Profiling Researchers Based on Features Extracted from Articles and Citations

B. Sc. Thesis

Erfan Loghmani
Supervisor: Dr. Abolfazl Motahari

Sharif University of Technology

Table of contents

Introduction

Question

How does Sharif University evaluates researchers that apply for academic position?

Why Evaluating Researchers

Universities and academic institutes have several decisions to make:

- Hiring decisions
- Promotion
- Salary decisions
- Performance reviews
- Allocation of research resources

Do Metrics Matter?[1]

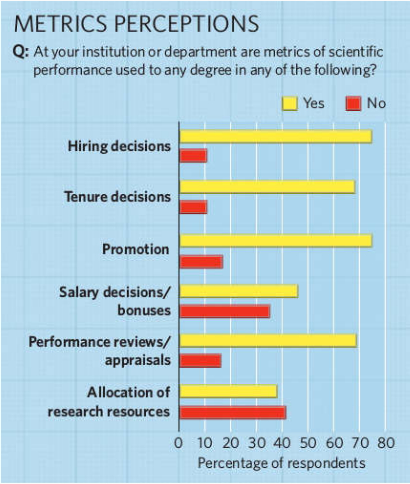


Figure 1: *Nature* poll results from 150 readers in 2000, from: Abbott, Alison, et al. "Metrics: Do metrics matter?." Nature News 465.7300 (2010): 860-862.

Metrics Side Effects

Figure 2: Some of readers opinions about how metrics affect their behaviour, from: Abbott, Alison, et al. "Metrics: Do metrics matter?." Nature News 465.7300 (2010): 860-862.

Literature Review

Sinatra et al. 2016

Quantifying the evolution of individual scientific impact [2]

A quantitative model, which systematically untangles the role of productivity and luck in each scientific career.

Observation: Random Impact Rule

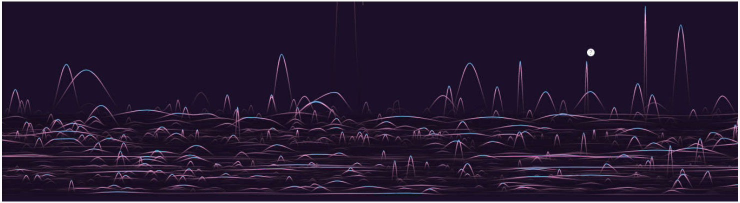


Figure 3: Visualizing the evolution of individual scientific impact, from: kimalbrecht.com

Highest-impact work can be, with the same probability, anywhere in the sequence of papers published by a scientist.

R Model

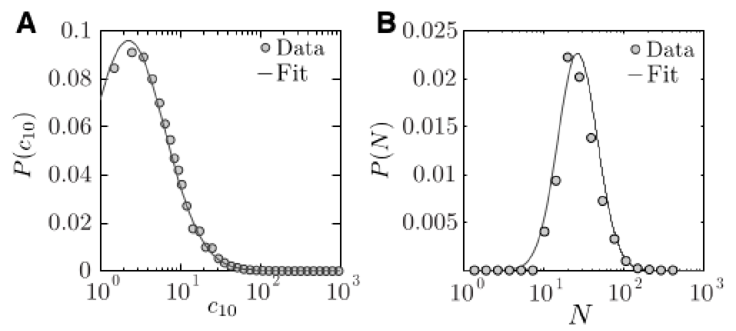


Figure 4: c_{10} and N distribution and best log-normal fit, from: Sinatra, Roberta, et al. "Quantifying the evolution of individual scientific impact." Science 354.6312 (2016): aaf5239.

Assume that each scientist publishes a sequence of papers whose impact is randomly chosen from the same impact distribution $P(c_{10})$.

Problems with R Model

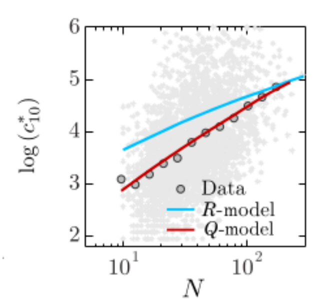


Figure 5: Citations of the highest-impact paper,c_{10} ,versus the number of publications N during a scientist's career, from: Sinatra, Roberta, et al. "Quantifying the evolution of individual scientific impact." Science 354.6312 (2016): aaf5239.

Problem 1

Productivity alone begets success

Problems with R Model (contd.)

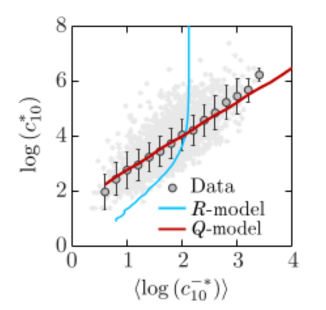


Figure 6: $log(c^*_{10})$ vs $< log(\hat{c^*_{10}}) >$, from: Sinatra, Roberta, et al. "Quantifying the evolution of individual scientific impact." Science 354.6312 (2016): aaf5239.

Problem 2

Divergent impact: The higher the average impact of a scientist's publications without the most-cited publication $< log(\overline{c^*_{10}}) >$, the higher the impact of the most-cited paper, $log(c^*_{10})$.

Q Model

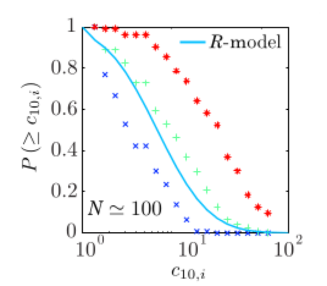


Figure 7: Citation distribution of authors with same productivity, , from: Sinatra, Roberta, et al. "Quantifying the evolution of individual scientific impact." Science 354.6312 (2016): aaf5239.

With the same productivity authors have different citation distributions.

Introducing equation:

$$c_{10,ia} = Q_i p_a, \tag{1}$$

Q Model (contd.)

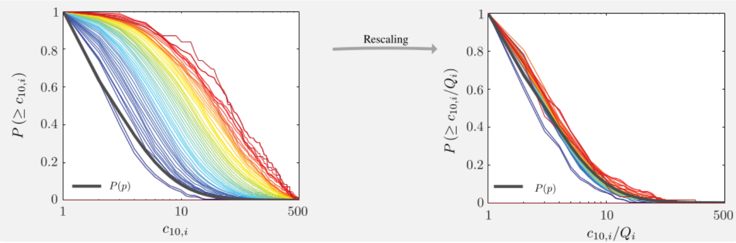


Figure 8: Universal behaviour after rescaling by q, , from: Sinatra, Roberta, et al. "Quantifying the evolution of individual scientific impact." Science 354.6312 (2016): aaf5239.

Dataset

DBLP Summery

Table 1: Dataset summery

Papers	3,079,007
Citations	25,166,994
Authors	1,766,547

With fields for each paper:

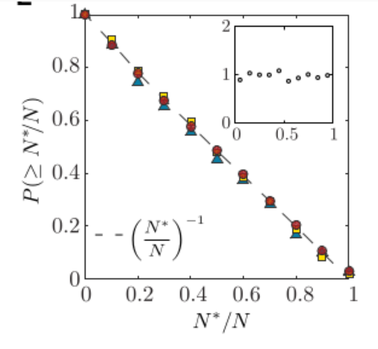
- Authors
- Abstract
- Venue
- Publish year
- References
- DBLP ID

```
{
    "abstract": "We consider a memoryless Gaussian interference channel (GIC) where  K  single-antenna users com
municate with their respective receivers using Gaussian codebooks. Each receiver employs a successive group deco
der with a specified complexity constraint, to decode its designated user. It is aware of the coding schemes emp
loyed by all other users and may choose to decode some or all of them only if it deems that doing so will aid th
e decoding of its desired user. For a GIC with predetermined rates for all transmitters, we obtain the minimum o
utage probability decoding strategy at each receiver which satisfies the imposed complexity constraint and revea
ls the optimal subset of interferers that must be decoded along with the desired user. We then consider the rate
allocation problem over the GIC under successive group decoding and design a sequential rate allocation algorith
m which yields a Pareto-optimal rate allocation, and two parallel rate allocation algorithms which yield the sym
metric fair rate allocation and the max-min fair rate allocation, respectively. Remarkably, even though the prop
osed decoding and rate allocation algorithms use ldquogreedyrdquo or myopic subroutines, they achieve globally o
ptimal solutions. Finally, we also propose rate allocation algorithms for a cognitive radio system.",
    "authors": [
        "Narayan Prasad",
        "Xiaodong Wang"
    ],
    "id": "53eda920-67db-4aa0-8652-6ad3238be775",
    "references": [
        "0d90f37b-3aaa-4fc2-b3aa-0b4b1de24c10",
        "1774d6ce-20bb-4010-ad23-361fa8e367be",
        "1eddb4e0-a074-4189-a370-e53724a96bbd",
        "28c88468-6c97-46b6-b551-4fa0be9f0b30",
        "6c5c8a75-af91-4f49-a9d7-b483a1fbe977",
        "8bb2c446-0081-4404-a944-56a0d5dc2f15",
        "90609f26-928e-47a1-80f5-0d5a8e3a4858",
        "a46640e2-63e8-40b1-a4c6-744e514936f3",
        "b286c0b1-55f7-4b9a-8252-27117bf82b77",
        "cac3fc3f-ffc3-4b8f-a77d-078358ea6e4c",
        "d03c481d-ce53-415b-b250-d4f745ecbf6d",
        "fdefadfa-5be7-497d-9ae7-766b7675c720",
        "ff56835b-e3b8-4a86-8510-917c6bb58d84"
    ],
    "title": "Outage Minimization and Rate Allocation for the Multiuser Gaussian Interference Channels With Succ
essive Group Decoding",
    "venue": "IEEE Transactions on Information Theory",
    "year": 2009
}
```

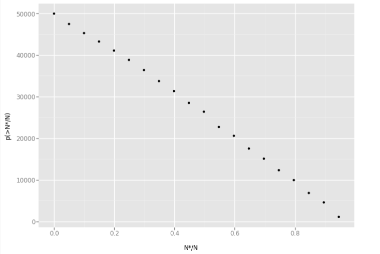

Figure 9: Document sample

Observations

Test Sinatra et al. Observations



(a) Sinatra et al.



(b) DBLP

Figure 10: Distribution of the highest-impact paper

Observation

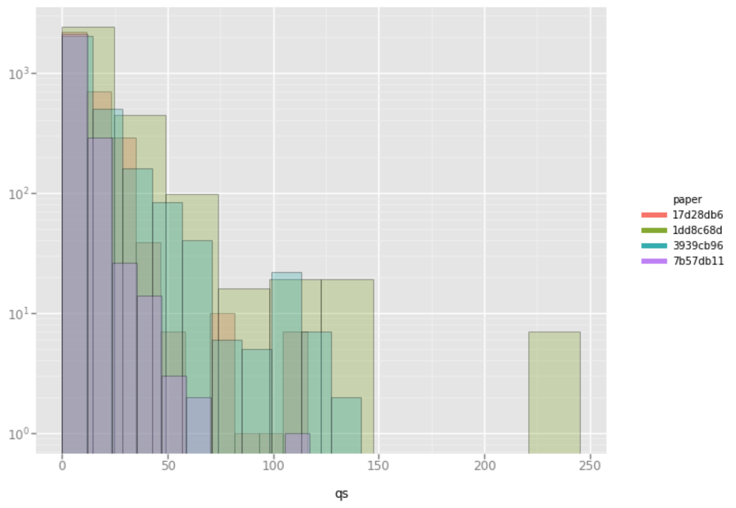


Figure 11: Quality distribution of authors cited papers with around 1000 c_{10} that published in 2000

Method

Q Model Problem

In Q model all paper citations count the same while if an author with more quality cites a paper we could value that citation more than a citation from low rank author.

Method

1. Set initial author qualities (Q) to 1
2. For each paper calculate v1, as average author qualities
3. For each paper calculate v2, with summing v1 of papers that cite this paper
4. For each author, calculate Q with Sinatra et al. equation for Q values using papers v2.
5. While Q changing is not stable go to 2

Results

Practical Convergance

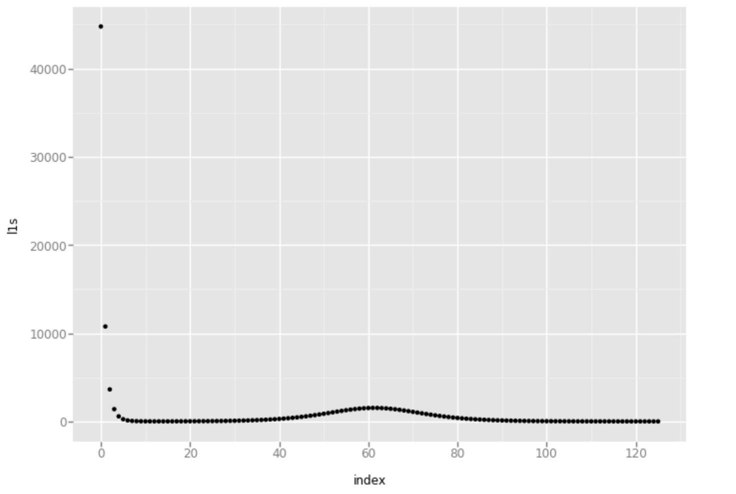


Figure 12: L1 distance of each iteration author qualities to previous iteration

Ranking change

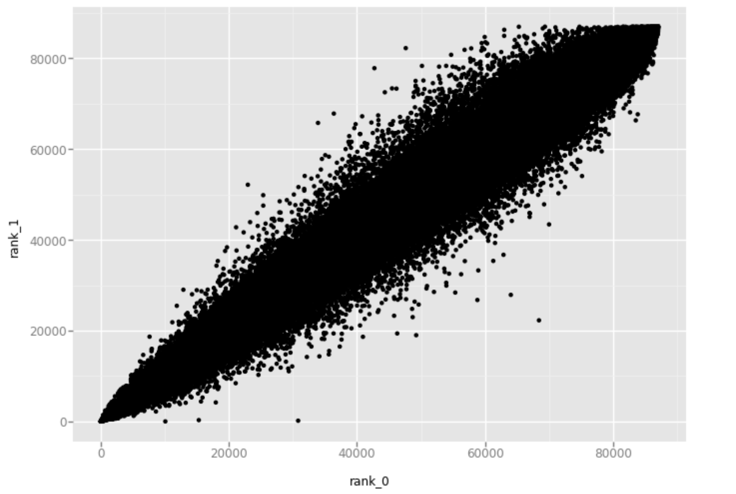


Figure 13: Ranks after Iterative refinement vs before it

Self Citing Authors

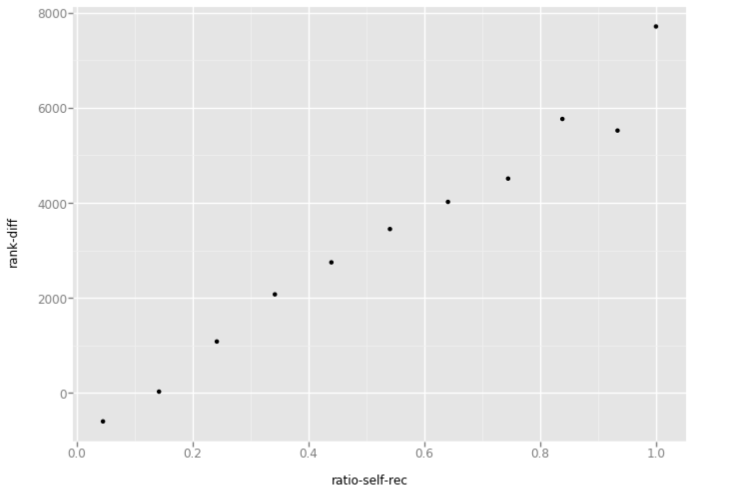


Figure 14: How our ranking, changes position of authors with different ratio of recieved self citations.

Questions?

- A. Abbott, D. Cyranoski, N. Jones, B. Maher, Q. Schiermeier, and R. Van Noorden.
  Metrics: Do metrics matter?
  Nature News, 465(7300):860–862, 2010.
- R. Sinatra, D. Wang, P. Deville, C. Song, and A.-L. Barabási.
  Quantifying the evolution of individual scientific impact.
  Science, 354(6312):aaf5239, 2016.