Learning Nuanced Cross-Disciplinary Citation Metric Normalization using the Hierarchical Dirichlet Process on Big Scholarly Data

Hafsah Umar Ognjen Arandjelović

School of Computer Science
University of St Andrews
St Andrews, KY16 9SX
Fife, Scotland
United Kingdom

ognjen.arandjelovic@gmail.com

ABSTRACT

Citation counts have long been used in academia as a way of measuring, *inter alia*, the importance of journals, quantifying the significance and the impact of a researcher's body of work, and allocating funding for individuals and departments. For example, the h-index proposed by Hirsch is one of the most popular metrics that utilizes citation analysis to determine an individual's research impact. Among many issues, one of the pitfalls of citation metrics is the unfairness which emerges when comparisons are made between researchers in different fields. The algorithm we described in the present paper learns evidence based, nuanced, and probabilistic representations of academic fields, and uses data collected by crawling Google Scholar to perform field of study based normalization of citation based impact metrics such as the h-index.

CCS Concepts

•**Computing methodologies** → Machine learning; •**Applied computing** → Computers in other domains;

Keywords

Academic; publication; publishing; quantification; university; index; science.

SAC'17, April 3-7, 2017, Marrakesh, Morocco

http://dx.doi.org/xx.xxxx/xxxxxxx.xxxxxxx

1. INTRODUCTION

Notwithstanding various lines of criticism they have received [1], various metrics of academic prestige, productivity, and quality have become pervasive in the research community. Publication and citation counts, in different forms, are widely used to quantify the impact that a researcher or a publication venue (such a conference or a journal) has. The *journal impact factor* is one of the most established and longest used metrics with journal citation reports (JCR) for peer reviewed journals being released annually since 1976 [11]. As its name suggests, the broad aim of this metric is to quantify the impact of articles appearing in a specific journal. Informally, it is often used as a proxy measure for the quality of a journal. In recent years, and the last decade specifically, citation based indexes for researchers have also become pervasive. The total citation count of all articles published by a researcher and the so-called h-index are amongst the most widely used indexes. Numerous others have also been described, including the e-index [23], g-index [8], z-index [18], and i10-index [12].

The primary argument against the use of the aforementioned metrics is that they do not assess directly the substance of a work itself. Instead they rely on proxy observations which, while certainly affected by the aforementioned substance, are confounded by numerous other factors including the visibility of the work, affected by where the work was published, what the researcher's institution is, the prior reputation of the researcher is etc. Another major criticism concerns the phenomenon of so-called honorary authorship [10] [1].

[1] Also see:
http://ijcai-16-pc.blogspot.co.uk/2016/04/the-increasing-practice-of-expanding-co.html
and
http://ijcai-16-pc.blogspot.co.uk/2016/04/

Considering the inherently subjective understanding of what 'impact' and 'quality' mean in the context of academic work and the lack of an objective basis (the 'ground truth') for assessing the fairness of a particular index, unlike different previous authors who have described and argued in favour of different indexes in this paper we do not aim to introduce a new index *per se*. Rather, accepting the pragmatic standpoint that for better or worse citation indexes are being increasingly used in academia [16], we show how a any citation count based index can be normalized to make it *prima fasciae* fairer when applied to a comparison of researchers in different fields.

1.1 Previous work

Different fields of academic research are characterized by different publication and citation dynamics. This is poignantly illustrated by considering the statistics summarized in Table 1 for fields of study recognized by the Institute for Scientific Information (ISI). Certain research areas attract more researchers, have a shorter work-to-publication turnover time, offer a greater number of peer reviewed publication venues, generate more articles etc. As can be observed from Table 1, the areas of medical and biomedical research are particularly prolific in this sense. This phenomenon has been widely acknowledged which is why already Hirsch warned against the use of the h-index for inter-disciplinary comparisons [14]. The subsequent analysis of Iglesias and Pecharroman demonstrated this convincingly using empirical data [15].

A major limitation with the approach of Iglesias and Pecharroman concerns the concept of a 'field of study' and the hard delineation between these fields [15]. To give an example, should an article published in a bioinformatics journal be considered as being in the field of medicine, computer science, or an entirely separate field of biomedicine? Similarly, it may be asked if, say, pattern recognition or computer vision should for the purposes of the problem at hand be considered separate fields, a single field, or indeed should they be both treated as belonging to computer science? We argue that the answer should be evidence and data driven, and demonstrate how this can be achieved.

The problem here is that the language used to describe different fields of study is not intended for use in rigorous formalizations such as this. The idea behind the present work is that rather than having manually crafted academic fields described using language not fit for purpose at hand, the large amounts of scholarly data can be leveraged through the use of sophisticated, automatic machine learning methods to learn nuanced descriptions of academic fields, which can then be used to perform field specific citation normalization and thus facilitate an inter-discipline normalization of research output metrics.

the-increasing-practice-of-expanding-co.html.

2. PROPOSED APPROACH

In this section we present the main technical contributions of the present work; experimental contributions are presented in the next section. Herein, following an overview of our approach, we summarize the key aspects of Bayesian non-parametric topic models, central to the proposed algorithm. Thereafter we describe how the hierarchical Dirichlet process based model – a particularly powerful Bayesian non-parametric topic model – can be used to infer automatically nuanced descriptions of academic fields of study and thus facilitate the normalization of citation based research output metrics, such as the h-index.

2.1 Overview

The work described in the present paper involved three key stages. The first of these concerns the collection of vast amounts of scholarly data from the Internet. In particular, as we describe in the next section, we developed a tool which given a small number of "seed" individuals, crawls Google Scholar and collects details of articles published by a large number of researchers. The second stage, which includes the key technical novelty of our work, extracts nuanced descriptions of academic fields of study from the collected data. In particular we adopt the use of Bayesian non-parametric topic models, originally introduced for text analysis. The novelty of our approach lies in the idea of treating each researcher as a "document" (to adopt the usual terminology from text analysis), and different publication venues (journals and conferences) as "terms" (or more colloquially "words"). The co-occurrence statistics of different publication venues across different researchers can then be used to infer topics – probability distributions over different publication venues – which represent the sought after academic fields. The key technical background on topic models is explained in the next section. Lastly, the inferred nuanced descriptions of academic fields are used for citation normalization by considering the likelihood of each article (i.e. the associated publication venue) corresponding to a specific field and the distribution of citation counts in that field.

2.2 Technical background: probabilistic topic modelling

In the last decade and a half so-called topic modelling has emerged as a powerful statistical paradigm for the automatic semantic analysis of large collections of documents. Topic models as their name suggests can be seen as formalizations of the colloquial understanding of 'topics' addressed in a piece of text. More specifically, in this context a topic becomes a probability distribution over a fixed vocabulary of words (or more generally terms). Using higher order semantic understanding, a human interpreting this formal representation of a topic may describe it as being related to a subject matter which has a high chance of co-occurrence of the words inferred as being most probable under the represen-

Table 1: Average numbers of citations for papers published in different Institute for Scientific Information (ISI) recognized fields of research (per annum and on average across the period 1995–2005).

Field	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	Average
Agricultural sciences	8.4	8.1	7.6	7.4	6.8	6.1	4.8	3.5	2.3	0.9	0.2	4.9
Biology & biochemistry	26.5	24.5	24.4	21.8	19.5	17.4	14.1	10.5	6.8	3.0	0.5	15.4
Chemistry	13.6	12.9	12.3	11.8	10.7	9.8	7.8	6.3	4.1	2.0	0.3	8.1
Clinical medicine	19.1	17.3	16.3	15.1	13.7	12.0	10.0	7.7	4.9	2.2	0.4	10.6
Computer science	5.0	4.9	4.8	4.7	4.0	3.3	3.1	2.6	1.2	0.5	0.1	2.5
Economics & business	9.2	7.5	7.2	6.1	5.1	4.2	3.2	2.4	1.3	0.5	0.1	4.2
Engineering	5.4	5.1	5.2	4.6	4.2	3.7	3.1	2.3	1.4	0.6	0.1	3.2
Environment & ecology	14.6	13.9	13.2	12.4	10.8	9.6	7.2	5.3	3.3	1.4	0.2	7.8
Geosciences	15.1	14.1	13.2	12.2	10.3	8.5	6.9	4.7	3.0	1.3	0.3	7.6
Immunology	34.1	30.7	28.8	28.2	24.2	22.2	18.5	13.7	8.8	4.2	0.6	19.6
Materials science	7.6	7.3	6.8	6.6	6.0	5.5	4.6	3.4	2.2	1.0	0.1	4.3
Mathematics	5.2	4.9	4.4	3.9	3.6	2.9	2.3	1.7	1.0	0.5	0.1	2.7
Microbiology	24.3	22.6	22.0	20.7	18.3	15.8	13.0	9.8	6.3	2.9	0.5	14.0
Molecular biology & genetics	42.7	39.8	38.3	35.9	32.4	28.1	23.3	17.5	11.2	5.2	0.8	24.6
Neuroscience & behaviour	30.0	27.0	25.7	23.8	21.6	18.8	15.6	11.3	6.8	2.9	0.4	16.4
Pharmacology & toxicology	16.1	14.5	14.4	13.0	12.4	11.1	9.4	7.4	4.6	2.0	0.3	9.4
Physics	12.3	11.8	10.8	10.2	9.4	8.6	7.1	5.4	3.7	1.9	0.4	7.2
Plant & animal science	11.2	10.6	9.8	8.8	7.9	6.9	5.6	4.1	2.6	1.2	0.2	6.2
Psychiatry & psychology	15.2	13.8	13.4	11.9	11.0	9.0	7.3	5.0	3.1	1.3	0.2	8.2
Social sciences	6.0	5.7	5.5	5.1	4.5	3.9	3.0	2.3	1.4	0.6	0.2	3.5
Sports science	18.7	17.6	17.9	16.1	16.9	12.3	12.3	8.4	6.7	3.2	0.6	11.6

tation (although it should be noted that such interpretation may not always be straightforward [6]).

2.3 Bayesian non-parametric topic models

Finite mixture models rely on the assumption that the observed data is generated by K clusters, each cluster being associated with the parameter ϕ_k and underlain by the probability density function $f(.|\phi_k)$. An observation x is assumed to be generated by first choosing a cluster k with probability π_k followed by a random draw from the corresponding distribution described by ϕ_k. Therefore the process can be summarized by the following:

$$p(x|\pi_{1:K}, \phi_{1:K}) = \sum_{k=1}^{K} \pi_k f(x|\phi_k). \quad (1)$$

In a Bayesian setting the model parameters (i.e. mixing proportions $\pi_{1:K}$ and component parameters $\phi_{1:k}$) are further endowed by priors. Typically the symmetric Dirichlet distribution is placed on top of $\pi_{1:K}$ and a prior on $\phi_{1:K}$ conjugate with $f(.|\phi_k)$ chosen for computational convenience.

2.3.1 Latent Dirichlet allocation

In the previous section we described how to model a group of data points with a Bayesian finite mixture model. Latent Dirichlet allocation adds a level of hierarchy on the mixing proportions to allow for the modelling of data points in groups that share a set of components.

Following the consensus in the literature we adopt the terminology used in the analysis of textual data – which is the context in which LDA was originally proposed [3] – and hereafter interexchangably refer to data points as words, their groups as documents, and mixture components as topics. The technical term 'topic' can be interpreted as formalizing and abstracting the colloquial notion of a topic which is understood at a higher semantic level. Therefore the modelling framework of LDA can be described by the following

generative process:

$$\phi_{1:K} \sim H, \quad (2)$$

$$\pi_j \sim \text{Dirichlet}(\alpha), \quad (3)$$

$$z_{ji}|\pi_j \sim \pi_j, \quad (4)$$

$$x_{ji}|z_{ji}, \phi_{1:K} \sim F\left(\phi_{z_{ji}}\right), \quad (5)$$

where H is the base distribution of topics, α the hyperparameter of the prior on the distribution of topics within a document, π_j the distribution of topics in document j, and z_{ji} the topic corresponding to the i-th word in the j-the document. The corresponding model likelihood is:

$$p(w_{ji}|\alpha) = \int_{\pi_j} \int_{\phi_{1:K}} \sum_{k=1}^{K} \pi_{jk} f(x|\phi_k)\, dP(\pi_j)\, dP(\phi_{1:K}), \quad (6)$$

Approximation techniques such as MCMC [13] and Variation Bayes [3] methods can be used for posterior inference.

2.3.2 Infinite mixture modelling

As mentioned earlier, LDA requires the number of topics to be fixed in advanced which is a serious limitation in practice. Choosing the number of topics is usually performed by examining how well the model fits a set of held-out documents. However, if a previously unseen topic has contributed in generating the held-out data, LDA is not able to infer correct parameters of that topic.

Bayesian non-parametric (BNP) methods place priors on the infinite-dimensional space of probability distributions and provide an elegant solution to this problem. Dirichlet Process (DP) [9] as the non-parametric counterpart of the Dirichlet distribution and the building block of BNP allows for the model to accommodate a potentially infinite number of mixture components. The generative likelihood for a data point x in infinite mixture model is:

$$p(x|\pi_{1:\infty}, \phi_{1:\infty}) = \sum_{k=1}^{\infty} \pi_k f(x|\phi_k). \quad (7)$$

$\text{DP}(\gamma, H)$ is defined as a distribution of a random probability measure G over a measurable space $(\Theta, \mathcal{B})$, such that for any finite measurable partition $(A_1, A_2, \ldots, A_r)$ of Θ the random vector $(G(A_1), \ldots, G(A_r))$ is a Dirichlet distribution with parameters $(\gamma H(A_1), \ldots, \gamma H(A_r))$. A DP generates imperfect atomic copies of its base measure H with a variance governed by its concentration parameter γ. An alternative view of the DP emerges from the so-called stick-breaking process which adopts a constructive approach using a sequence of discrete draws [20]. Specifically, if $G \sim \text{DP}(\gamma, H)$ then $G = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$ where $\phi_k \overset{iid}{\sim} H$ and $\boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty}$ is the vector of weights obtained by a stick-breaking process that is $\beta_k = v_k \prod_{l=1}^{k-1}(1 - v_l)$ and $v_l \overset{iid}{\sim} \text{Beta}(1, \gamma)$.

Owing to the discrete nature and infinite dimensionality of its draws, the DP is a highly useful prior for Bayesian mixture models. By associating different mixture components with atoms ϕ_k of the stick-breaking process, and assuming $x_i|\phi_k \overset{iid}{\sim} f(x_i|\phi_k)$ where $f(.)$ is the likelihood kernel of the mixing components, we can formulate the infinite Bayesian mixture model or Dirichlet process mixture model (DPM). Approximate methods are used for posterior inference [17].

2.3.3 Hierarchical Dirichlet process mixture models

While DPM is suitable for non-parametric clustering of exchangeable data in a single group, many real-world problems are more appropriately modelled as comprising multiple groups of exchangeable data. In such cases it is usually desirable to model the observations of different groups jointly, allowing them to share their generative clusters. This idea is known as "sharing the statistical strength" and it is naturally obtained by hierarchical architecture in Bayesian modelling.

Consider a collection of documents. DPM models each group with an infinite number of topics. However, it is desired for multiple group-level DPMs to share their clusters. Amongst different ways of linking group-level DPMs, HDP [21] offers an interesting solution whereby base measures of group-level DPs are drawn from a corpus-level DP. In this way the atoms of the corpus-level DP (i.e. topics in our case) are shared across the documents. Formally, if $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_J\}$ is a document collection where $\mathbf{x}_j = \{x_{j1}, \ldots, x_{jN_j}\}$ is the j-th document comprising N_j words:

$$G_0|\gamma, H \sim \text{DP}(\gamma, H) \quad (8)$$

$$G_j|\alpha_0, G_0 \sim \text{DP}(\alpha_0, G_0) \quad (9)$$

$$\theta_{ji}|G_j \sim G_j \quad (10)$$

$$x_{ji}|\theta_{ji} \sim F(.|\theta_{ji}) \quad (11)$$

Since G_j is drawn from a DP with base measure G_0, it takes the same support as G_0. Also the parameters of the group-level mixture components, θ_{ji}, share their values with the corpus-level DP support on $\{\phi_1, \phi_2, \ldots\}$. Therefore G_j can be equivalently expressed using the stick-breaking process as $G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}$ where $\boldsymbol{\pi}_j|\alpha_0, \gamma \sim \text{DP}(\alpha_0, \gamma)$ [22]. The posterior for θ_{ji} has been shown to follow a Chinese restaurant franchise process which can be used to develop inference algorithms based on Gibbs sampling [21].

2.4 Nuanced field of study inference and probabilistic representation

The key conceptual contribution behind our method concerns the way in which topic modelling can be used to infer probabilistic representations of academic fields of study from data. Although topic modelling was originally developed for (and is still predominantly used for) text analysis, ingenious analogies which treat non-textual data as 'words' or 'documents' have demonstrated their usefulness in a broad range of domains, such as computer vision and image analysis.

For example, by considering fixed length representations of super-pixels as 'words' and images as 'documents' containing these 'words', previous work has demonstrated that abstract visual topics such as 'sky', 'grass', or 'aeroplane' can be inferred directly from data in an unsupervised manner. We perform a similar paradigm abstraction in the present work too.

In particular, we treat different peer reviewed publication venues (journals and conferences) as 'terms' and each researcher's output as a 'document' with each article corresponding to a 'term' defined by the venue where it was published. By applying hierarchical Dirichlet process based inference on a large data set of researchers, nuanced representations of research areas can be extracted automatically as probability distributions over a 'vocabulary' over the most frequent publication venues (as usual, the highly uncommon 'terms' are excluded from the dictionary as they are deemed to provide unreliable evidence). After such representations are extracted, field based normalization is citation counts for each article can be achieved in a straightforward fashion. Specifically, the citation count of an article is first distributed across different 'topics' according to their likelihoods, each contribution scaled inversely proportionally to the average citation count for the topic, and then added together. Lastly, the result is multiplied with the average citation count over all articles, in order to produce a meaningful citation count (this merely adjusts the absolute scale of the normalized counts, without affecting their relative values).

3. EMPIRICAL EXPERIMENTS

In this section we turn our attention to the empirical aspects of the present work. We start by describing the workings of the tool we developed to crawl and collect automatically scholarly data from Google Scholar, the automatic data clean-up and pre-processing necessary to facilitate subsequent robust inference, and finally present and discuss examples of the normalization achieved by our algorithm.

3.1 Data collection

The methodology described in the previous section necessities the use and availability of a large amount of scholarly data. In particular, our algorithm requires data on the publications of a large number of authors, with the associated citation counts. While information on the publication output for specific researcher of interest is indeed easily available (e.g. using Google Scholar or Microsoft Academic), collecting it for a large number of researchers is challenging. In particular, Google does not provide an API to access Scholar, and there are no databases of Google Scholar user IDs of different researchers.

Our idea was to design a crawler which uses minimal user input to get started and thereafter amasses data at an increasing rate automatically itself. In particular, we initialize the crawler using a small number of "seed authors". These are particularly prolific researchers which we selected manually. From this point on the crawler performs a breadth first search. Firstly, the crawler scrapes publication data (publication venue names for each paper and the corresponding number of citations) of the researcher currently being considered, as well as the links to Google Scholar pages of all of the person's co-authors, adding them into the crawling queue. To avoid repetition we keep track of already processed researchers. In principle the crawling can then continue until the author queue is empty. However, for the purposes of the present paper we terminated the collection process earlier, and used in our experiments the data of 3466 researchers.

3.1.1 Data clean-up and pre-processing

As even a cursory examination of typical Google Scholar pages readily confirms, raw data collected by our crawler requires significant clean-up and pre-processing before it is used as input to a Bayesian model of a type described in the previous section. We perform two key stages at this point: (i) publication venue canonization, and (ii) rejection of invalid data. The two are explained next.

Publication venue canonization.

Google Scholar references to the same publication venue, such as a conference or a journal, exhibit a great amount of variability. For our model to extract meaningful academic field representations, it is crucial that this variability is eliminated, that is, that all references to the same venue have the same form. For example, "JMLR", "Journal of Machine Learning Research", and "J Mach Learn Res" should all map to the same entry i.e. the same 'term' in the context of our topic modelling algorithm. We accomplish this task using fuzzy logic as the underlying matching methodology and data in the form of different standard abbreviations for publication venues (these can be readily obtained from various public sources, e.g. http://library.stanford.edu/guides/find-journal-abbreviations). To perform robust matching between raw data and various standard forms of referring to the same venue (full title, ISO 4 abbreviated etc) we adopt a fuzzy matching strategy. In particular we impose a matching penalty for each permissible transformation of the name, such as a deletion of a word or a letter (e.g. so that a match between 'conf' and 'conference' can be made), word re-ordering (e.g. so that a match between 'computer vision, conference on' and 'conference on computer vision' can be made) etc.

Rejection of invalid publication entries.

In addition to the challenge posed by non-standardized references to publication venues, a further difficulty is presented by the presence of entries which can for the purposes of this work be considered invalid. These include references to

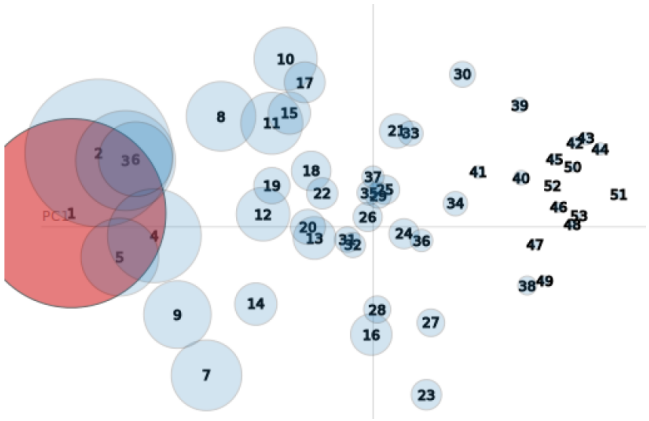

Figure 1: Inferred probabilistic representations of academic fields as topics, shown in the 2D principal component subspace using multidimensional scaling [4].

publication venues which are not peer reviewed or which are not considered credible (the clearest example being that of predatory journals). To remove such entries we remove all collected data which could not be matched (using the described fuzzy matching algorithm) with the comprehensive list of valid publication venues.

3.2 Results and discussion

In the experiments reporter herein we used the publications data of 3466 researchers. The number of peer reviewed articles ranged from 7 to over 1000 per researcher. The small number of articles found on the low end of the distribution in our data set can be easily seen to correspond to young academics whose publishing career is only starting. On the other hand the extremely high number of articles published by the individuals on the high end of the spectrum is highly indicative of the concerning phenomenon we noted earlier, that of so-called honorary authorship [10].

As our 'dictionary' of 'words' (publication venues) we selected the 1000 most common journals and conferences in the data set we collected. Applying our hierarchical Dirichlet process based inference on our data set, that is, treating researchers as 'documents' and their publications as 'words' associated with the corresponding publication venues, resulted in 64 topics i.e. probabilistic representations of fields of study. Observe that this is over three times the number of fields used by the ISI, which are shown in Table 1. The relationships between the inferred topics is illustrated in Figure 1 where topics are shown in the 2D principal component subspace constructed using a multidimensional scaling based data embedding algorithm [4]. As expected from the already noted unevenness in the number of publications in different fields, as illustrated by the sizes of blobs representing topics in Figure 1, most of the data is explained by relatively few topics (academic fields) with approximately 80% of the publications being in the first 20 inferred research fields. We observed a rough inverse power law distribution – observed frequently in nature across a wide range of phenomena [19, 7, 5, 2] – in the publishing output per research field.

Table 2: Randomly selected researchers from our database and the corresponding h-index before and after field based normalization using the proposed algorithm. † At the time of data collection (August 2016).

	h-index	
Author name	Original†	Field normalized†
Scott Shenker	136	41
Ramesh Govindan	82	32
Matei Zaharia	32	13
David Clark	53	29
Lixia Zhang	88	36
Ali Ghodsi	34	12
Vern Paxon	86	33
Randy Katz	108	36
Deborah Estrin	117	41

Finally, the performance of our method is illustrated on a set of typical and randomly selected examples of researchers working in different fields in Table 2, using the h-index as the baseline metric. As expected from theory, the broad range of values obtained using the original metric is drastically reduced with the application of the proposed normalization.

4. SUMMARY AND CONCLUSIONS

In this paper we addressed the problem of disciplinary bias exhibited by citation based metrics of research output, which has been widely recognized as their major limitation. This bias is a consequence of different publication dynamics characterizing different academic fields. Our starting point was an argument that the current state of the art, based around *ad hoc* and manually defined 'academic fields' is unprincipled and inadequate at capturing what is in reality a fuzzy rather than a crisp definition of a field. To address this problem in this paper we introduced the first truly data and evidence driven normalization approach which leverages so-called big data in the scholarly domain. In particular, (i) we created an automatic tool that crawls Google Scholar and collects researchers' publication information, and (ii) proposed a topic modelling based approach based on the hierarchical Dirichlet process that is able to extract automatically nuanced, probabilistic representations of academic fields. The latter stage was achieved by adopting a paradigm whereby each peer reviewed publication venue (journal or conference) is

considered a 'term', and a researcher's publication output as a 'document'. Using a large data set collected from Google Scholar, which will be made public following the publication of the present paper, we demonstrated the effectiveness of the proposed approach. We continue to collect more data and will make an implementation of the algorithm freely available for download, as well as provide an online tool which can be used to compute a variety of normalized metrics for researchers with Google Scholar profiles.

5. REFERENCES

[1] O. Arandjelović. Fairer citation based metrics. *Publishing Research Quarterly*, 32(3):163–169, 2016.

[2] A. Beykikhoshk, O. Arandjelović, D. Phung, S. Venkatesh, and T. Caelli. Using Twitter to learn about the autism community. *Social Network Analysis and Mining*, 5(1):5–22, 2015.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] I. Borg and P. Groenen. *Modern Multidimensional Scaling: theory and applications.* Springer-Verlag, New York, 2nd edition, 2005.

[5] J. P. Bouchaud and M. Mézard. Wealth condensation in a simple model of economy. *Physica A: Statistical Mechanics and its Applications*, 282(3):536–545, 2000.

[6] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei. Reading tea leaves: how humans interpret topic models. *Advances in Neural Information Processing Systems*, pages 288–296, 2009.

[7] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *Society for Industrial and Applied Mathematics Review*, 51(4):661–703, 2009.

[8] L. Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006.

[9] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

[10] A. Flanagin, L. A. Carey, P. B. Fontanarosa, S. G. Phillips, B. P. Pace, G. D. Lundberg, and D. Rennie. Prevalence of articles with honorary authors and ghost authors in peer-reviewed medical journals. *The Journal of American Medical Association*, 280(3):222–224, 1998.

[11] E. Garfield. The evolution of the science citation index. *International Microbiology*, pages 65–69, 2007.

[12] Google Scholar Blog. Google scholar citations open to all. 16 November, 2011.

[13] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Supplement 1):5228–5235, 2004.

[14] J. E. Hirsch. An index to quantify an individual's scientific research output. *In Proc. National Academy of Sciences of the United States of America*, 102(46):16569–16572, 2005.

[15] J. E. Iglesias and C. Pecharroman. Scaling the h-index for different scientific ISI fields. *Scientometrics*, 73(3):303–320, 2007.

[16] L. Meho. The rise and rise of citation analysis. *Physics World*, pages 32–36, 2007.

[17] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[18] A. M. Petersen and S. Succi. The z-index: A geometric representation of productivity and impact which accounts for information in the entire rank-citation profile. *Journal of Informetrics*, 7:823–832, 2013.

[19] L. F. Richardson. Variation of the frequency of fatal quarrels with magnitude. *Journal of the American Statistical Association*, 43(244):523–546, 1948.

[20] J. Sethuraman. A constructive definition of Dirichlet priors. Technical report, DTIC Document, 1991.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[22] Y. W. Teh, K. Kurihara, and M. Welling. Collapsed variational inference for HDP. *Advances in Neural Information Processing Systems*, pages 1481–1488, 2007.

[23] C. T. Zhang. The e-index, complementing the h-index for excess citations. *PLOS ONE*, 4(5):e5429, 2009.