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

(As Described by Wikipedia)

*Christian Johnson, Maddie Reeve, Theo Rode*

## 1 Overview

Wikipedia offers a unique and unparalleled collection of concepts, definitions, and general information on a broad range of topics. Typically, each Wikipedia page describes a single one of these concepts or definitions while also linking to other related Wikipedia pages. This page structure implicitly gives a natural interpretation of Wikipedia as a network: each page representing a node and edges representing a link between pages. As a link between pages inherently represents a form of similarity between pages, we aim to use network analysis techniques to determine the latent structure of the field of mathematics on Wikipedia, and to quantify the structural relationships between subcategories of math. Specifically, does mathematics—as described by Wikipedia—give rise to a latent network structure that illuminates central ideas, definitions, and theorems within the field, and can we draw further connections between hierarchical subcategories of mathematics? By exposing these central relationships in mathematics, we hope to give a better understanding of the interconnectedness of the field. This improved understanding could have influence in the classroom, aiding professors and students in understanding intricate relationships between concepts, but additionally in research, where bringing together related ideas may lead to new innovation.

In order to investigate this question, we will make significant use of a variety of centrality measures on our network. By analyzing these centrality measures not only on the entire graph, but also the subgraphs constructed by sub-fields of mathematics, we may gain significant insight into important nodes on the network. Additionally, assortativity measures and community detection will allow us to understand the (potential) complex relationships between categories of mathematics, and with the unique hierarchical structure of Wikipedia categories, we may gleam insights into the intersections of separate fields of mathematics. Furthermore, understanding the degree distribution of the entire mathematics network, and the category subgraphs, may allow us to understand how the network is growing, and pull out important distinctions found between subcategories. Through thorough investigation utilizing a wide array of network metrics, we hope to illuminate components of the latent network structure of mathematics on Wikipedia.

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 [10]. 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? They find significant structural differences between the different topic areas and the larger overall network with regards to degree distribution. Further, they show how certain centrality measures, when examined in the subgraphs versus the entire graph, illuminate nodes which are particularly connective between the different subjects. With their analysis only focuses on four subcategories, we aim to go deeper in our analysis with many more categories. Additionally, using the categories provided by Wikipedia, we may analyze the subcategories in regard to their hierarchy, allowing for a richer understanding of relationships between categories.

Kaptein and Kamps explored how the category structure of Wikipedia can be used for entity ranking [6]. Their preliminary goal was to explore how leveraging Wikipedia's inherent category and link structures could improve the process of ranking entities directly in response to a search query. They found that using Wikipedia's category data significantly improved the retrieval results. This motivates a continuation into a look of how the category structure of Wikipedia can reveal information about the structure of the pages. We will use the hierarchical nature of the Wikipedia categories to explore the structure of the Mathematics Wikipedia pages.

In further analysis of the structure on Wikipedia, Gabella investigated the influence of culture as seen through the network structure of Wikipedia [4]. In particular, Gabella investigated the network induced by Wikipedia when only considering the first link out of a page. This constructed a graph where each node only had a single outgoing edge, though particularly central graphs could have many incoming edges. It was found that depending the language the articles were written in, and thereby the culture associated with the articles, different nodes stood out as central in the network. In performing the analysis, Gabella investigated three main metrics: the core cycle, the degree distribution, and betweenness centrality. It was found that a majority of nodes in the graph where attracted to a single core cycle in the network, and the smaller cycles represented core

concepts in the network. Further, degree distribution was found to model a power law distribution, aligning with the characteristics of other scale-free networks. Finally, betweenness centrality proved useful in analyzing how close a particular node was towards the core cycle, and therefore how central it was to the entire network. We may build on this work by extending this analysis to the rich hierarchical category information, and specifically applying the analysis to mathematics.

## 2 Resources Needed

Central to our project is link and category data for mathematics pages on Wikipedia. We plan on collecting this data through a mixture of methods, both utilizing the frequent data dumps including all the necessary information about articles we require, as well as utilizing the open-access API provided to grab article data from Wikipedia. The API allows us to pull hierarchical category data from Wikipedia, and then pull data pertaining specifically to articles within each category. The data dumps will provide us with an alternative source of this data in the event rate limits become an issue (a useful backup). Additionally, when looking at the structure of the links between the pages within mathematics, we will have to filter out some pages, such as the page "Mathematics", that do not contain any key results or ideas but link to most other pages within the network. We are interested in the connections between concepts/ideas/theorems, so we will filter out the nodes that are not necessary. We expect our graph to be on the order of magnitude of hundreds of thousands of nodes—constrained by the about 6.5 million total Wikipedia articles—and therefore we believe that our laptops will be sufficient for running further analysis on our graphs. To perform our analysis, we will use the Julia programming language [2], alleviating many of the performance concerns with alternatives such as Python, and allowing us to perform many more experiments on our limited hardware. We will use the HTTP.jl package to get the data, Graphs.jl [3] to create and analyze our graph, and other Julia packages such as LinearAlgebra.jl to complete this project.

## 3 Tentative Workplan

Our project work plan consists of 3 major stages. The first stage involves gathering our data from the Wikipedia API and then putting this data into a graph for further analysis. After we gather the data, we will compile it into a graph using the Graph.jl package. We aim to complete this stage by November 9th.

The next stage of our plan is analysis. We will explore the graph we created to see what results we can arrive at. Specifically, we are preliminarily interested in looking at assortativity, centrality measures, and community detection. However, our approach to analysis will be iterative and we will further decide what to focus on/explore based on what seems to be the most interesting. In this stage we will begin to create figures to explore the structure of our graph. We aim to complete this stage by November 20th.

The final stage of our plan is writing up our results and polishing our figures for a presentable write-up. We will do this ideally by December 6th, but certainly by December 9th, the due date.

## 4 Vision and Contingencies

Given the successes of similar research previously conducted our group is optimistic in the final outcome of our project. In a full success scenario in which everything goes perfectly we will have several notable outcomes. Firstly we will have produced a visually rich network of mathematical definitions on Wikipedia. Additionally, our network will be informative allowing viewers to easily identify mathematical subfields and high-centrality nodes. Ideally the visualization for our network would have functionalities such as zooming into specific subfields to examine their internal structure. Another notable outcome for this full success scenario would be revealing non-obvious relationships between mathematical subcategories. This could be identifying clusters, or communities, of topics which had been previously considered conceptually disconnected.

Unfortunately, things rarely go perfectly as planned and there are some foreseeable challenges that could lead to a partial success scenario. One potential setback is that the community detection algorithms we use may fail to uncover meaningful clusters. This could be due to the inherent sparse nature [10] of the mathematics category of Wikipedia. If these circumstances lead to issues we could pivot our focus to analyzing results from centrality measures. This alternative method would still allow our group to analyze the importance of particular mathematical definitions or theorems. Another potential challenge is difficulties with retrieving complete data through the Wikipedia API. In the case that parts of the dataset are incomplete we could scale the scope of our project down. Narrowing the scope on a better documented subset of the network, such as calculus or algebra, could make analysis of the network more manageable while still providing valuable insights on the structure of mathematics.

## 5 Anticipated Learnings

Through this project, our group anticipates learning a wide range of skills which broaden our technical and theoretical knowledge. The implementation of centrality measures and other algorithms on the real world data will provide us with a better understanding of network science methods. This deeper understanding of network science methods such as centrality measures and community detection algorithms is one of the main outcomes from working on this project. Additionally, a technical skill we envision picking up from working on this project is a proficiency in Julia. Julia is a high level programming language [2] which boasts substantial performance gains over Python, and offers a host of powerful mathematics libraries that would be invaluable in our future mathematics work [1, 5, 7, 8, 9].

Further, we hope to come out of the project with a deeper understanding of the structure of mathematics. As mathematics students, this increased understanding may aid us in future work and give us a greater appreciation for the field. This may serve as a spark to learn a new concept that is interconnected to ones we already love, or give us reason to explore a topic entirely orthogonal to what we already know. We hope that this new look into mathematics allows us to see more into the inherent beauty of mathematics.

## References


[1] Gaurav Arya et al. "Automatic differentiation of programs with discrete randomness". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 10435–10447.

[2] Jeff Bezanson et al. "Julia: A fresh approach to numerical computing". In: *SIAM review* 59.1 (2017), pp. 65–98. URL: https://doi.org/10.1137/141000671.

[3] James Fairbanks et al. *JuliaGraphs/Graphs.jl: an optimized graphs package for the Julia programming language*. 2021. URL: https://github.com/JuliaGraphs/Graphs.jl/.

[4] Maxime Gabella. "Structures of Knowledge from Wikipedia Networks". In: *arXiv preprint arXiv:1708.05368* (2017).

[5] Shashi Gowda et al. "High-Performance Symbolic-Numerics via Multiple Dispatch". In: *ACM Commun. Comput. Algebra* 55.3 (Jan. 2022), pp. 92–96. ISSN: 1932-2240. DOI: 10.1145/3511528.3511535. URL: https://doi.org/10.1145/3511528.3511535.

[6] Rianne Kaptein and Jaap Kamps. "Exploiting the category structure of Wikipedia for entity ranking". In: *Artificial Intelligence* 194 (2013), pp. 111–129.

[7] Michael Lindner et al. "NetworkDynamics.jl—Composing and simulating complex networks in Julia". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 31.6 (2021), p. 063133. DOI: 10.1063/5.0051387. eprint: https://doi.org/10.1063/5.0051387. URL: https://doi.org/10.1063/5.0051387.

[8] Christopher Rackauckas and Qing Nie. "DifferentialEquations.jl – A Performant and Feature-Rich Ecosystem for Solving Differential Equations in Julia". In: *The Journal of Open Research Software* 5.1 (2017). Exported from https://app.dimensions.ai on 2019/05/05. DOI: 10.5334/jors.151. URL: https://app.dimensions.ai/details/publication/pub.1085583166%20and%20http://openresearchsoftware.metajnl.com/articles/10.5334/jors.151/galley/245/download/.

[9] Ali Ramadhan et al. "Oceananigans.jl: Fast and friendly geophysical fluid dynamics on GPUs". In: *Journal of Open Source Software* 5.53 (2020), p. 2018. DOI: 10.21105/joss.02018. URL: https://doi.org/10.21105/joss.02018.

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