Noname manuscript No.
(will be inserted by the editor)

Machine Learning Research at Alan Turing Institute: Bibliometric Overview and Trends

Waleed Iqbal[1] · Junaid Qadir[1] ·
Gareth Tyson[2] · Adnan Noor Mian[1 3] ·
Saeed-ul-Hassan[1] · Jon Crowcroft[3]

the date of receipt and acceptance should be inserted later

Abstract Machine learning is a major research discipline and an intersection of different major research domains. The field has been actively growing, in terms of both research and development, for the past hundred years. This study uses the article content and metadata of UK's most important machine learning research institute, Alan Turing Institute (ATI)—obtained using Scopus, for a 4-year period (2016–2019) to address important bibliometrics questions. In this study, we aim to track the co-evolution of trends in ATI publications. Our analyses of the machine learning literature include: (a) metadata analysis; (b) content-based analysis; and (c) citation analysis. In addition, we identify the significant trends and the most influential authors, institutes and countries, based on the publication count as well as article citations. Through this study, we are proposing a methodology and framework for performing a comprehensive bibliometric analysis on machine learning research.

Keywords Bibliometrics · Co-authorship Patterns · Computer Networking · Full-text · Social Network Analysis

1 Introduction

Bibliometric analysis of a literature is a crucially important source of objective knowledge and information about the quantity and quality of scientific work (Narin et al., 1994). In this work we perform a bibliometric analysis of the the literature of the field of machine learning which is a major research discipline and an intersection of different major research domains. This breadth-wise knowledge saves ample amount of time for researchers to get started with the

✉ Waleed Iqbal (E-mail: waleed.iqbal@itu.edu.pk)

[1] Information Technology University, Lahore, Pakistan
[2] Queen Mary University of London, United Kingdom
[3] Computer Laboratory, University of Cambridge, United Kingdom

research of a domain and helps inform about the major trends observed in machine learning publications.

Towards this end, we statistically analyze 4 years of accepted articles published in ATI publications, and examine the publication behaviors of several research entities and how these are affected by the elements of articles. We also analyze popular topics in periodicals on machine learning and the effects of several parameters on the citations of an article.

Our aim is to investigate changes in publication behavior and collaboration patterns of distinctive authors, institutes and countries in the various machine learning publications. Our goal is therefore to provide generalized insights into the publication trends in the field of machine learning. We also aim to answers questions such as the following: Which topics are popular in which regions of the world? What are the topics discussed by the top authors in their articles in the various publications? Which parameters affect the citations of an article?

The *key contribution of this article* is to develop a methodology and framework for performing a comprehensive bibliometric analysis on machine learning research and the public release of a comprehensive dataset.

The rest of this article is structured as follows. In section 2, we discuss related previous research work. The bulk of our investigations focus on the publication trends in machine learning publications in ATI (Sections 3–6). In Section 3, our dataset is described and our methodology is broadly outlined. A detailed bibliographic focused on analysis of ATI publications is presented in Sections 4, 5, 6 in which metadata analyses, content-based analyses, citation-based analyses are presented respectively. The paper is finally concluded in Section 7.

2 Related Work

In this section, we present related work and highlight the novelty of this article. Bibliometrics is an established field in which the major trends of research fields are studied rigorously. A number of bibliometrics studies have been conducted in various fields to gain useful insights through the analysis of authorship and publication trends of different research outlets and areas (Nobre and Tavares, 2017; Fernandes and Monteiro, 2017; Serenko et al., 2009; Chiu and Fu, 2010; Rajendran et al., 2011; Nattar, 2009; Yin and Zhi, 2017). These bibliometric analyses are not confined to the authorship based meta-data analysis of venues.

Some authors have also undertaken quantitative analysis on the top ACM conferences. The purpose of these studies is to determine the genre of the article and to understand the publication culture of these conferences (Flittner et al., 2018). These related studies do not explain which factors of the article affect the productivity parameters and the information about the correlation between important parameters required to analyze the productivity of different entities. Many previous works have performed an analysis on the content of various research areas using topic modeling (Paul and Girju, 2009) and keyword-based analysis (Choi et al., 2011).

A number of studies have used social networking analysis for social sciences and medical science research to find the most significant collaborating entities (Savić et al., 2017; Wagner et al., 2017; Didegah and Thelwall, 2018; Borgatti et al., 2009; Waheed et al., 2018), using social network analysis on generally social media data and altmetric data (Hassan et al., 2017). Social media analysis has not been used to determine the communities in computer networking research due to which we do not yet have complete insights into the collaborating patterns that exist in computer networking research.

Limited work has focused on using bibliometric or scientometric techniques to analyze the publication mores of the field of computer networks. Chiu et al. (Chiu and Fu, 2010) have performned an analysis of author productivity in computer networking venues in 2010. Our work is different in that we perform a detailed bibliometric analysis on the computer networking literature including an analysis of the effects of various features of article (such as the graphical and mathematical elements and the numbers of references) on the article's productivity metrics as defined in the field of bibliometrics.

Bibliometric analyses can also be utilized to see the extent of the incorporation of related research. Reference count in a article is the simplest way to observe the inclusion of related research and literature review. Different researchers analyzed referencing patterns in research articles to identify incorporation of the latest studies relating to a research article (Heilig and Voß, 2014) and citation analysis of the productivity of various research entities (Hamadicharef, 2012; Bartneck and Hu, 2009). These studies do not explain how the references are affected by the type of article venue.

We also have published another bibliometric study which solely focus on computer networking research. Detailed document of aforementioned study can be found in Iqbal et al. (2019).

3 Data Collection and Methodology

3.1 Dataset Collection

To perform the analysis of machine learning research, we used a collection of 350 articles, indexed over Scopus repository under ATI affiliation.

Data was obtained in CSV (Comma Separated Values) format from the aforementioned scientific repository. The CSV files contain bibliographic details such as authors' name, affiliation, citation count, publication year and references used in an article. Incomplete and irrelevant entries were removed from the dataset. These entries include messages from editors, entries without references, and entries without relevant metadata such as author names, institute names and indexed keywords.

Two further pre-processing tasks were performed on the extracted text: (a) calculation of number of metadata elements such as authors, institutes, countries; (b) Finding the number of references in an article cited and number of references in an article cited from the previous decade's published articles.

For references, we used an in-house formula script in Microsoft Excel and a python scripts as final step, which takes the list of all references for an article and outputs the total number of references, for the past decade.

3.2 Bibliometric Indicators

In this study, we used several bibliometric indicators in order to measure the impact of research published in ATI. Details of these bibliometric indicators are shown in Table 1. Here, we briefly list the methodologies we will use in the remainder of the paper.

Table 1: Bibliometric indicators used in this article

Dimension	Indicator	Definition
Metadata based Analysis	Publication count (P) per author	Number of articles published by an author
	Publication count (P) per institute	Number of articles published by an institute
	Publication count (P) per country	Number of articles published by a country
	h-index of an author	h-index of a researcher (h) shows us that h articles of a researcher have got h citations
	Reference count per article	Number of references used in an article
Content-based Analysis	Readability scores	Score indicates the difficulty level of language for intended audience
Citation based Analysis	Citation count per keyword	Total number of citation against a keyword
	Citation count per author	Total number of citation obtained by an author

- *Statistical Analysis*: There are a number of analyses that come under the umbrella of statistical analysis, but our focus, for the most part, will be on occurrence-based analysis (Weatherburn, 1949) in this study for finding significant entities either in terms of publications count or in terms of citation and h-index count.
- *Social Network Analysis*: Social network analysis is useful in finding connections and relations between various entities. These relations cannot be observed through statistical analysis. Social network analyses are useful in finding hidden communities within data, e.g., we used a modularity class-based clustering technique (Blondel et al., 2008) for finding various communities in our data. To find the significance of a single node, we used an average degree algorithm.

The rest of this paper will explore our datasets through the lens of the above analytical techniques. We performed analysis over journals' data explicitly in section 4, 5 and 6 respectively.

4 Metadata Analysis and Findings

We start our analysis by exploring the key metadata attributes associated with the publications. Specifically, we focus on metadata associated with publications authors and their respective institutes, before inspecting the structural elements of the articles (e.g., presence of figures). In this section we focus on analyzing these observations on publications of Alan Turing Institute.

4.1 Research Productivity of Authors and Countries

4.1.1 Author Based Productivity Analysis

First, we investigate the most important authors in Alan Turing Institute's publications. There are many parameters to analyze the significance of a researcher's published work. A simple measure would be publication count is listed in Figure 1. The h-index is also another widely used metric where h tells us that h articles of a researcher have h citations (Hirsch, 2005). Using the h-index of only paper published under ATI affiliation, we can observe which authors are publishing highly cited research in ATI.

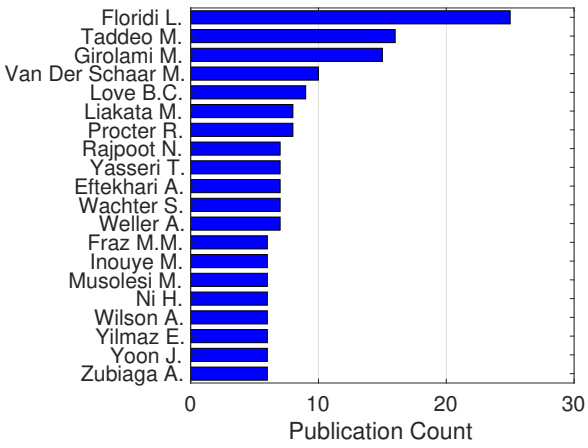

Fig. 1: Most-published authors during 2016–2019, according to article count.

Figure 2 shows the authors, publishing with ATI affiliation, with the highest h-index, and how the top five highest publication counts are from the top ten authors with the highest h-index. The data confirms that the top authors (measured by publication count) are the ones who have significant research contributions in terms of publication count as well as citation count.

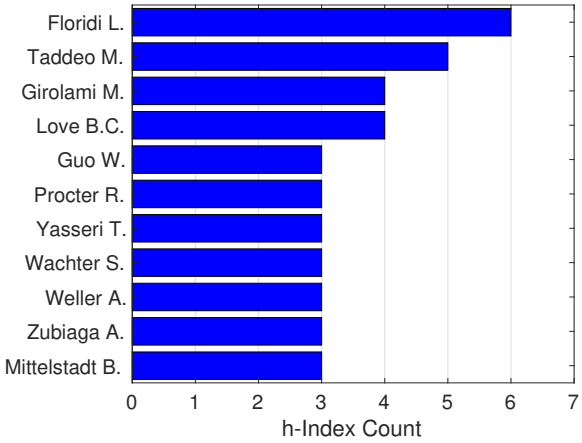

Fig. 2: Top authors with the highest h-index during 2016–2019. *Majority of The top most-published authors are the same as the authors with the highest h-index indicating a strong relationship between numbers of articles published and h-index.*

4.1.2 Country Based Productivity Analysis

In a research domain, some countries play a pivotal role in driving the ongoing advancements in that field. Figure 3 shows the distribution of published articles under affiliation of ATI from different countries using a global heat map. As expected, the United Kingdom is in the highest position in terms of publication count. Other top countries include USA, Italy, Germany, France, and Australia in that list.

Figure 4 shows the rank of different countries in COMST and TON based on published articles using a global heat map.

We next inspect the collaborations that took place between these countries. Figure 5 shows the co-authorship network of top countries under ATI affiliation. UK, USA, and Asian countries shown significant co-authorship pattern whereas European countries specially Baltic and Nordic regions shown significant co-authorship pattern.

4.2 Author Collaborations

4.2.1 General Co-Authorship Trends

Author collaborations is a key ingredient for research productivity (Iglič et al., 2017; Powell, 2018). We next explore the changing trends in co-authorship in

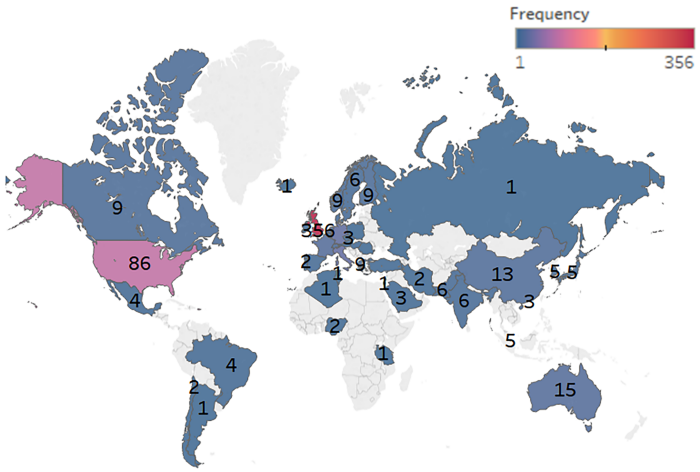

Fig. 3: Publication count of different countries under ATI affiliation.

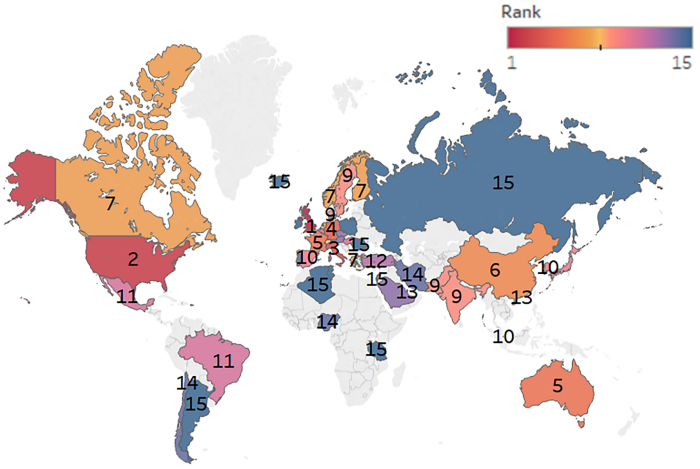

Fig. 4: Rank of different countries in under ATI affiliation based on publication count.

COMST and TON over the period 2016 to 2019. We explore how the distribution of collaborating authors changes over time; what kinds of authoring entities (foreign or local authors) have changed in collaborations over time; and whether influential authors tend to collaborate on publications. Note that we use the terms collaboration and co-authorship interchangeably, as it is impossible to identify the exact form of collaboration that took place during the preparation of an article.

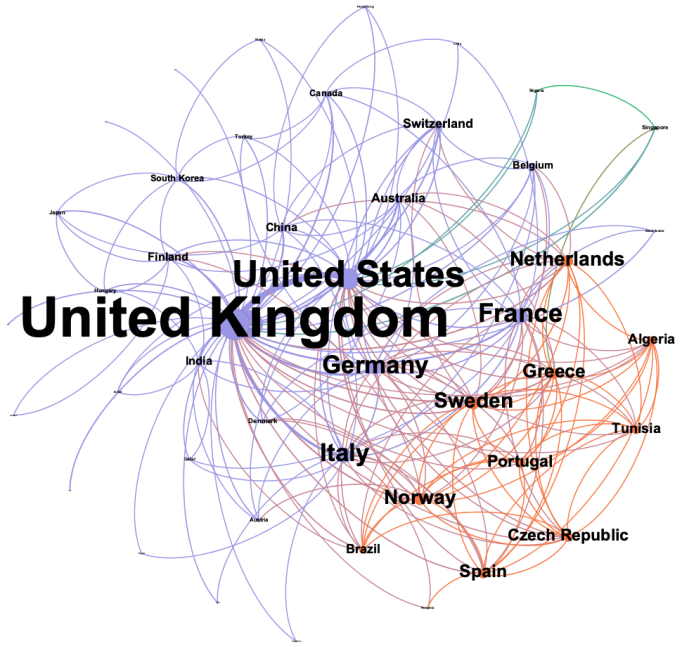

Fig. 5: Co-authorship network among top countries, based on publication count, under ATI affliation

Figure 6 shows the distribution of the number of authors per article in COMST per year. It is clear that the tendency for co-authorship is increasing; in 2016 the median number of authors is 3, compared to 4 in 2019. This increasing trends may be a result of several elements which include expanding the number of members in different graphical unit e.g. European Union, cross-country funding, and the arrival of increasing degrees of remote (skype/email) collaboration.

4.2.2 Institutional and Country Based Collaborations

This subsection presents the varying trends of collaborations among the institutes and countries in ATI affiliated publications over the period from 2016 to 2019. We will address several important questions relating to the collaboration patterns of institutes and countries; how the distribution of collaborating institutes and countries changes over time; the most influential institutes and nations in COMST and TON; and whether influential institutes and nations tend to work as collaborators. To observe collaborative relations among the top researchers in ATI affiliated publications, we generate undirected graphs of

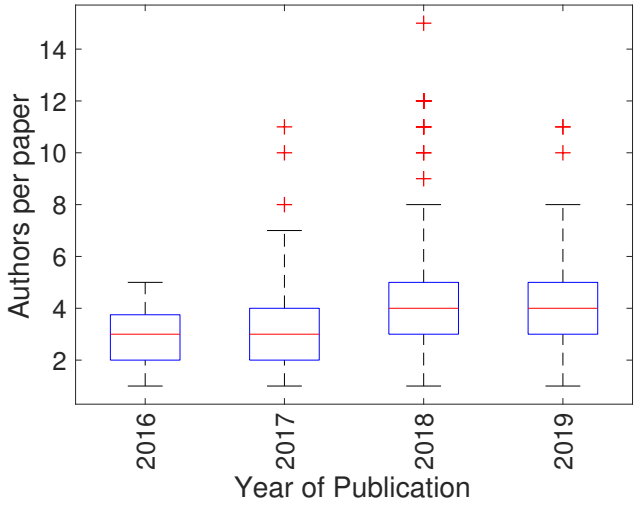

Fig. 6: Distribution of the number of authors per article in ATI affiliated publications throughout 2016–2019. *Tendency for co-authorship is increasing over time due to enhancing collaboration between institutes and authors.*

co-authors and identify clusters using modularity class partitioning. We used undirected graphs to remove duplicate links among publishing entities.

Figure 7 presents the clusters present in the network. We find 20 different clusters of authors in ATI publications.

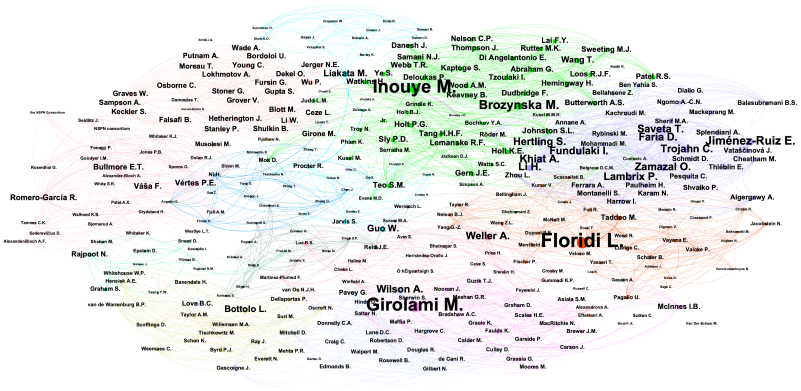

Fig. 7: Co-authorship network among top authors in ATI publications.

Published research is a crucial factor in determining the quality of education and research at any institute. Figure 8 shows a similar result for the

top institutes. We performed a clustering analysis using modularity class algorithm ATI publications. Figure 9 shows a similar result for ATI publications. The top publishing institutes are clustered into six groups according to their publishing behavior. In the ATI data, University of Oxford, University College London, Imperial College London, and University of Cambridge showed a significant co-authorship pattern.

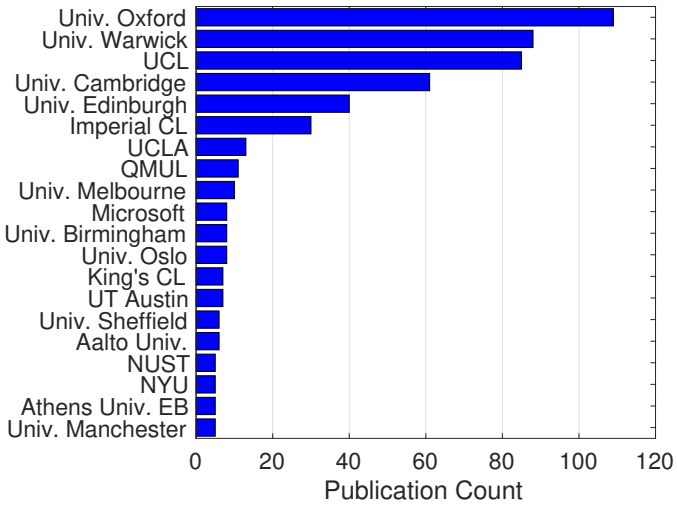

Fig. 8: Most-published institutes during 2000–2017, according to their article count.

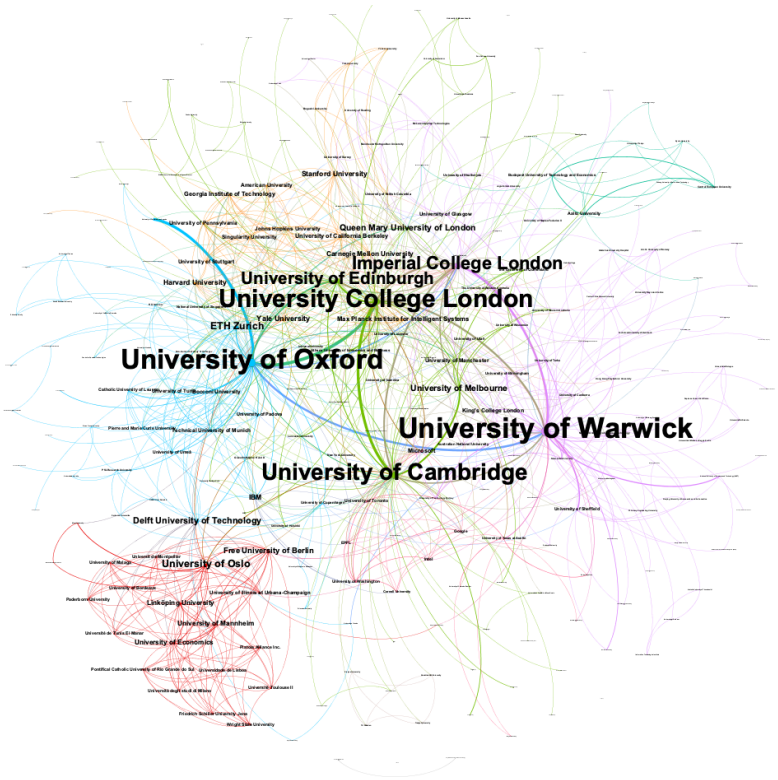

Fig. 9: Co-authorship network among top institutes in ATI publications. *In the ATI data, University of Oxford, University College London, Imperial College London, and University of Cambridge showed a significant co-authorship pattern.*

5 Content Based Analysis and Findings

This section contains two types of analysis of COMST and TON: (A) keyword-based analysis, based on index keywords; and (B) readability-based analysis. We address questions such as, what are the popular topics of computer networking research during each year? what topics are discussed by top authors in COMST and TON? and which types of articles are easiest to read?

5.1 Keyword-based analysis of articles

Investigating the popular topics is considered to be one of the best ways of studying the paradigm shifts in any research field. It is helpful in describing the research trends of a field. In this sub-section, we use publication dataset

of ATI to analyze the popular topics in ATI publications. We have described the top 10 popular topics discussed in ATI publications.

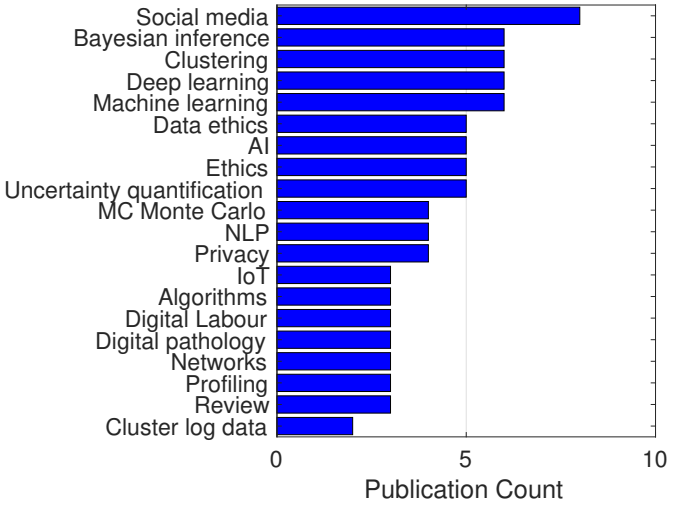

Fig. 10: Most popular topics in ATI publications and their article count during 2016–2019, in terms of article count (cf. Figure **??**, in which keywords of the most-cited articles are listed.)

Figure 10 represents the most popular topics in ATI publications.

5.2 Keyword co-occurrence analysis

Keyword co-occurrence analysis helps researchers to find a publication venue's most common topics. These analyses also help researchers to find topics and domains that are strongly related to each other. Figure 11 is the term co-occurrence map for ATI publications.

Terms in a larger font size have a higher co-occurrence than other keywords in the graphs. In ATI publications, frequently co-occurring terms are "Social Media", "Natural Language Processing", "Artificial intelligence", "Machine Learning", "Markov Chain Monte Carlo", "Uncertainty Quantification", and so on. Top keywords (measured on publication count) in ATI publications are clustered in the same groups and have stronger links with each other than with unpopular keywords. This trend shows that in ATI publications, there are only some top keywords (measured on publication count) which are discussed in most of the articles. The results also show that in most of the articles in ATI publications, top keywords co-occur with each other.

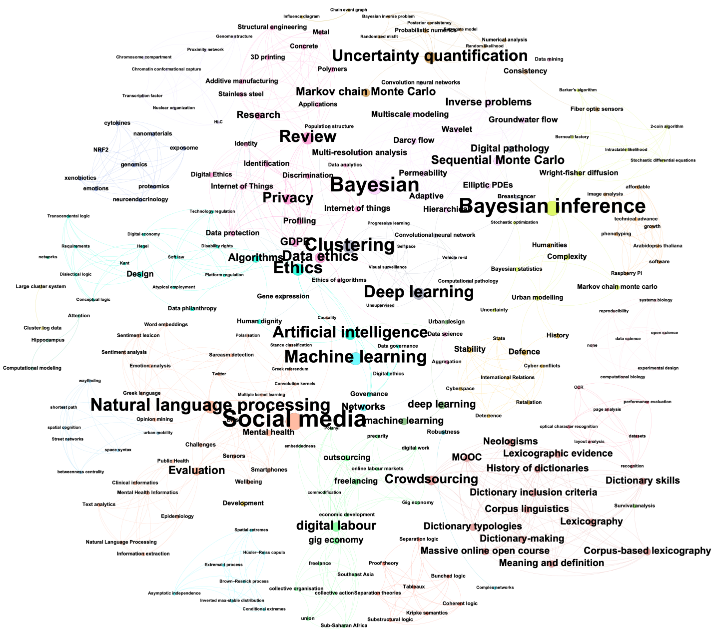

Fig. 11: Keyword co-occurrence network in which the node size indicates the number of links with other nodes and node color represents cluster membership.

6 Citation Based Analysis and Findings

Citations are used to investigate the contributions of an author, organization, country or publication venue. Citation analysis is an effective tool to rank the productivity of various research bodies. In this section, we address some important bibliometric questions using citation data from ATI articles, such as who are the most-cited authors in ATI publications; whether they have the same h-index as the most-published authors in ATI publications; whether increasing the number of authors affects the number of citations of an article; and the most-cited keywords in ATI publications.

6.1 Citation Based Analysis of Different Research Entities

In any field, some authors play more significant roles in advancements of the field than others. It is worth observing the impact and usability of their research. Figure 12 shows the most-cited authors in ATI publications from 2016–2019. From Figure 12 and Figure 1, it can be observed that the top most-published authors and the top most-cited authors in ATI publications are majorly different. Citations do not entirely represent the significance of the

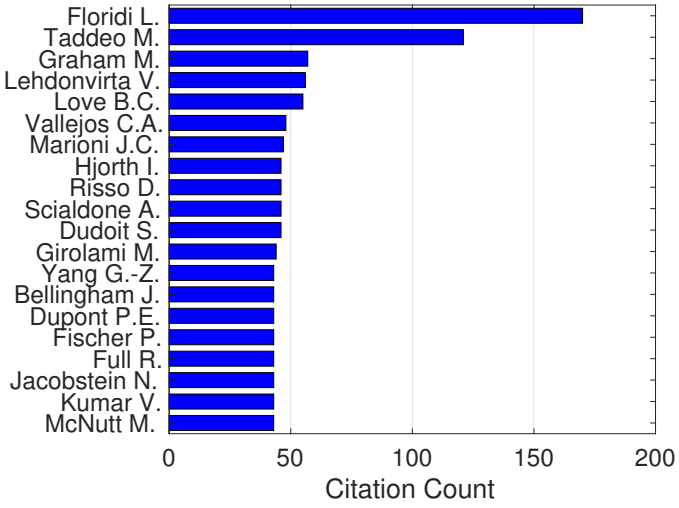

Fig. 12: Most-cited authors in ATI publications

research undertaken by a researcher. There are many parameters to analyze its significance, but the h-index is the most widely used, and it is a better measure of an author's significance in a field than a simple citation count.

Figure 2 shows the authors in ATI publications with the highest h-index, and how the top ten highest publication counts are from the top ten authors with the highest h-index in ATI publications. The data confirms that the top authors (measured by publication count) are the ones who have significant research contributions in terms of publication count as well as citation count.

7 Conclusions

In this paper, we have performed an in-depth bibliometric study of the publication trends in machine learning literature using article content and metadata of publications under ATI affiliation—gathered over the time period 2016–2019. Our work extends the state of the art in bibliometric analysis of machine learning literature by presented comprehensive analyses that shed light on the publication patterns in ATI publications, which different authors, institutes, and countries have been successful in these ATI publications (and how). Although we cannot make strong claims about causality or the parameters responsible for the acceptance/rejection of an article since we did not have access to missing data (rejected articles), we believe that our analyses provide an insightful look into the publication culture in the machine learning community and can help develop a more nuanced understanding of this research field especially in

the light of the limited existing bibliometric work that focused on the machine learning community.

References

Bartneck C, Hu J (2009) Scientometric analysis of the CHI proceedings. In: Proceedings of the SIGCHI conference on human factors in computing systems, ACM, pp 699–708

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008(10):P10,008

Borgatti SP, Mehra A, Brass DJ, Labianca G (2009) Network analysis in the social sciences. science 323(5916):892–895

Chiu DM, Fu TZ (2010) Publish or perish in the internet age: a study of publication statistics in computer networking research. ACM SIGCOMM Computer Communication Review 40(1):34–43

Choi J, Yi S, Lee KC (2011) Analysis of keyword networks in mis research and implications for predicting knowledge evolution. Information & Management 48(8):371–381

Didegah F, Thelwall M (2018) Co-saved, co-tweeted, and co-cited networks. Journal of the Association for Information Science and Technology

Fernandes JM, Monteiro MP (2017) Evolution in the number of authors of computer science publications. Scientometrics 110(2):529–539

Flittner M, Mahfoudi MN, Saucez D, Wählisch M, Iannone L, Bajpai V, Afanasyev A (2018) A survey on artifacts from CoNEXT, ICN, IMC, and SIGCOMM Conferences in 2017. ACM SIGCOMM Computer Communication Review 48(1):75–80

Hamadicharef B (2012) Scientometric study of the IEEE transactions on software engineering 1980-2010. In: Proceedings of the 2011 2nd International Congress on Computer Applications and Computational Science, Springer, pp 101–106

Hassan SU, Imran M, Gillani U, Aljohani NR, Bowman TD, Didegah F (2017) Measuring social media activity of scientific literature: an exhaustive comparison of scopus and novel altmetrics big data. Scientometrics 113(2):1037–1057

Heilig L, Voß S (2014) A scientometric analysis of cloud computing literature. IEEE Transactions on Cloud Computing 2(3):266–278

Hirsch JE (2005) An index to quantify an individual's scientific research output. Proceedings of the National academy of Sciences of the United States of America 102(46):16,569

Iglič H, Doreian P, Kronegger L, Ferligoj A (2017) With whom do researchers collaborate and why? Scientometrics 112(1):153–174

Iqbal W, Qadir J, Tyson G, Mian AN, Hassan Su, Crowcroft J (2019) A bibliometric analysis of publications in computer networking research. Scientometrics pp 1–35

Narin F, Olivastro D, Stevens KA (1994) Bibliometrics/theory, practice and problems. Evaluation review 18(1):65–76

Nattar S (2009) Indian journal of physics: A scientometric analysis. International Journal of Library and Information Science 1(4):043–61

Nobre GC, Tavares E (2017) Scientific literature analysis on big data and internet of things applications on circular economy: a bibliometric study. Scientometrics 111(1):463–492

Paul M, Girju R (2009) Topic modeling of research fields: An interdisciplinary perspective. In: Proceedings of the International Conference RANLP-2009, pp 337–342

Powell K (2018) These labs are remarkably diverse–here's why they're winning at science. Nature 558(7708):19

Rajendran P, Jeyshankar R, Elango B (2011) Scientometric analysis of contributions to journal of scientific and industrial research. International Journal of Digital Library Services 1(2):79–89

Savić M, Ivanović M, Surla BD (2017) Analysis of intra-institutional research collaboration: a case of a Serbian faculty of sciences. Scientometrics 110(1):195–216

Serenko A, Bontis N, Grant J (2009) A scientometric analysis of the proceedings of the McMaster world congress on the management of intellectual capital and innovation for the 1996-2008 period. Journal of Intellectual Capital 10(1):8–21

Wagner CS, Whetsell TA, Leydesdorff L (2017) Growth of international collaboration in science: revisiting six specialties. Scientometrics 110(3):1633–1652

Waheed H, Hassan SU, Aljohani NR, Wasif M (2018) A bibliometric perspective of learning analytics research landscape. Behaviour & Information Technology pp 1–17

Weatherburn CE (1949) A first course mathematical statistics, vol 158. CUP Archive

Yin Z, Zhi Q (2017) Dancing with the academic elite: a promotion or hindrance of research production? Scientometrics 110(1):17–41