21_03_2023

Changes on already written sections

- Modified papers cited on RQ3
- Re-compiling images to have inner legend
- Description of CORE / GGS rating
- Justification for using SJR based on Scopus

Writing progress

- Information sources
- Incl. / Excl. criteria
- Search Strategy
- Snowballing (methods)
- Added "Papers reviewed" appendix
- Data collection process
- Data items
- Study selection
- Results of Snowballing activities

Upcoming writing progress

- Study characteristics
- Analysis of literature networks generated so far
- Analysis and synthesis methods
- Theoretical foundations

Assessment of the established data collection items (on 39 papers)

Initial set of data items:

Data extraction form Item nr.	Item	Description	RQ	Mandatory
1	Year	Publication year	-	
2	Author(s)	Publication author(s)	-	
3	Title	Publication title	-	
4	Venue	Publication venue	-	
5	Topic labeling	Topic labeling approach(es)	RQ1	
6	Focus	Primary / Secondary focus on topic labeling	RQ1	
7	Type of contribution	Established / Novel approach for topic labeling	RQ1	
8	Underlying technique	Technique / Algorithm on which the topic labeling approach is based	RQ1	x
9	Topic labeling parameters	Parameter names and values used for topic labeling	RQ1	x
10	Approach details	Details of the employed approach	RQ1	
11	Motivation	What was the main motivator behind employing the topic labeling step?	RQ1	
12	Topic modeling	Underlying topic modeling approach(es)	RQ2	
13	Topic modeling parameters	Parameter names and values used for topic modeling	RQ2	
14	Label	Label description (e.g. single- multi-word) and nr of candidate labels per topic	RQ3	
15	Label selection	Selection approach(es) for label candidates	RQ3	x
16	Label quality evaluation	Quality metric(s) for label evaluation	RQ3	x
17	Assessors	Number and details of the assessors involved in the selection and evaluation	RQ3	x
18	Domain	Domain(s) of interest	RQ4	
19	Corpus	Origin, format, shape and content of the corpus	RQ4	
20	Document	Format of individual documents in the corpus	RQ4	
21	Pre-processing	Pre-processing steps performed on documents	RQ4	

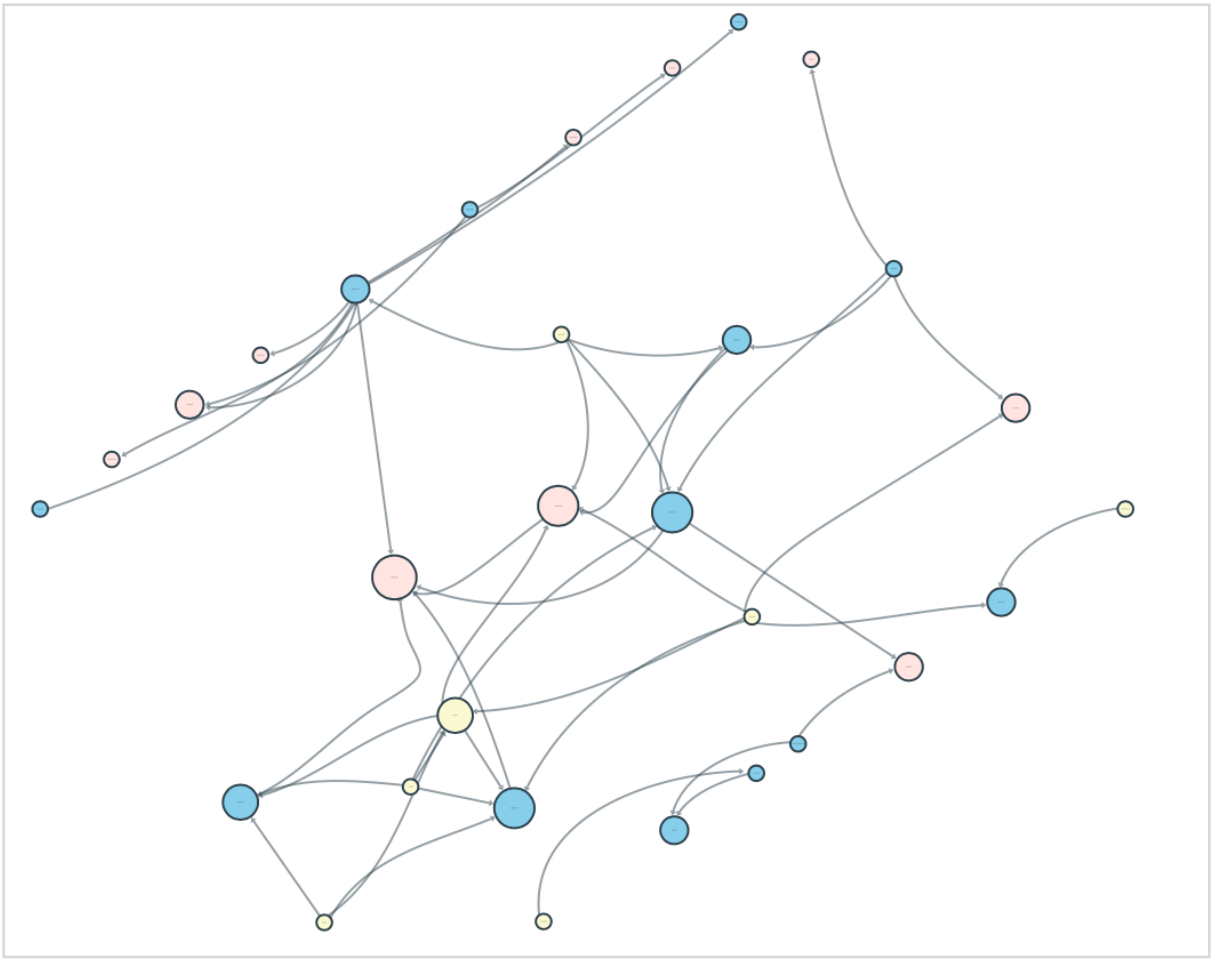
Updated set of data items:

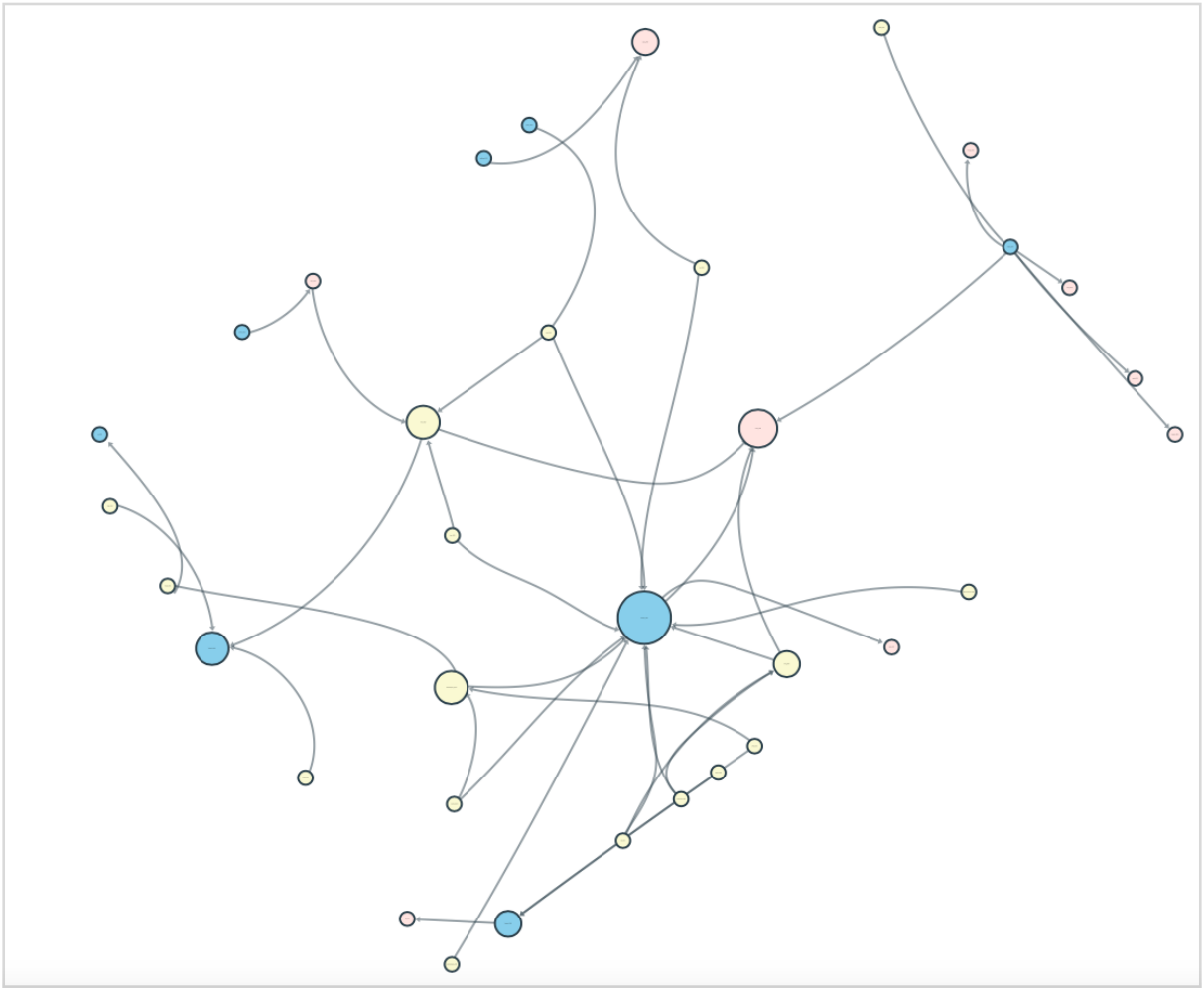
Data extraction form Item nr.	Item	Description	RQ	Mandatory
1	Year	Publication year	-	
2	Author(s)	Publication author(s)	-	
3	Title	Publication title	-	
4	Venue	Publication venue	-	
5	Topic labeling	Topic labeling approach(es)	RQ1	
6	Focus	Primary / Secondary focus on topic labeling	RQ1	
7	Type of contribution	Established / Novel approach for topic labeling	RQ1	
8	Underlying technique	Technique / Algorithm on which the topic labeling approach is based	RQ1	
9	Topic labeling parameters	Parameter names and values used for topic labeling	RQ1	x
10	Label generation	Label generation process related to the proposed technique	RQ1	x
11	Motivation	Motivation for applying a labeling step or for introducing an approach	RQ1	x
12	Topic modeling	Underlying topic modeling approach(es)	RQ2	
13	Topic modeling parameters	Parameter names and values used for topic modeling	RQ2	
14	Nr. of topics	Nr of topics generated from the corpus	RQ2	
15	Label	Label description (e.g. single- multi-word) and nr of candidate labels per topic	RQ3	
16	Label selection	Selection approach(es) for label candidates	RQ3	x
17	Label quality evaluation	Quality metric(s) for label evaluation	RQ3	x
18	Assessors	Number and details of the assessors involved in the selection and evaluation	RQ3	x
19	Domain	Domain(s) of interest	RQ4	
20	Problem statement	Summary of problem statement	RQ4	
21	Corpus	Origin and shape of the corpus	RQ4	
22	Document	Format of individual documents in the corpus	RQ4	x
23	Pre-processing	Pre-processing steps performed on documents	RQ4	

Paper graphs

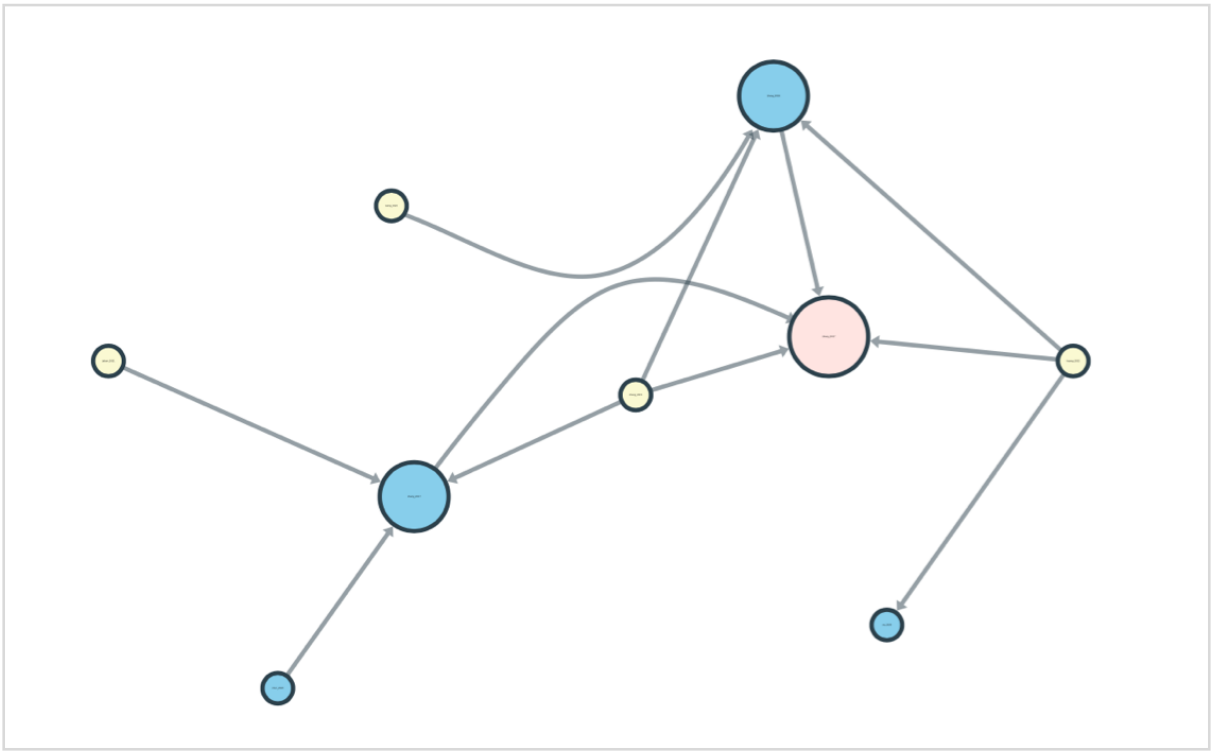
The initial graphs have been re-generated using the netgraph library, which offer a wider array of formatting capabilities and allows to more easily visualise distinct components within a given network.

For instance, if we take the graph related to the set of selected papers (including the ones selected with forward and backward snowballing), we obtain two large components (>20 nodes):

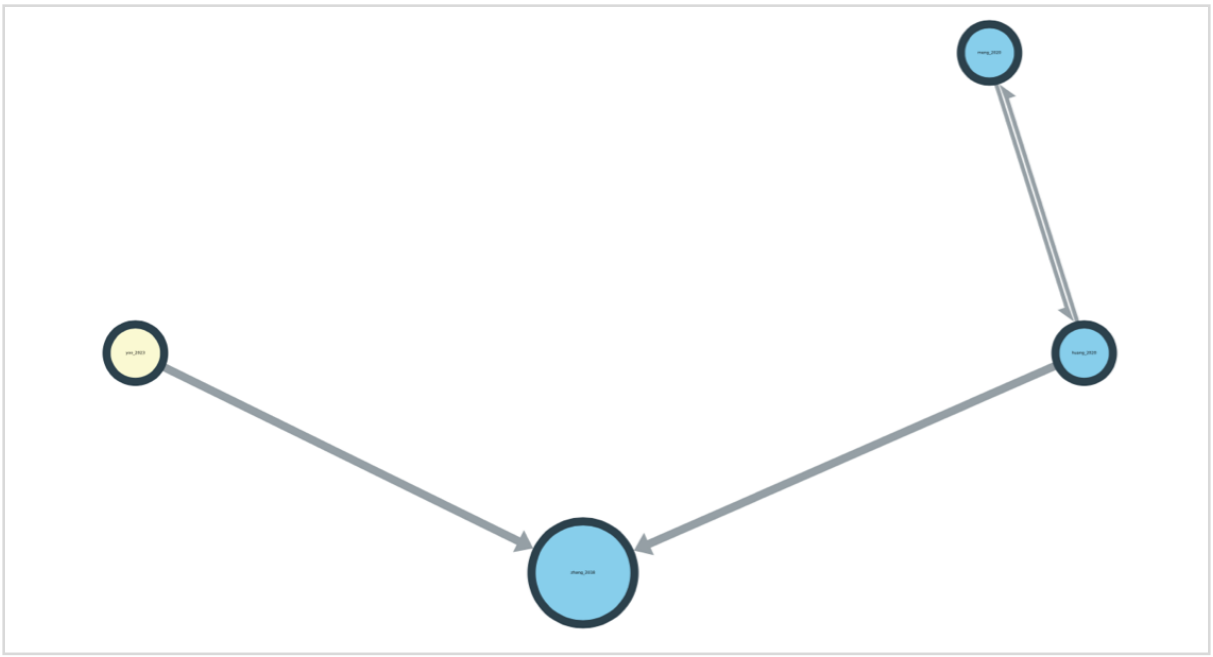


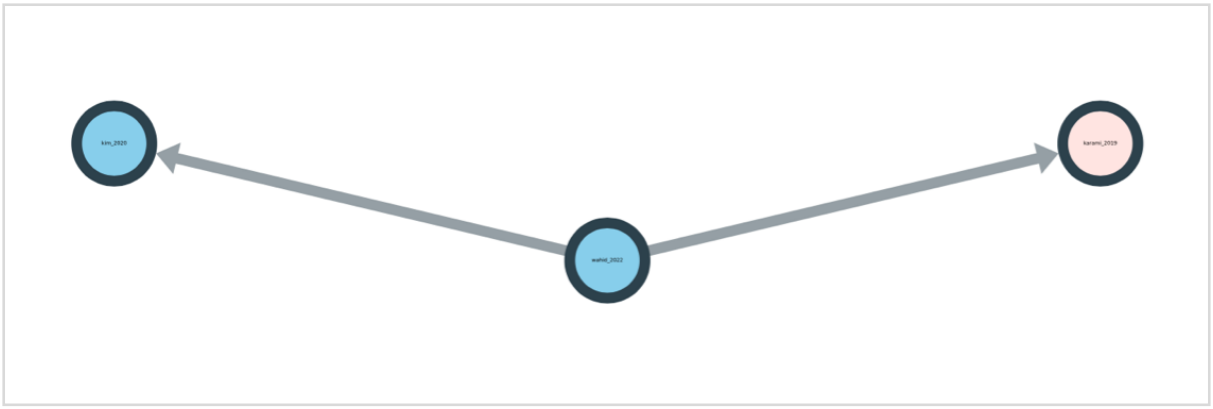


One medium sized component:

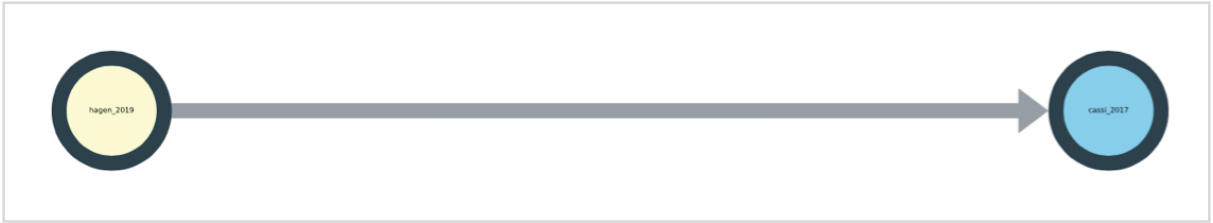


Two small components:





And 9 two-nodes isolates:



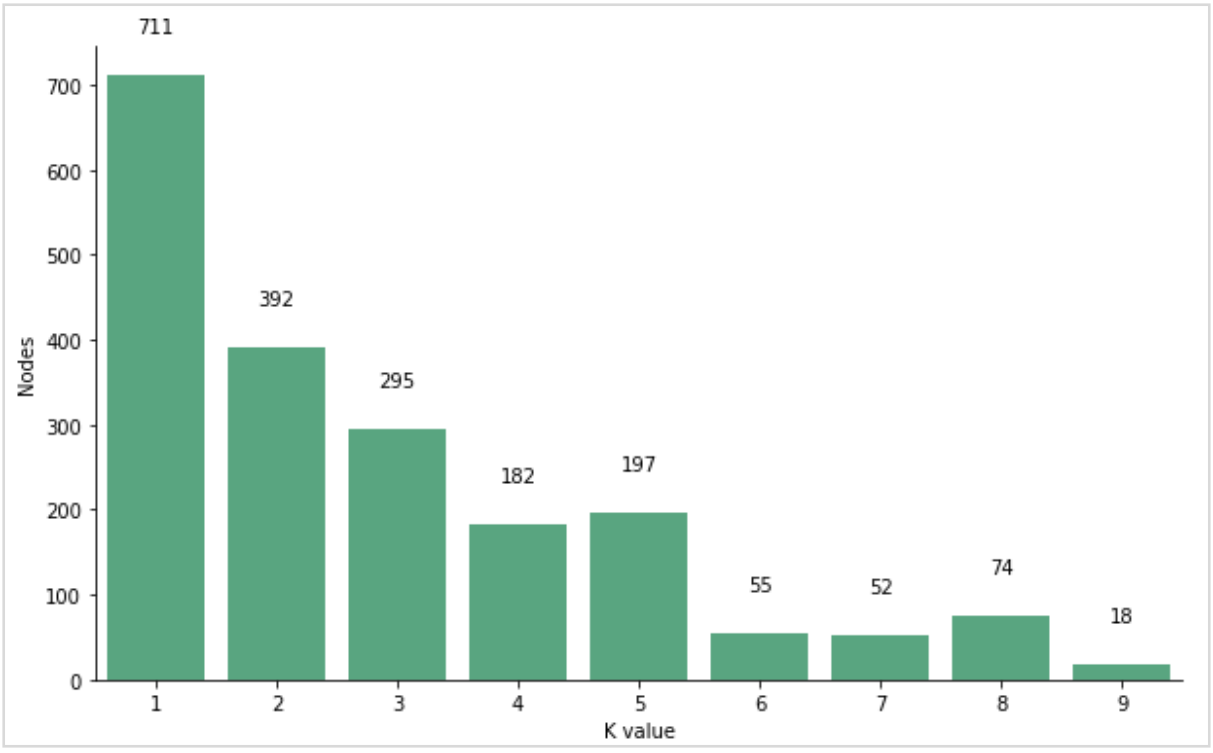
Three of which are authors citing their old paper:

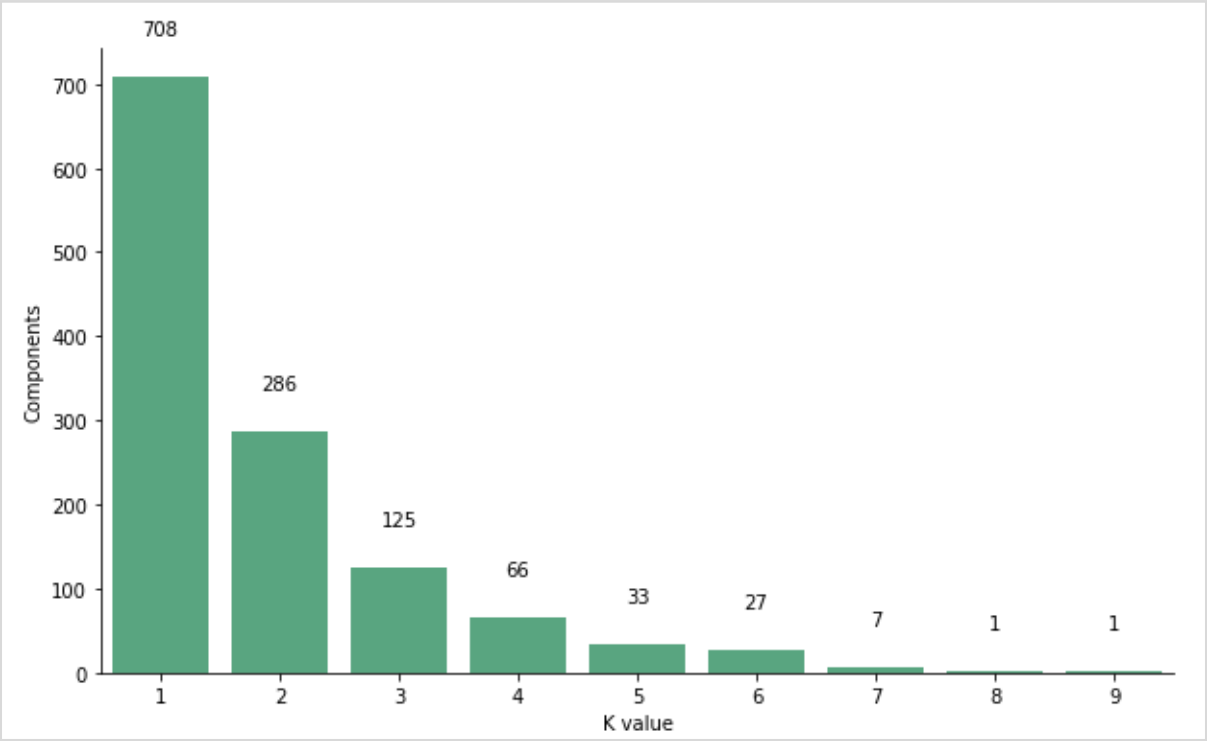
This division into components is much easier to visually analyse when generating the graph with netgraph as opposed to the visualisation created with NetworkX

k-cores

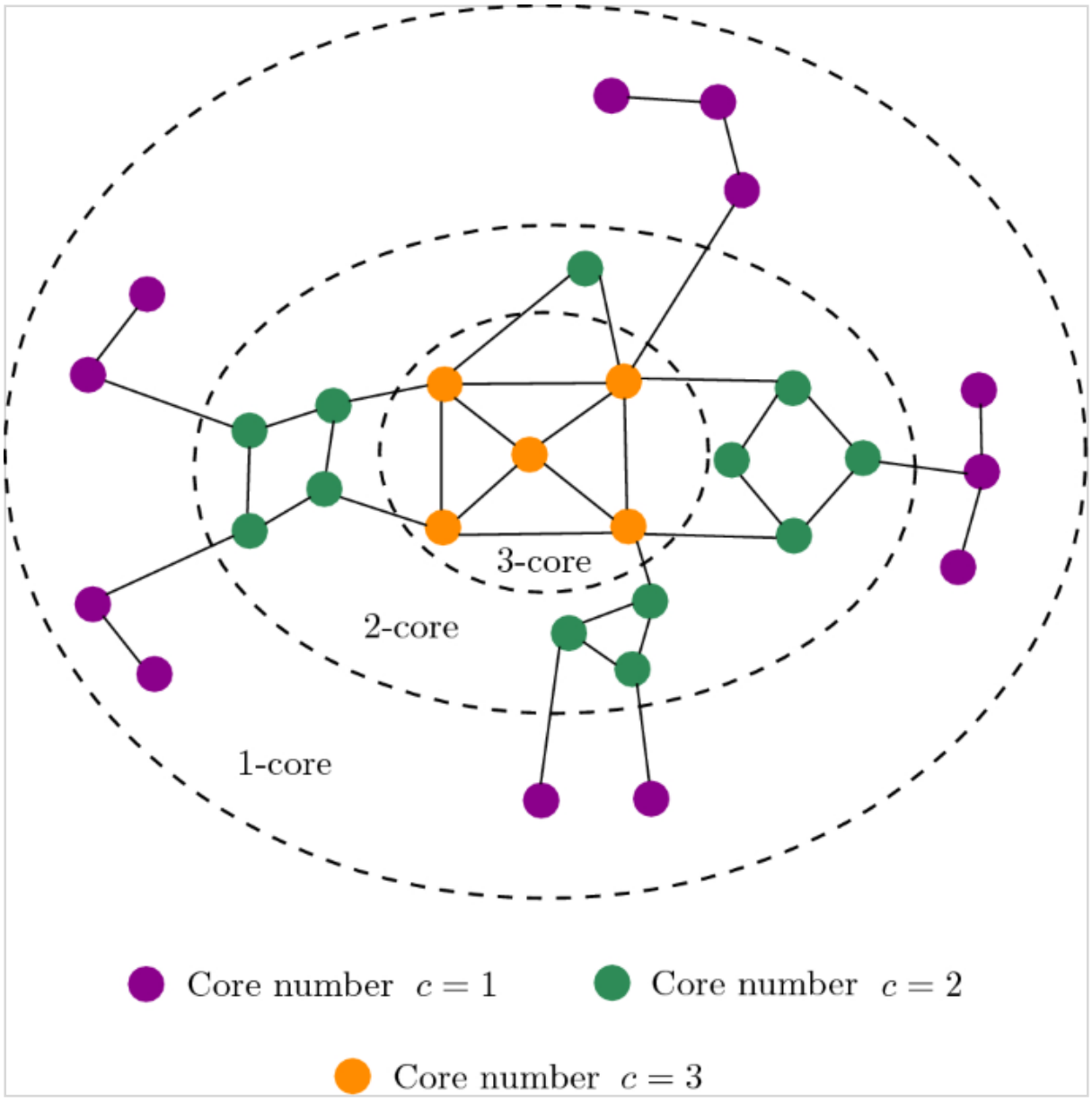
k_core – NetworkX 3.0 documentation returns all the (possibly disconnected) k-cores in the graph given a value for k.

For directed graphs the node degree is defined to be the in-degree + out-degree.





Initial idea - K-shell decomposition

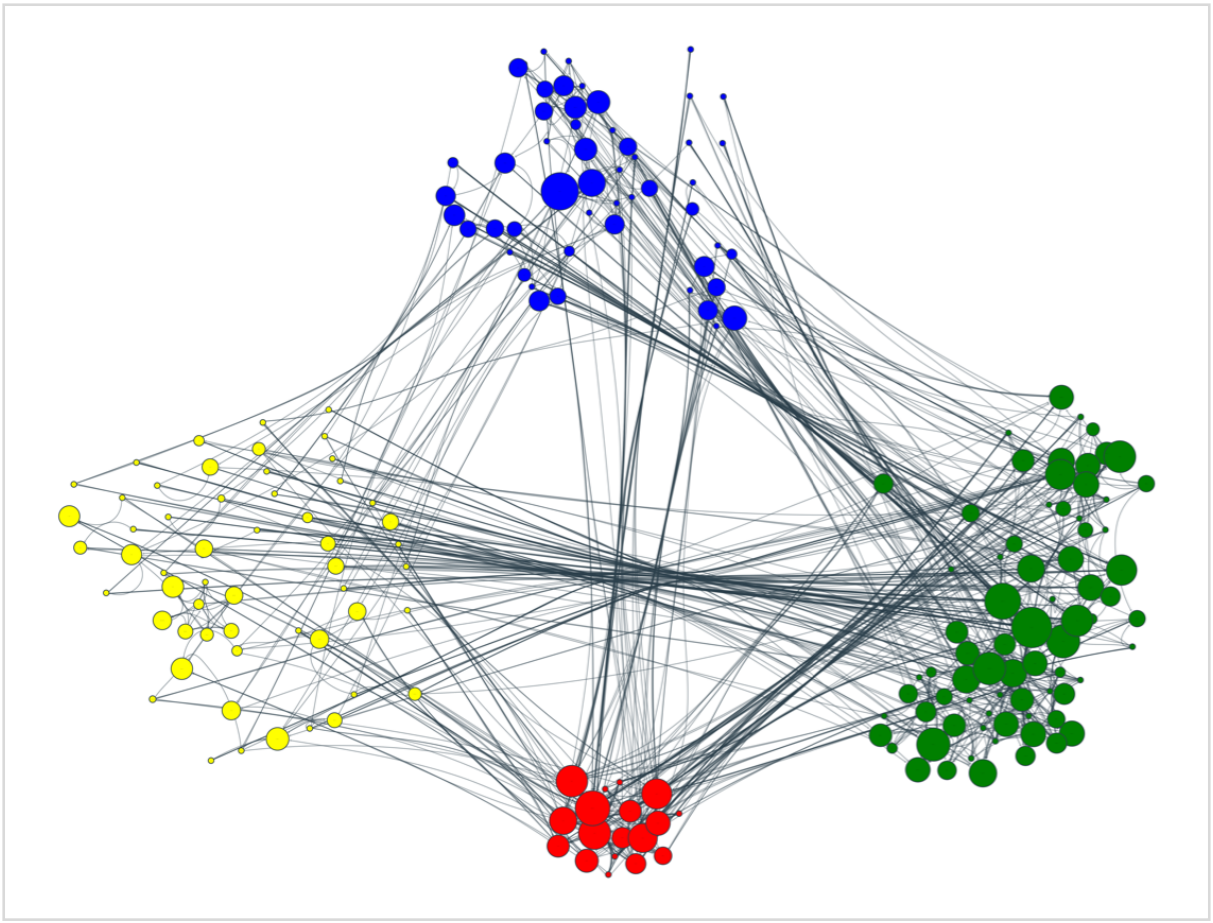


Current approach

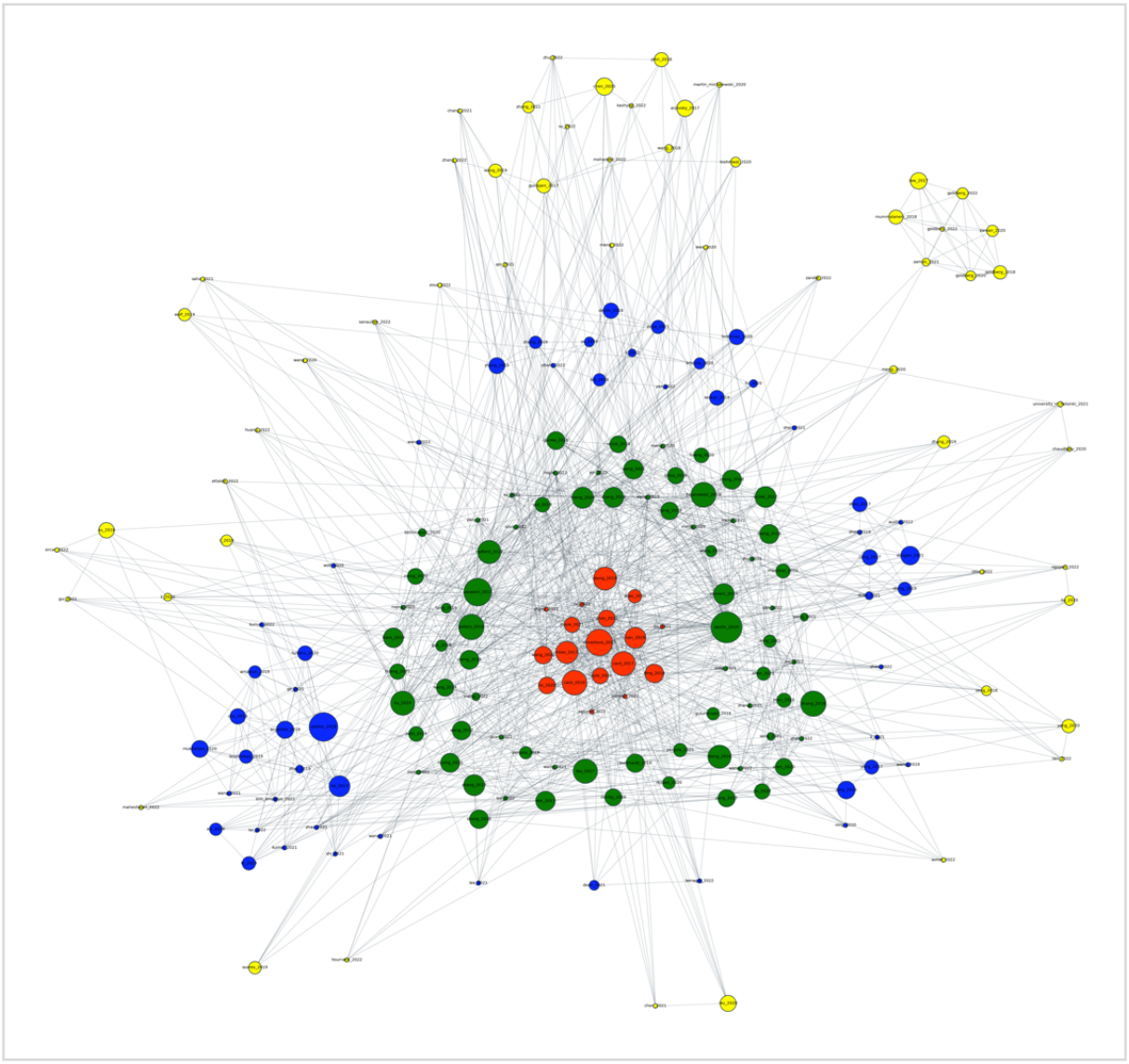
Treating different cores as communities and visualising them using the related netgraph community layout.

Analysing papers by k-core and by component.

Initial visualisation using the community layout:



Node positions adjusted using Gephi:



What to look at when analysing the distinct k-cores (and related sub-components)

- **Authors**: Are there repeating authors appearing within the core / component?
  - For example, in the 18 papers of the 9-core we have 5 authors appearing in 3 papers and 14 authors appearing in two
- **Domains**: What is the level of overlap between domains among the different papers?
  - Conceptually, once can infer the domain from:
    - Author defined keywords
    - Category assigned by the repository where the paper is hosted (e.g. Subjects in arXiv)
    - Named entities in title (and maybe abstracts)
- **Topic modeling**: Can we perform topic modeling (on titles and abstracts)

#Thesis/Weekly notes#