Project Report
CSE519: Data Science Fundamentals
Ranking of Research Papers

December 6, 2018

Abstract

The assessment of scholarly work of a scientist and evaluation of a quality of a journal or conference is of high importance as there is a benefit from obtaining an unbiased and fair criterion. In this project, we discuss on defining a scoring function that overcomes the disadvantages of the present metrics used for ranking scientists and journals. We intend to design a reach metric for the publication network and evaluate its efficiency/correctness against the established metrics for the varid dataset of publications. We also perform Time analysis of paper citation count against our metric system by performing topic modelling.

Keywords: H-Index; Eigen vector; Citation network; Page Rank; Centrality measure; Time Analysis; Topic Modelling; Network Graph;


#1
#2
#3
#4
#5
#6
#7
#8
Groundbreaking
Researdhers
#9
#10
#11
PosterMyWall.com

1 Objective

This Project aims at designing a metric tool using Data science principles in automatically evaluating the scholars and academic papers based on analyzing the user input.

2 Motivation

The present citation metrics outline the research quality by considering only the number of citations that research have made or the consistency of a particular research cited in other research but not on how strongly the research content of the paper effect that particular domain was our main motivation in designing a metric system that accurately evaluate a work of a researcher by considering all possible outliers.

3 Literature Survey

The H-index is an established bibliometric ranking function, but as the problem statement mentions the h-index has its limitations. The limitations are due to the fact that the h-index does not take into account of the importance of the cited paper in its ranking function.
The paper [1] discusses about the possible approaches for bibliometric ranking such as the citation count, balanced citation count and page rank. We can see that the first two methods again does not take into account of the importance of the citing paper and can be categorized as simpler approaches. The third method is the page rank and only in this we get to see the importance of the citing papers being used in the ranking function.
Disadvantages of current comparative features.

- H- index, Although its the most popular metric of comparison available, It doesn't rank the author or the paper it just attempts to measure both the productivity and citation impact of the publications of a scientist or scholar.
- Does not take into account of the importance of the citing paper.
- Does not take into account of the self citations impacting the rank.
- I-10 index, Again is just a measure of productivity of the author. It actually means the number of publications with at least 10 citations. It doesn't rank the authors.

4 Data Sources

Initially, we started exploring Citation Data Network (aminer.org) as our primary source of dataset. This data has been extracted from **DBLP, ACM, MAG** and other sources covering over three millions of data. We also explored other datasets such as **CiteSeer** and **Scopus** which is confined to one particular domain, i.e. Science and Technology. Exploring dataset from **Microsoft Academic Graph (MAG)** made us realize that the data is beyond the scope of our hardware computation. Even the Google cloud virtual API failed to effectively consolidate and outline even after pipelining the data into smaller chunks. Also, diving the dataset and working on them individually will not create a robust model. Finally, as suggested by the captain: Shouvik Roy, we explored and trained our model based on citation data network. We have also incorporated and explored other data sources such as **OAG, CORA** suggested by the evaluators of our project proposal before considering our primary source.

5 Data Limitations

All the datasets available are either not consistent with the current research, meaning its not up to date. Also the research data span across approximately across **300 millions** and its size of this dataset is approximately **150 GB**. Its nearly impossible to load such amounts of data and process for hence we have restricted our dataset to dblp, but have explored MAG dataset to work in for topic modelling and for ranking research papers across multiple domains. Also since the data is not in consistency with the current Google scholars data we have calculated our own hindex and other parametric measures.

6 Challenges Faced

- Citation Data Network has a lot of incomplete data within which made the task of structuring data difficult. Still, we were able to overcome this by processing the data in chucks and homogenized into single CSV.
- Creation of edge list between the cited paper and the citing paper to build a graph of research paper which is utilized to calculate the evaluation rank metric and centrality of the paper (reach function) was time consuming and more challenging as the created edge list was humongous and data was rancorous. Still, we able to optimize our algorithm by running DFS to calculate in-degrees (the list of papers citing the present paper) and saving the computations at every step using dictionary to avoid duplication.

7 Data Preprocesing

Data obtained for Citation Network Dataset requires munging and unit conversions of data objects. In our process of effectively preprocessing the data we are passing the data through the following pipeline that performs required action:

- **Plaintext Conversion:** As our first step, the loads of imported data should be flattered to extract the json fields. This is done by integrating the json data as files into our pipeline.
- **Dataframe formulation:** The file pointers referring to flattened json data chucks were consolidated to form a single dataframe utilizing pandas module.
- **Data typecast:** Conversion of date objects to datetime format, Unnamed objects to strings, float as it is required to train the model.
- **Imputation:** Data with incomplete attributes were filled with NULL values to effectively use the data in training the model.
- **Outlier Detection:** Removal of all the values signalling negative year, date and month attributes.

8 Dataset Transformation

The Unified preprocessed data from the above pipeline is obtained where each row represents a research work of scientist(s) belong to a particular domain. Each row has unique paper id, so no research component is repeated. As one author/co-author can have multiple works, the author names can be repeated with different paper-ids. Also, as one paper can have multiple authors, the tuple (paper-id, author) is not unique whereas paper-id alone is unique.

The columns under our preprocessed data looks like this:
`[[Year, Abstract, Authors, Id, n_citation, References, Title, Venue]]`

As we need to compute and cite for researchers and research work separately, we decided to calculate them separately by duplicating the records.

Columns in our transformed author rank records:
`[[Author, Cum_Page_Rank, num_citation, centrality, hindex, Metric_Rank, Citation_Rank, Centrality_Rank, hindex_Rank]`

Columns in our transformed research paper records:
`[[Author, Cum_Page_Rank, num_citation, centrality, hindex, Metric_Rank, Citation_Rank, Centrality_Rank, hindex_Rank]`

9 Feature Engineering

We have build a citation network defined as a graph with each research paper representing a node and citations representing a edges in the graph. The edges here are directed ones each being directed from citing node to cited node to cited node. With the help of the graph and their visualisations, we extracted the following features:

- **Cum_Page_Rank:** The new metric we have created to ranks the author is summation of all the page ranks of all the authors published papers divided by the number of authors who have published. Here the weight of the paper which is the pagerank has been distributed to all the authors who have published the paper.
- **Num_Citation** Total number of citations of the author.
- **Centrality:** Centrality is the reach function of an Author it basically shows the reach of the paper.
- **H_index:** H_index is a consistency measure, it removes the weight for a single paper, H Index has to be calculated as data is not consistent.
- **Metric_Rank:** Metric Rank is the rank of the author according to Cum_Page_Rank.
- **Citation_Rank:** Citation Rank is the rank of the author according to number of citations of the author.
- **Centrality_Rank:** Centrality Rank is the rank of the author according to the centrality measure of the author.
- **Hindex_Rank:** hindex_Rank is the rank of the author according to h index of the author.

10 Feature Importance: Reach

Reach is is roughly translates into the maximum number of nodes which can be reached from that node. In our case the maximum number of papers which reference our paper, or reference a paper which reference our paper of interest. To determine the reach of paper we have first crated a Author graph where in an edge is added when author has referenced an other author, we have identified and used three different metrics as below. Please refer to **Figure 1**

- **Eigenvector Centrality:** Eigenvector centrality is a measure of the influence of a node in a network. Relative scores are assigned to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. A high eigenvector score means that a node is connected to many

Author	Cum_Page_Rank	num_citation	centrality	hindex	Metric_Rank	Citation_Rank	Centrality_Rank	hindex_Rank	diff
Chih-Chung Chang	0.000170	33441	0.000000	3	67	195	973324.5	360338	360271
John Ross Quinlan	0.000169	20830	0.000612	3	69	641	8330.5	360338	360269
Siavash M. Alamouti	0.000193	16708	0.000000	6	40	1036	973324.5	173376	173336
J. Ross Quinlan	0.000208	11676	0.000400	7	35	2038	15001.5	144746	144711
Leo Breiman	0.000362	47250	0.000290	8	6	67	22519.5	123186	123180
Emre Telatar	0.000169	15394	0.000033	22	70	1217	177134.0	30437	30367
Fred D. Davis	0.000209	52206	0.008978	24	34	53	16.0	26516	26482
Andrew P. Witkin	0.000162	36676	0.001292	30	76	153	2373.0	18297	18221
M. E. J. Newman	0.000169	30142	0.002877	33	71	260	459.0	15474	15403
Sally Floyd	0.000158	32779	0.001282	33	78	210	2415.0	15474	15396
John R. Koza	0.000188	18557	0.004463	34	46	841	136.0	14646	14600
Jyh-Shing Roger Jang	0.000151	17403	0.000182	34	91	967	38641.0	14646	14555
Paul A. Viola	0.000145	46951	0.001267	37	98	69	2482.5	12550	12452
Rodney A. Brooks	0.000198	21044	0.001834	39	38	631	1218.5	11319	11281
Charles E. Perkins	0.000171	53295	0.000874	40	66	48	4776.5	10779	10713
Yoav Freund	0.000153	42716	0.007321	41	86	92	32.0	10275	10189

Figure 1: Top 10 Authors based on Reach function.

nodes who themselves have high scores. We have calculated based on the eigen network centrality of our author network graph.

- **Katz Centrality:** Katz centrality is a measure of centrality that computes the relative influence of a node within a network by measuring the number of the immediate neighbors (first degree nodes) and also all other nodes in the network that connect to the node under consideration through these immediate neighbors

- **Indegree centrality:** Indegree as the name itself suggest, it is the total count of incidents on a node, in our case total number of papers referring the main paper.

11 Feature Importance: Topic Modelling

To identify the abstract models that occur in a collection of documents, we chose **Title** of every paper which is rich in content and in-fact the most efficient way in identifying semantics in the text body. Although, abstract has more key words, the dataset we were using had many **NULL** values. To efficacy the data modelling, we followed three metrics to our model as discussed below. Please refer to **Figure 12, Figure 13, Figure 14, Figure 15** for more insights.

- **Parallel Dots API:** The API snippet we used:

```
>>>>>import paralleldots
```

```
# Setting your API key
paralleldots.set_api_key( "YOUR API KEY" )

# Viewing your API key
>>>>> paralleldots.get_api_key()
```

- **Rake Nltk:** RAKE short for Rapid Automatic Keyword Extraction algorithm, is a domain independent keyword extraction algorithm which tries to determine key phrases in a body of text by analyzing the frequency of word appearance and its co-occurance with other words in the text.

```
from rake_nltk import Rake

r = Rake()
r.extract_keywords_from_text(<text to process>)
r.extract_keywords_from_sentences(<list of sentences>)
r.get_ranked_phrases()
r.get_ranked_phrases_with_scores()
```

- **LDA with Gensim:**Latent Dirichlet Allocation is a widely used topic modelling technique which we used to convert set of research papers to a set of topics.

```
import gensim
NUM_TOPICS = x
ldamodel = gensim.models.ldamodel.LdaModel(corpus,
num_topics = NUM_TOPICS, id2word=dictionary, passes=15)
ldamodel.save('model5.gensim')
topics = ldamodel.print_topics(num_words=4)
```

12 Feature Importance: Time Analysis

For the time analysis of the paper citation, we set out to find the popular technologies over the years and collected the data. Using those data, we found out the number of citations those technologies had using Google scholar between an interval of 5 years. We collected this data over repeated spans across the years. This helped us to plot and visualize the data which conveyed the top/trending technologies which had large citation counts for a particular span of time. **Figure 8 Figure 9, Figure 10, Figure 11** describes the time analysis on the research papers from published between 2000 and 2018.

13 Feature Importance: New Rank Metric

The Rationale behind including the new ranking metric was to divide the weight or the measure of the pagerank of papers across all the authors.

- Increases rank of researchers with greater influence based of single paper published but not substantial amount of consistent research work.
- Also this reduces the rank of researchers who research is highly collaborative, whose papers have a lot of co-authors.
- This reduces the rank of researchers who research is highly derivative, whose work references others work.

14 Feature Importance: Arxix Future Popularity

For finding the papers on arxiv that could be popular, we used our metric which we could apply and predict heuristically based on the metadata of the paper such as the number of citations, number of authors, etc. We have also set a threshold for the metric which being calculated on the metadata, which essentially determines whether a paper becomes popular or not in the future. And based on the evaluations,the following are the papers that we predict to become popoular in the future This result also reflects the time analysis we have which shows that in between the

paper	citations	author
Generative adversarial nets	6139	Ian Goodfellow;Jean Pouget-Abadie;Mehdi Mirza;...
Wasserstein gan	1111	M Arjovsky; S Chintala; L Bottou
Self-normalizing neural networks	303	Günter Klambauer; Thomas Unterthiner;Andreas M...
Densely connected convolutional networks	2280	G Huang; Z Liu; L Van Der Maaten; KQ Weinberger
Dynamic routing between capsules	337	Sara Sabour;Nicholas Frosst;Geoffrey E. Hinton
Deep residual learning for image recognition	15606	Kaiming He; Xiangyu Zhang; Shaoqing Ren; Jian Sun
Inception-v4 inception-resnet and the impact o...	1521	C Szegedy; S Ioffe; V Vanhoucke; AA Alemi
Attention is all you need	901	Ashish Vaswani;Noam Shazeer;Niki Parmar;Jakob ...
Mask r-cnn	1143	Kaiming He;Georgia Gkioxari;Piotr Dollár;Ross ...

Figure 2: Arxiv future popularity.

year 2015-2018, the technology having the larger number of citations is machine learning and deep learning combined. Thus, we expect our prediction to hold true and the papers mentioned will stay relevant/popular in the coming years.

$$AverCitations = NumCit/PublishedYear - 2018$$

$$Thresold = \sum_{Top20papers} AverCitations$$

For Each paper in ArXiv Dataset if the *AverageCitation* of the paper is greater than the *Thresold* then we select the paper which might be popular in the future.

15 Baseline Model

We reduced the problem to compare the rankings of the author with their H-Index against our ranking metric. A simple regression model was used to predict the rank based on the columns preprocessed with the above techniques.

- **Features:**
  [[Author, Cum_Page_Rank, num_citation, centrality, hindex, Metric_Rank, Citation_Rank, Centrality_Rank, hindex_Rank]
- **Results:** Please refer to Figure 3 for the correlation.

hindex	Author
90.000	Jiawei Han
90.000	Anil K. Jain
82.500	Philip S. Yu
76.875	Andrew Zisserman
76.250	Thomas S. Huang
75.625	Scott Shenker
73.125	Hector Garcia-Molina
70.625	Sebastian Thrun
70.625	Michael I. Jordan
70.000	Christos Faloutsos

Figure 3: H-Index and Our Ranking Metric Comparison.

- **Drawbacks:** The major flaw in this model is that it ignores the number of citations to each individual article over the other (Reach is missing). Also, Once a paper belongs to the top h papers, its subsequent citations no longer count. Hence, in order to give more weight to highly-cited articles, we are comparing our ranking metric with other metric systems.

16 Advance Model

To factuate the effectiveness of the research paper, we planned to design a model that outputs the best possible rank by considering the various metric systems and normalizing them including our metric system to output the best possible ranking for every research metric.

$$Noramilzation = \frac{dataset - dataset.min()}{dataset.max() - dataset.min()} * 100$$

Minimizing the effect of H Index and maximizing the effect of Cummulative Page Rank.Normalized and weighted average of all the paramters.

$$NormScr = \sum 1.5 * CumPageRank + numCit + InDegCentrlty + KatzCentrlty + 0.9 * HInd$$

- **Features:**
  `[[Author, Cum_Page_Rank, num_citation, centrality, hindex, Metric_Rank, Citation_Rank, Centrality_Rank, hindex_Rank, eigen_vector_centrality, karz_centrality, indegree_centrality]`
- **Results:** Please refer to **Figure 5** and **Figure 4** which bettered the model from before.

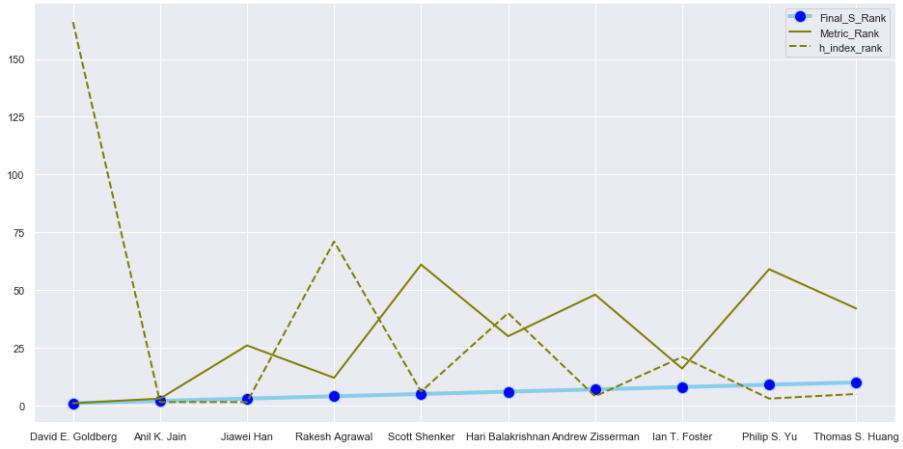

Figure 4: Comparison Model.

17 Evaluation and Validation Metric

Considering the data and tasks in hand, the present model will evaluate the potential of the paper and author by outputting cum_page_rank. To explain the accuracy of the output is compared against the other evaluation techniques such as H_Index,

Author	Final_Normalised_Score	Final_S_Rank
David E. Goldberg	480.063661	1.0
Anil K. Jain	432.198799	2.0
Jiawei Han	327.869374	3.0
Rakesh Agrawal	308.383265	4.0
Scott Shenker	296.799046	5.0
Hari Balakrishnan	294.660663	6.0
Andrew Zisserman	290.066639	7.0
Ian T. Foster	288.800695	8.0
Philip S. Yu	285.545117	9.0
Thomas S. Huang	270.635846	10.0

Figure 5: Top 10 authors based on our final metric.

page_rank and from the explanations (Future_Importance) above, we can deduce that cum_page_rank is the best metric. To explain the performance of our model, we have run our algorithm on a large dataset containing 65 millions of diversified dataset and we have outputed the comparison of our metric with other metrics. As suggested by project captain: Shouvik Roy and other emulators in our project, we have validated our metric system against different ranking metrics. Figure 6 and Figure 7 describes the better placement of our metric with others by correlating higher compares to others.

18 Observations

- Figure 8 This describes the Time Analysis of Research papers against Technology between 2000 and 2005.
- Figure 9 This describes the Time Analysis of Research papers against Technology between 2005 and 2010.
- Figure 10This describes the Time Analysis of Research papers against Technology between 2010 and 2015.
- Figure 11 This describes the Time Analysis of Research papers against Technology between 2015 and 2018.

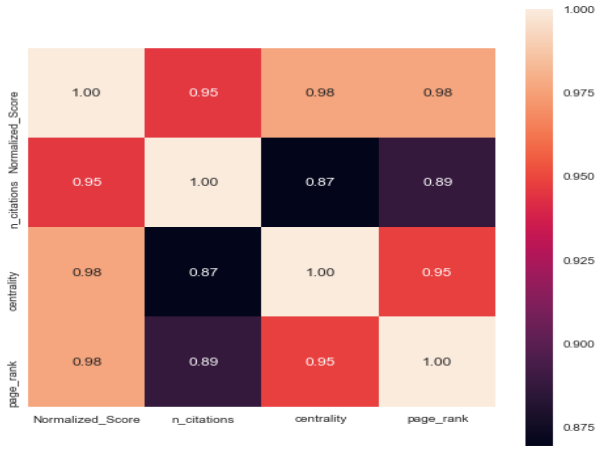

Figure 6: Correlation between our metric and others.

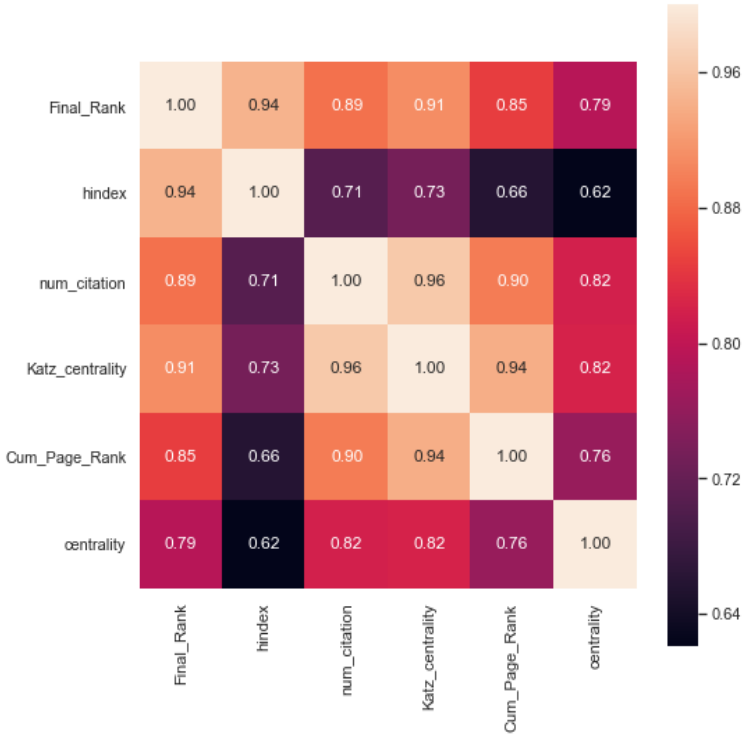

Figure 7: Correlation between our metric and others.

- Figure 12 Topic Modelling explanation as required.
- Figure 13 Topic Modelling explanation as required.

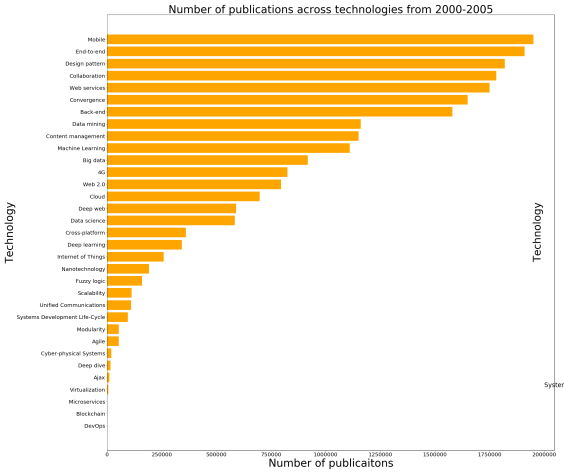

Figure 8: Time Analysis: 1

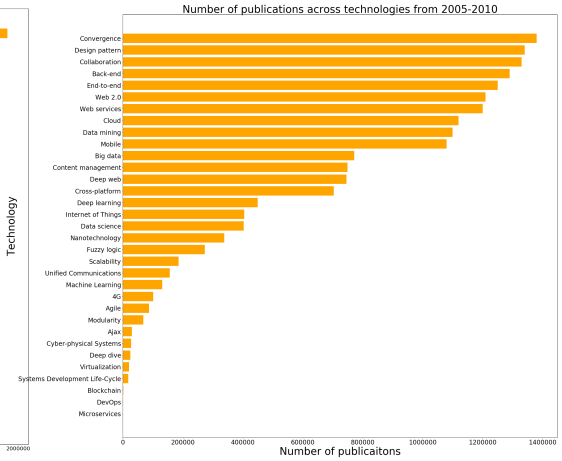

Figure 9: Time Analysis: 2

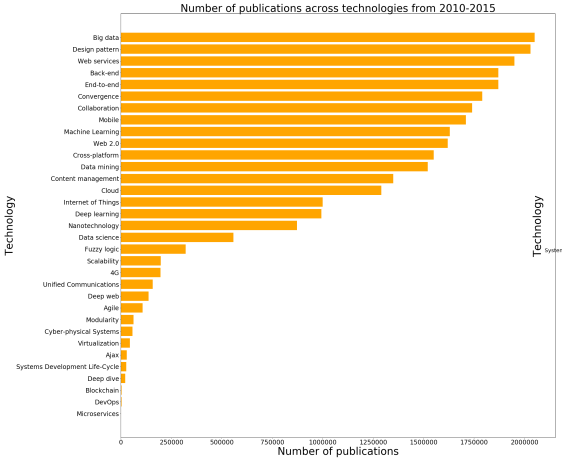

Figure 10: Time Analysis:3

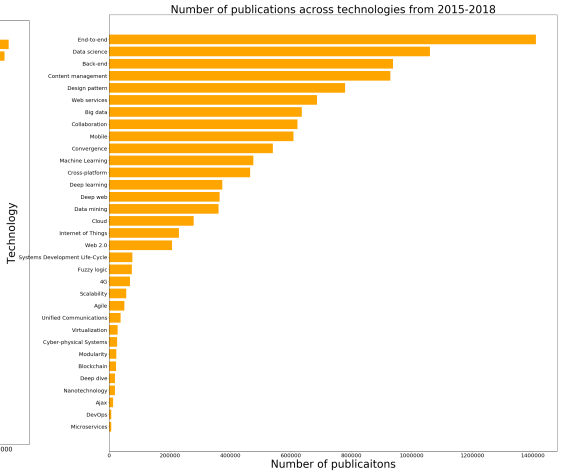

Figure 11: Time Analysis: 4

- Figure 14 Topic Modelling (please watch out for red circle).
- Figure 15 Topic Modelling for Research citation (please watch out for red circle).

19 Answers to the questions asked in course projects section:

- **1.** You must identify top 100 ranked researchers from multiple disciplines based on your ranking metric and see how it stacks up against their respective

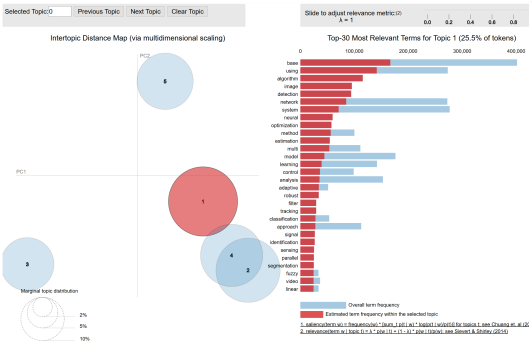

Figure 12: Topic Modelling 1

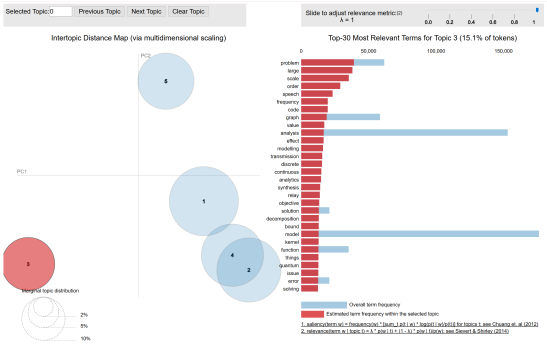

Figure 13: Topic Modelling 2

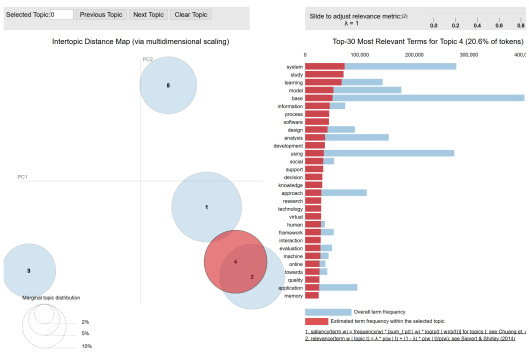

Figure 14: Topic Modelling 3

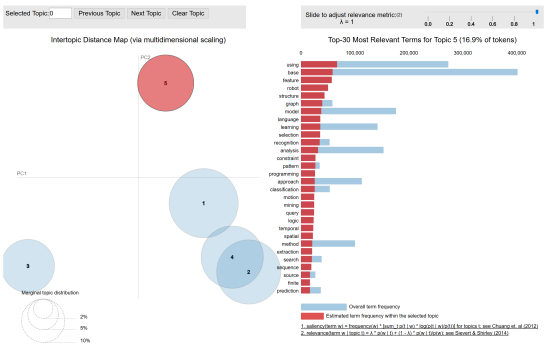

Figure 15: Topic Modelling 4

h-index.

- **Solution 1:** please refer to **Figure 16 and Figure 17** for top 1-20 and top 80-100 authors rank according to our metric.
- **2.** Using a citation network graph, devise a reach function which can identify the degree of a papers influence, which can either be localized in a domain or have a more global inter-disciplinary effect.
- **Solution 2.** Please refer to **Section 10** for the explanation of Reach functionality.
- **3.** Identify examples of researchers with substantial differences between your metric and h-index. Develop an evaluation to decide who is better in these cases, and use it to improve your metric.
- **Solution 3.** After considering the features such as quality of paper, number of citation, number of authors, most popular paper, most popular author, our

Author	Final_Normalised_Score	Final_S_Rank
David E. Goldberg	480.063661	1.0
Anil K. Jain	432.198799	2.0
Jiawei Han	327.869374	3.0
Rakesh Agrawal	308.383265	4.0
Scott Shenker	296.799046	5.0
Hari Balakrishnan	294.660663	6.0
Andrew Zisserman	290.066639	7.0
Ian T. Foster	288.800695	8.0
Philip S. Yu	285.545117	9.0
Thomas S. Huang	270.635846	10.0
Christos Faloutsos	268.969105	11.0
David G. Lowe	266.075354	12.0
Adi Shamir	264.061599	13.0
Takeo Kanade	259.241407	14.0
Wei Wang	259.088771	15.0
Bernhard Schölkopf	256.057925	16.0
Michael I. Jordan	254.929011	17.0
Jitendra Malik	252.152331	18.0
David E. Culler	251.572877	19.0
Geoffrey E. Hinton	251.382710	20.0

Figure 16: Rank 1-20

Moshe Y. Vardi	167.267367	80.0
Pietro Perona	166.982143	81.0
Alberto L. Sangiovanni-Vincentelli	166.850627	82.0
Andrew McCallum	166.684768	83.0
Prabhakar Raghavan	166.634410	84.0
Demetri Terzopoulos	166.338961	85.0
Anantha P. Chandrakasan	165.906202	86.0
Josef Kittler	165.525023	87.0
Alan C. Bovik	165.410224	88.0
Kang G. Shin	164.948867	89.0
Wolfram Burgard	164.296189	90.0
Michael Wooldridge	163.899183	91.0
Carl Kesselman	163.603237	92.0
Jennifer Widom	163.148992	93.0
Bing Liu	162.962089	94.0
Wei Li	162.830122	95.0
Guillermo Sapiro	162.178298	96.0
Dieter Fox	161.923745	97.0
Rui Zhang	161.269525	98.0
Dan Boneh	160.719934	99.0
Stéphane Mallat	160.204448	100.0

Figure 17: Rank 80-100

metric system definitely stands out better than H-Index which is merely based on consistency of the citation. In the below image, you can see that, Author, **Chih-Chung Chang** ranked higher by our metric even though his H-Index is huge because his research work towards Machine Learning, Data science is commendable although he has published very less papers single-authored. Please refer to **Figure 18** for the better picture.

- **4.** Do a time analysis of paper citation count against your metric. Is there any relevance to the topic of the paper against the time it was published. Maybe a paper got too popular due to its time of publishing coinciding with a recent technology interest which the said paper covered. Ensure your metric incorporates that to smooth out papers of high h-index which got higher citation counts because of factors other than solely the quality of material presented.

- **Solution 4.** Please refer to **Section 12** and **Figure 8 Figure 9, Figure 10, Figure 11** for the clear explanation and effort in achieving this.

- **5.** Can you identify papers on Arxiv which should become popular or important?

- **Solution 5.** Please refer to **Section 14** and **Figure 8 Figure 2** for the clear

Author	Cum_Page_Rank	num_citation	centrality	hindex	Metric_Rank	Citation_Rank	Centrality_Rank	hindex_Rank	diff
Chih-Chung Chang	0.000170	33441	0.000000	3	67	195	973324.5	360338	360271
John Ross Quinlan	0.000169	20830	0.000612	3	69	641	8330.5	360338	360269
Siavash M. Alamouti	0.000193	16708	0.000000	6	40	1036	973324.5	173376	173336
J. Ross Quinlan	0.000208	11676	0.000400	7	35	2038	15001.5	144746	144711
Leo Breiman	0.000362	47250	0.000290	8	6	67	22519.5	123186	123180
Emre Telatar	0.000169	15394	0.000033	22	70	1217	177134.0	30437	30367
Fred D. Davis	0.000209	52206	0.008978	24	34	53	16.0	26516	26482
Andrew P. Witkin	0.000162	36676	0.001292	30	76	153	2373.0	18297	18221
M. E. J. Newman	0.000169	30142	0.002877	33	71	260	459.0	15474	15403
Sally Floyd	0.000158	32779	0.001282	33	78	210	2415.0	15474	15396
John R. Koza	0.000188	18557	0.004463	34	46	841	136.0	14646	14600
Jyh-Shing Roger Jang	0.000151	17403	0.000182	34	91	967	38641.0	14646	14555
Paul A. Viola	0.000145	46951	0.001267	37	98	69	2482.5	12550	12452
Rodney A. Brooks	0.000198	21044	0.001834	39	38	631	1218.5	11319	11281
Charles E. Perkins	0.000171	53295	0.000874	40	66	48	4776.5	10779	10713
Yoav Freund	0.000153	42716	0.007321	41	86	92	32.0	10275	10189

Figure 18: Why our metric is better.

explanation and effort in achieving this.

20 Acknowledgement

We would like to thank Professor Steven Skiena for his support and encouragement. We would also like to thank our project captain Shouvik Roy for his efforts in helping us whenever we are stuck with our project.

21 References

- https://en.wikipedia.org/wiki/PageRank
- https://en.wikipedia.org/wiki/Centrality
- https://www.guru99.com/database-normalization.html