中国科学院大学
University of Chinese Academy of Sciences

博士学位论文

大规模图数据中的聚集性群体异常模式挖掘

作者姓名：冯文杰

指导教师：程学旗 研究员　中国科学院计算技术研究所

学位类别：工学博士

学科专业：计算机系统结构

培养单位：中国科学院计算技术研究所

2020 年 8 月

Collective Anomaly Pattern Mining in Large-Scale Graph Data

A dissertation submitted to the

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in Computer Architecture

By

Feng Wenjie

Supervisor: Professor Cheng Xueqi

Institute of Computing Technology, Chinese Academy of Sciences

August, 2020

中国科学院大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。本人完全意识到本声明的法律结果由本人承担。

作者签名：

日　　期：

中国科学院大学
学位论文授权使用声明

本人完全了解并同意遵守中国科学院大学有关保存和使用学位论文的规定，即中国科学院大学有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：　　导师签名：

日　　期：　　日　　期：

摘 要

物联网、社交媒体、电子商务等技术和应用的发展，产生和积累了海量的图式关联型数据，这些数据对象之间天然地具有相互依赖、长距离相关的特征且包含各种其他信息，激发了对此类数据的高效处理和知识挖掘的巨大需求；其中，以此为背景的群体性异常模式挖掘和行为规律分析业已成为金融风控、网络安全、社交网络等众多领域的核心，蕴含着巨大的科学和市场价值。

大规模图数据的群体异常模式挖掘仍需要解决离群点及群体异常的分布多样性、图拓扑结构中聚集异常的度量统一性、多属性图中群体异常的结构复杂性以及算法的可扩展性等方面带来挑战。本文重点研究在大规模图数据中的聚集性群体异常模式，分析不同数据形态表示下群体异常的复杂特性和行为表现，提出快速有效、可扩展的检测算法，并应用于大规模真实场景中不同群体异常行为模式的发现，这些模式包括具有同步性和密集性特征的共谋欺诈、恶意群体攻击、有趣学术合作团体等。本文的主要贡献如下：

首先，本文提出了图特征空间中基于视觉引导的聚集性群体异常检测方法。基于大规模图数据的节点特征表示，构建特征相关直方图映射来反映非欧空间中节点的分布表现和聚集性质，并分析微簇结构所对应的群体异常模式。文中提出了基于视觉引导的 EagleMine 算法来识别直方图中的类簇结构，结合词典模型和统计检验技术对直方图进行描述和总结，并由可疑性度量识别异常微簇。在大规模数据上的实验表明，EagleMine 识别到与人眼视觉感知结果相一致的类簇，以更加简洁的方式描述直方图数据；EagleMine 在群体异常检测中具有比基于图的检测方法更优的性能，也能用于时序事件中同步转发行为模式的检测等。

其次，本文提出了针对拓扑关联密集异常的统一图谱检测方法，研究了在不同目标设定中以稠密子图为对象的拓扑密集性关联及相关的群体异常行为模式。基于对多种实际场景中与稠密子图检测密切相关问题的理论分析和对比，包括图的最小割集、稀疏割社区、对比密度子图、最密可疑子图等模式的检测问题，文中提出了一种用于最密子图检测的统一形式化框架 (GenDS) 和图谱分析优化理论，结合大规模图的谱性质设计了快速、可扩展的检测算法 SpecGreedy。在 40 个真实网络上实验结果表明，SpecGreedy 在最密子图检测任务中比基线方法

的速度提高 58.6× 倍且子图密度更优；SpecGreedy 算法随图的大小线性可扩展，并在大规模学术共同作者网络中检测到具有突现合作行为的群体异常模式。

最后，本文提出了多属性图中层次结构感知的群体异常模式发现算法。本文以稀疏张量建模多维关联数据，基于统一的子张量密度度量和稠密子张量检测问题的形式化，提出一种新的框架 CatchCore，能够快速有效地检测在连接结构和密度分布上具有层次性特征的稠密子张量模式，并给出了基于最小描述长度准则的结果评价指标。在大规模真实数据上的实验结果表明，CatchCore 在稠密子张量和异常模式的检测中性能最优；在社交网络、网络攻击和科研合作的真实场景中，CatchCore 能够更加准确地识别出具有可解释性的、层次性稠密的群体异常行为模式；此外，CatchCore 的复杂度与张量的各方面指标呈线性相关。

综上所述，本文从相关性特征、拓扑连接结构、多元层次结构三方面研究了大规模图数据中的聚集性群体异常模式，并有针对性地设计了高效的检测算法，在真实数据上验证了算法的性能，并根据不同应用场景对识别到的异常模式进行了分析和解释。

关键词：图挖掘；群体异常检测；模式识别；算法设计；复杂行为

Abstract

The development of technologies and applications related to the Internet of Things, social media, and e-commerce, etc. have generated and collected massive amounts of the relational data, which contain objects that are naturally inter-dependent and have long-range correlations and rich attributes. Thus, there is a great demand for efficient-processing and knowledge-discovery for such complex relational data. Moreover, the collective anomaly pattern mining and behavioral analysis in these situations have become the core in many fields and applications, such as financial-risk control, network security, and social networks; it also contains great scientific and marketing values.

However, researchers are facing a number of challenges for the massive relational data, including the diversity of collective anomalies beyond outliers, the unify measurement design for anomalies on the topology, the complex structure of anomalies in the multi-attribute graph, and the scalability requirements for graph-based algorithms. This thesis focuses on the collective anomaly pattern mining in the large-scale graph data, it analyzes the complex characteristics and behavior of collective anomalies in different data representation, and proposes fast, effective, and scalable detection methods, which are applied to large-scale real scenes and used to spot some interesting and explainable anomaly patterns with synchronized and lockstep characteristics, including collusion-type of fraud, malicious attacks, academic collaborations, etc. Main contributions are summarized as follows:

Firstly, proposing a vision-guided collective anomaly detection method in the graph feature space. Based on the node feature representation of large-scale graph data, one constructs a histogram for correlated features to show the distribution and aggregation properties of nodes of non-Euclidean space, and analyzes the collective anomaly patterns corresponding to the micro-clusters. The proposed detection method, EagleMine, identifies the clusters in the histogram, and describes it with a vocabulary-based summarization model and statistical test. Experiments on the large real-world data show that EagleMine recognizes clusters that are consistent with the visual recognition of the

human eye and describes the histogram in a more concise manner; EagleMine has better performance in anomaly detection than the graph-based detection methods; it can also be used to detect synchronous fraud-retweeting behavior patterns in retweet-events.

Secondly, proposing a unified detection method based on spectral theorem for the dense-connected anomaly pattern in the graph, which studies the different dense subgraph detection goals in various backgrounds of the graph and related collective anomaly behaviors. Based on the theoretical analysis and comparison of some problems closely related to the dense subgraph detection in various practical scenarios, including the MinQuotientCut of a graph, community with sparse cut, contrast dense subgraph detection and suspicious densest subgraph detection, etc., the thesis proposes a unified problem formulation (GENDS) for generalized dense subgraph detection, analyzes the optimization properties based on spectral theory, and designs a fast and scalable detection algorithm SPECGREEDY based on the properties of the large graph. The experimental results on 40 real-world networks show that SPECGREEDY runs 58.6× faster than the baselines for the densest subgraph detection and achieves subgraphs with better quality; the SPECGREEDY algorithm is linearly scalable with the size of the graph and spots the sudden-appear collaboration groups in a large time-evolving co-authorship network.

Finally, proposing a hierarchical-awareness detection method for collective anomaly patterns in the attribute graphs. With the sparse tensor to model the multi-dimensional correlational data and based on the unified formulation for the density of subtensor and the dense subtensor detection problem, the thesis proposes a novel framework CATCHCORE to fast and effectively detect the hierarchical dense subtensors, provides the evaluation metric with the principle of minimum description length. The experimental results on the large real datasets show that CATCHCORE has the best performance in detecting dense subtensors and anomaly patterns; it can identify explainable hierarchical dense anomaly patterns more accurately in real scenarios, such as social networks, cyber-attacks, and research cooperation; and it is linearly scalable to all aspects of the tensor.

In summary, this thesis explores and analyses the collective anomaly patterns in large graphs from three aspects, i.e., correlated feature space, topological connection structure, and multi-aspect multi-layer structures. Also, the thesis devises efficient and

scalable detection algorithms for those patterns, verifies their superior performance on many large real-world datasets; it also provides semantic analysis and interpretation for these detected suspicious patterns according to different application scenarios.

Keywords: graph mining; collective anomaly detection; pattern recognition; algorithm design; complex behavior

目 录

图形列表

表格列表

第 1 章　引　言

随着社交媒体、电子商务、物联网等技术及各类应用的快速发展，使天然具有相互依赖、长距离相关等特征且包含富信息的图式关联数据以爆炸式增长的方式产生和积累，由此带来了对这种图式数据的高效处理、潜在知识的有效挖掘的巨大需求；在拥有海量关联数据的真实场景中，准确地识别和检测具有群体性特征的异常模式成为众多研究和不同领域的关注焦点，并应用于包括金融风控、工业制造、网络安全、医疗健康等场景；其中，关联数据形态、异常表现及用户行为等方面所呈现的复杂、多样性特征给传统的检测技术带来了前所未有的挑战。本文重点研究对大规模图数据在不同数据形态中群体异常特性的探索和刻画、实现对多种聚集性群体异常模式的高效准确检测，并在大量真实数据中验证检测技术的应用效果、分析识别到的各种有趣模式。

本章旨在简要阐述相关研究背景，总结本文的研究问题和所面临的挑战，并概括本文研究的主要目标和内容、主要贡献和基本概念，最后给出章节安排。

1.1　研究背景

信息时代中的海量数据成为当代新科技革命中的无色金子，并催生了以知识发现和智能认知为发展目标的数据科学领域。随着物联网、移动互联网、社交媒体服务、电子商务平台等的快速发展，形式多样的各类应用业务正在以高效数字化的手段实现万物互通互联。社交媒体中的 Facebook[1]、微信[2] 等提供了联系好友的社交服务，Twitter[3]、新浪微博[4] 等提供了短消息分享、名人关注与互动的博客服务；在线电子商务中的 Amazon[5]、淘宝[6] 等提供了连接买卖双方的网购服务；互联网金融中 Paypal[7]、蚂蚁金服[8] 等提供了在线便捷支付及其他金融服务；此外，还有各大互联网公司推出的各类智能 IoT 平台、兴趣推荐服务以及其他的

[1]脸谱，美国社交网络服务网站: http://www.facebook.com

[2]微信，中国腾讯公司推出、现有用户规模最大的通讯和社交软件: http://weixin.qq.com

[3]推特，美国微博客服务网站: http://www.twitter.com

[4]新浪微博，新浪网推出的微博客服务网站: http://www.weibo.com

[5]亚马逊，美国最大的网络电子商务公司：https://www.amazon.com

[6]淘宝，阿里巴巴集团投资创建的网络零售服务、电子商务平台：https://www.taobao.com

[7]贝宝，全球最大的在线支付平台：http://www.paypal.com

[8]蚂蚁金服，中国领先的互联网金融产品：https://www.antfin.com

信息服务系统等。这些广泛存在的应用平台和服务产生和积累了大量的关联型数据，包括社交网络、用户行为、系统日志等，这里的数据对象之间存在相互依赖、长距离的相关性同时包含许多富信息 (包括对象的属性信息、关系的描述内容等，如时间、权重、类别)，对其背后蕴含的潜在价值的挖掘已成为关注焦点，并用于广告投放、内容推荐、状态监控、反欺诈等不同场景，为服务商带来了不菲的市场盈利；另一方面，群体异常模式挖掘和行为数据分析已逐渐成为众多领域和应用中的核心需求和重要工具，如金融风控、推荐系统、安全保护等；同时，它们对国民生产和国家安全也具有重大意义，包括工业制造控制、网络舆情监控、通信安全保障、金融市场监管等；而且层出不穷、不同形式的群体异常对个人和社会有直接影响，如在电信通讯、金融保险、社交媒体、网络安全等实际在线场景中出现的电信诈骗、信用风险、作弊欺诈、洗钱交易与恶意刷单、网络恶意攻击 (DDos 攻击、病毒传播) 等都与我们的生活息息相关。

奇虎 360 和腾讯公司的研究数据表明 [1]，2016 年中国国内共发生 5 亿次电话诈骗，造成经济损失达 164 亿美元，且其中只有不到 3% 的案件得到处理和解决；据 Skymind 最新估计[9]，欺诈交易每年使美国银行损失高达 110 多亿美元；同样 UK Finance 在“Fraud the Facts 2020”的统计报告中也显示[10]，2019 年通过欺诈或诈骗手段造成的英国金融损失超过 12 亿英镑，欺诈仍然构成重大威胁，且犯罪手段也愈加高级；2019 年微软委托 Frost & Sullivan 公司进行的一项研究显示[11]，网络安全事件给亚太地区造成的潜在经济损失可达 1.745 万亿美元。此外，针对电商卖家的专职羊毛党通过恶意操作致使某淘宝店家一夜破产，损失 700 万人民币的新闻报道[12]更是骇人听闻。因此，以复杂关联形式出现、形式各异的群体异常模式和恶意行为，直接影响着个人、社会、企业及国家的诸多方面，对这些关联性异常模式的检测研究的重要性不言而喻。

按照数据特征及行为表现，异常一般分为由单个数据实例形成的点异常、由环境内容定义的上下文异常以及多数据实例构成的群体异常。在前面所述的场景中，这些关联型数据是海量规模的，在数据类型、对象关系、表现形式等方面具有复杂特性，其中包含了在结构、拓扑、内容、行为等多方面、多形态、多类

[9]Skymind 统计数据来源：https://skymind.ai/wiki/fraud-detection

[10]UK Finance 报告：https://www.ukfinance.org.uk/policy-and-guidance/reports-publications/fraud-facts-2020

[11]Frost & Sullivan 公司研究数据来源：https://cybersecurityventures.com/cybersecurity-almanac-2019/

[12]新闻来源：http://www.nbd.com.cn/articles/2019-11-07/1384423.html

型的异常模式，它们是点异常和上下文异常所不能准确刻画和描述的，例如，网络中的共谋欺诈、协作模式、同步行为等；另外，传统的用于处理无结构型数据对象 (欧式空间多维点数据或序列性数据) 的方法并不适用这些复杂关联数据；即使可以通过特征提取的方式予以转换，也难以满足后者对数据的独立同分布基本假设；监督学习类模型和方法通常受限于有标注数据的可获得难度、数据假设的适用性以及模型可扩展性等问题；而图建模的方式具备特有的优势，包括其强大的表达能力、对数据自然属性和问题关联属性的刻画及其健壮机制等，其直接建模和刻画关联数据的复杂依赖关系及富信息内容、反映不同类模式的代表特征。因此，本文通过图的方式建模关联型数据，探索和分析其中包含的复杂模式，挖掘和检测海量数据中具有聚集性特征的群体异常行为。

该课题除了对实际应用的重大意义外，也具有非常重要的科学研究价值。

首先, *基于图数据的异常模式与传统异常模式检测、行为分析相比更加复杂*。相比与传统异常检测关注的欧式空间下无结构的多维向量数据或具有上下文的序列型数据，刻画具有相互依赖、长距离相关性的关联型图数据时带来了数据呈现和问题定义的复杂性、搜索空间指数增长等新挑战。真实场景下的实用研究需要处理具有规模庞大、动态变化以及包含丰富内容等特征的复杂关联数据：以社交媒体为例，2019 年统计结果显示 Facebook 的月活跃用户超过 24.1 亿；通常用户的行为和相互关系会随时间动态新增、删除或更新；对信息的描述和刻画通常会包含多种类型的，如基本属性信息、超链接以及连边上附带的时间戳、评分与评论、表情符号等内容；同时，庞大的用户群在复杂环境下 (移动设备、台式终端等) 的行为及交互数据也承载了更大的内容复杂度。因此，如何进行数据的高效的表示、处理和分析，设计具有性能保障、快速可扩展性的算法是关联数据知识挖掘区别于其他传统研究工作的重点，需要借助图谱分析、近似计算、随机算法、流式模型、分布式处理等多种分析技术予以解决。

其次，*大规模实用异常检测研究需要交叉学科知识的支撑*。由于异常的定义复杂性和实际场景的多样性，关联数据中的异常检测依赖于统计学、机器学习、信息论、数据挖掘等相关领域的指导和明确化；而实用检测技术的探索和应用需要结合应用领域知识，如疾病诊疗、金融安全等领域。此外，实现可解释性异常模式的分析、描述和应用，需要结合经济学、社会学、行为学和传播学等众多领域的知识，用以帮助对用户行为规律和特征的理解，间接地对人和网络系统的行

为的建模和学习产生影响。例如，共谋欺诈、入侵攻击行为的分析中，异常与正常的差异产生机理在于：这些群体异常一般只有有限的可控资源 (如黑色产业所提供的代理服务器、登录账户、IP 地址等)，因此需要充分的资源复用来降低运营成本 [2]；同时为了追求收益最大化，它们需要经常完成相应的任务以达到规模效应 (Economy of Scale)；在其他的实际情况中，也需要确定具有恶意企图或以欺诈为目的的用户会产生怎么样的行为表现？这些可疑行为与正常用户有什么明显差别等？一般异常模式如何进行逆向解释和溯源，明确与真实场景中的对应行为等？只有融合交叉学科知识和多种数据挖掘技术，才能实现在大规模关联数据中异常模式的高效检测、挖掘和分析。

最后，异常检测和模式挖掘是实现基于关联数据的应用算法的重要构成。一方面，基于关联数据的应用多种多样，可疑行为检测可以为反欺诈、反垃圾传播、刑侦与风控等安全问题提出解决方案；而新颖模式的检测可以为确定复杂规律、探索潜在价值等提供支撑。另外，异常检测和模式分析是其他相关应用的第一步，包括系统评估、推荐系统、预测分析、模拟推断等。比如，想要知道在线商品的质量和效果，需要综合考虑商品属性、销售状况和用户评价反馈等，只有移除了其中具有可疑性的恶意评价、刷单刷分记录等异常模式的影响之后才能更加准确地反映真实状况；再比如，在金融信贷业务中，想要评价某一行业、某种资产健康状况，需要进行征信分析、挖掘可疑性产品或记录，才能进行准确的模型构建、训练和预测分析。比如，在生物病理诊断中，需要在蛋白质交互网络中识别出群体异常模式，才能确定其产生过程和致病机理，同时帮助实验模拟、反映正常功能团间之间作用关系等。由此可见，群体异常检测是实际应用算法中的重要部分，是构建和完善相应系统必不可少的环节。

1.2 本研究工作面临的主要挑战

综上分析，海量数据规模、关联复杂性、异常模式多样性等方面给关联型数据中异常模式的挖掘带来了诸多挑战，现有的方法和技术仍不能有效处理复杂关联数据、准确发现复杂异常模式，难以满足在性能、可扩展性等方面的要求，在异常描述、实际应用分析等方面仍存在严重不足。这里，我们分析了在图的不同数据形态下的群体异常模式检测所面临的挑战。

1. 离群点及群体异常的分布多样性。大规模图中节点的行为、相互关系存

在很大的差异性，同时受拓扑连接的结构约束；图中包含单节点的奇异性、关联约束下局部相似性等表现的多种不同异常模式。如何形象地刻画节点及异常的复杂表现、识别具有显著区别的群体异常是大规模图分析所面临的重要挑战；从机器学习角度，基于特征的图表示实现了欧式空间中对非欧数据的刻画，而特征空间中节点和异常的表现与结构有怎样的相对应关系、常规离群点外的多样群体异常模式具有怎样的分布表现是可解释性异常模式挖掘需要解决的核心问题。

2. 图拓扑结构中聚集异常的度量统一性。依据不同的问题设定和应用需求，图结构约束下的群体异常模式存在着相关且多样的表现；形式各异的模式定义和检测手段设计给相似模式间的分析和区别带了较大的困难，统一的度量的探究是图中异常多样性问题研究的重点；在大规模图背景下，与统一度量相匹配的快速异常检测算法的设计是避免重复工具构建的重要方面，也是具有广泛实用性技术需要解决的难点。

3. 多属性图中群体异常的结构复杂性。通过引入多维内容进一步刻画了关联数据本身的特性，结合结构约束形成了复杂的多属性图。因此，具有多维结构的异常模式的表现和特性是多属性图中挖掘中的重要挑战；在大规模数据中，不同粒度的多维结构分析、对应关联的异常行为模式研究是精细化群体异常检测需要探索和解决难题；同时需要为不同的实用场景，如网络攻击、团队协作等，挖掘意义明确的模式和特征。

此外，针对图数据的海量规模和问题搜索复杂性，具有近似线性或亚线性的高效、可扩展算法是大规模关联数据背景下能够实用的性能保证 [3]。真实世界中的数据往往规模巨大，如 Facebook、Youtube 平台的用户数超过 20 亿[13]，Amazon 商品数超过 5 亿种等；对于具有 10 亿节点的图一个时间复杂度为 $O(n^2)$ 的算法，在单机上的运行时间需要 310 亿年，即使花费 10 万亿美元召集全美国 100 亿台机器也需要运行 3 年的时间，这显然是不可接受的。因此，大规模关联数据对算法和模型的设计提出了更加严峻的挑战。

1.3　本文的主要工作

针对上述挑战，本文确定了如下研究内容，之后讨论了各研究内容间的关系，并概括了本文主要贡献。最后为行文方便，列出了文中常用的基本概念说明。

[13]统计数据来源：https://datareportal.com/

1.3.1 研究内容

本文重点研究了大规模图数据中的聚集性群体异常模式，分析其在不同数据形态下的复杂特性和行为表现，设计具有可扩展性的检测工具，并应用于实际场景中异常模式的分析和解释。因此，本文的研究内容包括：

1. 图特征空间中的微簇聚集群体异常：研究以节点特征作为图数据的刻画方式、在欧式空间中对应的点分布表现和聚集性质，以优化的聚类方式准确、有效地检测离群点之外的群体异常模式。

2. 拓扑关联结构中的稠密子图聚集群体异常：研究在不同应用场景中以稠密子图结构为代表的群体异常模式，以统一形式化表达的方式解决异常定义的多样性问题，设计高效的检测算法。

3. 多属性图中的层次核结构的群体异常：针对关联数据形式及异常模式的结构复杂性，研究多属性图中的多维稠密聚集模式的形式化表示，探究具有结构和密度层次化嵌套的群体异常模式，设计准确高效地检测算法。

此外，文本的研究着力于可扩展检测算法设计与真实场景的应用分析。针对群体异常的分析与应用，在社交网络、学术合作、购物评分、网络攻击等多种场景的大规模真实数据中，对检测算法的性能进行验证，并从结构特征、用户行为等方面对此类异常模式进行溯源分析和解释。

1.3.2 研究内容间的关系分析

针对大规模图数据中的聚集性群体异常模式，上面的研究工作分别从三个不同方面（即相关性特征、拓扑连接结构、多元层次结构）就其表现形式、建模量化、检测方法、实用分析进行了探究；这三个方面在数据模型、模式特征存在着一定的互补关系，并针对性地设计了不同的高效算法进行检测和分析。

通过可疑微簇定义的异常模式以点类簇形式反映了非欧式数据（关系图中具有相似行为的对象）在特征空间投影的直方图中的聚集性表现，这些对象在图拓扑结构上可能表现为稠密子图、具有伪装性的连接关系、具有相似连接目标的节点组（而其连接可能不稠密、目标对象不一定很重要）等，因此检测的类型更加丰富、结果更加准确，弥补了只通过检测图中密集子图来识别群体异常的缺陷和不足；而基于统一形式化框架和检测算法的最密子图模式系统性地分析和挖掘多种类别的密集连接（如可疑子图、对比模式、最小割集等），一定程度上

解决了具有问题依赖的群体异常定义的多样性，而基于图谱性质的检测算法保证了性能和准确性，同时可以识别跨图的异常模式，这是特征方法不能处理的问题；另外，多属性图是拓扑结构在复杂数据上的模态扩展，张量是矩阵表示在数据模型上的高阶扩展；而稠密子张量模式反映了多维数据的聚集性异常（包括属性、拓扑等）以及更加丰富的语义特征，它们是通过张量中各元素的联合形式来构成，而在张量某一切片或者由属性聚合之后检测的结果不一定具有相同的特征（导致基于稠密子图检测的方法失效）；此外，层次化的稠密模式捕获了更加复杂的结构特性，检测到了不同于一般的稠密子图和子张量所对应的群体异常模式，设计的梯度优化策略和检测算法能够高效准确地进行大规模多维数据分析和模式检测。

因此，这三个研究内容在模式特性、检测算法、溯源分析等方面既相互区别又紧密联系，综合实现了大规模图数据中的聚集性群体异常的模式识别、高效检测和实用研究等。

1.3.3　本文主要贡献

针对上述研究内容，本文研究工作的主要贡献包括：

1. 图特征空间中基于视觉引导的聚集性群体异常检测

为了研究图式关联数据的复杂连接在欧式空间中的表现以及聚集性群体异常行为的特征，在基于节点特征的大规模图数据表示下，本文通过多特征组合构建二维 (或高维) 相关性直方图映射，用以反映在欧式空间下图节点的分布表现和聚集性质，并根据以微簇结构为代表的稠密异常分析原图在结构特征和隐含模式 (如社区结构、欺诈攻击、共同协作等) 等方面的群体异常行为。针对直方图中微簇分布的复杂性，本文提出了基于视觉引导的检测算法 EagleMine 来识别和总结其中的节点组、并检测异常微簇。借助于具有多层分辨率感知的 water-level tree 结构，EagleMine 以层次化的方式发现直方图中的类簇；用基于统计分布的词典模型对各类簇进行总结，利用统计检验对树形结构进行探索与检测，以确定用于直方图描述的最优类簇组合；通过对各类簇可疑性的度量，EagleMine 能够识别其中的异常微簇结构。在大规模数据上的实验表明，该分析方法能够解决具有不同区分度的多类簇聚类识别，能够识别到与人眼视觉感知结果相一致的类簇，对整个直方图的总结能够得到简洁的描述；在注入实验及微博数据的群体异常检测实验中，验证了 EagleMine 在图特征表示下的检测性能优于基于图的检测

方法；此外，EagleMine 算法还可用于检测时序转发事件中的同步转发行为模式。

2. 针对拓扑关联密集异常的统一图谱检测算法

为了探究关联数据中异常定义的问题多样性，本文研究了图数据在不同目标背景中以稠密子图为对象的拓扑密集关联及相关的群体异常行为模式的高效检测。本文从理论分析的角度比较和对比了许多实际场景中与稠密子图检测密切相关的一些问题，包括图的最小割集、社区发现、对比密度子图、最密可疑子图等模式的检测；并提出了一种用于最密子图检测的统一形式化框架 (GenDS)，利用大规模图的谱分布性质与幂率特征，结合贪心策略设计了一种简单、高效的检测算法 SpecGreedy 来解决此广义问题；并在 40 个来自不同领域的真实图数据上进行了广泛的验证。实验结果表明，对比与其他基线方法，在最密子图检测中 SpecGreedy 算法可将检测速度提高 58.6× 倍且得到具有密度更大或者近似相同的子图结果；此外，SpecGreedy 算法是随着图的大小线性可扩展的；并在实际数据中用于检测群体异常以证明了算法的有效性：在大规模的随时间变化的学术共同作者网络中发现了突现的团结构合作模式。

3. 多属性图中层次结构感知的群体异常模式发现

针对关联数据的多元异构形式以及异常模式的结构复杂性，本文研究了在多属性图背景下的多维稠密聚集模式以及具有层次性的群体异常行为模式的检测。以稀疏张量建模多维关联数据，本文提出一种新的框架 CatchCore，能够对以稠密子张量为代表的异常进行表示和分析，并有效地发现在连接结构和密度上具有层次性分布的稠密子张量模式。文中首先形式化地提出了一种针对稠密子张量检测的、可以通过梯度优化方式计算的统一密度指标，并囊括现有的绝大多数密度衡量指标，具有很好的可扩展性和解释性；以此为基础，CatchCore 通过层次交替优化的方式检测具有层次性稠密的子张量，算法的收敛性和可扩展性有理论保证；最后，给出了基于最小描述长度准则 MDL 的评价指标来度量检测结果的质量，并用以选择最优的层次化稠密子块。在大规模真实数据上的实验结果表明，CatchCore 在稠密子张量检测和异常模式识别问题中性能都优于目前最优的对比方法；同时，CatchCore 能够更加准确地识别出具有可解释性的有趣模式，包括可疑的好友关系连接、带周期性的网络入侵攻击模式以及 DBLP 中具有层次性紧密合作行为的科研学术团体等群体异常；此外，CatchCore 的复杂度与张量的各方面指标呈线性相关，保证了快速、可扩展的模式检测性能。

1.3.4　基本概念

这里，我们给出文中使用的一些关键概念与说明。

图 (graph): 一个图或者 网络 (network) 是用于建模关系的一种数据结构。数学上，一个图 $\mathcal{G}$ 对应一个集合的有序对 (V, E)，其中 E 中的每个元素表示一对 V 中的元素。V 中每个元素称为一个节点，E 中每个元素 $e = \{u, v\}$ 称为节点 $u, v \in V$ 之间的连边。例如，在实际场景中

- 社交网络：节点集 V 表示人 / 用户，边集 E 表示他们之间的朋友关系；
- 网络安全：节点集 V 表示主机或 IP 地址，边集 E 表示 IP 之间的连接 (HTTP 访问、数据传输等)。

二部图 (bipartite graph): 对于给定的图 $\mathcal{G}$，如果其节点集可分为两个不相交的集合 V 和 W，使得 $\mathcal{G}$ 中的每条边连接某个 V 中节点和某个 W 中节点，则 $\mathcal{G}$ 称为二部图。例如在电子商务的实际场景中，图的节点集 V 和 W 分别表示用户和商品，边集 E 表示用户和商品之间的购买或评分关系。

节点邻居 (neighbor)、度 (degree): 当两个节点 u 和 v 之间存在连边 ($\{u, v\} \in E$)，则称 u 是 v 的一个邻居 (或 u 与 v 相邻接)；对任一节点 v，其所有邻居表示为 $\mathcal{N}_v$；节点 v 的 度定义为其邻居数 ($|\mathcal{N}_v|$)。

子图 (subgraph)、团 (clique): 图 $\mathcal{G}' = (V', E')$ 是图 $\mathcal{G} = (V, E)$ 的一个子图，满足 $V' \subset V$ 且 $E' \subset E$；如果 V 的一个子集 V' 是一个团，则 V' 中的每对节点都由 E 中的一条边相连接，即唯一性的边数为 $\frac{|V'|(|V'|-1)}{2}$。

导出子图 (induced subgraph): 给定一个图 $\mathcal{G}$，对于节点子集 V' 的导出子图 $\mathcal{G}'$, 它是由 V' 作为节点集、$\mathcal{G}$ 中连接 V' 中节点的所有连边组成。

张量 (tensor): 一个张量表示实体的一个多维数组，一个张量的阶 (order) (也称为 模 (mode)) 指的是其维度大小。一个 N 阶 (N-way) 的、大小为 $I_1 \times \cdots \times I_N$ 张量表示为 $\mathcal{X}$，其中每个实体为 $x_{i_1 \ldots i_N}$，它第 n 个索引 i_n 的取值范围是 $[1, I_n]$，I_n 为第 n 模的维度 (dimension) 大小。当 $N = 1$ 时，$\mathcal{X}$ 表示的是一维数组；$N = 2$ 时，$\mathcal{X}$ 表示的是二维矩阵。张量比图具有更强大的表示能力，连边具有多属性的图可以表示成多维张量。如实际场景中

- 社交网络：一个 3 阶的张量 $\mathcal{X}$ 的各个模分别代表人，人 和 日期；则其中的实体 $x_{i_1 i_2 i_3}$ 表示 第 i_1 个人与第 i_2 个人在第 i_3 个日期成为了朋友，否则 $x_{i_1 i_2 i_3} = 0$。

- 网络安全：一个 4 阶的张量 $\mathcal{X}$ 的各个模分别代表IPs，IPs, 协议 和 时间戳；则其中的实体 $x_{i_1 i_2 i_3 i_4}$ 表示 第 i_1 个 IP与第 i_2 个 IP 通过第 i_3 种协议在第 i_3 个时间点建立了连接；
- 电子商务：一个 3 阶的张量 $\mathcal{X}$ 的各个模分别代表用户，商品 和 时间；则其中的实体 $x_{i_1 i_2 i_3}$ 表示 第 i_1 个用户在第 i_3 个时间点购买了第 i_2 个商品。

子张量 (subtensor): 如果一个 N 阶张量 $\mathcal{Y}$ 是 N 阶张量 $\mathcal{X}$ 的子张量，则 $\mathcal{Y}$ 是通过移除 $\mathcal{X}$ 中的某些切片得到的，其中，N 阶张量的切片是通过固定某一模的索引得到的 $(N-1)$ 阶张量，第 n 模切片是通过固定第 n 个模索引得到。

因此，在之后的章节中利用图和张量来表示关联型数据 (区别在于是否存在多维属性)；并在上下文不引起歧义时，二者等价使用。

群体异常 (collective anomaly): 数据中相对整个数据集具有异常性表现的实例集合；单独考察某个实例可能是正常类型，多个实例同时出现则表现为异常。

聚集性群体异常: 具有 (局部) 密集性特征表现的群体异常，其具体表现形式可能对应于稠密聚集、同步行为等模式，在实际应用场景中有不同的语义解释。

本文中图的聚集性模式涵盖了在不同数据形式下具有密集性特征的多维点的集合 (微簇)、由连接结构定义的子图以及多维数据在张量模型下的子张量 (子块) 等形式，在第二章中会对异常进行详细的说明和特征分析，而其中一类就是具有密集性特征的模式，且在实际场景中对应不同的具体含义，它们可能对应于团或社区结构、群体欺诈行为等。

1.4 本文的组织结构

本文共包含六章内容，各个章节的组织安排如下。

第一章为引言，主要介绍针对大规模图数据中的群体异常模式挖掘的研究背景和重要意义，概括了当前研究工作所面临的挑战；然后总结了本文的研究目标和主要贡献点，并对相关的重要概念作了说明。

第二章对异常检测和图数据异常分析研究进行综述。首先，对异常定义和性质进行了总结描述，对异常检测研究所面临的挑战做了系统性的概括，并对现有的异常检测方法、采用的基本假设以及评价标准等方面进行了归纳和概述；然后着重关注基于图的异常检测部分，分析了图建模带来的优势、图异常检测的基本问题和特有挑战，并对现有的图异常检测方法进行了总结，依据图数据、检测方

法类型分类介绍了相关技术；最后与本文的研究工作进行了对比总结。

第三章研究图特征空间中基于视觉引导的聚集性群体异常检测。本章中首先阐述了通过图的相关特征构造直方图并以直观方式进行模式挖掘的动机和优势，展示了直方图中的数据分布特征，概述了本章的研究成果；然后分析和总结了本章的相关研究方法；接着给出了相关概念和所用的总结模型；随后详细描述了本文所提出的 EagleMine 算法及理论分析；最后在多个大规模真实数据上对算法的性能和群体异常检测效果进行了验证、评估和分析。

第四章研究针对拓扑关联密集异常的统一图谱检测算法。本章首先概述了稠密子图检测问题在广泛应用场景中的多样性形式，对现有检测方法进行了概括和分析，并总结了本章的研究成果；然后介绍了一些与本章相关的研究工作和进展；随后给出了广义最密子图检测 GenDS 的形式化定义、总结了与其他相关问题的对应关系；在给出 GenDS 问题优化与图谱性质的理论分析之后，描述了本文提出的 SpecGreedy 高效检测算法及复杂度分析；最后在大量真实图数据上验证了算法检测的性能、分析了算法检测群体异常的结果。

第五章研究多属性图中层次结构感知的群体异常模式发现。本章首先阐述了稠密子张量模式检测的实际意义、分析了现有检测方法的基本假设和缺陷、概要性总结了本章的研究成果；然后对本章相关的研究方法进行了概述；在给出了以张量建模的多维关系的概念与表示之后，提出了最密子张量和层次化稠密子张量检测的形式化定义及分析；随后详细描述了本文提出的 CatchCore 检测算法；最后在多个真实数据上对 CatchCore 算法的性能进行了验证，并分析了在不同场景中算法检测群体异常行为的结果。

第六章对本文的结果研究工作进行了总结，概述了本文的贡献和创新点，并对未来研究工作进行了展望。

第 2 章　研究现状与相关工作

异常检测是数据挖掘中一类具有广泛应用场景和实用价值的重要任务，应用于包括电信、金融、安全、医疗等众多领域；随着英特网、物联网、社交媒体、电子商务以及多样行为数据等各种相关应用的发展，针对关联型图数据的异常检测技术吸引了极大的关注。区别于传统的、无结构的多维点集数据及对应的异常检测技术，图刻画了对象间具有相互依赖、长距离相关的性质及其他丰富信息，也由此带来了新的挑战，并诞生了许多新颖的图异常检测方法。

在本章中，首先系统地总结了异常的特征、与相关概念的联系，分析了异常检测的挑战、不同技术的基本假设和结果的评价标准等；之后针对图数据的异常检测，文中概括了其优势与挑战，对典型的相关研究方法和成果进行文献综述。

2.1　异常检测的性质与挑战

通俗地讲，“异常检测”(anomaly detection) 目的在于发现数据中不符合期望行为 *(非标准)* 的模式。这里的“非标准模式”通常包含不同应用领域中的异常 *(anomalies)*、离群点 *(outliers)*、不一致的观测 *(discordant observations)*、例外 *(exceptions)*、畸形 *(aberrations)*、惊喜 *(surprises)*、特殊 *(peculiarities)* 及污染 *(contaminants)* 等 [4]。在异常检测文献中使用最多的是 异常和 离群点，有时这两者是可互换的。其中，文献 [4] 对异常检测做了系统性的全面总结：从异常的性质和特征、问题定义、应用场景等方面做了详细地分类说明，并分析了各类别中所采用技术的基本假设、区别及优劣势，提供了对各类方法更好的理解。

2.1.1　何为“异常”?

虽然上面给出了异常检测问题的简要描述，但这一定义并不是唯一的，因为词语“异常”或“离群点”的一般含义往往是不明确的，而只有在给定上下文或在具体应用中其定义才具有明确的实际意义。

Hawkin [5] 于 1980 年首次给出关于离群点的定义，其描述为：

定义 2.1 (Hawkin 对离群点的定义, 1980)**.** “An outlier is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism.”

显然，Hawkin 给出的是一种非常一般的定义和概念，由此也使得对应的检测任务成为一个开放性问题。因此，在不同的上下文和实际应用中，异常检测问题存在多种相对应的定义及称谓，像离群点 (outlier)、异常 (anomaly)、爆发 (outbreak)、事件 (event)、变化 (change)、欺诈 (fraud) 等等。虽然异常产生的原因可能各不相同，但其中一个共同特点在于它们对于分析者来讲是有趣的 (interesting)；尽管具有一定的主观性，但异常的“趣味性”(interestingness) 或与现实生活相关性是异常检测的关键特征。数据挖掘的异常检测任务对应的众多应用场景包括：计算机网络中由于挟持产生的异常流量模式、信用卡转账记录或电话通信中的欺诈行为、MRI 图像中的恶性肿瘤影像、传感器中的异常信号以及软件源码中异常的系统接口调用等。

另一个与之相关的话题是“新奇检测”*(novelty detection)*，其目的在于检测数据中之前未观测的或未出现的、新颖模式，它们的产生与正常数据具有相同的机制。区别于异常模式，这种新奇模式在被发现之后通常会成为正常模式的一部分参与后续分析 [4]。而异常检测的方法也经常用于新奇检测，反之亦然 [6]。

2.1.2 异常检测挑战

针对前面的广义异常定义，对应最直接、最简单的异常检测方法是识别和定义“正常”行为的区域，然后将数据中不属于正常区域的数据实例归类为异常。但实际中由于问题定义的复杂性和数据本身特征，使得这一简单的问题描述和检测方法面临着如下因素带来的诸多挑战 [4, 7]。

首先，针对异常检测问题本身，其复杂性和挑战在于如下的几个方面。

1. 准确、完备地定义正常区域非常困难，而且正常与异常行为的边界并不精确；同时很多领域中正常行为处于不断变化中，导致当前认为的正常行为概念可能在未来并不具有代表性；

2. 不同应用领域中确切的异常概念是不同的，因此不存在一种普适的、应用无关的检测技术；由于领域的容错性、鲁棒性的差异，即使针对相同的数据也存在很大差异的定义。如医疗领域和股票市场针对较小数据波动的不同敏感度；

3. 由于异常定义本身的相对性，会存在新异常的产生与进化。针对动态场景，尤其在传感数据中，之前属于正常的数据在设备工艺的改善或产品精度的提升之后的检测分析中可能归为异常；对于由恶意行为产生的欺诈类异常，当多数欺诈者掌握检测算法的工作原理之后，会改变操作策略或伪装成正常行为以逃

避检测。因此，对新异常的动态适应和对抗检测也存在巨大挑战；

4. 异常的语义依赖于检测阶段后的“解释”，同时也需要检测结果具有直观可解释性，这包括找出产生该模式的根本原因，并通过用户友好的形式呈现结果以便进一步的分析；此外，需要对异常的“原因”和“表现方式”给出一个合理、连贯的故事说明；而目前许多现有的检测技术虽然性能很好，但是完全省去这一描述和归因阶段，导致其实际检测结果难以理解或令人信服；

5. 缺少用于检测模型训练、验证的标签数据，且手动数据标注难度很大，同时人工标注存在噪声和偏差；手动标注的质量受到很多因素的影响，即使相同的数据在不同的背景下标注的结果也不一定相同，同时交叉验证成本较大；

6. 类别严重不均衡且分析误差的影响具有不对称性。通常异常样本的数量在整体数据中占比很少，同时在不同应用中对正常样本和异常样本的误标产生的代价是不同的，如疾病诊断中癌症病例的漏标或错误标注等；

此外，对于数据的复杂特征方面，大数据处理中的海量、流式和复杂数据在体积、速度、多样性等方面带来的挑战对异常检测问题也同样适用。其中，

1. 数据规模与动态性：随着技术和设备的发展，使得收集的数据规模庞大、增长速度惊人，并且数据的产生有时是实时、流式的。如社交网络中的用户群和消息、快速增长的多媒体数据、金融交易记录等；
2. 数据复杂：数据形态和关系呈多样化，针对图类型数据，可能包含不同类型的关系、节点和连边属性 (如用户描述及兴趣、时间戳和评分) 等。

上面的这些挑战，使得在广义定义下的异常检测问题很难被解决。而现有的异常检测技术也只能解决这一问题的某种特殊形式或子问题；而这种形式的确定和算法的设计依赖于数据的自然属性、数据标签的可获得性、待检测异常的类型、问题的输出形式、具体的应用需求和约束等因素；同时会结合和采用来自不同学科的概念和技术以应用于特殊的问题和设定中，包括概率论、数理统计、机器学习、数据挖掘、信息论、矩阵论（如扰动理论、谱理论）等。这里，数据的自然属性包括数据类型，如数据样本的属性 (离散型、连续型、类别型)、维度 (单变量、多变量)、私密性 (原始数据是否可直接获得)，和数据实例关系，如记录型 (离散点类型)、序列型 (存在序关系)、关联型 (空间数据、图关系数据) 等。

因此，常见的异常可分为如下几种类型 [4, 7--9]：

- 点异常 *(Point Anomalies)*：单个数据实例相对于数据中其他实例是异常的，

则该实例被视为点异常。如个人单次信用卡消费记录中的巨额开销;

- 上下文异常 *(Contextual Anomalies)*：在特定的上下文环境中，某个数据实例具有明显异常表现，则被视为上下文异常 / 条件异常；每个数据实例需要一组上下文属性来确定其上下文或近邻信息以及一组行为属性。以某餐馆为例，其 2020 年二月份的月营业额数据几乎等同于淡季水平，但与历史同期相比的话则属于异常情况（农历春节期间应该属于餐饮业旺季），这是因为“CONVID-19”新冠肺炎疫情下聚集用餐限制所引起的；
- 群体异常 *(Collective Anomalies)*：如果相关数据实例的集合相对于整个数据集是异常的，则称为群体异常；其中的各个数据实例本身可能不是异常，但它们作为集合一起出现是异常的。如人类心电序列中由于心脏过早收缩而产生的连续长时间的低值信号、电话诈骗中的团伙作案行为等。

其中，点异常类型可能存在于任何类型的数据中；通过融合其他上下文信息，点异常和群体异常可转换为上下文异常。

2.1.3 异常检测方法

根据数据中有标签样本的情况，异常检测技术可划分为：

1. 有监督异常检测：训练数据中存在有标签的正常数据实例及异常类别，由此学习一个预测或判别模型，通过未见数据与模型的比较来确定其所属类别。此类方法面临样本标注数据难获得、标注有噪声和类别分布不均衡等问题；
2. 半监督异常检测：训练数据中仅存在正常样本的标注数据、没有异常标注实例，由此构建针对正常样本的模型，并识别测试数据中的异常；
3. 无监督异常检测：不需要训练数据，依据的隐含假设在于“数据中正常样本远多于异常样本”，并通过对异常性的不同定义和度量发现异常实例。如果隐含假设不满足的话，则会导致较高的误报率。

另外，异常检测方法的结果报告方式通常有：

1. 异常分值：检测方法对每个数据实例输出一个分值用于衡量其异常度或者结果的确信度。由此可构造一个有序的异常对象列表，可从中选择排名前几的实例确定为异常，也可以通过阈值进行筛选；
2. 类别标签：检测方法对每个测试数据实例赋值一个标签 (正常 / 异常)，以确定其所属类别。

对于无监督方法，还需要输出检测到的异常模式，如子序列、子图等。这样，基于输出异常分值类的方法，分析者可以依据领域知识进行进一步选择具有相关性的其他异常实例；而仅输出标签的方法，分析者则可以通过方法中相关参数的选择和调节来间接控制筛选过程。

针对不同类型异常的检测技术，Varun 等 [4] 和 Goldstein 等 [9] 对较早期的异常检测方法进行了详细的分类和总结，并简要地对比了各类检测方法的相对优势与缺点等；Salehi 等 [10] 对变化数据上的异常检测方法进行了概述。按类型划分，各类异常检测方法的总结如下 (具体检测算法和技术请参考 [4, 7, 9])：

- 基于分类的异常检测方法 [11]：分类模型有基于规则、神经网络、贝叶斯网络及支持向量机等的不同类型，基本假设在于给定数据特征空间中学到的分类器能够区分正常和异常类别；
- 基于近邻的异常检测方法：根据 k 近邻之间距离 [12, 13] 或者相对密度 [14, 15] 来确定异常，基本假设在于正常数据实例出现在稠密的、范围小的邻域内，而异常则与其邻居偏离很远；
- 基于聚类的异常检测方法 [16, 17]: 包含了如下三种不同假设，分别是：① 正常数据实例属于数据中的某一类簇，而异常则不属于任意一个类簇，此类方法并不直接优化用于发现异常；② 正常数据实例分布于靠近与之最近的类簇中心周围，而异常则与其最近的类簇中心偏离很远，当异常数据实例本身构成类簇时，则此类检测方法将会失效；③ 正常数据实例属于大的稠密类簇、而异常则属于小的或者稀疏的类簇，此类方法通过给定的大小或密度的阈值来区别异常；
- 基于统计的异常检测方法 [18]：通过设计有参数或无参数的统计模型和技术来刻画数据的分布特征，应用统计推断测试来确定未见的实例是否属于拟合模型，基本假设在于正常数据实例出现在随机模型的高概率区域，而异常则出现在低概率区域。有参数模型针对数据假设了底层分布知识 (如高斯模型、回归模型、极值统计等) 并通过给定的数据估计相应参数，而无参数模型没有对底层分布做任何假设或先验，直接由数据本身性质确定 (通过直方图、核密度估计等)；
- 基于信息论的异常检测方法 [19]：利用不同的信息论测度 (如熵、条件熵、相对熵、Kolomogorov 复杂度、最小描述长度等) 来分析数据中的信息内容，基本假设在于数据中的异常引起数据集信息内容的不规则变化；
- 基于谱子空间的异常检测技术 [20, 21]：使用数据中具有大量可变性的属

性组合来找到原数据的低秩近似表示，以确定一种能够用以区分异常实例的子空间 (嵌入、投影等)，基本假设在于数据可嵌入到低维子空间，其中的正常实例和异常出现显著的差异。此类方法包括主成分分析 (PCA)、紧促矩阵分解 (CMD)、非负矩阵分解 (NMF) 等；

- 基于深度学习模型的异常检测方法 [6]：结合深度模型的更强表示能力进行更复杂的数据表示和异常检测，例如文本、图像、视频等数据。基本假设在于深度模型能够得到对数据更加丰富的表示和特征抽取，使得在新的数据表示下的异常和正常数据实例能更好的区分。模型需要结合分类层或子空间方法，实现异常检测或对检测性能的提升。

2.1.4 异常检测的结果评价

为比较不同异常检测方法的性能，目前发明了如下一些验证和评价手段 [22]。

对于有标签数据和有监督检测模型，可以在测试数据集上根据算法的输出结果与真值构造混淆矩阵，并得到真正例数 (TP)、真负例数 (TN)、伪正例数 (FP)、伪负例数 (TN)，可以用准确率 Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$ 来评测模型的分类性能；为了避免由于类别的严重不均衡导致该指标不能更加准确反映结果 [1] 的情况，可以用代价矩阵 (cost matrix) [23] 来代替混淆矩阵，以规定类别错分之后产生的代价或损失，改善算法的检测性能。此外，还可以构造代价敏感型指标：精确率 (Precision $= \frac{TP}{TP+TN}$)、召回率 (Recall $= \frac{TP}{TP+FN}$) 和 F-measure ($=\frac{2TP}{2TP+FP+FN}$)、以及接受者操作特征曲线 (Receiver Operating Characteristic curve，ROC 曲线) 和 AUC \ ROC (the Area Under the Curve) 等。

对于无监督检测方法，现有的相关研究中涉及到的评测方式总结如下，

1. 内部验证：该方法对异常分值（检测算法对数据实例的赋值）进行极值统计度量 (如计算在异常分值经验分布的 p 值)。具体的分值依赖于检测方法，可能是距离、概率似然或者压缩表示的代价等，也可能不直接用于异常检测目的；

2. 定性评估：采用定性分析的非正式过程进行评估，例如描述检测到异常所对应的实际场景，或者依赖于领域知识的解释 (如在医药、健康领域)；

3. 外部来源验证：结合与识别到的异常相一致的额外多源信息 (未使用的) 进行验证。例如通过行为数据、用户资料、文本内容等信息来检验只利用图的结构检测到的可疑模式 (用户、观点、连接关系等)；

[1] 假设正、负样本数为 9990 和 10，即使模型预测所有的结果为类型 0，则得到的准确率为 $\frac{9990}{10000} = 99.9\%$。

4. 人造数据验证：通过人造生成的数据进行验证，例如通过偏好连接 [24]、幂率图 [25, 26]、Kronecker 图 [27] 等图生成模型来产生理想图数据；然后直接注入对应类型的异常或者通过随机重连、属性修改等方式做改变以得到真值，并用有监督的方式评价检测算法对所创建异常的恢复性能，如精确率、召回率等；此种方法便于评价图的性质对算法的影响，如图的大小、直径、节点度分布等；

5. 注入异常检验：类似于上一种方法，此处需选择在真实世界的图数据中进行异常注入。但此种方法在评价时受到原始数据中可能存在的与注入异常相类似的实例 / 模式的影响。

这里，上面的后三种评价方式在基于图的异常检测性能评价中被广应用。

2.2 基于图的异常检测

尽管过去的几十年中已经发明了很多异常检测方法，尤其是对于无结构的、相互独立的多维数据点类型 (点异常)。然而，很多实际应用场景中，包括物理、生物、社会科学及信息系统等，数据对象之间是互相关联依赖的，因此在异常分析和检测任务中有必要考虑这种关系、并解决由此带来的复杂性挑战。

图提供了一种有力的机制用于有效地捕获这种关联的数据对象间的长距离相关关系。正如第一部分所述，由图产生的群体异常检测逐渐得到了广泛的关注。[28] 对近几年图数据中的异常检测技术提供了详细的综述，从检测方式（无监督与半监督）、数据形态（静态与动态、多属性与普通图）等角度出发对现有方法的有效性、可扩展性、可泛化性及健壮性等方面进行分析和比较，并展示了不同领域的多种实际应用。[29] 对基于动态网络中的异常检测技术做了综合性比较和总结，包含了针对节点、连边、子图及变化这四种异常模式的不同检测方法。此外，还有关于针对网络入侵、欺诈等以图数据为背景的应用场景和方法的概述工作 [30, 31] 等。此外，随着深度学习技术的快速发展，也出现了在异常检测领域的应用，[6, 32] 总结了现有的一些深度学习模型，包括基于受限玻尔兹曼机的深度置信网络、深度神经网络、循环神经网络等，用于图数据的异常检测。

2.2.1 为什么需要“图”?

这里我们给出了基于图的异常检测的必要性及其优势分析：

- 关联型数据的自然属性：如前所述，数据对象之间存在相互依赖和关联关系。大多数的关系数据都是内部依赖的，在异常检测中需要考虑这种关

联对象及其影响，如网页链接、通信网络、社交网络、电子商务、蛋白质交互等；

- 图的强大表达能力：通过引入连边，图自然地表示了数据对象之间的内在依赖以及约束关系，对象之间的多路径也有效地捕获了它们之间的长距离相关性或者不同强度的关联性。在图表示下也可以通过添加节点和连边的属性、类型、权重等内容以表达丰富的数据信息；
- 问题之间的关联属性：在群体异常模式中，异常数据实例之间存在关联性；而这些模式只有通过关系的整体表示和建模才能显现，并通过相应方法进行识别和检测。如电信业务中有组织的团伙诈骗、电商服务中的薅羊毛行为、计算机网络中的恶意攻击等场景。
- 具有健壮机制：图能够提供一种对对抗鲁棒的建模工具；如在欺诈检测系统中，登录时间、位置 (IP 地址)、用户属性等信息或特征往往是容易被修改，从而使欺诈者躲避检测；但他们往往没有针对网络的整体视角 (各个网络的连接关系和状态等) 以供其操纵，如电信通信、电子邮件、评分网络等；在不了解整体特征和动态操作的情况下，欺诈者很难通过尽可能与该网络相适应的方式来逃避检测。

根据 Hawkin [5] 对于离群点的通用定义，图中的广义异常检测问题定义为：

定义 2.2 (广义的图异常检测)**.** 给定一个图关系数据 (普通图/带属性图、静态图/动态图)，找到图中的一些罕见对象 (包括节点、连边、子结构等)，它们与图中其他大多数的参考对象存在明显差异。

实际检测中，可以利用一般异常检测方法中的假设予以区分，即异常对象对应于一个或一组数据实例，其满足 **罕见性** (如罕见的属性值组合、低概率样本等)、**孤立性** (如基于与近邻或类簇中心的距离或密度指标的度量)、**新奇性** (数据实例不满足统计模型或者理想假设) 或者 **具有较大的描述复杂度** (如基于信息论方法中的熵、最小描述长度原则) 等特征。

对于图中异常子结构来讲，包含团 (clique)、星型图 (star)、二部核 (bipartite core)、稠密子图 (dense subgraph) 及其他特殊的连接结构、甚至在不同场景下这些子结构的出现、消失、变化的行为表现等。另一方面，社区检测、图的分割等是与之密切相关的研究问题，但是直观上它们的目标并不直接用于异常检测，在结果评价及解释等方面也并不相同；目前已经有很多的针对社区检测类的研究

工作，[33, 34] 等对网络社区检测进行了详细的综述和概括，[35] 对社区检测算法进行了比较分析；此外当前还有一些基于深度学习、表示学习方法的社区发现工作 [36--38]。同时，在图异常检测中也有许多基于社区的方法，由此得到的异常模式包括：不属于任何社区的桥边或者节点、社区内与其他成员的连接或属性等存在很大差异的节点或连边、时序图中出现的社区结构及其动态演变规律等。另外，稠密子图类模式的检测通常直接评价不同定义下的子图密度 [39]，如平均度密度、几何密度等，而社区检测通常的优化指标或质量评价标准有：模块度 (modularity)、(平均) 度 (degree)、嵌入度 (embeddedness)、(平均) 内部 / 外部度、外部连接密度、边割率 (cut ratio)、传导率 (conductance) 等 [40]。因此，社区检测与异常检测相互联系，前者算法也可以作为后者技术的一部分来使用。

2.2.2 图异常检测挑战

除了继承于前面一般异常检测所面临的共性难题之外，图异常检测所特有的挑战包括：

1. *内在关联的数据对象*：数据的关联性质给图对象 (节点、连边、子结构等) 的异常性的量化带来了挑战。区别于传统异常检测中数据的独立同分布 (i.i.d.) 性质 / 假设，图数据中对象间的长距离相关性使得在异常度度量时需要考虑“扩展激活与传播”以及“有罪关联”(Guilt By Associations，GBA 原则) 特性；
2. *异常定义的多样性*：由于图及包含模式具有丰富的表现形式，图中的异常定义比传统的离群点检测更具多样性；与图结构相关的新异常类型对很多应用来讲都是感兴趣的，如交易网络中的洗钱环路模式等；
3. *搜索空间尺度巨大*：与复杂异常 (如图的子结构) 相关的主要挑战在于巨大的搜索空间，其中很多与图论中图搜索问题相关 (通常是指数型的)。而对可能的子结构的罗列枚举是组合问题，使得找出其中的异常更加困难；如果进一步考虑对象属性的话，由此得到的可能搜索空间会更大。

结合大规模数据的实际应用需求，对于图异常检测问题需要设计和发明具有有效性、高效性、可扩展性等性能要求的方法，也需要对流式、时变场景得到实时的快速响应等。

此外，基于图异常的描述 (description) 是解决实际应用中缺少真值标签的重要手段，用以对检测算法及其输出结果的诠释和说明。一方面，可以将具有解释

友好型属性和设计思想直接融入到传统的图异常检测算法，以使得单个实例的检测有更强的可解释性，如矩阵 / 张量分解类算法；另外，在给定一组初始可疑对象后 (如在检测算法返回的有序列表的前几个实例)，根据对象的关联性质找到并刻画它们之间的关系，以便于更好地理解此类异常的根本原因；此处，我们还可以利用交互式图查询 [41, 42]、图压缩总结技术 [43, 44] 等。

2.2.3 图异常检测方法综述

图按照性质划分包括静态图 / 动态图，普通图 / 属性图，对应于不同的异常检测方法，它们可利用的给定信息包括：连接结构、节点/连边属性 (如位置、分值、类型、权重)、实体行为等，并根据异常模式的不同表现予以识别和检测。

2.2.3.1 静态普通图的异常检测

对于给定的普通图，唯一可用的信息就是图的拓扑连接结构。因此，对应的检测方法利用其结构性质和特点找到和识别其中的异常模式。

基于结构特征：包括基于特征和接近度 (proximity) 的方法。对于特征类的方法，通过抽取与图实体 (包括节点、三角形、Egonets 等) 或整个图相关的特征进行量化，如节点级特征 [45--49]、节点组级特征 [50, 51]、图级特征 [52]；然后通过点异常检测方法或相应的统计规律判断异常部分。[53] 等给出计算机系统中基于特征空间的异常检测方法, OddBall [49] 抽取了与 egonet 相关的特征 (密度、权重、排序和特征值等)，并找到了大多数图的 egonet 对应特征服从的模式，由此识别到一些异常的节点及其 egonet 子图；EigenSpokes [54] 找到了图中的奇异值向量所表现出的一种特殊“spoke”模式，并提出 SpokEn 算法能够找到紧密社区模式；[55] 分析了真实图中的 k-core 特征所服从的分布规律，提出了针对大规模图的快速统计方法，并用于检测异常模式 (如 core-periphery、社区结构等)；根据图节点特征空间中的模式，CatchSync [46] 算法用于捕获大规模图中的同步性可疑行为；Net-Ray [56] 考虑成对节点特征构成的相关图，并提出了利用可视化工具来分析百万级节点图中的性质等。另一类检测方法则利用图结构来衡量图中对象之间的接近度，以捕获它们间的自相关性，并将相似度接近的节点归为同一类。图的接近度度量方法有随机游走、PageRank、置信传播 (Belief Propagation, BP) 等，Gyöngyi 等 [57] 采用 TrustRank 方法实现网络中垃圾信息的检测，NetProbe [58] 利用置信传播的方式在二部图中检测在线拍卖网站 eBay 中

具有欺诈共犯的核结构，[59] 中提出了 FaBP 对图中的 GBA 原则相关的带重启的随机游走 (RWR)、半监督学习以及置信传播三种方法统一表示成矩阵求逆问题，并给出了快速计算方法；zooBP [60] 利用 BP 快速确定异构网络中节点的标签，并找到其中的异常结构；Shah 等 [61] 提出 fBox 方法在 Twitter 网络中检测可疑的链路行为。

基于聚类与社区：通过图中稠密连接的、相近的一组节点来识别其中的稠密连接模式以及不属于任何社区的桥边 /点。Sun 等 [62] 研究了图中节点的接近度并用于识别二部图中的异常，[63] 以最小描述长度准则 MDL 为指标分析图的邻接矩阵，将相似的近邻节点进行聚类实现无参数的图分隔，并将其他不属于任何结构的边视为异常; Sun 等 [64] 提出一种新的聚类框架——gSkeletonClu，能够在图聚类之外找到中心点即离群点；[65] 利用基于非负矩阵分解的图社区算法检测其中的异常模式 (如端口扫描、DDos 行为等)；Beutel 等在 Facebook 的“喜欢”页面行为数据中，提出 CopyCatch 方法检测由密集性组攻击 [66] 构成的稠密子图结构；HiDDen [67] 直接检测图中的层次化稠密子图并用于金融异常检测；GetScoop [45] 通过局部搜索的方式来发现二部图中的相关稠密子图；[68] 解释了子空间聚类在密集子图挖掘中的作用；Fraudar [69] 将节点和连边可疑度与最密子图检测相融合，用于检测具有伪装性的欺诈行为。[70, 71] 在二部图中统计 bipartite-core 并用于发现稠密子图。

2.2.3.2 静态属性图的异常检测

对于给定的数据，可能包含丰富的图表示形式，其中节点和连边拥有不同的属性，例如社交网络中的用户兴趣和位置、交易网络中的金额与时间、在线评论中的用户评分或文本信息等。因此，对应的异常模式可能是具有多属性的子结构或者具有不一致属性的节点 / 连边。

基于结构特征：目的在于找到图中在连接方面和属性方面罕见的子结构。[72] 用基于属性和结构的联合聚类方式识别大规模属性图中的内聚性子组; [73] 基于 MDL 检测非频繁的罕见结构以找到给带属性图或者图集合中的异常子图结构；[74] 基于修改、插入、删除对三种异常做形式化，并用来检测那些与正常结构相似但不完全相同的子结构，对应于那些通过对正常数据实例作微小改变以避免被检测的入侵模式；Rev2 [75] 通过计算用户、评分、商品的相互依存的内在质量指标及迭代算法来检测评分系统中的欺诈用户。

基于聚类与社区：将社区中具有与其他节点属性值显著差异的节点视为社区离群点。[76] 将基于图的社区离群点检测与其他相关问题进行了区分，即仅考虑节点属性的全局离群点检测、仅考虑连接的结构异常检测和仅考虑直接邻居属性值的局部离群点检测，并提出了一种概率模型同时检测社区和识别社区离群点；Müller 等 [77] 提出了一种属性图中节点离群点排序技术，通过相关属性的子集 (子空间) 揭示复杂异常；[78] 通过特征向量找到图中基于密度的子空间聚类，将子空间模式和子图聚类方法高效结合起来，可以在检测结果之后筛选离群点；[79] 提出的 FocusCO 算法能找到用户驱动的类簇或属性图中的社区离群点：给定由用户提供的初始节点集后，算法首先找到给定节点具有一致性的属性子集，然后再由此找到图中稠密连接的、属性相同的节点类簇，最后将结构上属于类簇但属性上有偏差的节点定义为异常；Tsourakakis [80] 提出通过质量保障的约束项来抽取优化的近似闭环，使得这种子图比起最密子图更有意义；Akoglu 等 [81] 从在线评论中利用网络效应检测欺诈性评论。

基于关系学习：对应于有标签的检测方法，通过利用对象间的关系学习分类算法。[82, 83] 探索了关系型分类方法以处理不同形式的输入 (节点及邻居标签、节点和邻居属性)；Chakrabariti 等 [84] 利用朴素贝叶斯模型建模对象的局部属性及邻居节点的标签; Fang 等 [85] 通过稀疏正则化和核扩展方法实现基于图的学习；其他模型还包括马尔科夫模型、支持向量机、循环信念传播等。

基于张量表示：用张量来建模多属性关联关系，用于检测其中的稠密子张量等结构或感兴趣区域。依据张量的 CP 分解以及 HOSVD [86] 能被用来检测检测稠密子张量；MultiAspectForensics(MAF) [87] 利用分解结果检测网络流量日志中的端口扫描异常; [88] 利用 CP 分解进行动态张量分析；[89] 提出了一种针对多模数据中稠密块的通用可疑性指标和直观原理，给出的 CrossSpot 异常块检测算法从一个种子块开始，然后逐渐调整直至找到一个局部最优值；此外，M-Zoom [90] 通过对稠密子图检测 [91] 的扩展，以贪心选择的方式检测多维数据中的稠密子块，D-Cube [92] 进一步扩展到对于大规模数据集和分布式文件系统的场景，DenseAlert [93] 能够实时的处理多维流式数据，并检测其中的稠密块和异常行为；Ban 等在 [94] 中考虑图的不同属性对于稠密子块评价的不同贡献程度来检测异常稠密子块；Turbo-SMT 将矩阵-张量的耦合分解的计算提升了 200 倍，并应用与人脑神经元的行为分析 [95]。

2.2.3.3 动态图的异常检测

Ranshous 等对动态网络中的异常检测、相关应用及实例等进行了总结和概括 [29]。此时的异常模式包括图中节点、连边和子结构异常以及事件、变化等，例如边权重突然变化、子结构的突然出现或消失等。[96] 对图的流式和随机算法进行了概述。

对基于特征类方法，同样抽取一些特征将图映射为向量表示 (如一个图对应一个实数值)，然后经过异常检测器的判断输出对应的异常分值或者标签。[97] 考虑了基于社区的六种变化方式，即收缩、发展、合并、分割、诞生与消亡，并根据规则检测其中的异常；[98] 以节点间的带权路径来表示核心集，通过度量节点移除对核心集带来的变化确定节点的异常分值，并选择其中的离群点；[99] 结合 MDL 和递增的张量分解方法，识别数据中周期性的动态社区；SpotLight [100] 通过素描 (sketching) 的方式选择部分节点以实现图的向量化表示，并结合随机割森林 (Random Cut Forest) 来检测图流数据中的异常子图；SedanSpot [101] 设计了基于带重启的随机游走（RWR）度量和采样的技术实时地检测连边流式数据中的异常；[102] 根据结构和边权变化的两类动态图中不同的异常行为以及对递增 RWR 分值度量所引起变化来检测异常行为；Graphscope [103] 通过比较连续时间图的分割带来的编码长度变化，来确定流式图中的变化点；TimeCrunch [104] 在 VoG [44] 的基础上根据结构在时间上的变化特点，引入时间特征词汇来总结动态图，并发现其中的异常模式。

对基于矩阵-张量分解类方法，[105] 利用紧促矩阵分解 (CMD) 对矩阵进行低秩近似，应用到流式图的各邻接矩阵上并根据重构误差进行事件检测；[106] 基于图的稀疏主成分分解来定位异常和识别对应的节点；[107] 通过张量 PARAFAC 分解的方式来识别异常；SMF [108] 通过随时间变化数据的矩阵分解，并明确建模时序关系来捕获其中的周期性模式。

此外，还有基于概率模型的检测方法来建模连边或事件发生的概率，将其中的小概率事件视为异常，其中包括识别邮件中的异常 [109]、基于 MRF 的交易欺诈检测 [58] 以及隐马尔可夫模型和扫描统计的图局部异常识别 [110] 等。

2.3 本章小结

本章对异常检测、基于图的异常分析等研究内容进行了综述，对基本研究问题和研究现状做出清晰、系统地描述和概括，分别就一般异常的基本性质、异常检测的挑战进行了总结和分析，并对检测方法和评价准则进行了分类概要说明；重点对基于图的异常检测进行了详细分析，包括图表示的必要性、优势与挑战等，并对不同类型场景下的异常检测方法按类别进行了总结和对比。

总的来说，异常检测是数据挖掘中一类重要任务，但由于数据复杂度、问题定义的不确定性和多样性以及实际应用的特点等因素带来了很大的挑战，因此不存在一种统一、全面且可操作的定义方式以及具有普适性的检测手段，而更为实用的策略在于与具体应用相结合，根据数据的实际特征、针对异常的可操作性具体假设来设计有效的检测方法。

基于图的异常检测是其中重要且更复杂的一类问题。以图的方式表示相关关系，能够充分、有效地反映数据之间的关联及其中的异常模式、便于以更加健壮的方式进行识别和检测；同时图的引入也带来了新的挑战，体现在复杂的建模因素、异常的度量与定义、问题搜索空间等方面。因此，需要根据图的性质、模式的特征等来设计有效的、可扩展的检测方法。根据可利用信息的不同，现在方法针对不同类型的图、不同场景设计了丰富的检测方法，但对其中的具体模式和问题，这些现有方法也并不能彻底的、完美的予以检测和解决，在检测准确度、算法性能、不同应用适配等方面存在进一步提升的空间。

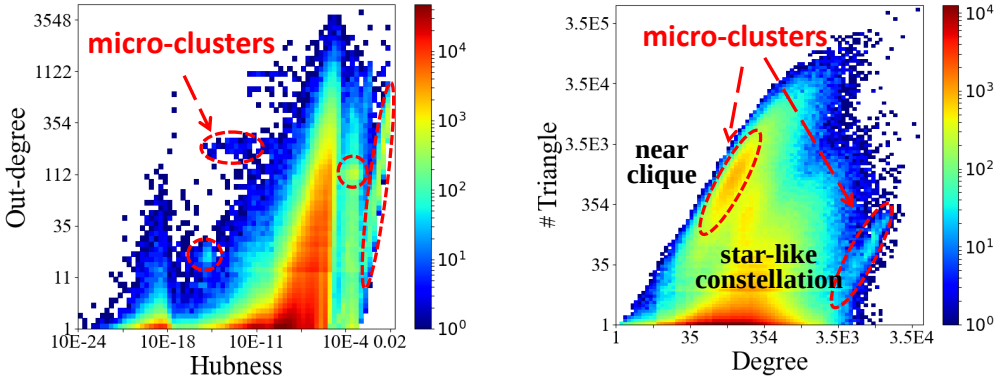
本文的研究主要针对图数据中的聚集性群体异常模式的挖掘，探索在不同数据场景下群体异常的具体表现形式，着力于分析实际应用中的共谋欺诈、垃圾传播、恶意攻击等可疑行为的同步性和密集性特征，并提出快速有效、可扩展的检测方法，并应用于大规模实际应用中不同的有趣模式的发现。在多场景应用实践中，着重对异常行为的意图动机和产生规律进行分析和解释。与当前方法比较，本文的方法具有准确率高、运行速度快、可扩展性和可解释性强等优势。

第 3 章　图特征空间中基于视觉引导的聚集性群体异常检测

本章考察大规模关联数据的复杂连接以及聚集性群体异常在欧式空间中表现和分布特征。基于大规模图的节点特征表示，识别在投影欧式空间下（直方图表示）与图的连接结构相对应的微簇，并应用于群体异常模式的检测。通过结合人类视觉识别和大脑感知的特性，以直观的方式检测直方图中的微簇结构以及异常模式；利用设计精细的、可扩展的检测算法，解决传统的、基于聚类的方式在检测异常时出现的识别有效性问题——源于在复杂数据分布下多类簇的不同区分度，并弥补基于图结构类异常检测方法的假设缺陷和性能不足，提升大规模数据中异常检测效果和结果的直观可解释性。

对于给定的包含数百万点的直方图，这些节点的分布具有什么样的模式呢？我们该如何以类似于人视觉识别检测的功能来识别这些模式，同时分离出那些具有可疑性的点簇呢？更一般地讲，我们该如何在直方图中识别一些微簇结构并检测某些有趣的模式呢？

为了实现上述目标，我们提出了一种基于视觉引导的算法，EagleMine, 来识别和总结给定直方图中的节点组，这里的直方图的构成可以是大规模图中节点的相关特征组合，也可以是时序应用中的时间相关特征张成投影空间，或者更加泛化的一般点云的密度分布图。借助于一种根据视觉识别原理所设计的、具有多层分辨率感知的 water-level tree 结构，EagleMine 能够以层次化的方式发现这些节点组——它们在直方图中形成一些相互分离且内部稠密联通的区域。通过对该树结构的探索与检查，EagleMine 利用统计假设检验的方式确定最优的节点组，同时用基于统计分布为词典的方式进行总结；另外，通过对各节点组可疑性的度量，EagleMine 能够识别一些异常的微簇 (micro-cluster)，这部分节点具有相似的且不同于多数正常节点的异常行为。在真实数据上的实验结果表明，EagleMine 能够以直观的方式识别到与人视觉检测预期相一致的节点组，在直方图分布总结方面也具有比对比方法更优的性能；针对异常模式检测，微博数据集上的实验结果表明 EagleMine 的识别准确率相对当前已有最好的、基于图的检测方法有显著的提升。此外，EagleMine 也能用于其他不同的应用场景，比如在时序转发事件中检测同步行为模式等。



(a) Microblog Sina Weibo **(b) Tagged [112]**

图 3.1 不同实际应用数据中由图节点的相关特征构造的二维直方图示例，分别对应于：(a). 新浪微博数据中用户-消息的转发场景下，用户节点的 Out-degree 与 Hubness 值特征空间；(b). Tagged 数据中用户间的好友关系，节点的 #Triangle 与 Degree 特征空间。

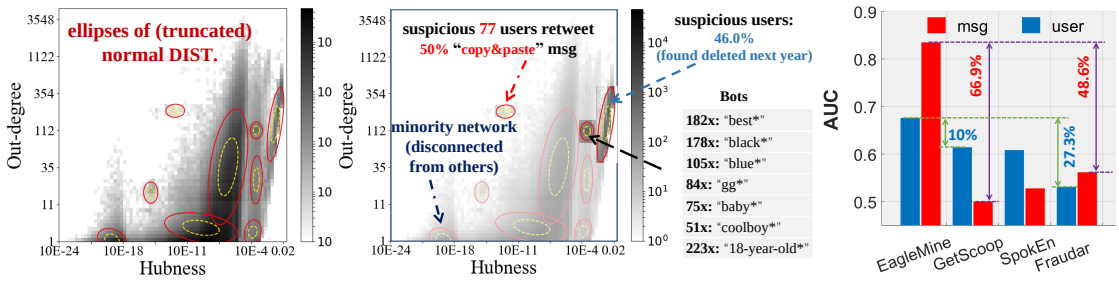
Figure 3.1 Heatmap of correlation plots for graph node in different applications. (a) Out-degree vs. Hubness for users in retweets relation of Sina Weibo. (b) # Triangle vs. Degree of user in friendship of Tagged (color figure online).

3.1 本章引言

对真实世界中包含数百万节点和连边的关系图，能够最直观的探索其性质和状态的方式就是根据图中节点的关联特征构造相关图 (Correlation plot) [111]，即以代表图中节点的散点在二维空间的分布所构成的直方图 (热度图) $\mathcal{H}$ 来刻画散点的聚集状态和密度分布。在这些直方图中，人可以从中直观地识别出一些聚集成不连通稠密区域的节点群组 (如图 3.1 所示)，以可解释强的特点用于帮助发现一些有趣的模式 (如小规模好友社区、共同合作者的协作团体等) 以及某类可疑行为等 (如欺诈、攻击或离群点)。

这里的图可以表示 Facebook 中表示好友关系、Amazon 中用户对商品的评分关系以及 Twitter 上用户和消息的转发关系等，这些图甚至可能是动态变化的。我们能够从图或其快照（snapshots）中为节点提取很多可以快速计算的关联性特征，如节点的入度和出度、参与三角形数目、PageRank 分值等，这些特征之间的组合可生成多种如前所述的相关图。但是依靠纯人工手动方式来分析这些热度图是非常耗时费力的，尤其对于那些动态图场景。另外，如果包含更多组特征的话，高维数据的可视化和模式分析都是非常困难的。

由此产生的一个问题是，给定一个由图节点的某一特征空间所构造的直方



(a) EagleMine 对图中节点分布的总结描述结果

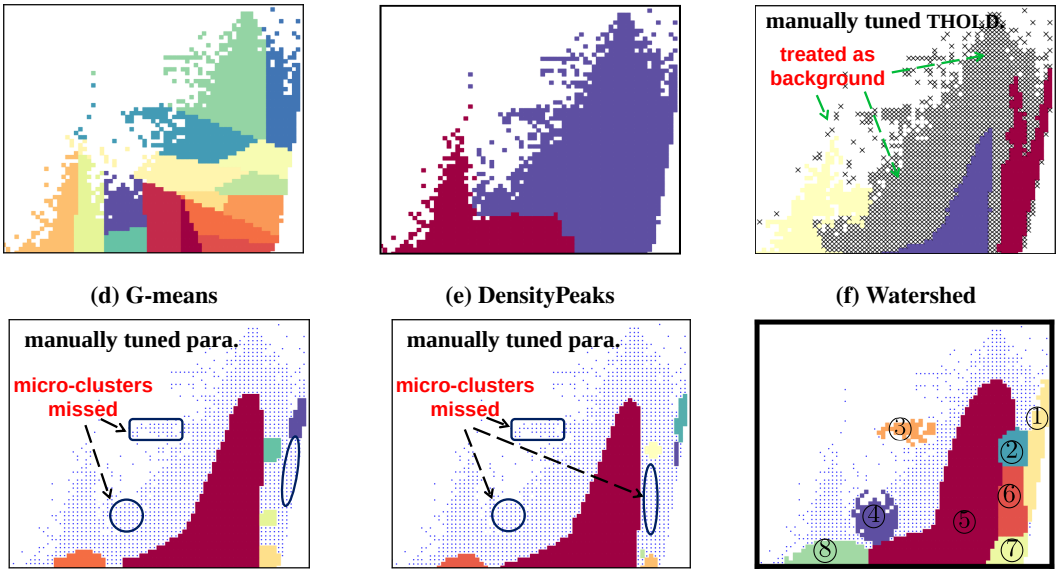
(a) Graph summarization for Figure 3.1a

(b) 部分突出标注的微簇以及由账户名列表所示的异常模式示例说明

(b) Micro-clusters highlighted and anomaly pattern.

(c) 异常用户和消息的检测准确率 (AUC)

(c) AUC for suspicious users and msgs



(d) G-means

(e) DensityPeaks

(f) Watershed

(g) DBSCAN

(h) STING

(i) EagleMine

图 3.2 EagleMine 在新浪微博数据上的检测结果示例。(a) 展示了 EagleMine 利用截断高斯分布对图 3.1a 所示的特征空间的节点分布进行描述总结的结果. (b) 突出显示了一些微簇结构，包括一个非连通的子网络以及最可疑的一些微簇。(c) 对比与其他检测算法，EagleMine 在检测微博数据中的异常用户和消息时得到最优的 AUC 准确率结果。(d)-(i) 是 EagleMine 和其他相关聚类方法在识别节点组时的效果比较。其中 Watershed、DBSCAN 和 STING 算法展示的效果是由手动调节参数达到相对最好的结果。所以视觉展示结果表明 EagleMine 明显优于其他方法。

Figure 3.2 Our proposed EagleMine achieves effective results on Sina Weibo data. (a) EagleMine summarizes the graph nodes in a feature space of Figure 1a with truncated Gaussian distributions. (b) highlights some micro-clusters, including a disconnected small network, and very suspicious micro-clusters. (c) EagleMine achieves the best AUC for detecting suspicious users and messages on Sina Weibo compared to the state-of-the-art competitors. (d)-(h) are the comparison between the baselines and EagleMine for recognizing node groups. Watershed, DBSCAN, and STING are manually tuned to have relatively better results. EagleMine outperforms others by visual comparison.

图 $\mathcal{H}$，如何设计一种新的有效算法能够自动地完成

- 以类似与人类视觉感知的方式识别和检测图中的节点组；
- 总结该特征空间中节点及类簇的分布情况和形式；
- 识别其中包含的一些可疑微簇团 micro-clusters 。

直方图中“微簇”结构指的是那些在特征空间中所表现的行为非常相似的图节点，它们可能对应于原图所表示的实际应用场景中的少部分节点，例如用户、消息、物品等。图 3.1 中展示了部分可能的特征空间所构成的直方图：

- out-degree vs. hubness ——从图 3.1a 中可以识别那些特殊的可疑性微簇，其中包含的节点具有较大的 out-degree 但是很低的 hubness 分值 [113]，它们表示那些被付费购买的欺诈用户连接了许多不重要的节点，即转发了大量的低质量消息（如活动广告、促销内容等）；
- #triangle vs. degree ——从图 3.1b 中可以捕获那些具有不同连接结构的团体，其中有近似于全连接的子图 clique (此部分包含的节点的度虽小，但其与邻居之间都构成紧密连接的三角形)、呈星云状分布的星型子图 (此部分节点的度虽大，但其邻居之间几乎都相互独立，没有构成连接) [114]。

为此本文提出了一种新颖的、基于树的挖掘方法 EagleMine 来识别和总结直方图中的节点组，也能检测其中的可疑微簇。在多个大规模图数据上的实验结果表明，EagleMine 明显优于其他聚类算法（基于密度聚类及层次化聚类等方法），直观上能够得到与人类视觉感知相一致的识别结果，同时以最优编码长度 MDL 为指标的量化评估也表明 EagleMine 能够以更加简洁的方式总结热度图的点分布；此外，算法也能检测到真实的微博数据中由数百个机器人账号形成的微簇，在特征空间中表现出类似的行为模式（如图 3.1a 所示），这些账户具有明显的非正常登录名，如‘best*’，‘black*’以及‘18 岁的 *’等，其他的账户属性信息内容也是相同或类似的。EagleMine 可以扩展适用于除了图数据之外的其他应用场景中以检测和发现其中的有趣模式。

总的来说，本章所提出的 EagleMine 算法具有如下优势和性质:

- **自动的分布总结：**EagleMine 能够对由相关特征构造的直方图进行自动地总结（图 3.2a），并以与人类视觉识别预期相一致的方式识别出直方图中的不连通节点组（图 3.2i）。
- **有效性：**EagleMine 能够识别到具有可解释性的节点组，在量化实验中性

能优于其他对比方法，定性分析也优于经过手动调参的聚类算法（图. 3.2d-3.2h）。

- **实现异常检测：**EagleMine 能在实际大规模图中识别一些结构可解释的异常模式；对比于其他基于图的异常检测算法，EagleMine 在合成数据及微博数据上有更高的识别准确率（图 3.2c）。此外，EagleMine 也可检测图应用场景之外对应于某些可疑模式的微簇结构，如时序数据中的同步行为分析。
- **具有可扩展性：**EagleMine 具有线性可扩展的时间复杂度，与图中节点数成线性相关；也能应用于由多维特征空间构造的热度图。

可复现性：本文提供了开源的程序 (https://github.com/wenchieh/eaglemine) 和大部分在线可获得的真实数据。

本章后续内容中，先对相关研究方法进行了分析和总结，接着给出了相关概念和所用的总结模型，随后详细描述了所提出的 EagleMine 算法及理论分析，之后在多个大规模真实数据上对算法的性能和群体异常检测进行了验证、评估和结果分析，最后对本章的工作进行总结和概括。

3.2 相关工作

本小节中从三个方面概括了相关研究工作，即聚类方法、基于视觉分析的数据挖掘和异常检测。

3.2.1 聚类方法

多维特征空间中的类簇识别：在以多维特征为表示的数据分析中，对于具有高斯型的数据类簇，K-means、X-means [115]、G-means[116] 以及 BIRCH [117] (适用于球形的类簇) 算法都对离群数据点敏感，这些方法都是基于点的距离来判断类簇的，导致其更倾向于将特征空间中距离近的点进行聚类以形成球形区域。基于密度的方法，如 DBSCAN [16] 和 OPTICS [118] ，是对噪声点鲁棒的，也能检测任意形状的和数据分布类型的类簇，但是 DBSCAN 的聚类性能以及 OPTICS 从 reachability-plot 中得到类簇的效果都严重依赖于用户给定的密度参数。RIC [119] 作为一种可以与其他算法相结合的聚类框架，利用最小描述语言作为评价准则来选择拟合分布和分离噪声，从而增强其他聚类算法的效果。STING [120] 层次化地合并位于低层每四个格子的信息，并根据给定某一密度找到满足阈值的类簇。DensityPeaks [121] 定义的类簇中心点为那些具有较高局部密度同时与其他类中心相距较远的点，并根据决策图（decision graph）来选择某些

类簇中心，然后将其他点归类到与之最近的类簇中，其缺点在于需要计算各点对间的距离，所以不适合于大规模的数据，同时还需要手动地确定类簇中心的数目。由分水岭变换（watershed transformation） [122] 衍生出的聚类算法 [123] 将位于不同 watershed 之间的区域视为同一类簇，但它仅关注最后的分割效果而忽略了类簇间的层次结构。Campello 等 [124] 比较了不同的聚类算法，并在 DBSCAN 的基础上提出了一种层次化的“HDBSCAN”算法，但该算法过高的计算复杂度使其不能处理大规模数据 (如大图数据)，而且它所定义的“outlierness”分值和本文的异常模式分析的预期也不相符。此外，基于模块度的聚类方法、社区检测算法 [35] 以及基于图割的方法 [125] 通常很难处理具有百万节点的大规模图，或者在用于图聚类时并不能给出直观的、具有可解释性的结果。

3.2.2 基于视觉分析的数据挖掘

由人类视觉感知理论所支持，如视觉显著性 (visual saliency)、颜色敏感度、深度感知以及视觉系统的注意机制 [126]，可视化技术 [127, 128] 和人机交互工具能够帮助我们更好地了解数据 [129--132]。Scagnostic [129, 133] 能够从散点图中诊断异常，[134] 提出通过由基于图理论度量的统计特征进一步提升了检测性能。Net-Ray [56] 对大规模图的邻接矩阵和散点图进行可视化和挖掘，发现了一些有趣的模式。[135] 中由多个特征构造 focus-plot（特征对之间的图形）并提出了 LookOut 算法用于给出对异常多维行为的图示化解释结果。GraphVis [131] 是一种灵活的、基于 web 的网络实时可视化分析平台，其融合了交互式可视化和分析技术，用于揭示一些重要模式和决策支持等。

3.2.3 异常检测

针对异常检测，Varun 等 [4] 对异常检测做了系统性的全面总结：从异常的性质和特征、问题定义、应用场景等方面做了详细地分类说明；针对图中的异常也存在多种检测方法 [28]。群体性异常也在多种不同的应用中进行了分析，包括移动服务 [136]、社交网络 [137, 138] 和网络流量 [139] 等。在图异常检测中，[45, 54] 等工作通过图谱子空间的图表示发现其中的社区和可疑类簇。SpokEn [54] 考虑在有图的特征向量对构成的 EE-plot 中的“EigenSpokes ”模式，并将其泛化到异常检测中；Shah 等基于 SVD 分解结果的重构方式提出 fBox 方法以检测 Twitter 网络中可疑的链路行为 [61]; GetScoop [45] 通过局部搜索的方式来发现二部图

表 3.1 EagleMine 与相关方法的比较

Table 3.1 Comparison of EagleMine and relevant methods. ✓denotes `supported'.

	DensityPeaks [121]	*DBSCAN [16]*	*STING [120]*	*Watershed [122, 123]*	*SpokEn [54]*	*GetScoop [45]*	*Fraudar [69]*	***EagleMine***
parameter free					✓	✓	✓	✓
non-spherical cluster	✓	✓	✓	✓				✓
anomaly detection		✓	✓		✓	✓	✓	✓
summarization			✓					✓
linear in #nodes				✓			✓	✓

中的相关稠密子图。在最近的研究工作中，稠密子图和子块的检测用于识别异常模式和可疑行为 [69, 140, 141]。其中的一个代表性工作，Fraudar [69] 提出了一种最密子图检测的方法，在算法优化过程中融合了节点和连边的可疑性分数用以定义和检测欺诈行为。

表 3.1 总结了 EagleMine 和上述一些主要的相关方法的比较。本文提出的 EagleMine 是唯一一个满足所有指标的方法。

3.3 相关定义与总结模型

考虑一个由节点集合 V 和连边集合 E 构成的图 $\mathcal{G} = (V, E)$, 它可以是一个用来表示用户间的好友与关注关系的节点同质图，也可以是描述用户对餐馆的评分关系的二部图。

对于给定图 $\mathcal{G}$ 可能存在的无数特征中，应该选择哪些特征来刻画图的节点呢？在大规模图应用中，直观上我们希望选择的特征应该满足 **(a)** 能够快速计算；**(b)** 能够描述一些模式或大多数节点所遵循的规律 (一些异常节点除外)。这样，一些可能的有用特征包括节点度 degree (出度、入度)、参与的三角形数 # triangle、核值 coreness、谱向量以及 PageRank 中的对应 score 等，这些特征本质上都是节点在图中的不同重要度指标。在其他的场景中也可以抽取与时间相关的特征，如

总的时长、时间间隔等。

根据抽取到的关联特征，可以通过不同特征之间的组合构成子空间，将目标对象 (如图节点) 投影到该空间中以形成多维直方分布图，从而反映对象的分布聚集特征。在图 3.1 所示的直方图中，那些具有相似颜色（密度值）的单元格及其近邻在视觉直观判断上是属于同一组的，即形成了一个类簇（节点组）。在关系图场景中，具有同步性行为的对象在关联特征空间中表现出相似性的特点，在直方图中形成具有局部聚集性且人眼视觉容易辨别的微簇结构，它们与正常实体所构成的较大的类簇存在很大偏移和差别。

正如引言部分所述，对于给定的一个由某些特征空间构造的直方图，我们的目标是识别出与人视觉感知预期相一致性的节点组，并通过优化模型的拟合度 (Goodness-of-Fit) 实现对节点组中点分布的准确刻画，同时根据正常和异常行为的不同表现和分布性质，识别其中的微簇结构并进行可疑度评估，以用于检测某些可疑模式。

对于直方图 (热度图) $\mathcal{H}$，其维度表示为 F，非空单元格数为 $nnz(\mathcal{H})$；用 $\boldsymbol{b}$ 表示直方图 $\mathcal{H}$ 中的任一单元格，h 表示一个单元格内所包含点的数目。

在给定的特征空间中，为了更加简洁地总结所对应的直方图 $\mathcal{H}$，本文利用一些统计分布函数作为词汇来描述 $\mathcal{H}$ 中的节点组内点的分布密度和随机性特征，这些统计分布会包含一些对应的特征描述参数。因此，本文提出基于词汇的直方图总结模型，其包括如下几项，

- 可参数配置的词项: 统计分布 $\mathcal{Y}$ 以用于 $\mathcal{H}$ 中点的分布；
- 赋值指派变量: $\mathcal{S} = \{s_1, \cdots, s_C\}$ 用于指定各节点组（共 C 个）描述时所用词项的类型；
- 模型参数: $\Theta = \{\theta_1, \cdots, \theta_C\}$ 表示用分布词项描述每个节点组时的参数配置，如高斯分布中的均值和方差等；
- 离群点: $\mathcal{O}$ 表示 $\mathcal{H}$ 中那些不能为词项所刻画的单元格。

对于可配置的词汇 $\mathcal{Y}$，它可包含任意合适的或依赖于应用的不同类型分布，例如均匀分布、高斯分布、拉普拉斯分布以及指数分布等，或其他适用于待描述数据特征的分布函数。

需要注意的一点是，由于受图节点间连接的稀疏性和高维空间中分布函数的几何特性复杂度 [142, 143] 的约束，建议所选的特征维度（即直方图维度）$F \leq 5$。

3.4 算法与分析

本文的算法 EagleMine 是从仿生学的角度，受人类视觉和认知系统的如下机制和特征所启发。

> 特性 1. 人类视觉通常检测一些连通性的分量，尽管存在很大的外观变化，但它们能够被人眼快速地识别 [144, 145].

这启发我们确定并识别一个热度图中那些内部连通的紧密连接区域作为节点组。由于不同节点组之间的不相交性质，可指导检测算法中后续用于平滑的 refinement 步骤。

> 特性 2. 具有自顶向下识别和层次化分割的特点 [146]。通过将某些基本元素（例如单词、形状、视觉感知区域）组织成更高阶的组，人的认知系统能够生成和表示在认知和视觉空间域中更加复杂的层次性结构。

这一性质说明，算法应该将不同的连接节点组按层次化结构的方式进行组织和搜索。最近的一项研究工作 [147] 利用神经网络模型探究人眼视觉对重要度的不同感知能力，并将结果用于辅助数据可视化和图形化设计。

算法 Alg. 1 描述了 EagleMine 的整体框架。给定一个直方图 $\mathcal{H}$，EagleMine 算法首先由 WaterLevelTree 算法构建一个层次化树结构 $\mathcal{T}$ 用于组织在 $\mathcal{H}$ 中检测到的类簇结构（节点组），然后由 TreeExplore 算法对 $\mathcal{T}$ 进行搜索并计算 $\mathcal{H}$ 的最优的分布总结，并得到目标的总结参数，包括模型参数 Θ、对节点组的赋值指派变量 S 以及离群点 $\mathcal{O}$。后续的小节中对该算法的各步骤做详细说明。

算法 1 EagleMine Algorithm

Input: Histogram $\mathcal{H}$ of specific feature spaces.

Output: Summarization result $\{S, \Theta, \mathcal{O}\}$.

1: Build a hierarchical tree structure $\mathcal{T}$ of node groups for $\mathcal{H}$ (WaterLevelTree 2).

2: Explore $\mathcal{T}$ and search the optimall summarization for $\mathcal{H}$ (TreeExplore 3).

3.4.1 Water Level Tree 算法

将直方图 $\mathcal{H}$ 中那些具有正值且相连接的单元格 ($h > 0$) 所构成的连通区域视为岛屿，其他单元格视为水域。这样，可以想象通过水面升高的方式（不同阈值的筛选过程）逐渐淹没岛屿部分，对于某一特定的海拔高度 r，使得岛屿区域

算法 2 WaterLevelTree Algorithm

Input: Histogram $\mathcal{H}$.

Output: Water-Level tree $\mathcal{T}$.

1: $\mathcal{T} = \{$positive bins in $\mathcal{H}$ as root$\}$.

// Raw tree construction.

2: **for** $r = 0$ to $\log h_{max}$ by step ρ **do**

3: $\mathcal{H}^r$: assign $h \in \mathcal{H}$ to zero if $\log h < r$.

4: $\mathcal{H}^r = \mathcal{H}^r \circ \mathbf{E}$. ▷ binary opening to smooth

5: islands $\mathcal{A}^r = \{$jointed bin areas in $\mathcal{H}^r\}$.

6: link islands in $\mathcal{A}^r$ to its parent at level r_{prev} in $\mathcal{T}$.

7: **end for**

// Tree refinement steps.

8: *Contract* $\mathcal{T}$: iteratively remove each single-child island and link its children to its parent.

9: *Prune* $\mathcal{T}$: heuristically remove noise nodes.

10: *Expand* islands in $\mathcal{T}$.

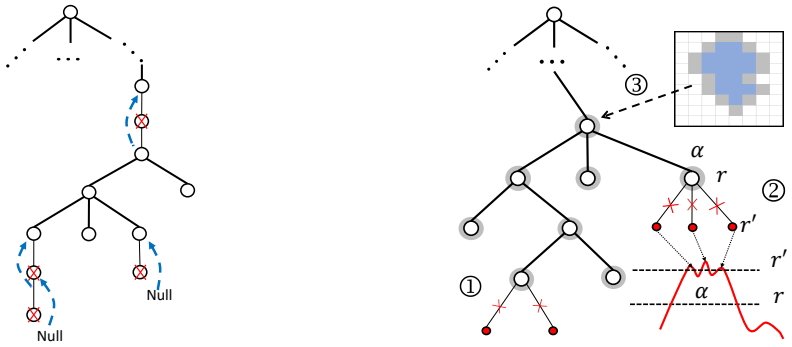
11: **return** $\mathcal{T}$

在 r 以下的单元格 $(h < r)$ 消失，即设置这些单元格的值 $h = 0$。在当前水面高度为 r 的条件下，由剩余的、具有正值的单元格构成了新的岛屿。

为了组织不同水面高度下所形成的岛屿，本文提出了一种 water-level tree 结构。树中每个节点表示一个岛屿，连边表示不同岛屿之间的隶属关系，即子节点所表示的高水位下的岛屿来自于位于更低水面高度的对应父节点岛屿。水位高度 r 从 0 开始升高的过程对应于在树中从根节点逐渐下移到叶子节点的过程。

在一个二维的直方图中，这些岛屿对应由特性 1 所产生的候选节点组，上述的淹没过程直观地反映了如特性 2 所描述的人眼在彩色的直方图中层次化地捕获这些不同目标的过程。例如，图 3.1 中按不同梯度变化的颜色变化对应与不同水位下的连通岛屿。

WaterLevelTree 算法对应于 Alg. 2。以整个直方图 $\mathcal{H}$ 作为 water-level tree 的根节点，水位 r 以特定的步长 ρ 从 0 开始升高到 $\log h_{\max}$，为了适应与大量数据观测中 h 服从类似幂率分布的特点，此处采用对数形式，其中 $h_{\max} = \max \mathcal{H}$；并



(a) 树的节点收缩 **(b) 树剪枝与节点扩展**

(a) Contract **(b) Tree prune and island expand**

图 3.3 EagleMine 算法中优化步骤图示说明。(a)、(b) 对应于 WaterLevelTree 中的 contract, prune 和 expand 的过程。

Figure 3.3 Refinement steps in the WaterLevelTree Algorithm. (a), (b) correspond to contract, prune, and expand for the WaterLevel tree respectively.

且在 Line 4 处利用图像处理技术中常用的形态学操作 opening (∘) [148] 对每个联通的岛屿进行平滑，并分离那些存在弱连接的岛屿：该步骤由一个特定的结构元素 **E** 对 $\mathcal{H}$ 进行扫描和过滤。然后，把在位于当前水位 r_{curr} 时检测到的每个岛屿连接到树中对应的位于低水位 r_{prev} 的父节点之后 (Line 6)。最后，当 r 达到最高可达高度 $\log h_{\max}$ 时，这个水位淹没的过程结束。这一过程如图 3.3a 所示。

此外，通过以下几个步骤对上一步得到的原始树 $\mathcal{T}$ 进行调整和优化，如图 3.3 所示。

Contract: 当前树 $\mathcal{T}$ 中包含许多节点只有单一的子节点（即单链连接）。这意味着水位在后续上升过程中不会从此节点处分裂出现新的岛屿，而仅仅是面积的缩小，所以它们是冗余部分。因此，此处对树进行深度优先搜索，一旦发现一个这样的节点，就将其移除并把它的后继节点连接到它的父节点处。

这个过程如图 3.3a 所示，其中的带箭头的蓝色点线显示了将单一子节点收缩连接到其父节点的过程，灰色的细连接线将会移除。

Prune: 剪枝的目的在于平滑由于某些岛屿顶部的相邻单元格内数值不稳定特点以及波动噪声带来的影响。根据总的面积大小来裁剪这样的子分支及其后继节点，即如果一个节点的所有子节点的单元格内数值总和与该节点的单元格数值之和的比率小于 95% 的话，就将其子节点及后继节点移除。

这个剪枝过程如图 3.3b 所示 (① 和 ②)。以节点 ② 为例，位于水面 r 的岛

屿 α 在其顶部包含一些扰动噪声。当水面达到 r' 时，这些噪声点会将 α 分割成三个小的岛屿，并连接到 $\mathcal{T}$ 中它们的父节点 α 上。因此，在剪枝过程中这些小的岛屿将会被移除。

Expand：由于水位 r 上升中选用的是对数形式下的统一步长 ρ，为了消除由此可能带来的影响（单元格损失）、避免学习分布参数时的过拟合现象，将树中每个节点进行扩展以包含其近邻范围内更多的单元格。这里所采用的启发式策略为：每次针对树中同一层的岛屿，以一个单元格的距离向外扩展以包含周围其他有正值的单元格，直到这些岛屿相邻接或者一个岛屿扩展的单元格数（面积）是原始岛屿的两倍大小。

$\mathcal{T}$ 中每个节点的扩展结果如图 3.3b 中的阴影圆圈所示。对于节点 ③，旁侧的不规则蓝色区域示意性地描述了原始的岛屿大小和形状，外围灰色的单元格表示在经过一步扩展后得到的结果，类似的扩展过程会持续进行直到达到上述的终止条件为止。

与之相对比，分水岭变换方法 (Watershed formalization) [122] 为了实现聚类目的，在 $\mathcal{H}$ 中的前景被定义为集水盆地 (catchment basins)，能够用于捕获在分割时类簇间的边界。在后续的实验部分 (第 3.5.2 小节以及图 3.2) 可以看到，由分水岭产生的分割效果近似对应于 $\mathcal{T}$ 中位于某一水位下同一层次的所有岛屿，此水位值由背景阈值所定义；因此，该方法不能完美地分离那些位于不同水位层次的岛屿，也不能处理多维数据的情况。在另一种基于网格的聚类算法 STING [120] 中，它以直方图单元格为索引构建了一种多分辨率的树结构以便于快速查询：在这种树形索引中，类簇可以通过以密度阈值 c 为参数的查询过程直接获得，但是这些得到的类簇也同样仅对应于 water-level tree 中位于同一层次的岛屿。此外，在我们的场景中并没有可以利用的先验知识来设定最优阈值参数，因此即使 DBSCAN 算法可以进行聚类，但是也存在参数无法确定的缺点。然而，EagleMine 不需要调节参数，其通过对 water-level tree 进行搜索发现具有最优组合的岛屿，它们可能来自不同的水位层次，能够有效地避免前述相关方法的缺点 (见第 3.4.3节)。

3.4.2 岛屿描述词汇

描述岛屿的词表词项 $\mathcal{Y}$ 可以包含任意合适的用户自定义分布。针对本文关心的特征空间，此处采用多元高斯作为其中一个词项。由于图节点的许多特征，如节点度、#triangle 等，通常是服从幂率分布的，导致靠近直方图边缘的这些

单元格会包含大量的点 (如图 3.1 的底部区域)，所以，截断的高斯分布 [149] 是一种描述此种截断椭圆类型的更好选择。同时，由于 $\mathcal{H}$ 中的单元格是离散化的，所以选用的分布词项也需要是被离散化后的结果，其表示每个单元格的概率函数而非连续的概率密度函数。因此，本文定义了如下形式的离散化多元截断高斯分布 DTM Gaussian (*discretized, truncated, multivariate* Gaussian distribution) 作为 $\mathcal{Y}$ 中一项，它的参数配置包括均值向量 $\boldsymbol{\mu}$ 和协方差矩阵 $\boldsymbol{\Sigma}$。

定义 3.1 (DTM Gaussian)**.** 根据单元格 $\boldsymbol{b}$ 的边界，$\boldsymbol{b}$ 中的概率函数定义为

$$P(\boldsymbol{b};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\beta}) = \int\cdots\int_{\boldsymbol{\beta}} \psi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\beta})d\boldsymbol{x}$$

其中，$\boldsymbol{x}$ 是一个维度为 F 的随机变量，$\boldsymbol{\beta} \in \mathbb{R}^{F\times 2}$ 表示截断分布的边界 (由每一维度的上下界所确定)，$\psi(\cdot)$ 是一个均值为 $\boldsymbol{\mu}$、协方差为 $\boldsymbol{\Sigma}$ 的截断标准高斯分布的概率密度函数。

在二维的直方图中，$\boldsymbol{\beta} = [[0,+\infty];[0,+\infty]]$。DTM Gaussian 是一种灵活的模型，能够刻画呈现不同形状分布的类簇，如线状、圆形、椭圆及上述形状的部分截断形状等。图 3.2a 中的红色和黄色标注的椭圆圈分别描述了具有 $1.5\cdot\Sigma$ 和 $3\cdot\Sigma$ 的 DTM Gaussian 的分布轮廓，代表节点类簇分布情况。

考虑到某些岛屿呈现多模 (multi-mode) 分布特点 (如图 3.1a 中所示的倾斜三角状岛屿)，因此，采用混合型 DTM Gaussian 分布作为另一种分布词项。在我们的数据分析中，这种三角形的岛屿包含了多数原图中节点且存在于很多不同特征构造的直方图中，例如，图 3.1a 表示的新浪微博消息转发应用中，用户在 out-degree 和 hubness 空间中分布，由于节点的出度特征的幂律分布性质使得分布密度随着垂直轴方向地递减；同时具有相似出度的用户也具有类似的 hubness 值，从而在水平轴方向形成了近似正态分布。因此，大部分的用户在该特征空间下聚集成三角状的岛屿形式，可以用混合 DTM Gaussian 予以描述。

一般来讲，为确定每个岛屿描述时所用的分布，可以用一些分布无关的统计假设检验的方法，如皮尔逊卡方检验等，或者其他与特定分布相关的方法来确定词项的指派变量 S。

词项指派确定后，可通过极大似然估计的方式来学习模型参数，其中对于某一岛屿 α，其对应的参数为 $\theta_\alpha = \{\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha, \tilde{N}_\alpha\}$，这里 $\tilde{N}_\alpha = \sum_{h\in\alpha}\log h$。为了方便表示，用 $DistributionFit(\alpha, s_\alpha)$ 函数表示参数学习过程，其返回结果为参数 θ_α。

算法 3 TreeExplore Algorithm

Input: WaterLevelTree $\mathcal{T}$

Output: summarization $\{\mathcal{S}, \Theta, \mathcal{O}\}$.

$\Theta = \varnothing$.
$\mathcal{S}$ = decide the distribution type s_α from vocabulary for each island in $\mathcal{T}$.
Queue $\boldsymbol{Q}$ = root node of $\mathcal{T}$.
while $\boldsymbol{Q} \neq \varnothing$ **do** ▷ breath-first search (BFS).
    $\alpha \leftarrow$ dequeue of $\boldsymbol{Q}$.
    $\theta_\alpha = DistributionFit(\alpha, s_\alpha)$ ▷ determine parameters
    Hypothesis test $\boldsymbol{H}_0$ = bins of island α come from distribution s_α.
    if $\boldsymbol{H}_0$ is rejected **then**
        enqueue all the children of α into $\boldsymbol{Q}$.
        $\mathcal{S} = \mathcal{S} \setminus \{s_\alpha\}$
    else
        $\Theta = \Theta + \{\theta_\alpha\}$
    end if
end while
// Post-processing.
Stitch and *replace* promising distributions in $\mathcal{S}$, then update Θ.
Decide outliers $\mathcal{O}$ deviating from the recognized groups.
return summarization $\{\mathcal{S}, \Theta, \mathcal{O}\}$.

3.4.3 Tree Explore 算法

根据上述调整后的层次化 water-level tree 结构及描述词典，之后利用由 Alg. 3 所描述的树搜索算法来确定最优的节点组以及其描述总结结果。在 TreeExplore 算法中，在确定每个岛屿（树中节点） α 所用的描述词项 s_α 之后，以广度优先搜索 BFS 方式遍历 $\mathcal{T}$，然后以某一准则选择最优的岛屿，最后通过粘合 (stitch) 过程来改善最终结果。图 3.4 展示了上述搜索过程和最终选择结果。

此处，通过一种启发式的方式来确定描述岛屿的分布词项的赋值指派，即在同一水位 r 得到的所有岛屿中，用混合 DTM Gaussian 描述的岛屿是：选择父节

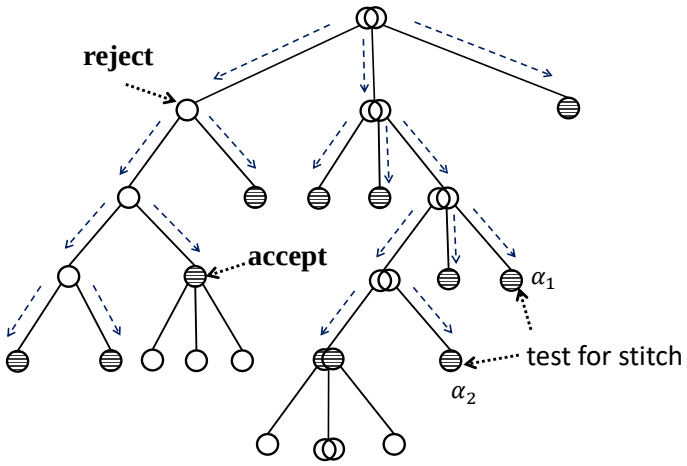


图 3.4 算法 TreeExplore 中的最优岛屿搜索及粘合后处理过程示意图。图中与连边相并行的带箭头的虚线表示了广度优先搜索过程，带阴影的圆圈表示通过假设检验后最终确定的最优描述岛屿候选集。

Figure 3.4 The Optimal islands search and the stitching post-process in TreeExplore Algorithm. The dashed lines with an arrow along with the edges of the tree denote the BFS search, the dashed circles denote the final optimal combination of node groups selected by the statistical hypothesis test.

点由混合 DTM Gaussian 所描述、且其所分裂产生的子节点中包含图中节点数最多的岛屿；其他岛屿则简单地用 DTM Gaussian 来描述。

3.4.3.1 广度优先搜索与选择标准

之后，在算法的 Line 4- 14 步中通过 BFS 遍历树 $\mathcal{T}$ 以搜索最优的类簇组合。从树的根节点开始，根据某一准则决定是否继续搜索某一节点的子节点及后继。

在统计学和机器学习中，AIC、BIC 等指标能够作为在正则框架下从有限模型集合中进行模型选择的准则，并挑选得分最低者作为最优模型 [150]。这样，可以利用这些准则来确定最优的目标岛屿，即如果某一节点的得分小于其所有子节点分值之和，则选择该节点且搜索过程终止，否则的话，会进一步对其子节点进行探索。

其他的另一类指标——统计假设检验，通过度量与理想零假设 (Null hypothesis) 的统计显著性来选择模型 [116, 151]，其中包括与分布无关的 Pearson's χ^2 检验、K-S 检验以及用于高斯模型的 Anderson-Darling 检验等。这里，搜索 $\mathcal{T}$ 中岛

屿 α 的子节点所对应的零假设为：

$$\boldsymbol{H}_0：岛屿\ \alpha\ 所包含的二单元格来自于分布\ s_\alpha.$$

如果 $\boldsymbol{H}_0$ 没有被拒绝，即停止进一步向下搜索；否则，将继续探索 α 的子节点。

由于直方图单元格内数值的奇异性，导致通过 Pearson's χ^2 检验、AIC 和 BIC 准则选择的结果并不稳定，此处采用 Anderson-Darling 统计检验。更明确来讲，根据岛屿中的单元格所构成的二值图像来对此岛屿进行检验，即关注于该岛屿的形状是否符合截断高斯或其混合模型所表示的分布特征。类似于投影追踪和 G-means 所用的方式，这里将单元格的数据分别投影到不同的维度并进行检验，根据其截断的特征实现了二次方类的上尾部 Anderson-Darling 统计检验[1] [152, 153]，并选择 1% 作为显著性水平。只有当各个维度上投影后的检验结果都为真时，才接受零假设；如果其中任意一个被拒绝，则 $\boldsymbol{H}_0$ 将会被拒绝。最终在 BFS 结束后，就能得到用于总结整个直方图的最优节点组候选集。图 3.4 中的带箭头虚线展示了上述搜索路径，带阴影的圆圈表示最终选择的最优节点组候选集。

3.4.3.2 粘合过程 (Stitch)

另外，由于某些来自于不同父节点的岛屿，其实际物理位置相互靠近 (如图 3.4 中的 α_1 和 α_2)，但是由于两者中间的弱联通扰动或噪声的影响而造成相互分离，这种情况下这些岛屿可能会被同一个分布所描述。因此，在算法第 15 步中利用了粘合过程，即根据假设检验的结果来确定是否将其合并，这个过程会一直进行直到最终没有任何变化产生为止。当同一时刻有多个岛屿组可以合并时，从中选择一组使得合并前后对数似然减少最小的岛屿组进行粘合，即

$$(\alpha_{i^*}, \alpha_{j^*}) = \arg\min_{i,j} \frac{\mathcal{L}_i + \mathcal{L}_j - \mathcal{L}_{ij}}{\#\text{points of } \alpha_i \text{ and } \alpha_j},$$

其中 α_{i^*} 和 α_{j^*} 是将被合并的两个岛屿，$\mathcal{L}_{(\cdot)}$ 表示用分布拟合一个岛屿所得到的对数似然值，$\mathcal{L}_{ij}$ 表示 α_{i^*} 和 α_{j^*} 粘合之后描述产生的对数似然值。

3.4.3.3 离群点与异常值分数

离群点是由那些与已识别节点组所描述的分布偏离较远的单元格所组成 (即属于该分布的概率小于 10^{-4})。

[1] 该检验用于度量左侧阶段的高斯分布的拟合度 GoF。

根据先验知识，直方图中包含绝大多数节点的主岛（major island）是属于正常部分；因此，可将一个岛屿与主岛的带权 KL-divergence 距离作为其异常分数。

定义 3.2 (Suspiciousness 异常分数)**.** 给定主岛的描述参数为 θ_m，由参数 θ_i 描述的岛屿 i 的异常分数定义为：

$$\kappa(\theta_i) = \log \bar{d}_i \cdot \sum_{\boldsymbol{b} \in \alpha_i} N_i \cdot \mathrm{KL}\left(P(\boldsymbol{b} \mid \theta_i) \,||\, P(\boldsymbol{b} \mid \theta_m) \right),$$

其中，$P(\boldsymbol{b} \mid \theta)$ 是单元格 $\boldsymbol{b}$ 在以 θ 参数的分布中的概率值，N_i 是岛屿 i 中所包含点的数目；并采用指数形式的 $\bar{d}_i$ 作为其权重，$\bar{d}_i$ 为岛屿 i 中所有点的平均度。

基于领域知识可知，其他特征相同的情况下，度越大的节点其可疑度越大。

3.4.4 算法复杂度分析

给定与图节点集合 V 的相关联特征，生成直方图 $\mathcal{H}$ 的时间复杂度为 $O(|V|)$。

令 C 表示 $\mathcal{H}$ 中类簇数目 ($C \ll nnz(\mathcal{H})$)，假定 Alg. 3 中由 $DistributionFit(\cdot)$ 函数通过梯度下降方式拟合参数时需要的迭代次数为 T，该项依赖于初始值与最优目标函数值之间的差异；$h_{max} = \max \mathcal{H}$，那么

定理 3.1. *算法 EagleMine 的时间复杂度为：* $O(\frac{\log h_{max}}{\rho} \cdot nnz(\mathcal{H}) + C \cdot T \cdot nnz(\mathcal{H}))$。

证明. 对算法 EagleMine，为了构建层次化树结构 $\mathcal{T}$，在 Line 3 处 WaterLevelTree 和 $\mathcal{H}$ 中的所有 $nnz(\mathcal{H})$ 个非空单元格进行比较，然后进行二值化的 opening 操作；在 Line 4 处通过检验非空单元格来移除那些小的噪声节点，这两步操作的复杂度均为 $O(nnz(\mathcal{H}))$；算法在 Line 5 - 6 将子节点 (岛屿) 连接至对应的父节点，这里岛屿的数目远小于 $nnz(\mathcal{H})$。所以，其时间复杂度等于构建的树 $\mathcal{T}$ 中的连接数目，即 $O(nnz(\mathcal{H}))$。这样 Line 2 - 7 的整个迭代需要 $O(\tau \cdot nnz(\mathcal{H}))$，其中 $\tau = \frac{\log h_{\max}}{\rho}$。最终得到的 $\mathcal{T}$ 是一棵高度为 τ 且宽度最多为 $nnz(\mathcal{H})$ 树，其中总的连接数小于 $\tau \cdot nnz(\mathcal{H})$。之后的优化调整过程中，树的收缩操作复杂度为 $O(\tau \cdot nnz(\mathcal{H}))$；在树的每一层中，每个岛屿内被总结的单元格数目小于 $nnz(\mathcal{H})$，树的剪枝和树节点的扩展复杂度也是 $O(\tau \cdot nnz(\mathcal{H}))$。所以，总体而言，构建 water-level tree 树 $\mathcal{T}$ 的复杂度为 $O(\tau \cdot nnz(\mathcal{H}))$。

在 Alg. 3 的最优节点组搜索过程中，分布拟合函数 $DistributionFit(\cdot)$ 的复杂度为 $O(T \cdot nnz(\mathcal{H}))$。由于算法在停止时找到 C 个类簇，所以在 $\mathcal{T}$ 中由 BFS 搜索访

问的节点构成的子树共包含 C 个叶节点；由于在 WaterLevelTree 算法中 Line 8 的收缩操作，所以子树中每个非叶节点至少有两个子节点，因此子树最多会包含 $2 \cdot C$ 个节点，这也意味着在算法 Line 4 - 14 的步骤中最多进行 $2 \cdot C$ 次来选择其中的最大岛屿、进行 $DistributionFit(\cdot)$ 以及应用统计检验。在每个树中节点 (岛屿) 上进行统计假设检验的复杂度与岛屿内包含的单元格数目线性相关，即小于 $nnz(\mathcal{H})$ 。在后处理的粘合过程中，算法仅检验那些在平面内相互邻近的岛屿，其复杂度小于在整个 $\mathcal{T}$ 进行的粘合尝试。

综上所述，算法 EagleMine 的时间复杂度为：

$$\begin{aligned} & O(\tau \cdot nnz(\mathcal{H}) + 2C \cdot (T \cdot nnz(\mathcal{H}) + nnz(\mathcal{H}))) \\ = \quad & O(\frac{\log h_{max}}{\rho} \cdot nnz(\mathcal{H}) + C \cdot T \cdot nnz(\mathcal{H})). \end{aligned}$$

■

3.5 实验验证与分析

本文设计完成不同的实验以回答如下的问题：

- *Q1* 基于视觉效果的定性分析：EagleMine 能够准确地识别到与人眼视觉判断相一致的类簇吗？
- *Q2* 定量分析：EagleMine 能否对直方图分布的总结方面带来显著提升？
- *Q3* 异常检测准确性：对比于其他已有的最新方法，EagleMine 在异常检测任务上效果表现如何？基于视觉启发信息的识别结果能带来多大的性能提升？
- *Q4* 应用可扩展：EagleMine 能够用于除图数据之外场景并发现其中的有趣模式吗？
- *Q5* 可扩展性：EagleMine 对数据大小是线性可扩展的吗？

3.5.1 实验设置

3.5.1.1 实验数据集

实验中所用的数据集统计信息如表 3.2 所示。其中的 Amazon [154]、Android [155]、BeerAdvocate [156] 和 Yelp [157] 数据集描述“用户-物品”之间的评分关系，其中的物品分别包括商品、手机应用程序、啤酒和食品；Flickr [158] 数据是关于用户与关系群组的隶属关系；Youtube [159] 描述用户间通过关注关系构建的同质图。Tagged 数据集收集自社交网站 Tagged.com，原始数据包含用户

表 3.2 真实数据集的信息总结

Table 3.2 Summary of the real world dataset.

数据集名称	图中节点数	图中边数	图表示关系类型
Amazon [154]	(2.14M, 1.23M)	5.84M	评分
Android [155]	(1.32M, 61.27K)	2.64M	评分
BeerAdvocate [156]	(33.37K, 65.91K)	1.57M	评分
Yelp [157]	(686K, 85.54K)	2.68M	评分
Flickr [158]	(2.30M, 2.30M)	33.14M	用户所属的群组
Tagged [112]	(2.73M, 4.65M)	150.8M	匿名的连接关系
Youtube [159]	(3.22M, 3.22M)	9.37M	用户间关注关系
Sina weibo	(2.75M, 8.08M)	50.1M	用户转发消息

间 7 种不同的匿名关系，实验中选用了边数最多、类型为 Type-6 的连接关系图。新浪微博数据集 Sina weibo 是从 weibo.com 网站中上爬取的 2013 年 11 月份的消息转发数据，构成“用户转发消息”的二部图结构。

3.5.1.2 实现

我们选用了许多聚类算法作为与 EagleMine 的对比方法，包括 X-means [115]、G-means [116]、DBSCAN [16]、STING [120] 以及 DensityPeaks [121] 等，这些算法的参数设置如下所列。对于 HDBSCAN [124] 聚类算法，复杂度为 $O(an^2)$，其中 a 是被描述物体的属性数目，n 是数据点的数目 (即图中节点数)，即 $\sum_{h\in\mathcal{H}} h^2$。故不适用于大规模图。

- X-means: 由 K-means 初始化并设置类簇为 5;
- G-means: 设置 $max\_depth = 5$，即约束最多划分的类簇数为 16 个，以避免太多类簇产生过度的分隔；设置 p-value = 0.01，实验结果对该参数不敏感;
- DBSCAN: 设置 Eps = 1, 采用单元格间的 manhattan 距离作为邻近距离度量；实验中 MinPts 参数是在直方图的单元格内点数的平均值到最大值之间以大小 50 为步长进行搜索，并手动地选择一组使得聚类效果与人眼视觉预期结果最一致的参数[3];

[2]现有作者提供的开源程序包 [160, 161] 及其他开源实现都不支持对于带权重类型数据的聚类分析。

[3]由于 DBSCAN 是手动调参得到的，因此不需要用 OPTICS 进行参数搜索。

- STING: 初始化 $c \approx \frac{\text{Minpts}+1}{\pi \text{Eps}^2}$，其中 MinPts 和 Eps 是 DBSCAN 聚类得到最优参数，之后通过微调提升可视化分析结果;
- DensityPeaks: 采用 Gaussian 核并根据决策图 (``decision graph") [121] 选择最优的类簇数目;
- EagleMine_DM: 以离散多元高斯 (DM Gaussian) 代替 DTM Gaussian 作为描述分布词汇实现的 EagleMine 算法。

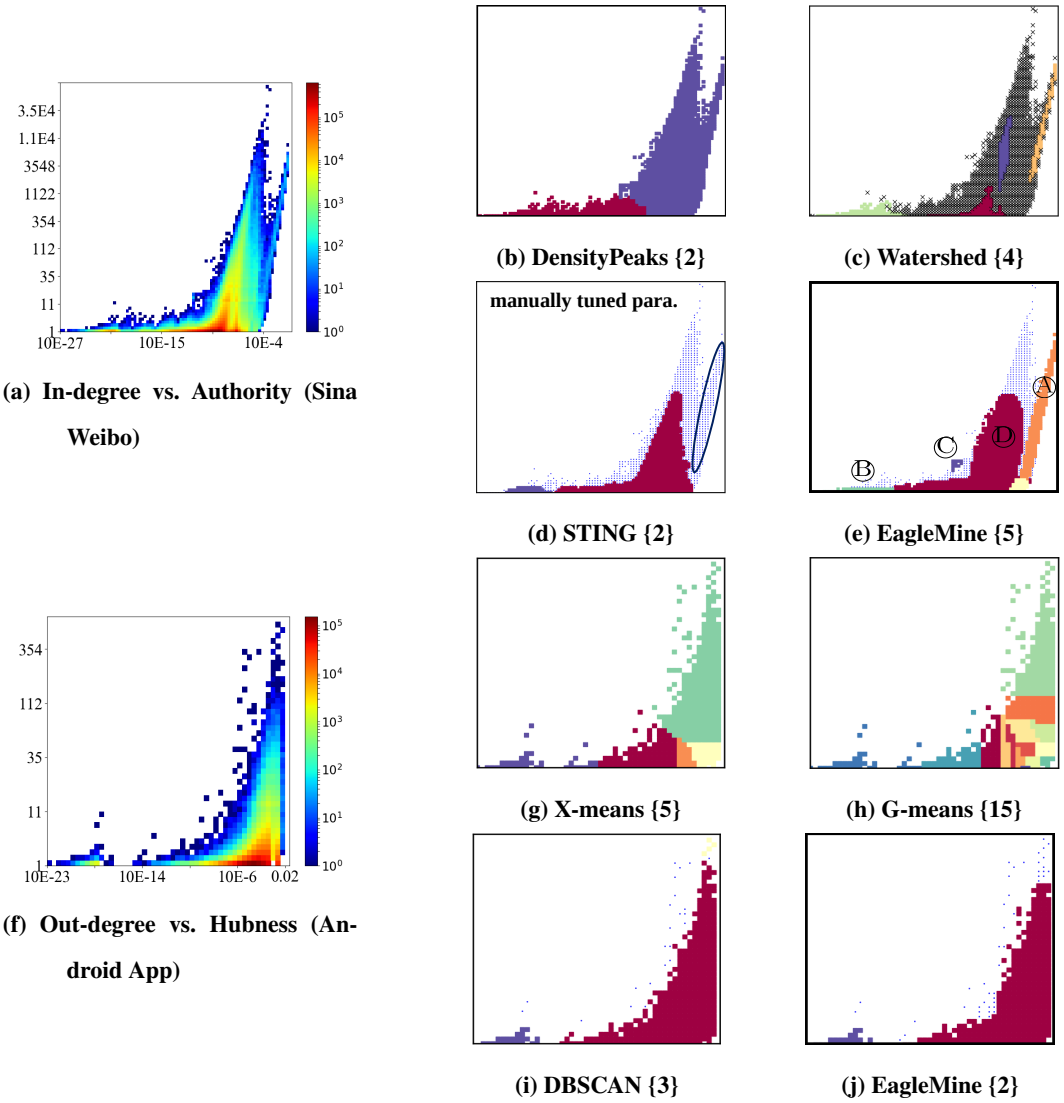
实验中选择的图节点特征空间如下：Tagged 数据集为 degree vs. PageRank 和 #triangle vs. degree 特征；其他数据集采用 in-degree vs. authority 以及 out-degree vs. hubness 特征。

针对不同的节点特征构造直方图，每种特征的带宽 (单元格大小) 可根据 plug-in 的方法或者核密度估计 [162] 的方式确定。在本文的实验设置中，采用了一种类似于 [46] 的启发式的规则，对于离散型特征，如 degree、#triangle 等，将其按对数形式将单元格划分成固定带宽；对于连续性特征，如图谱特征 (hubness 和 authority)，将其在对数形式下划分成与另一维特征数目相同的单元格。另外，实验中设置的水位升高步长统一为 $\rho = 0.2$（如果此数值太大可能会丢失某些微簇结果），在多个数据上验证表明当前设置是普遍有效的。

3.5.2 Q1. 定性实验分析

本小节中，基于可视化分析对二维直方图中类簇识别效果进行定性对比分析。由于空间限制，仅展示了在不同特征空间和不同对比方法的部分实验结果，其中对比方法包括 X-means、G-means、DensityPeaks、DBSCAN、STING 和 Watershed，特征空间为 out-degree vs. hubness、in-degree vs. authority 以及 #triangle vs. degree。

图 3.2d-3.2i 中显示了在“用户-转发-消息”的微博数据上的总结结果，采用特征为用户的 out-degree 和 hubness 值，其中 hubness 反映了用户所转发消息的重要度。对比方法中，DensityPeaks 不能很好地检测到类簇结果，原因在于其未能找到所有的微簇及离群点。通过移除直方图中作为背景的低密度单元格后，Watershed 算法可以把所有的节点组分成一个或两个大的节点组，因此通过手动调节用于过滤背景的阈值参数，来得到可视化效果最好的结果，最终的最优结果如图 3.2f所示，其中的背景为灰色区域，该算法识别到的节点组类似于 water-level tree 中处于同一层次的节点组；从中可以看到，Watershed 仅能识别一些高密度的节点组而不能正确地分离出所期望的其他微簇结构。EagleMine 则通过一种直观



(a) In-degree vs. Authority (Sina Weibo)

(b) DensityPeaks {2}

(c) Watershed {4}

(d) STING {2}

(e) EagleMine {5}

(f) Out-degree vs. Hubness (Android App)

(g) X-means {5}

(h) G-means {15}

(i) DBSCAN {3}

(j) EagleMine {2}

图 3.5 直观相比于其他对比方法，在不同的特种空间中 EagleMine 更好地识别的节点组。图中标注了检测结果对应的算法以及识别到的节点组数目。(a)-(e): 是新浪微博数据中对应于图 3.2 中用户的消息节点。(f)-(g): 是 Android 应用程序 Apps 的评分数据。

Figure 3.5 EagleMine visually recognizes better node groups than baselines in qualitative comparison for different feature spaces. The label plots reflect the node groups recognized by each algorithm. (a)-(e): Sina weibo data (msg. nodes), (f)-(j): Android rating data.

地方式识别到了被 DBSCAN 和 STING 所遗漏的部分，其中在遗漏区域①和③中有非常高比例的账户被删除，即被反作弊系统所识别并由系统管理员将其注销。另外，③和④微簇包含了一些出度大但 hubness 值小的用户，即他们转发了大量的不重要消息（如广告等）。因此，EagleMine 能够以一种与人眼视觉识别预期相一致的方式自动地识别一些非常有用的微簇结构。

另外，图 3.5 和 图 3.6 展示了四组不同数据集和不同特征空间的微簇识别的

可视化结果。原始的直方图在每组左侧第一列给出，之后分别是不同检测算法得到的带标注的聚类效果（算法名后标明了得到的类簇数）。离群点单元格有蓝色点或 `×' 所标识，结果中不同颜色代表不同算法检测到的不同类簇（节点组）。从图示结果中可以看到，G-means 和 X-means 过度分割了期望的类簇并产生了更多的节点组，并不符合视觉识别的预期结果；DensityPeaks 则没有能识别出很多类簇和离群点，从而识别效果不佳；对于 DBSCAN 和 STING，即使通过手动调参能够捕获每个直方图中主要的密集区域，但它们仍忽略了一些可疑的微簇，例如，如图 3.5e 所示的 Ⓐ 和 Ⓒ 部分。EagleMine 算法能准确地识别一些微簇结构，展示了其在识别节点组方面的优势，尤其是对微簇结构的识别。

3.5.3 Q2. 定量评价

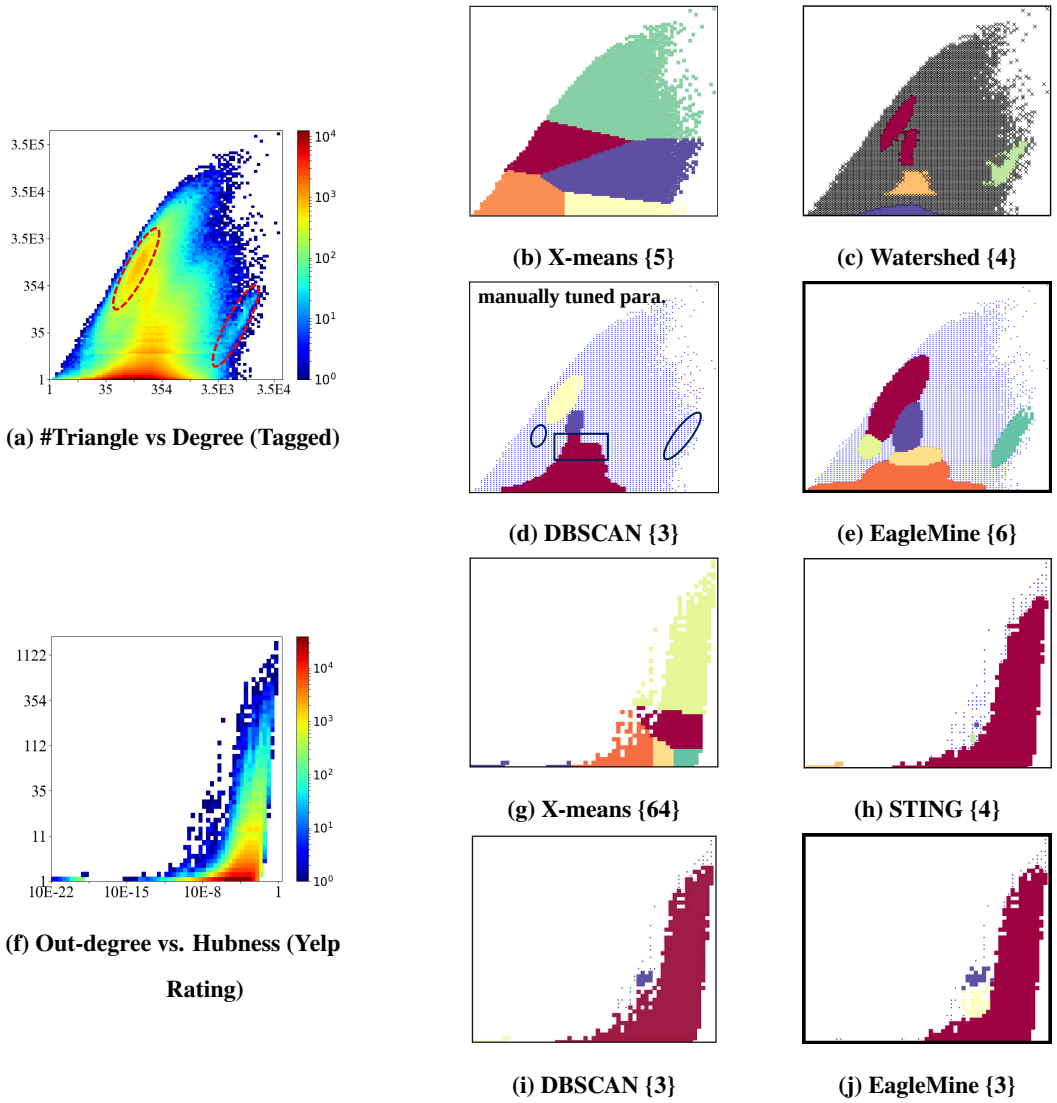
将聚类识别视为数据压缩问题，则可以利用最小描述长度准则 MDL 来度量算法的聚类总结性能（类似于 [119]）。简而言之，MDL 的假设在于如果模型能够更好地对数据进行压缩，就能够刻画和表示其中蕴含的真实模式，因此最优的模型的描述编码长度最短。基于 MDL 准则，EagleMine 的描述长度表示为：

$$L = \log^*(C) + L_S + L_\Theta + L_\mathcal{O} + L_\epsilon. \tag{3.1}$$

这里的模型描述包括如下的几项：

- 编码类簇数目需要 $\log^*(C)$ bits[4]；
- 编码 C 个节点组的分布词汇赋值指派 S 需要 $L_S = C \cdot \log(|\mathcal{Y}|)$ bits；
- 每个 DTM Gaussian 需要 $|\theta| = F + \frac{(1+F)F}{2} + 1$ 自由参数。对于两维特征 ($F = 2$)，则对应的二维分布 $|\theta| = 6$。因此参数的编码长度为 $|\theta| \cdot l_0$ bits，其中 l_0 为浮点数的编码长度，在我们的设置中采用 4×8 bit。因此，总的参数编码需要 $L_\Theta = C|\theta| \cdot l_0$ bits；
- 对于离群点 $\mathcal{O}$，共需要 $L_\mathcal{O}$ bits 来编码每一个单元格索引；
- 编码模型误差需要 L_ϵ bits。对于在岛屿 α_i 中的一个单元格 $\boldsymbol{b}$，按照模型所得到的期望点数为 $\tilde{h} = \left\lfloor 2^{\tilde{N}_i \cdot P(\boldsymbol{b}|\theta_i)} \right\rfloor$；因此，$\boldsymbol{b}$ 中原来的真实值能够准确地恢复为 $h = \tilde{h} + \epsilon$。这样，总的误差编码长度为 $L_\epsilon = \sum_{\boldsymbol{b}} (\log^*(h - \tilde{h}) + 1)$，其中 1 是对符号位的编码。

[4]这里，$\log^*$ 是对整数的通用编码长度，定义为 $\log^*(x) \approx \log_2(x) + \log_2 \log_2(x) + \cdots$，在求和式中只计算那些正值的项 [163]。



(a) #Triangle vs Degree (Tagged)

(b) X-means {5}

(c) Watershed {4}

(d) DBSCAN {3}

(e) EagleMine {6}

(f) Out-degree vs. Hubness (Yelp Rating)

(g) X-means {64}

(h) STING {4}

(i) DBSCAN {3}

(j) EagleMine {3}

图 3.6 EagleMine 可视化定性性能比较。(a)-(e): 对应于 Tagged 数据集的同质图 (#triangle vs. degree)，(f)-(g): 为 Yelp 评分数据集 (out degree vs. hubness)。

Figure 3.6 EagleMine visually recognizes better node groups in qualitative comparison. (a)-(e) : are for the homogeneous graph from Tagged website (# triangle vs. degree). (f)-(j) : use Yelp rating data (out degree vs. hubness).

对于其他的对比方法，基于相同的原则并按照类似 [119, 163, 164] 等的方式来计算对应的 MDL 编码长度。

对于二维直方图，图 3.7 展示了基于 MDL 指标的量化对比结果。从中可以看出 EagleMine 检测结果的描述编码长度最短，标志着该算法是一种更加简洁的数据总结方式。对比与 STING，DBSCAN，X-means 和 G-means 方法，分析各个数据集上的平均性能可以看出，EagleMine 缩短的编码长度比例分别超过 81.6%、79.0%、65.5% 和 20.2%；同时，对比于 EagleMine_DM 的结果，EagleMine 由于

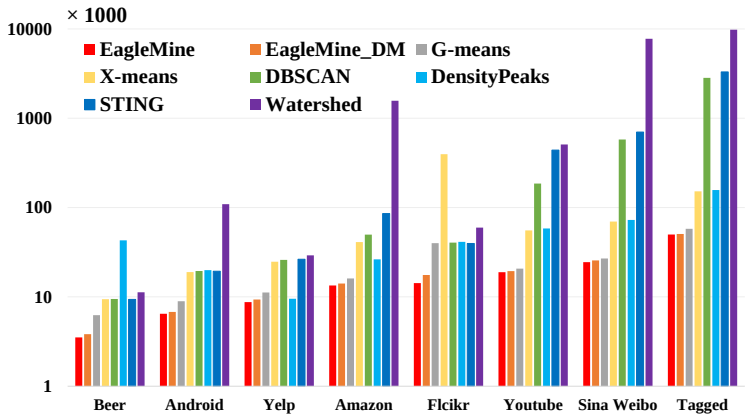


图 3.7 EagleMine 和其他聚类方法的量化性能比较。不同数据集上最小描述长度 MDL 指标的比较。EagleMine 产生的最短的描述编码长度，意味着能够实现对数据的更精简的描述总结，并且性能优于其他所有的对比方法。

Figure 3.7 Quantitative performance comparison for EagleMine and baselines. MDL is compared on different real-world datasets. EagleMine achieves the shortest description code length, which means concise summarization, and outperforms all other baselines.

选择了更加合适的描述词项而产生了 6.34% 的性能提升。所以 EagleMine 通过识别节点组能够以最优编码长度的方式对直方图进行总结描述。

表 3.3 EagleMine 对于多维特征空间的总结结果

Table 3.3 EagleMine's summary for multi-dimensional feature spaces.

所选特征	数据集	特征空间维度		
		3	4	5
out-degree + top-k hubness	Amazon	71,029	106,029	201,696
	Yelp	41.391	46,971	59,267

对于多维特征的情况，由在 Amazon 和 Yelp 数据集上提取的 3-、4-和 5-维特征构造不同直方图并应用 EagleMine 进行分析，这里的特征采用 out-degree vs. top-k hubness (图邻接矩阵 SVD 分解后的前 k 个的左奇异向量)。简单起见，此处仅采用 EagleMine_DM 检测算法，表 3.3 中列出了算法总结的 MDL 结果。因此，EagleMine 能够处理高维特征直方图且选择特征越多所产生的描述代价越大。

表 3.4 微博数据中检测到的微簇结构的可疑性排名结果

Table 3.4 Suspiciousness ranking of found micro-clusters in Sina Weibo dataset.

特征空间	异常得分 $\kappa(\cdot)$ 的排序结果
Out-degree vs Hubness (Fig. 3.2i)	①, ②, ③, ④, ⑥, ⑦, ⑧, ⑤
In-degree vs Authority (Fig. 3.5e)	Ⓐ, Ⓒ, Ⓑ, Ⓓ

3.5.4 Q3. 异常检测

为了说明 EagleMine 能够有效地检测图中的异常模式，本小节在合成数据和真实数据上与其他已有的最新检测算法进行了对比分析实验，对比算法包括 GetScoop [45]，SpokEn [54] 及 Fraudar [69]。

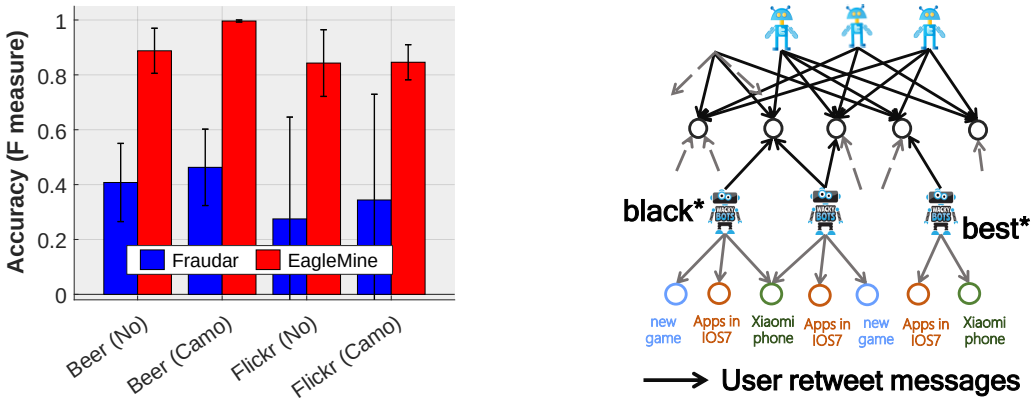
在合成测试数据中，由选择不同大小和不同密度的连接子图表示的欺诈攻击行为的真值进行注入检测分析。实验中考虑到更加狡猾的攻击策略，也对是否存在伪装型的欺诈情况进行了注入检测，其中增加的随机性伪装比例设置为 50%，即选择与攻击目标同样数量的对象进行伪装。选用了表 3.2 中所示的真实数据作为注入背景，BeerAdvocate 中注入块大小为 $1K \times 500$ 和 $2K \times 1K$，注入密度为 0.05，Flickr 中注入块的大小为 $2K \times 1K$ 和 $4K \times 2K$，注入块的密度有 0.05，0.1，0.2。以 F score 作为评价指标来度量对注入子图中节点的检测准确率，图 3.8a 中比较了各数据集上所有测试样例的实验结果平均指标的结果。其中，GetScoop 和 SpokEn 未能检测到任何的注入目标，所以其结果在图所省略。从结果中很明显的可以看出，EagleMine 的识别效果一致性地优于 Fraudar，并且不论是否存在伪装行为 EagleMine 检测结果的方差变化也更小。

在真实数据中的异常检测中，实验验证了 EagleMine 算法能够准确地发现新浪微博数据中的异常目标。通过手动标注一些用户和消息节点用以校验，这些节点来自于从对比方法的检测结果以及在 EagleMine 检测到的可疑微簇中采样得到的部分节点[5]。标注异常节点的规则依据类似于 [69]，即：

- 被微博平台所删除的用户账户或消息[6];
- 具有不寻常的公共账户名前缀的账户，以及其他可疑性信号：近似相同的

[5]实际操作中无法采集和标注全部数据。

[6]存在状态是在 3 年 (2017 年 5 月) 之后通过新浪微博提供的 API 接口进行的检验



(a) 检测注入欺诈行为的准确率 * (b) 异常的图示化展示

(a) Accuracy for detecting injected fraud (b) Graphical view of anomalies.

图 3.8 EagleMine 异常检测性能。(a) EagleMine 在 BeerAdvocate 和 Flickr 数据的注入欺诈行为检测中得到最优检测效果。* 其中 GetScoop 和 spokEn 未能检测出任何一注入块所以未显示其结果；(b) 在新浪微博数据中检测到的异常“Jellifish”子图连接结构。

Figure 3.8 EagleMine's performance for anomaly detection . (a) EagleMine achieves best accuracy for detecting injected fraud for BeerAdvocate ("Beer" as the abbr.) and Flickr data. *Note that GetScoop and SpokEn are omitted for failing to catch any injected object. (b) The anomalous "Jellyfish" anomaly pattern identified in Sina weibo.

注册时间、发帖数、朋友数和粉丝数等;

- 关于广告类、转发促销以及一些具有大量复制粘贴文本内容的微博消息；

因此，最终标记结果中共计 5,474 个异常账户，4,890 条可疑微博消息。

EagleMine 算法返回了各个微簇的异常分值，表 3.4 中列出了在 图 3.2i 和 图 3.5e 中 EagleMine 所识别的微簇结构，结果根据得分按降序排列。

采用 AUC 指标（ROC 曲线下方面积）来度量每个算法输出的有序检测结果的质量；从 EagleMine 微簇中采样的节点，按照 hubness 或 authority 值降序排列。图 3.2c 展示了各算法异常检测的性能对比。实验结果表明，EagleMine 对异常用户和消息的检测结果明显地一致优于其他对比方法，准确率提升分别达到 10% 和 50%。而且，FRAUDAR 和 SPOKEN 所检测到的异常用户分布在 图 3.2i 中的 ① 微簇中，这些基于图的检测算法仅关注于最密子图结构；而 EagleMine 则能够在整个特征空间中检测和评价所有的类簇，并发现其中显著微簇中的异常对象。因此，对比基线方法，EagleMine 能够得到更多、更精确的检测结果，例如，包括特征空间中的 ②, ③ 和 ④ 等微簇结构。

3.5.5 Q3. 实例研究与异常模式

如前所述，图 3.2i 所示的 ③、④微簇中包含的用户经常转发一些不重要的消息。另外，实验着重分析了位于主岛右侧的 ①、② 微簇内的用户行为模式，其中超过半数的用户已被系统管理员所删除，而剩余存在的用户登录名具有相同前缀模式，即如 图 3.2b 右侧列表所示的“best*”、“baby*”和“18-year-old*”等（列表中 182x:“best*”表示有 182 个账户名的前缀都是“best”）。

在实际应用场景中，算法发现了一些什么样的可疑模式呢？图 3.8b 示例性地展示了由图 3.2i 中 ① 和 ② 微簇内的可疑用户及他们转发的微博所形成的子图连接模式，由此构成了一种 `Jellyfish' 水母型结构。`Jellyfish' 的头部是由 ① 构成的稠密子图，这些用户与转发的消息之间形成非同寻常的密集连接，导致其具有较高的 hubness 值。这些可疑的机器账户以复制粘贴的方式疯狂地多次转发大量广告类消息——涉及的主题包括“新游戏”，“IOS7 应用”和“小米手机”等，这些消息构成 `Jellyfish' 的尾部。②中所表示的机器人账户采用了不同的攻击策略，使得其 hubness 值低于①中用户，他们被其他基于密度检测的方法所遗漏。

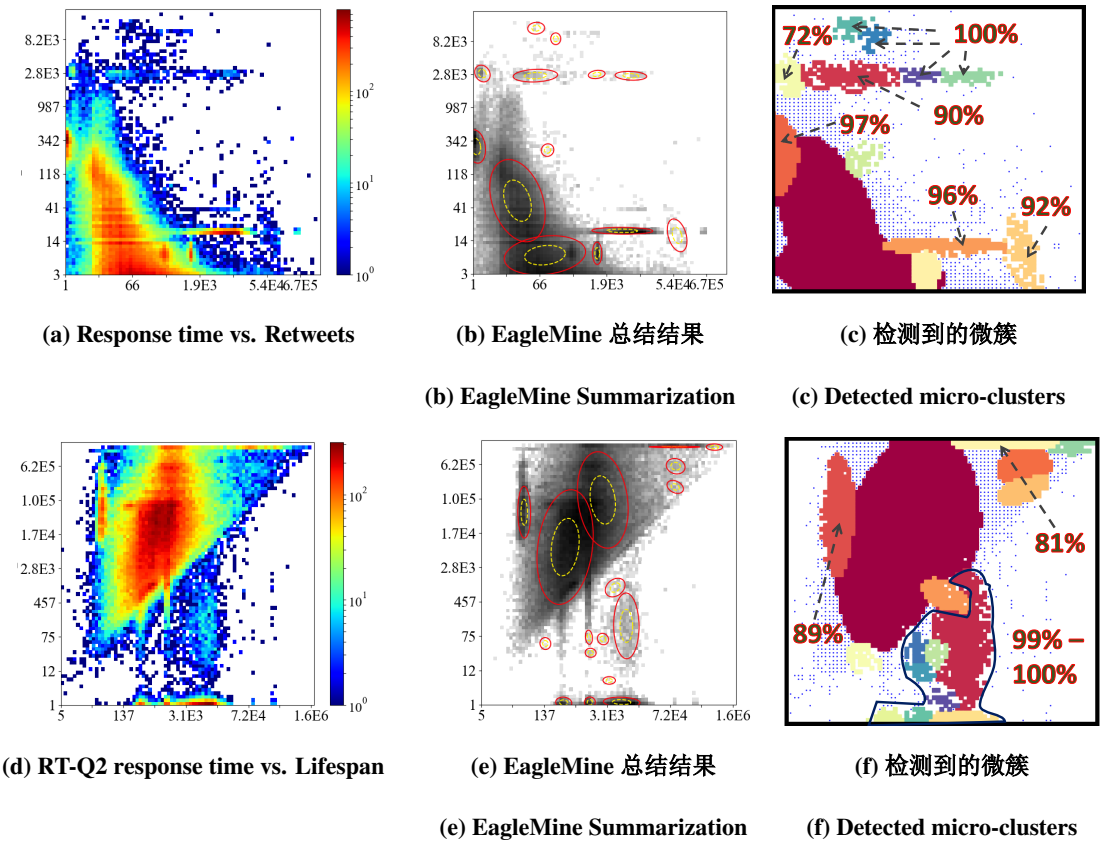
3.5.6 Q4. 应用扩展分析

本小节研究了 EagleMine 在图数据之外的不同场景下检测其他有趣模式的性能。

给定一组由许多用户参与的转发行为数据集，我们该如何检测集体性异常和同步行为呢 [46]？ND-SYNC [165] 公布了一组关于转发异常检测的 Twitter 数据，根据从转发线索序列中抽取的与时间相关特征来定义此类异常行为，由此反映了异常行为相比于正常用户的同步性特征。该数据集共包括 298 个用户以及他们发布的 134,022 条转发线索数据，其中正常用户行为数据有 83,587 条，其他为异常用户数据，每条线索都有人工验证的标签。

此处测试了 EagleMine 在由如下特征构造的直方图上的性能，

- Retweets: 转发数目；
- Response time: 从某一 tweet 的发布到第一条转发所经过的时间；
- Lifespan: 从 tweet 的第一次转发到观察期内最后一次转发所经过的时间；
- RT-Q2 response time: 从一条 tweet 的发布之后到其产生半数总转发次数转发所经过的时间。



(a) Response time vs. Retweets　(b) EagleMine 总结结果　(c) 检测到的微簇

(b) EagleMine Summarization　(c) Detected micro-clusters

(d) RT-Q2 response time vs. Lifespan　(e) EagleMine 总结结果　(f) 检测到的微簇

(e) EagleMine Summarization　(f) Detected micro-clusters

图 3.9 EagleMine 在转发行为的时序异常检测结果，转发线索数据被映射到不同的特征空间中。(a) 为 Response time vs. Retweets，(d) 为 RT-Q2 response time vs. Lifespan；(b)-(e) 显示了 EagleMine 以 DTM Gaussian 为词表词汇得到的总结结果；(c)-(f) 展示了检测到的微簇，其中的由箭头标注的大部分微簇都包含很高比例的异常结果。

Figure 3.9 EagleMine performance for temporal fraudsters retweet activity. The retweet threads are mapped to different feature space. (a) Response time vs. Retweets. (d) RT-Q2 response time vs. Lifespan. (b), (e) show the summarization of EagleMine with DTM Gaussian similar to Figure 3.2a and 3.2i show the detected cluster. Most of the micro-clusters contain a high percentage of fraudsters (marked with text).

图 3.9 展示了由 Response time vs. Retweets 和 RT-Q2 response time vs. Lifespan 不同特征子空间所构造的直方图结果，从中我们可以看到点分布的多样性和不规则特征，并存在大量微簇结构和散点（离群点）分布。从检测结果可以看出，EagleMine 仍能够从复杂分布中准确地识别出大多数类型各异的微簇结构，并给出直观的总结结果。这些检测到的微簇反映了在不同特征空间下转发异常用户的行为模式以及他们的同步时序行为。如图 3.9c和图 3.9f所示，大多数离主岛较远的节点组内包含了更大比例的异常用户，在图中标注了不同微簇中包含异常

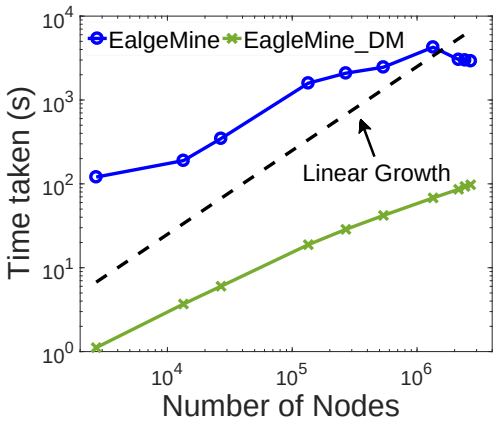


图 3.10 EagleMine 是线性可扩展的。在双对数坐标下蓝色曲线显示了随着图中节点数的增加 EagleMine 和 EagleMine_DM 的运行时间的线性变化情况。

Figure 3.10 EagleMine is scalable. EagleMine and EagleMine_DM scale (sub-) linearly with the number of nodes in graph (or the total value in histogram).

数据的比例；其中最少的一个包含 72%，其他很多微簇中有 99-100% 的点都属于异常。对照原直方图，图中显示的异常在各特征上的表现都较大差异，如具有响应时间较短的 (约 14sec)，也包含 300sec 或者 2800sec 等长间隔等特点。

因此，EagleMine 能够用于一般的直方图分析中，同时识别出其中的微簇结构，并挖掘实际应用中的其他有趣模式。

3.5.7 Q5. 可扩展性分析

图 3.10 中显示了 EagleMine 的运行时间与图的节点数呈近似线性的比例关系。此处选用新浪微博数据，从中根据转发不同时间段构成不同阶段的转发子图快照，以作为不同节点规模的图数据，其中选取的分别是在前 3 天，前 6 天，…，前 30 天内产生的转发关系。图中的黑色虚线表示线性增长趋势。

3.6 本章小结

在本章中，以基于特征表示的方式分析大规模关联数据的复杂连接以及聚集性群体异常在欧式空间中的表现和分布特征、识别在特征直方图中与图连接结构相对应的微簇，并应用于群体异常模式的检测。通过结合人类视觉识别和大脑感知的特性，本章提出了一种基于树的 EagleMine 算法来挖掘和总结与大规模图数据对应的直方图，并检测其中的微簇结构。EagleMine 算法根据 water-level tree 结构和统计假设检验找到最优的类簇，并由一个基于可配置的词典模型对类

簇特征进行描述，同时得到对直方图分布性质的总结。通过衡量微簇的异常度，算法可检测其中的一些包含群体异常的可疑微簇，在实际场景中对应团伙欺诈、密集社区以及同步性异常行为等模式。因此，EagleMine 算法有以下的性质：

- **自动地分布总结：**EagleMine 利用一些统计分布词汇对由关联特征构造的直方图进行自动总结，并以与人类视觉识别相一致的方式识别出其中的节点组。
- **有效性：**EagleMine 能够识别具有可解释性的节点组，同时对直方图有最优的总结效果。在定性可视化结果分析和定量评价实验比较中，其性能都优于其他对比方法。
- **异常检测：**EagleMine 能在真实大规模图中识别一些结构可解释的异常模式，对比于其他基于图的异常检测算法，EagleMine 在合成数据及微博数据上有更高的识别准确率。此外，EagleMine 也可检测图应用场景之外的某些可疑模式，如时序数据中的同步性异常行为。
- **具有可扩展性：**EagleMine 具有线性可扩展的时间复杂度，其复杂度与直方图中的数值总和（对应于图中节点数）成正比，也能处理多维特征空间构造的直方图。

此外，对于多维特征的直方图的可视化分析和总结仍然是一个挑战性问题，文中给出的变种 EagleMine_DM 算法以离散多元高斯作为总结词汇，为多维直方图分析提供了一种可行的方法，其具有可观的总结效果和运行时间（计算复杂度），与 EagleMine 有可比的性能；同时，有许多针对高维统计的具有理论保证的假设检验方法，可以用于高维分布分析和决定直方图中的最优类簇组合。

第 4 章　针对拓扑关联密集异常的统一图谱检测算法

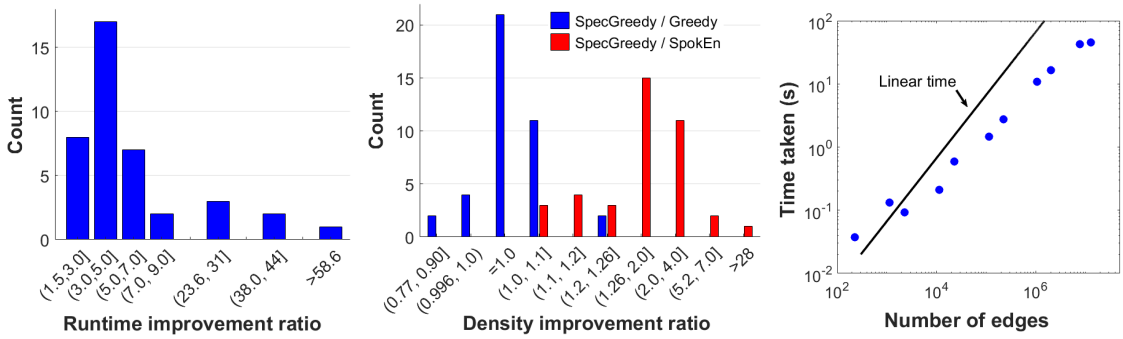
本章考察以稠密子图为对象的拓扑密集关联模式和对应聚集性群体异常行为的挖掘，探究关联数据中异常定义的问题多样性。针对大规模图数据在不同目标背景中的稠密子图检测问题，分析和对比了各类问题之间的区别与联系，提出了一种广义的统一形式化框架和谱理论优化分析，结合大规模图的谱分布性质与幂率特征设计了快速高效的检测算法，进一步提升现有不同检测方法的性能和效果，并用于其他有趣群体异常模式的检测。

我们应该如何有效地识别在线服务平台中的虚假评论或具有欺诈性的连接呢？如何基于用户的交互关系识别其中突然出现的社区或协同性的群体行为模式呢？在大规模图中我们该如何高效地找到其中的最小割集呢？所有上述的问题都与稠密子图检测问题密切相关，它是图数据分析中的一类重要的基本问题，在许多不同的领域有广泛应用。

本章中关注于真实的大规模图数据中稠密子图问题的统一形式化与检测问题，从理论分析的角度比较和对比了许多与之密切相关的问题；并提出了一种用于最密子图检测的统一形式化框架 (GenDS)，利用图的谱性质以及贪心优化策略设计了一种简单、高效计算的 SpecGreedy 算法来解决此广义问题。在来自不同领域的 40 个真实图数据 (其中最大网络的边数达到 14.7 亿条) 上大量实验分析和验证的结果表明，在最密子图检测中，对比与其他基线方法，SpecGreedy 算法可将检测速度提高 58.6× 倍且得到子图密度更大或者近似相同的检测结果；而且，SpecGreedy 算法是随着图的大小 (节点数和边数) 线性可扩展的；并在实际应用中证明了算法的有效性，例如，在大规模的随时间变化的学术共同作者网络中发现了突现的合作模式 (形成较大的团结构)。

4.1　本章引言

我们如何在时序图或动态图中捕获随时间变化的对比度最大的群组或社区，比如在科研社区中突然出现的热门话题或者学术合作关系等？我们如何在大规模图中高效地确定其对应的最小割集？如何在欺诈检测任务中基于用户行为找到最可疑的用户或者找到针对有争议事件 (民意调查、堕胎等) 找到具有一致性



(a) 最密子图检测加速比统计结果

(a) Speedup statistic for the densest subgraph detection.

(b) 最优的最密子图质量比较

(b) Optimal densest subgraph density quality comparison.

(c) SpecGreedy 的线性可扩展性

(c) The linear scalability of SpecGreedy.

图 4.1 本文提出的 SpecGreedy 算法是快速、有效和可扩展的。(a) 在得到具有相同或可比密度的最密子图结果时，本文算法在真实世界数据上比广泛使用的 Greedy 算法有可观的速度提升，最大提升比例达到 58.6× 倍；(b) 对比于 Greedy 和 SpokEn 算法，在最密子图检测任务上，SpecGreedy 得到密度更大或者近似可比的密度质量。算法在所有图上的检测结果一致性地优于 SpokEn 算法的结果，并能够得到密度超过 28× 倍的子图；与 Greedy 算法的结果相比，算法在多数图上的检测结果的密度相同或更密 (有超过 1.26× 倍)，其中有 4 个图密度非常接近 (≥ 0.996×)，另外 2 个图的密度提升小于 0.9；(c) SpecGreedy 算法的运行时间与图的大小 (图的边数) 呈线性关系。

Figure 4.1 Proposed algorithm SpecGreedy is fast, effective, and scalable. (a) Our proposed method detects the densest subgraphs with the same or comparable density quality up to 58.6× faster than the widely-used Greedy algorithm for various real-world datasets. (b) SpecGreedy has better or comparable density quality compared with Greedy and SpokEn algorithm in the densest subgraph detection. It consistently outperforms SpokEn for all graphs and finds up to 28× denser subgraph; it obtains the same or denser (more than 1.26×) optimal density for most graphs compared with Greedy, and 4 graphs with very close densities (≥ 0.996×) and only 2 graphs with less than 0.9 density improvement. (c) The time taken of SpecGreedy grows linearly with the size of graph (# of edges).

观点的最大团体呢？上述这些看似差别很大的实际问题都与最密子图的检测任务紧密相关。

图中稠密模式的挖掘是一项关键的基础任务，可以用来抽取有用信息并捕捉大量关系数据中的潜在原理等，在很多不同的场景中被广泛应用 [39]，包括发现生物组织中的功能团 [166]，人类行为数据中的迁移模式 [167]、社交网络中的社区结构 [168] 以及金融网络中的欺诈犯罪等 [28]。其中的最密子图问题在实践

中引起了大家极大的兴趣，原因在于该问题是能在多项式时间内精确求解或者在几乎线性时间内近似求解。最密子图检测的目标是在给定图中找到使得特定密度定义最大化的子图结构，根据应用场景的特征和问题约束，对输入数据和密度表示有不同的假设。Goldberg 等提出的最大流算法 [169] 和 Charikar 提出的基于线性规划的算法能够给出精确结果，同时 Charikar 证明了依据 [170] 中给出的贪心删除策略衍生的简单贪心算法能够保证得到 $\frac{1}{2}$ 近似最优的近似解，且算法复杂度与图的大小呈线性关系。在不同的问题设定中，贪心剥离策略及类似方法因其优良的性能得到广泛应用。然而由于未考虑真实数据的性质，这些算法在处理现代科学应用中出现的大规模图数据时仍然面临计算代价过高的问题。

据我们所知，目前还没有相关工作梳理和总结前述不同实际问题。因此，本章分析和对比了一些知名的相关问题之间的区别与联系，包括检测具有稀疏割集的社区结构、可疑性最密子图等，并提出了一种统一的形式定义广义最密子图检测问题（GenDS），它可以囊括多种不同的应用问题，这种统一表示能够从形式上清晰、明确地突出不同问题的差异，同时有助于设计一致性的解决方法；借助于图谱性质和贪心剥离优化策略，本文提出了一种高效的可扩展算法 SpecGreedy 来求解该统一形式化的问题。在大小不同的 40 个真实世界网络上的广泛实验验证结果证明，SpecGreedy 算法是快速、高效和线性可扩展的（见图 4.1）；在最密子图检测问题中，SpecGreedy 将检测速度提升 58.6× 倍并得到密度更大或近似相同的稠密子图，能够有效地处理大图数据（如有 14.7 亿条连边的网络）。此外，在 DBLP 的论文共同作者网络中算法检测到一些有趣的对比稠密子图模式。

本章的主要贡献在于：

- **理论和对应总结：**提出了广义的最密子图检测问题（GenDS），以统一的形式表示了许多相关问题，并总结了不同实际问题之间的对应关系，同时根据图谱理论分析了问题的优化性质。
- **算法：**发明了一种快速、高效和可线性扩展的 SpecGreedy 算法以求解统一的 GenDS 问题。
- **实验：**在多个不同领域的真实世界的图数据上进行了详尽的实验分析，用于验证 SpecGreedy 算法的有效性和线性可扩展性。此外，算法在学术合作数据中找到了一些规模很大的对比性稠密子图模式。

可复现性：本文提供了开源的程序 (https://github.com/wenchieh/specgreedy)

和在线可获得的真实数据。

本章后续内容中，先介绍了一些相关的研究工作和进展，随后给出了广义最密子图检测 GenDS 的形式化定义、总结了与其他相关问题的对应关系，在给出问题优化与图谱性质的理论分析之后，描述了所提出的 SpecGreedy 高效检测算法及复杂度分析，之后在大量真实图数据上验证了算法的检测性能、分析了群体异常检测任务中的结果，最后对本章的工作进行总结和概括。

4.2 相关工作

本小节总结了与最密子图问题相关的研究工作以及不同应用中与稠密子图模式检测相关的方法和进展。

4.2.1 最密子图检测

从大规模输入图中找到最密子图是一个已经被广泛研究的问题 [39]。广义地讲，这一问题的目标在于找到给定输入图中的一个节点子集以最大化某一密度指标。通常所指的*最密子图问题 (DSP)* 是找到一个子图最大化度密度（degree density），即子图中节点的平均度（= 边数 / 节点集大小或者 边权重之和 / 节点集大小）。当连边权重为非负值时，最密子图可以通过最大流算法 [169] 在多项式时间内准确地优化得到。然而，尽管在近些年的研究中有相关理论的进展，但通过最大流的方式获得准确解仍需要高昂的计算代价，因此这种方法并不适用于大规模图。Charikar [91] 引入了该问题的线性规划的形式化表示，并且证明了由 Asashiro 等 [170] 提出的贪心算法能够在线性时间内得到满足 $\frac{1}{2}$ 最优密度的近似结果。Rossi 等 [171] 提出了一种针对稀疏图进行快速并行的最大团检测的算法。Mitzenmacher 等 [172] 利用一种对输入图的采样方法实现最密子图的稀疏化，并在得到的稀疏图上计算最密子图。[173] 中提出了一种用于社区检测的优化模型作为最密子图检测问题的扩展。在最近的一项研究中，[174] 提出了 Greedy++ 算法在 Charikar 的贪心剥离方法的基础上借助凸优化中迭代方法的启发进一步提高了检测输出的子图的质量。然而，当边权可能为负值时，上面的问题成为 **NP** 难的 [175]。当对输出子图的大小有指定的约束时，对应的大小为 k 的最密子图检测问题 (DkS) 是 **NP** 完全的 [176, 177]，因此，在合理的复杂度假设下不存在多项式可解的近似算法 (PTAS)。

4.2.2 稠密子图模式

另一个相关的研究方向是对比图模式的挖掘，其目标在于发现两个图中具有显著差异的子图。Yang 等 [178] 提出了密度对比子图的检测问题，其等价于在一个“差异”(difference) 图上挖掘最密子图，并采用局部搜索的方式找到对应解。Tsourakakis 等 [175] 在一个具有较小负边权重的图中抽取具有 Risk-aversion 模式的稠密子图，并将贪心算法扩展到这种场景。对于符号网络，[179] 从中挖掘了对应“finding the gang in war”问题的稠密子图模式，通过求解一个二次优化问题检测其中的“大小为 k 的对立内聚团”(k-OCGs)。同时，稠密子图模式也被用作检测社区结构 [141, 166] 以及发现异常 [69] 等。其中，FRAUDAR [69] 提出了一种融合节点和连边可疑度的贪心优化方案用以解决带伪装的欺诈检测问题，SPOKEN [54] 利用由图的特征向量对构成的 EE-plot 中的“eigenspokes”模式来检测社区结构，并将其泛化到异常检测的应用中。

此外，也有很多研究工作利用图的谱性质来检测社区 [168] 和稠密子图 [180, 181]，或者用于图的分割任务 [182] 等。

4.3 问题形式化与对应关系

4.3.1 预设与定义

在本章中，向量用粗体小写字母表示（如 $\mathbf{x}$），矩阵用粗体大写字母表示（如 $\mathbf{A}$），集合用一般大写字母表示（如 S, V），运算符 $|\cdot|$ 表示集合的势或者一个向量中非零元的个数（nnz），$\|\cdot\|$ 表示一个向量的 l_2 模，且 $[x] \equiv \{1, \ldots, x\}$。表 4.1 中给出了本章中所用的符号列表。

考虑一个无向图 $\mathcal{G} = (V, E)$，其节点集大小为 $|V| = n$。令 S 表示节点集的子集，即 $S \subseteq V$, $E(S)$ 表示由 S 导出的子图 $\mathcal{G}(S)$ 的边集，即 $E(S) = \{e_{ij} : v_i, v_j \in S \wedge e_{ij} \in E\}$。非负矩阵 $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{n \times n}$ 为图 $\mathcal{G}$ 的邻接矩阵（$a_{ij} \geq 0$）。

给定一个与 S 相对应的指示向量 $\mathbf{x}$，作为最密子图检测问题中常用的密度度量指标，由 Charikar [91] 给出的子图 $\mathcal{G}(S)$ 的平均度密度可形式化表示为：

$$g(S) = \frac{E(S)}{|S|} = \frac{1}{2} \cdot \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}, \mathbf{x} \in \{0, 1\}^n, \tag{4.1}$$

同时约束 $|\mathbf{x}| \geq 1$ 以避免出现平凡解（空图）。在更一般的情况中，Hooi 等在 [69] 中提出在表示图的总质量时考虑子集 S 中节点权重值（由常量所定义）的贡献。此

表 4.1 符号表与定义

Table 4.1 Symbols and Definitions

符号	定义
$\mathcal{G}=(V,E)$	由节点集 V 和边集 E 构成的无向图
$\hat{\mathcal{G}}=(L\cup R,E)$	由节点集 L 和 R，边集 E 构成的二部图
$\mathcal{G}_r=(V,E_r)$	有节点集 V 和残差边集 E_r 构成的正残差图
$g(\cdot)$	密度度量指标
$\mathbf{x},\mathbf{y}$	节点子集的选择指示向量
$\mathbf{u},\mathbf{v}$	特征向量或奇异值向量
$\mathbf{A},\mathbf{L}$	图的邻接矩阵和拉普拉斯矩阵
$\mathbf{d},\mathbf{D}$	节点的度向量及度的对角矩阵，$d_i=\sum_j a_{ij}$
$\mathbf{I}$	大小为 $n\times n$ 的单位阵
$\mathbf{D_x}$	以向量 $\mathbf{x}$ 为对角元素的对角矩阵

时，$\mathcal{G}(S)$ 的密度表示为：

$$g(S)=\frac{|E(S)|+\sum_{i\in V}c_i}{|S|}=\frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{2\cdot\mathbf{x}^T\mathbf{x}}+\frac{\mathbf{x}^T\mathbf{D_c}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}=\frac{1}{2}\frac{\mathbf{x}^T(\mathbf{A}+2\mathbf{D_c})\mathbf{x}}{\mathbf{x}^T\mathbf{x}},\ \mathbf{x}\in\{0,1\}^n,\quad(4.2)$$

其中 $c_i\in\mathbb{R}$ 表示节点 i 的权重，$\mathbf{D_c}$ 是由权重向量 $\mathbf{c}=\left[c_1,\ldots,c_n\right]$ 定义的对角阵。

除了单个图中的最密子图模式之外，不同图之间的对比模式也是一类重要模式，即在两个具有相同节点集的图中存在的一个节点子集，其对应的两个导出子图包含具有显著差别的连边数或连边权重。该场景可能出现在动态图中相邻的两个快照中，或表示节点间具有相反观点的关系图中。

4.3.2 广义最密子图检测问题 GenDS

因此，本文提出了一种广义的最密子图检测问题 GenDS，它能够概括许多不同的实际问题，其具体形式化定义为：

表 4.2 问题 GenDS 的对应关系总结

Table 4.2 Summary for correspondence to problem GenDS.

	方法	matrix **P**	matrix **Q**	约束条件
1	MinQuotientCut [182]	$\mathbf{A} - \mathbf{D} = -\mathbf{L}$	$\mathbf{I}$	$\lvert\mathbf{x}\rvert < n$
2	Charikar [91]	$\mathbf{A}$	$\mathbf{I}$	$\lvert\mathbf{x}\rvert \geq 1$
3	FRAUDAR [69]	$\mathbf{A} + 2\,\mathbf{D_c}$	$\mathbf{I}$	
4	SPARSECUTDS [1] [173]	$\mathbf{A} - \frac{2\cdot\alpha}{2\alpha+1}\mathbf{D}$	$\mathbf{I}$	
5	TEMPDS [166]	$\mathbf{A}_t$	$\mathbf{A}_{t-1} + 2\,\mathbf{I} = \tilde{\mathbf{A}}_{t-1}$	
6	Risk-averse DS [175]	$\mathbf{A}^+ + \lambda_1 \mathbf{I} = \tilde{\mathbf{A}}^+$	$\mathbf{A}^- + \lambda_2\mathbf{I} = \tilde{\mathbf{A}}^-$	
	GENDS [2]	$\mathbf{A} + 2\,\mathbf{D_c}$	$\mathbf{A}' + \gamma\,\mathbf{I} = \tilde{\mathbf{A}}'$	

[1] [183] 中提出的对比子图模式等价于 $\alpha = 1$ 的情况；而 $\alpha = \frac{1}{2}$ 在 [184, 185] 中用于社区检测问题。

[2] 根据引理 4.8 中的形式化方法可以将二部图转换成一个无向图形式，应用于统一形式化表示中。

问题 1 (GenDS: 广义最密子图检测). **给定**图 $\mathcal{G} = (V, E)$ 及其对比图 $\mathcal{G}' = (V, E')$，且节点集 $|V| = n$；

找到一个最优的子集 $S^* \subseteq V$ 且 $|S^*| \geq 1$ **满足**如下条件

$$S^* = \underset{S \subseteq V, |S| \geq 1}{\arg\max}\ g(S; \mathbf{P}, \mathbf{Q}) = \underset{\mathbf{x} \in \{0,1\}^n, |\mathbf{x}| \geq 1}{\arg\max} \frac{\mathbf{x}^T \mathbf{P} \mathbf{x}}{\mathbf{x}^T \mathbf{Q} \mathbf{x}}, \tag{4.3}$$

其中，矩阵 $\mathbf{P}$ 和 $\mathbf{Q}$ 分别与 $\mathcal{G}$ 和 $\mathcal{G}'$ 相关，并定义为 $\mathbf{P} = \mathbf{A} + 2\mathbf{D_c}$ 且 $\mathbf{Q} = \mathbf{A}' + \gamma \mathbf{I}$。

这里，将矩阵 $\mathbf{Q} = \tilde{\mathbf{A}} = \mathbf{A}' + \gamma \mathbf{I}_n$ 定义为图 $\mathcal{G}'$ 的增广邻接矩阵，这样公式 (4.3) 的分母部分同时考虑了对应子图中的连边数和节点子集大小。特别的情况中，如果对比图 $\mathcal{G}'$ 是一个空图（empty graph），则 $\mathbf{Q}$ 退化为由 γ 缩放的单位阵 $\mathbf{I}$，即在 GenDS 中仅考虑子图的大小。另一方面，如果节点的权重都相同的话（即 $c_i = c > 0$），则矩阵 $\mathbf{P}$ 对应与图 $\mathcal{G}$ 的增广邻接矩阵。

正如定理 4.1 所述，GenDS 问题是一种更普遍的形式，许多与稠密子图相关的问题都是它的某一特例。

定理 4.1. 广义最密子图检测问题 GenDS 是一个通用的框架，可包括的问题涉及：*MinQuotientCut*、最密子图检测、稠密可疑子图检测的 Fraudar、时序密度子图 TempDS、一致稠密子图的 *Risk-averse DS* 等。

下面的几个推论给出了 GenDS 与许多常见问题的详细实例化说明。表 4.2 中对不同设置条件进行了总结，并给出了对齐的等式表示以强调和比较它们与 GenDS 之间的对应关系。

推论 4.2. *[MinQuotientCut]* 最优 quotient cut ratio 问题的目标在于将图分割为使割集最小化的两部分。令 $cut(S)$ 表示与子集 $S \subseteq V$ 对应的割集，即 $cut(S) = \{(u, v) \in E | u \in S, v \in V \setminus S\}$，则割集的大小可表示为：

$$|cut(S)| = \sum_{e_{ij} \in E} a_{ij} (\mathbf{x}_i - \mathbf{x}_j)^2 = \mathbf{x}^T (\mathbf{D} - \mathbf{A}) \mathbf{x} = \mathbf{x}^T \mathbf{L} \mathbf{x}.$$

当 $i \in S$ 时 $\mathbf{x}_i = 1$，否则 $\mathbf{x}_i = 0$。与 S 对应的 *cut ratio* 表示为 $\frac{|cut(S)|}{\min\{|S|, |V \setminus S|\}}$。不失一般性地，假设 S 是其中较小的集合（对比于其补集 $V \setminus S$），那么，可以通过最大化 $-\frac{\mathbf{x}^T \mathbf{L} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$ 得到最小的 *cut ratio*，这里该问题对应于 GenDS 中 $\mathbf{P} = -\mathbf{L}$ 且 $\mathbf{c} = -\frac{\mathbf{d}}{2}$、$\mathbf{Q} = \mathbf{I}$ 的情况，即令 $\mathbf{A}' = \mathbf{0}$ 且 $\gamma = 1$ [1]。

[1] 另一种设置中 $\mathbf{Q} = \mathbf{D}$，该设置等价于令 $\mathbf{P} = -\mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2}$、$\mathbf{Q} = \mathbf{I}$，即 $\mathbf{P}$ 为图 $\mathcal{G}$ 的归一化拉普拉斯矩阵。

推论 4.3. *[Charikar]* 在忽略常数因子后，由公式 (4.1) 形式化定义的标准最密子图检测问题对应于 GenDS 中 $\mathbf{P}=\mathbf{A}$、$\mathbf{Q}=\mathbf{I}$ 的设置。另外，*[186, 187]* 利用基于 *domain-set* 的优化方法探索由不同 γ 定义下的 $\mathbf{P}=\tilde{\mathbf{A}}$ 矩阵，分析最终的稠密子图的大小和密度之间的平衡关系。

推论 4.4. *[Fraudar]* 在最密可疑子图检测问题中，图中节点和连边的权重被视为其可疑度分值，即 c_u 和 a_{ij} 分别表示节点 u 和连边 e_{ij} 各自的可疑分数。正如公式 (4.2) 所示，其对应于 $\mathbf{P}=\mathbf{A}+2\mathbf{D_c}$、$\mathbf{Q}=\mathbf{I}$ 的设置，在忽略常数因子之后，其分子项 $\mathbf{x}^T\mathbf{P}\mathbf{x}$ 表示对应子图的可疑度总分值。

此外，也可以定义一个子图相对与图 $\mathcal{G}_0$ 的可疑度分值，其中 $\mathcal{G}_0$ 对应于一个参考图或由某种标准的图生成模型所产生。$\mathcal{G}_0$ 中对应子图的连边数（或分值）$\mathbf{x}^T\mathbf{A}_{\mathcal{G}_0}\mathbf{x}$ 可直接得到，而不需要进行矩阵-向量的乘法运算。例如，可以通过以 $\mathbf{x}^T\mathbf{x}$ 为参数的广义帕雷托分布（*Generalized Pareto*，*GP*）来估计任一子图中包含的边数 *[188]*，或者可以由一个随机图模型或偏好连接模型来计算其边数。

推论 4.5. *[SparseCutDS]* SparseCutDS 目标在于找到一个内部紧密连接且与图中剩余部分稀疏连接的社区结构，并通过最大化密度同时最小化平均割的大小来优化求解 *[173]*。根据推论 *4.2* 中 $|cut(S)|$ 的表示，SparseCutDS 的最大化目标可形式化为

$$g_\alpha(S)=\frac{|E(S)|-\alpha\cdot|cut(S)|}{|S|}=\frac{\mathbf{x}^T\left((\frac{1}{2}+\alpha)\mathbf{A}-\alpha\mathbf{D}\right)\mathbf{x}}{\mathbf{x}^T\mathbf{x}}=c\cdot\frac{\mathbf{x}^T\left(\mathbf{A}-\frac{2\alpha}{2\alpha+1}\mathbf{D}\right)\mathbf{x}}{\mathbf{x}^T\mathbf{x}},$$

其中，α 控制割集大小对应的权重，$c=\frac{1}{2}+\alpha$ 为常数因子。因此，GenDS 与之对应的设置为 $\mathbf{P}=\mathbf{A}+2\mathbf{D_c}$、$\mathbf{Q}=\mathbf{I}$，其中 $\mathbf{D_c}=-\frac{\alpha}{2\alpha+1}\mathbf{D}$。

推论 4.6. *[TempDS]* TempDS 问题检测由节点集 $S\subseteq V$ 定义的稠密子图，该子图在时间 t 突然出现或包含稠密连接，而在前一时间 $t-1$ 时则图中只有很少的连边 *[166]*。令 $\mathbf{A}_t$、$\mathbf{A}_{t-1}$ 分别表示动态图中相邻时间 t 和 $t-1$ 的图快照的邻接矩阵，这样，$\mathbf{x}^T\mathbf{A}_t\mathbf{x}$ 和 $\mathbf{x}^T\mathbf{A}_{t-1}\mathbf{x}$ 分别表示由指示向量 $\mathbf{x}$ 在不同时间导出的子图中所含连边数的两倍。再考虑子集 S 大小的贡献，TempDS 问题的目标可形式化为：

$$g(S)=\frac{\mathbf{x}^T\mathbf{A}_t\mathbf{x}}{\mathbf{x}^T(\mathbf{A}_{t-1}+2\mathbf{I})\mathbf{x}}=\frac{\mathbf{x}^T\mathbf{A}_t\mathbf{x}}{\mathbf{x}^T\tilde{\mathbf{A}}_{t-1}\mathbf{x}}.$$

推论 4.7. *[Risk-averse DS]* 给定一个图 $\mathcal{G}$，其邻接矩阵 $\mathbf{A}$ 中正值的元素 a_{ij} 代表连边 (u_i,u_j) 的期望回报 *(expected reward)*，负值元素代表的连边表示其风险 *(risk)*，

而绝对值 $|a_{ij}|$ 表示对应的强度。因此，$\mathbf{A}$ 可表示为 $\mathbf{A}=\mathbf{A}^{+}-\mathbf{A}^{-}$，其中 $\mathbf{A}^{+}$ 是由 $\mathbf{A}$ 中所有正值连边构成的回报网络，即其元素为 $\mathbf{A}^{+}_{i,j}=\max\left(a_{ij},0\right)$，而 $\mathbf{A}^{-}$ 为相反的风险网络，包含的元素为 $\mathbf{A}^{-}_{i,j}=|\min\left(a_{ij},0\right)|$。

Risk-averse 密度子图检测问题目标是找到一个具有较大正值权重的平均度且较小负值权重平均度的子图 *[175]*。在 GenDS 的框架表示中，对应于 $\mathbf{P}=\mathbf{A}^{+}+2\mathbf{D}_{\mathbf{c}}$ 且 $\mathbf{Q}=\mathbf{A}^{-}+\gamma_2\mathbf{I}$，其中 $\mathbf{c}=\frac{\gamma_1}{2}\mathbf{1}$，通过考虑节点集大小 $|S|$ 的贡献并由 $\gamma_1,\gamma_2\geq 0$ 控制子图的大小。

针对二部图 $\hat{\mathcal{G}}$ 中的最密子图检测，也可以通过下面的方式将其转化为等价的单部图形式，并规约到 GenDS 的框架中。

引理 4.8. 给定一个二部图 $\hat{\mathcal{G}}=(L\cup R,E)$ 且 $|L|+|R|=n$，图 $\hat{\mathcal{G}}$ 的最密二部子图的检测问题对应于如下的设置：令 $\mathbf{y}\in\{0,1\}^{|L|},\mathbf{z}\in\{0,1\}^{|R|}$ 表示两个指示向量，$\mathbf{x}=\left[\mathbf{y},\mathbf{z}\right]\in\{0,1\}^{n}$，且矩阵 $\mathbf{P},\mathbf{Q}\in\mathbb{R}^{n\times n}$ 满足如下定义：

$$\mathbf{P}=\begin{bmatrix}\mathbf{D}_{\mathbf{c}_L} & \frac{\mathbf{A}}{2}\\ \frac{\mathbf{A}^T}{2} & \mathbf{D}_{\mathbf{c}_R}\end{bmatrix}=\frac{1}{2}\begin{bmatrix}\mathbf{0} & \mathbf{A}\\ \mathbf{A}^T & \mathbf{0}\end{bmatrix}+\begin{bmatrix}\mathbf{D}_{\mathbf{c}_L} & \mathbf{0}\\ \mathbf{0} & \mathbf{D}_{\mathbf{c}_R}\end{bmatrix},\ \mathbf{Q}=\begin{bmatrix}\mathbf{I}_{|L|} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{|R|}\end{bmatrix} \tag{4.4}$$

其中，$\mathbf{c}_L$ 和 $\mathbf{c}_R$ 分别对应于 L 和 R 中节点的权重，$\mathbf{I}_{|L|}$ 和 $\mathbf{I}_{|R|}$ 表示大小为 $|L|\times|L|$ 和 $|R|\times|R|$ 的单位阵。

证明. 对于指示向量 $\mathbf{x}$，定义函数 $\delta(\mathbf{x})=\{i:\mathbf{x}_i=1,i\in\lceil n\rceil\}$；则被选择的节点子集可表示为 $S=\delta(\mathbf{x})=\delta(\mathbf{y})\cup\delta(\mathbf{z})$，且 $L_S=\delta(\mathbf{y})$，$R_S=\delta(\mathbf{z})$。因此，

$$\begin{aligned}\mathbf{x}^T\mathbf{P}\mathbf{x}&=\mathbf{y}^T\mathbf{D}_{\mathbf{c}_L}\mathbf{y}+\mathbf{z}^T\mathbf{D}_{\mathbf{c}_R}\mathbf{z}+\mathbf{y}^T\frac{\mathbf{A}}{2}\mathbf{z}+\mathbf{z}^T\frac{\mathbf{A}^T}{2}\mathbf{y}\\&=\mathbf{y}^T\mathbf{D}_{\mathbf{c}_L}\mathbf{y}+\mathbf{z}^T\mathbf{D}_{\mathbf{c}_R}\mathbf{z}+\mathbf{y}^T\mathbf{A}\mathbf{z}\\&=\sum_{i\in S}c_i\quad+\sum_{(i,j)\in E\wedge i\in L_S\wedge j\in R_S}a_{ij},\end{aligned}$$

其中的分母项 $\mathbf{x}^T\mathbf{Q}\mathbf{x}=\mathbf{y}^T\mathbf{I}_m\mathbf{y}+\mathbf{z}^T\mathbf{I}_n\mathbf{z}=|L_S|+|R_S|=|S|$。如果 $\mathbf{D}_{\mathbf{c}_L}=\mathbf{0}$ 且 $\mathbf{D}_{\mathbf{c}_R}=\mathbf{0}$，则 $\mathbf{x}^T\mathbf{P}\mathbf{x}=\mathbf{y}^T\mathbf{A}\mathbf{z}=|E(S)|$，即无向图 $\hat{\mathcal{G}}(S)$ 中包含的边数。 ■

对于带权无向图，为了避免由于单条边的权重太大而产生的一些平凡解，可以引入矩阵的列权重用以平衡，通过某一函数 h 来构造得到 $\hat{\mathbf{A}}=\mathbf{A}\cdot diag(\frac{1}{h(\mathbf{1}^T\mathbf{A})})$。例如，函数 h 的可能形式有 $h(x)=x^{\alpha}$（$\alpha\in\mathbb{R}^{+}$）或 $h(x)=\log(x+c)$（其中 c 是一个小的常量值以避免分母为零的情况）。此外，也可以利用基于某种 motif 定义的高阶图 [189] 以发现其中包含的复杂有趣的稠密模式。

4.4 理论与分析

在本小节中，给出了 GenDS 问题优化与图谱理论之间的关系，说明了通过真实世界图的一些偏态分布性质有助于对问题的近似分析和高效求解。

对于给定的图 $\mathcal{G}$ 及其对比图 $\mathcal{G}'$，此处构造了如下定义的“正残差”图 $\mathcal{G}_r = (V, E_r)$，其中的边集为 $E_r = \{(u,v)|(u,v) \in E \wedge (u,v) \notin E'\}$，其邻接矩阵表示为 $\mathbf{A}_r = (\mathbf{P}-\mathbf{Q})^+$。因此，在 $\mathcal{G}_r$ 图中的最密子图检测意味着其目标在于使得子图在 $\mathcal{G}$ 中的密度最大化同时在 $\mathcal{G}'$ 中的密度最小化，且只考虑正值元素的影响。这样，由公式 (4.3) 定义的目标函数可重新形式化为：

$$S^* = \arg\max_{|S|\geq 1} g(S;\mathbf{P},\mathbf{Q}) = \arg\max_{\mathbf{x}\in\{0,1\}^n,|\mathbf{x}|\geq 1} \frac{\mathbf{x}^T(\mathbf{P}-\mathbf{Q})^+\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \arg\max_{\mathbf{x}\in\{0,1\}^n,|\mathbf{x}|\geq 1} \frac{\mathbf{x}^T\mathbf{A}_r\mathbf{x}}{\mathbf{x}^T\mathbf{x}}. \quad (4.5)$$

由此，公式 (4.3) 中的矩阵 $\mathbf{Q}$ 被简化为单位阵 $\mathbf{I}$。在后续的优化分析中将利用这种形式的定义。

4.4.1 GenDS 问题优化与图谱分析

一个 $n\times n$ 对称矩阵的一个重要性质是：存在 n 个实数型特征值（不一定完全不同）并可以由对应特征向量构建的正交基将其对角化。考虑如下定义中 (4.6) 给出的，在实空间中以瑞利比值 （Rayleigh ratio） 形式表示的优化问题：

$$R(\mathbf{A}_r,\mathbf{x}) = \frac{\mathbf{x}^T\mathbf{A}_r\mathbf{x}}{\mathbf{x}^T\mathbf{x}}, \mathbf{x}\in\mathbb{R}^n, \mathbf{x}\neq\mathbf{0}. \quad (4.6)$$

其中，$\mathbf{A}_r \in \mathbb{R}^{n\times n}$ 且 $\mathbf{A}_r = \mathbf{A}_r^T$。而由公式 (4.5) 定义的 GenDS 优化目标是其二值变量的一种特殊情况。图论中的 Rayleigh−Ritz 定理 [182] 给出了由矩阵 $\mathbf{A}_r$ 特征值所确定的公式 (4.6)的最优性，也就是，

定理 4.9 (Rayleigh−Ritz 定理[2]). *假定 $\mathbf{A}_r$ 是一个 $n\times n$ 的对称阵，其特征值为 $\lambda_1 \geq \ldots \geq \lambda_n$，与之关联对应的特征向量为 $\mathbf{u}_1,\ldots,\mathbf{u}_n$，存在如下的性质：*

$$\begin{aligned} \lambda_1 &= \max_{||\mathbf{x}||_1=1} \mathbf{x}^T\mathbf{A}_r\mathbf{x} = \max_{\mathbf{x}\neq\mathbf{0}} R(\mathbf{A}_r,\mathbf{x}) \iff \mathbf{x}=\mathbf{u}_1, \\ \lambda_n &= \min_{||\mathbf{x}||_1=1} \mathbf{x}^T\mathbf{A}_r\mathbf{x} = \min_{\mathbf{x}\neq\mathbf{0}} R(\mathbf{A}_r,\mathbf{x}) \iff \mathbf{x}=\mathbf{u}_n. \end{aligned} \quad (4.7)$$

在一般情形下，对于 $1\leq k\leq n$，令 $\mathcal{S}_k$ 表示由 $\mathbf{u}_1,\ldots,\mathbf{u}_k$ 张成的空间并定义 $\mathcal{S}_0 = \mathbf{0}$，$\mathcal{S}_k^\perp$ 表示 $\mathcal{S}_k$ 的正交补空间，则

$$\lambda_k = \max_{||\mathbf{x}||_1=1, \mathbf{x}\in\mathcal{S}_{k-1}^\perp} \mathbf{x}^T\mathbf{A}_r\mathbf{x} = \max_{\mathbf{x}\neq\mathbf{0}, \mathbf{x}\in\mathcal{S}_{k-1}^\perp} R(\mathbf{A}_r,\mathbf{x}) \iff \mathbf{x}=\mathbf{u}_k. \quad (4.8)$$

[2]Rayleigh−Ritz 定理的证明参见文献 [182].

这意味着在补空间 $\mathcal{S}_k^{\perp}$ 中，$R(\mathbf{A}_r, \mathbf{x})$ 的最大值即为 λ_k。

Achiya Dax [190] 将矩阵的特征值和奇异值进行类比，并将上述定理推广到矩形矩阵，得到了与瑞利商矩阵类似的最优性质。这里为了避免矩阵的特征值中存在绝对值很大的负数的情况 [191]，将利用矩阵的奇异值和奇异向量来代替其特征值分解的结果。

矩阵 $\mathbf{A}_r$ 的奇异值分解表示为 $\mathbf{A}_r = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$，其中 $\mathbf{U} = \left[\mathbf{u}_1, \ldots, \mathbf{u}_r\right]$、$\mathbf{V} = \left[\mathbf{v}_1, \ldots, \mathbf{v}_r\right]$ 分别表示左、右奇异向量矩阵，对角阵 $\boldsymbol{\Sigma} = diag(\sigma_1, \ldots, \sigma_r)$ 对应于奇异值 $\sigma_1 \geq \cdots \geq \sigma_r > 0$。对于属于半正定对称阵的拉普拉斯矩阵 $\mathbf{L}$，可利用 Fiedler 向量（第二个特征（奇异）向量）求解推论 4.2 中图的最小割集问题；同时，还可以得到如下的引理。

引理 4.10. *对于 GenDS 问题中由公式 (4.3) 定义的最优解，可以表示为*

$$S^* = \underset{\mathbf{x}\in\{0,1\}^n, |\mathbf{x}|\geq 1}{\arg\max} \frac{\mathbf{x}^T\mathbf{A}_r\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \underset{|S|\geq 1}{\arg\max} \frac{1}{|S|} \sum_{i=1}^{n} \sigma_i \left(\sum_{j\in S} \mathbf{u}_{ij}\right)\left(\sum_{j\in S} \mathbf{v}_{ij}\right), \tag{4.9}$$

其中 $\mathbf{u}_{ij}$ $(\mathbf{v}_{ij})$ 表示奇异向量 $\mathbf{u}_i$ $(\mathbf{v}_i)$ 的第 j 个元素；对应的最优密度满足 $g_{opt} \leq \sigma_1$.

引理 4.11. *对于矩阵 $\mathbf{A}$ 中每一个与特征值 σ_i 对应的特征向量 $\mathbf{u}_i$，$\mathbf{u}_i$ 也是增广矩阵 $\tilde{\mathbf{A}} = \mathbf{A} + \gamma\mathbf{I}_n$ 对应于特征值 $\lambda_i + \gamma$ 的特征向量。*

当矩阵 $\mathbf{Q}$ 属于正定对称矩阵时，GenDS 问题是更为一般的情形，此时可以将公式 (4.3) 定义为广义瑞利商形式：

$$R(\mathbf{P}, \mathbf{Q}; \mathbf{x}) = \frac{\mathbf{x}^T\mathbf{P}\mathbf{x}}{\mathbf{x}^T\mathbf{Q}\mathbf{x}}, \ \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{0},$$

该问题的最大化规约为广义特征分解问题，同时与定理 4.9 类似的广义最小最大 定理保证广义特征值的最优性质。但考虑到真实图对应矩阵的奇异性，此处采用了由公式 (4.5) 表示的正残差图的问题定义形式。

另外，对于非对称矩阵 $\mathbf{A}_r \in \mathbb{R}^{m\times n}$，可以定义如下与之相关的二次优化问题

$$R(\mathbf{A}_r, \mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T\mathbf{A}_r\mathbf{y}}{\mathbf{x}^T\mathbf{x} + \mathbf{y}^T\mathbf{y}}, \ \mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n, \ \mathbf{x} \neq \mathbf{0}, \ \mathbf{y} \neq \mathbf{0}. \tag{4.10}$$

该问题的形式类似于前面公式 (4.6) 定义的瑞利商；本文同样证明了存在如下类似于定理 4.9 的最优性定理，这种形式有助于避免针对二部图构造一个大的表示矩阵 （$\mathbb{R}^{(m+n)\times(m+n)}$）。

定理 4.12 (Bi-graph Spectral)**.** 假定 $\mathbf{A}_r$ 为 $m \times n$ 的矩阵，对应的奇异值分解为 $\mathbf{A}_r = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$。对于任意的向量 $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{y} \in \mathbb{R}^n$ 存在如下的性质，

$$\sigma_1 = \max_{||\mathbf{x}||_1=||\mathbf{y}||_1=1} \mathbf{x}^T\mathbf{A}_r\mathbf{y} \geq \max_{\mathbf{x}\neq\mathbf{0},\mathbf{y}\neq\mathbf{0}} 2 \cdot R(\mathbf{A}_r, \mathbf{x}, \mathbf{y}) \iff \begin{matrix} \mathbf{x} = \mathbf{u}_1 \\ \mathbf{y} = \mathbf{v}_1 \end{matrix}, \tag{4.11}$$

其中，$\mathbf{u}_i \subseteq \mathbf{U}$、$\mathbf{v}_i \subseteq \mathbf{V}$ 分别表示与奇异值 σ_i 所对应的第 i 个左奇异向量和第 i 个右奇异向量。更为一般的情形，对于 $1 \leq k \leq r$，$\mathcal{S}_k^U$ 和 $\mathcal{S}_k^V$ 分别表示由 $\mathbf{u}_1, \ldots, \mathbf{u}_k$ 和 $\mathbf{v}_1, \ldots, \mathbf{v}_k$ 张成的空间，且 $\mathcal{S}_0^U = \mathbf{0}, \mathcal{S}_0^V = \mathbf{0}$，则

$$\sigma_k = \max_{\substack{||\mathbf{x}||_1=||\mathbf{y}||_1=1 \\ \mathbf{x}\perp\mathcal{S}_{k-1}^U, \mathbf{y}\perp\mathcal{S}_{k-1}^V}} \mathbf{x}^T\mathbf{A}_r\mathbf{y} \geq \max_{\substack{\mathbf{x}\neq\mathbf{0},\mathbf{y}\neq\mathbf{0} \\ \mathbf{x}\perp\mathcal{S}_{k-1}^U, \mathbf{y}\perp\mathcal{S}_{k-1}^V}} 2 \cdot R(\mathbf{A}_r, \mathbf{x}, \mathbf{y}) \iff \begin{matrix} \mathbf{x} = \mathbf{u}_k \\ \mathbf{y} = \mathbf{v}_k \end{matrix}.$$

证明. 给定一个非对称矩阵 $\mathbf{A}_r \in \mathbb{R}^{m\times n}$，对应的奇异值分解表示为 $\mathbf{A}_r = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \sum_{i=1}^K \sigma_i \mathbf{u}_i \mathbf{v}_i^T$，如果 $0 < i \neq j \leq K$ 则 $\mathbf{u}_i^T\mathbf{u}_j = \mathbf{v}_i^T\mathbf{v}_j = 0$，否则其值为 1。

给定的非零向量 $\mathbf{y} \in \mathbb{R}^m$ 和 $\mathbf{z} \in \mathbb{R}^n$，对于与 $\mathbf{A}_r$ 对应的二次优化问题为 $R(\mathbf{A}_r; \mathbf{y}, \mathbf{z}) = \frac{1}{2}\mathbf{y}^T\mathbf{A}_r\mathbf{z}$，下面仅考虑 $||\mathbf{y}||_1 = ||\mathbf{z}||_1 = 1$ 的情况。如果 $||\mathbf{y}||_1 \neq 1$ 或 / 和 $||\mathbf{z}||_1 \neq 1$，通过归一化后的向量 $\hat{\mathbf{y}} = \frac{\mathbf{y}}{\sqrt{\mathbf{y}^T\mathbf{y}}}$ 和 $\hat{\mathbf{z}} = \frac{\mathbf{z}}{\sqrt{\mathbf{z}^T\mathbf{z}}}$，存在如下的性质：

$$R(\mathbf{A}_r; \mathbf{y}, \mathbf{z}) = \frac{\mathbf{y}^T\mathbf{A}_r\mathbf{z}}{\mathbf{y}^T\mathbf{y} + \mathbf{z}^T\mathbf{z}} = \frac{\sqrt{\mathbf{y}^T\mathbf{y}}\sqrt{\mathbf{z}^T\mathbf{z}}}{\mathbf{y}^T\mathbf{y} + \mathbf{z}^T\mathbf{z}}\hat{\mathbf{y}}^T\mathbf{A}_r\hat{\mathbf{z}} \leq \frac{\hat{\mathbf{y}}^T\mathbf{A}_r\hat{\mathbf{z}}}{2} = R(\mathbf{A}_r; \hat{\mathbf{y}}, \hat{\mathbf{z}}).$$

令向量集 $\mathcal{U} = \{\mathbf{u}_i; i \in \lceil K \rceil\}$、$\mathcal{V} = \{\mathbf{v}_i; i \in \lceil K \rceil\}$；对于定理 4.12 通过如下的三种情况进行讨论。

情形 1. 如果 $\mathbf{y} = \mathbf{u}_i \in \mathcal{U}$, $\mathbf{z} = \mathbf{v}_j \in \mathcal{V}$，则

$$R(\mathbf{A}_r; \mathbf{y}, \mathbf{z}) = \begin{cases} \frac{\sigma_i}{2} & i = j, \\ 0 & i \neq j. \end{cases}$$

因此，$R(\mathbf{A}_r; \mathbf{y}, \mathbf{z}) \leq \frac{\sigma_1}{2}$，且当 $i = j = 1$ 是等式条件满足。

情形 2. 如果 $\mathbf{y} = \mathbf{u}_i \in \mathcal{U}$, $\mathbf{z} \notin \mathcal{V}$，则

$$R(\mathbf{A}_r; \mathbf{y}, \mathbf{z}) = \frac{1}{2}\sigma_i\mathbf{v}_i^T\mathbf{z} < \frac{1}{2}\sigma_i\mathbf{v}_i^T\mathbf{v}_i = \frac{\sigma_i}{2} \leq \frac{\sigma_1}{2}.$$

对另一种对称情形 $\mathbf{y} \notin \mathcal{U}$, $\mathbf{z} = \mathbf{v}_j \in \mathcal{V}$，可由类似的证明过程能得到相同的结论。

情形 3. 如果 $\mathbf{y} \notin \mathcal{U}$ 且 $\mathbf{z} \notin \mathcal{V}$，对应于 $\mathbb{R}^m$ 空间的基向量集合定义为 $B_m = \mathcal{U} \cup \tilde{\mathcal{U}}$，其中 $\tilde{\mathcal{U}} = \{\mathbf{u}_{K+1}, \ldots, \mathbf{u}_m\}$ 是由矩阵 $\mathbf{A}_r^T$ 的零空间 (Null space) 向量构成，

对于任意的 $i,j \in \lceil m \rceil$ 满足，如果 $i \neq j$ 则 $\mathbf{u}_i^T \mathbf{u}_j = 0$，否则其值为 1；类似地，$\mathbb{R}^n$ 空间对应的基向量集合为 $B_n = \mathcal{V} \cup \tilde{\mathcal{V}}$，其中 $\tilde{\mathcal{V}} = \{\mathbf{v}_{K+1}, \ldots, \mathbf{v}_n\}$ 是由矩阵 $\mathbf{A}_r$ 的零空间向量构成，对于任意的 $i,j \in \lceil n \rceil$ 满足，如果 $i \neq j$ 则 $\mathbf{v}_i^T \mathbf{v}_j = 0$，否则其值为 1。与零空间对应的矩阵奇异值均为 0。

利用基向量集合 B_m 和 B_n，向量 $\mathbf{y}$、$\mathbf{z}$ 可分别表示为 $\mathbf{y} = \sum_{j=1}^{m} s_j \mathbf{u}_j$、$\mathbf{z} = \sum_{l=1}^{n} t_l \mathbf{v}_l$，其中的系数 s_j 和 t_l 表示对应基下的坐标, 满足 $\sum_{j=1}^{m} s_j^2 = \sum_{l=1}^{n} t_l^2 = 1$。由此，存在如下的结论：

$$
\begin{aligned}
R(\mathbf{A}_r; \mathbf{y}, \mathbf{z}) &= \frac{1}{2} \sum_{i=1}^{K} \sigma_i (\mathbf{y}^T \mathbf{u}_i)(\mathbf{z}^T \mathbf{v}_i)^T \\
&= \frac{1}{2} \sum_{i=1}^{K} \sigma_i \left((\sum_{j=1}^{m} s_j \mathbf{u}_j^T) \mathbf{u}_i \right) \left((\sum_{l=1}^{n} t_l \mathbf{v}_l^T) \mathbf{v}_i \right)^T \\
&= \frac{1}{2} \sum_{i=1}^{K} \sigma_i \left(s_i + \sum_{j=K+1}^{m} s_j \mathbf{u}_j^T \mathbf{u}_i \right) \left(t_i + \sum_{l=K+1}^{n} t_l \mathbf{v}_l^T \mathbf{v}_i \right)^T \\
&= \frac{1}{2} \sum_{i=1}^{K} \sigma_i s_i t_i \leq \frac{1}{2} \sum_{i=1}^{K} \sigma_i |s_i| |t_i| \\
&\leq \frac{1}{2} \max_{i \in \lceil K \rceil} \sigma_i = \frac{\sigma_1}{2}.
\end{aligned}
$$

其中，最后一项不等式成立是根据性质

$$
\sum_{i=1}^{K} |s_i| |t_i| \leq \frac{1}{2} (\sum_{j=1}^{K} s_j^2 + \sum_{l=1}^{K} t_l^2) \leq 1.
$$

因此，综上所述，可以得到的结论为，对于任意的 $\mathbf{y} \in \mathbb{R}^m$、$\mathbf{z} \in \mathbb{R}^n$，满足 $R(\mathbf{A}_r, \mathbf{y}, \mathbf{z}) \leq \frac{\sigma_1}{2}$，当且仅当 $\mathbf{y} = \mathbf{u}_1$ 且 $\mathbf{z} = \mathbf{v}_1$ 时，不等式成立。上面的证明过程可以以类似的方式扩展到针对 σ_k 的一般情形。 ■

对于二部图 $\hat{\mathcal{G}} = (L \cup R, E)$，其邻接矩阵为 $\mathbf{A} \in \mathbb{R}^{|L| \times |R|}$，存在如下的性质：

引理 4.13. FRAUDAR 所定义的最密二部子图检测问题中，如果 $\mathbf{P} = diag(\left[\mathbf{A}/2, \mathbf{A}^T/2\right])$ 且 $\mathbf{x}^T \mathbf{P} \mathbf{x} = |E(S)|$，则对应的最优解可表示为

$$
\begin{aligned}
S^* &= \arg\max_{\substack{\mathbf{x} \in \{0,1\}^n \\ |\mathbf{x}| \geq 1}} \frac{\mathbf{x}^T \mathbf{P} \mathbf{x}}{\mathbf{x}^T \mathbf{x}} = \arg\max_{\substack{\mathbf{y} \in \{0,1\}^{|L|}, \mathbf{z} \in \{0,1\}^{|R|} \\ |\mathbf{y}| > 0, |\mathbf{z}| > 0}} R(\mathbf{A}_r, \mathbf{y}, \mathbf{z}) \\
&\leq \arg\max_{S = \delta(\mathbf{y}) \cup \delta(\mathbf{z}), |S| \geq 1} \frac{1}{|S|} \sum_{i=1}^{k} \sigma_i \left(\sum_{j \in \delta(\mathbf{y})} \mathbf{u}_{ij} \right) \left(\sum_{j \in \delta(\mathbf{z})} \mathbf{v}_{ij} \right),
\end{aligned}
\tag{4.12}
$$

其中 $\mathbf{u}_{ij}, \mathbf{v}_{ij}$ 分别为奇异向量 $\mathbf{u}_i$ 和 $\mathbf{v}_i$ 中的第 j 个元素；最优密度满足 $g_{opt} \leq \sigma_1$.

4.4.2 真实世界图的性质

真实世界的网络，包括万维网、社交网络、电子商务应用、在线评论和推荐系统等，具有许多不同的特征，连边稀疏性和幂律分布是其中的关键特性，利用这些重要性质可用于图分析中（时间和空间）的高效存储和计算以及对现实中网络的综合建模。许多研究 [27, 192] 表明许多真实世界的图存在很多满足统计显著的幂律分布的性质，包括节点度分布、“二部核”(对应社区) 的大小分布、邻接矩阵和拉普拉斯矩阵的特征值、奇异值的截断部分的分布等等。另外，图的邻接矩阵中与顶部的几个较大特征值相对应的特征向量元素——用于指示网络价值 (``network value") ——的分布是严重偏态的 [193]。

因此，在最密子图检测问题中，基于 GenDS 问题的图谱形式化、真实世界图的奇异值及奇异向量元素的分布的偏态性，仅需要考虑前几个大奇异值以及对应的特征向量中较大值的一些元素就可以有效地构建一些稠密子图的候选集，并以此进行最优结果的检测。本章将在下面的算法部分对此作详细的介绍。

4.5 算法与分析

本节中展示了本章提出的 SpecGreedy 算法用于解决广义的最密子图检测问题 GenDS，并给出了对其性质的分析。

此处首先回顾一下密切相关的 Charikar 贪心剥离算法，Alg. 4 展示了算法的整体框架。该算法以整个原始输入图作为起点，然后重复贪心地选择并删除一个节点——该节点移除后使得剩余节点子集对应的子图密度最大，从而产生一个大小逐渐收缩的节点集序列，最后输出其中使得密度指标 g 值最大的子图。对于具有相同度的节点，以随机选择的方式删除。算法保证最后返回的结果至少是理论最优密度的一半，即 $g^* \geq \frac{1}{2} g_{opt}$。如果给定的 $\mathcal{G}$ 是一个无向图，则算法中 Line 10 对应于选择当前子图 X_{t-1} 中度最小的节点进行删除。

在算法 Alg. 4 中，采用一种优先树的数据结构管理贪心删除过程中的图节点集，它是一棵二叉树且包含所有 $|V|$ 个图节点元素作为其最低层的叶子节点，其内部节点指示它的哪一个子节点具有最大的优先次序。该优先树结构能够保证查询节点 i^* (Line 8) 以及更新相关元素 (Line 9) 的复杂度是 $O(\log|V|)$，从 Line 8 到 Line 10 的代码片段共执行 $|V|$ 次。最终，整个过程需要对节点的优先次序更新 $|E|$ 次，每次针对一条边。因此，整个算法的复杂度是 $O(|E|\log|V|)$。

算法 4 GREEDY Algorithm

Input: Graph $\mathcal{G}$; density metric g.

Output: A dense subgrpah of $\mathcal{G}$.

if G is bipartite **then**
    $V \leftarrow L \cup R$
else if G is monopartite **then**
    $V \leftarrow V$
end if
Construct the priority tree T from V
$X_0 \leftarrow V$
for $t = 1, \ldots, |V|$ **do**
    ▷ *Greedly remove the vetex from X_{t-1} and its adjacent edges*
    $i^* \leftarrow \arg\max_{i \in X_i} g(X_{t-1} \setminus \{i^*\})$
    Update the priorities in T for all neighbors of node i^*
    $X_t \leftarrow X_{t-1} \setminus \{i^*\}$
end for
return $\arg\max_{X_i \in \{X_0, \ldots, X_{|V|}\}} g(X_i)$

最密的子图通常是嵌入在一个大规模图的背景中且相对很小的一部分。因此，为了找到一个近似解或者最优解的候选集，算法 Alg. 4 的贪心剥离过程中需要进行多次的迭代，这些前期的逐步删除、更新过程实际是非常低效的。

4.5.1 理论分析结果

引理 4.10 和引理 4.13 说明了理论最优密度的上界 $g_{opt} \leq \lambda_1$，且由定理 4.9 和定理 4.12 说明了 σ_k 是在与 $\mathcal{S}_{k-1}$ 正交的实空间中检测的最优值上界，而最优解 S^* 的展开形式强调了实值的特征向量与最密子图检测之间的联系，即在顶部较大奇异值对应的奇异向量中属于 S^* 的这些节点应该具有更高的重要度。

考虑到顶部几个奇异向量中元素的偏态分布，在最密子图检测时可以构造如下的一些较小的候选节点子集，从而避免从整个图开始检测。这一候选集表示为 $S_C = \{S_1, \ldots, S_k\}$，其中 $1 \leq k < n$，其中的元素为：

$$S_i = \{k; \mathbf{u}_{ik} > \Delta_L, k \in \lceil |L| \rceil\} \cup \{k; \mathbf{v}_{ik} > \Delta_R, k \in \lceil |R| \rceil\},$$

算法 5 SpecGreedy Algorithm: General dense subgraph detection

Input: Matrix $\mathbf{A}_r$ of the positive residual graph $\mathcal{G}_r$;
density metric g; top approximation rank k.

Output: The densest subgraph.

1: $[\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}] = \mathbf{SVD}(\mathbf{A}_r, k)$ ▷ Top-k spectral decomposition of $\mathbf{A}_r$
2: $S = \varnothing$
3: **for** $r \leftarrow 1, \ldots, k$ **do**
4: Construct the candidate node subset S_r based on $\mathbf{u}_r$ and $\mathbf{v}_r$, i.e.
$S_r = \{i : \mathbf{u}_{ri} > \frac{1}{\sqrt{|L|}}, i \in L\} \cup \{j : \mathbf{v}_{rj} > \frac{1}{\sqrt{|R|}}, j \in R\}$
▷ Greedily remove nodes to maximize the density metric g.
5: $S_r^* \leftarrow \textsc{Greedy}(\mathcal{G}(S_r), g)$
6: **if** $g(S_r^*) > g(S)$ **then** ▷ $g(S) = g_{cur}^*$
7: $S \leftarrow S_r^*$
8: **end if**
9: **if** $g(S) > \sigma_{r+1}$ **then** ▷ Spectral early-stopping condition
10: **break**
11: **end if**
12: **end for**
13: **return** $\mathcal{G}(S)$.

即由第 i 个较大奇异向量 $\mathbf{u}_i$ 和 $\mathbf{v}_i$ 所定义，Δ_L 和 Δ_R 为预定义的截断阈值；由 S_i 所导出子图 $\mathcal{G}(S_i)$ 检测得到的最密子图密度为 $g_i \leq \sigma_i$ 。这里，根据公式 (4.9) 和公式 (4.12) 中对最优解的形式化，确定的截断阈值为 $\Delta_L = \frac{1}{\sqrt{|L|}}$ 和 $\Delta_R = \frac{1}{\sqrt{|R|}}$（如果矩阵 $\mathbf{A}_r$ 是公式 (4.9) 定义的对称阵时，$|L| = |R| = n$ 且 $\Delta_L = \Delta_R = \frac{1}{\sqrt{n}}$）。

4.5.2 算法

因此，本文利用图谱性质和贪心剥离策略发明了新算法 SpecGreedy 来求解 GenDS 问题，算法 Alg. 5 总结了该算法的结构。

给定一个正残差图 $\mathcal{G}_r$ 的邻接矩阵 $\mathbf{A}_r$、密度度量指标 g 以及用于控制最大候选集大小的顶部特征向量的近似秩 k，算法 SpecGreedy 首先找到矩阵的 top-k 的谱分解近似 (Line 1), 然后基于这些顶部奇异向量检测可能的最密子图。在每

一轮检测中，根据截断的奇异向量 $\mathbf{u}_r$ 和 $\mathbf{v}_r$ 构造候选子集 S_r, 然后利用 Greedy 算法检测对应 $\mathcal{G}(S_r)$ 的最密子图实现最大化密度指标 g；算法根据后一个奇异值与当前已检测得到的最优密度值的关系检查是否满足提前终止条件 (Line 8)；最后返回其中具有最大密度的子图。

那么，算法中待检查的候选子图需要多少个呢？假设 $\mathcal{G}(S_r)$ 是通过某些现成的最密子图检测算法（如上面所用的 Greedy 算法）检测得到的当前最优密度，如果存在 $1 < j \leq n$ 满足 $g^*_{cur} \geq \sigma_j$，由于按递减排列的奇异值 ($\sigma_j > \sigma_{j+1}$) 和前面提到的理论最大密度上界 ($g_i \leq \sigma_i$) 关系，而且基于奇异向量所构造的子图能检测到的最优密度即为 g^*_{cur}。值得一提的是，由于真实世界图的特征值、奇异值服从的幂率分布以及检测算法得到的解的理论界的保证（如准确解算法或者 $\frac{1}{2}$-最优的近似算法）使得候选集的大小会很有限。

除了算法中 Line 1 所示的直接计算 top-k 谱分解方法外，通过幂迭代方法或 Krylov 子空间方法 (如 Lanczos 方法 [194] 等)，也可以用一种惰性的或者延迟在线的方式计算矩阵的第 $r+1$ 个最大的奇异值分解结果。实验中采用一种递增分解的方式来实现：首先计算 top-l 的奇异值与奇异向量，如果 Line 8 中的提前终止条件不满足的话，根据递增步长 s 继续分解 top-$(l+s)$ 的奇异值与奇异向量。这种逐步分解的方式会不断迭代进行直到 $l + s > k$ 或符合提前终止条件。此外，为了进一步提升检测结果的质量，也可以考虑将 Line 5 的贪心法替换为其他的最密子图检测算法，如 Greedy++ [174] 或线性规划算法等。

定理 4.14 (算法时间复杂度). SpecGreedy 算法的复杂度为：

$$O(K \cdot |E| + K \cdot |E(\tilde{S})| \log |\tilde{S}|),$$

其中 $\tilde{S} = \max_{|S_i|} S_i$，$K$ 为近似秩的大小。

理想条件下，$K = \min\{k, r_{opt} + 1\}$ 其中 k 为输入的参数 r_{opt} 为产生最优密度 g^* 时对应的秩。在稀疏图中计算一个顶部特征向量或奇异向量的复杂度是线性的（$O(|E(V)|)$）；对应于子图 $\mathcal{G}(S)$，算法在 Line 5 中贪心检测算法的复杂度为 $O(|E(S)| \log |S|)$。由于真实世界图的奇异向量中顶部元素的偏态性，通常会有 $|\tilde{S}| \ll |V|$，因此，SpecGreedy 算法的复杂度与图中的边数是线性的。

4.6 实验验证与分析

本文设计完成了不同实验以回答下面的问题：

- *Q1*. 算法高效性: SpecGreedy 算法在检测最密子图任务时，与当前最优算法相比性能如何？
- *Q2*. 检测有效性: SpecGreedy 在真实数据集上的检测效果怎么样？对于注入的不同密度的子图以及对对比稠密子图能否有效的检测？
- *Q3*. 可扩展性: SpecGreedy 是否随着输入图的大小可线性扩展？

表 4.3 实验中所用数据集信息的总结

Table 4.3 Statistics information of real-world networks used in experiments

Name	$\|V\|$	$\|E\|$	Content
soc-twitter [195]	41.7M	1.47B	Social network
soc-Sinaweibo [196]	58.7M	261M	Social network
com-Orkut [197]	3.07M	117M	Social network
twitter-ASU [198]	11.3M	85.3M	Social network
livejournal-MPI [199]	5.28M	76.9M	Social network
ca-DBLP-NET [196]	1.31M	19.0M	Co-authorship
ego-gplus [197]	108K	12.2M	Social network
as-Skitter [197]	1.7M	11.1M	Internet topology
web-BerkStan [197]	685K	6.65M	Web
soc-Flickr [198]	80.5K	5.90M	Social
road-CA [197]	1.97M	5.53M	Road Net
com-WikiTalk [197]	2.39M	5.02M	Communication
web-Google [197]	876K	4.32M	Web graph
ca-Aminer [200]	1.56M	4.26M	Collaboration
road-TX [197]	1.38M	3.84M	Road Net
road-PA [197]	1.97M	3.08M	Road Net
soc-Youtube [197]	1.13M	2.99M	Social network
web-Stanford [197]	282K	2.31M	Web graph
ca-DBLP2012 [196]	317K	1.05M	Collboration

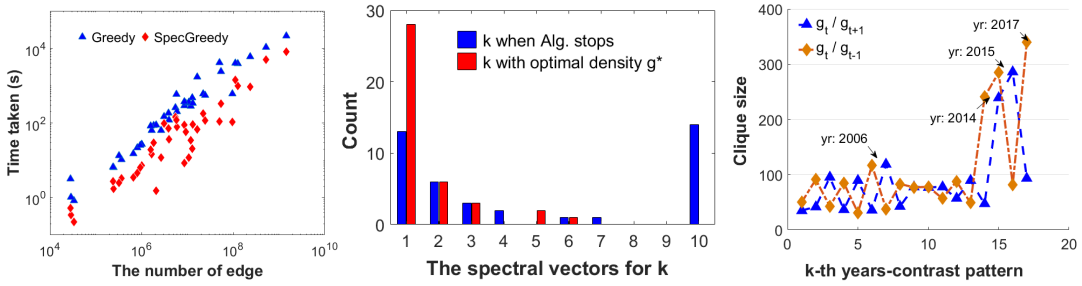
表 4.3 实验中所用数据集信息的总结

Name	$\|V\|$	$\|E\|$	Content
com-Amazon [197]	548K	926K	Community
twitter-ICWSM [195]	820K	835K	Social network
soc-Slashdot0902 [197]	82.2K	504K	Social network
soc-Slashdot0811 [197]	77.4K	469K	Social network
soc-Epinions [197]	75.9K	406K	Social network
blogcatalog [198]	10.3K	334K	Blog
ca-AstroPh [197]	18.7K	198K	Collaboration
email-Enron [197]	36.7K	183K	Communication
ca-HepPh [197]	12K	118K	Collaboration
soc-Hamsterster [196]	2.4K	16.6K	Social network
ca-GrQc [197]	5.2K	14.5K	Collaboration
*ca-Patents-AM [200]	2.08M	11.5M	Co-authorship
*ego-twitter [197]	81.3K	2.42M	Social network
livejournal-group [199]	10.7M	112M	Social network
cit-Patents-AM [200]	6.84M	54.0M	Citation
cit-Patents [197]	3.77M	16.5M	Citation
yelp-business [197]	86.4K	3.22M	Rating
beerAdvocate [197]	33.4K	65.9K	Review
*weibo-retweet	10.8M	50.1M	Social network
*amazon-Good [197]	3.38M	5.84M	Rating
*amazon-Art [197]	28.3K	28.0K	Rating

*: 对于带标记的部分数据集我们也考虑其带权图形式

机器配置：本章所有的实验都在一台有 2.4GHz Intel(R) Xeon(R) CPU 和 64GB 内存配置的机器上测试。

数据集：此处使用了来自多个的开源数据仓库中收集到的不同真实图进行实验，其中包括 Stanford's SNAP database [197]、ASU's Social Computing Data Repository [198]、Network Repository [196]、AMiner scholar datasets [200] 以及来



(a) 所有图上最密子图检测结果

(a) The densest subgraph detection for all graphs.

(b) 谱向量秩 k 的统计信息

(b) The statistic of spectral vectors for k.

(c) DBLP 共同作者图中的对比模式

(c) Contrast patterns for the DBLP co-authorship graphs.

图 4.2 SpecGreedy 算法在真实世界图上的检测性能。(a) 在所有网络上的最密子图检测中，SpecGreedy 运行速度都快于 Greedy 且得到密度相同或可比的结果；其中，对 *ca-DBLP2012*、*soc-twitter* 的大规模图产生的加速比分别达到 58.6× 倍和 3× 倍。(b) 关于谱向量的秩 k 的统计信息。对于多数数据集通常在第一个奇异向量 ($k = 1$) 上得到具有最优密度的最密子图结果。其中的蓝色柱状图显示了对于给定参数 $k = 10$ 时，算法终止时对应的实际 k 值的统计结果。(c) 显示了根据 2000 年至 2017 年的 DBLP 共同合作者数据构造的正残差图中检测得到的对比子图模式 (都具有团结构)，其中在 2017 年、2015 年和 2014 年都包含规模极大的团，它们对应与罕见的合作模式。

Figure 4.2 The performance of SpecGreedy for the real-world graphs. (a) SpecGreedy runs faster than Greedy algorithm in all networks for detecting the densest subgraph with the same or comparable density, and achieves 58.6× speedup for *ca-DBLP2012* and about 3× for the largest graph *soc-twitter*. (b) The statistic information about the spectral vectors for k. The densest subgraphs with optimal density $g*$ are achieved in the first singular vector ($k = 1$) for most of the datasets. The blue bars show the statistics of k when algorithm stops given the parameter $k = 10$. (c) The contrast patterns (all cliques) for DB LP co-authorship data in $2000 - 2017$ with the positive residual $\mathcal{G}_r$. There are some extremely large cliques in 2017, 2015, and 2014.

自 Koblenz Network Collection [195] 等。表 4.3 中列出了所用的真实世界的图数据的详细信息，表中的第一部分包含 32 个单部图，第二部分是一些二部图数据。所有图中的复边、自环已经移除，同时忽略有向图的连边方向；其中最大的无权图是包含约 1.47B 连边的 soc-twitter 的关系图，最小的无权图包含约 14.5K 的连边。实验中也对表 4.3 中标注的部分带权图进行了分析和检测。

实现：本章实现了与其他高效检测稠密子图算法的比较，包括 Greedy [91], SpokEn [54]、和 Fraudar [69]。其中，SpokEn 只通过对特征向量的直接截断以

检测最密子图。在所有实验中，SpecGreedy 中顶部特征向量的近似秩 k 设置为 10，逐步 SVD 分解中的参数设置为 $l = s = 3$。

4.6.1 Q1. 算法高效性

为了回答问题 Q1，本小节在 40 个无权网络上对比了 SpecGreedy 和基线方法 Greedy 的性能，并对它们的运行时间做了比较。图4.1a 中显示了在最密子图检测中，SpecGreedy 相较于 Greedy 方法带来的运行时间的提升比的统计信息；图4.2a中展示了这两种方法检测所花时间的详细信息：给出了每个网络的大小和两个方法的运行时间。

观察 1. SpecGreedy 算法比 Greedy 有更快的检测速度，并能够得到近似相同或更密的最优密度 (如图4.1b所示)。这些大小不同的数据集上，SpecGreedy 算法的运行结果中，有 17 个数据集的加速比为 3.0 – 5.0× 倍，8 个加速比为 1.5 – 3.0× 倍，还有 7 个图的加速比为 5.0 – 7.0× 倍；对 ca-DBLP2012 图的加速比超过 58.6× 倍；同时可以看出，SpecGreedy 算法对于大规模图是高效的，比如对 ca-DBLP-NET、cit-Patents 和 soc-twitter 得到的加速比分别达到 30×、25× 和 3× 倍。

针对表 4.3 中标注的其他 5 个带权图的实验也得到类似的结果：SpecGreedy 对其中的 3 个图产生的加速比为 24 – 39× 倍，另外两个的加速比为 11 – 17× 倍。而 Greedy 算法对带有少数较大连边权重的图性能表现很差，原因在于它需要从整个图开始逐步剥离掉所有的边才能得到最终的最优解。

图 4.2b 总结了计算得到最优密度 g^* 时对应的谱向量 k 值以及算法终止时（达到最大参数值设置或者满足提前终止条件）对应的实际 k 值的统计结果；k 值越大意味着需要更多的时间完成奇异值分解、对候选子图进行检测以及验证终止条件等步骤。

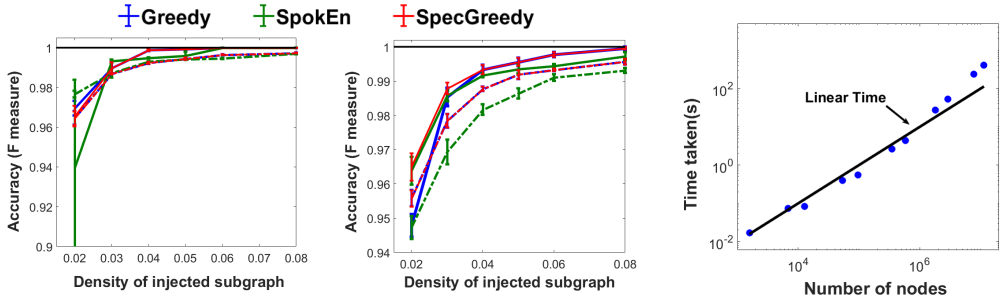
观察 2. 从图所示的实验结果可以看出，对于大多数的数据集，检测到具有最优密度 g^* 的最密子图是对应于第一个奇异向量，有 5 个图对应于第二个奇异向量，且仅有 3 个图需要检查超过 5 个的奇异向量；另外，在 26 个图的检测中 SpecGreedy 算法根据提前终止条件结束，其他的需要检查所有的 10 个奇异向量（由预设的参数决定，实际最密子图对应的 $k < 10$），其原因在于这些图得到的最优密度较小 (最密子图并不显著) 或者奇异值的幂律分布较为平坦。

针对需要检查多个奇异向量的图的检测结果中，其他由前 top-$(k-1)$ 个奇异

向量检测到的最密子图结果也对应于与最优密度相近似但规模较小的团。因此，上述观察结果以及奇异值 / 特征值的幂率分布特性使得算法 SpecGreedy 的高效性有所保证，且选择较小的 k 值就可以得到质量足够好的结果。

4.6.2 Q2. 检测有效性

本小节验证了 SpecGreedy 算法能够在真实世界图中检测到高质量的最密子图，可以准确地检测到不同密度的注入子图；而且算法在一个大规模的 DBLP 学术合作网络中检测到具有显著对比性的稠密子图模式。



(a) 无伪装型的注入实验

(a) Injection without camouflage.

(b) “随机伪装”的注入实验

(b) Injection with random camouflages.

(c) SpecGreedy 的线性可扩展性

(c) The linear scalability of SpecGreedy.

图 4.3 SpecGreedy 算法在合成数据上的性能与扩展性分析。(a)-(b) 对应于不同伪装策略下、注入密度变化时最密子图检测的准确率结果。其中的实线和虚线对应于 amazon-Art 数据的不同子集。SpecGreedy 的性能与 Greedy 算法相似，且优于 SpokEn 。(c) SpecGreedy 算法的运行时间与图的大小 (图的节点数) 呈线性关系。

Figure 4.3 SpecGreedy performance for the synthetic graphs and scalability analysis. (a)-(b) Detection accuracy for the injected dense subgraphs with varied injection densities. The solid and dash lines correspond to two different subsets of amazon-Art data. SpecGreedy achieves similar accuracy as the Greedy algorithm and outperforms SpokEn. (c) The time taken of SpecGreedy grows linearly with # of nodes in graph.

4.6.2.1 密度改善的结果

按照问题 Q1 中描述的实验设置，图 4.1b 展示了在最密子图检测中 SpecGreedy 算法相对于基线方法 Greedy 和 SpokEn 算法找到最密子图的最优密度的改善比的统计信息。

观察 3. 从图所示的实验结果中可以看出，SpecGreedy 的性能一致性地超过 SpokEn，在真实世界数据集中检测到了具有更大密度的最密子图结果，且在 soc-

twitter 图上得到的密度超过 28.3× 倍；同样，在大多数图上，相对于 GREEDY 算法，SPECGREEDY 也能够得到近似相同或者具有更大密度的最优检测结果，得到的最大密度提升比超过 1.26× 倍；其中，在 4 个图上 SPECGREEDY 检测的最优密度稍小于但非常接近 GREEDY 检测的结果 (即密度比 $\geq 0.996\times$)，另外有 2 个图的结果较差，提升比小于 0.9。

因此，利用最密子图的谱分布性质，SPECGREEDY 能够在大多数情况下提升 GREEDY 算法检测结果的质量：根据奇异向量中元素值的偏态分布一定程度上避免 GREEDY 算法在节点度相等时随机删除节点方式可能对最终结果带来的影响。

4.6.2.2 注入子图检测

本小结通过以注入子图作为真值进一步评估了算法 SPECGREEDY 在合成数据上的检测性能。考虑到实际场景的设置，此处在被选注入节点和其他剩余的未选节点之间也添加了一些额外的连边作为“伪装”。以 F score $= \frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$ 作为度量指标，比较了 SPECGREEDY、GREEDY（FRAUDAR）和 SPOKEN 算法检测注入模式的性能，并报告了在 5 次试验上 F score 的平均值结果。实验中，在 amazon-Art 评论数据的一个大小为 $4K\times 4K$ 的子图中注入了一个大小为 600×600 的、具有不同连边密度的子图；同时为了消除被选子图特殊性的影响，分别选了两种背景密度分别为 $2.7E-5$ 和 $3.4E-5$ 的子图作比较。图 4.3a-4.3b 显示了每种算法在不同注入密度下的检测准确率结果，可以看出 SPECGREEDY 能得到与 GREEDY 近似的高准确率，同时性能优于 SPOKEN 算法。

4.6.2.3 实例分析

作为一个实例分析，此处将 SPECGREEDY 算法应用于 DBLP 中从 2000 年到 2017 年的论文共同合作者 (co-authorship) 数据网络 [200][3]，用来识别其中有趣的对比子图模式。图 4.2c 展示了在逐年构造的正残差图中 SPECGREEDY 检测到的对比稠密子图模式的结果。这些最密对比子图都对应于不同大小的团，意味着这些形成团的连接仅出现在 $\mathcal{G}_t$ 而不是 $\mathcal{G}_{t-1}$ 或 $\mathcal{G}_{t+1}$ 中。从实验结果可以看出，分别在 2017 年，2015 年和 2014 年存在 3 个规模很大的团，其大小都超过 280 个节点，他们合作发表的成果分别是关于“脑网络与疾病”、“神经科学与医药”及“物理学”，这些大规模合作团体是由来自于不同学科的科研人员所构成。

[3]数据中作者数为 1.59M，时间为 2000 年到 2017 年，共包含 27.58M 连边。

4.6.3 Q3. 算法可扩展性

图 4.1c 和 4.3c 分别展示了随着图的连边数和节点数变化，SpecGreedy 算法运行时间的线性可扩展特点。这里选用 ca-Patents-AM 图进行最密子图检测，并以随机降采样方式从中选出不同比例的边数以得到不同大小的子图。图中与对角线平行的斜线表示线性增长的趋势。

4.7 本章小结

本章中探究了复杂关联数据中群体异常定义的问题多样性，研究了以稠密子图为对象的拓扑密集关联模式和对应聚集性群体异常行为的高效挖掘。针对大规模图数据在不同目标背景中的稠密子图检测问题，本章从理论分析的角度比较和对比了各类问题之间的区别与联系；提出了一种广义的统一形式化框架 GenDS，用以检测最密子图；基于谱理论的优化分析，结合大规模图的谱分布性质与幂率特征，设计了快速、高效的检测算法 SpecGreedy 来解决此广义问题，进一步提升现有不同检测方法的性能和效果，并用于其他有趣群体异常模式的检测。因此，本章的主要贡献点包括：

- **理论和对应总结：**对来自不同应用的最密子图检测问题进行统一的形式化，利用图谱理论分析了问题的最优化性质。
- **算法：**发明了一种快速、高效和可扩展的 SpecGreedy 检测算法以求解 GenDS 问题。
- **实验：**在 40 个不同大小的真实世界图数据上验证了 SpecGreedy 算法的效率，算法在应用实际中有显著效果，并在实际数据中发现了一些由 clique 表示的突然出现的学术合作关系；而且算法的复杂度是与图的大小线性相关的；

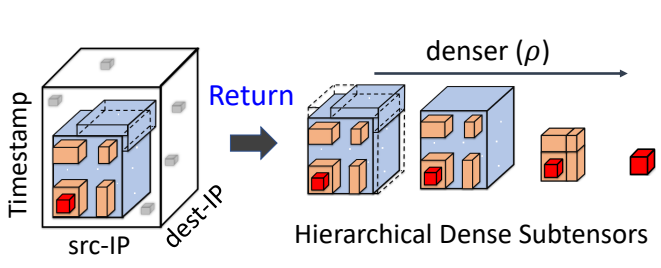
可复现性：本文提供了开源的程序 (https://github.com/wenchieh/specgreedy) 和在线可获得的真实数据。

第 5 章　多属性图中层次结构感知的群体异常模式发现

本章考察关联数据的多元异构形式以及异常模式的结构复杂性，研究了在多属性图背景下的多维稠密聚集模式以及具有层次性特征的群体异常行为的检测。利用稀疏张量模型，设计了具有统一形式的子张量密度指标和稠密子张量一致性表示；分析了在结构和密度具有层次性嵌套的复杂稠密子块模式，从梯度迭代优化的角度设计了具有线性可扩展的检测算法；并在大规模真实数据中验证检测方法的有效性，识别出实际应用中有趣的、可解释性的群体异常行为模式，弥补现有属性图中异常检测的缺陷和性能不足。

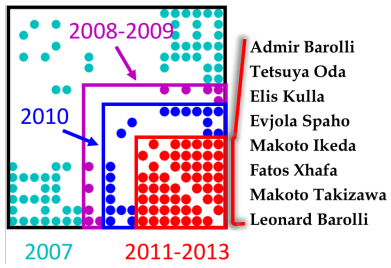
在以张量为表示的多属性数据中，稠密子张量检测在识别异常模式和欺诈行为等应用中都取得了显著的效果，包括社交网络分析和事件流检测等。现有的方法都采用直接分离的方式检测最密子张量，并假设不同的稠密子张量之间是相互独立且互斥的。而真实场景中的张量数据通常会出现层次化的性质，比如，core-peripheral 结构及网络中的动态社区等。

因此，本文提出一种新的框架 CatchCore，能够有效地发现具有层次性的稠密子张量。本章首先形式化地提出了一种可以通过梯度迭代更新方式进行优化的统一指标用于检测稠密子块；以此为基础，CatchCore 通过层次交替优化的方式检测具有层次性稠密的子张量；最后，给出了基于最小描述长度准则 MDL 的评价指标来度量检测结果的质量，并用以选择最优的层次化稠密子块。在合成数据和真实数据的大量实验结果表明，CatchCore 在稠密子张量检测和异常模式识别问题中性能都优于其他目前最优的对比方法；而且 CatchCore 在 DBLP 的数据中检测到一个具有层次性紧密合作行为的科研团体，同时 CatchCore 的算法复杂度与张量的各方面指标呈线性相关。



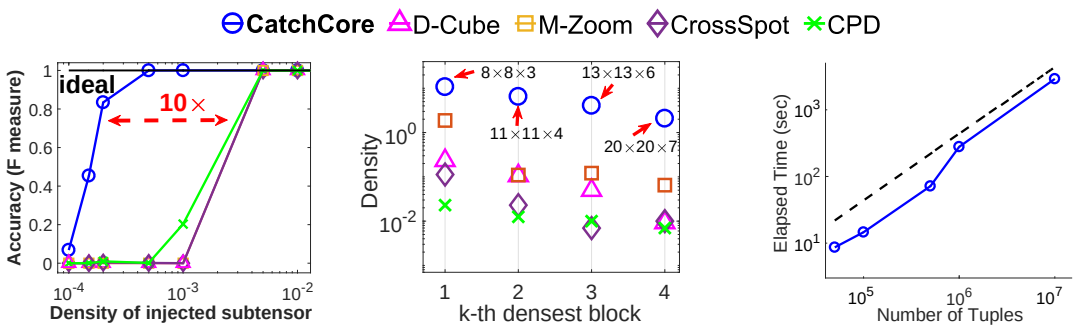
(a) 层次化稠密子张量的图例描述

(a) Pictorial depiction of hierarchical dense subtensors.



(b) DBLP 数据中的科研合作团体

(b) Researcher group in DBLP.



(c) 合成数据实验结果

(c) synthetic data result.

(d) DBLP 中密度前 4 的稠密子张量

(d) Top 4 blocks in DBLP.

(e) CatchCore 算法可扩展性

(e) Scalability w.r.t $nnz(\mathcal{R})$

图 5.1 层次化稠密子张量示例说明及 CatchCore 性能展示。(a) 给出了在 TCP DUMP 场景下层次化稠密子张量图示及检测流程。(b) DBLP 数据中检测到的多层稠密核，其包含 20 位共同合作者的研究团体。其中由红色标注的最密块由 8 位作者构成（如右侧列表所示），其密切合作关系持续 3 年时间（2011 - 2013），有不同颜色显示的外层稠密块由其他研究人员组成，他们之间的相对较弱合作关系发生于不同的时间段内。(c) 在合成数据上检测单个最密块实验中 CatchCore 性能优于其他方法，能够检测到的子块密度阈值比其他方法检测结果低 10 倍。(d) 对比其他方法，CatchCore 在 DBLP 数据中检测到的密度更大的稠密块，其中显示的密度在前 4 的子块密度，其结构对应于图 5.1b 所示。(e) CatchCore 的运行时间与张量中非零元素数呈线性关系。

Figure 5.1 Examples and CatchCore performance overview. (a) Example and workflow of hierarchical dense subtensors detection. (b) shows the detected dense co-authorship researcher group (a multi-layer core) of 20 users in DBLP. The densest block (red) lasts 3 years (2011-2013) containing 8 authors as the list shows, the outer hierarchies (with different colors) include other researchers and exist in various time ranges (text labeled). (c) CatchCore outperforms competitors for detecting injected blocks in synthetic data, it achieves lower detection bound than others. (d) CatchCore detects dense subtensors with higher density compared with baselines for the top four densest blocks in DBLP. These blocks correspond to a hierarchical group as Figure 5.1b shows. (e) CatchCore is linearly scalable w.r.t the number of tuples in tensor.

5.1 本章引言

稠密子图和子张量检测已被成功应用于许多不同领域，比如，在社交媒体及评分网站中检测异常或欺诈行为 (如同步性行为、提高评分等级等) [69, 89] 、在网络访问日志及流式数据中识别恶意攻击 [90, 93] 以及在生物网络中的基因社区的突变检测 [166] 等。

现有的许多最密子张量和子块检测算法 [89, 90, 92, 93] 均是独立地检测各个最密子块，即以“检测-移除-再检测”的方式逐个识别，其背后的假设在于所检测的子张量之间是分开且互斥的。然而，现实世界的多维张量数据通常会呈现出层次化的结构，如网络中的 core-peripheral 结构 [201] 以及社交网络中的动态社区 [140] 等。因此，现有方法不能识别稠密块中的细致结构 (如多层核 multi-layer core) 和不同块之间的关系 (如重合、包含等)。而传统的社区检测 [141, 202, 203] 或者密度子图检测 [67, 69, 204] 并不能应用于多维数据的分析场景。

由此引出的一个挑战性问题是，在多维数据中如何高效地检测具有层次化稠密性质的子张量模式。图 5.1a 中给出了以 TCP DUMP 攻击为例的图示描述。由于网络入侵攻击的攻击强度以及所选的对象和范围 (不同的 IP 地址或主机) 在不同时间段存在很大差别，因此整个攻击行为构成了多层次的高密度核 (core)。所以，层次化的检测方式有助于识别此类异常以及所对应的行为模式。

因此,本文提出了一种用于多维数据中检测层次化稠密核的算法框架 CatchCore。首先本章设计了一种可囊括各种子张量密度度量的通用指标，其目标函数可以通过梯度迭代更新的方式进行优化；基于此度量指标，CatchCore 算法能够按层次交替更新优化的方式高效地检测层次化稠密核结构，且算法收敛性也有理论性保证。大量实验表明，CatchCore 的性能明显优于其他基线方法，在真实数据中检测到多种重要的模式；在稠密子块检测中 CatchCore 算法具有最优性能，即使在低注入密度下也能有效地检测到注入子块 (图 5.1c)；在层次化稠密子张量的检测中，CatchCore 算法性能也优于其他对比方法；算法能捕获到 DBLP 数据中的科研合作团体 (图 5.1b) 及其他真实数据中的异常模式；借助一种信息论标准—— MDL 准则以评价检测效果的质量，并以此选择最优的层次化稠密子张量结果。

总的来说，本文的主要贡献如下所述:

- **统一的密度度量与检测算法：**设计了一种能够通过梯度迭代更新方式进

行优化的统一密度度量指标来检测稠密子张量，提出了具有理论保证的 CatchCore 算法来检测层次化稠密核结构以及基于最短描述长度的度量评价指标。

- **准确性：** CatchCore 能够更加准确地检测最密块以及层次化稠密核，在合成数据及真实数据上的检测结果明显优于目前其他对比算法（图 5.1c-5.1d）。
- **有效性：** CatchCore 在真实数据中成功地检测出了多种异常模式，包括可疑的好友关系连接、带周期性的网络入侵攻击以及具有层次化稠密结构的学术合作团体 (图 5.1b)。
- **可扩展性：** CatchCore 是高可扩展的，其运行时间与输入张量的各方面指标呈成线性关系 (图 5.1e) 。

可复现性：本文提供了开源的程序（https://github.com/wenchieh/catchcore），实验中的真实数据均在线可获得。

本章后续内容中，先对相关的研究方法进行了概述，在给出了以张量建模的多维关系的概念与表示之后，提出了最密子张量和层次化稠密子张量检测的形式化定义及分析，随后详细描述了本文提出的 CatchCore 检测算法，之后在多个真实数据上对 CatchCore 算法的性能进行了验证，并分析了在不同场景中算法检测群体异常行为的结果，最后对本章的工作进行总结和概括。

5.2 相关工作

5.2.1 张量中的稠密子图和子张量检测

稠密子图的检测已经被广泛研究 [91, 205, 206]，许多解决此类 NP 难问题的算法用以社区结构发现 [67, 141, 207]、异常检测 [28, 69]，并扩展到多维数据的情况 [90, 92]。CrossSpot [89] 通过贪心的方式调节种子直至收敛到一个局部最优解，以检测可疑的稠密子块；张量分解方法，如 HOSVD 和 CP 分解 [86]，也用于识别密集子张量；M-Zoom [90] 和 D-Cube [92] 采用有质量保证的贪心近似方法在大规模张量中检测稠密子张量，DenseAlert [93] 利用增量式算法在流式张量中检测稠密子块。所有这些现有方法都没有考虑不同子块间的结构以及它们之间的关系，也不能跟踪稠密张量的变化过程以及层次化结构。

5.2.2 层次化模式挖掘

社区结构广泛存在于各类图中 [141, 201]，[140, 203, 208, 209] 等在不同场景下对社区的进化行为以及层次化结构进行了分析和探索。[210] 提出了一种框架

表 5.1 CatchCore 与相关方法的比较

Table 5.1 Comparison of CatchCore and related methods. (✓, △ denote `supported' and `partly supported' respectively.)

	D-Cube [92]	M-Zoom [90]	CrossSpot [89]	DSBM [209]	HiDDen [67]	Fraudar [69]	**CatchCore**
Multi-Aspect Tensors	✓	✓	✓				✓
Hierarchical Structure					✓		✓
Detection Evaluation	△	△		✓		✓	✓
Linear Scalability	✓	✓	△		✓	✓	✓

对社区的重叠结构和活动周期进行联合学习。[211] 通过层次化时序关联挖掘机制检测多媒体应用中的视频事件。HiDDen [67] 检测了图中层次化稠密子图结构并发现了金融欺诈行为。[204] 通过剥离算法利用 k-core 分解来计算稠密子图的层次。但上述这些方法不能用于多维数据中层次化结构的分析。

5.2.3 异常和欺诈检测

综述文章 [28] 对图数据中异常检测方法给出了一个结构性的概述和总结。稠密子图或子张量通常包含一些可疑的模式，例如社交网络中的欺诈行为 [67, 69, 166]，网络分析中的端口扫描活动攻击 [89, 92] 以及同步行为等 [89, 92, 93]。

表 5.1 给出了 CatchCore 和上述部分相关方法在几个方面的对比，本文的 CatchCore 算法是唯一能满足所有要求的方法。

5.3 概念与符号

在本章中，所有的向量用粗体小写字母表示 (如 $\mathbf{x}$)，标量用小写字母表示 (如 c)，且 $[x] \equiv \{1, \ldots, x\}$。表 5.2 中列出了后续部分常用的符号和定义。

令 $\mathcal{R}(A_1, \ldots, A_N, C)$ 表示一个由 N 维属性 $\{A_1, \ldots, A_N\}$ 所构成的关系，其中的非负度量属性为 $C \in \mathbb{N}^{\geq 0}$。用 $\mathcal{R}_n$ 表示属性 A_n 中的不同值的集合，即其元素为 $a_k \in \mathcal{R}_n$。对于每一个条目（元组） $t \in \mathcal{R}$ 以及任意 $n \in [N]$，$t[A_n]$ 和 $t[C]$ 分别用于表示 A_n 和 C 中对应于 t 的值, 也就是 $t[A_n] = a_n$，$t[C] = c$。这样，多元

表 5.2 符号表总结

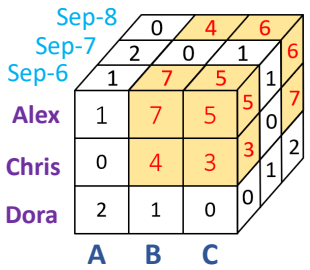
Table 5.2 Summary of symbols.

符号	定义
$\mathcal{R}(A_1,\dots,A_N,C)$	由张量表示的多元关系
$t(a_1,\dots,a_N,c)$	张量 $\mathcal{R}$ 中的一个元组 (实体)
N	张量 $\mathcal{R}$ 的模的数目
A, a	分别表示 $\mathcal{R}$ 的属性和某一属性取值
$\mathcal{R}_n$	$\mathcal{R}$ 中属性 A_n 的不同值的集合
$M_{\mathcal{R}}, V_{\mathcal{R}}, S_{\mathcal{R}}$	分别为 $\mathcal{R}$ 的质量、体积和张量势
$nnz(\mathcal{R})$	$\mathcal{R}$ 中非零元组的数目
$\rho(\mathcal{B})$	子张量 $\mathcal{B}$ 的密度度量
$\mathbf{x}_n \in \{0,1\}^{\lvert\mathcal{R}_n\rvert}$	$\mathcal{R}$ 中的对应第 n 模的指示向量
$\mathbf{X}_{\mathcal{B}}$	与子张量 $\mathcal{B}$ 对应的指示向量集
$\mathcal{B} \preccurlyeq \mathcal{R}$	表示 $\mathcal{B}$ 是 $\mathcal{R}$ 的一个子张量
$\mathcal{R} \bar{\times} \mathbf{X}$	$\mathcal{R}$ 与 $\mathbf{X}$ 的 full-mode 乘积
K	检测层次数的最大值
λ	正则项参数
$p\ (>0)$	子张量中缺失元组的惩罚因子值
$\eta\ (>1)$	相邻两层间的密度比值

关系 $\mathcal{R}$ 可以表示为大小为 $|\mathcal{R}_1| \times \cdots \times |\mathcal{R}_N|$ 的N-way 张量；对于张量中的每一元组的值，当 t 存在时为 $t[C]$，否则为 0。令 $\mathcal{R}(n, a_n) = \{t \in \mathcal{R};\, t[A_n] = a_n\}$ 表示 $\mathcal{R}$ 中属性 A_n 的值固定为 a_n 时的所有元组，由此形成一个张量的 *n-mode* 切片（即 $(N-1)$-way 张量）。此处，定义 $\mathcal{R}$ 的质量 (mass) 为其中所有元组的度量属性值之和，即 $M_{\mathcal{R}} = \sum_{t\in\mathcal{R}} t[C]$，$\mathcal{R}$ 的体积 (volume) 为 $V_{\mathcal{R}} = \prod_{n=1}^{N} |\mathcal{R}_n|$，张量 $\mathcal{R}$ 的势 (cardinality) 为所有维度大小之和，即 $S_{\mathcal{R}} = \sum_{n=1}^{N} |\mathcal{R}_n|$。

一个由 $\mathcal{R}$ 中属性集的子集构成的子张量定义为关系 $\mathcal{B} = \{t \in \mathcal{R};\, t[A_n] \in \mathcal{B}_n, \forall n \in \lfloor N \rceil\}$，即 $\mathcal{B}$ 中的元组的每一个属性 A_n 的值都属于 $\mathcal{B}_n$。用张量的形式表示，$\mathcal{B}$ 构成 $\mathcal{R}$ 中一个大小为 $|\mathcal{B}_1| \times \cdots \times |\mathcal{B}_N|$ 的子块（“block”）。文中

User	Item	Date	#Count
Alex	A	Sep-6	1
Alex	B	Sep-6	7
Chris	B	Sep-8	4
Dora	A	Sep-6	2
Alex	C	Sep-7	1
⋮	⋮	⋮	⋮



$\mathbf{User}: \mathbf{x}_1 = [1, 1, 0]$
$\mathbf{Item}: \mathbf{x}_2 = [0, 1, 1]$
$\mathbf{Date}: \mathbf{x}_3 = [1, 0, 1]$

(a) 3 维关系 $\mathcal{R}$. (着色的为部分记录) **(b) 张量形式表示的 $\mathcal{R}$** **(c) 子张量 $\mathcal{B}$ 的指示向量集**

(a) 3 dimensions relation $\mathcal{R}$. (only a small part are colored) **(b) Tensor Representation of the relation $\mathcal{R}$** **(c) Indicator vectors for attributes of $\mathcal{B}$**

图 5.2 多维数据的张量表示图例化示例。(a) 表示 $\mathcal{R}$(user, item, date, #count) 的多维数据。(b) 用 3-way 张量形式化表示 $\mathcal{R}$，其中的着色区域表示由 (a) 中部分着色记录构成的子张量 $\mathcal{B}$ 。(c) 为对应的各属性维度的指示向量，表示 $\mathcal{B}$ 的指示向量集为 $\mathbf{X}_{\mathcal{B}} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$，体积为 $V_{\mathcal{B}} = \prod_{i=1}^{3} ||\mathbf{x}_i||_1 = 2 * 2 * 2 = 8$。

Figure 5.2 Pictorial description of Example 5.1. (a) Relation $\mathcal{R}$(user, item, date, #count). (b) 3-way tensor representation of $\mathcal{R}$, the colored region indicates a subtensor $\mathcal{B}$ formed by some colored tuples in relation $\mathcal{R}$. (c) The indicator vectors collection representation for subtensor $\mathcal{B}$ can be denoted as $\mathbf{X}_{\mathcal{B}} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ and $V_{\mathcal{B}} = \prod_{i=1}^{3} ||\mathbf{x}_i||_1 = 2 * 2 * 2 = 8$.

用 $\mathcal{B} \lesssim \mathcal{R}$ 表示“$\mathcal{B}$ 是 $\mathcal{R}$ 的子张量”的关系；类似 $\mathcal{R}$ 的方式，可以定义子张量 $\mathcal{B}$ 的质量、体积和张量势等。数学上，对于任意的 $n \in \lfloor N \rceil$，可以用一个指示向量 $\mathbf{x} \in \{0,1\}^{|\mathcal{R}_n|}$ 来表示任一属性 $a_n \in \mathcal{R}_n$ 是否属于 $\mathcal{B}_n$，并且 $\mathbf{x}[a_n] = 1$ 当且仅当 $\mathcal{B}(n, a_n) \subseteq \mathcal{R}(n, a_n)$。由此，$\mathcal{B}$ 与 $\mathcal{R}$ 的包含关系可以用一个指示向量集来表示，即 $\mathbf{X}_{\mathcal{B}} = \left\{\mathbf{x}_n \in \{0,1\}^{|\mathcal{R}_n|}; \forall n \in \lfloor N \rceil\right\}$。一些特殊的情况中，$\mathbf{X_0} = \left\{\mathbf{0}^{|\mathcal{R}_n|}; \forall n \in \lfloor N \rceil\right\}$ 对应于空张量 (Null tensor：$\varnothing$)，而 $\mathbf{X_1} = \left\{\mathbf{1}^{|\mathcal{R}_n|}; \forall n \in \lfloor N \rceil\right\}$ 对应于张量 $\mathcal{R}$。

给定一个张量 $\mathcal{R}$ 的指示向量 $\mathbf{x} \in \{0,1\}^{|\mathcal{R}_n|}$，则对应于第 n 个属性由值 $\{a; \mathbf{x}[a] = 1, a \in \mathcal{R}_n\}$ 所构成的子张量可以表示为 $\mathcal{R} \times_n \mathbf{x}$。其中“$\times_n$”是张量和向量的 *n-mode* 乘积，即张量 $\mathcal{R}$ 与向量 $\mathbf{x}$ 之间元素间的乘积表示：

$$(\mathcal{R} \times_n \mathbf{x})_{i_1 \ldots i_{n-1} i_{n+1} \ldots i_N} = \sum_{i_n=1}^{|\mathcal{R}_n|} t(i_1, \cdots, i_N, c) x_{i_n}.$$

另外，$\mathcal{R}$ 中对应于指示向量集 $\mathbf{X}_{\mathcal{B}}$ 的子张量 $\mathcal{B}$ 可通过 $\mathcal{B} = \mathcal{R} * (\mathbf{x}_1 \circ \cdots \circ \mathbf{x}_N)$ 计算得到，其中“$*$”是张量的 *Hadamard* 乘积，“$\circ$”是向量间的外积。

例 5.1 (Review History). 假定一种多元关系为 $\mathcal{R}$(user, item, date, #count)，它的三个属性为 { user, item, date }，另一个 C = #count 是度量属性。$\mathcal{R}$ 中的每一个元

组 $t=(u,i,d,c)$ 表明用户 u 在日期为 d 的时间访问条目 i 的总次数为 c。

图 5.2a 给出了关系 $\mathcal{R}$ 的一个示例描述。其中 $\mathcal{R}_1=\{Alex,Chris,Dora\}$，$\mathcal{R}_2=\{A,B,C\}$，$\mathcal{R}_3=\{Sep-6,Sep-7,Sep-8\}$。图 5.2b 中的着色部分展示了一个子张量 $\mathcal{B}\lesssim\mathcal{R}$，其构成属性为 $\mathcal{B}_1=\{Alex,Chris\}$，$\mathcal{B}_2=\{B,C\}$，$\mathcal{B}_3=\{Sep-6,Sep-8\}$。利用 $\mathcal{R}$ 和指示向量集 $\mathbf{X}_{\mathcal{B}}=\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3\}$ (如图 5.2c所示) 可以对 $\mathcal{B}$ 进行表示和计算：以指示向量 $\mathbf{x}_1$ 为例，由于 $\mathcal{B}_1$ 中不包含 $\mathcal{R}_1$ 中的“Dora”，所以 $\mathbf{x}_1=[1,1,0]$；根据张量体积的定义，$V_{\mathcal{B}}=\prod_{i=1}^{3}\left|\left|\mathbf{x}_i\right|\right|_1=2*2*2=8$。

5.4 问题形式化与框架

本章节中提出一种能够用梯度更新方式来优化的统一度量指标，以用于检测稠密子张量，然后借此给出层次化稠密子张量检测的问题形式化定义。

5.4.1 最密子张量检测框架

令 $\mathcal{R}$ 是一个 N-way 张量，$\mathcal{B}$ 是 $\mathcal{R}$ 中由指示向量集 $\mathbf{X}_{\mathcal{B}}$ 定义的的子张量，那么 $\mathcal{B}$ 质量 $M_{\mathcal{B}}$ 可表示为：

$$M_{\mathcal{B}}=\mathcal{R}\bar{\times}\mathbf{X}_{\mathcal{B}}=\mathcal{R}\times_1\mathbf{x}_1\cdots\times_N\mathbf{x}_N. \tag{5.1}$$

其中 *full-mode* 乘积 $\bar{\times}$ 按维度逐个应用前述的 n-mode 张量-向量乘积。以类似于 [80] 中稠密子图的密度定义方式，此处提出了如下的统一指标：

定义 5.1 (Entry-Plenum)**.** 假定 $\mathbf{X}$ 是一个指示向量集，ϕ 是一个正值常数，给定任意两个严格递增的函数 g 和 h，则 entry-plenum 定义为：

$$f_{\phi}(\mathbf{X})=\begin{cases}0 & \mathbf{X}=\mathbf{X_0},\\ g(M_{\mathbf{X}})-\phi\cdot h(S_{\mathbf{X}}) & \text{otherwise.}\end{cases} \tag{5.2}$$

其中，$M_{\mathbf{X}}$ 和 $S_{\mathbf{X}}$ 分别表示由 $\mathbf{X}$ 定义的 $\mathcal{R}$ 中子张量的质量和大小，$S_{\mathbf{X}}$ 可以是 $V_{\mathbf{X}}$，$S_{\mathbf{X}}$ 或其他形式。

大多数现在常用的子张量密度度量指标 [89, 90, 92] 都能纳入到上述形式化框架中。例如，

- 令 $g(x)=\log x$，$h(x)=\log\frac{x}{N}$，$\phi=1$ 且 $S_{\mathbf{X}}=S_{\mathbf{X}}$，$f_{\phi}(\mathcal{B})$ 等价于算术平均质量 (*arithmetic average mass*)，$\rho_{ari}(\mathcal{B})=\frac{M_{\mathcal{B}}}{S_{\mathcal{B}}/N}$。

- 令 $g(x) = h(x) = \log x$ 且 $S_{\mathbf{X}} = V_{\mathbf{X}}$，如果 $\phi = 1$，则 $f_\phi(\mathcal{B})$ 对应于体积密度 (*volume density*)，即 $\rho_{vol}(\mathcal{B}) = \frac{M_{\mathcal{B}}}{V_{\mathcal{B}}}$；如果令 $\phi = \frac{1}{N}$，则 $f_\phi(\mathcal{B})$ 等价于几何平均质量 (*geometric average mass*) $\rho_{geo}(\mathcal{B}) = \frac{M_{\mathcal{B}}}{V_{\mathcal{B}}^{1/N}}$。

从上述 entry-plenum 定义形式来看，原则上第一项 $g(M_{\mathbf{X}})$ 倾向于检测质量较大的子张量，而第二项 $-\phi \cdot h(S_{\mathbf{X}})$ 则起到正则约束的作用，惩罚体积大的块。由此，在 entry-plenum 密度指标下，最密子张量检测问题能够形式化定义为:

> 问题 2 (最密 (g, h, ϕ)-entry-plenum 子张量检测). **给定**一个 N-way 张量 $\mathcal{R}$，一个正的常数 $\phi > 0$，递增函数 g 和 h；**找到**一个指示向量集 $\mathbf{X}^*$，**使得**对任意可能的 $\mathbf{X} = \{\mathbf{x}_n \in \{0,1\}^{|\mathcal{R}_n|} : \forall n \in \lfloor N \rceil\}$ 均满足 $f_\phi(\mathbf{X}^*) \geq f_\phi(\mathbf{X})$。
> 那么，由指示向量集 $\mathbf{X}^*$ 导出的子张量就称之为张量 $\mathcal{R}$ 中的最密 *(g, h, ϕ)-entry-plenum* 子张量。

在给定某些密度指标情况下，通常最密子张量的检测问题是 NP 难的 [90, 176]，对于大规模张量而言是几乎不可解的。出于可扩展性的考虑，现有的方法 [69, 90--92] 一般采用贪心近似的算法，依据某一以比例方式定义的密度指标，用迭代的方式从候选集中选择局部最优解。本文提出了一种可以通过梯度优化的角度来求解的最密子张量检测问题。给定某一初始的指示向量集 $\mathbf{X}$，将其作为一个块变量，能够使得上述问题在引入一定松弛约束后可以采用块变量的非线性 *Gauss-Seidel (GS)* [212] 方法进行优化，算法最终会收敛到某一确定解；而且在此定义下 $M_{\mathbf{X}}$ 和 $S_{\mathbf{X}}$ 对每一个指示向量都是可导的。因此，只要函数 g 和 h 可微，就可以采用基于梯度优化的策略进行变量更新。正如下面的 CatchCore 算法所示，这个计算过程的时间复杂度是线性的。

5.4.2 层次化稠密子张量

实际应用中，张量 $\mathcal{R}$ 中的不同稠密子块之间可能存在重叠或者包含关系，而不是像现有的稠密子块检测方法 [89, 90, 92] 所假设的那样：它们是分开且互斥的。在下面的实际例子和定义中明确地说明了多维数据中存在的层次化结构。

例 5.2 (网络入侵检测)**.** DARPA 数据集中包含约 60% 的日志记录是被标注为异常攻击类的，大多数是以不常见的爆发的方式出现，其中主要有 9 类不同的攻击，如 neptune，smurf，satan 等。这些不同入侵攻击有不同的攻击强度、攻击对象和爆发活动行为，产生了在不同时间段内有区分性的稠密模式。

给定一个 N-way 张量 $\mathcal{R}$，本章的目的在于找到几个在多层次上呈现不同密度的稠密子张量。此处用 $\rho(\mathcal{B})$ 表示子张量 $\mathcal{B}$ 的密度，$\mathcal{B}^k$ 表示 $\mathcal{R}$ 中位于第 k 层的稠密子张量。为了找到一些具有意义的模式，同时避免在不同层次上找到相同的结果，这里给出如下的定义和约束，令张量 $\mathcal{B}^0 \leftarrow \mathcal{R}$，常数 $K \in \mathbb{N}^+$，

定义 5.2 (层次化稠密子张量 (HDS-tensors))**.** 对任意 $k \in \lfloor K \rceil$, 子张量 $\mathcal{B}^{k-1}$, $\mathcal{B}^k$ 位于邻近两层，它们需要满足：

1. **密度约束**: 两个子张量的密度存在显著区别，即存在某一常数 $\eta > 1$，使得 $\rho(\mathcal{B}^k) \geq \eta\rho(\mathcal{B}^{k-1})$。

2. **结构约束**: 在高层次的张量结构更加紧致 (“close-knit”)，形成多层稠密核 $\mathcal{B}^k \preccurlyeq \mathcal{B}^{k-1}$，即 $\mathcal{B}_n^k \subseteq \mathcal{B}_n^{k-1}, \forall n \in \lfloor N \rceil$。

这样，在 K 层上找到的所有子张量构成了**层次化稠密子张量**。

需要注意的是,利用现有的稠密子张量检测算法直接优化来寻找 HDS-tensors 的方法是不可行的，因为它们并没有考虑不同块之间的关系；即使假设可能，它们通常会返回一些平凡解 (也就是不同层上的子张量是相同的)；另外，如何设计一种能够通过递归的启发式方法以优化整体目标函数目前仍未明确定义。

由 $\mathbf{X}^k$ 表示稠密子张量 $\mathcal{B}^k$，则 HDS-tensors 检测问题可形式化定义为：

问题 3 (层次化稠密子张量 HDS-tensors 检测). **给定** (1) 输入的 N-way 张量 $\mathcal{R}$, (2) 预期的邻近层次间的密度差异比 η^a, (3) 检测的最多层次数 K；**求解**指示向量集 $\{\mathbf{X}^1, \ldots, \mathbf{X}^r\}, r \leq K$ 层具有显著差异的层次化稠密子张量，**满足**：$\rho(\mathbf{X}^r) \geq \eta\rho(\mathbf{X}^{r-1}) \geq \cdots \geq \eta^{r-1}\rho(\mathbf{X}^1)$ 且 $\mathbf{X}^r \preccurlyeq \mathbf{X}^{r-1} \preccurlyeq \cdots \preccurlyeq \mathbf{X}^1$。

[a] 对于更一般的情况，可以对不同层之间的密度差设置不同的值而不是一个统一的固定值。

方便起见，此处定义一种 3 级下标的索引结构 (k, n, i) 来检索指示向量集，即 $\mathbf{X}_{(k,n,i)}$ 表示 $\mathbf{X}^k$ 中第 n 个指示向量 $\mathbf{x}_n$ 的第 i 个标量元素 x_i。$\mathbf{X}_{(k,\cdot,\cdot)}$ 和 $\mathbf{X}_{(k,n,\cdot)}$ 分别表示 $\mathbf{X}^k$ 和 $\mathbf{X}^k$ 中的指示向量 $\mathbf{x}_n$。

5.5 算法与分析

在本小节中提出了一种基于优化策略的 CatchCore 算法检测层次化稠密子张量，并给出了关于算法的性能分析。

5.5.1 基于优化的稠密子张量检测

此处，基于体积密度定义，给出一种具有可解释性的 entry-plenum 度量实例化。根据表示子张量 $\mathcal{B}$ 的指示向量集 $\mathbf{X}_{\mathcal{B}}$，$\mathcal{B}$ 的密度表示为 $\rho_{\mathcal{B}} = \frac{M_{\mathcal{B}}}{V_{\mathcal{B}}} = \frac{M_{\mathcal{B}}}{\prod_{\mathbf{x}\in\mathbf{X}_{\mathcal{B}}}||\mathbf{x}||_1}$，其中体积 $V_{\mathcal{B}}$ 是指示向量集中各维度的大小的乘积，即 $\prod_{\mathbf{x}\in\mathbf{X}_{\mathcal{B}}}||\mathbf{x}||_1$，也就等价于子张量中包含零元素在内的所有可能的元组。

为了找到稠密子张量，如果直接最大化密度指标 $\rho_{\mathcal{B}}$ 会产生一些平凡解，即所有度量属性最大的元组就能最大化密度指标，而在 0-1 张量中任意一个元组都可满足；另外，如果通过优化 $\mathbf{X}_{\mathcal{B}}$ 最大化由公式 (5.1) 定义的子张量质量时，产生的结果是另一个平凡解，即张量 $\mathcal{R}$ 本身，这是由于此定义中没有将大小或体积等因素考虑在内所导致的。

直观上讲，稠密子张量优化中需要最大化子张量中包含的元组的质量，同时最小化其中出现的缺失元组，因此，本文提出如下的优化目标：

$$\max_{\mathbf{X}:\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}} \mathcal{F}(\mathbf{X}) = (1+p)\mathcal{R}\bar{\times}\mathbf{X} - p\prod_{\mathbf{x}_n\in\mathbf{X}}\left|\left|\mathbf{x}_n\right|\right|_1 \quad \textbf{s.t. } \mathbf{x}_n \in \{0,1\}^{|\mathcal{R}_n|},\ \forall n \in \lfloor N \rceil. \tag{5.3}$$

其中，$p > 0$ 是给定的惩罚项参数。

上述定义背后原理在于,在最终的子张量中每个非零元组 t 对目标函数 $\mathcal{F}(\mathbf{X})$ 的贡献即为 $t[C]$；而其他每一个缺失的元组 $\tilde{t}$ 则由 p 表示惩罚（可视为 $\tilde{t}[C] = -p$）。通过这种方式，该目标函数能够最大化目标子张量的总质量同时最小化缺失元组的惩罚。而且该目标也对应于最密 (g, h, ϕ)-entry-plenum 子张量的一个实例化，此时对应的参数设置为 $g(x) = h(x) = x, \phi = \frac{p}{1+p}$，且 $S_{\mathbf{X}} = V_X$。

由于公式 (5.3) 中的指示向量元素都是二值变量，因此，目标函数 $\mathcal{F}(\mathbf{X})$ 是一个由组合优化产生的 NP 难问题。这里，将其中的 $0-1$ 整数规划问题的约束条件放松为一个多项式时间可解的线性规划问题，即满足 $\mathbf{0}^{|\mathcal{R}_n|} \le \mathbf{x}_n \le \mathbf{1}^{|\mathcal{R}_n|}$；松弛后的约束条件表示切片 $\mathcal{R}(n, a_n)$ 属于目标子张量的概率，最终只有那些概率为 1 的属性会被选择。

5.5.2 层次化稠密子张量检测

基于上面的最密子张量优化的形式化定义，求解 K 层的层次化最密子张量的最直观方法就是在每层中优化公式 (5.3)，即最大化 $\sum_{k=1}^{K}\mathcal{F}(\mathbf{X}^k)$，同时增加之前所述的关于 HDS-tensors 定义中的约束条件。根据密度约束 1 ，在第 k 层稠密

子张量的密度可以表示为 $\rho_{\mathbf{X}^{k+1}} \geq \eta\rho_{\mathbf{X}^k}$，这里 $\eta > 1$ 为密度倍数值；根据结构约束 2 ($\mathcal{B}^{k+1} \preccurlyeq \mathcal{B}^k \preccurlyeq \mathcal{B}^{k-1}$)，同时为了避免找到的相同的结果，指示向量集需要满足 $\mathbf{X}_{(k+1,n,\cdot)} \leq \mathbf{X}_{(k,n,\cdot)} \leq \mathbf{X}_{(k-1,n,\cdot)}, \forall\, n \in \lfloor N \rceil$，这里假定 $\mathbf{X}^0 = \mathbf{X_1}$，$\mathbf{X}^{K+1} = \mathbf{X_0}$。由此，得到的整体优化目标函数为：

$$\begin{aligned} &\max_{\mathbf{X}^1,\ldots,\mathbf{X}^K} \sum_{k=1}^{K} \mathcal{F}(\mathbf{X}^k) \\ &\text{s.t. } \rho_{\mathbf{X}^{h+1}} \geq \eta\rho_{\mathbf{X}^h},\ \mathbf{X}_{(h+1,n,\cdot)} \leq \mathbf{X}_{(h,n,\cdot)} \leq \mathbf{X}_{(h-1,n,\cdot)}.\ \ \forall\, h \in \lfloor K \rceil;\ n \in \lfloor N \rceil. \end{aligned} \tag{5.4}$$

很明显，这个带有界约束的多元非线性规划（BMV-NLP）优化问题是非凸的，不存在解析形式的解。因此，可以通过下面的方式将其中的约束条件进行松弛变换，使其成为目标函数的一个正则项：令 $d^{k-1} = \eta\rho_{\mathbf{X}^{k-1}}$ 表示对 $\mathbf{X}^k$ 的一个约束条件，对应的正则项可表示为 $\mathcal{G}(\mathbf{X}^k) = \mathcal{R} \bar{\times} \mathbf{X}^k - d^{k-1} \prod_{n=1}^{N} \left|\left|\mathbf{X}_{(k,n,\cdot)}\right|\right|_1$. 可以看到，该项同样具有 entry-plenum 形式。

因此，HDS-tensors 检测中带松弛条件的优化目标为：

$$\begin{aligned} &\max_{\mathbf{X}^1,\ldots,\mathbf{X}^K} \sum_{k=1}^{K} \mathcal{F}(\mathbf{X}^k) + \lambda \sum_{j=2}^{K} \mathcal{G}(\mathbf{X}^j) \\ &\text{s.t. } \mathbf{X}_{(h+1,n,\cdot)} \leq \mathbf{X}_{(h,n,\cdot)} \leq \mathbf{X}_{(h-1,n,\cdot)}. \quad \forall\, h \in \lfloor K \rceil;\ n \in \lfloor N \rceil. \end{aligned} \tag{5.5}$$

其中，参数 λ 控制正则项部分的重要度。进一步展开之后可以得到

$$\bar{\mathcal{F}}(\mathbf{X}^k) = (1 + p + \lambda)\mathcal{R} \bar{\times} \mathbf{X}^k - (p + \lambda d^{k-1}) \prod_{n=1}^{N} \left|\left|\mathbf{X}_{(k,n,\cdot)}\right|\right|_1 .$$

公式 (5.5) 展示的目标函数可重新表示为 $\mathcal{F}(\mathbf{X}^1) + \sum_{k=2}^{K} \bar{\mathcal{F}}(\mathbf{X}^k)$. 从中可以看出，对于 $k \geq 2$，$\bar{\mathcal{F}}$ 中对第 k 层中缺失元组的赋予一个比第 $(k-1)$ 层中缺失元组更大的惩罚参数，其目的在于使得不同层之间的密度差异更显著。

5.5.3 优化算法

本文提出 CatchCore 算法来求解上述优化问题，其主要结构如 Alg. 6 所示。针对利用规划方法解决 BMV-NLP 优化问题，由于公式 (5.5) 是一个非凸目标函数，同时对每一个指示向量集 $\mathbf{X}_{(k,\cdot,\cdot)}$ 都有高阶的有界约束。因此，采用交替迭代更新的方式进行求解，即在每一轮迭代中，固定当前更新的指示向量集之外的其他所有层次的指示向量集变量，对每个维度对应的指示向量 $\mathbf{X}_{(k,n,\cdot)}$ 更新时，也采用类似的方式交替更新。根据结构约束条件，对于任一维度 $n \in \lfloor N \rceil$，其可

算法 6 CatchCore Algorithm: HDS-tensors detection

Input: (1) the N-way tensor: $\mathcal{R}$

(2) the maximum number of hierarchies: K

(3) the penalty value for each missing entry: p

(4) the density ratio between two adjacent hierarchies: η

(5) the regularization parameter: λ

(6) maximum number of iterations: $t_{\max}$

Output: The dense subtensors indicator vector collections: $\{\mathbf{X}^1, \cdots, \mathbf{X}^r\}$.

initialize $\mathbf{X}^1, \cdots, \mathbf{X}^K$ as $\mathbf{X}^{\text{init}}_{(k,n,\cdot)}$

compute $\left|\left|\{\nabla^P_{(\mathbf{x}^k_n)_{\text{init}}} f\}\right|\right|_2$ $(\forall\, n \in \lfloor N \rceil, k \in \lfloor K \rceil)$ ▷ initial norm

$t \leftarrow 1$, $r \leftarrow 1$

while $t \leq t_{\max}$ **and** Eq. (5.9) is not satisfied **do** ▷ stop criteria

// Gauss-Seidel method updating

  for $k \leftarrow 1 \ldots K$ **do** ▷ for the k^{th} hierarchy

    for $n \leftarrow 1 \ldots N$ **do** ▷ for the n^{th} dimension

      ▷ block coordinate descent

      $\mathbf{x}^k_n \leftarrow$ OneWayOpt $(\mathbf{x}^k_n)$

    end for

    update $\mathbf{X}^k$

  end for

  $t \leftarrow t + 1$

end while

while $r \leq K$ **do** ▷ select significant subtensors

  $\mathcal{S} = \{\mathbf{X}_{(r,n,\cdot)}; \max \mathbf{X}_{(r,n,\cdot)} < 1, \forall n \in \lfloor N \rceil\}$

  if $\mathcal{S} \neq \varnothing$ **then**

    break ▷ no significant subtensors for hierarchies $> r$

  else:

    $r \leftarrow r + 1$

  end if

end while

return the resultant r indicator vector collections $\{\mathbf{X}^1, \cdots, \mathbf{X}^r\}$.

行解 $\mathbf{X}_{(k,n,\cdot)}$ 在高维空间中被邻近的指示向量 $\mathbf{X}_{(k-1,n,\cdot)}$ 和 $\mathbf{X}_{(k+1,n,\cdot)}$ 所约束，因此优化时可以将这个约束松弛为 $\mathbf{0}^{|\mathcal{R}_n|} \leq \mathbf{X}_{(k,n,\cdot)} \leq \mathbf{X}_{(k-1,n,\cdot)}$，并可以按序依次得到 $\mathbf{X}^1, \mathbf{X}^2, \cdots, \mathbf{X}^K$。

具体求解过程为：首先基于 $\{\mathbf{0}^{|\mathcal{R}_n|} \leq \mathbf{X}_{(1,n,\cdot)} \leq \mathbf{1}^{|\mathcal{R}_n|}, \forall n \in \lfloor N \rceil\}$ 的约束，同时忽略其他 $\mathbf{X}^k$ 中的变量约束条件，可得到 $\mathbf{X}^1$；然后根据由第一步的结果所导出的约束 $\{\mathbf{0}^{|\mathcal{R}_n|} \leq \mathbf{X}_{(2,n,\cdot)} \leq \mathbf{X}_{(1,n,\cdot)}, \forall n \in \lfloor N \rceil\}$，同时也忽略其他约束条件，可以进一步得到 $\mathbf{X}^2$。用这个方法就可以逐层地求解 K 层稠密子张量的检测子问题。具体实现时，本章采用 Trust-Region 方法 [213] 来求解每一个边界约束的非线性规划子问题，可以将公式 (5.5)所述的优化问题重新表示为：

$$\begin{aligned} \min_{\mathbf{X}^1,\ldots,\mathbf{X}^K} f(\mathbf{X}^1,\ldots,\mathbf{X}^K) &= -(1+p)\mathcal{R} \bar{\times} \mathbf{X}_{(1,\cdot,\cdot)} + p \prod_{n=1}^{N} \left|\left|\mathbf{X}_{(1,n,\cdot)}\right|\right|_1 \\ &- (1+p+\lambda) \sum_{k=2}^{K} \mathcal{R} \bar{\times} \mathbf{X}_{(k,\cdot,\cdot)} + \sum_{k=2}^{K} (p + \lambda d^{k-1}) \prod_{n=1}^{N} \left|\left|\mathbf{X}_{(k,n,\cdot)}\right|\right|_1 \\ \text{s.t. } & \mathbf{X}_{(h+1,n,\cdot)} \leq \mathbf{X}_{(h,n,\cdot)} \leq \mathbf{X}_{(h-1,n,\cdot)}. \qquad \forall h \in \lfloor K \rceil;\ n \in \lfloor N \rceil. \end{aligned} \tag{5.6}$$

这里利用一种简单有效的交替投影梯度下降的方法 [214, 215] 求解该优化问题。对于任一维度 n，对公式 (5.1) 中的变量 $\mathbf{x}_n$ 计算梯度可得：

$$\nabla_{\mathbf{x}_n} M_{\mathcal{B}} = \mathcal{R} \bar{\times}_{(-n)} \mathbf{X}_{\mathcal{B}} = \mathcal{R} \times_1 \mathbf{x}_1 \cdots \times_{(n-1)} \mathbf{x}_{(n-1)} \times_{(n+1)} \mathbf{x}_{(n+1)} \cdots \times_N \mathbf{x}_N.$$

因此，目标函数 $f(\cdot)$ 对于 $\mathbf{x}_n^1$ $(\mathbf{X}_{(1,n,\cdot)})$ 和 $\mathbf{x}_n^k$ $(\mathbf{X}_{(k,n,\cdot)}, k \geq 2)$ 的梯度分别是：

$$\begin{aligned} \nabla_{\mathbf{x}_n^1} f &= -(1+p) \nabla_{\mathbf{x}_n^1} M_{\mathbf{X}^1} + p \prod_{\mathbf{x}_n \in \mathbf{X}^1 / \{\mathbf{x}_n^1\}} \left|\left|\mathbf{x}_n\right|\right|_1 \mathbf{1}, \\ \nabla_{\mathbf{x}_n^k} f &= -(1+p+\lambda) \nabla_{\mathbf{x}_n^k} M_{\mathbf{X}^k} + (p + \lambda d^{k-1}) \prod_{\mathbf{x}_n \in \mathbf{X}^k / \{\mathbf{x}_n^k\}} \left|\left|\mathbf{x}_n\right|\right|_1 \mathbf{1} \end{aligned} \tag{5.7}$$

其中，$\mathbf{1}$ 表示一个大小为 $|\mathcal{R}_n|$ 的全一向量 (其大小与 $\mathbf{x}_n^1$、$\mathbf{x}_n^k$ 相同)。令 $\mathbf{x}_n$ 表示在投影梯度更新方法中第 k 层上的迭代更新变量，新的迭代值由 $\tilde{\mathbf{x}}_n = P(\mathbf{x}_n - \alpha \nabla_{\mathbf{x}_n} f)$ 的更新规则给出，其中 $P(\cdot)$ 表示将该向量投影到之前所述的有界可行域范围内；$-\nabla_{\mathbf{x}_n} f$ 表示与梯度相反的搜索方向，α 为更新步长，可以通过 Armijo 步长选择策略计算得到。在以 Armijo 规则的线性搜索方法中，通过多次计算直至以下条件得到满足来确定一个恰当的步长 α，其停止条件为：

$$f(\cdot, \mathbf{X}^k_{\mathbf{x}_n \to \tilde{\mathbf{x}}_n}, \cdot) - f(\cdot, \mathbf{X}^k, \cdot) \leq \sigma \cdot (\nabla_{\mathbf{x}_n} f)^T (\tilde{\mathbf{x}}_n - \mathbf{x}_n) \tag{5.8}$$

其中 $\mathbf{X}^k_{\mathbf{x}_n \to \tilde{\mathbf{x}}_n}$ 表示由更新后的指示向量 $\tilde{\mathbf{x}}_n$ 代替 $\mathbf{X}^k$ 中的指示向量 $\mathbf{x}_n$ 得到的结果，参数 σ $(0 < \sigma < 1)$ 常用设置为 0.01。这样就可以交替地更新当前更新的第 k 层的指示向量集中的每一个指示向量 $\mathbf{x}_n$，具体更新算法 OneWayOpt 如 Alg.7所示。

本文提出了 CatchCore 算法解决由公式 (5.6) 定义的规划优化问题。算法中，首先对指示向量集依概率进行初始化，即在第 1 层中切片 $\mathcal{R}(n, a_n)$ $(\mathbf{X}^{\text{init}}_{(k,n,i)})$ 被选择的概率为 0.5，其他层次的初始概率为 0.01，这样可以一定程度上避免得到一些平凡解。基于收敛性的考虑，即使得问题的解最终接近于某一固定点，迭代过程中除了有总迭代次数 $t_{\max}$ 的约束条件外，此处也采用了一种对有界约束优化问题方法常用规则作为终止条件：

$$\left\| \left\{ \nabla^P_{\mathbf{x}^k_n} f; \forall n,\ k \right\} \right\|_2 \leq \epsilon \left\| \left\{ \nabla^P_{\mathbf{X}^{\text{init}}_{(k,n,\cdot)}} f; \forall n,\ k \right\} \right\|_2, \tag{5.9}$$

其中，$\nabla^P_{\mathbf{x}^k_n} f$ 是元素级的投影梯度操作，对其中第 i 个元素计算方式为：

$$(\nabla^P_{\mathbf{x}^k_n} f)_i = \begin{cases} \min(0, (\nabla^P_{\mathbf{x}^k_n} f)_i) & \text{if } \mathbf{X}_{(k,n,i)} = \mathbf{X}_{(k+1,n,i)}, \\ (\nabla^P_{\mathbf{x}^k_n} f)_i & \text{if } \mathbf{X}_{(k+1,n,i)} < \mathbf{X}_{(k,n,i)} < \mathbf{X}_{(k-1,n,i)}, \\ \max(0, (\nabla^P_{\mathbf{x}^k_n} f)_i) & \text{if } \mathbf{X}_{(k,n,i)} = \mathbf{X}_{(k-1,n,i)}. \end{cases}$$

然后，CatchCore 调用 OneWayOpt 算法，逐层地交替更新所有层次中每维度对应的指示向量直到收敛。最后，只选择 K 层中最显著的 r 层子张量作为最终结果返回（第 9 – 14 行所示）。

如 Alg. 7 所示的 OneWayOpt 算法，用以更新 $\mathbf{X}^k$ $(k \in \lfloor K \rceil)$ 中任意一个指示向量 $\mathbf{x}_n$。Armijo 更新中搜索一个恰当的更新步长 α 的过程是其中最耗时的步骤；因此，这里利用一种更加灵活的线搜索 (line search) [215] 的方式来减少对由公式 (5.8) 给出的终止条件的检验次数。具体来讲，令 α 表示更新 $\mathbf{x}_n$ 时的步长大小，并假定它与 $\tilde{\mathbf{x}}_n$ 的更新步长相似；因此，在更新 $\tilde{\mathbf{x}}_n$ 时，用 α 作为其初始值，然后每次增大或减小原来的 β 倍 $(0 < \beta < 1)$，当满足公式 (5.8) 时，即找到了最优的步长大小。Alg. 7 中的 Line 8 - 12 对应于高效搜索 α 的核心步骤，在实验中设置 $\beta = \frac{1}{10}$。

5.5.4 参数选择与结果评价

对于缺失元组的惩罚参数 p，其控制着最终子张量的最低密度，而密度约束 η 则影响最终检测到的层次数以及不同块之间的密度多样性。因此，在无监

算法 7 OneWayOpt Algorithm: one indicator vector updating

Input: (1) the indicator vector $\mathbf{x}_n$ to be updated for $\mathbf{X}^k$

(2) the initial update step size: α

(3) the ratio of decreasing the step size: $\beta \in (0,1)$ (default $\frac{1}{10}$)

(4) the linear search parameter: $\sigma \in (0,1)$

Output: the new updated indicator vector $\tilde{\mathbf{x}}_n$.

compute the gradient $\nabla_{\mathbf{x}_n} f$ as the Eq. (5.7).
compute the objective function $f(\cdot, \mathbf{X}^k, \cdot)$ with $\mathbf{x}_n$ (as the Eq. (5.6) shows)
initial $\alpha \leftarrow 1$ ▷ step size for updating
while not satisfy the Armijo's condition **do**
    $\tilde{\mathbf{x}}_n = P(\mathbf{x}_n - \alpha \nabla_{\mathbf{x}_n} f)$ ▷ the new $\mathbf{x}_n$
    compute the objective function $f(\cdot, \mathbf{X}^k_{\mathbf{x}_n \to \tilde{\mathbf{x}}_n}, \cdot)$ with $\tilde{\mathbf{x}}_n$
    adjust the step size α with β based on the condition in Eq. (5.8).
    if the condition in Eq. (5.8) is not satisfied **then**
        $\alpha \leftarrow \beta \cdot \alpha$ ▷ decrease α
    else
        $\alpha \leftarrow \frac{\alpha}{\beta}$ ▷ increase α
    end if
end while
return the final $\tilde{\mathbf{x}}_n$.

督应用中，如何选择合适的参数以及对某一给定参数配置下检测结果的评价就是一个挑战性问题。本文提出基于最短描述长度 MDL 的方式来衡量不同参数下的返回结果。原理上讲，通过利用层次化稠密子张量的结果来计算编码整个张量 $\mathcal{R}$ 时产生的比特数来选择最优的参数 (模型)，而最优的模型能够得到最短的编码长度。直观上，缺失元组越少、检测的层次越准确，就可以用最少的比特数以一种无损的方式对张量 $\mathcal{R}$ 进行编码表示。

对于指示向量 $\mathbf{X}_{(k,n,\cdot)}$ ($k \in \lfloor K \rceil$, $n \in \lfloor N \rceil$)，可采用哈夫曼编码来表示这个二进制字符串，其形式化表示为由一个二项式随机变量 X 控制的实现序列；由于 $\mathbf{X}_{(k,n,\cdot)} \leq \mathbf{X}_{(k-1,n,\cdot)}$ 条件存在，因此只需要考虑其中的重合部分即可，以避免对 0 元素进行重复编码；其中的重合部分可表示为：$\tilde{\mathbf{x}}^k_n = \{\mathbf{X}_{(k,n,i)}; \mathbf{X}_{(k-1,n,i)} = 1, \forall i \in$

$\lfloor |\mathcal{R}|_n \rceil\}$。指示向量 $\mathbf{x}$ 的熵表示为 $H(\mathbf{x}) = -\sum_{q\in\{0,1\}} P(X=q)\log P(X=q)$，其中 $P(X=q) = \frac{n_q}{||\mathbf{x}||_1}$，$n_q$ 表示 $\mathbf{x}$ 中取值为 q 的元素个数。因此，指示向量集 $\mathbf{X}^k$ 的描述长度为[1]：

$$L(\mathbf{X}^k) = \sum_{n=1}^{N} \left(\log^* \left|\left|\mathbf{X}_{(k,n,\cdot)}\right|\right|_1 + \left|\left|\mathbf{X}_{(k-1,n,\cdot)}\right|\right|_1 \cdot H(\bar{\mathbf{x}}_n^k)\right).$$

假定 $\mathbf{X}^{K+1} = \mathbf{X_0}$，$\mathbf{X}^k$ 对应于稠密子张量 $\mathcal{B}^k$。由于层次化子张量检测中的包含关系 $\mathcal{B}^{k+1} \lesssim \mathcal{B}^k$，因此仅需要基于某一概率分布对其中存在差异部分的元组进行编码表示，即 $\bar{\boldsymbol{B}}^k = \mathcal{B}^k - \mathcal{B}^{k+1}$。对于任一元组 $t \in \bar{\boldsymbol{B}}^k$，如果 $t[C] \in \{0,1\}$，t 建模为从一个二项分布中采样结果；如果 $t[C] \in \mathbb{N}^{\geq 0}$，可以通过以 $\bar{\boldsymbol{B}}^k$ 的密度 $\rho_{\bar{\boldsymbol{B}}^k}$ 为参数的泊松分布 [89] 来建模数据。因此，$\bar{\boldsymbol{B}}^k$ 的编码长度表示为：

$$L(\bar{\boldsymbol{B}}^k) = -\sum_{q\in\{t[C];t\in\bar{\boldsymbol{B}}^k\}} n_q \cdot \log P(X=q) + C_{para},$$

其中，$P(X=q)$ 表示概率分布函数 $\mathbf{P}$ 中取值为 q 的概率，C_{para} 表示 $\mathbf{P}$ 中参数的编码长度（如泊松分布中的均值）。

至于整体张量的剩余部分 $\bar{\boldsymbol{R}} = \mathcal{R} - \mathcal{B}^1$，考虑到数据分布的稀疏性和离散型元组值，这里采用哈夫曼编码策略来编码其中的元组，其编码长度为 L_ϵ。

将上述各项综合起来，利用检测到的 K 层层次化稠密子张量的指示向量集来表示整个张量 $\mathcal{R}$ 的编码长度为：

$$L(\mathcal{R};\mathbf{X}^1,\ldots,\mathbf{X}^K) = \log^* K + \sum_{n=1}^{N}\log^*|\mathcal{R}_n| + \sum_{k=1}^{K} L(\mathbf{X}^k) + L(\bar{\boldsymbol{B}}^k) + L_\epsilon. \tag{5.10}$$

为了得到最优的参数，可以用启发式的方法在可能的参数范围内进行网格搜索，选择最优的参数配置以最小化 MDL 。在后面的实验中会看到，由 MDL 原则选择出的参数对应于检测 HDS-tensors 的最优结果。

5.5.5 算法分析

本节中给出 CatchCore 算法的收敛性、时间和空间复杂度的理论分析。

如下的引理 5.1 给出了 CatchCore 中梯度迭代算法的收敛性质。

引理 5.1. CatchCore 算法结果收敛到某一驻点。

[1] $\log^* x$ 是对整数 x 的通用编码长度 [216]。

证明. 对于由公式 (5.3) 定义的子张量检测，CatchCore 的交替更新过程对应分块非线性 *Gauss-Seidel* 方法。

基于 Quasi-Convex 目标函数的收敛性结论 [212]，每一维度的指示向量被封闭的凸集所约束，OneWayOpt 算法中采用的 Armijo 线性搜索会对各维度生成一个点的系列，其中至少存在一个有限的点 $\tilde{\mathbf{x}}$，满足 $(\nabla_{\mathbf{x}} f)^T(\tilde{\mathbf{x}} - \mathbf{x}) \geq 0$，即 $\tilde{\mathbf{x}}$ 是一个关键点（驻点）。即使对目标函数是否为凸不做任何假设，这一收敛性仍然满足。

为了检测层次化子张量，CatchCore 按层更新各指示向量，作为该问题的子问题，在每一层中检测子张量时，仍然满足公式 (5.3)，其收敛性仍然保持。因此，CatchCore 算法最终会收敛到某一驻点。 ■

定理 5.2中描述了 CatchCore 算法的时间复杂度与参数 K，N 及 $\mathcal{R}$ 中的非零元素数 $nnz(\mathcal{R})$ 呈线性相关的关系。算法的空间复杂度由定理 5.3给出。

定理 5.2 (最坏情况下的时间复杂度). *令 t_{als} 表示 OneWayOpt 算法中为了更新某一指示向量所采用的 Armoji 线搜索中的迭代次数，CatchCore 算法最坏情况下的时间复杂度为：$O(K \cdot N \cdot t_{\max} \cdot t_{\text{als}} \cdot (nnz(\mathcal{R}) + c \cdot S_{\mathcal{R}}))$。*

证明. 对于计算 $M_{\mathbf{X}^k}$ 和 OneWayOpt 中的 $\nabla_{\mathbf{x}_n} f$，其复杂度都依赖于 full-mode 乘积 $\bar{\times}$ 或 $\bar{\times}_{(-n)}$，进行张量-向量之间元素的 n-mode 乘积操作的复杂度最多是 $O(nnz(\mathcal{R}))$；在对目标函数 f、各个向量梯度以及终止条件的计算中，包含对指示向量的模的计算，其复杂度为 $O(\sum_{n=1}^{N} |\mathcal{R}_n|)$。这样，对于 t_{als} 的搜索迭代中，OneWayOpt 的时间复杂度为：$O(t_{\text{als}} \cdot (nnz(\mathcal{R}) + c \cdot S_{\mathcal{R}}))$。

CatchCore 中所有变量的交替迭代更新的次数最多为 $t_{\max}$，所以 OneWayOpt 算法调用的总次数为 $K \cdot N \cdot t_{\max}$；在选择显著子张量时，K 层中的所有指示向量最多被检测一次，总次数为 $O(K \cdot \sum_{n=1}^{N} |\mathcal{R}_n|)$。因此，CatchCore 在最坏情况下的复杂度为：$O(K \cdot N \cdot t_{\max} \cdot t_{\text{als}} \cdot (nnz(\mathcal{R}) + c \cdot S_{\mathcal{R}}))$。 ■

定理 5.3 (内存空间需求). *CatchCore 算法运行中所需的内存空间大小为:*

$$O(nnz(\mathcal{R}) + 2K \cdot S_{\mathcal{R}}).$$

证明. CatchCore 用稀疏形式存储张量 $\mathcal{R}$；对指示向量集 $\mathbf{X}^k$，在计算 $M_{\mathbf{X}^k}$、$\nabla_{\mathbf{x}_n^k} f$ 时需要存储所有的指示向量和对应的梯度向量。因此，算法总的内存空间需求为 $O(nnz(\mathcal{R}) + 2K \cdot S_{\mathcal{R}})$。 ■

参数分析：显著层次的最大值 $K_{\max} = \log_{\eta}(\frac{\max(\mathcal{R})}{\rho_{\mathcal{R}}})$，其中 $\max(\mathcal{R})$ 表示 $\mathcal{R}$ 中度量属性 C 的最大值。此外，通常有下面的结论保证 CatchCore 算法的高效性。

观察 4. 在实际的数据分析中，相关参数满足：

- $nnz(\mathcal{R}) \gg S_{\mathcal{R}}$;
- $K \ll K_{\max}$，即密度差异显著层次的数目较少;
- $t < t_{\max}$, 即迭代可以提前终止;
- t_{als} 较小，意味着搜索一个恰当的更新步长仅需要少数几步迭代即可.

同时，在并行环境下算法 Line 5 表示各个维度在更新时是可以独立求解的。

5.6 实验验证与分析

本文设计完成了不同实验以回答下面的问题：

- *Q1.* 准确性: CatchCore 算法在合成数据和真实数据中检测层次化稠密子张量的准确率如何？基于 MDL 的评价指标是否选择到了最优的参数配置？
- *Q2.* 模式识别与异常检测: CatchCore 在真实数据集中找到了哪些有意义的模式？检测到的异常具有怎样的行为模式？
- *Q3.* 可扩展性: CatchCore 算法对输入数据大小是线性可扩展的吗？

5.6.1 实验设置

5.6.1.1 对比方法

本文选择多个当前最先进的、用于稠密子块检测的方法作为基线，包括 D-Cube [92]、M-Zoom [90]、CrossSpot [89] 和 CP Decomposition (CPD) [86]。在所有实验中，为了有效利用内存空间，数据均以稀疏张量的形式存储；对于 D-Cube 和 M-Zoom 采用 ρ_{ari} 和 ρ_{geo} 密度指标；CrossSpot 中也改用上述指标，并用 CP Decomposition (CPD) 的结果作为其初始种子节点的选取方式 [90]。

5.6.1.2 数据集

表 5.3 中总结了实验中所用的真实数据的统计信息。在 *Rating* 类别中，数据构成 4-way 张量，表示 *(user ,item, timestamp, rating)* 关系，元组值为评论的次数；在 *Social network* 类别中，构成 3-way 张量，表示 *(user, user, timestamps)* 关系，元组值为交互（添加收藏夹或者共同合作者关系）的次数；*DARPA* 是表示 TCP DUMP 的 3-way 张量，表示关系为 *(source IP, target IP, timestamps)*, 元组值为两

表 5.3 真实数据集的统计信息

Table 5.3 Summary of real-world datasets.

Name	Size	$card(\mathcal{R})$	$nnz(\mathcal{R})$
Ratings: users × items × timestamps × rating → #reviews			
Android [217]	1.32M × 61.3K × 1.28K × 5	1.38M	2.23M
BeerAdvocate [156]	26.5K × 50.8K × 1,472 × 3	78.7K	1.07M
StackOverflow [217]	545K × 96.7K × 1,154 × 1	643K	1.30M
Social network: users × users × timestamps → #interactions			
DBLP [218]	1.31M × 1.31M × 72	2.63M	18.9M
Youtube [199]	3.22M × 3.22M × 203	6.45M	18.5M
TCP dumps: IPs × IPs × timestamps → #connections			
DARPA [219]	9.48K × 23.4K × 46.6K	79.5K	522K
TCP dumps: duration × protocol × service × ⋯ → #connections			
AirForce [220]	3 × 70 × 11 × 7.20K × 21.5K × 512 × 512	39.7K	863K

个 IP (主机) 之间建立的连接的数目；*AirForce* 是由 U.S Air Force LAN 收集的网络入侵日志数据集，构成了 7-way 张量，其中的属性分别为：*(protocol, service, src_bytes, dst_bytes, flag, host_count, src_count, #connections)*。对于时间为属性，DARPA 数据的时间戳单位是分钟，在 Rating 和 Youtube 数据中时间戳以天为单位，在 DBLP 中以年为单位。

5.6.2 Q1. 准确性验证

实验中比较了不同方法在合成数据和真实数据中检测注入稠密块的准确性，这里采用 F score $= \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$ 作为准确度衡量指标。

对于合成数据，以随机均匀分布的采样方式构造了一个大小为 $5K \times 5K \times 2K$、密度为 $3 \cdot 10^{-6}$ 的 3-way 张量 $\mathcal{R}$，并在其中注入了一个大小为 $200 \times 300 \times 100$、密度不同的稠密块，用于测试各个算法检测到的稠密块的密度下界。从图 5.1c 所示的结果看出，对于检测出稠密块的密度，CatchCore 算法检测的结果是其他对比方法检测结果最低密度的 $\frac{1}{10}$，这意味着通过 CatchCore 的有效检测能使那

表 5.4 BeerAdvocate 数据集中注入的层次化稠密子张量的检测结果准确率

Table 5.4 Hierarchical subtensors detection results for BeerAdvocate dataset.

K	Injected Densities	H1				H2					H3				
		CC	D/M	CS	CPD	**CC**	D	M	CS	CPD	**CC**	D	M	CS	CPD
2	0.05 + 0.01	**1**	1	1	0.999	**1**	0.121	0.041	0	0.080					
	0.1 + 0.05	**1**	0	0.185	0.185	**1**	1	1	1	0.999					
	0.2 + 0.05	**1**	0	0.986	0.986	**1**	1	1	1	0.998			--		
	0.2 + 0.1	**1**	0	0.236	0.236	**1**	1	1	1	0.997					
	0.7 + 0.1	**1**	0	0.325	0.325	**1**	1	1	1	0.519					
	1 + 0.05	**1**	0	0.784	0.784	**1**	1	1	1	0.998					
	1 + 0.1	**1**	0	0.751	0.751	**1**	1	1	1	0.998					
	1 + 0.2	**1**	0	0.795	0.795	**1**	1	1	1	0.999					
3	0.2 + 0.1 + 0.05	**1**	0	0.265	0.265	**1**	0	0	0	0	**1**	1	1	1	0.980
	1 + 0.2 + 0.01	**1**	0	0	0	**1**	0.999	1.0	1.0	0.223	**1**	0.444	0.431	0.746	0.850
	1 + 0.7 + 0.2	**1**	0	0	0	**1**	0	0	0.981	0.981	**1**	1	1	1	0.999

* 算法名称简写：CC: CatchCore， D: D-Cube， M: M-Zoom， CS: CrossSpot 。

* 不同注入子张量的大小及对应密度值: 0.01: $1K \times 800 \times 15$, 0.05: $800 \times 600 \times 10$, 0.1: $500 \times 500 \times 5$, 0.2: $300 \times 300 \times 5$, 0.7: $200 \times 100 \times 2$, 1: $100 \times 80 \times 1$。

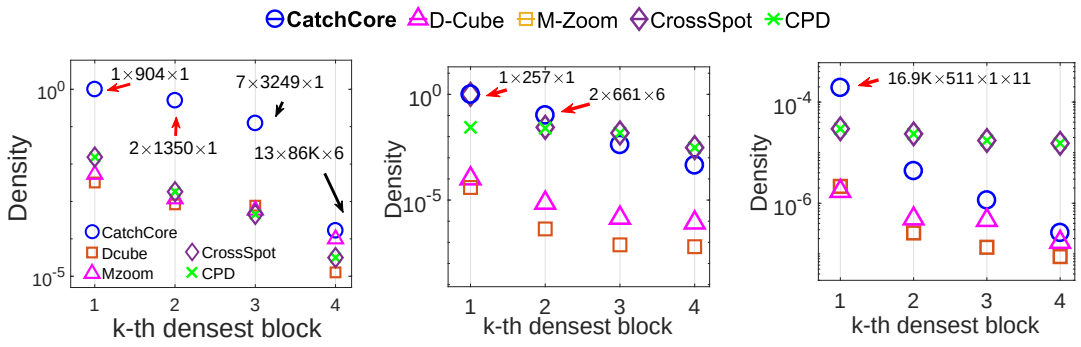
表 5.5 合成数据集中注入的层次化稠密子张量的检测结果准确率

Table 5.5 Accuracy of hierarchical subtensors detection for Synthetic dataset.

K	Injected Densities	H1				H2				H3				H4			
		CC	D/M	CS	CPD	**CC**	D/M	CS	CPD	**CC**	D/M	CS	CPD	**CC**	D/M	CS	CPD
2	0.01 + 0.001	1	1	0.14	0.14	1	0.183	0.89	0.89								
	0.1 + 0.01	1	1	0.25	0.25	1	1	0.89	0.89			--					
	0.25 + 0.1	1	1	0.35	0.35	1	0.257	1	1							--	
3	0.1 + 0.01 + 0.001	1	1	0.17	0.17	1	1	0.20	0.20	1	0.321	0.74	0.87				
	0.25 + 0.1 + 0.01	1	0	0.98	0.98	1	1	0.20	0.20	1	1	0.85	0.98				
4	0.25 + 0.1 + 0.01 + 0.001	1	0	0.96	0.96	1	1	0.51	0.51	1	1	0.19	0.19	1	0.359	0.79	0.95

* 算法名称简写：CC: CatchCore， D: D-Cube， M: M-Zoom， CS: CrossSpot 。

* 对应不同注入密度的子张量大小: 0.001: $200 \times 250 \times 100$, 0.01: $200 \times 100 \times 60$, 0.1:$80 \times 80 \times 30$, 0.25: $50 \times 60 \times 10$。



(a) Youtube (b) StackOverflow (c) Android

图 5.3 CatchCore 在真实数据中检测到的层次化稠密子张量。(a)-(c) 分别对应于 Youtube、StackOverflow、Android 数据中密度为前 4 的检测结果，CatchCore 的检测效果优于其他对比方法。(a) 中密度前 2 的子块中包含了可疑性行为，表示一个用户在一天的时间内创建了大量的好友关系。

Figure 5.3 The subtensors detected by CatchCore achieve higher densities. In each plot, points represent the density of k-th of top 4 blocks found by the methods. And the size of some blocks are labeled in text. CatchCore catches suspicious patterns for the datasets.

些欺诈者需要付出更大作弊成本，因此具有更强的反作弊能力。

在 BeerAdvocate 和合成数据中，以层次化的方式注入了不同大小、不同密度、不同 K 值的 K 层稠密子块。表 5.4 中列出了在真实数据上各算法的检测结果，其中 H1 表示密度最大或者第一个由某个算法检测到的子块，从 H1 到 H3 的子张量的密度递减或者被检测到的顺序递增，表格下方列出了不同注入块的大小和密度信息。从结果中可以看出，CatchCore 能够准确地检测所有注入的在不同层次、不同大小和密度的稠密子张量，其结果一致地超过其他对比方法——它们检测准确率较低或者遗漏了至少一个稠密块。表中除了有颜色标注的几个示例外，D-Cube 和 M-Zoom 算法有相似的准确性，它们不能识别稠密块之间的依赖结构，使得某些最稀疏或者最稠密的注入块被遗漏或者被其他体积更大的块所淹没。CrossSpot 和 CP Decomposition (CPD) 也不能准确地找到层次化的稠密块。在如表 5.5 所示的合成数据上的检测结果呈现类似的结论。

此外，采用不同的注入策略和注入块的形态变化的实验结果也表明，CatchCore 能够准确地检测一些非重叠的或部分重合的稠密子块，或者一些有空洞 (hollow) 的子张量的稠密部分（基于张量维度属性的可交换性，此时将转换为部分重合的情况）。

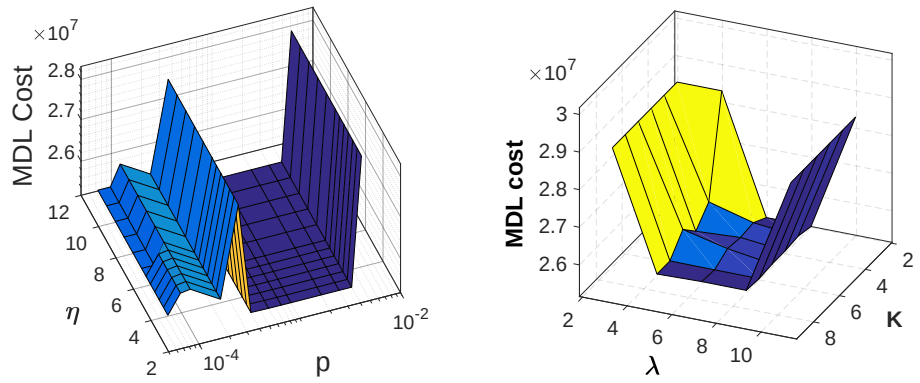


图 5.4 CatchCore 算法的参数敏感性分析，算法检测结果的 MDL 代价随着参数 p、η、λ、K 的不同取值的变化情况。最优检测结果的 MDL Cost 最小，CatchCore 算法在较大的参数取值范围内都能找到最优的层次化稠密子张量。

Figure 5.4 The parameter sensitivity analysis for CatchCore. The optimal hierarchies achieve the lowest MDL cost w.r.t the parameter p、η、λ、K. CatchCore can obtain optimal hierarchical dense subtensors for wide parameters range to some extend

真实数据中的稠密块：这里利用 CatchCore 检测其他真实数据中存在的模式。由于一些密度更大的子块通常大概率地包含着某种有趣的模式，所以这里衡量检测到稠密块的密度值而非其质量，以避免出现一些平凡解的结果。图 5.1d 和 图5.3 给出了不同数据集中检测到的密度前 4 层的检测结果，从中可以明显地看到，CatchCore 能够识别出各数据中密度更大的块，且性能一致性优于其他对比算法，稠密块所包含的模式在后续部分的分析中给出。

基于 MDL 的结果评价：这里通过衡量不同参数设置下由 CatchCore 检测结果的质量来评估 MDL 度量指标的有效性。该实验中，在 BeerAdvocate 数据中注入了 3 层层次化稠密子张量 (即表 5.4 中 $K = 3$ 时第 2 行的数据)，比较了参数 p 和 η 在一定范围内变化时检测结果的 MDL 代价，其结果如图 5.4 所示。对于不同的 λ、K，p 和 η，最优层次的检测结果的 MDL 代价最小，并且参数在一个较大范围内变化时，算法都能得到最优的结果。

5.6.3 Q2. 模式识别与异常检测验证

5.6.3.1 异常检测

在 TCP DUMP 应用中，CatchCore 以高准确率识别带标注的 DARPA 和 Air-Force 数据集中的网络入侵行为。表 5.6 比较了不同算法的检测准确率 (F score)。在 DARPA 数据集上，CatchCore 的结果超过其他对比方法，并在 H1 - H3 层中

表 5.6 CatchCore 在真实的 TCP DUMPs 数据中准确的识别其中的网络攻击。以 F1 作为准确度指标，左侧表内 CatchCore 的检测结果最好或者与现有最好方法的效果近似可比，右侧表内列出了在 DARPA 数据集上检测出了层次化稠密子张量信息。

Table 5.6 CatchCore identifies network attacks from the real-world TCP Dumps dataset with best or comparable performance in F score (Left); it spots different hierarchical dense blocks with interesting patterns for DARPA (Right).

	DARPA	AirForce
CPD	0.167	0.785
CrossSpot	0.822	0.785
M-Zoom	0.826	0.906
D-Cube	0.856	**0.940**
CatchCore	**0.877**	0.902

H	块大小	异常比例
1	$1 \times 1 \times 96$	100%
2	$1 \times 1 \times 100$	100%
3	$1 \times 1 \times 274$	100%
4	$16 \times\ 5\ \times 24.7K$	87.0%
5	$171 \times 15 \times 29.2K$	85.4%

检测到 Neptune 攻击的层次化行为模式，即一对 IP 之间在不同时间创建的连接。图 5.5 显示了 1998 年 6 月 18 日上午 7 时到 8 时之间对应的一个攻击模式片段，其密度 (即每分钟的攻击强度) 在不同层之间存在很大差别，其中 H1 的密度超过 5K，而 H3 的其他部分密度约 3K 左右；同时也可以看到这类攻击构成以 5 分钟为单位的周期行为。在 AirForce 数据上 CatchCore 的检测效果与其他方法近似可比；因为在密度较低的层次内，CatchCore 检测到的稠密子块虽然包含了更多的异常连接 (其召回率为 98%)，但同时也包含了一些伪正例样本。

在其他真实数据中，CatchCore 也找到了一些密度更大、可疑性更高的子块。如图 5.3a 所示的 Youtube 数据集上的检测结果中，这些由 CatchCore 找到的稠密块被其他对比方法所遗漏，而其中最密的子块 (H1) 表示有一个用户在一天之内与另外的 904 个用户成为好友关系，而 H2 中包含的另一个用户则在相同时间内与另外的 475 个用户成为好友，因此，这两个用户的行为很大程度上是一种欺诈或者机器人账号导致的；如图 5.3b 所示，在 StackOverflow 数据集中的最密子块表明一个用户在一天之内收藏了 257 条消息帖子，这也是一种非正常的行为；对于 Android 数据集，CatchCore 检测到的第一个稠密子块的密度是其他方法检测结果的 6 倍。

此外，CatchCore 在检测过程中优化的目标函数是由多层稠密子张量构成

的整体，而其中的最密子块只是一个部分而非直接目标，在体积密度指标下会趋向于选择非空的稠密块，因此可能导致在某些结果中最密的层次是 1 维的张量切片 (但可能并不是在整个张量中最密的切片)，如果选用其他不同的密度度量可以缓解这种类似的问题。

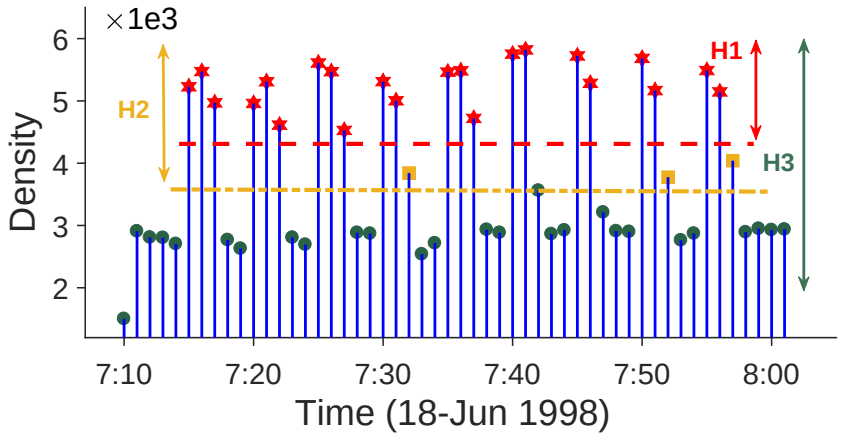
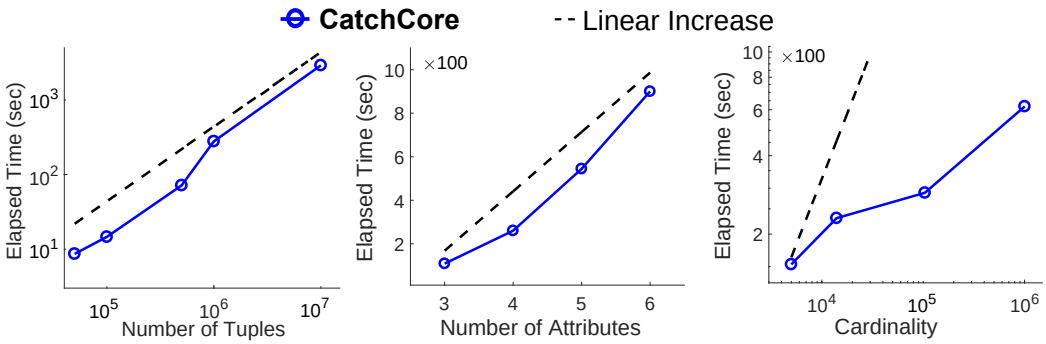


图 5.5 CatchCore 异常模式检测结果：在 TCP DUMP 数据中检测到的异常模式片段。1998 年 6 月 18 日的 DARPA 数据中，一对 IP 之间存在的层次化网络入侵攻击行为模式。

Figure 5.5 The network attack snapshot of TCP DUMP found by CatchCore. Hierarchical network intrusion behavior between a pair of IPs on June 18, 1998 in DARPA dataset.

5.6.3.2 演变的社区模式

如图 5.1d，5.1b 中显示了在 DBLP 中由前 4 层的稠密张量构成的层次化演化共同合作者的社区结构，其中包含了 20 位来自不同国家的科研人员在 2007 年到 2013 年之间的论文合作关系，图 5.1d 中给出了不同层次的密度值。H1 表示的最密子块大小为 8×8×3，是由 Leonard Barolli，Makoto Takizawa 和 Fatos Xhafa 等在 2011 年到 2013 年在“算法及分布式系统”领域的合作关系构成，每年的平均合作连接数超过 10.7 次 (即每年共同合作的文章数约 11 篇之多)，形成以了一个高密度合作团体；其他层次的子张量包括其他的共同合作者 (包括他们的学生等) 和不同的年份，内部的连接数相对少于 H1 层，最外层 H4 的稠密块的密度也超过 2。其他对比方法并没能成功检测出 H1 - H4 所代表的稠密子张量。因此，CatchCore 适合用于以层次化的方式在不同尺度上检测演变的社区结构。



(a) 张量中元素数目 ($nnz(\mathcal{R})$) **(b) 张量属性数 (N)** **(c) 张量属性的势 ($V_{\mathcal{R}}$)**

(a) Number of Tuples ($nnz(\mathcal{R})$) **(b) Number of Attributes (N)** **(c) Cardinalty of Attributes ($V_{\mathcal{R}}$)**

图 5.6 CatchCore 算法的可扩展性。(a)-(b) 展示了算法复杂度与张量 $\mathcal{R}$ 中的元组数目、属性数目（维度）呈线性关系，(c) 说明其复杂度与张量 $\mathcal{R}$ 的势成压线性关系。

Figure 5.6 CatchCore is scalable. (a)-(b) CatchCore scales linearly with the number of tuples and the number of attributes of $\mathcal{R}$. (c) CatchCore scales sub-linearly with the cardinalities of attributes of $\mathcal{R}$.

5.6.4 Q3. 可扩展性检验

本小节通过实验证明 CatchCore 算法对输入数据的各个大小度量呈线性或亚线性复杂度，即元组数目、维度数以及维度属性的势 (张量势)。

为了度量各个因子对可扩展性的影响，实验在随机生成的张量 $\mathcal{R}$ 中检测注入的两层不同大小、不同密度的层次化稠密子张量：生成的 3-way 张量 $\mathcal{R}$ 包含的 100 万元组，其势为 10 万；两个注入块大小分别为 $100 \times 100 \times 2$ 和 $50 \times 50 \times 1$，对应的密度分别为 0.2 和 1.0；然后，每次改变其中一个度量因子并记录算法检测时的总运行时间。如图 5.6 所示，可以看到 CatchCore 的运行时间与元组数目和维度数呈线性比例关系，与维度属性的势呈亚线性比例关系；由此说明，实际算法的可扩展性比定理 5.2 所述的时间复杂度理论分析结论会更好。

5.7 本章小结

在本章中考察了多属性图中具有多维稠密的聚集模式以及具有层次性的群体异常行为的检测，以解决关联数据的多元异构形式以及异常模式的结构复杂性问题。利用张量建模多维关联数据，本章提出一种用于稠密子张量检测的可解释性统一框架 CatchCore，根据具有统一表示的稠密子张量密度指标和形式化检测问题，CatchCore 算法通过层次交替梯度优化的方式高效地检测在连

接结构和密度上具有层次嵌套的稠密子张量模式。在大规模真实数据中验证了CatchCore算法的性能，并多种实际应用中识别出有趣的、有可解释性的群体异常行为模式。因此，本文的主要贡献在于：

- **统一的密度度量与算法：**设计了一种能够通过梯度迭代更新方式进行优化的统一密度度量指标来检测稠密子张量，提出了具有收敛保证的CatchCore算法以检测层次化稠密核结构。
- **准确性：** CatchCore能够更加准确地检测最密子块以及层次化稠密核，在合成数据及真实数据上具有最优的性能。
- **有效性：** CatchCore在真实多维数据中地发现了多种异常模式，如可疑的好友关系连接、带周期性的网络入侵攻击以及合作密切的研究团体等。
- **可扩展性：** CatchCore是高可扩展的，其运行时间与张量的各方面指标呈线性正比关系。

可复现性：本文提供了开源的程序（https://github.com/wenchieh/catchcore）并且实验中的真实数据均在线可获得。

第 6 章　总结与展望

本文研究了大规模图数据中的聚集性群体异常模式的挖掘。在这一章中，对全文内容进行总结，并对未来研究工作进行展望。

6.1　研究工作总结

随着互联网、物联网、社交网络、电子商务等相关应用的发展，产生和积累了海量的关联数据，其中的数据对象之间具有相互依赖、长距离相关且包含很多富信息等复杂特征，由此带来了对此类图式数据高效处理、潜在知识有效挖掘的巨大需求。在金融、安全、电信、网络安全、社交媒体等诸多领域及应用中，群体性异常模式挖掘和行为数据分析成为核心需求和重要手段，能够为企业带来可观的经济效益的同时，也深刻影响着个人、社会和国家的诸多方面。

当前，大规模图数据中群体异常模式检测的研究仍需要解决的一些重要挑战包括：(1) 离群点及群体异常的分布多样性; (2) 图拓扑结构中聚集异常的度量统一性; (3) 多属性图中群体异常的结构复杂性。因此，本文重点研究大规模图数据中具有聚集性特征的群体异常模式，分析在不同数据形态表示下群体异常的复杂特性和行为表现，提出了快速有效、可扩展的检测方法，并应用于大规模实际应用中不同有趣模式的发现，并在真实场景中用于识别具有同步性和密集性特征的共谋欺诈、恶意攻击、学术合作等群体异常行为。

首先，本文提出了图特征空间中基于视觉引导的聚集性群体异常检测方法：研究图式关联数据的复杂连接在欧式空间中的表现以及聚集性群体异常行为的特征。基于节点特征的大规模图数据表示，文中通过特征组合方式构建相关性直方图映射，以反映在欧式空间下结构化图数据中节点的分布表现和聚集性质，在具有广泛适用性的直方图中，分析了以微簇结构为代表的异常模式及其对应的特殊图结构 (如团、社区及星型结构) 和隐含行为模式 (共谋欺诈、协作合作) 等方面的群体异常。针对直方图中微簇分布的复杂性 (多类簇混合下具有不同区分度的复杂分布)，文中提出了基于视觉引导的检测算法 EagleMine 来识别和总结其中的节点组、并检测异常微簇。EagleMine 以层次化的方式检测直方图中出现的类簇结构，并利用具有多层分辨率感知的 water-level tree 树结构来组织和管理

识别结果，利用基于统计分布的词典模型对各类簇进行总结，以统计假设检验为标准对树进行探索和检测，并最终选择出能够简洁描述直方图的最优类簇组合；通过对类簇可疑性的度量，确定其中的异常微簇结构。在大规模数据上的实验表明，该分析方法能够解决具有不同区分度的多类簇聚类识别问题，算法能够识别到与人眼视觉感知预期相一致的类簇，对整个直方图的总结能够得到简洁的描述；在注入实验及微博数据的群体异常检测实验中，验证了 EagleMine 在图特征表示下的检测性能优于基于图的检测方法；此外，EagleMine 算法还可以用于时序转发事件中的同步性转发异常模式的检测。

第二，本文提出了针对拓扑关联密集异常的统一图谱检测算法：探究复杂关联数据中异常定义的问题多样性。本文研究了图数据在不同目标背景中以稠密子图为对象的拓扑密集关联及相关的群体异常行为模式的高效检测。基于对许多实际场景中与子图模式检测相关问题的对比和总结，文中提出了一种广义最密子图检测的统一形式化定义 (GenDS)，从图谱理论角度分析了该问题的优化性质，利用大规模图的谱分布性质与幂率特征，结合贪心策略设计了一种简单、高效的检测算法 SpecGreedy 来解决 GenDS 问题，并在来自不同领域的 40 个真实图数据上进行了广泛的性能评估与验证。实验结果表明，对比与其他基线方法，SpecGreedy 算法可将最密子图的检测速度提高 58.6× 倍且得到具有密度更大或者近似相同的子图结果；而且 SpecGreedy 算法是随着图的大小线性可扩展的；此外，在实际数据中验证了算法检测群体异常模式有效性，即在大规模随时间变化的学术共同作者网络中发现了突现的团结构合作模式。

最后，本文提出了多属性图中层次结构感知的群体异常模式发现：建模具有多元异构形式的多维关联数据，探索具有复杂结构的异常模式，即多属性图背景下的多维稠密聚集模式以及具有层次性的群体异常模式的检测。在张量作为多维关联数据的建模表示下，本文提出一种新的框架 CatchCore，能够对稠密子张量进行表示和高效计算，并有效地发现在连接结构和密度上具有层次性分布的稠密模式。针对稠密子张量检测问题，文中提出了一种可以通过梯度优化更新方式计算的、囊括现大多数密度衡量指标的统一密度形式化表示，并以此为基础，进一步提出了基于层次交替优化的更新算法求解层次性稠密子张量，且算法的收敛性和可扩展性有理论保证；此外，由最小描述长度准则给出了检测结果质量的量化评价指标，并可用于确定层次化稠密子块检测的最优的参数配置。在大

规模真实数据上的实验结果表明，CatchCore 在稠密子张量检测和异常模式识别问题中的性能都优于当前最优的对比方法；同时，CatchCore 能够更加准确地识别出具有可解释性的有趣模式，包括可疑的好友关系连接、带周期性的网络入侵攻击模式以及 DBLP 中具有层次性紧密合作行为的科研学术团体等群体异常；此外，CatchCore 的复杂度与张量的各方面指标呈线性相关，保证了快速、可扩展的检测性能。

6.2 研究工作展望

本文主要针对大规模图数据背景下具有聚集性的群体异常的检测，分析其在不同数据形态下的复杂特性和行为表现，设计具有可扩展性的检测算法，并应用与实际场景中异常模式的分析和解释，并取得了一定的成果。作者认为，未来需要进一步研究的内容包括：

1. *基于复杂数据形态下的多样异常模式探索与挖掘*。本文主要研究静态关联数据中稠密模式的表现特征和有效挖掘，然而，多数实际场景中的数据具有增量、动态、流式的性质，因此需要对应检测算法具有实时响应的性质，如传感器网络、在线交易监控等，其中涉及到对局部子图结构变化的及时度量和准确检测、在此场景下有效地融合具有不完全可靠标签以进行异常识别等实际问题；此外，对于动态属性图中异常的定义和识别、实际中对应的场景设计和应用分析等仍然是一个开放性的探索领域。
2. *基于行为特征的规律发现、建模与异常模式挖掘*。本文针对关联数据在不同形态下以密集性为特征的异常进行了分析，而实际行为数据中包含了各种形式的特征和模式：社交媒体真实应用中多元耦合关系、多模信息内容使得用户行为具有动态性、多面性；物联网、移动互联网中设备异构、关联噪声、信息动态异质使得行为刻画具有复杂性、随机性等。因此，这些因素和场景导致行为表现具有更加独特的性质，其潜在规律也会随着环境、状态等因素的改变而变化；如何深度挖掘行为的复杂特征、分析和识别潜在的产生规律，对于行为的刻画、建模和预测具有重要意义；此外，复杂数据下群体性异常模式检测、解释和应用拓展也需要对其原理有深刻的认识；同时政府、社会、企业也需要基于行为分析为依据的决策支持和解决方案设计。

3. 智能认知系统中的异常诊断和检测。现在的异常检测方法都侧重于从数据角度出发，定义、分析其中的模式并设计相应的检测算法。随着机器学习、深度学习成为数据建模的有力工具并得到众多可用的各类模型，它们大多数用于常规的知识学习、表示和应用，而训练数据中异常实例对于模型的稳定性、有效性具有重要影响；因此以模型为对象的异常检测和评估将变得愈加重要，其中包括健壮学习模型的设计、模型本身的鲁棒性验证、模型表现异常性的追踪等；此外，确定模式是否被异常数据所污染以及被污染模型的修正等都是在数据匿名和隐私保护场景下需要解决的重要问题。
4. 具有集成性的关联数据分析系统与异常检测工具。以关联数据为对象的网络数据科学需要提供一套具有一定通用性的图化表示、分析和挖掘系统，结合大数据分析工具，为实际应用给出有效的解决方案，同时为基础数据分析和问题发现提供有力支持。针对异常应用的广泛性，在多形态数据表示下，通过高效地集成具有可扩展性的分析算法，构建复合型检测工具实现使用目标场景的综合数据分析和规律发现，为决策支持提供简明扼要、易于实施的系统方案。

参考文献

[1] Yang Y, Xu Y, Sun Y, et al. Mining fraudsters and fraudulent strategies in large-scale mobile social networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2019.

[2] CAO Q, YANG X, YU J, et al. Uncovering large groups of active malicious accounts in online social networks[C]//Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. 2014: 477-488.

[3] MILLER G L. 15-859n: spectral graph theory with applications to ml, spring 2020[EB/OL]. 2020. http://www.cs.cmu.edu/afs/cs/academic/class/15859n-s20.

[4] CHANDOLA V, BANERJEE A, KUMAR V. Anomaly detection: A survey[J]. ACM Comput. Surv., 2009, 41:15:1-15:58.

[5] HAWKINS D M. Identification of outliers: volume 11[M]. Springer, 1980.

[6] CHALAPATHY R, CHAWLA S. Deep learning for anomaly detection: A survey[J]. arXiv preprint arXiv:1901.03407, 2019.

[7] AHMED M, NASER MAHMOOD A, HU J. A survey of network anomaly detection techniques[J/OL]. J. Netw. Comput. Appl., 2016, 60(C):19–31. https://doi.org/10.1016/j.jnca.2015.11.016.

[8] IBIDUNMOYE O, HERNáNDEZ-RODRIGUEZ F, ELMROTH E. Performance anomaly detection and bottleneck identification[J/OL]. ACM Comput. Surv., 2015, 48(1). https://doi.org/10.1145/2791120.

[9] GOLDSTEIN M, UCHIDA S. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data[J/OL]. PLOS ONE, 2016, 11(4):1-31. https://doi.org/10.1371/journal.pone.0152173.

[10] SALEHI M, RASHIDI L. A survey on anomaly detection in evolving data: [with application to forest fire risk prediction][J]. ACM SIGKDD Explorations Newsletter, 2018, 20(1):13-23.

[11] ABE N, ZADROZNY B, LANGFORD J. Outlier detection by active learning[C]// Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006: 504-509.

[12] ANGIULLI F, PIZZUTI C. Outlier detection for high dimensional data[C]//2002.

[13] WANG Y, PARTHASARATHY S, TATIKONDA S. Locality sensitive outlier detection: A ranking driven approach[C]//2011 IEEE 27th International Conference on Data Engineering. IEEE, 2011: 410-421.

[14] BREUNIG M M, KRIEGEL H P, NG R T, et al. Lof: identifying density-based local outliers

[C]//Proceedings of the 2000 ACM SIGMOD international conference on Management of data. 2000: 93-104.

[15] PAPADIMITRIOU S, KITAGAWA H, GIBBONS P B, et al. Loci: Fast outlier detection using the local correlation integral[C]//Proceedings 19th international conference on data engineering (Cat. No. 03CH37405). IEEE, 2003: 315-326.

[16] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise.[C]//KDD. 1996.

[17] HE Z, XU X, DENG S. Discovering cluster-based local outliers[J]. Pattern Recognition Letters, 2003, 24(9-10):1641-1650.

[18] ŠALTENIS V. Outlier detection based on the distribution of distances between data points [J]. Informatica, 2004, 15(3):399-410.

[19] ANDO S. Clustering needles in a haystack: An information theoretic analysis of minority and outlier detection[C]//Seventh IEEE International Conference on Data Mining (ICDM 2007). IEEE, 2007: 13-22.

[20] LIU B, XIAO Y, CAO L, et al. Svdd-based outlier detection on uncertain data[J]. Knowledge and information systems, 2013, 34(3):597-618.

[21] KELLER F, MULLER E, BOHM K. Hics: High contrast subspaces for density-based outlier ranking[C]//2012 IEEE 28th international conference on data engineering. IEEE, 2012: 1037-1048.

[22] CAMPOS G O, ZIMEK A, SANDER J, et al. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study[J]. Data Mining and Knowledge Discovery, 2016, 30(4):891-927.

[23] ELKAN C. The foundations of cost-sensitive learning[C]//International joint conference on artificial intelligence: volume 17. Lawrence Erlbaum Associates Ltd, 2001: 973-978.

[24] BARABÁSI A L, ALBERT R. Emergence of scaling in random networks[J]. science, 1999, 286(5439):509-512.

[25] CHAKRABARTI D, ZHAN Y, FALOUTSOS C. R-mat: A recursive model for graph mining [C]//SDM. 2004.

[26] AKOGLU L, FALOUTSOS C. Rtg: a recursive realistic graph generator using random typing[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2009: 13-28.

[27] LESKOVEC J, CHAKRABARTI D, KLEINBERG J, et al. Kronecker graphs: An approach to modeling networks[J]. JMLR, 2010, 11(Feb):985-1042.

[28] AKOGLU L, TONG H, KOUTRA D. Graph based anomaly detection and description: a survey[J]. Data mining and knowledge discovery, 2015, 29(3):626-688.

[29] RANSHOUS S, SHEN S, KOUTRA D, et al. Anomaly detection in dynamic networks: a survey[J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2015, 7(3):223-247.

[30] BHUYAN M H, BHATTACHARYYA D K, KALITA J K. Network anomaly detection: methods, systems and tools[J]. Ieee communications surveys & tutorials, 2013, 16(1):303-336.

[31] PHUA C, LEE V, SMITH K, et al. A comprehensive survey of data mining-based fraud detection research[J]. arXiv preprint arXiv:1009.6119, 2010.

[32] KWON D, KIM H, KIM J, et al. A survey of deep learning-based network anomaly detection [J]. Cluster Computing, 2017:1-13.

[33] PAPADOPOULOS S, KOMPATSIARIS Y, VAKALI A, et al. Community detection in social media[J]. Data Mining and Knowledge Discovery, 2011, 24:515-554.

[34] Chintalapudi S R, Prasad M H M K. A survey on community detection algorithms in large scale real world networks[C]//2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom). 2015: 1323-1327.

[35] LANCICHINETTI A, FORTUNATO S. Community detection algorithms: a comparative analysis[J]. Physical review E, 2009, 80(5):056117.

[36] CAVALLARI S, ZHENG V W, CAI H, et al. Learning community embedding with community detection and node embedding on graphs[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017: 377-386.

[37] CHEN P Y, HERO A O. Deep community detection[J]. IEEE Transactions on Signal Processing, 2015, 63(21):5706-5719.

[38] YANG L, CAO X, HE D, et al. Modularity based community detection with deep learning. [C]//IJCAI: volume 16. 2016: 2252-2258.

[39] LEE V E, RUAN N, JIN R, et al. A survey of algorithms for dense subgraph discovery[C]// Managing and Mining Graph Data. 2010.

[40] SHI C, CAI Y, PHILIP S Y, et al. A comparison of objective functions in network community detection[C]//2010 IEEE International Conference on Data Mining Workshops. IEEE, 2010: 1234-1241.

[41] TONG H, FALOUTSOS C. Center-piece subgraphs: problem definition and fast solutions [C]//Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006: 404-413.

[42] CHAU D H, AKOGLU L, VREEKEN J, et al. Tourviz: interactive visualization of connection pathways in large graphs[C]//Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. 2012: 1516-1519.

[43] LIU Y, SAFAVI T, DIGHE A, et al. Graph summarization methods and applications: A survey[J]. ACM Comput. Surv., 2016, 51:62:1-62:34.

[44] KOUTRA D, KANG U, VREEKEN J, et al. Vog: Summarizing and understanding large graphs[C]//Proceedings of the 2014 SIAM international conference on data mining. SIAM, 2014: 91-99.

[45] JIANG M, CUI P, BEUTEL A, et al. Inferring strange behavior from connectivity pattern in social networks[C]//PAKDD. 2014.

[46] JIANG M, CUI P, BEUTEL A, et al. Catchsync: catching synchronized behavior in large directed graphs[C]//SIGKDD. 2014.

[47] BONACICH P, LLOYD P. Eigenvector-like measures of centrality for asymmetric relations [J]. Social networks, 2001, 23(3):191-201.

[48] HENDERSON K, GALLAGHER B, ELIASSI-RAD T, et al. Rolx: structural role extraction & mining in large graphs[C]//Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. 2012: 1231-1239.

[49] AKOGLU L, MCGLOHON M, FALOUTSOS C. Oddball: spotting anomalies in weighted graphs[C]//PAKDD.

[50] NEWMAN M E. Modularity and community structure in networks[J]. Proceedings of the national academy of sciences, 2006, 103(23):8577-8582.

[51] ANDERSEN R, CHUNG F, LANG K. Local graph partitioning using pagerank vectors[C]// 2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06). IEEE, 2006: 475-486.

[52] KANG U, MCGLOHON M, AKOGLU L, et al. Patterns on the connected components of terabyte-scale graphs[C]//2010 IEEE International Conference on Data Mining. IEEE, 2010: 875-880.

[53] IDÉ T, KASHIMA H. Eigenspace-based anomaly detection in computer systems[C]// Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 2004: 440-449.

[54] PRAKASH B A, SRIDHARAN A, SESHADRI M, et al. Eigenspokes: Surprising patterns and scalable community chipping in large graphs[C]//PAKDD. 2010.

[55] SHIN K, ELIASSI-RAD T, FALOUTSOS C. Corescope: Graph mining using k-core analysis —patterns, anomalies and algorithms[J]. 2016 IEEE 16th International Conference on Data Mining (ICDM), 2016:469-478.

[56] KANG U, LEE J Y, KOUTRA D, et al. Net-ray: Visualizing and mining billion-scale graphs [C]//PAKDD. 2014.

[57] GYONGYI Z, GARCIA-MOLINA H, PEDERSEN J. Combating web spam with trustrank [C]//Proceedings of the 30th international conference on very large data bases (VLDB). 2004.

[58] PANDIT S, CHAU D H, WANG S, et al. Netprobe: a fast and scalable system for fraud detection in online auction networks[C]//WWW.

[59] KOUTRA D, KE T Y, KANG U, et al. Unifying guilt-by-association approaches: Theorems and fast algorithms[C]//ECML-PKDD. 2011: 245-260.

[60] ESWARAN D, GÜNNEMANN S, FALOUTSOS C, et al. Zoobp: Belief propagation for heterogeneous networks[J]. Proceedings of the VLDB Endowment, 2017, 10(5):625-636.

[61] SHAH N, BEUTEL A, GALLAGHER B, et al. Spotting suspicious link behavior with fbox: An adversarial perspective[C]//ICDE. 2014: 959-964.

[62] SUN J, QU H, CHAKRABARTI D, et al. Neighborhood formation and anomaly detection in bipartite graphs[C]//Fifth IEEE International Conference on Data Mining (ICDM'05). IEEE, 2005: 8-pp.

[63] CHAKRABARTI D. Autopart: Parameter-free graph partitioning and outlier detection[C]// PKDD. 2004.

[64] SUN H, HUANG J, HAN J, et al. gskeletonclu: Density-based network clustering via structure-connected tree division or agglomeration[C]//2010 IEEE International Conference on Data Mining. IEEE, 2010: 481-490.

[65] TONG H, LIN C Y. Non-negative residual matrix factorization with application to graph anomaly detection[C]//Proceedings of the 2011 SIAM International Conference on Data Mining. SIAM, 2011: 143-153.

[66] BEUTEL A, XU W, GURUSWAMI V, et al. Copycatch: stopping group attacks by spotting lockstep behavior in social networks[C]//WWW. 2013: 119-130.

[67] ZHANG S, ZHOU D, YILDIRIM M Y, et al. Hidden: hierarchical dense subgraph detection with application to financial fraud detection[C]//SDM'17. SIAM, 2017.

[68] GUNNEMANN S, FARBER I, BODEN B, et al. Subspace clustering meets dense subgraph mining: A synthesis of two paradigms[C]//2010 IEEE International Conference on Data Mining. IEEE, 2010: 845-850.

[69] HOOI B, SONG H A, BEUTEL A, et al. Fraudar: Bounding graph fraud in the face of camouflage[C]//SIGKDD. 2016: 895-904.

[70] SARIYÜCE A E, PINAR A. Peeling bipartite networks for dense subgraph discovery[C]// WSDM. 2016.

[71] SANEI-MEHRI S V, SARIYÜCE A E, TIRTHAPURA S. Butterfly counting in bipartite networks[C]//KDD. 2017.

[72] AKOGLU L, TONG H, MEEDER B, et al. Pics: Parameter-free identification of cohesive subgroups in large attributed graphs[C]//SDM. 2012.

[73] NOBLE C C, COOK D J. Graph-based anomaly detection[C]//Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. 2003: 631-636.

[74] EBERLE W, HOLDER L. Discovering structural anomalies in graph-based data[C]//Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007). IEEE, 2007: 393-398.

[75] KUMAR S, HOOI B, MAKHIJA D, et al. Rev2: Fraudulent user prediction in rating platforms[C]//WSDM. 2018.

[76] GAO J, LIANG F, FAN W, et al. On community outliers and their efficient detection in information networks[C]//Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. 2010: 813-822.

[77] MÜLLER E, SÁNCHEZ P I, MÜLLE Y, et al. Ranking outlier nodes in subspaces of attributed graphs[C]//2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW). IEEE, 2013: 216-222.

[78] GÜNNEMANN S, BODEN B, SEIDL T. Finding density-based subspace clusters in graphs with feature vectors[J]. Data mining and knowledge discovery, 2012, 25(2):243-269.

[79] PEROZZI B, AKOGLU L, SÁNCHEZ P I, et al. Focused clustering and outlier detection in large attributed graphs[C]//KDD. 2014.

[80] TSOURAKAKIS C, BONCHI F, GIONIS A, et al. Denser than the densest subgraph: extracting optimal quasi-cliques with quality guarantees[C]//SIGKDD. 2013: 104-112.

[81] AKOGLU L, CHANDY R, FALOUTSOS C. Opinion fraud detection in online reviews by network effects.[J]. ICWSM, 2013:2-11.

[82] JENSEN D, NEVILLE J, GALLAGHER B. Why collective inference improves relational classification[C]//Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 2004: 593-598.

[83] LU Q, GETOOR L. Link-based classification[C]//Proceedings of the 20th International Conference on Machine Learning (ICML-03). 2003: 496-503.

[84] CHAKRABARTI S. Dynamic personalized pagerank in entity-relation graphs[C]// Proceedings of the 16th international conference on World Wide Web. 2007: 571-580.

[85] FANG Y, WANG R, DAI B, et al. Graph-based learning via auto-grouped sparse regularization and kernelized extension[J]. IEEE Transactions on knowledge and data engineering, 2014, 27(1):142-154.

[86] KOLDA T G, BADER B W. Tensor decompositions and applications[J]. SIAM review, 2009.

[87] MARUHASHI K, GUO F, FALOUTSOS C. Multiaspectforensics: Pattern mining on large-scale heterogeneous networks with tensor analysis[C]//2011 International Conference on Advances in Social Networks Analysis and Mining. IEEE, 2011: 203-210.

[88] SUN J, TAO D, FALOUTSOS C. Beyond streams and graphs: dynamic tensor analysis[C]// Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006: 374-383.

[89] JIANG M, BEUTEL A, CUI P, et al. A general suspiciousness metric for dense blocks in multimodal data[C]//ICDM'15. 2015.

[90] SHIN K, HOOI B, FALOUTSOS C. M-zoom: Fast dense-block detection in tensors with quality guarantees[C]//ECML-PKDD'16. 2016.

[91] CHARIKAR M. Greedy approximation algorithms for finding dense components in a graph [C]//APPROX'00. 2000.

[92] SHIN K, HOOI B, KIM J, et al. D-cube: Dense-block detection in terabyte-scale tensors[C]// WSDM'17. ACM, 2017.

[93] SHIN K, HOOI B, KIM J, et al. Densealert: Incremental dense-subtensor detection in tensor streams[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017: 1057-1066.

[94] BAN Y, LIU X, HUANG L, et al. No place to hide: Catching fraudulent entities in tensors [C]//WWW. 2018.

[95] PAPALEXAKIS E E, FALOUTSOS C, MITCHELL T M, et al. Turbo-smt: Accelerating coupled sparse matrix-tensor factorizations by 200x[J]. Proceedings of the ... SIAM International Conference on Data Mining. SIAM International Conference on Data Mining, 2014, 2014:118-126.

[96] MCGREGOR A. Graph stream algorithms: A survey[J/OL]. SIGMOD Rec., 2014, 43(1): 9-20. http://doi.acm.org/10.1145/2627692.2627694.

[97] CHEN Z, HENDRIX W, SAMATOVA N F. Community-based anomaly detection in evolutionary networks[J]. Journal of Intelligent Information Systems, 2012, 39(1):59-85.

[98] JI T, YANG D, GAO J. Incremental local evolutionary outlier detection for dynamic social networks[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2013: 1-15.

[99] ARAUJO M, PAPADIMITRIOU S, GÜNNEMANN S, et al. Com2: fast automatic discovery of temporal (‘comet’) communities[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2014: 271-283.

[100] ESWARAN D, FALOUTSOS C, GUHA S, et al. Spotlight: Detecting anomalies in streaming graphs[C]//KDD. 2018.

[101] ESWARAN D, FALOUTSOS C. Sedanspot: Detecting anomalies in edge streams[J]. 2018 IEEE International Conference on Data Mining (ICDM), 2018:953-958.

[102] YOON M, HOOI B, SHIN K, et al. Fast and accurate anomaly detection in dynamic graphs with a two-pronged approach[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2019: 647-657.

[103] SUN J, FALOUTSOS C, PAPADIMITRIOU S, et al. Graphscope: parameter-free mining of large time-evolving graphs[C]//Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. 2007: 687-696.

[104] SHAH N, KOUTRA D, ZOU T, et al. Timecrunch: Interpretable dynamic graph summarization[C]//KDD. 2015.

[105] SUN J, XIE Y, ZHANG H, et al. Less is more: Compact matrix decomposition for large sparse graphs[C]//Proceedings of the 2007 SIAM International Conference on Data Mining. SIAM, 2007: 366-377.

[106] JIANG R, FEI H, HUAN J. Anomaly localization for network data streams with graph joint sparse pca[C]//Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. 2011: 886-894.

[107] KOUTRA D, PAPALEXAKIS E E, FALOUTSOS C. Tensorsplat: Spotting latent anomalies in time[C]//2012 16th Panhellenic Conference on Informatics. IEEE, 2012: 144-149.

[108] HOOI B, SHIN K, LIU S, et al. Smf: Drift-aware matrix factorization with seasonal patterns [C]//SDM. 2019.

[109] HUANG Z, ZENG D D. A link prediction approach to anomalous email detection[C]//2006 IEEE International Conference on Systems, Man and Cybernetics: volume 2. IEEE, 2006: 1131-1136.

[110] PRIEBE C E, CONROY J M, MARCHETTE D J, et al. Scan statistics on enron graphs[J]. Computational & Mathematical Organization Theory, 2005, 11(3):229-247.

[111] KOUTRA D, JIN D, NING Y, et al. Perseus: an interactive large-scale graph mining and visualization tool[J]. VLDB, 2015.

[112] FAKHRAEI S, FOULDS J, SHASHANKA M, et al. Collective spammer detection in evolving multi-relational social networks[C]//KDD '15: SIGKDD. ACM, 2015.

[113] KLEINBERG J M. Authoritative sources in a hyperlinked environment[J/OL]. J. ACM, 1999, 46(5):604-632. http://doi.acm.org/10.1145/324133.324140.

[114] KANG U, MEEDER B, FALOUTSOS C. Spectral analysis for billion-scale graphs: Discoveries and implementation[C]//PAKDD. 2011.

[115] PELLEG D, MOORE A W, et al. X-means: Extending k-means with efficient estimation of the number of clusters.[C]//ICML. 2000: 727-734.

[116] HAMERLY G, ELKAN C. Learning the k in k-means[J]. NIPS, 2004.

[117] ZHANG T, RAMAKRISHNAN R, LIVNY M. Birch: An efficient data clustering method for very large databases[C]//SIGMOD. 1996: 103-114.

[118] ANKERST M, BREUNIG M M, KRIEGEL H P, et al. Optics: ordering points to identify the clustering structure[C]//SIGMOD. 1999: 49-60.

[119] BÖHM C, FALOUTSOS C, PAN J Y, et al. Robust information-theoretic clustering[C]// KDD.

[120] WANG W, YANG J, MUNTZ R, et al. Sting: A statistical information grid approach to spatial data mining[C]//VLDB. 1997: 186-195.

[121] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191):1492-1496.

[122] VINCENT L, SOILLE P. Watersheds in digital spaces: an efficient algorithm based on immersion simulations[J]. PAMI, 1991:583-598.

[123] ROERDINK J B, MEIJSTER A. The watershed transform: Definitions, algorithms and parallelization strategies[J]. Fundamenta informaticae, 2000.

[124] CAMPELLO R J G B, MOULAVI D, ZIMEK A, et al. Hierarchical density estimates for data clustering, visualization, and outlier detection[J]. ACM TKDD.

[125] SCHAEFFER S E. Graph clustering[J]. Computer science review, 2007, 1(1):27-64.

[126] HEYNCKES M. The predictive vs. the simulating brain: A literature review on the mechanisms behind mimicry[J]. Maastricht Student Journal of Psychology and Neuroscience, 2016, 4.

[127] WARE C. Color sequences for univariate maps: theory, experiments and principles[J]. IEEE Computer Graphics and Applications, 1988:41-49.

[128] BORKIN M, GAJOS K, PETERS A, et al. Evaluation of artery visualizations for heart disease diagnosis[J]. IEEE Trans. on Visualization and Computer Graphics, 2011:2479-2488.

[129] TUKEY J W, TUKEY P A. Computer graphics and exploratory data analysis: An introduction [J]. Nat Computer Graphics Association, 1985.

[130] AKOGLU L, CHAU D H, KANG U, et al. Opavion: Mining and visualization in large graphs [C]//SIGMOD. 2012: 717-720.

[131] ROSSI R A, AHMED N, ZHOU R, et al. Interactive visual graph mining and learning[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2018, 9:1 - 25.

[132] PEROZZI B, AKOGLU L. Discovering communities and anomalies in attributed graphs: Interactive visual exploration and summarization[J/OL]. ACM Trans. Knowl. Discov. Data, 2018, 12(2). https://doi.org/10.1145/3139241.

[133] BUJA A, TUKEY P A. Computing and graphics in statistics[M]. Springer-Verlag New York.

[134] WILKINSON L, ANAND A, GROSSMAN R. Graph-theoretic scagnostics[J]. Proceedings - IEEE Symposium on Information Visualization, INFO VIS, 2005:157-164.

[135] GUPTA N, ESWARAN D, SHAH N, et al. Beyond outlier detection: Lookout for pictorial explanation[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2018: 122-138.

[136] WENG Y, LIU L. A collective anomaly detection approach for multidimensional streams in mobile service security[J]. IEEE Access, 2019, 7:49157-49168.

[137] YAN J, SHI L, TAO J, et al. Visual analysis of collective anomalies using faceted high-order correlation graphs[J]. IEEE transactions on visualization and computer graphics, 2018.

[138] MIZ V, RICAUD B, BENZI K, et al. Anomaly detection in the dynamics of web and social networks using associative memory[C]//The World Wide Web Conference. 2019: 1290-1299.

[139] AHMED M. Collective anomaly detection techniques for network traffic analysis[J]. Annals of Data Science, 2018, 5(4):497-512.

[140] KUMAR R, NOVAK J, TOMKINS A. Structure and evolution of online social networks [M]//Link mining: models, algorithms, and applications. Springer, 2010.

[141] CHEN J, SAAD Y. Dense subgraph extraction with application to community detection[J]. IEEE Trans. on knowledge and Data Engineering, 2010.

[142] ZAKI M J, WAGNER MEIRA J. Data mining and analysis: Fundamental concepts and algorithms[M]. Cambridge University Press, 2014.

[143] BLUM A, HOPCROFT J, KANNAN R. Foundations of data science[J]. Vorabversion eines Lehrbuchs.

[144] DICARLO J J, ZOCCOLAN D, RUST N C. How does the brain solve visual object recognition?[J]. Neuron, 2012.

[145] LIU X M, JI R, WANG C, et al. Understanding image structure via hierarchical shape parsing [C]//CVPR. 2015.

[146] ARBELAEZ P, MAIRE M, FOWLKES C, et al. Contour detection and hierarchical image segmentation[J]. PAMI, 2011:898-916.

[147] BYLINSKII Z, KIM N W, O'DONOVAN P, et al. Learning visual importance for graphic designs and data visualizations[C]//Proceedings of the 30th Annual ACM Symposium on User Interface Software & Technology. 2017.

[148] GONZALEZ R C, WOODS R E. Image processing[J]. Digital image processing, 2007.

[149] THOMPSON H R. Truncated normal distributions[J]. Nature, 1950, 165:444-445.

[150] BISHOP C M, NASRABADI N M. Pattern recognition and machine learning[J]. J. Electronic Imaging, 2007, 16:049901.

[151] CUBEDO M, OLLER J M. Hypothesis testing: a model selection approach[C]//2002.

[152] CHERNOBAI A, RACHEV S T, FABOZZI F J. Composite goodness-of-fit tests for left-truncated loss samples[M/OL]//LEE C F, LEE J C. Handbook of Financial Econometrics and Statistics. New York, NY: Springer New York, 2015: 575-596. https://doi.org/10.1007/978-1-4614-7750-1_20.

[153] STEPHENS M A. Edf statistics for goodness of fit and some comparisons[J]. Journal of the American statistical Association, 1974:730-737.

[154] Amazon ratings network dataset -- KONECT[EB/OL]. 2017. http://konect.uni-koblenz.de/networks/amazon-ratings.

[155] MCAULEY J, PANDEY R, LESKOVEC J. Inferring networks of substitutable and complementary products[C]//KDD. 2015.

[156] MCAULEY J J, LESKOVEC J. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews[C]//WWW. 2013.

[157] Yelp dataset challenge[EB/OL]. 2017. https://www.yelp.com/dataset_challenge.

[158] MCAULEY J, LESKOVEC J. Image labeling on a network: using social-network metadata for image classification[J]. ECCV, 2012:828-841.

[159] MISLOVE A, MARCON M, GUMMADI K P, et al. Measurement and analysis of online social networks[C]//SIGCOMM. 2007.

[160] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: Machine learning in python[J]. Journal of machine learning research, 2011, 12(Oct):2825-2830.

[161] MCINNES L, HEALY J, ASTELS S. hdbscan: Hierarchical density based clustering[J]. The Journal of Open Source Software, 2017, 2(11):205.

[162] WASSERMAN L. All of nonparametric statistics (springer texts in statistics)[M]. Springer-Verlag New York, Inc., 2006.

[163] ELIAS P. Universal codeword sets and representations of the integers[J]. IEEE trans. on information theory, 1975:194-203.

[164] CHAKRABARTI D, PAPADIMITRIOU S, MODHA D S, et al. Fully automatic cross-associations[C]//SIGKDD. 2004: 79-88.

[165] GIATSOGLOU M, CHATZAKOU D, SHAH N, et al. Nd-sync: Detecting synchronized fraud activities[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2015: 201-214.

[166] WONG S W, PASTRELLO C, KOTLYAR M, et al. Sdregion: Fast spotting of changing communities in biological networks[C]//SIGKDD'18. 2018.

[167] LIU Y, ZHU L, SZEKELY P A, et al. Coupled clustering of time-series and networks[C]//SDM. SIAM, 2019: 531-539.

[168] SHEN H W, CHENG X Q. Spectral methods for the detection of network community structure: a comparative analysis[J]. J STAT MECH-THEORY EXP, 2010.

[169] GOLDBERG A V. Finding a maximum density subgraph[C]//1984.

[170] ASAHIRO Y, IWAMA K, TAMAKI H, et al. Greedily finding a dense subgraph[J]. Journal of Algorithms, 2000, 34(2):203-221.

[171] ROSSI R A, GLEICH D F, GEBREMEDHIN A H, et al. Fast maximum clique algorithms for large graphs[C]//WWW. 2014: 365-366.

[172] MITZENMACHER M, PACHOCKI J, PENG R, et al. Scalable large near-clique detection in large-scale networks via sampling[C/OL]//KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015: 815-824. http://doi.acm.org/10.1145/2783258.2783385.

[173] MIYAUCHI A, KAKIMURA N. Finding a dense subgraph with sparse cut[C]//CIKM. 2018.

[174] BOOB D, GAO Y, PENG R, et al. Flowless: Extracting densest subgraphs without flow computations[J]. 2019.

[175] TSOURAKAKIS C E, CHEN T, KAKIMURA N, et al. Novel dense subgraph discovery primitives: Risk aversion and exclusion queries[J]. ArXiv, 2019, abs/1904.08178.

[176] ANDERSEN R, CHELLAPILLA K. Finding dense subgraphs with size bounds[C]// WAW'09.

[177] KHULLER S, SAHA B. On finding dense subgraphs[C/OL]//ICALP ' 09: Proceedings of the 36th International Colloquium on Automata, Languages and Programming: Part I. Berlin, Heidelberg: Springer-Verlag, 2009: 597–608. https://doi.org/10.1007/978-3-642-02927-1_50.

[178] YANG Y, CHU L, ZHANG Y, et al. Mining density contrast subgraphs[C]//ICDE. IEEE, 2018: 221-232.

[179] CHU L, WANG Z, PEI J, et al. Finding gangs in war from signed networks[C]//KDD. ACM, 2016: 1505-1514.

[180] PAPAILIOPOULOS D, MITLIAGKAS I, DIMAKIS A, et al. Finding dense subgraphs via low-rank bilinear optimization[C]//ICML. 2014: 1890-1898.

[181] ANDERSEN R, CIOABA S M. Spectral densest subgraph and independence number of a graph.[J]. J. UCS, 2007, 13(11):1501-1513.

[182] Chung F R K. Spectral graph theory[M/OL]. 1996. https://academic.microsoft.com/paper/1578099820.

[183] LIU S, HOOI B, FALOUTSOS C. A contrast metric for fraud detection in rich graphs[J]. TKDE, 2018, 31(12):2235-2248.

[184] LI Z, ZHANG S, WANG R S, et al. Quantitative function for community detection[J]. Physical review E, 2008, 77(3):036109.

[185] LI Z, ZHANG S, WANG R S, et al. Erratum: Quantitative function for community detection [J]. Physical Review E, 2015, 91(1):019901.

[186] WANG Z, CHU L, PEI J, et al. Tradeoffs between density and size in extracting dense subgraphs: A unified framework[C]//ASONAM. IEEE, 2016: 41-48.

[187] PAVAN M, PELILLO M. Dominant sets and pairwise clustering[J]. IEEE Trans, on pattern analysis and machine intelligence, 2006, 29(1):167-172.

[188] HOOI B, SHIN K, LAMBA H, et al. Telltail: Fast scoring and detection of dense subgraphs [C]//AAAI. 2020.

[189] YIN H, BENSON A R, LESKOVEC J, et al. Local higher-order graph clustering[C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017: 555-564.

[190] DAX A. From eigenvalues to singular values: a review[J]. Advances in Pure Mathematics, 2013, 2013.

[191] TSOURAKAKIS C E. Fast counting of triangles in large real networks without counting: Algorithms and laws[C]//2008 Eighth IEEE International Conference on Data Mining. IEEE, 2008: 608-617.

[192] EIKMEIER N, GLEICH D F. Revisiting power-law distributions in spectra of real world networks[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017: 817-826.

[193] CHAKRABARTI D, ZHAN Y, FALOUTSOS C. R-mat: A recursive model for graph mining [C]//SDM. SIAM, 2004: 442-446.

[194] GOLUB G H, VAN LOAN C F. Matrix computations: volume 3[M]. JHU press, 2012.

[195] KUNEGIS J. Konect: the koblenz network collection[C]//WWW. 2013: 1343-1350.

[196] ROSSI R A, AHMED N K. The network data repository with interactive graph analytics and visualization[C/OL]//AAAI. 2015. http://networkrepository.com.

[197] LESKOVEC J, KREVL A. SNAP Datasets: Stanford large network dataset collection [EB/OL]. 2014. http://snap.stanford.edu/data.

[198] ZAFARANI R, LIU H. Social computing data repository at ASU[EB/OL]. Arizona State University, School of Computing, Informatics and Decision Systems Engineering, 2009. http://socialcomputing.asu.edu.

[199] MISLOVE A, MARCON M, GUMMADI K P, et al. Measurement and analysis of online social networks[C]//Internet Measurement on SIGCOM. 2007.

[200] WAN H, ZHANG Y, ZHANG J, et al. Aminer: Search and mining of academic social networks[J]. Data Intelligence, 2019, 1(1):58-76.

[201] LESKOVEC J, LANG K J, DASGUPTA A, et al. Statistical properties of community structure in large social and information networks[C]//WWW '08. 2008.

[202] EDLER D, BOHLIN L, et al. Mapping higher-order network flows in memory and multilayer networks with infomap[J]. Algorithms, 2017, 10(4):112.

[203] YANG B, DI J, LIU J, et al. Hierarchical community detection with applications to real-world network analysis[J]. DKE, 2013.

[204] SARIYüCE A E, PINAR A. Fast hierarchy construction for dense subgraphs[J]. VLDB'16.

[205] GIBSON D, KUMAR R, TOMKINS A. Discovering large dense subgraphs in massive graphs [C]//VLDB'05. VLDB Endowment, 2005.

[206] JETHAVA V, MARTINSSON A, BHATTACHARYYA C, et al. Lovász ϑ function, svms and finding dense subgraphs[J]. The Journal of Machine Learning Research, 2013.

[207] BALALAU O D, BONCHI F, CHAN T H H, et al. Finding subgraphs with maximum total density and limited overlap[C]//WSDM'15. 2015.

[208] PAPADIMITRIOU S, SUN J, FALOUTSOS C, et al. Hierarchical, parameter-free community discovery[C]//ECML-PKDD'08. Springer, 2008.

[209] YANG T, CHI Y, ZHU S, et al. Detecting communities and their evolutions in dynamic social networks--a bayesian approach[J]. Mach. Learn., 2011.

[210] GOROVITS A, GUJRAL E, PAPALEXAKIS E E, et al. Larc: Learning activity-regularized overlapping communities across time[C]//SIGKDD'18. ACM, 2018.

[211] SIDDIQUE B, AKHTAR N. Temporal hierarchical event detection of timestamped data[J]. ICCCA'17, 2017.

[212] GRIPPO L, SCIANDRONE M. On the convergence of the block nonlinear gauss-seidel method under convex constraints[J]. Oper. Res. Lett., 2000.

[213] COLEMAN T F, LI Y. An interior trust region approach for nonlinear minimization subject to bounds[J]. SIAM Journal on optimization, 1996.

[214] LIN C J. Projected gradient methods for nonnegative matrix factorization[J]. 2007.

[215] LIN C J, MORÉ J J. Newton's method for large bound-constrained optimization problems[J]. SIAM J. on Optimization, 1999.

[216] RISSANEN J. A universal prior for integers and estimation by minimum description length [J]. The Annals of Statistics, 1983.

[217] MCAULEY J, PANDEY R, LESKOVEC J. Inferring networks of substitutable and complementary products[C]//SIGKDD'15. ACM, 2015.

[218] ROSSI R A, AHMED N K. The network data repository with interactive graph analytics and visualization[C]//AAAI'15. 2015.

[219] LIPPMANN R P, FRIED D J E. Evaluating intrusion detection systems: The 1998 darpa off-line intrusion detection evaluation[C]//DARPA Information Survivability Conference and Exposition, 2000. DISCEX'00. IEEE, 2000.

[220] Kdd cup 1999 data[EB/OL]. 1999. https://kdd.ics.uci.edu/databases/kddcup99/.

致　谢

弹指一挥间，匆匆六载已逝。曾记得往昔峥嵘岁月，当年的豪情壮志，言犹在耳，而今逝者如斯，不禁感慨万千。在博士攻读期间，夙兴夜寐，靡有朝矣，在科研之路上奋力攻坚克难、追求自我突破、忙碌而充实；个中艰辛与成长、付出与收获、坚持与积累终将沉淀为个人的宝贵财富。在博士论文付梓之际，向一路走来给予我帮助、支持的各位老师、同学、亲友表示衷心的感谢。

感谢我的导师程学旗研究员对我博士期间的悉心指导。针对科研工作开展和个人职业发展，程老师都给出了高屋建瓴的建议和温暖诚挚的关怀。他在科研工作中的深度思考和批判以及积极刻苦的工作精神使我在做人和做学问方面获益颇丰。感谢他提供的优良工作环境和充分有力的支持。

感谢我的直接指导老师刘盛华副研究员对我课题选择、论文工作的精心指导。从引路科研探索到提升学术能力到拓宽学术视野，从严谨求实的治学态度到周密细致的逻辑表达，从俛俛从事的工作精神到平易谦和的处世风格，他都给予我耐心的指导和有力的支持。在四年的研究生活中，他亦师亦友，给予我不断进步的动力，相信这些都将为我之后的学习和工作带来积极的影响。

感谢沈华伟研究员在科研、工作和生活中给予的指导和帮助。在组会讨论、工作交流中，他的深入思考、敏锐洞察、精辟见解和耐心指导对我科研的启蒙和成长、问题的分析和思考都起到了重要的影响；他平易近人的人格魅力、严谨求实的科研精神、认真负责的工作态度都使我受益匪浅，也是我学习的楷模。

感谢在论文开题、中期和终期阶段中给出宝贵建议和意见的各位老师，包括袁野教授、石川教授、郭嘉丰研究员、靳小龙研究员、兰艳艳研究员、罗平副研究员、张志斌副研究员，他们指导和意见促成了最终论文的不断精练和完善。

感谢中国科学院大学的孙翼老师在我的科研、学习和生活中给予的莫大鼓励、支持、分享和帮助。

感谢大图数据挖掘组的各位师弟师妹：刘财政、周斌、张嘉宝、周厚铨、曾四为、赵鑫、丁泉以及其他过往的实习学生，感谢大家在组会和课程学习中的分享和付出以及对我科研、生活的帮助和支持。

感谢曾经 NASC 研究组以及现在基础研究部的各位老师和同学。感谢欧阳

文涛、徐力、兰艳艳三位老师，从每次组会的讨论、建议或质疑中使我受益匪浅；感谢孙晓茜师姐、刘雅辉师姐、王永庆师兄、高金华师兄、孙冰杰师兄、王昊师兄、段雪野师兄、施发斌师兄、李思莹师姐等在科研、工作和生活上给予的慷慨帮助和建议。尤其在我的科研过程中，高金华师兄不厌其烦地参与某些问题的讨论并不吝提供有用的建议。还有组内的曹琦、吕珊珊、赵睿卓、高浩、岑科廷、徐冰冰、程亮、黄俊杰、王兆慧等师弟师妹，感谢你们在科研和生活上的帮助与分享。感谢傅川老师的帮助和支持，感谢程苏琦师姐、曾玮师姐、齐雅婷师姐、康鹏师兄的无私帮助和真诚建议。

感谢实验室和研究所的其他老师、员工和学生提供的支持和付出的工作。感谢刘悦老师在工作和生活中给予的帮助和支持；感谢王莹、崔连军、张冬、宋茵、常晶等各位老师在科研与生活中提供的支持和便利；感谢冯刚、张平等老师在工作中的付出和支持。感谢在论文开题、答辩过程的担任秘书工作的同学。

感谢我的朋友们的支持、帮助和分享，这些人包括但不限于：陆杰、李震、李雪琦、马玉卓、余显、李家宁、官赛平、牛颂杰等，感谢好友兼合作者王晋东同学，我们在生活和科研中相互帮助、共同进步，开始独立的科研探索之路。感谢我们的饭团小伙伴：李家宁、官赛平、钟巧灵、侯中妮、软件所王少将同学等。

在此，还特别感谢我论文工作中其他合作作者，他们是美国卡内基梅隆大学的 Christos Faloutsos 教授，美国密歇根大学的 Danai Koutra 教授，新加坡国立大学的 Bryan Hooi 教授、美国伊利诺伊大学芝加哥分校的 Philip S. Yu 教授等，谢谢你们的付出和鼓励。

本研究工作受到国家 973 计划项目、国家自然科学基金、中科院先导计划、北京市自然科学基金的资助，特此感谢！

感谢我的父母、兄弟和其他亲人朋友在我博士攻读期间给予的支持、鼓励和关怀。

作者简历及攻读学位期间发表的学术论文与研究成果

作者简历

姓名：冯文杰　性别：男　出生日期：1990.12　籍贯：山西阳泉

通讯地址：中国科学院计算技术研究所　邮编：100190

E-mail：wenchiehfeng.us@ict.ac.cn

教育经历

2014 年 9 月–2020 年 6 月，中国科学院计算技术研究所，攻读博士学位

2010 年 9 月–2014 年 6 月，北京交通大学，获得工学学士学位

已发表的学术论文

[1] **Wenjie Feng**, Shenghua Liu, Danai Koutra, Huawei Shen, and Xueqi Cheng. SpecGreedy: Unified Dense Subgraph Detection. The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD'20), 2020. (**Best student paper award**)

[2] **Wenjie Feng**, Shenghua Liu, Christos Faloutsos, Bryan Hooi, Huawei Shen, and Xueqi Cheng. Vison-Guided Micro-cluster Recognition and Anomaly Detection. Future Generation Computer Science (FGCS), 2020.

[3] **Wenjie Feng**, Shenghua Liu, and Xueqi Cheng. CatchCore: Hierarchical Dense Subtensor Detection. The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD'19), 2019.

[4] Jindong Wang, Yiqiang Chen, **Wenjie Feng**, Han Yu, Meiyu Huang, and Qiang Yang. Transfer Learning with Dynamic Distribution Adaptation. ACM Transactions on Intelligent Systems and Technology (TIST) 11, no. 1 (2020): 1-25.

[5] **Wenjie Feng**, Shenghua Liu, Christos Faloutsos, Bryan Hooi, Huawei Shen, and Xueqi Cheng. Beyond outliers and on to micro-clusters: Vision-guided anomaly detection. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 541-554. Springer, Cham, 2019.

[6] Jiabao Zhang, Shenghua Liu, Wenjian Yu, **Wenjie Feng**, and Xueqi Cheng. EigenPulse: Detecting Surges in Large Streaming Graphs with Row Augmentation. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 501-513. Springer, Cham, 2019.

[7] **Wenjie Feng**, Shenghua Liu, Christos Faloutsos, Bryan Hooi, Huawei Shen, and Xueqi Cheng. EagleMine: Vision-Guided Mining in Large Graphs. *Outlier Detection De-constructed (ODD) v5.0, KDD 2018.*

[8] **Wenjie Feng***, Jindong Wang*, Yiqiang, Chen, Han Yu, and Philip S Yu. Visual Domain Adaptation with Manifold Embedded Distribution Alignment. In Proceedings of the 26th ACM international conference on Multimedia, pp. 402-410. 2018.

[9] Jindong Wang, Yiqiang Chen, Shuji Hao, **Wenjie Feng**. Balanced Distribution Adaptation for Transfer Learning. In 2017 IEEE International Conference on Data Mining (ICDM), pp. 1129-1134. IEEE, 2017.

在审/待投论文

[1] **Wenjie Feng**, Shenghua Liu, and Xueqi Cheng. Hierarchical Dense Pattern Detection in Tensor. ACM Transactions on Knowledge Discovery from Data (TKDD), 2020.

[2] Chaohui Yu, Jindong Wang, Chang Liu, **Wenjie Feng**, TieYan Liu. Learning to Match Distributions across Domains.

申请或已获得的专利

[1] **冯文杰**, 刘盛华, 刘财政, 程学旗, 一种基于视觉引导的大图微簇检测系统. 国家发明专利 (已提交)

[2] 陈益强, 王晋东, **冯文杰**, 忽丽莎. 一种基于流形迁移学习的数据标定方法和系统. 国家发明专利 (已受理并公开). CN108960270A

[3] 傅川, **冯文杰**, 程学旗, 张国清, 靳小龙, 梁英. 一种基于二维码的影视剧互动的方法和设备. 国家发明专利. CN106060576A

[4] 傅川, **冯文杰**, 程学旗, 张国清, 靳小龙, 梁英. 一种基于二维码的综艺节目互动的方法和设备. 国家发明专利. CN106060577A

攻读博士学位期间参加的科研项目

1. 国家自然科学基金面上项目: “大规模多属性图中的异常模式挖掘”
项目号：61772498 ，时间：2018 年 1 月 - 2021 年 12 月

2. 中科院先导科技专项：“地球大数据科学工程”.
项目号：XDA19020400，时间：2017 年 - 2022 年

3. 973 研究项目：“社交网络分析与网络传播的基础研究”.
项目号：2013CB329602，时间：2016 年 9 月 -- 2017 年 9 月

4. 863 研究项目：“动态媒体业务支撑平台与应用系统”.
项目号：Y610061000，时间：2015 年 3 月 -- 2017 年 9 月

攻读博士学位期间的获奖情况

1. 2020 年获得 ECML PKDD 2020 数据挖掘最佳学生论文奖
2. 2019 年获得中科院计算所所长优秀奖
3. 2019 年获得网络数据科学与技术重点实验室，天玑团队优秀学生
4. 2019 年获得 PAKDD 2019 学生奖学金
5. 2018 年获得网络数据科学与技术重点实验室优秀学生奖
6. 2016 年-2018 年获得中国科学院大学一等学业奖学金
7. 2017 年获得中国科学院大学``三好学生" 荣誉称号