GESIS Guides to Digital Behavioral Data #3

Expert Insights into Concepts of Network Analysis

An Interview with David Schoch

Publication date: December 24, 2023

Networks are ubiquitous, or, to put it a bit longer: almost all our daily behavior and the digital traces of it are embedded in structures that we can interpret as networks. But what kind of measurement instruments do we need to carve out and 'understand' the social structure we call networks?

We asked David Schoch about this. David leads the team "Transparent Social Analytics" in the Computational Social Science department at GESIS. His main research interest is in the field of (social) network analysis where he has made technical and methodological contributions to topics such as network centrality, signed networks, and two-mode networks.

David told us about how he was drawn into network analysis, why he thinks the field can add substantially to (computational) social science research and shared his views on both the successes and pitfalls of working with network concepts.

The interview was conducted by Indira Sen and Leon Fröhling on April 26, 2023. The transcript was edited for clarity and length.

Keywords: network analysis, centrality measures, centrality indices, network structures, information networks, disinformation campaigns, coordinated behavior, social explanation, social structure

GESIS: Hello David, thank you for permitting some insights into your research with this interview. What are you working on now, particularly in the context of network analysis and centrality measures?

David Schoch: In terms of networks, I am always interested in coming up with new methods for network analysis or trying to understand existing measures. That is something that I have been now working on in the last months: understanding what centrality measures are measuring and if they are actually measuring something in networks. There are so many of these indices that have been defined over the years, but

typically very little effort goes into trying to understand the technicalities of these indices. I am interested in this technical background and also the more substantive perspective of centrality measurement: Do centrality indices operationalize some form of measurement or are they just picking up noise in the network?

GESIS: How did you end up studying networks?

David Schoch: I studied math and got really interested in graph theory, basically the technical foundation of network analysis. I wrote a thesis on clustering (or community detection) in networks and that's when my interest in studying networks really was born. I started my PhD with a desire to study communities in networks, but my supervisor suggested centrality measures as the more exciting topic. Back then, there were very few established guidelines for centrality indices, making it difficult to determine which one to use for a given empirical context. There was neither a technical nor a substantive definition that clearly defined what a centrality index was supposed to measure. While we knew that a central node was important for the network in some way, this definition was too vague, allowing for multiple definitions to emerge, leading to a plethora of centrality indices to choose from.

It became clear that simply throwing ten different indices onto the network and choosing the one that produced the most favorable outcome was not a sustainable approach.

As a result, I found myself facing the problem of having to choose from dozens of centrality indices, each with their own unique set of assumptions and limitations. It became clear that simply throwing ten different indices onto the network and choosing the one that produced the most favorable outcome was not a sustainable approach.

In response to this problem, I began to explore the technical side of centrality measures, attempting to identify the commonalities between various indices. Through this process, I was able to establish a small core of technical features that all centrality indices shared, allowing me to give a more precise technical definition of centrality indices.

However, I quickly realized that this technical definition alone was not enough to address the substantive questions I was interested in. I continued to work on turning my technical results into something that could be used in real-world research, but after ten years, I still do not have a satisfying answer to the question of which centrality index to choose for a given analysis. This is a problem that I am still working on today.

GESIS: During this period, did you have any kind of major turning points?

David Schoch: I have worked on many studies where centralities were supposed to be used, but unfortunately, it is still in a very technical stage that is hard to put into empirical

work. It is still a major problem. However, I think it is the right direction to go, as the data-driven approach is still ongoing.

From an explanatory perspective, we would rather ask […] "What characteristics does the position of a spreader in a network need?"

If we want to explain things rather than predict them, it is not the right approach to use any subset of indices and then see what happens. To give a concrete example: our study could be about spread of misinformation on a social network, and we want to identify the central spreaders. From a more predictive perspective, or rather: data-driven approach, we turn this into a classification task and test which centrality index ranks known spreaders the highest and is best to predict if a user is a spreader. It does not give us an explanation *why* this user is a spreader, just that they might be one. From an explanatory perspective, we would rather ask the questions "What characteristics does the position of a spreader in a network need?" and "How can we operationalize this as a network measure?" to catch all those spreaders.

I moved away from studying technicalities of indices themselves towards trying to understand what impact different network structures have on indices. […] This is a step towards guiding empirical research by assessing the network's structure and determining if it allows for proper centrality analysis.

So, I moved away from studying technicalities of indices themselves towards trying to understand what impact different network structures have on indices. What features does a network have to have in order for indices to give the same, or similar, results and what features make the results vary significantly? We have identified some networks where any index gives exactly the same result but also came up with a process which creates networks where any pair of indices can give completely different results. This is an important result that shows how dependent indices are on the network structure. We cannot say that two indices are similar or different, when there are networks where they both completely disagree or completely agree. The implication for empirical research is that it is hard to reason about centrality by means of indices. For instance, when the network structure only permits one centrality outcome – that is, all centrality indices produce the same result – then this outcome is due to the structure itself and we cannot say that index x measures centrality in this network.

This is a step towards guiding empirical research by assessing the network's structure and determining if it allows for proper centrality analysis. We defined measures that allow us to assess how predetermined a network is in terms of centrality. The closer we are to a completely predetermined network, the less we can expect different indices to give us any meaningfully different outcome, because there is essentially only one.

GESIS: You already mentioned the importance and the implications of your theoretical work. Taking a step back, could you expand on the importance of networks and centrality in the context of computational social science (CSS) research?

Basically, all our digital traces are kind of embedded in networks.

David Schoch: Stepping a bit back from centralities and thinking about networks in a broader sense – basically, all our digital traces are kind of embedded in networks. Specifically, when we think about social media − we have friendship networks, follower networks, retweet networks, mention networks, and all kinds of naturally occurring network structures in the digital world. The challenge that we have in computational social science when it comes to network analysis is that most measures were not designed for large networks. They were introduced 60, 70 years ago with very small networks in mind, and now we have these massive networks with millions of nodes, and all our tools are not designed to work on them. For example, a naive algorithm for computing betweenness centrality could take years on a million-node network. So, we have algorithmic challenges to scale up what we have.

But apart from that, there are a lot of interesting application areas in CSS for networks, specifically in the context of social media. I have already mentioned the case of identifying spreaders of disinformation, trying in general to understand information flows on a social networking site, or using clustering to uncover different subgroups of users.

GESIS: Could you tell us a bit more about different centrality measures and what each of these measures do in different types of networks?

David Schoch: I think the simplest one is degree centrality. So just counting how many neighbors a node has in the network. This is usually used as a kind of a popularity measure. The more friends you have, the more popular you are. The next simple one is closeness which measures the distance from each node to all other nodes, and you are central in the sense of closeness if you have very short connections to everyone else. In the information spreading context, this node can spread information the quickest to everyone. Betweenness centrality is always regarded as a dual measure to closeness. What betweenness measures is: how often does a node lie on the shortest path between nodes? This measure is an information control measure. So, if a lot of information must go through you, you can of course also control how you spread this information.

Degree, closeness, betweenness, and forms of PageRank, like eigenvector centrality, can be considered the standard centrality indices.

Centrality also plays a huge role for search engines. In the beginning, Google was using an algorithm called PageRank to rank their search results based on the hyperlink network of

websites. Naively, one could say that a webpage should be ranked high if a lot of other webpages link to them. But this is easy to manipulate. Simply create thousands of empty websites that only link to you. Instead, for having a high PageRank score, you need many other websites with high PageRank scores to link to you. This concept can easily be translated into other contexts: A node is only important if other important nodes are connected to it.

Degree, closeness, betweenness, and forms of PageRank, like eigenvector centrality, can be considered the standard centrality indices that are most often used in the literature.

GESIS: What are some of the common misunderstandings, misperceptions, and misuses of centrality indices that you encounter in today's literature on network analysis?

David Schoch: I mentioned the impact of network structure on indices − that they can give the same result or different results. And this is something that is still not being understood because when people use two indices on the network and they give very different results then people always assume this is because these two indices just in general behave differently, but that is not always the case. You will always be able to find a network where these two indices end up measuring the same thing, giving you exactly the same result. The differences in indices usually stem from the network structure, not the indices. I think this is by far the biggest misunderstanding of centrality.

GESIS: If I were a researcher who is working with network data and is interested in finding central actors, what would your recommendations or guidelines be, based on your work so far and the misunderstandings that you mentioned?

David Schoch: It is important to understand the network structure before applying centrality indices. I have developed measures that can help determine whether or not using centrality indices is appropriate for a given network [1]. If we are close to a network where only one centrality ranking exists, then it may not be useful to use centrality indices since they all give the same result anyway. Instead, one can simply use the existing ranking. By the way, by ranking, I mean the ranking of nodes that is induced by the values of a centrality index. However, if the network structure is more complex, one must consider the research hypothesis and what mechanisms may be at play. I also introduced an alternative approach to using centrality indices, which involves assigning probabilities to nodes to determine how likely they are to be central in general. While this approach has not been used in empirical research yet, it is an interesting alternative approach which is also feasible for larger networks since good approximation methods exist to assign probabilities to a node's centrality.

GESIS: That already brings us back into the realms of computational social science. There is one example where you recently used some of your methods for an empirical

study which was on detecting coordinated disinformation campaigns online. Could you explain how you approached that study?

David Schoch: Yes, it is actually interesting because when I started working on the project, I was in the middle of finishing my PhD on centrality. I thought that centrality would be the ultimate solution to all problems, including understanding how actors who spread disinformation online should have very central positions within the network. However, we could not come up with a measure to operationalize this, so we took a step back and thought more about how campaigns are being organized to spread disinformation.

We realized that it is not about single accounts but rather accounts that coordinate in some way. So, we are not interested in specific positions of nodes in the network, like centrality, but rather in finding components of the overall network that behave very similarly. We are exploring something related to clustering, not centrality. The details are a bit long, so I skip them for now, but you can read more about it in our paper [2].

GESIS: What method did you use for this work, and what were the implications of using this method?

David Schoch: In the world of Artificial Intelligence, the method we used for identifying disinformation campaigns might seem boring, but it was actually a great example of how putting effort into thinking about the problem from a social science perspective can lead to a simple and effective solution. We realized that we did not need a complex method to identify disinformation campaigns. Instead, we created a network of social media users based on whether they posted the same message within a short time window. This helped us identify coordination patterns and isolate the accounts involved in a disinformation campaign. While some noise can occur, such as people posting about a celebrity's birthday, we filtered this out and were able to identify over 40 campaigns across various geopolitical settings. It is surprising how well this simple method worked from a network analytic perspective.

GESIS: You mentioned that this method was specifically developed for detecting coordinated disinformation campaigns online. Do you think this method could be used for similar research tasks, or is it limited to detecting coordinated disinformation? Have you seen other researchers using this method for their work?

David Schoch: The problem with the method we used to detect coordinated disinformation is that it is quite restrictive. We only consider messages that are basically copy-pasted with the exact same wording, so it is a very blunt way of detecting coordination. However, the method has been further developed to measure similarities between content and create similarity networks.

I have not seen many other use cases for this method, but I think it would make sense to use it in systems where coordination is expected to happen. The definition of coordination may differ across contexts, but the method would still be applicable. The umbrella term is always coordination in some sense.

Some of the things that do not make it into official papers are things that did not work out. So, were there any such examples in this work?

David Schoch: Yeah, many examples. Initially, we thought that centrality would be the key factor, but it quickly became clear that it would not work. So, we started looking at other network measures, such as structural features, but it was a matter of finding the right theory and operationalizing it. We needed to determine the time window that counts as copy-pasting, and we settled on using exact matches instead of similarities because it worked well in our cases.

GESIS: What are some additional resources on the topic of network analysis in general or your work on centrality more specifically that you would like to point other researchers to, for them to really get a deep understanding of the topic?

David Schoch: There are some good introductory books on networks, such as Mark Newman's book "Networks: An Introduction" [3] and "Doing Social Network Research" [4], which is a very non-technical book focused on the substantive side. These provide a good overview of network analysis from a technical and theoretical point of view. From a more graph theoretic perspective, the 'Bible' of network analysis is Wasserman and Faust's "Social Network Analysis" [5] which is an old book from the 1990s but still relevant.

> ***… think about mechanisms that can be operationalized into proper centrality measurement.***

GESIS: To conclude, in this topic of network analysis and centrality, if you could make a wish to the universe for a research artifact, like one package, one theory or one research agenda, what would it be?

David Schoch: People should stop inventing new indices and stop throwing a lot of indices at networks, but rather think about mechanisms that can be operationalized into proper centrality measurement.

GESIS: David, thank you for this interview!

References

1 Schoch, D., Valente, Th. D., & Brandes, B. (2017) Correlations among centrality indices and a class of uniquely ranked graphs, *Social Networks*, 50, 46-54, DOI: 10.1016/j.socnet.2017.03.010

2 Keller, F. B., Schoch, D., Stier, S. & Yang, J. (2020) Political Astroturfing on Twitter: How to Coordinate a Disinformation Campaign, *Political Communication, 37*:2, 256-280, https://www.tandfonline.com/doi/abs/10.1080/10584609.2019.1661888

3 Newman, M. E. J. (2010). *Networks: An introduction*. Oxford University Press.

4 Robins, G. (2015). *Doing Social Network Research*. SAGE Publications Ltd.

5 Wasserman, S., & Faust, K. (1994). *Social network analysis methods and applications*. Cambridge University Press.

Suggested citation

Schoch, D. (2023). *Expert Insights into Concepts of Network Analysis. An Interview with David Schoch* (GESIS Guides to Digital Behavioral Data, 3). Cologne: GESIS – Leibniz Institute for the Social Sciences.

Series editor

Maria Zens

Publisher

GESIS Leibniz-Institute for the Social Sciences, Cologne

License

Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0)