Fundamentals of (Online ((Social) Media)) Network Analysis

Lecture 1

Online Social Networks, Elements of Networks, Network Measures, Data Sources

---

Who is this guy?

Why are you here?

Who are you?

Why did you choose this course?

What are your expectations for this day?

The Plan

1. Online (Social) Media Network Fundamentals
2. Network fundamentals
3. *break*
4. Network Analysis Methods
5. Data Mining Possibilities and Difficulties

Afterwards:

Practical on Data Collection and Exploratory Analysis with Descriptive Statistics in Python

---

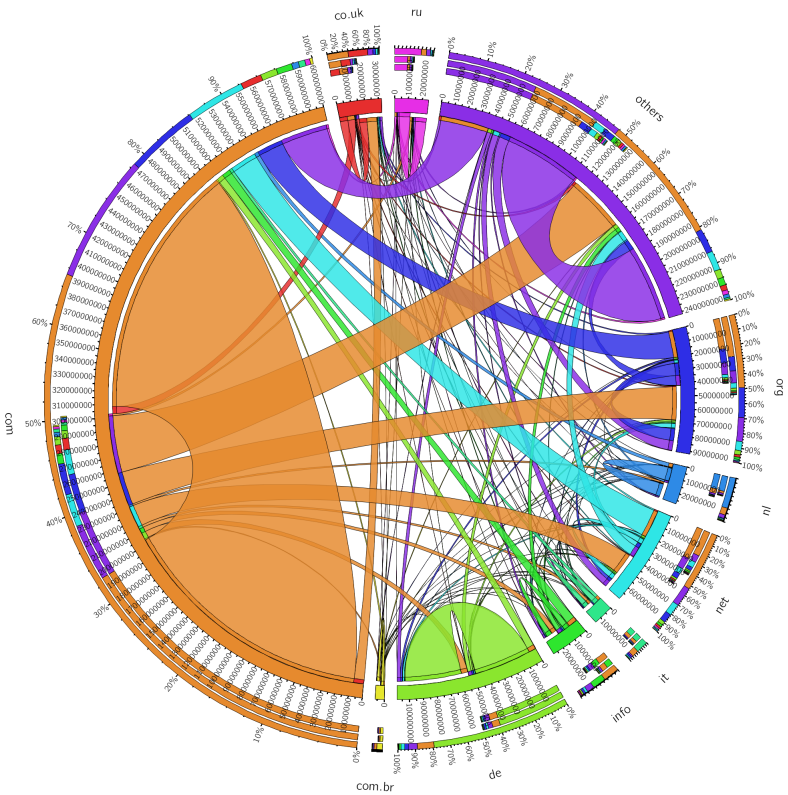
Online ((Social) Media) Networks

--

Not Technical Infrastructure Networks

--

Mostly not Hyperlink Networks



Links between top level domains in 2012 ("Topology of the WDC Hyperlink Graph", http://km.aifb.kit.edu/sites/webdatacommons/hyperlinkgraph/topology.html)

--

Influence (Some/Most?) Information Diffusion

#Sydneysiege vs #illridewithyou vs \Brexit petition

--

#illridewithyou

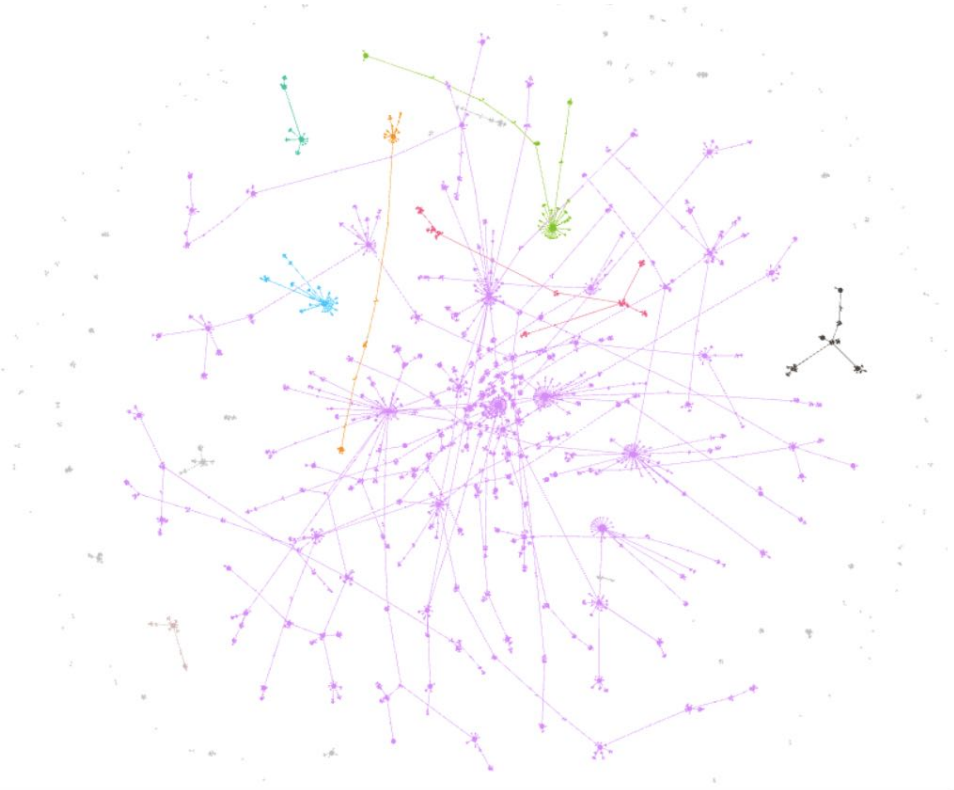


Figure 5.14: Force-directed visualisation of the diffusion tree network of the hashtag illridewithyou for the first 10 000 accounts using it, coloured by weakly-connected components

--

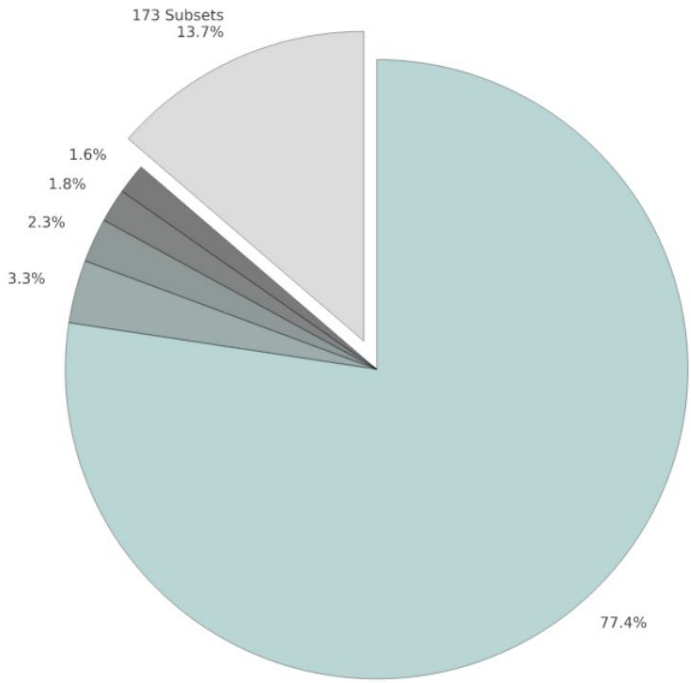


Figure 5.15: Percentages covered by the largest weakly-connected components of the diffusion tree network for the first 10 000 accounts tweeting the hashtag illridewithyou

--

#sydneysiege

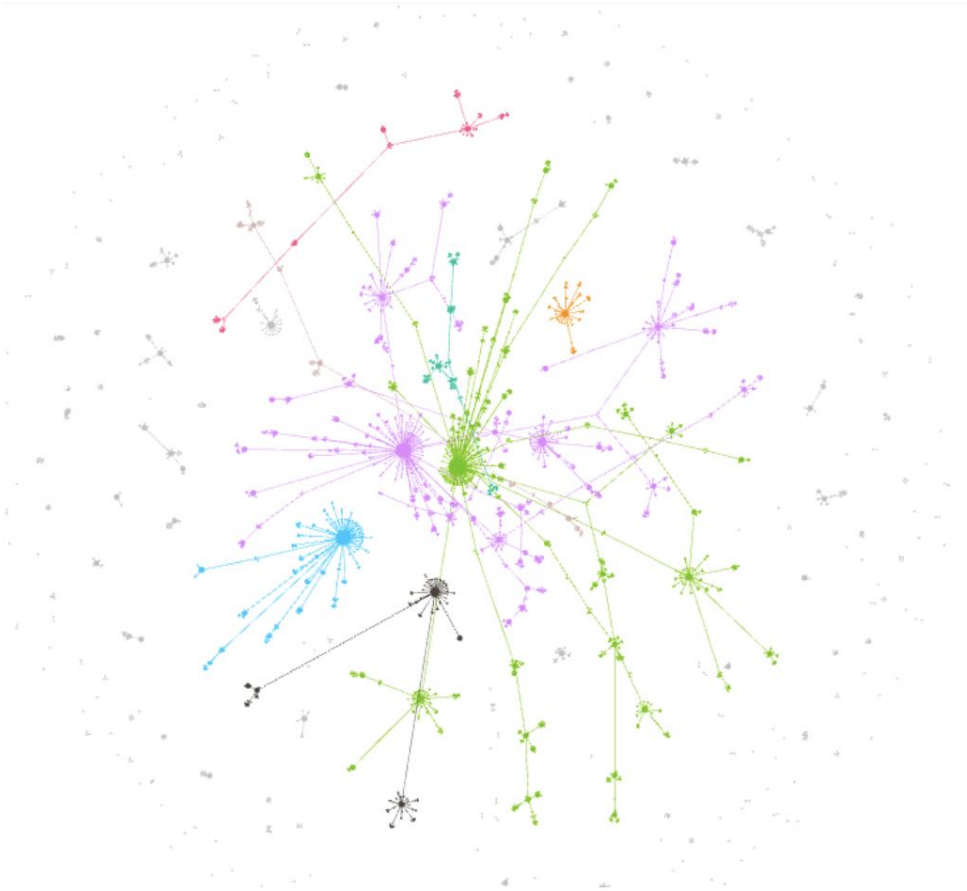


Figure 5.16: Force-directed visualisation of the diffusion tree network of the hashtag sydneysiege for the first 10 000 accounts using it, coloured by weakly-connected components

--

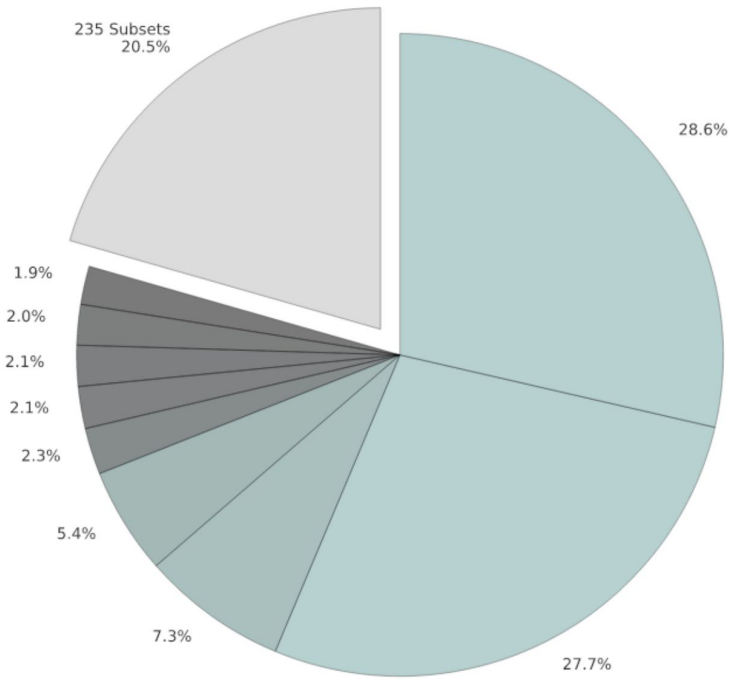


Figure 5.17: Percentages covered by the largest weakly-connected components of the diffusion tree network for the first 10 000 accounts tweeting the hashtag sydneysiege

--

Anti-Brexit petition

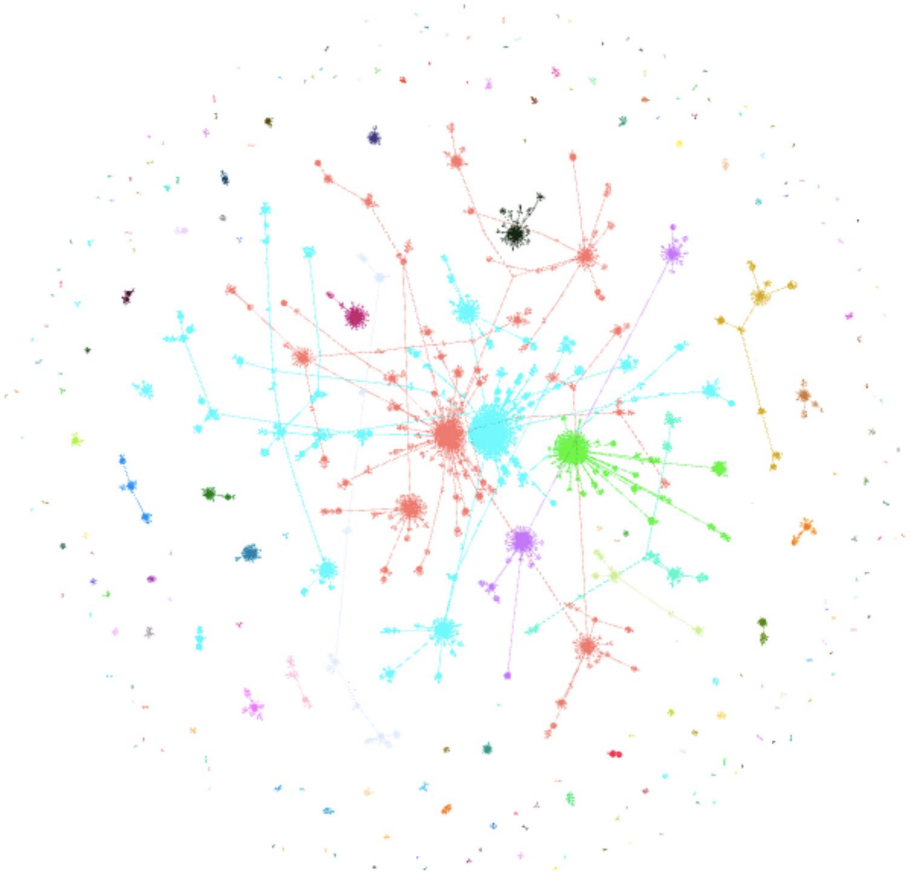


Figure 5.18: Force-directed visualisation of the diffusion tree network of the petition link for the first 10 000 accounts tweeting it, coloured by weakly-connected components

--

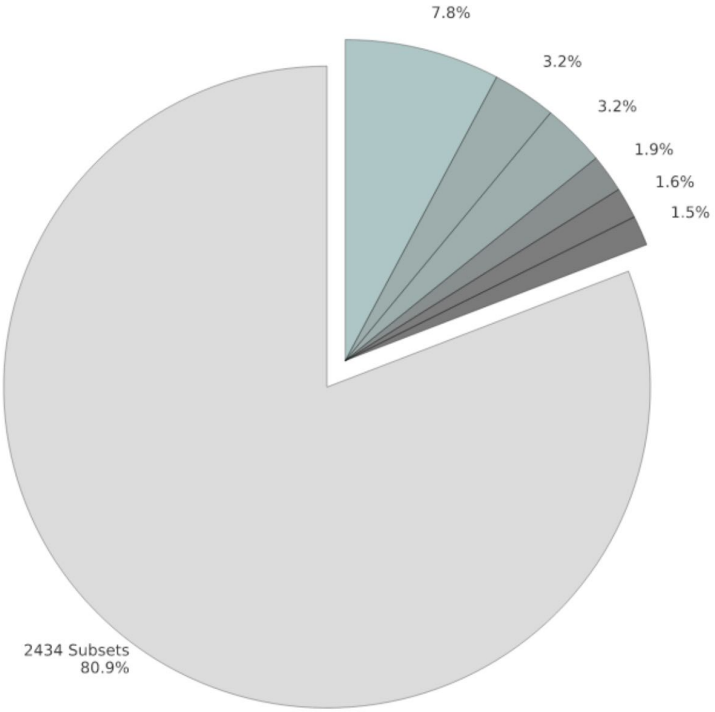


Figure 5.19: Percentages covered by the largest weakly-connected components of the diffusion tree network for the first 10 000 accounts tweeting the link to the petition

--

The number and size of connected components indicates the influence of the network compared to outside sources

--

Lead to Classifiable Communication Patterns

Himelboim, I., Smith, M. A., Rainie, L., Shneiderman, B., & Espina, C. (2017). *Classifying Twitter topic-networks using social network analysis*. 1–38. https://doi.org/10.1177/2056305117691545

--

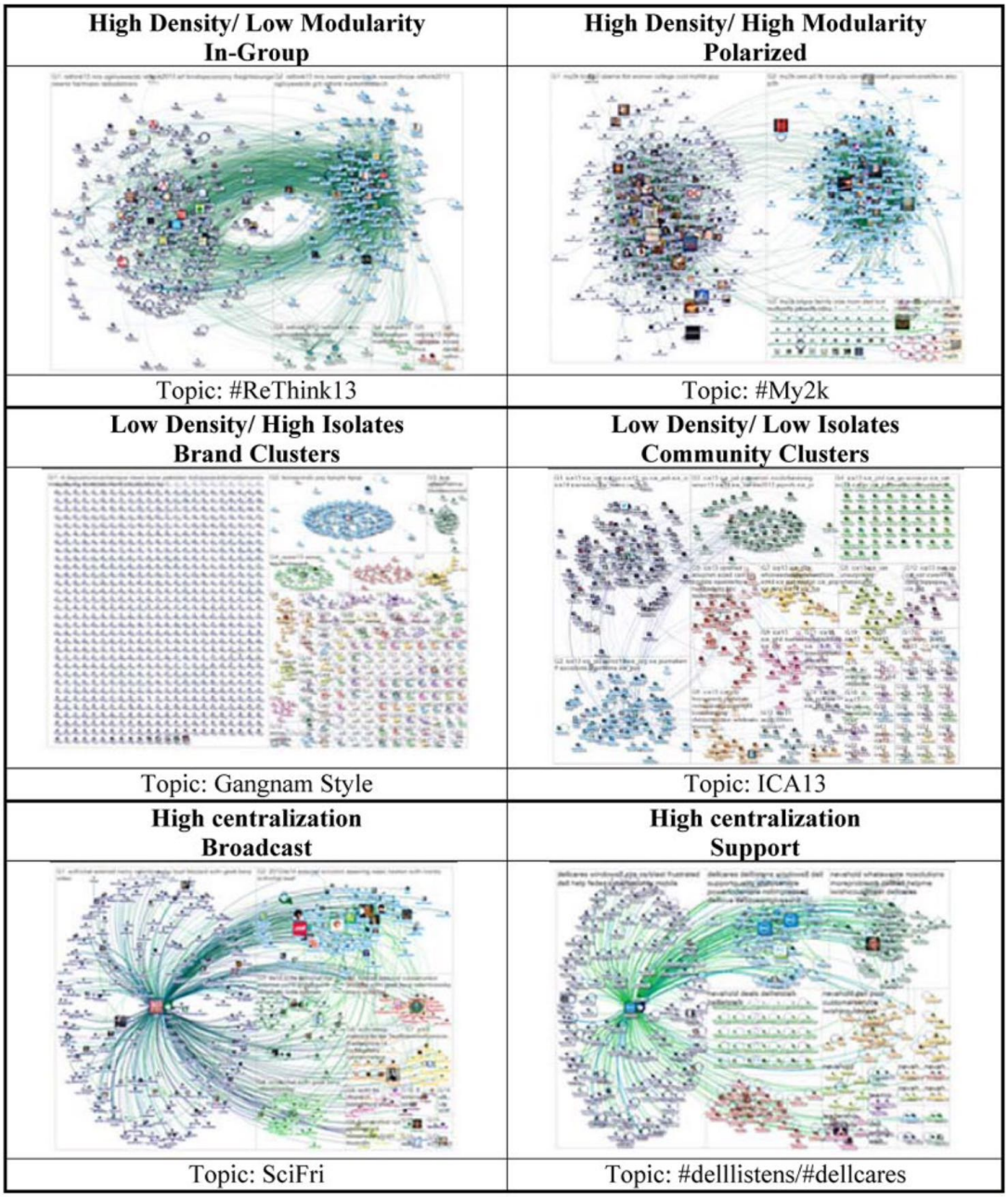
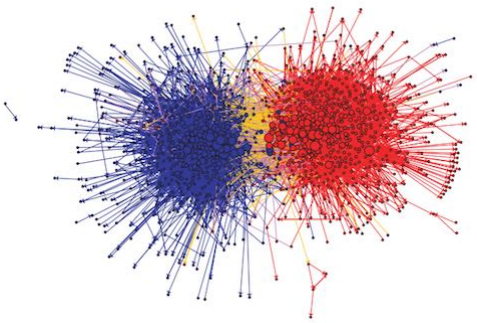


Figure 3. Network visualization by topic-network category.

--

Example: Polarisation



Lada A. Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 U.S. election: divided they blog. In Proceedings of the 3rd international workshop on Link discovery (LinkKDD '05). Association for Computing Machinery, New York, NY, USA, 36–43. https://doi.org/10.1145/1134271.1134277

--

Reflect Long-Term Structured Systems of (Parts of) Society

Bruns, A., Moon, B., Münch, F. V., & Sadkowsky, T. (2017). The Australian Twittersphere in 2016: Mapping the follower/followee network. *Social Media + Society*, *3*(4). https://doi.org/10.1177/2056305117748162

Münch, F. V. (2019). *Measuring the Networked Public – Exploring Network Science Methods for Large Scale Online Media Studies* [PhD thesis, Queensland University of Technology]. https://doi.org/10.5204/thesis.eprints.125543

Münch, F. V., & Rossi, L. (2020, October 5). A Tale of Two Twitters? Identifiying Bridges Between Language Based Twitterspherees. *AoIR Selected Papers of Internet Research*. https://doi.org/10.5210/spir.v2020i0.11283

Münch, F. V., & Rossi, L. (2020). *Bootstrapping Follow Networks of Influential Twitter Accounts*. IC2S2. https://vimeo.com/431470176

Münch, F. V., Thies, B., Puschmann, C., & Bruns, A. (2021). Walking Through Twitter: Sampling a Language-Based Follow Network of Influential Twitter Accounts. *Social Media + Society*. https://doi.org/10.1177/2056305120984475

--

Australian Twittersphere

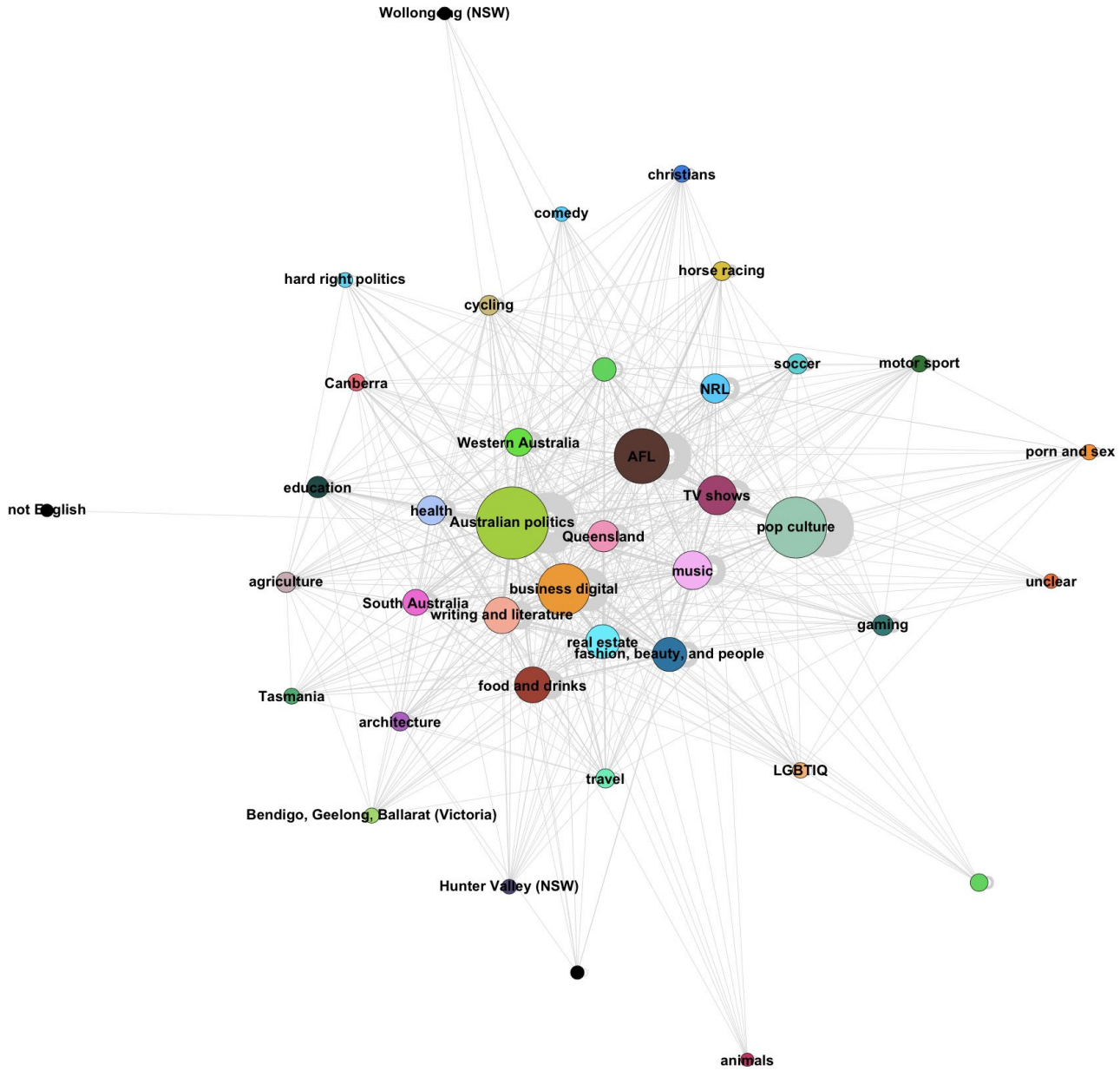


Figure 6.13: Force-directed visualisation of the undirected community graph, filtered for an edge-weight of 10 000 connections; self-loops included; thickness of edges represents weight; size of nodes represents weighted degree; labels are based on keyword analysis of tweets; unlabelled communities are either using a language other than English or no keyword could be determined.

--

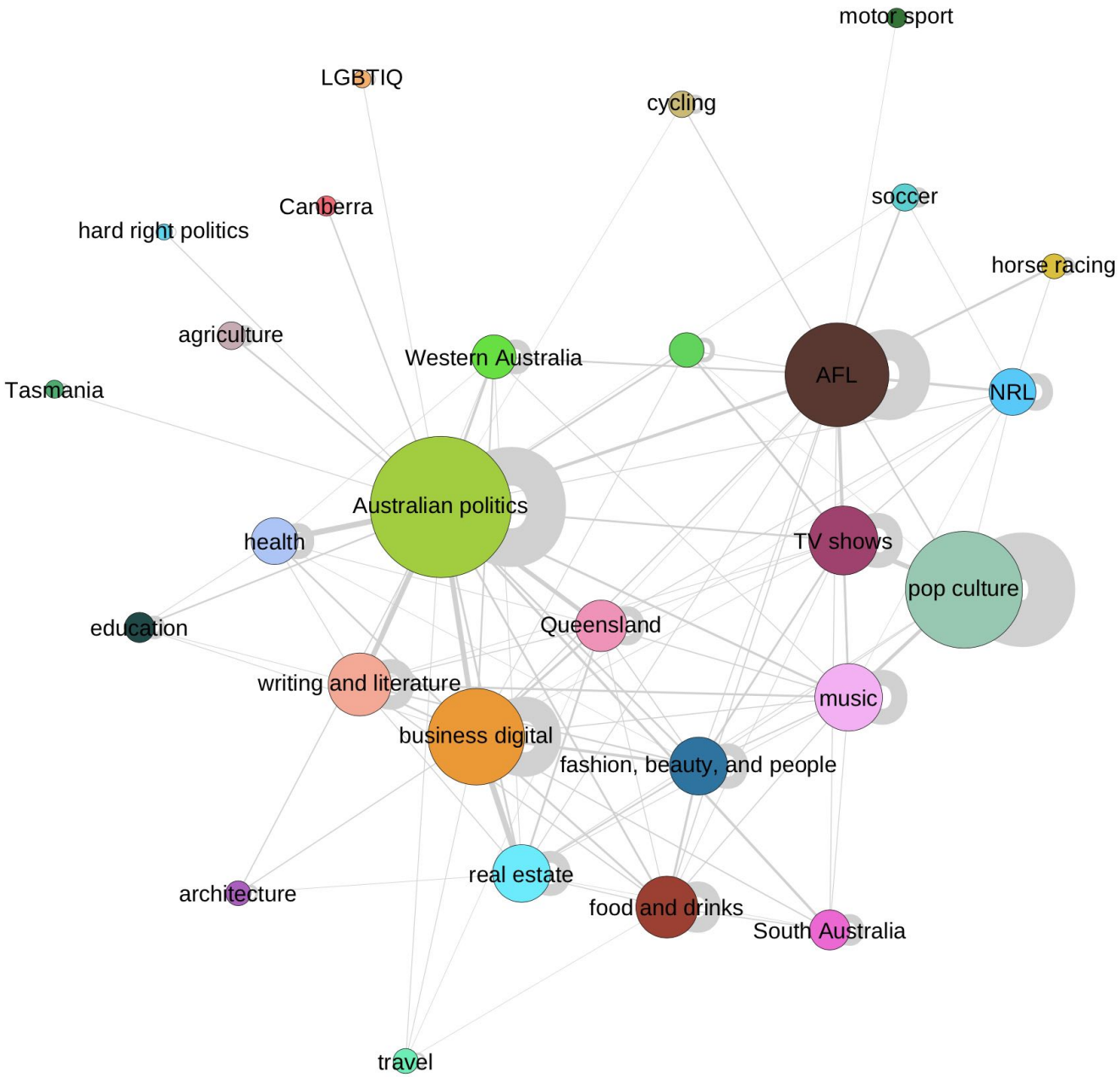


Figure 6.14: Force-directed visualisation of the undirected community graph, filtered for an edge-weight of 100 000 connections; self-loops included; thickness of edges represents weight; size of nodes represents weighted degree; labels are based on keyword analysis of tweets; unlabelled communities are either using a language other than English, or no keyword could be determined.

--

German Twittersphere

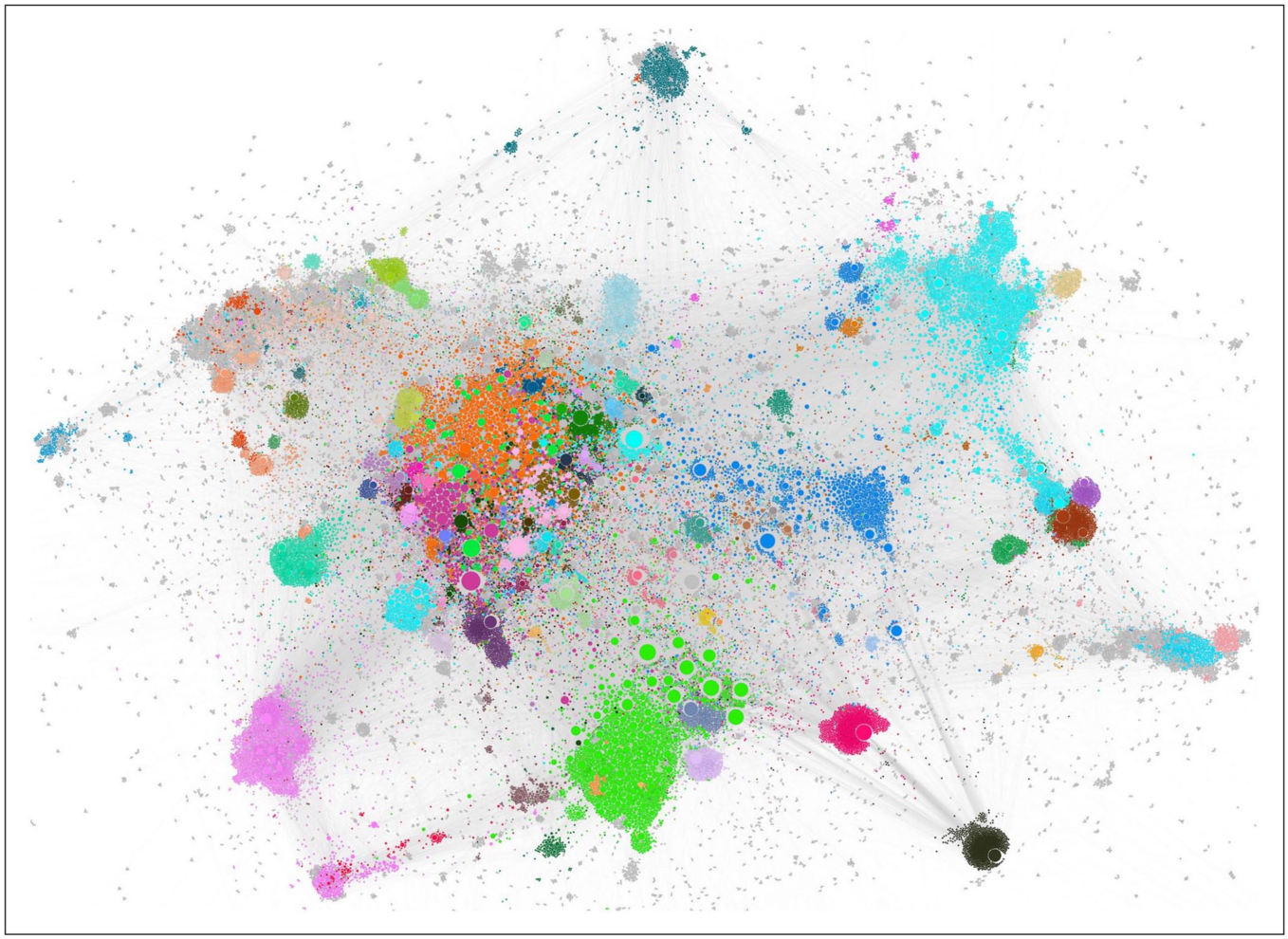


Figure 12. Central communities in the 3-core of our sample network; colored by largest communities detected with the Infomap community detection algorithm (Rosvall & Bergstrom, 2008; Rosvall et al., 2009); node size represents Page Rank (Brin & Page, 1998); layout done with Force Atlas 2 in Gephi (Bastian et al., 2009); (colored version available online).

--

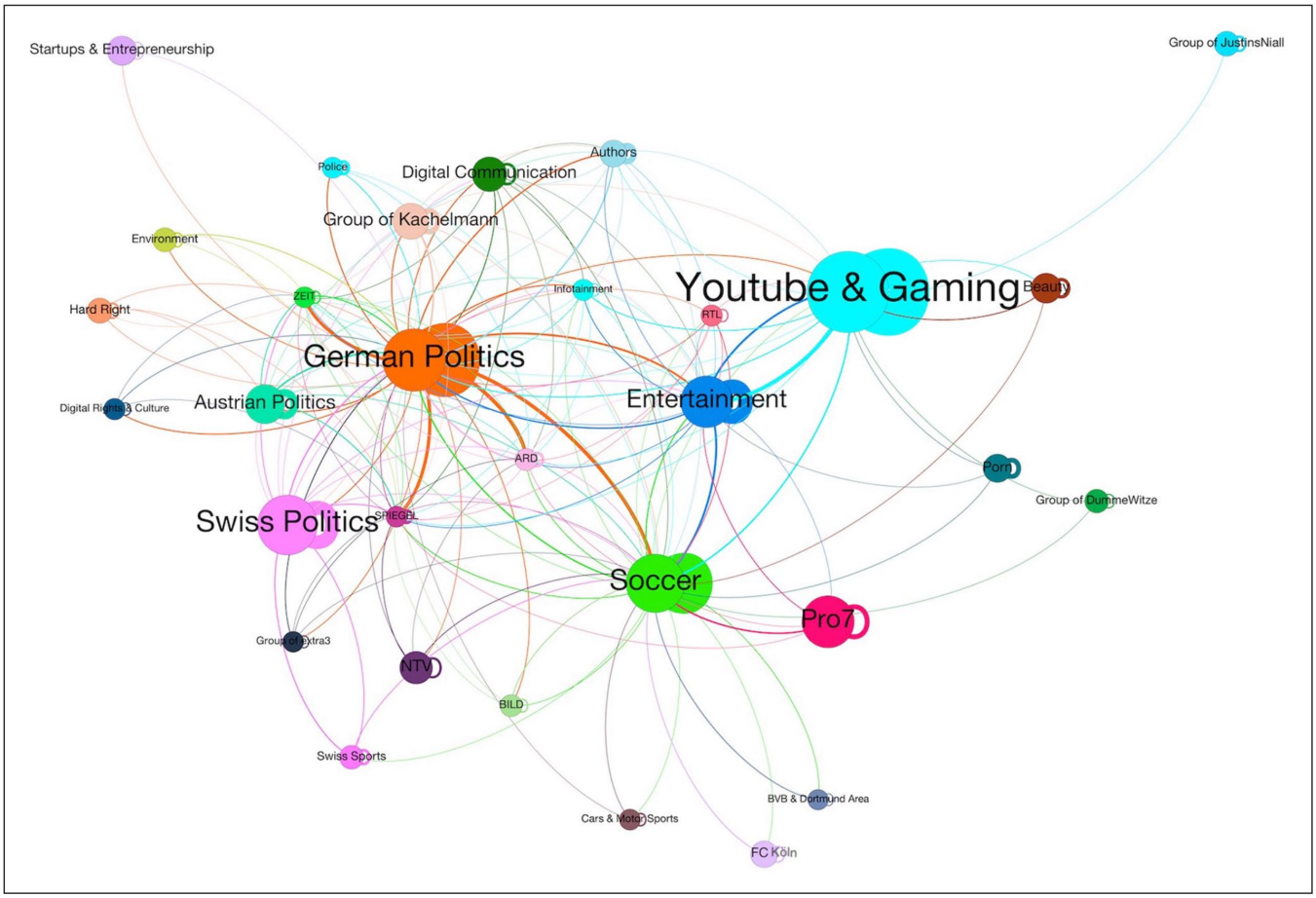


Figure 13. Community graph of communities in the 3-core of our sample with over 300 accounts, at least 80 active accounts during the examined timeframe, and edges with a weight of at least 150; edge width represents weight; edge direction follows clockwise curvature; edges colored by source node; node size represents the number of accounts in each community; node colors correspond with Figure 12; node labels based on interpretation of keywords and top accounts (see Supplemental Material); (colored version available online).

--

German-Italian Twittersphere

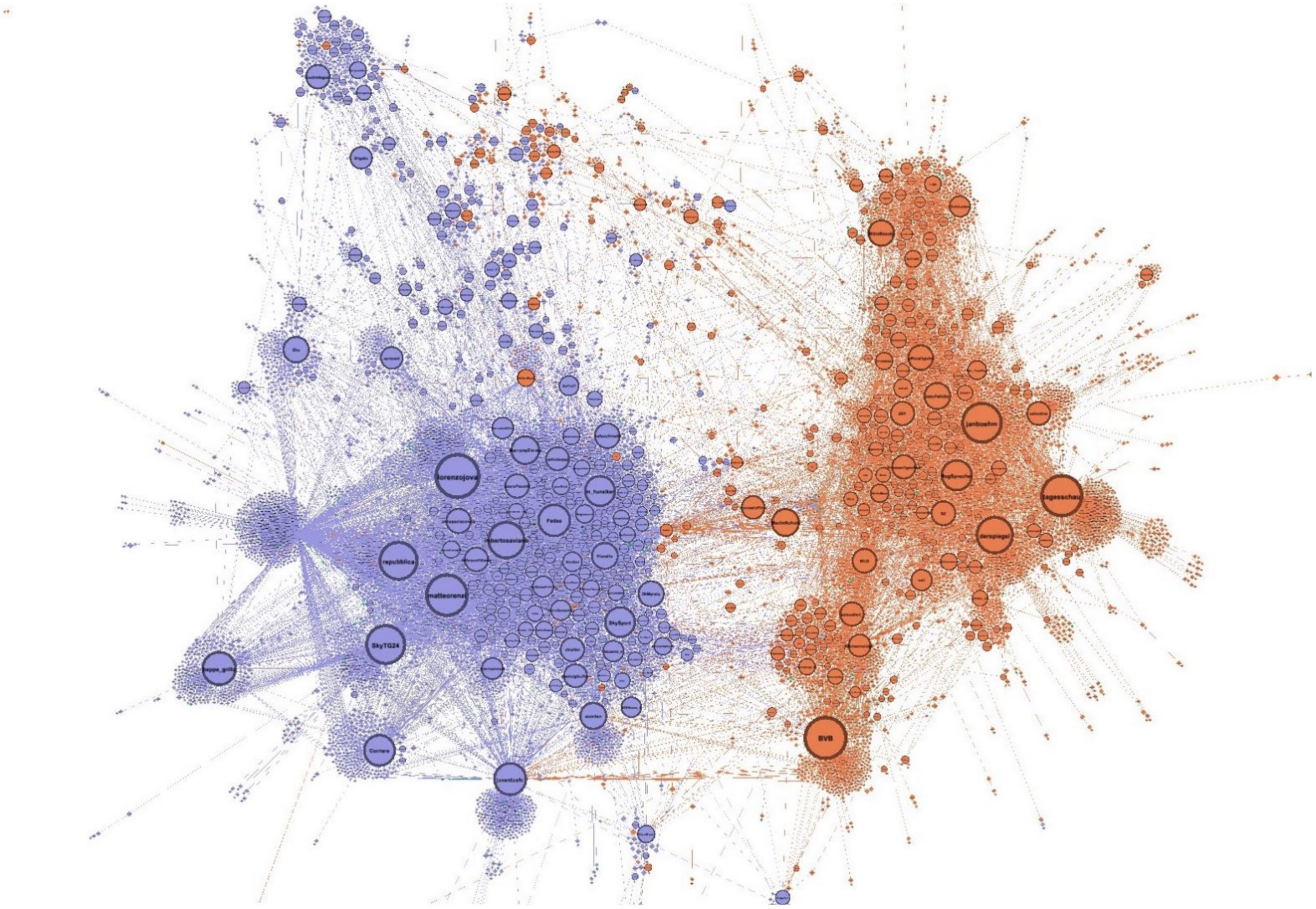


Figure 1: *Force directed layout (Force Atlas 2 (Bastian et al., 2009)) of the Italian (purple) - German (orange) follow network sample. Nodes sized by betweenness centrality (Brandes, 2001).*

--

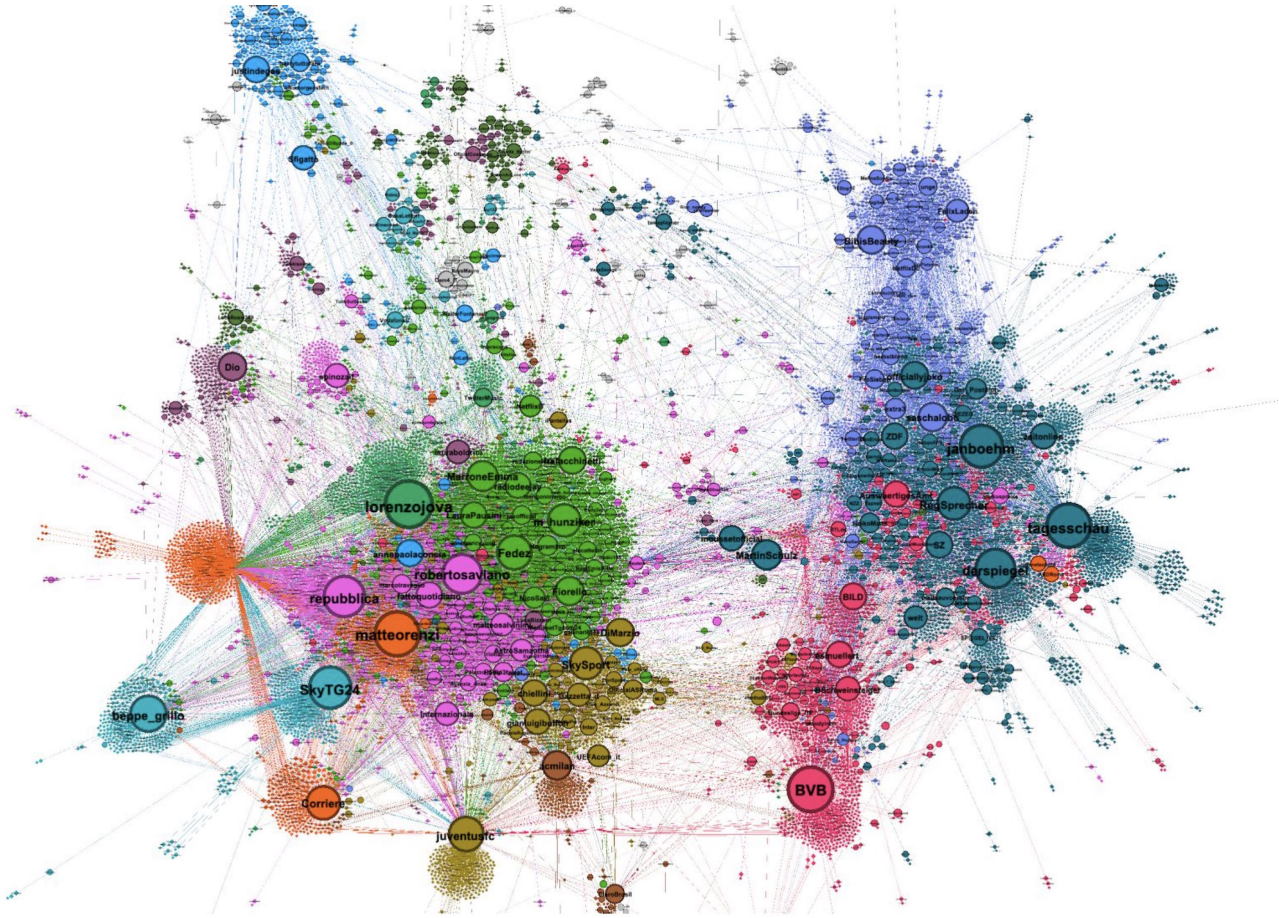


Figure 3: *Force directed layout (Force Atlas 2 (Bastian et al., 2009)) of the Italian (left) - German (right) follow network sample. Nodes sized by betweenness centrality (Brandes, 2001). Coloured by modularity maximising clusters.*

--

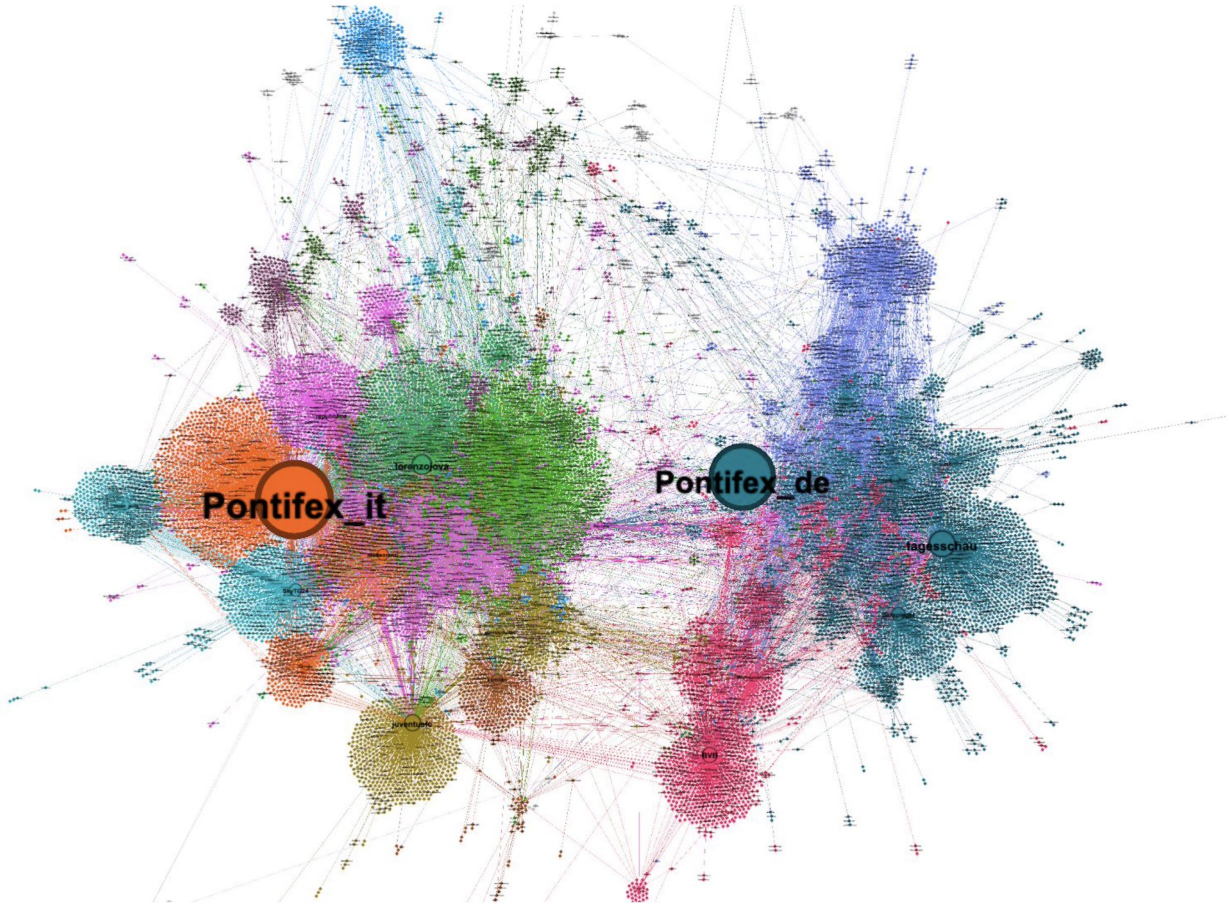


Figure 2: *Force directed layout (Force Atlas 2 (Bastian et al., 2009)) of the Italian (left) - German (right) follow network sample. Nodes sized by Page Rank (Brin & Page, 1998). Coloured by modularity maximising clusters.*

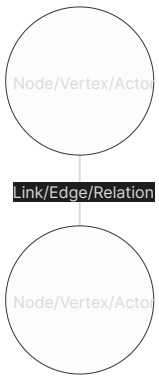
Network Analysis Fundamentals

--

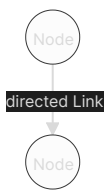
Elements and Properties of Networks

--

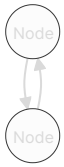
Dyads



--

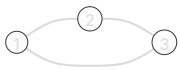


--



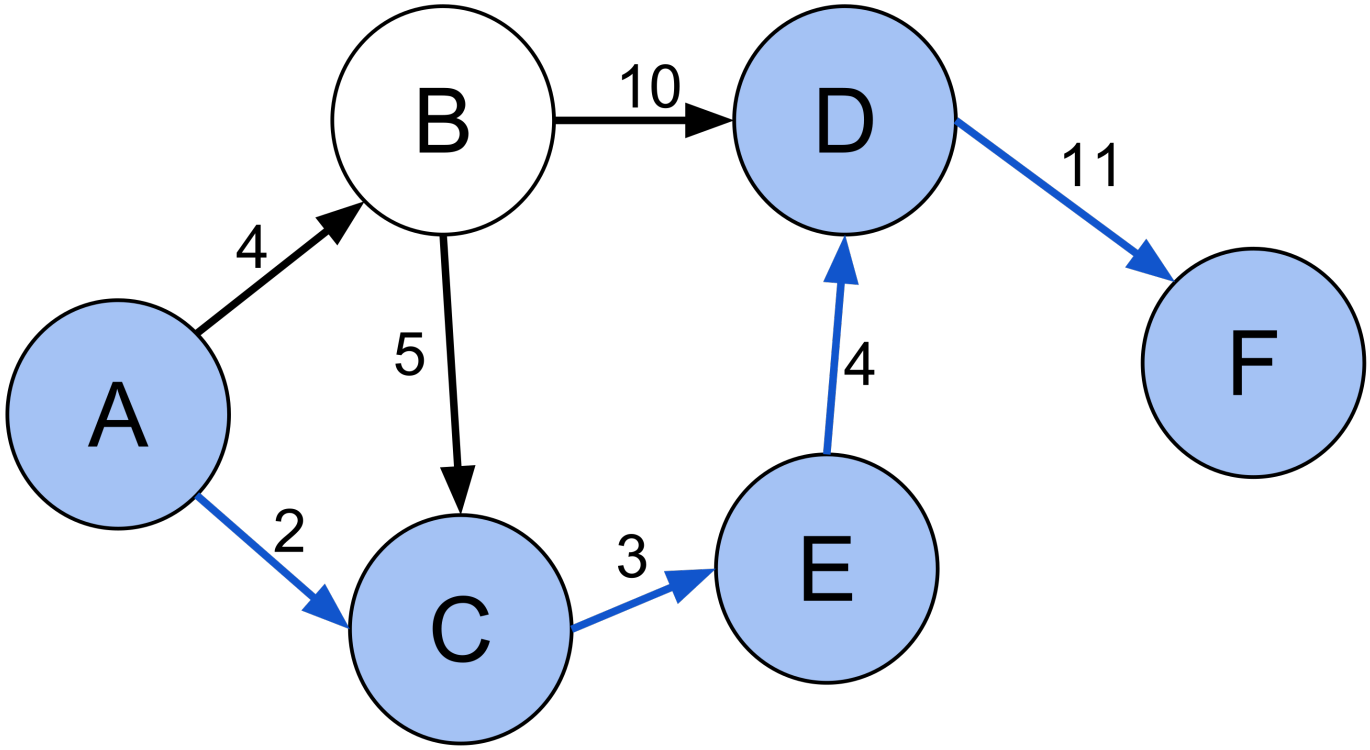
--

Triads



--

Weighted Links & (Shortest) Paths



Network Analysis Methods

--

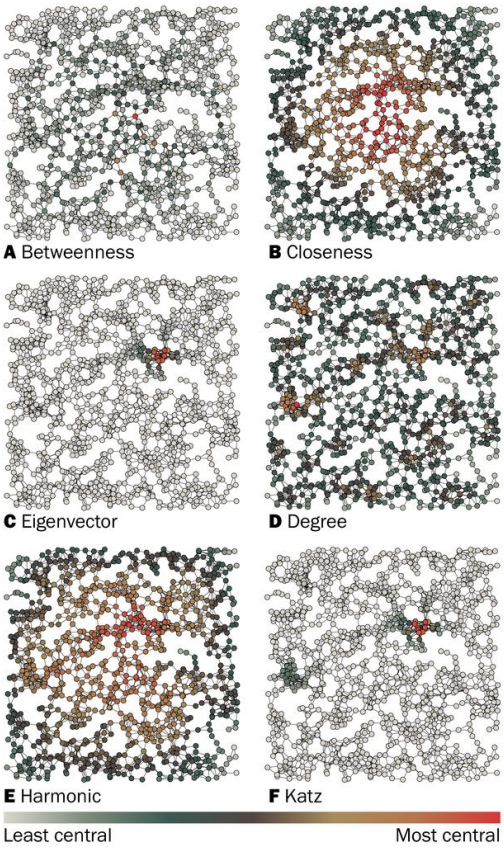
Measurements of Networks/Graphs and their Elements

--

Node Measures

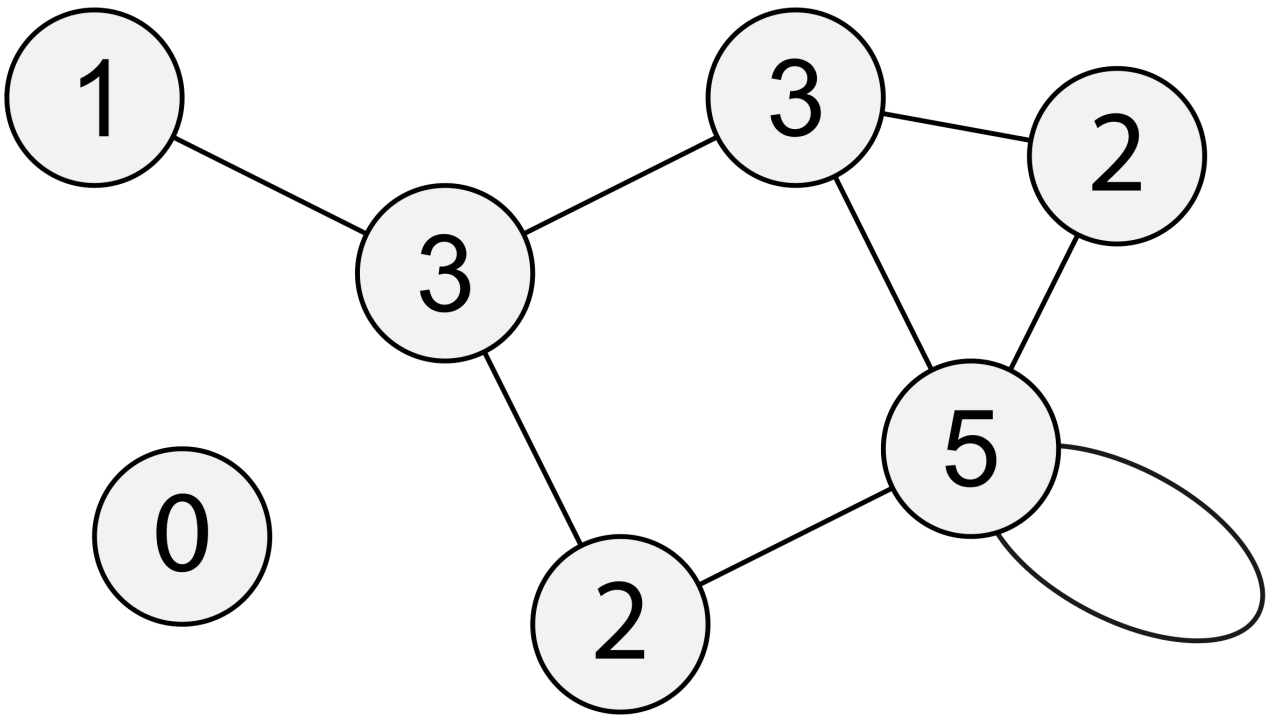
--

Important Centrality Measures



--

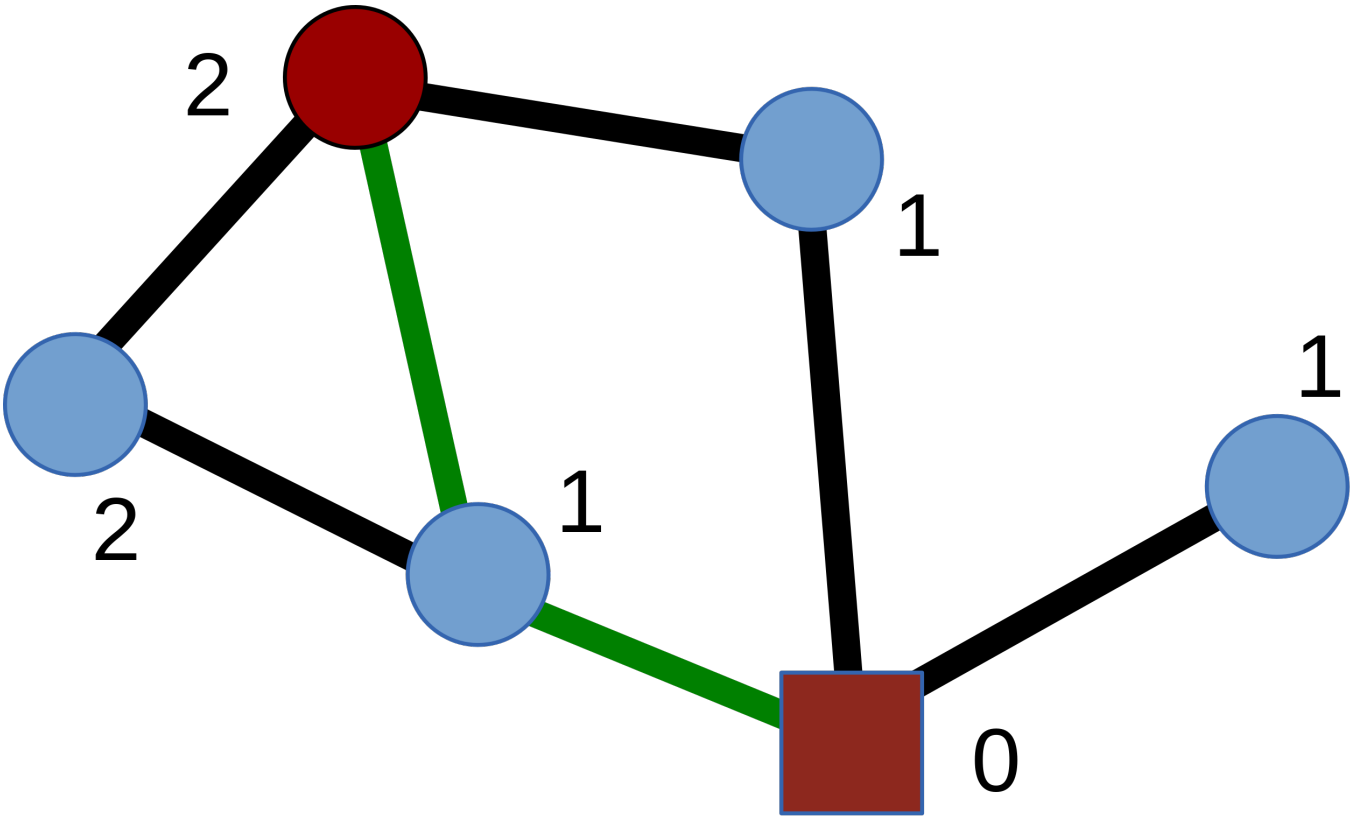
(In/Out-)Degree Centrality



--

Closeness Centrality

$$C(x) = \frac{N-1}{\sum_y d(y,x)}$$

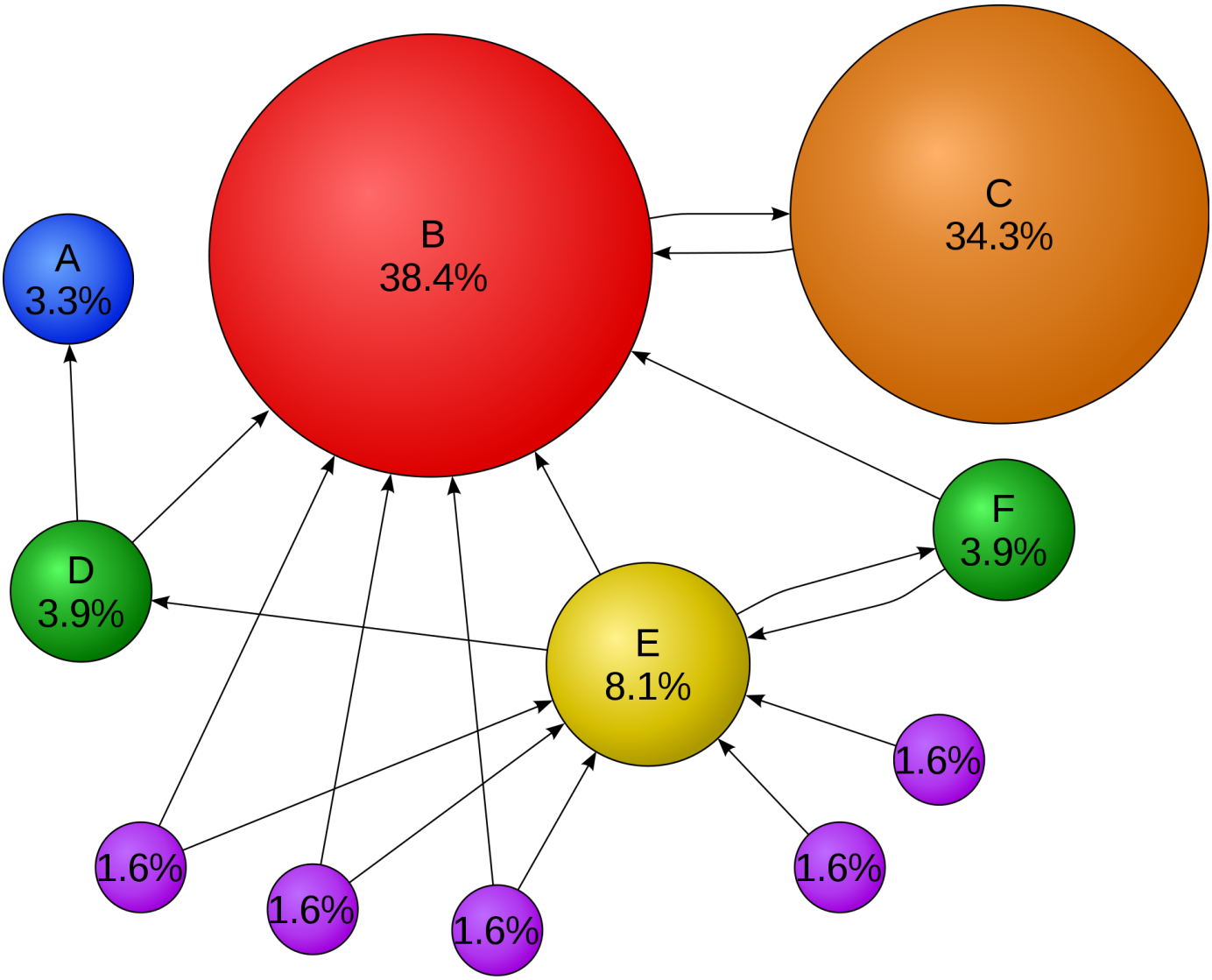


The red square node has (normalised) closeness centrality $\frac{(6-1)}{1+1+1+2+2}$

--

Eigenvector Centrality and Page Rank

Both based on the so-called Eigenvalue/Eigenvector equation of the adjacency matrix. In non-math terms: **nodes who have many high ranking nodes as neighbours rank high**.



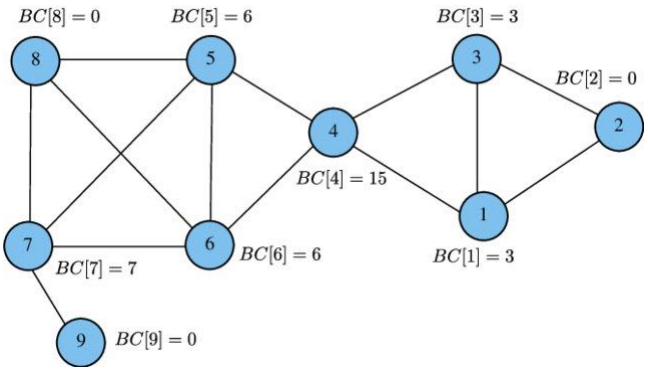
Simple Illustration of the PageRank Algorithm

Image Public Domain, https://commons.wikimedia.org/wiki/File:PageRanks-Example.svg

--

Betweenness-Centrality

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$



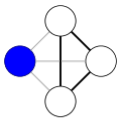
The more shortest paths are going through a node, the higher its betweenness.

Image Source: McLaughlin, Adam & Bader, David. (2015). Scalable and High Performance Betweenness Centrality on the GPU. International Conference for High Performance Computing, Networking, Storage and Analysis, SC. 2015. 572-583. 10.1109/SC.2014.52.

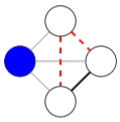
--

Local Clustering Coefficient

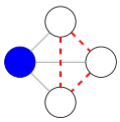
The number of realized edges divided by the number of possible edges between neighbouring nodes



c = 1



c = 1/3



c = 0

Note:
Image is public domain (https://commons.wikimedia.org/wiki/File:Clustering_coefficient_example.svg)

--

Important Global Network Measures

Global Clustering Coefficient $= \frac{\text{number of closed triplets}}{\text{number of all triplets}}$

Diameter: longest shortest path of the network

Density: $= \frac{\text{number of links}}{\text{number of possible links}}$

Average Shortest Path Length

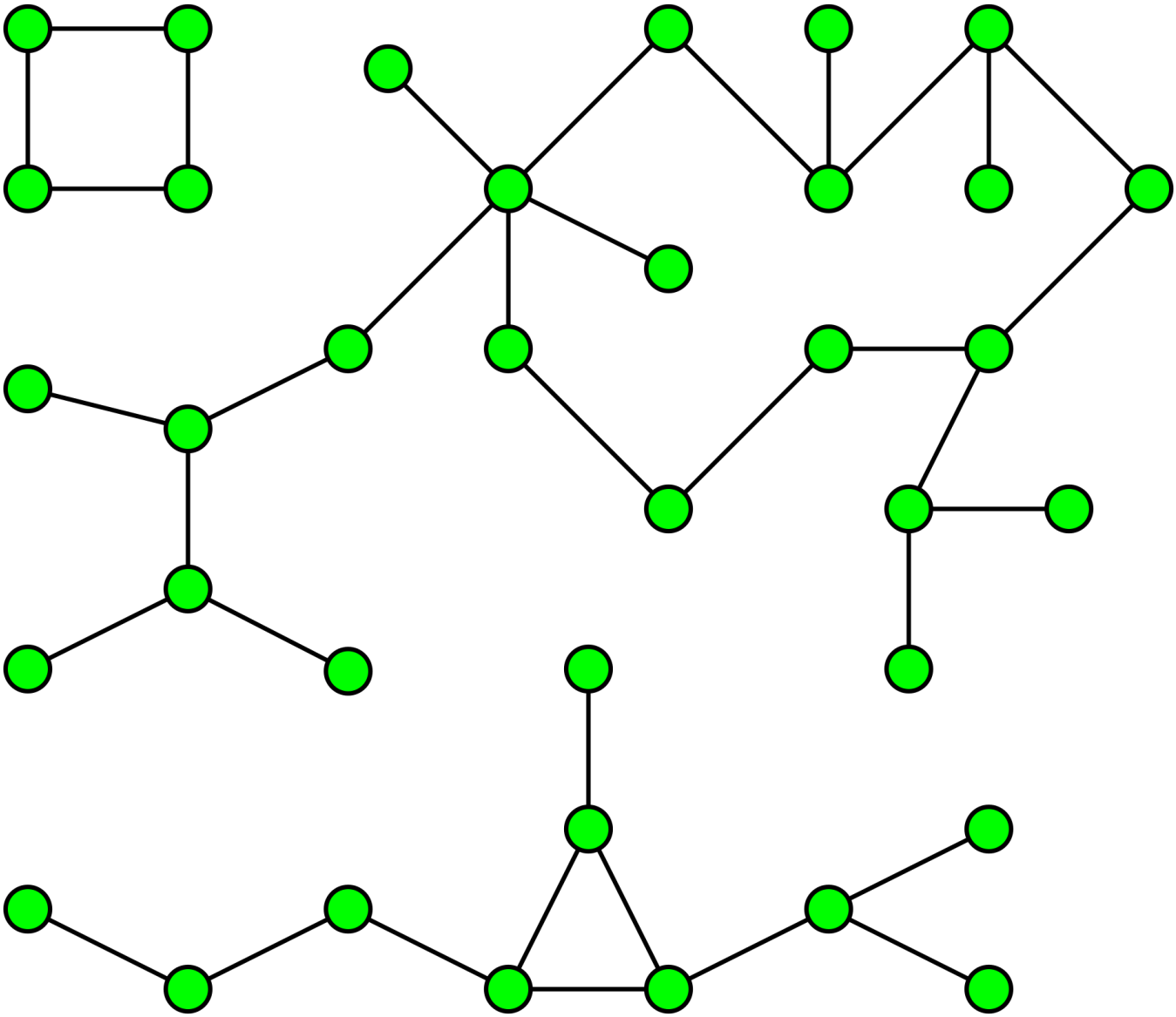
& Averages of most node measures (e.g., average degree, betweenness, closeness, ...)

--

Networks within Networks

--

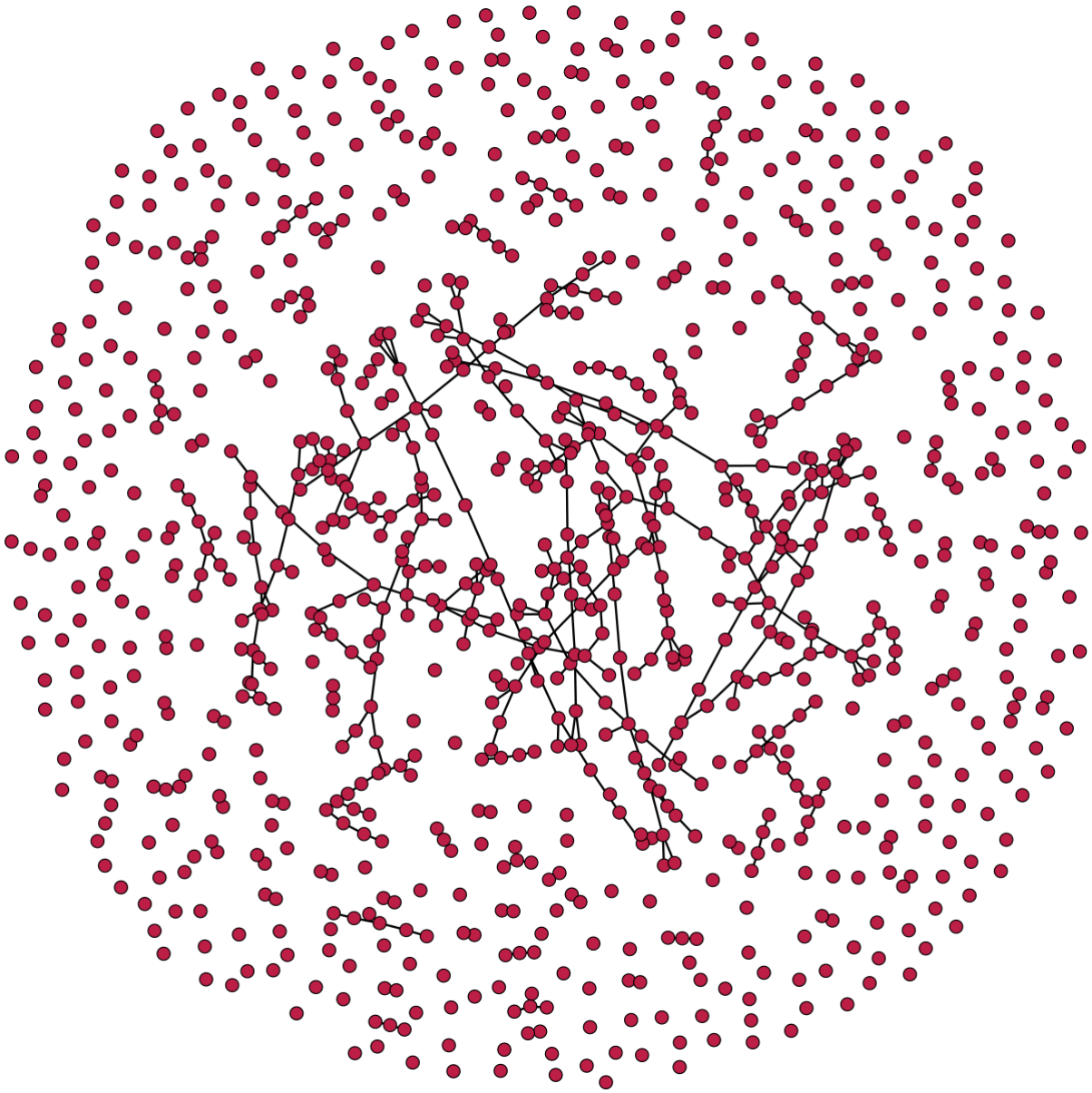
(Weakly) Connected Components



note:
Image Public domain https://en.wikipedia.org/wiki/Component_(graph_theory)#/media/File:Pseudoforest.svg

--

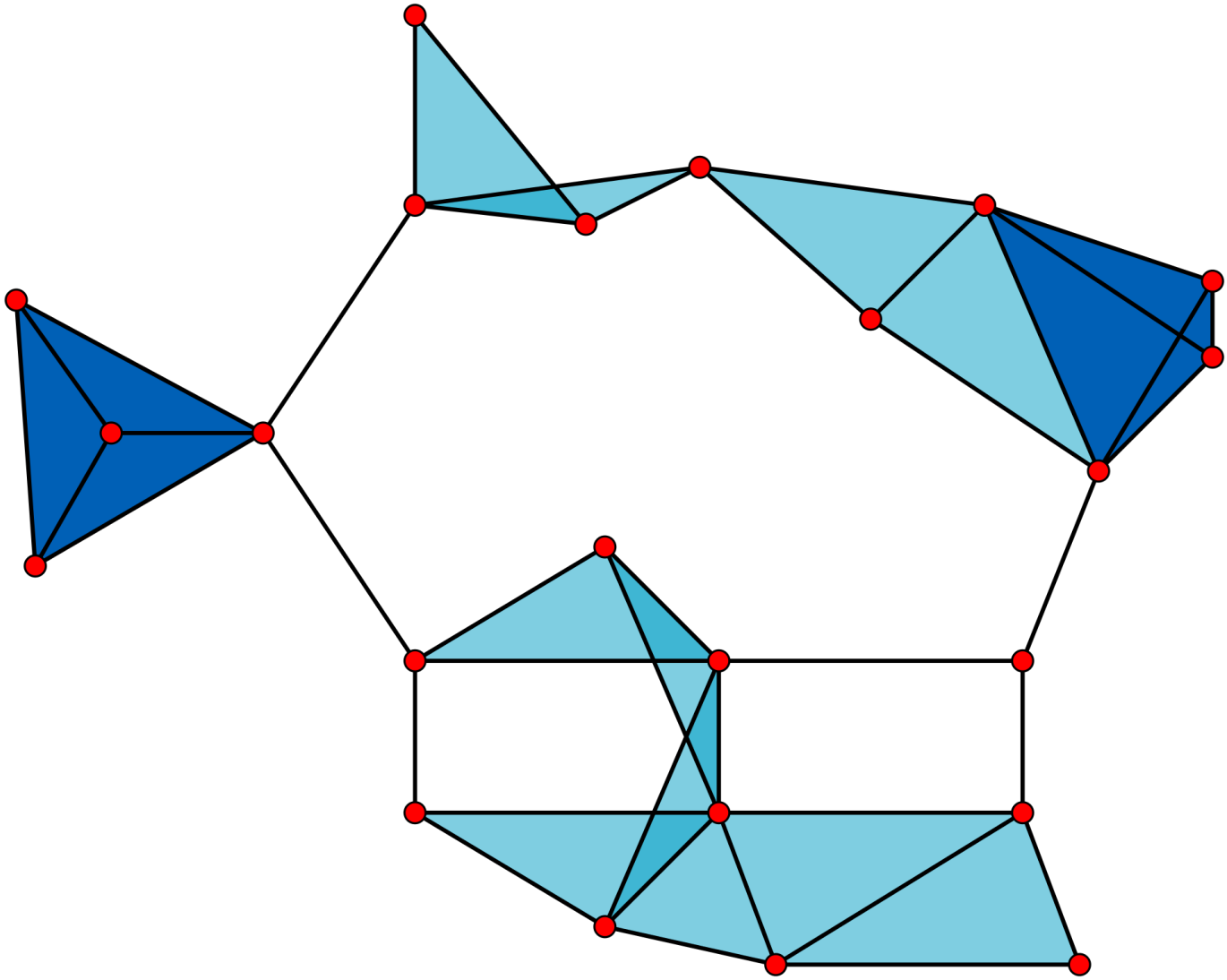
Giant Component



note:
Image CC0, https://commons.wikimedia.org/wiki/File:Critical_1000-vertex_Erd%C5%91s%E2%80%93R%C3%A9nyi%E2%80%93Gilbert_graph.svg

--

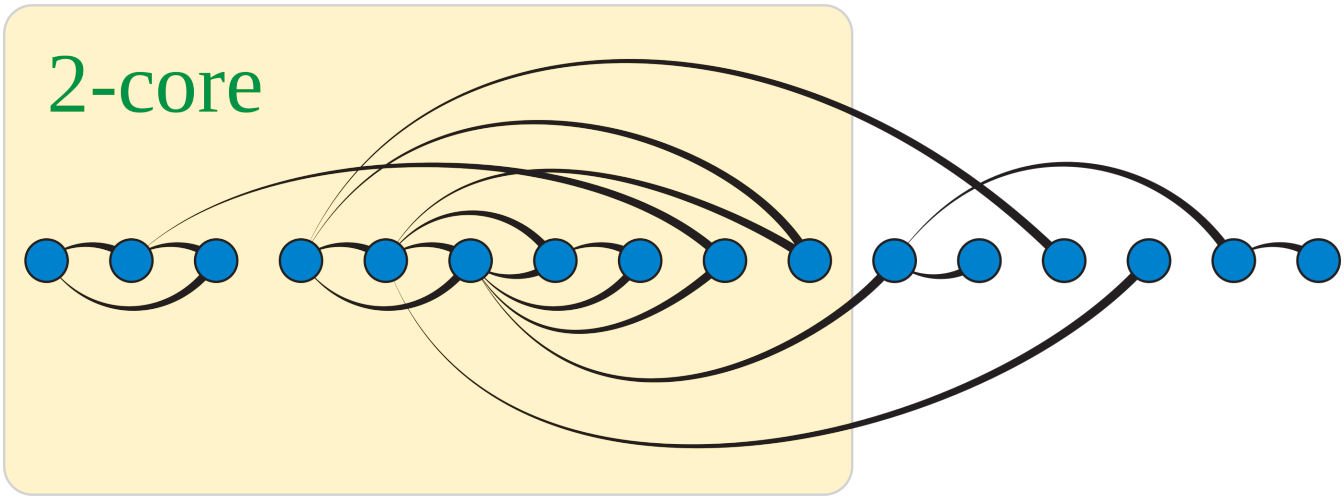
Cliques



note:
Image is Public Domain: https://en.wikipedia.org/wiki/Clique_(graph_theory)#/media/File:VR_complex.svg

--

K-Cores



(can also be used as a node centrality measure)

note:
The graph that remains after iteratively removing every node with less than k links.
Image is CC0/public domain: https://en.wikipedia.org/wiki/Degeneracy_(graph_theory)#/media/File:2-degenerate_graph_2-core.svg

--

Communities/Clusters

Depend on the detection algorithms used. Two of the most popular are

- ***Modularity Maximisation (mostly in the Louvain implementation)*** based on the relative density of in-/out-group edges

and

- ***Map Equation (infomap)*** based on the lenght of stay of random walks in certain regions of the network (technically the minimization of the description length of its path)

--

Flat communities

- simplify complex systems
- easy to understand
- can oversimplify
- inherently have a resolution limit/arbitrary resolution

--

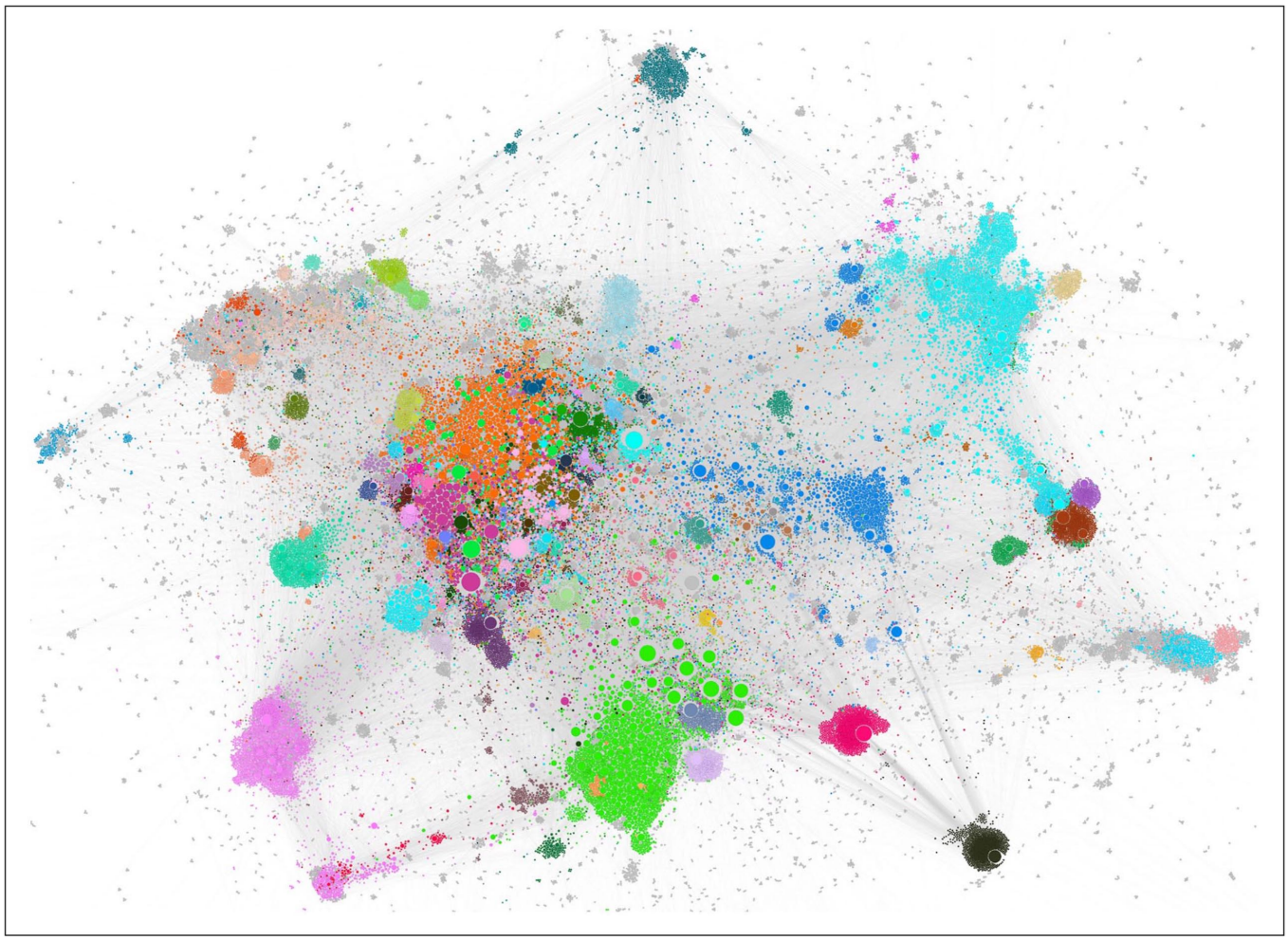


Figure 12. Central communities in the 3-core of our sample network; colored by largest communities detected with the Infomap community detection algorithm (Rosvall & Bergstrom, 2008; Rosvall et al., 2009); node size represents Page Rank (Brin & Page, 1998); layout done with Force Atlas 2 in Gephi (Bastian et al., 2009); (colored version available online).

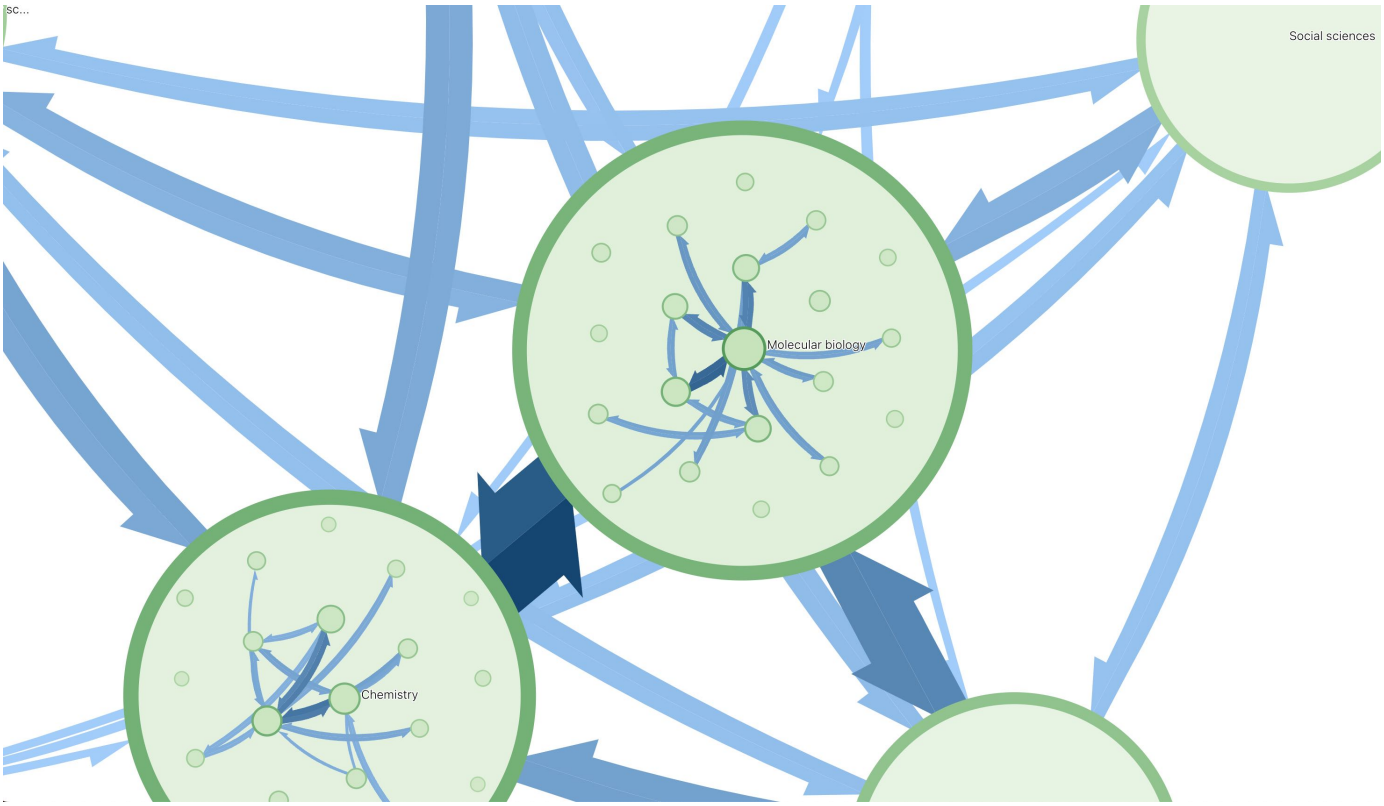
Münch, F. V., Thies, B., Puschmann, C., & Bruns, A. (2021). Walking Through Twitter: Sampling a Language-Based Follow Network of Influential Twitter Accounts. *Social Media + Society*. https://doi.org/10.1177/2056305120984475

--

Hierarchical and Overlapping Communities

- Possible, e.g. with infomap
- often closer to reality
- often easier to interpret
- harder to analyse and visualise

--



https://www.mapequation.org/navigator/

--

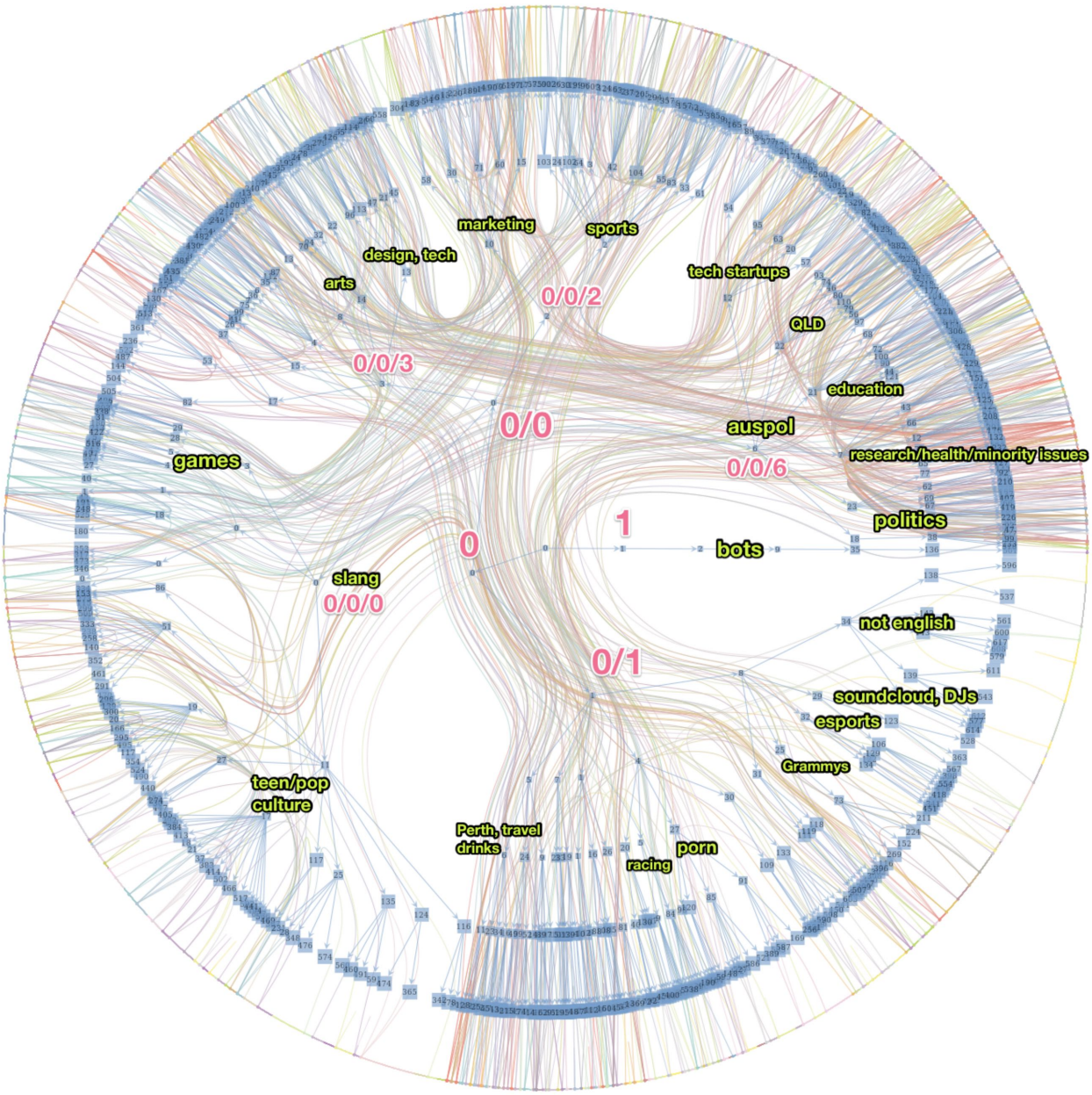


Figure 6.22: Annotated visualisation of the nested stochastic block model inferred for the filtered Australian follow network, edges sampled down to 1000 edges; nodes on outer circle represent accounts; labels based on clearly interpretable results from keyword extraction for blocks detected on level 2; labelling not complete and for overview only; pink labels used for references in text. (High resolution version available online.)

Münch, F. V. (2019). *Measuring the Networked Public – Exploring Network Science Methods for Large Scale Online Media Studies* [PhD thesis, Queensland University of Technology]. https://doi.org/10.5204/thesis.eprints.125543

--

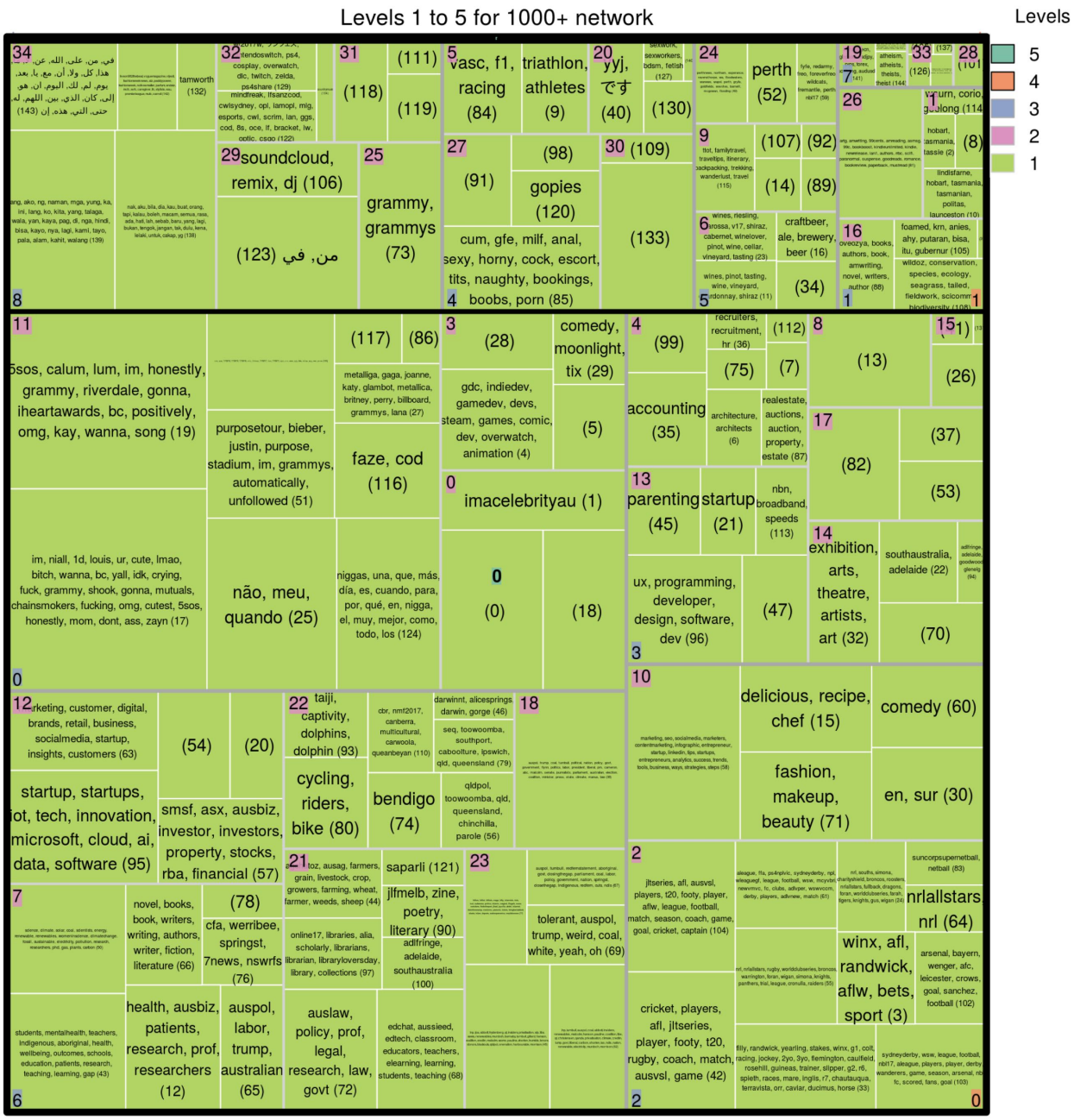


Figure 6.34: Treemap of blocks at levels 1 to 5, as inferred based on nested stochastic block models in the filtered Australian follow network; size of rectangles represents number of active accounts in a block; colour shows the level of label numbers as per legend; most distinctive keywords used by at least 5 percent of the accounts in blocks at level 1 are shown. (Higher resolution version available online.)

Münch, F. V. (2019). *Measuring the Networked Public – Exploring Network Science Methods for Large Scale Online Media Studies* [PhD thesis, Queensland University of Technology]. https://doi.org/10.5204/thesis.eprints.125543

--

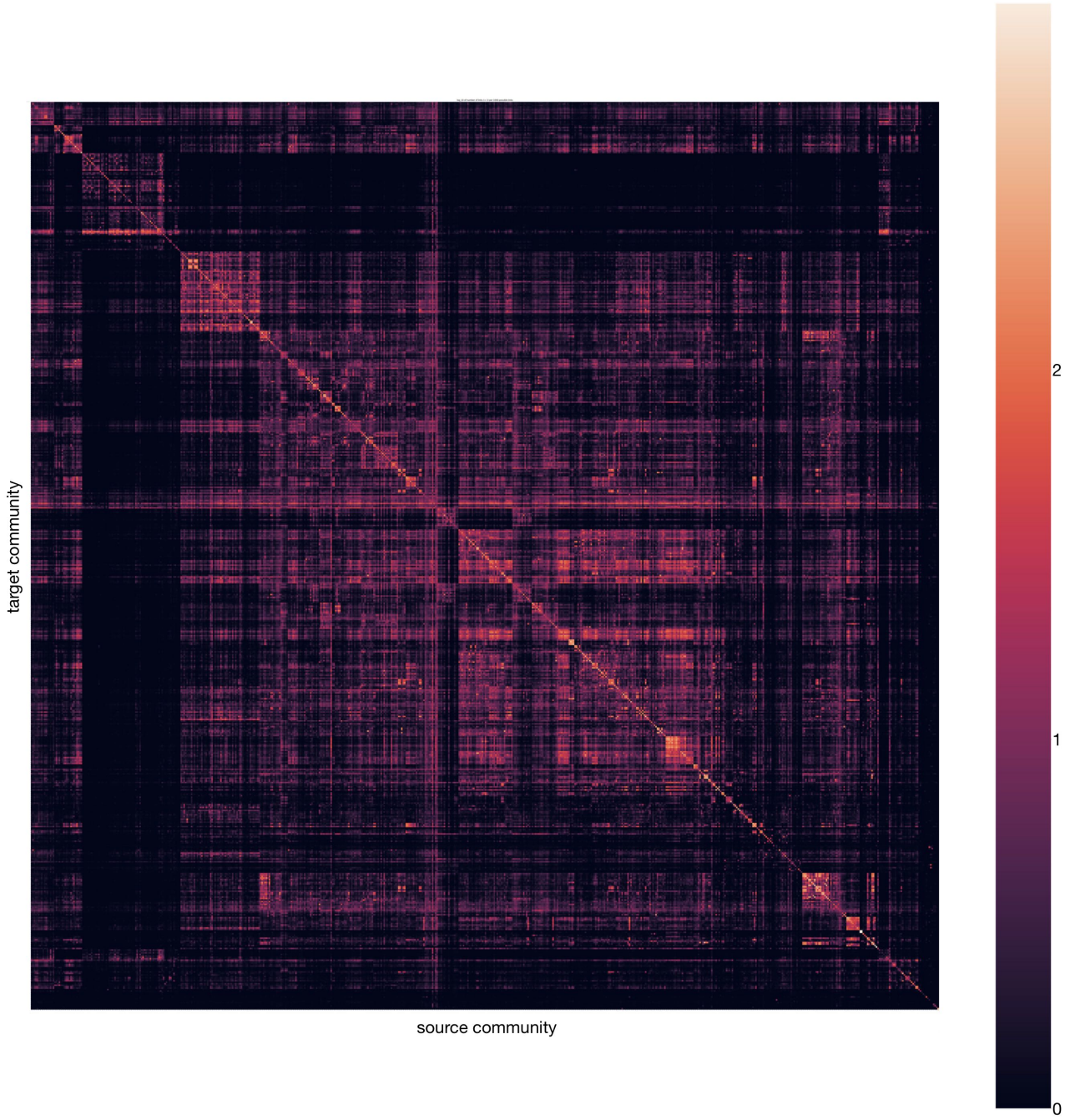


Figure 6.40: Visualisation of the adjacency matrix of the directed block graph of the filtered Australian follow network at level 0, as inferred based on nested stochastic block models; colour shows the logarithm to base 10 of the number of connections + 1 per 1000 possible connections between the source block on the horizontal, and the target block on the vertical axis.

Münch, F. V. (2019). *Measuring the Networked Public – Exploring Network Science Methods for Large Scale Online Media Studies* [PhD thesis, Queensland University of Technology]. https://doi.org/10.5204/thesis.eprints.125543

Data Sources for (Online ((Social) Media)) Networks

--

Repositories

e.g.:

- Netzschleuder: https://networks.skewed.de/,
- SNAP datasets: https://snap.stanford.edu/data/index.html
- Network Repository: https://networkrepository.com/
- and many more

--

Repositories

Pro	Contra
Easy to access	Old data

Pro	Contra
Fewer legal and ethical problems	Already studied, harder to find new topics
Good for meta studies	Need to trust the data collector
Good for method testing	Available info not tailored to your needs

--

API

Pro	Contra
New/live data	Often not historical data
More control over what to collect	Ethics and data protection considerations apply
Relatively stable machine readability	Often vetting and acceptance of Terms of Service (TOS)
Legally quite safe	Rate limits and accessible data shape research question
Sometimes access to additional metadata	Can be deprecated/shut down

--

Webscraping

Pro	Contra
New/live data	Kind of unstable machine readability
More control over what to collect	More Ethics and data protection considerations apply
No vetting or acceptance of TOS	Active countermeasures by platforms
No rate limits	Technically often more complex setup
In best case WYSIWYG	

--

Also possible, but less common for large networks:

- Surveys
- Questionaires
- Data Donations
- "Manual" Collection
- ...

Questions?

@flxvctr