实验结果.md

# 四、实验结果

## 4.1 数据集

### 4.1.1 数据集描述

我们的原始数据集是包含 UserId、phone*number1、phone*number2、email... 等总共15个字段节点的自然人图，若自然人图中某一字段的信息为空，则将其值设置为"null"。自然人图的拓扑结构假设按照图4.1的拓扑结构事先设定。

图4.1

自然人图的节点总数为 15， 按照事先设定的拓扑来进行边的构造，总边数为{边数}，其中，每个节点都包含了该节点的值，例如，phone\_number1 的值为 13960888959，表示为自然人的手机号码，为了方便数据的处理，我们在实验中统一将节点值的类型设置成字符串类型。

### 4.1.2 数据集子图划分

数据集划分的目的是将原先自然人图划分成若干个连通子图，把划分好的连通子图自然人的数据分发到各个参与方，后续协调端将会聚合各个参与方之间的子图，把分散在参与方之间的自然人子图聚合成一个包含所有自然人完整信息的大图，等待来自查询方的查询。

## 4.2 算法实验

### 4.2.1 Metircs (评价指标)

1. 时间消耗

* // todo：这里可以贴一点公式和定义什么的，但是我不知道该怎么找

1. 准确率

* //todo：同上

### 4.2.2 算法实验

我们的算法FedQuery采用邻接链表的数据结构表示自然人图，假设参与方 A 拥有自然人图 G1， 参与方 B 拥有自然人图 G2，服务器通过合并邻接链表的方法将这两张图合并起来，得到的一个大的连通图 G。若两张图没有相同的属性字段，则两张图无法合并，则只在服务器保存 G1 和 G2 即可。若后续参与方想要和服务器中的图合并，则遍历服务器中的每张图，若可以与之合并，就将其与大图合并，得到一个新的更大的图。

### 4.2.3 对比算法实验

对比算法的假设前提是：在理想条件下自然人的所有信息全部位于在服务器上，即服务器的数据库中拥有所有自然人的信息，对比实验采用暴力查询的方法，即给定查询（p1, p2），判断（p1, p2）是否为同一个自然人，我们将p1的属性字段和p2的属性字段在数据库中对所有自然人进行搜索匹配，查询的时间复杂度为O(N)，其中 N 为数据库中自然人的条目数量。

## 4.4 实验结果

我们得到的实验结果如下，我们在算法耗费的查询时间和算法最终得到的查询准确率进行了详细的测算，结果如图4.2 所示。由于对比算法采用的方法是暴力查询的方法，在服务器拥有全部自然人信息的前提下，准确率毫无疑问是100%。我们可以看到在数据分布在不同参与方的条件下，我们算法FedQuery最终取得的查询准确率和理想条件下数据全部分布在服务器上的查询准确率相差不大。查询时间方面，我们算法FedQuery的查询时间也和理想条件下的查询时间相差不大。