D1.1 - Description of Work (DOW)

Evolution over time of the structure of social graphs

Student Leonardo Serilli
Supervisors Małgorzata Sulkowska, Nicolas Nisse, Frédéric Giroire

Year 2021/2022

Abstract

Many natural and human made system can be represented by networks, that's it, graphs, sets of nodes and edges, for example the world wide web is just a set of hosts interconnected by data links and social network are made of people and their relationship. Those structures **respect similar mathematical properties** like the power-law distribution, which intuitively says that the number of connections of a node grows proportionally with the time. It is just the way nature want those kinds of networks to evolve: in a **self-organizing way**.
Finding this peculiar characteristic on data we can collect and analyze can brings us to the development of tool to study them, and even to predict, with high accuracy, their future evolution.
The scope of this project is to build a **network of scientific authors and their collaborations**, collected from the **Scopus database**, and to analyze the distribution of their collaborations over time.

Contents

1 General Project Description

1.1 Framework/Context

This project is part of a larger one involving researchers in various field, such as economics, sociology and computer science; it is focused on the evaluation of the impact of funding on scientific research.
As example LabEx and IdEx are french funding programs which scope is to promote collaborations involving different research field.

The purpose of this project is to analyse the evolution of nodes degree in a collaboration network built upon scientific publications extracted from Scopus Database, that is, **the vertex trajectory of the network**.

1.2 Motivations

Many system can be represented as a network, both natural as human built, such as the world-wide web, social networks, collaborations of actors in movies, or even the interaction among molecules. Each of this systems can be viewed as a set of nodes, routers, computers or people, and a set of edges connecting them, data-link among computers, social relations among people or, as in our case, **collaborations between scientific researchers**.
Those system are of practical interest and **attract researchers in different study fields**, since we reached the computational power to deal with the large amount of data and since we have mathematical models to describe their evolution over time. This can bring to a wide range of tools and applications to analyze the structure of those systems and ways to predict their evolution.

A common property of those networks is that they use to be **scale-free**, it mean that the probability distribution of degrees of nodes over the network follows a **power-law distribution**. The insight on it is that a quantity, the degree of a node in our own case, varies proportionally to another quantity, which may be the time. Power-law distribution can be seen in many other phenomena such as the frequencies of word in most languages, the sizes of craters on the moon, of Solar flare and so on. This feature is a consequence of the fact that a new node of the network sites preferentially to nodes that already have a huge number of connections. An intuitive example can be that, in social networks, the probability of having new friends is higher for people who already have a lot of them.

All those example shows that "**the development of large networks is governed by robust self-organizing phenomena that go beyond the particulars of the individual systems**"[1].

1.3 Challenges

In this work is analyzed a collaboration network build upon data collected from **Scopus database**. A node in the network represents an author and there is an edge between two authors if they have collaborated at least once on a Scientific paper.
The data is composed of **258145 french computer science authors** and the amount of collaborations they had **since 1990 until 2018**.

The first challenge would be to **avoid working on misleading data**, there can be authors who have published just once in their carrier or other that had a skyrocketing number collaborations for a couple of years before disappearing from the network and so on.
All of this data can bring to a model diverging from the expected one. Another challenge concern dealing with the **lack of temporal step**, from the collected data the only information that can be used for this purpose is the year in which a collaboration

appears, and there are present only 29 years of collaborations.

The final purpose would be be to have bunch of **good metrics to build vertex trajectory** upon the yet cited network.

1.4 Goals

The first goal is to take confidence with the collected data, building a **meaningful dataset** and **analyzing vertex trajectories**, that is, the evolution of node degrees over time. Using simple metrics like the average of the amount of collaborations for authors who started to publish in the same year.

Another goal is about understanding how the structure of the collaboration network varies for authors with a similar vertex trajectory when a subset of them get a funding. The expected behavior is that the funded author will have an increasing in the number of collaborations, but this is not what is usually observed in real data.

The final Goal of the project is to have methods to **build meaningful vertex trajectories** and to extract useful data from them.

2 State of the Art

Let $G = (V, E)$ be a graph of $|V| = n$ nodes, and let $n_k$be the number of nodes of degree k then Power-law distribution $P_k$is defined as follows:

$$P_k = \frac{n_k}{n}$$

$$P_k \sim Ck^{-\lambda}$$

where $\lambda > 0$ is an exponential parameter and $C > 0$ a scaling constant.
A network whose probability distribution of degrees of nodes respects the power law distribution is said to be **Scale-Free**.
In past decades researchers from different fields have worked for establish the scale free properties of networks. Observing this property of real-world network makes possible to develop theoretical models to study those networks.

An example of possible applications can be a tool to generate random networks having the same structure of the observed one, as the **Barabàsi–Albert[1] model** does, which is used to generates scale-free networks. In order to built his model Barabàsi mapped the topology of a portion of the Web observing that some nodes, called hubs, has an higher number of connections, and with it, an higher probability to develop connections with new nodes in future.

Recent studies shows that scale-free networks are not so widespread as thought, the majority follows a **power-law distribution with an exponential cutoff** of the form:

$$P_k \sim Ck^{-\lambda}\gamma^k$$

where $0 \leq \gamma \leq 1$ is a constant parameter of the distribution.

3 Workplan, Tasks and Milestones

3.1 Tasks

- Retrieving the collaboration dataset for computer science researchers since 1990 until 2018.
- Filter out inactive authors, indeed those that have no publication for 3 years since their first publication .
- Experiment with similar metrics as the one described above in order to refine the definition of active author.
- Splitting the dataset in 28 subset, one for each year, each containing all authors that have their first collaboration in the corresponding years.
- Computing the average vertex trajectory for each dataset obtained in the previous step.
- Visualize the data obtained in previous steps to bring out inconsistencies in the refining process.

3.2 Workplan and Milestones

- Collaboration dataset
- Refined Collaboration datasets
- Dataset splitted by starting publication year
- Average vertex trajectories for authors who started collaborating in the same year
- Plots about obtained data

References

[1] A.-L. Barabasi and R. Albert, "Albert, R.: Emergence of Scaling in Random Networks. Science 286, 509-512," Science (New York, N.Y.), vol. 286, pp. 509–12, Nov. 1999, doi: 10.1126/science.286.5439.509.

[2] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," Phys. Rev. E, vol. 70, no. 6, p. 066111, Dec. 2004, doi: 10.1103/PhysRevE.70.066111.

[3] "Scopus preview - Scopus - Welcome to Scopus." https://www.scopus.com/home.uri.