Journal of Scientometric Res.2019;8(2s):s44-s73
http://www.jscires.org

SES-RREF: The Machine Learning Approach to Credible Metrics of Scholastic Evidence via Recursive Referencing

Archana Mathur[1], Snehanshu Saha[2], Saibal Kar[3], Gouri Ginde[4], Ankit Sinha[5]

[1]Department of Information Science and Engineering, Nitte Meenakshi Institute of Technology, Bangalore, Karnataka, INDIA.
[2]Department of Computer Science and Engineering, PES University, Bangalore, Karnataka, INDIA.
[3]Centre for Studies in Social Sciences, Calcutta, West Bengal, INDIA.
[4]University of Calgary, CANADA.
[5]Scibase, Center for Applied Mathematical Modelling and Simulation.

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. Despite this recognition, citation network is mainly used for judging the contribution of an author or a journal for the scientific community. Analyzing citations of an article or of the journal to which it belongs, follows a bottom-up approach and provides a varying degree of information. These include the pattern of spread and the influence it has in academics, *per se*. However, analysis of references provides a top down approach. The present paper introduces the concept of reference network analysis with the objective of measuring the extent of scholarly inculcation of knowledge and effort while pursuing specific research work. Such reference networks can examine how variegated a research is (diversity) and intensity of the concepts studied (depth) by a researcher. We prove that both these aspects play crucial roles in generating recognition by not relying on citations explicitly. The paper uses these features to devise article-level and author-level metrics,like Scholastic Evidence Score and Trust Score. Using two different case studies of highly reputed scholars, we further demonstrate that Trust score of reputed and reliable authors do not fluctuate noticeably with time. On a broader spectrum, the durability of citation might reflect the depth of a scientific contribution. Our contribution imparts multi-dimensional approach to scholarly influence and creates avenues for future explorations in journal credibility study.

Keywords: Scholastic Evidence Score (SES), Recursive References (RREF), Depth, Diversity, Convex Optimization.

Correspondence
Ankit Sinha
B-214, VRR Nest Apartment, Konappana Agrahara, Electronic City, Bangalore - 560100, Karnataka, INDIA.
Email: ankit.sinha2000@gmail.com

Received: 15-02-2019
Revised: 30-04-2019
Accepted: 11-06-2019
DOI: 10.5530/jscires.8.2.24

INTRODUCTION

In recent times, the researchers and evaluators of scientific contributions have shifted their attention from Journal level to Article level metrics, as part of a paradigm change. The main reason behind this transition is perhaps owing to the fact that the number of issues published by a journal and the number of articles per issue often obfuscates the importance and contributions of a specific article. After all, the observed impact of a given journal might not do justice to the quality of a specific article published by it owing to the overall score being usually calculated as a weighted average. There should be little doubt that within a given journal there is always substantial heterogeneity in the quality and impact of articles published. Therefore, formation of a standard opinion about quality of all published papers based on the overall metrics of the respective journal might be construed as a disservice, particularly to those which perform better according to measures of individual metrics. Indeed, for a precision-loving scientific community opinions formed according to the grand scores attached to a journal is tantamount to enduring statistical discrimination towards each article published by it. The subjective-ness involved in assessing the true quality of an article is another stumbling block towards isolating one from the other when published by the same journal. The prevalence of statistical discrimination arising from asymmetric information between different agents is quite common in commodity markets[18] and labor markets[19-22] of any society and economy. Despite many screening and signaling devices created to

overcome the undue standardization, the pooling equilibrium continues as a vexing problem. To attain a quality-wise separating equilibrium, a published article must therefore match exactly with the appropriate journal and that this must become the character of scientific publications, *en masse*. For all practical purposes, homogenization of quality within a journal and strict heterogeneity across journals are both difficult to arrive at. Of course, the distinction between a top rated journal and a poor quality journal, despite all the subjectivity, has some validity in the common perception. Finally, whether the established quality of a journal facilitates the visibility of an article or the quality of an article helps to attract readership for a journal is a moot question, equally applicable to many other aspects around us. The ability to isolate the quality and persistent impact of an article, seems one of paramount importance if we are to draw useful conclusion in this matter. This makes the article level estimation/ measurements of scholastic indicators of individual articles imperative. Traditional metrics like Article Influence Score[17] and PLoS Article Level Metrics[7] are subject to doubts as these are based on citations and could therefore be biased. Furthermore, these might not reveal a truthful and reliable measure of the quality of articles. One possible point of entry into this matter is driven by the steps or procedures followed during a research work. It is generally agreed that considerable amount of preparation and collection of relevant information lies at the core of identifying a potential research question, and subsequently developing an output. Often, the quality of research output submitted for publication depends on the quality of these prerequisites, and the effort exerted by the researcher in attaining a desirable quality. An integral part of this performance involves looking up potential references to be cited. Naturally, many more than those actually cited may have to be looked up in order to find the appropriate connections. This process is often recursive in the sense that a given paper may cite *n* different papers, some of which are directly useful in the respective context, but these citations might as well have used *n* number of papers, which indirectly influence the present research and need to be studied. Thus, the whole set of recursively referred to papers cumulatively determine the quality of a given paper. The process of obtaining this network of recursive references, though tedious and time consuming, can be of great help for the researcher and his/her collaborators, i.e. collectively for an entire group of researchers. Scholastic Evidence Score (SES), described later, can help researchers realize the type of foundation required to produce an effective paper by focusing on diversity and depth of such prerequisites. A relevance score can be attached to the collective set of references which may be utilized in order to complete the study with a proper direction. It may also be possible to identify community of researchers working on the topic that the researcher is interested in. The strength and reliability of an article, as well as its scientific/ scholastic value depend on many factors. The authors of the manuscript are of the opinion, which is established later, that a recursive citation network, built through sophisticated technological tools, may go a long way in ascertaining and determining those factors. The study, carefully done, aided and abetted by a novel mathematical model, is complex and interwoven between the Recursive Reference Network (RREF) and the evidence score of articles, SES. The larger picture is to construct a set of reliable authors and articles domain wise. SES serves as the conceptual framework laying the foundation for defining and computing a new metric for articles (and authors eventually), trust value. RREF acts as the necessary tool that replenishes this framework and helps achieve the goal of building the reliable corpus of articles and authors. It must not be lost in translation that RREF is a software designed by the authors which should serve as an extremely useful tool. This has not been attempted before and should stand out in the event the academic community decides to discount the other contributions of the manuscript. The study is motivated by the following factors:

- citations alone may not capture the true essence of an article
- an article with lean citation corpus may be rich and path-breaking; therefore associating fat citation index with the quality of article may be misleading
- diversity in background preparation may be revealed by reference patterns where apparently disconnected articles may be referred by an article, and with a reasonably good justification. This is different from "erratic" and "motivated" reference patterns, as demonstrated by a few seminal articles chosen for our study.
- quality and reliability of articles and trustworthiness of references is a complex issue. Existing article metrics do not capture the breadth and depth of the scenario.
- the problem mentioned above demands a multi-faceted study linking citations received by articles with intrinsic characteristics of the article such as depth and diversity of background study while preparing the article.
- provide young researchers with a trustworthy reading list by mining references of articles passing the test of quality as defined by complex quality considerations.
- Collaborative or community behavior (copious citations) is often observed.[5] Such behavior, though unethical, is unfortunately visible even among well known researchers. Searching for trustworthy articles, free from such tendencies, is a value added service to the community as it provides some principled guidelines to scholarly publishing. The authors believe that, trustworthy articles written to propagate the selfless advance of knowledge are revelations

of years of hard work and therefore manifest true intrinsic scholarly evidence such as diversity and depth. These articles are not designed to game citations and therefore, reference study of such scholarly work is indeed a worthy exercise.

We intend to stress the fact that our study initiated with the sole purpose of creating the reference pool and development of a software (RREF) to access those references in quick and efficient manner. This was the primary motivation as we strove to build a platform for the immediate benefit of early career researchers **(RREF- section 4)**. It is well known that searching for references is painstaking and a test of patience. However, the process of building the software laid open a plethora of questions which led to critical analysis of the reference network. We reiterate that the proposed model for scholarly evidence and trust is a derivative of the initial study. However, the outcome has been too overpowering to ignore and has hijacked the center-stage of the manuscript.

STATEMENT OF THE PROBLEM

When an "early career" researcher looks for materials to begin an active research, an important contribution or significant help to extend would be to provide him/her with a decent set of articles and a knowledge base. An article with a trustworthy/reliable list of references provides the knowledge base. On the contrary, if the article itself suffers from its own myopic inhibitions, driven by the goal of gaming citations, is of little value. The list of references should be a balanced mix of two types: one that is pertinent with the theme of the main article i.e. depth computed using machine learning techniques and the second belongs to the type of articles that are apparently dissimilar to the theme of the citing article, defined as diversity. The latter, if reliably compiled by the source article, holds the key to the diversity spectrum of the background study undertaken to write such a scholarly piece. Along with the software which provides the ready references, these two metrics not only speak of the intrinsic quality of the article, they do offer inspirational value to the early career researcher. These metrics speak volumes for intellectual and esoteric virtues of the article and the level of commitment. These metrics, diversity and depth, combined to create some measure of trust, set examples for the researcher.

The problem under investigation aims to bring out the nuanced aspects of scholarly, hitherto unexplored. Usually, quality metrics are determined "post-facto". Our endeavor is to propose quality measures which are devoid of external factors such as peer review and map these kernel measures with post-facto metrics for a complete evaluation. In a nutshell, we investigate and construct founding principles of knowledge base by navigating through sophisticated analytical and computational techniques.

1. What citations and semantic networks don't capture?

Lean citation corpus don't get highlighted except within niche peer groups. However, there are several such niche groups in theoretical statistics, Astroinformatics, Mathematical Analysis, to name a few, where h-index/citations don't accurately reflect the quality of articles. Semantic network may elucidate the knowledge flow but fails to bring out the scholarly preparation required for such articles, which often do not get the attention they deserve.

Moreover, the rich body of information (diversity, depth and Trust Value) is usually restricted within the niche peer group. Out article intents to disseminate the information and make it available in public domain. The article put forward the following hypothesis:

- Diversity (background preparation) is a quality metric that should not be ignored.
- Diversity needs to be computed from the reference network not citation network.
- An article with few citations, may be trustworthy and free from esoteric information and knowledge, if a good mix of diversity and depth is evident.
- Diversity should complement depth.
- A range of diversity is suitable for the article as far as the receiving citations is concerned.
- By computing diversity, depth, associated citations and Trust Score, we intend to show the trustworthiness of articles, there-by providing a knowledge base for young researchers and guidelines to mine the reference network of the chosen articles. This exercise mitigate the concern of possibly having to go through "erratic" reference patterns. Erratic reference patterns may boost the diversity score but is not credible. This is the reason to investigate Trust score and its global optima (to analyze the deviation from the global maxima of the trust value curve of celebrated scholars)
- We offer an insightful baseline study with regard to the quality and credibility parameters, never investigated before.

OUR CONTRIBUTION

Any reference network that stipulates an inclusion of strikingly different domains referred by a scholar can open a realm in which an article's as well as author's potential, integrity and truthfulness can be measured. The section introduces concepts and definitions of new metrics which evolved during this research. The web application and visualization tool are few important outcomes explained in section 3.2.

1. Concepts, models and quantifiers introduced

Nested Reference Network (NRN): It is a reference network represented as a graph. The nodes of this graph represent articles and the directed connections between the nodes represent the referenced relationship. A referred article has connections present with articles referenced by it, providing the nested structure and thus justifying the name Nested Reference Network. This is termed as Recursive Referencing Framework (RREF) throughout the remainder of the literature (refer to Section 4.3).

Diversity and Depth: The concept of NRN is deployed to derive new metrics like Diversity and Depth. Diversity in reference network is a metric which reflects the preparation of an author while performing an investigation. The diversity score is a quantitative value that signifies a scholar's preparedness in his/her domain. Depth score, on the other hand, deals with computing the depth the subject field has traversed. Since citation count alone is not enough, both scores together with citation count can be used as a good quality/reliability indicator at article and consequently at author level (refer to Section 5).

Scholastic Evidence Score (SES): SES is a combination of diversity and depth computed from Recursive referencing framework (RREF). Diversity and depth are a pair and complement each other while contributing to Scholastic Evidence Score (SES) (refer to Section 5.2). SES of articles are computed by using Machine learning techniques such as Latent Dirichlet Allocation (LDA), similarity measures etc (LDA is explained in detail as a part of Case study for Dr. Terrence Tao's Reference Network).

Trust Score : A credibility metric, which analyzes and quantifies the extent and quality of scholastic preparedness of peer reviewed articles. This is an exercise complementary to citation analysis of those articles over time. Trust Score is a metric of article/author trustworthiness by correlating the citations he/she receives (which is an indicator of the value of articles written by him/her as endorsement from his/her peers) coupled with SES. A novel growth function for Trust Value is presented later. Conditions for global optimization of Trust score are derived and studied in detail in section 6.

- **Predictive Model:** A binomial model to predict article's diversity from historical data is presented (please see section 5.1).

2. Technical contribution and Implementation

- **Web-Scraping API's:** A collection of web-scraping API's are written to retrieve article-level, journal-level and author-level information from websites like IEEE Xplore[8] and ACM Digital Library (DL). In our current work, the API's are used to sweep an article's references and data is stored in JSON structure.
- **Creation of a Visualization Toolkit-RREF:** To assist in the undergoing study, a visualization toolkit has been created which helps an individual or a scholar see the underlying relation between data in a much more comprehend-able form. The toolkit allows the user to select the data and see the underlying relation or distribution. [1]
- **Knowledge discovery from data:** There are challenges associated with JSON structure in terms of time and space complexity. Moreover, a JSON structure is not interpretable as a graph. To get the better of these problems, an n × n adjacency matrix of references is built, on which graph theory algorithms are applied to extract pertinent information for reference networks. This is elaborated in appendix A3.

Figure 1 highlights the various techniques used to scrape, process and analyze a Reference network. It also indicates various outputs/results obtained as these techniques are put into effect. The remainder of the paper is organized as follows. The next section (section 4) discusses the importance of Reference Network and highlights Graph Theory based techniques to extract details on articles scraped from IEEE website such as most referred, most important article, chronological growth of a network etc. Section 5 defines SES, Scholastic Evidence Score, and brings out article's diversity and depth for its computation by using LDA and cosine similarity concepts. Relationship between diversity and citations, a key component in computing Trust Score, is also explored in same section. Section 6 examines Trust score model and establishes the suitability of CES production function for computation of article's (and correspondingly author's) trust score. Authors propose a novel model, "Additive Trust Model", wherein the model's input parameters are endowed with different elasticity of substitution. The effectiveness of the new model is also discussed in same section.

The following sections (section 7 and section 8) are case studies on articles of Dr. Vidyasagar and Dr. Terrence Tao. Each case study reveals in-depth analysis of author's Reference Network and computes trust score based on the proposed model. Section 9 concludes the manuscript and is followed by Appendix where JSON structure explaining information storage and retrieval are elaborated. The following sections (section 7 and section 8) are case studies on articles of Dr. Vidyasagar and Dr. Terrence Tao. Each case study reveals in-depth analysis of author's Reference Network and computes trust score based on the proposed model. Section 9 concludes the manuscript and is followed by Appendix where JSON structure explaining information storage and retrieval are elaborated.

REFERENCE NETWORK : THE BACKBONE FOR DIVERSITY AND DEPTH COMPUTATION

There exists various kinds of net works[3] related to scholarly publishing information, which are diverse in nature and usage such as Collaboration networks, semantic networks and publication citation networks. In this study, we introduce 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 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. The ramification of building such a network unfolded in terms of completely

1. Network of citations and references: A note on key distinctions

In order to truly understand the emphasis on references, let us formally begin by appreciating the difference between citations and references.[2] References are a list of articles referred by the authors of a paper. This is a list of sources one (authors) has/have cited. Generally, the references are listed in APA style. In fact, every bibliographic item listed in references has to be cited in the main text. Every source listed in references should be accessible by others who read the paper. It is like a paper trail or footprints one leaves for readers to help them compile an optimal reading list eventually.

Citation is a specific source that is mentioned in the body of the paper. Most of the popular metrics are built on citations. This triggers a wide area of research on citation analysis and citation networks. Citation per se, is a reference to scholarly

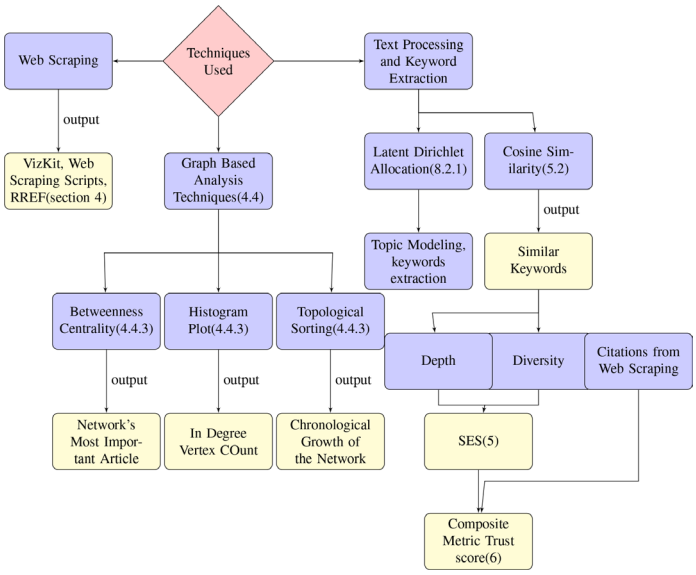


Figure 1: Flowchart showing different techniques and outcomes in the manuscript: internal section referencing is done within the boxes

works to give due credit to earlier contribution. Regardless to say, all promising metrics at article, author and Journal level are built on citations. On a close analytical evaluation, citations alone can not work as an independent ingredient for evaluation of an author's intellect. They can be manipulated through practices like injudicious self citations, coercive and copious citations. Nevertheless, citation network and reference network can be seen as two sides of the same coin. The former concentrates only on building a graph of crucial citations across authors, whereas the latter, builds a network of articles which researcher would have explored. Nodes in citation network are articles represented as author name and publication year (author, year of publication). Edges are citations which exist when one articles cites another. In order to keep visualizations manageable, only those connected articles that are cited more than a certain threshold are shown. Contrary to this, a reference network shows all articles of reference list where every node (i.e. every article) is assigned an identifier. The authors have restricted the number of levels to 4 with so that the visualization is explorable, within a feasible limit.

A reference network thus, represents a graph; the nodes represent articles and the directed connection between the nodes represent the referenced relationship forming a Nested Reference Network (NRN). Analysis of this network can help exploring the history of many influential article of a journal. Identifying various important articles in the reference network of such an article, using graph theory, helps identify the path breaking articles which contributed to 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. 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 graph databases, such as Neo4j, 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.

2. Why Analyze References?

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 may become extremely narrow and streamlined. This can vary from one research domain to another. There can be multiple reasons why one refers an article. As explained by Eugene Garfield[4] citations in scholastic work may influence other work. Citation is a means for acknowledging prior

work, identifying methodology, equipment. A sensible list of citations provides useful background reading, creating an environment of scientific temper and room for constructive criticism. Moreover, references indicate breadth of the research domain. It could measure diverse topics authors consulted before proposing the solution. Reference network could be a good indicator of depth of the topics related to a domain as well. Let us first analyze the network at conceptual level.

3. Theoretical Analysis of Reference network

Figure 2 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, D, C, E, F, G and L articles in its references list. Out of all the articles in the graph, J, K and L are the oldest referenced articles and A is the latest article. Article A is the successor of all the articles in the graph. Article J, K and L are the articles, which are predecessor articles, comprise of the most initial work in that research domain. Figure 3 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, betweenness 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 as shown above.

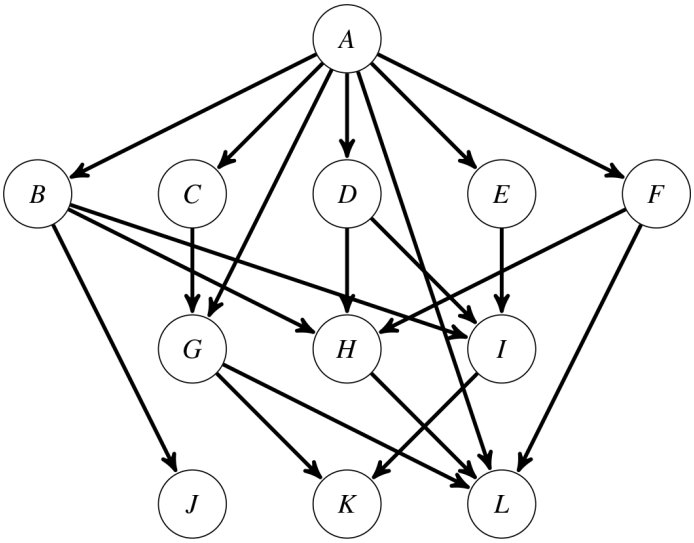


Figure 2: Reference network

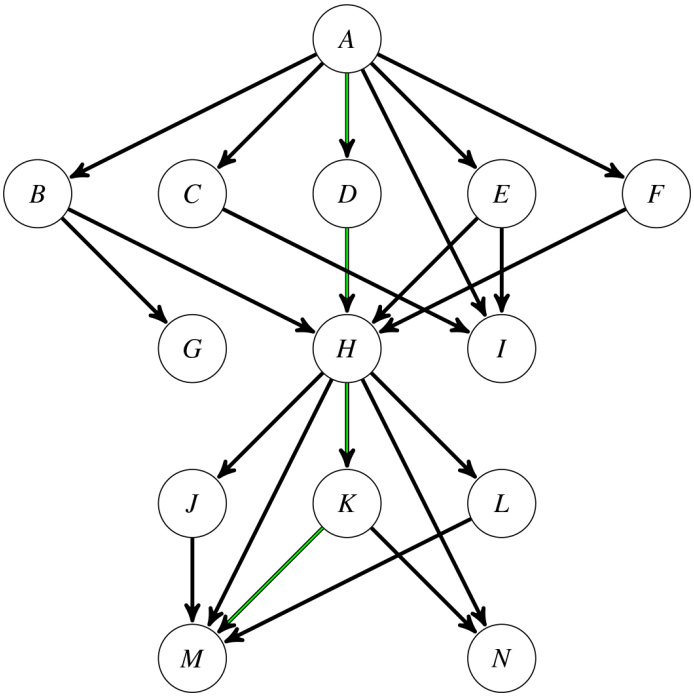


Figure 3: Directed Acyclic Graph of reference network

4. Reference Network Analysis using Graph Theory

a. Reliable Data Acquisition

Data is an imperative factor in any sort of analysis. A reliable source and acquisition method must be used for reliable analysis. A route of non subscription based data accumulation methodology from IEEE Explore was executed for this research, which was primarily achieved through web scraping. In order to scrape data for further processing, articles containing relevant information were identified.Web-scraping scripts using python and it's libraries were run which allowed automated access of the web-pages containing the required data. BeautifulSoup, a html/xml parser available in python, was used to locate the required fields of data and extract information from them. The raw data gathered was then stored as JSON files for further processing. To ease the process of analysis of data, it is often required to perform transformation on the obtained data. In this study, scraped data contained unwanted characters which were removed. In case of accented characters due to Unicode, the text was decoded to support only the standard alpha-numeric characters supported by ASCII. The article format was then restructured for easy access of related fields and articles. Figure 20 (in Appendix C), shows a processed sample article JSON format obtained at the end.

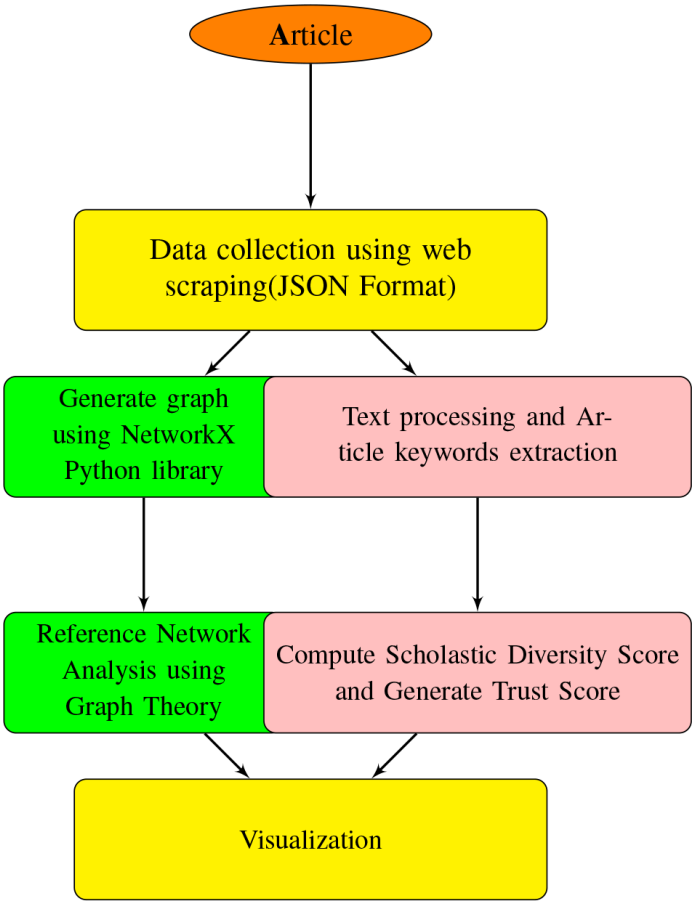


Figure 4: Overall system architecture

b. System architecture

Figure 4 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 using web-scraping python script. As indicated earlier, the script utilizes HTML parsing tools, BeautifulSoup, to extract references and other details and stores in JSON file. The raw data gathered for constructing the reference network contains articles of IEEE journals only. We extracted reference 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 LDA for topic modeling.

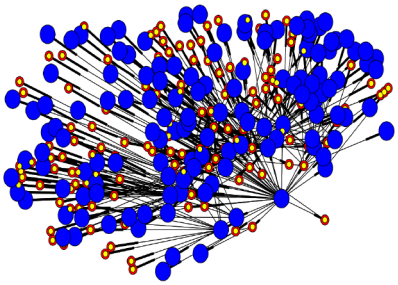
c. Article reference network analysis

The references graph network analysis of the highly cited articles of a research scholar can provide very interesting 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 article id:6656866 from IEEE Journal of Translational Engineering in Health and Medicine for analysis. Figure 5(a) 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.

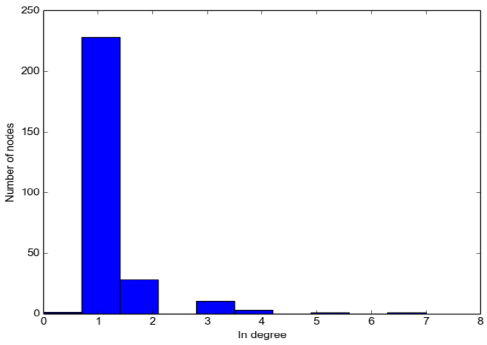
Graph theory based algorithms can be now easily used on this network which can yield interesting results.

- **A Network's Most Referred Article:** A node's *In-degree count* is the number of articles, the node is referred by, in a given reference network. Figure 5(b) shows the number of in-degrees vs number of articles in a histogram plot. As shows there are a very few articles(just 8 out of 250) in the network which have in-degree greater that 3. These 8 articles can be termed as the most referred articles of this reference network.
- **A Network's Most Important Articles:** 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 5(d) shows the all the most important articles of the network computed using betweenness centrality Table 1 (All tables are in Appendix C) shows the details of the top two of these articles.

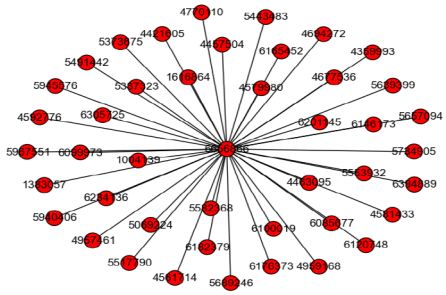
  Analysis: On analyzing the article Id's from this exercise we could find following two articles which stood out of all the articles in the reference network which appear to be very influential.
- **Chronological Growth of a Network A topological sort** is a non unique permutation of the nodes such that an edge from u to v implies that u appears before v in the topological sort order.[14] 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 5(c) shows the topological sort starting root node with article id:6656866.



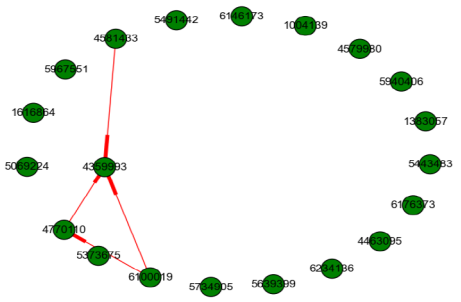
(a) IEEE reference network of level 2, article id: 6656866.



(b) Vertex count histogram.



(c) IEEE topology sort for an article



(d) IEEE most important articles

Figure 5: Reference Network Analysis Using Graph Theory

- **Longest path in the Network:** In order to find the longest directed path in the network, topological ordering is found in a Directed Acyclic Graph (DAG).[12] The length of the longest path ending at v is computed by keeping track of incoming neighbor list and adding one to the maximum length recorded for those neighbor. For each vertex v of the DAG, if v has no incoming neighbors, the length of the longest path ending at v is set to zero.

  For the reference network built using scraped IEEE articles, longest path: is 6656866– > 5657094– > 5553932– > 4359993– > 1634506 as shown in the Figure 6. In this picture the nodes represent the IEEE article id's.

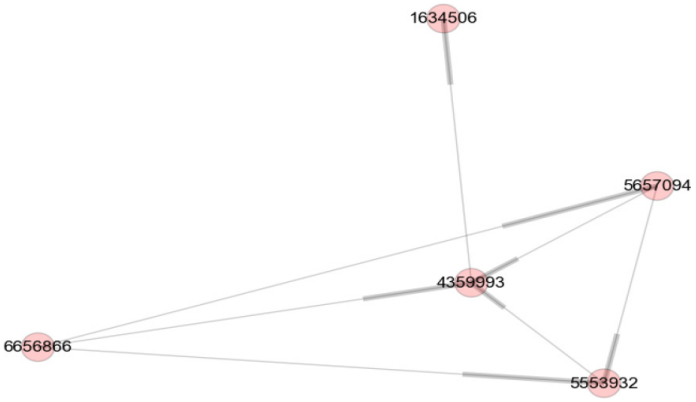


Figure 6: IEEE longest path

The current section explored various graph based techniques to find the most crucial articles in a Reference Network, the following section uses Reference Network of articles of acclaimed author Dr. Mathukumalli Vidyasagar, to illustrates the computation of Scholastic Evidence score (SES) by incorporating diversity and depth of articles. Diversity in background preparation is a proof of author's preparedness while writing his/her research. Authors believe and in later section, establish the fact that more diverse the research is, more likely the article may receive citations.

SCHOLASTIC EVIDENCE SCORE(SES): OFFSPRING OF RREF

Two apparently different topics may be connected and when we make a "topical" discovery, it *springs shock* to most of us, including the critical audience. This is the triumph of diversity. It is well known that measuring quality of a published article is a difficult task and various measures proposed toward that goal invited criticism. We introduce a new measure, Diversity in Background Preparation without tall claims. 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. 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. In this section, we integrate these quality parameters with a novel metric, Scholastic Evidence

Score. The score is a derivative of an author's reference network, which is network consisting of an article along with it's references and their references and so on in a nested or recursive manner. We later illuminate upon the relationship between Recursive Referencing (RREF) and Diversity.

SES is computed as a combination of diversity and depth. As already highlighted in the Introduction, these are one of the factors on which strength and reliability of an article, as well as its scientific/scholastic value depends. Lets look at their definitions before proceeding for SES computation.

Diversity: The definition of diversity is borrowed from Moody's Investors Service, used in a slightly modified form. Moody's created a measure, Diversity Score, to estimate the diversification in a portfolio, related to a collateralized debt obligation (CDO). We compute diversification score by the extent of industrial diversification of a portfolio. "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".[15]

Hence, diversity score measures the number of uncorrelated research domains studied by a scholar while writing manuscript. A low Diversity Score could imply not enough research in terms of variety, while an extremely high Diversity Score could suggest insufficient reading for the central concept. 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. Diversity Score is thus computed from *RREF*, the *nested reference paths associated* with an article-the proof of potential of which is 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. Algorithm 1 calculates diversity score of an article, similarity score is computed with decreasing effect at every level. So, at each 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 works in favor of diversity score. Since the keywords are mutually exclusive in nature, it implies that the scholar has diverse field of readership.

Depth: Authors believe that quality and trustworthiness must involve a measure of depth, extent and insight to which the subject domain is studied. Depth, a complementary metric to diversity, is defined as a measure of topical similarity i.e. if d_n is the diversity score of an article at level n, than, s_n, similarity score or Depth score is defined as: $s_n = 1 - d_n$.

1. Calculation of Scholastic Evidence Score

The first step in computation of SES is Topic Modeling. It is a text-mining tool to discover the abstract "topics" in Reference Network and is executed via Latent Dirichlet Allocation (LDA). Once the results of topic modeling are obtained, semantic similarity between the article keywords and the keywords of articles referenced at different levels is computed (using cosine similarity metric used in Algorithm 2). Modified theory says neither a too strong semantic similarity nor a too weak semantic similarity score is optimal.

Due to levels of nesting as we proceed in the network from one level to another, we impose a penalty according to the level. Similarity score computed at each level has a weightage inversely proportional to the level of reference in the final score. We propose an increase in penalty by $1/n_i$ at each level, where i is level and n_i is number of referred article at level i. Thus at level 1, there's no penalty. At level 2, $1/n_2$ of 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 R_n be the set of all articles referred at level n. Then a set L_n is the set of all articles from R_n which are not part of any R_i such that $i < n$.

Let K be the set of all keywords for the main article the score for which is being calculated.

Algorithm 1 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 ← 0
    tot_similarity ← 0
    for article ∈ ref_articles do
        tot_article ← tot_article + 1
        similar_word ← 0
        tot_word ← 0
        for keyword ∈ keywords do
            tot_word ← tot_word + 1
            for ref_keyword ∈ article[keywords] do
                if synonym(keyword; ref_keyword) then
                    similar_word ← similar_word + 1
                    break
                end if
            end for
        end for
        article_similarity ← similar_word/tot_word
        tot_similarity ← tot_similarity + article_similarity
    end for
    similarity ← tot_similarity/tot_article
    diversity ← 1− similarity
    return diversity
end procedure
```

Let s_n be the similarity score for the article with respect to articles referenced at level n only. Let there be x_i keywords in a referred article a_i. Let γ_i of the x_i keywords be *semantically similar* to words in K. Let there be M_n articles in L_n. Then the similarity s_n is calculated as follows:

$$S_n = (1/M_n)\sum_{i=1}^{M_n}(\gamma_i / x_i)$$

The respective diversity score d_n can be calculated as follows:

$$d_n = 1 - S_n$$

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

$$diversity = (1/1) * d_0 + (1/n_1) * d_1 + (1/n_2) * d_2 + \cdots + (1/n_{n-1}) * d_{n-1}$$

2. Predictive model on Binomial Distribution predicting diversity from historical data:

Binomial distribution function is a good predictor of 'k' successe by using total trials N and probability of success in each trial. 'k successes is equivalent to the number of weakly correlated topic (or diversity) out of a total of N trials (topics or keywords). Give the history of an author, we may find it interesting to predict th diversity of newly submitted/published article and based on th past history of correlation between diversity Score and depth, may facilitate predicting diversity in a fixed time window. Th probability mass function of the binomial distribution is given a

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

where:
N is the number of topics
k is the number of weakly correlated topics
p is the diversity score of the article
q is the depth score.

Remark: Diversity Score lies between 0 and 1 with 1 as the maximum diversity and 0 as minimum.

3. Prerequisite for calculating diversity

For each article scraped, a corpus of keywords was created by extracting keywords of the article under examination, creating a *root list*, along with the keywords of articles present in the set of it's nested references, creating a *reference list*. 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. Also, LDA is performed for topic modeling, discussed in section 6.2 below.

Algorithm 2 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[][] ← −1
    for kwd_1 ∈ root_list do
        for kwd_2 ∈ ref_list do
            score ← Cosine_Similarity(kwd_1,kwd_2)
            if score >= 0:6 then
                score_matrix[i][ j ] ← α * score + (1 − α) * 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 ← (kwd_2; score_matrix[i][ j])
            end if
        end for
    end for
    return sorted_keyword_list
end procedure
```

Cosine Similarity Metric: Similarity metric is a textual based metric returning the extent of similarity or dissimilarity (distance) between two pieces of text (strings). It returns a floating point number indicating the magnitude of similarity on the basis of lexicographic match. For example, similarity between the strings arrange and range is considered significantly greater than the strings apple and orange. Cosine similarity, an often used metric, is a vector based similarity measure. Cosine of two vectors a, b can be derived by using the inner product formulation.

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

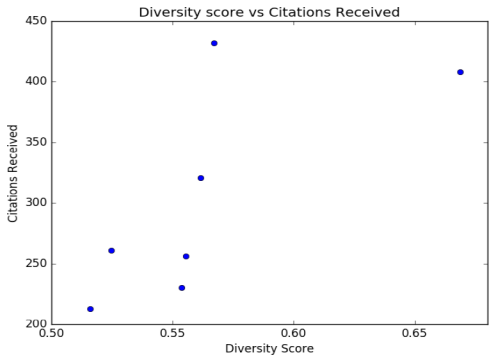
Where, θ represents the angle between a and b.

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

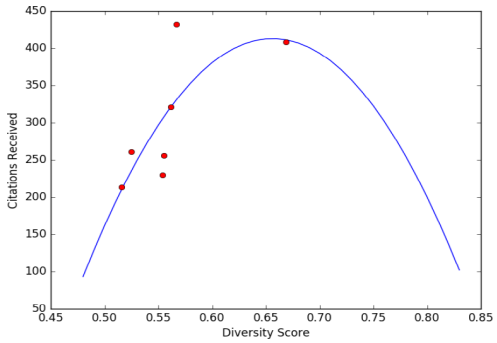
where A_i and B_i are components of vectors **A** and **B** respectively.

4. Relationship between Diversity Score and Citations

Rationale: Erratic citation/reference pattern may augment diversity score but could be insignificant towards the true diversity in background preparation. It may also suffer from the allegation that the references are motivated and therefore suffer from lack of credibility. Therefore, we investigate a range for diversity score which may contribute to the growth



(a) Diversity Vs Citation graph



(b) Citations vs Diversity Score: the plot shows an increase in citations for diversity ranging between 0.4 to 0.6. Too little or too much diversity may affect the quality of an article and citations consequently.

of citations. At this moment, this is a possibility and subject to empirical study. Therefore, we have chosen scholars whose integrity and intellect are beyond reasonable doubt. The authors believe that there might be some insight in the relationship between diversity of articles and citations received by the articles. this is different from attributing quality of articles based on citations only. Rather, a visual inspection backed by analytical approximation of the relationship may reveal significant information. This is achieved by quadratic least square fitting.

The scatter-plot of Diversity Score vs Citations for the seven articles of Table 2 (in Appendix C) is shown in Figure 7(a).

The above figure suggests that a linear relationship between Citations and Diversity Score might not exist. In order to determine the relationship between these, quadratic regression was performed on diversity and citation values on Vidyasagar's article using Table 2 (in Appendix C). The quadratic regression is a method of finding equation of parabola that best fits the data points.

- Let x be the diversity and y be the citations
- The Quadratic Regression equation in terms of diversity(x) and citation(y) is given by:

$$y = Ax^2 + Bx + C, \tag{1}$$

(2)

where

A = −10292.6

B = 13508.5

C = −4019.4

- The coefficients A,B and C are obtained from the standard statistical equations: $\bar{x} = \frac{\sum x_i}{n}, \bar{y} = \frac{\sum y_i}{n}, \bar{x}^2 = \frac{\sum x_i^2}{n}$
- Interim Variables are generated:
  - $S_{xx} = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{\sum x_i^2}{n} - \bar{x}^2$
  - $S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\sum x_i y_i}{n} - \bar{x}\bar{y}$
  - $S_{xx^2} = \frac{\sum (x_i - \bar{x})(x_i^2 - \bar{x}^2)}{n} = \frac{\sum x_i^3}{n} - \bar{x}\bar{x}^2$
  - $S_{x^2x^2} = \frac{\sum (x_i^2 - \bar{x}^2)^2}{n} = \frac{\sum x_i^4}{n} - \bar{x}^2\bar{x}^2$
  - $S_{x^2y} = \frac{\sum (x_i^2 - \bar{x}^2)(y_i - \bar{y})}{n} = \frac{\sum x_i^2 y_i}{n} - \bar{x}^2\bar{y}$
- Trend line is obtained: $y = Ax^2 + Bx + C$, where:

$$A = \frac{S_{x^2y} S_{xx} - S_{xy} S_{xx^2}}{} \tag{3}$$

$$B = \frac{S_{xy} S_{x^2x^2} - S_{x^2y} S_{xx^2}}{S_{xx} S_{x^2x^2} - (S_{xx^2})} \tag{4}$$

$$C = \bar{y} - B\bar{x} - C\bar{x}^2 \tag{5}$$

Figure 8(b) shows the graph obtained. The graph clearly indicates that for article's diversity values ranging between 0.4 to 0.6, the article may receive very high citations. Forvalues beyond 0.6 and less than 0.4, citations may not be as high.

5. A composition of Diversity, Depth and Citations

As discussed in section 4 of the paper, citations are inadequate representations of an article's quality. Correspondingly, diversity alone cannot work as an independent element to judge a uthor's preparedness. Extremely low or high values of diversity may indicate poor depth in research as citations grow only for certain range of diversity values. Figure 7(b) reflects an increase in citations only at diversity values ranging from 0.4 to 0.6. [colback=red!5!white,colframe=red!75!black] Utilizing these metrics, authors take the privilege of introducing Trust Score Model, which uses a growth function and analyze article's as well as author's trustworthiness by correlating citations he/she receives, its depth and diversity. The model provides an overall score to authors' profile by analysis of their work overtime. It judges every work of an author at micro level and scores the author's complete profile at macro level, giving an insight whether it's worthwhile to analyze an author's profile further.

A COMPOSITE CREDIBILITY METRIC

Trust Score evaluates an article's and author's trust value and validates its credibility in its research domain. The score could be researched upon further and later introduce Scholastic Integrity Profile. Given the humongous size reference network, a filter preventing most of the calculations could be more than desirable. Diversity and depth are novel metrics and journal database do not contain this data, we attempt to map a universally accepted quality metric, citations per article under consideration. The section begins by introducing CES production function and its suitability in deriving Trust Score. The motivation behind using CES function, its optimization, the curvature characteristics of the function and Stochastic Gradient Ascent algorithm for computing elasticity values are examined in detail. A variation of Trust Score Model called Additive Trust Model is introduced and its worthiness is investigated for computing article's score. These models are used for computing Trust scores for articles authored by Dr. Vidyasagar and Dr. Terrence Tao.

1. Trust Score Model and Metric: A novel model

The information about credible and high impact research is usually limited to peer groups/communities. Exceptional cases always exist but the usual norm is that, the information about a particular research domain and remarkable articles in that domain are not disseminated fast enough outside the peer group. The quantification of trust is not an absolute necessity within the peer group as reasonable information about quality and trustworthiness of articles and authors are available via grapevine. However, the restrictive peer group culture also doesn't encourage the flow of information across domains in the way as desired. Quantification of trust and scholarly value of articles should be available to anyone, irrespective of the nature scientific/research alignment. As diversity is the cornerstone of this exercise, we strongly believe that a "Trust quantifier" will go a long way in helping young researchers and people from heterogeneous research domains identify articles as beacon indicators or good starting points in their endeavor in that particular field. We begin by borrowing a model from production economics in our efforts to build a trust metric and model.

a. Introduction to CES Production function

The Constant Elasticity of Substitution Production Function belongs to the family of neoclassical production functions that displays constant change in output as a result of any change in input parameters. Algebraically, CES production function for two inputs can be shown as

$$Q(L,K) = (\alpha L^{\rho} + (1-\alpha)K^{\rho})^{1/\rho} \quad (6)$$

where Q = Quantity of output; L, K are Labor and capital, respectively, $\rho = \frac{s-1}{s}; s = \frac{1}{1-\rho}$ is the elasticity of substitution and is share parameter[23].

Consider a case where an article's trust score is to be computed. While writing the article, author uses reference network N that is built from articles cited at various levels. Diversity Score (V) and Depth score (P) for the article can be computed by using algorithm1 and algorithm2. If C shows the citation count (number of citations the article receives), then the CES production function, in order to calculate the trust score of the article, can be written as

$$T_{ar} = f(V,P,C) = (V^{\rho} + P^{\rho} + C^{\rho})^{\frac{1}{\rho}} \quad (7)$$

where

T_{ar}: Article Diversity; V: Diversity score; P: Depth score and C: Citation count.

Let m be the upper bound of trust score constrained by

$$w_1 V + w_2 P + w_2 C = m \quad (8)$$

where w_1, w_2 and w_3 are penalties on the diversity, depth score and citation count.

b. MOTIVATION OF THE CES FUNCTION:

We pose the following questions:

- What is the optimal strategy for choice of citation for the academic community and young researchers in particular, when encountering large number of articles?
- what is the most appropriate model of production, with optimal trust for cited articles when quality of published materials become increasingly questionable?

We shall show, empirically, that for constant elasticity of scale production functions, the trust is maximized at low levels of elasticity. This is one of the guiding principles behind utilizing the Constant Elasticity of Substitution (CES) functions that helps to derive conditions for trust maximization. This elasticity of substitution is constant for the CES production function. To be precise, the elasticity of substitution measures the percentage change in the factor ratio to the percentage change in the technical rate of substitution. Holding output remains fixed during the measurement process.

Case I: Linear production function:

ρ = 1 makes the production function assume the form: $y = K + L$, where capital and labor as inputs are perfect substitutes.

Case 2: Cobb-Douglas production function:

When ρ tends to 0, i.e. $\lim_{\rho \to 0,} y$, the isoquants of the CES production function appear similar to those of the Cobb-Douglas

production function. This can be shown in a variety of ways. The technical rate of substitution turns out to be most convenient. Standard isoquants are derived from the two inputs (in this case, imperfect substitutes).

Case 3: Leontief Production function:

As ρ tends to $-\infty$, i.e. $\lim_{\rho\to-\infty} y$, the isoquants become L-shaped, associated with inputs as perfect complements. The model proposed show that articles register maximum trust score when the elasticity of substitution is low, and close to 0.1. Flexibility of the CES function should also allow us to use more inputs for trust estimation. The estimation of optimal levels is theoretically consistent with the input parameters. A detailed analysis of the input vector and its role in trust score optimization for the academia is therefore imperative and timely.

c. The Problem of Trust Optimization

The problem of trust score maximization is perceived as: max $f(C, V, P)$ subject to m where m is needed as a constraint to bounded maxima. The following values of V, P and C, thus obtained, are the values for which the article has maximum trust score.

$$C^* = \frac{mw_1 \frac{1}{\rho-1}}{w_1^{\frac{\rho}{\rho-1}} + w_2^{\frac{\rho}{\rho-1}} + w_3^{\frac{\rho}{\rho-1}} + w_4^{\frac{\rho}{\rho-1}}} \quad (9)$$

$$V^* = \frac{mw_2 \frac{1}{\rho-1}}{w_1^{\frac{\rho}{\rho-1}} + w_2^{\frac{\rho}{\rho-1}} + w_3^{\frac{\rho}{\rho-1}} + w_4^{\frac{\rho}{\rho-1}}} \quad (10)$$

$$P^* = \frac{mw_2 \frac{1}{\rho-1}}{w_1^{\frac{\rho}{\rho-1}} + w_2^{\frac{\rho}{\rho-1}} + w_3^{\frac{\rho}{\rho-1}} + w_4^{\frac{\rho}{\rho-1}}} \quad (11)$$

These results are proved in Appendix B.

d. Curvature Characteristics of CES

We study curvature properties in order to to ascertain the global maxima/minima properties of CES, without which we cannot model the credibility metric and guarantee its maxima.

CES function models the trust score, y as:

$$y = (C^\rho + V^\rho)^{\frac{1}{\rho}}$$

$$\frac{\partial y}{\partial C} = \frac{1}{\rho}(C^\rho + V^\rho)^{\frac{1}{\rho}-1}\rho C^{\rho-1}$$

$$\frac{\partial y}{\partial V} = \frac{1}{\rho}(C^r + V^\rho)^{\frac{1}{\rho}-1}\rho V^{\rho-1}$$

$$\frac{\partial y}{\partial C\partial V} = \rho C^{\rho-1}V^{\rho-1}(C^\rho + V^\rho)^{\frac{1}{\rho}-2}$$

$$\frac{\partial y}{\partial V\partial C} = \rho C^{\rho-1}V^{\rho-1}(C^\rho + V^\rho)^{\frac{1}{\rho}-2}$$

$$\frac{\partial^2 y}{\partial^2 C} = \rho(C^\rho - 1)^2(C^\rho + V^\rho)^{\frac{1}{\rho}-2} + (\rho-1)C^{\rho-2}(C^\rho + V^\rho)^{\frac{1}{\rho}-1}$$

$$\frac{\partial^2 y}{\partial^2 V} = \rho(V^{\rho-1})^2(C^\rho + V^\rho)^{\frac{1}{\rho}-2} + (\rho-1)V^{\rho-2}(C^\rho + V^\rho)^{\frac{1}{\rho}-1}$$

Hessian Matrix

$$\begin{bmatrix} \rho(C^\rho-1)^2(C^\rho+V^\rho)^{\frac{1}{\rho}-2} + (\rho-1)C^{\rho-2}(C^\rho+V^\rho)^{\frac{1}{\rho}-1} & \rho C^{\rho-1}V^{\rho-1}(C^\rho+V^\rho)^{\frac{1}{\rho}-2} \\ \rho C^{\rho-1}L^{\rho-1}(C^\rho+V^\rho)^{\frac{1}{\rho}-2} & \rho(V^{\rho-1})^2(C^\rho+V^\rho)^{\frac{1}{\rho}-2} + (\rho-1)V^{\rho-2}(C^\rho+V^\rho)^{\frac{1}{\rho}-1} \end{bmatrix}$$

$$\Delta_1 = (C^\rho + V^\rho)^{\frac{1}{\rho}-1}C^{\rho-1}\left(\frac{\rho C^{\rho-1}}{C^\rho+V^\rho} + \frac{\rho-1}{C}\right)$$

As $C, V, \rho > 0$ $\Delta_1 > 0$;

$$\Delta_2 = \rho(\rho-1)C^\rho - 1)^2(V^{\rho-2})C^\rho + V^\rho)^{\frac{2}{\rho}-3} +$$

$$\rho(\rho-1)(V^{\rho-1})^2(C^{\rho-2})(C^\rho+V^\rho)^{\frac{2}{\rho}-3} + (\rho-1)^2(C^{\rho-2})(C^\rho+V^\rho)^{\frac{2}{\rho}-2}$$

$\Delta_2 \geq 0$ in case $\rho \geq 1$

As $\Delta_1 \geq 0$ and $\Delta_2 \geq 0$ in case $\rho \geq 1$. It will produce concave graph.

When $\rho < 1$, $\Delta_1 \geq 0$ and $\Delta_2 \geq 0$. It is neither concave or convex.

2. Positivity of Δ_1 of CES Hessian Matrix

Let us now explain the reason why Δ_1 from previous section is always positive.

Considering Δ_1 value again:

$$\Delta_1 = (C^\rho + V^\rho)^{\frac{1}{\rho}-1}C^{\rho-1}\left(\frac{\rho C^{\rho-1}}{C^\rho+V^\rho} + \frac{\rho-1}{C}\right)$$

Δ_1 will be negative if below two conditions are satisfied.

1) $\rho < 1$

2) $\frac{\rho-1}{C} \geq \frac{\rho C^{\rho-1}}{C^\rho+V^\rho}$

$$\frac{\rho-1}{C} \geq \frac{\rho C^{\rho-1}}{C^\rho+V^\rho}$$

$$\Rightarrow (\rho-1)C^\rho + V^\rho) \geq \rho C^\rho$$

$$\Rightarrow \frac{\rho-1}{\rho} \geq \frac{C^\rho}{C^\rho+V^\rho}$$

$$\Rightarrow 1-\frac{1}{\rho} \geq \frac{C^\rho}{C^\rho+V^\rho}$$

$$\Rightarrow 1-\frac{C^\rho}{C^\rho+V^\rho} \geq \frac{1}{\rho}$$

$$\Rightarrow \frac{C^\rho}{C^\rho+V^\rho} \geq \frac{1}{\rho}$$

$$\rho V^\rho \geq C^\rho+V^\rho$$

$$\Rightarrow (\rho-1) \geq \frac{C^\rho}{V^\rho}$$

$$\Rightarrow \rho-1 \geq \left(\frac{C}{V}\right)^\rho$$

As C, $V > 0$, hence $\left(\frac{C}{V}\right)^\rho$ will be always positive.

$$(\rho-1) > 0$$
$$\Rightarrow \rho > 1$$

Which is contradicting the first condition $\rho < 1$. Hence Δ_1 will be always positive.

3. Elasticity of Substitution

The theorem proves that marginal rate of change (σ) of elasticity is almost linear.

Theorem: The constant elasticity of substitution, ρ for the CES Production Function is approximated by

$$\sigma \cong 1+\rho \text{ where } 0 < \rho < 1 \tag{12}$$

Proof : CES function (Trust Score Model) is represented as

$$T_{ar} = (C^\rho + V^\rho)^{\frac{1}{\rho}} \tag{13}$$

Marginal rates of change are computed as

$$T_C = \frac{\partial T}{\partial C} = \frac{1}{\rho}(C^\rho+V^\rho)^{\frac{1}{\rho}-1} \cdot \rho C^{\rho-1} \tag{14}$$

$$T_V = \frac{\partial T}{\partial V} = \frac{1}{\rho}(C^\rho+V^\rho)^{\frac{1}{\rho}-1} \cdot \rho V^{\rho-1} \tag{15}$$

We need to define the rate at which citations may be substituted by the diversity. Let us define the elasticity as

$$\sigma(x) = \frac{xf'(x)}{f(x)} = \frac{\frac{df}{f}}{\frac{dx}{x}} \tag{16}$$

Change of Variable:

$$x = e^4 \qquad f(x) = e^v$$
$$u' = \frac{1}{x} \qquad v' = \frac{f'(x)}{f(x)}$$

such that

$$\frac{v'}{u'} = \frac{\frac{f'(x)}{f(x)}}{\frac{1}{x}} = \sigma(x) \tag{17}$$

Therefore,

$$\ln\left(\frac{T_V}{T_C}\right) = \ln\left(\frac{\frac{1}{\rho}(C^\rho+V^\rho)^{\frac{1}{\rho}-1} \cdot \rho V^{\rho-1}}{\frac{1}{\rho}(C^\rho+V^\rho)^{\frac{1}{\rho}-1} \cdot \rho C^{\rho-1}}\right)$$
$$= \ln\left(\frac{V}{C}\right)^{\rho-1}$$
$$= (\rho-1)\ln\left(\frac{V}{C}\right)$$
$$= (1-\rho)\ln\left(\frac{C}{V}\right)$$
$$\ln\left(\frac{C}{V}\right) = \frac{1}{1-\rho}\ln\left(\frac{T_V}{T_C}\right)$$
$$\sigma = \frac{1}{1-\rho}$$
$$\sigma = (1-\rho)^{-1}$$
$$= 1+\rho+\rho^2+\rho^3\ldots$$

Ignoring the higher order terms

$$\sigma \cong 1+\rho \tag{18}$$

4. Global Maxima For Trust Value Maximization

To ensure the maximum value of trust score, we have employed Gradient Ascent, which determines the optimum trust value when the diversity score and citation count is known. In other terms, the elasticities of citation count and diversity score is identified, for which trust value attains a maximum. Writing the CES function:

$$T_{ar} = (C^\rho + V^\rho)^{\frac{1}{\rho}}$$

Differentiating with respect to the elasticity of substitution, we get

$$\frac{\partial T_{ar}}{\partial \rho} = \frac{(C^\rho+V^\rho)^{\frac{1}{\rho}}}{\rho^2}\ln(C^\rho+V^\rho)$$
$$(C^\rho \ln C + V^\rho \ln V)$$

The partial derivative is used in gradient ascent method for trust value maximization.

Gradient Ascent Algorithm:

1. **procedure** GRADIENT ASCENT( )
2. $\frac{\partial T_{ar}}{\partial \rho} = \frac{(C^\rho+V^\rho)^{\frac{1}{\rho}}}{\rho^2}\ln(C^\rho+V^\rho)$
   $(C^\rho \ln C + V^\rho \ln V)$

3. **repeat**
4. $\rho_{n+1} \leftarrow \rho_n + \delta \frac{\partial T_{ar}}{\partial \rho}$
5. $\rho_n \leftarrow \rho_{n+1}$
6. **until** $(\rho_{n+1} > 0)$
7. **end procedure**

Using the above algorithm, the optimal values of elasticity and trust have been computed.

a. Stochastic Frontier

A production function will be called frontier when it gives the maximum possible output for a given set of inputs. All the production units of a frontier function will be fully efficient. Now, efficiency can be explained in two ways: technical and allocative.The technical efficiency can be further modeled by either deterministic or by stochastic frontier production function. The deterministic frontier model explains the shortfall from the frontier, which is the maximum output by technical inefficiency, whereas the stochastic model includes the random shocks to the frontier function.[28] There arises a need to address the stochastic nature of production function which is nothing but uncertainty or shock associated with trust value. The CES production frontier can be written as:

$$y = f(K,L)TE$$

where TE is the technical inefficiency, the ratio of observed output to maximum possible output. If TE=1, the trust achieves maximum value. This production frontier is deterministic as the entire deviation from maximum feasible output is attributed to technical inefficiency. It does not consider random shocks, which is not beyond control of production function. To address the random shocks, the production frontier function can be redefined as below:

$$y = f(K,L)TE\exp(v)$$

where v is the stochastic variable which defines the shocks, uncertainty, luck etc. Let us consider the linear logarithmic form of stochastic frontier production function.

$$\ln y = \frac{1}{\rho}\ln(C^{\rho} + V^{\rho}) + v - u \quad (19)$$

where y = output trust value
C = Citation count
V = Diversity score
v = random shocks
u = technical inefficiency

$$\rho = n \quad (20)$$

CRS: n = 1* Constant returns to scale*
IRS: n > 1* Increasing returns to scale*
DRS: n < 1* Decreasing returns to scale*

b. Stochastic Gradient Ascent

While using the partial derivative equation $\frac{\partial T_{ar}}{\partial \rho}$, it is often observed that either the equation is not solvable or, if it can be solved, it suffers from slow convergence. Stochastic Gradient Ascent is a well known method and used in many different fields to achieve optimal value. The algorithm can be used to estimate maximum trust score, ensuring quick convergence of elasticity. This was done for two reasons: to be able to break free from the in-built library functions, and to devise a sensitive method which would mitigate oscillatory nature of Newton-like methods around the local minima/maxima. There are many methods which use gradient search including the one proposed by Newton. Although theoretically sound, algorithmic implementations of most of these methods faces convergence issues in real time due to the oscillatory nature. Stochastic Gradient Descent was used to find the minimal value of a multivariate function, when the input parameters are known. We tried to identify the elasticity for depth, diversity and citations received. We do this to compute the trust score for which the maximum value is attained under certain constraints. We have employed a modified version of the descent, a Stochastic Gradient Ascent algorithm, to calculate the optimum trust score and the elasticity values ρ. As opposed to the conventional Gradient Ascent/Descent method, where the gradient is computed only once, stochastic version recomputes the gradient for each iteration and updates the elasticity values. Theoretical convergence, guaranteed otherwise in the conventional method, is sometimes slow to achieve. Stochastic variant of the method speeds up the convergence, justifying its use in the context of the problem (the size of data, i.e. the number of articles is increasing every day).

Output elasticity of CES Trust function is the accentual change in the output in response to a change in the levels any of the inputs. ρ is the output elasticity of depth, diversity and citations. Accuracy of ρ values is crucial in deciding the right combination for the optimal trust score, where different approaches are analyzed before arriving at final decision.

c. Computing Elasticity via Stochastic Gradient Ascent

Gradient Ascent algorithm is used to find the values of ρ. Gradient Ascent is an optimization algorithm used for finding the local maximum of a function. Given a scalar function $F(x)$, gradient ascent finds the $\max_x F(x)$ by following slope of the function $\frac{\partial F}{\partial x}$. This algorithm selects initial values for the parameter x and iterates to find the new values of x which maximizes $F(x)$, the trust score. Maximum of a function $F(x)$ is computed by iterating through the following step,

$$x_{n+1} \leftarrow x_n + \chi \frac{\partial F}{\partial x}, \quad (21)$$

where x_n is an initial value of x, x_{n+1} the new value of x, $\frac{\partial F}{\partial x}$ is the slope of function $Y = F(x)$ and c denotes the step size, which is greater than 0 and forces the algorithm make a small jump (descent or ascent algorithms are trained to make small jumps in the direction of the new update). Stochastic variant thus mitigates the oscillating nature of the global optima, a frequent malaise in the conventional Gradient Ascent/Descent and Newton-like methods, such as fmincon used in.[27] At this point of time, without further evidence of recorded/measured parameters, it may not be prudent to scale up the trust score model by including more parameters other than the ones mentioned already. But if it ever becomes a necessity (to utilize more parameters), the algorithm will come in handy and multiple optimal elasticity values may be computed fairly easily.

SGA algorithm

- Choose an initial vector of parameters ρ and randomly select learning rate δ
- $\frac{\partial T_{ar}}{\partial \rho} = \frac{(C^\rho + V^\rho)^{\frac{1}{\rho}}}{\rho^2} \ln(C^\rho + V^\rho)$
  $(C^\rho \ln C + V^\rho \ln V)$
- Repeat
- Rather than calculating the gradient once, which happens in conventional gradient algorithm, here for each iteration the gradient being recalculated and added to the updated ρ
- $\rho_{n+1} \leftarrow \rho_n + \delta \frac{\partial T}{\partial \rho}$
- $\rho_n \leftarrow \rho_{n+1}$

******The iteration will continue till ρ is greater than 0 ******

- Until $\rho_{n+1} > 0$
  Stop when the convergence conditions are met
- Calculate the Trust Score by putting ρ in the trust score function.

5. Additive Trust Score: A Novel Model

Different input parameters should not have the same elasticity as diversity and citations are related non linearly. This may create a problem while using CES model. Hence we propose a new method, where diversity, citation and depth are endowed with different elasticity of substitution. We call the proposed model as Additive Trust Model where trust value for an article is modeled as the following: $T_v = X^\alpha + Y^\beta + Z^\gamma$; where:

- T_v: is the Trust Value of the article
- X: is the Diversity score of the article
- Y: is the Depth Score
- Z: is the Citation count of the article
- α, β, γ are the elasticities of the modeled growth function. The reason this growth function is chosen, for the first time, to the best of the authors' knowledge are inadequacy of the other models such as Cobb Douglas (product model)[25],[27] and CES production functions.[26] Cobb Douglas model will render the overall trust score as zero if any of the input parameters is not known/measured and therefore assumed as zero. CES production function assumes the elasticity associated with each input parameter identical. This assumption is not realistic enough unless the parameters themselves are statistically dependent. As desired, the method obtains a global maxima by satisfying the conditions of concavity as proved in the theorem below.

Theorem: Conditions for concavity: Given the input parameters diversity, x; depth, 1-x; and citations y, the trust score model,

$$y = x^\alpha + (1-x)^\beta + y^\gamma$$

will satisfy conditions for concavity, i.e., it will have a global maxima if the following conditions hold: $0 \le \alpha \le 1$; $0 \le \gamma \le 1$; $\beta \ge 1$; $\alpha + \beta + \gamma \ge 1$

Proof: Given the model,

$$f = x^\alpha + (1-x)^\beta + y^\gamma$$

The first order conditions are:

$$\frac{\partial f}{\partial x} = \alpha x^{\alpha-1} - \beta(1-x)^{\beta-1}$$

$$\frac{\partial^2 f}{\partial^2 x} = \alpha(\alpha-1)x^{\alpha-2} + \beta(\beta-1)(1-x)^{\beta-2}$$

$$\frac{\partial f}{\partial y} = \gamma y^{(\gamma-1)}$$

$$\frac{\partial^2 f}{\partial^2 y} = \gamma(\gamma-1)y^{(\gamma-2)}$$

$$\frac{\partial^2 f}{\partial x \partial y} = 0; \qquad \frac{\partial^2 f}{\partial y \partial x} = 0$$

Constructing the Hessian Matrix:

$$\begin{bmatrix} \alpha(\alpha-1)x^{\alpha-2} + \beta(\beta-1)(1-x)^{\beta-2} & 0 \\ 0 & \gamma(\gamma-1)y^{(\gamma-2)} \end{bmatrix}$$

$$\Delta_1 = \alpha(\alpha-1)x^{\alpha-2} + \beta(\beta-1)(1-x)^{\beta-2}$$

$$\Delta_1 = \gamma(\gamma-1)y^{(\gamma-2)}$$

$$\Delta_2 = [\alpha\gamma(\alpha-1)(\gamma-1)x^{(\alpha-2)}y^{(\gamma-2)}] + [\beta\gamma(\beta-1)(\gamma-1)(1-x)^{(\beta-2)}y^{(\gamma-2)}]$$

The conditions for the functions to be concave are $\Delta_1 \le 0$ and $\Delta_2 \ge 0$. Δ_1 is less than 0 when $\alpha(\alpha-1)x^{\alpha-2} + \beta(\beta-1)(1-x)^{\beta-2}$ is less than 0 and $\gamma(\gamma-1)y^{(\gamma-2)}$ is also less than 0. Lets examine different cases to see if both the above conditions meet.

Case I: when $\alpha(\alpha-1)x^{\alpha-2} \leq 0$ and $\beta(\beta-x)^{\beta-2} \leq 0$. Both the terms can be made negative by considering just $(\alpha-1) \leq 0$ and $(\beta-1) \leq 0$; since other factors α and β are elasticities, and can never be negative, and $x \geq 0$ and $(1-x) \geq 0$ holds true for various values of diversities. From this, we obtain: $\alpha \geq 0$, $\beta \geq 0$, $\alpha \leq 1$, $\beta \leq 1$.

Case II: when $\alpha(\alpha-1)x^{\alpha-2} \geq 0$, $\beta(\beta-x)^{\beta-2}0$ and $(\beta-1) \geq (\alpha-1)$. This means that the first additive term is positive, second is negative and second is larger than first. These condition holds iff $(\alpha-1) \geq 0$ and $(\beta-1) \leq 0$ is true, implying $\alpha \geq 1$ and $\beta \leq 1$. But $(\beta-1) \geq (\alpha-1)$ denotes $\beta \geq \alpha$, which can not be true by any means. Hence case II can be discarded.

Case III: when $\alpha(\alpha-1)x^{\alpha-2} \leq 0$, $\beta(\beta-x)^{\beta-2} \geq 0$ and $(\alpha-1) \geq (\beta-1)$. These conditions can be true iff $(\alpha-1) \leq 0$ and $(\beta-1) \geq 0$ holds. This shows the values of $\alpha \leq 1$, $\beta \geq 1$. And also, $(\alpha-1) \geq (\beta-1)$ indicates $\alpha \geq \beta$. All three conditions of case III cannot hold simultaneously and hence, should also be discarded.

Having considered above three cases, it can easily be shown that $\gamma(\gamma-1)\gamma^{(\gamma-2)} \leq 0$ is true for $(\gamma \leq 1) \leq 0$ i.e. for $\gamma \leq 1$. Thus, for Δ_1 to be negative, the following should hold $0 \leq \alpha \leq 1$; $0 \leq \beta \leq 1$; $0 \leq \gamma \leq 1$. Concavity of a function also demands Δ_2 to be positive, which means $[\alpha\gamma(\alpha-1)(\gamma-1)x^{(\alpha-2)}\gamma^{(\gamma-2)}] + [\beta\gamma(\beta-1)(\gamma-1)(1-x)^{(\beta-2)}\gamma^{(\gamma-2)}]$ must be positive. The first additive term of this equation is positive because $(\alpha-1)(\gamma-1)$ are both negative. Likewise, second additive term is positive since $(\beta-1)(\gamma-1)$ terms are negative. The added $(1-x)$ is apparently positive. Thus, Additive trust model satisfies both the conditions for concavity (i.e. Δ_1 being negative and Δ_2 positive) and can attain global maxima for the above stated conditions of elasticities. Figure 8 shows a maximum value of trust score reached at certain values of α and β; $\alpha+\beta+\gamma \geq 1$

Remark: Concavity conditions guarantee global maxima enabling the Trust Model to achieve maximum trust score for each article. Citations are normalized to arrest asymptotic growth at low values of elasticity, thus mitigating one problematic aspect of the additive trust model.

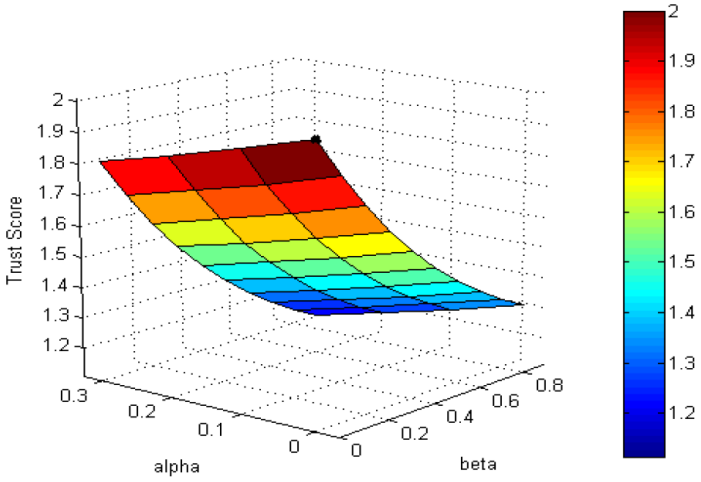


Figure 8: 3D plot for trust score vs. Elasticity : The plot shows different trust values at varying α and β values

MATHUKAMALLI VIDYASAGAR

This is not a thumb rule or generic guideline. We pick authors, based on intuition and then investigate the quality metrics of the article written by those authors. We match our model with the perception. But more often then not, the perception is limited to small communities like Partial Differential Equation (Tao), Control system (Vidyasagar). Our effort is to transcend the knowledge discovery to researchers in other domains. This is in agreement with the general principals elucidated in Motivation and Objective section.

1. Analysis: Dr. Mathukumalli Vidyasagar

Authors' use real-time data of Professor Vidyasagar's articles for their research by computing diversity, depth and citations from recursive reference network. Vidyasagar is an Indian-American scholar who is a leading control theorist. At present, he focuses in the area of compressed sensing and in applying ideas from machine learning to problems in computational biology with emphasis on cancer. He became a fellow of the Royal Society in 2012 and won the Rufus Oldenburger Medal the same year. In 2017, he was accepted into the International Federation of Automatic Control. He also boasts two as his Erds number and three as Einstein number.

The reason for choosing Mathukumalli Vidyasagar doesn't beg detailed explanation. Along with being a highly cited author, he was named as 125 "People of Impact" during the 125th anniversary of the Department of Electrical Engineering, University of Wisconsin and his extent of scholarship is widely regarded. His profile provides access to articles useful for breath-wise and depth-wise analysis.

a. Trust score Calculation

This subsection demonstrate the implementation of CES Production Function to compute optimal trust score for Vidyasagar's articles. The input parameters are diversity (V), depth (P) and citations(C). Table 2 (in Appendix C) shows summary of his articles which were scraped and processed to extract diversity score, citation count and reference count. Depth for the articles can be computed by

$$P_n = 1 - V_n \tag{22}$$

All input parameters are fed into CES function represented by equation

$$T_{ar} = f(V,P,C) = (V^{\rho} + P^{\rho} + C^{\rho})^{\frac{1}{\rho}} \tag{23}$$

where

T_{ar}: Article Diversity

V: Diversity score

P: Depth score

C : Citation count

The trust score is studied for various values of ρ by using fmin-con function in MATLAB. Three constraints applied to the function are:

$\rho < 1$

$\rho = 1$

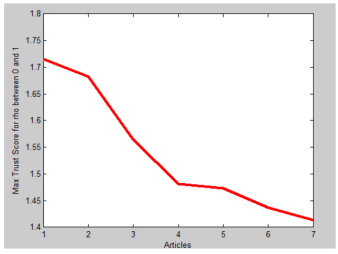
$\rho > 1$

Case 1: $\rho \leq 1$ Applying the constraints $\rho \leq 1$ and $\rho \geq 0$ to the function, $f = (V^{\rho} + P^{\rho} + C^{\rho})^{\frac{1}{\rho}}$ and using fmincon function, the values of elasticity are obtained for which the trust value is maximum. For a specific case ρ = 0.9, the trust values of all articles is shown in the Figure 9(a). The figure reflects a decrease in score from article 1 to 7. It can be validated from the table data that citations and diversity both decreases in the same pattern. Arguably, with decrease in diversity, the depth increases. The overall effect on trust is a proportionate increase in its value.

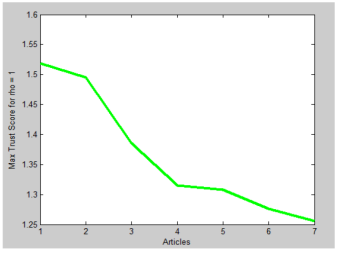
Case 2: For ρ = 1, the function finds the trust score for all articles and plot is shown in figure 9(b). The value of ρ is directly kept into the function, along with other input parameters to calculate Trust. The plot shows a sudden descent of trust for article 3 and 4. which is credited to the drop in citation value for these articles, irrespective of the diversity and depth values being close.

Case 3: For $\rho > 1$ and $\rho < 2$, the optimum value of ρ is 1.1, as the Trust is maximum at this value for all articles. The plot shown in Figure 9(c) reflects articles and their trust score. Despite a decrease in citations from article 1 to 2, an increase in trust value is evident from figure and this is attributed to the increase in diversity for article 2. This implies the inclusion of diversity as a crucial ingredient in computing trust score for articles.

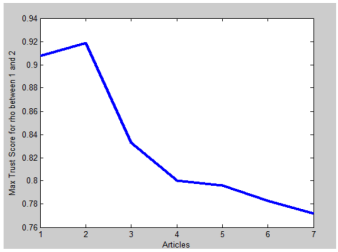
Various plots are drawn and examined between Trust score and values of diversity, depth and ρ. When Trust score is plotted against ρ values ranging between 0.1 to 0.9, the largest value is seen at ρ = 0.1. Correspondingly, for ρ = 1.1, the largest trust score is obtained, as indicated in Figure 10(b). Figure 11, a 3D surface plot for Trust score, diversity and citations, shows that for a specific value of diversity and citation, a maximum trust score (global maxima) is achieved. Table 3 (in Appendix C) contains an article summary with trust score values computed by using "Additive Trust Model". The equation is represented as: $T_v = X^{\alpha} + Y^{\beta} + Z^{\gamma}$; where T_v: is the Trust Value of the article; X: is the Diversity score of the article; Y: is the Depth Score; Z: is the Citation count of the article. Figure 12



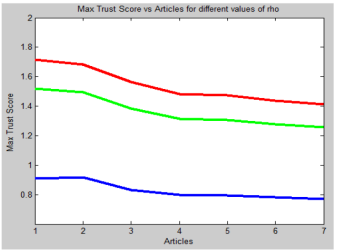
(a) Articles vs Trust score for $\rho < 1$



(b) Articles vs Trust Score $\rho = 1$: graded decrease, if we choose articles and authors carefully, random fluctuation in TV is almost non-existent.

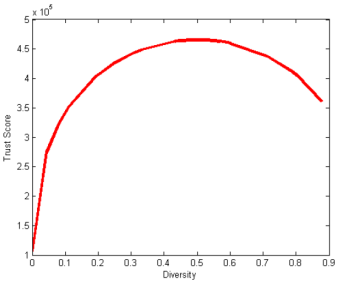


(c) Articles vs Trust Score $\rho > 1$: A slight increase in trust value in seen, which is attributed to increase in div, this fortifies the fact that diversity is as important as citations received by an article and thus must be attributed as a key quality parameter



(d) Articles vs Trust Score for all three cases combined

Figure 9: Articles vs rho



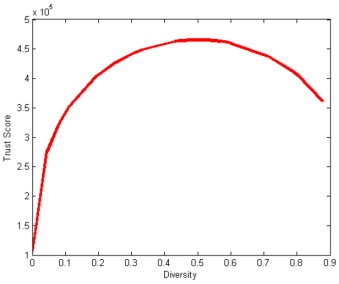


Figure 10: Trust Score vs Elasticity.

The Trust score graph of the articles mentioned in table 3 (in Appendix C), is indicated in Figure 13. The plot is drawn between trust score and articles arranged in chronological order. It establishes the fact that trust value of genuine and trustworthy authors does not deflect or deviate and trust score tends to remain stable with time.

CASE STUDY: TERRENCE TAO REFERENCE NETWORK ANALYSIS

Terence Tao, a mathematician of Australian-American descent has worked in various areas of mathematics. His most notable contribution is in the areas of harmonic analysis and partial differential equations. He is currently the James and Carol

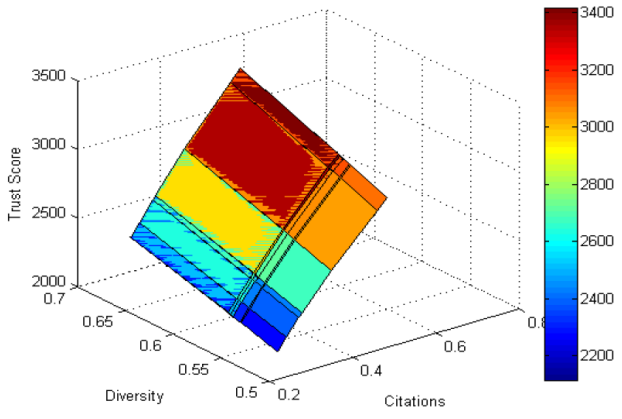


Figure 11: 3D plot for Diversity Citations and Trust Score: the plot shows how trust score reaches its largest value at certain values of divand citations.

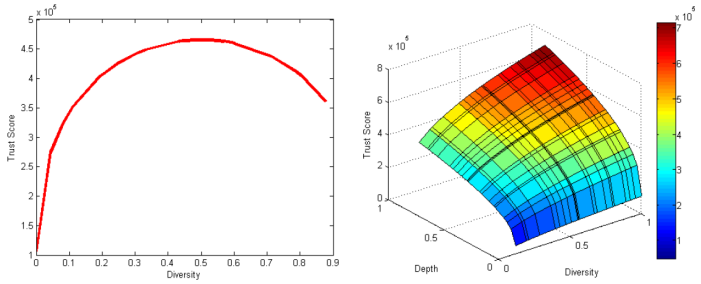


Figure 12: Analysis of Trust Score, Diversity Depth and Citations.

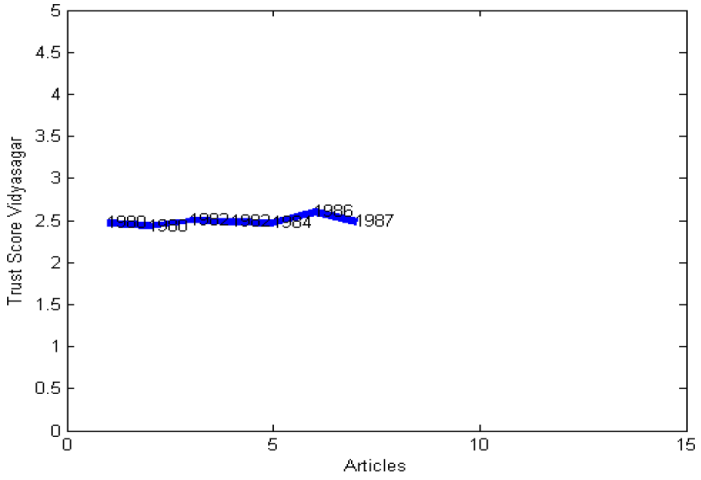


Figure 13: 2D Plot-Trust score vs articles of Prof. Vidyasagar. The steady trend in Trust score is observed.

Collins chair in Mathematics at the University of California, Los Angeles. Tao received 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 2 (Table 4 in Appendix C). Next we will explain the application of graph theory algorithms and text analysis algorithms on the reference network of these articles and the findings.

1. Graph theory based analysis

Figure 14(a) shows the complete reference network of three levels for the first article from Table 4.

The red largest size node is the root node which corresponding to a first article in Table 4 (in Appendix C). 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 root node. However, the figure 14(b) shows the strongly connected component of this complete graph where each node displays year of publication as its property A graph is said to be strongly connected if every vertex can be reached from every other vertex. Tarjan showed that, the strongly connected components of an arbitrary directed graph form a partition into subgraphs that are also strongly connected.[13]

Figure 14(c) 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. 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.

2. Text analysis on keywords

a. Latent Dirichlet Allocation

Latent Dirichlet Algorithm (LDA) is a statistical model designed to extract topics from words from documents. LDA was reported by David Blei *et al.*[6] in 2003. LDA assumes the following for each document w in a corpus D:

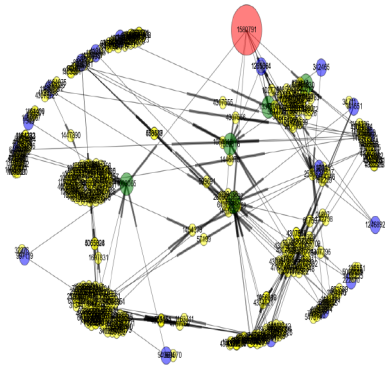
- N words following Poisson distribution(e).
- θ following Dirichlet distribution, Dir(α).
- For each of the N words w_n:

– Choose a topic z_n following a Multinomial distribution, Multinomial(θ).

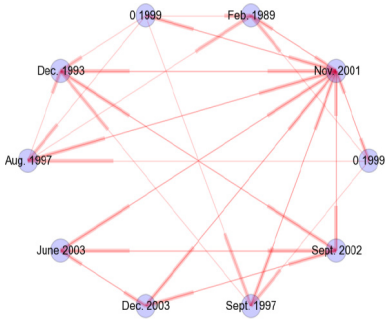
– Choose a word w_n from $p(w_n — z_n, \beta)$, a multinomial probability conditioned on the topic zn.

Given the parameters α and β, the joint distribution defines contribution of a topic mixture θ, a set of N topics z, and a set of N words w in the following way:

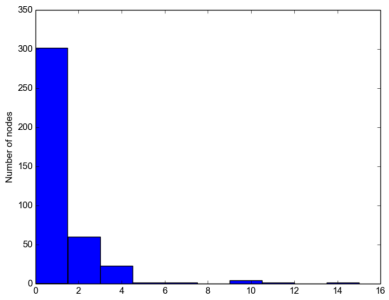
$$p(\theta, z, w \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta),$$



(a) Terrence Tao reference network



(b) TTao's strongly connected biggest component



(b) Number of articles Vs In degree count

Figure 14: Terrence Tao Reference Network Analysis Using Graph Theory

where $p(z_n—\theta)$ is θ_i for the unique i such that $z_n^i = 1$.

In other words, LDA assumes that documents are produced from a mixture of topics. Those topics then generate words based on their probability distribution. LDA analyzes a given document in the following way:

- It determines the number of words in a document.
- It determines the mixture of topics in that document and gives each topic a priori for the complete document.
- Using each topic's multinomial distribution, LDA produces words to fill the document's work slots, i.e, based on the priori, ratio of the total words in the document is filled with a particular topic.

With the knowledge of above composition of any document, LDA backtracks and tries to figure out what topics would create such a document in a first place.

3. Using LDA

We have used gensim[10] which is the most robust, efficient and hassle-free piece of software to realize unsupervised semantic modeling from plain text. LDA was implemented in the following manner:

- **Importing Documents:** Text from documents to be analyzed was extracted and stored for easy processing.
- **Cleaning Documents:** Documents are cleaned for accurate and smoother processing. This was achieved by 3 subroutines:
  - **Tokenization:** Text from documents was converted in it's atomic elements or words.
  - **Stopping:** Words which were, more so, meaningless are removed.
  - **Stemming:** Words, equivalent in meaning, are merged.
- **Construction of Document term-matrix:** The resultant of cleaning phase is a tokenized, stopped and stemmed list of words for each document. These steps assign a unique ID to each unique token in the list while also collecting word counts and relevant statistics. The resultant is matrix where each element is an ordered (x_1, y_1) such that x_1 is an unique ID while y_1 is frequency count. This is a corpus of words for all documents used for further processing.
- **Applying the LDA model:** Subsequently, a document-term matrix or a corpus was obtained, allowing us to generate an LDA model. The model is achieved using the LDA class of gensim library and considers the following three parameters:
  - **num_topics:** An LDA required user to determine the number of topics to be generated.
  - **id2word:** This is the hash-map mapping each ID found in a corpus to it's string.
  - **passes:** This is an optional parameter specifying the number of laps model will take through the corpus.
- **Obtaining results:** After analyzing, the LDA model returns a list of ordered pairs representing a topic found, along with it's presence (or percentage composition) in the document.

We obtained 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 15 shows output of this process. As shown in the figure, 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 3, topics listed in Figure 15 are discovered. Finally, all the JSON files (data set) are processed to create a list of all 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 done 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 15.

Algorithm 3 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 f iles)
for i ∈ path_to_files do
    for j ∈ i do
        kwdList.append(j['keywords'])
    end for
    kwdCorpus.extend(kwdList)
    end for
            ▷ use corpora function from gensim library
    dictionary ← corpora.Dictionary(kwdCorpus) ▷
  convert the dictionary to a bag of words corpus
    for text ∈ kwdCorpus do
        corpus.append(dictionary.doc2bow(text))
    end for
   lda ← LdaModel(corpus, num_topics = 5; id2word =
   dictionary)
return lda.showtopics( )
end procedure
```

a. 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 article DOI : 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.[11] Cosine similarity distance metric has been employed in generating this clustering. We have used various python packages from Scikit-learn[9] 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 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 16(b) shows the output from algorithm 5.

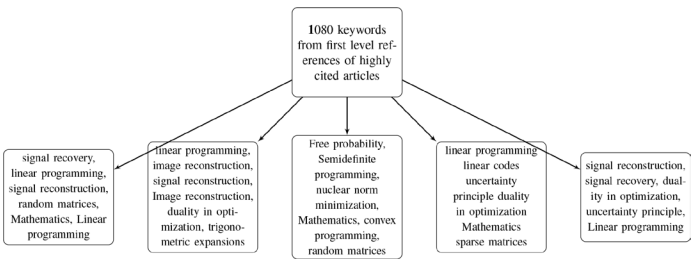


Figure 15: LDA output

4. Keyword frequency histogram of all the articles

For Table 4, (Appendix C) 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 4 are extracted in one list and all the keywords for 2 levels of reference nesting were extracted into another list. 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 16(a). There were total 31823 keywords from 1531 articles.

5. Trust Score for Articles of Terence Tao

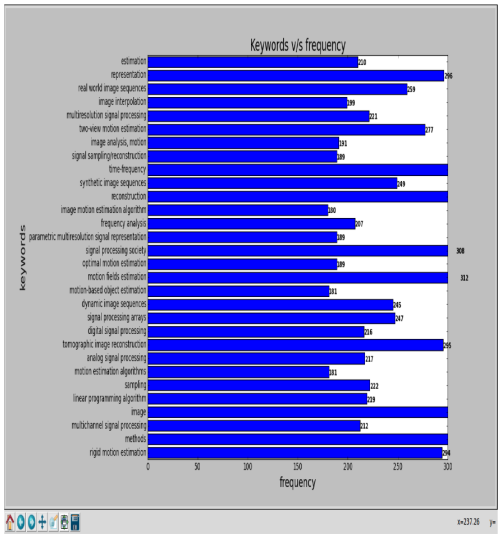
Two articles authored by Dr Tao and its applicable information like citation count, reference count (up to level 4) is extracted from RREF network and shown in table 5 (Appendix C). Diversity and depth is computed by performing LDA and cosine similarity on article keywords (section 5). Trust score associated with each article is quantified using Additive trust model. Input parameters are loaded and run on matlab to obtain elasticity and final trust score for each article, reflected in table 6 (Appendix C). Figure 17 shows year wise trust score for both articles.

Algorithm 4 Plotting Histogram for most frequent and similar keywords from Terence Tao data set

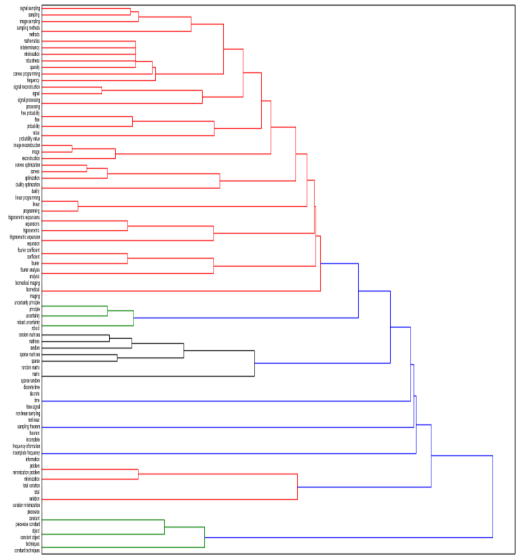
Input: root keyword list root list,referenced keyword list reference list, freq dict for keywords in referenced list.
Output: Keyword v/s frequency graph for top 30 most frequent and similar keywords in referenced articles
procedure CREATE_FREQ_DICT(*root_list; ref_list; freq_ dict*)
    count ← 0
    for *keyword*_1 ∈ *root_list* **do**
        if *keyword*_1*notin f req*_dict **then**
            f req_dict[*keyword*_1] ← 1
        else
            for *key*, *value in f req_dict* **do**
                score ← *Cosine_Similarity*(*key*,*keyword*_2)
                if *score* > 0.7 **then**
                    count ← *f req_dict*[*key*]
                    count ← *count* +1
                    f req_dict[*key*] ← *count*
                else
                    f req_dict[*keyword*_2] ← 1
                end if
            end for
        end if
    end for
    return f req dict
end procedure
procedure SORTED_FREQUENCY_DICT(*f req_dict*)
    sorted_list[] *sorted*(*f req_dict*[])
    for *keywords in sorted_list*[: 30 :] **do**:
        sorted_new_list[] ← *keywords*
        sorted_f requency_dict ← *dict*(*sorted_new_list*[])
    end for
    return *sorted_f requency_dict*
end procedure

Algorithm 5 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 ← *T f id fVectorizer*( )
    X ← *vectorizer*. *f it_transform*(*KwdList*)
    C ← 1 – *cosine_similarity*(*X.T*) ▷ define the linkage matrix using ward clustering
    linkage_matrix_ward(*C*)
    ax = *dendrogram*(*linkage_matrix*) **return** *ax*
end procedure



(a) Keywords Vs Frequency bar chart



(a) Agglomerative clustering

Figure 16: Terrence Tao Article processing

DISCUSSION AND FUTURE WORK

Article trust value is not JUST the merit of the article quantified but it does offer some insight into the credibility of the references listed in the article, and therefore makes it worthwhile to explore those references.

There has been a shift of focus in research from Journal level to Article level metrics. Traditional metrics like Article Influence Score and PLoS Article Level Metrics are subject to doubts as they are based on citation and may be biased or may not represent the quality of articles faithfully. Even if citations are not manipulated, they may not reflect the groundwork laid before the article is published. They may be inconsistent and credibility of references cited in articles is difficult to

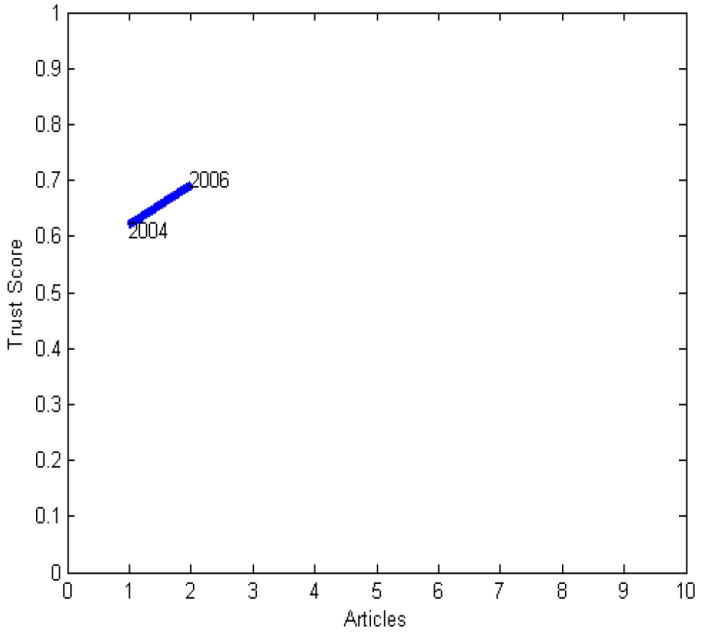


Figure 17: Terrence Tao Trust Score: Trust score increases, expectedly.

judge. It is important to understand whether the list of references is an accurate representation of the background preparation or some of the references are just a gesture of good faith. Some articles in niche areas may have very few references. This should not dampen the credibility of the article or the authors. However, some articles may have a very long list of references. But, it is not prudent to rule out the probability of articles citing references under coercion or camaraderie or copious nature. If a young researcher decides to mine all listed references of an article useful to him, it would be disheartening for him/her to be convinced about the futility of such exercise. The credibility of references and importance of reference study is extremely critical for knowledge dissemination and shaping future of young researchers. Trust Score model of an article wishes to serve the purpose by hypothesizing that : If an article has received and continues to receive reasonable number of citations in proportion to diversity and/or depth score, the article and the authors are trustworthy and mining references of these articles makes sense. The mining would actually provide some indicators of reliability/trust for the article (billions of such articles) and eventually for the authors. Therefore, trust score, a novel metric, intends to capture the credibility of reference literature and thereby may establish the connection between diversity, depth and trust score of the article. Final trust score is a model based on diversity, depth and citations received, the citations are devoid of self-citations, cross citations etc. This approach is fresh and has not been exploited by any of the standard services such as Google scholar, semantic scholar researchgate etc.

Future work: Authors intend to build a corpus of trustworthy articles and very importantly, a list of credible authors. Some may express their opinion/perception about leading authors in different fields, a perception based list. These two lists then may be matched to vindicate the efficacy of the proposed model. On the microlevel, this has already been done on two authors, Vidyasagar and Terrence Tao, and it has been found that our model matches perception. Future work also includes computation of trust value of authors and identifying journals where they publish. A new metric, Journal Index of Scholastic Reliability (JISR), derived from Journal trust value, JTV will be brought into the system. JIV can be computed by averaging the depth, diversity score and citations of all articles of a journal in one year. This score will be computed for multiple years and any growth will be tracked, which will be a symbol of a journal's trustworthiness. JISR will be formulated as a combination of JIV and fair fractional scholastic contribution (FSC). Raw FSC will be computed by the fraction of articles contributed by the listed scholars and the total number of articles published by a journal annually. Fair FSV shall be computed by subtracting the associated cost of publishing in a journal from the raw FSV.

Suppose journal A publishes good papers (measured by citations), has made a long-standing contribution to the subject (number of years in operation, and/or from a place of repute measured by some index, say from MIT ranked as 1), and has spread (data about subscriptions by countries, number of institutions, etc). There are B....Z journals which are ranked according to this criterion (it may also be available already). Now consider an author Alpha. He/she considers publishing a paper. To get the maximum visibility in the field and gather maximum reputation/impact, he/she chooses A as a possible outlet. If the paper gets accepted the person becomes famous. The acceptance would mean that the journal considers this paper worthy of their reputation. If the paper does not get published, Alpha goes down the rank of the journals in that order. Question is, when a reader approaches this paper if published in the top outlet, it must be because the journal serves the first point of attraction. If the same paper was published in journal Z, it would not attract adequate attention. This means that the character of the journal and the character of the paper must follow 'positive assortative matching' in line with papers which test Gary Becker's theory of assortative matching.[24] If it were negative assortative matching, it would cause loss of welfare (i.e. less value addition to the subject because not many people would read it, and therefore, not many people would work on it and therefore less knowledge creation [as a measure of welfare] would be the outcome). If the article was befitting for the journal, it would draw citations; several papers in great journals do not get as many citations!

There is a cost associated with finding the right match for Alpha's paper. It could be search cost, i.e., I do not know what journal will fit my paper best, so I have to search, or typical cost of submission. This might cause negative assortative matching in the end. The following may be pertinent to ask:

Research question: Do journal articles satisfy positive assortative matching?

We intend to pose another research question: Does an author with high SES have high self citations? In other words, we intend to establish negative correlation between SES and fraction of self citations over time. Future work would also focus on building nested article and author trust value and a trust graph! We intend to perform a detailed factor analysis of citation corpus as well.[16]

CONCLUSION

Bibliometricians often want to understand the pattern in which science and knowledge grows. One 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 have been 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 Networks on IEEE Journal's articles. 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. Betweenness 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 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 data set 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 and depth of background reading indicative of the *versatility and 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. The website developed for this purpose, http://sahascibase.org/rref[1], gives an elegant visual account.

Creating knowledge network and evaluating qualities of knowledge network is a very complex task since blanket rules don't exist. It is not available for free in the public domain. Citations network are complex and not free from 'coterie' network citations and copious citations. Argument may be presented in defense of such practices, particularly for niche areas and with limited audience. However, knowledge metrics and rich indicators proposed in the manuscript may nullify such defensive arguments. Cornerstone of the proposed approach is to build and validate rich indicators of quality such as diversity and depth leading to credibility of authors and the nested references used in their work. Our work relies on analyzing and quantifying the range of background preparation od scholars leading to RREF. Moreover, RREF produces astonishing insights and leads to related studies such as depth, citation, trust score- very novel characterizations of scholastic and article quality. Semantic networks don't investigate quality from this perspective. This renders RREF uniqueness and makes our effort different from semantic network and knowledge network. We established diversity as a credible metric.

REFERENCES

1. Scibase "http://sahascibase.org/".
2. IEEE - WEBSITE. "https://drsaraheaton.wordpress.com/2013/10/18/ whats-the-difference-between-a-citation-and-a-reference".
3. Webpage. "http://www.infotoday.com/online/may12/Belter-Visualizing-Networks-of-Scientific-Research.shtml".
4. Whitepaper. "http://docplayer.net/14505628-Whitepaper-a-guide-to-evaluating-research-performance-scientific.html".
5. Matthew L. Wallace and Vincent Larivire and Yves Gingras. A small world of citations? The influence of collaboration networks on citation practices. Eprint-arXiv:1107.5469, Doi:10.1371/journal.pone.0033339.
6. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research,* 3(Jan):993-1022,2003.
7. Website A Comprehensive Assessment of Impact with Article-Level Metrics (ALMs). "https://www.plos.org/article-level-metrics".
8. IEEE Xplore. "http://ieeexplore.ieee.org/Xplore/home.jsp". [Online; accessed 16-September-2016].
9. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blon-del, 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.
10. R. Rehurekand P.Sojka. Software Frameworkfor 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.
11. Wei. Zhang. Graph Degree Linkage: Agglomerative Clustering on a Directed Graph. *Journal of machine Learning research*,pages 428-441,2012.
12. Gabow Harold. Finding long paths, cycles and circuits. *International Symposium on Algorithms and Computation.*
13. Robert Tarjan. Depth-first search and linear graph algorithms. *SIAM Journal on Computing*,1972.
14. Arthur B. Kahn Topological sorting of large networks. *Communications of the ACM*,1962.
15. Moody's Investors Service. "https://www.moodys.com/sites///products/ProductAttachments/CDOGlossary.pdf" [Online; accessed 04-June-2017].
16. Jie Yang, and Abyuday Mandal. D-optimal Factorial Designs under Generalized Linear Models. *Journal of Communications in Statistics - Simulation and Computation*, Volume 44, pages:2264-2277,2015.
17. Article Influence and Eigenfactor ""http://journalinsights.elsevier.com/journals/0004-3702/article-influence" [Online; accessed 28-June-2017].
18. George A. Akerlof. The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of* Economics, Volume 84, pages: 488,1970.
19. Michael Spence. Job Market Signaling. *The Quarterly Journal of* Economics, Volume 87, pages:355,1973.
20. Hamid Beladi, and Saibal Kar. Skilled and Unskilled Immigrants and Entrepreneur-ship in a Developed Country. *Review of Development Economics,* Volume 19, pages:666-682, 2015
21. Saibal Kar and Hamid Beladi. A Model of Smuggling and Trafficking of Illegal Im-migrantswith aHostCountryPolicy. *ReviewofDevelopmentEconomics*,Volume21, pages:698-712, 2017.
22. Eliakim Katz and Oded Stark. International Migration Under Asymmetric Information. *The Economic Journal,* Volume 97, pages:718, 1987.
23. K. J. Arrow, H. B. Chenery, B. S. Minhas and R. M. Solow. Capital-Labor Substitution and Economic Efficiency. *The Review of Economics and Statistics,* Volume 43, pages: 225,1961.
24. Siow, A. Testing Beckers theory of positive assortative matching. *Journal of Labor Economics,* Volume 33, pages:409-441,2015.
25. Gouri Ginde, Snehanshu Saha, Archana Mathur and Sukrit Venkatagiri, Sujith Vadakkepat, Anand Narasimhamurthyand B.S. DayaSagar. ScientoBASE: A Framework and Model for Computing Scholastic Indicators of non-local influence of Journals via Native Data Acquisition algorithms. *Scientometrics,* Volume 107, pages: 151,2016.
26. Jyotirmoy Sarkar, Bidisha Goswami, Saibal Kar and Snehanshu Saha. Revenue Forecasting in Technological Services: Evidence from Large Data Centers, Working Paper, DOI: 10.13140/RG.2.2.21103.027282016
27. Snehanshu Saha, Jyotirmoy Sarkar, Avantika Dwivedi, Nandita Dwivedi, Anand M. Narasimhamurthy, Ranjan Roy A Novel revenue optimization model to address the operation and maintenance cost of a data center, *Journal of Cloud Computing,* Volume 5 (1), pages: 1-46,2016.
28. Camilla Mastromarco, Sucharita Ghosh Foreign capital, human capital, and efficiency: A stochastic frontier analysis for developing countries, *World Development, Volume37(2), pages:489-502, 2009*

APPENDIX A

RREF: The software framework: The Terrence Tao Dataset contains the referenced network of Terence Tao's 4 most cited articles. This data was scraped and stored in JSON format. Each article has its own reference network stored in separate JSON files. Figure 18 displays a sample of the JSON structure.

The JSON Structure: Implementation We used one of these articles to create a visualization of the referenced network, using Web technologies.

Article Id: 5452187

Article Title: The power of convex relaxation: Near-optimal matrix completion

We used Node.js server as our back-end, HTML, CSS and JavaScript as our front-end and enhanced the user experience by the use of Web Sockets. The web architecture was build upon MVC framework. Handlebars view-engine was used at the front-end. We used an external JavaScript based library, vis.js, for displaying the graphical structure using nodes and edges.

From the sample of the JSON structure mentioned above, our prime concern for creating a visualization lies only with the keys *details* and *referenced articles*. Information such as Article Id, Authors, DOI is fetched from *details*. *referenced articles* is an array of objects which contain the articles being referenced by the current article. Each object in the array has the same structure as the current article, making the complete structure recursive. We created two different types of visualization for the dataset. A detailed structure, and another a graphical structure. We shall discuss both of them separately.

Detailed Structure In the detailed structure we fetch all information that is present in the dataset, relevant to out reference network. As the request for the page arrives,we fetch first article's details, such as Title, Authors, DOI, Article ID, and limited details such as Article Id and Title for each of the articles being referenced. This extracted information is sent to the front-end where, with the use of front-end technologies, the data is displayed. Here is an image of the view.

While fetching the referenced articles, a unique id for each of them is dynamically created and is packed with the rest of the details and sent to the front-end. This unique Id has a format of x-y. Where y denotes the array index of the article in the referenced articles list of xth article. As an example an Id such as 0-2-3 would point to the 4th article in the referenced list of 3rd article in the referenced list of root article.

Based on the selection of user, we fetch the unique Id and apply the following algorithm to arrive at the required article.

Graphical Structure: Similar to the Detailed Structure, the Graphical Structure requires fetching of referenced articles. Here, our focus lies only in fetching the Article Id and Title.

Algorithm 6 Fetching referenced articles of an article

```
Input: Unique Id, Root article
Output: Referenced articles list
procedure findReferencedArticles(id,article)
    heirarchy ← id.split('–')
    list ← { }
    for i ∈ heirarchy do
        article ← article.referenced_list[i]
    end for
    for ref ∈ article.referenced list do
        list ← list ∪ ref
    end for
    return list
end procedure
```

Upon arrival of request for the page, the first article's details are fetched and sent. Using vis.js, root article node is created. The same method of assigning unique id is followed here as well. On selection of the article node, a request is sent to the server along with the corresponding unique id. Algorithm 1 is applied to form a list referenced articles, and response is sent with the formed list. On the front-end, with the help of vis.js, article nodes are created with the text as the Article Id, and edges are created between the selected article node and each of the newly created referenced article nodes.

THE JSON STRUCTURE : CHALLENGES: Using the JSON structure of the referenced network for achieving the visualization, put forward several problems. These are described in brief below:

- **Complexity:** Due to the huge size of the referenced network, the JSON data is too difficult to understand just by reading. For the purpose of knowing the structure of the data, one must program. Since the structure is recursive, it becomes fairly easy to understand the structure.
- **In-Memory Storage:** As stated, the best way to understand the data would be through the code written. But, while doing so, the data would have to be stored in the main memory. For systems with low computing capacity and less main memory size, this becomes difficult, and eventually impossible to understand the data.
- **Deriving Outcomes:** With the data being a recursive JSON structure, outcomes such as obtaining the most referenced article or the centrality becomes burdensome. The JSON format cannot be interpreted as a graph and hence, fails to achieve the basic requirement of various graph theory algorithms.
- Time Complexity: JSON structure is computationally intensive for mining purposes.

THE MATRIX STRUCTURE : Knowing the challenges proposed by the JSON structure, we came up with the MATRIX structure for understanding the referenced network. The JSON data was traversed and MATRIX was created from it. The MATRIX is an n x n adjacency matrix, where n is the number of distinct articles. Unlike the JSON Structure, here our primary key or the unique identification for an article is the article id itself.

At first a 2-dimensional array is created with rows and columns as distinct article IDs. As soon as a new distinct article is discovered in the data, it is appended to both row and column lists. After creating the 2-dimensional array, another traversal is made through the structure to identify references and create edges in the MATRIX, by placing '1' at the correct row-column pair. Once the MATRIX is created, it is stored in CSV format for ease of use. Further, a list containing the article IDs corresponding to row index is appended to the CSV file, at the beginning.

THE MATRIX STRUCTURE: IMPROVING THE TASK COMPLEXITY

The query destination for finding out the referenced articles for a given article has now shifted from JSON to MATRIX. This results in decreasing the size of the query destination by a huge margin. For the second article of the dataset, it was noted that the size dropped to around 5% of the JSON data. The decrease in size appears because most of the irrelevant data, in context to our purpose, is cropped out and only the referenced network is taken into consideration. Additionally, the reference is now denoted just by placing '1' in the appropriate intersection of row and column.

Decrease in size of the file solves the issue of **In-Memory Storage,** put forward by the JSON structure. Being small in size, the data fits in most of the modern day systems' main memory. Finding out the referenced articles for a given article using this structure is fairly easy than with the JSON structure. Unlike JSON structure, we don't assign unique Ids to the articles, but work primarily with the article ID itself. Since no two articles will have the same article ID, it serves our purpose to identify the parent article in the referenced network MATRIX.

Algorithm 7 demonstrates querying for the referenced articles of a given article ID from the MATRIX structure. First of all, the generated CSV file is read. Because the file contains both, adjacency matrix and a list of article Ids for respective rows and columns, they need to be separated to work upon. Once the adjacency matrix is separated it can now be traversed like any other matrix and the referenced articles list for a given article Id can be generated as shown in the algorithm.

Algorithm 7 Fetching referenced articles of an article using MATRIX

```
Input: article ID
Output: Referenced articles list
procedure findReferencedArticles(articleId)
    refMat ← read_csv("ReferencedMatrix:csv")
    articles ← list(refMat)
    matrix ← toMatrix(refMat)
    index ← articles.index(articleId)
    list ← { }
    for i ∈ matrix[index].length do
        if matrix[index][i] = 1 then
            list ← list ∪ articles[i]
        end if
    end for
    return list
end procedure
```

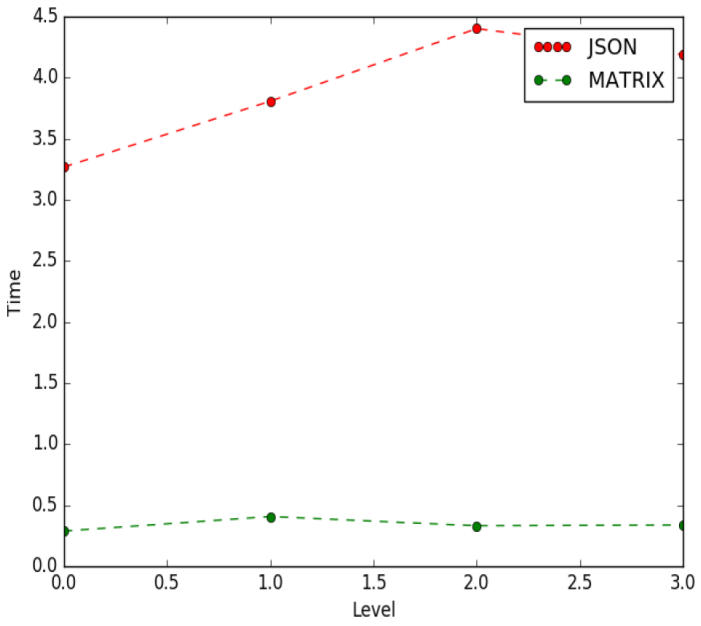


Figure 18: A Time-Level graph: The graph shows time taken to find referenced articles of articles at level 0-3, for JSON and MATRIX

THE MATRIX STRUCTURE: IMPROVING THE TIME COMPLEXITY

When considering Time Complexity, we not only consider the time to fetch the referenced articles but also the time to read the file containing the data. Accepting the fact that MATRIX is a lot lesser in size than JSON, it is obvious that fetching from MATRIX would take lesser time than JSON. Figure 19 is a Time-Level graph for the referenced network of article with article Id **5452187.**

THE MATRIX STRUCTURE: INFORMATION DISCOVERY

With the MATRIX Structure in use for storing reference networks, it becomes uncomplicated to derive outcomes of significance. Extracting the adjacency matrix out of the structure, and applying simple graph theory algorithms on it, important information can be obtained.

As an example, calculating the row-wise sum and finding out the maximum out of them, will result in obtaining **most referencing article** in the network. Similarly, maximum of column-wise sum will provide most referenced article in the network.

APPENDIX B: PROOF OF TRUST SCORE MAXIMIZATION

The Lagrangian function for optimization problem is:

$$L = \gamma - \lambda(w_1 C + w_2 V + w_3 P - m)$$

$$L = (C^\rho + V^\rho + P^\rho)^{\frac{1}{\rho}} - \lambda(w_1 C + w_2 V + w_3 P - m)$$

The first order conditions are:

$$\frac{\partial L}{\partial C} = (C^\rho + V^\rho + P^\rho)^{\frac{1}{\rho}-1} C^{\rho-1} - \lambda w_1 = 0 \quad (24)$$

$$\frac{\partial L}{\partial V} = (C^\rho + V^\rho + P^\rho)^{\frac{1}{\rho}-1} V^{\rho-1} - \lambda w_2 = 0 \quad (25)$$

$$\frac{\partial L}{\partial P} = (C^\rho + V^\rho + P^\rho)^{\frac{1}{\rho}-1} P^{\rho-1} - \lambda w_3 = 0 \quad (26)$$

$$\frac{\partial L}{\partial \lambda} = (C_1 S + V_2 I + w_3 P - m) = 0 \quad (27)$$

Dividing (25),(26),(27) by (24)

$$\frac{w_2}{w_1} = \left(\frac{V}{C}\right)^{\rho-1}$$

$$\frac{w_3}{w_1} = \left(\frac{P}{C}\right)^{\rho-1}$$

Similarly,

$$V = \sqrt[\rho-1]{\frac{w_2}{w_1} C} \quad (28)$$

$$P = \sqrt[\rho-1]{\frac{w_3}{w_1} C} \quad (29)$$

(30)

Substituting these values in equation ((27)), we obtain

$$w_1 C + w_2 \sqrt[\rho-1]{\frac{w_2}{w_1} C} + w_3 \sqrt[\rho-1]{\frac{w_3}{w_1} C} - m = 0$$

```
{   "references":[ ... ],
      "details":{
         "title":"Evaluating Innovative In-Ear Pulse Oximetry for ...",
         "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 ...",
         "date_publication":"Thu Aug 08 00:00:00 EDT 2013",
         "doi":"10.1109/JTEHM.2013.2277870",
         "issue_date":"2013",
         "publisher":"IEEE",
         "authors":[
         "Boudewijn Venema",
         "Johannes Schiefer",
        ],
      },
      "keywords":[ ... ],
      "citations":[ ... ],
      "referenced_articles" : [{
                "references":[ ... ],
              "details":{
                 "title":"Evaluating Innovative In-Ear Pulse Oximetry for ...,
                 "journal_title":"IEEE Journal of Translational ...",
                 "date_current_version":"Thu Sep 05 00:00:00 EDT 2013",
                 "issn":"2168-2372",
                 "abstract":"Homecare is healthcare based on the principle...",
                 "date_publication":"Thu Aug 08 00:00:00 EDT 2013",
                 "doi":"10.1109/JTEHM.2013.2277870",
                 "issue_date":"2013",
                 "publisher":"IEEE",
                 "authors":[
                 "Boudewijn Venema",
                 "Johannes Schiefer",
                ],
              },
                "keywords":[ ... ],
                "citations":[ ... ],
                "referenced_articles" : [ ... ]
      },{
        ...
      }, ...
      ]


  }
```


Figure 19: Sample of scraped data in JSON format

$$C = \frac{mw_1 \frac{1}{\rho - 1}}{w_1^{\frac{\rho}{\rho-1}} + w_2^{\frac{\rho}{\rho-1}} + w_3^{\frac{\rho}{\rho-1}}} \quad (31)$$

Similarly

$$V = \frac{mw_2 \frac{1}{\rho - 1}}{w_1^{\frac{\rho}{\rho-1}} + w_2^{\frac{\rho}{\rho-1}} + w_3^{\frac{\rho}{\rho-1}}}$$

$$P = \frac{mw_3 \frac{1}{\rho - 1}}{w_1^{\frac{\rho}{\rho-1}} + w_2^{\frac{\rho}{\rho-1}} + w_3^{\frac{\rho}{\rho-1}}} \quad (33)$$


```
{   "Article":{  We have applied various
    processing methodologies for ...",
      "references":[
          "5290134",
          "6189752",
      ],
      "details":{
          "title":"Evaluating Innovative In-
              Ear Pulse Oximetry for ...",
          "journal_title":"IEEE Journal of
              Translational ...",
          "date_current_version":"Thu Sep 05
               00:00:00 EDT 2013",
          "issn":"2168-2372",
          "abstract":"Homecare is healthcare
               based on the principle ...",
          "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 20: Sample of scraped data in JSON format

Table 1: Most influential articles of a sample IEEE reference 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

Table 2: Mathukumalli Vidyasagar Article Summary Sample: We observe that diversity is within the range 0.4-0.6; this may facilitate increased citations

Title	Citation Count	References Count (LVL 1-4)	Diversity Score
Algebraic design techniques for reliable stabilization	432	972	0.567
Optimal rejection of persistent bounded disturbances	412	1962	0.669
Algebraic and topological aspects of feedback stabilization	321	1816	0.562
Robust linear compensator design for nonlinear robotic control	262	1249	0.525
The graph metric for unstable plants and robustness estimates for feedback stability	256	1469	0.555
On the stabilization of nonlinear systems using state detection	230	733	0.554
Decomposition techniques for large-scale systems with non additive interactions: Stability and stabilizability	213	1750	0.516

Table 3: Mathukumalli Vidyasagar Article's Trust Score Value

Title	Year	Citation Count	Diversity	Depth	Trust score
Algebraic design techniques for reliable stabilization	1982	0.5194	0.567	0.433	2.5
Optimal rejection of persistent bounded disturbances	1986	0.4953	0.669	0.331	2.6
Algebraic and topological aspects of feedback stabilization	1982	0.3959	0.562	0.438	2.48
Robust linear compensator design for nonlinear robotic control	1987	0.315	0.525	0.475	2.48
The graph metric for unstable plants and robustness estimates for feedback stability	1984	0.3078	0.555	0.445	2.47
On the stabilization of nonlinear systems using state detection	1980	0.2765	0.554	0.446	2.47
Decomposition techniques for large-scale systems with non additive interactions: Stability and stabilizability	1980	0.2561	0.516	0.484	2.44

Table 4: Terence Tao Articles data set 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

Table 5: 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

Table 6: Terrence Tao Article's Trust Score Value

Title	Year	Citation Count	Diversity	Depth	Trust score
Decoding by linear programming	2004	0.6799	0.59	0.41	2.62
Near-optimal signal recovery from random projections: Universal encoding strategies	2006	0.733	0.68	0.32	2.69