Exploratory & Descriptive Analysis for Math Topics

Sam Ly Nathan Brown Jacob Lembach

November 9, 2025

1 Dataset Overview

The dataset used was the Wikipedia Math Essentials dataset. This dataset was sourced from Kaggle and represents the relationships between math articles on Wikipedia. From the dataset, a directed graph is created with nodes being articles and edges being links from on article to another. This yields a graph with 1,068 nodes and 27,079 edges.

This dataset is relevant because studying the relationships between topics in mathematics gives us a better understanding of the overall field. We can identify understudied topics, find new research areas, etc. Studying this dataset also has important educational applications, as it allows educators to make more informed course content decisions. Determining which topics to include in courses is difficult, and having a quantitative measure of importance can be instrumental to making that decision.

2 Methods Summary

The dataset was formatted as a large JSON object, so reading and using the dataset was relatively easy. The dataset was already cleaned, so no cleaning methods were needed.

The nodes were given an integer ID, which could then be mapped back to the article name. Then, the edges were given as a large list of node IDs. Duplicate edges, ie. multiple links from and to the same articles, were handled for us, and were given to us as edge weights. The final graph built must be directed because links between articles are directed.

3 Results and Interpretation

3.1 Primary Metrics

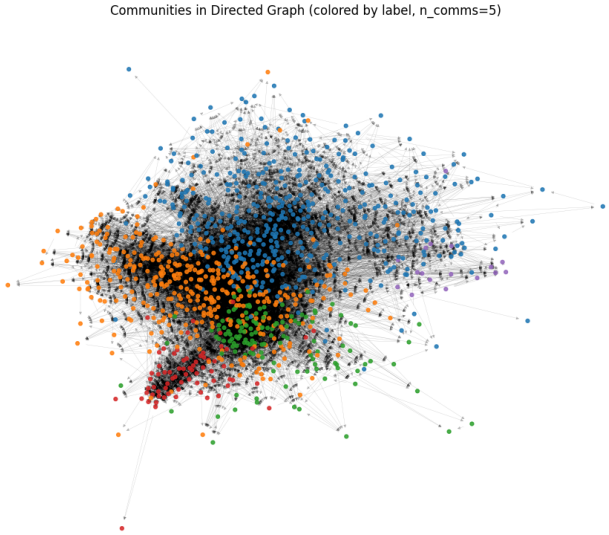
As stated before, our graph contains 1,068 nodes and 27,079 edges in total, meaning it tracks 1,068 articles and 27,079 links. However, we notice that our

density is extremely low, at 0.02. This isn't out of the ordinary, since this is not a social network. Intuitively, we know that each individual article tends to have a small amount of links relative to every other article out there.

One notable feature of this graph is that there are no cycles. This can be shown when finding the strongly connected components. There are no components with size > 1. This means that no two nodes have a "two-way connection". This also makes sense, as a chain of articles linking back on itself isn't very useful. This also means that the graph is **acyclic**, thus there exists valid topological sorts of this graph. Although not used not, the topological sort of the graph may provide valuable insights.

3.2 Communities and Clustering

In theory, finding the communities of nodes within our graph would give us insight on the relationships between the branches of mathematics.



However in practice, this graph tells us more about how Wikipedia has ended up structuring the articles for math, rather than the underlying structure of math knowledge. Further experimentation is needed to find the information we are looking for.

By running a greedy modularity search, we find 5 primary communities. We define the hub of each community to be the node with the highest out degree. We will see later that the degree distribution of the out degree follows that of a scale-free graph, so the degree centrality is actually a viable measure.

We find the hubs and their corresponding out degree to be:

```
Mathematics 533
Real number 290
Calculus 197
```

```
Statistics 160
Prisoner's dilemma 19
```

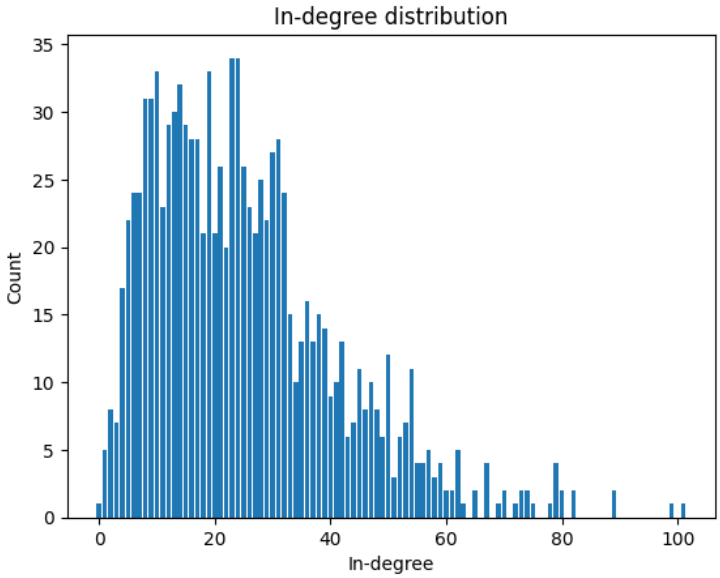
Here we see some issues with our dataset. Extremely broad articles like "Mathematics" have isn't really a "math topic" per se, and can get in the way of actually useful information. Also, definition-heavily articles like "Real number" end up seeming like important topics because of their high out degree, despite not necessarily meetin the criteria for an actual topic. The "Calculus" and "Statistics" nodes make the most sense.

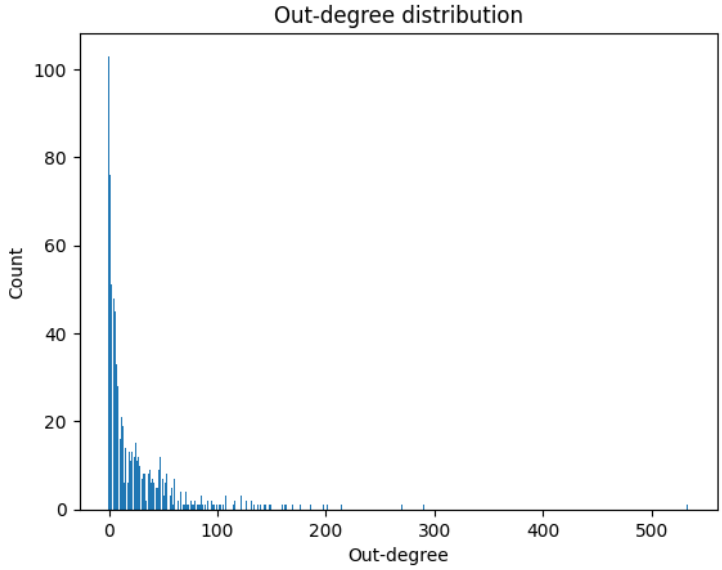
An artifact of our clustering algorithm is the "Prisoner's dilemma" being falsely considered a hub. When looking at the other nodes in its community, we see that the actual hub should be the article "Game theory", however it was somehow lost when we computed the communities.

3.3 Degree Distribution

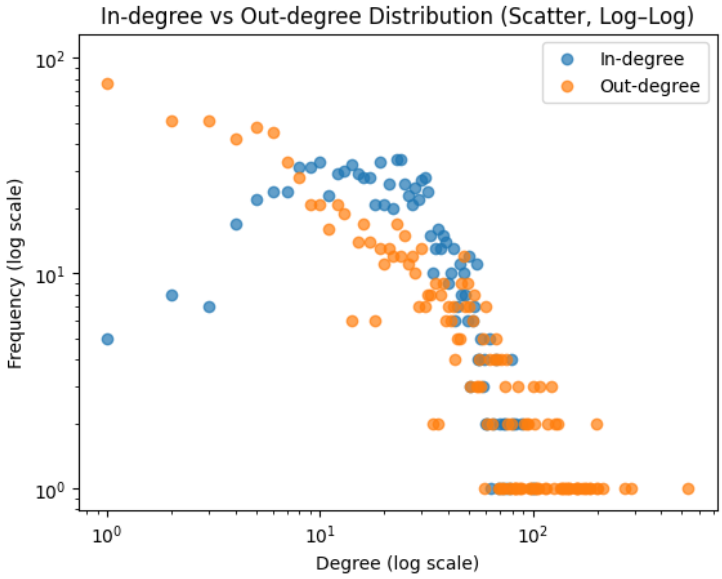
First, because our graph is directed, there are two distict degree distributions: the in degree and out degree.

The in degree distribution seems to follow a Poisson distribution, while the out degree distribution follows the degree distribution of a scale-free network.





This is more clearly shown on a log-log plot.



This pattern emerges because articles about broad topics tend to link outwards to many other articles. At the same time, there are fewer broad topics out there to write articles on. So the scale-free distribution emerges as broad articles are "preferred" to link outwards.

Counterintuitively, inwards links seem to be more random, hence the Poisson distribution. Although we may falsely assume that broad articles like "Mathematics" should also be preferred as link targets, in reality, articles for broad

topics don't get linked to any more than other articles. An obvious example of this would be that literally every node in the dataset semantically relates to the "Mathematics" article, but it doesn't mean that every single article written needs to refer to "Mathematics". Another way to view this is from the perspective of the writer. When referring to other articles, the usefulness of said article is independent from its "broadness". Thus, the likelihood of an article receiving a new inwards link is practically random.

4 Visualization Discussion

Write about the visual

Referring back to the figure, ...

5 Reflection and Next Steps

write this out in paragraph form

What new questions or hypotheses emerged from your analysis?

- How can we quantitify article "broadness"?

What analyses might you do next (e.g., centrality, diffusion, or community detection)?

- The dataset includes time-series data. Thus, we can see if a diffusion model makes sense.
- Experiment with other community detection algorithms.

Note any technical challenges or surprises you encountered.

- Finding communities in articles gets "weird".