# 基于相似度算法的 facebook 网络数据链路预测

## 一、研究目的

本文通过利用 Facebook 用户好友数据，对原始数据按不同比例划分训练集和测试集，使用基于相似度的算法分别对数据计算局域指标、全局指标、准局域指标的相似性得分，并通过特征曲线下面积（AUC）和精确度（Precision）两个评估指标来评估算法的准确度，从而检测几种链路预测方法的稳定性和敏感性，判断每种相似度计算的适用情况。

## 二、facebook 网络数据的来源与说明

本文所使用的的数据集来源于斯坦福大学的 snap 网络项目中的 ego-Facebook 数据集，这个数据集由 10 个子网络构成，每个子网络代表不同用户 ID 的 Facebook 好友网络。该数据共有 4039 个节点，88234 条边，点从 0 开始计数，为无向网络。

## 三、facebook 网络数据的描述统计分析

为方便理解和计算，分别做出如下处理：（1）将节点计数全部加一，更改为从 1 开始计数；（2）为提高计算效率，降低运算时长，截取节点计数在 100 以内（含 100）的观测，共计 275 条观测。

## （一）单个用户的 facebook 网络数据的描述统计分析

10 个用户 facebook 好友网络的描述统计如下表 1 所示，10 个子网络的平均聚类系数都达到的 0.5 以上，表明用户的好友都大致可划分为几类，而网络密度都在 0.2 以下，较为稀疏，平均度在 6 到 81 之间，可以看做是每个用户的每个好友的好友数。这些描述统计情况表明 10 个子网络符合现实用户的好友相互关注情况。

**表 1 10 个用户的 facebook 网络数据的描述统计**

| 用户 ID | 边数 | 结点数 | 聚类系数 | 平均聚类系数 | 平均度 | 网络密度 |
|---|---|---|---|---|---|---|
| 0 | 2866 | 348 | 0.2827 | 0.6546 | 16.4713 | 0.0475 |
| 107 | 27794 | 1046 | 0.4329 | 0.5756 | 53.1434 | 0.0509 |
| 348 | 3419 | 228 | 0.4554 | 0.6162 | 29.9912 | 0.1321 |
| 414 | 1852 | 160 | 0.5658 | 0.6862 | 23.1500 | 0.1456 |
| 686 | 1826 | 171 | 0.4101 | 0.6230 | 21.3567 | 0.1256 |
| 698 | 336 | 67 | 0.4942 | 0.7504 | 10.0299 | 0.1520 |
| 1684 | 14816 | 793 | 0.3544 | 0.5365 | 37.3670 | 0.0472 |
| 1912 | 30780 | 756 | 0.6646 | 0.6624 | 81.4286 | 0.1079 |
| 3437 | 5360 | 548 | 0.2581 | 0.6237 | 19.5620 | 0.0358 |
| 3980 | 205 | 60 | 0.3016 | 0.6557 | 6.8333 | 0.1158 |

图 1 展示了 4 个用户的好友网络图，结点均表示用户，连线表明用户间存在好友关系，从图 1 可以看到 4 种不同的网络特征。ID 为 107 的用户的好友数众多，并且关系较为紧密；ID 为 689 的用户好友数少且关系松散；ID 为 348 的用户好友数不多，联系不算紧密；ID 为 414 的用户好友明显可划分为 3 个社区。

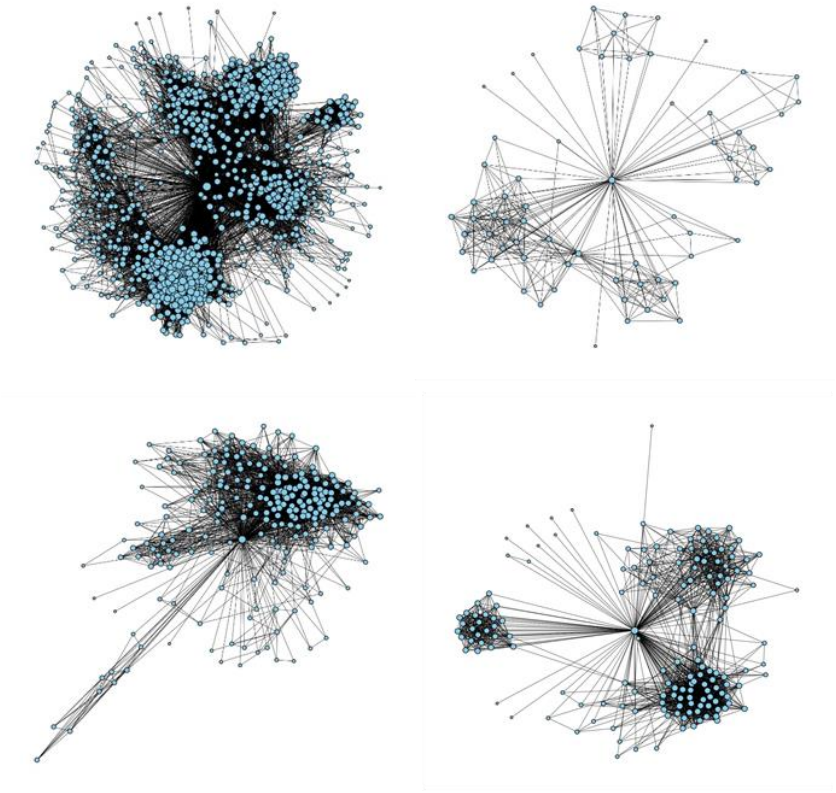

图 1 ID 为 107（左上）、689（右上）、348（左下）、414（右下）的用户的 facebook 网络图

## （二）合并后的 facebook 网络数据的描述统计分析

10 个用户合并后的网络图如图 2 所示，每个点都代表一个用户，中心点为多人关注的用户，连线表明用户间存在好友关系，未产生连接的点为不登陆的用户或者是与现选用户不联系的用户。从图 2 中可以发现有 7 个比较密集的社区。

表 2 为合并后的网络的描述统计情况，合并后的网络有 4039 个节点数，说明数据中一共涉及 4039 个用户，边数为 88243 条，平均度为 43.69，说明每个用户的平均好友数为 43.69 个。聚类系数为 0.6055，聚类情况比较明显，类之间朋友的朋友有很大可能性是朋友。

表 2 合并的 facebook 网络数据的描述统计

| 边数 | 结点数 | 聚类系数 | 平均聚类系数 | 平均度 | 网络密度 |
|---|---|---|---|---|---|
| 88234 | 4039 | 0.5191 | 0.605547 | 43.691013 | 0.01082 |

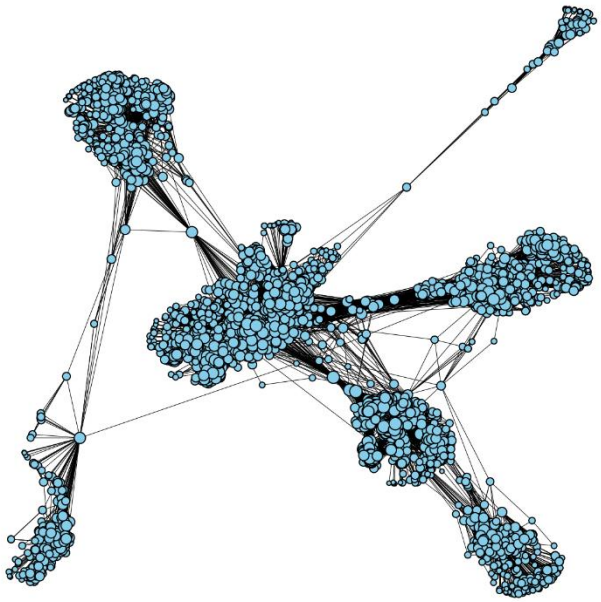

图 2 合并后的 facebook 网络数据的网络图

## 四、基于相似度算法的指标比较

### （一）度量标准的选择

衡量链路预测算法精度的指标主要有 AUC 和 Precision。它们对预测精度的衡量的侧重点不同：AUC 从整体上衡量算法的精确度，Precision 只考虑对排在前 L 位的边是否预测准确。

AUC 可以理解为在测试集中的边的分数值有比随机选择的一个不存在的边的分数值高的概率，也就是说，每次随机从测试集中选取一条边与随机选择的不存在的边进行比较，如果测试集中的边的分数值大于不存在的边的分数值，就加 1 分；如果两个分数值相等，就加 0.5 分。独立地比较 n 次，如果有$n'$次测试集中的边的分数值大于不存在的边的分数，有$n''$次两分数相等，则 AUC 定义为：

$$AUC = \frac{n' + 0.5n''}{n}$$

显然，如果所有分数都是随机产生的，AUC=0.5。因此 AUC 大于 0.5 的程度

衡量了算法在多大程度上比随机选择的方法准确。

Precision 定义为在前 L 个预测边中被预测准确的比例。如果有 m 个预测准确，即排在前 L 的边中有 m 个在测试集中，则 Precision 定义为：

$$\text{Precision} = \frac{m}{L}$$

显然，Precision 越大预测越准确（本例中，L 选择 10）。

**（二）三种类型指标的对比**

1.数据预处理

本文采用合并后的 Facebook 数据来对三种类型的指标进行对比，首先将数据分别按 0.9,0.8,0.6 的比例切分训练集与测试集，这是为了测试几种链路预测方法的稳定性和敏感性，判断每种相似度计算的适用情况。再将训练集和测试集数据变为邻接矩阵的形式，然后分别计算出数据中未存在的链接集合和训练集中观测的链接集合。

2.基于相似度算法的三类指标

本次链路预测的基于相似度算法的指标比较中一共采用了 8 种方法。

其中，局域指标 5 个，共同邻居(Common Neighbor, CN)，索尔顿(Salton)指标（也叫余弦相似性）、雅卡尔(Jaccard)指标、索伦森(Sorenson)指标、AA 指标、RA 指标(资源分配指数)。

全域指标有 2 个，分别为基于网络随机游走过程的指标中的平均通勤时间(Average Commute Time, ACT)、有重启的随机游走(Random Walk with Restart, RWR)。

准局域指标只使用了局部路径指标(Local Path, LP)。

3.以 AUC 为度量标准的指标对比

表 3 为三类指标共 8 个指标在数据不同划分比例下的 AUC 值。图 3 为表 3 的可视化，其中实线为局域指标，长虚线为全域指标，短虚线为准局域指标。可以看出，训练集比例越大，整体上指标的 AUC 越好。局域指标中， Sorenson 指标在数据的 3 个不同划分下 AUC 值最大，局域指标之间的差异性较大；在全域指标中，RWR 指标表现最好，与 Sorenson 指标不相上下，而 Karz 指标的表现则差强人意;准局域指标 LP 指标表现不错,仅次于 Sorenson 指标和 RWR 指标。

**表 3 以 AUC 为度量标准的指标比较**

| 指标类型 | 相似度指标 | 划分比例/AUC | | |
|---|---|---|---|---|
| | | 0.9 | 0.8 | 0.6 |
| 局域指标 | CN | 0.992 | 0.872 | 0.843 |
| | Jaccavrd | 0.99 | 0.81 | 0.723 |
| | Sorenson | 0.993 | 0.974 | 0.951 |
| | AA | 0.99 | 0.924 | 0.831 |
| | RA | 0.834 | 0.821 | 0.807 |
| 全域指标 | Katz | 0.607 | 0.591 | 0.542 |
| | RWR | 0.992 | 0.967 | 0.937 |
| 准局域指标 | LP | 0.961 | 0.946 | 0.928 |

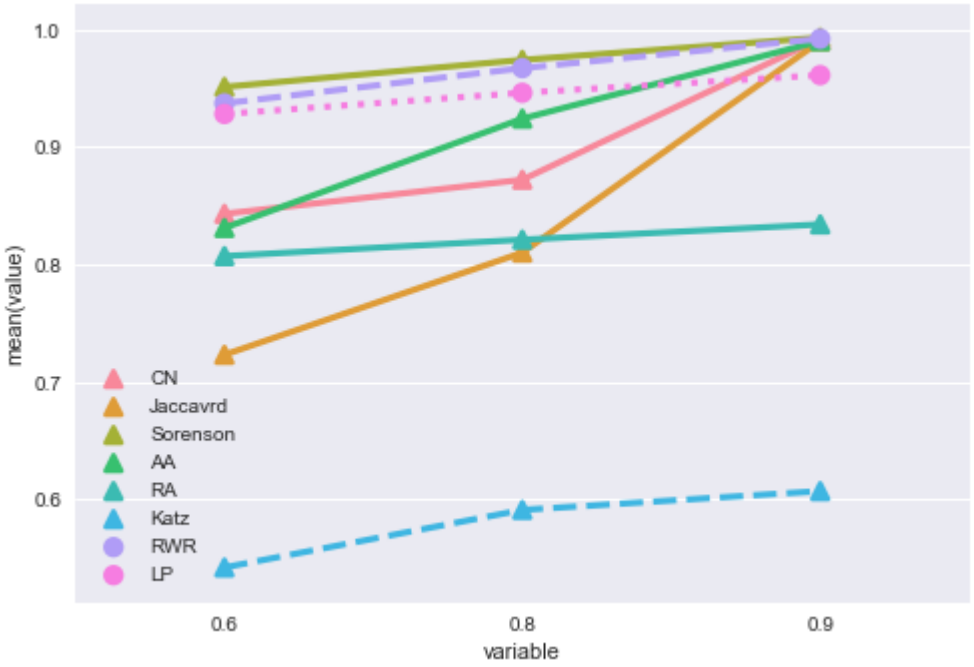


**图 3 以 AUC 为度量标准的指标对比折线图**

4.以 precision 为度量标准的指标对比

表 4 为三类指标共 8 个指标在数据不同划分比例下的 precision 值。图 4 为表 4 的可视化，其中实线为局域指标，长虚线为全域指标，短虚线为准局域指标。可以看出，训练集比例越大，整体上指标的 precision 值越低。局域指标中，Sorenson 指标在数据的 3 个不同划分下 precision 值最大，局域指标之间的差异性较大，RA 指标表现最糟糕；在全域指标中，RWR 指标与 Karz 指标在训练集比例为 0.8 时差异较大，而比例为 0.6 和 0.9 时较为相似；准局域指标 LP 指标表现不错，与 Sorenson 指标不相上下。

**表 4 以 precision 为度量标准的指标比较**

| 指标类型 | 相似度指标 | 划分比例/precision | | |
|---|---|---|---|---|
| | | 0.9 | 0.8 | 0.6 |
| 局域指标 | CN | 0.592 | 0.635 | 0.642 |
| | Jaccavrd | 0.497 | 0.657 | 0.724 |
| | Sorenson | 0.611 | 0.744 | 0.756 |
| | AA | 0.473 | 0.65 | 0.703 |
| | RA | 0.038 | 0.125 | 0.225 |
| 全域指标 | Katz | 0.276 | 0.542 | 0.554 |

| | | | | |
|---|---|---|---|---|
| | RWR | 0.295 | 0.336 | 0.542 |
| 准局域指标 | LP | 0.548 | 0.728 | 0.775 |

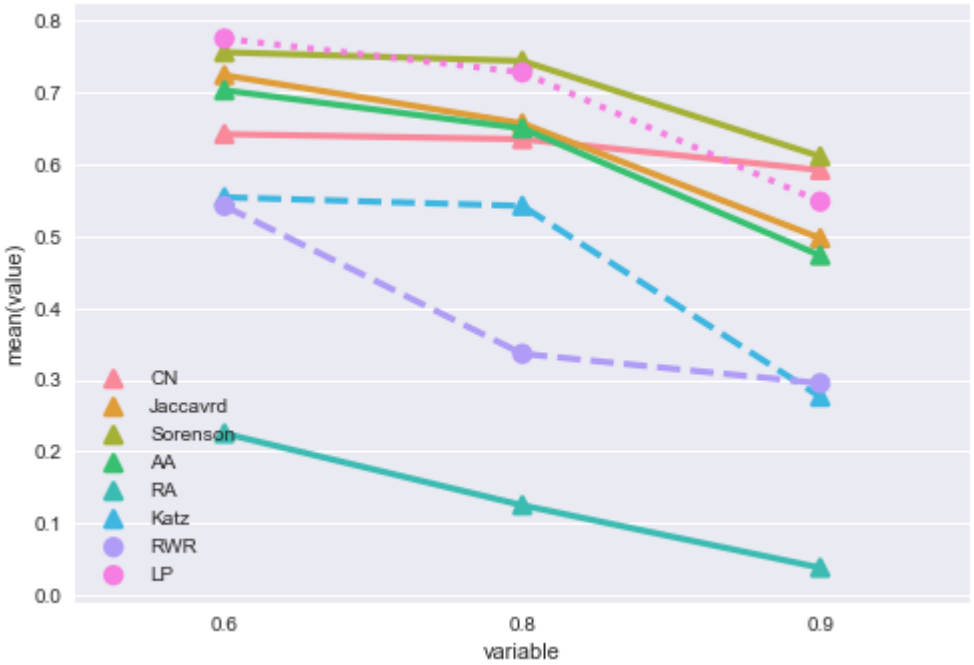


图 4　以 precision 为度量标准的指标对比折线图

## 五、结论

对于合并后 Facebook 的网络数据，局域指标中的 Sorenson 指标预测确度上的表现和准局域指标中的 LP 指标的表现均不错，且随着划分比例变化而改变较小，稳定性好。