MATH189AI Short Essay 4:
An Exploration of the Structure of Mathematics

(As Described by Wikipedia)

Theo Rode, Maddie Reeve, Christian Johnson

Wikipedia offers a unique look into a large component of human knowledge. In particular, by constructing a network of pages (nodes are pages, edges are links between them) we can gain a structural understanding of the knowledge embedded in Wikipedia. This structure could illuminate topics of particular importance and clusters of similar ideas. For this project, I aim to ask the question: **does the structure of definitions in mathematics, as descried by a network of Wikipedia pages for these definitions, illuminate latent structure in the field of mathematics?** By examining mathematics in this way, we may draw out interesting correlations between subfields of mathematics, and find concepts/definitions/theorems that are particularly central to specific fields or mathematics as a whole. It is common to hear about a particular theorem that is "fundamental" in some way, but is that actually the case? This is additionally a unique way of examining connection between subfields in mathematics as we can use our network to measure similarity based on overlapping concepts directly. In particular, it can provide a direct numerical comparison for the similarity between different sub-categories of mathematics.

Investigation into the network structure of Wikipedia is not uncommon. Similar to the planned analysis outlined here, Silva et at. investigated the Wikipedia structure for the categories of Mathematics, Biology, Physics, and Medicine [1]. In particular, they aimed to answer the question: how does the network structure change between the four topic areas, and what conclusions can be drawn from these differences? In their paper, they find that Mathematics, Physics, and Medicine all adhere to a power law degree distribution, while Biology does not. They note that this, in addition to the power law distributions having different values of γ, suggest that the growth of these categories is different. Further, they note that the structure inherent to each group is distinct, with Biology and Medicine sharing a "high modular pattern," while Mathematics is sparse, and Physics has a "dense core" [1]. They further compare all of these to the greater Wikipedia network, demonstrating that the inherent structure of these categories is distinct from that of the larger network, giving important pause if using characteristics of the larger Wikipedia network to analyze categories within Wikipedia. Similar to this work, my goal is to analyze structure inherent to a network constructed by Wikipedia pages. Using many of the same metrics and measurement techniques as Silva et al. would likely be incredibly productive in my analysis of particularly the mathematics category. However, distinct from this paper, my goal would be to analyze mathematics specifically, and therefore there will be many more sub-categories of mathematics to analyze separately. This will allow for a more thorough investigation of similarity between categories, where Silva et al. was only considering 4 different categories. In particular, an interesting question is how categories/subfields of mathematics are related, which will involve analysis on the similarity of collections of nodes—this is something that Silva et al. were unable to fully consider due to the limited set of categories [1].

It will be important to examine many forms of centrality measures in order to analyze how certain definitions in particular sub-categories of mathematics are rated in terms of relevance. Further, similar Silva et al. [1], it will be interesting to examine how these centrality measures change when we move from the centrality measure within sub-categories to the centrality measures on the full network. This can give us a picture on how certain definitions are more important in the context of multiple categories, and therefore act as "connections" between categories. In addition to this, in order to examine how well "grouped" mathematics is, it will be important to use community detection algorithms. These could be used in order to determine if the self-assigned Wikipedia categories line up with found communities, and/or if certain categories are close enough to be grouped together. Furthermore, as discussed by Newman, community detection allows us to simplify networks into just communities, where we could theoretically use this representation to analyze the similarity between the different sub-categories of mathematics. As for computational needs, I don't expect the size of this dataset to be large enough to be a concern. As there are only about 6.5 million Wikipedia articles, and we are only considering the subset relevant to mathematics, the size of the network will not be overly massive. For procuring the data, Wikipedia has a simple API that should allow us to retrieve the necessary data easily.

One inherent risk to this project is the possibility that there either isn't enough data on Wikipedia to make conclusions about the structure of mathematics, or the structure simply isn't interesting. However, as Silva et al. found interesting structure in their network [1], I find it unlikely that this will happen—especially considering the size of Wikipedia has almost doubled since. It is also possible, falling under the same umbrella, that community detection on this network fails to give us insight into sub-topics of mathematics. In which case, this analysis will differ considerably less from Silva et al. However, if successful, this project will give us insight into mathematics more broadly. In particular, for undergraduates studying math especially, it will be valuable to see how many of the concepts applied in classes come together to form a coherent and interconnected field. Hopefully, we also are able to construct a visually appealing visualization that will simply be cool to look at!

References

[1] Filipi Nascimento Silva et al. “Investigating relationships within and between category networks in Wikipedia”. In: *Journal of informetrics* 5.3 (2011), pp. 431–438.