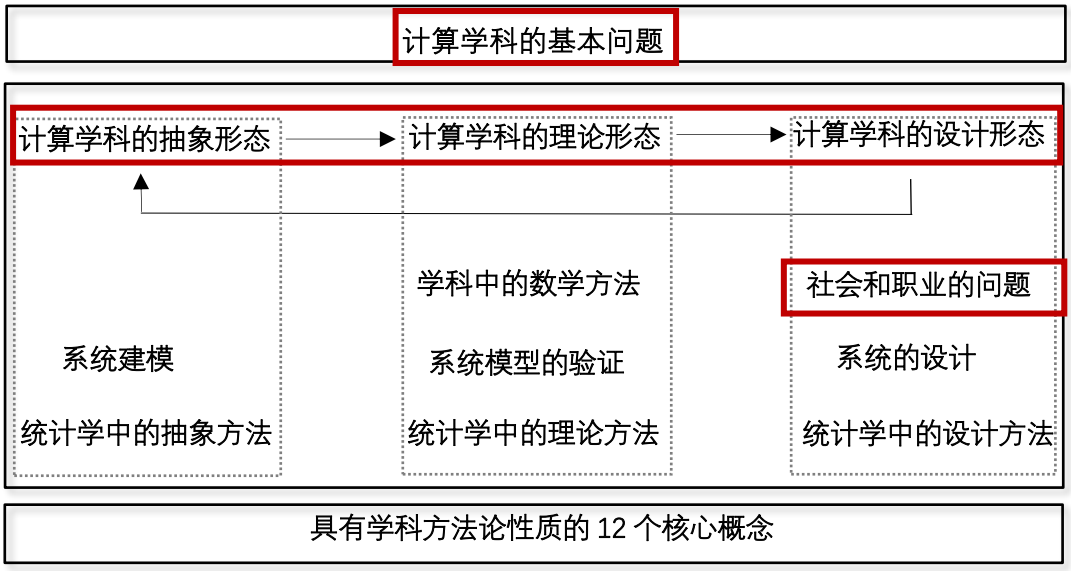
图数据分析洞察关联关系

邓明森　贵州财经大学
msdeng@mail.gufe.edu.cn

能力的评估：本案例能够置于 Bloom 分类法知识维度的“**元认识知识**”位置，学生学习后能够达到 Bloom 分类法认知过程维度的“**创造**”层次。

一、本案例课程思政的关注点

1、本案例内容在计算学科课程思政总体框架中的位置



2、本案例是一个典型的综合性案例，它涉及到的内容可以包括如何从纷繁复杂的表象中进行抽象建立模型，并对系统模型进行验证，并实现系统的设计。它是培养学生解决复杂工程问题的一种尝试。学生学习该案例后，能否对抽象、理论和设计三个学科形态如何进一步区分将有进一步的认识，对于计算学科的基本问题和解决问题的方式有一定程度的把握，对实现“0 到 1”的突破种下科学思维的种子。

3、在本案例中，要求教师将“知识”或者“技能”与具体的背景结合起来，让学生“更好”或者“正确”对待“知道是什么”，从而促使学生“知道怎么做”。并将 CC2020 报告中 11 个品行元素中的“目标导向、创造性、适应性”与该案例绑定在一起进行可操作性解释。**目标导向：**结合当前社会热点问题所暴露出来的问题，能够有效进行分析，找到所需要实现的目标；**创造性：**要求学生换道思考，既然基于财务指标的指标体系存在已知的问题，那么能否在大数据背景下利

用公开的非财务数据建立基于非财务数据的指标体系？并根据可能的情况提出前瞻性解决方案；**适应性**：资本逐利本质与监管体系之间永远是矛与盾之争。通过案例学习，让学生认识到没有一个完美的监管指标体系，只有不断发展的监管指标体系。在设计系统时永远保持开放的态度，永远保持灵活、敏捷和适应事物的发展变化。

二、本案例的具体内容

1、背景介绍

广东某个成立于 1997 年的药业公司在 2001 年于 A 股上市，是一家以中药饮片、化学原料药等为主导，集药品生产、研发等为一体的现代化大型医药企业。在 2018 年 5 月公司股价达到了历史最高点，市值超过了 1390 亿元，在漫漫熊市中成了股民们心中的大白马和摇钱树。然而出人意料的是，该药业公司受网络媒体质疑与其深圳关联公司涉嫌内幕交易，2018 年 10 月其股票盘中突然跌停，次日再度闪崩跌停。12 月底，证监会受理调查该公司并要求其自查。但让股民感到振奋的是，该药业公司在 2019 年初兑付了一笔 20 多亿元人民币的融资券。随后股民大幅加仓了该公司股票，导致 2019 年第一季度的股东数相比 2018 年第四季度还增加了 6 万户，股价又翻了一倍。

然而好景不长，在证监会的调查下，2019 年 4 月该药业公司发布了会计差错公告。在公告中该药业公司宣称由于财务数据出现会计差错导致 2017 年营业收入多计入了 80 多亿元，同时营业成本多计入了 70 多亿元等，总计多计入了近 300 亿元。因此这个差错报告对 2017 年的财务报告作出重大调整而导致账面近 300 亿现金不翼而飞。次日，上交所向该药业公司发布问询函。次月证监会正式通报其三年财务报告具有重大虚假的事实。2019 年 5 月，该公司股票正式被特别处理(ST)，从而拉开了对 A 股历史上最大的一起造假案进行处理的序幕。截止 2022 年 8 月，该公司仍然属于“摘星”未“脱帽”状态。

此处需要提出的第一个问题是，在监管层面有无可能对上市公司特别是将会发生信用风险的企业进行风险预测，从而避免类似的重大风险发生。按照我国相关监管规定，如果上市公司出现财务或者其他状况异常，导致该股票存在终止上市风险或者让投资者无法判断公司前景而致使投资者权益可能受到损害，证券交易所会对该公司股票交易实施特别处理(即 ST)，以告诫股民投资该公司股票的

风险系数较高。我国《公司法》还规定，上市公司必须依照法律、行政法规的规定，公开其财务状况、经营状况及重大诉讼，在每会计年度内半年公布一次财务会计报告。相关条例规定上市公司的年度报告必须经过独立的第三方会计师事务所进行审计，防止出现报表错误、漏洞、偷税漏税等问题，也保护投资者的经济利益。也就是说，在监管层面，实际上已经努力营造更好的投资环境，对上市公司的财务数据进行第三方监管。

抛开稍微久远的 2001 年轰动全球的安然造假事件，2021 年春节期间，某重要会计师事务所员工公布了该事务所协助客户财务造假的幻灯片，展示了在过去 4 年该事务所的各种不合规事件，涉事企业包括多家上市公司，涉事员工层次复杂，甚至包括合伙人。实际上，在上述药业公司的造假案中，财务数据的会计差错也是重要的导火索。

因此，实际上已经提出了第二个问题，既然财务数据并不完全可靠，在当前信息技术飞速发展、产业数字化高度渗透的环境下，是否有可能发展出一套基于非财务数据的企业风险预测与监管体系？

2、问题求解

（1）从现象到本质的归纳

虽然基于关联方的关联交易有正面影响，如降低成本，改善业绩、优势资源配置、提高企业的市场竞争力等等，但一般意义上，关联交易容易导致企业财务造假，这也是安然造假案发生之后监管机构极力防范的造假途径。首先通过仔细阅读近 5 年涉事药业公司的年度报告，并对其报告中提到的主要交易进行跟踪，研判是否存在关联交易的可能。其年度报告中提到的人和事件非常复杂，涉及到了商品采购与出售、关联担保、资金拆借等等。利用图算法中的抽象方法，可以将涉及到的人或者企业均定义为 1 个顶点（Vertex），它们之间的交易或者其他关联均定义为 1 条边（Edge）；如它们之间没有任何关联，它们之间就没有边相连。因此可以将这种复杂的相互关联关系抽象为一张图（Graph）$G(V,E)$。这个图可以用类似于图 1 的这种结构来表示。

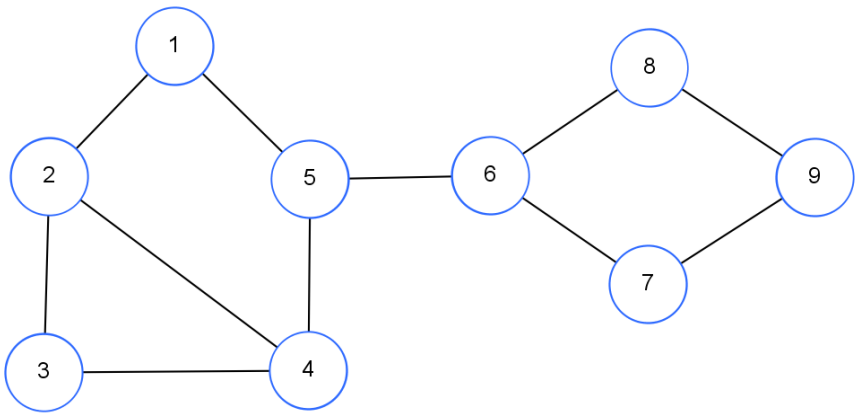


图 1 企业之间的关联示意图$G(V,E)$

（2）图结构的数学表示和计算机存储

图 1 对应的图可以用邻接矩阵(adjacent matrix)表示为：

$$A=\begin{pmatrix}0&1&0&0&1&0&0&0&0\\1&0&1&1&0&0&0&0&0\\0&1&0&1&0&0&0&0&0\\0&1&1&0&1&0&0&0&0\\1&0&0&1&0&1&0&0&0\\0&0&0&0&1&0&1&0&0\\0&0&0&0&0&1&0&1&1\\0&0&0&0&0&1&0&0&1\\0&0&0&0&0&0&1&1&0\end{pmatrix} \quad (1)$$

图 1 中包含 9 个顶点的图被表示为这个9×9的矩阵之后，首先要考虑的就是它在计算机中的存储。每个矩阵元素假定需要 4 Byte 的存储空间，那么图 1 所示的图就需要$9\times9\times4=324$Byte.可能我们对这个存储需求并不敏感。然而截止 2019 年 12 月，A 股一共有 3378 家上市公司，根据这些上市公司的股东、董事会、监事会、高管等公开的非财务数据构建起来的图结构超过了 7 亿个顶点。仅假定这个图的顶点数为 1 亿，按照这种存储方式需要的存储空间为$10^9\times10^9\times4$Byte=$10^6\times10^9\times4$KB=$10^3\times10^9\times4$MB=$10^9\times4$GB=$10^6\times4$TB=$10^3\times4$PB（为了计算方便，我们简单将 1024byte≈1000byte=1KB，并以此类推）。仅仅是存储一个类似于图 1 的超过 1 亿个顶点的图的邻接矩阵就需要如此大的存储空间，这显然带来了更为巨大的 IO 与计算开销。

通过仔细观察矩阵(1)，可以发现这个矩阵的元素绝大部分都是 0，而只有少数部分为 1，是一个名副其实的“稀疏”矩阵。如果我们采用一个数组，仅仅存储元素值为 1 的顶点对，需要的存储容量会大幅度下降。实际上，稀疏矩阵的存储和高效使用在当前数据科学、机器学习等领域有着十分广泛的应用。稀疏矩阵

的存储和计算方法还有很多种，同学们可以自行学习。

（3）A 股上市公司关联方图的建立

通过获取 A 股所有上市公司的相关数据以及这些上市公司的股东、董事会、监事会、高管等公开的非财务数据，就能得到围绕目标企业的关联方图（如图 2 所示）。假定中心的黑点为目标企业，我们可以定位围绕目标企业的 n 度关联方，即与目标企业直接发生关联的叫 1 度关联方，与 1 度关联方直接关联的叫 2 度关联方，等等。

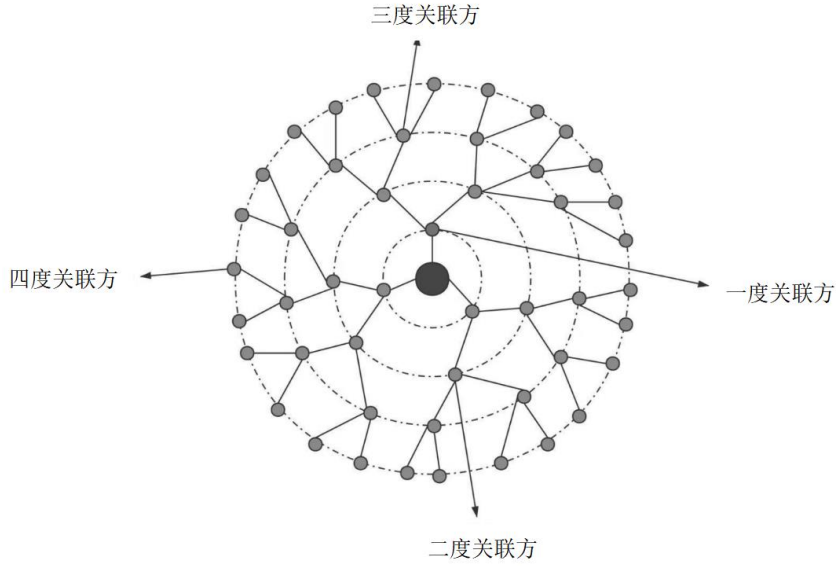


图 2 上市公司关联方图示意图（中心点为目标企业）

在这个图中，通过董事会等“管理”或者股权等“投资”公开信息就能将 A 股全部的上市公司以及跟它们关联的公司全部链接起来了。

（4）如何在上市公司关联图与企业风险之间建立关联

建立关联图之后，最重要的问题是如何根据这个图得到描述它与企业风险之间的关联。这就必须找到一些描述它的关键变量或者关键指标。事实上，在图算法中很容易找到如度、度分布、聚类系数、最短路径等各种描述图中节点重要性和边重要性的变量。从图(1)可以很直观的看到，顶点 5 的度为 5。通过邻接矩阵可以看到，度的计算就是将邻接矩阵中相应的行或者列相加：

$$k_5 = \sum_{j=1}^{9} A_{5j} = 3 \tag{2}$$

其他的量也可以类似得到。但除此之外是否还有其他更为综合性的指标呢？答案自然是肯定的，可以列出一些指标或者关键变量。如：

$$\rho_i = \frac{c_i}{n_i(n_i-1)/2} \tag{3}$$

其中c_i是企业 i 度关联方之间直接相连的边数，n_i是企业 i 度关联方数量。显然公式(3)表示的是在某一个关联方层次内的连接密度，它展示的是这些企业之间的关联紧密程度。这一变量在其他场合可能并不一定有特别具体的意义，但在企业关联方这一场合是有具体的意义的。实际上我们还可以根据实际场景构建更多的具有现实意义的指标。

由此，得到上市公司关联图相关联的形式模型为：

$$\text{Related-partyGraph}=\langle i,j,l,N,a_{ij},e_{jl},v_i,n_i,k_i,C_i,c_i,\sum c_{ij},\rho_i,\beta_i,\cdots\rangle$$

其中，

（1） $i,j,l=\{1,2,\dots\dots\}$表示关联图维度变量

（2） $N=\{0,1,2,\dots\}$为关联图的维度

（3） $a_{ij}=\begin{cases}1，& \text{存在关联}\\ 0，& \text{不存在关联}\end{cases}$，表示关联图的邻接矩阵

（4） $e_{jl}\in E$表示关联图中的边

（5） $v_i\in V$表示关联图中的顶点

（6） $n_i=\{0,1,2,\dots\dots\}$表示目标企业i度关联方的数量

（7） $k_i=\sum_{j=1}^{N}A_{ij}$表示目标企业的度

（8） $C_i=\frac{2|e_{jl}:v_j,v_j\in V,e_{jl}\in E|}{k_i(k_i-1)}$为目标企业的聚类系数

（9） $c_i=\sum_{j=1}^{N}m_{ij}$表示企业 i 度关联方之间直接相连的边数，其中m_{ij}表示企业 i 度关联方向下连出的边数

（10） $\rho_i=\frac{c_i}{n_i(n_i-1)/2}\in[0,1]$表示企业 i 度关联方内的连接密度

（11） $\beta_i\in[0,1]$表示企业 i 度关联方向下连出的密度

如果将这些所有的指标放在一起，就构成了一个可能描述企业风险的非财务指标的指标池。定义这个指标池为J，任意选择特征集合的一个特征子集$J^{(0)}$用于初始建模，将初始模型中系数不为 0 的特征标记为 1，系数为 0 的特征标记为 0，即：

$$J^{(0)}=\left(J_1^{(0)},J_2^{(0)},\cdots,J_M^{(0)}\right)\in\{0,\ 1\} \tag{4}$$

其中M表示的是指标总数。对这个指标池进行筛选就可以得到按照重要性排序的

指标体系。进行指标筛选实际上需要在一个指标池J中选择一组最优的指标组合，这是另一个算法难题。假定最终指标池有N个特征，仅仅只考虑它们的线性组合的情况，特征组合总数也会达到2^N个，从而成为一个 *NP* 难的问题。因此必须考虑更为有效的筛选方法。当前，Gibbs 抽样为此提供了一个很好的办法。

总结起来可以得到上市公司财务预测的步骤可以表示如下：

第一步：获取上市公司的注册、高管、股东等公开信息，并将其作为图的顶点V;

第二步：通过这些公开信息进一步挖掘得到与上市公司关联的其他企业的关联信息，并将其作为图的顶点v;

第三步：判断图的顶点集合V与v中的任意两个顶点是否存在关联，得到关联图的邻接矩阵A;

第四步：按照表 1 及公式 3 等构建特征池；

第五步：采用 Gibbs 抽样方法进行指标筛选。

根据这一流程最终得到一个最优的基于非财务指标的上市公司财务风险预测结果。我们将这套流程工程化为一套系统“Nofira”.

因此，基于非财务指标的上市公司财务风险预测形式模型为：

$$\text{Nofira} =< V, N, i, j, A, M, J, P(J) >$$

其中，

（1） $V = (V, E)$表示上市公司关联图的顶点集。

（2） $N = \{0,1,2, \ldots\}$表示邻接矩阵的维度，即关联方的总数，$1 \le i, j \le N$表示第i, j个顶点。

（3） $A = (a_{ij})_{N \times N}$表示关联图的邻接矩阵，$a_{ij} = \begin{cases} 1, & \text{存在关联} \\ 0, & \text{不存在关联} \end{cases}$。

（4） M表示指标总数。

（5） $J = (J_1, J_2, \cdots, J_M)$表示描述企业风险的非财务指标的指标池。

（6） $P(J)$表示采用 Gibbs 抽样方法进行指标筛选并计算的抽样转移概率，它可以通过$P(J_s = 1|J_{-s}) = \frac{P(J_s=1, J_{-s})}{P(J_s=1, J_{-s}) + P(J_s=0, J_{-s})}$计算得到，其中$J_s$表示第$s$个特征，$J_{-s}$表示除第$s$个特征之外的其他所有特征的组合。。

（5）还有更好的指标体系吗？

上市公司的健康发展对国民经济具有十分重要的意义，不断更新的监管规则和监管政策使得更好监管上市公司的经营行为成为可能。例如在 2008 年金融危机之后，巴塞尔协议 III 规则下信贷资本的计算要求与融资、保证金、流动性和信贷(XVAs)的计算复杂度相同。然而这永远是“矛与盾之争”，上市公司期望通过不断逃避监管获利，而监管部门则希望不断更新监管规则从而保障市场的健康发展。因此，如何利用大数据构建更好的上市公司风险预测特别是能够预测当前企业正在发生的风险永远是一个开放性的问题。此外，本案例中虽然重新构造了一些指标体系，这些指标体系展现的是图结构的一些局域信息，但是还有一些基于局域感知的指标并没有揽括在内。而图结构本身也具有超大规模，在上面进行一次全局变量的调用需要耗费较长的时间，基于分布式并行计算的图计算框架的工程化系统也需要进一步探索。

三、教学体会

随着计算机科学的不断发展和演变，计算学科涉及到的范围在不断拓展。近 10 年来，以数据为核心的计算不断丰富计算科学的内涵并改变了人们的生产和生活方式。

1、关联关系的表达

让学生认识到直接的 A 导致 B 的因果关系在当前日益纷繁复杂的现象中已经变得不再容易获得。通过相关性建立二者之间的相互关系实际上已经成为一种常规的分析手段。图结构是用来表达关联关系的最有效的方法之一。对图数据的处理已经是当前分布式计算领域最重要的方向之一。在许多行业中，一些非图结构的数据也常常被转换为图数据进行分析。

2、在系统设计过程中需要充分平衡计算效率（efficiency）与计算开销(cost)

随着半导体技术的不断进步，计算机的计算能力在不断增强。大部分在校学生在设计和实现计算系统时，常常默认计算资源总是无限的，无论是 CPU、还是内存亦或是存储资源。通过本案例让学生们认识到计算资源不是无限的，在计算中如何采用更好的策略降低计算开销，无论是从数学上还是本身对计算机体系结构的运用如局部性原理的运用，都是十分必要的。

四、激励、唤醒和鼓励同学们向上的途径

1、养成并保持终身学习的习惯

通过本案例的学习可以看到在貌似复杂的问题背后可以找到简单的解决方案，简单的解决方案背后实际上需要考虑到的问题却又需要平衡多方面的因素，涉及到多方面的知识和技能。在图算法的实现过程中，需要充分考虑稀疏矩阵这种数据结构。隐含在背后还有需要考虑对数据的存储和大规模矩阵的处理需要充分考虑计算机本身的限制，充分利用计算机的性能等因素，它们需要学生自主考虑降低计算开销的方法和策略。只有养成和保持终身学习的习惯才能应对变化的“计算”对象的不断演化。

2、敢于实现从“0 到 1”的突破

在系统设计过程中，可以让学生看到，一些新的变量实际上是根据需要创造出来的，而这些新创造出来的变量在指标体系设计过程中扮演了重要的角色。以此让学生敢于创新、敢于胜利。

五、习题

1、你还能找到哪些稀疏矩阵的存储方法？在稀疏矩阵的存储过程中如果运用局部性原理？实际上局部性原理是各类算法优化的基石。

2、能不能找到一套适用于所有企业风险评估的非财务指标体系？

3、指标的重要性如何去衡量？

致谢：感谢贵州省省级教学内容和课程体系改革项目（编号：2021099）的支持。

参考文献

1．X. Yu, Q. Yang, R. Wang, R. Fang and M. Deng, Data cleaning for personal credit scoring by utilizing social media data: An empirical study, IEEE Intelligent Systems，35(2):7-15, 2020.

2．R. Magdalena, Y. Dananjaya, Effect of related-parties transactions to the value of enterprises listed on Indonesian stock exchange. European Journal of Business and Management, 7(6): 47-57, 2015.

3．J. Yan, F. An, R. Wang, L. Chen, X. Yu, M. Deng, Social network analysis of coauthor networks in inclusive finance in China, Y. Wang et al.(Eds): Data Science. ICPCSEE 2022, CCIS 1628, pp.111-122, 2022.

4．G. Yuan, and H. Wang, The general dynamic risk assessment for the enterprise by the hologram approach in financial technology, International Journal of Financial Engineering, 6(1):1950001,2019.

5. P. Chen, R. Lei, and M. Deng, The influence-maximizing node selection algorithm based on local isolated centrality in large-scale networks implemented by distributed graph computing, 2022 IEEE 24th Int. Conf. on High Performance Computing & Communications, 2022, 1234-1240.