ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

NTDS 2017 Project

Exploring the Crunchbase Dataset to detect high potential startups

Timothée Bronner, Alexandre Carlier, Saleh Gholam Zadeh & Clément Luneau

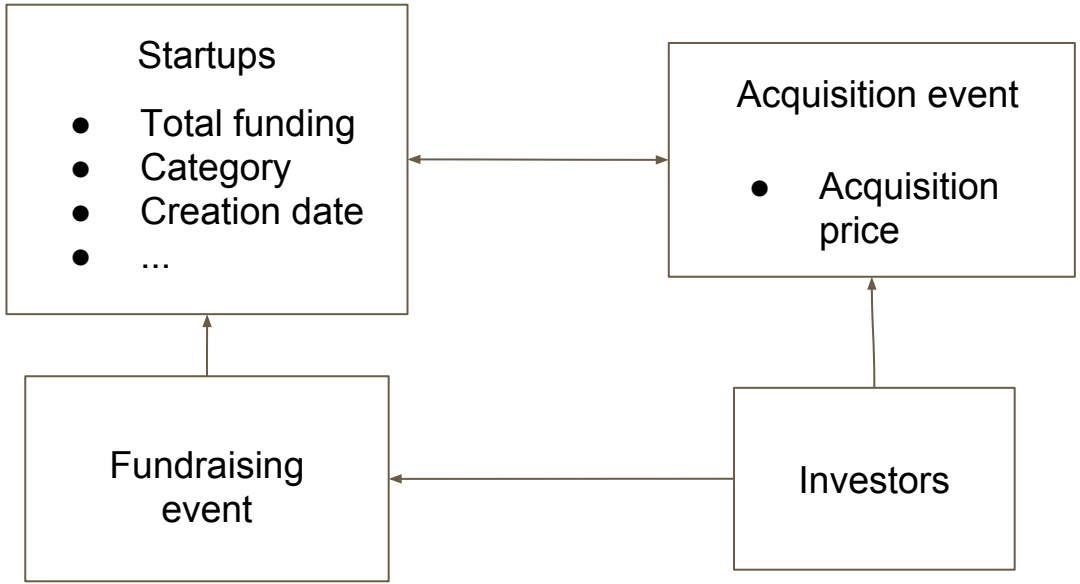
Motivation

- New startups are appearing every day
- Venture Capitalists seek to invest in the most profitable ones
- Crunchbase 2013 Snapshot:
  - 200,000 companies
  - 80,000 investments

Goal: Advise investors thanks to Data Science!

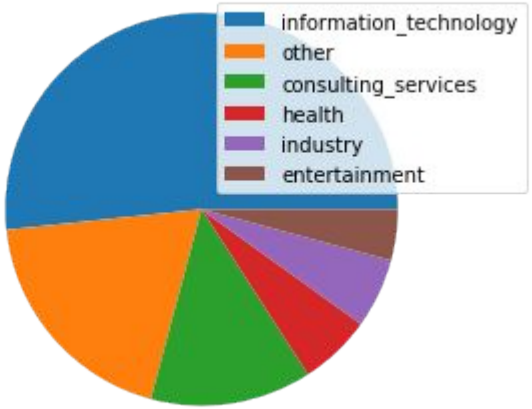
Data exploration

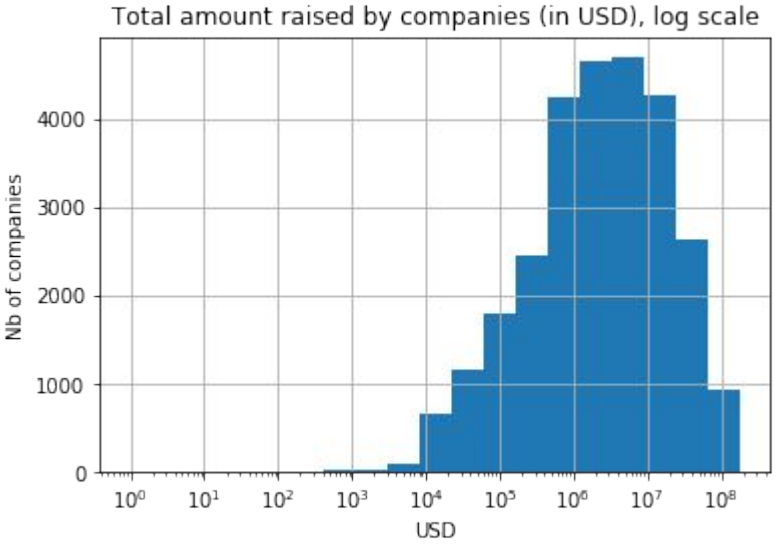
Agents : companies, financial organizations, persons (angel investors).



Data exploration

42 categories

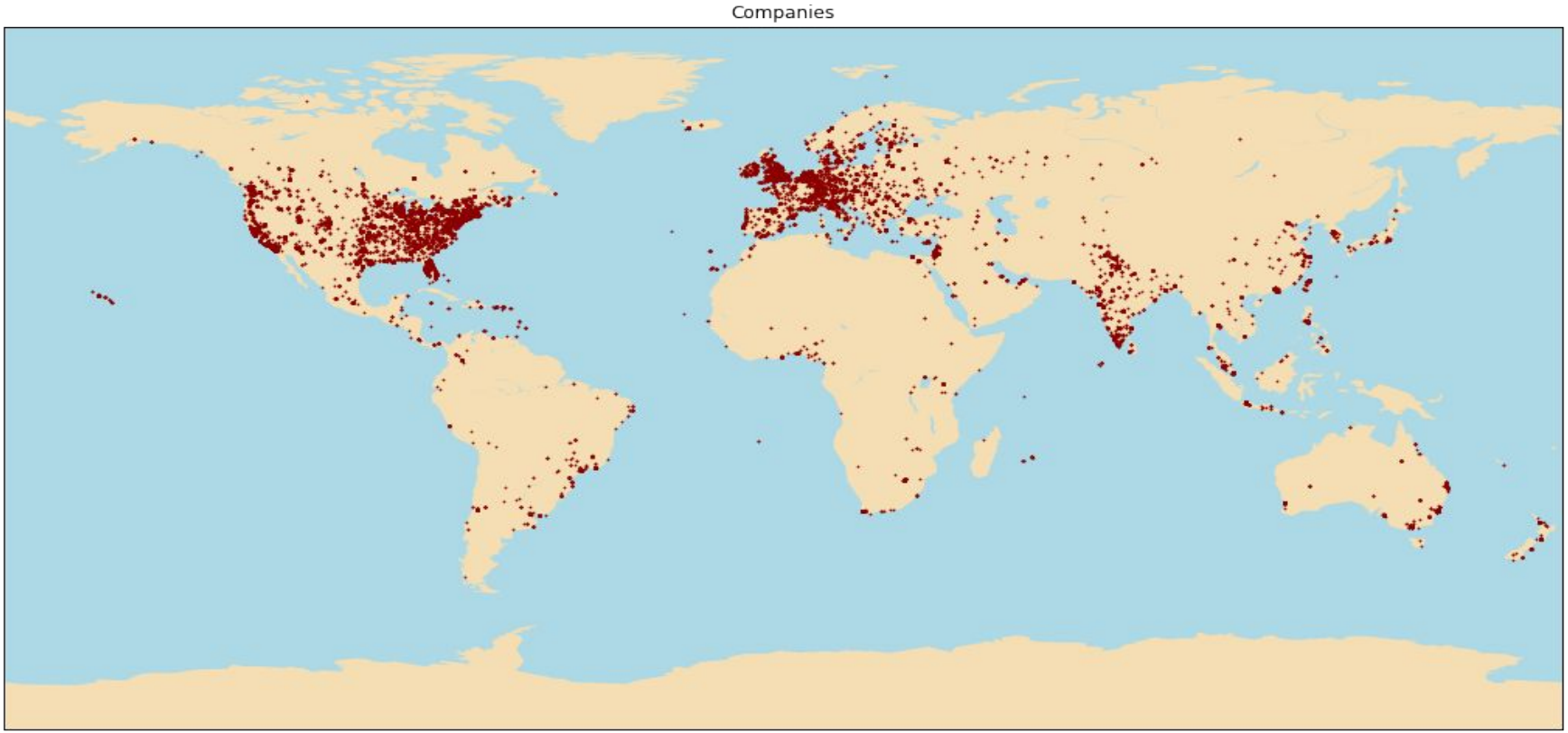




How can we find successful companies to invest

- Find good companies to invest in.
  - Good means that we can sell our shares with a large capital gain (high ROI).
- We have many features in our dataset but:
  - Many of them are irrelevant or difficult to exploit
  - Some of them are raw and we should combine them
- We added these features to our dataset:
  - Return on investment (acquisition/total funding)
  - Latitude and longitude via the Google Maps Geocoding API
- What are our important features:
  - Return on investment (it is known for the some companies)
  - Number of common investors
  - Total funding of the company
  - Location maybe has some effect

Geographical distribution

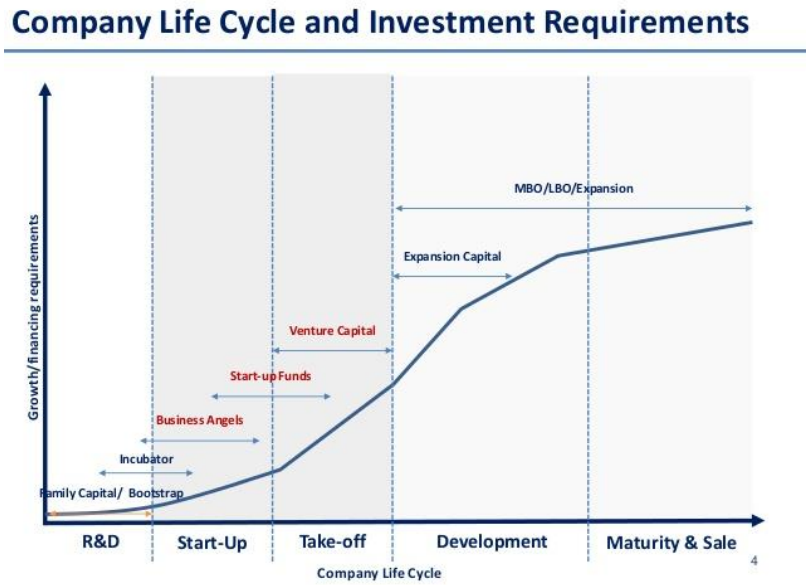


Data cleaning

- Remove companies with missing information:
  - Total funding
  - Location

- Also remove the ones with insufficient funding

From the 80,902 initial investments, 40,963 are kept.



Source: www.slideshare.net/manishvirgo/introduction-to-private-equity-venture-capitalist-fund-42180988

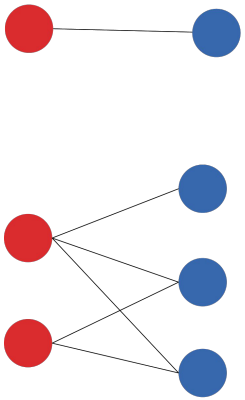
Graph creation

Too few features to create similarity graph.
Instead:

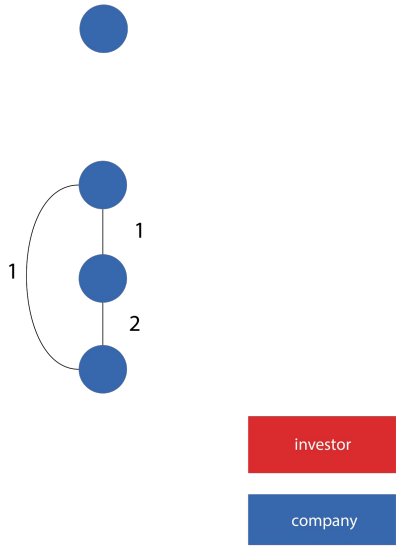
1. Bipartite graph: link each investor to the companies he invested in
2. Investment graph: link the companies with a common investor, edge weight is the # of common investors

Solely the investment graph is used in the data exploitation.

Bipartite Graph



Investments Graph

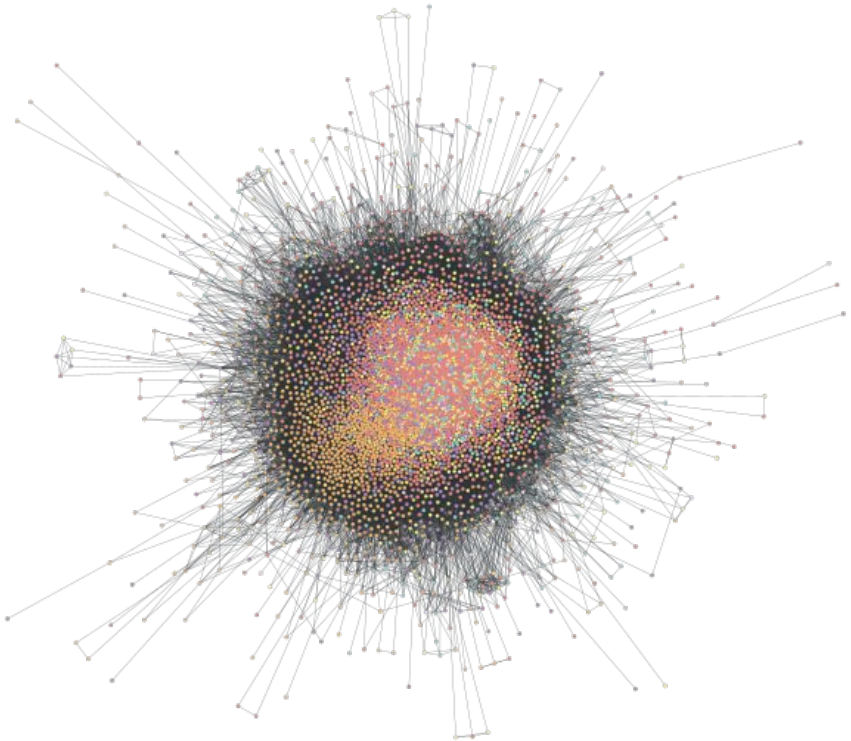


Giant component

Some companies are isolated.

Focus on the connections => keep only giant component

- **Diameter** 8
- **Nb edges** 473,351
- **Nb nodes** 5,933



Obtain the Return On Investment (ROI)

Depends on the status of the company:

- **Acquired** Get ROI directly from total funding and acquisition price
- **Closed** ROI set to 0
- **Operating** Set to a risk factor (arbitrary set to 1 for now)
- **Initial Public Offering (IPO)**

  Valuation obtained with:
  - CSV files provided by Crunchbase
  - Call of the API of “Intrinio”

Data exploitation

Goal: Predict companies with a high return on investment.

Assumptions:

- we know a few companies that have a high return on investment;
- someone who invested in a company acquired with a high ROI has a good insight on the market, i.e. the other investments (s)he made might be good too.

Idea: Use the Graph Fourier transform to spread onto the graph a signal first localized on successful companies.

Heat propagation

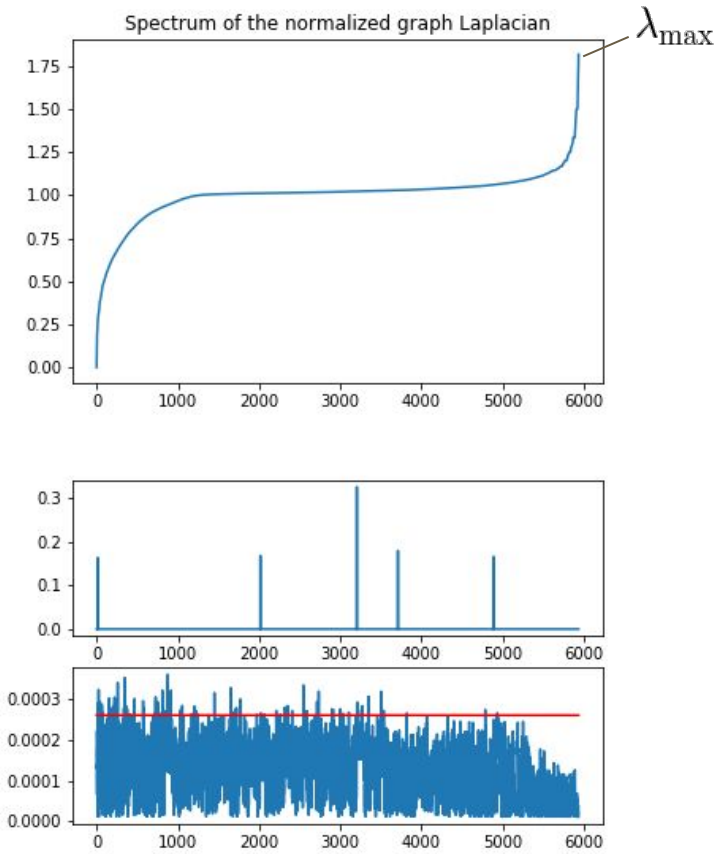
- **Signal at time t=0**
  Sum of weighted deltas localized on 5 vertices corresponding to companies with the highest ROI.

- Diffusion with a heat kernel:

$$\hat{g}(\lambda_\ell) = \exp\left(\frac{-\tau\lambda_\ell}{\lambda_{\max}}\right), \quad \tau = 50 \ .$$

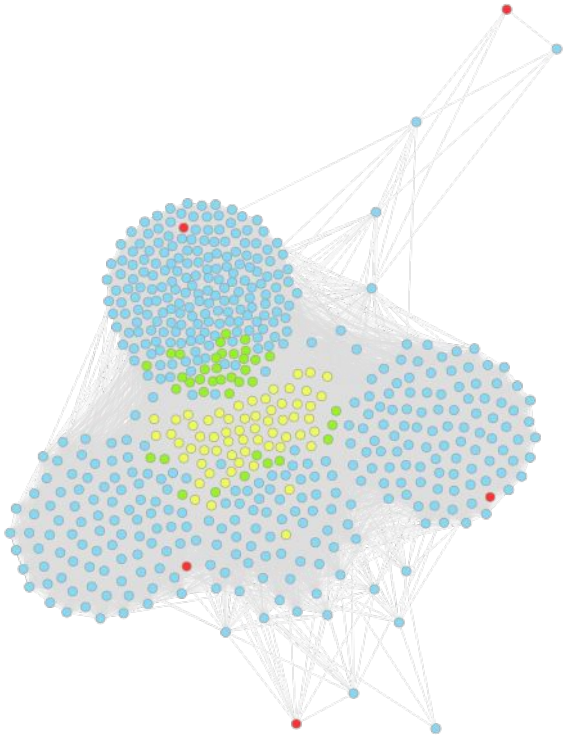
- From a given vertex, the signal propagates more towards neighbors with many common investors.

- Advise investing in the companies (here 100) where the signal is the strongest.

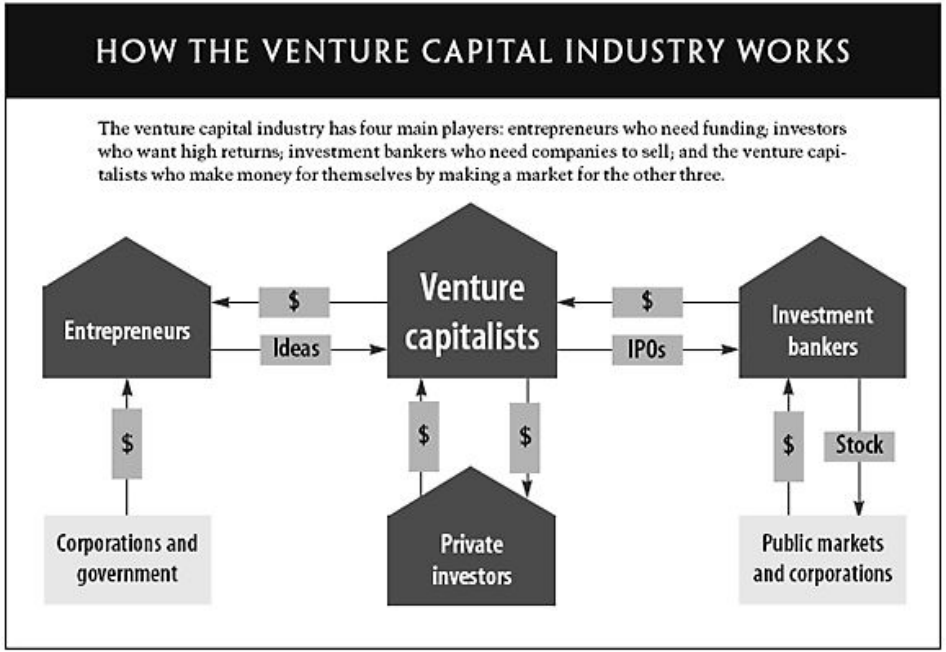


Picked companies on the graph

- **Red** Starting companies (high ROI)
- **Green** Chosen companies being neighbor of one of the starting companies
- **Yellow** Chosen companies being not direct neighbor of a starting company
- **Blue** Neighbors of a starting companies that are not chosen



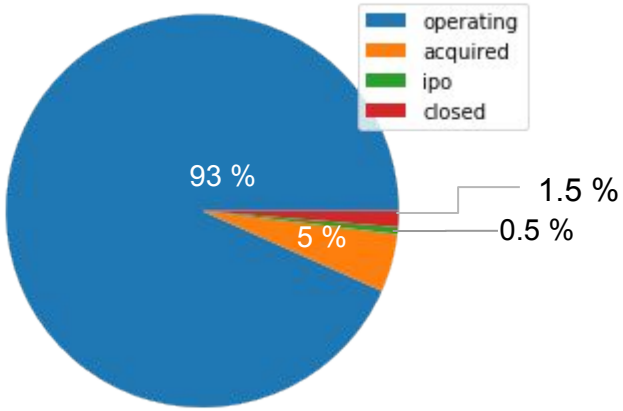
Results



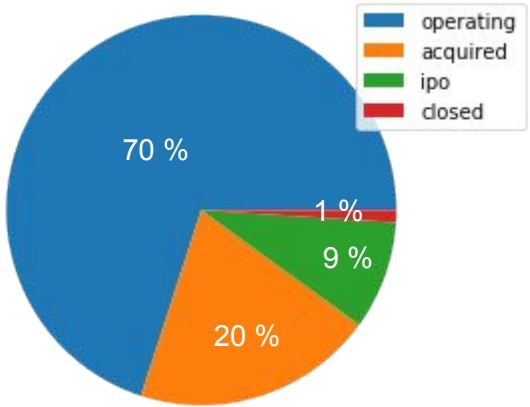
Source: hbr.org/1998/11/how-venture-capital-works

Results

Distribution on the entire dataset



Distribution on the selected companies



Some “discovered” companies

Spotify

airbnb

UBER

KAYAK

stripe

$40M → $9B

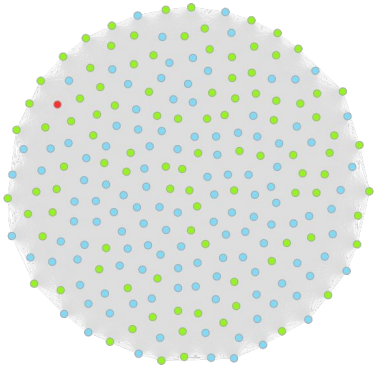
DOLLAR SHAVE CLUB

$22M → $1B

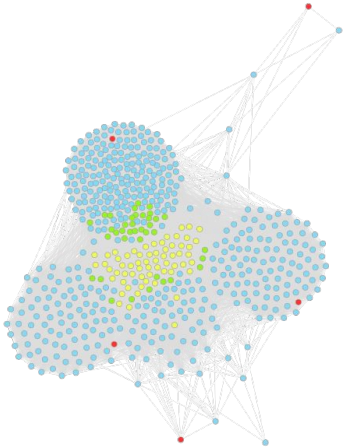
Tuning the hyperparameters

Number of starting companies

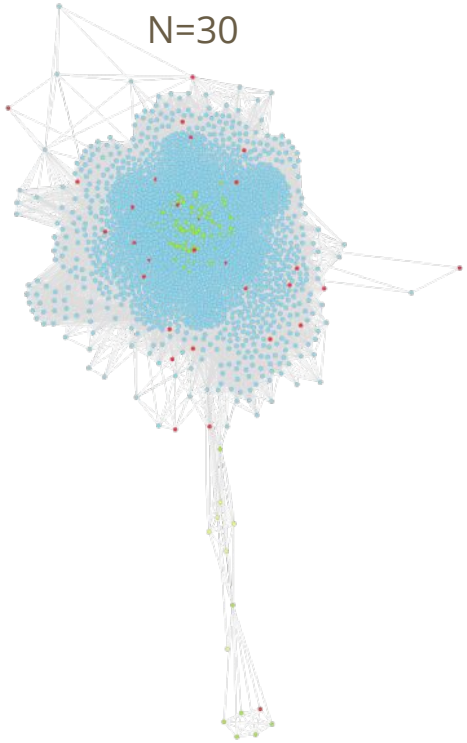
N=1



N=5



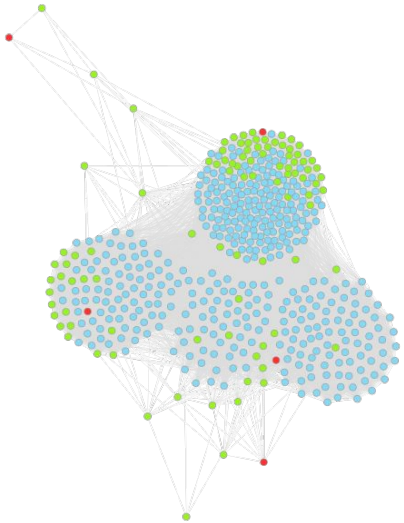
N=30



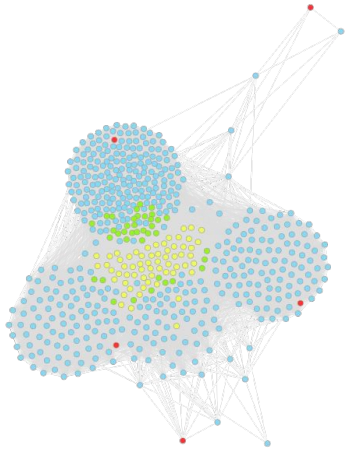
Tuning the hyperparameters

Scaling parameter of the heat kernel

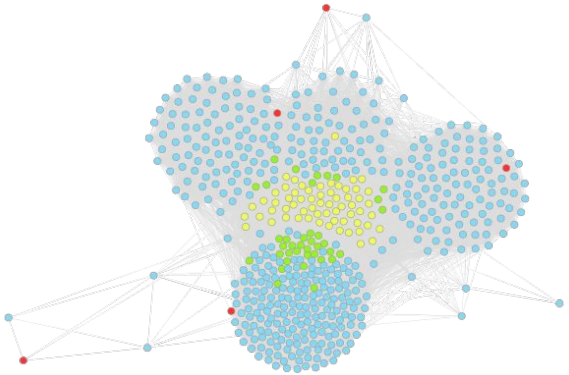
$\tau = 1$



$\tau = 50$

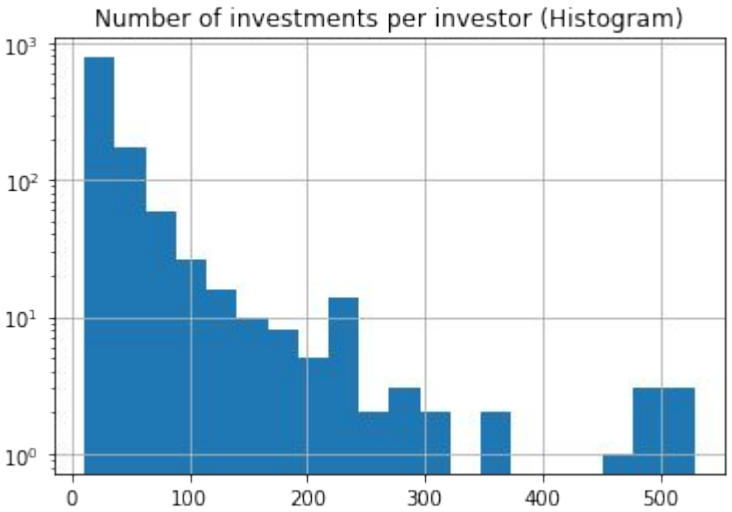


$\tau = 1000$



Tuning the hyperparameters

Number of selected companies



=> Average number of investments per investor: 40

Returns On Investment for different scenarios

Nb. selected companies / Nb. starting companies	10	40	100
1	4.3	4.0	3.2
5	4.3	4.1	3.2
30	4.3	4.1	3.0

Risky !

Conclusion

- Graph built from a low number of features, edges/weights come from the investments.

- Use of Graph Fourier transform to look for potentially successful companies. Promising results but...
  - Other kernel could be considered in the future;
  - Method evaluated on a snapshot of the market;
  - A lot of other factors have to be considered by VCs.