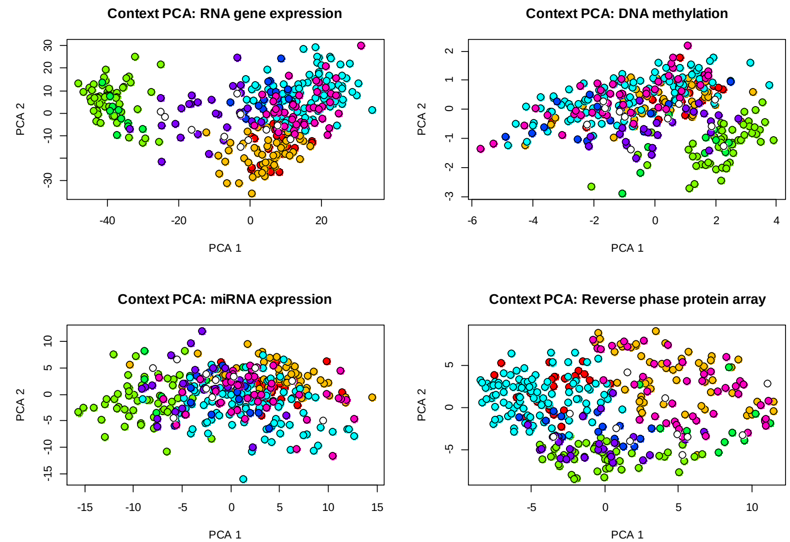
Data science with F#: Analysing social networks

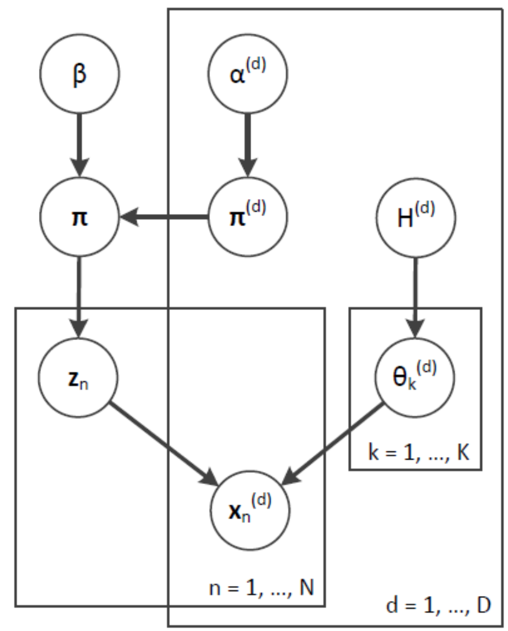
Evelina Gabasova

Twitter @evelgab

Blog evelinag.com



$$
\begin{aligned}
& p\left(\boldsymbol{\pi}^{(d)} \middle| \alpha_d, \beta, \boldsymbol{\pi}\right) \\
& \quad \propto p\left(\boldsymbol{\pi}^{(d)} \middle| \alpha_d\right)\; p\left(\boldsymbol{\pi} \middle| \boldsymbol{\pi}^{(1)}, \ldots, \boldsymbol{\pi}^{(D)}, \beta\right) \\
& \quad = \left\{ \frac{\Gamma\left(K^{(d)}\alpha_d\right)}{\left(\Gamma(\alpha_d)\right)^{K^{(d)}}} \prod_{k=1}^{K^{(d)}} \left(\pi_k^{(d)}\right)^{\alpha_d - 1} \right\} \times \\
& \qquad \times \left\{ \frac{\Gamma\left(\sum_{i_1=1}^{K^{(1)}} \cdots \sum_{i_D=1}^{K^{(D)}} \left[\beta \times \pi_{i_1}^{(1)} \times \cdots \times \pi_{i_D}^{(D)}\right]\right)}{\prod_{i_1=1}^{K^{(1)}} \cdots \prod_{i_D=1}^{K^{(D)}} \Gamma\left(\beta \times \pi_{i_1}^{(1)} \times \cdots \times \pi_{i_D}^{(D)}\right)} \prod_{i_1=1}^{K^{(1)}} \cdots \prod_{i_D=1}^{K^{(D)}} \left(\boldsymbol{\pi}_{i_1,\ldots,i_D}\right)^{\beta \pi_{i_1}^{(1)} \ldots \pi_{i_D}^{(D)} - 1} \right\} \\
& \quad = \frac{\Gamma\left(K^{(d)}\alpha_d\right)}{\left(\Gamma(\alpha_d)\right)^{K^{(d)}}} \times \frac{\Gamma(\beta)}{\prod_{i_1=1}^{K^{(1)}} \cdots \prod_{i_D=1}^{K^{(D)}} \Gamma\left(\beta \times \pi_{i_1}^{(1)} \times \cdots \times \pi_{i_D}^{(D)}\right)} \times \\
& \qquad \times \left\{ \prod_{k=1}^{K^{(d)}} \left(\pi_k^{(d)}\right)^{\alpha_d - 1} \right\} \left\{ \prod_{i_1=1}^{K^{(1)}} \cdots \prod_{i_D=1}^{K^{(D)}} \left(\boldsymbol{\pi}_{i_1,\ldots,i_D}\right)^{\beta \pi_{i_1}^{(1)} \ldots \pi_{i_D}^{(D)} - 1} \right\} \\
& \quad \propto \frac{\Gamma\left(K^{(d)}\alpha_d\right)}{\left(\Gamma(\alpha_d)\right)^{K^{(d)}}} \left\{ \prod_{k=1}^{K^{(d)}} \left(\pi_k^{(d)}\right)^{\alpha_d - 1} \right\} \left\{ \prod_{i_1=1}^{K^{(1)}} \cdots \prod_{i_D=1}^{K^{(D)}} \frac{1}{\Gamma\left(\beta \times \pi_{i_1}^{(1)} \times \cdots \times \pi_{i_D}^{(D)}\right)} \left(\boldsymbol{\pi}_{i_1,\ldots,i_D}\right)^{\beta \pi_{i_1}^{(1)} \ldots \pi_{i_D}^{(D)} - 1} \right\}
\end{aligned}
\tag{18}
$$

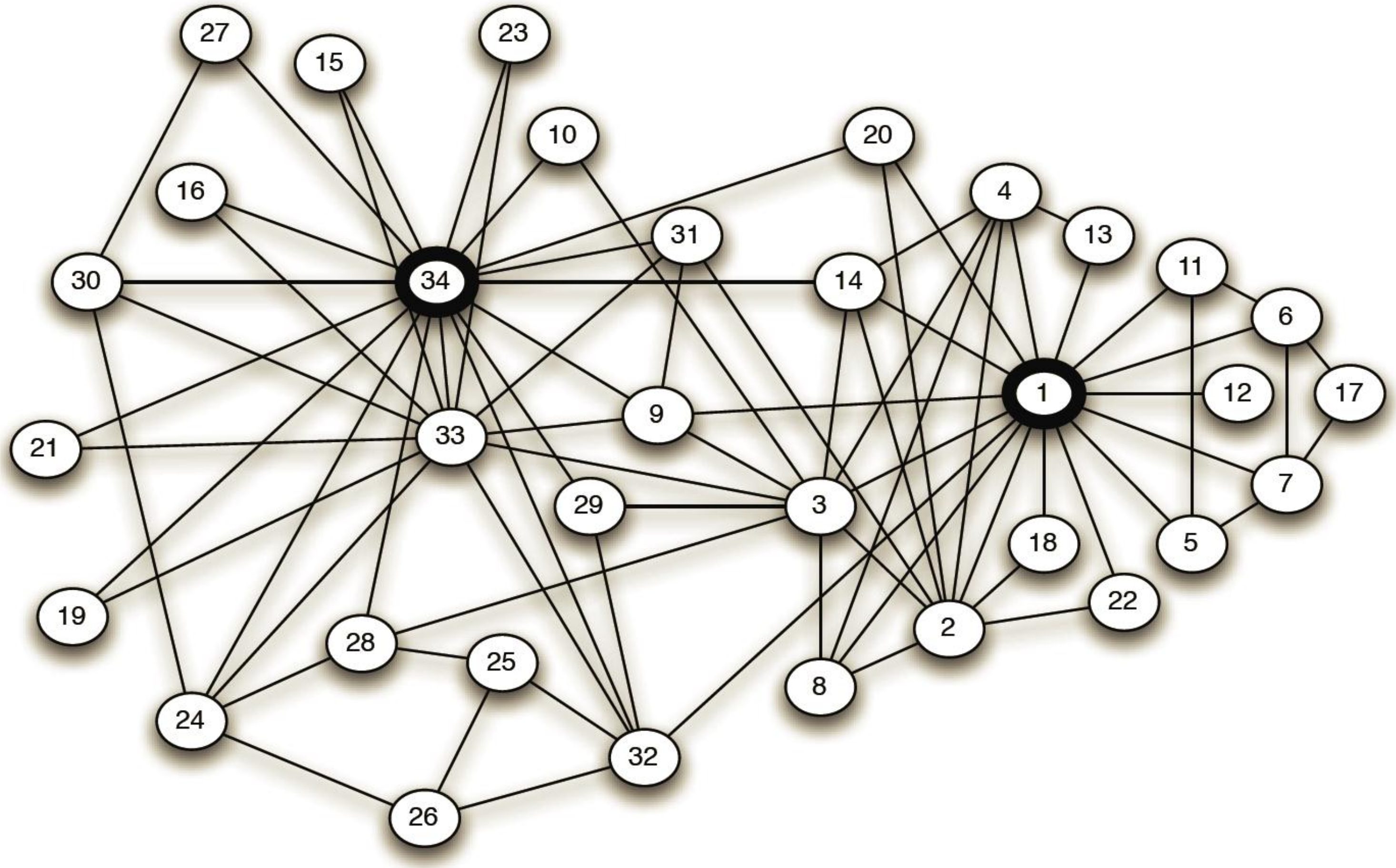


Why network science

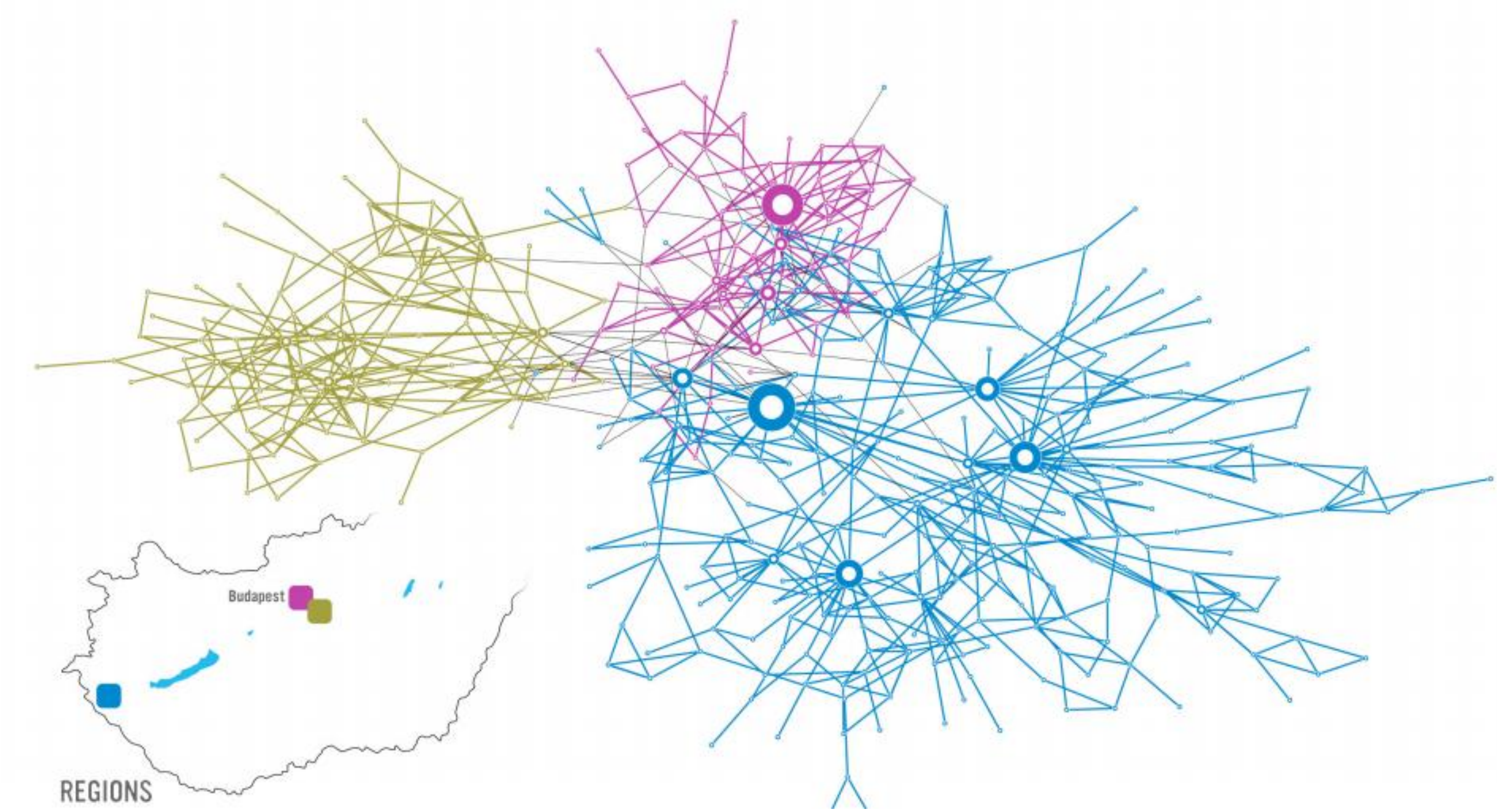
- gene interaction networks
- disease spreading
- cascading failures in power grids
- brain connections
- social networks

Social network analysis

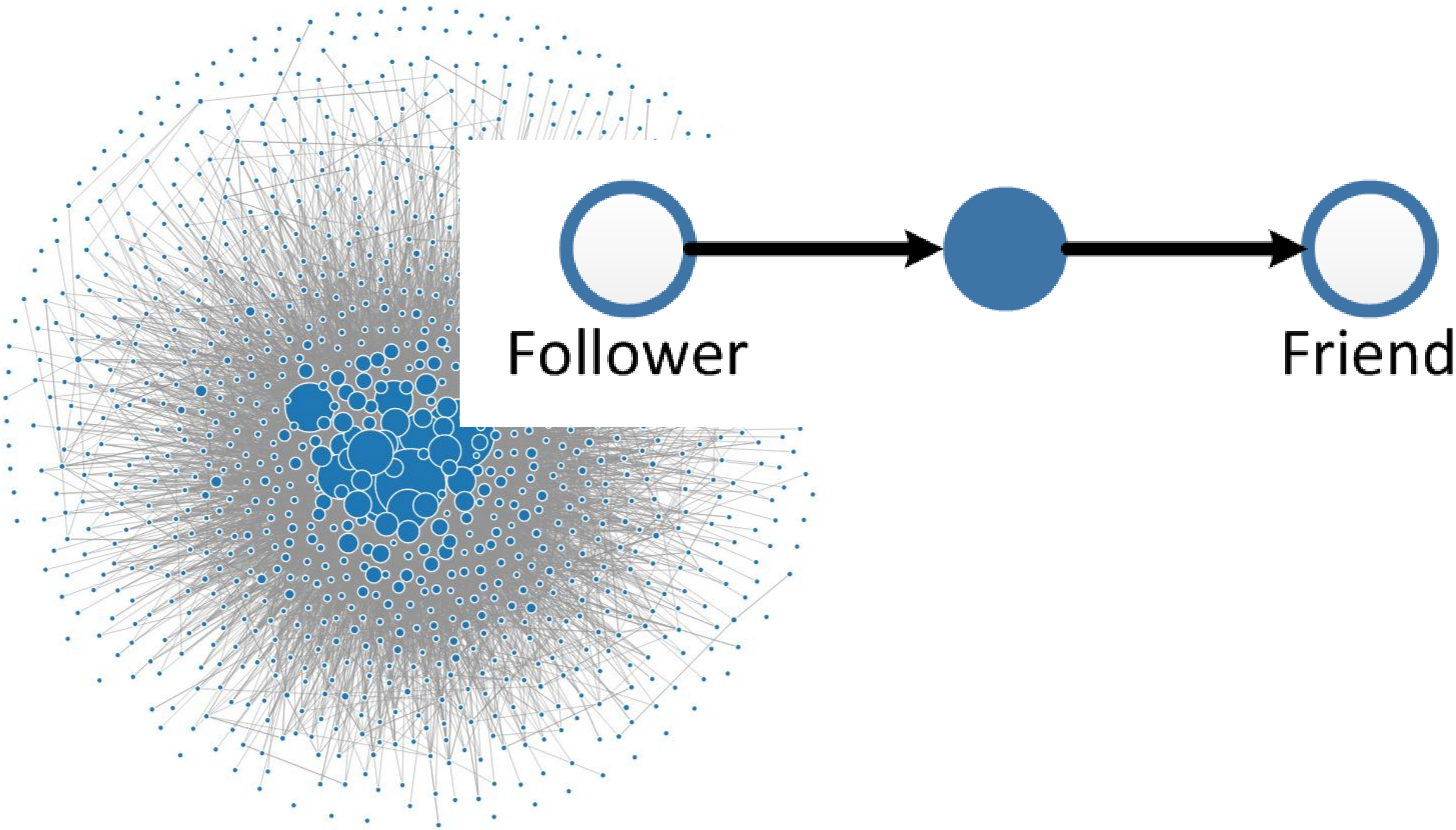
Karate club network



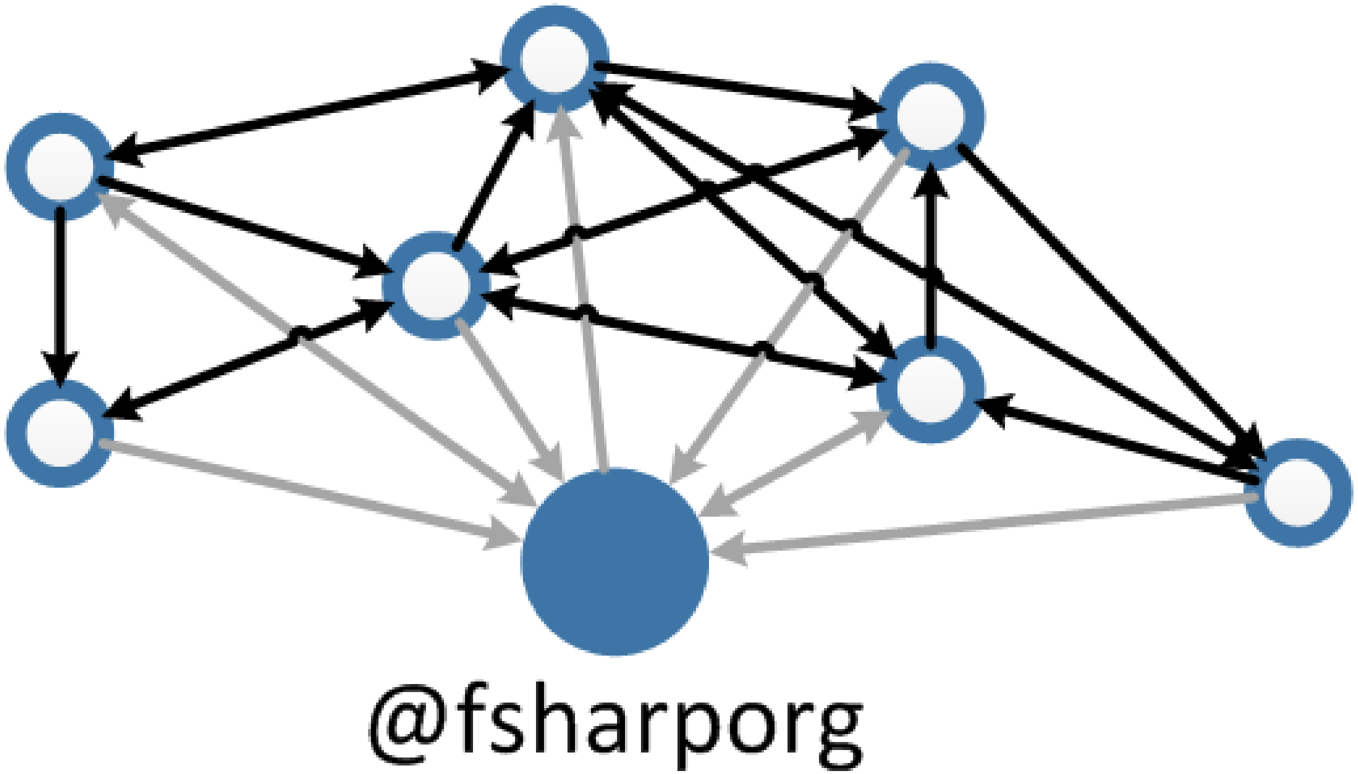
Social network analysis



Twitter networks



Ego network



Downloading data from Twitter

1) List of nodes
2) Connections between nodes

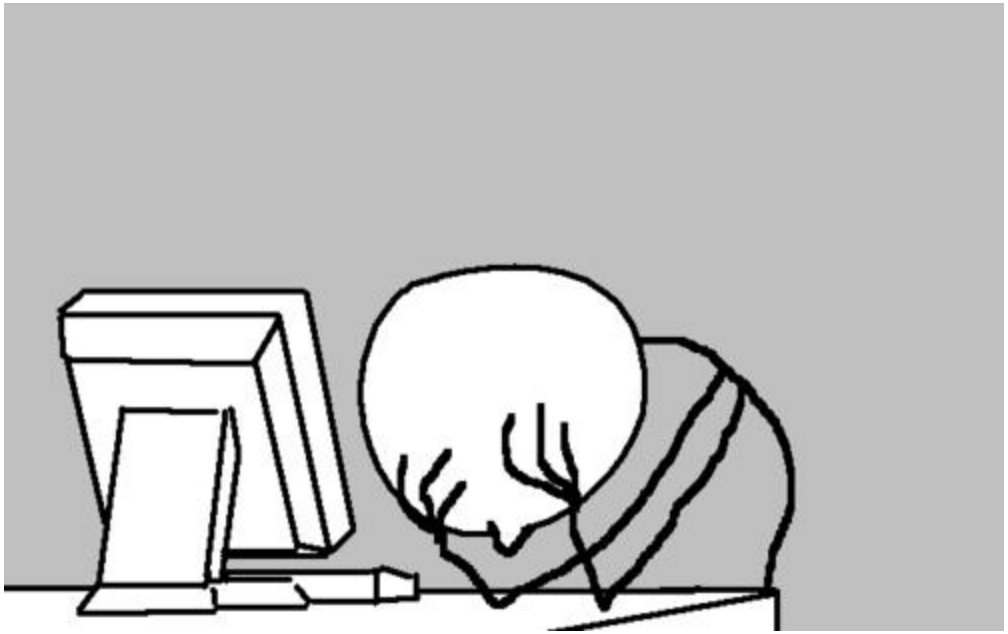
Twitter API allows only 15 requests every 15 minutes to list connections.

DOWNLOADING DATA FROM TWITTER

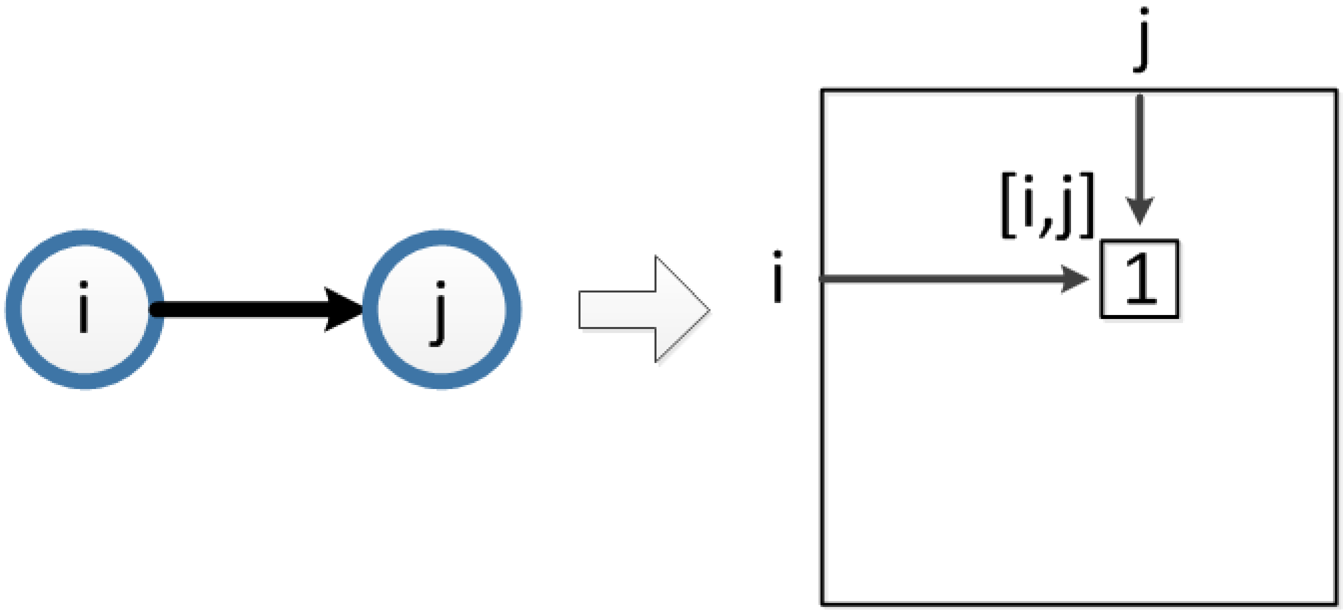
Downloading data from Twitter

Twitter is not consistent

and networks are dynamic

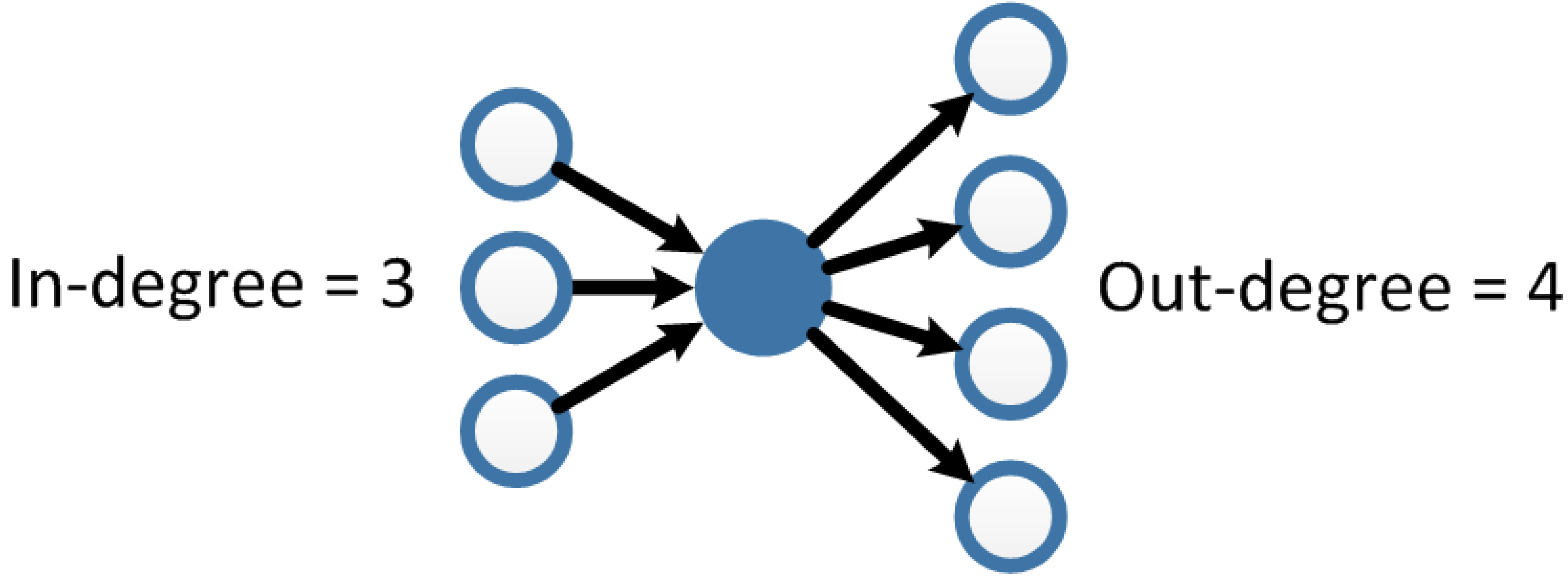


Adjacency matrix

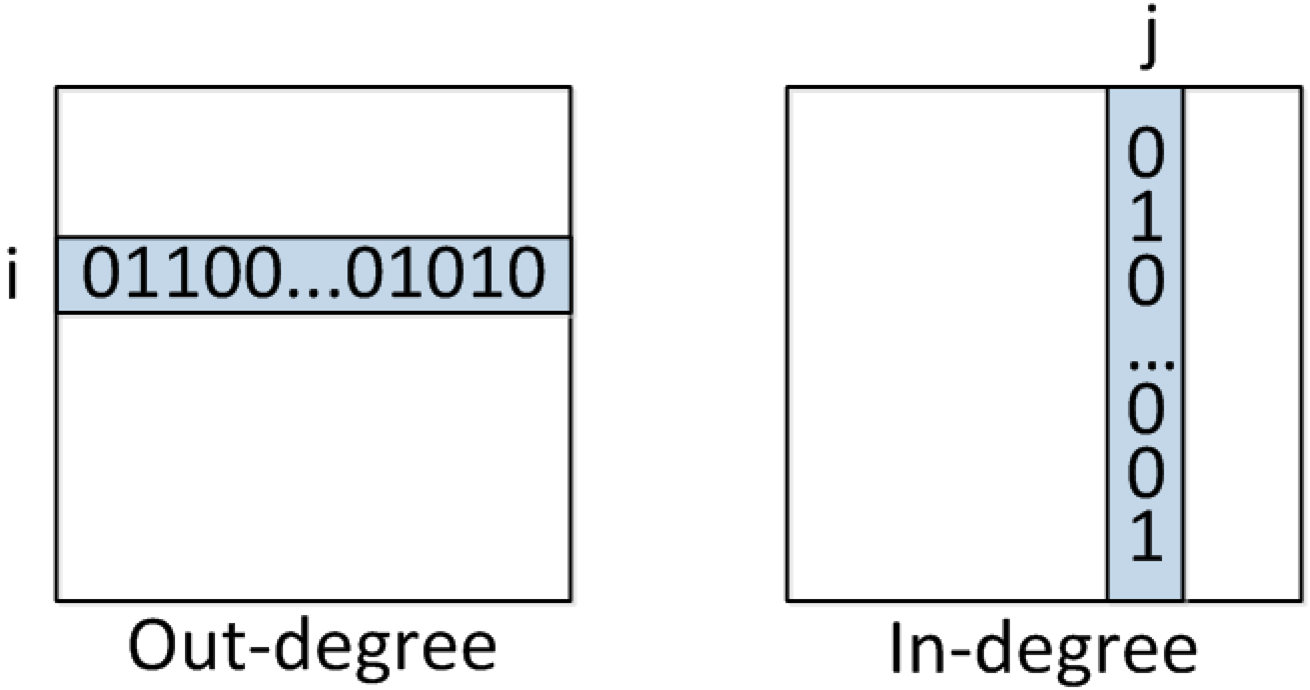


ADJACENCY MATRIX

Degrees

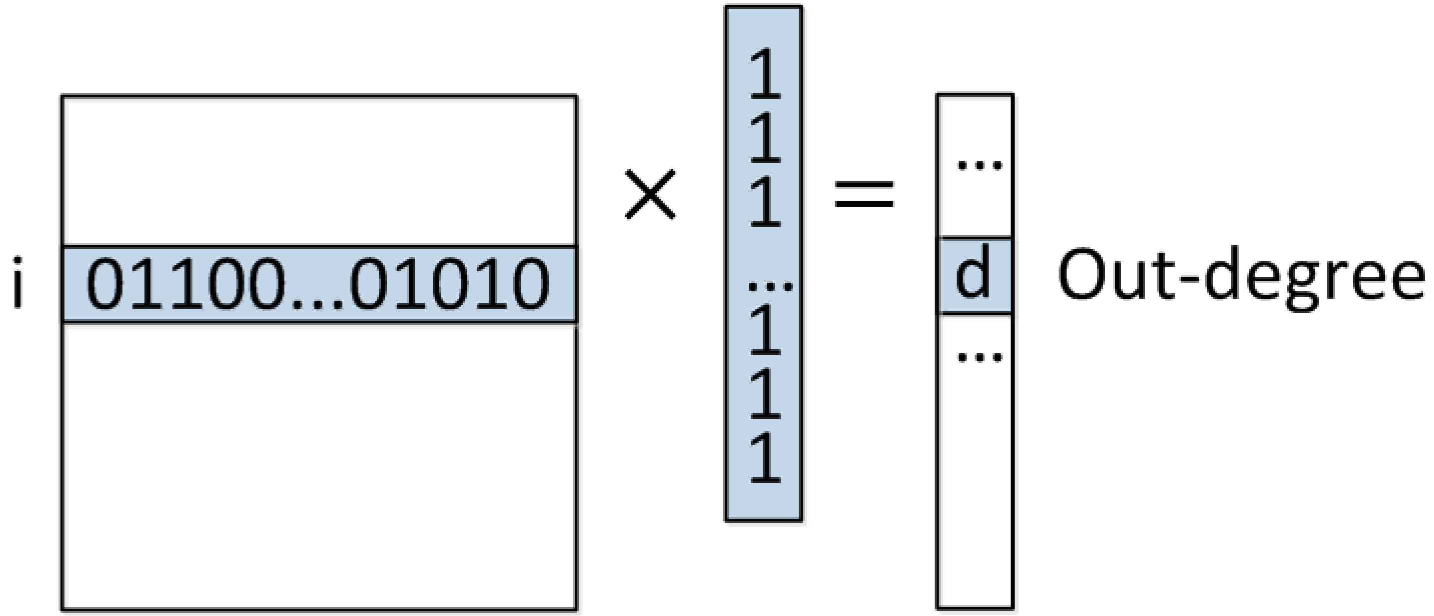


Degrees

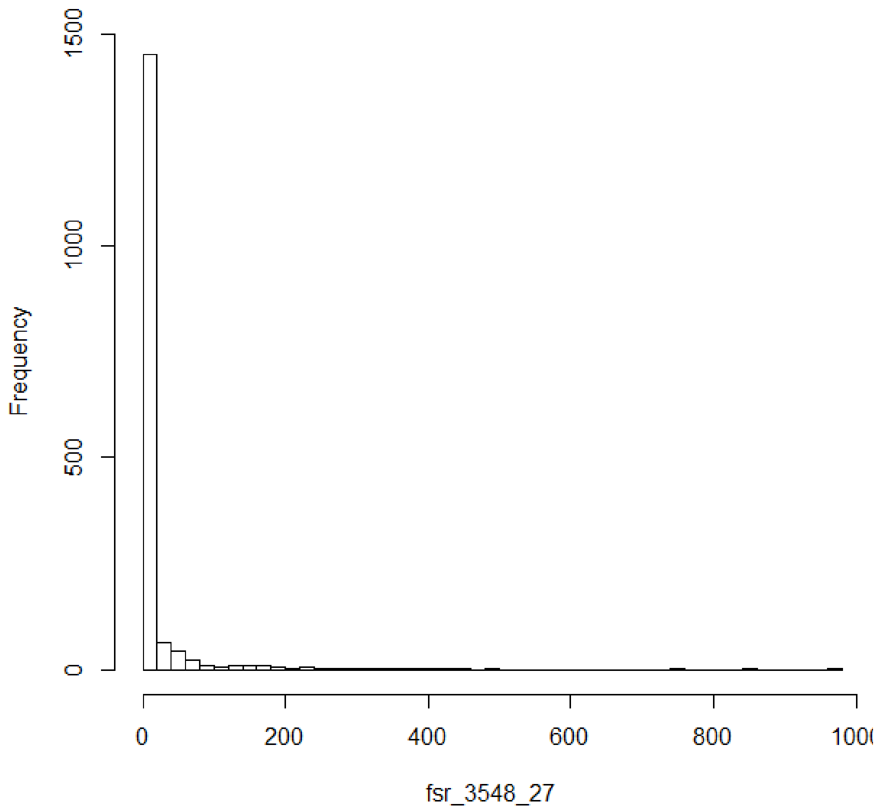


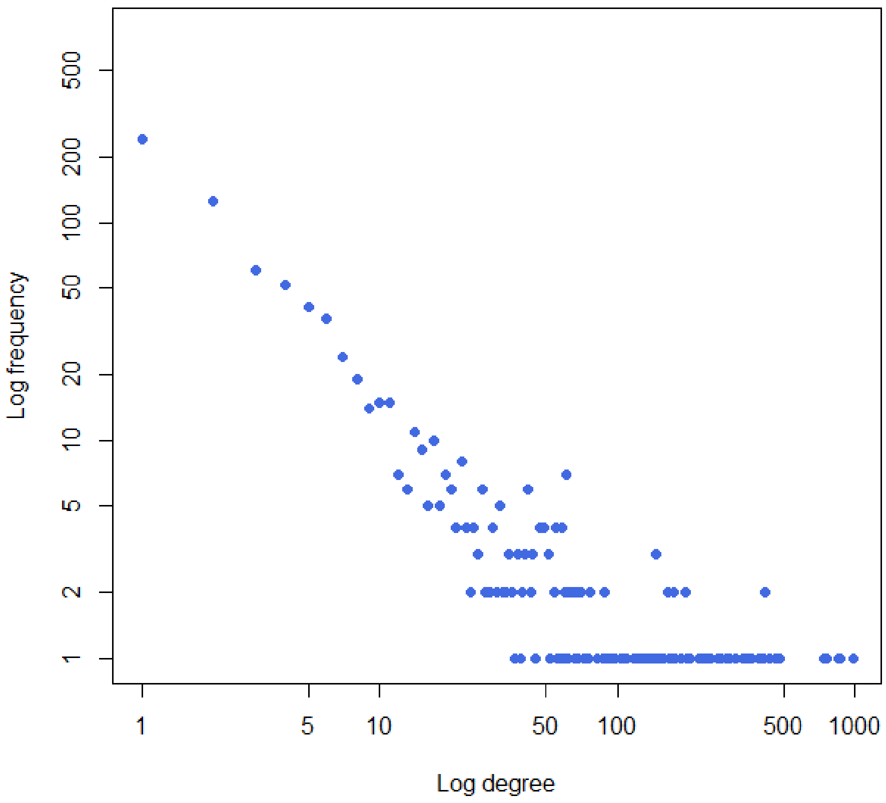
DEGREES

Degrees



Degree distribution



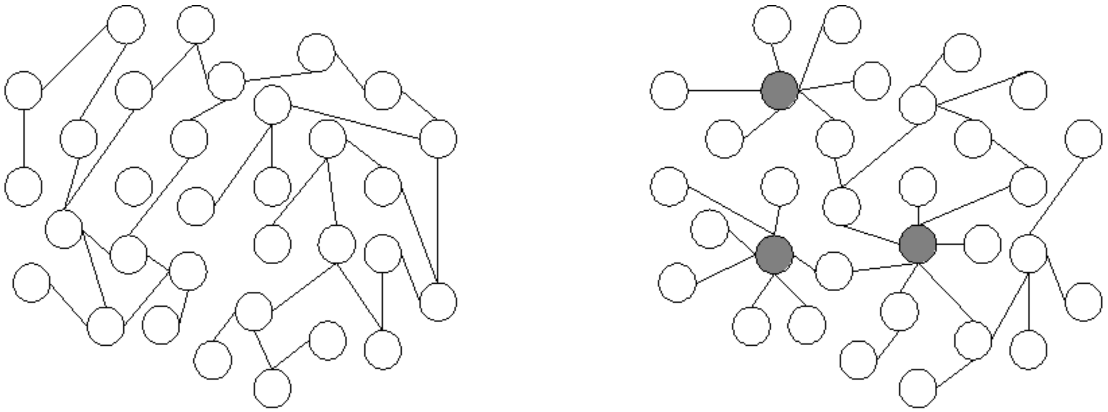


Scale-free networks

- Power law

$$P(d) \sim d^{-\gamma}$$

- Networks growing over time with preferential attachment



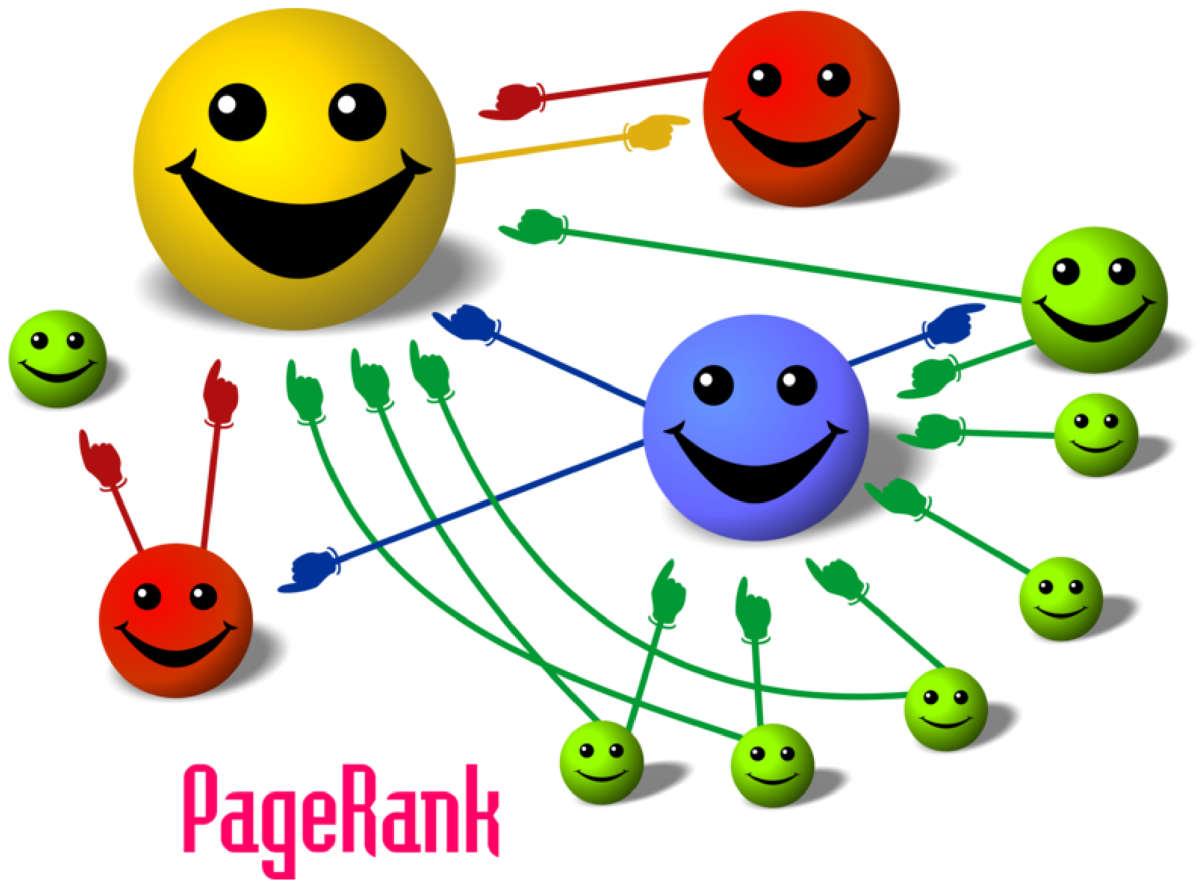
- Hubs
- Robustness

- Your friends have more friends than you do.

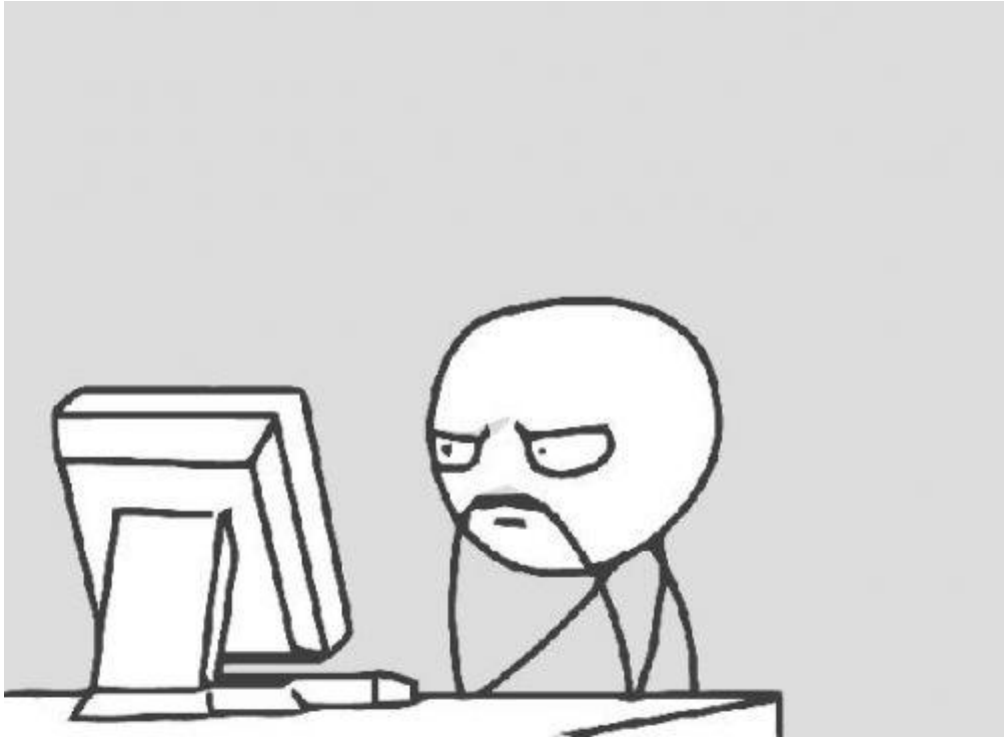
TOP RANKING USERS

Centrality with PageRank

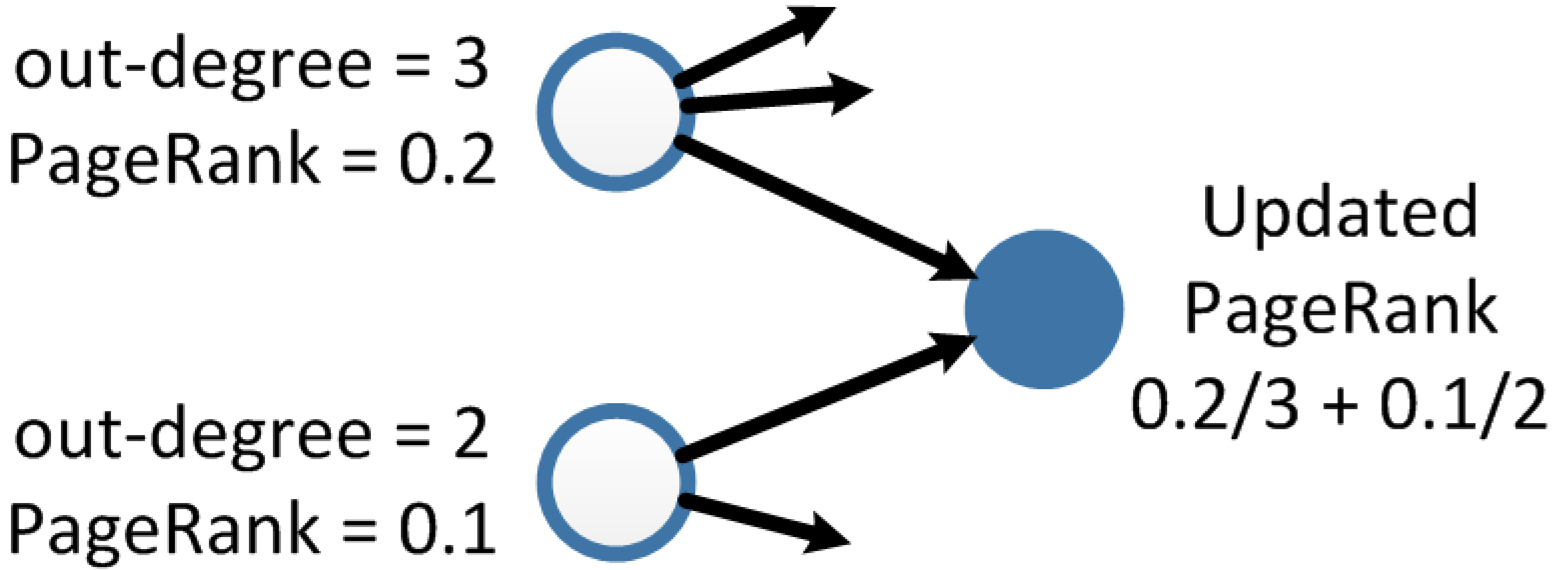
Your followers are not created equal.



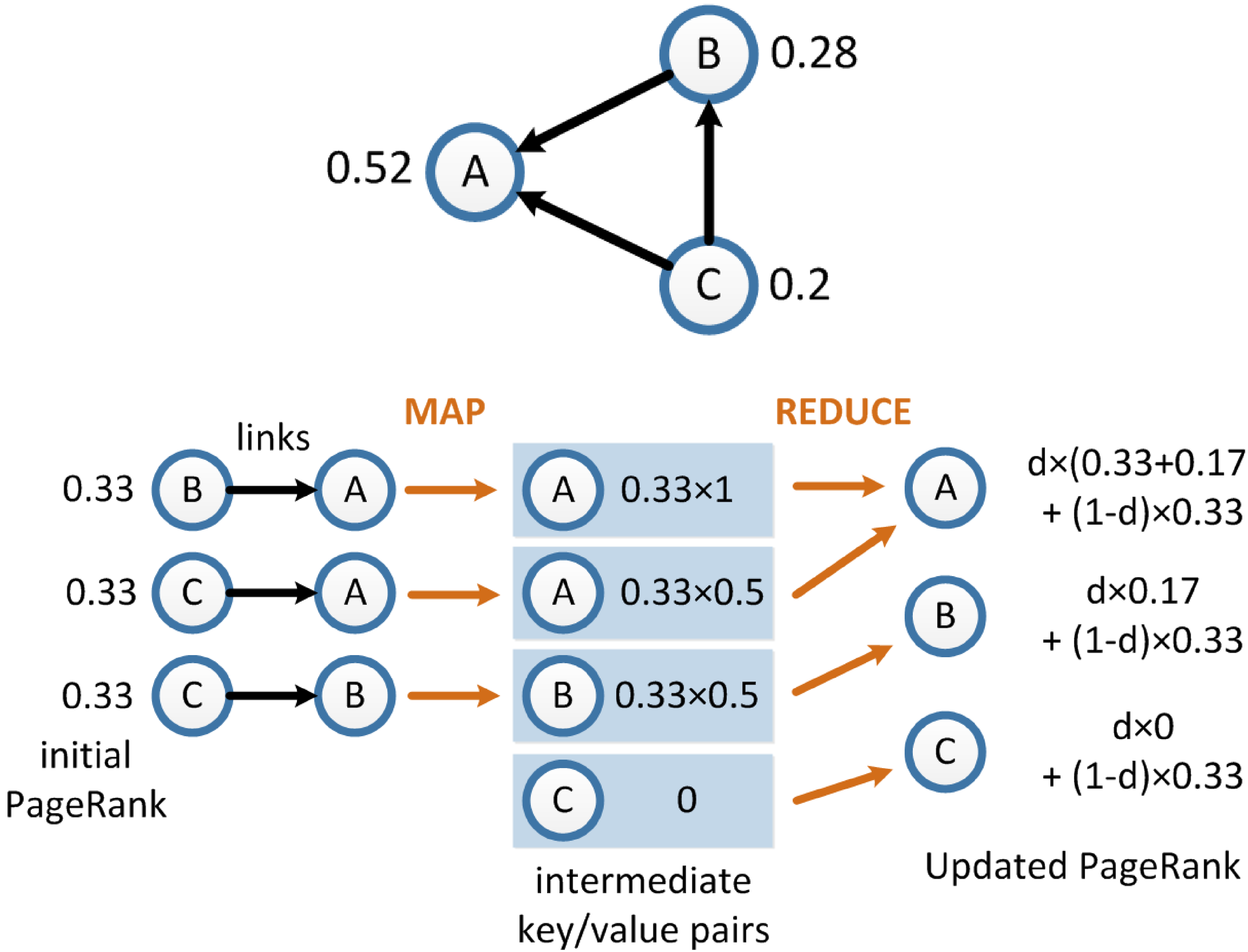
Random surfer model



Centrality with PageRank



+ random jumps



B
0.28
0.52
A
C
0.2
MAP
REDUCE
links
0.33
B
A
A
0.33×1
A
d×(0.33+0.17)
+ (1-d)×0.33
0.33
C
A
A
0.33×0.5
B
d×0.17
+ (1-d)×0.33
0.33
C
B
B
0.33×0.5
C
d×0
+ (1-d)×0.33
initial
PageRank
C
0
intermediate
key/value pairs
Updated PageRank

CENTRALITY WITH PAGERANK

PageRank changes

February

1. migueldeicaza (0.033130)
2. dsyme (0.032783)
3. tomaspetricek (0.027756)
4. LincolnAtkinson (0.021993)
5. VisualFSharp (0.020233)
6. c4fsharp (0.019720)
7. rickasaurus (0.019189)
8. ptrelford (0.018099)
9. 1tgr (0.016525)
10. sforkmann (0.014970)

September

1. dsyme (0.028640)
2. migueldeicaza (0.024808)
3. VisualFSharp (0.024479)
4. tomaspetricek (0.021066)
5. c4fsharp (0.019612)
6. rickasaurus (0.014272)
7. sforkmann (0.013471)
8. 1tgr (0.012768)
9. ptrelford (0.012669)
10. FSPowerTools (0.012113)

VISUALISATION WITH D3.JS

So who’s my most central follower?

1) dsyme
2) tomaspetricek
3) rickasaurus
4) ptrelford
5) sforkmann
6) brandewinder
7) sergey_tihon
8) rachelreese
9) ScottWlaschin
10) 7fsharp9

So who's my most central follower?

1) dsyme
2) tomaspetricek
3) rickasaurus
4) ptrelford
5) sforkmann
6) brandewinder
7) sergey_tihon
8) rachelreese
9) ScottWlaschin
10) 7fsharp9

Thank you!