser-advisee relationships we first classified each author of each publication as either a junior or senior author. We began by computing an upper bound on each scientist's age (year of first publication). Initially we defined a scientist's age to be the minimum of the year of their first publication, year of their first last author publication minus five years, and year of their first grant minus five years. Next we defined junior authors as those authors of documents within three years of their age. Finally we iteratively computed the number of junior authors of

each paper, and recomputed the age of each scientist by computing the minimum of the year of their first publication, year of their first non-junior, or senior, author publication minus five years, and year of their first grant minus five years, until convergence.

Next we computed each scientist's adviser by finding the scientist with the most co-publications, of publications with at most two senior authors, within three years of the each scientist's age. In cases of ties among multiple potential advisers, we linked scientists to multiple advisers, and illustrate our lower confidence in these adviser-advisee relationships in the web-based interface using dashed lines and asterisks.

2.3.2 Collaboration Relationships

We assigned collaboration relationships to all pairs of scientists with at least 1 co-publication where both scientists were senior authors. Furthermore, we noted the confidence of each collaboration relationship as the number of such co-publications; we display this confidence as the opacity of edges in each each scientist's collaboration graph.

2.4 Scientist Digital Author Identifier

We assigned each scientist a unique and stable digital identifier equal to the smallest PubMed ID of their publications concatenated with their position in the author list of that paper. Each time we synchronize our primary dataset with PubMed and rerun our analysis, this identifier points to the same scientist. Finally we use this identifier to provide permanent links to our web-based interface.

2.5 Scientist Profile

We computed each scientist's email and title as that associated with their most recent grant or last author publication. Second we computed each scientist's research topics to be the most frequency MeSH headings of their publications. Next we computed each scientist's current department and institution as that associated with their most recent, in decreasing priority, 1) single author grant, 2) multiple author grant, or 3) last author publication. We similarly computed each scientist's graduate department and institution.

2.6 Interdisciplinarity

2.6.1 Publications

First we computed the frequency of each MeSH heading, and of each pair of MeSH headings over the 7,614 publications. Next we computed the minimum of the mutual information over each pair of MeSH headings for each publication. Finally we defined the interdisciplinarity of each publication to be the negative of this quantity.

2.6.2 Scientists

Similarly, we computed the interdisciplinarity of each scientist. First we computed the most frequent MeSH headings for each scientists as described above. Next we computed the frequency of each MeSH heading, and of each pair of MeSH headings over the set of scientists. Finally we computed the negative average of the mutual information between all pairs of MeSH headings for each scientist, and defined the interdisciplinarity of each scientist to be the negative of this quantity. Additionally we defined the interdisciplinarity of each scientist's advisees and collaborators to the average of that of their advisees and collaborators.

2.7 Productivity

We investigated scientist productivity, defined as a the numbers of publications and grants authored by a scientist, as a function of their interdisciplinarity by plotting histograms of scientist productivity versus scientist, advisees, and collaborators interdisciplinarity, as shown in Figure (3).

2.8 Web-based Interface

Finally, we built the web-based interface illustrated in Figure (2). The web-based interface summarizes the basic properties and education of each inferred scientist, displays a visualization of each scientist's genealogical tree and collaboration relationships, and list each scientist's publications and grants. The web-based interface was built in PHP, and uses MySQL to store the dataset, GraphViz to layout the genealogy and collaboration graphs, GD to display the genealogical tree, and JpGraph to plot the number of publications and grants of each scientist versus time (not shown).

3 Results

3.1 Scientists

Using the machinery described above we clustered 23,469 author names into 21,270 scientists. For the few instances where we knew a scientist's true publication record, such as for the scientists illustrated in Figure (2), we found the results to be highly accurate. For example, we correctly clustered all 28 publications written by the scientist Bernhard Ø Palsson published under the author names B O Palsson, Bernhard Ø Palsson and Bernhard O Palsson into a single scientist. We found the computed profiles of our inferred scientists equally accurate. For example, we correctly identified that Nathan Price is a professor of Chemical and Biological Engineering at UIUC, that he attended UCSD a Bioengineering

SCIENCE GENEALOGY - VIEW SCIENTIST

- Profile
  - Basic Info
  - Education
  - Achievement
- Statistics
- Genealogy
- Collaborators
- Publications
- Grants

Profile

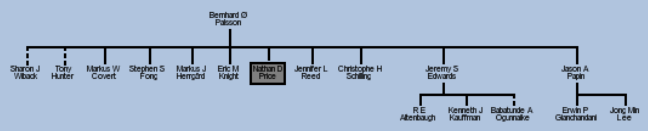
Basic Info

ID	12466293-2 [PermaLink]
Name	Nathan D Price
Aliases	Nathan D Price
Email	ndprice@uiuc.edu
Department	Chemical And Biomolecular Engineering
Institution	University Of Illinois Urbana-champaign, Office Of Sponsored Programs & Research Admin, Champaign, Il 61820
Research	Computational Biology

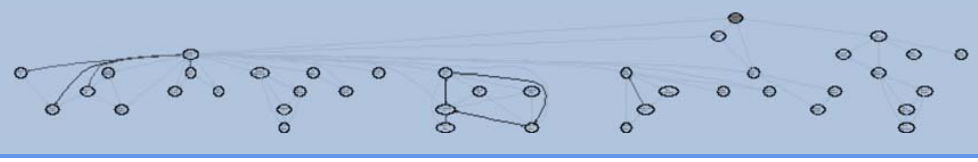
Education

Adviser	Bernhard Ø Palsson
Labmates	Markus W Covert, Jeremy S Edwards, Markus J Herrgård, Tony Hunter, Eric M Knight, Jason A Papin, Jennifer L Reed, Sharon J Wiback
Institution	Department of Bioengineering, University of California-San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0412, USA

Genealogy [larger]



Collaborators [dot | neato | twopi | circo | larger]



Publications

1. **Nathan D Price**, Greg Foltz, Anup Kumar Madan, Leroyay Hood, Qiang Tian (2008). Systems biology and cancer stem cells. *Journal of cellular and molecular medicine*. 12(1): 97-110. [View Details | Similar articles & grants | PubMed]
2. **Nathan D Price**, Ilya Shmulevich (2007). Biochemical and statistical network models for systems biology. *Current opinion in biotechnology*. 18(4): 365-70. [View Details | Similar articles & grants | PubMed]
3. Erwin P Gianchandani, Jason A Papin, **Nathan D Price**, Andrew R Joyce, Bernhard Ø Palsson (2006). Matrix formalism to describe functional states of transcriptional regulatory systems. *PLoS computational biology*. 2(8): e101. [View Details | Similar articles & grants | PubMed]
4. Jason A Papin, Joerg Stelling, **Nathan D Price**, Steffen Klamt, Stefan Schuster, Bernhard Ø Palsson (2004). Comparison of network-based pathway analysis methods. *Trends in biotechnology*. 22(8): 400-5. [View Details | Similar articles & grants | PubMed]
5. Jason A Papin, **Nathan D Price**, Sharon J Wiback, David A Fell, Bernhard Ø Palsson (2003). Metabolic pathways in the post-genome era. *Trends in biochemical sciences*. 28(5): 250-8. [View Details | Similar articles & grants | PubMed]
6. Jason A Papin, **Nathan D Price**, Bernhard Ø Palsson (2002). Extreme pathway lengths and reaction participation in genome-scale metabolic networks. *Genome research*. 12(12): 1889-900. [View Details | Similar articles & grants | PubMed]

Grants

1. **Nathan D Price** (2008). Identifying Network Pertubation Using Secreted Protein Profiles In Glioblastoma. National Cancer Institute #1K99CA126184-01A2. [View Details | Similar articles & grants | CRISP]

Last updated Dec 8 2008.

Figure 2 – Screen shot of web-based interface. Scientist view summarizes each scientist's basic information and education, plots their productivity versus time (not shown), displays their academic lineage and collaboration relationships graphically with higher confidence relationships illustrated in bold, and lists each publication and grant they have authored.

graduate student, and that he has published 6 articles and has won 1 grant on the topic computational biology.

3.2 Professional Relationships

We identified 12,784 adviser-advisee and 3,528 collaborator relationships. For the few instances where true adviser-advisee and collaboration relationships were known we found the computed adviser-advisee and collaboration relationships to be highly accurate. For example, each of the high confidence, genealogical relationships shown in bold in Figure (2) from Bernhard Palsson to his advisees are correct.

3.3 Interdisciplinarity

Perhaps most interestingly we found that the average number of publications and grants per scientist correlates very strongly with our metric of a scientist's or their collaborators interdisciplinarity, but does not correlate with the interdisciplinarity of a scientist's advisees. We believe this suggests that the best strategy for a scientist, that is to maximize their productivity, is to participate in interdisciplinary work as a senior scientist, but to focus on particular topic as a student, first trying to master a single field.

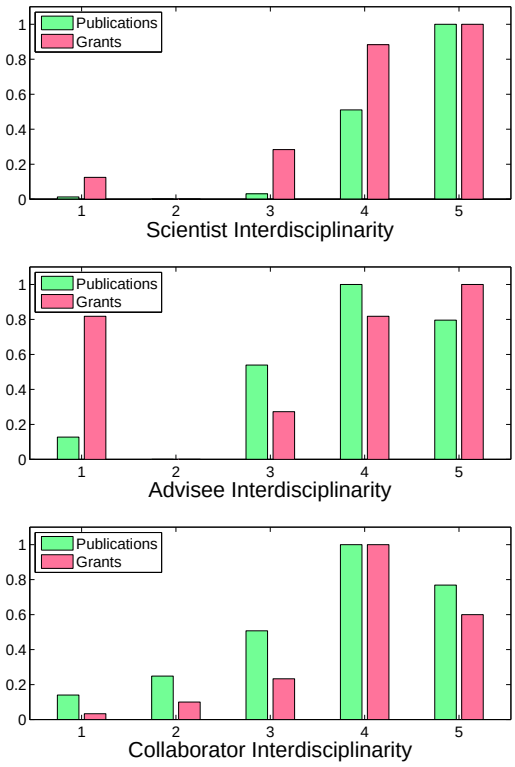


Figure 3 – Productivity versus interdisciplinarity. Number of scientists, average number of grants and publications per scientists versus scientist (top), scientist's advisees (middle), and scientist's collaborators interdisciplinarity (bottom).

4 Discussion

We describe several significant advances. First we automatically inferred the identity of real scientists from the NIH database of author names, and provide a unique and permanent digital author identifier for scientists using the PubMed database. Second, we automatically inferred adviser-advisee and collaborator relationships among scientists. We believe that both results will help senior scientists identify and network with potential collaborators, as well as help junior scientists understand the research interests of their new colleagues, and the professional relationships among them.

Third, we created a metric of a scientist's interdisciplinarity, applied it to each scientist, and found that a scientist's productivity correlates strongly with their interdisciplinarity and that of their collaborators, but correlates poorly with that of their advisees. This suggests that the optimal strategy for a young scientist is to focus their research interests and master a particular field. Furthermore, this suggests that the optimal strategy shifts to broad collaboration as a scientist establishes him/herself. We hope this provides valuable insight for both young and old scientists in setting their research goals and planning their careers.

In the future we plan to apply the method described here to the all 17 million publications and 2 million grants indexed by PubMed and CRISP.

Authors' contributions

JRK, JJH, and TKL conceived the study, wrote the manuscript, and built the distinct author name inference and publication/grant clustering engine. JRK obtained the primary data, and built the engines for inferring scientific relationships and scientists' properties and calculating interdisciplinarity. JRK built the web-based interface.

Acknowledgments

We thank Professor Andrew Ng for introducing us to the subject of machine learning, and encouraging us to apply machine learning concepts to the current study. We thank Laurie Burns for valuable feedback.

Funding: This work has been supported by a NDSEG Fellowship to JRK, a Stanford Bio-X Fellowship to JJH, and a Stanford Graduate Fellowship to TKL.

Conflict of Interest: none declared.