RREF:Reference nesting in Scientific Literature and Scholastic Diversity Score: A graph mining approach *

Gouri Ginde†
Dept. Computer Science and Engineering
PESIT Bangalore South Campus, Bangalore
Karnataka, India
gouriginde@pes.edu

Snehanshu Saha‡
Dept. Computer Science and Engineering
PESIT Bangalore South Campus, Bangalore
Karnataka, India
snehanshusaha@pes.edu

Aditya Agarwal§
Dept. Computer Science and Engineering
PESIT Bangalore South Campus, Bangalore
Karnataka, India
adityag2511@gmail.com

Arijit Mukherjee¶
TCS Research and Innovation
Kolkata
West Bengal, India
mukherjee.arijit@tcs.com

Archana Mathur
Dept. Computer Science and Engineering
PESIT Bangalore South Campus, Bangalore
Karnataka, India
archanamathur@pes.edu

ABSTRACT

Citation network analysis of scholarly articles and journals has already been explored in depth and the subtlety of the differences between citations and references has also been recognized. The articles listed under the references section of an article contribute to the citation count of the referenced article. Analyzing citations of an article or a journal of that article, which is a bottom up approach, provides a varied degree of information, such as, patterns of spread and influence of them in the academic world. However, analysis of references provides a top down approach. The reference network is represented as a graph; the nodes of this graph represent articles and the directed connection between the nodes represent the referenced relationship forming a Nested Reference Network (NRN). The reference network analysis can help in exploring the history of any famous or influential article of a journal. Identifying various important articles in the reference network of such an article, using graph theory, helps pin point the path breaking articles which contributed in the subject/domain evolution. Text Analysis on keywords of multiple reference network of many highly cited articles of a scholar is used in generating readership profile for that scholar. Further, text analysis and natural language processing is used in introducing and computing Scholastic Diversity Score, a novel concept. An interface, which provides diversity, readership profile and history of information mined through graph theory and text analysis, should be a handy tool for young researchers looking for a range of background material in the early stages of his/her research career. This tool can be easily scaled up using Neo4j graph database for data storage and mining in future. Scholastic Diversity Score, a potentially rich discovery from data that may turn out to be inspirational and could feature prominently in the Scientometrics literature in future.

Keywords

Reference network, , latent Dirichlet allocation (LDA), Graph Theory, Scholastic Diversity Score, Machine Learning.

1. INTRODUCTION

State of the Art study forms the basis of any relevant research and a researcher in his/her academic tenure spends a considerable amount of time and effort in performing this step. Often, while studying an academic paper, a researcher finds it useful to look up the references cited by the author, in order to understand the context. This process is often recursive in the sense that the original paper may cite n different papers, some of which are useful in the context in which the researcher is working, and within each of those n papers, there may be many more relevant references which the researcher may have to look up. The process is tedious and time consuming, but the final reference tree or path, if created, can be of great help for the researcher and his/her collaborators, i.e. collectively for a group of researchers. A relevance score can be attached to the collective set of references which can be utilized in order to complete the study with a proper direction. Pioneering works in the area can be identified which may be of great help for the researcher. It may also be possible to identify community of researchers working on the topic

that the researcher is interested in. During the course of research, every researcher manually performs these set of steps manually as there exists no tool which can automatically scan through this recursive set of references to build the complete reference tree given a paper that the researcher is interested in.

1.1 Difference between citations and references

There is a evident difference between the citations and references[1]. References are a list of articles referred by the authors of a paper. This is a list of articles which appear in the references section of a paper. This is a list of the the sources one (authors) has/have cited. Generally the references are listed in APA style, this is not just a list of works consulted. In fact every source that is listed in references also needs to be cited in the body of the paper.
Every source listed in references should be accessible by others who read the paper. It is like a trail of breadcrumbs that one leaves for readers to show them where they can go to find the original source material for themselves.

Citation is a specific source that is mentioned in the body of the paper. The format of the citation may change depending on the style one uses (e.g. MLA and APA). The basic elements of the citation that needs to be include are:

- Name of the author(s)
- Year of publication
- Page number or page range

1.2 How references network is different from citation network?

There exists various kinds of networks [2] related to scholarly publishing information, which are diverse in nature and usage. Collaboration networks can identify the most productive and highly cited authors, institutions, and countries in the publication set. Semantic networks can identify the concepts central to all publications in the set as well as those central to the various research directions represented in the set. Publication citation networks can identify the quantity and citation impact of publications in each of these research directions and disciplines. In this study we focus on reference network, which is created from referred articles nested at various levels. Reference network is the most effective method of describing and evaluating a scientific publication. Performing analyses on a all the referenced set of publications for a particular research article of a scholar provides a great deal of information about the structure and direction of research being done on that topic. By extracting relationships among publications, bibliometric mapping offers a method of quickly summarizing and then visualizing the structure inherent to a set of publications bibliometric maps which shows the existing relationships among publications within a research topic.

For every path breaking work, the authors of that paper generally perform a lot of research on the prior work with utmost importance. It gets highlighted in the reference list of the paper. The reference list can get diverse to the highest degree or can get very narrowed and streamlined. This can vary from one research domains to other. There can be multiple reasons why one cites the articles. As described in Eugene [3] Garfields reasons for citing a paper, the citations in academic papers create a record of influence. Motivations for citing a prior work can include:

- Paying homage to pioneers.
- Giving credit for related work (homage to peers).
- Identifying methodology, equipment, and the like.
- Providing background reading.
- Correcting ones own work
- Correcting the work of others.
- Criticizing previous work.
- Substantiating claims.
- Alerting researchers to forthcoming work.
- Providing leads to poorly disseminated, poorly indexed, or un-cited work.
- Authenticating data and classes of fact (such as physical constants).e
- Identifying original publications in which an idea or concept was discussed.
- Identifying the original publications describing an eponymic concept or terms.
- Arguing against the work or ideas of others.
- Disputing the claims of others to have been first with their work.

Please note that the authors have deliberately excluded literature from citation network study as the objective of this manuscript is not aligned with previous work. In this research, we will endeavor to find out various patterns and other evident information which may work in favor of the authors of the paper and not the cited articles. This is essentially a top-down approach with focus on the root node of the tree.

2. THEORETICAL ANALYSIS OF REFERENCE NETWORK

Figure 1 represent a typical references graph network. In this graph the nodes at different depth represent articles published in different time lines. Node A is the article node which has B,F,C,D,Q,X and R articles in it s references list. Out of all the articles in the graph, N, O and T are the oldest referenced articles and A is the latest article. Article A is the successor of all the articles in the graph. Article N, O and T are the articles, which are predecessor articles, comprise of the most initial work in that research

domain. Figure 2 is another toy set of reference network which highlights how a path from the root vertex, a paper, to the article which is at the last level of reference network, the most preliminary work, which influenced the scholars for their research. Article H, which is at the 3rd level of the reference network is the most crucial article as it connects level 4 to level 2 and above, way back to root paper. However, this can be practically proved only when we confirm that the directed graph of reference network is always acyclic. Various graph theory algorithms such as, computation of strongly connected components, between-ness centrality, longest path in directed acyclic graph, vertex count etc can provide a lot of information on the structure of the network and valuable insights to information in it. On the flip side, if we also have textual data of the articles then, application of natural language processing technology can yield insights on diversity in research arena of a research scholar. This diversity, when quantified, can provide a diversity score, which can be used to explain the huge citation count that a paper incurs. We intend to do this study on the real data, explained in the following sections.

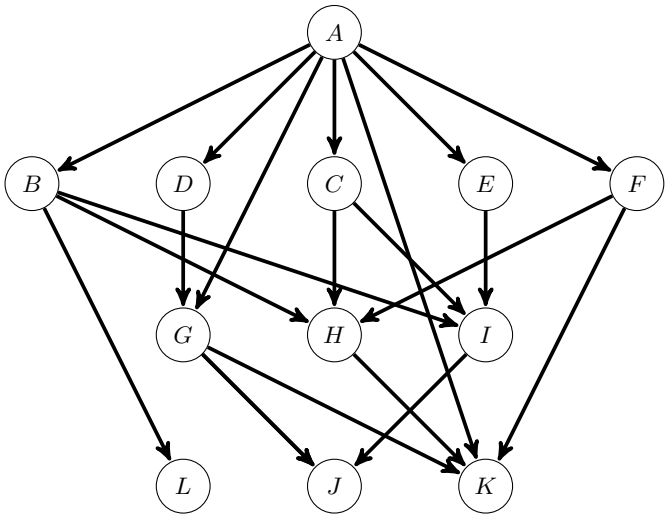

Figure 1: Reference network

3. IEEE JOURNAL'S ARTICLE REFERENCE NETWORK ANALYSIS

First step towards reference network analysis of the real data is to acquire it from authentic sources. We took a route of non subscription based data accumulation methodology, which was primarily through web scraping.

3.1 System architecture

Figure 3 shows the overall system architecture employed for this study. The article is first identified for analysis and then the required data is scraped from IEEE website [5] using web-scraping python script. This python script utilizes various HTML parsing tools such as Beautiful Soup to extract the references and other details from the webpages of respective articles into JSON format files. The raw data gathered for constructing the reference network

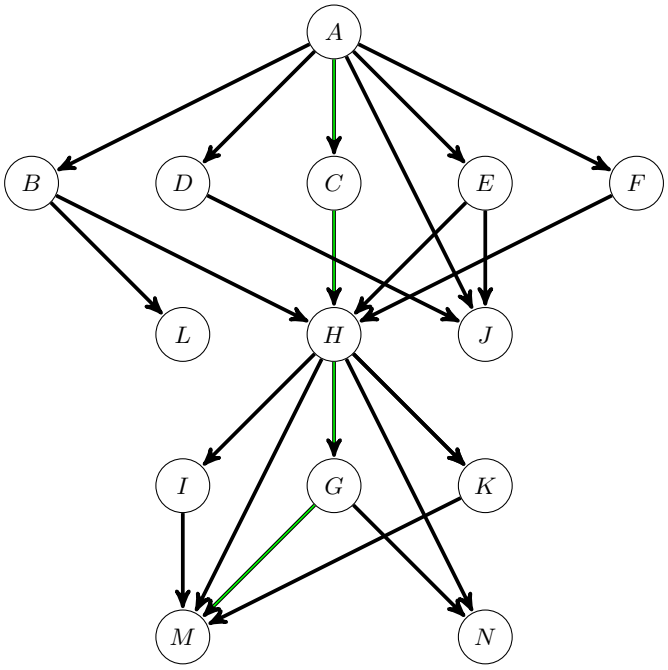

Figure 2: Directed Acyclic Graph of reference network

contains the articles that belong to IEEE journals only. We extracted references information for the depth of 2 to 4 based on requirement. Further on, we used two mutually exclusive approaches to analyze this data. Mainly using graph theory based algorithms and Text analysis based algorithms such as keyword clustering and latent Dirichlet allocation (LDA) for topic modeling.

3.2 Data acquisition and pre processing

Data accumulation, a first step in data curing, is an arduous task for any study. We chose to use the non subscription based approach to gather the information. Hence web scraping methodology has been used to gather the data [1, 4]. First the webpage is analyzed and the HTML tags corresponding to the data of interest is then specified in the python script to extract information in JSON format. Next, accumulated data is pre-processed, which is an intermediate task, where the data is cured and made ready for further analysis. The scraped data is first trimmed of any unwanted characters. Then, data is cleansed of Unicode characters.

In order to provide interesting visualization patterns, few of the parameters had to be derived from accumulated data. Pre-processing also involves computation of these scholastic indicators/parameters. The scraped data is in the pure textual form, hence, cosine similarity string metric has been used for text comparison in place of pure string comparison operation for better results. Figure 4 is the processed sample data in JSON format.

3.3 Article reference network analysis

The references graph network analysis of the highly cited articles of a research scholar can provide very interesting

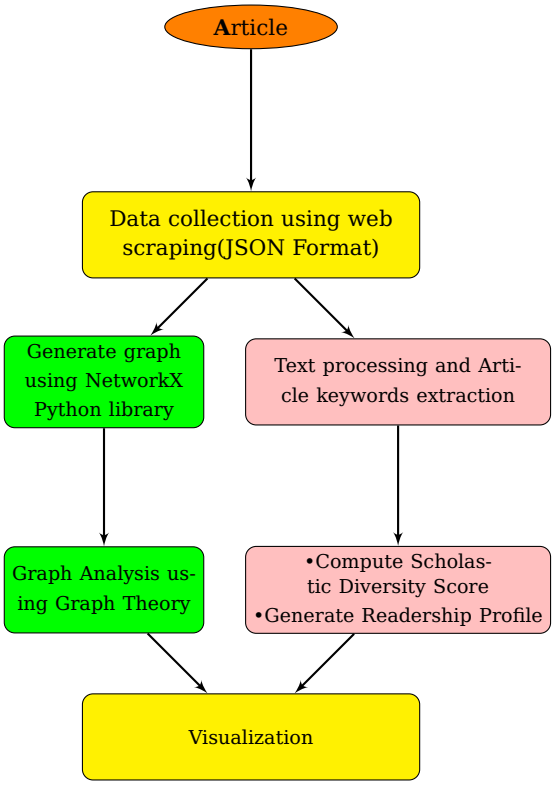

Figure 3: Overall system architecture

insights to various aspects of his/her research work. A directed graph network can be generated using the nested references. In this graph root node is the article under study and the children at first level are the articles listed in the references section of the root article. The second level of nodes are the articles from the references section for each one of the nodes in the first level. This nested network with higher depth shows exponential growth in size of the graph.
We have used a random article titled "Eye Tracking and Head Movement Detection: A State-of-Art Survey" with IEEE article id:6656866 and DOI:10.1109/JTEHM.2013.2289879 from IEEE Journal of Translational Engineering in Health and Medicine for analysis. Figure 5 shows network of this article for 2 level references nesting. The blue color nodes represent the directly referred articles at level 1 and yellow nodes represent the articles at level 2. The interconnections between the nodes represent the directed referred relationship between these nodes.

Graph theory based algorithms can be now easily used on this network which can yield interesting results.

- **In degree count histogram:** Figure 6 shows the number of in-degrees vs number of articles in a histogram plot. As shows there are a very few articles in the network which have in-degree greater that 3. Just about 8 articles out of 250 articles. All these 8 articles can be termed as the most referred articles of this reference network.


```
{   "Article":{  We have applied various
    processing methodologies for author names,
    references and the keywords extraction for
    this research.
s
      "references":[
         "5290134",
         "6189752",
      ],
      "details":{
         "title":"Evaluating Innovative In-Ear
             Pulse Oximetry for Unobtrusive
             Cardiovascular and Pulmonary
             Monitoring During Sleep",
         "journal_title":"IEEE Journal of
             Translational Engineering in Health
             and Medicine",
         "date_current_version":"Thu Sep 05 00:00
             :00 EDT 2013",
         "issn":"2168-2372",
         "abstract":"Homecare is healthcare based
              on the principle .................
             with sufficient accuracy.",
         "date_publication":"Thu Aug 08 00:00:00
             EDT 2013",
         "doi":"10.1109/JTEHM.2013.2277870",
         "issue_date":"2013",
         "publisher":"IEEE",
      },
      "keywords":[
         "assisted living",
         "cardiovascular system",
         "ear",
         ".........
         "heart rate dynamics",
         "homecare",
         "in-ear sensor",
         ],
      "authors":[
         "Boudewijn Venema",
         "Johannes Schiefer",
        ],
      "citations":[
         "6827738",
         "7299367",
         "7193056",
         "7279735"
      ],
      "arnumber":"6576858"


   }}
```


Figure 4: Sample of scraped data in JSON format

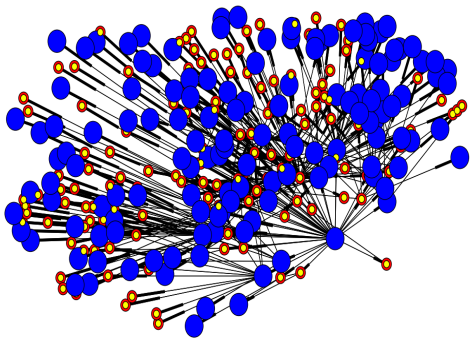

Figure 5: IEEE reference network of level 2, article id: 6656866, DOI:10.1109/JTEHM.2013.2289879.

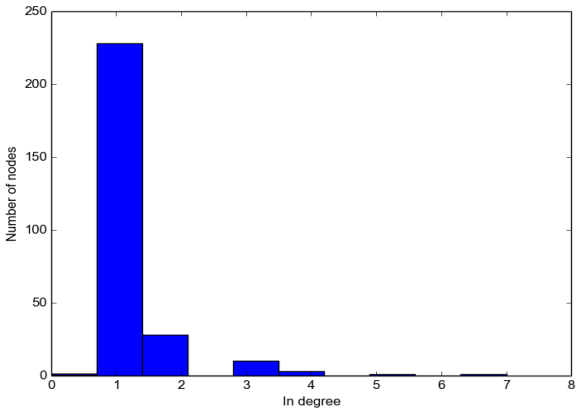

Figure 6: Vertex count histogram.

- **Identify Most important Articles in the network:** The most important nodes of the network for one such article under study provides us a list of articles who have the maximum impact on the research quality of this particular paper with article id:6656866. Centrality measures can be used to find out such nodes in the network. We have used Betweenness centrality measure. The betweenness focuses on the number of visits through the shortest paths. If a walker moves from one node to another node via the shortest path, then the nodes with a large number of visits have a higher centrality. Figure 7 shows the all the most important articles of the network computed using betweenness centrality. Table 1 shows the details of the top two of these articles.

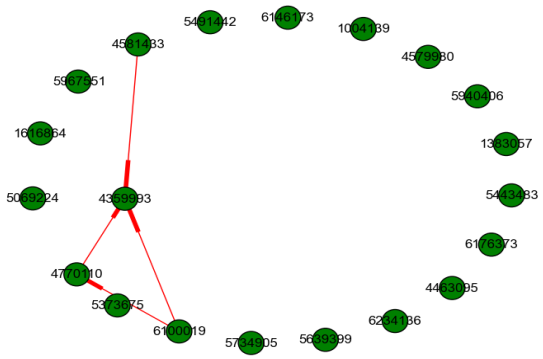

Figure 7: IEEE most important

Analysis: On analyzing the article Id's from this exercise we could find following two articles which stood out of all the articles in the reference network which appear to be very influential.

- **Topological sort** A topological sort is a non unique permutation of the nodes such that an edge from u to v implies that u appears before v in the topological sort order. [12]. Using topological sort and the corresponding details of year of publication, we can easily find the chronological order of growth in the subject area. Figure 8 shows the topological sort starting root node with article id:6656866, DOI:10.1109/JTEHM.2013.2

- **Longest path:** In order to find the longest directed path in the network first topological ordering is found in a Directed Acyclic Graph (DAG) [10]. For each vertex v of the DAG, in the topological ordering, the length of the longest path ending at v is computed by looking at its incoming neighbors and adding one to the maximum length recorded for those neighbors. If v has no incoming neighbors, the length of the longest path ending at v is set to zero. In either case, this number is recorded so that later steps of the algorithm can access it.

Table 1: Most influential articles of the network

Title	Id	Year	Citations	In degree
Human-computer interaction using eye-gaze input (IEEE Transactions on Systems, Man, and Cybernetics)	44068	2002	424	5
Novel Eye Gaze Tracking Techniques Under Natural Head (IEEE Transactions on Biomedical Engineering) Movement	435993	2007	295	7

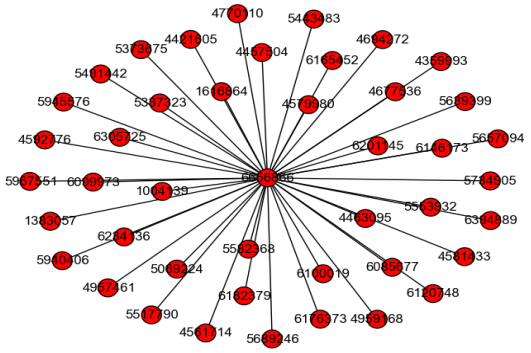

Figure 8: IEEE topology sort for an article, DOI:10.1109/JTEHM.2013.2289879

For the reference network in Figure 5, longest path: is $6656866- > 657094- > 5553932- > 4359993- > 1634506$ as shown in the Figure 5.In this picture the nodes represent the IEEE article id's.

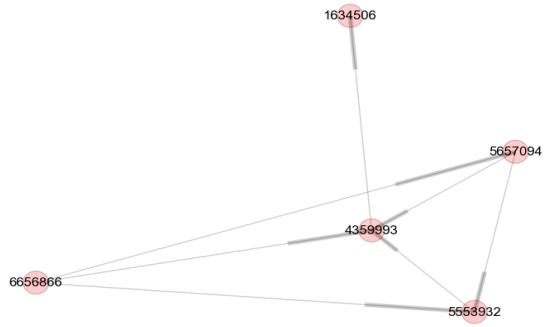

Figure 9: Longest path for reference network in Figure 5

4. CASE STUDY: TERRENCE TAO REFERENCE NETWORK ANALYSIS

Terence Tao is an Australian-American mathematician who has worked in various areas of mathematics. He currently focuses on harmonic analysis, partial differential equations, algebraic combinatorics, arithmetic combinatorics, geometric combinatorics, compressed sensing and analytic number theory. As of 2015, he holds the James and Carol Collins chair in mathematics at the University of California, Los Angeles. Tao was a co-recipient of the 2006 Fields Medal and the 2014 Breakthrough Prize in Mathematics. We have scraped data for the four of his top cited articles published in IEEE journals only, using methodology explained in section 4.1. Next we will explain the application of graph theory algorithms and text analysis algorithms on the reference network of these articles and the findings.

4.1 Graph theory based analysis

Figure 10 shows the complete reference network of three levels for the first article with DOI:10.1109 /TIT.2005.862083 from Table 2. The red largest size node is the root node

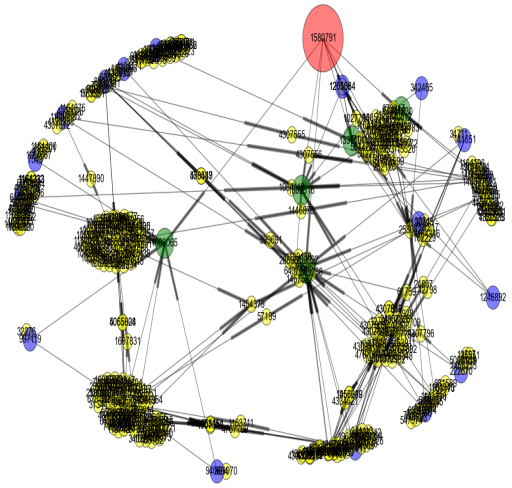

Figure 10: Terrence Tao reference network article id: 1580791, DOI:10.1109 /TIT.2005.862083

which corresponding to a first article (DOI: 10.1109 /TIT.2005.86208 in Table 2. The light green color nodes represent the first level reference nodes with respective article id's. Blue color nodes represent second level article nodes and yellow ones are the 3rd level nodes.

This network is directed cyclic graph, hence we could not easily find out the longest route in the network starting

Table 2: Terence Tao Articles dataset summary

Title	ID(DOI)	Year	Citations	Size of network	Keywords count
Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information	10.1109 /TIT.2005.862083	2006	10692	5016	82211
Near-optimal signal recovery from random projections: Universal encoding strategies?	10.1109 /TIT.2006.885507	2006	4963	2688	54188
Decoding by linear programming	10.1109/ TIT.2005.858979	2005	4603	3521	67682
The power of convex relaxation: Near-optimal matrix completion	10.1109 /TIT.2010.2044061	2010	992	15	31795

root node. However, the figure 11 shows the strongly connected component of this complete graph where each node displays year of publication as its property In the mathematical theory of directed graphs, a graph is said to be strongly connected if every vertex is reachable from every other vertex. The strongly connected components of an arbitrary directed graph form a partition into subgraphs that are themselves strongly connected. [11]

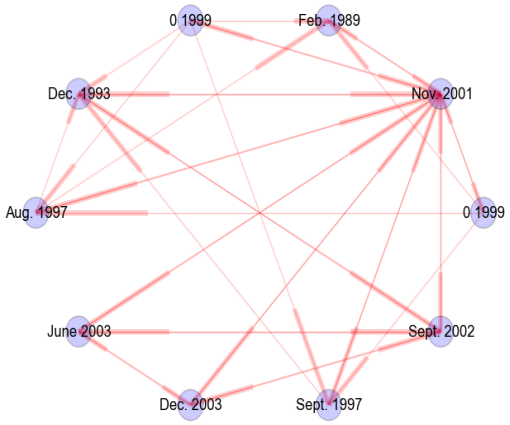

Figure 11: Strongly connected biggest component for reference network in Figure 10

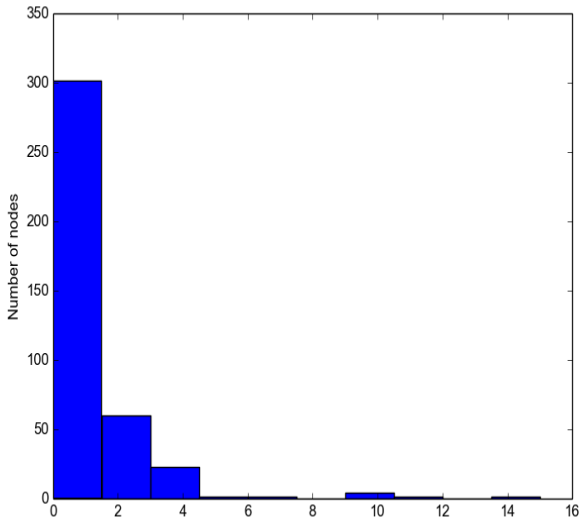

Figure 12: Number of articles Vs In degree count

Figure 12 shows the histogram of article counts Vs the In degree count of each one of the vertices in the reference network for article id:1580791 (DOI: 10.1109 /TIT.2005.862083). There are 7 articles in the network who have in degree greater than 7. These are the most referenced articles of this network. One article is with in degree 15.
Article id:495957
Title: A fast and accurate Fourier algorithm for iterative parallel-beam tomography

Year of publication: 2002.
Journal: IEEE Transactions on Image Processing Journal.

4.2 Text analysis on keywords

4.2.1 Latent Dirichlet Allocation and Readership profile

We have a huge corpus of keywords for each one of the referenced articles from Table 2.

We have used latent Dirichlet allocation (LDA), a generative statistical model in natural language processing, that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. LDA is an example of a topic model and was first presented as a graphical model for topic discovery by David Blei et al [4] **latent Dirichlet allocation (LDA)** The goal is to find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks such as classification, novelty detection, summarization, and similarity and relevance judgments.. We have used gensim [7] which is the most robust, efficient and hassle-free piece of software to realize unsupervised semantic modeling from plain text.

We have a set of documents which need to be scanned and key topics need to be modeled. A supervised topic classification is not welcome since we don't prefer fixing topics a priori, rather discover topics as we go. Essentially a clustering problem of keywords and associated topics where documents could exhibit multiple topics. LDA is a probabilistic model where each document is generated by a generative process. The topic is a distribution over a fixed vocabulary.

A distribution over topics is randomly chosen and then, for each word in the document a topic from the distribution over topics is randomly chosen. We then choose randomly a word from the corresponding topic. It is to be noted that words are generated independently of other words. Once a joint distribution of hidden and observed variables is formulated to identify the plates which indicate repetition of topics where the parameters of the Dirichlet distribution are used to compute distribution over vocabulary for topic and topic proportion for a particular topic in a document. Posterior estimates are used to discover most frequent topics. Figure 13 shows output this process. As shown the top 5 lists of keywords which can describe the topics broadly are discovered. These topics define the readership profile of the scholar, Terrence Tao. Topics are limited to tokens contained within the text corpus. Using algorithm 1 topics listed in Figure 13 are discovered. In the initial steps, all the JSON files(data set) are processed to create a list of list of all the keywords that belong to each one of the articles at first level of reference nesting only. Next a dictionary is created from the tokens in the entire text corpus. Then, a word frequency for each document is created in this step. Each document in the text corpus will be transformed into list of tuples $[[(token_id, doc_freq), (token_id, doc_freq), (token_id, doc_freq)]...]$ Each list of keywords is iterated to create this set. Conversion from a dictionary to a bag of words corpus is down for reference. Finally the LDA model is input with this corpus and the related parameters. This returns a list of words containing words describing various topics as shown in figure 13.

Algorithm 1 Topic discovery using Latent Dirichlet Allocation (LDA) library from GenSim

Input: Path to a directory containing JSON files of articles
Output: List of topics
procedure Discover_topics($path\_to\_files$)
  for $i \in path\_to\_files$ **do**
    for $j \in i$ **do**
      $kwdList.append(j['keywords'])$
    end for
    $kwdCorpus.extend(kwdList)$
  end for
  ▷ use corpora function from gensim library
  $dictionary \leftarrow corpora.Dictionary(kwdCorpus)$ ▷ convert the dictionary to a bag of words corpus
  for $text \in kwdCorpus$ **do**
    $corpus.append(dictionary.doc2bow(text))$
  end for
  $lda \leftarrow LdaModel(corpus, num\_topics = 5, id2word = dictionary)$
return $lda.showtopics()$
end procedure

4.2.2 Agglomerative clustering of keywords

We have used hierarchical clustering analysis on the huge pool of keywords extracted from all the articles that belong to reference network of IEEE article id : 1580791. Agglomerative hierarchical clustering, which is a "bottom up" approach, where: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. [9] Cosine similarity distance metric has been employed in generating this clustering. We have used various python packages from Scikitlearn [6] to generate the clusters. The input corpus of keywords is first transformed into list of tuples $[[(token_id, doc_freq), (token_id, doc_freq), (token_id, doc_freq)], ..]$. This is done by iterating through the text corpus. Next step is to convert a this text corpus to a matrix of TF-IDF features. Finally linkage_matrix is defined using ward clustering on pre-computed cosine similarity distances before plotting the dendogram. Figure 15 shows the output from algorithm 2

4.3 Keyword frequency histogram of all the articles

For Table 2 it is evident that the keyword corpus that has acquired is big. As a study we tried to find out a frequency of these keywords. For this keywords for 1st article from Table 2 are extracted in one list and all the keywords for

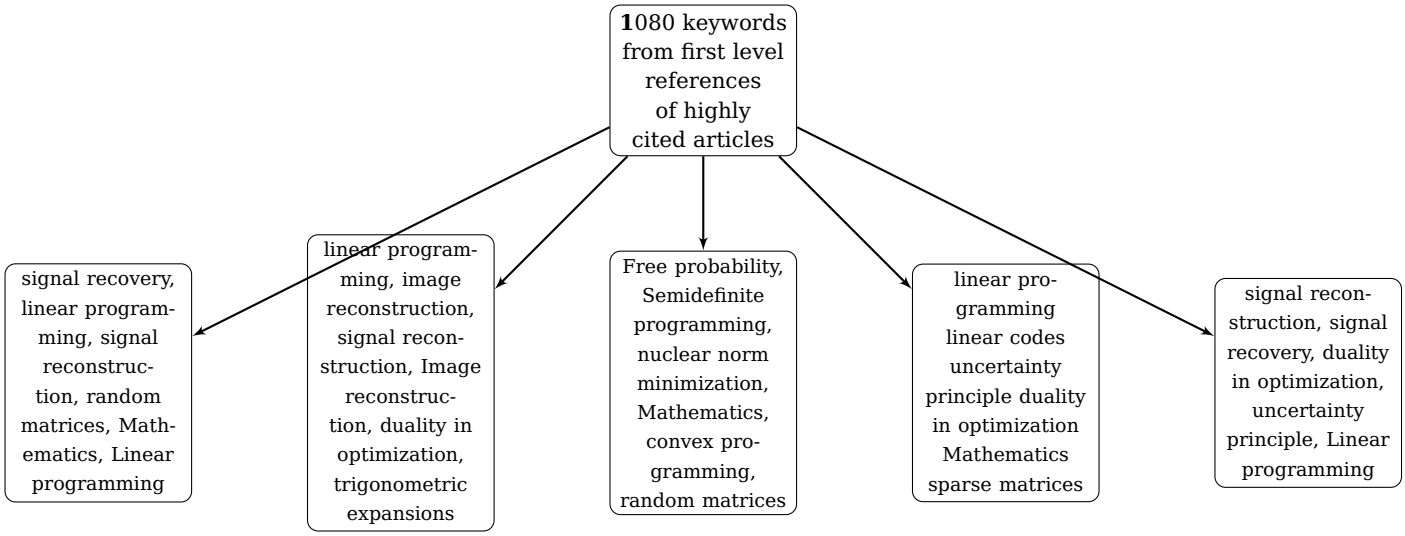

Figure 13: LDA output

Algorithm 2 Agglomerative clustering of keywords

Input: List of all the keywords from the article dataset: $KwdList$
Output: clusters of similar words.
procedure Keyword_Clustering($KwdList$)
$vectorizer \leftarrow TfidfVectorizer()$
$X \leftarrow vectorizer.fit\_transform(KwdList)$
$C \leftarrow 1 - cosine\_similarity(X.T)$ ▷ define the linkage_matrix using ward clustering
$linkage\_matrix \leftarrow ward(C)$
$ax = dendrogram(linkage\_matrix)$ **return** ax
end procedure

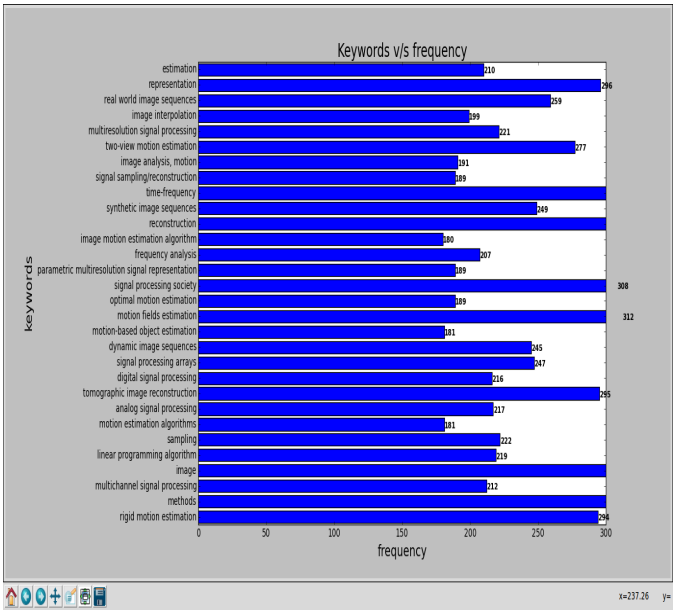

Figure 14: Keywords Vs Frequency bar chart

2 levels of reference nesting were extracted into another list. Using algorithm 3, Keywords from both the lists are then compared using cosine similarity and a final score is obtained. Top 30 keywords with maximum frequency are plotted in the Figure 14. There were total 31823 keywords from 1531 articles.

Algorithm 3 Plotting Histogram for most frequent and similar keywords from Terence Tao dataset

Input: root keyword list root_list,referenced keyword list reference_list, freq_dict for keywords in referenced_list.
Output: Keyword v/s frequency graph for top 30 most frequent and similar keywords in referenced articles
procedure create_freq_dictionary($root\_list, ref\_list, freq\_dict$)
  $count \leftarrow 0$
  for $keyword\_1 \in root\_list$ **do**
    if $keyword\_1 notin freq\_dict$ **then**
      $freq\_dict[keyword\_1] \leftarrow 1$
    else
      for $key, value\ in freq\_dict$ **do**
        $score \leftarrow Cosine\_Similarity$(key,keyword_2)
        if $score > 0.7$ **then**
          $count \leftarrow freq\_dict[key]$
          $count \leftarrow count + 1$
          $freq\_dict[key] \leftarrow count$
        else
          $freq\_dict[keyword\_2] \leftarrow 1$
        end if
      end for
    end if
  end for
  return $freq\_dict$
end procedure
procedure sorted_frequency_dict($freq\_dict$)
  $sorted\_list[] \leftarrow sorted(freq\_dict[])$
  for $keywords\ in\ sorted\_list[:30:]$ **do**:
    $sorted\_new\_list[] \leftarrow keywords$
    $sorted\_frequency\_dict \leftarrow dict(sorted\_new\_list[])$
  end for
  return $sorted\_frequency\_dict$
end procedure

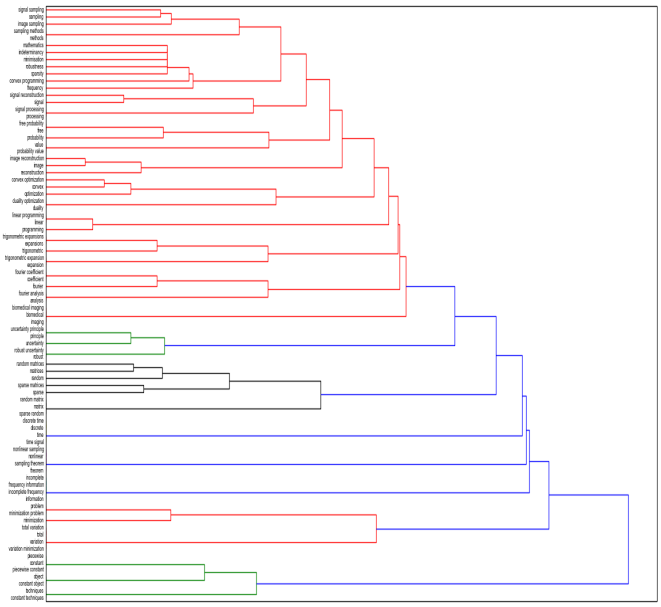

Figure 15: Agglomerative clustering

5. DIVERSITY IN BACKGROUND PREPARATION: DISCOVERY FROM DATA

Measuring quality of a published article is a difficult task and various measures proposed towards that invited criticism. We introduce a new measure, *Diversity in Background Preparation* without tall claims. We intend to gauge the breadth of Terrence Tao's diversity in readership and background preparation. The authors believe that good number of citations(contextual) received by an article is a testimony of authors' preparation level and diversity in background reading leading up to the final manuscript. The reason for choosing Terrence Tao doesn't beg detailed explanation. He is not only a highly cited author and a Fields Medal winner, his breadth and depth of scholarship is widely regarded. It is inspirational for a young researcher, we believe, to be aware of these traits, to have a list of diversity tokens handy for an article and be able to relate the impact of the article with the true scholarship. We integrated these quality parameters with a novel metric, Scholastic Diversity Score in the next subsection. The score is a derivative of his reference network (up to 2 levels of nesting). We hope that there is enough clarity regarding the relationship between *RREF* and *scholastic diversity*.
We begin with the corpus of keywords extracted for the all the four article from Table 2. Algorithm 4 was used and we extracted keywords from 4 articles of Terence Tao along with the keywords from its referenced articles up to 2 levels of nesting. The frequency of each key word in the referenced articles is first stored in a dictionary. Then, keywords from both the lists are compared using cosine similarity and a final score is computed. Figure 16 shows the lists the diversity range of Terrence Tao.

Algorithm 4 Diversity and range of Terence Tao **Most frequent and similar keywords in referenced articles**

Input: root keyword list root_list,referenced keyword list ref_list, freq_dict for keywords in referenced_list.
Output: Top 20 most frequent and similar keywords in referenced articles
procedure calc_smlrity_scr($root\_list, ref\_list, freq\_dict$)
  $score\_matrix[][] \leftarrow -1$
  for $kwd\_1 \in root\_list$ **do**
    for $kwd\_2 \in ref\_list$ **do**
      $score \leftarrow Cosine\_Similarity$(kwd_1,kwd_2)
      if $score >= 0.6$ **then**
        $score\_matrix[i][j] \leftarrow \alpha * score + (1 - \alpha) * freq\_dict[j]$
        ▷ where i and j are index values of kwd_1 and kwd_2 respectively. α is the convex relation weight.
        $keyword\_list \leftarrow (kwd\_2, score\_matrix[i][j])$
      end if
    end for
  end for
  return $sorted\_keyword\_list$
end procedure

Cosine Similarity Metric: Similarity metrics are the class of textual based metrics resulting in a similarity or dis-

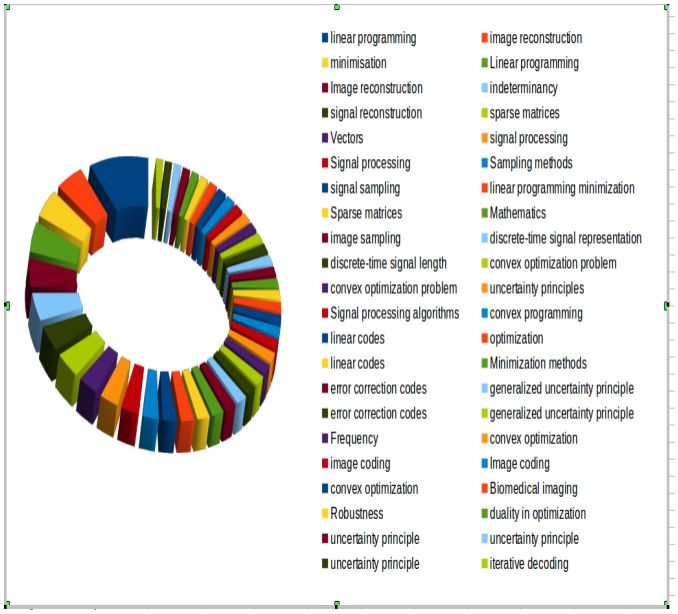

Figure 16: Diversity range of Terrence Tao

similarity (distance) score between two text strings. A string metric provides a floating point number indicating the level of similarity based on plain lexicographic match. For example, similarity between the strings orange and range can be considered to be much more than the string apple and orange by using Similarity metrics. Cosine similarity is a vector based similarity measure. Cosine of two vectors a, b can be derived by using the Euclidean dot product formula.

$$a.b = |a||b|cos\theta$$

Where, θ represents the angle between a and b.

$$similarity = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

where A i A and B i B are components of vector A A and B B respectively.

5.1 Scholastic Diversity Score

DEFINITION: Moody's Investors Service created a measure, Diversity Score, to estimate the diversification in a portfolio, related to a collateralized debt obligation (CDO). The calculation methodology for a diversification score takes into account the extent to which a portfolio is diversified by industry. "Technically speaking, the diversification score measures the number of uncorrelated assets that would have the same loss distribution as the actual portfolio of correlated assets". [8] The definition of *Scholastic Diversity Score* is borrowed from Moody's at a conceptual level.

Once the results of topic modeling*(via LDA)* are obtained, semantic similarity between the article keywords and the keywords of articles referenced at different levels is computed. Strong semantic similarity is conceptualized as good portfolio and weak similarity is equated to bad portfolio. *" Bad news is good news"* for us! We thus define Good Portfolio and Bad Portfolio. Lower similarity score implies higher diversity score. This is called a "Good Portfolio". Higher similarity score implies low Diversity Score. This is called a "Bad Portfolio". Topics which are detected to be weakly semantically similar indicate diversity of readership as the authors prepare the manuscript and is also a measure of coupling between apparently dissimilar topics. Thus, Scholastic Diversity Score is computed from *RREF, the nested reference paths* associated with an article–*the central theme of our work*. It is defined as an index which can measure degree of diversity in subject areas to which the referenced articles for a particular article belong to. The calculation methodology for this score considers the extent of diversity in subject areas for a scholar. Alternately, the diversity score measures the number of uncorrelated subject area that would have the same loss distribution as the actual portfolio of correlated subject areas. The scholastic diversity score is completely based on the reference network. Algorithm 5, similarity score is computed at every level with decreasing effect at every level. So, at every level if the semantic match between keywords of the root article and the referring is successful then it has very less effect on the diversity score. Conversely, unsuccessful match is works in favor of diversity score. Since the keywords are mutually exclusive in nature, it implies that the scholar has diverse field of readership.

5.1.1 Calculation of Scholastic Diversity Score

For a scholar, Scholastic Diversity Score is computed by comparing similarity/dis-similarity between keywords from the scholar's articles with the key words from referenced articles in the reference network. In order to improve the accuracy of this metric, semantic similarity is first computed with multiple levels of referenced articles, i.e. articles referenced by a referred by a referenced article will be used to in order to calculate similarity for level 2 and so on. Due to levels of nesting as we proceed in the network from on level to another, we impose a penalty according to the level.

In order to compute diversity for an article we use algorithm 5, we first compute similarity between keywords of an article and keywords of the referred articles at multiple levels. Similarity score computed at each level has a weight-age inversely proportional to the level of reference in the final score. We propose a increase in penalty by 50% at each level. Thus at level 1, there's no penalty. At level 2, half the dis-similarity score is added in the final diversity score and so on. In order to compute diversity score, we first compute similarity score at each level.
Let $\mathbf{R}_n$ be the set of all articles referred at level n. Then a set $\mathbf{L}_n$ is the set of all articles from R_n which are not part of any $\mathbf{R}_i$ such that $\mathbf{i} < \mathbf{n}$.

Let **K** be the set of all keywords for the main article the score for which is being calculated.
Let $\mathbf{s}_n$ be the similarity score for the article with respect to articles referenced at level **n** only. The respective diversity score $\mathbf{d}_n$ can be calculated as follows :

$$d_n = 1 - s_n$$

Let there be $\mathbf{x}_i$ keywords in a referred article $\mathbf{a}_i$. Let $\mathbf{y}_i$ of the $\mathbf{x}_i$ keywords be **semantically similar** to words in **K**. Let there be $\mathbf{M}_n$ articles in $\mathbf{L}_n$. Then the similarity $\mathbf{s}_n$ is calculated as follows:

$$s_n = (1/M_n) \sum_{i=1}^{M_n} (y_i/x_i)$$

The final diversity score for an article with 'n' levels of referred articles is calculated as follows :

$$diversity = (1/2^0) * d_0 + (1/2^1) * d_1 + (1/2^2) * d_2 + ... + (1/2^{n-1}) * d_{n-1}$$

Algorithm 5 Calculating Diversity Score for an article

Input: Keywords of article, Keywords of referred articles
Output: Dis-similarity/Diversity Score for the article
procedure $calc\_div\_score(keywd, ref\_articles)$
  $tot\_article \leftarrow 0$
  $tot\_similarity \leftarrow 0$
  for $article \in ref\_articles$ **do**
    $tot\_article \leftarrow tot\_article + 1$
    $similar\_word \leftarrow 0$
    $tot\_word \leftarrow 0$
    for $keyword \in keywords$ **do**
      $tot\_word \leftarrow tot\_word + 1$
      for $ref\_keyword \in article[keywords]$ **do**
        if $synonym(keyword, ref\_keyword)$ **then**
          $similar\_word \leftarrow similar\_word + 1$
          $break$
        end if
      end for
    end for
    $article\_similarity \leftarrow similar\_word/tot\_word$
    $tot\_similarity \leftarrow tot\_similarity + article\_similarity$
  end for
  $similarity \leftarrow tot\_similarity/tot\_article$
  $diversity \leftarrow 1 - similarity$
  return $diversity$
end procedure

Table 3 shows data obtained from web or calculated for two of the articles considered :

From the Figure 17, it is observed that there is a definite linear relation between Diversity Score and citation count. However, it is too early to predict that, the relationship between Scholastic Diversity Score and citation count is strictly monotonically increasing. But it is worth investigating further!

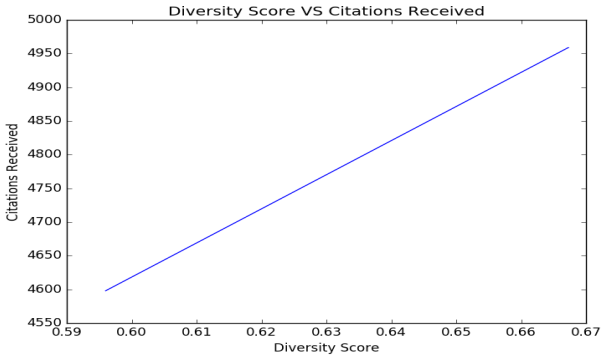

Figure 17: Diversity Vs Citation graph

Binomial model:
This relationship may also be interpreted as a failure modeling exercise. We can use binomial distribution to find the 'k' failures out of N trials. 'k' failures is equivalent to the number of weakly correlated topics out of a total of N topics (trials). Given the history of an author, we may find it interesting to predict the diversity of newly submitted/published article and based on the past history of correlation between Scholastic Diversity Score and the number of citations obtained, it may facilitate predicting the number of citations in a fixed time window.

$$P(A) = \sum P(\{(e_1, \ldots, e_N)\}) = \binom{N}{k} \cdot p^k q^{N-k}$$

is defined as the modeled Scholastic Diversity Score.

6. CONCLUSION

Bibliometricians often want to understand the pattern in which science and knowledge grows. Best way to perceive how and in what capacity the research has progressed is by building a network of articles and its references. Citation Network and Reference Networks are analyzed in the past to discover patterns that reflect growth and development of science. An Article Reference Network provides an understanding of the extent to which a scholar has progressed in his/her domain. Authors, in this paper, performed analysis of Reference Network on an IEEE Journal's article. The article's details are scraped using python script and stored in JSON files after being parsed by Beautiful Soup parser. Initially, a graph of references is build from the root node, which expands, as references are added at different levels. Once the graph is ready, graph theory algorithms have been used to find structures and patterns for extracting information. Between-ness centrality is used to determine most informative articles of the network. Topological sorting has been used to find paths from the root article to every other article at different levels in the network. In degree Vertex count returns the highly influential article of the network since it received the largest number of references from other articles. Second phase of the study was to carry out Natural Language Processing on huge keyword corpus that was built through web scraping. Keyword frequency analysis investigates

Table 3: Terence Tao Article Summary Sample

Title	Citation Count	References Count (LVL 1-4)	Diversity Score
Near-optimal signal recovery from random projections: Universal encoding strategies	4959	937	0.68
Decoding by linear programming	4598	1055	0.59

the occurrence of keywords in the entire network. Broadly, the high frequency keywords may define the subject area for the reference network. One of the major breakthrough of our work is the introduction of a score that measures dimensions (*spectrum or degrees of freedom*) of a scholar's research. The score, termed as Scholastic Diversity Score is an indication of how diverse a scholar's portfolio is. It is computed by comparing semantic similarity between keywords from scholar's articles with keywords from referenced articles. Similar the keywords are, less diverse is the scholar's readership profile. This score can be used to describe the spectrum of subject domains a scholar is proficient in, pertaining to his/her research interest.

The authors believe that domain proficiency and diversity estimated from a toy dataset is indicative of a trend and stronger validation and conjectures shall emerge as the size of the data set increases. It is worthwhile to note that the work presented here is markedly different from the approaches usually adopted in Scientometrics literature. The authors haven't investigated the coupling or co-citation networks to arrive at some conclusion. Rather, the focus is on the path of references up to a certain level (constrained by computational limitations) and scrutiny/identify articles which are old(chronologically) and still relevant. We intend to put forward the theory that the number of citations should not be the only criteria to measure the scholarship of authors. The authors must get some credit for the diversity of background reading indicative of the *intensity* of their preparation before writing a manuscript. Reading from various scholarly sources is a good practice to follow as we all know but *never has been quantified*, to the best of our knowledge. However, we don't claim that high diversity score should be called as a "golden rule" but nonetheless is a good exercise, especially for researchers in the early stages of their career. Finally, the observation and data discovery should help us build a *tool* where author profiles of institutions will be stored that will feature the *citations, breadth count of subject areas, Scholastic Diversity Score and nested reference links* for all the articles written by them.

7. REFERENCES

[1] IEEE - WEBSITE. "https://drsaraheaton.wordpress.com/2013/10/18/whats-the-difference-between-a-citation-and-a-reference".

[2] Webpage. "http://www.infotoday.com/online/may12/Belter-Visualizing-Networks-of-Scientific-Research.shtml ".

[3] Whitepaper. "http://docplayer.net/14505628-Whitepaper-a-guide-to-evaluating-research-performance-scientific.html ".

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[5] IEEE. Ieee - website, 2016. [Online; accessed 16-September-2016].

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[7] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA.

[8] D. S. M. D. Score. Diversity score moody's diversity score, 2016. [Online; accessed 7-Nov-2016].

[9] Wikipedia. Hierarchical clustering — wikipedia, the free encyclopedia, 2016. [Online; accessed 16-September-2016].

[10] Wikipedia. Longest path problem — wikipedia, the free encyclopedia, 2016. [Online; accessed 22-October-2016].

[11] Wikipedia. Strongly connected component — wikipedia, the free encyclopedia, 2016. [Online; accessed 17-August-2016].

[12] Wikipedia. Topological sorting — wikipedia, the free encyclopedia, 2016. [Online; accessed 6-September-2016].