Exploiting Relational Structure to Understand Publication Patterns in High-Energy Physics*

Amy McGovern, Lisa Friedland, Michael Hay, Brian Gallagher, Andrew Fast, Jennifer Neville, David Jensen
Knowledge Discovery Laboratory, University of Massachusetts Amherst, Amherst, MA 01003
amy,lfriedl,mhay,bgallag,afast,jneville,jensen@cs.umass.edu

ABSTRACT

We analyze publication patterns in theoretical high-energy physics using a relational learning approach. We focus on four related areas: understanding and identifying patterns of citations, examining publication patterns at the author level, predicting whether a paper will be accepted by specific journals, and identifying research communities from the citation patterns and paper text. Each of these analyses contributes to an overall understanding of theoretical high-energy physics.

1. INTRODUCTION

We identify interesting patterns and relationships in the theoretical high-energy physics publishing community (*hep-th*). We focus on several high-level questions:

- Can we predict why some papers receive more citations than others? What are the trends?
- What factors contribute to an author's influence? Can we predict potential award winners?
- What factors are important for predicting whether a paper will appear in a journal?
- Can we identify schools of thought or communities in theoretical high-energy physics? Who are the most authoritative authors for each community?

These questions and others are answered in Sections 3 through 6. Findings include:

- Approximately 26% of the people in *hep-th* wrote the papers that received 80% of the citations.
- Edward Witten is the most influential author in theoretical high-energy physics.
- Papers with a single author are less likely to be published in journals than papers with more authors.
- Authors tend to prefer particular journals, that is, a journal is *autocorrelated* through authors.
- Authors tend to publish within topics (i.e., topics are also autocorrelated though authors).

2. DATA REPRESENTATION

We use a relational representation where the data is represented by an attributed graph, $G = (V, E, A(V), A(E))$. Objects, such as authors, journals, and papers, are represented as vertices in the graph. Relations between these objects, such as *published-in(paper, journal)*, are represented by edges between the objects. For a given relation $r(o_1, o_2)$, $o_1, o_2 \in V$ and $r \in E$. Attributes are associated with objects, $A(V)$, such as *author.last-name*, or edges, $A(E)$, such as *authored.rank*.

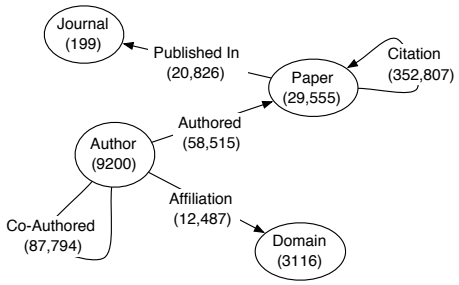


Figure 1: Schema extracted from abstracts and citations. Objects are represented by vertices and relations by edges; numbers in parentheses are object and relation counts.

Figure 1 shows the objects and relations, along with their counts in the database. Details on the attributes and how we extracted them from the *hep-th* data are given in Appendix A. The process of author consolidation (i.e., determining if the *John Smith* who wrote paper 1 is the same person as the *J. Smith* who wrote paper 2) was greatly facilitated by the relational structure [1] (details in Appendix B).

3. CITATION ANALYSIS

Our first analysis considers the papers and citation relations between them. We identify patterns and correlations in the data and use them to understand why some papers are more popular than others. We also build a relational model to predict popular papers.

3.1 Citation Graph Analysis

The citation graph comprises 1,928 separate connected components. The largest contains 27,400 papers, while the others contain 10 or fewer papers. The growth in popularity of *arXiv* and *hep-th* (1397 papers submitted in 1992 and 3312 in 2002) and the limited time frame of the data set cause edge effects on the early and late years (Figure 2a). We often concentrate on the more stable middle years. We break both bibliographic references and citations into self and non-self categories. A self citation or reference means that the two papers share at least one author. Eighteen percent of the citations in *hep-th* are self citations. An average of 28% of

*SIGKDD Explorations, December 2003, Volume 5, Issue 2, pages 165-172. Winning entry to the open task for KDD-Cup 2003.

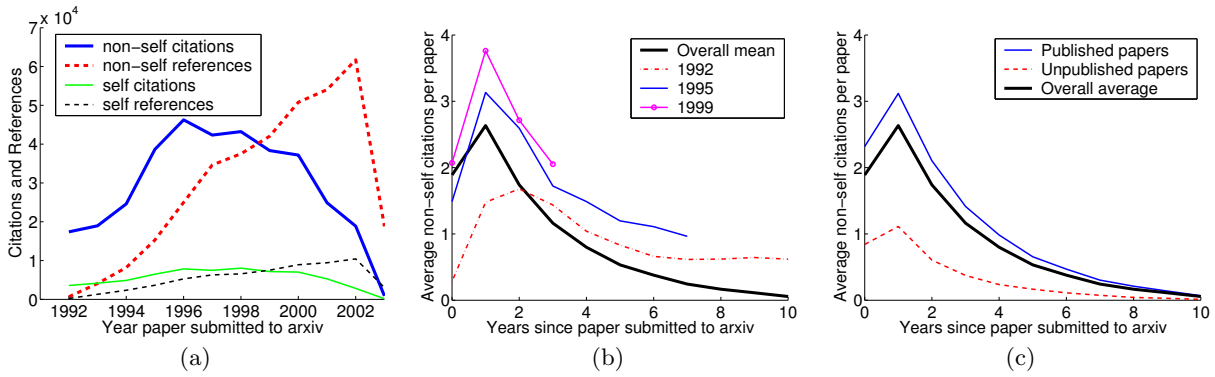


Figure 2: Temporal citation and reference patterns for papers submitted to *arXiv*. (a) Total number of non-self and self citations and references by year. (b) Citations patterns since papers were submitted to *arXiv*. (c) Citation patterns for published versus unpublished papers.

Author of authority papers	**Num. in top 10**	**Num. in top 50**	**Non-self citations**
Edward Witten	4	14	18716
Juan M. Maldacena	2	6	8076
Steven S. Gubser	2	4	5067
Igor R. Klebanov	1	4	5843
Leonard Susskind	1	4	5526
Joseph Polchinski	1	4	5535
Paul K. Townsend	1	3	4991
Stephen H. Shenker	1	2	2300
Michael R. Douglas	0	5	5787
Nathan Seiberg	0	3	9911
Cumrun Vafa	0	3	8594
Andrew Strominger	0	3	6480
Petr Horava	0	2	1936
Daniel Z. Freedman	0	2	1874

Table 1: Authors of the top 10 and top 50 most authoritative papers and the total number of non-self citations that these authors have received in *hep-th*.

a paper's references cite its authors' past work and 34% of a paper's citations are from its authors.

Because papers are often submitted to *hep-th* before they are published in a journal, we hypothesized that papers might receive citations in two peaks. A paper could be cited by other papers in *hep-th* as soon as it was submitted to *arXiv* and again after being published in a journal. Figure 2b shows the number of citations that each paper received following its submission to *arXiv*. Looking at the overall mean, papers generally receive the most citations in the year following submission to *arXiv*. This peak likely coincides with journal publication as the average time from a paper appearing on *arXiv* to journal publication is one year. Papers also receive an average of two citations in the year prior to journal publication, demonstrating *arXiv*'s effectiveness at disseminating results quickly. The pattern of citations for papers submitted to *arXiv* in 1992 peaks two years after submission. This delay can be explained by *arXiv*'s growing popularity as use of the Internet grew. The number of citations increases more quickly in later years due to the larger number of authors with Internet access.

Figure 2c shows the average number of non-self citations for papers that have been published in a journal versus unpublished papers. Published papers have a significantly higher average non-self citation rate than papers that appear only on *arXiv*. This indicates that either journal publication is still important in increasing a paper's visibility or authors writing highly cited papers still seek journal publication.

Author of hub papers	**Num. in top 1%**	**Num. in top 5%**	**Non-self references**
Igor R. Klebanov	15	31	5843
Arkady A. Tseytlin	10	29	5352
Steven S. Gubser	9	28	5067
Ofer Aharony	8	19	2307
Clifford V. Johnson	6	21	1615
Alberto Zaffaroni	6	13	1369
Washington Taylor IV	6	7	2115

Table 2: Authors of the top 1 and top 5 percent hub papers and the total number of non-self references that these authors have made.

The hubs and authorities algorithm [5] was used on the citation graph to identify authoritative papers and potential review papers. A *hub* points to many authorities. This is likely to be a review paper. An *authority* is pointed to by many hubs. Once we identified the most authoritative papers, we examined the authorship for these papers. Table 1 shows the authors who have written at least two of the top 10 and top 50 most authoritative papers. As many of these names appear again when we study influential authors, we discuss their specifics in section 4. The authors of these highly authoritative papers include a number of award winners and they hail from prestigious institutions.

We were interested in the question of whether some authors write mostly review papers. Table 2 shows authors who have written top hub papers. No author has written more than one of the top 10 or top 50 hub papers but several authors appear as frequent authors in the top 1% and top 5% of hub papers. Table 2 contains no major award winners and represents a slightly different list of institutions than Table 1. The top three authors on this list are frequent co-authors.

3.2 Citation Data Dependencies

To better understand what makes papers popular, we examined correlations in the citation data. For discrete attributes, we used corrected contingency coefficients; for continuous attributes we used correlation coefficient[11]. Ta-

Attribute 1	Attribute 2	Score
For paper		
Paper authority score	# of citations	0.85
arXiv area	References (binned)	0.68
Hub score	# of references	0.62
# downloads first 60 days	# of citations	0.57
Is paper published	Citations (binned)	0.46
For author		
# of publications	# of distinct coauthors	0.85
# of distinct coauthors	# of non-self citations	0.59

Table 3: Selected correlation scores between attributes.

Attribute	Through	Score
arXiv area of paper	Author	0.72
Journal name	Author	0.69
# downloads first 60 days	Author	0.55
Clustered topic of paper	Author	0.54
Coauthor authority score	Paper	0.74
arXiv area of cited paper	Paper	0.70
# of coauthors	Paper	0.45
# downloads first 60 days	Journal	0.42

Table 4: Selected autocorrelation scores.

bles 3 and 4 list significant correlations in the data. Results from these tables are discussed throughout the paper. All reported correlations are significant ($p < 0.0001$).

For example, the number of times that a paper is downloaded is correlated with the number of non-self citations of that paper. This is not surprising as one expects more frequently downloaded papers to be cited more frequently.

In addition to correlations among variables of a single object, we also measured *autocorrelation* throughout the data graph [2]. Autocorrelation is a statistical dependency between the values of the same variable on related objects, also known as homophily [6]. For example, the number of downloads of a paper is autocorrelated through authors. This means that if one of an author's papers is frequently downloaded, other papers by the same author are likely to be downloaded.

3.3 Predicting Popular Papers

We used relational probability trees (RPTs) [9] for several classification tasks. The resulting relational models enhance our understanding of the publication patterns in *hep-th*. For each task, we sampled papers temporally, training the model on papers from one year and testing on the following year's papers. To avoid edge effects, we considered only papers from 1995 to 2000. The model considered features of papers, their referenced papers, authors, and other past papers written by the authors. Example attributes include the number of pages and the author's number of past co-authors and number of past publications. Attributes were calculated for each temporal sample. To predict the class label on a paper submitted in 1997, the model considered the history of related objects through 1996.

The first modeling task involved predicting the number of non-self citations a paper will receive. We categorized the number of non-self citations into quartiles: {0-1, 2-5, 6-14, >14}. Default classification accuracy is approximately 25%. Over 5 training/test splits, RPT models achieved an average accuracy of 44%. Although 44% is not a high accuracy, it is notable that we could achieve this based solely on the information in *hep-th*, which does not account for such important factors as paper quality.

One reason we chose to use RPT models is their selectivity. We can examine the features chosen by the trees and identify the most relevant features for the classification task. The RPT models estimated that a paper has a 0.85 probability of receiving more than 14 non-self citations if 1) the paper has more than 8 references, 2) the authors have at least 2 past papers with more than 8 non-self citations, 3) the authors have at least 25 past papers (each at least 15 pages long), and 4) at least 30% of the cited work is unpublished.

4. AUTHOR ANALYSIS

The second part of our analysis focuses on authors. We examine the overall structure of the author subgraph and extend this understanding to identifying influential authors. We define several measures of influence and build a relational model to identify and predict award-winning authors. Finally, we identify potential award winners.

4.1 Co-Author Graph Analysis

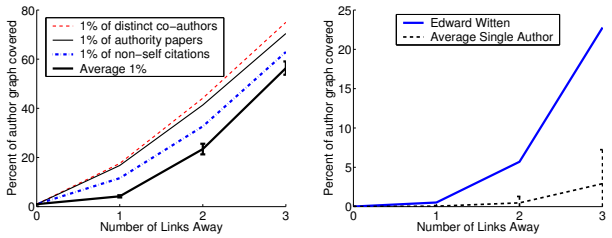
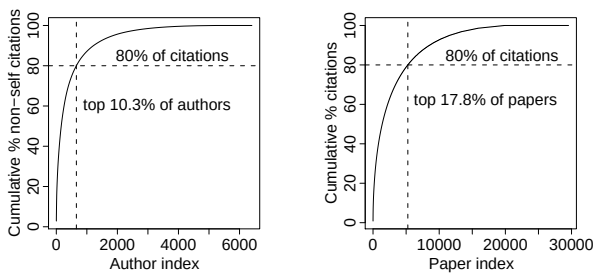


Figure 3: (a) Percent of the author graph that is one, two, and three links away from several sets of the top 1% of authors as well as from a random sampling of 1% of authors. (b) Percent of the author graph that is 1, 2, and 3 links away from Edward Witten versus the average author.

We found that the high-energy physics community is tightly knit. In the graph of authors linked by co-authored relations, 7304 of the total 9200 authors belong to a single connected component. As with the paper graph, other components are all small (15 or fewer authors). When we narrowed this set of authors to authors who wrote the top 1%, 5% and 10% of the authoritative papers, we found that the vast majority of the authors remained connected. This provides evidence for the idea that influential scientists train the up-and-coming influential scientists in their labs, either as students or post-doctoral fellows [4].

We also found that authors whose papers are highly cited or have many distinct co-authors are more central to the author graph than randomly selected authors. Figure 3 shows the percentage of authors who are 1, 2, and 3 links away from authors who wrote the top 1% of authority papers, authors who have received the top 1% of non-self citations and the top 1% of authors who have co-authored with different people. These numbers are compared to 10 random samplings of 1% of the authors. Each of these sets of influential authors is linked to a higher percentage of authors through co-authored relations than the random baseline. We also show that the average degree of separation from Edward Witten, who consistently shows up as the most influential author in *hep-th*, is significantly lower than the average author.



(a) (b)

Figure 4: Cumulative percent of non-self citations received (a) per author and (b) per paper.

Before building a quantitative measure of an author's influence, we examined the data for general trends. From 1995 through 2000, 6405 authors submitted papers to *hep-th*. These authors each wrote 5 papers on average with a median of 2. Sergei Odintsov (with 92 papers) and H. Lu and C.N. Pope (each with 84) topped the distribution. Of the papers submitted to *arXiv* in this period, each author published an average of 4 papers in journals. Authors recieved an average of 76 non-self citations, with a much lower median of 7. The top 10% of authors averaged 140 non-self citations. As seen in Table 5, the top authors produce high numbers of papers by co-authoring widely and frequently. The average number of distinct co-authors is 5.5.

The 80/20 rule or Pareto's Principle states that, in power law distributions, 80% of the mass is generally due to only 20% of the values (whether in science or other domains)[10; 7]. We investigated this rule in theoretical high-energy physics by examining the number of non-self citations per paper and per author. In the *hep-th* data, 80% of the non-self-citations go to 17.8% of the papers and 26.3% of the authors wrote these papers. Counting by author, 10.3% of the authors received 80% of the non-self citations. Both of these distributions are shown in Figure 4.

4.2 Author Data Dependencies

Tables 3 and 4 summarize the trends and dependencies for authors. The number of an author's publications is correlated with the number of citations that the author receives. Authors who have more citations may publish more frequently or people who publish more papers may receive more citations. Perhaps more surprising is that an author's number of publications is correlated with the number of distinct co-authors of that the author. This indicates that frequently published authors do not tend to work repeatedly with the same set of co-authors but with new people.

Contrary to our expectation, authors who write authoritative papers do not write other authoritative papers. A paper's authority score is not autocorrelated through author which means that most authors will write only a few authoritative papers.

Information about the research styles of authors can be gained from autocorrelation scores. For instance, the number of distinct coauthors is autocorrelated through papers, which means that if you publish with other authors who publish with many distinct people you are also likely to publish with many distinct people. Also, an author who publishes a paper in a particular journal is likely to publish other papers in that journal.

(a) Overall co-authorships

Author	Count
C.N. Pope	337
H. Lu	325
S.D. Odintsov	296
Sergio Ferrara	233
Mirjam Cvetic	231

(b) Distinct co-authorships

Author	Count
Cumrun Vafa	63
Gary W. Gibbons	60
Jan de Boer	56
Sergio Ferrara	55
Antoine Van Proeyen	55

Table 5: (a) Authors who most frequently co-author. (b) Authors who frequently co-author with different people.

1. Number of non-self citations received
2. Total number of citations received
3. Number of papers written
4. Number of papers published in journals
5. Number of papers with over 12 citations
6. Number of co-authorships
7. Number of distinct co-authors
8. Average non-self citations per paper
9. Maximum non-self citations received on any paper
10. Percentage of papers published
11. Percentage of papers with over 12 citations

Table 6: Measures of author influence

4.3 Analyzing Author Influence

We hypothesized that author influence (overall reputation and impact) could be estimated using the measures shown in Table 6.

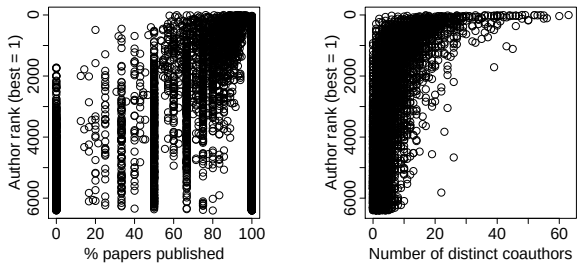
We evaluated the measures by using each to rank the authors who submitted papers from 1995 to 2000 and counting the number of award winners listed in the top 100 authors. We hand-identified 55 winners of prestigious awards, including Nobel prize winners, MacArthur Foundation fellows, Dirac fellows, Guggenheim fellows, Fields medal winners, and Alfred P. Sloan Foundation winners. Most of the measures performed about equally well, finding around 10 award winners. Measures 1 and 2 did best, with 14 winners. We chose measure 1 to be our canonical influence measure[1]. Table 7 shows the top authors under measure 1 and their citation counts. Heading the list, Edward Witten is a MacArthur Foundation fellow, a Fields medalist, and a Dirac fellow. Juan Maldacena, also a MacArthur Foundation fellow, is a younger researcher and looks quite likely to become the most cited author as he continues his research.

Surprisingly, measures 10 and 11, which indicate an author's consistency of success, performed poorly in our validation, identifying 2 or fewer winners. Closer inspection shows that perfectionism is not the key to success. The percentage of papers published in journals varied widely among award-winners, from 100% to 0%, although the top 50% of influential authors did have a higher rate (88%) of journal publication than the bottom half (67%). Figure 5a shows a scatter-plot of measure 10 versus influential authors. The high variance at 100% explains the poor performance of this measure. This also occurs with measure 11. In both cases, the problem is that one out of one paper satisfying the mea-

[1]This is also used by the popular research tools Citeseer: http://citeseer.nj.nec.com/mostcited.html and ISI Essential Science Indicators: http://www.in-cites.com.

Author	Non-self citations	# papers	Won award?
Edward Witten	13806	59	Y
Juan M. Maldacena	7334	39	Y
Cumrun Vafa	6578	55	Y
Nathan Seiberg	6258	45	Y
Andrew Strominger	5371	44	N
Michael R. Douglas	5089	24	Y
Igor R. Klebanov	5063	51	N
Joseph Polchinski	4815	25	Y
Steven S. Gubser	4812	31	Y
Ashoke Sen	4201	51	N

Table 7: Top-cited authors, based on papers 1995-2000



(a) (b)

Figure 5: (a) Author influence vs. percent of papers published. (b) Author influence vs. distinct co-authors

sure yields a higher rank than 19 out of 20 papers satisfying it.

Figure 5b examines the correlation between measures 1 and 7. Authors with high citation counts write both frequently and widely. Collaborating with 10-15 other people is typical. Anyone with over 30 co-authors is almost certain to be in the top 10% of influential authors; presumably one must be extremely well-regarded to attract that kind of demand by collaborators. In the top 10% of influential authors, no one writes alone, and of the top 100 authors, only Donam Youm has fewer than 10 distinct coauthors. Table 5 displays authors with high co-author counts.

We wondered if a different combination of features could better separate award-winners from other authors. To investigate this, we built an RPT using the set of 55 award winners and a random sample of 55 non-award winners. We performed 10-fold cross validation and achieved an average accuracy of 78% with an area under the ROC curve (AUC) of 0.75. The tree chosen most frequently is shown in Figure 6. The first split in the tree, the author's authority score, is based on the score assigned by the hubs and authorities algorithm over the undirected co-author graph.[2] This roughly indicates authors who co-author frequently and whose co-authors also co-authored frequently.

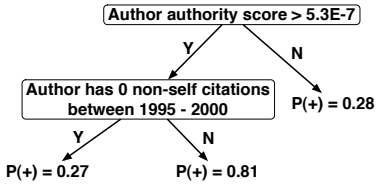
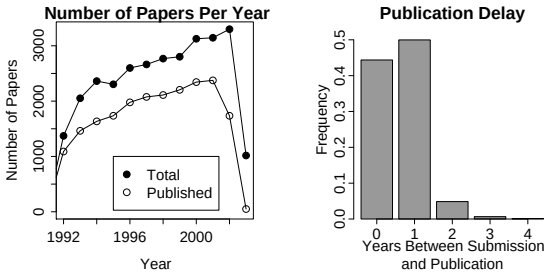


Figure 6: RPT built to predict award-winning authors.

Informed by the features in the tree as well as by our other analyses, we conjecture that some of the following highly cited authors may soon receive recognition: Andrew Strominger, Igor R. Klebanov, Ashoke Sen, Arkady A. Tseytlin, Paul K. Townsend, Gregory Moore, and Hirosi Ooguri.

5. PUBLICATION ANALYSIS

Influential authors are more likely to have their papers accepted by a journal, as discussed above. It is also clear from Figure 2 that published papers receive more citations. The third part of our analysis studied other factors that affected journal acceptance.



(a) (b)

Figure 7: (a) Number of published and unpublished papers submitted to *arXiv* each year. (b) Number of years between a paper's submission to *arXiv* publication in a journal.

Approximately 70% of the papers in *arXiv* have been published in a journal. Figure 7a shows the total number of papers submitted to *arXiv* each year for both published and unpublished papers. Although the total number of papers increases each year, the proportion of published and unpublished papers remains relatively constant. Figure 7b shows the number of years between a paper's submission to *arXiv* and its publication in a journal. Most papers, if published at all, are published within one year of submission to *arXiv*. A small number are published up to 3 years later.

We analyzed the differences between the published and unpublished papers and discovered significant effects. Several of these effects are shown in Figure 8. The most surprising difference is that published papers usually have more than one author while unpublished papers are more frequently written by a single author. This is an example of *degree disparity*[3], where the number of relations differs significantly between objects with different class labels. Unpublished papers have fewer references than published papers and published papers have more pages than unpublished ones. This correlates with the finding that published papers are revised more frequently. As a paper is revised, additional text is added and the number of pages grows. Journals may induce degree disparity with page limits as seen in Figure 8d. Most papers published in *Physics Letters B* are between 5 and 15 pages in length while the unpublished papers have varying lengths.

[2]Hub and authority scores are equivalent on undirected graphs.

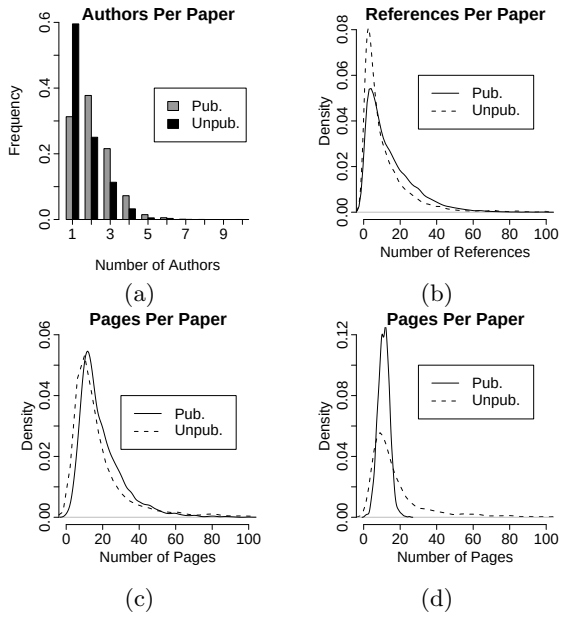


Figure 8: Characteristics that differentiate published and unpublished papers. Panels a, b, and c are from all published and unpublished papers from 1995 to 2000 inclusive while d is from a sample of 1500 papers from *Physics Letters B* and 1500 unpublished papers.

5.1 Predicting Publication

For this task, we trained two types of relational models, RPTs and relational multiple-instance learning [8] (RMIL), to predict whether papers submitted to *hep-th* from 1995 to 2000 will be published in a journal.

We trained an RPT to differentiate between unpublished papers and papers published in *Physics Letters B*, the most common publication venue for *hep-th* papers. We sampled a set of 500 papers per year (3000 total), with equal proportion of published and unpublished papers. Given the difficulty of this task, the RPT performed well, with an average of 68% accuracy and 0.75 AUC.

The model selected four attributes that discriminate between unpublished and published papers: the number of authors, the number of references, the paper's length and the paper's filesize. Figure 9a shows an example RPT. The model used the degree disparity examples discussed above. For example, the RPT predicts that papers over 16 pages in length and at least 13K in size were unlikely to be published (P(+)=0.03). Browsing a subset of these papers on *arXiv* shows that the unpublished papers are either workshop papers (short papers, few references) or theses (long papers, a single author).

We also trained an RPT on the entire set of published and unpublished papers, and had moderately successful results (0.70 AUC). The sample for each year had between 2300 and 3100 papers, and approximately 75% of the papers each year are published. The algorithm learned similar trees to the previous task. As shown in Figure 8c, paper length is not as discriminative in this larger sample, which may explain the lower performance on this larger set.

For RMIL, we created random samples of 200 papers (100 published and 100 unpublished papers) per year. RMIL achieved an accuracy of 61% with an average AUC of 0.61. RMIL identified that papers with at least 2 authors, papers that cited papers published in *Nuclear Physics B*, or papers that were cross-posted to areas other than *hep-th* were all more likely to be published.

5.2 Predicting Publication Venue

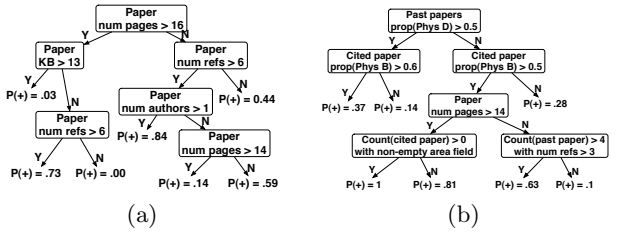


Figure 9: (a) RPT to predict whether a paper will be published in *Physics Letters B*. (b) RPT to differentiate between two popular journals.

We also applied RPTs to a related task, differentiating between papers published in one of two popular journals. We expected this task to be challenging because approximately 55% of the papers were written by authors who have published in both journals.

For each year, we sampled a set of 480 published papers, half of which were published in *Nuclear Physics B* and half in *Physical Review D*. For this task, RPTs achieved an average accuracy of 73% and an average AUC of 0.81. An example tree is shown in Figure 9b. The authors' publication history, the publication venue of cited papers, and paper length are useful features to differentiate between papers published in these two journals. If over 50% of an author's past papers were published in *Physics Letters D*, and less than 60% of cited papers were published in *Nuclear Physics B*, then the paper is unlikely to be published in *Nuclear Physics B* (P(+)=0.14).

6. COMMUNITY ANALYSIS

Our final analysis focused on identifying schools of thought, or research communities, and the influential authors and journals associated with these communities.

6.1 Topic detection

We conjectured that journals tend to represent distinct topics and treated journals as topic markers. Clustering based only on the paper text did not yield useful clusters. Instead, we made use of the rich information available in the citations by using a spectral clustering method. This method was based on previous work by [13] on spectral partitioning algorithms. We used the citation graph to cluster papers but weighted the strength of citation relationships by the cosine similarity between paper abstracts. For this approach, we clustered a sample of 833 papers from 1995-2000 that each had more than 50 non-self citations. This choice of sample set was made to identify a small set of authoritative papers that are likely to define topics. This approach identified 14 topics with 2 to 285 papers each.

6.2 Topic cluster evaluation

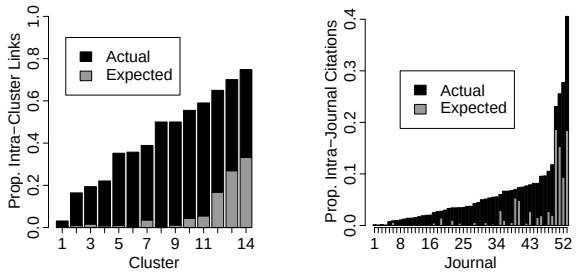
If the approaches identified useful topics, we hypothesized that authors would cite papers within their preferred topic more than papers outside of the topic. To measure the citation rate, we calculated the actual and expected proportion of intra-cluster citations for each cluster. We define the actual proportion of intra-cluster citations for a cluster, C, as:

$$\frac{\text{the total \# of citations from papers in C to papers in C}}{\text{the total \# of citations from cluster C}}.$$

We define the expected proportion under uniform clustering of intra-cluster citations for a cluster, C, as:

$$\frac{\text{the total \# of papers in cluster C}}{\text{the total \# of papers in the collection}}.$$

We also expected that authors would publish within a relatively small set of research communities. This was measured by examining the autocorrelation of the topic clusters through authors and journals.



(a) (b)

Figure 10: Expected and actual intra-clustering citation ratios for (a) spectral clustering (b) journal clusters.

Figure 10 shows the expected and actual intra-cluster citation proportions for spectral clustering (a) and journal-based topics (b). In both cases, the actual intra-citation values deviate significantly from the expected values. Both sets of topics were autocorrelated through authors (see Table 4). In addition, spectral topics are both correlated with journal name (corr=0.58) and autocorrelated through journals (corr=0.56). For identifying the communities, we focus on the spectral topic clusters as they can identify topics for unpublished papers.

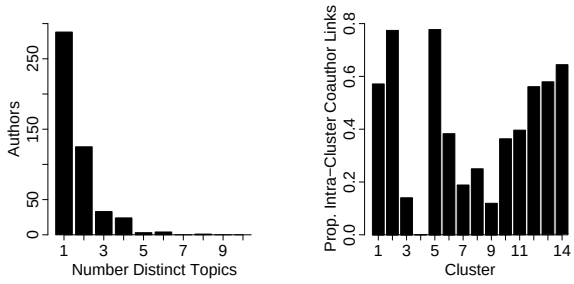
6.3 Research communities

Because topics are autocorrelated through authors, we used the clusters to partition authors into communities. Each author was assigned to their most prevalent cluster based on authorship. Table 8 includes randomly selected paper titles from four example clusters for subjective evaluation. Each cluster was provided with a topic title from a high-energy physics graduate student. The most authoritative author and journal for each cluster is also included.

We expected that most scientists have a focused area of research and will publish within only a small number of research communities. Figure 11a shows the number of topic clusters that each author is associated with and supports this hypothesis. We further hypothesized that authors are more likely to collaborate with other authors within their research community. To evaluate this, we examined whether

Cluster 2 : Sumit R.Das (251), *Physical Review D* **Topic:** Black hole approach to string theory Absorption of Fixed scalars and the D-brane Approach to Black Holes; Universal Low-Energy Dynamics for Rotating Black Holes; Interactions involving D-branes; Black Hole Greybody Factors and D-Brane Spectroscopy
Cluster 7 : Gary T.Horowitz (588), *Physics Letters B* **Topic:** Black hole approach to string theory On D-Branes and Black Holes in Four Dimensions; The Black Branes of M-theory; Counting States of Near-Extremal Black Holes; Internal Structure of Black Holes
Cluster 10 : Juan M. Maldacena (1924), *Journal of High Energy Physics*, **Topic:** Tachyon condensation Field theory models for tachyon and gauge field string dynamics; Super-Poincare Invariant Superstring Field Theory; Level Four Approximation to the Tachyon Potential in Superstring Field Theory; SO(32) Spinors of Type I and Other Solitons on Brane-Antibrane Pair
Cluster 13 : Ashoke Sen (4683), *Nuclear Physics B* **Topic:** AdS/CFT offshoots Dynamics of Anti-de Sitter Domain Walls; Gravitational Stability and Renormalization-Group Flow; String Theory on AdS_3; The Holographic Bound in Anti-de Sitter Space

Table 8: Example research communities found by spectral clustering.



(a) (b)

Figure 11: (a) Association of authors to topic clusters (b) Intra-cluster coauthor frequency.

the proportion of coauthor relations within topic clusters was higher than expected. Figure 11b shows the actual proportion of intra-cluster coauthor relations. The expected proportions are too small to be visible. Collaboration is significantly higher with these clusters than would be expected by chance. Both results further validate the spectral clusters as research communities.

7. CONCLUSIONS

Based on our analysis, theoretical high-energy physics appears to be a healthy scientific community. Both the citation and authorship graphs reflect a pattern of tightly knit communication via the formal and informal scholarly literature. The community publishes a large numbers of papers, and the temporal pattern of citations indicates the rapid uptake and use of relevant new work. Despite the existence of "super-stars," the papers of individual authors can vary greatly in their authority scores, indicating that papers are cited more for their innovative content than the pre-existing prominence of their author.

This analysis raises the possibility, already explored by the field of scientiometrics [12], of assessing and comparing the

health of different scientific communities and subcommunities. The statistical techniques under development within relational learning offer an improved toolbox for the study of scientific networks, particularly as reflected in patterns of publication, citation, and downloading. Central to our analyses in this paper were: 1) measures that use a combination of attributes and structure of relational data; and 2) algorithms for learning statistical models that search a vast space of possible structures and parameter values to select those features most predictive of an attribute of interest. Both of these allowed simultaneous consideration of multiple object and relation types, rather than only a single object and relation type, as is common in much prior work in citation analysis. Finally, consolidation of authors was important to the analysis above, and the relational structure was a strong contributor to how authors were consolidated.

Acknowledgments

We thank Hannah Blau for her comments and Daniel Larson for domain knowledge contributions. This effort is supported by DARPA, AFRL, and NSF under contract numbers F30602-00-2-0597, F30602-01-2-0566,and EIA9983215. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements either expressed or implied, of DARPA, AFRL, NSF or the U.S. Government.

8. REFERENCES

[1] H. Goldberg and T. Senator. Restructuring databases for knowledge discovery by consolidation and link formation. In *Proc of the 1st Intl Conf on Knowledge Discovery and Data Mining*, pages 136–141. AAAI Press, 1995.

[2] D. Jensen and J. Neville. Linkage and autocorrelation cause feature selection bias in relational learning. In *Proc of the 19th Intl Conference on Machine Learning*, pages 259–266. Morgan Kaufmann, 2002.

[3] D. Jensen, J. Neville, and M. Hay. Avoiding bias when aggregating relational data with degree disparity. In *Proc of the 20th Intl Conf on Machine Learning*, pages 274–281, 2003.

[4] R. Kanigel. *Apprentice to Genius: The Making of a Scientific Dynasty*. Johns Hopkins University Press, 1993.

[5] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[6] P. Lazarsfeld and R. Merton. Friendship as social process: A substantive and methodological analysis. In *M. Berger et al. (eds.), Freedom and Control in Modern Society*. Octagon, New York", 1964.

[7] A. Lotka. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16:317–323, 1926.

[8] A. McGovern and D. Jensen. Identifying predictive structures in relational data using multiple instance learning. In *Proc of the 20th Intl Conf on Machine Learning*, pages 528–535, 2003.

[9] J. Neville, D. Jensen, L. Friedland, and M. Hay. Learning relational probability trees. In *9th ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining*, pages 625–640, 2003.

[10] V. Pareto. *Le Cours d'Economie Politique*. Macmillan, London, 1897.

[11] L. Sachs. *Applied Statistics*. Springer-Verlag, 1982.

[12] Scientometrics: An international journal for all quantitative aspects of the science of science, communication in science and science policy. Kluwer Academic Publishers, Dordrecht, The Netherlands.

[13] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

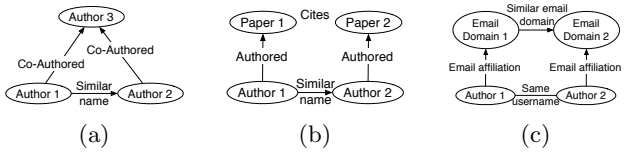


Figure 12: Relational evidence of duplicate authors. (a) Authors with a similar name who have co-authored with the same third-party. (b) Authors who have cited a paper written by an author with a similar name. (c) Authors with similar email domains and the same username.

APPENDIX

A. CREATING THE SCHEMA

The data available for 2003 KDD Cup task 4 was in the form of LaTeX files, text abstracts, and paper citations. From the abstract files, we extracted paper properties such as title, file size, journal reference, and submission dates. We used the earliest of the revision dates and the SLAC date as the best estimate of authorship date. Author names and institutions were parsed out of the Authors field, and the email address of the submitter was associated with the best-matching author name. We used the domain name of the submitter email address as a surrogate for institution. Journals were consolidated by hand. The email domains were given similarity links based on matching suffixes to facilitate identifying distinct institutions, and for use during author consolidation. We performed a nominal amount of hand data cleaning to correct for spelling errors or formatting problems.

B. AUTHOR CONSOLIDATION

Many *hep-th* authors publish under variants of the same name where the number of distinct identities was unclear (e.g. "J. Adams" and "J.A. Adams"). We began with the assumption that no two people had submitted papers under the exact same name, although we did find rare instances of such doubles when hand-checking. We labeled pairs as *similar* if the last names and the first initial of the first names matched. Of the initial 13,185 distinct author names, over 7500 had candidate matches.

Possible evidence for duplicate authors came from several sources. First, authors had to have *similar* names, and co-authors could never be consolidated. Another piece of evidence arose from author email addresses: using the same email address for multiple papers meant the authors were likely to be the same person. This was not conclusive evidence, because we found instances of people sharing email addresses. If a candidate pair's last name was rare in *hep-th*, this boosted the evidence.

We also identified evidence for duplicate authors based on the relational neighborhood of the authors, as depicted in Figure 12. If two authors with similar names had each coauthored with the same third person, the two were likely to be the same person. Similarly, since people frequently cite their own work, we reasoned that if an author cites someone with a similar name, the two may well be the same person. Last, if two authors had the same username at similar email domains, this was comparable to using the same email address.

Using these guidelines, we iteratively identified and consolidated duplicate authors until quiescence. Because evidence involving third-party authors was often not available until the third parties had themselves been merged correctly, this took five rounds of consolidation. At completion, we had 9200 distinct authors. Due to the noisy nature of the data, the final author set is not perfect but, it correctly merged all eight variations of the name "Ian Kogan," and consolidated the top ten authors from Table 7 from 28 authors to 11 (e.g., one mistake).