Exploring the Crunchbase dataset to detect high potential startups EE558 - Network Tour of Data Science

Timothée Bronner
Alexandre Carlier
Saleh Gholam Zadeh
Clément Luneau

November 28, 2017

1 Aim of the project

Since only a few years, the startup ecosystem has totally disrupted the way we think of innovation. Agile methodologies and lean management have enabled the rapid launch of new services or products to a very wide public and the total amount of money spent on fundraising by venture capitalists has skyrocketed.

In this context, it seems crucial for investors to be able to spot high potential startups at an early stage. By analyzing data of previous years, we may be able to find out some patterns that could predict which startups will become the next billion dollar company (unicorn).

Crunchbase is a database consisting of startups, investors and incubators, which stores several thousands of fundraisings. It was initially part of the tech industry media TechCrunch and has become a private entity in 2015. Thus, it seems particularly interesting to study this database for our project.

2 Data acquisition

By registering to the Basic Access program, we can easily use the Crunchbase API, which provides daily updated data. What's more, a complete replica of the whole dataset of Crunchbase as of December 2013 can be downloaded with the Basic Access credentials (2013 Snapshot). This snapshot contains more than 200,000 startup profiles along with 50,000 fundraising events. Thus, we will mainly exploit this complete database and use the API in order to get more up-to-date data if needed.

3 Data exploitation

The database contains precious information about startup profiles, such as the year of creation, the geographical location, the total amount of money raised, the list of investors, etc. From this information, several network structures can be drawn:

- a geographical network: by displaying startups on a map, we may be able to identify clusters, i.e. geographical regions that facilitate venture creations,
- an investor network: in this network, startups are linked if they share common investors. Thus, we can segment the different profiles of venture capitalists based on the startups in which they have invested,
- a similarity network: by choosing well-adapted features, we can define a distance between startups and identify similar startups.