Understanding Relationship between Scholars' Breadth of Research and Scientific Impact

Shiyan Yan[1] and Carl Lagoze[2]

[1] *shiyansi@umich.edu*
School of Information, University of Michigan, Ann Arbor

[2] *clagoze@umich.edu*
School of Information, University of Michigan, Ann Arbor

Abstract

Many existing metrics to evaluate scholars consider their scientific impact without considering the importance of breadth of research. In this paper, we define a new metric for breadth of research based on the generalized Stirling metric that considers multiple aspects of breadth of research. We extract research topics in computer science using concept extraction and clustering from the literature in the ACM dataset. We then assign authors a distribution over these research topics, from which we calculate scores of breadth of research for each author. We design five simulation experiments that evaluate the ability of a metric to measure breadth of research and use these experiments to compare our new metric to traditional metrics. The results show how these metrics perform in different experiments, concluding that no metric consistently outperforms the others. We test the relationship between our new metric and scientific impact and find a weak correlation between them. Finally, we find that the variation of the metric over time illustrates a possible publication pattern for scholars.

Conference Topic

Indicators

Introduction

An increasing number of scholars are engaged in interdisciplinary research (Porter, Cohen, David Roessner, & Perreault, 2007; Wagner et al., 2011). Some of this is due to the emergence of new scholarly "disciplines" that are inherently multi-disciplinary such as information science, while some arises from scientific problems such as climate change that require expertise from multiple fields. Meanwhile, scholarly impact and influence continues, by and large, to be measured by indices that ignore breadth of research and may even penalize scholars who diversify their research portfolio. For example, H-index, which is used extensively to measure scholarly impact, and which has been criticized for its limited focus (Weingart, 2005), may be unfair when comparing scholars with different degrees of breadth of research. Ultimately, a metric or a set of metrics is needed that accounts for breadth of research, so that breadth of research can be measured and be included in an evaluation system of scholars' scientific influence.

In this paper we describe research that explores the area of scholarly impact metrics and breadth of research. The contributions of our work are as follows. We design a new metric to measure scholars' breadth of research that builds on traditional metrics. We develop a multi-stage method for extracting topics from a corpus (in our case computer science papers) and calculate the scores of breadth of research for authors who have published papers in computer science conferences. We design five simulation experiments that compare the relative performance of existing metrics and our new metric for measuring breadth of research. We measure the relationship of breadth of research and H-index for scholars who are authors in our corpus. Finally, we explore the variation of breadth of research for scholars over time to observe their paper publication behavior over their careers.

The structure of this paper is as follows. The next section describes related work in the areas relevant to our work. Following that, we report on the dataset we used in our research. We then describe our process of dictionary extraction, topic extraction, paper assignment and author assignment to topics. In the subsequent section we illustrate our new metric and compare it to traditional metrics. The penultimate section describes simulation experiments to show the performance of the new metric, the relationship between the new metric and metrics of research impact, and the variation over time of breadth of research for scholars. Our conclusions and possible future work are listed in the final section.

Related Work

There is a variety of existing literature relevant to the area of breadth of research. The areas covered by this literature include topic extraction, topic relationship extraction, metrics design and the relationship between different aspects of research evaluation systems.

There are many methods to associate topics to publication. The simplest one is to use the classification codes in a dataset, such as ISI subject categories in Web of Science, as the set of topics. But these categories are too coarse-grained and hide intra-disciplinary variability. Another method is to use unsupervised learning algorithms to extract some topics according to the content of papers or the citation network of papers. Topic modelling (Blei, Ng, & Jordan, 2003) is one of the popular unsupervised learning algorithms based on content of papers. This model has been used to identify the disciplines that comprise interdisciplinary work funded by NSF (Nichols, 2014). The ACT model (author-conference-topic) (Li et al., 2010) is an adaptation of Blei's model. Another approach is to use community detection in networks as a basis for finding topics. One example is the use of two-round clustering (Rosvall & Bergstrom, 2008) over the citation network to extract topic-associated communities (Velden & Lagoze, 2013). Another method using both the citation network and the word distribution of abstracts (Jo, Hopcroft, & Lagoze, 2011) finds temporally-ordered topics from a corpus of scientific literature, such as the ACM dataset.

Understanding the relationship between topics is also an important step after topic extraction, because the calculation of the similarity of topics is necessary for understanding the breadth of research. Some researchers have extracted the relationships and used information visualization techniques to represent the relationship between different topics. For example, Yan (2013) detects the path between different disciplines to find the evolution of some areas. Another paper describes a new method to find the diversity subgraph in a multidisciplinary scientific collaboration network (He, Ding, Tang, Reguramalingam, & Bollen, 2013). An interesting visualization method leverages the circle of science to visualize the relationship between disciplines in one dimension (Boyack & Klavans, 2009).

Many metrics have been designed to measure factors related to scientific influence. The most common metrics are impact factor and H-index, which measure the number of citations of scholars' papers. Although these metrics have many problems such as lack of universality between different disciplines (Kaur, Radicchi, & Menczer, 2013), they are still widely used in systems like Google Scholar. Some alternative metrics also use the number of citations to measure the scientific influence of scholars (Ruscio, Seaman, D'Oriano, Stremlo, & Mahalchik, 2012). They offer advantages over simple metrics such as H-index, but they also focus solely on the citation count of papers. Other metrics based on the centrality of scholars in a network (e.g., co-authorship) like PageRank and betweeness centrality (Bollen, Van de Sompel, Hagberg, & Chute, 2009) are also widely used. However, the correspondence of centrality to actual influence is unknown.

As mentioned earlier, commonly used metrics of scholarly influence fail to consider breadth of scholars' research. In response a number of researchers have created some metrics for the degree of interdisplinarity and more generally breadth of research. The report of quantitative metrics and context in interdisciplinary scientific research (Wagner et al., 2011) is a good survey for metrics for interdisciplinarity. Specialization and integration (Porter et al., 2007) are good metrics of interdisciplinarity because they consider similarity between disciplines when measuring interdisciplinarity. They can be modified easily in the context of a diversity of research topics. Some papers discuss different dimensions of interdisplinarity (Rafols & Meyer, 2010; Rafols, Leydesdorff, O'Hare, Nightingale, & Stirling, 2012): diversity, coherence and intermediation. They define diversity as a combination of variety, balance and disparity. Coherence means link strength between different disciplines. Intermediation is based on the network structure and is measured by betweenness centrality, clustering coefficient and average similarity. Other papers describe metrics based on these dimensions. Cassi, Mescheba, and de Turckheim (2014) divides the Stirling metric into "within component" and "between component" to measure the diversity of articles. Jensen & Lutkouskaya (2013) defines six indicators based on the dimensions and measure the breadth of research at two levels (article and laboratory). Karlovčec and Mladenić (2014) defines a new diversity metric based on Generalized Stirling. The metric incorporates connectedness of the citation graph into the original metric and applies it in exploratory analysis of the research community in Slovenia. Roessner, Porter, Nersessian, and Carley (2012) validates the interdisciplinarity metrics with ethnographic materials (field observations and unstructured interviews).

Finally, some research has focused on the relationship between breadth of research and other factors considered in scientometrics (not just scientific influence). One interesting paper finds that the papers with an average degree of interdisciplinarity will get higher impact than papers with too high or too low degree of interdisciplinarity (Sternitzke & Bergmann, 2008). The results are convincing but metrics used in this paper are quite simple (Jaccard similarity and cosine similarity). Two papers find that interdisciplinary papers have potentially lower impact than more focused papers. One of them finds that multidisciplinary papers are not frequently cited in contrast to the disciplinary papers (Levitt & Thelwall, 2008). The other explains how high-ranked journals suppress interdisciplinary research (I Rafols & Meyer, 2010). Other papers describe some factors that can encourage researchers to be involved in interdisciplinary research work (Carayol & Thi, 2005; van Rijnsoever & Hessels, 2011). They provide some theories to explain why scholars choose interdisciplinary projects. Some findings support that there are no correlations between citation ranks and ranked interdisciplinarity indices (Ponomarev, Lawton, Williams, & Schnell, 2014). In contrast, other researchers confirm that the degree of interdisciplinarity is strongly correlated with the impact factor (Silva, Rodrigues, Oliveira, & da F. Costa, 2013).

Dataset

We extract abstracts, full text and other metadata from the ACM digital library for proceedings of major conferences in computer science. From these proceedings we select authors whose names are unambiguous and who have published at least five papers. The standard for unambiguity is whether using the full name as the query sent to Google Scholar returns only one researcher profile with the same name. We extract the citation numbers and H-indexes by crawling over Google Scholar. Overall we crawled H-indexes and citation numbers for 8911 authors from Google Scholar in August 2014. We also used the Wikipedia dataset to extract important terms in computer science.

Topic Extraction and Assignment

Both traditional metrics and the new metric designed in this paper require a distribution over different topics or areas for authors. In order to generate topic distributions, we leverage the text data in the papers of ACM digital library and implement three steps to form distributions: dictionary extraction, topic extraction and author assignment.

Dictionary Extraction

How to define topics is the first problem to be solved in the topic extraction and assignment. In our work, we extract a dictionary of n-grams in computer science and cluster them into topics using the Affinity Propagation algorithm (Frey & Dueck, 2007). Three different sources of dictionaries are used in this paper: grams that are frequently used in papers, grams that can be matched to their abbreviations in the papers, and entries in Wikipedia.

Dictionary extraction follows these steps:

1. Extract bigrams and trigrams that occur frequently in papers using a threshold of more than 10 times for bigrams and more than 5 times for trigrams. The threshold helps to eliminate noisy grams with low frequency.
2. Extract grams from papers that conform to the pattern "n-grams (abbreviation)", e.g. machine learning (ML).
3. Intersect the results of step 1 and step 2 (3816 terms in total).
4. Build a network of entries in Wikipedia according to hyperlinks between them in the website.
5. Make use of grams in step 3 and search their neighbours in the network of Wikipedia terms. If their neighbours also occur frequently in papers (with frequency higher than the thresholds mentioned above), add the terms into the final dictionary (6100 terms)

The top 5 bigrams and top 5 trigrams in the final dictionary are shown in Table 1:

Table 1. Grams with top frequency

Grams	*Frequency*
User Interface	2372
Software development	2102
Programming language	2042
Software engineering	1988
Operating system	1761
Wireless sensor network	586
World wide web	467
Graphical user interface	305
Support vector machine	300
Discrete event simulation	287

Topic Extraction and Assignment

After extracting the dictionary, we count the co-occurrence measure for every pair of terms. We then calculate the similarity between different terms by:

$$Sim_{ij} = \log \frac{Cooccur_{ij} + 1}{Max(Cooccur_{ij}) + 2}$$

The logarithm calculation makes the distribution of similarity more uniform and avoids the influence of outliers of co-occurrence numbers. We weight co-occurrences of terms in abstracts of papers more than those in full text based on the intuition that abstracts generally have a stronger “topic signal”. Using the computed similarity matrix of terms, we then run Affinity Propagation to cluster together similar terms and choose an exemplar for every cluster. The benefits of Affinity Propagation are that there isn’t a need to parameterize the number of clusters and that the exemplars for every cluster provide a straightforward explanation of what these clusters are about. More than two hundred clusters, or topics, are generated. Here are two examples of the clustering results:

Exemplar: digital library

Terms:
citation analysis, citation index, community building, digital earth, digital library, digital library software, digital preservation, digital reference, discourse analysis, dublin core.
Exemplar: machine learning
Terms:
active learning, adaptive control, bayes classifier, belief propagation, clinical trial, computational learning theory , concept learning, conditional random field.

We then assign every paper a probabilistic assignment to the different topics according to their respective frequency of n-grams associated with the particular topic. Therefore, every paper will have a distribution over topics.

Author Assignment

Using the clusters of grams in computer science and the topic distributions for every paper, we assign authors into different topics according to their papers. Every author is represented by a distribution over topics, which are used to calculate scores of metrics. There does not exist a "gold standard" list of researchers that ranks breadth of research that we can use to evaluate how reasonable our topic assignments are. We list below some topic distributions for well-known computer scientists to demonstrate our assignment.

John Koza	
1 genetic programming	0.567
2 programming language	0.083
3 knowledge base	0.063
Peter Denning	
1 memory management	0.107
2 computer systems	0.093
3 information systems	0.050
Eric Horvitz	
1 user interface	0.082
2 information retrieval	0.067
3 machine learning	0.051
4 speech recognition	0.047

Breadth of Research Measurement

With the author distribution of topics established, the key question is how to translate this into a measure of breadth of research for authors. As mentioned in the section describing related work, many metrics have been used to measure the "degree of interdisciplinarity". Compared to previous metrics to measure breadth of research, we design a new metric that considers the topic distribution, similarity distribution and coherence within research topics.

Summary of Old Measurements

There are many measurements of diversity or interdisciplinary, like entropy (Weaver, 1949), Simpson's index (Simpsons, 1949) and generalized Stirling (Stirling, 2007). Each of these is computed as follows. Denote p_i as the probability of topic distribution for an author over topic i, d_{ij} as the distance between topic i and topic j.

$$Entropy = \sum_{i=1}^{n} -p_i \times \log_n (p_i)$$

$$Simpson = 1 - \sum_{i=1}^{n} {p_i}^2$$

$$Generalized\ Stirling = \sum_{i,j} d_{ij}^{\alpha}\ (p_i \times p_j)^{\beta}$$

Comparing them, only generalized Stirling considers not only the distribution of topics but also the similarity between topics. The further the distance between topics in which an author publishes papers, the more diverse will the author's research interest be. However, the traditional metrics do not consider the notion of differing *coherence* between different research topics. And the degrees of influence of topics with small proportions are very limited. The new measurement is a modified version of the generalized Stirling metric and it incorporates the coherence of topics and value of *minor topics* (topics with small proportions).

New Measurement

The new metric for breadth of research is defined as follows.

Denote d_{ij} as the distance between two topics, which are defined as the average distance (inverse of similarity defined above) between terms in the two topics, p_i as the probability of an author's paper belong to topic i, coh_i as the *coherence* of topic i. Coherence of each topic is the proportion of authors for whom the respective topic is their major research topic, which is an important signal to illustrate whether a research topic concentrate on some core research questions. Parameters α, β, γ are used to control the relative weights of different components.

$$Breadth\ of\ Research = \sum_{i,j} d_{ij}^{\alpha}\ (p_i + p_j)^{\beta}(Coh_i \times Coh_j)^{\gamma}$$

We modify the product of p_i and p_j in generalized Stirling to summation of p_i and p_j because the summation will give minor topics more chances to be counted into the measurement of breadth of research. We add the coherence term into the metric because different topics have different "density" within themselves. For example, some topics like digital library are less coherent topics because there are many diverse subtopics in these topics. But for topics like operation systems, researchers concentrate on several narrow subtopics. A researcher focusing on digital library should have larger breadth of research than operating systems researchers if other variables are controlled (so the gamma should have a negative value).

The new metric leverages properties of papers (topic distribution), properties of topics (coherence) and properties of relationship (topic similarity). The tunable parameters give the metric more flexibility to balance between different aspects of breadth of research.

Experiments

Simulation Experiment

There is no established standard for determining the quality of metrics of breadth of research. Furthermore, there is no ground truth to show the rankings of scholars' breadth of research with which to validate the various metrics. We propose an alternative evaluation method based on a set of axioms concerning breadth of research and then test how the metrics perform according to these axioms.

In addition to the definition of d_{ij} and coh_i defined in the previous section, the following definitions relate to the axioms.

- Denote A_i as the article i, $C=\{ A_1, A_2 ...\}$ as a *collection* of articles, and N_C as the number of articles in collection C.
- Denote t_i as the topic i, $D_A(t)$ as the topic distribution of article A over topic t. $(\sum_t D_A(t) = 1)$

- Denote $D_C(t)$ as the topic distribution of collection C over topic t. $D_C(t)$ $=\frac{1}{N_C}\sum_{A_i \in C} D_{A_i}(t)$. $(\sum_t D_C(t) = 1)$
- Denote *score(C)* as the score of a metric over the collection of articles C

Axiom1: Publish in Old Topics

If an author publishes a paper in a topic in which she has published many papers before, her breadth of research should decrease.

Choose t, s.t. $t = arg\ max_t\ D_C(t)$, construct a new article A_{new}, s.t. $D_{A_{new}}(t) = 1$. $C' = C \cup \{A_{new}\}$. Then *score(C') < score(C).*

Axiom2: Publish in New Topics

If an author publishes a paper in a new topic in which she has never published, her breadth of research should increase.

Choose t, s.t. $D_C(t)=0$, construct a new article A_{new}, s.t. $D_{A_{new}}(t) = 1, C' = C \cup \{A_{new}\}$.

Then *score(C') > score(C).*

Axiom3: Publish in New Topics Twice

If an author publishes papers in two new topics in a sequence, the increase of breadth of research in the second time should be smaller than the increase of that in the first time.

Choose t_1 and t_2, s.t. $D_C(t_1)=0$, $D_C(t_2)=0$, $t_1 \neq t_2$,construct two new articles A_{new1} and A_{new2}, s.t. $D_{A_{new1}}(t) = 1$ and $D_{A_{new2}}(t) = 1$. $C' = C \cup \{A_{new1}\}$, $C'' = C' \cup \{A_{new2}\}$. Then *score(C')-score(C) > score(C'')-score(C').*

Axiom4: Publish in Close Topics

If an author publishes a paper in a new topic close to the author's research interest, the improvement of her breadth of research should be less than that of publishing a new paper in a randomly chosen topic.

Randomly Choose t_1 s.t. $D_C(t_1)=0$, construct a new article A_{new1}, s.t. $D_{A_{new1}}(t_1) = 1$. $C' = C \cup \{A_{new1}\}$. Choose t_2 s.t. $D_C(t_2)$=0 and $arg\ min_t(inf_{t_0 \in \{t|D_C(t)>0\}} d_{t_0 t_1})$. Construct a new article A_{new2}, s.t. $D_{A_{new2}}(t_2) = 1, C'' = C' \cup \{A_{new2}\}$. Then *score(C'') < score(C')*

Axiom5: Publish in Coherent Topics

If an author publishes a paper in a new topic with high coherence, the improvement of her breadth of research should be less than that of publishing a new paper in a randomly chosen topic.

Randomly Choose t_1 s.t. $D_C(t_1)=0$, construct a new article A_{new1}, s.t. $D_{A_{new1}}(t_1) = 1$. $C' = C \cup \{A_{new1}\}$. Choose t_2 s.t. $D_C(t_2)=0$ and $t_2 = arg\ max_t\ (Cohe_t)$. Construct a new article A_{new2}, s.t. $D_{A_{new2}}(t_2) = 1, C'' = C' \cup \{A_{new2}\}$. Then *score(C'') < score(C').*

We implemented five simulation experiments based on the original dataset with 8911 authors to test how the traditional metrics and our new metric conform to the axioms. The results are shown in Table 2.

Table 2. Probability that metrics satisfy of the axioms

	Entropy	*Simpson's*	*GL Stirling*	*New Metric*
			($\alpha = 2; \beta = 0.3$*)*	*(*$\alpha = 1, \beta = 0.5, \gamma = -0.5$*)*
Axiom1	0.99	0.99	0.97	0.88
Axiom2	0.89	0.97	0.86	0.86
Axiom3	0.97	0.94	0.50	0.50
Axiom4	0	0	0.76	0.70
Axiom5	0	0	0.54	0.62

The results show that entropy and Simpson's perform well in the first three axioms because they don't consider distances between topics and introduce less noise. Because every new topic will be regarded equally for these metrics, they cannot follow Axiom4 and Axiom5. Generalized Stirling and our metric perform reasonably well in Axiom1 and Axiom2, but worse than entropy and Simpson's. They perform relatively badly in Axiom3 because relatively bad performance on publishing a paper in new topic (Axiom2) will aggregate when testing the performance of publishing two papers in two new topics. But they perform well in Axiom4 because of the consideration of distances. Also we find our metric performs better than generalized Stirling in Axiom5, which means coherences of topics and greater weights on minor topics are beneficial when we consider variation of metrics when people publish in topics with different coherence levels.

Parameter Sensitivity

The performance of new metric is influenced by the value of parameters α, β and γ. We tested the performance of the new metric with different settings. The results are shown in Table 3, Table 4 and Table 5.

Table 3. Average Prob of satisfying the axioms with different α.

	$\alpha = 0.1$	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$
Axiom1	0.40	0.42	0.48	0.62
Axiom2	0.33	0.38	0.44	0.55
Axiom3	0.34	0.32	0.24	0.22
Axiom4	0.38	0.57	0.66	0.64
Axiom5	0.63	0.61	0.57	0.52

Table 4. Average Prob of satisfying the axioms with different β.

	$\beta = 0.1$	$\beta = 1$	$\beta = 10$	$\beta = 100$
Axiom1	0.86	0.67	0.30	0.08
Axiom2	0.69	0.57	0.24	0.16
Axiom3	0.40	0.40	0.29	0.05
Axiom4	0.57	0.57	0.59	0.53
Axiom5	0.61	0.61	0.59	0.52

Table 5. Average Prob of satisfying the axioms with different γ.

	$\gamma = 0.1$	$\gamma = 1$	$\gamma = 10$	$\gamma = 100$
Axiom1	0.58	0.47	0.45	0.45
Axiom2	0.24	0.39	0.47	0.48
Axiom3	0.09	0.26	0.34	0.38
Axiom4	0.49	0.57	0.59	0.59
Axiom5	0.62	0.66	0.58	0.53

The tables show that the metric is very sensitive to the α, β and γ. In order to find the best parameter setting, we calculated the average performance over five different simulation experiments for every parameter settings. We selected the settings with highest average performance and a minimum threshold of at least 0.5 in every experiment. The best setting for Generalized Stirling is $\alpha = 2, \beta = 0.3$. The best setting for the new metric is $\alpha = 1, \beta = 0.5$ and $\gamma = -0.5$. They are used in the comparison of metrics in Table 2.

Summation Modification

One of important modifications of our metric is the replacement of product with summation in the second term of metric. We test the effect of this. If we control the distance term and coherence term in the metric to be the same for every topic and set $\beta = 1$. The metric using summation will definitely follow Axiom2 but not follow Axiom1 and Axiom3.

Let *n* represents the number of topic.

Axiom1: Publish in Old Topics

$$score(C) = \sum_{i,j} d^{\alpha}\ (p_i + p_j)(\mathrm{coh} \times \mathrm{coh})^{\gamma} = (n-1)d^{\alpha}(coh)^{2\gamma}$$

$$= \sum_{i,j} d^{\alpha}\ (p_i{}' + p_j{}')(\mathrm{coh} \times \mathrm{coh})^{\gamma} = score(C')$$

Axiom2: Publish in New Topics

$$score(C) = \sum_{i,j} d^{\alpha}\ (p_i + p_j)(\mathrm{coh} \times \mathrm{coh})^{\gamma} = (n-1)d^{\alpha}(coh)^{2\gamma}$$

$$< \sum_{i,j} d^{\alpha}\ (p_i{}' + p_j{}')(\mathrm{coh} \times \mathrm{coh})^{\gamma} = (n)d^{\alpha}(coh)^{2\gamma} = score(C')$$

Axiom3: Publish in New Topics Twice

$$score(C) = (n-1)d^{\alpha}(coh)^{2\gamma}$$

$$score(C') = (n)d^{\alpha}(coh)^{2\gamma}$$

$$score(C'') = (n+1)d^{\alpha}(coh)^{2\gamma}$$

$$score(C'') - score(C') = score(C') - score(C)$$

From the derivation above, the performance of new metric in Axiom 1 and Axiom 3 should be worse than the metric with product. The performance of Axiom 2 should be better than the metric with product. So we construct a metric using product in the second term and compare the performance of it with the new metric in different parameter settings.

$$Breadth\ of\ Research = \sum_{i,j} d_{ij}^{\alpha}\ (p_i \times p_j)^{\beta}(Coh_i \times Coh_j)^{\gamma}$$

The results in Table 6 shows that the metric using summation outperforms product in Axiom 2, and metric using product outperforms summation in Axiom1, which is consistent with the results of derivation. But the results for the other three axioms are close between the two metrics, which means the interaction between different terms in the metric (distance term, distribution term and coherence term) will influence the results of simulation.

Table 6. Comparison between metric with summation and production.

Metric	*Parameter setting*	*Axiom1*	*Axiom2*	*Axiom 3*	*Axiom4*	*Axiom5*
Production	$\alpha = 0.1\ \beta = 0.1 \gamma = -0.1$	0.99	0.85	0.45	0.22	0.59
	$\alpha = 100\ \beta = 1 \gamma = -1$	0.82	0.62	0.47	0.69	0.53
	$\alpha = 1\ \beta = 1 \gamma = -10$	0.83	0.40	0.39	0.55	0.76
Summation	$\alpha = 0.1\ \beta = 0.1 \gamma = -0.1$	0.97	0.89	0.45	0.22	0.59
	$\alpha = 100\ \beta = 1\ \gamma = -1$	0.69	0.69	0.50	0.69	0.55
	$\alpha = 1\ \beta = 1 \gamma = -1$	0.69	0.47	0.41	0.54	0.77

Relationship between breadth of research and scientific impact

We tested the Pearson correlation between metrics of breadth of research and H-indexes of scholars. Our results (Table 7) show that some metrics have a positive relationship with H-index. Others have weak negative relationship. Because publication numbers may influence the correlation between breadth of research and scientific impact i.e. the increase of numbers of publications may bring increase of breadth of research and increase of H-index simultaneously to make them positively correlated to each other, we test the partial correlation between metrics of breadth of research to H-index controlling publication numbers (Table 7). They are weaker than Pearson correlations. And all the weak partial correlation scores don't illustrate strong correlation between metrics for breadth of research and H-index for scholars.

Table 7. Correlation between breadth of research and H-index.

	Pearson Corr.	*Partial Corr.*
Entropy v.s. H-index	-0.1722	-0.0769
Simpson's v.s. H-index	0.2102	0.0922
GL Stirling v.s. H-index	0.4283	0.1820
New Metric v.s. H-index	0.4337	0.1832

The Variation of metrics over publication years

We illustrate in Figure 1 the variation of average scores of metrics for all the scholars over publication years. Simpson's, generalized Stirling and our new metric initially increase and then level off, which explains a possible publication pattern of scholars: scholars' breadth of research may increase with the increase of publications in the early stage of their career. But because of accumulation of publications, their accumulative breadth of research will not change dramatically in the late years. For the entropy metric with base n, it is normalized by topic number. So it keeps in a stable level over year, which shows a different pattern compared to other metrics.

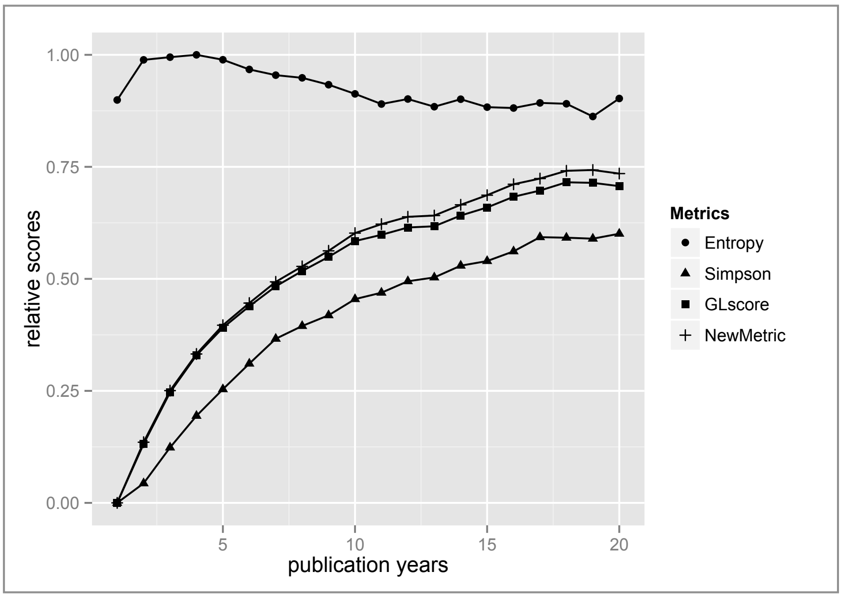

Figure 1. Variation of metrics over publication years.

Conclusion and Future Work

In this paper, we describe a new metric based on generalized Stirling to evaluate breadth of research for scholars in computer science. The metric makes use of topic distribution, similarity between topics, and coherence of topics and it can capture the diversity aspects of breadth of research. The simulation experiments show that traditional metrics can perform well in some axiom, but they don't perform well when coherence within topics and similarity between topics are considered. In contrast, generalized Stirling metric and the new metric for breadth of research work better in the simulation related to similarity between topics and coherences but perform worse in the experiments of adding new topics. It is a trade-off between the simplicity of metrics and the concern of topic similarity and coherence.

With the new metric for breadth of research, we find the correlation between breadth of research and scientific metrics are weak, especially when we control publication numbers. From our study, there's no evidence to show whether the increase of breadth of research will influence the impact of scholars' publication. Also, after testing the variation of the new metric over years, we find a possible publication pattern of scholars: Breadth of research increases in the beginning with the increase of publications. But they increase slowly when publications have been accumulated.

There are a number of research questions that arise from the work described in this paper. The first one is finding alternative methods to generate research topics. Unsupervised learning models based on both text contents and citation information may be helpful to extract topics and show topic variation for authors. The second question is how to improve the simulation results for the new metric. The new metric performs better than general Stirling and other traditional metrics in some aspects. But if more information from co-author and citation network can be incorporated into the metric, the performance may be better and interpretable.

Acknowledgments

The research is funded by NSF 1258891 EAGER: Collaborative Research: Scientific Collaboration in Time.

References

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PloS One*, *4*(6), e6022. doi:10.1371/journal.pone.0006022

Boyack, K., & Klavans, R. (2009). Measuring multidisciplinarity using the circle of science. *From WRK1: Tracking and Evaluating Interdisciplinary Research, Workshop at ISSI*, *87122*.

Carayol, N., & Thi, T. (2005). Why do academic scientists engage in interdisciplinary research? *Vasa*.

Cassi, L., Mescheba, W., & de Turckheim, É. (2014). How to evaluate the degree of interdisciplinarity of an institution? *Scientometrics*. doi:10.1007/s11192-014-1280-0

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, *315*(5814), 972–6. doi:10.1126/science.1136800

He, B., Ding, Y., Tang, J., Reguramalingam, V., & Bollen, J. (2013). Mining diversity subgraph in multidisciplinary scientific collaboration networks: A meso perspective. *Journal of Informetrics*, 1–18.

Jensen, P., & Lutkouskaya, K. (2013). The many dimensions of laboratories' interdisciplinarity. *Scientometrics*, *98*(1), 619–631. doi:10.1007/s11192-013-1129-y

Jo, Y., Hopcroft, J., & Lagoze, C. (2011). The web of topics: discovering the topology of topic evolution in a corpus. *Conference on World Wide Web*, 257–266.

Karlovčec, M., & Mladenić, D. (2014). Interdisciplinarity of scientific fields and its evolution based on graph of project collaboration and co-authoring. *Scientometrics*. doi:10.1007/s11192-014-1355-y

Kaur, J., Radicchi, F., & Menczer, F. (2013). Universality of scholarly impact metrics. *Journal of Informetrics*, *7*(4), 924–932. doi:10.1016/j.joi.2013.09.002

Levitt, J. M., & Thelwall, M. (2008). Is multidisciplinary research more highly cited? A macrolevel study. *Journal of the American Society for Information Science*, *59*, 1973–1984. doi:10.1002/asi.20914

Li, D., He, B., Ding, Y., Tang, J., Sugimoto, C., Qin, Z., Dong, T. (2010). Community-based topic modeling for social tagging. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management - CIKM '10*, 1565. doi:10.1145/1871437.1871673

Nichols, L. G. (2014). A topic model approach to measuring interdisciplinarity at the National Science Foundation. *Scientometrics*, 741–754. doi:10.1007/s11192-014-1319-2

Ponomarev, I. V., Lawton, B. K., Williams, D. E., & Schnell, J. D. (2014). Breakthrough paper indicator 2.0: can geographical diversity and interdisciplinarity improve the accuracy of outstanding papers prediction? *Scientometrics*, 755–765. doi:10.1007/s11192-014-1320-9

Porter, A. L., Cohen, A. S., David Roessner, J., & Perreault, M. (2007). *Measuring researcher interdisciplinarity*. *Scientometrics* (Vol. 72, pp. 117–147). doi:10.1007/s11192-007-1700-5

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., & Stirling, A. (2012). How journal rankings can suppress interdisciplinary research: A comparison between innovation studies and business & management. *Research Policy*, *41*(7), 1262–1282. doi:10.1016/j.respol.2012.03.015

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics*, 1–28.

Roessner, D., Porter, A. L., Nersessian, N. J., & Carley, S. (2012). Validating indicators of interdisciplinarity: linking bibliometric measures to studies of engineering research labs. *Scientometrics*, *94*(2), 439–468. doi:10.1007/s11192-012-0872-9

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(4), 1118–23. doi:10.1073/pnas.0706851105

Ruscio, J., Seaman, F., D'Oriano, C., Stremlo, E., & Mahalchik, K. (2012). Measuring scholarly impact using modern citation-based indices. *Measurement: Interdisciplinary Research & Perspective*, *10*(3), 123–146. doi:10.1080/15366367.2012.711147

Silva, F. N., Rodrigues, F. a., Oliveira, O. N., & da F. Costa, L. (2013). Quantifying the interdisciplinarity of scientific journals and fields. *Journal of Informetrics*, *7*(2), 469–477. doi:10.1016/j.joi.2013.01.007

Simpsons, E. H. (1949). Measurement of Diversity. Retrieved October 9, 2014, from http://www.nature.com/nature/journal/v163/n4148/abs/163688a0.html

Sternitzke, C., & Bergmann, I. (2008). Similarity measures for document mapping: A comparative study on the level of an individual scientist. *Scientometrics*, *78*(1), 113–130. doi:10.1007/s11192-007-1961-z

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society, Interface / the Royal Society*, *4*(15), 707–19. doi:10.1098/rsif.2007.0213

Van Rijnsoever, F. J., & Hessels, L. K. (2011). Factors associated with disciplinary and interdisciplinary research collaboration. *Research Policy*, *40*(3), 463–472. doi:10.1016/j.respol.2010.11.001

Velden, T., & Lagoze, C. (2013). The extraction of community structures from publication networks to support ethnographic observations of field differences in scientific communication. *Journal of the American Society for Information Science and Technology*, *64*(12), 2405–2427.

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J.T., Boyack, K. W., Keyton, J., Rafols, I., & Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics*, *5*(1), 14–26. doi:10.1016/j.joi.2010.06.004

Weaver, W. (1949). Recent Contributions to the Mathematical Theory of Communication 1 Introductory Note on the General Setting of the Analytical Communication Studies.

Weingart, P. (2005). Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics*, *62*(1), 117–131.

Yan, E. (2013). Finding knowledge paths among scientific disciplines. *arXiv Preprint arXiv:1309.2546*, (812), 1–31.