UNIVERSITY OF AMSTERDAM

ASSIGNMENT 1 - EXPLORING REAL-WORLD NETWORKS

# Comparative Network Analysis of UvA Co-authorship: Structure, Models, and Influence

November 9, 2025

*Student:*
Sabrina Liu
14244861

*Lecturer:*
Mike Lees

*Course:*
Model-Based Decision Making

*Course code:*
5404MBDM6Y

One thing I've learnt in the past two weeks is that Mike and Vitor sure have done, are doing, and—one can only assume–will be doing a lot of work together. I began to ponder, "Hmm, what the network of UvA publications would look like".

This report presents an analysis of a real-world academic co-authorship network, constructed from publication data sourced from UvA's Digital Academic Repository (UvA-DARE). The dataset was obtained via a simple custom web scraper, which collected author lists and publication years of records (UBA, 2025).

The raw dataset initially contained 210,879 records. The cleaning process involved standardising the year formats, removing exact and near-duplicate entries, filtering out publications with fewer than a minimum degree (in this case, $k_i \geq 8$), and standardising author names to a `SurnameInitials` format. This whittled the clean dataset to $9,416$ publications, spanning from 1977 through to 2026. From this, a pairwise authorship network was built where nodes represent individual authors, and an undirected edge connects two authors if they have co-authored at least one paper. Edge weight represents the number of collaborative publications.

The final analysed network consists of $57,343$ authors (nodes) and $591,464$ co-author connections (edges). This was the trade-off I was willing to compromise at, in spite of what you may have predicted the very challenging upcoming computational costs.

## 1 Descriptive Statistics of the Real Network

The network has a massive Largest Connected Component (LCC) containing $55,700$ nodes, which encompasses 97.1% of all authors in the filtered dataset. This is indicative of a highly interconnected community where most researchers are connected through some chain of collaboration. I would speculate the remaining components are likely isolated research groups, or perhaps interdisciplinary collaborations not yet integrated into the main network.

The network density is extremely low (0.000360). While the network is large, any two randomly selected authors are very unlikely to have collaborated directly.

The average clustering coefficient sits really high at 0.8741, a feature of the small-world structure. This means the collaborators of an author are highly likely to collaborate with each other.

The average shortest path length within the LCC is approximately 4.83. This "small-world"

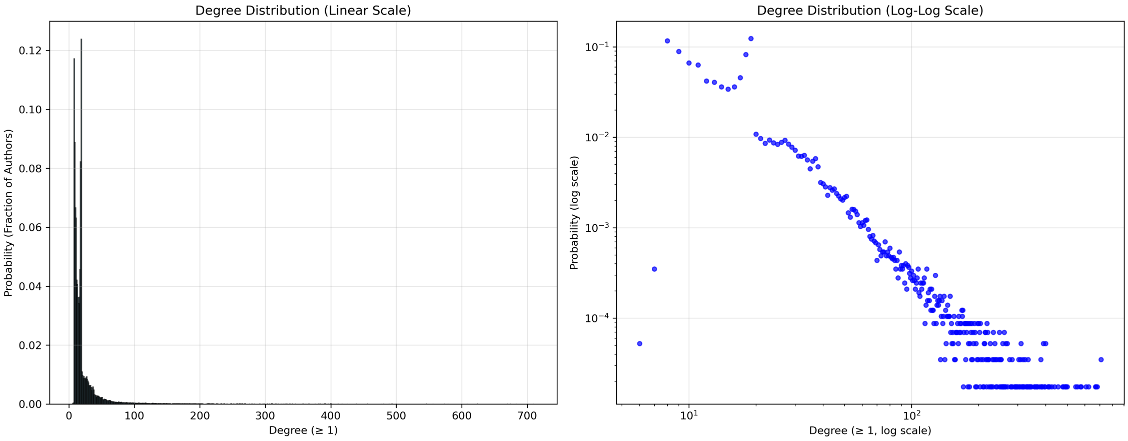

Figure 1: Degree distribution of the real network

property suggests that despite the scattered nature of the network, information can travel as efficiently across the entire UvA research community in just a few steps.

Academic collaboration networks often show disassortative mixing by degree, meaning highly connected hubs tend to connect with less-connected nodes (e.g., senior professors collaborating with PhD students).

## 2 Comparison with Model Networks

To contextualise the structure of the UvA network, it was compared against three canonical network models—Erdős-Rényi (ER, random), Watts-Strogatz (WS, small-world), and Barabási-Albert (BA, scale-free)—generated with a similar number of nodes ($N$) and average degree ($< k >$) as the LCC.

Table 1: Comparison of network properties between real and model networks (on LCC).

| model | $N$ | $E$ | avg_degree | density | avg_clustering | avg_path_len |
|---|---|---|---|---|---|---|
| Real (UvA) | 55,700 | 581,153 | 20.867253 | 0.000375 | 0.327535 | 4.825467 |
| Erdős-Rényi | 55,700 | 580,321 | 20.837379 | 0.000374 | 0.000368 | 3.884734 |
| Watts–Strogatz | 55,700 | 557,000 | 20.000000 | 0.000359 | 0.515286 | 5.241847 |
| Barabási–Albert | 55,700 | 556,900 | 19.996409 | 0.000359 | 0.002430 | 3.516671 |

The Erdős-Rényi (ER) random model matches the real network's density and average degree, however fell drastically short at the average clustering. Its path length is shorter, but this could be seen as one of the consequences of its random structure (Erdös & Rényi, 2011). The degree distribution follows the Poisson distribution, unfortunately lacking the hubs present in the real network.

The Watts-Strogatz (WS) small-world model, by design, captures both high clustering and short path lengths (Watts & Strogatz, 1998). The WS model we generated has a higher clustering and a longer path length than the real network. Its homogeneous degree distribution caused it to fall short of a real network's highly heterogeneous distribution with prominent hubs.

The Barabási-Albert (BA) scale-free model heavy-tailed degree distribution matches that of the real network, as we can visually see in Fig. 2. Its short average path length is also consistent with the small-world phenomenon. However, it fails to generate the high level of clustering seen in the real model. The "rich-get-richer" mechanism generates hubs but not the local triangles needed that define collaborate teams (Barabási & Albert, 1999).

Therefore, the UvA co-authorship network is a hybrid structure, combining the scale-free property of a BA network (presence of hubs) with the high clustering of a WS network. No single classic model fully captures its topology. It is a scale-free network with strong small-world

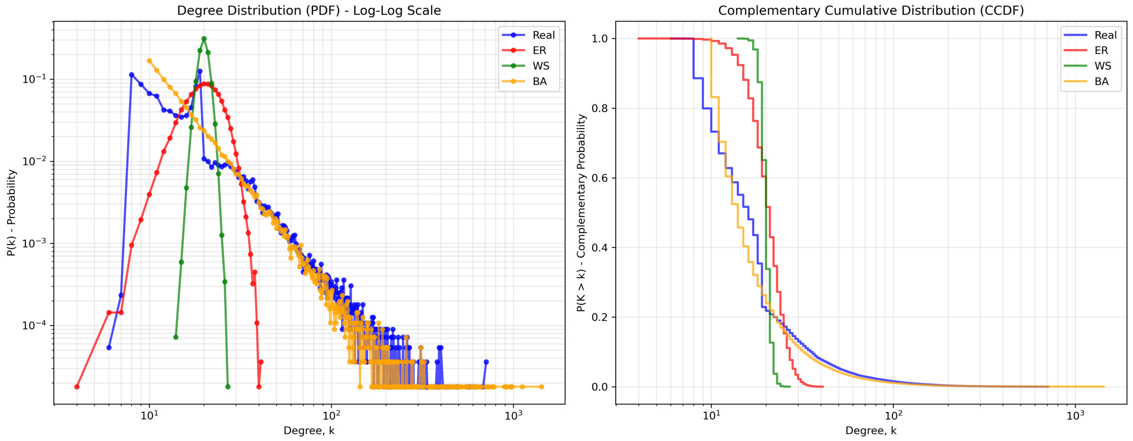

Figure 2: Comparison of the degree distribution of real vs model networks.

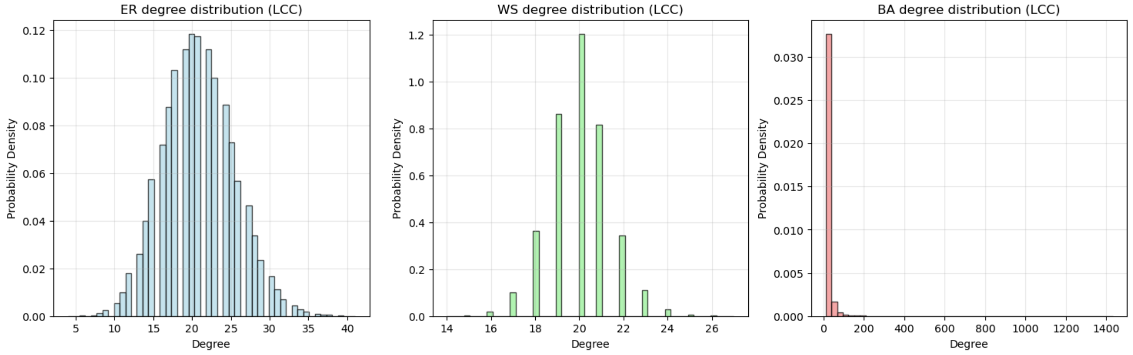

Figure 3: Erdős-Rényi, Watts-Strogatz & Barabási-Albert Degree Distributions

characteristics.

## 3 Centrality Analysis and Real-World Interpretation

Centrality measures were calculated to identify the most well-connected and thereby influencial authors within the network.

Looking at the degree centrality, we know that `HesselsJWT`, `WijersRAMJ` and `KouveliotouC` are the network's super collaborators. These individuals have an exceptionally high number of direct co-authors, possibly as senior researchers, or perhaps the 'social-glue' figure of multiple projects. Their high degree suggests a central role in the flow of ideas and coordination of large-scale research.

The Betweeness Centrality identifies the broker authors, a bridge to otherwise disconnected parts of the network. They likely belong to interdisciplinary fields, or collaborate with dintinct, seperate communities. Their roles are critical for network connectivity; if removed, the network would fracture into more isolated components.

Many authors scored the maximum value of 1.0 in the closeness centrality. In theory, this means they can quickly reach all the other authors in the network. They can be thought of as "in the thick of things.", efficient for gathering and disseminating information.

The Eigenvector Centrality tells us about the number of connections but also the quality of them. It's remnant of a schoolyard popularity contest, where who you're friends with boosts your score. These authors are not only well-connected themselves, but embedded within the core of the network's most influential collaborative circles.

As these measures go, it's about what they tell you based on what you need. The diver-

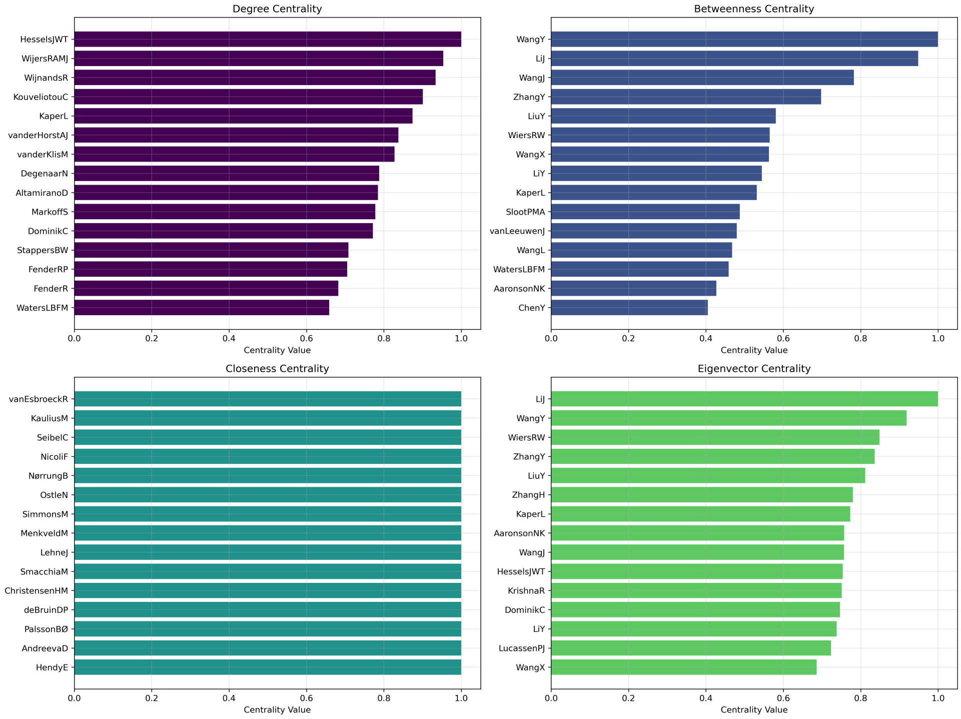

Figure 4: Centrality measures of the real network.

gence in these lists also show how being prolific in one area is not the same as the same in another. Although a truly influential academic might appear high on several of these lists.

## 4 Discussion and Conclusion

The network is sparse, highly clustered, has short overall path lengths, and possesses a heavily right-skewed degree distribution featuring prominent hubs. It is best described as a hybrid small-world and scale-free network; a product of both cumulative advantage and localised, team-based project work.

The BA structural property points to the preferential attachment mechanism common in academia: established, well-connected researchers are more visible and thus more likely to attract new collaborators. Large, multi-institutional projects (e.g., in astronomy or physics), naturally create hubs with an extremely high number of co-authors.

The WS property is driven the work done in labs, departments, research groups, where everyone collaborates with everyone else, creating dense clusters of interconnected nodes. A small number of which act as “connectors” to other tightly-knit clusters, creating the high clustering coefficient short average path length effect.

In terms of its implications for the dynamics of the UvA research community, the network is robust to random failures (like a death, I guess), but vulnerable to targeted attacks on its major hubs. Should a ‘super-collaborator’ depart, some of the network, perhaps students writing their thesis preferring a supervisor, would feel the loss.

The small-world property facilitates the rapid dissemination of ideas, methods and resources across disparate fields within the university, and interdisciplinary research will only make these bonds stronger.

The hub-and-spoke structure can reinforce existing inequalities. A small core of hyper-connected authors enjoy disproportionate influence and access to diverse ideas, while the long tail of those on the periphery may find it harder to gain visibility. Perhaps this boosts people aiming to achieve full professorship in the Dutch system, but makes the goal even more out of reach for others.

# References

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks [Publisher: American Association for the Advancement of Science]. *Science*, *286*(5439), 509–512. https://doi.org/10.1126/science.286.5439.509

Erdös, P., & Rényi, A. (2011, December 31). On the evolution of random graphs. In *The structure and dynamics of networks* (pp. 38–82). Princeton University Press. https://doi.org/10.1515/9781400841356.38

UBA, D. P. C. (2025). *Digital academic repository - university of amsterdam*. Retrieved November 9, 2025, from https://dare.uva.nl/

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks [Publisher: Nature Publishing Group]. *Nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918

# 5 Limitations

As I don't know much about the world of academia, all my inferences are speculations. I would love an interpretation from someone on the inside :-).

There are many flaws in my code, many of which were done in the interest of computational expense (such as using BFS in the `base_stats` function when calculating the centrality measures). Despite my efforts, there still exist duplicates in the dataset, as some entries were in both Dutch and English.

I have used LLMs to assist in the code, as well as taken much from the course material and Mike's git repository.

The full dataset and all my code is available in my repository https://github.com/sabrina-liu/model-based-decision-making.

All plots are normalised.

It was at this point I realised I built an Erdos calculator but for UvA.