基于网络结构模型和非监督学习的指数增强策略

Vince Ji

2020.7

摘要

研究背景

传统金融模型中通常运用相关性和协方差矩阵等模型来对股票间相互作用的关系进行建模。但是现有的模型被发现对估计误差极其敏感，虽然各类研究提出了各种调整的方法使预测值尽量逼近真实值，一些常见的应用，如均值方差组合优化，在实际操作中还是会有样本内表现不俗、样本外表现差强人意的情况。本文的主要目的是从传统金融建模的线性思维中更进一步，利用工程学的网络模型与金融系统的相似性进行非线性高维度的建模。通过非线性高维度的模型获得信息优势，从而更好的构建投资组合并获得更高的收益。

网络结构模型

我们用图结构 G(V,E)里的节点 V 来表示每一个在股票池里的股票，用边 E 来表示股票之间的关系。对于每条边的长度，我们用基于互信息度量$I_S(X,Y)$的互信息距离$\mathrm{d(X,Y)} = 1 - \sqrt{1 - \exp(-2I_S(X,Y))}$来衡量。通过计算每对节点的互信息度距离，构造图结构(graph)的邻接矩阵。我们可以通过对互信息距离设定一个阈值来对完全连接的图结构进行初步剪枝，从而得到一个复杂度略低的高度连接网络。接下来我们更进一步运用最小生成树算法对网络进行进一步化简，在保证系统基本结构的情况下、剔除尽可能多的冗余信息。最后，我们基于最小生成树计算每个节点的中心性，并用中心性构建投资组合。

基于网络结构模型的指数增强策略

我们选用沪深 300 作为基准，使用网络结构的紧密中心性、度中心性和介数中心性分别选取沪深 300 成分股中心性前 20%的“中心资产”构建组合。我们选用的三种中心性构建的组合相对于基准沪深 300 的表现都有所增强。其中介数中心性在 A 股市场表现最好，自 2013 年 7 月至 2020 年 7 月，获得了 187.96%的累计收益，0.22 的夏普率，16.31%的年化收益，相对于基准沪深 300 获得了 5.05%的年化超额收益和 53.57%的月度胜率

目录

1. 研究背景 3
2. 构建网络结构模型 3
2.1 基于互信息（mutual information）的节点间关系及邻接矩阵构建 4
2.2 最小生成树（minimum spanning tree） 5
3. 基于网络结构模型的指标 6
3.1 中心性的量化衡量 7
3.2 中心性与市场机制关系的探索 8
4. 基于网络结构模型的指数增强策略 12
4.1 基于不同中心性度量组合优化的回测表现 12
5. 总结 13
6. 附录 14

图 1：完全连接网络 6
图 2：最小生成树网络 6
图 3：介数中心性-高中心性组和低中心性组均值的时序变化 8
图 4：介数中心性-高中心性组和低中心性组标准差的时序变化 9
图 5：度中心性-高中心性组和低中心性组均值的时序变化 9
图 6：度中心性-高中心性组和低中心性组标准差的时序变化 10
图 7：紧密中心性-高中心性组和低中心性组均值的时序变化 10
图 8：紧密中心性-高中心性组和低中心性组标准差的时序变化 11
图 9：基于中心性的指数增强策略 12
图 10：基于中心性的组合相对基准沪深 300 的回测 13

表 1：各中心性未来 1 个月平均收益 11
表 2：沪深 300 中心性指数增强回测表现 13

1. 研究背景

股票价格的波动并不是独立的，而是会相互作用的。传统金融模型中运用相关性和协方差矩阵来对这种相互联系、相互作用的关系进行建模。在这种传统建模思维下，主流的应用包括马科维茨的均值方差模型（mean-variance），利用组合内股票价格间的协方差矩阵来优化组合配置权重，以期望获得更优的风险和期望回报平衡。目前业内主流的组合优化工具如 Barra 和 Northfield 等也是基于以上建模思想进行拓展和延申的。随着研究的深入，现有的模型被发现对估计误差极其敏感，虽然各类研究提出了各种调整的方法使预测值尽量逼近真实值，但是在实际操作中还是会有样本内表现不俗、样本外表现差强人意的情况。比如 DeMiguel et al(2009)就展示过等权配置的组合在样本外就能超过组合优化后的表现。除此之外，基于传统金融模型主要集中于对线性关系的捕捉，而资产之间的关系并不局限于线性关系，也有非线性关系。因此我们试图寻找合适的方法建模来弥补传统模型的缺点。

学术研究表明，作为复杂系统之一的金融系统通常是层级结构的(Simon,1962)。网络结构模型(network-based model)作为层级结构模型的代表因为和金融系统的相似性，由 Mantegna(1999)首次提出并运用于金融系统的建模。但在当时网络结构模型虽然可以很完整的呈现系统中复杂的关系，由于模型的复杂程度，研究也仅限于描述性的定性分析。而随着网络理论(network theory)的发展，我们获得了更高效的算法和工具(Borgatti, 2005)，如互信息度量、最小生成树的非监督学习层级聚类算法和网络中心性指标，可以对大规模网络结构进行更深入的量化分析，而不仅限于描述性研究。通过将网络结构模型的建模思想和网络理论的工程工具相结合，我们尝试获取传统模型的线性思维所不能发掘的非线性信息。

本文的主要目的是从传统金融建模的线性思维中更进一步，利用工程学的高维度网络模型与金融系统的相似性进行非线性高维度的建模。通过非线性高维度的模型获得信息优势，从而更好的构建投资组合并获得超额收益。下文首先介绍了基于互信息度量的股票间相互作用关系（包括线性和非线性关系）以及基于此关系构建的网络结构模型；其次介绍了机器学习中非监督学习的最小生成树算法，用于降低之前所构建的高度连接网络模型的复杂度；然后介绍了基于最小生成树的中心性度量以及探索了中心性与市场机制间的相互作用关系；最后检验了基于网络模型中心性的指数增强策略。

2. 构建网络结构模型

因为股票市场构建网络模型时最基本的数据结构是图 G(V,E)，其中 V 代表顶点，E 代表边。我们用图结构里的顶点来表示每一个在股票池里的股票，用边来表示股票之间的关系（若两股票之间有关则代表两股票的顶点之间有边相连，若没有关系则两顶点之间没有边）。两顶点间的距离或边的长度（权重）取决于顶点所代表的股票回报率时间序列之间的关系。在实际操作中我们可以对边的权重设一个阈值。在阈值低于一定水平时，我们可以剪除这条边，忽略连接的顶点之间的关系。

2.1 基于互信息（mutual information）的节点间关系及邻接矩阵构建

首先我们需要定义节点与节点间的关系。常用的 Pearson 相关性系数只适合用于量化线性的相关性，受限于 Pearson 相关性系数的计算原理，非线性相关性是很难被其捕捉到的。而在金融系统中随着我们研究越来越深入，我们发现单纯的线性关系是很难完全描述复杂的金融系统之间的关系的。事实上，越来越多的非线性关系在量化研究中被发掘出来，因此如果能量化非线性相关性，能够给我们带来更进一步的信息优势。这里我们使用互信息率这一指标来描述包含线性和非线性的关系。且因为收益率之间的相关性函数并不已知，互信息率这样的非参指标就非常适合用于描述这样的关系。简单来说互信息率可以量化两个随机变量之间共享了多少信息，从而很好的捕捉到随机变量间的线性和非线性关系。互信息率是从信息理论中熵的概念延伸出来的，熵的表达形式如下：

$$\mathrm{H(X)} = -\sum_i p_x(x_i)\log p_x(x_i)$$

其中$p_x()$是 x 的概率密度函数。此时若有另一个随机变量 Y，令$p_{x,y}()$为 X 和 Y 的联合概率密度函数，那么 X 和 Y 的联合熵为：

$$\mathrm{H(X,Y)} = -\sum_i p_{x,y}(x_{i,}, y_{i,})\log p_{x,y}(x_i, y_{i,})$$

那么我们可以用熵的表达式来定义互信息率：

$$I_S(X,Y) = H(X) + H(Y) - H(X,Y)$$

$$I_S(X,Y) = \sum_{y\in Y}\sum_{x\in X} p_{x,y}(x,y) log\frac{p_{x,y}(x,y)}{p_x(x)p_y(y)}$$

通过以上步骤计算出的互信息率是非负的，当且仅当 X 和 Y 相互独立时为 0。在有了互信息率这个指标之后，我们就可以着手准备构造邻接矩阵了。在构造邻接矩阵之前，我们需要知道每一个顶点之间的距离。基于互信息率，我们可以计算出这个互信息距离 d(X,Y)，其表达形式如下：

$$\mathrm{d(X,Y)} = 1 - \sqrt{1 - \exp(-2I_S(X,Y))}$$

以上的互信息率距离 d(X,Y)取值在 0 到 1 之间。和互信息率相反的是，距离 d(X,Y)越小说明 X 和 Y 所包含的相似信息越多，因此在网络结构中的距离就越近，反之亦然。d(X,Y)=0 时说明 X 和 Y 完全相关，而 d(X,Y)=1 时，说明 X 和 Y 互相不会有任何影响。

在定义了节点与节点之间的关系之后，我们就可以开始构建网络模型了。模型中最关键的部分是确定任意两节点间是否有边相连。此时我们可以用邻接矩阵来表示节点与节点之间的关系，1 为连接，0 为不连接。最简单的方法是构建一个完全连接的模型，即邻接矩阵里所有的元素都为 1。这里我们可以更进一步，对一些非常弱的互信息做选择性剪除。我们可以事先设定一个阈值d_f，令 m 为邻接矩阵里的元素，x 和 y 为系统中任意两个节点，则：

$$m_{x,y} = \begin{cases} 1, d(x,y) < d_f \\ 0, \text{Otherwise} \end{cases}$$

2.2 最小生成树（minimum spanning tree）

取决于阈值的设定和邻接矩阵和互信息距离，我们可以构建出一个高度甚至完全连接的网络模型。但是一个有 n 个节点的完全连接网络有 n(n-1)/2 条边，假如我们以沪深 300 的成分股来建模，此时我们就有 44850 条边。如此复杂的系统会使我们接下来的分析变得极为困难，因此我们要想办法降低模型的复杂程度。这里我们考虑使用非监督学习中的最小生成树算法，通过把高度连接的网络剪枝，舍弃一部分连接，从而获得一个树状结构的网络是网络分析中常用的做法。我们获得的树状模型的每个节点还是能通过其他节点互相连接的，而节点与节点之间的距离之和是我们能获得的最小值。从另一个角度来说，我们对整个复杂的网络剔除冗余的信息，在不改变其结构本质的情况下保留了尽可能多的信息。

最小生成树的数学表达如下。用图 G=(V,E)来表示一个完全连接的网络，其中 V 是顶点，E 是边，d 是边的权重（距离）。此时我们要找到一个新的生成树 Gt=(Vt,Et)使顶点与顶点间的总距离最小。令：

$$x_e = \begin{cases} 1, & e \in Et \\ 0, \text{Otherwise} \end{cases}$$

则寻找最小生成树转化为以下优化问题：

$$\begin{aligned} \underset{x_e}{\text{minimize}} \quad & \sum_{e \in E} d_e x_e \\ \text{subject to} \quad & \sum_{\varepsilon \in E} x_e = n-1 \\ & \sum_{e \in (S,S)} x_e \leq |S| - 1, \forall S \subset V, S \neq \varnothing, S \neq V \\ & x_e \in \{0,1\}, \forall e \in E \end{aligned}$$

其中（S，S）表示所有的边，|S| 表示 S 的基数。第二个约束条件表示生成树不能有闭环。

为了便于可视化，我们在沪深 300 成分股中随机取 20 支股票组成的网络举例，图 1 是完全连接的网络模型，图 2 是基于完整网络得到的最小生成树。

图 1：完全连接网络

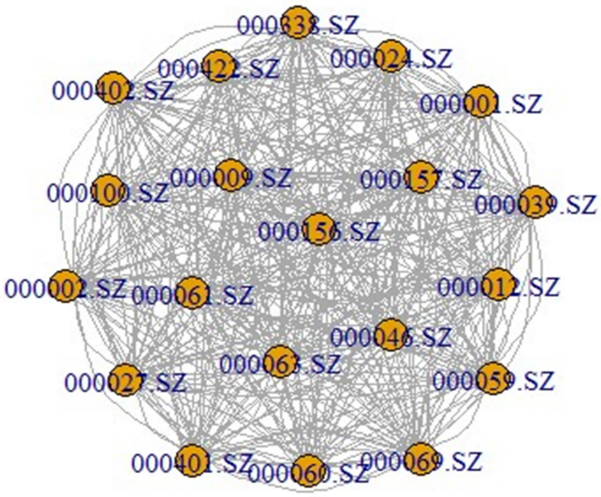
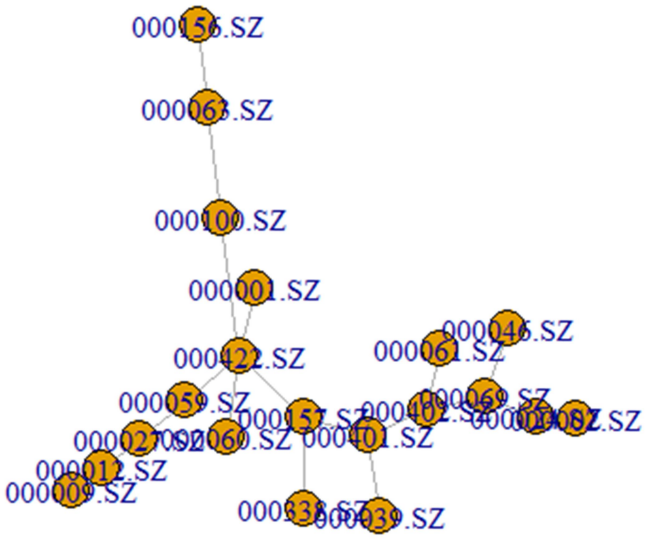


图 2：最小生成树网络



3. 基于网络结构模型的指标

我们构建网络模型的目的是为了更好的描述股票之间的关系，从而帮助我们更有效的构建投资组合。因此，如果我们的分析仅仅止步于最小生成树，对最小生成树做描述性的分析是不行的。我们需要通过最小生

成树获得一些可量化的指标帮助我们描述网络模型里股票间的关系。在网络分析中，最常用的一类量化指标就是中心性。中心性描述的是给定节点相对于其他节点的中心程度，通常周边连接了越多节点的节点其中心性越高，而位置越靠边缘的节点中心性越低。引申到股票网络里来，我们可以认为中心性衡量的是给定股票对整个系统的重要程度。中心性越高说明该股票对整个系统的重要性越大，因此我们可以称这类中心性高的股票为“中心资产”。反之，所处位置越边缘，对系统的影响程度越小，我们可以认为这类股票为“边缘资产”。在网络分析领域，有一些常用的中心性指标(Borgatti, 2005)，经过长期的实践检验可以很好的量化网络系统里的中心性。这里我们选取了如下三个指标帮助我们进一步的分析。

3.1 中心性的量化衡量

度中心性（degree centrality）

度中心性是最常用的中心性，它衡量的是一个节点与其他节点发生直接联系的程度。如果一个节点与其他很多节点发生直接联系，那么这个节点就处于中心地位。即节点的关系越广，相邻节点越多，那么节点也就越重要。通常为了便于比较或者进行其他计算，需要将度中心性进行标准化。标准化的方式通常是每个顶点的度除以途中可能的最大度数，即 n-1，令 degree(i)为顶点 i 的度，其表达式如下

$$C_d(\mathrm{i}) = \frac{\mathrm{degree(i)}}{n-1}$$

紧密中心性（closeness）

紧密中心性反应的是某个节点与其他节点之间的接近程度。如果一个节点离其他节点越近，那么他影响其他节点的能力就越强。这个点的紧密中心性基于该点到网络中其他所有节点的最短路径之和，等价的我们也可以通过求这个节点到其他所有节点的平均最短距离来表示。一个节点的平均最短距离越小，那么该节点的紧密中心性越大，令 *l(i,j)*为节点 i 和节点 j 之间的最短路径长度，则为了方便比较我们用平均最短距离的倒数定义为该节点的紧密中心性，表达式如下:

$$C_c(i) = \frac{n-1}{\sum_{i \neq j} l(i,j)}$$

介数中心性（betweenness）

介数中心性是指某节点出现在其他节点之间的最短路径的个数。如果这个节点的介数中心性高，那么它对整个网络结构的转移会有很大的影响，考察的是节点对其他节点信息传播的控制能力。介数中心性的求解过程可以分为三个部分：1.计算每对节点 (j,k)之间的最短路径，以及记录该路径所经历的节点； 2. 对每个节点 i 判断出现在上一步中(j,k)间的最短路径集合中的次数占最短路径总数的比例；3. 最后对所有节点累加节点 i 在第二步中的比例从而获得节点 i 的介数中心性，表达式如下:

$$C_b(\mathrm{i}) = \sum \frac{d_{j,k}(i)}{d_{j,k}}$$

为了方便和其他中心性对比以及之后的计算，我们要进行归一化，归一化后的表达式如下：

$$C_b(i) = \sum \frac{d_{j,k}(i)/d_{j,k}}{(n-1)(n-2)}$$

3.2 中心性与市场机制关系的探索

我们基于沪深 300 建模，自 2013 年 7 月至 2020 年 7 月，每一期使用过去 360 天的交易数据计算了以上 3 个指标。经过分析我们发现，基于 A 股的中心性表现出很强的“头部效应”，即中心性最强的前 20%股票的平均中心性要远远高于剩下的 80%，且剩下中心性较弱的 80%股票的平均中心性相差不大。基于以上的实证观察我们可以认为，A 股的网络系统由少数“中心资产”组成系统的“骨架”，而其他的“边缘资产”通过“中心资产”与其他股票以及整个系统进行互动。

我们把沪深 300 成分股按 20%/80%分为高中心性和低中心性两组，它们的平均时序变化如下图 3-图 8 所示：

图 3：介数中心性-高中心性组和低中心性组均值的时序变化

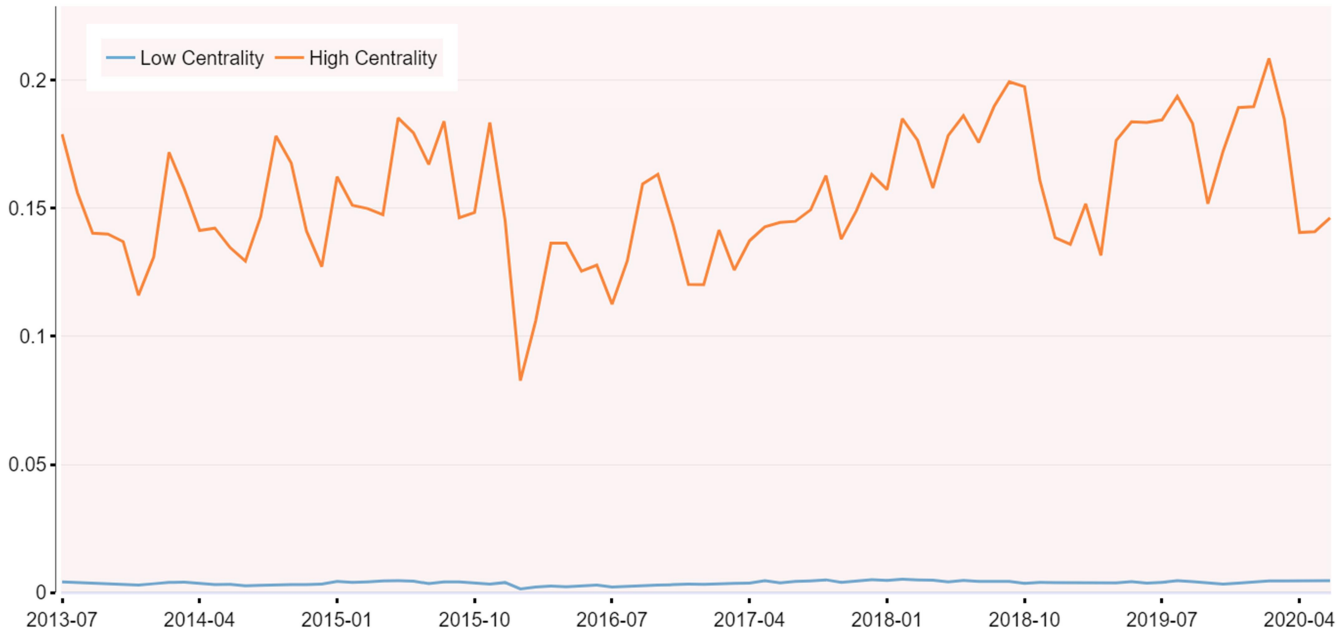


图 4：介数中心性-高中心性组和低中心性组标准差的时序变化

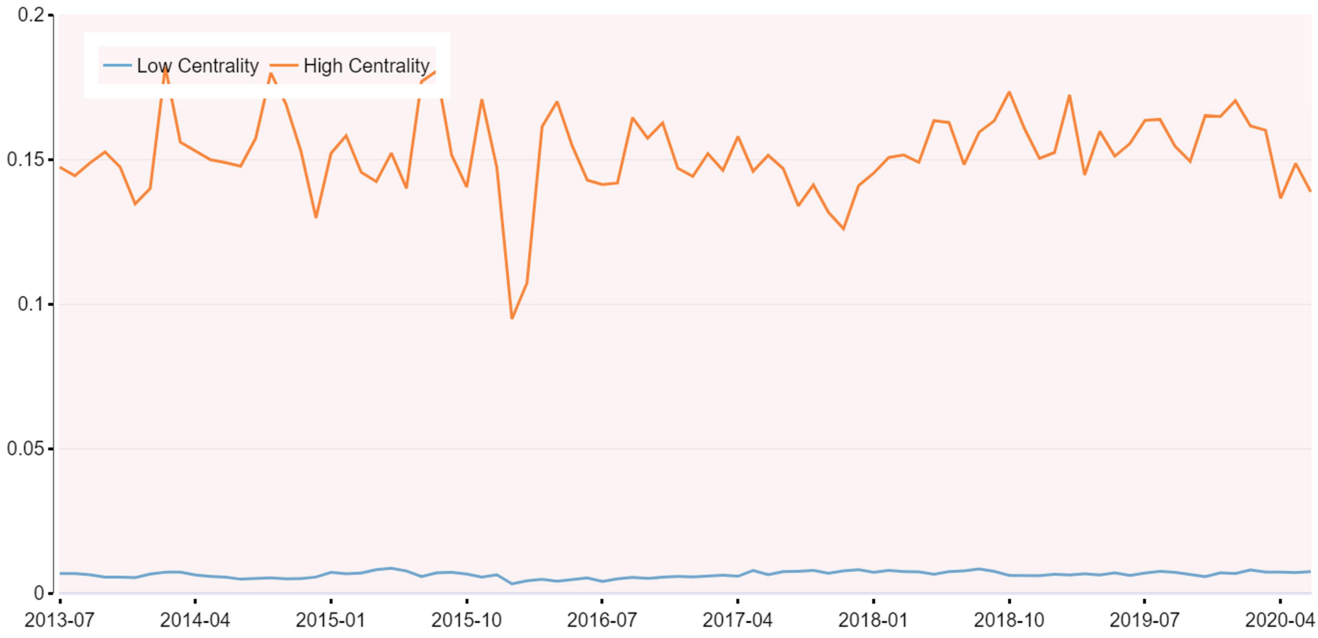


图 5：度中心性-高中心性组和低中心性组均值的时序变化

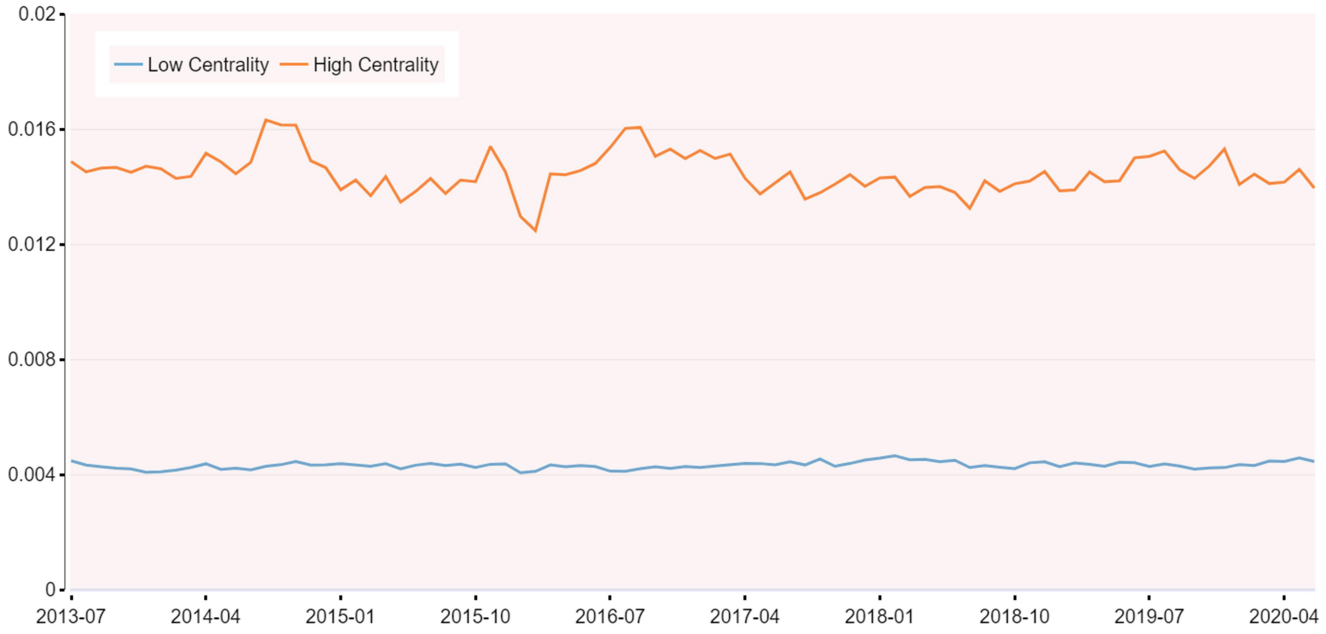


图 6：度中心性-高中心性组和低中心性组标准差的时序变化

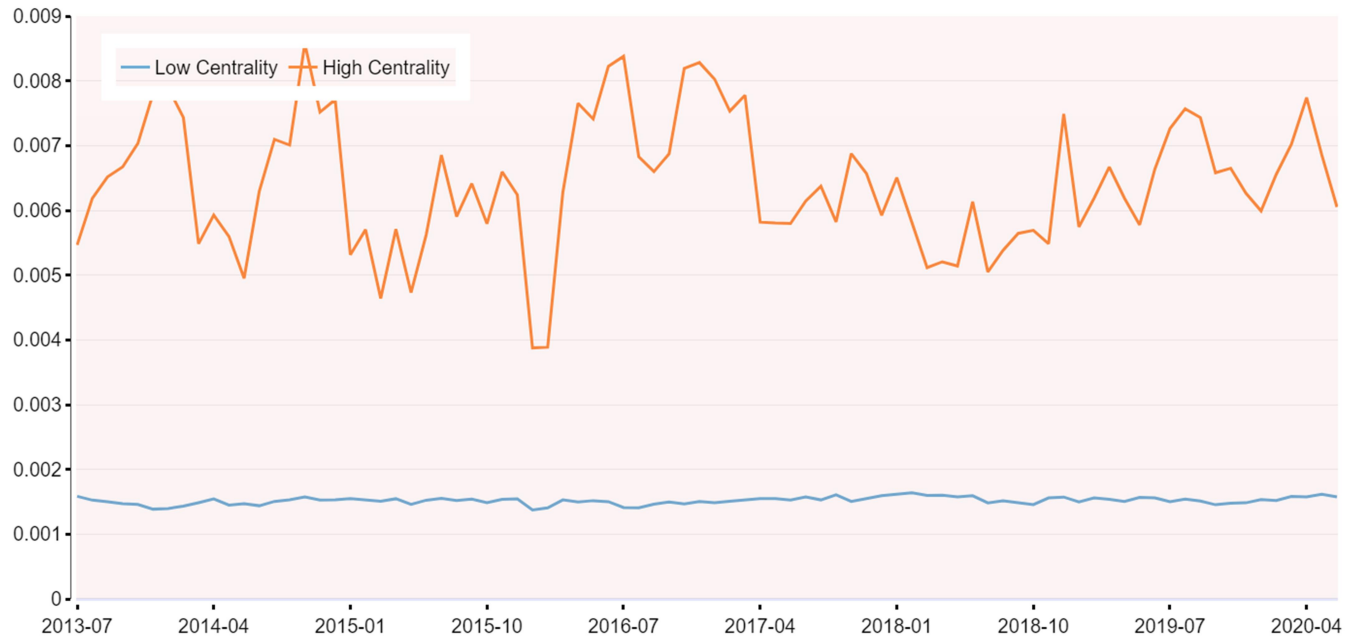


图 7：紧密中心性-高中心性组和低中心性组均值的时序变化

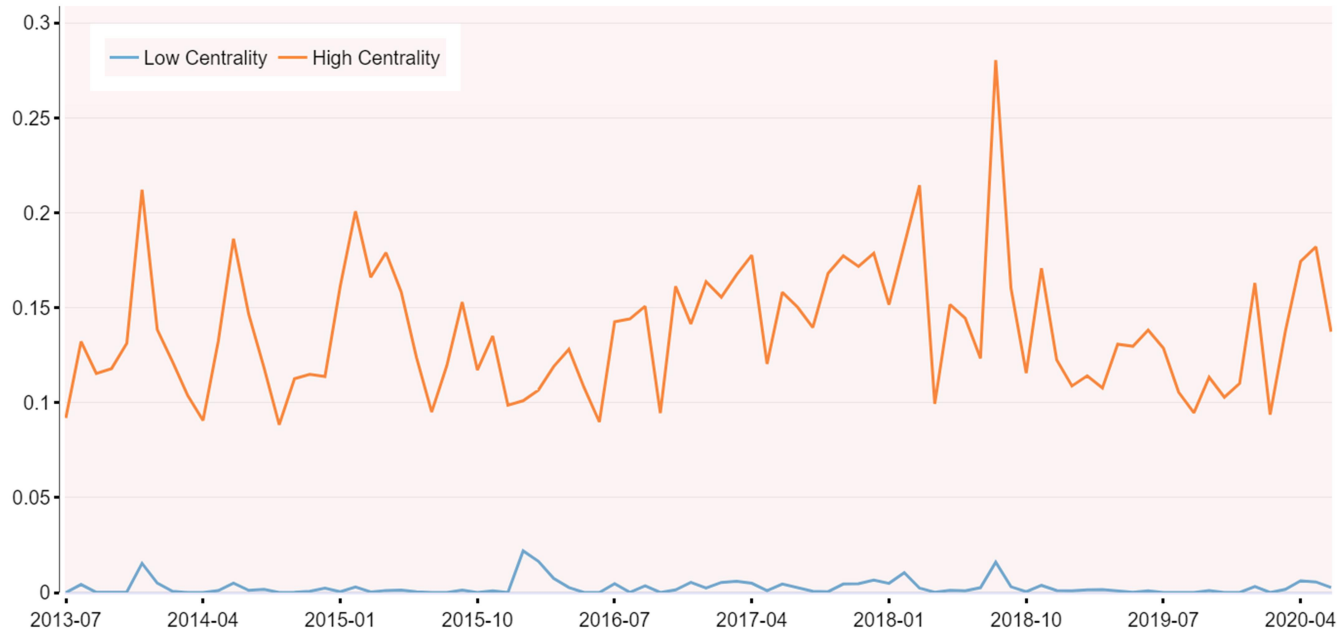
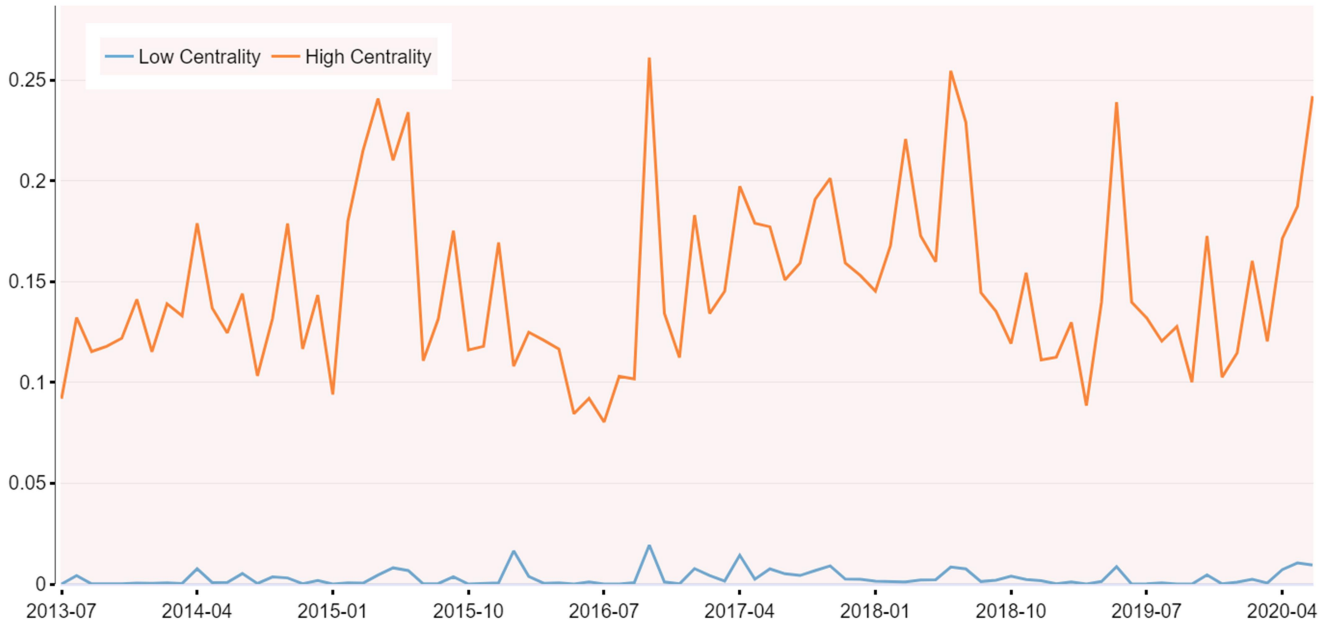


图 8：紧密中心性-高中心性组和低中心性组标准差的时序变化



通过观察我们还发现，基于这三种度量的低中心性组时序变化相对平稳，而高中心性组时序变化波动较大。我们认为这可能是由于投资者对不同主题的追逐导致市场结构性变化导致的。Kaya(2015)在美股的高中心性组中发现了相对于低中心性组的风险溢价，因此基于我们对 A 股市场的中心性变化的解读，我们可以大胆假设在 A 股中也可能存在这种高中心性风险溢价。

为了验证我们对于中心性和股票回报率之间关系的假设，我们把中心性分为两部分，头部 20%的高中心性组和尾部 80%的低中心性组，来分析中心性的风险溢价。

表 1：各中心性未来 1 个月平均收益

	紧密中心性	度中心性	介数中心性
高中心性	0.260%	0.322%	0.365%
低中心性	-0.065%	-0.047%	-0.089%

从表 1 我们发现对应紧密中心性、度中心性和介数中心性的高中心性组的平均收益分别为 0.260%、0.322%、0.365%，而低中心性组的平均收益分别为-0.065%、-0.047%、-0.089%。因此我们可以认为从实证的角度来说高中心性的股票相对于低中心性的股票有更高的风险溢价。

4. 基于网络结构模型的指数增强策略

4.1 基于不同中心性度量组合优化的回测表现

基于以上对于中心性的观察，即中心性集中在约 20%的“中心资产”，且“中心资产”相对于“边缘资产”存在风险溢价，为了获得超过基准的表现，我们应该希望在组合里加入“中心资产”。顺着上述思路，最直接的办法就是使用中心性的大小前 20%的股票等权重构建组合。这里我们选用沪深 300 作为基准，分别使用紧密中心性、度中心性和介数中心性选择沪深 300 成分股中心性前 20%的股票构建组合。组合的表现如图 9、图 10 和表 1 所示。

图 9 、 图 10 中 HS300 代表沪深 300 指数，Degree 代表度中心性组合，Closeness 代表紧密中心性组合，Betweenness 代表介数中心性组合。

图 9：基于中心性的指数增强策略

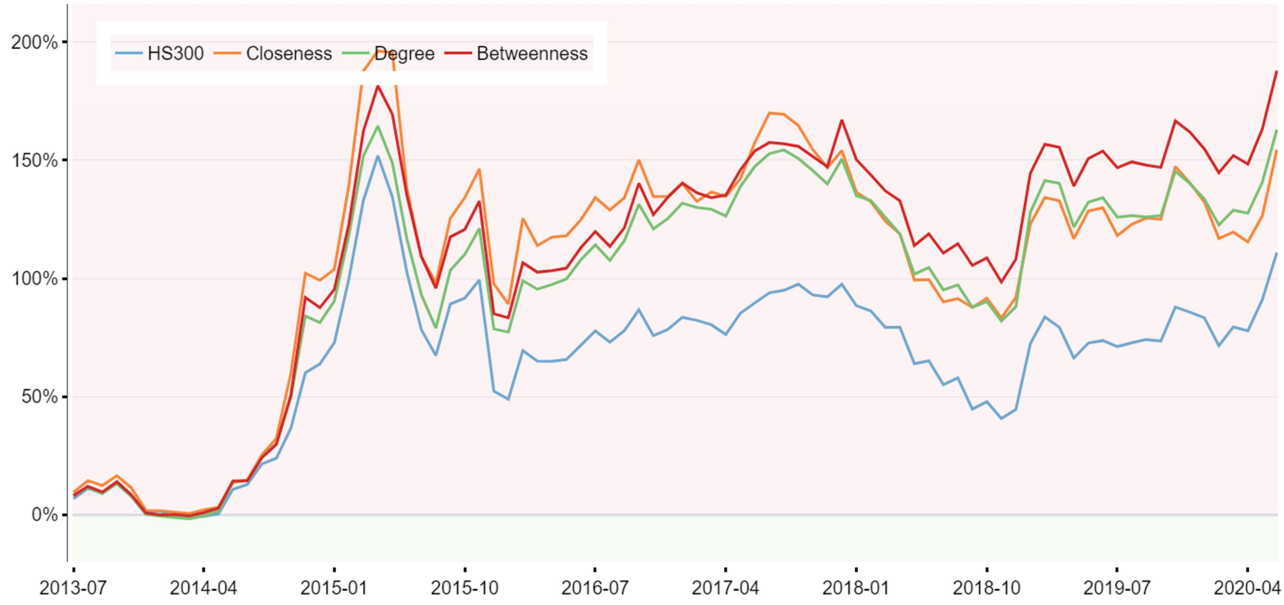


图 10：基于中心性的组合相对基准沪深 300 的回测

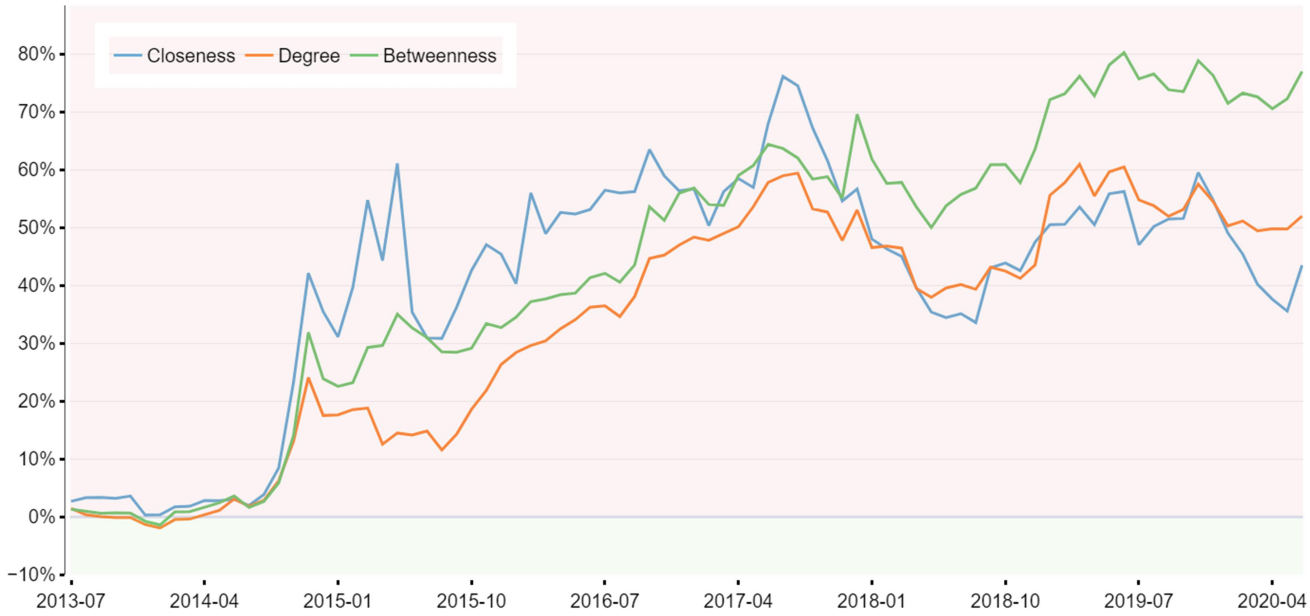


表 2：沪深 300 中心性指数增强回测表现

	沪深 300	紧密中心性	度中心性	介数中心性
累计收益	111.00%	154.49%	162.97%	187.96%
年化收益	11.30%	14.28%	14.81%	16.31%
夏普率	0.17	0.18	0.20	0.22
超额收益	-	3.02%	3.56%	5.05%
月度胜率	-	52.38%	59.52%	53.57%

5. 总结

本文介绍了通过网络模型研究股票市场的方法。在构建网络阶段，通过互信息距离度量构建邻接矩阵，再运用非监督学习的最小生成树算法获得降低了复杂度的网络模型，最后运用三种中心性度量，即度中心性、紧密中心性、介数中心性，选出“中心资产”，构建组合增强策略。我们选用的三种中心性构建的组合相对于基准沪深 300 的表现都有所增强。其中介数中心性在 A 股市场表现最好，自 2013 年 7 月至 2020 年 7 月，获得了 187.96%的累计收益，0.22 的夏普率，16.31%的年化收益，相对于基准沪深 300 获得了 5.05%的年化超额收益和 53.57%的月度胜率。

综上所述，网络结构模型运用高维度非线性建模思维与网络理论的工程工具相结合，构建的基于此模型的指数增强应用在 A 股获得了不错的表现。我们将持续探索网络结构在金融建模中的应用，并在后续研究中持续报告相关的研究成果。

6. 附录

参考文献

[1] Mantegna, R. N. 1999. “Hierarchical Structure in Financial Markets.” The European Physical Journal B: Condensed Matter and Complex Systems 11 (1): 60 − 70.

[2] Borgatti, S. P. 2005. “Centrality and Network Flow” Social Networks 27: 55 − 71.

[3] Prim, R. C. 1957. “Shortest Connection Networks and Some Generalizations.” Bell System Technical Journal 37: 1389 − 1401.

[4] Pozzi, F., T. Di Matteo, and T. Aste, 2013, Spread of risk across financial markets: better to invest in the peripheries, Nature Scientific Reports 3:1665, 1 − 7.

[5] Kaya, H. 2015. “Eccentricity in Asset Management.” Journal of Network Theory in Finance 1 (3): 1 − 32.

[6] Peralta, G., & Zareei, A. 2016. A network approach to portfolio selection. Journal of Empirical Finance, 38, 157 − 180.

[7] Baitinger, E., and J. Papenbrock. 2017. “Interconnectedness Risk and Active Portfolio Management: The Information- Theoretic Approach.” Journal of Network Theory in Finance 3 (4): 25 − 47.

[8] Simon, H. A. 1962. The architecture of complexity. In Proceedings of the American Philosophical Society, volume 106, pages 467{482.