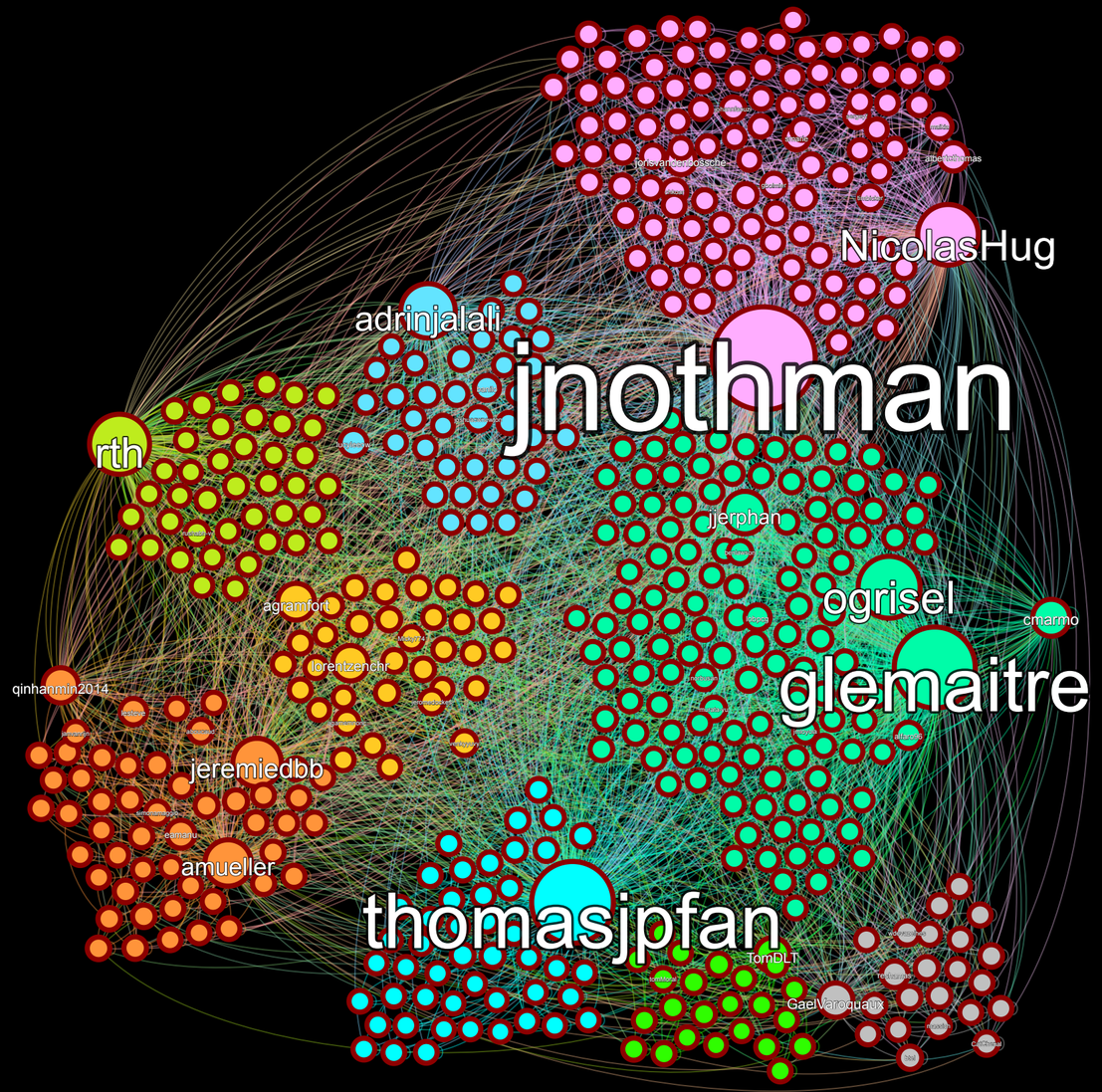
Community detection in large open-source projects

Analyzing scikit-learn contribution graph



German Arutyunov 08.04.2022

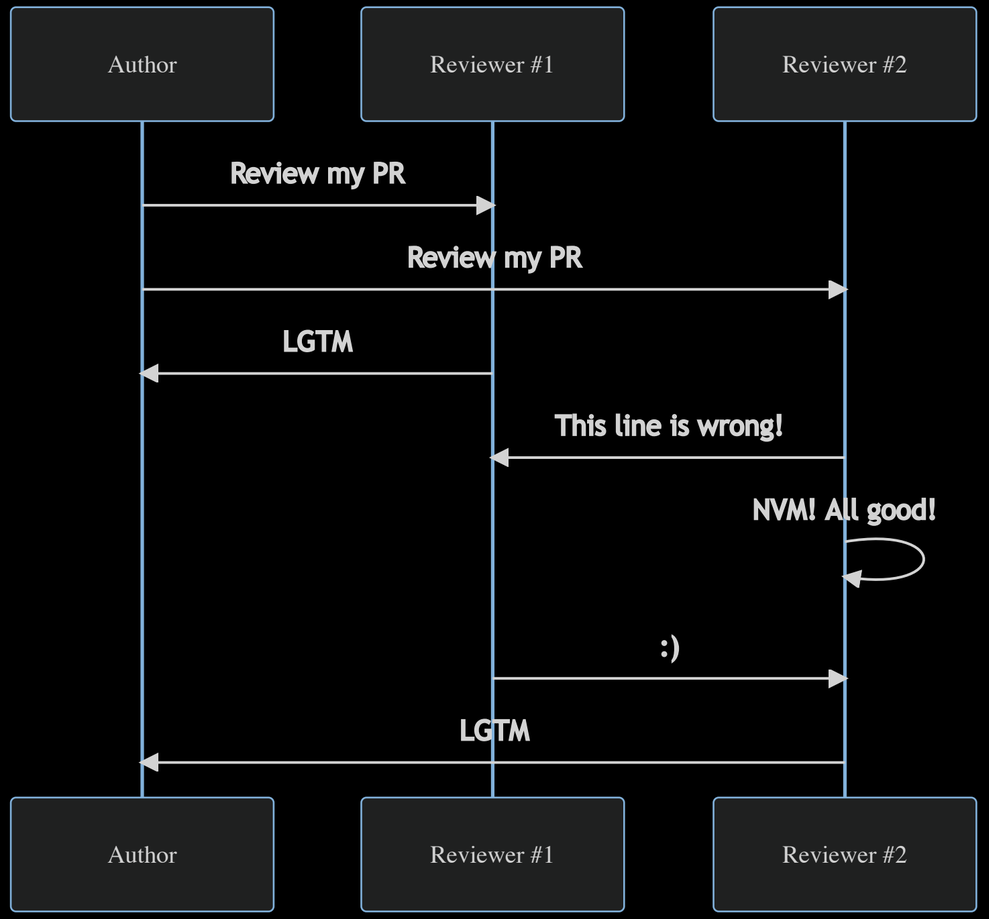
GitHub Data

- Pull requests: 990
- Reviews: 18905
- Comments: 29077
- Contributors: 509

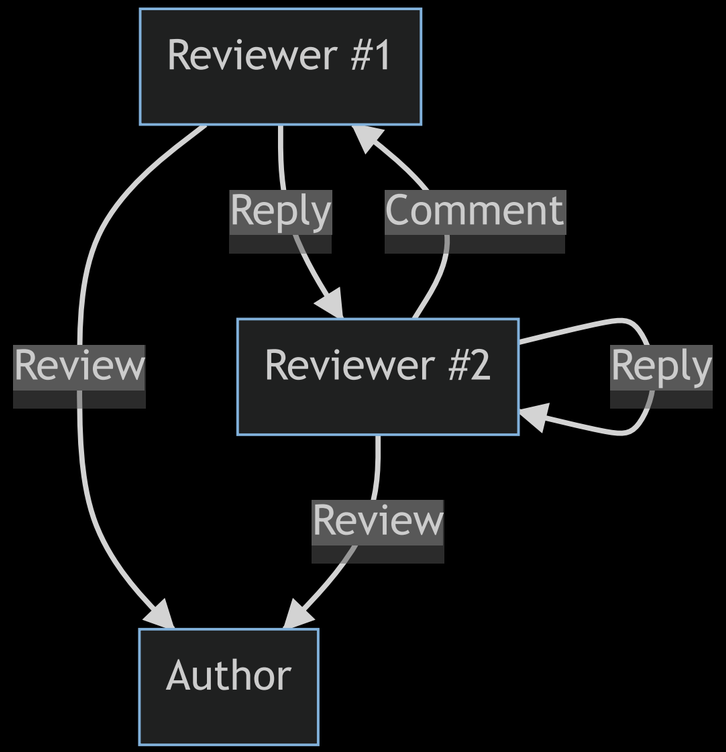
Graph summary

- Directional homogeneous unweighted graph with self loops
- Nodes: 503
- Edges: 3182

Communication



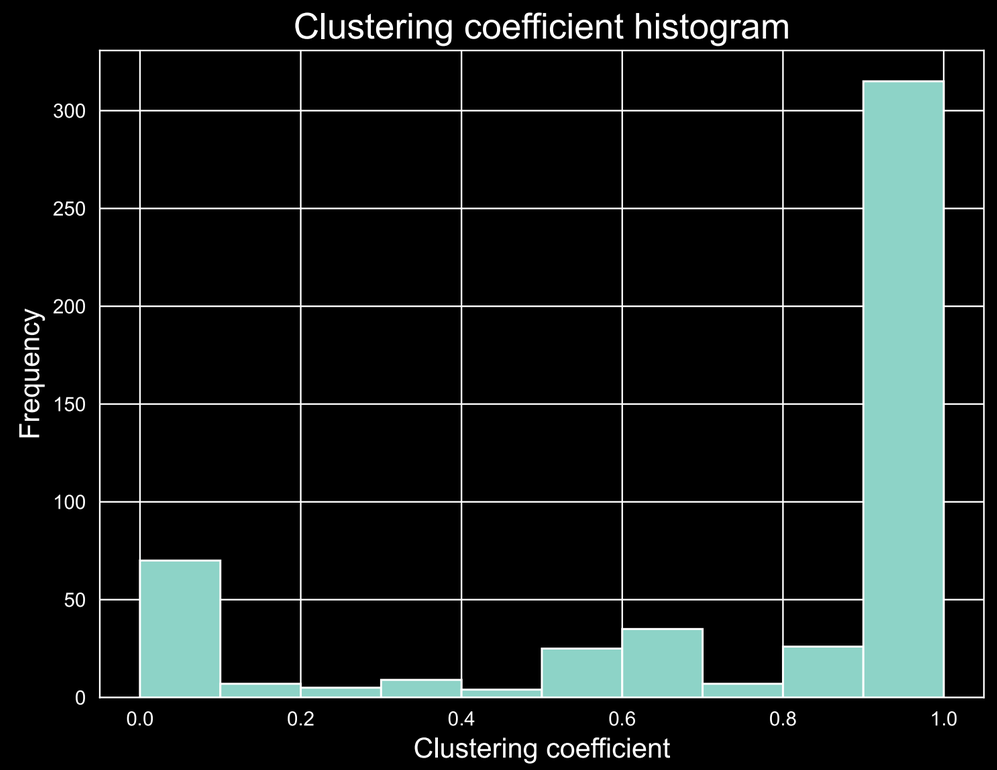
Resulting Graph



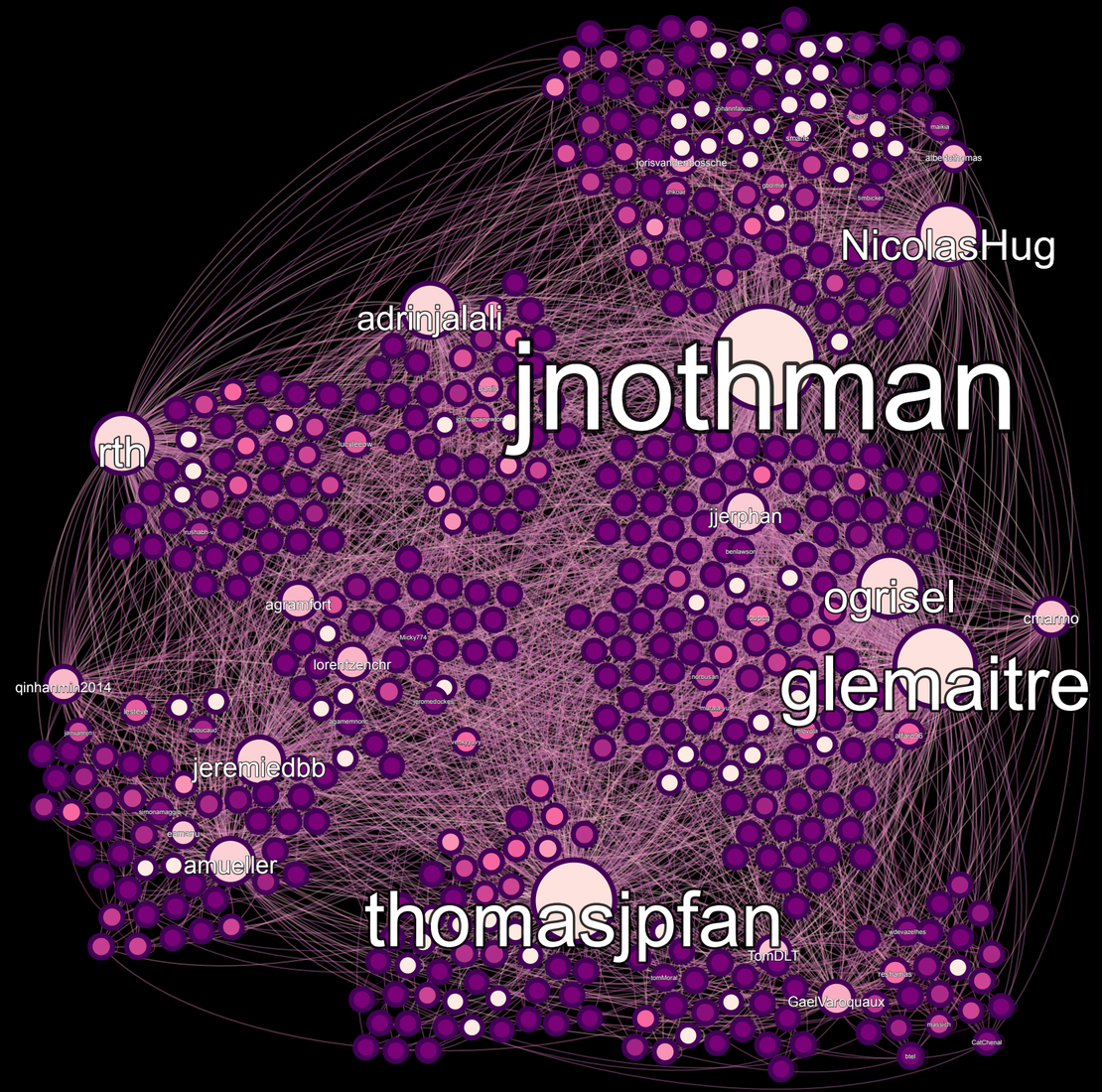
Hypothesis

- H1: High clustering coefficient
- H2: Hubs with high degrees and centralities
- H3: Short average path length
- H4: Power-law like distribution
- H5: Few tight communities

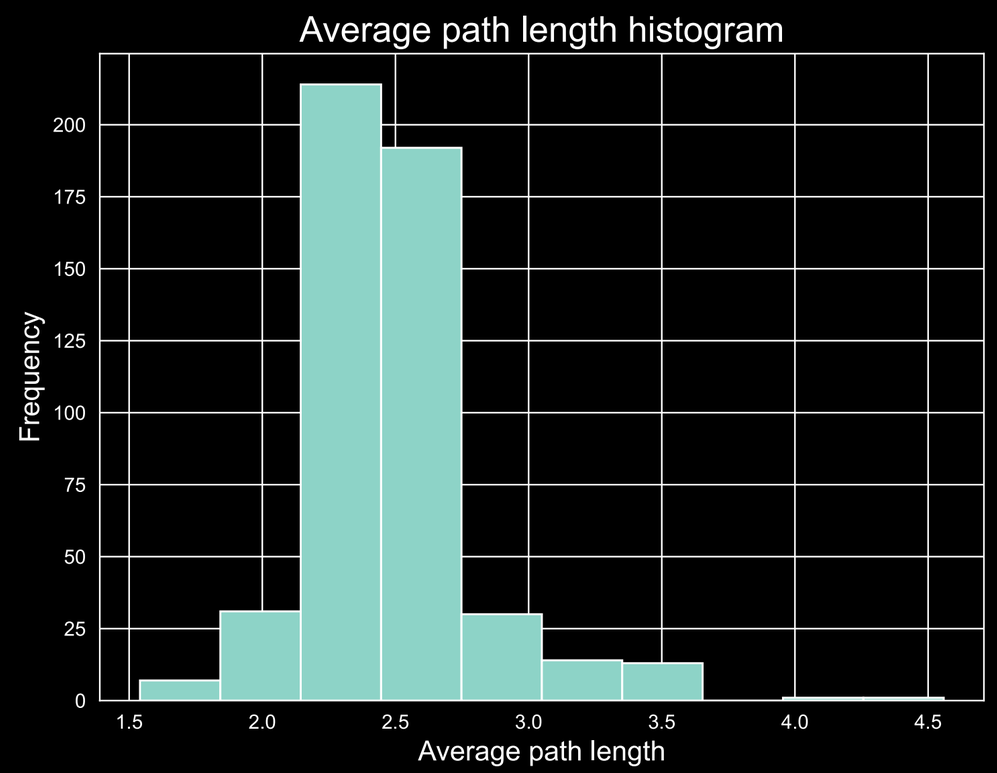
Clustering coefficient

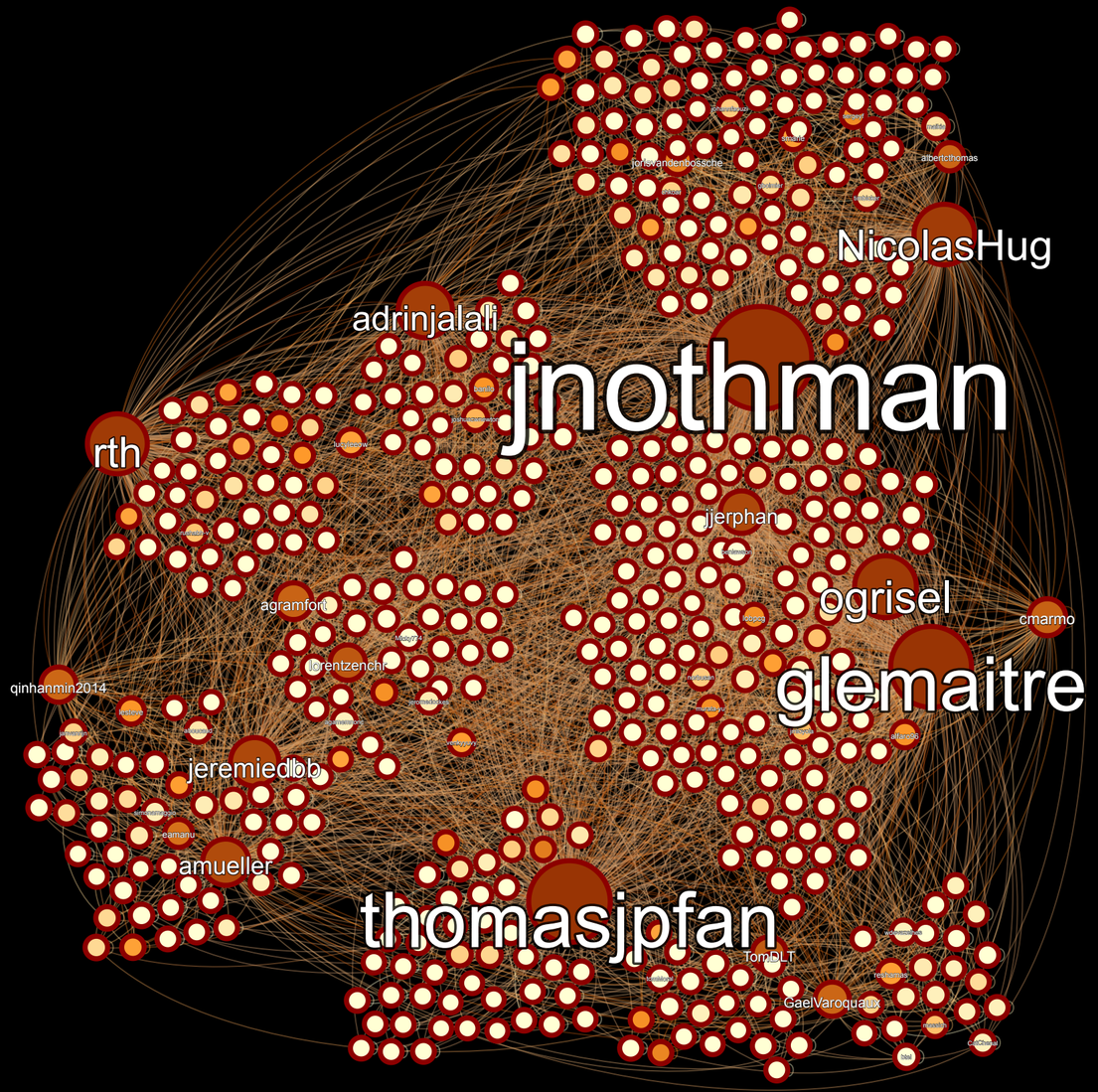


Average clustering coefficient: 0.77



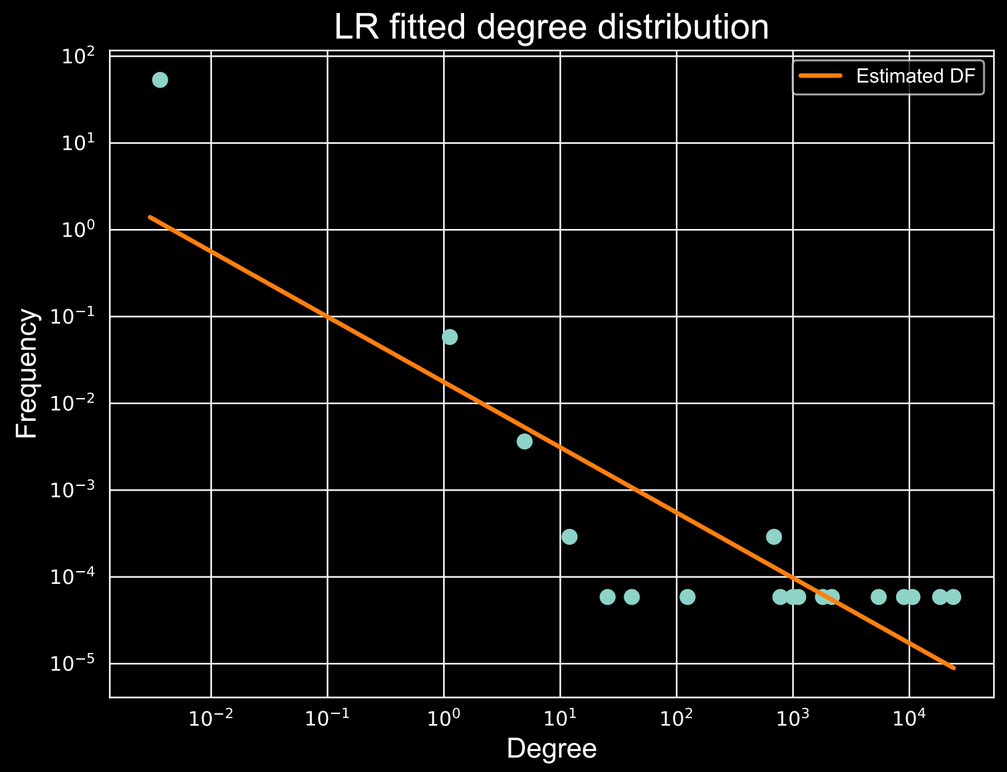
Betweenness centrality and shortest paths

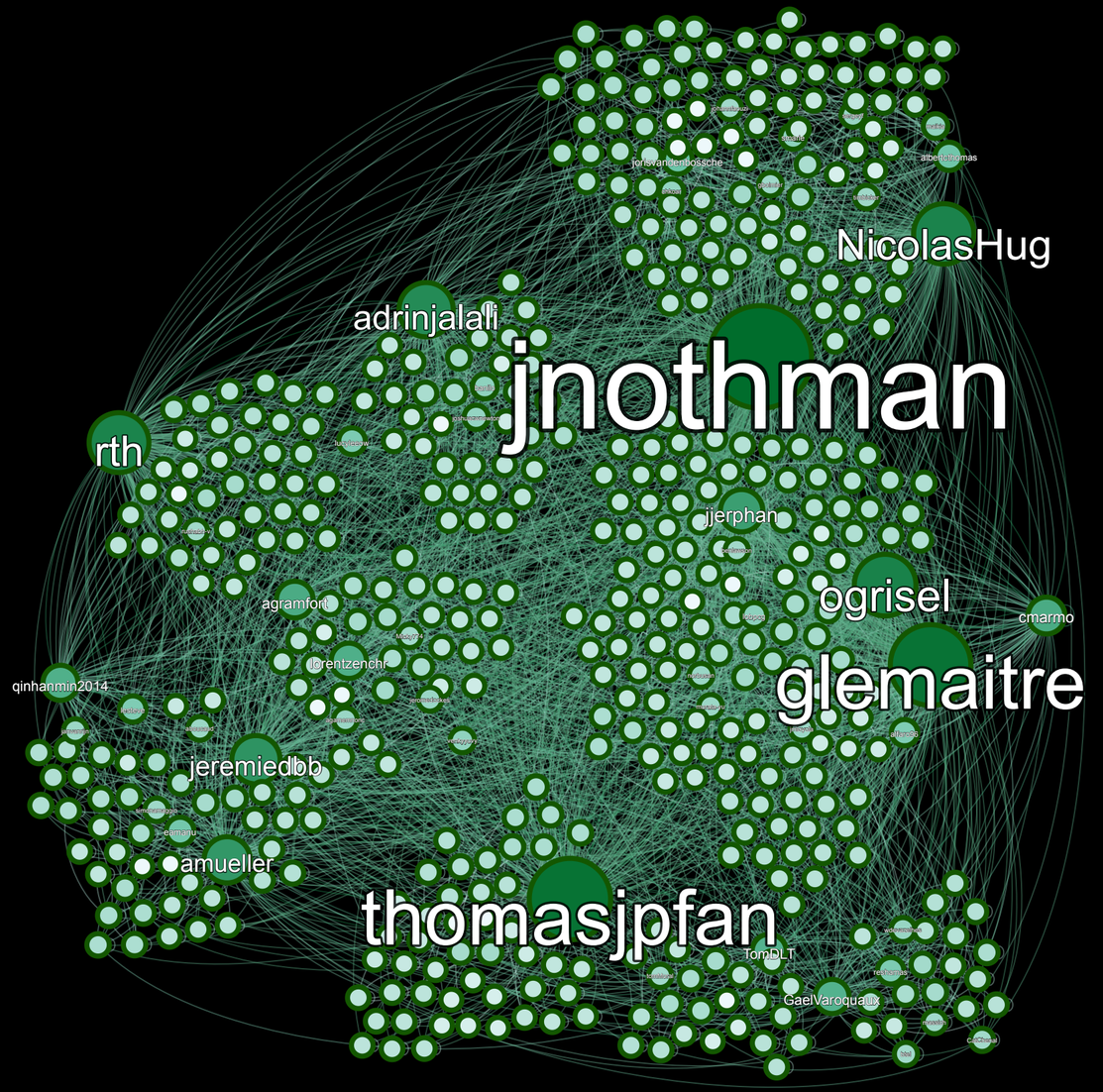




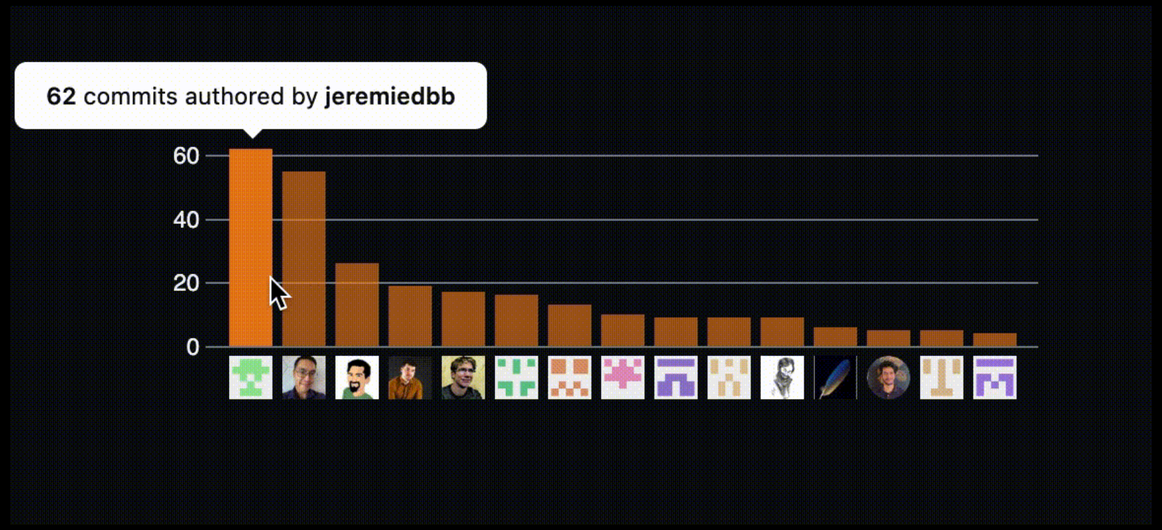
Average path length: 2.48

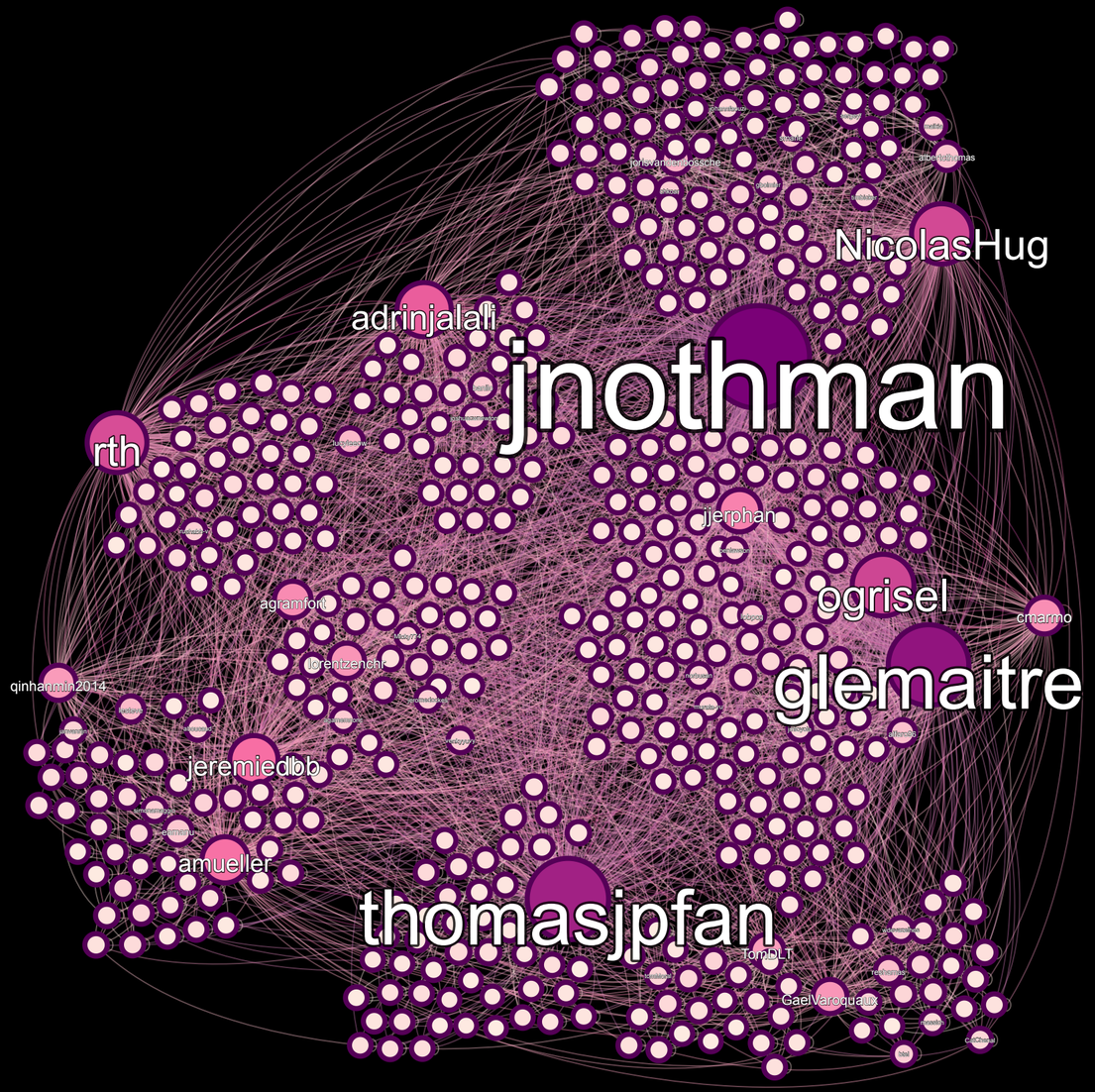
Power law?


LR fitted degree distribution
Estimated DF
Frequency
Degree
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}


NicolasHug
adrinjalali
jnothman
rth
jjerphan
ogrisel
cmarmo
agramfort
glemaitre
lorentzenchr
qinhanmin2014
jeremiedbb
amueller
thomasjpfan
TomDLT
GaelVaroquaux

Who are the hubs?

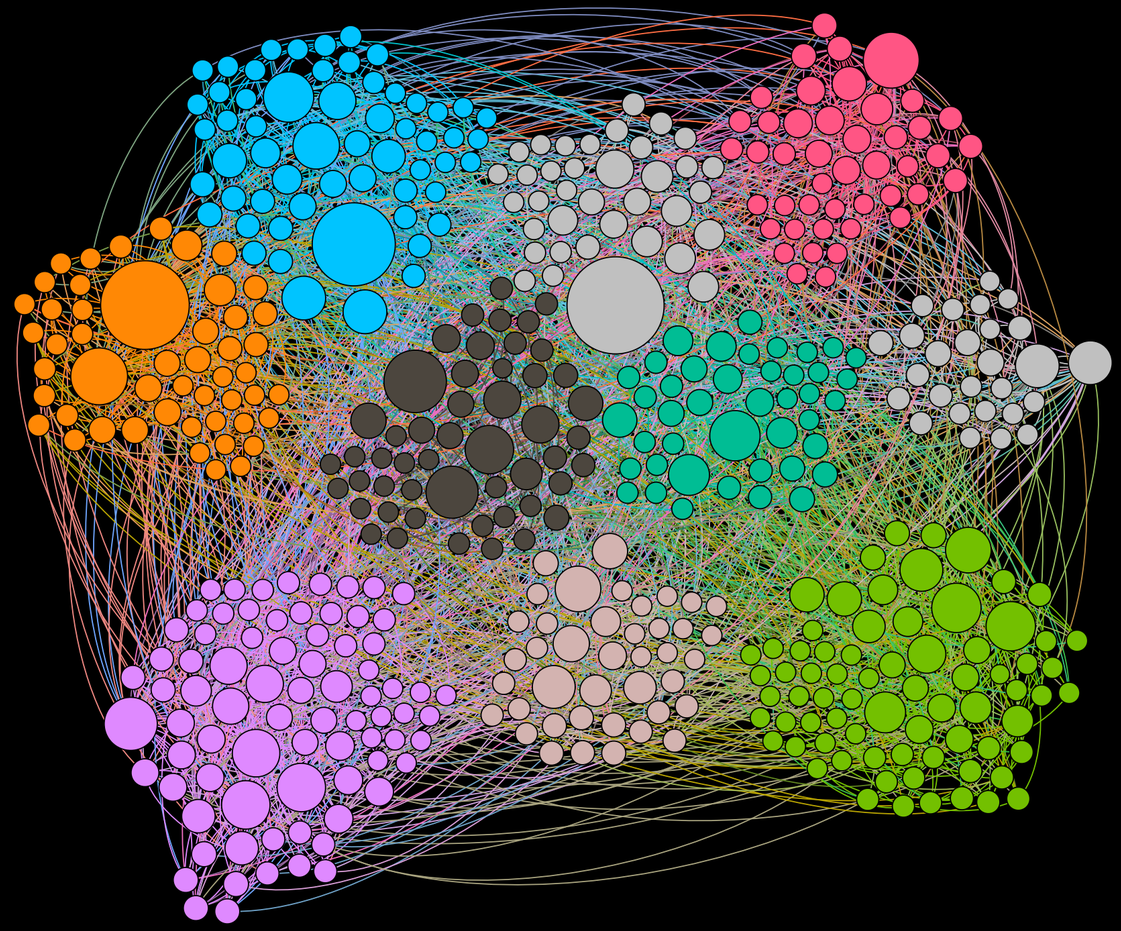




Barabasi & Albert Model

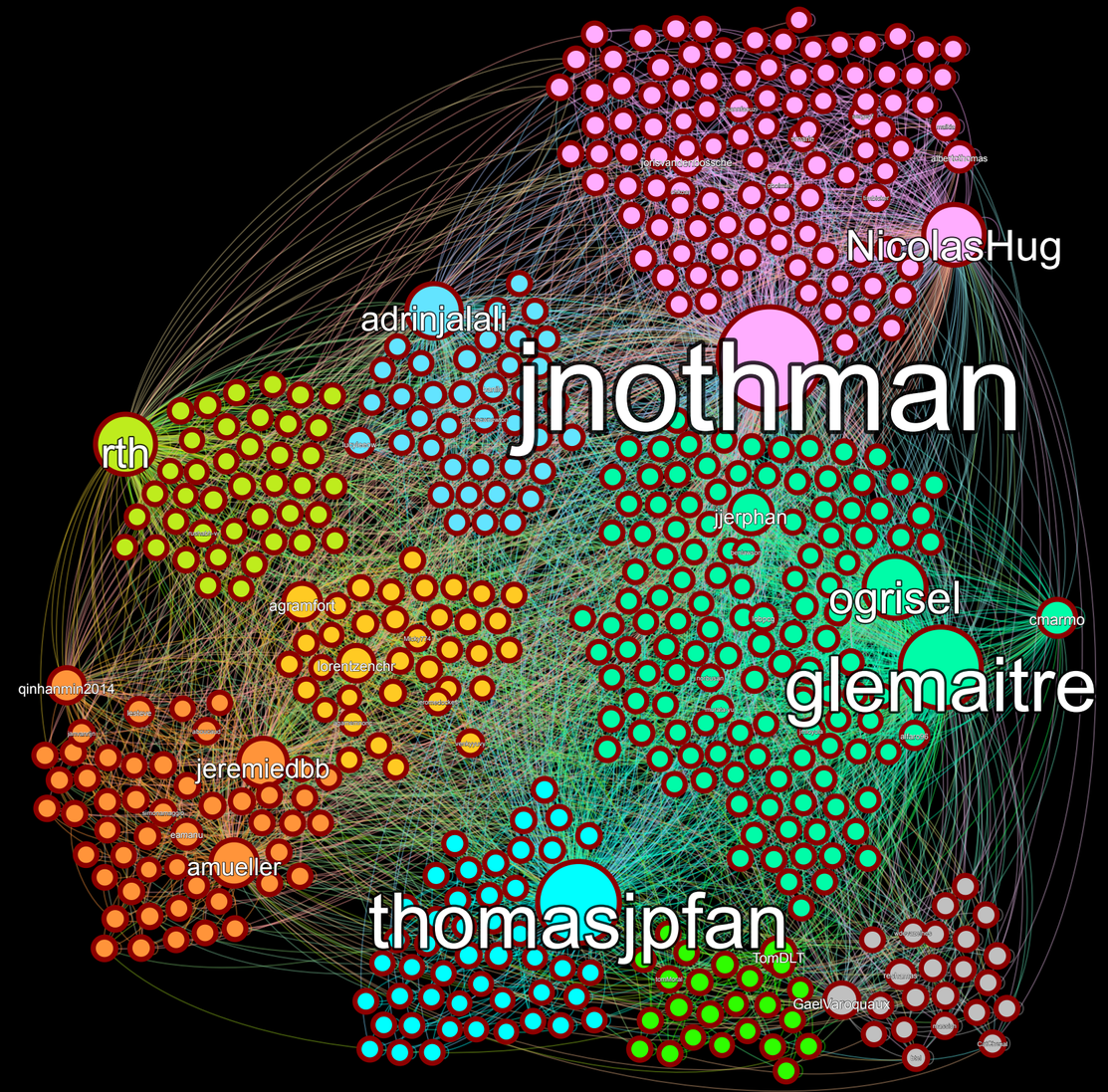
- Nodes: 503
- Edges: 3163

Metrics	Real	BA
Avg. clustering	0.77	0.06
Modularity	0.29	0.26
Avg. path length	2.56	2.78
Diameter	6	4
Radius	3	3



Community detection

- Method: Louvain
- Number of communities: 8
- Modularity: 0.291



Thank you for your attention!