The swiss scientific social network

Vanessa Braglia

May 6, 2018

Contents

1 Introduction

A social network consists of a set of objects connected to each other by social relations. The best way to model social networks is using graphs (see an example in Figure 1): the objects (entities) are represented as nodes and the connections as edges between two different nodes. The most common example we can take is the World Wide Web (WWW) where we have web pages as nodes connected by hyperlinks, the edges.
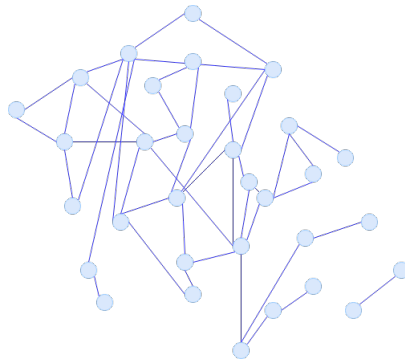

Figure 1: Example of social network graph

The goal of this project is to create a social network of computational science authors belonging to Swiss institutions and then analyze the relative graph. The first step is retrieving all necessary information for the construction of the social network: names of the authors and relationships between each other. I crawled the 2015,2016 and 2017 PASC Conferences, interdisciplinary conferences which brings together research across the areas of computational science, high-performance computing, and various domain sciences, and SIAM conferences of several years, selecting the topics relevant to us (e.g. Optimization, Parallel Computing,...).

The second step is the information analysis to find out relations between institutions but also between members belonging to the same institution.

With PageRank algorithm we obtain a "ranking" of all conferences' partici-

pants: PageRank is an algorithm used by Google Search to rank websites in their search engine results but it can be applied to any social network. In this project I use the algorithm to measure the importance of institutions' members considering the number and quality of their collaborations.

I use Graph Partitioning to "invert" the process and obtain the institutions from members collaborations: probably members of the same institution collaborate more between each other than with other institutions' representatives.

I also analyze the institutions' connectivity matrices and their structure: looking at the cliques (i.e. a sub matrix where every two distinct members collaborate with each other; this means that all entries of the sub matrix are ones) present in the matrices we can for example detect the different research areas of the institutions and the connection between them.

The results will provide an interesting picture of the different research scenarios in Switzerland and how they interact with each other.

2   PageRank Algorithm

PageRank is an algorithm used by Google Search to rank websites in their search engine results; it was developed by Larry Page and Sergey Brin (the two founders of Google) in 1996 as part of their research project at Stanford University.

PageRank is used, together with other algorithms, to measure the importance of web pages and sort them by popularity in the result; it can be used in any graph to compute the importance of each node with its PageRank value. Larry Page and Sergey Brin algorithm works by counting the number and quality of links to a page to compute a value which represents its importance; more popular a page is, more likely it receives links from other websites and more likely from important websites.

As we can see in Figure 2, node C has higher PageRank than E even if it
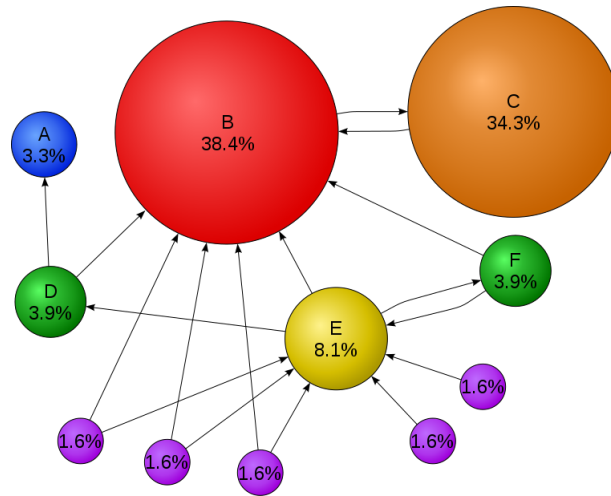
4

Figure 2: PageRank example

has lower number of inner links; this is because the only inner link C has is from the most important node of the graph B so it is considered more than all the inner links of E from purple lower nodes.

Imagine surfing websites, going from one page to another: to avoid problems with pages without outer links we have two different methods to choose the next page. Our surfer can:

- randomly choose an outgoing link from the page,

- simply choose a random page from the Web.

We assume that with probability p (typically p=0.85) our surfer follows the first option and with probability $1 - p$ the second.

The probability that an infinite random surfer visits a specific website it's called its PageRank; a page will have high rank if other pages with high rank link to it.

2.1 How to compute PageRank values?

To apply PageRank, from a graph of n nodes we construct a n-by-n connectivity matrix G: $g_{ij} = 1$ if and only if node i and j are connected. The number of non-zeros in G is the total number of connections in the graph. The row r_i represents the number of inner-links of page i, instead column c_i the number of outer-links. So we can define:

- In-degree of page i: $\qquad r_i = \sum\limits_{j=1}^{n} g_{ij}$

- Out-degree of page i: $\qquad c_i = \sum\limits_{j=1}^{n} g_{ji}$

Let p be the probability that the surfer follows a link and $1-p$ the probability that it chooses a random page.
$\delta = \frac{1-p}{n}$ is the probability that a particular random page is chosen.
We construct a transition probability matrix A where all elements are positive, smaller than one and sums of columns are equal to one; column j represents the probabilities to go from page j to all other pages. So we set:

$$
a_{ij} = \begin{cases} \frac{pg_{ij}}{c_j} + \delta & \text{if } c_j \neq 0 \\ \frac{1}{n} & \text{if } c_j = 0 \end{cases}
$$

If the page j has no outer links it means that column j of A assigns equal probability $\frac{1}{n}$ to all its elements.
We can now compute PageRanks values solving the homogeneous linear system:

$$
x = A\,x
$$

The transition matrix A can be written as

$$
A = pGD + ez^T
$$

where D is a diagonal matrix with the reciprocals of the out-degrees

$$d_{jj} = \begin{cases} \frac{1}{n} & \text{if } c_j \neq 0 \\ 0 & \text{if } c_j = 0 \end{cases}$$

z is the vector formed by

$$z_j = \begin{cases} \delta & \text{if } c_j \neq 0 \\ \frac{1}{n} & \text{if } c_j = 0 \end{cases}$$

and e is the vector of length n with all ones.

We can write the linear system

$$(I - A)x = 0$$

as

$$(IpGD)x = \gamma e$$

where we take $\gamma = z^T x = 1$. WHY?

To conclude, we find the solution of

$$(IpGD)x = e$$

and then rescale the solution x so that $\sum_{i=1}^{n} x_i = 1$.

3   Graph Partitioning

Cliques[1]

4   Results

4.1   Information Retrieval

4.2   PageRank

4.3   Graph Partitioning

5   Bibliography

- https://en.wikipedia.org/wiki/PageRank

- Assignment1 1 Course Numerical Computing 2107/2018

- Assignment2 1 Course Numerical Computing 2107/2018

---

[1]Some text in a footnote.