对加州大学尔湾分校的社交网络分析报告

```
knitr::opts_chunk$set(echo = TRUE,
                      comment = "")
```

摘要

本报告的主要目的是建立社交网络结构，并再此基础上进行分析，提取不同群体结构特征，个体特征以及用户中信息传播方式。分析结果显示，整个网络中的重要成员是编号为105、103、32和9的成员，这些成员无论是在度中心性还是接近中心性上都排在全部节点的前5。另外校园网站的社交网络中存在明显的社群划分，大的社群特征和小的社群特征之间存在较为明显的差异。因此，用户之间的影响力和信息传播方式也根据网络结构的不同而不同。

1. 背景介绍

近几十年来，”网络科学“已经发展成为一门繁荣的社会科学研究领域。而随着互联网技术的蓬勃发展，各种各样的社交网络工具也在最近几年呈现出爆发式的增长。在大学生群体中，除了大众社交工具之外还有一个重要的网络交流工具便是”校园网站（论坛）“。如果能够有效的提取社交网络中的各种数据并对用户行为、群体特征、用户间的信息传播等进行分析，掌握用户的行为模式和社交网络中的信息传播模式，不仅能帮助网站运营商全面掌握用户需求从而提供更好的服务，还能为有关部门提供对网络舆情的合理监管和干预提供理论依据。

本案例报告将介绍：建立社交网络结构，并再此基础上进行分析，提取不同群体结构特征，个体特征以及用户中信息传播方式等。

该数据集由加州大学尔湾分校社交网络上发送的个人消息组成。用户可以在网络上搜索其他人，然后根据个人信息发起对话。通过社交网络分析可以初步知道出谁是这个群体的实际控制者，谁是这些成员中有影响力的人，哪些成员更倾向于聚集在一起。

2. 数据描述

数据来源：本报告所用的数据都来自公开数据库（http://snap.stanford.edu/data/CollegeMsg.html），一共1899个节点，59835条边，时间跨度为193天。

数据说明：一个节点表示一个用户，一条边（u,v,t）意味着用户u在某时刻t向v发送了一条私人消息。

变量表：

变量名	含义
SRC	用户节点的id
TGT	用户发消息的目标对象id
UNIXTS	发送消息的时间戳

3. 网络的基本描述

3.1 数据预处理

```
# 清除工作环境
rm(list = ls())
# 加载必要包
library(igraph)
library(plyr)

data = read.table("../A2_data/final/CollegeMsg.txt",col.names =
c("SRC","DST","UNIXTS") )

data["SRC1"] = apply(data[,c(1,2)], 1, min)
data["DST1"] = apply(data[,c(1,2)], 1, max)
g.data = ddply(data,.(SRC1,DST1),nrow)

# 构建网络
g <- graph_from_data_frame(g.data,directed = F)

# 迭代删除节点度小于10的节点
print(paste("删除之前网络的节点数为：",length(V(g))))
while (any(degree(g) <10)) {
        g = delete.vertices(g,V(g)[degree(g)<10])
}
# 计算节点数、边数
print(paste("删除之后网络的节点数为：",length(V(g))))
print(paste("删除之后网络的边数为：",length(E(g))))
print(paste("网络的密度为：",graph.density(g)))
```

```
[1] "删除之前网络的节点数为： 1899"
[1] "删除之后网络的节点数为： 659"
[1] "删除之后网络的边数为： 9740"
[1] "网络的密度为： 0.0449239199118126"
```

通过观察后对数据进行简单的预处理：由于A给B发送消息与B给A发送消息都等价于A,B之间存在交流关系，而本案例关注的是用户之间的甲流关系已经社群特征，所以将相同成员之间（A给B发送消息与B给A发送消息都看作相同两个成员）在不同时间发送消息的边合并为一条边，将累积发送次数设置为边权重。因此本网络结构考虑为无向有权网络。

除此之外，为了更细致地展示校园网站成员的社区结构，本案例通过提取所以成员社交网络的核心网络：通过不断删除网络中度小于10的节点直到网络不再变化。最后得到一个由659个节点，9740条边构成的校园网站社交网络。且该社交网络的密度为：0.0449。

3.2 网络的基本描述

3.2.1 节点度

```
# 计算度
degree.all = degree(g,mode = "all")
# 绘制直方图

par(mar=c(9,4,2,4))
hist(degree.all,xlab = "节点的度",ylab = "频数",main = "")
title("图1  节点度分布直方图", line = -22, outer = TRUE)
```

根据度的分布直方图可以发现，该网络结构的度呈现严重右偏，即大部分节点的度都较小，只有少部分的节点度值较大，其中大部分节点的度都小于50，只有少量节点的度大于50，甚至个别节点的度达到150-200。结合实际情况，校园网站社交网络中大部分成员的社交圈数量都处于较低水平（<50），只有少部分成员的社交圈朋友数量较高（>150）。因此，很可能这些节点对应的成员便是该社交网络中的重要节点。

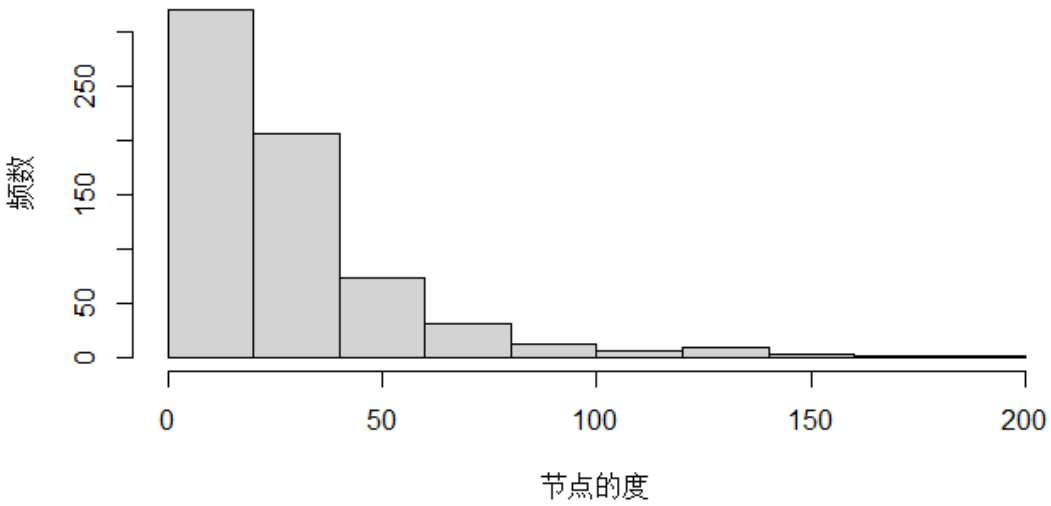


图1 节点度分布直方图

3.3 中心性指标

3.3.1 度中心性

```
degree.center = sort(degree(g,mode = "all",normalized = T),decreasing = T)
round(degree.center,6)[1:10]
```

```
     103      105       32        9      249      400      194        3
0.281155 0.264438 0.232523 0.229483 0.212766 0.208207 0.199088 0.197568
     638     1283
0.196049 0.196049
```

计算所有成员的度中心性，从整体上看，排名前十的成员其度中心性几乎都接近0.2或大于0.2。在排名前10的成员之间，差距相对较小，其中度中心性排名第一的是编号103的成员（度中心性：0.281），其次为编号105（度中心性：0.264）和编号32的成员（度中心性：0.233）。

度中心性衡量的是节点对促进网络中传播过程发挥的作用。在本案例中是一种识别社交网络中最"重要"成员的指标，说明编号103的成员在其他节点之间充当了重要的"桥梁"，其次是编号105的成员。

根据度中心性度量指标，标号为103、105、32、9、249、400、193、3、638和1283的成员比较重要，说明这些成员与所有其它成员相联系的程度较高，在网络中的参与度高。

3.3.2 接近中心性

```
closeness.center = sort(closeness(g,mode = "all",normalized = T),decreasing = T)
round(closeness.center,6)[1:10]
```

```
     105      103       32        9        3      194      249      638 
0.574672 0.567241 0.564807 0.563356 0.553872 0.552941 0.552477 0.548333 
    1283       42 
0.547877 0.546058 
```

计算所有成员的接近中心性，从整体上看，排名前十的成员其度中心性几乎都大于0.54。在排名前10的成员之间差距相对较小，其中接近中心性排名第一的是编号105的成员（度中心性：0.575），其次为编号103（度中心性：0.567）和编号32的成员（度中心性：0.565）。

根据接近中心性度量指标，标号为105、103、32、9、3、194、249、638、1283和42的成员比较重要，说明这些节点与其他节点的"亲密"程度高。

综合上述两个评价指标，编号为105、103、32和9的节点在两个个指标中都排进了前5，是重要节点。

3.4 可视化

```
set.seed(100)
plot(g,layout = layout.fruchterman.reingold,
     vertex.size = degree(g)*0.1,
     vertex.shape = 'circle',
     vertex.color = 'gold',
     vertex.label = "" ,
     edge.label = "",
     edge.width=0.1,
     asp=-2,margin=-0.05)
title("图2  全节点的网络结构图", line = -22, outer = TRUE)
```

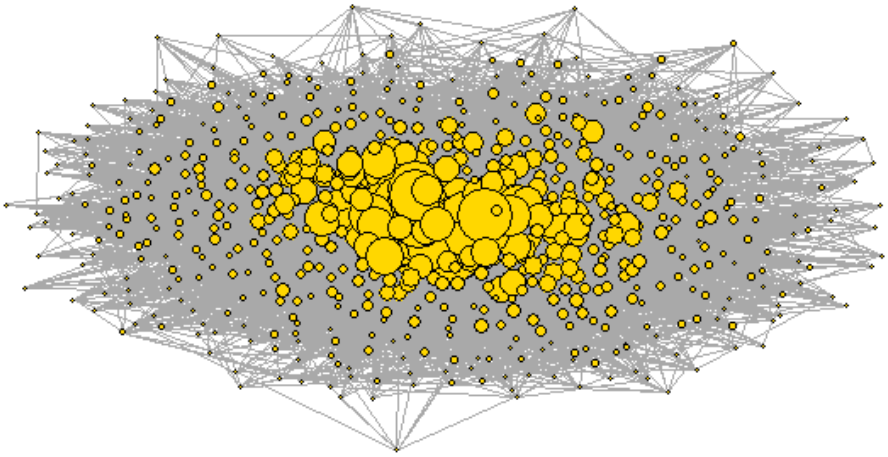


图2 全节点的网络结构图

绘制全部节点的网络结构图，节点的度通过节点的大小呈现。可以发现网络中呈现明显的聚集趋势，社交网络中节点度较高的只集中在中心较少的节点，而大部分节点都较小的分布在周围。因此进一步可以通过聚类的方式对该社交网络进行社区发现，进一步探索社群特征。

4. 网络的社区发现

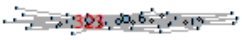
4.1 社区发现

```
mc <- multilevel.community(g,weights = E(g)$V1)
sizes(mc)
op = par(no.readonly = T)
par(mfrow=c(4,3))
for(node in groups(mc)){
    nodes = as.vector(unlist(node))
    g.sub = induced.subgraph(g,nodes)

    g.sub_order = degree(g.sub)[order(degree(g.sub),decreasing = T)][1]
    g.sub_ordername = names(g.sub_order)
    V(g.sub)[g.sub_ordername]$label = g.sub_ordername
    V(g.sub)[g.sub_ordername]$label.color = "red"

    plot(g.sub,layout=layout.fruchterman.reingold,
     vertex.size = degree(g.sub)*0.5,
     vertex.shape = 'circle',
     vertex.color = 'skyblue3',
     edge.width=0.3,
     asp=-2,
     margin=0,
     sub = paste0("子网络密度：",round(graph.density(g.sub),4)))
}
par(op)
```

子网络密度：0.1002

子网络密度：0.1049

子网络密度：0.0976



子网络密度：0.1233

子网络密度：0.0917

子网络密度：0.1402

子网络密度：0.1106

子网络密度：0.0882

子网络密度：0.1127

子网络密度：0.1195

子网络密度：0.0803

```
Community sizes
 1  2  3  4  5  6  7  8  9 10 11
70 61 78 49 99 28 63 50 50 41 70
```

关于社区发现的算法较多，例如点连接、随机游走、自旋玻璃、中间中心度、标签传播等。在R语言的igraph包中也提供了大量的函数进行社区发现。算法的区别与对比不是本案例的重点，故本案例仅采用了multilevel.community()函数进行社区发现。**（由于导出图片的格式问题，组图长宽比例被压缩，且像素点降低）**

结果如上，一共发现了11个社区，社区节点数最大的为第5个社区（99个节点），最小的为第6个社区（28个节点），同时，第6社区也是唯一一个节点数小于40的社区。大部分社区的节点数都集中在50-90之间。分析原因，节点数超过90的节点极大可能包含了重要节点，且可能该社群在学校社交中承当着一定地位，可能是学生社团或学生会群体。该社区对整个校园网社交网络的结构特征和交流传播方式起到关键作用。因此，对该社群的分析至关重要。另外，全网络的最小社区仅由28个节点，该社区的成员可能是学校的最低一级的组织单位，或许是一个班级也获取是一个社团或者学生会部门。因此，通过进一步分析该社交网络中的最大社区和最小社区即可了解到该社交网络中的大部分重要特征。

另外一方面，通过计算各个子网络的密度，密度最大的是第6个社区（网络密度：0.1402，节点数：28），其次为第4社区（网络密度：0.1233，节点数：49），密度最小的是第11社区（网络密度：0.0803，节点数：70）。可以发现，节点最小的社区反而网络密度最高，而节点较高的社区，网络密度反而较低。说明虽然在一个大的社区中，但大部分社区成员的交流圈并没有变大。而在一个小的社区中的成员反而能够有广泛的交流。

4.2 最大社群

```
max_index = which.max(lengths(groups(mc)))
subg.maxnodes = groups(mc)[max_index]
subg.maxnode = as.vector(unlist(subg.maxnodes))
subg.max = induced.subgraph(g,subg.maxnode)
```

4.2.1 权重直方图

```
par(mar=c(9,4,2,4))
hist(E(subg.max)$V1,breaks = 50,xlab = "权重",ylab = "频数",main="")
title("图3  边权重分布直方图", line = -22, outer = TRUE)
```

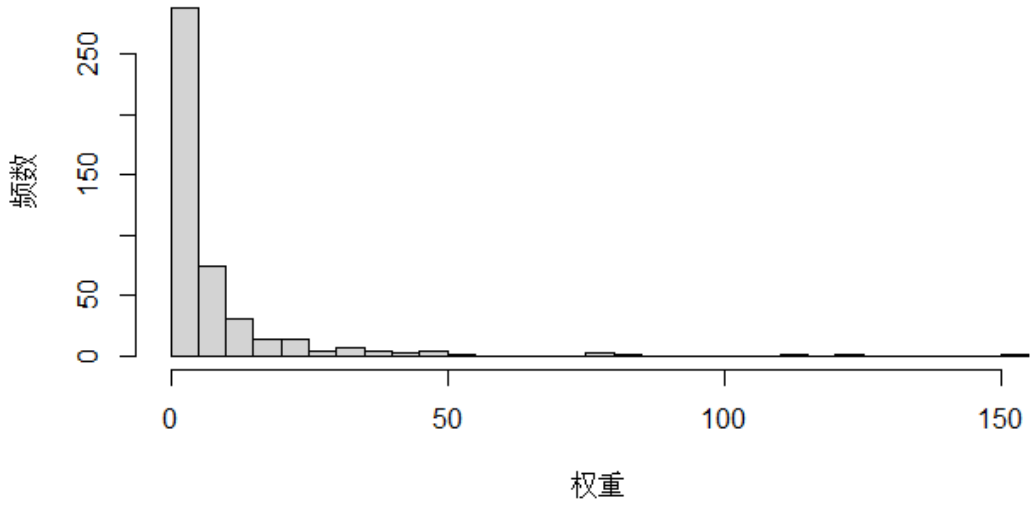


图3 边权重分布直方图

报告的前文已经说明，边的权重表明在时间跨度（193天）期间，两个成员之间累积发短信次数，也即交流次数。通过绘制边权重直方图可以发现，最大社区的大部分成员之间的交流次数都为1-2次，整个社区的主要交流次数都小于20次。但是也存在交流次数大于100次的，甚至大于150次，几乎达到平均一天一次的频率。当然这种情况有多种可能性，例如情侣之间的交流，或者网站重要管理员，亦或是学生会社团的管理者之间的交流。毕竟某个组织相关的工作交流存在每天都需要沟通的情况。因此，对于这些高交流频率的用户可能便是社交网络的重要节点。

4.2.2 可视化

```
set.seed(100)
g.sub_ordername = names(degree(subg.max)[order(degree(subg.max),decreasing = T)]
[1])
#设置节点属性
V(subg.max)$size = degree(subg.max)*0.2
V(subg.max)$color = 'gold'
V(subg.max)$shape = 'circle'
V(subg.max)$label.cex = 0.8
V(subg.max)$label.dist = 0.5
```

```
V(subg.max)$label.color = 'grey2'
V(subg.max)[g.sub_ordername]$color = "blue"
V(subg.max)[g.sub_ordername]$label.color = "green"

#设置连边属性
E(subg.max)$lty = 2
E(subg.max)$label.cex = 0.8
E(subg.max)$color = "grey"
E(subg.max)[E(subg.max)$V1 == max(E(subg.max)$V1)]$lty = 1
E(subg.max)[E(subg.max)$V1 == max(E(subg.max)$V1)]$color = "grey3"
E(subg.max)[order(E(subg.max)$V1,decreasing = T)[1:5]]$label =   E(subg.max)
[order(E(subg.max)$V1,decreasing = T)[1:5]]$V1
E(subg.max)$label.color = 'red'

# 绘制网络图，图布局选取力导向布局
plot(subg.max,layout=layout.fruchterman.reingold,
      edge.width=0.3,asp=-2,margin=-0.05)
title("图4  最大社区的网络结构图", line = -21, outer = TRUE)
```

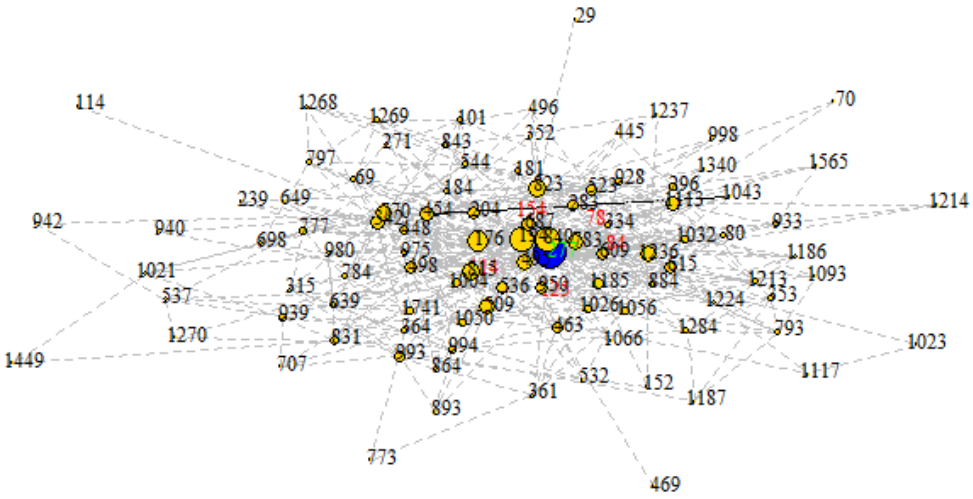


图4 最大社区的网络结构图

绘制最大社区的网络结构图，节点度通过节点的大小体现，其中最大的节点已经标记为蓝色节点（编号277），而权重最大的边用黑色实线标记出来了（权重：154），红色标签为排名前5的权重。可以发现，权重最大的边并不是度最大的节点，但是排名前5的权重都是在度较大的节点附近。可以得到编号277这个成员即为该社区的重要节点。

4.2.3 子网络社区发现

```
mc.max <- multilevel.community(subg.max,weights = E(subg.max)$V1)
sizes(mc.max)
plot(mc.max,subg.max,layout=layout.fruchterman.reingold,
     edge.width=0.3,asp=-2,margin=-0.05)
title("图5  最大社区的社群发现", line = -21, outer = TRUE)
```

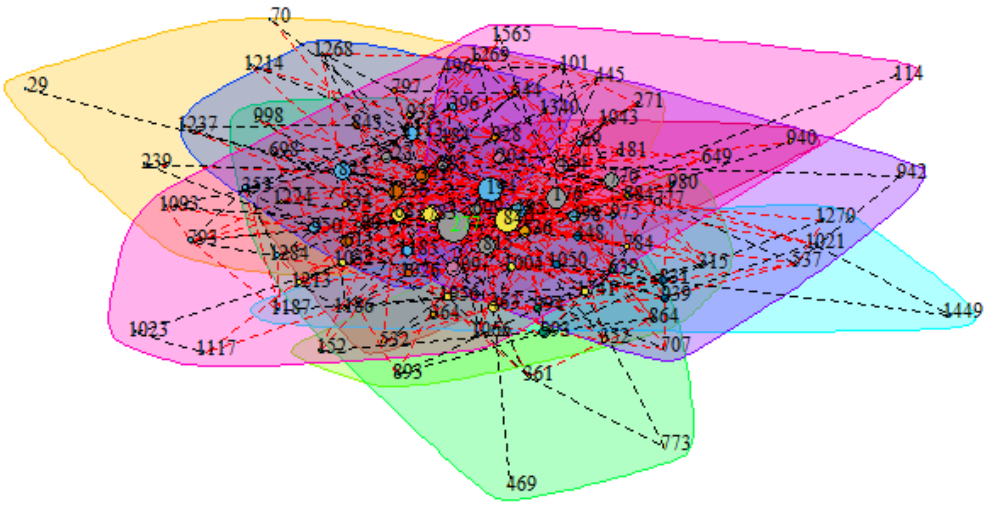


图5 最大社区的社群发现

```
Community sizes
 1  2  3  4  5  6  7  8
 8 17  8 17  8 10 10 21
```

通过对最大社区的社区发现，可以将该社区进一步划分为8个社区。但是社区之间存在相互交叠的现象，说明两两成员之间相互交流的情况较为常见。而且度较大的几个节点几乎被所有社区重叠。通过进一步的社区发现可视化可以得出，最大社区中的度节点较大的几个成员便是该社交网络的重要成员。

4.3 最小社群

```
min_index = which.min(lengths(groups(mc)))
subg.minnodes = groups(mc)[min_index]
subg.minnode = as.vector(unlist(subg.minnodes))
subg.min = induced.subgraph(g,subg.minnode)
```

4.3.1 权重直方图

```
par(mar=c(9,4,2,4))
hist(E(subg.min)$V1,breaks = 50,xlab = "权重",ylab = "频数",main="")
title("图6  边权重分布直方图", line = -22, outer = TRUE)
```

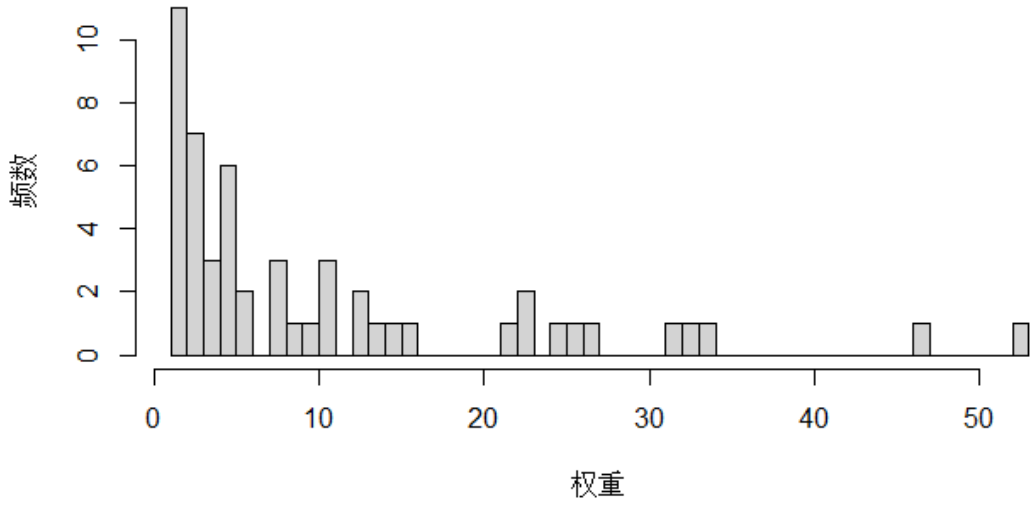


图6 边权重分布直方图

同样的，对最小社群进行相同处理，通过绘制边权重直方图可以发现，最小社区的主要交流次数都小于10次。但是在大于10次的分布情况上，明显要高于最大社区的情况。进一步说明在小社群中虽然交流的总次数没有最大社群多，但是密度明显高于最大社群。

4.3.2 可视化

```
set.seed(100)
#设置节点属性
V(subg.min)$size = degree(subg.min)*0.8
V(subg.min)$color = 'gold'
V(subg.min)$shape = 'circle'
V(subg.min)$label.cex = 0.8
V(subg.min)$label.dist = 0.5
V(subg.min)$label.color = 'grey2'
V(subg.min)[max(degree(subg.min))]$color = "blue"
V(subg.min)[max(degree(subg.min))]$label.color = "blue"

#设置连边属性
E(subg.min)$lty = 2
E(subg.min)$label.cex = 0.8
E(subg.min)$color = "grey"
E(subg.min)[E(subg.min)$V1 == max(E(subg.min)$V1)]$lty = 1
E(subg.min)[E(subg.min)$V1 == max(E(subg.min)$V1)]$color = "grey3"

E(subg.min)[order(E(subg.min)$V1,decreasing = T)[1:5]]$label = E(subg.min)
[order(E(subg.min)$V1,decreasing = T)[1:5]]$V1
E(subg.min)$label.color = 'red'

# 绘制网络图，图布局选取力导向布局
plot(subg.min,layout=layout.fruchterman.reingold,
     edge.width=0.3,asp=-2,margin=-0.05)
title("图7  最小社区的网络结构图", line = -22, outer = TRUE)
```

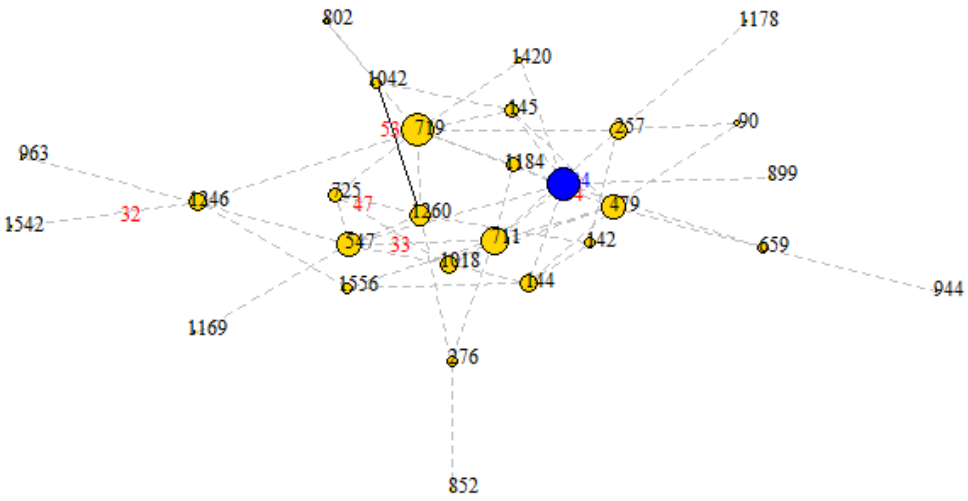


图7 最小社区的网络结构图

按照最大社群的可视化方式，对最小社群进行同样的可视化。最大节点为编号644的节点，边权重最大的达到了53次。但是改变对应的节点度相对较小，所以推测该节点对应的成员之间可能是情侣。另外最小社区和最大社区的另一个明显不同点在于：排名前五的边权重零散的分布在网络图的各处，并不是集中在大节点附近。

4.3.3 子网络社区发现

```
mc.min <- multilevel.community(subg.min,weights = E(subg.min)$V1)
sizes(mc.min)

plot(mc.min,subg.min,layout=layout.fruchterman.reingold,
     edge.width=0.3,asp=-2,margin=-0.05)
title("图8  最小社区的社群发现", line = -22, outer = TRUE)
```

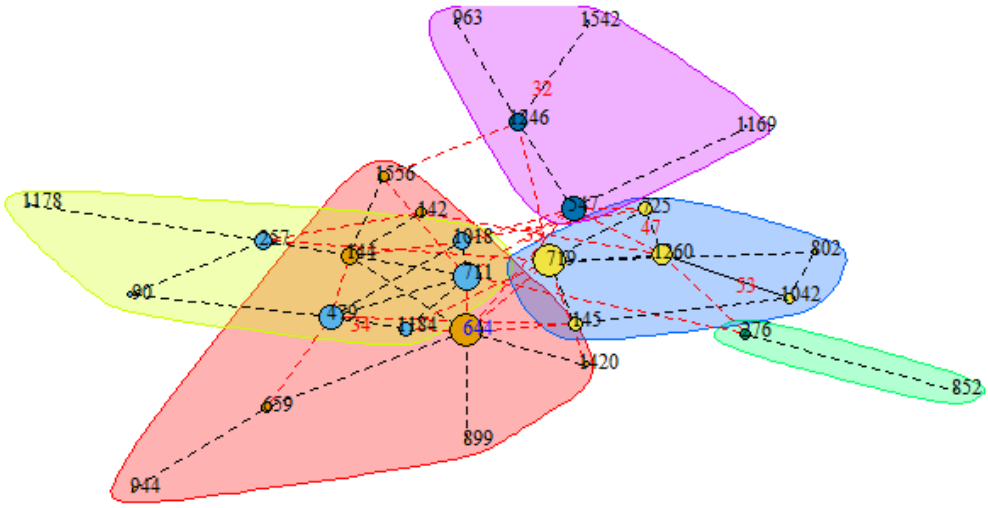


图8 最小社区的社群发现

```
Community sizes
1 2 3 4 5
8 7 2 6 5
```

通过对最小社区的社区发现，可以将该社区进一步划分为5个社区。社区的划分边界之间较为清晰，最小社区成员仅有两个成员节点，这两个节点在最小社区中仅进行了相互交流。且不同子社区之间的交流较少，主要集中在子社区内部之间的交流。这样的结果也与实际情况符合，一个班级或者社团之间内部的交流明显是要强于与外部的交流频率。

5. 结论与建议

随着在线社交网络的发展和线上用户的急剧增长，以交友、信息共享等为目的的社交网络迅速成长为人们传播信息、推销商品、表述观点、产生影响力的理想平台。在线社交网络中的影响力分析和建模是社交网络分析的重要内容, 通过分析人们相互之间的影响模式和影响力传播方式，能够从社会学角度加深对人们社会行为的理解。

本案例通过分析加州大学尔湾校区校园网社交网络，首先对整个社交网络结构进行了简单描述和可视化分析，然后通过社区发现11个不同的子社区，进一步选择分析最大的子社区和最小的子社区，找到了该社交网络的基本特征。本报告的主要结论归纳如下：

根据分析的结果，整个网络中的重要成员是编号为105、103、32和9的成员，这些成员无论是在度中心性还是接近中心性上都排在全部节点的前5。另外校园网站的社交网络中存在明显的社群划分，大的社群特征和小的社群特征之间存在较为明显的差异。不同的社群之间的交流关系也根据社群大小而不同。本案例在数据预处理时，通过删除边缘节点来提取核心节点，因为在社交网络节点数量众多情况下，用户

之间形成的关系也非常复杂，因此在这方面的分析很难厘清影响力和其它因素之间的关系，因此还存在较大的改善空间。

参考文献

> [1]尹雅丽.社交网络数据采集方法研究及社团结构分析[J].现代计算机(专业版),2016(08):31-34.

> [2]许进,杨扬,蒋飞,金舒原.社交网络结构特性分析及建模研究进展[J].中国科学院院刊,2015,30(02):216-228.

> [3]吴信东,李毅,李磊.在线社交网络影响力分析[J].计算机学报,2014,37(04):735-752.

> [4]刘晓曼. 社交网络数据获取与结构分析系统的设计与实现[D].安徽大学,2014.

> [5]郭琛. 社交网络分析与信息传播研究[D].复旦大学,2012.

> [6]Pietro Panzarasa, Tore Opsahl, and Kathleen M. Carley. "Patterns and dynamics of users' behavior and interaction: Network analysis of an online community." Journal of the American Society for Information Science and Technology 60.5 (2009): 911-932.