模式识别实验 2

210810508 - 彭珂
1094451830@qq.com

Contents

1 数据预处理与可视化

1.1 数据介绍

本实验的数据来源于 Mathlib4 开源项目 [1]，该项目是一个数学证明库，其中包含了大量的数学定理和证明。在 Mathlib4 中，相关的数学定理和证明被组织成了单个 lean4 文件，每个 lean4 文件中包含了一个或多个数学定理和证明。在本实验中，我们将 lean4 文件看作是一个节点，节点之间的依赖关系看作是一条有向边，这样我们就可以将 Mathlib4 中的数学定理和证明组织成一个有向图。

1.2 数据预处理

Mathlib4 中不仅有数学定理的证明，还包括一些 lean4 的库文件、编译脚本、证明策略等等。在本实验中，我们筛选出了数学定理和证明，而去除了其他文件。然后进行依赖的提取。

为了快速地得到 Mathlib4 中数学概念的大致依赖关系，我们把每个 lean4 文件中直接使用 import 关键字引入的其他 lean4 文件看作是该文件的一个依赖，而没有使用编译器前端或者 lsp 等进行依赖分析。但实际上，由于 lean4 灵活的命名空间，lean4 文件之间的依赖关系可能会比我们所统计的要复杂。另一方面，单个 lean4 文件也可能包含相当多的数学定理和证明，这具体取决于代码贡献者是否严格遵循 Mathlib4 的编程规范以及仓库管理者是否对代码贡献者的代码进行了严格地审核。因此，我们的数据只能作为 Mathlib4 中数学定理和证明的一个大致依赖关系的参考。

按照上面的方法，我们把 Mathlib4 中的文件的依赖关系提取出来之后，共有 4910 个节点和 13858 条边，我们把数据保存为一个 json 文件，以便之后使用。

1.3 数据可视化

把上文中提取出的数据进行可视化，我们可以得到 Mathlib4 中各个学科的依赖图。图中的每个节点代表一个 lean4 文件，节点之间的边代表依赖关系，节点的颜色取决于其所在的子学科，边的颜色与被依赖的节点相同。下图 1 展示了 Mathlib4 中的各个子学科的依赖关系。

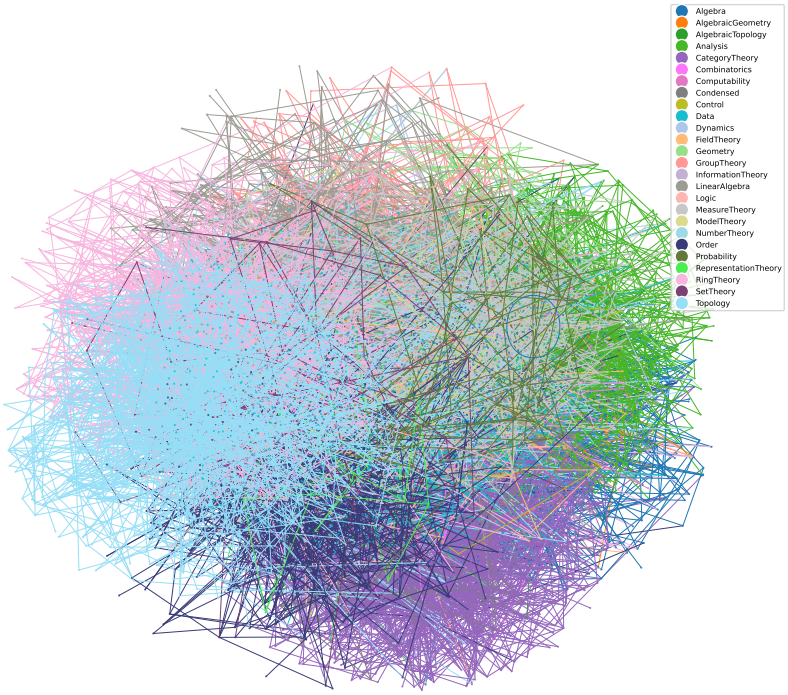

Figure 1: Dependency graph of Mathlib4

2 社群识别与谱聚类

2.1 数据分析

在进行社群识别之前，我们首先对数据进行一个粗略的分析。我们统计了每个子学科的节点数、任一端点位于学科内的边的数量、边的两个端点均属于该学科的边的数量及比例以及学科节点占总节点的比例。统计结果如表 1 所示。

Subject	Nodes	Edges	EdgesInSameSubject	EdgesRatio	NodesRatio
Algebra	894	6058	1939	0.32	0.18
AlgebraicGeometry	70	303	100	0.33	0.01
AlgebraicTopology	43	171	53	0.31	0.01
Analysis	488	2708	972	0.36	0.10
CategoryTheory	591	3351	1403	0.42	0.12
Combinatorics	107	429	112	0.26	0.02
Computability	18	66	15	0.23	0.00
Condensed	25	121	33	0.27	0.01
Control	25	112	28	0.25	0.01
Data	576	3144	801	0.25	0.12
Dynamics	23	95	16	0.17	0.00
FieldTheory	52	292	71	0.24	0.01
Geometry	80	311	109	0.35	0.02
GroupTheory	119	649	157	0.24	0.02
InformationTheory	1	1	0	0.00	0.00
LinearAlgebra	233	1411	402	0.28	0.05
Logic	50	383	55	0.14	0.01
MeasureTheory	196	1001	350	0.35	0.04
ModelTheory	29	118	41	0.35	0.01
NumberTheory	149	643	168	0.26	0.03
Order	209	1281	331	0.26	0.04
Probability	61	221	85	0.38	0.01
RepresentationTheory	15	83	18	0.22	0.00
RingTheory	368	2169	638	0.29	0.07
SetTheory	46	216	55	0.25	0.01
Topology	442	2379	820	0.34	0.09

Table 1: Number of nodes, edges and edges in the same subject for each subject

我们知道，对于一张随机图，当 NodeRatio 较小时，EdgesRatio 应大致为 NodesRatio 的一半，而实际上 EdgesRatio 远大于 NodesRatio 的一半，说明这张图中确实存在社群结构，且该结构一定程度上被学科分类所反映。我们下面尝试使用谱聚类的方法来找到这些社群。

References

[1] Mathlib Community. Mathlib4: A lean mathematical library. https://github.com/leanprover-community/mathlib4, 2023.