Analysis of Self-Feeding Communities in the Medical Research World

Lior Ziv - lior.ziv2@mail.huji.ac.il - liorz

Tair Shauli - tair.shauli@mail.huji.ac.il - tairsha

Amir Bar - amir.bar@mail.huji.ac.il - amirbar

Table of Contents

Description

Do rich *really* get richer? We hypothesise that researchers of enriching communities, communities with high intra-citation frequencies, are more likely to be cited. We wanted to examine this question from a different angle than it is usually examined. The medical research world is a society with variety of highly competitive researchers, with an organisational structure where each head investigator is measured by the amount of publications; This is a world where more citations translates to more success. We wanted to know whether there exists a phenomena of communities of researchers, citing each other in a "scratch my back and i'll scratch yours" sort of behaviour, or not- there might be just a few weak (poorly intraconnected) communities which feed one another. In order to address this question, we have built a directed graph of researchers, where each edge means that there is a citation connecting the two researchers, and the weights are the number of citations. We then applied clique percolation as a method to for community detection (this was done on an undirected graph), filtered for communities which stood out for being highly self-enriching, and applied statistical tests in order to validate our hypothesis.

Data

The generation of the researchers graph required two types of data. First, we needed to retrieve a summary of all the articles published in PubMed Central (PMC) for creation of nodes in the graph (authors of the articles). Afterwards, we needed to extract all the citing articles according to the article IDs. The articles summary data is composed of multiple XML files, each containing many article tags. Each article tag contains information such as the title, authors, journal, dates, institute, abstract (not for all entries) and many more. The list of citing articles to each article ID is a simple XML file (one for each article) containing ID tags of all citing articles. The data was downloaded from PubMed using FTP and Web queries (using python) and contained 1,200,000 articles written by ~200,000 researchers, parsed into python hand-tailored Article objects, and saved in binary.

Method

Graph construction

We built a weighted directed graph where the vertices are researchers and there is an edge from one researcher to another iff there exists an article where the first's name is signed on and the article cites an article that the second researcher is signed on. We are aware of the issue of duplicate names per researcher, and duplicate researchers per name. We could not think of an elegant way to bypass this without introducing a significant amount of complexity (and work hours). We hope, and honestly believe, that this is not a huge bias given the size of our data. The weight for each edge is the amount of times a researcher was citing his partner.

Algorithm parameters selection

To decide which value of K we needed to use for the clique percolation, we have analyzed the distribution of the amount of communities generated by the algorithm in respect to possible values of K (figure 1) and used our prior knowledge of the likely significant size of researcher cliques in the medical research world. This has led us to use the value $K = 4$, which, happily, resulted in 960 communities; a size sufficiently large to test our hypothesis on.

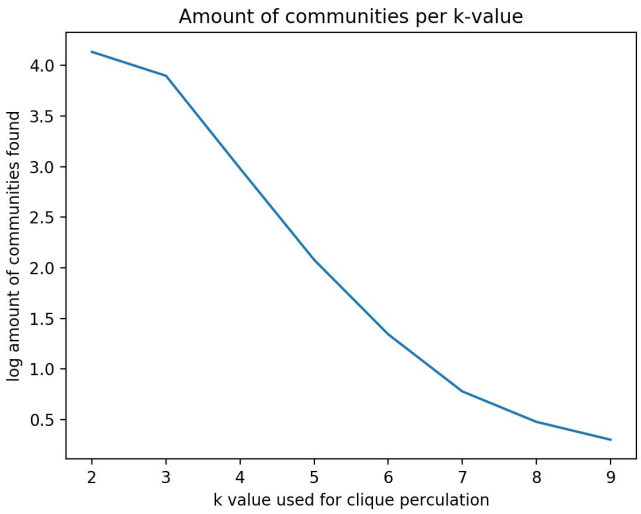

Figure 1 The distribution of local cluster coefficient for each community. The number of communities decays logarithmically as the average grows.

For each community, we calculated two measures to estimate whether it enriches itself, or not. The first measure is the probability of a community to cite articles which were written by someone who is part of the community. The second measure is the average local clustering coefficient of the nodes in the community. After looking at the distributions (figure 2) we decided to use a threshold of 1 for the average cluster coefficient and a threshold of 0.67 (1 STD above the expectation) for the citation probability to distinct communities from enriching communities. We defined a community as enriching community if both measures were above the threshold.

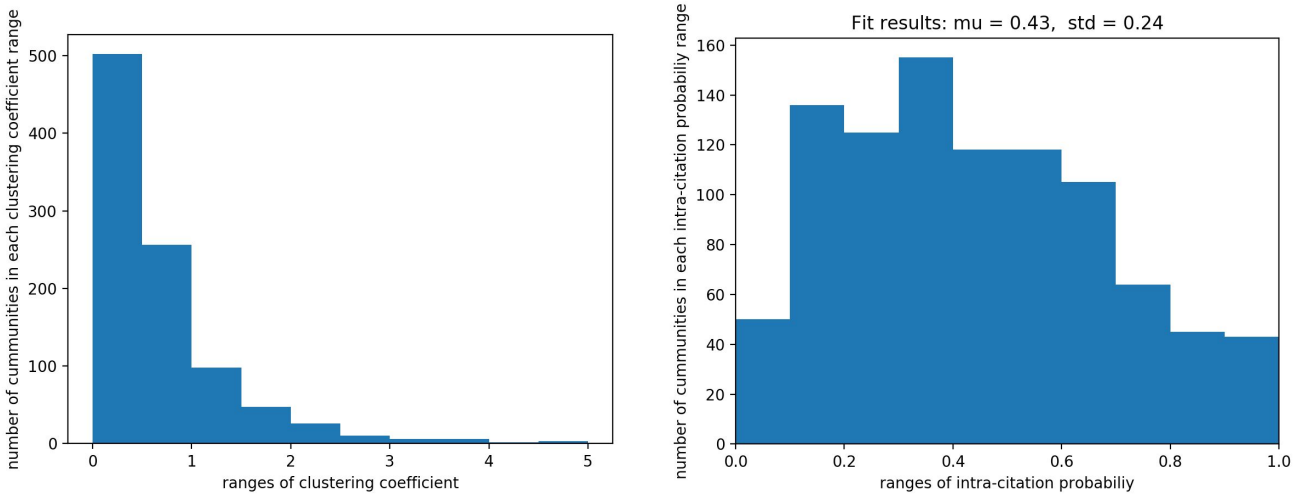

Figure 2 (a) Amount of communities with respect to the initial clique size. This analysis helped us to select the value k, we used for the clique percolation algorithm. (b) Distribution of inner citation probability over the communities. As we can see, the distribution is normal with $\boldsymbol{\mu}$- 0.43 and std - 0.24.

The last parameter we had to decide about for our model is what threshold rank we would use for defining a researcher as *rich*. We used the number of incoming citations as its rank, and analyzed how the ranks are distributed

(figure 3). We decided to use the 90th percentile as our definition of who is rich and therefore we used the value 4 as the threshold.

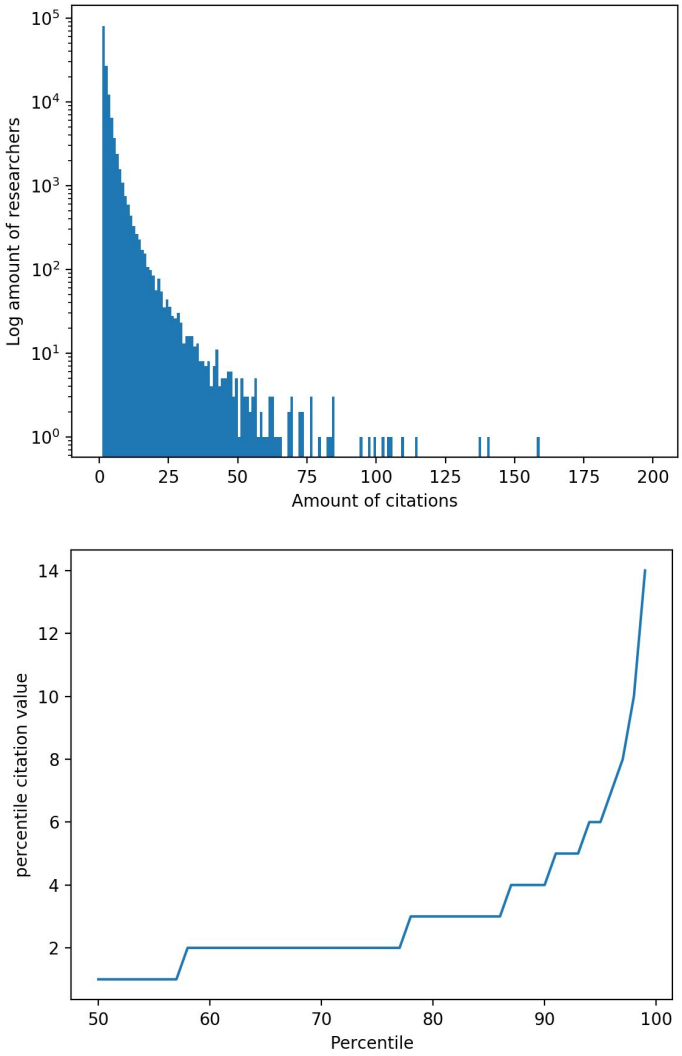

Figure 3 (a) Distribution of researchers amount per number of citations (log scale), as we can see the graph depicts that most of the researchers were cited once. In order to find the rich ones we will take the upper 10% percentile and define them as rich. (b) The amount of researchers per Percentile percentage from 50% up to 100% .

Results

Statistical results

Applying our method on the dataset have resulted in 960 communities where 16 out of them (1.6%) were enriching. The researchers fraction of the different groups have resulted in 36 rich researchers within enriching community, 18,996 rich researchers outside an enriching community, 41 not rich researchers within enriching community, and 183,151 researchers that were not rich and outside an enriching community. With those partitions we have calculated the probabilities

$$P(researcher = rich \mid inside\ enriching\ community) = 0.467$$

$$P(researcher = rich \mid outside\ enriching\ community) = 0.103$$

and applied fisher's exact test to verify the significance of the results and received a p-value of 2.518e-17. We also compared the distribution of researchers ranking in enriching communities to the rest (figure 4). This comparison shows that the median rank of researchers in enriching community is at 3 while the median rank of researchers that are not in enriching community stands at 1. Although those results look promising, we can see that among researchers outside an enriching community there are many outliers. This has led us to believe that although enriching communities is a method to increase a researcher rank there are other methods that researchers use to increase the rank (except of writing an exceptional article).

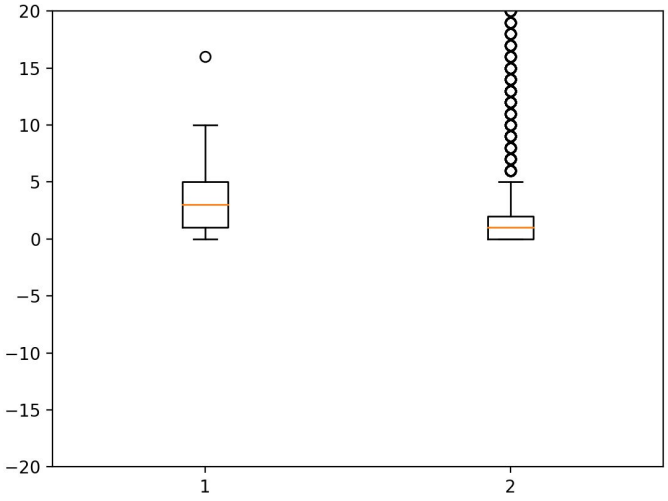

Figure 4 Researchers ranking distribution. (1) In enriching community (2) Outside of enriching community. In orange we see the medians of each of the groups and the box represents upper and bottom boundaries represents the first and the third quartile of the data in each group. The whiskers represents the boundaries ∓ 1.5 the interquartile range (IQR). This graph was trimmed at rank 20 so it would be possible to see the difference in the median although there were higher ranks (as we can see at figure 3).

Communities identified

To see if our method succeeded we have also looked at some of the enriching communities members and have seen an interesting observation. At three out of three of the communities we checked there is strong geographic relation. One community composed of 7 researchers, 6 out of them are from germany and 4 of them are currently working at the same university. Looking deeper at their articles we have seen that most of them have published articles together on multiple occasions (it is possible that they have cited an article they also published together). Another community was composed of 4 researchers from the US (different universities) and another one was composed of 7 researchers 5 out of them are currently from the US and 2 of them from Finland and Taiwan.

Discussion

Although the results support our hypothesis, we are still not convinced that our method is completely accurate. The first concern comes from the size of the data we used for the analysis. At first we thought we will be able to use all the articles since 1970, not realizing that we don't have the space on the disk or RAM memory to store it. That have led us to use only a small fraction of the dataset (all articles between august 1st 2011 to may 24th 2012). As a result of this, we take our results with a grain of salt since the graph might miss many connections due to the relatively short time period. Another issue is our method to select the thresholds. For that task we decided to rely mostly on the distribution of the data although it might not be optimal (or even wrong) and maybe another approach such as parameter estimation would show better and accurate results.

Overall we come to believe that our original hypothesis is true and would like to improve it by running it on a larger scale of data and applying additional analysis such as a comparison of citation count of similar articles between the groups. More intriguing questions we haven't deal with and would like to is the prediction of how the current communities will change over time, how the enriching communities articles citations accumulate in comparison to others articles and what is the quickest way to get cited the most.

Appendix A - Enriching communities

Community 1

- Helmut Kewes - 71461, University of **Cologne ,Germany**
- Nadine Scholz - Neumann - 109447, CEVEC Pharmaceuticals GmbH, **Cologne, Germany**.
- Daniel Neumann - 56373, department of NanoBiophotonics, Max Planck Institute for Biophysical Chemistry, Göttingen, **Germany**.
- Jens Wölfel - 30620, CEVEC Pharmaceuticals GmbH, Gottfried-Hagen-Straße 62, 51105, **Cologne, Germany**.
- Ruth Essers - 30621, CEVEC Pharmaceuticals GmbH, **Cologne, Germany**.
- Corinna Bialek - 133758, CEVEC Pharmaceuticals GmbH, **Cologne, Germany**.
- Sabine hertel - 71460, Waltham, MA, United States

Community 2

- John F Gamble - 104229, department of Anesthesiology, Duke University Medical Center, DUMC 3094, Durham, NC 27710, USA
- Kavi P Patel - 54157, Department of Medicine (Division of Endocrinology, Metabolism and Diabetes), College of Medicine, University of Florida, Gainesville, FL 32611, USA.
- Patricia A Stewart - 54158, Stewart Exposure Assessments, LLC, Arlington, USA.
- Kenny Crump - 33309, Ruston, LA, USA
- Yen-Yuan Chen - 14997, Department of Medical Education, National Taiwan University Hospital bNational Taiwan University College of Medicine cGraduate Institute of Medical Education & Bioethics, National Taiwan University College of Medicine, Taipei, Taiwan.
- Debra T Silverman - 14998, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health.
- Pasi I Jalava - 66621, Department of Environmental and Biological Sciences, University of Eastern Finland, Kuopio, Finland.

Community 3

- Karl W Butzer - 7011, Department of Geography and the Environment, University of Texas, Austin, TX 78712, USA.
- Sheryl Luzzadder-Beach - 7012, Department of Geography and Geoinformation Science, George Mason University, Fairfax, VA 22307, USA.
- B L Turner - 2438, Clinical Associate Professor, Department of Oral and Maxillofacial Surgery, School of Dentistry, University of North Carolina, Chapel Hill, NC.
- Nicholas P Dunning - 2439, Departments of Anthropology, University of Cincinnati, Cincinnati, OH 45221, USA