一：建模

首先，我们以论文为点，设 a 被 b 引用为 a 到 b 有边建模。因为论文大多数按照时间顺序写作，也就是论文引用一般是引用比其先发表的文章，因此论文网络一般是无环的（这组数据的确也是无环的）。

其次，我们以作者为点，以作者之间合作关系为权 1 的无向边，建模。

二：算法

在论文网络里，我们使用 brandes 算法计算了论文的介数中心度，来获取论文的重要性。并且我们将作者的所有论文的介数中心度求和，来判断作者的重要性。介数中心度越高，代表作者学术水平和影响力越大。我们在可视化部分输出了（前 600 个教授中）前五教授的介数中心度和姓名，可以用于选择导师。

在作者网络里，我们计算了作者的紧密中心度，来了解作者间关系以及获取人脉广的作者。我们在可视化部分输出了（前 600 教授中）每个强连通分量紧密中心度最小的五个作者的紧密中心度和姓名，可以用于找到人脉最广的教授，可以用于联系其余教授。

在作者网络里，我们用 dijikstra 算法计算了任意两点间最短路，并利用先前的最短路结果获取了强连通分量。可以用于从一个已知教授联系到你想找到的教授。

圈子算法：

圈子是对论文作者的划分，一个圈子内部的作者关系紧密，圈子之间的作者关系较为疏远。将作者划分为一个一个圈子后，我们可以进一步将一个圈子收缩为一个新的结点，在新图中做进一步的分析。在尽可能维持原图的关系的前提下，缩减图的规模。

对数据进行分析后，我们认为两个作者的亲密程度可以由以下几个因素确定：

1.两个人之间是否合作发表过论文。

2.两个人所写论文是否发生过引用。

因此，我们为两个作者之间建立了亲密度这一指标。亲密度的大小正比于两个人合作的论文数量和发生过的引用关系数量。如果两个人之间的亲密度大于一阈值，则认为这两个作者在一个圈子里。同时，如果 A 与 B 在一个圈子、B 与 C 在一个圈子，则认为 A 与 C 在一个圈子中。我们使用了并查集来判断两个人是否在一个圈子中。

求解出圈子后，我们发现了一些有趣的结论。Raya, L.、Diaz, F.、Sanchez, A.这三个人之间的亲密度十分的高，在数据中分列前三名。Archambault, D.和 Scheidegger, C.之间，有比较高的亲密度。

三：可视化：

我们使用 d3 的力导向图进行可视化，用 html 网页进行显示，只进行了前 600 个节点的可视化，代表近几年发表论文的作者。

我们进行了作者网络及作者网络中的最短路，介数中心度，紧密中心度与强连通分量，还有圈子的可视化。并且，我们显示了最短路的具体路径，介数中心度最大的五个作

者，同一个强连通分量里紧密中心度最小的五个作者，以及圈子的具体作者。

四：具体贡献：

沈冠霖完成了基本数据处理（csv 的读入和 json 的输出），介数中心度算法，最短路算法，还有一部分可视化

孙骏博完成了圈子算法，一部分可视化和网页 html 交互部分

五：引用：

介数中心度：http://algo.uni-konstanz.de/publications/b-fabc-01.pdf

D3：www.ourd3js.com/wordpress/555/

https://github.com/d3/d3/wiki/%E5%8A%9B%E5%B8%83%E5%B1%80

https://bl.ocks.org/mbostock/4062045

感谢汪元标同学，傅禹泽同学的帮助