Open-Source Software Network

Seoul Lee

Department of Management and Organizations, Northwestern University | seoul.lee@kellogg.northwestern.edu

Northwestern | Kellogg

Questions

A key characteristic of the open-source software (OSS) ecosystem is that different OSS projects share contributors with one another. That is, OSS projects are linked through contributors who contribute to multiple projects. The degree of cohesion between different OSS projects can vary, with some being more cohesively connected than others. Can we identify some interesting patterns within this network of OSS projects?

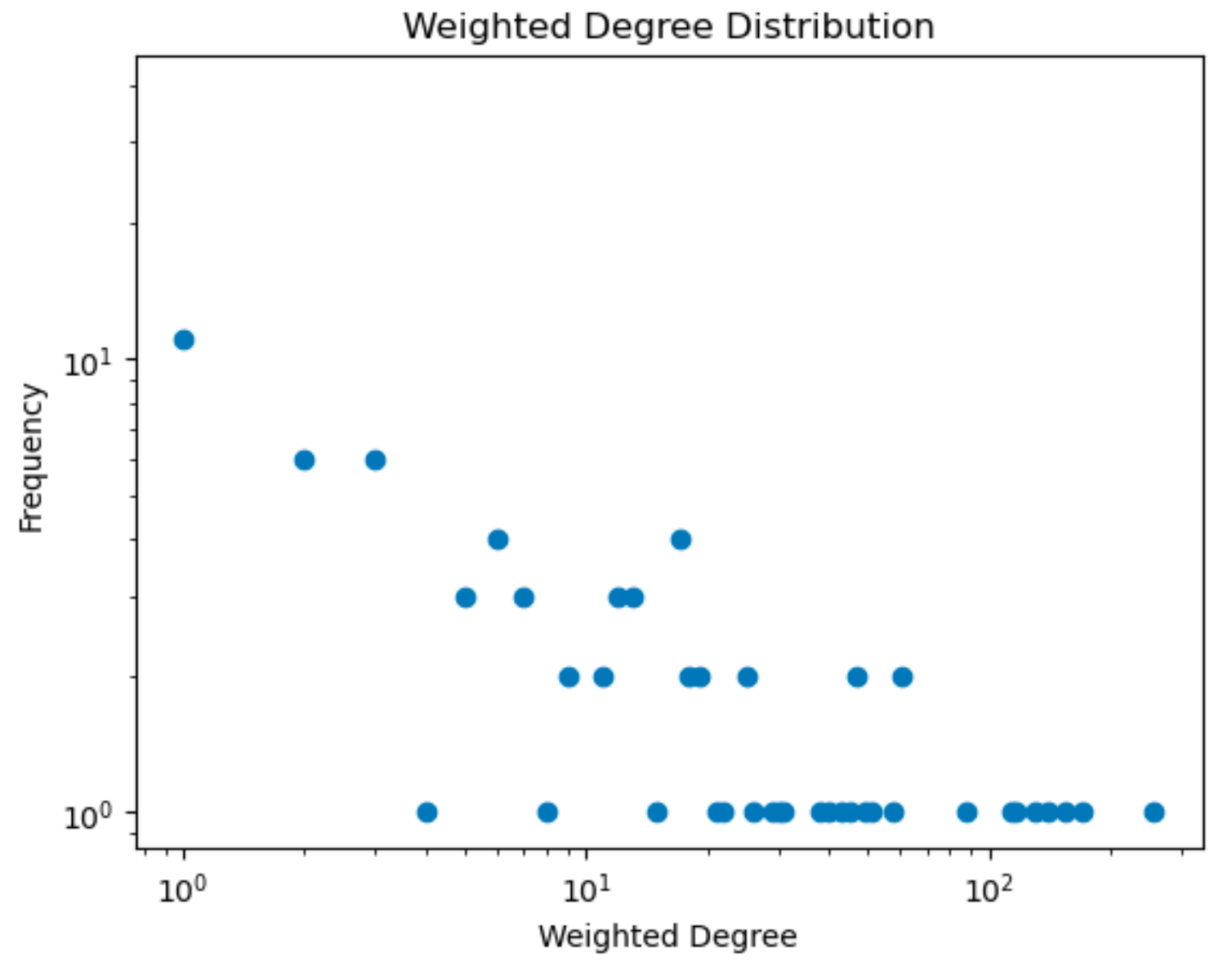
1) How are OSS projects clustered?
2) Do OSS projects within the same cluster or those sharing similar groups of contributors exhibit similar characteristics?
3) Do network properties predict OSS projects' governance structure?

How to construct the network

- **Nodes:** 120 open-source software projects associated with NumFOCUS, Inc. (Examples: Pandas, NumPy, NetworkX, Scikit-learn, Julia, Jupyter notebook)
- **Links (weighted):** the number of contributors who are part of the top 10% contributors in both projects
- **Data collection:**
  - GitHub Rest API (e.g., list of contributors, age, programming language, and stars)
  - Manual collection from each project's website (e.g., governance documents)
- **Programs:**
  - Python (NetworkX) – Data preprocessing and network construction
  - Gephi – Visualization and community detection

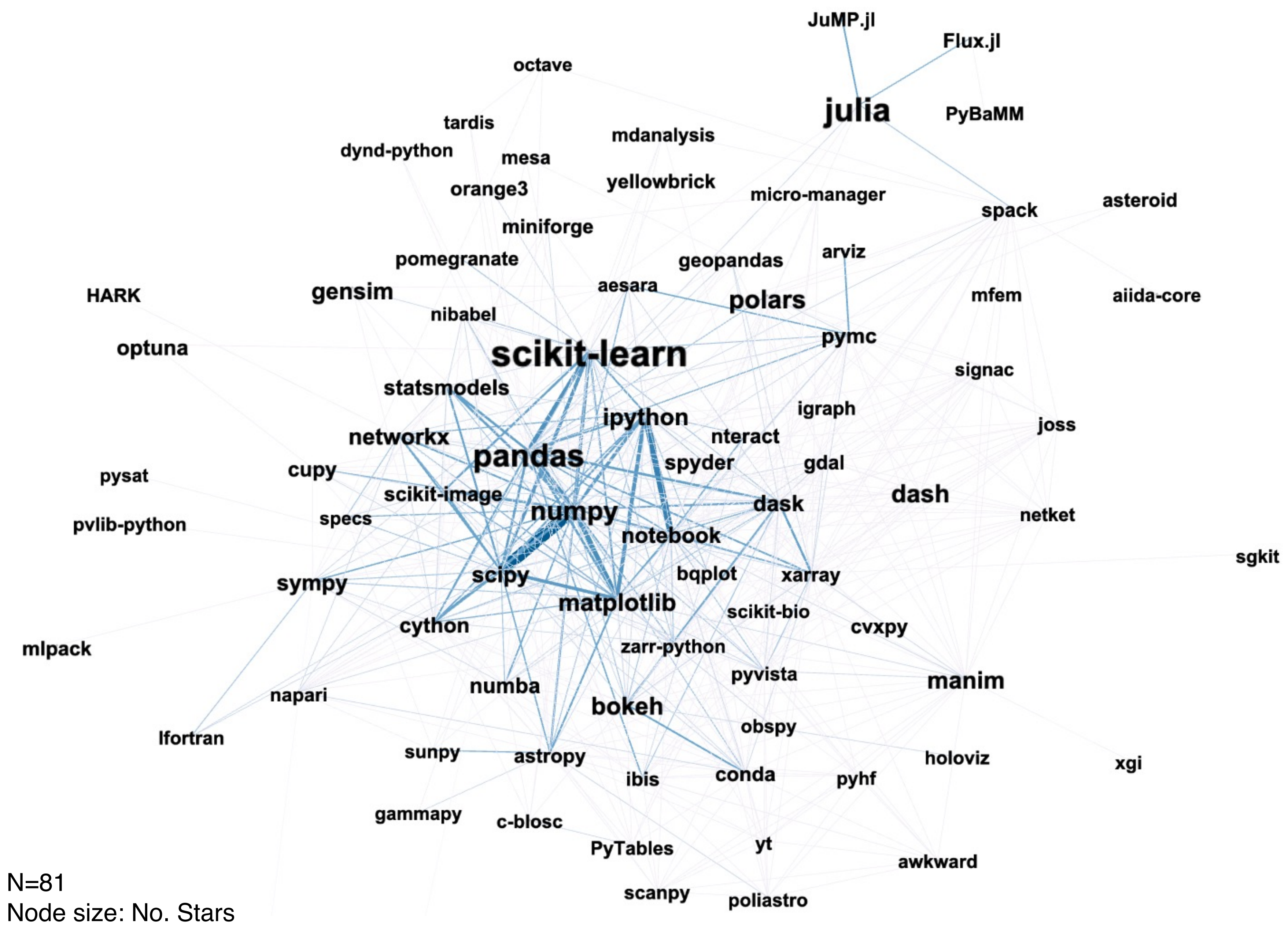
Network summary

Number of nodes	81*
Number of links	504
Average path length	2.23
Network Diameter	5
Average clustering coefficient	0.68
Minimum degree	1
Maximum degree	42
Average degree	12.44
Average weighted degree	28.44



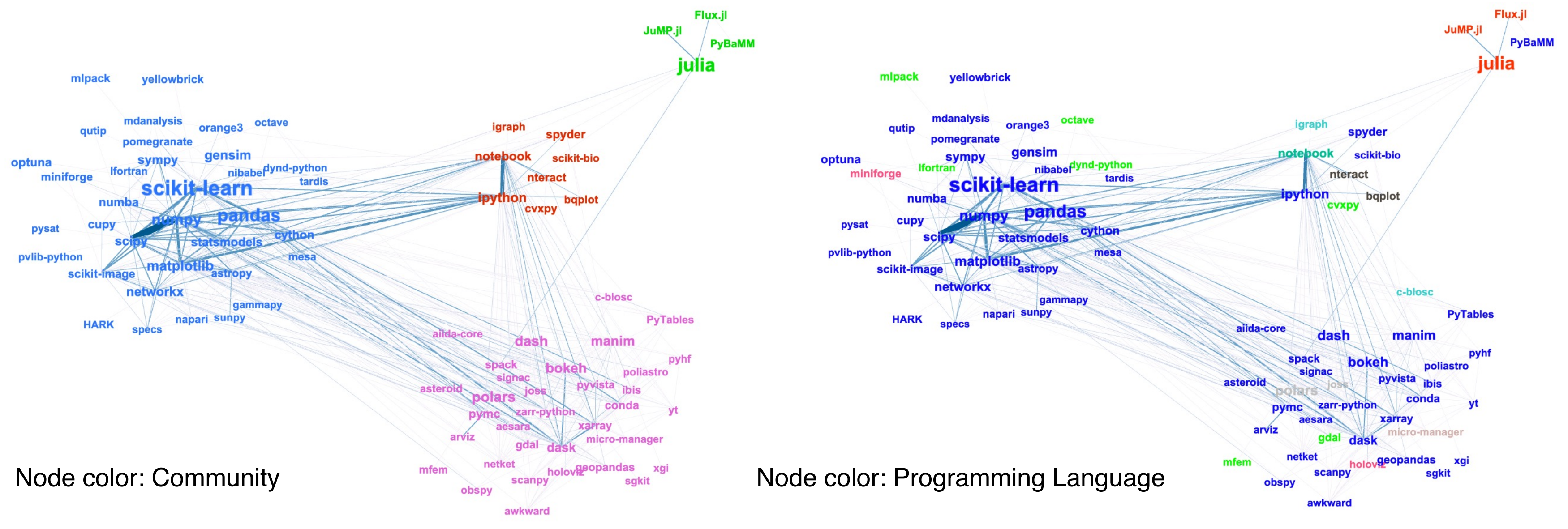
* Only nodes with at least one link are included.

Network visualization



N=81
Node size: No. Stars
Link width: Weight (No. Overlapping key contributors)

Community detection



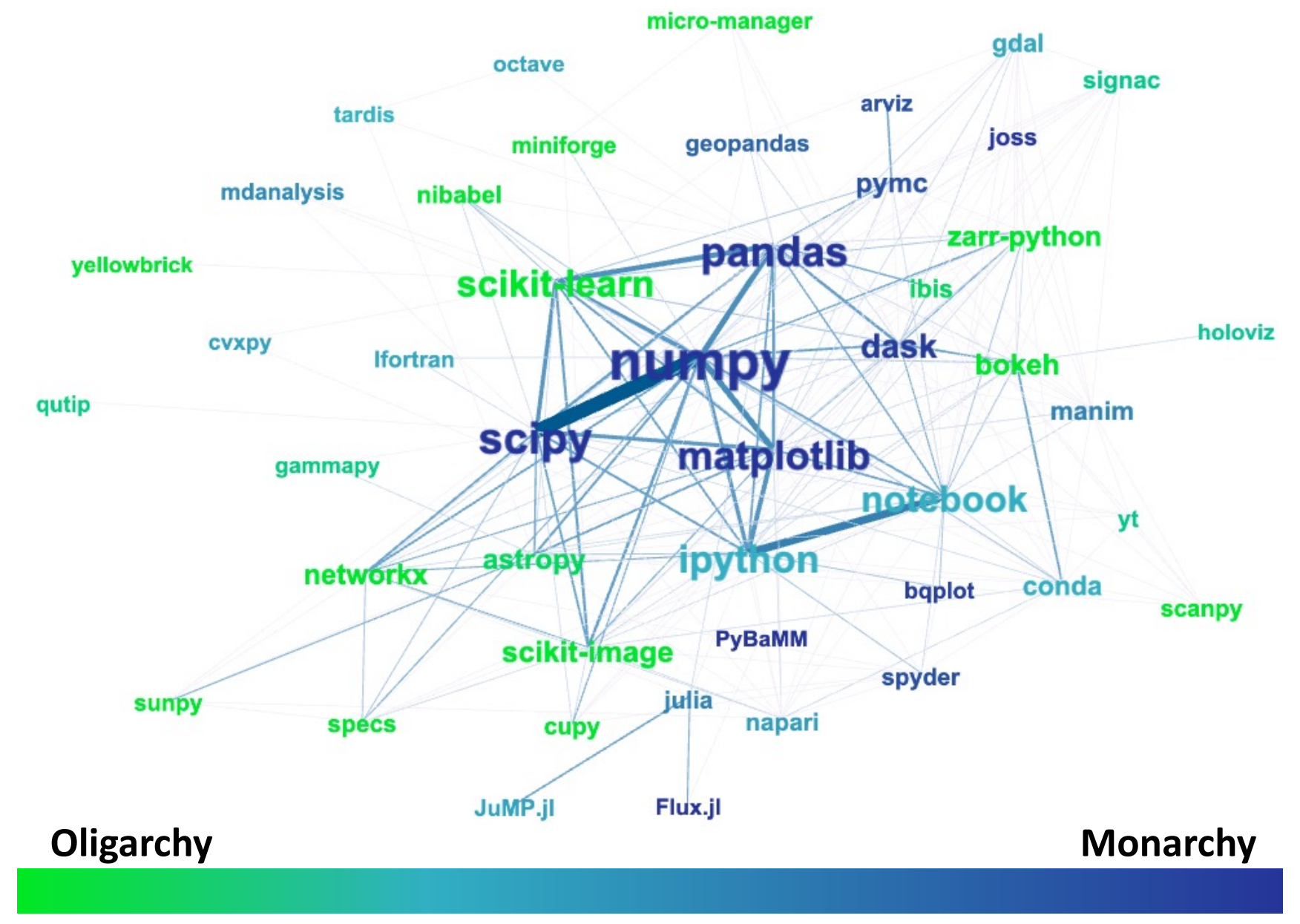
Node color: Community

Node color: Programming Language

Gephi's community detection algorithm identified four distinct communities within the dataset. The leftmost community appears to encompass general-purpose data science projects, while the bottom one is predominantly composed of specialized-purpose data science projects. On the right top, there is a cluster largely comprising Julia-based projects, and in the middle, there is a community consisting mainly of Jupyter-related projects.

Predicting governance structure

The purpose of this section is to see if network properties predict OSS projects' governance structure. To quantify the governance structure of each project, I performed LDA topic modeling on the governance documents of 63 projects, selected from the initial sample of 120 projects that possessed governance documents. Two distinct topics, namely "Oligarchy" and "Monarchy," were identified, and the predominant topic for each project is visually represented by the node color. Notably, nodes with a higher weighted degree appear to exhibit a higher monarchy score. While weighted degree is correlated with other variables such as the number of contributors, the number of stars, and the age of the project, the number of stars and project age did not show significant correlations with the monarchy score. Although the number of contributors displayed a moderate level of correlation, the significance level was lower than that of the weighted degree.



N=47
Node size: Weighted degree
Node color: Governance structure

Correlations between OSS project attributes

	Monarchy	(1)	(2)	(3)
(1) Weighted degree	**0.38****			
(2) No. Contributors	0.30*	0.74***		
(3) No. Stars	0.21	0.59***	0.84***	
(4) Age (days)	0.11	0.42**	0.45***	0.37***

* p<0.05, ** p<0.01, *** p<0.001

Discussion

- **How are OSS projects clustered?**
  ➔ 4 communities detected (general-purpose data science projects, Jupyter-related projects, Julia-based projects, and specialized-purpose data science projects)
- **Do OSS projects within the same cluster or those sharing similar groups of contributors exhibit similar characteristics?**
  ➔ Purpose, function, or language
- **Do network properties predict OSS projects' governance structure?**
  ➔ Weighted degree appears to be associated with the monarchy structure.
  ➔ The openness (i.e., connectivity or blurry boundaries) of communities may demand a more centralized governance structure.