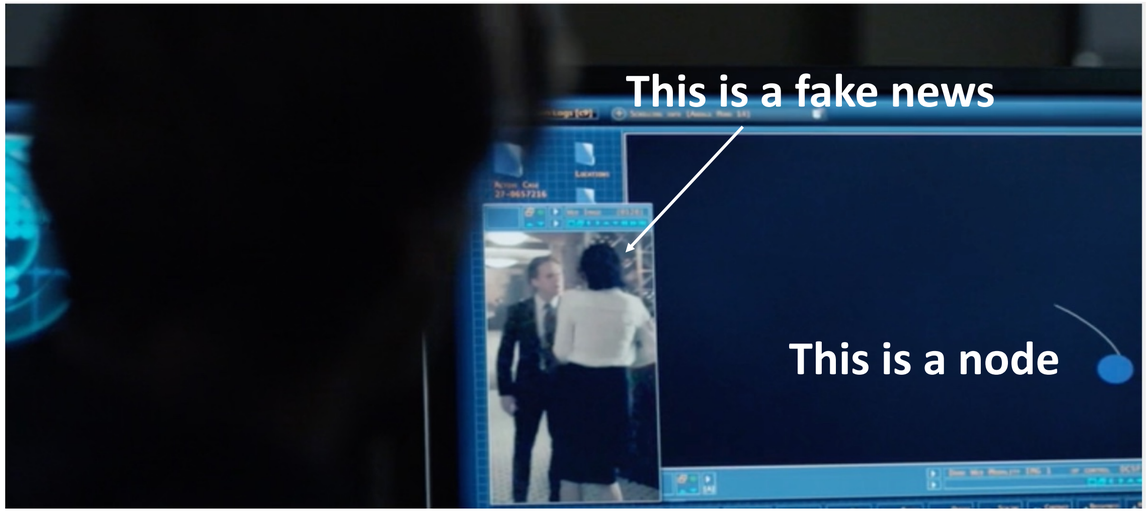
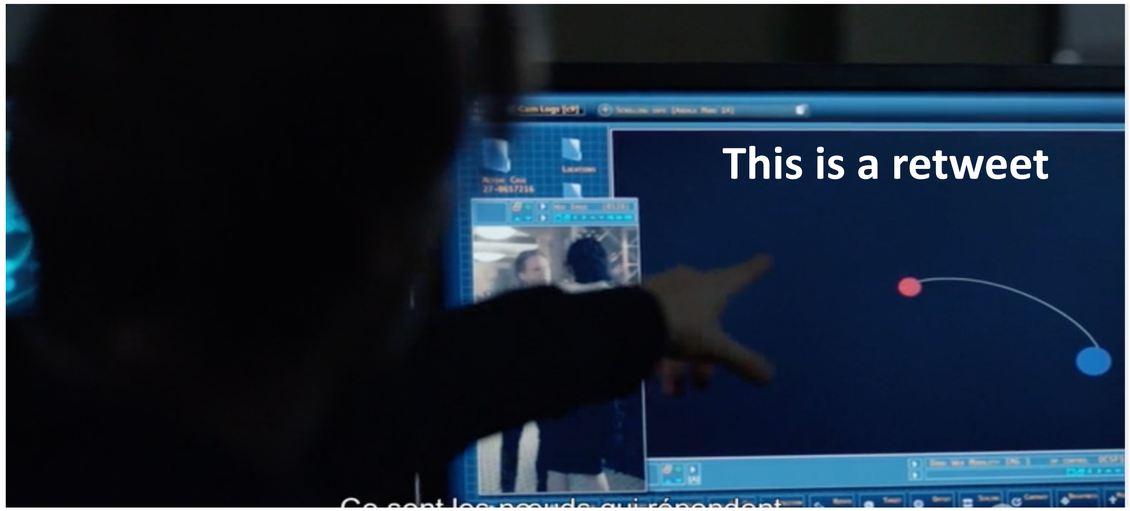
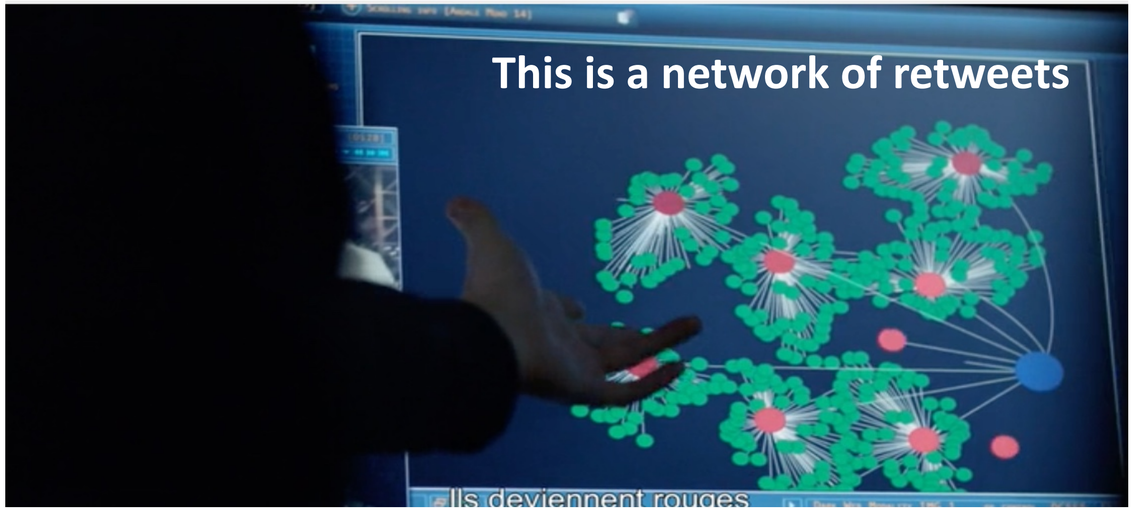
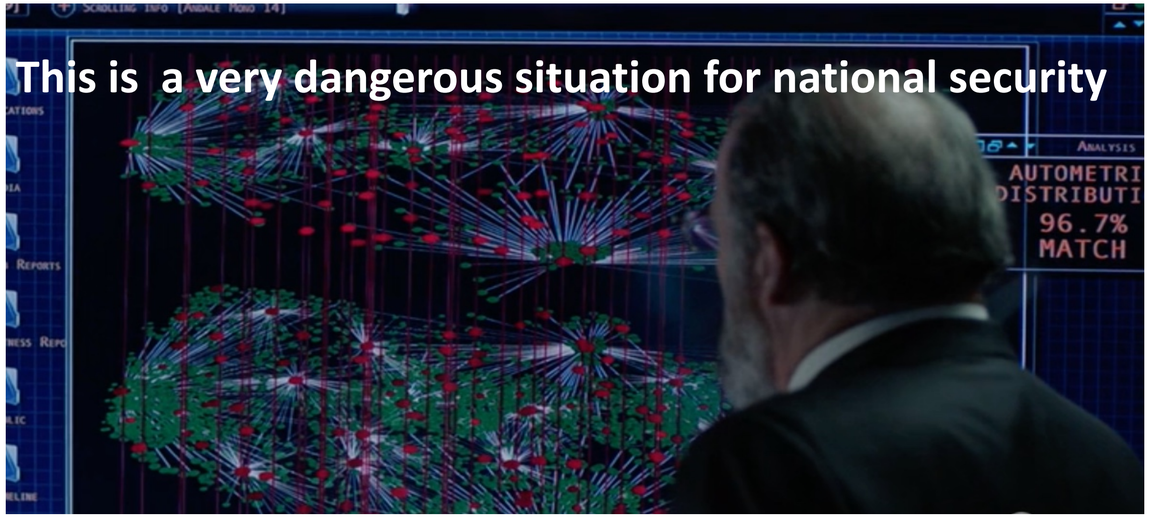
Crawling Twitter Data

Why crawling data on Twitt









What we'll cover today?

01 **Get your Twitter Credentials**

02 **Perform simple queries with R on Twitter**

03 **Community Detection in the @RLadiesBerlin Twitter account**

04 **A small Exercise**

01

Get your Twitter Credentials

01 Get your Twitter Credentials

1 - Register your app with your Twitter account here: https://apps.twitter.com/

2 - Create an app:

Website: Anything starting with **http://** and **not a twitter URL**

Callback: not necessary

01 Get your Twitter Credentials

Berengere_2

Details | Settings | Keys and Access Tokens | Permissions

Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	
Consumer Secret (API Secret)	
Access Level	Read and write (modify app permissions)
Owner	bergautier
Owner ID	2412334885

01 Get your Twitter Credentials

« Generate your Access Token »

Your Access Token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access Token	
Access Token Secret	
Access Level	Read and write
Owner	bergautier
Owner ID	2412334885

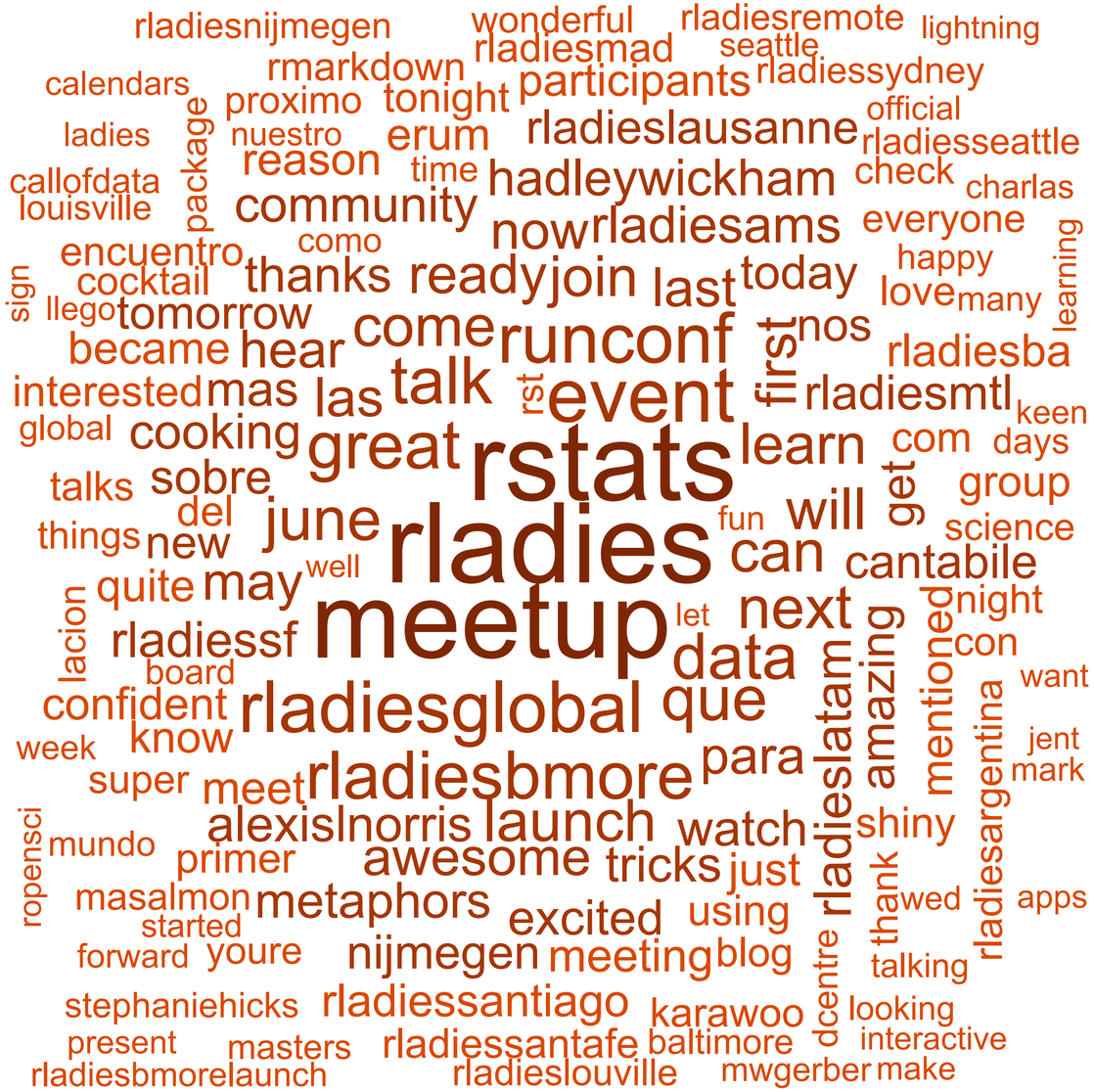
01 Get your Twitter Credentials

key = your_consumer_key
apisecret = your_consumer_secret
accesstoken = your_access_token
accesstokensecret = your_access_token_secret

02

Perform a simple query on Twitter with R

02 Perform a simple query on Twitter with



02 Perform a simple query on Twitter with I

TF-IDF for: Term frequency–Inverse document frequency

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

number of times a term t occurs in document d.

is a measure of how much information the word provides, that is, whether the term is common or rare across all documents

$$= \log \frac{N}{|\{d \in D : t \in d\}|}$$

With:

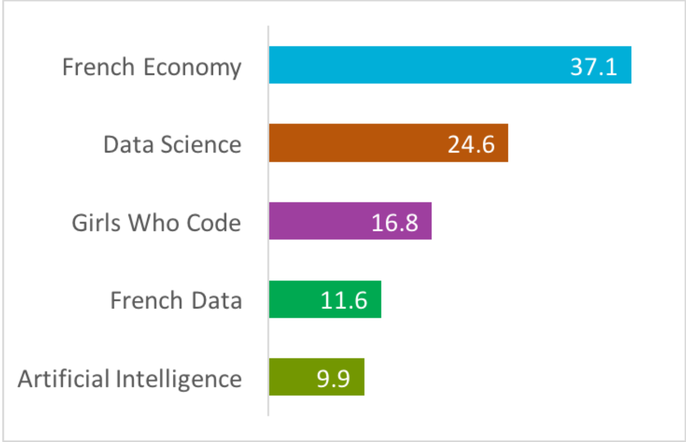
N: total number of documents / Number of documents in our dataset that contains the term

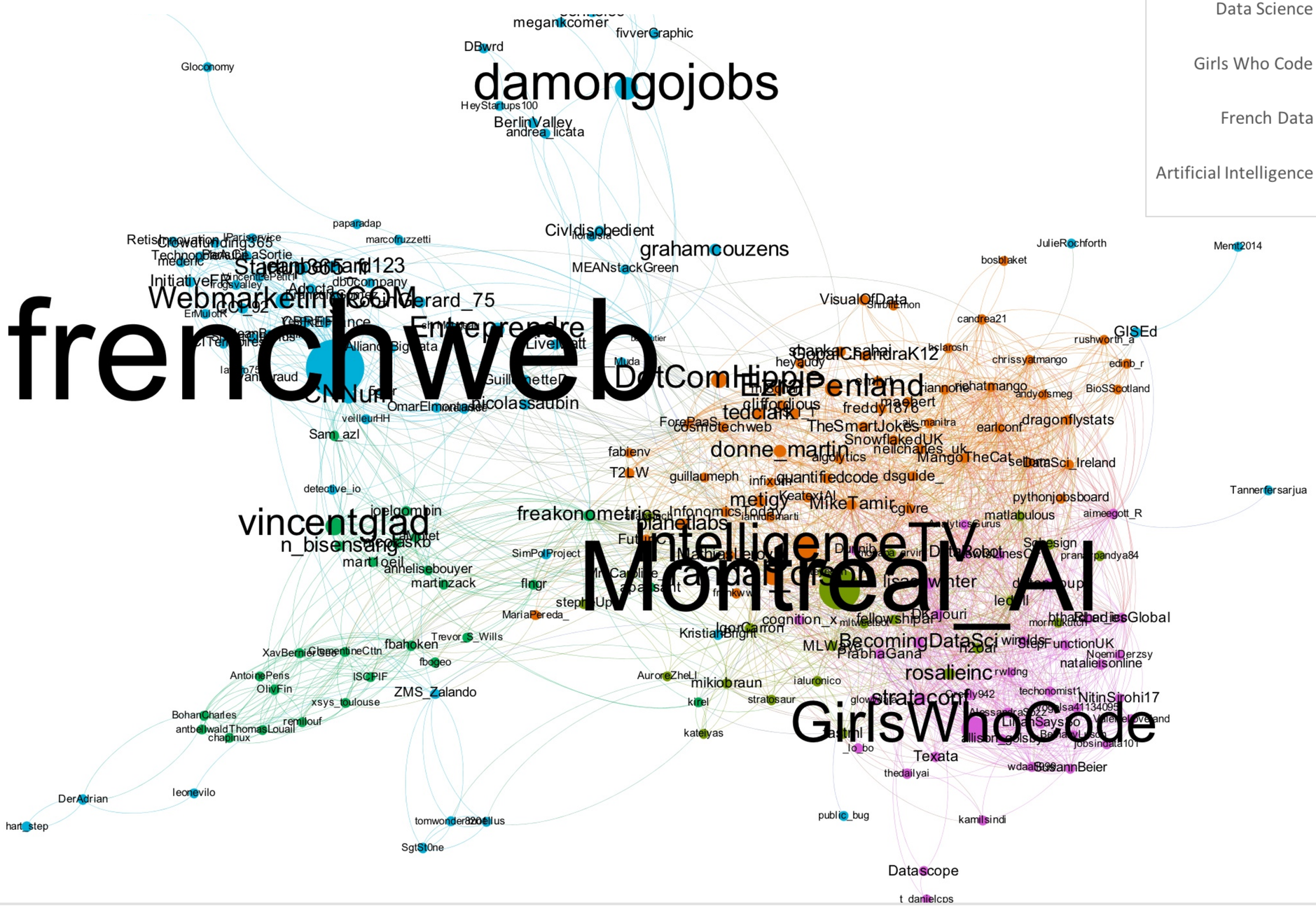
Log to dampen the effect of the idf function

03

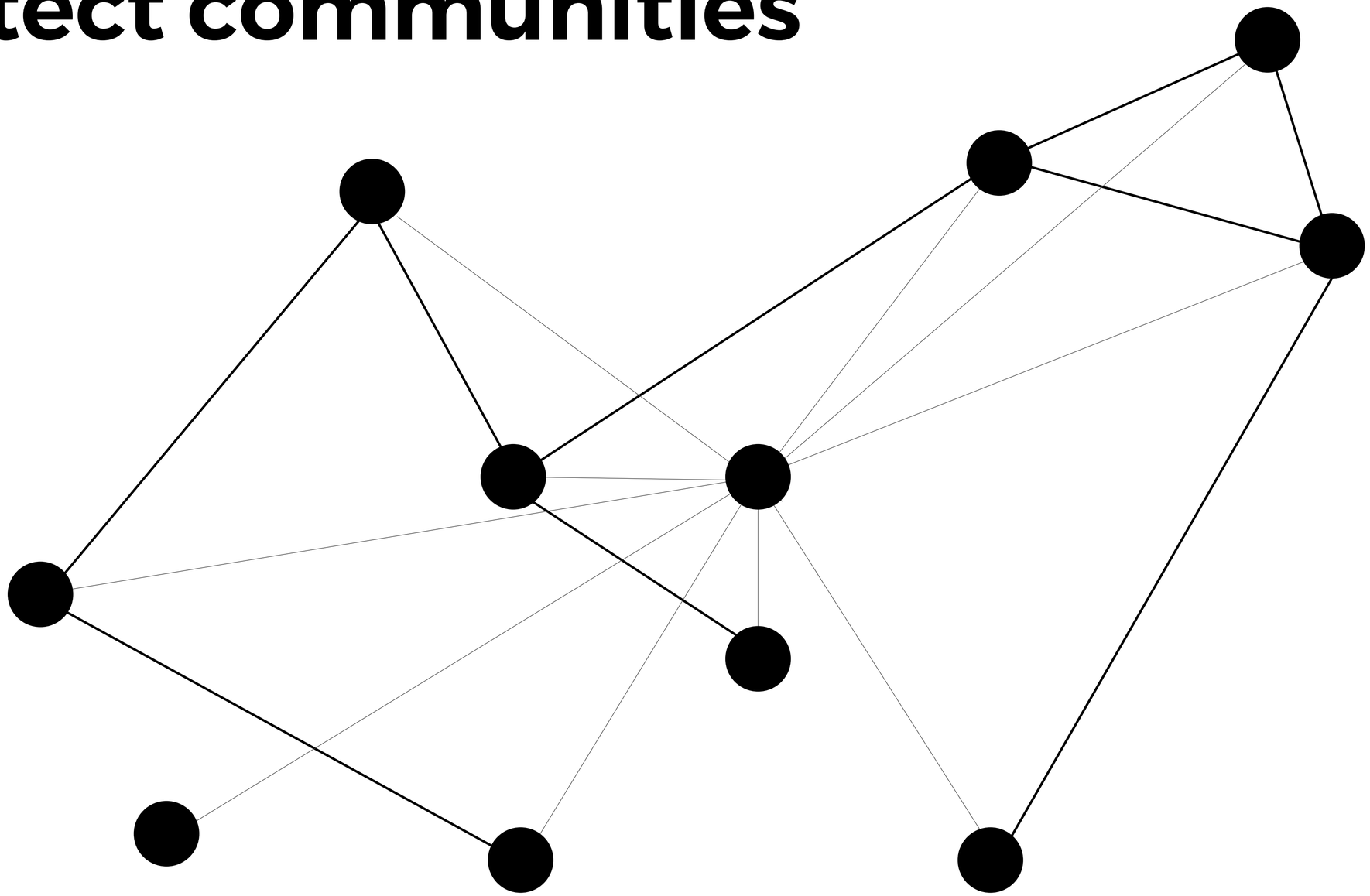
Detect communities in @RLadiesBerlin Twitter users network

03 Detect communities





03 Detect communities



03 Detect communities

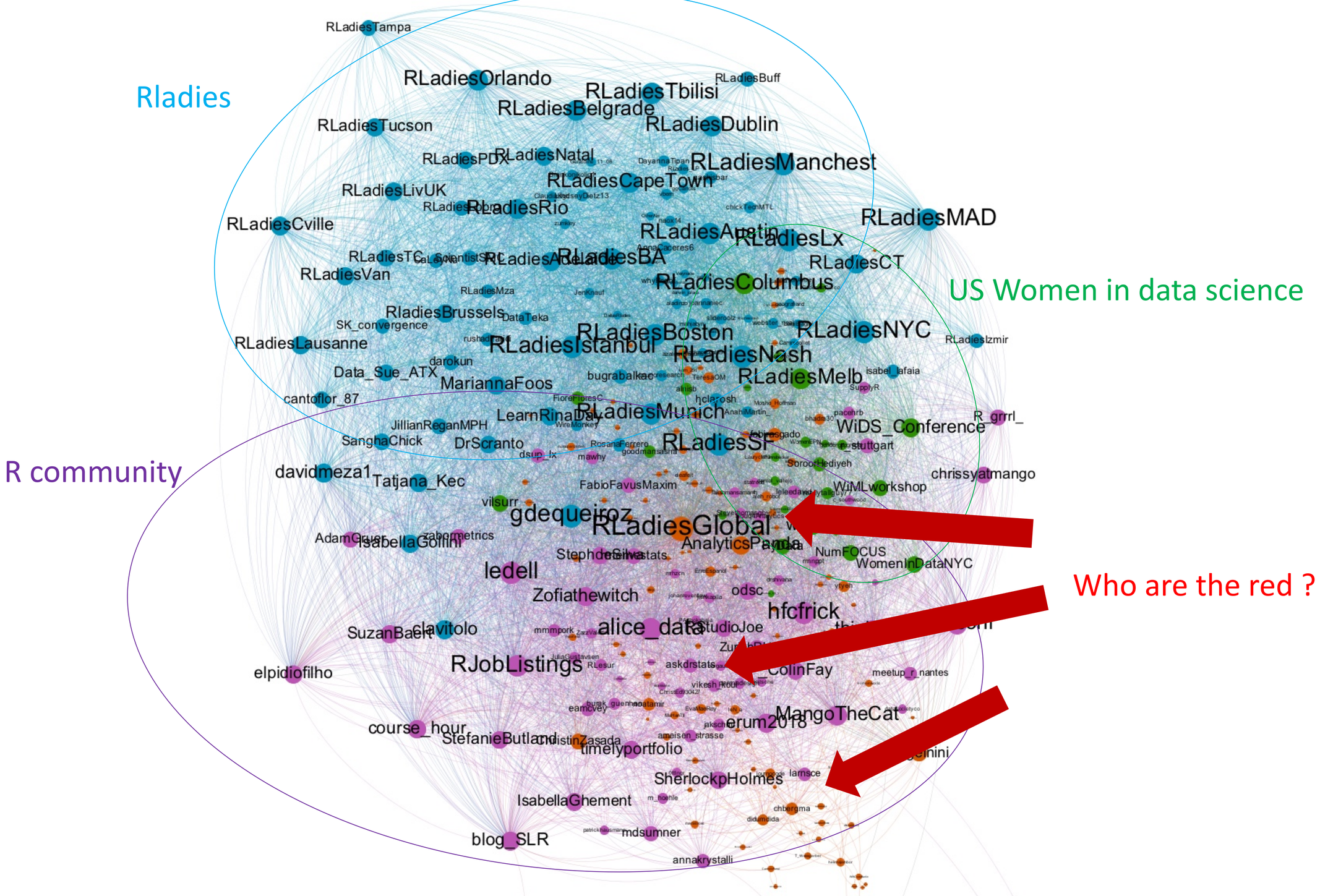
„Modularity“ algorithm (Louvain Method)

- Modularity=fraction of the edges that fall within the given groups minus the expected fraction if edges were distributed at random. [-1,1]
- Initially, each node is assigned to a community on its own.
- In every step, nodes are re-assigned to communities in a local, greedy way: each node is moved to the community with which it achieves the highest contribution to modularity.
- When no nodes can be reassigned, each community is considered as a node on its own, and the process starts again with the merged communities.
- The process stops when there is only a single node left or when the modularity cannot be increased any more in a step.

0

4

Ex: Who are the red people?


Rladies
US Women in data science
R community
Who are the red ?
RLadiesTampa
RLadiesOrlando
RLadiesTbilisi
RLadiesBuff
RLadiesBelgrade
RLadiesTucson
RLadiesDublin
RLadiesPDX
RLadiesNatal
RLadiesManchest
RLadiesCapeTown
RLadiesLivUK
RLadiesRio
RLadiesMAD
RLadiesCville
RLadiesAustin
RLadiesLx
RLadiesTC
RLadiesBA
RLadiesCT
RLadiesVan
RLadiesColumbus
RladiesBrussels
SK_convergence
RLadiesBoston
RLadiesNYC
RLadiesIzmir
RLadiesLausanne
RLadiesIstanbul
RLadiesNash
darokun
Data_Sue_ATX
MariannaFoos
RLadiesMelb
cantoflor_87
RLadiesMunich
LearnRinaDay
JillianReganMPH
WiDS_Conference
R_grrrl_
SanghaChick
DrScranto
RLadiesSF
stuttgart
davidmeza1
Tatjana_Kec
FabioFavusMaxim
WiMLworkshop
chrissyatmango
vilsurr
gdequeiroz
RLadiesGlobal
AdamGruer
IsabellaGollini
zabormetrics
AnalyticsPanda
NumFOCUS
WomenInDataNYC
ledell
Zofiathewitch
odsc
hfcfrick
SuzanBaert
clavitolo
alice_data
StudioJoe
elpidiofilho
RJobListings
askdrstats
ColinFay
meetup_r_nantes
MangoTheCat
course_hour
StefanieButland
ChristinZasada
timelyportfolio
SherlockpHolmes
IsabellaGhement
blog_SLR
mdsumner
annakrystalli

04 Ex: Who are the red people?

- (1) Get the followers of @RLadiesBerlin
- (2) Get their „Description“
- (3) Merge with the community information
- (4) For each community, create a clean text corpus
- (5) For each cleaned text corpus, create a wordcloud
- (6) Who could be the followers in the red