The united counties of arXived mathematics

Samuel N. Cohen
samuel.cohen@maths.ox.ac.uk
Mathematical Institute, University of Oxford

October 18, 2019

1 Introduction

This short summary describes the process by which I constructed this 'map' of mathematics. The aim of this was a simple one – to try and describe, using data from real papers and authors' own identification – the most prominent interconnections between areas of mathematics, and then to present it in an somewhat interesting manner.

By using the subject codes for research papers, this approach particularly tries to identify which areas of mathematics are sharing current research questions, rather than what the historical links are between them, or what area is foundational for another.

While the approach given here seems broadly sensible (to me!), this project was mainly for fun, rather than to attempt a detailed or rigorous analysis of the discipline of mathematics.

2 Data

The data used comes (as the title suggests) from arXiv.org. Using arXiv's OAI API (`https://arxiv.org/help/oa`), I extracted the metadata from all papers listed in '.math', which stretches back to 2007. This gives approximately 425k papers to mid-September 2019 (when I extracted the data). This data is provided by authors when they submit papers to the database.

From each paper, I extracted the MSC codes as listed in the 'subject' of the metadata. These codes indicate, in a moderately precise way, the mathematical topic of the paper. For each paper, I reduced down to only the first two characters of the MSC code, which indicates the general mathematical field. Around 47.8% of papers in mathematics on arXiv have at least one subject area listed, while around 26.7% are listed as belonging to more than one subject area.

From this data, I computed two statistics:

1. The number of papers submitted in each subject area since 2013. This recent date was chosen due to the fact that the adoption of arXiv has not been uniform among different areas. As arXiv was started by physicists, the areas closest to mathematical physics appear to be significantly over-represented earlier in the data, but this effect is lessened more recently.

When more than one subject area was listed, the weight was split evenly between them, to prevent closely related areas from artificially boosting each other (i.e. if every paper in an area was listed in both subjects, the total weight would remain the same).

There is still noticable bias in this data as a representation of mathematical research (which is why the title is 'arXived mathematics', rather than just 'mathematics'); this is particularly noticeable for mathematics which is interacting with other disciplines. Computer Science, Statistics, Quantitative Finance and Quantitative biology all have their own arXiv sections, the social sciences use another repository (as does biology, chemistry, ...) , and an informal straw poll of Oxford's mathematicians working in applied mechanics indicates that many people are publishing through engineering conference proceedings, rather than using preprints on arXiv.

Nevertheless, this statistic gives an indication of the scale of each subject area.

2. The total number of papers (over the whole dataset) which connect each pair of subject areas. Given we are going to rescale this statistic in what follows, it is less important to correct for the subject-specific biases.

   Of the 63 active MSC codes, 91.7% of pairs of subjects had at least one paper relating these two areas, while only 30.7% of pairs had at least 100 papers relating these areas.

3 Generating a graph

To generate an interesting graph, there are various techniques that can be used. In this data set, some subjects have significantly more papers than others, and there are a large number of pairs of topics with few papers linking them. To address this, and produce a readable and informative graph, I modified the connections as follows:

1. I first rescaled the connections, so that each row and column of the connection matrix had a total equal to the square root of the number of papers in that area. This used a variant of the algorithm by Paul Slater (arxiv:0904.4863). The square root of the total number of papers seemed to give a good compromise between allowing bigger subjects to have interactions with many others, while still emphasising the connections from smaller subjects.

2. I then removed all but the top 10% of connections, with the rule that if this would create an orphaned node (ie a subject with no links), then the link from this node with the largest weight should be included.

Using these transformations to the matrix counting the numbers of papers between two subjects, I obtained a directed weighted graph, with a reasonable degree of sparsity.

4 Communities/Counties of mathematics

Now that we have our graph, it is an easy process (automated in `R` by the `igraph` package), to attempt to detect communities within the graph. After experimenting, I settled on the use of the 'louvain' community detection algorithm (Blondel, Guillaume, Lambiotte and Lefebvre, arXiv:0803.0476). This algorithm detected some broadly-realistic-looking communities within the subject classes. I decided to call these 'counties', as a Count is surely the most mathematical regional administrator.

I obtained a rough initial layout using the Fruchterman–Reingold layout algorithm (doi:10.1002/spe.4380211102), as implemented in `igraph`, which I then extensively tweaked by hand. The resulting graph visualisation is shown below (Figure 4). In this plot the area of a node corresponds to the number of papers in that area, while the colour corresponds to the community detected. Graphically close nodes are *typically* strongly connected, but visual simplicity was preferred to optimality.

5 Design, lettering, etc...

Using the design above as a basis, I sketched out an initial layout for the 'map' image. Individual nodes were converted into building designs, which very roughly correspond to the size of the nodes in the computer generated design. Communities were translated into a variety of topographical boundaries (i.e. rivers, cliffs, fences...)

In order to fit, subject names needed to be shortened relative to the full MSC specification, for which I used some judgement, trying to indicate the feel of the area even if some subfields weren't specified (so, for example, "Game theory, economics and behavioural sciences" became "Games and social sciences").

Lettering was done in a style loosely based on classical Celtic work (e.g. the Book of Kells or the Lindisfarne Gospels, but with nowhere near their elegance...). Text and outlining was done using Winsor & Newton waterproof Indian ink, colouring using Royal & Langnickel gouache.

Thanks

Thanks to Renaud Lambiotte and Mason Porter for useful comments and suggestions, and to my children for putting up with me taking over 'their' painting table for a few weeks.

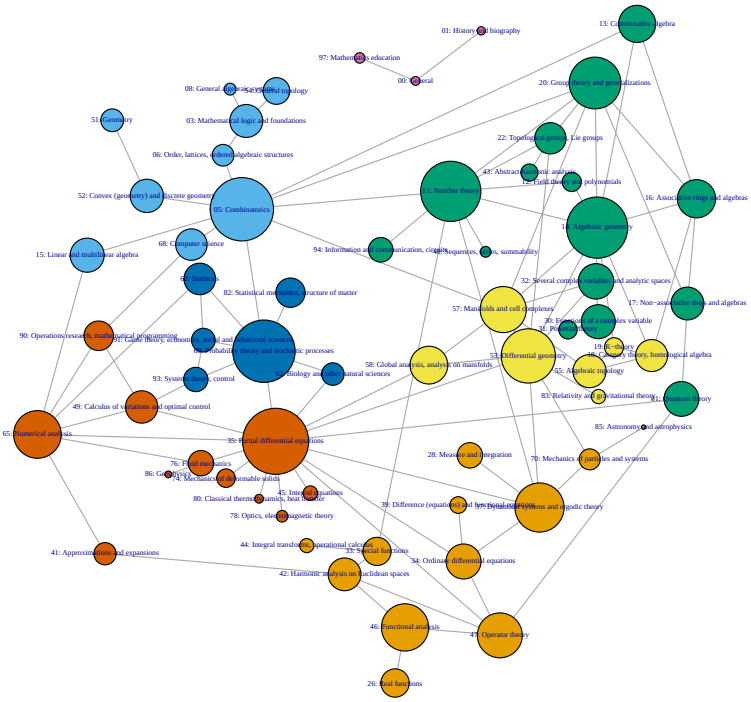

Figure 1: Automatially clustered graph of subject classifications, with manually tweaked position of nodes.