1 网络的建立

一个以球员为节点建立的拓扑结构, 除了点外边也是非常重要的量, 有边相连的两个节点在网络中被看作是相邻的, 而边的长度 d(也就是路径长度) 在本问中表示的是两球员 (三球员) 之间传球的难易程度, 通过给出的比赛数据从提取信息进而通过合理的构造获得相应的距离 l.

定义每一个球员的传球成功率 α_i, 其中 i 为球员编号取值为 1-11. 传球次数为 w_{ij}, 为 i 号球员传给 j 号球员的次数.

对于一般的无权网络 (相当于所有路径等长) 相关的参数如下:

特征路径长度 (characteristic path length): 在网络中, 任选两个节点, 连通这两个节点的最少边数, 定义为这两个节点的路径长度, 网络中所有节点对的路径长度的平均值, 定义为网络的特征路径长度. 这是网络的全局特征.

聚合系数 (clustering coefficient): 假设某个节点有 k 条边，则这 k 条边连接的节点（(k个) 之间最多可能存在的边的条数为 $\frac{k(k-1)}{2}$，用实际存在的边数除以最多可能存在的边数得到的分数值，定义为这个节点的聚合系数. 所有节点的聚合系数的均值定义为网络的聚合系数. 反映了相邻两个人之间存在配合的队员的重合度, 即与该点存在配合的两点同时也存在配合的程度, 聚合系数是网络的局部特征. 聚类系数越多，表示该点也越容易参与组成一个多元结构.

这两个参数都表征了网络的一些特征, 在此问题中可以此衡量队伍的合作. 但是容易知道, 实际问题并不是无权网络. 需要找到合适的权重下面详细阐述.

1.1 简单模型

题目给出的文献”Defining a historic football team: Using Network Science to analyze Guardiola’s F.C. Barcelona” 给出了距离 l 和传球距离的关系，由于通过的网络是加权的 (权为球员之间传球次数)，文章考虑链路的权值的不同，即权值越大，两个节点之间的拓扑距离越短。两个参与人 i 和 j 之间的链路的拓扑长度 l_{ij} 定义为链路权值的倒数，$l_{ij} = \frac{1}{w_{ij}}$. 在此距离意义下, 使用 Dijkstra 算法计算所有参与人对之间的最小最短路径 p_{ij}。定义整个

团队的平均最短路径 d 为:

$$d = \frac{2}{N(N-1)} \sum_{i \neq j} p_{ij}$$

其中 N = 11 为团队总人数。

当对网络进行加权时, 还要考虑链路权重是如何分布的. 对传球网络的情况, 不同球员之间的传球次数不同. 注意到这个问题, 文章使用加权聚类系数 $C_w(i)$ 来衡量与第 i 个球员相连接 (存在配合) 的球员中之间连接的可能性，和参与组成多元结构的可能性.

$$C_w(i) = \frac{\sum\limits_{j \neq k} w_{ij} w_{jk} w_{ik}}{\sum\limits_{j \neq k} w_{ij} w_{ik}}$$

其中 j 和 k 表示球队中的任意两个球员,w_{ij} 和 w_{ik} 表示第三个球员 i 和他们之间的传球次数。最后，通过将所有参与人的 $C_w(i)$ 平均得到整个网络的聚类系数，即 $C = \frac{1}{N} \sum\limits_{i=1}^{N} C_w(i)$. 需要注意的是，聚类系数的加权版本表征了团队在参与者之间形成平衡三角形的趋势，是对局部鲁棒性的一种度量.

1.2 模型改进

事实上用传球次数的倒数衡量拓扑网络的距离是不够细致的, 我们还应该考虑一些具体的因素譬如, 某个球员的传球水平和技术都对拓扑距离有重要的影响. 为了考虑问题先介绍两个球员的传球过程. 两个队员之间传球的过程大致可以分成两部分. 由 A 发起传向 B 的过程,B 带球在场上移动或进行下一步操作. 这样的过程可以运用于整个足球网络.

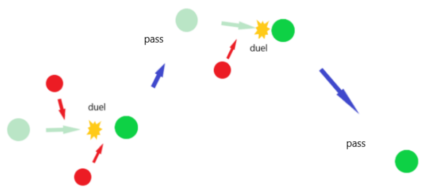

图 1: 两球员之间传球过程演示

1.2.1 pass

前面已经定义过 i,j 队员之间传球的次数 w_{ij}, 相应的在两球员之间应该存在传球的成功率 α_{ij}, 这样就能够计算得到有效传球次数 (可以直接从 passingevents 得到, 而传球的成功率需要从 fullevents 得到) 同时在传球方面需要考虑的因素还有传球队员之间的空间距离, 传球的种类. 下面我们将计算等效的传球次数, 等效的传球次数将包含上面两个因素的信息.

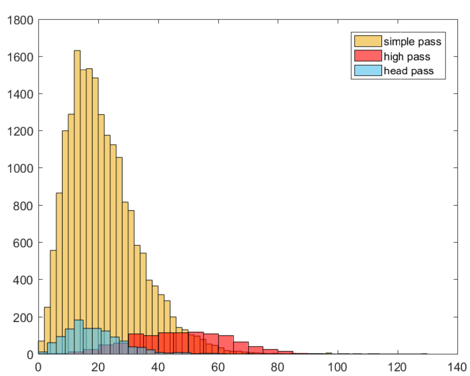

图 2: 三种重要类型分布直方图

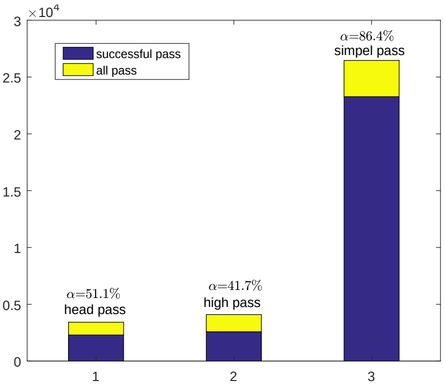

图 3: 三种类型成功率

由于在所有传球信息中只有 simple pass,high pass, head pass 的占比较大, 其余均在 1% 左右，因此在计算等效的传球次数时可以将其他种类的传球忽略. 从分布直方图可以看出 simple pass 的次数最多,head pass 和 high 的次数相近. 但是 high pass 的传球距离最长其次是 simple pass, head pass 最短 (由门将发出). 同时由图 3 的成功率可以发现 simple 类型成功率最高而 high 类型成功率最低. 在等效传球次数时，应该满足某传球的距离越长等效的传球次数越多. 而成功率越低等效的传球相应的就越小. 定义等效的传球次数:

$$\overline{w}_{ij} = \sum_{n=1}^{3} \alpha_{ij}^{(n)} \frac{\overline{L}^{(n)}}{\overline{L}} w_{ij}^{(n)}$$

下面解释其含义:$\overline{w}_{ij}$ 表示 i 与 j 的等效传球次数,$\alpha_{ij}^{(n)}$ 表示不同种类传球的成功率其中 n 取 1，2，3 分别对应 head pass,high pass 和 simple pass, 下同理.$\overline{L}^{(n)}$ 表示某一种类传球的平均距离.$\overline{L}$ 表示所有类型传球的平均距离. 等效的传球次数可以看成是不同种传球次数按照距

离和成功率的加权求和, 这样定义符合上面的要求.

1.2.2 duel

下面考虑一对一动作的影响. 因为某个球员参与一对一的次数越多, 说明在进攻中受到的阻力越大, 其在运动过程中容易失去球权. 而在一对一争斗中成功率又能反应球员的能力. 因此定义第 i 个球员在一对一胜利的几率为 β_i 在上面的一个带球传球过程中可以定义两人间的几率 $\beta_{ij} = \frac{\beta_i+\beta_j}{2}$.

1.2.3 拓扑距离

考虑了 pass 的种类和 duel 后可以给出调整后的拓扑距离

$$l_{ij} = \frac{1-\beta_{ij}}{\overline{w}_{ij}}$$

$1-\beta_{ij}$ 表示在一对一中失败的几率, 失败会导致拓扑距离增长, 同时等效的传球次数使得拓扑距离减少, 这里采用失败几率的原因是防止因子在分母处变化导致 1 的变化剧烈.

1.2.4 相关参数的修改

由于拓扑参数的修改, 导致前面提到的相关的系数也要做出一定程度的修改.

居于距离定义的团队平均最短距离不需要修改, 而聚类系数由于 w_{ij} 传球次数在此不再具有意义, 因而我们从距离出发, 利用原先 w_{ij} 与距离 l_{ij} 的倒数关系将公式中的 w_{ij} 用 $\frac{1}{l_{ij}}$ 代替. 同时 l_{ij} 的计算方法采用 1.2.3 给出的公式计算.

$$C_l(i) = \frac{\sum_{j\neq k} \frac{1}{l_{ij}l_{jk}l_{ik}}}{\sum_{j\neq k} \frac{1}{l_{ij}l_{ik}}}$$