École Polytechnique Fédérale
de Lausanne

Center for Digital Education

Detect EPFL Activities related to Sustainability

Author
Florine Réau

Teachers
Patrick Jermann
Francisco Pinto

Spring 2021

Contents

Abstract

This report is about graph theory and how it is used to recommend collaboration between researchers working on similar subject but not collaborating. We based this project on graphsearch website and using its database we construct a graph for future collaborations recommendations. In order to do so, we used to metrics, shortest path and modified pagerank, allowing us to cluster the researchers that already collaborated. We then simply recommend the new researchers closest to this cluster.

The code is available at : https://github.com/Floumzi/sustainability.

1 Introduction

The study of social networks through graph theory to understand and predict human relationship is field of recent interest that keeps growing. The networks is represented through a graph where nodes represent people and links represent ties between them.

This project offers a study of collaboration network, linking researcher of a same university, EPFL, and trying to predict further collaboration between researchers.

The co-authorship recommendation problem is well known in social networks analysis, in 2017, Claudel et al. [3] started by analysing the collaboration patterns of faculty at the Massachusetts Institute of Technology. Researchers also started to studied the recommendation part by proposing innovative algorithms to offer co-authorship recommendation, for a example, in 2014 J. Li et al. proposed a random walk algorithm [2].

In this project, we will then study the recommendation problem for co-authorship, on EPFL researchers, trying to answer the following question : **Which researchers are doing research on sustainability, overlapping but not collaborating ?**

2 Implementation

The structure of the implementation is as follows :

First we create the graph with concepts related to Sustainability. Then, we add the researchers who work on these concepts. Finally, for a given researcher, we explore how pre-existing collaborators exist in the graph, and look for similar nodes.

2.1 Data

The original dataset used in this project is an Arango Database named Campus_Analytics. The underlying data is organized as follows :

First we have the nodes :

- The concepts nodes
- The person nodes
- The publication nodes

And we also have the edges between nodes type :

- The edges between concepts
- The edges between persons and concepts
- The edges between persons and publications
- The edges between publications and concepts

This data is extracted via AQL queries using *epflgraph* and *pyarango* libraries.

2.2 Graph Sampling

In order to sample the graph around a given central concept we proceed as follows : Searching the Arango database, we look for all concepts that are neighbours of our central concept, keeping only those whose connecting edge has a normalized score greater than 0.2.

To include concepts which are not neighbours of our central concept, we apply Dijkstra's algorithm [1], and thus start by recomputing edge weights such that they are more representative of distance.

$$e_{Weight} = max(0.2, 1 - e_{NormalizedScore}) \tag{1}$$

The lower floor of 0.2 exists to force Dijkstra's algorithm's accumulator to increase every time we traverse a new edge. The maximum distance from the central concept is set to $\alpha = 1$.

Neighbouring researchers are then added to our graph, along with their incident edges. The resulting sub-graph is the sample that we will then work from.

The choice of α is important as we observe that when setting a small α, we tend to have more researchers than concepts in the graph, while setting a bigger α, we have more concepts than researchers, meaning that we will have more paths from one researcher to another.

The ratio between researchers and concepts is important, as having more concepts allows our graph topology to account for more complex paths between different researchers. Setting $\alpha = 1.0$, we obtain 2760 researchers and 5107 concepts.

2.3 Metrics

In order to measure the suitability of collaboration between researchers, we develop two different metrics. For a given researcher, we compute scores measuring suitability for all others. Note that although the first metric is commutative, the second is not.

Furthermore, our metrics are purposefully independent from pre-existing collaborations. This allows us to see how researchers with which one has collaborated map into the two-dimensional metric space, and then make recommendations about future collaborations from this.

2.3.1 Shortest Path Metric

From our sample, we isolate the sub-graph containing all concept and researcher nodes , we can compute our first metric for the recommendation. To do so, we once again applied Djikstra's algorithm. The function *dijkstra_nx* take as parameters a graph, a researcher and the list of all other researchers in the graph. For the given researcher it returns the cost of the shortest path to each of the remaining researchers. The comparatively low edge weight (and hence high edge cost) of edges between concepts and researchers, compared to that of edges between concepts, privileges concepts over researchers in building these paths between researchers. This metric thus gives us information on the similarity of the research : the smaller it is, the closer the two researchers' domains are.

2.3.2 Modified PageRank Metric

Now let's compute our second metric using a modified version of PageRank algorithm. As with the first metric, this metric is calculated for all other researchers, with regards to a given researcher. Below is the original definition of PageRank[5] :

$$PR(i) = \gamma \sum_{j \in V} \frac{a_{ji}}{d_j^{(out)}} PR(j) + \frac{1-\gamma}{n} \tag{2}$$

The modified version of PageRank that we use is defined below[4]. Let us note how it differs from the original formulation by allowing for edge weights w_{ji}, overall edge importance θ, and node weights β_i.

$$\phi(i) = \gamma \sum_{j \in V} (\theta \frac{w_{ji}}{s_j^{(out)}} + (1-\theta)\frac{a_{ji}}{d_j^{(out)}})\phi(j) + \frac{(1-\gamma)\beta_i}{\sum_{i \in V} \beta_i} \tag{3}$$

Our implementation of PageRank uses as default parameters :

- $\gamma = 0.85$ (damping factor)
- $\theta = 0.5$ (overall edge importance)
- $\beta_i = 10^8$ * 1{i is central node} (node weights)

Furthermore, while all edges are normally replaced by two directional edges of the same weight but of opposite direction, edges incident to the given

researcher are replaced with a single incoming edge (directed towards the given researcher). Combined with the low damping factor and large β_i for the given researcher's node, this boosts the PageRank of nodes in the neighborhood of our given researcher. This is how we adapt PageRank algorithm for different professors (and why this metric is non-commutative). Also, replacing these single directed edges with double edges, as is done with all other graph edges, would cause numerical instability if all other parameters would stay the same, and our algorithm would not converge.

3 Results and Analysis

Looking at the following figures, we can see that previous collaborators tend to cluster to the lower limit of our scatter-plots in the shortest distance dimension : this is consistent with researchers wanting to work with collaborators whose work closely resembles their own.

Furthermore, we notice how the long tail in the PageRank dimension does not follow the point distribution of along the shortest distance axis: it is biased downwards. We can see here the effect of our large β_i for the given researcher's node: the nodes with the best PageRank score thus tend to be closer than expected to the given researcher's node.

Within the cluster of past collaborators which hugs the lower limit of our scatter-plots in the shortest distance dimension, we can see that according to the given researcher, the cluster stretches more or less along the PageRank dimension.

To compute our recommendations, we calculate the geometric mean of points representing past collaborators within this 2-dimensional metric space, and return the closest researchers to that point with whom the given researcher has not yet collaborated.

Figure 1: Recommendation plot for researcher 120480

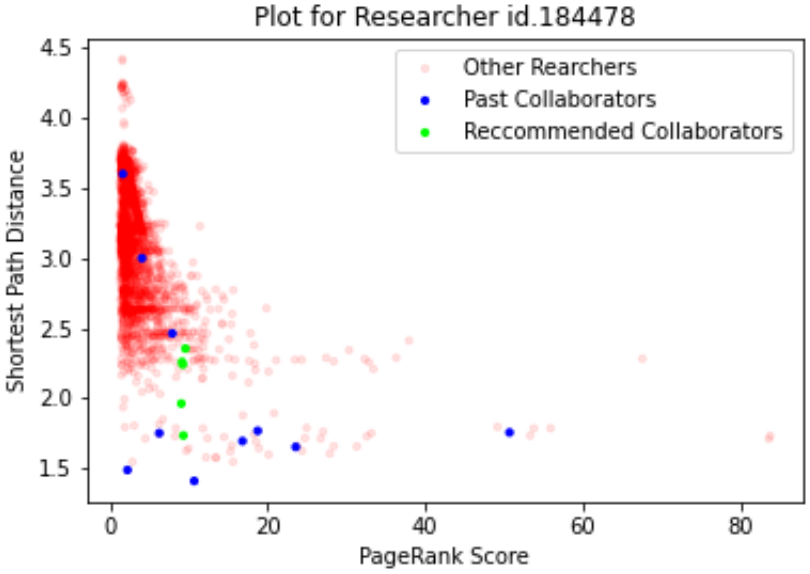

Figure 2: Recommendation plot for researcher 184478

Figure 3: Recommendation plot for researcher 191291

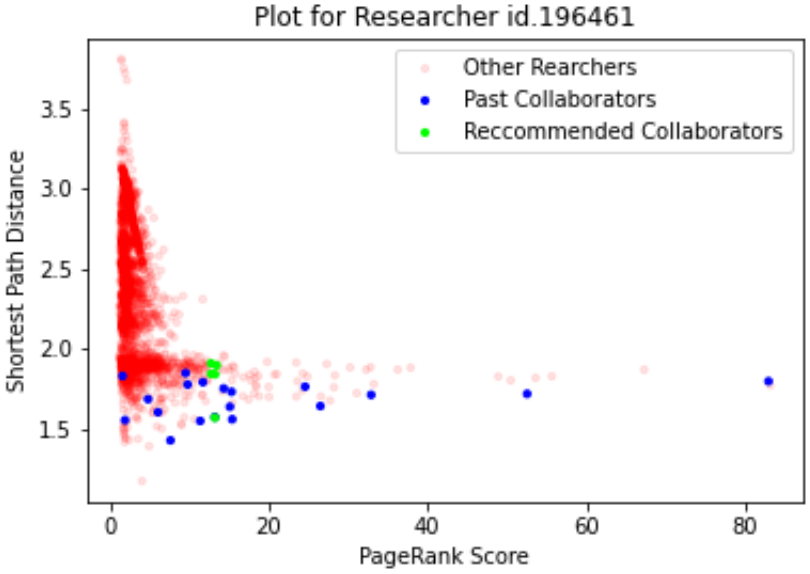

Figure 4: Recommendation plot for researcher 196461

Figure 5: Recommendation plot for researcher 270291

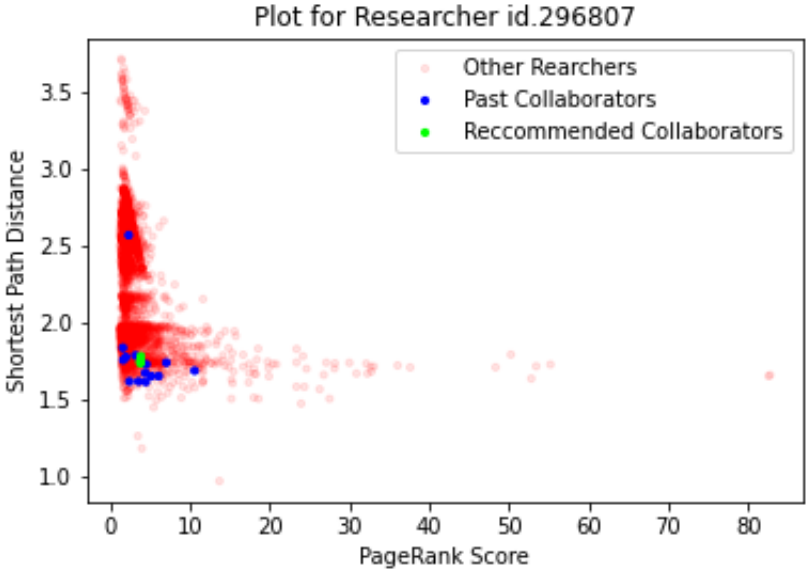

Figure 6: Recommendation plot for researcher 296807

4 Limitations and Further work

The AQL queries are long and slowing down the code and not every information is contained within the DB. For example to know if two researchers have collaborated before, we need to check their edges with publications nodes and then check if some of those nodes have the same id. One could imaging creating edges between the researchers and setting as the score the number of time they have collaborated.

Another information that could be interesting to add to this project would be to know whether the researchers are on the same laboratory or not. This would allow to recommend extra-laboratory collaboration, promoting inter disciplinary researches.

References

[1] Wikipedia contributors. *Dijkstra's algorithm — Wikipedia, The Free Encyclopedia*. 2021.

[2] W. Wang J. Li, F. Xia and H. Jiang. *Acrec: A co-authorship based random walk model for academic collaboration recommendation*. 2014.

[3] P. Santi F. Murray M. Claudel, E. Massaro and C. Ratti. *An exploration of collaborative scientific production at MIT through spatial organization and institutional affiliation*. 2017.

[4] Jun Yanc Panpan Zhanga, Tiandong Wangb. *PageRank centrality and algorithms for weighted, directed networkswith applications to World Input-Output Tables*. 2021.

[5] Lawrence Page Sergey Brin. *The Anatomy of a Large-Scale Hypertextual Web Search Engine*. 1998.