arXiv:2512.18918v1 [cs.SI] 21 Dec 2025

大海捞针：运用取证网络科学揭露内幕交易

Gian Jaeger[1*]，Wang Ngai Yeung[2] 和 Renaud Lambiotte[3]1*牛津大学牛津互联网研究所，英国牛津。[2]复杂连接实验室、网络科学研究所，英国伦敦。[3]牛津大学数学研究所，英国牛津。

*通讯作者。电子邮箱：gian.jaeger@oii.ox.ac.uk；贡献作者：
j.yeung@northeastern.edu；
renaud.lambiotte@maths.ox.ac.uk；

摘要

尽管近年来反金融犯罪调查的自动化和数字化取得了显著进展，内幕交易的侦测仍然面临独特挑战，部分原因在于标注数据有限。为应对这一挑战，我们提出采用一种数据驱动的网络方法，标记出报告协同交易、且其行为表明可能涉及内幕交易的公司内部人士群组。具体而言，我们利用公司内部人士（高管、董事会成员和主要股东）在2014年至2024年间向美国证券交易委员会（SEC）报告的290万笔交易数据。我们提出的算法根据内部人士在这十年间交易的时间相似性，在他们之间构建加权边。随后，我们在该网络中通过聚焦中心节点和异常子图，揭示表明内幕交易的趋势。为凸显本方法的有效性，我们参照两个零模型评估研究结果；这两个零模型是通过在合成的、经经验校准的数据集和打乱的数据集上运行算法生成的。结果表明，该方法可用于检测行为显示其可能涉及内幕交易和／或市场操纵的内部人士配对或群组。

关键词：网络分析、异常检测、图推断、金融图、内幕交易

1 引言

最常见的内幕交易发生于公司内部人（即员工、董事或管理人员）利用其在组织中的职务，获取公众无法获得的信息，并据此获利交易。据估计，这种被称为传统内幕交易的内幕交易形式，伴随五分之一的并购交易和二十分之一的盈利公告（来源）。尽管这种现象十分普遍，但它仍以难以检测而著称；2024年，美国证券交易委员会（SEC）仅起诉了34起案件（Jones 等，2024）。大量研究致力于探讨其不利影响，相关研究表明，内幕交易会降低金融市场的流动性，并对管理层激励产生不利影响（Fishman 和 Hagerty，1992；Bhattacharya，2014）。因此，检测并遏制内幕交易显然受到政府和公众的共同关注。

为遏制此类行为并提高金融市场的透明度，美国与大多数其他国家一样，依法要求高层公司内部人士在交易发生后两个工作日内，通过提交 Form 4申报，向美国证券交易委员会报告购买或出售其所属公司发行的任何金融证券（证券和交易委员会，未注明年份）。这包括高管（例如首席执行官和首席财务官）、董事会成员以及主要股东。需要注意的是，这些交易通常并不违法；公司内部人士往往以股票和股票期权的形式获得相当一部分薪酬。在特定时间窗口内，他们可以出售这些证券或购买额外证券；当内部人利用其职位所能接触到的重大非公开信息进行交易时，该交易才构成违法（Fried，2013）。

监管机构使用多种工具来发现内幕交易，包括跨市场监控系统（综合审计追踪系统，CAT）、定向数据请求（电子蓝单）以及通过美国证券交易委员会举报人计划报告的举报和转介（Arif 等，2022）。此外，Form 4披露也被用于筛查影响市场的事件前后的集群交易、逾期申报，以及相关内部人士之间的协同行为模式（Tamersoy 等，2014年）。标准做法是基于事件的异常检测，即调查人员筛查影响市场的事件前后短时间窗口内获利异常的交易和交易量的异常激增（Priyadarshi 和 Kumar，2024）。随后，调查人员会优先对在这些窗口内被标记的账户进行后续调查。

这种方法的主要局限在于侧重识别个人。这使得有效起诉变得困难，因为当局必须证明被告能够接触并使用重大非公开信息，才能达到证明其具有主观故意的举证要求；而被告则可以令人信服地将收益归因于技能、运气或公开信息。为弥补这一不足，我们提出一种基于网络的筛查方法，用于标记那些在交易模式上呈现持续且在统计学上难以解释的时间协调行为的公司内部人士配对及簇。从检察官的角度看，该方法为现有方法提供了补充证据，使团体层面的起诉成为可能；连带责任和交叉印证则提高了至少有一名内部人士配合的可能性。

我们利用了2014年至2024年间美国公司内部人士向美国证券交易委员会报告的290万笔交易数据，构建了一个网络，其中边权重表示

公司内部人士交易模式的时间相似性（证券交易委员会，未注明日期）。我们的方法基于这样一个前提：如果两名或更多内部人士持续在同一时间范围内报告卖出和买入，那么他们可能是在相互交换信息，或者依据双方都能获取的内幕信息进行交易。这里的区别在于，前者涉及某种形式的沟通，而后者只要求能够获取相同信息的内部人士利用这些信息进行类似交易（Cohen et al. 2012）。在这一网络中，只有在特定时间点隶属于同一家公司的两名个人之间才能建立联系，因此网络不可避免地相对分散。然而，如果一名内部人士同时在多个董事会任职，或在这10年期间曾为多家公司工作，那么连通分量可能包含来自不同公司的节点；这两种情况都是企业网络中信息传播的重要驱动因素（Ahern 2017）。

我们的算法具有以下特点：(i) 可解释，(ii) 可扩展，(iii) 可泛化，因为它能够应用于任意一组带时间戳的交易，从而推导出协同行为的基于图的表示。为评估其有效性，我们将生成的网络与两个生成式零模型进行比较。第一个是经过经验校准的合成数据集，第二个是实际数据的打乱版本。因此，我们在两个零模型下均保持显著的经验观测网络特征，代表的协同行为既不能归因于监管合规，也不是对外生信号的偶然响应。随后，我们参照中心节点和异常自我网络检查实际网络中的可疑活动，发现了一些有趣的现象，包括两个由家庭成员组成、持续协同交易的簇。通过展示数据驱动的网络如何在较长时间跨度内捕捉高阶关系，我们的研究为反金融犯罪调查提供了创新框架。该方法无需带标签的训练数据，即可针对异常行为提供可解释、可扩展的洞见。

2 相关研究

网络分析为表示交易数据提供了自然且有效的框架。因此，许多学者强调，网络分析有助于增强金融情报机构的行动能力，尤其是在打击洗钱、恐怖主义融资和欺诈检测方面（Basu 等，2018；Li 等，2022；Oliveira 等，2025）。与此相应，越来越多的研究开始关注公司内部人士网络。这一领域的研究主要集中在两个方面。其一，研究内部人士在更广泛的公司环境中的连接与其交易盈利能力之间的相关性（El-Khatib 等，2015；Goergen 等，2019）。这些研究探讨的是：与公司影响力较小的高管相比，那些人脉广泛、能够获取同一行业多家公司的信息的董事会成员，是否会报告系统性地更为盈利的交易。其二，研究公司内部人士在公司中的角色如何影响交易披露后的股价（Berkman 等，2020；Afzali 和 Martikainen，2021）。这些研究基于这样一个前提：公司高管公开披露交易，可能会对股价产生重大影响（El-Khatib

等，2021）。随后，这些研究着重评估某些内部人士报告的交易是否会产生更显著的影响（例如，首席技术官的交易与首席执行官的交易相比）。

据我们所知，此前只有两项研究利用基于图的数据挖掘技术，来识别这些数据中的内幕交易趋势（Tamersoy 等人，2014；Kulkarni 等人，2017）。这两项研究都使用 Form 4申报文件，根据个人在同一天进行交易的频率，构建个人之间的加权边。我们的工作遵循相同的思路，但针对其方法中的若干局限进行了改进。首先，我们不要求交易日期完全匹配，而是采用周线性核函数，对相隔数日的交易给予部分匹配分，从而使度量能够抵御申报延迟和轻微的时间差异。其次，我们不再对日期集合中的精确重叠进行计数，而是使用基于匹配的相似度函数，通过有界周核函数将每个交易日与最接近的对应日期对齐，从而避免重复的近似巧合造成组合数量膨胀，并得到一个连续得分： $[0, 1]$。第三，我们分别计算基于买入和卖出的相似度，并根据交易活动加权后合并，使任一方向的对齐都能为总体边权重作出贡献。最后，鉴于内幕交易窗口期因公司而异，每年持续 120 至 160 天，我们采用严格的边创建标准，以减少偶然共同发生的情况（Guay 等人，2021）。

此外，由于以往这些尝试仅将其算法应用于经验观测数据，并未进一步验证其方法的统计显著性或可靠性，因此无法做到这一点。在非法活动的真实标签几乎不存在的情况下，这一局限尤为关键，因为这使得标准的监督式评估指标无法适用（Yeung et al. 2025）。为解决这一问题，我们不再局限于描述性分析，而是将研究结果与两种互补的零模型进行基准比较。首先，我们使用校准生成模型，控制交易窗口、内部人任期和公司存续期等结构性约束。借此，我们可以评估协同行为是否完全能够由监管时机解释。其次，我们对底层数据进行受约束的时间置换，在保留公司层面活动的精确序列的同时随机打乱内部人士身份，从而评估同步现象是否仅仅反映了对内幕交易以外外生因素的共同反应。因此，结合现有文献，我们的贡献可概括如下：

- 我们提出了一种基于网络的筛查方法，该方法具有可解释性、可扩展性和良好的泛化能力，并以带有时间戳的 Form 4披露为依据，对公司内部人士之间的协同交易行为进行量化。

- 我们使用两种互补的零模型验证了这一方法，证明网络结构反映的是真实的协同行为，而非偶然的重叠时段。通过控制系统性约束和随机巧合，该方法为在噪声较大且未标注的数据集中识别可疑活动提供了统计依据，且适用范围更广。

- 我们将该方法应用于290万笔美国内部人士交易（2014–2024年），发现了此前未被报告的协调行动内部人士簇，证明了其作为《法证》工具的实际价值。

3 方法

3.1 数据

美国联邦法律要求公司内部人士在交易发生后的两个工作日内，通过 Form 4申报向美国证券交易委员会报告其证券受益所有权的所有变更。这一要求具体适用于上市公司的董事、高管以及持有超过10%股权的股东（证券交易委员会 n.d.）。为分析股票买入和卖出，我们聚焦于该表格的第1部分，该部分可通过美国证券交易委员会的电子数据收集与检索系统（EDGAR）公开获取。我们使用了2014年至2024年间申报的约290万笔交易数据进行分析。为聚焦于个人，并避免高频交易和大额交易量对分析造成影响，我们剔除了机构投资者申报的交易（包括有限责任公司、有限合伙企业和股份有限公司）。此外，我们将涉及同一内部人士、发行人、方向和日历日的多条项目合并为单一事件。数据的汇总统计结果见表1。

统计量	所有交易	按日汇总
内部人士数量	70,941	—
公司数量	9,426	—
交易次数	2,735,932	461,785
买入	1,326,228	243,640
卖出	1,409,704	218,145
内部人平均交易次数	38.6	6.51

表1：原始内部人交易数据（2014–2024）的汇总统计。买入和卖出的规模相近，按日汇总后，观测值数量约减少至原来的六分之一。

3.2 构建网络

个体之间的时间对齐

我们的方法旨在推导边权重，以捕捉所研究的 10 年期间公司内部人士交易模式的相似度。首先，我们生成一个空网络 $G = (V, E)$，其中 V和 E 分别表示节点集合以及节点之间连接的集合。随后，我们通过比较来自 9,426 家公司的内部人士所有可能配对的交易日期，引入节点和边。令 $X_C = \{x_1, \ldots, x_m\}$和 $Y_C = \{y_1, \ldots, y_n\}$表示公司 C两位内部人士的交易日期。由于我们定义的是不同节点之间的边，因此严格要求 $X_C \neq Y_C$。我们使用非负权重 $w(d)$ 来衡量时间邻近性，该权重会随绝对时间间隔 $d = |x - y|$的增大而减小。具体而言，我们使用线性“同周”核函数

$$w(d) = \begin{cases} 1 - \dfrac{d}{7}, & d \le 7 \\ 0, & d > 7 \end{cases}. \tag{1}$$

如此一来，相隔超过一周的事件不会计入相似度得分，而相隔一天的事件会获得权重 $6/7$，同一天发生的事件则获得权重 1。在我们的语境中，线性衰减提供了一种透明且易于解释的方式，可降低时间上相距较远的交易之间的相似度，同时仍聚焦于近乎同时发生的活动。定义好核函数 (1) 后，我们现在可以计算内部人士之间的匹配程度。对于每个 $x \in X_C$，我们在 Y_C中找到与其最匹配的事件，即

$$s_{X|Y} = \frac{1}{|X_C|} \sum_{x \in X_C} \max_{y \in Y_C} w(|x - y|). \tag{2}$$

这是在 X_C 范围内， Y_C 中可获得的接近度的平均值。同样，

$$s_{Y|X} = \frac{1}{|Y_C|} \sum_{y \in Y_C} \max_{x \in X_C} w(|x - y|). \tag{3}$$

因为 $s_{X|Y} \neq s_{Y|X}$ 如果一名内部人士交易更频繁，我们便通过对两个方向取平均来定义一个对称指标。对于任意一对不同的内部人士（其中 $X_C \neq Y_C$）：

$$S(X_C, Y_C) = \frac{1}{2} \left[\frac{1}{|X_C|} \sum_{x \in X_C} \max_{y \in Y_C} w(|x - y|) + \frac{1}{|Y_C|} \sum_{y \in Y_C} \max_{x \in X_C} w(|x - y|) \right]. \tag{4}$$

我们分别计算了（4）买入（A）和卖出（D）：

$$S^{(A)} = S\big(X_C^{(A)}, Y_C^{(A)}\big), \qquad S^{(D)} = S\big(X_C^{(D)}, Y_C^{(D)}\big). \tag{5}$$

随后，我们通过活动加权将二者合并，

$$S_{\text{combined}} = \frac{|X_C^{(A)}| + |Y_C^{(A)}|}{T} S^{(A)} + \frac{|X_C^{(D)}| + |Y_C^{(D)}|}{T} S^{(D)}, \tag{6}$$

其中

$$T = |X_C^{(A)}| + |Y_C^{(A)}| + |X_C^{(D)}| + |Y_C^{(D)}|, \tag{7}$$

表示这两名内部人士进行的交易总数。

解释与性质

最终得分 $S_{\text{combined}} \in [0, 1]$ 是 $S^{(A)}$ 和 $S^{(D)}$的凸组合，表示从合并日志中随机抽取一笔交易时，预期的周核函数匹配率。$S_{\text{combined}} = 1$ 表示每笔交易都有同日对应交易，$S_{\text{combined}} = 0$ 表示一周内没有发生任何匹配。该度量具有对称性、不受顺序影响；若其中一项不存在，则退化为另一项活跃的组成部分。这意味着未匹配的申报会降低得分，而匹配的同周申报会提高得分。收紧周

核函数会降低相似度。尽管该度量有界且对称，但它是相似度，而非度量。在补充材料中，我们提出了一种边检测的替代方法，该方法强制交易之间进行一对一匹配。这种更严格的表述避免了“突发膨胀”现象，即同一日志中的多笔交易与同一个对应项匹配；同时也提供了稳健性检验，表明我们的发现并非更宽松的最佳匹配方案所造成的假象。

算法1 构建内部人相似性网络

创建空图 $G = (V,\ E)$
对于每个公司 C 执行
    对于内部人配对（u、v）在 C 其中 $u \neq v$ 执行
        获取交易日期 $Buys_u$，$Sells_u$ 内部人为 u
        获取交易日期 $Buys_v$，$Sells_v$ 内部人为 v
        计算 Sim_{Buy} 使用 BestMatch($Buys_u$，$Buys_v$)
        计算 Sim_{Sell} 使用 BestMatch($Sells_u$，$Sells_v$)
        // 根据交易量合并得分
        将 $Score$ 设为 Sim_{Buy} 和 Sim_{Sell}的加权平均
        如果 (u 处于活动状态) $\wedge$ (v 处于活动状态) $\wedge$ (u 有 ≥ 8 笔交易) $\wedge$ (v 有 ≥ 8 笔交易) $\wedge$ (重叠 ≥ 4 周) $\wedge$ ($Score > 0.8$) 则
            添加边 (u, v) 到 G，权重为 $Score$
        结束条件判断
    结束循环
结束循环
返回 G

相似度指标用于比较同一公司中每对内部人士在给定时间点所报告的完整交易日期集合。我们认识到，由于公司内部人士获准交易的时间窗口每年约为 120 至 160 天，因此部分重叠可能是偶然发生的（Guay 等人，2021）。这意味着，同一组公司内部人士在某个时间点偶然于同一个每周时间窗口内进行交易，是可以预期的。因此，方程（4）仅在满足以下条件时才构建一条边：(i) 他们至少报告了 8 笔交易；(ii) 他们报告的交易在至少 4 个不同周存在重叠；(iii) 相似度得分 $S(X_C, Y_C)$ 大于 0.8。我们的补充材料中包含一项证明，说明这一阈值是合理的；该证明表明，达到这一阈值所需的重叠程度在统计上不太可能偶然出现。

3.3 零模型

稳健的零模型构建有助于衡量协同交易纯粹由偶然因素导致的可能性。为此，我们采用两种不同的方法生成底层数据的合成版本，然后将我们的算法应用于这些数据。首先，我们实施了校准生成模型，以控制监管伪影。

通过将合成交易均匀分布在可允许的开放窗口内（考虑禁售期和任期），该模型建立了一个基准，用于衡量仅由共享交易约束产生的边权重（Gauvin et al. 2022）。其次，为了将特定的社会协调与环境同质性（例如，对上市公司新闻的集体反应）区分开来，我们采用了受约束的时间置换。通过在固定的公司—时间区间内随机置换内部人士身份，同时保留精确的交易时间戳，该方法能够分离出超过同期公司活动背景水平的二元同步。

3.3.1 校准生成模型

第一个零模型构成了一个随机化参考模型，该模型受经验活动时段的约束，用以判断观测到的时间模式在“随机但可比”的世界中是否更可能出现（Gauvin et al. 2022）。在法证金融领域，这种方法是标准做法：通过生成预期“随机共现”的分布，并将其与经验观测值进行比较，从而区分技能（或信息）与监管导致的巧合（Cohen et al. 2012；Fagioloet al. 2013）。就本研究而言，校准的零模型模拟了一个合法合理但彼此不协调的内幕交易世界，其中（a）每位内部人的交易频率和交易跨度都与现实相同；（b）每家公司都保留经验数据中观测到的相同总活动量和存续期。

完整的校准参数列表见补充材料。从概念上说，公司存续期和内部人任期以其在2014—2024年时间线上的实际首次和末次交易日期为锚点。鉴于内部人可能在这些观测界限之前或之后也依法有资格交易，因此这应被视为一种保守的代理。在这些锚定时段内，每个内部人—公司配对都保持实际交易次数和任期时长不变，同时也保留买入与卖出的整体比例。交易会均匀分配到法律上合理的开放交易日，即每个91天季度的前30个工作日。我们认识到，由于合法交易窗口由公司选择，其长度约为25至40天不等；我们参考了Guayet al. (2021)的研究，因此选择30天；该研究报告称，美国公司的交易日平均为30至35天。接下来，假定所有公司的每个季度都在相同的日历日期开始。这不会影响单家公司内部的共同交易，但会略微增加在多个董事会任职的董事之间偶然出现重叠的可能性。因此，该模型描述了一个符合法律规定、以经验为依据但彼此不协调的公司交易世界，为提出以下问题提供了恰当的基准：从时间重叠的角度看，内部人实际的交易模式是否比随机独立活动下更具结构性?

3.3.2 受约束的时间重排 g

合成随机参考模型为评估网络中观测到的协同行为的统计显著性提供了有用的基准。尽管如此，该模型忽略了这样一个事实：金融活动往往受外生公共信号（例如时事、经济政策和社会趋势）驱动，而这些信号会同时影响同一家公司的所有内部人士（Cline等（2017））。因此，交易簇可能看起来

具有协同性，但这并非串通所致，而可能只是因为多名内部人士各自独立地以相似方式对同一公共事件作出反应。为考虑这种非平稳性，我们采用了受约束随机参考模型，也称为区块约束配置模型（Gauvin 等，2013；Casiraghi，2019）。在方法上，这种对底层数据进行的分层随机置换通过固定所有交易事件的精确时间线，同时在季度层面随机置换每家公司中与交易时间戳相关联的身份，从而生成一段反事实历史。

形式化地说，我们将所有交易的集合记为 $\mathcal{E} = \{(u_i, f_i, t_i, d_i)\}$，其中，$u$ 是内部人，f 是公司，t 是精确时间戳，d 是交易方向。我们根据公司 f 和粗粒度时间区间 τ（例如日历月），将 $\mathcal{E}$ 划分为互不相交的子集 $S_{f,\tau}$。在每个子集 $S_{f,\tau}$ 内，我们定义内部人士身份向量 $\mathbf{U}_{f,\tau} = [u_1, u_2, \ldots, u_k]$。我们通过对 $\mathbf{U}_{f,\tau}$ 应用随机置换 σ 来生成零模型数据集，使得零模型中的第 j 笔交易被分配给内部人 $u_{\sigma(j)}$，而时间戳 t_j 和方向 d_j 保持不变。该模型保留了公司的精确时间交易量分布（包括可能由外生因素导致的所有活动突发），但破坏了特定行动者之间的同步关系。这种方法借鉴了 Crane 等（2019）的研究；他们使用零模型，在控制投资者所属行业和规模偏好的同时，随机化投资者之间的联系，以更好地理解信息共享。

3.4 网络分析

在构建观测网络，并参照上述两个零模型评估其结构后，我们使用接近中心性、特征向量中心性和 OddBall 算法考察网络中的人群。OddBall 算法由 Akoglu 等人 (2010)开发，用于检测加权图中的异常自我网络（即由节点的直接邻居构成的子图）。我们首先根据所有节点的接近中心性进行排名，以从基本层面找出最“可疑”的内部人士。参照Goergen 等人 (2019)，我们还根据节点的特征向量中心性进行排名，以识别在更广泛的网络中嵌入程度较高的内部人士；这种方法不仅考察有多少节点与目标节点相连，也考察这些节点在更广泛网络中的重要性（即每个相连节点的度）。

最后，为了识别加权边图中的异常节点——这正是我们所处情境——Akoglu 等人(2010) 首先概述了四种经验模式或幂律关系，他们认为这些通常是现实世界加权图的固有特征。随后，他们提出了一种算法，用于识别其自我中心网络偏离这一结构的中心节点（节点）。该方法适用于我们的情境，因为我们希望在更广泛的结构中识别异常配对或内部人士群组。为确认这一方法的有效性，补充材料提供了统计检验和可视化结果，证明我们的数据满足全部四项条件。根据 Akoglu 等人(2010) 使用的符号约定，离群线分数定义如下：

$$\textit{out-line}(i) = \frac{\max(y_i, Cx_i^{\theta})}{\min(y_i, Cx_i^{\theta})} \cdot \log\left(|y_i - Cx_i^{\theta}| + 1\right), \tag{8}$$

其中， y_u 表示中心节点 u的自我中心网络中的边数，而 $f(x_u)$ 表示根据幂律拟合得出的预期边数，其中自我中心网络 u 包含 x_u 个节点。该公式实质上是根据所有节点与预期拟合值的距离进行排名。按照Akoglu 等人(2010) 的方法，随后通过结合 u 的离群线分数（如上所定义）与局部离群因子（LOF），得出其最终异常值分数。局部离群因子是一种基于密度的度量，可标记图中相对稀疏区域的异常值。这有助于避免将远离大多数点、但仍接近幂律趋势的中心节点误判为正常。随后对所有节点的异常值分数进行排名，再对排名靠前的节点（即异常自我中心网络）进行单独深入评估。

4 结果

4.1 网络参数

根据第 3.2 节所述的相似度函数构建的网络包含 4,650 个节点和 7,007 条边。我们只纳入至少有一条边的节点；由于大多数公司内部人士报告的交易与其他内部人士的交易重叠很少或没有重叠时段，因此不会与其他节点相连。结果是，我们观察到的节点数量远少于原始数据中的人数。

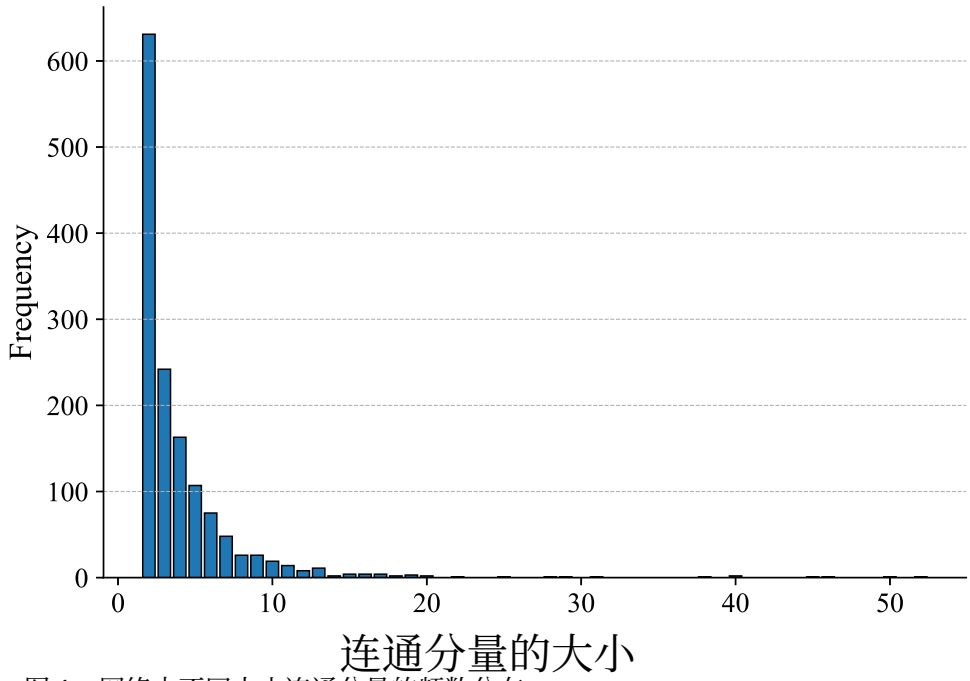


图 1：网络中不同大小连通分量的频数分布。

图 1 显示，该网络非常稀疏：1,313 个连通分量中，大多数仅包含两个节点，对应的密度为 0.000648。这表明，大多数公司内部人士的交易时间段并不相似，或者至少相似度不足以达到我们提出的相似度函数所规定的连边标准。尽管如此，图 1 还显示了一些较大的连通分量，其中包括 10 个由 30 多名公司内部人士组成的簇。补充材料中提供了一张表，按所含公司数量列出了连通分量的分布。值得注意的是，所有连通分量中有 93% 仅由与同一家公司的个人组成。其余 7% 的连通分量则包含来自最多九家不同公司的内部人士。

4.2 零模型

若不为随机或偶然活动建立基准，就无法判断上述模式究竟代表真正的串通，还是仅仅反映了监管约束和／或对外生事件的共同反应。为此，我们的零模型框架对独立执行假设下预期的网络拓扑进行基准评估。对于两种零模型，我们生成 10^3 合成数据集。随后，我们对每个合成数据集应用相似度函数，并将其结构属性与经验网络的结构属性进行比较。

网络指标	观测值 (X_{obs})	结构零模型 Z 分数（均值 [5%，95%]）	(Z_{struc})	置乱零模型（均值 [5%，95%]）	Z 分数 (Z_{shuff})
Global Connectivity					
节点	4,650	21.9 [15.0, 29.0]	1,075***	779.6 [744, 813]	178.6***
边	7,007	19.1 [14.0, 24.0]	2,271***	1,449.6 [1, 401, 1, 499]	186.1***
平均度	3.01	1.76 [1.48, 2.13]	6.1***	3.72 [3.57, 3.88]	-7.3***
Network Structure					
连通分量	1,313	8.99 [6.0, 12.0]	651.8***	258.9 [244, 274]	112.2***
巨型连通分量	40	4.42 [3.0, 6.0]	33.7***	41.0 [41.0, 41.0]	–†
多公司内部人士	23	0.59 [0.0, 2.0]	23.8***	3.95 [3.0, 7.0]	15.4***

表 2：网络拓扑与校准后的零模型及打乱空模型的比较分析。该表报告了经验网络指标（$N = 4$，650 个节点，$M = 7$，007 条边），以及根据 10^3次模拟得出的校准后的零假设和置乱零模型的均值与 5–95% 置信区间。经验值（X_{obs}）与结构基线（H_0^{struc}）进行比较，以判断拓扑是否由交易量约束所致；同时与打乱基线（H_0^{shuff}）进行比较，以判断其是否由同质性所致。 Z分数表示经验值与空分布之间的标准化距离。显著性以星号表示（*** $P < 0.001$）；对于 †，零假设方差为零，因此 Z 分数未定义。

表2 量化了观测到的内部人网络与理论基准之间的结构差异。我们采用 Z得分对这些偏差的幅度进行标准化，从而能够直接比较尺度迥异的各项指标。此外，我们报告的是 95% 置信区间，而非标准差，因为网络中的零分布通常呈偏态（Squartini 和 Garlaschelli （ 2011 ））。如表 2 所示，经验网络的连通性和凝聚性明显偏离反事实情景。观测到的活跃节点数量（4,650）和边数量（7,007）比结构基线预期值高出逾 1,000 个标准差（$Z_{struc} > 1,000$），证实观测到的聚类并非公司上市期间造成的机械性结果。关键的是，即使在更严格的“置乱零模型”下（$Z_{shuff} \approx 170$），这种显著性依然存在。这意味着，我们在第4.3 节讨论的异常网络结构中，至少有一部分是由特定的、有意识的协调行为所致，而非整体市场趋势。补充材料中提供了零分布的详细描述性统计数据。

尽管几乎所有指标都显著高于两种零模型基线，但置换模型的最大连通分量值在所有置乱中始终保持不变，为 40。这个数字凸显了在我们的研究情境中使用两种零模型的重要性。事实上，校准后的零假设显示，大小为 40 的连通分量几乎不可能出现，其 95% 置信区间上限为 6。相反，在置乱零模型中，该连通分量的大小在所有 10^3 模拟中始终保持不变，这意味着该公司内部的 40 名内部人士都在协同行为，因此对他们进行置乱不会影响我们的算法对其行为的解读。我们将在第 4.3 节中更深入地讨论这个簇。

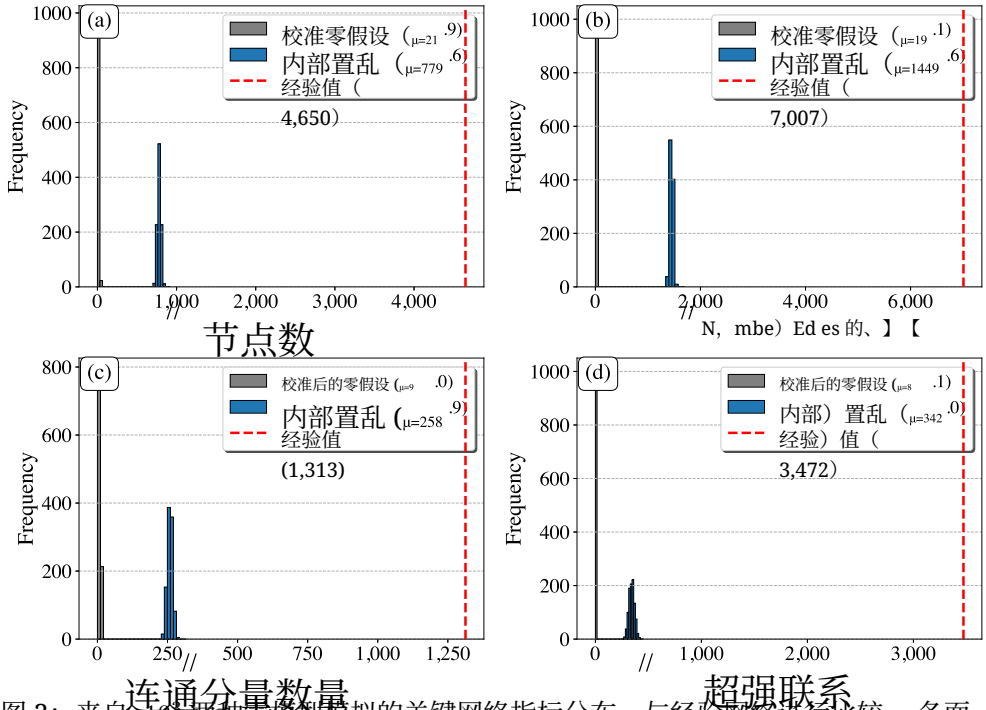


图 2：来自 10^3 两种零模型模拟的关键网络指标分布，与经验网络进行比较。各面板显示：(a) 节点数量，(b) 边数，(c) 连通分量数，以及 (d) 超强关系（相似度为 > 0.9 的边）。横轴上的 // 表示刻度中的断裂，用于将高度集中的零分布与相距较远的经验值同时可视化。

图 2 展示了校准后的零假设（结构基线，灰色）和内部洗牌10^3 模拟所生成的四项关键拓扑指标的分布，并与经验观测值（红色虚线）进行对比。结果表明，观测到的协同行为水平与反事实基线之间存在根本性的差异。事实上，我们所有指标的经验值——这些指标反映了活跃网络的规模（节点、边）及其凝聚性（连通分量、超强关系）——都远远超出零分布。例如，观测到的网络包含 3,472 条超强关系（$w_{ij} > 0.9$），比置乱预期（$\mu \approx 340$）高出一个数量级。这证实了协同交易的高密度并非月度季节性或行业整体动量造成的假象，因为置乱保留了这些因素。

4.3 实证考察

4.3.1 案例研究 1：接近中心性

本节对经验推导出的网络进行可解释且易于理解的考察。我们概述了分析网络结构的三种方法，并运用这些方法对六个最值得关注的自我中心网络进行定性讨论。我们的目标是说明金融监管机构如何利用我们的算法生成网络，从而识别值得进一步审查的公司内部人士群体。我们首先考察接近中心性，因为按定义，接近中心性得分最高的节点能够以最有效的方式到达网络中的其他节点。

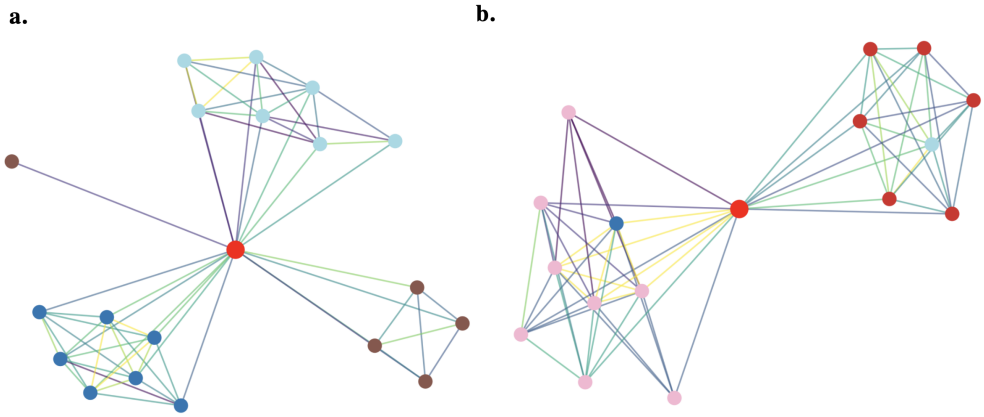


图 3：按接近中心性衡量的中心性最高的两名内部人士的自我中心网络。节点颜色表示公司隶属关系，边的颜色表示边权重：黄色代表强联系，绿色代表中等偏强联系，蓝色代表中等联系，紫色代表弱联系。

图3 展示了接近中心性得分最高的两名内部人士。具体而言，图 3a 展示了一名在四家中型市值公司（市值在40亿至120亿美元之间）担任董事会席位的个人；该个人与其隶属的四家公司中的其他高管存在大量重叠。接下来，图 3b展示了一名在所考察的10年期间，曾在五家不同公司担任多个高管职位的内部人士。例如，与其有强联系的那个大型粉色簇代表一家《财富》美国500强公司的董事会成员。还需注意的是，这两名个人都只与各自董事会中的部分董事保持联系，且边权重差异显著。这种差异表明，观测到的簇不能归因于内部人士在同一时间获得股权薪酬。

我们将排名前 100 位的中心节点与一份全面的联邦证券欺诈刑事起诉数据库进行了交叉比对。结果发现，三名中心性得分极高的内部人士曾是 AVEO Pharmaceuticals 的董事会成员。该公司因隐瞒……而卷入备受瞩目的欺诈诉讼。

美国食品药品监督管理局的反馈。至关重要的是，后续执法行动并未起诉纳入我们网络的具体董事会成员，而是针对首席执行官、首席财务官和首席营销官；他们最终就民事指控达成和解，或被裁定犯有证券欺诈（美国证券交易委员会 2018）。这意味着，尽管我们的算法未能准确锁定作案者，却成功识别出该企业网络是信息不对称风险极高的区域。

4.3.2 案例研究 2：特征向量中心性

接近中心性能够识别出可以有效触达网络中其他成员的个体，但无法反映相邻节点的重要性。特征向量中心性通过向与连接良好的同行建立联系的内部人士赋予更高得分，弥补了这一局限。因此，它特别适用于识别具有结构性影响力的群体，这些群体的行为可能会在网络中传播，或放大协同交易模式。在我们的案例中，特征向量中心性得分的分布高度偏斜（见补充材料）：大多数内部人士处于网络边缘位置，而少数成员则表现出显著更高的中心性。

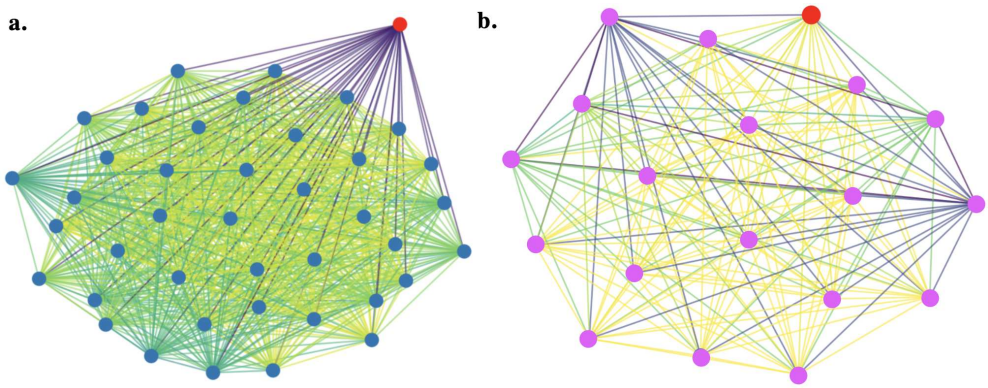


图 4：通过特征向量中心性测得的两个最具中心性的内部人士的自我中心网络。节点颜色表示公司隶属关系，边的颜色表示边权重，其中黄色表示强联系，绿色表示中强联系，蓝色表示中等联系，紫色表示弱联系。红色节点表示当前考察的个体。

图 4 展示了特征向量中心性得分最高的两个节点的自我中心网络。此处展示的第一个连通分量共包含41名公司内部人士。有趣的是，其中36人属于同一个家族，并持续在相近的时间段内交易大笔股权。该家族同步进行的交易引发了人们对信息优势、串通行为以及公司治理结构中监督不足的担忧。由于市值不到10亿美元，此类交易也可能对实际股价产生重大影响（Gomez，2024）。有趣的是，图4b 还展示了一簇属于某个大家族的内部人士，他们曾交易一家前赛车娱乐公司的股票

该公司于2019年10月被NASCAR收购（International SpeedwayCorporation 2019）。

4.3.3 案例研究 3：异常自我网络

接下来，我们使用前文讨论的 OddBall 算法来考察异常自我网络。OddBall 能识别局部连接模式偏离网络预期结构关系的节点，从而找出交易行为似乎异常的群体。

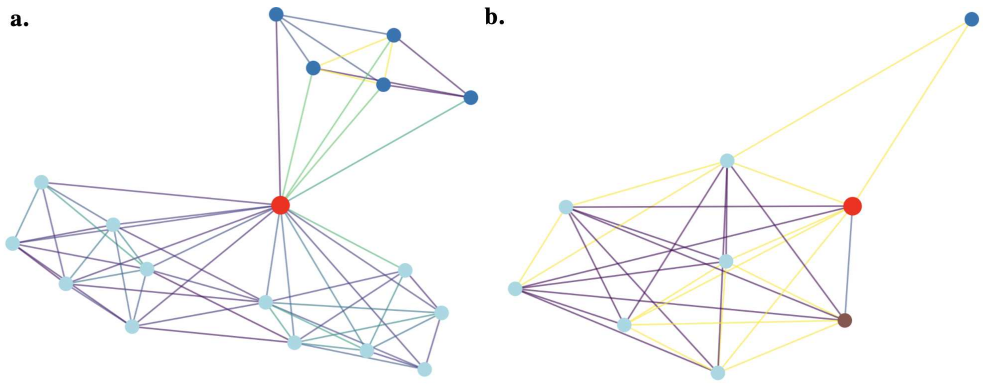


图 5：使用 *OddBall*算法识别出的异常程度最高的四名个人的自我网络。节点颜色表示公司隶属关系，边的颜色则代表边权重：黄色表示强联系，绿色表示中强联系，蓝色表示中等联系，紫色表示弱联系。红色节点表示正在考察的个人。

令人瞩目的是，将异常程度最高的 100 个自我网络对应的姓名与过往起诉案件数据进行交叉比对后，我们发现，拥有异常程度最高自我网络的公司内部人曾因一起重大会计丑闻相关的内幕交易而遭到起诉。具体而言，图 5a 中的自我网络展示了 Plug Power Inc. 的首席执行官；他是向美国纽约南区联邦地区法院提起的一起合并集体诉讼的被告。起诉指控描绘了这样一幅图景：先是进行典型的财务工程，随后趁机抛售内部人士持有的股票（Schubert Jonckheer& Kolbe LLP 2022）。值得注意的是，起诉状称，在这些财务报表重述公布前不久，首席执行官和首席财务官出售了约$4500 万美元的个人持股。我们的算法捕捉到了这起大规模清仓事件，在两人之间生成了一条强联系，因此在运行 OddBall 时触发了“异常”标记。

接下来，图 5b 描绘了一位声名显赫、来自加利福尼亚州的风险投资家。他担任多个董事会的成员，并报告称，其交易活动与另外三家公司的高管存在重叠时段。值得注意的是，这位内部人士所关联的六名董事会成员中，有四人与其存在显著的重叠时段。进一步观察他们的交易可以发现，这四名内部人士都曾多次在两天的时间窗口内卖出大宗股票（金额超过$50,000）。这种模式可能表明他们交易时使用了共享信息。另一种可能是，

鉴于该公司的市值不足 $500 million，六名内部人士可能会在这些时间窗口内故意协调卖出，然后再向美国证券交易委员会报告，从而避免单个内部人士的交易过早披露可能引发的价格影响。考虑到众所周知，金融市场会对公司高层内部人士披露的大额交易作出反应，这种协同行为并非不可能（Fidrmuc 等，2008；Gomez，2024）。

5 讨论

传统上，内幕交易检测侧重于识别个体异常，监管机构和合规官会筛查那些产生异常收益的独立账户（Goergen 等，2019；Priyadarshi 和 Kumar，2024）。我们的研究表明，这种原子化视角忽视了金融不当行为的一个基本维度，即信息流的社会性和协调性。我们提出了一种算法，通过周核函数和严格的重叠阈值构建稳健的相似性网络，将在相同短时间窗口内反复交易的内部人士联系起来。将该算法应用于2014年至2024年间报告的290万笔交易后，我们得到一张稀疏但结构高度复杂的图，其中包含众多异常集群，集群内的内部人士协调紧密。

将研究结果与零模型进行比较，对于结果的解读至关重要。在缺乏标注数据的情况下，这些模型提供了必要的基准，用以判断该网络捕捉到的究竟是真实的协调行为，还是对市场波动的共同反应。首先，校准生成模型保留了内部人士—公司任职期限、公司存续期和总体交易数量，同时将交易均匀重新分配到允许交易日上。这些合成数据构成了一个保守基准，代表“在法律上合理但未经协调”的内部人士活动。在该模型下，我们发现，观察到具有相近数量的活跃节点、边或超强关系的网络的概率极低（Guay 等人，2021）。其次，置乱零模型证实，这些连接并非由对外生因素的集体反应所驱动。当我们保留某家公司的确切交易量分布、但随机改变交易者身份时，网络结构便会瓦解。这意味着，至少有一些规模庞大的内部人士群组，尤其是包含来自多家公司的个人的群组，体现了交易行为中隐蔽的协同。

从更细致的层面来看，对中心性较高和异常自我网络的实证考察，展示了监管机构和合规团队能够通过所提出的算法识别出的模式类型。例如，特征向量中心性较高的内部人士通常身处密集的簇中，多个董事会成员在相近的短时间窗口内系统性地进行交易。我们发现的家族集群尤为重要，其中大批有关联的内部人士反复协同实施大额交易；这类交易往往发生在小市值公司中，因此很可能对价格产生显著影响。OddBall 分析进一步印证了这些定性观察，标记出其加权结构明显偏离更广泛网络中所观测到的幂律规律的自我中心网络。在实践中，这些中心节点及其直接邻域自然会成为深入开展法证调查跟进的对象。

至关重要的是，我们的方法识别出的最异常自我中心网络对应着一名曾因内幕交易而被起诉的内部人士，这在带标签的真实情况基本不可得的领域中，提供了一种事后验证。虽然我们的方法并非旨在预测个人是否应承担责任，但这一真阳性结果表明，基于网络的信号所捕捉的行为，与监管概念中的非法协同存在实质性的一致。

在这些发现的背景下，需要强调的是，我们的框架旨在作为一种筛查和优先排序工具，而不是作为独立认定责任的机制。我们的零模型有助于在合理的交易约束下排除纯粹的统计巧合，但无法确切区分对重大非公开信息的非法利用，与对公开信息或内部治理决策作出的合法共同反应。因此，对我们方法的全面实施还应纳入股票价格数据，以评估网络“标记”的个人是否报告了异常收益。鉴于Cziraki和Gider（2021）观察到，个人通常不会从事内幕交易，除非潜在利润足以证明相关风险是合理的，这一点无疑具有重要意义。因此，未来的研究应对网络中处于核心位置或异常的内部人士所报告交易的盈利能力开展补充分析。这将为内幕交易的起诉提供更有力的证据。

我们的工作存在三个值得注意的局限性。首先，我们的分析仅限于Form 4申报文件，这意味着我们只能观测到内幕交易的“可见”层面。那些通过离岸账户或空壳公司进行交易以规避申报要求的最恶劣违规者，自然不会出现在这一数据集中。其次，我们没有考虑金融衍生品，包括股票期权和认股权证。这一点很重要，因为高层高管通常会以股票期权的形式获得大额激励性薪酬方案（Augustin等，2019年）。值得注意的是，由于这些工具也必须由公司内部人士向美国证券交易委员会申报，因此可以将其纳入我们的方程式，或作为独立网络进行建模。

第三，我们基于匹配的相似度会按比例处理重叠时段，因此 9/10 与 18/20 的比例等同。然而在实践中，重叠时段的绝对规模可能包含有关协同行为严重程度的信息。我们设定的三项边创建条件通过要求达到最低活动水平并存在多个重叠周，在一定程度上缓解了这一问题；但未来的研究可以纳入严重程度评分，同时对相对重叠和绝对重叠赋予权重，从而更好地捕捉交易同步性的强弱。最后，从方法论角度看，未来研究还应更加重视规模虽小但联系紧密的连通分量。事实上，我们的分析主要着眼于识别协同交易的内部人士群组。然而，考察仅包含两个或三个紧密相连节点的连通分量，也与研究目标相关，因为这些内部人士报告的交易高度协同，且与公司内其他未被认为具有足够相似行为、因而未建立边的人员有所不同。Sun 和 Gorduza (2024) 在这方面开展了一些颇具前景的研究，利用基于图的机器学习，在标普500指数公司的网络中识别内幕交易模式。

6 结论

我们的目标是通过检测交易在时间上的重叠异常，识别那些依据重大非公开信息进行交易、且有两方或多方能够获取该信息的公司内部人士。从检察官的角度看，这种方法为传统的异常盈利能力分析提供了补充证据。从概念上说，我们的研究对三个研究领域有所贡献。第一，在内幕交易检测文献中，我们的方法补充了传统的公告前筛查和基于利润的异常检测，将关注点从个别异常值转向行为的关系模式。第二，从更广义的取证网络科学角度看，我们的分析表明，精心构建的零模型能够区分由制度性约束（例如交易窗口、任期、公司存续期）造成的结构特征，与可能由协同决策驱动的特征。第三，在方法论方面，我们证明，在真实标签稀缺或缺失、噪声极大的金融环境中，无监督、无标签的异常检测也能够付诸实践。

声明

作者贡献

Gian Jaeger (G.J.)、Wang Ngai Yeung (W.N.Y.) 和 Renaud Lambiotte (R.L.) 是本文作者。G.J.构思了这项研究，开展了实验/模拟，并撰写了论文初稿。R.L. 和 W.N.Y. 负责监督该项目。所有作者均参与了论文的修订，并共同完成了最终版本。

资金支持

R.L. 感谢英国工程与自然科学研究委员会通过资助项目 EP/V013068/1、EP/V03474X/1 和 EP/Y028872/1 提供支持。

利益冲突

作者声明不存在任何利益冲突。

伦理审批和参与同意

不适用。

数据可用性

本研究使用的所有数据均可从 https://www.sec.gov/search- filings 公开获取。

材料可用性

本文附有补充材料。

代码可用性

所有代码将在文章正式发表后提供。

参考文献

Afzali M、Martikainen M（2021）内幕交易的网络中心性与价值相关性：来自欧洲的证据。《金融评论》56(4)：793–819；Ahern KR（2017）信息网络：来自非法内幕交易线索的证据。《金融经济学杂志》125(1)：26–47；Akoglu L、McGlohon M、Faloutsos C（2010）OddBall：在加权图中发现异常。载于：《知识发现与数据挖掘进展：第14届亚太会议，PAKDD 2010，印度海得拉巴，2010年6月21日至24日。会议论文集。第二部分14》，施普林格，第410–421页；Arif S、Kepler JD、Schroeder J等（2022）审计过程、私人信息与内幕交易。《会计研究评论》27(3)：1125–1156；Augustin P、Brenner M、SubrahmanyamMG（2019）收购公告发布前的知情期权交易：内幕交易？《管理科学》65(12)：5697–5720；Basu K、Zhou C、Sen A等（2018）监测恐怖主义网络的新型图分析方法。载于：2018 IEEE国际并行与分布式处理及其应用、普适计算与通信、大数据与云计算、社会计算与网络、可持续计算与通信会议（ISPA/IUCC/BDCloud/SocialCom/SustainCom），IEEE，第1159–1166页；Berkman H、Koch P、Westerholm PJ（2020）董事网络内部：董事何时交易或持有内幕股票、交叉持股股票和非关联股票。《银行与金融杂志》118：105892；Bhattacharya U（2014）内幕交易争议：文献综述。《金融经济学年度评论》6(1)：385–403；Casiraghi G（2019）区块约束配置模型。《应用网络科学》4(1)：123；Cline BN、Gokkaya S、Liu X（2017）机会主义内幕交易的持续性。《金融管理》46(4)：919–964；Cohen L、Malloy C、Pomorski L（2012）解码内幕信息。《金融学杂志》67(3)：1009–1043；Crane AD、Koch A、Michenaud S（2019）机构投资者派系与治理。《金融经济学杂志》133(1)：175–197；Cziraki P、Gider J（2021）内幕交易带来的美元利润。《金融评论》25(5)：1547–1580

El-Khatib R、Fogel K、Jandik T（2015）首席执行官网络中心性与并购绩效。《金融经济学杂志》116(2)：349–382；El-Khatib R、Jandik D、Jandik T（2021）网络中心性、连接与社会资本：来自首席执行官内幕交易收益的证据。《金融评论》56(3)：433–457；Fagiolo G、Squartini T、Garlaschelli D（2013）经济网络的零模型：世界贸易网络案例。《经济互动与协同杂志》8(1)：75–107；Fidrmuc J、Goergen M、Renneboog L（2008）内幕交易、新闻发布与所有权集中度。载于：《内幕交易》。CRC出版社，第309–370页；Fishman MJ、Hagerty KM（1992）内幕交易与股票价格的有效性。《兰德经济学杂志》，第106–122页；Fried JM（2013）通过公司进行内幕交易。《宾夕法尼亚大学法律评论》162：801；Gauvin L、Panisson A、Cattuto C 等（2013）活动时钟：人际接触时序网络上的传播动力学。《科学报告》3(1)：3099

Gauvin L, G´enois M, Karsai M 等（2022）时序网络的随机参考模型。《美国工业与应用数学学会评论》64(4):763–830

Goergen M, Renneboog L, Zhao Y（2019）内幕交易与网络关联董事。《公司金融杂志》56:152–175

Gomez EA（2024）强制性披露传播对投资者间信息不对称的影响：来自EDGAR系统实施的证据。《会计评论》99(1):235–257

Guay WR、Kim S、Tsui D（2021）内幕交易窗口期的决定因素。International Speedway Corporation（2019）International Speedway Corporation 宣布与NASCAR Holdings, Inc. 达成合并协议。sec.gov/edgarJones MT、Magee JB、Kernisky A（2024）三只法国母鸡？并非如此。美国证券交易委员会加大对内幕交易、影子交易及公平披露条例的执法力度。Holland & Knight《第二意见》博客，Season’s Readings Series。Kulkarni A、Mani P、Domeniconi C（2017）用于内幕交易检测的基于网络的异常检测。Li R、Liu Z、Ma Y 等（2022）基于图学习的互联网金融欺诈检测。《IEEE计算社会系统汇刊》10(3):1394–1401。Oliveira R MA、Sant’Anna AMO、Ferreira PH（2025）用于反洗钱金融交易的复杂网络异常检测。《法证》

Science International：《数字调查》55:302005

Priyadarshi P, Kumar P（2024）利用人工智能进行内幕交易检测的综合综述。《计算社会科学杂志》7(2):1645–1664

Schubert Jonckheer & Kolbe LLP（2022）股东警示：Plug Power Inc.（PLUG）的高管和董事因涉嫌虚假陈述及内幕交易而接受调查。
https://www.prnewswire.com/news-releases/shareholder-alert-plug-power-inc，新闻稿，美通社

Securities U, Commission E（无日期）内幕交易及 3、4 和 5 号表格

Squartini T, Garlaschelli D（2011）用于检测真实网络中模式的解析最大似然法。《新物理学杂志》13(8):083001

Sun Y, Gorduza D（2024）通过高度互联的核心人物网络与图机器学习预测内幕交易行为。载于：2024年第六届机器人、智能控制与人工智能国际会议（RICAI），IEEE，第775–779页

Tamersoy A、Khalil E、Xie B 等（2014年）大规模内幕交易分析：模式与发现。《社会网络分析与挖掘》4：1–17

美国证券交易委员会（2018）美国证券交易委员会针对 AVEO Pharmaceuticals 前高管获得最终判决。网址https://www.sec.gov/enforcement-litigation/litigation-releases/lr-24062

Yeung WN, Di Clemente R, Lambiotte R（2025）垃圾进，垃圾出？数据质量对犯罪网络干预的影响。《EPJ数据科学》14(1):37

补充材料

采用 0.8 边阈值的依据

设定与符号约定

我们考虑同一家公司中的两名内部人士 C，其交易日期多重集为 $X_C = \{x_1, \ldots, x_m\}$ 和$Y_C = \{y_1, \ldots, y_n\}$，其中 $X_C \neq Y_C$时间邻近性通过以下方式衡量：周核函数

$$w(d) = \begin{cases} 1 - \dfrac{d}{7}, & d \le 7, \\ 0, & d > 7, \end{cases} \qquad d = |x - y|. \tag{9}$$

有向最佳匹配相似度为

$$s_{X|Y} = \frac{1}{m} \sum_{x \in X_C} \max_{y \in Y_C} w(|x - y|), \qquad s_{Y|X} = \frac{1}{n} \sum_{y \in Y_C} \max_{x \in X_C} w(|x - y|),$$

对称最佳匹配相似度为

$$S(X_C, Y_C) = \frac{1}{2}\Big(s_{X|Y} + s_{Y|X}\Big). \tag{10}$$

边预筛选条件。
在全文中，我们采用与正文相同的边筛选条件：每位内部人士在研究期间至少有 8 笔交易，且这对内部人士在至少 4 个不同周内存在重叠时段。这些筛选条件强化了下文的结论，但并非我们推导概率界所必需。

偶然重叠零模型

令 A 表示十年期限内允许的交易日数量。我们固定有 $A = 120 \times 10 = 1200$ 个允许的交易日（关于为何这是保守做法，请参阅备注）。在独立随机时序零假设下，给定交易笔数（m， n），每位内部人士都从 A 个允许的交易日中独立、均匀地选择交易日（有放回抽样），且不同内部人士之间相互独立。

保守化简。
对于任意固定的 $x \in X_C$,

$$\max_{y \in Y_C} w(|x - y|) \;\le\; \mathbf{1}\{\exists\, y \in Y_C :\ |x - y| \le 7\}.$$

因此，$m\, s_{X|Y}$ 随机占优于一个二项分布随机变量，该变量统计了有某个 周邻近交易落在 Y_C 中的 x 笔交易。记

$$p_n \;=\; \mathbb{P}\big(\exists\, y \in Y_C :\ |x - y| \le 7\big) = 1 - \Big(1 - \frac{14}{A}\Big)^n, \tag{11}$$

我们有

$$m\, s_{X|Y} \;\preceq\; \mathrm{Binomial}\big(m, p_n\big), \qquad n\, s_{Y|X} \;\preceq\; \mathrm{Binomial}\big(n, p_m\big). \tag{12}$$

这一替代方案有意采取保守做法：指示函数将任何周内邻居都计为完整权重 1，而式（9）中的线性核会降低相隔 1–7 天的配对权重，因此在随机情况下，出现高相似度的可能性严格更低。

尾部界与族错误率校准

对于 $\tau \in (0,\ 1)$ 和 $K \sim$二项分布$(k,\ p)$,

$$\mathbb{P}\left(\frac{K}{k} \ge \tau\right) \;\le\; \exp\big\{ - k\, D(\tau \| p)\big\}, \qquad D(a\|b) = a \log \frac{a}{b} + (1 - a) \log \frac{1 - a}{1 - b}. \tag{13}$$

由于 $S = \frac{1}{2}(s_{X|Y} + s_{Y|X}) \geq \tau$强制要求 两个 方向均至少为 τ，(12)–(13) 得出

$$\mathbb{P}\big(S \geq \tau\big) \;\leq\; \exp\big\{ - m\, D(\tau \| p_n) - n\, D(\tau \| p_m) \big\}. \tag{14}$$

令 $N_{\text{pairs}} = \sum_C \binom{n_C}{2}$ 为公司内部人配对的数量。在 9,426 家公司中共有 70,941 名内部人士，经验平均值为每家公司 $\bar{n} \approx 7.53$ 名内部人士，因此

$$N_{\text{pairs}} \;\approx\; 9{,}426 \times \binom{\bar{n}}{2} \;\approx\; 2.31 \times 10^5.$$

我们通过要求 $\alpha = 0.05$，将数据集层面的族错误率（FWER）控制在

$$\mathbb{P}(S \geq \tau_\star) \;\leq\; \frac{\alpha}{N_{\text{pairs}}} \quad \text{for every eligible pair.}$$

在筛选条件最宽松的情况下（$m = n = 8$），(11) 给出

$$p_8 \;=\; 1 - \Big(1 - \frac{14}{1200}\Big)^8 \;=\; 0.08961.$$

根据 (14)，只需选择 $\tau_\star$，使得

$$D(\tau_\star \| p_8) \;\geq\; \frac{\log(N_{\text{pairs}}/\alpha)}{m+n} \;=\; \frac{\log(2.31 \times 10^5 / 0.05)}{16} \;=\; 0.959 \;\; \text{(to 3 s.f.)}.$$

求解二元 KL 不等式可得族错误率校准的临界相似度

$$\boxed{\tau_\star \approx 0.652 \;\text{（独立随机时序下的 FWER 为 } \leq 5\%\text{）。}} \tag{15}$$

关于保守性的说明

1. 核函数的保守性。 用 0–1 窗口替代线性周核函数会严格提高零假设下的相似度，因此上述所有尾概率都是 上界。2. 活跃度筛选条件。 要求 ≥ 4 个不同的重叠周并未用于界限推导，且能进一步抑制孤立突发造成的虚假对齐。3.较高的活跃次数。 对于每人交易次数超过 8 次的内部人士，式（14）中的界限会通过 KL 项 $m + n$呈指数级改善， $D(\tau \| p)$，从而使假阳性更不可能出现。4. 允许交易日时间范围。 我们固定 $A = 1200$ （10 年内每年 120 个可交易日历日），这是一个保守的设定：有些公司每年可提供 120–180 天，即 A 最高可达 1800 天。较大的 A 会降低 p_n 式（11）中的值，使偶然重叠更少，并降低校准后的 $\tau_\star$，从而强化式（15）。

解读

在一个透明且保守的零假设下，该假设根据研究时间范围（10 年， $A = 1200$ 允许天数）和我们的边筛选条件（每位内部人士有 ≥ 8 笔交易）进行校准，族错误率校准的临界相似度为 $\tau_\star \approx 0.652$，如式（15）所示。因此，将 $S \geq 0.8$ 用作边阈值，是一种非任意且高度保守的选择，可在数据集规模上有效排除偶然边。

B A用于边检测的替代性指派方程

令 $X_C = \{x_1, \ldots, x_m\}$ 和 $Y_C = \{y_1, \ldots, y_n\}$ 表示公司 C 中两位内部人士的事件时间（即 Form 4 交易日期）（买入或卖出）。时间邻近性通过周核函数进行评分

$$w(d) = \begin{cases} 1 - \dfrac{d}{7}, & d \le 7, \\ 0, & d > 7, \end{cases} \qquad d = |x - y|. \tag{16}$$

因此，同日事件的权重为 1，间隔一天的权重为 $6/7$，间隔超过一周则权重为零。

一对一匹配。
我们构建一个关于边权重$w(|x-y|)$的最大权重二分匹配 $M^\star \subseteq X_C \times Y_C$，并满足以下条件：(i) 每个事件最多匹配一个对应事件（单射约束）；(ii) 只有符合 $d \le 7$ 的事件对才具备资格；符合 $d > 7$ 的事件对不纳入考虑。这种一对一约束可防止突发膨胀，确保一笔交易不会同时被用来解释另一名内部人士的多笔交易。

窗口级资格。
我们不按事件总数进行归一化，而是按可能发生协同行为的时间窗口数量进行归一化。设 $\mathcal{W}_X$ 和 $\mathcal{W}_Y$ 分别表示内部人士X 和 Y 进行交易的日历周集合，并定义可能存在重叠的周集合

$$\mathcal{W}_{XY} = \mathcal{W}_X \cap \mathcal{W}_Y. \tag{17}$$

在.....周内 仅有一名内部人士进行交易的周被视为不符合条件，且不会计入分母 r.

方向性得分。
我们将 X–给定–Y 方向性得分定义为

$$s_{X|Y} = \frac{1}{|\mathcal{W}_{XY}|} \sum_{(x,y)\in M^\star} w(|x-y|), \tag{18}$$

w按照约定， $s_{X|Y} = 0$ 当且仅当 $\mathcal{W}_{XY} = \varnothing$。 Y–给定–X 得分的定义类似 ,

$$s_{Y|X} = \frac{1}{|\mathcal{W}_{XY}|} \sum_{(x,y)\in M^\star} w(|x-y|). \tag{19}$$

这种归一化意味着，同一重叠周内的多笔交易既不会抬高也不会拉低得分：一旦两名内部人士在某个给定周内都活跃，该周内的额外活动不会产生影响。

按类别划分的对称相似度。
我们将对称相似度度量定义为

$$S_{\text{assign}}(X_C, Y_C) = \tfrac{1}{2}\big(s_{X|Y} + s_{Y|X}\big), \tag{20}$$

其满足 $S_{\text{assign}} \in [0,1]$ 在（16）条件下。

结合买入和卖出（按活动加权）。
我们分别计算卖出（20）和买入（A）的D:

$$S^{(A)}_{\text{assign}} = S_{\text{assign}}(X_A, Y_A), \qquad S^{(D)}_{\text{assign}} = S_{\text{assign}}(X_D, Y_D). \tag{21}$$

Let

$$T_A = |X_A| + |Y_A|, \qquad T_D = |X_D| + |Y_D|, \qquad T = T_A + T_D.$$

我们定义平均值 总体相似度为按活动加权的平均值

值

$$S_{\text{combined}}(X,Y) = \frac{T_A}{T}\, S^{(A)}_{\text{assign}}(X,Y) + \frac{T_D}{T}\, S^{(D)}_{\text{assign}}(X,Y). \tag{22}$$

解释与性质。

• $S^{(\cdot)}$ 分配 $\in [0,1]$ 和 $S_{\text{combined}} \in [0,1]$。• $S^{(\cdot)}$ 分配$= 1$ 如果在每个重叠周内，内部人士的交易在时间上完全一致。

• 一对一匹配可消除突发膨胀：单笔交易不会产生多个匹配项。
• 按重叠周进行归一化，可确保周内密集交易不会影响得分，而非重叠周内的活动则会降低得分。

• （22）中的活动加权组合反映了协同行为主要发生在买入还是卖出中。

C 校准的零模型参数

表 3：校准的零模型参数

参数	描述	数值 / 分布
N_f	公司数量	9,381（经验）
N_i	内部人士数量	65,001（经验）
C_{if}	每个内部人—公司配对的交易次数	经验整数计数
T_f	公司活动跨度	经验首末交易日期（已锚定）
T_{if}	内部人士—公司任职时长	以天计的经验跨度
p_{buy}	购买概率（A）	0.528（经验值）
时间范围	模拟窗口	2014–01–01 至 2024–12–31
Q	季度长度	91个日历日
D_{open}	每季度营业日	30
W_{open}	全球开放日历	每个91天季度的前30个工作日

D 连通分量按所含公司数量分布

表4：连通分量按所含公司数量分布。

公司数量	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
分量（%）	93.3%	4.6%	0.9%	0.5%	0.3%	0.3%	0.1%	0.0%	0.1%

E 特征向量分布

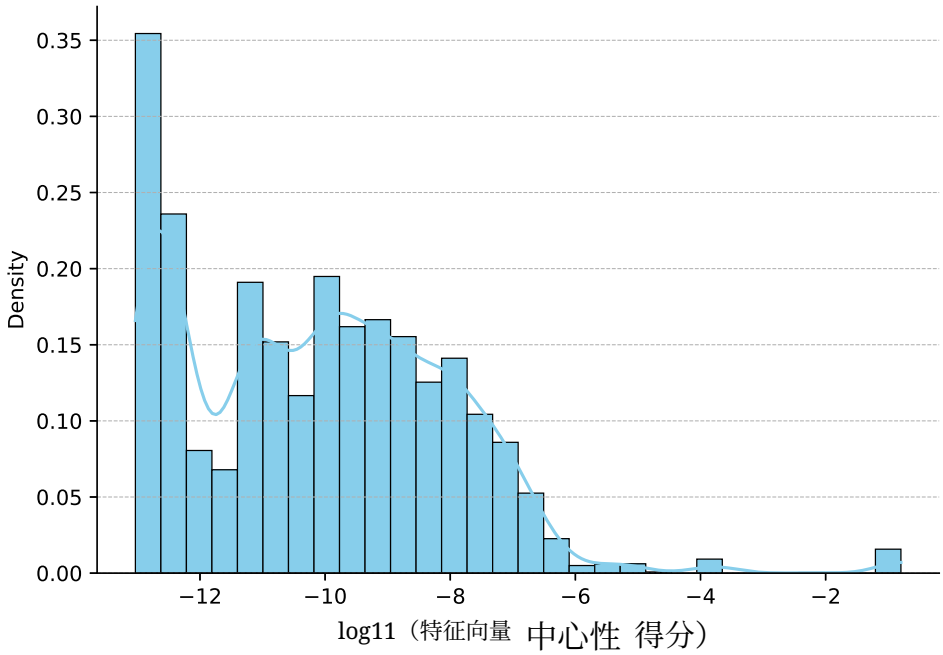


图 6：特征向量分布（对数刻度 ale)

F 零模型的详细拓扑统计数据

指标	观测值	零假设均值	零假设标准差	Z 分数	P 值
节点（N）	4,650	21.6	4.6	1,012.6	< 0.001
边（E）	7,007	18.9	3.4	2,042.6	< 0.001
连通分量	1,313	8.9	2.1	630.9	< 0.001
最大连通分量的大小	40	4.3	1.0	36.1	< 0.001
平均边权重	0.904	0.892	0.012	1.0	0.30
平均度（$\langle k \rangle$）	3.01	1.77	0.17	7.3	< 0.001
多公司内部人士	23	0.6	0.9	24.6	< 0.001
多公司内部人士比例	3.2×10^{-4}	0.1×10^{-4}	0.1×10^{-4}	22.4	< 0.001
超强关系	3,472	7.9	1.6	2,199.1	< 0.001

表 5：校准生成式零模型的详细拓扑统计数据。该模型将公司和内部人士活动固定在其经验生命周期内，但会随机化具体的交易日期。结构指标（节点、边、连通分量）中极大的 Z 分数证实，观测网络结构极不可能由仅受任期约束的非协同交易活动产生。P 值为 0.0 表示 $p < 10^{-100}$。

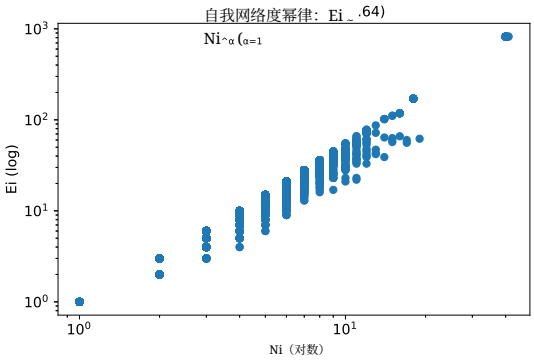
指标	观测值	零假设均值	零假设标准差	Z 分数	P 值
节点（N）	4,650	776.2	22.8	169.7	< 0.001
边（E）	7,007	1,446.1	31.0	179.3	< 0.001
连通分量	1,313	257.9	9.7	109.0	< 0.001
最大连通分量的大小	40	41.0	0.0	$-\infty$	< 0.001
平均边权重	0.904	0.869	0.001	25.5	< 0.001
平均度（$\langle k \rangle$）	3.01	3.73	0.09	-7.6	< 0.001
多公司内部人士	23	3.9	1.3	14.2	< 0.001
多公司内部人士比例	3.2×10^{-4}	0.6×10^{-4}	0.2×10^{-4}	14.2	< 0.001
超强关系	3,472	340.7	23.7	132.0	< 0.001

表6:约束置乱零模型（季度分层）的详细拓扑统计数据。该模型保留了每个公司—季度分层内交易的确切交易量和时间安排，但对内部人士身份进行了置乱。与这种“环境混杂”基准相比，观测网络仍表现出显著更强的结构性（更高的聚类程度和更强的联系）。请注意，零模型中的最大连通分量规模具有很强的确定性（标准差 ≈ 0）。

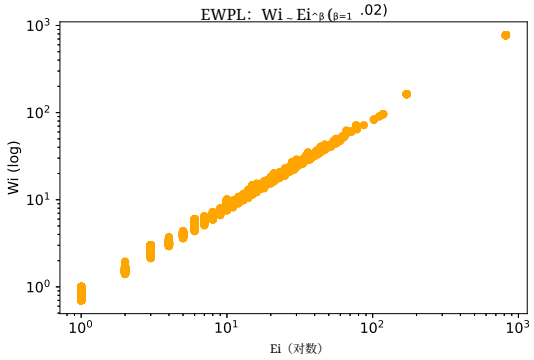
G 加权图中通过 OddBall 进行异常检测的基础条件

表7：自我中心网络幂律条件的验证。每个条件都表示自我中心网络结构特征之间的幂律关系。该数据集满足全部四个条件。

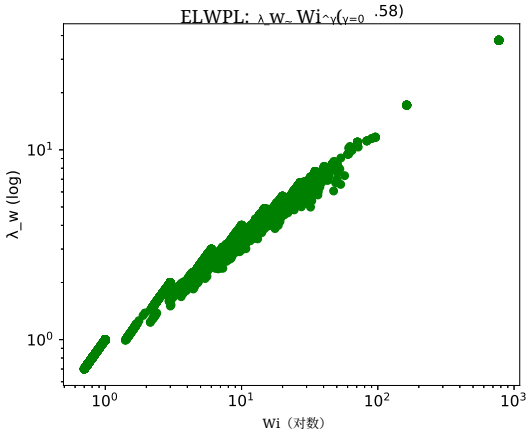
条件	变量关系	拟合值指数	理论值范围	条件 Met?
自我网络密度幂律	$E_i \propto N_i^{\alpha}$	$\alpha = 1.52$	$1 \leq \alpha \leq 2$	✓
自我网络权重幂律	$W_i \propto E_i^{\beta}$	$\beta = 1.21$	$\beta \geq 1$	✓
自我网络加权特征值幂律	$\lambda_{w,i} \propto W_i^{\gamma}$	$\gamma = 0.75$	$0.5 \leq \gamma \leq 1$	✓
自我网络秩幂律	$W_{i,j} \propto R_{i,j}^{\theta}$	$\theta = -1.15$	$\theta \leq 0$	✓



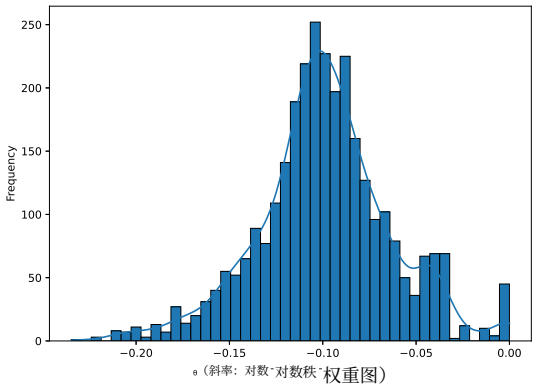
(a) 自我网络度幂律： $E_i \propto N_i^{\alpha}$



(b) EWPL： $W_i \propto E_i^{\beta}$



(c) ELWPL： $\lambda_{w,\ i} \propto W_i^{\gamma}$



(d) ERPL： $W_{i,j} \propto R_{i,j}^{\theta}$

图7：在双对数图上验证自我网络幂律条件。每个子图展示了针对不同结构性自我网络属性的幂律拟合。各缩写对应表4中列出的条件。