职业资格考试异常代报名行为检测算法-基于图网络半监督学习

薛如冰[1]，郭文星[2]，王纬[3]
1.山东大学，山东 济南 250100
2.埃塞克斯大学，英国
3.山东财经大学，山东 济南 250014

摘　　要： 本研究针对大规模考试数据中的代报名与作弊行为识别问题，提出了一种基于图结构的多阶段综合分析方法。首先，通过Louvain算法对数据构建的加权图进行社区划分，利用最大化模块度Q来发现显著的社区结构，从而缩小分析范围、减少噪声干扰，并聚焦于高风险考生群体。接着，运用图卷积神经网络（GCN）在社区层面进行特征学习与标签预测，有效挖掘节点间的非线性关系与复杂交互，从而弥补传统线性模型与规则分析的不足。最后，通过加权标签传播算法，将预测结果扩散至完整数据集，确保标注信息在全局层面的一致性与覆盖性。此外，通过构建图数据库替代传统关系数据库，避免了多表关联查询的低效性，加速了数据检索与处理过程。该方法有助于在海量、复杂的考试数据中高效识别潜在高风险考生群体，提高考试的公正性与管理效能。

关 键 词： 标签传播；社区划分；图数据库；图卷积神经网络

The Detection Algorithm for Abnormal Proxy Registration in Professional Qualification Examinations: A Semi-Supervised Learning Approach Based on Graph Networks

Xue Rubing[1], Guo Wenxing[2], Wang Wei[3]
1. Shandong University, Jinan, Shandong 250100
2. University of Essex，UK;
3. Shandong University of Finance and Economics, Jinan, Shandong 250014

Abstract: This study addresses the identification of proxy registration and cheating behaviors in large-scale examination data by proposing a multi-stage comprehensive analysis method based on graph structures. First, the Louvain algorithm is employed to partition the weighted graph constructed from the data into communities, maximizing the modularity Q to discover significant community structures. This approach narrows the analysis scope, reduces noise interference, and focuses on high-risk candidate groups. Next, a Graph Convolutional Network (GCN) is used for feature learning and label prediction at the community level, effectively capturing the nonlinear relationships and complex interactions between nodes, thereby overcoming the limitations of traditional linear models and rule-based analysis. Finally, a weighted label propagation algorithm is applied to diffuse the prediction results to the entire dataset, ensuring global consistency and coverage of the label information. Additionally, a graph database is utilized in place of the traditional relational database, avoiding the inefficiencies of multi-table join queries and accelerating data retrieval and processing. This method facilitates the efficient identification of potential high-risk candidates in large and complex examination datasets, enhancing the fairness and management efficiency of the examination process.

Keywords: label propagation; community partitioning; graph database; graph convolutional network

引言

在当今数据驱动的世界，如何有效地分析和挖掘大规模数据集中的有用信息是一个重要的研究课题。传统方法，如全局特征提取（Global Feature Extraction）或基于规则的分析（Rule-Based Analysis）一般是直接分析所有数据，常常需要处理大量噪声和不相关信息。这不仅容易受到随机节点的干扰，进而降低模型的预测准确性，还增加了计算成本，降低了分析效率[1]。此外，在缺乏明确社交

结构或关联的情况下，这些方法往往难以识别出潜在的行为模式或特征，使得在作弊识别等关键任务中无法有效聚焦于高风险考生。此外，许多传统模型在分析特征时依赖于线性特征提取或简单的分类方法，如 Logistic 回归、线性支持向量机（Linear SVM）等 [2]，这些方法通常假设特征之间的关系是线性的。这在处理复杂的社交网络或考生行为数据时，往往无法捕捉到潜在的非线性关系。同时，图的结构一般来说是十分不规则的，可以认为是无限维的一种数据，所以它没有平移不变性。每一个节点的周围结构可能都是独一无二的，这种结构的数据，就让传统的 CNN（Convolutional Neural Network，卷积神经网络）、RNN（Recurrent Neural Network，循环神经网络）失去效果。

在国家组织的大规模考试中，代报名行为的组织化日益严重，这种现象与考生之间的社区特征有着密切的联系。组织化的代报名和团伙作弊行为通常通过特定的培训机构或第三方组织集中安排考生进行代报名和统一作答，严重威胁到考试的公正性。这一代报名特点可以通过对报名数据进行深入分析揭示出来。首先，大规模考试的一个显著特点是考生基数庞大，因此相关考生的数据量也非常庞大。这往往会为特征提取过程引入大量的噪声和不相关信息，使得传统方法难以有效捕捉潜在的作弊行为。其次，考生报名数据的维度众多，涉及的特征复杂且相互关联。在这种高维数据环境中，各个因素的权重衡量变得极为复杂，简单的线性回归模型难以在高维图数据分析中取得良好的效果 [3]。此外，识别和扩散考生之间的潜在特征也是一项重要的挑战。我们需要从考生的关系网络中提取这些特征，并确保能够覆盖所有考生，以避免遗漏任何潜在的高风险考生。这一过程要求有效地处理节点之间的关系信息，确保所有考生的标签能够准确反映其在社交网络中的位置和相互影响。最后，传统的 SQL 查询方式由于依赖于大量的联表操作，导致查询效率极低，并受到硬件设备的限制，无法满足快速检索和实时分析的需求。这些问题促使我们探索更加高效和精准的数据分析方法，以维护考试的公平性与诚信性。

为了解决这些问题，我们提出了一种基于图（graph）结构的多阶段综合分析方法，以提高对高危考生的识别能力和考试管理的有效性。本文首先对数据集采用 Louvain 算法 [4] 进行社区划分，通过最大化网络的模块度 Q（modularity）来识别图中的社区结构，可以专注于具有明显社区特征的数据。通过缩小研究范围，能够有效减少噪声干扰。对于这些数据采用 GCN [5]（图卷积神经网络，Graph Convolutional Network）进行标签预测，可以解决许多现有模型依赖线性特征提取或简单分类方法的问题，通过其强大的特征学习能力，能够更好地反映节点间的复杂交互。并且，GCN 创新性的设计了一种从图数据中提取特征的方法，让我们可以使用这些特征去对图数据进行节点分类（node classification）、图分类 [6]（graph classification）、边预测（link prediction），还可以顺便得到图的嵌入表示（graph embedding）。在此基础上，通过加权标签传播算法 [7]，对所有数据进行标签更新，进一步识别潜在的社区关系和标签信息。这一过程能够在减小运算量的同时还能够将结果传播到完整的数据集上，增强了对数据整体性的理解，使得模型的分析结果更具可信度。此外，传统的关系数据库在研究数据之间的关系时，往往需要使用 JOIN 操作。这种方法在数据量大或关系复杂时，计算性能可能显著下降，导致查询效率低下。为了解决这一问题，构建图数据库以图结构（节点、边和属性）来表示数据，成为了一种有效的替代方案。

一、研究方法

（一）社区划分

图数据库是一种专门用于存储和处理图形数据的数据库管理系统 [8]，它以图结构，包括节点 (Nodes)、边 (Edges) 和属性来表示数据及其关系 [9]。我们的算法是在图数据库的基础上，首先进行社区划分，用于缩小数据范围，找出具有明显社区特征的数据，然后通过 GCN 进行特征的获取与识别，并且通过标签预测和染色的形式输出结果。[10]

设 $G=(V,E,A,K)$ 表示一个加权图。其中 $V=v_1,v_2,...,v_N$ 表示大小为 $|V|=N$ 的节点的集合，节点 (Nodes) 是图中表示实体的基本单位。在许多应用中，节点可以表示不同的对象，如用户、商品、地点等，例如用 v_i 表示第 i 个节点。A 是一个 $N\times N$ 的邻接矩阵，邻接矩阵 (Adjacency Matrix)：是一个方阵，用于表示图中节点之间的连接关系。对于一个有 N 个节点的图，邻接矩阵 A 的大小为 $N\times N$。矩阵元素 $A_{ij}=1$ 表示节点 i 和节点 j 之间存在边，$A_{ij}=0$ 表示节点 i 和节点 j 之间不存在边。在加权图中，邻接矩阵的元素是边的权重，即 $A_{ij}=w_{ij}$。同时，E 表示所有边的集合，边 (Edges) 是连接两个节点的线段，表示节点之间的关系或交互。边可以是有向的（表示关系有方向）或无向的（表示关系没有方向），例如 e_{ij} 表示节点 i 和节点 j 之间的边，我们这里假设 E 是 m 维的，$m=1/2\sum_{ij}A_{ij}$。K 表示度，节点的度 (Degree) 是指与该节点直接连接的边的数量，加权图的情景下度表示与该节点直接连接的边的加权和。通常用 k 表示，例如节点 i 的度用 k_i 表示，$k_i=\sum_j A_{ij}$。[6]

社区划分算法 [11] 主要分为局部最优节点移动和社区合并两个迭代部分，旨在通过最大化网络的模块度 Q 来发现图中的社区结构。我们从有 N 个节点的加权图 $G=(V,E,A,K)$ 开始研究问题，模块度的定义如下：

$$Q=\frac{1}{2m}\sum_{i,j}\left(A_{ij}-\frac{k_ik_j}{2m}\right)\delta(C_i,C_j)$$

C_i，C_j：表示节点 i 和 j 的社区标识。

$\delta(C_i,C_j)$：δ 是克里尔函数，如果节点 i 和 j 属于同一个社

区，则 $\delta(C_i,C_j)=1$，否则为 0。

局部最优节点移动 [12]：首先，我们为网络中的每个节点分配一个不同的社区。因此在初始的分区中，节点与社区的数量一样多。然后，对于每个节点 i，我们考虑 i 的邻居节点 j，将节点 i 移动节点 j 所在的社区中，到计算增益（$\triangle Q$），若增益是正的就在原本社区中移除 i，让它保持在 j 的社区里。也就是说，依据这个规则，我们会计算节点 i 的所有邻居节点的增益，通过

$$arg\max_{\delta}\frac{1}{2m}\sum_{i,j}\left(A_{ij}-\frac{k_ik_j}{2m}\right)\delta(C_i,C_j)$$

将此节点放到增益最大的社区中，如果不可能获得积极的增益，我就留在它原来的社区里。这个过程被重复和顺序地应用于所有节点，直到没有进一步的改进，然后第一阶段完成。当达到模块化的局部极大值时，即当没有单个移动可以改善模块化时，停止第一阶段。

社区合并：在节点移动达到局部最优之后，将每个社区视为一个超级节点（supernode），该算法的第二阶段需要建立一个新的网络，其节点现在是在第一阶段发现的社区。[13] 为此，新节点之间的链接的权值由对应的两个社区中节点之间的链接的权值之和给出。同一社区节点之间的链接被视为该社区在新网络中的自环边。第二个阶段完成之后就可以将算法的第一个阶段重新应用到所得到的加权网络上并进行迭代。原始社区的数量在每次迭代时都会减少，因此大部分的计算时间花费在第一阶段中。这个算法让人想起了复杂网络的自相似性质，并很自然地包含了层次结构的概念，因为“社区的”社区是在这个过程中建立的。所构造的层次结构的高度由通过的次数决定，通常是一个很小的数目 [14]。

（二）标签预测与标签传播

基于社区划分结果，我们可以筛选出有明显的社区结构的数据，过滤掉模块度较低的和组成的人数较少的社区。接下来，我们对于这些数据采用 GCN 进行标签预测以及标签传播 [15]。在上述图 $G=(V,E,A,K)$ 的基础上，我们介绍一些重要的符号如下：

1. $\mathbf{X}=[x_1,x_2,\ldots,x_N]^T\in R^{N\times p}$ 是一个观测到的大小为 $N\times p$ 的输入特征矩阵，其中 $x_i\in R^{N\times p}$ 是一个 $p\times 1$ 维的特征向量。

2. $\mathbf{Y}=(y_1,y_2,\ldots,y_C)^T$ 是响应变量，y_i 是一个分类标签，C 表示类别的数量，可以是二分类也可以是多分类。

图卷积网络（GCN）的核心思想是利用图结构中的邻接关系对节点特征进行聚合和更新。每一层的 GCN 操作都会聚合节点的邻居信息，以此更新每个节点的表示，直到最后一层产生用于分类或预测的节点表示。设总共有 L 层图卷积神经网络第 l 层的节点表示为 $H^{(l)}$，$H^{(l)}\in R^{N\times D_l}$ 是第 l 层的节点特征表示矩阵，D_l 是第 l 层的输出特征维度。同理，第 $l+1$ 层的节点可以表示为 $H^{(l+1)}$，$H^{(l)}\in R^{N\times D_{l+1}}$，这里需要说明一点，每一层的输出特征维度 D_l 可以不同，通常由超参数决定，这种设计使得模型具有灵活性，并能够适应不同任务的需求。例如，对于图分类任务，可能需要较大的特征维度来捕捉复杂的图结构特征；而在节点分类任务中，可能在中间层使用较小的特征维度，通过后续层逐步增强节点的表示能力。当 $l=0$ 时，$H^{(0)}$ 其实就是输入层 $\mathbf{X}$。我们对特征矩阵的迭代如下：

$$H^{(l+1)}=\sigma\left(\hat{A}H^{(l)}W^{(l)}\right),$$

其中 $W^{(l)}\in R^{D_l\times D_{l+1}}$ 是第 l 层的可学习权重矩阵。$\hat{A}$ 是是归一化后的邻接矩阵，$\hat{A}=K^{-1/2}(A+I)K^{-1/2}$

其中 I 是单位矩阵，K 是节点度矩阵，它是一个对角矩阵，其中 $K_{ii}=k_i$。σ 是一个非线性的激活函数，可以是 Relu，softmax 等。

将其展开后，每个节点的更新函数如下：

$$h_i^{(l+1)}=\sigma\left(b^{(l)}+\sum_{j\in N(i)}\frac{1}{c_{ij}}h_j^{(l)}W^{(l)}\right)$$

其中，$N(i)$ 是目标节点 i 的邻居节点集合，c_{ij} 是 i 和 j 对应节点度数的平方根的乘积，即 $c_{ij}=\sqrt{N(i)N(j)}$，σ 为激活函数。

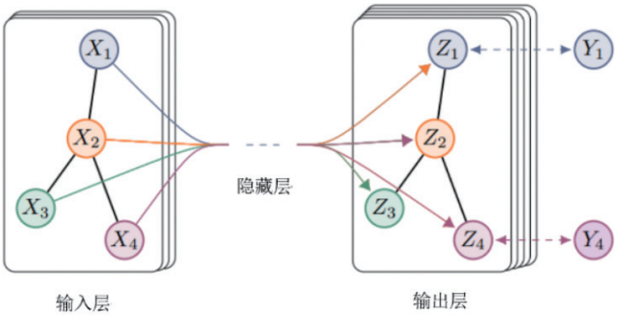
假设我们进行一个多分类问题，那么标签预测通常用 softmax 函数将每个节点的最终表示映射到标签空间，得到每个类别的概率。softmax 函数将原始的 logits（即 h_iW_{out}）转换为概率分布，使得每个类别的输出值都在 $[0,1]$ 之间，并且所有类别的概率之和为 1。softmax 可以理解为对每个类别的预测概率进行归一化处理。假设有 C 个类别，节点 i 的最终表示是 h_L，记预测标签 $\hat{y}$ 为 Z_i，则

$$Z_i=softmax(h_iW_{out})$$

为了训练模型，使用交叉熵损失函数（cross-entropy loss），该损失函数用于计算预测标签与真实标签之间的差异。对于每个节点 i，交叉熵损失计算为：

$$L=-\sum_{i\in Y_N}\sum_{c=1}^{C}Y_{ic}lnZ_{ic},$$

Z_{if} 为节点 i 的预测标签。[16]



> 图 1：GCN 结构示意图

在每次迭代中，通过前向传播计算 Z_i，再计算损失 L，这一步会得到一个标量损失值，用于量化当前模型的预测效果。然后使用反向传播算法计算损失 L 对模型参数（权重矩阵 $W^{(l)}$）的梯度 $\nabla LW^{(l)}$。反向传播通过链式法则，从输出层逐层回传到输入层，从而计算每一层权重 $W^{(l)}$ 对损失的偏导数。之后使用梯度下降（或者 Adam 优化器）更新每一层的权重矩阵 $W^{(l)}$：

$$W^{(l)}\leftarrow W^{(l)}-\eta\nabla LW^{(l)}$$

其中 η 是学习率，控制每次迭代的更新步长。更新后的权重将用于下一次迭代的前向传播。重复上述步骤直到损失函数收敛

（即损失值不再显著下降）或达到预设的最大迭代次数。

同时，softmax函数的定义为：

$$Z_{ic}=\frac{e^{(h_L W_{out})_c}}{\sum_{j=1}^{C} e^{(h_L W_{out})_j}}$$

由此，我们得到了每个类别的预测概率，最终的标签预测是选择概率最大的类别，[17]即：

$$Z_i = arg\ \max_c Z_{ic}$$

GCN可以用于监督学习，也可以用于半监督学习。在监督学习的情况下，GCN利用所有带标签的节点信息进行训练，通过优化损失函数（如交叉熵损失）来尽量准确地预测每个节点的标签。而在半监督学习中，GCN能够有效地通过少量已标记节点的信息和图结构的传播机制，将标签信息扩散至未标记节点，从而实现对未知标签的预测。因此，GCN 在处理具有复杂关系的图数据时，具备良好的灵活性和适应性。

（三）标签传播

基于以上从社区数据中预测并且重新打标签的结果，我们要将标签结果推广到完整的数据集中，这个过程要采用一种带权标签传播。标签的传播过程会参考节点之间的边权重，每条边的权重通常代表两个节点之间关系的强弱。具体来说，当一个节点的标签需要更新时，它不仅会参考邻居节点的标签，还会根据邻居节点的边权重来加权这些标签的影响。例如，若两个节点的互动频率较高（即边的权重较大），则标签传播时，频繁互动的邻居节点的标签对该节点的更新影响更大。

假设节点 i 的标签是通过其邻居的标签来决定的，带权标签传播的标签更新可以表示为：

$$y_i = arg\ \max_y \sum_{j\in N(i)} w_{ij}\cdot\delta\left(y_j, y\right)$$

其中，y_i 是节点 i 的标签。$N(i)$ 是节点 i 的邻居集合。w_{ij} 是节点 i 和 j 之间的边的权重。$\delta\left(y_j, y\right)$ 是克里尔函数，当邻居节点 j 的标签 y_j 等于候选标签 y 时，返回1，否则返回0。此公式表示节点 i 选择其邻居中标签最多的标签，且邻居的标签影响力会根据边权重 w_{ij} 进行加权。

通过引入边权重，可以让标签传播更集中在强关联的节点之间，标签的更新会更偏向于那些具有较大边权重的邻居。同时，噪声数据（如偶然连接的节点）在标签传播过程中可能会导致错误传播。通过权重控制，噪声节点（通常具有较小的边权重）的标签影响力会减弱，因而预测更为准确。

二、算法

我们提出了一种基于图结构的多阶段分析方法，旨在提高对高危考生的识别能力。在建立好图数据库后，通过最大化模块度 Q 来识别社区结构，从而缩小研究范围。然后，对这些具有明显社区特征的数据，使用GCN进行标签预测。接下来，通过加权标签传播算法，对所有节点进行标签更新，使得结果可以传播到完整数据集上。

表1 Louvain社区划分算法

算法 1：Louvain社区划分算法
输入：加权图 $G=(V,E,A,K)$，节点 V，边集 E，邻接矩阵 A，度集 K。 输出：图的社区划分结果 1：初始化：将每个节点 $v\in V$ 分配为一个单独的社区。 第一阶段－模块度优化： 2：for每个节点 $v\in V$，遍历其邻居节点集合 $N(v)$： 3：$\arg\max_{\delta}\frac{1}{2m}\sum_{i,j}\left(A_{ij}-\frac{k_i k_j}{2m}\right)\delta\left(C_i,C_j\right)$， 4：if Q>0 then 5：将节点 v 移动到该社区。 6：else 保持节点社区不变。 7：return局部模块度优化结果 第二阶段－聚合社区： 8：初始化：将局部模块度优化结果中的每个社区合并为一个超级节点 w，$w\in W$。 9：for每个节点 $v\in V$， 10：repeat 步骤1-7。 10：return社区划分结果。

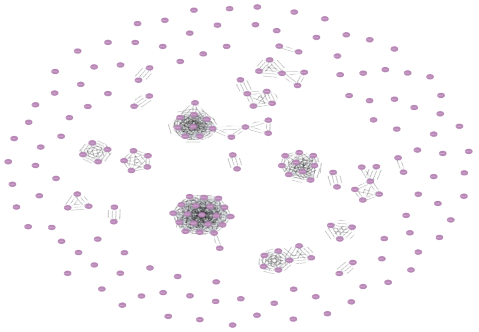
表2 标签预测与标签传播算法

算法 2：标签预测与标签传播算法
输入：特征矩阵 $\mathbf{X}$，标签矩阵 $\mathbf{Y}$，加权图 $G=(V,E,A,K)$，学习率 η。 输出：标签预测概率矩阵 $H^{(L)}$，标签传播结果 $\hat{\mathbf{Y}}$，训练后的权重 $\left\{W^{(l)}\right\}_{l=1}^{L}$。 1：初始化：将节点特征表示设为 $H^{(0)}=\mathbf{X}$。 2：for每一层 $l=1,...,L$： 3：归一化邻接矩阵 $\hat{A}=K^{-1/2}(A+I)K^{-1/2}$， 4：计算节点嵌入表示 $H^{(l)}=softmax\left(\hat{A}H^{(l-1)}W^{(l-1)}\right)$， 5：计算 $L=-\sum_{i\in Y_N}\sum_{c=1}^{C}Y_{ic}lnH_{ic}^{(L)}$， 6：逐层计算权重矩阵 $W^{(l)}$ 的梯度 $\nabla LW^{(l)}$， 7：使用梯度下降更新权重矩阵 $W^{(l)}$： $W^{(l)}\leftarrow W^{(l)}-\eta\nabla LW^{(l)}$ 8：repeat 2-7， 9：until L 足够小或达到迭代次数。 10：return标签预测概率矩阵 $H^{(L)}$，训练后的权重 $\left\{W^{(l)}\right\}_{l=1}^{L}$。 11：计算概率最大的标签值作为标签预测结果： $\hat{y}_i=arg\ \max_c H_{ic}^{(L)}$ 12：根据标签结果进行权重标签传播： $y_i=arg\ \max_y \sum_{j\in N(i)} w_{ij}\cdot\delta\left(y_j,\hat{y}_i\right)$ 13：return标签传播结果 $\hat{\mathbf{Y}}$。

三、实证分析

为了检验以上的方法的有效性，我们对2023年的某地（我们称为J市）实际考试情况进行了分析。数据的考生个人信息已经经过加密，所以没有泄露隐私等问题。J市一共893个考生，划分成了403个社区，其中社区人数低于5人的有392个，高于5人的有11个社区，剔除低风险的人之后剩余408个人，模块度为0.706。

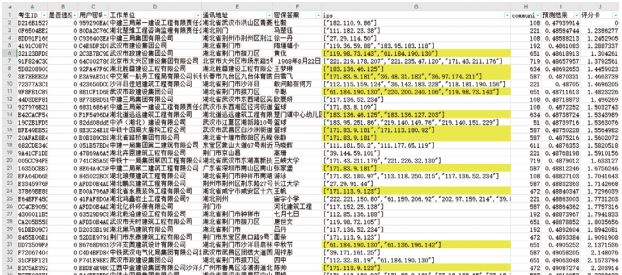
下面这张图展示了J市考生的网络在社区划分后的结构特征，其中存在多个密集的社区，这些社区代表着相互关联较强的考生群体。从网络结构来看，密集的连接意味着这些考生之间可能存在某种特殊的联系，比如通过同一渠道报名、共享资源，或者存在某种组织关系。尤其是一些大的社区，节点之间联系非常紧密，这可能意味着这些考生属于一个由代报名代理人或团体组织的集体。它们可能在报名时间、地点，甚至个人信息方面存在高度的相似性，从而形成了这种网络中的核心团伙。



> 图1 J市考生的网络

此外，图中的许多外围节点则代表着联系较少的独立考生，这些考生的特征比较分散，说明他们在报名过程中并未体现出与其他考生有显著的关联。这样的孤立节点通常可以理解为个人报名，行为上没有明显的群体特征，因此也相对不太容易被代报名团伙所控制。

从社区的整体分布来看，图中的大规模社区比如编号为587,651，192社区内的人数众多，ip和通讯地址，工作单位重复度高，联系紧密，尤其值得关注。这些社区可能代表着某种非正常的组织行为，即代报名活动。比如，社区中规模越大，节点之间联系越紧密，说明该团体内部的考生之间关联性越强——这在正常的考试报名中是不太合理的，因为独立的考生之间并不应有如此密集的交互。因此，这些大的社区可以视为风险较高的重点群体，可能是代报名团伙的关键部分。



> 图2 通过我们的算法得出的J市考生结果

这张图展示了通过我们的算法得出的J市考生的数据，结果显示出一些考生具有较低的评分，这些低评分的考生确实揭示了潜在的异常行为。结果包含多个特征列，如考生ID、用户密保、工作单位、通讯地址，密保答案等。其中，我们发现了多次重复出现的典型风险IP，这些IP在多个不同考生中反复出现，表明这些考生可能通过相同的网络环境进行报名，这很可能与代报名活动有关。我们将这些典型IP被标记为黄色。例如，“171.83.9.181”和“171.113.9.123”这样的IP在不同考生中多次出现，暗示这些考生很可能使用了同一网络位置进行报名。

> 图3 587的社区的结果

这个表格展示了Louvain算法划分出的编号为587的社区的结果，社区587一共有84个成员表现出多种关联特征，包括共用IP、共用密码，以及多人存在作弊行为，这表明这个社区内考生的行为是高度协作的。并且这些考生在网络结构中彼此联系紧密，符合算法对社区划分的标准。而这些紧密联系的行为并非随机发生，而是有特定的行为模式和目的。

1.社区587内有55个考生共用“171.83.9.181”这个IP地址，高达65%。这表明这些考生很可能是在同一个网络环境中报名的，这种情况通常发生在公共网络设施（例如网吧、代理机构）或者某个特定的组织内部。正常情况下，考生报名的网络环境应该是分散的，除非存在特定的理由（如住址相同或公司内部报名），否则共用同一IP是非常不寻常的。因此，这种共用IP的现象可能是某种集体报名的证据，特别是当该IP连接的考生人数众多时，进一步增强了作弊或代报名的怀疑。

2.在社区587中，有14人共用了同一个密码。这在一个正常的报名流程中也是极为不合理的现象。每个考生的密码通常是唯一且保密的，用于确保个人信息的安全性，而共用密码显然违背了这一原则。这很可能是因为这些考生的账户是由同一个人或组织统一管理的，这样才能解释为什么会出现同样的密码。这也极大地提示了存在一个代报名代理，或者这些考生之间存在某种组织上的从属关系。

3.在这个社区内一共84名考生中，有6名被抓到了作弊行为，这比例远高于正常水平。这进一步说明社区587是一个高风险群体，这些考生之间可能共享资源、信息，甚至通过类似手段参与作弊。作弊行为的高发与共用IP和共用密码的特征结合，几乎可以确认这个社区内部的人员具有某种关联，他们可能利用集中控制来操纵考生，形成某种作弊团伙。

4.在考后对该社区的进一步分析中，我们对错同率进行了检测，结果再次证明了社区587的高度风险性。33名考生在错同率上出现重大问题，错同率异常意味着这些考生在考试中的答案具有高度一致性，这通常意味着考生之间可能存在信息共享、集体作弊或作弊手段类似等问题。整体来看，社区587中错同率异常的考生达到了76人，占了社区总人数的44%。相比之下，社区587在研究的考生总体中仅占20%，但其错同率问题占据了更高的比例，这种显著差异表明，社区587内部存在不正常的考试行为。

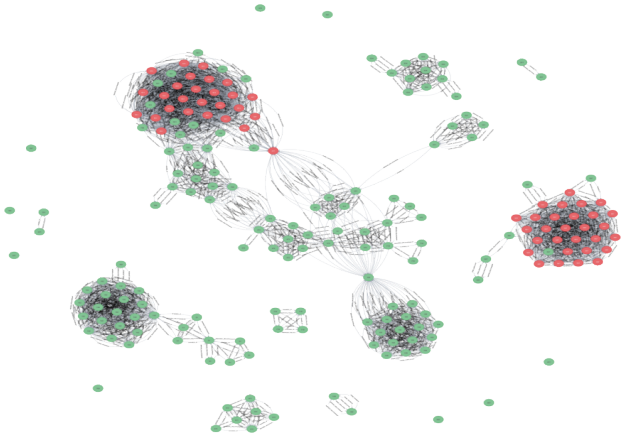
这些考生在考前存在共用密码、共用IP的现象，考后又表现出错同率异常，表明这个社区内部可能存在系统化、集中化的作

弊组织行为，具有较强的关联性。社区 587 被证明是一个高风险社区，不论是在考前的报名信息分析，还是在考后的考试结果分析中，这些考生都显示出高度的作弊和关联行为。

考生ID	是否违纪(工作单位	ips	community预测结果	评分卡
37B69BE8E	0 湖北省永晨装饰工程有限公司	["171.113.9.123"]	472	0.488403 1.729604
8455B09E1	0 荆门市东泰建筑工程有限公司	["171.113.9.123"]	472	0.489338 1.909191
E2C7A8E5F0	0 江西中金建设集团有限公司沙洋分公司	["171.113.9.123"]	472	0.490873 2.203914
F208C3596	0 中铁二十二局集团有限公司	["171.113.9.123"]	472	0.492917 2.596496
5B5824DC2	0 武汉亚云盛物业管理有限公司	["171.113.9.123"]	472	0.504409 4.804018
390428B3AE	0 江西中金建设集团有限公司沙洋分公司	["171.215.35.149","171.215.30.86","171.113.9.123","171.215.35.123"]	472	0.508414 5.573207
24CB5C423	0 焦作市投资集团有限公司	["171.113.9.123","171.113.180.92"]	472	0.523108 8.39577
8B8C76317	0 武汉虹信技术服务有限责任公司	["171.113.9.123","171.43.149.74"]	472	0.527468 9.23336
B2F03ABF2	0 湖北省工业建筑集团有限公司	["171.113.9.123"]	472	0.5350597 10.89094
F663890A6	1 湖北远升房地产集团有限公司	["171.113.9.123","171.43.149.74"]	472	0.539273 11.50086
08B154E9C	0 湖北亿鼎建设工程有限公司	["171.113.9.123"]	472	0.542544 12.12921
B1AA5BF37	0 江西中金建设集团有限公司沙洋分公司	["171.113.9.123"]	472	0.5827605 19.85515
D3239216I	0 中交二航局第六工程分公司	["171.113.9.123"]	472	0.5985345 22.84789
4B074D10E	0 武汉永泰建设工程监理有限公司	["111.173.182.247","111.173.180.168","171.113.9.123"]	472	0.6053355 24.19437
670AE2241	0 湖北投资担保开发有限公司	["171.113.9.123","27.27.55.127"]	472	0.632278 29.46233
6045227C1	0 中铁十一局集团第四工程有限公司	["171.113.9.123","111.183.52.58"]	472	0.643821 31.58311
E0672ADB1	0 江西中金建设集团有限公司沙洋分公司	["112.20.86.110","171.113.9.123"]	472	0.644863 31.78315
3A5DF8FDE	0 中交二航局第六工程分公司	["171.113.9.123"]	472	0.6545566 33.64098
168741AC8	0 武汉永泰建设工程监理有限公司	["171.113.9.123"]	472	0.678888 38.31902
CF9A43F23	0 荆门市东泰建筑工程有限公司	["171.113.9.123","171.43.149.74","113.57.114.198"]	472	0.6966557 41.7322
BEC9E83A2	0 江西中金建设集团有限公司沙洋分公司	["171.113.9.123"]	472	0.803885 62.32916

> 图4 171.113.9.123 ip的考生结果

在这张表格中表现了所有使用 171.113.9.123 ip的考生。他们普遍获得较低的评分。这意味着在使用该 IP 地址进行报名的考生中，存在一些明显的异常行为模式，这些行为和作弊的高风险特征相符。评分低的原因很大程度上是由于该 IP 地址下有一个考生已经存在明确的作弊行为，这直接增加了对其他使用同一 IP 考生的怀疑程度。模型根据数据特征（如共享 IP、社区关系等）进行评分，表明这个 IP 的风险已经被放大，进而影响了其他使用同一 IP 的考生评分。



> 图5 标签扩散后的 J 市结果

上图为对社区结构进行染色后的网络图，其中节点被设置为绿色，风险考生被染成了红色。图中可以看到多个由红色节点组成的大型簇，这些簇显示出高度的内部联系性，这些考生很可能是通过某种形式紧密联系的。尤其是在社区 651 和 587 内部，许多节点被染成红色，表明这些考生在之前的分析中被识别为高风险个体，可能参与了代报名、集体作弊等行为。这些社区内部的红色节点之间边的密集性表示他们有较强的关联，可能通过共享 IP、共用密码、同时出现在特定报名地点等方式连接在一起。这种群体行为恰恰是代报名团伙的典型特征。

社区 408 和 587 在图中展现为核心的红色聚集地，这些社区的节点不仅内部联系紧密，而且还通过一些边连接到其他节点，这些外部连接通常指向一些绿色节点。这些红色社区内部的节点连接模式显示出高度组织化的特征，表明它们很可能受到某个中心化的管理或者某个代理的控制。例如，社区 587 之前的分析中就揭示出存在多人共用密码、多次使用相同 IP 的问题，以及考后错同率的显著异常，图中显示的红色进一步验证了社区 587 的高风险性。红色节点与绿色节点的交互：部分红色节点与外围的绿色节点有连接，表明这些绿色节点可能与高风险考生存在某种间接联系，例如认识或者某些信息上的共享，虽然这些绿色节点没有表现出直接的异常行为，但它们也可能受到影响，需要适度关注。

整个网络中，红色节点相对集中于几个特定社区（如 408 和 587），这表明高风险考生并非随机分布，而是具有较强的集中趋势，这种趋势往往来源于集体行动，例如通过同一个中介报名或者以群体形式进行作弊。

参考文献

[1] L. C. Thomas, J. N. Crook 和 D. B. Edelman, Credit Scoring and Its Applications. SIAM., 2017.

[2] E. E, J. Webber 和 R. I, Graph Databases: New Opportunities for Connected Data., O'Reilly Media., 2015.

[3] G. Rossetti 和 R. Cazabet, "Community Discovery in Dynamic Networks: A Survey.," *ACM Computing Surveys (CSUR),* pp. 1-37, 2018.

[4] B. V. D, G. J. L and L. R, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment,* 2008.

[5] T. Kipf 和 M. Welling., "Semi-Supervised Classification with Graph Convolutional Networks," *Proceedings of the International Conference on Learning Representations (ICLR),* 2017.

[6] F. Zhou, T. Li, H. Zhou, H. Zhu 和 J. Ye, "Graph-Based Semi-Supervised Learning with Non-ignorable Non-response," *In Advances in Neural Information Processing Systems 32 (NeurIPS 2019),* 2019.

[7] Z. Wu, S. Pan, F. Chen, G. Long 和 C. Zhang, "A comprehensive survey on graph neural networks"., *IEEE Transactions on Neural Networks and Learning Systems, 32(1), 4-24.,* 2020.

[8] 张云雷，《社区发现方法及应用》，清华大学出版社，2018.

[9] A. L, T. H 和 K. D, "Graph based anomaly detection and description: a survey," *Data Mining and Knowledge Discovery,* pp. 626-688, 2015.

[10] B. M, H. S 和 J. M, "Gephi: An open source software for exploring and manipulating networks," *Proceedings of the Third International ICWSM Conference.,* 2009.

[11] 马慧芳，《网络社区发现与搜索》，科学出版社，2019.

[12] 王建民，《复杂网络的社区结构与检测算法》，电子工业出版社，2017.

[13] S. Pandit, C. D. H, S. Wang 和 F. C, "NetProbe: a fast and scalable system for fraud detection in online auction networks," *Proceedings of the 16th international conference on World Wide Web.,* 2007.

[14] S. Fortunato, "Community detection in graphs.," *Physics Reports,* pp. 75-174, 2010.

[15] K. T. N 和 W. M, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907.,* 2016.

[16] R. U. N, A. R 和 K. S, "Near linear time algorithm to detect community structures in large-scale networks," *Physical Review E,* 2007.

[17] X. Zhu 和 Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," *Technical Report CMU-CALD-02-107,* 2002.