
 vs.  python™  

# Users for Data Science

COSC2671 Assignment 2 Presentation

Team: a small world (Chi Ting Low, Kanru Wang & Yong Kai Wong)

Date: 23 May 2018

# Overview

- Goal:
  - to detect communities of R and Python users for data science on Twitter
- Why is it important?
  - To find unicorn who is bilingual in R and Python
  - To find like-minded users to collaborate in some projects
  - To understand the trend in the data science field
- Data scrapping: from a greedy to a narrow scope



reddit



stackoverflow

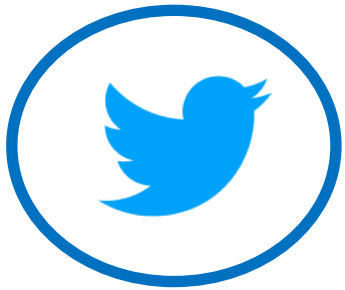
LinkedIn

kaggle



# Overview

- Goal:
  - to detect communities of R and Python users for data science on Twitter
- Why is it important?
  - To find unicorn who is bilingual in R and Python
  - To find like-minded users to collaborate in some projects
  - To understand the trend in the data science field
- Data scrapping: from a greedy to a narrow scope



# Which did we find?

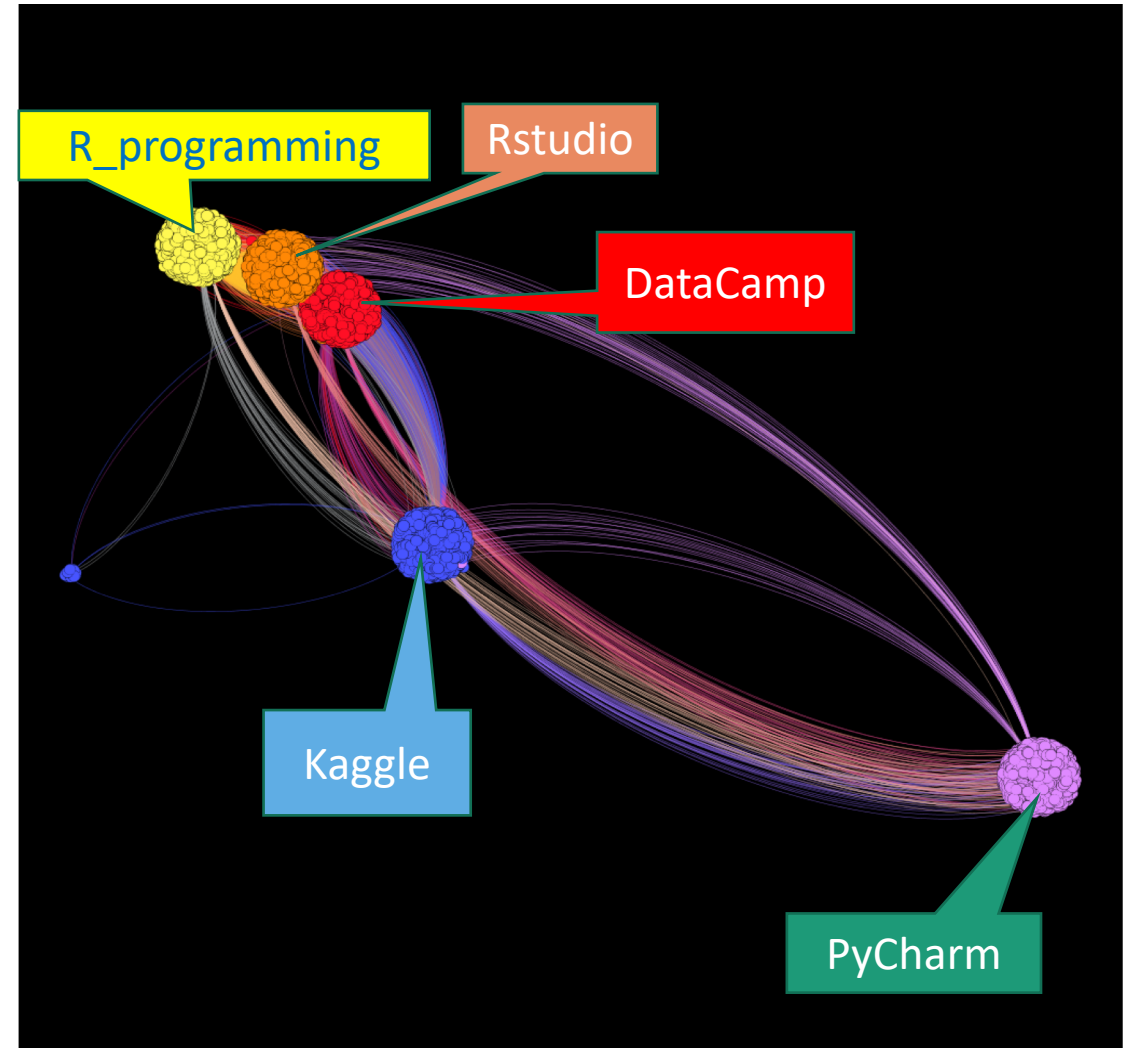
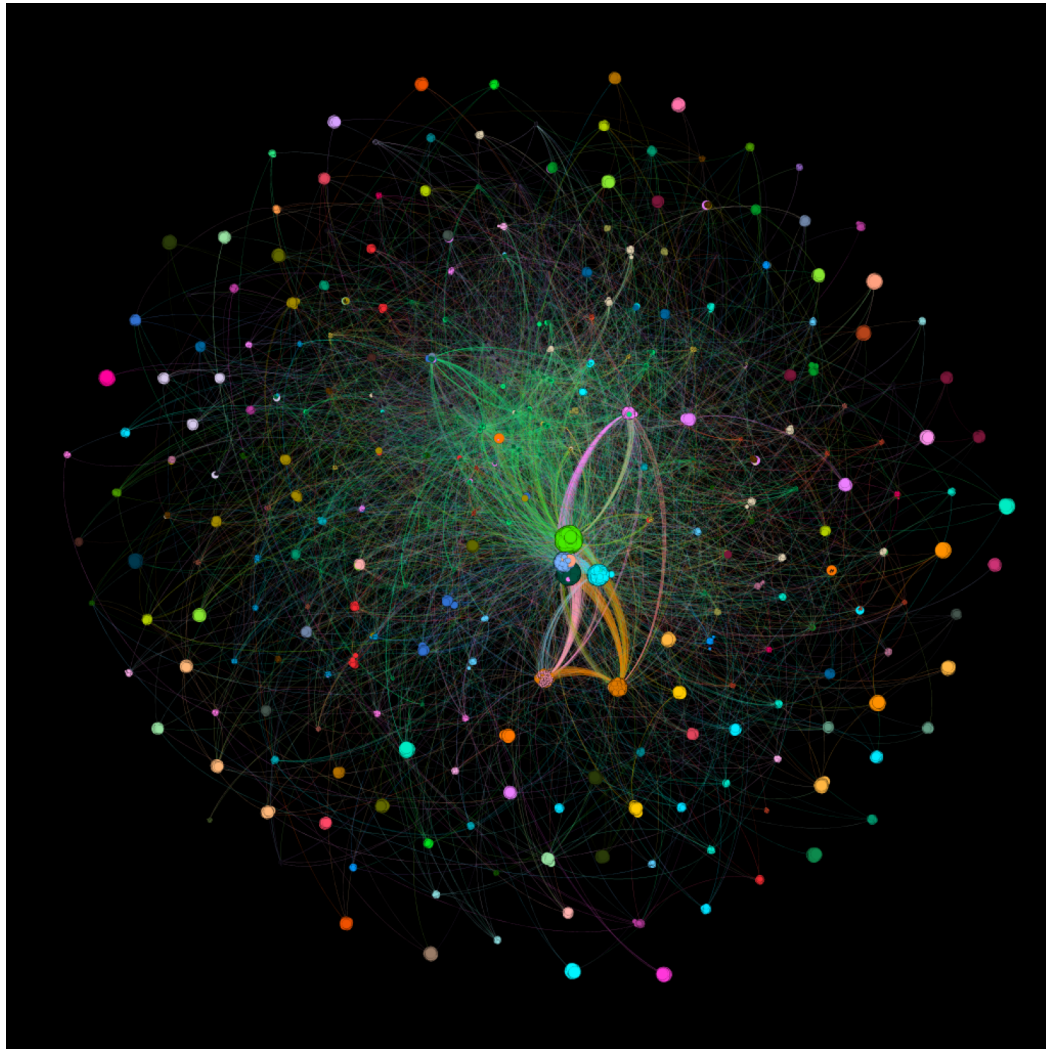
- Data: 48869 nodes (Twitter users) and 56098 edges
- Top 5 most central Twitter users

Degree centrality	R_programming (R and Py Tips), DataCamp, Kaggle, rstudio, pycharm
Eigenvector centrality	Rprogramming, rstudio, hadleywickham, RLangTip, kaggle

- Communities

k-CPM	k = 3 yields 19 communities (5 % of nodes) k = 4 yields 3 communities (< 1 % of nodes) K = 5 yields 3 communities (1 % of nodes)
Louvain's Method	81 groups where top 5 groups account for 45 % of nodes

# Louvain Communities



# Conclusion

- Kaggle attracts users from R and Py
- No clear community which shows “bilingualism” in R and Py
- Louvain (group-based) method yields more interpretable groups
- Caveat: following ‘central’ Twitter account does not mean “fluent” in each language

# FAQ Session