Universidad Politécnica de Yucatán

Gabriel Islas Peraza

Data 8ºB

Social Network Analysis

Project 1

Teacher: Didier Omar Gamboa Angulo

Abstract

We delve into the structural characteristics and implications of a co-authorship network of scientists in network theory, compiled by Mark Newman in 2006. The network's average shortest path length suggests moderate connectivity, with implications for the efficiency of information flow. The clustering coefficient indicates tightly-knit research groups and a decently collaborative research environment. Barabasi acts as a really important researcher, not only with his connections but the influence past them, while Newman's general network research acts more as the core of network science. Community division shows consistency in clear divisions between them, but a variety of reasons for communities to form, from regions of research to specialized fields of study.

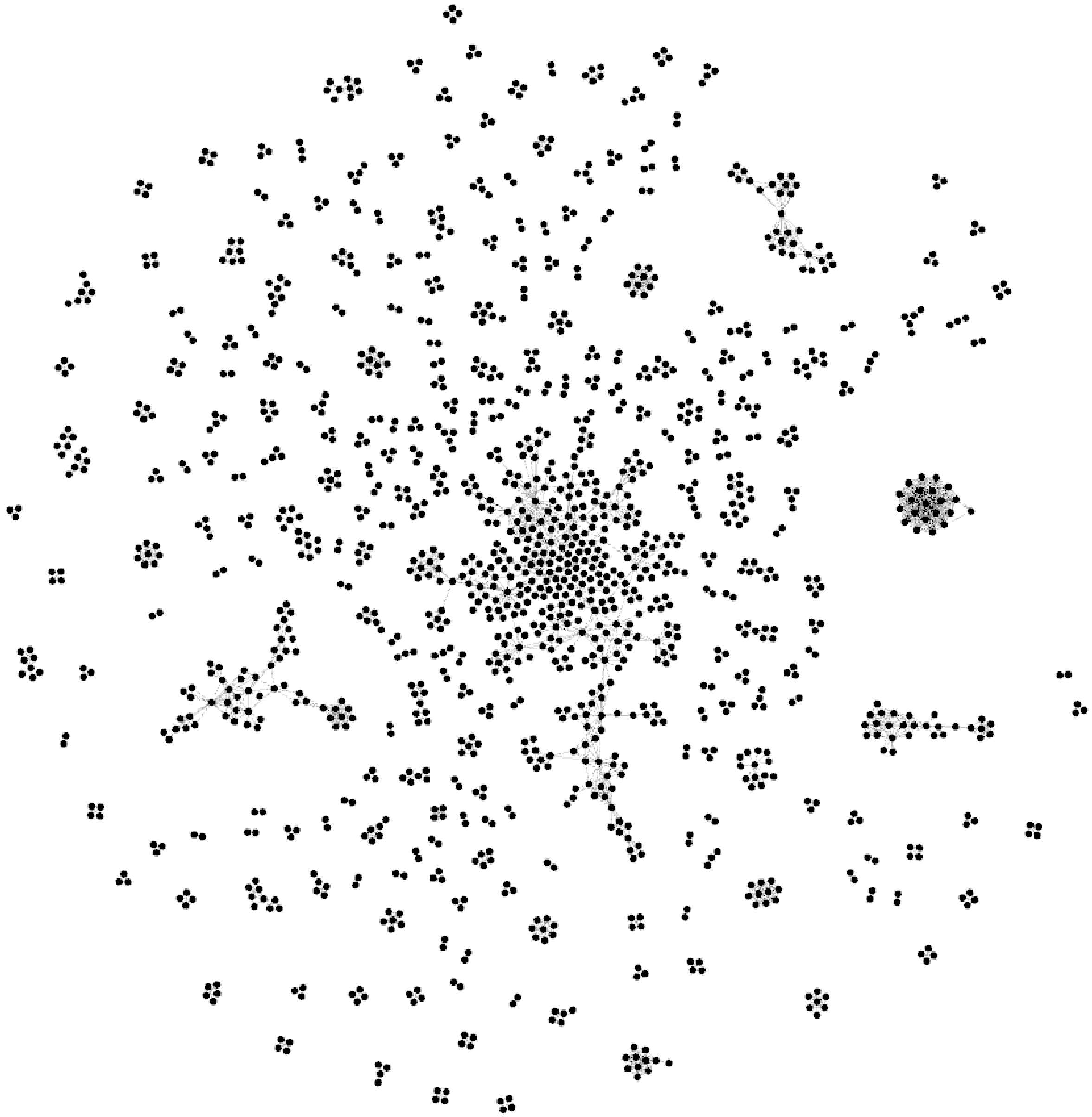
Introduction

This report examines a network obtained through the KONECT project website, which compiles network datasets together with some analysis, making them available for everyone. The network chosen for this report is a network of co-authorships among scientists specializing in network theory and experimentation. Through the data source link in the page, it can be found that this network was compiled by Mark Newman, a physicist known for his contributions to the study of networked systems. He compiled this network for his article "Finding community structure in networks using the eigenvectors of matrices" published in the "Physical Review" journal back in 2006, and it was used to provide multiple examples throughout the article.

The dataset was created using two prominent reviews on network research, one authored by Newman himself, and the other one by Stefano Bocaletti, another name known in network science. Nodes in the network represent individual authors cited in these reviews, and edges connect pairs of authors who have co-authored in at least one paper together from the cited ones.

Co-authorship networks for such an interdisciplinary field reveal how researchers in different fields collaborate and integrate different perspectives into their work, and it also shows important channels of knowledge exchange; these networks can also allow to identify leading researchers and the impact of their contributions in the network science field, as well as their productivity; in more recent examples it could also aid in identifying emerging areas of interest which could help in funding or finding important collaborators for new fields, allowing forecasting and manipulation of rising methodologies, collaboration scenarios, community platforms, and idea propagation.

This network is undirected, given that a co-authorship relationship just goes for both authors; it is not planar and it is unweighted. The interesting classification that was considered the most important, since it defines how analysis is done, is that the network is not connected, and it is in fact formed by 268 components. While networkx was mostly used to do analysis and most visualizations, an overall visualization of the entire network was more approachable through Gephi using the ForceAtlas 2 layout. To see the visualization in your browser, click here.



In this image you can appreciate that there are two components way larger than the other ones, the one in the center, and the one to the left of it. While some metrics can be appreciated for the entire network, most of the analysis will focus on the largest component, as it is the most relevant one given its size, and it is the only of which I could actually find data on who the authors are in this component subgraph, as the actual dataset that is available for download does not have metadata per node, and just identifies them with a simple ID number. The Jupyter Notebook included together with this document will include metrics for the second largest component as well.

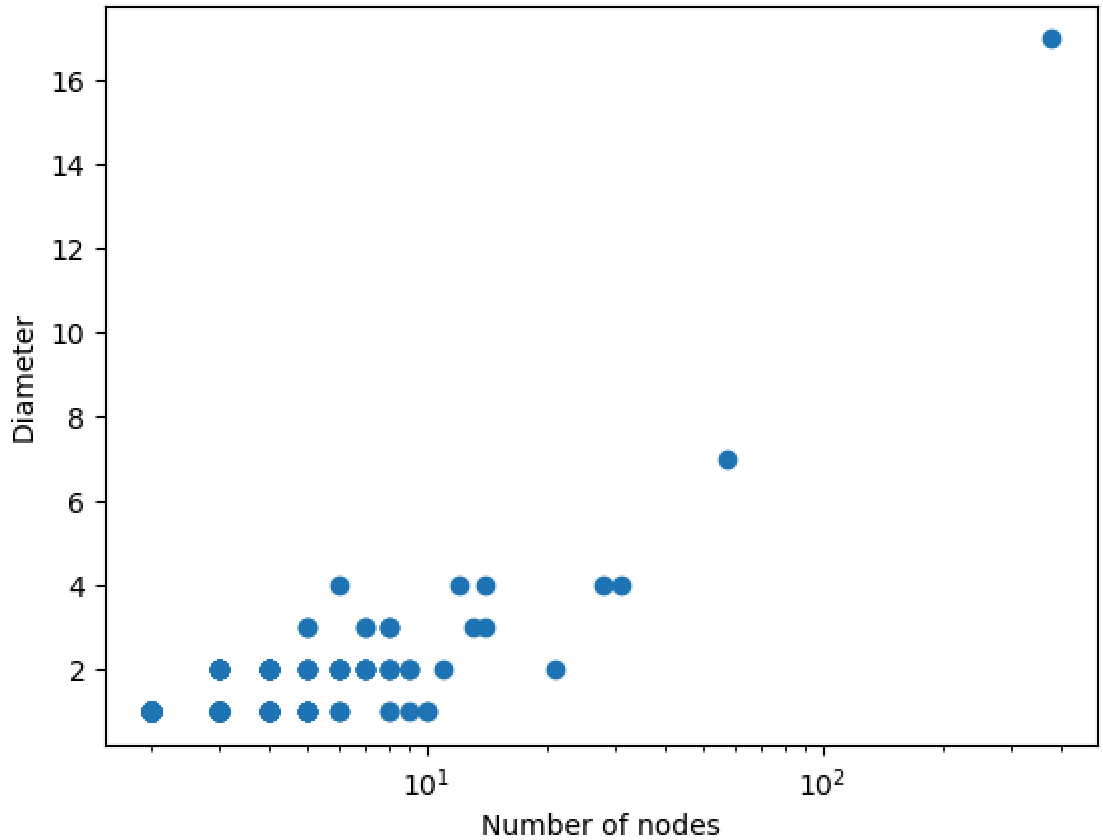
Network characteristics

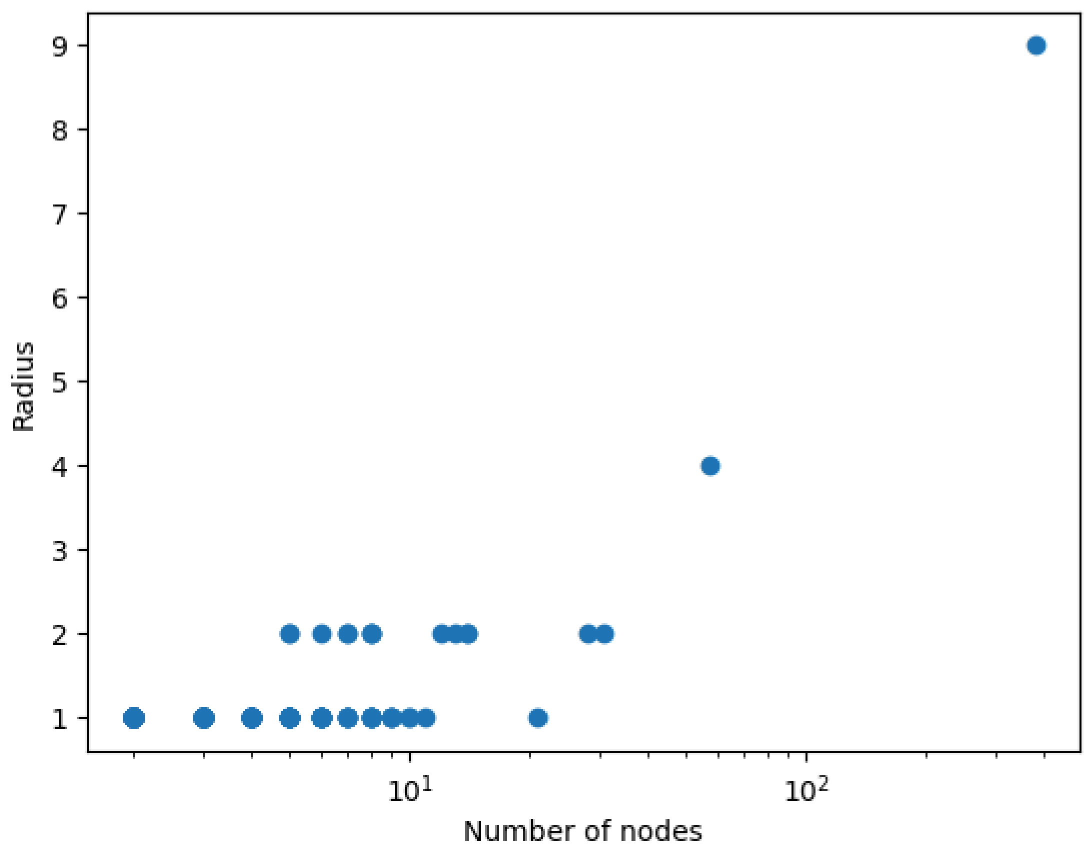
The network is composed of 1461 author nodes and 2742 co-authorship links, where the largest component has 379 author nodes and 914 links and the second largest has 57 nodes and 149 links, which is a lot less than the first one. From there, node amounts rapidly decrease, to the point that most common node amounts in components range from 2 to 3 nodes. While it is obvious that for components made of 2 nodes, only 1 link is possible, one interesting observation is that for components of 3 nodes and 4 nodes, almost all of them are complete, meaning each node is connected to every other vertex. This makes sense; most components of this size probably represent one paper with a specific focus, and if it's multiple papers, then they probably all are on the same subfield, and as component size grows, the more broad the subjects of collaboration become, meaning connections are less related with each other.

The average shortest path length for the first component is 6.04, which seems to be about average of other fields, but on the higher side of it, meaning it is not high, but it is also not low, implying a moderate level of connectivity in this component, implying there are some acting as intermediaries between them and the presence of various communities in this network.

The clustering coefficient of 0.693 for the entire network shows that in general, while the network is not fully connected, most of its components are pretty well connected individually, and showing there might be lots of community structures that frequently collaborate in the larger components.

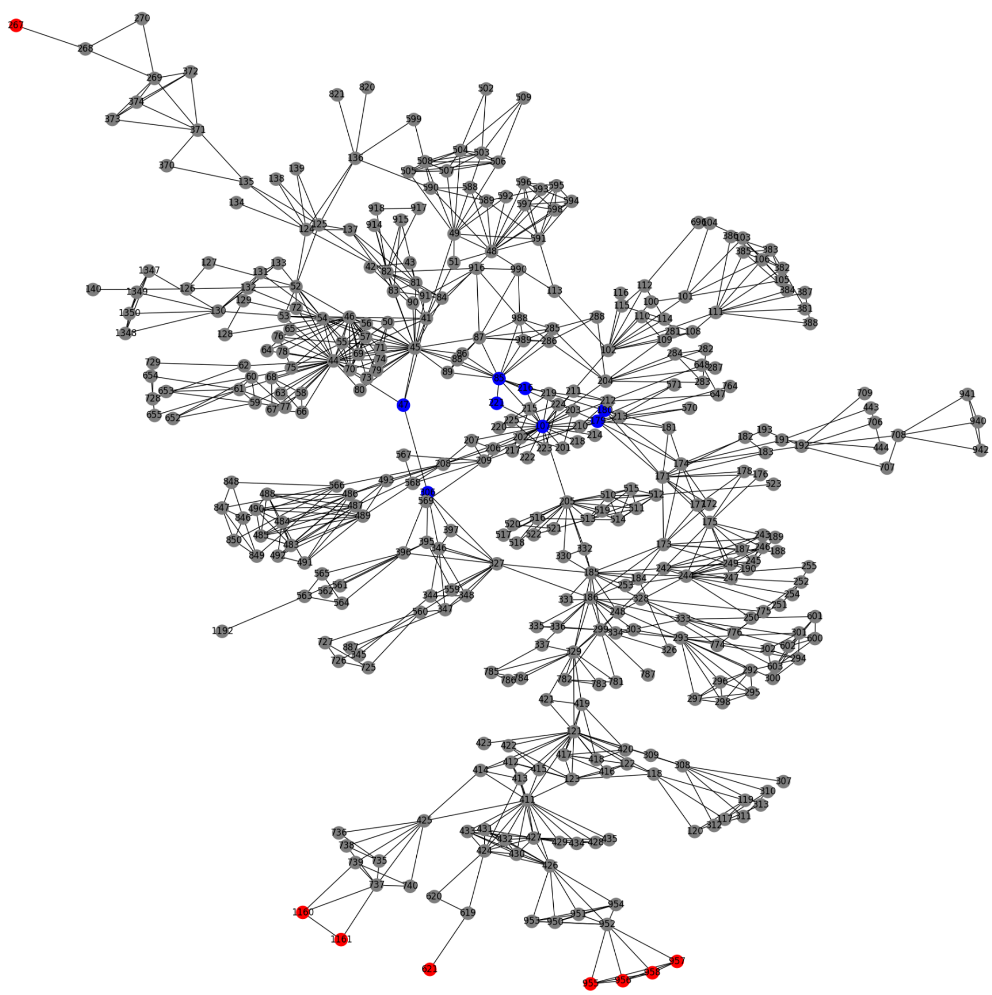
The following are scatter plots where the x-axis represents the number of nodes, and the y-axis represents the diameter and the radius of the network respectively.


Diameter
Number of nodes
16
14
12
10
8
6
4
2
10^1
10^2


Radius
Number of nodes
9
8
7
6
5
4
3
2
1
10^1
10^2

As expected, the two largest components are pretty much outliers compared to every other component in regards to eccentricity values, which makes sense given their size. As more authors are involved in a co-authorship component, the more broad subjects will be on the extremes and even on the smallest values of eccentricity.

The following image is a visualization of the periphery and center of the main component of the network. It can be seen in your browser [here.](#)

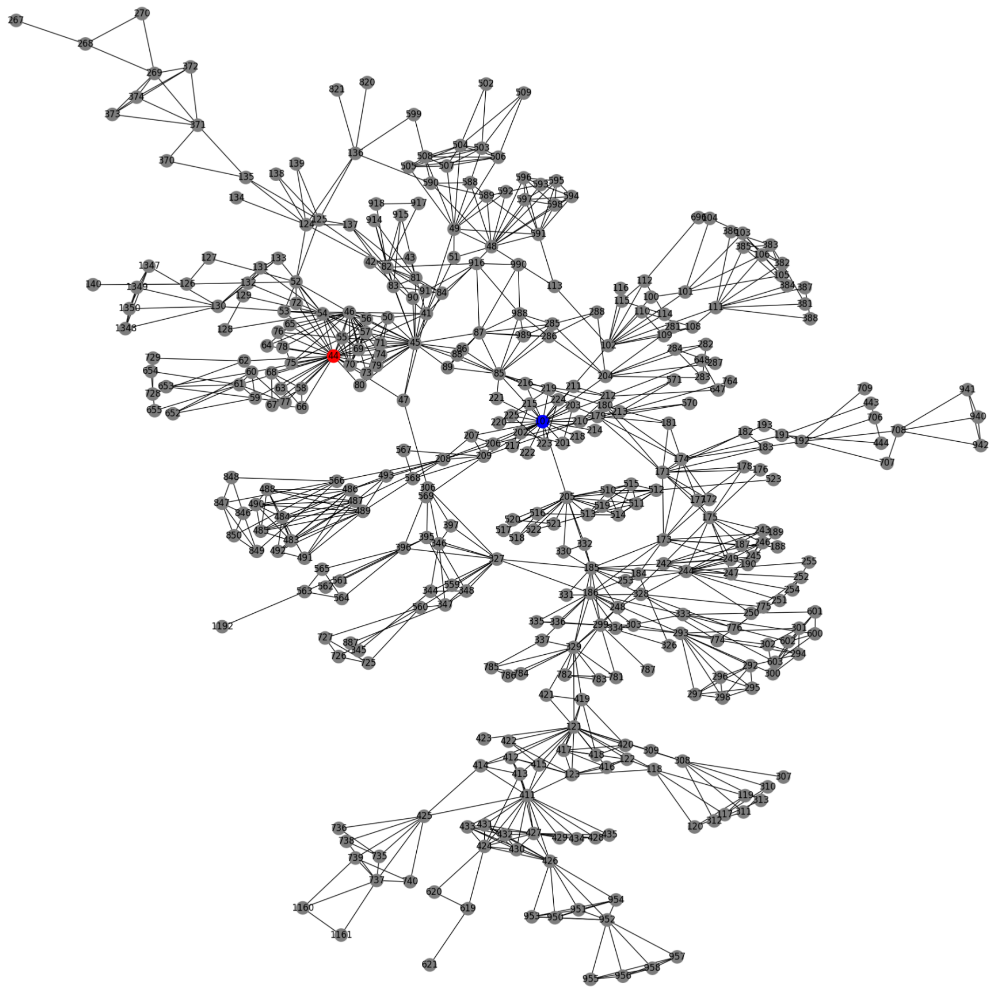


The periphery of the network represents authors that had few collaborations, either by limited connections or due to specialized research in the field. It could also be due to an early career or geographical isolation. Using the [author data available for the first component](#), it is not difficult to identify the authors. For example, the one in the top left is Marco Baiesi, whose only connection to the network is through Maya Paczuski; a quick search for both names leads us to the article “Scale-free networks of earthquakes and aftershocks”, which does seem to be an specialized application of network science.

Authors in the center have pivotal roles; in a sense, they have important connections, they probably are influential in their fields, but at the same time are also important as interdisciplinary bridges. Newman himself is part of the center for example, as well as Ginestra Bianconi, who has known work in network theory, multilayer and higher-order networks, but mostly for the Bianconi-Barabasi model, creating a bridge with Albert-Laszlo Barabasi, a very important author, with many other authors.

Centrality measures

The following image represents four centrality measures, even if just two nodes are colored. The red node is Barabasi, who has the maximum degree centrality and the maximum eigenvector centrality, while the blue node is Newman, who has the maximum closeness centrality and the maximum betweenness centrality.

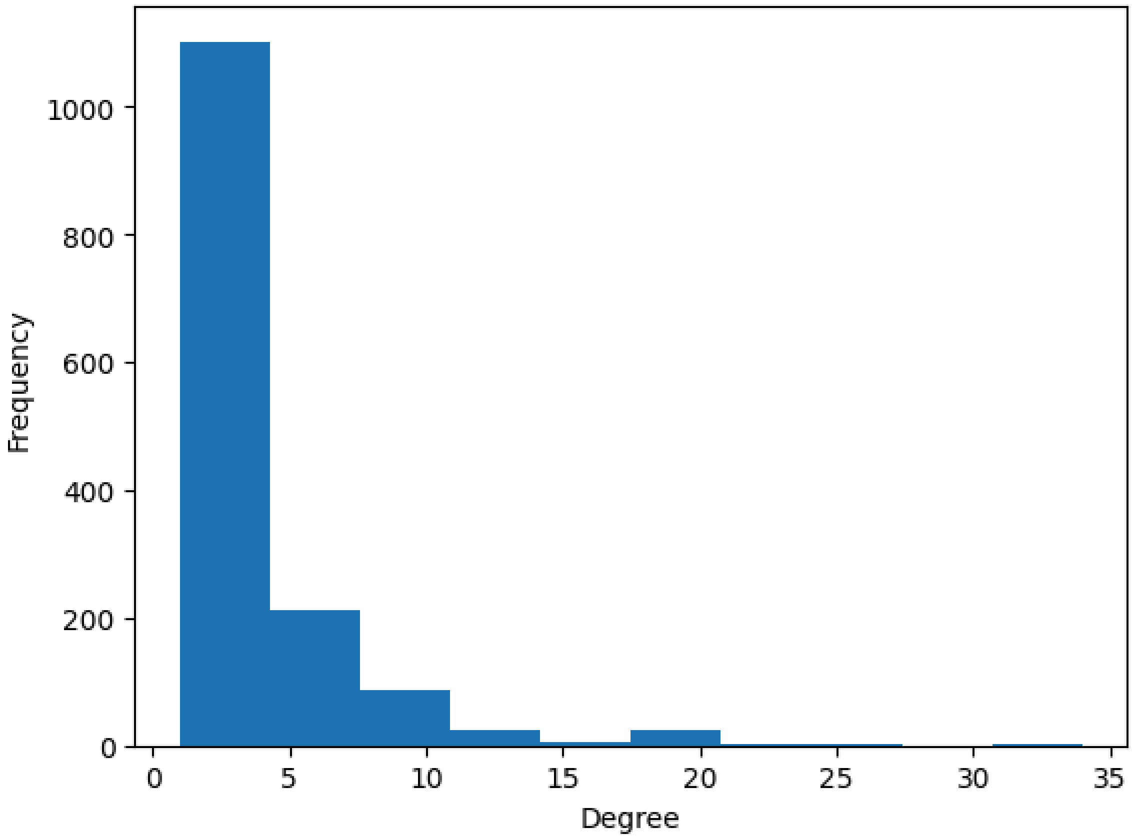


The people involved in these metrics are not surprising given how influential their work is in network science. Barabasi has the highest degree centrality probably due to how influential scale-free networks and the Barabasi-Albert model are, and given the breakthrough of such a model, not only is he highly connected, but he is highly influential even further than just his co-authorships, which is represented by the eigenvector centrality; however, that does not make him as central as Newman.

While his work is clearly influential, it can be thought to have less general applications than Newman's work, which is more generally applied, like the Girvan-Newman algorithm for community detection, his general research into communities, and a lot of random network theory, meaning he is probably the most important bridge in the network.

Degree distribution

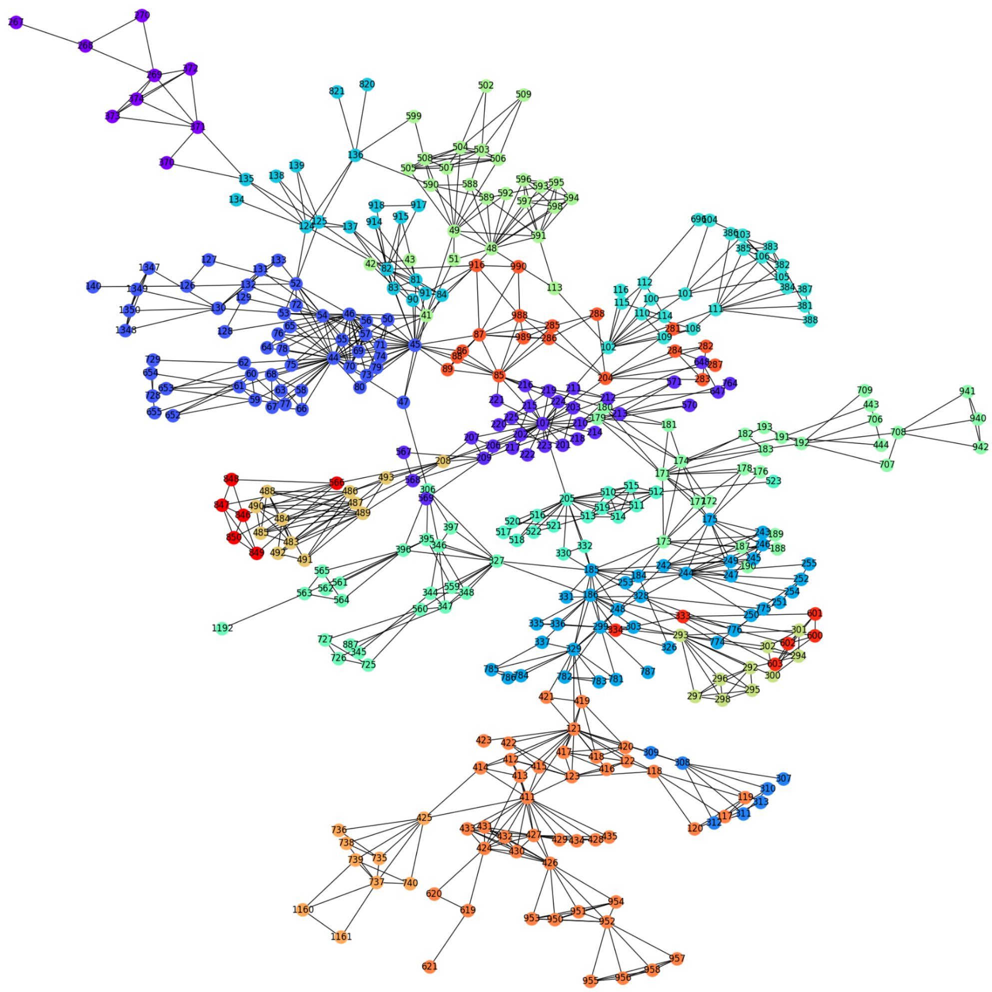
The following is a plot of the degree distribution of this network



The distribution follows a Power-Law distribution, meaning there is a very small number of authors that are hubs, and the majority have just a few collaborations; when new researchers keep working in the field, the more effort they put in can make it more likely that they will collaborate with already established and well-connected authors, reinforcing this distribution.

Community detection

The following image showcases the results of finding the partition with the best modularity using the Girvan-Newman algorithm. The result was a division of 18 communities with a modularity of 0.844



At first glance, it can be noticed that the blue community on the left is the Barabasi community, while the purple community in the center is the Newman community. Most of these communities have common characteristics between them, meaning that the division was successful. The purple community on the top left is very varied in subjects of research; the previous Baiesi work has already been mentioned, but there is also a lot of very specific research into small-world networks and vortex avalanches; the article that seems to unite most of the people in this community is one that defines the concept of Gradient Networks. The cyan community right below it is mostly about growing networks, which comes with an application to proteins in one of the articles. The mint green community next to it is a very curious case; they have a lot of variety in subjects but with some research there is something clearly in common: it is composed of mostly Italian researchers from Sicily.

Conclusion

Co-authorship networks provide insight on the structure of research, and the influence of different authors, fields, and even institutes as bridges of knowledge between different communities of research.

This of course applies to the field of network science as well, which up to 2006 had already gone through very important innovations creating a very large component subgraph of varied research.

Periphery authors represent specialized fields of research or very important newcomers of the field, and central authors represent important bridges and central authors of important subjects in the network. If I had to guess, the growth and extension of the network science field has probably only helped in the increase of periphery authors, and rather than new central authors, maybe there has been shifts on who is a central author as the focus of researchers moves into new subjects.

In the main component, Barabasi and Newman are the giants of network science; Barabasi with his contribution to the modeling of social networks considering all their expected characteristics sees the most co-authorship links and the most influence, given how many real life applications such a breakthrough has; Newman, with more general research on the inner workings of common network structures and community detection in general acts more as the main connection and center in all of the network science field. Given the exponential growth in the interest of social networks, I would not be surprised if a modern version of this network sees Barabasi closer to stealing all of the centrality measurements.

The degree distribution of the network follows the Power-Law distribution, which goes with the mechanics of cumulative advantage found in social interactions, where there will be a preference of attachment for already well-connected nodes, which makes sense in research as well, given that most newcomer authors that put effort into their work would likely be honored to work with influential authors on influential subjects.

Community detection shows many types of divisions, either by subjects, institutes, author influence, and sometimes, simply defined by one article with many collaborators on it, showing how varied the field truly is; one takeaway in this work is that community analysis is really hard to do without node metadata as it takes a long time to discern what each characteristic of the communities are. In general, doing network analysis without any extra data about the nodes other than their purpose is difficult. Ideally, co-authorship networks should include their author data and maybe even article data in their links for an easier analysis, which would have made for an easier analysis, but the fact that conclusions were drawn by just looking at some of the data partially shows how powerful this type of network is for analysis.

References

Network science. (2024). Konect.cc. http://konect.cc/networks/dimacs10-netscience/

10th DIMACS Implementation Challenge. (2024). Gatech.edu. https://sites.cc.gatech.edu/dimacs10/archive/clustering.shtml

Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. Physical Review E, 74(3). doi:10.1103/physreve.74.03610

Community Centrality. (2024). Umich.edu. https://websites.umich.edu/~mejn/centrality/

Nascimento, M. A., Sander, J., & Pound, J. (2003). Analysis of SIGMOD's co-authorship graph. *Sigmod Record*, *32*(3), 8–10. https://doi.org/10.1145/945721.945722

Ghazal Kalhor, Amin Asadi Sarijalou, Niloofar Sharifi Sadr, & Behnam Bahrak. (2022). A new insight to the analysis of co-authorship in Google Scholar. *Applied Network Science*, *7*(1). https://doi.org/10.1007/s41109-022-00460-4

Baiesi, M., & Paczuski, M. (2004). Scale-free networks of earthquakes and aftershocks. *Physical Review. E, Statistical, Nonlinear and Soft Matter Physics*, *69*(6). https://doi.org/10.1103/physreve.69.066106

Professor Ginestra Bianconi. (2023). Qmul.ac.uk. https://www.qmul.ac.uk/maths/profiles/bianconig.html

Mark Newman. (2014). Umich.edu. https://websites.umich.edu/~mejn/

Albert-László Barabási. (2019). Google.com. https://scholar.google.com/citations?user=vsj2slIAAAAJ&hl=en

Mohler, G., & Stroud, D. (2024). Study of Vortex Avalanches using the Bassler-Paczuski Model. ArXiv.org. https://arxiv.org/abs/cond-mat/9905125

Kozma, B., Hastings, M. B., & G. Korniss. (2004). Roughness Scaling for Edwards-Wilkinson Relaxation in Small-World Networks. Physical Review Letters, 92(10). https://doi.org/10.1103/physrevlett.92.108701

Toroczkai, Z., Kozma, B., Bassler, K. E., W, H. N., & Korniss, G. (2024). Gradient Networks. ArXiv.org. https://arxiv.org/abs/cond-mat/0408262

Krapivsky, P. L., Rodgers, G. J., & Redner, S. (2001). Degree Distributions of Growing Networks. Physical Review Letters, 86(23), 5401–5404. https://doi.org/10.1103/physrevlett.86.5401

Krapivsky, P. L., & Redner, S. (2001). Organization of growing random networks. Physical Review. E, Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics, 63(6). https://doi.org/10.1103/physreve.63.066123

Bosiljka Tadić, Thurner, S., & Rodgers, G. J. (2004). Traffic on complex networks: Towards understanding global statistical properties from microscopic density fluctuations. Physical Review. E, Statistical, Nonlinear and Soft Matter Physics, 69(3). https://doi.org/10.1103/physreve.69.036102

Kim, J., Krapivsky, P. L., Kahng, B., & Redner, S. (2002). Infinite-order percolation and giant fluctuations in a protein interaction network. Physical Review. E, Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics, 66(5). https://doi.org/10.1103/physreve.66.055101

G. Ergün, & Rodgers, G. J. (2002). Growing random networks with fitness. Physica. A, 303(1-2), 261–272. https://doi.org/10.1016/s0378-4371(01)00408-3

Caruso, F., Latora, V., Pluchino, A., Rapisarda, A., & Tadic, B. (2006). Olami-Feder-Christensen model on different networks. ˜the ɶEuropean Physical Journal. B, Condensed Matter Physics/European Physical Journal. B, Condensed Matter and Complex Systems, 50(1-2), 243–247. https://doi.org/10.1140/epjb/e2006-00110-5

Cosenza, S., Crucitti, P., Fortuna, L., Frasca, M., Manuela La Rosa, Stagni, C., & Usai, L. (2005). From Net Topology to Synchronization in HR Neuron Grids. Mathematical Biosciences and Engineering, 2(1), 53–77. https://doi.org/10.3934/mbe.2005.2.53