唐德权,黄金贵,史伟奇.一种新的犯罪团伙挖掘算法[J].湖南科技大学学报(自然科学版),2023,38(2):80-87. doi:10.13582/j.cnki.1672-9102.2023.02.011

TANG D Q, HUANG J G, SHI W Q. A New Criminal Gang Mining Algorithm[J]. Journal of Hunan University of Science and Technology (Natural Science Edition),2023,38(2):80-87. doi:10.13582/j.cnki.1672-9102.2023.02.011

一种新的犯罪团伙挖掘算法

唐德权[1*],黄金贵[2],史伟奇[1]

(1.湖南警察学院 信息技术(网监)系,湖南 长沙 410138; 2.湖南师范大学 信息科学与工程学院,湖南 长沙 410081)

摘 要:为了利用图模式挖掘犯罪情报网络中的核心团伙和核心人物,提高犯罪网络威胁预测和识别的效率,提出一种新的核心团伙挖掘算法(Core Gang Mining Algorithm, CGMA).对海量的犯罪情报网络数据集建立相应的无向简单图模型,通过改进图挖掘方式,构建候选核心团伙集的数据结构,并提出由 k-团伙通过连接和扩展 2 种操作得到(k+1)-团伙,从各个不同的图数据中统计其频度,最后在模拟数据集和真实数据集上验证算法 CGMA 的准确性和时间复杂度.该算法避免了传统的图模式挖掘中的子图同构问题,同时也优于其他常用的犯罪团伙挖掘算法.试验结果表明:该算法能对犯罪核心团伙信息进行有效预测.

关键词:图模式;核心团伙;图挖掘;连接和扩展;子图同构

中图分类号:TP311.2 **文献标志码**:A **文章编号**:1672-9102(2023)02-0080-08

A New Criminal Gang Mining Algorithm

TANG Dequan[1], HUANG Jingui[2], SHI Weiqi[1]

(1. Department of Information Technology, Hunan Police Academy, Changsha 410138, China;
2. College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China)

Abstract: In order to utilize graph patterns to mine core gangs and key characters in criminal intelligence networks and improve the efficiency of threat prediction and identification in criminal networks, a new core gang mining algorithm (CGMA) is proposed. Data set based on vast amounts of criminal intelligence network, the undirected simple graph model is established, by improving the mining method and building candidate core gangs set data structure. It obtains the k-gangs by join and extension of two (k+1)-gang from different graph data statistics of the frequency. Finally, the accuracy and time complexity of the algorithm CGMA are verified on the real data set, the algorithm avoids the problem of subgraph isomorphism in traditional graph pattern mining, which is better than other common mining algorithms of criminal gangs. Experimental results show that the algorithm can effectively predict the information of criminal core groups.

Keywords: graph pattern; core gang; graph mining; join and extension; subgraph isomorphism

随着社会经济的迅猛发展,各国安全部门都更加重视犯罪数据的收集与挖掘,大型团伙犯罪数据库挖掘已经成为新的研究热点.据文献统计,在跨国集团犯罪案中,团伙犯罪所占的比重不仅很大,而且犯罪手

收稿日期:2021-12-01 **修改日期**:2023-03-31

基金项目:湖南省教育科学“十四五”规划课题资助项目(XJK23BGD034);湖南警察学院高层次人才科研启动专项基金资助项目(2022KYQD03);国家自然科学基金资助项目(61471169)

* **通信作者**,E-mail: tdq525@126.com

段的隐蔽性和科技程度也越来越高[1].目前,对犯罪团伙的数据挖掘问题归为两类:第一类是已知嫌疑人,挖掘与特定嫌疑人有关联的其他同伙;第二类是未知嫌疑人,挖掘潜在未知的犯罪团伙.第一类问题的求解相对简单,本文关注第二类问题的求解.近年来,国内外学者都对第二类问题进行了研究.DWIVEDI 等[2]研究数据挖掘在网络犯罪中的应用,对数字取证工具和技术进行比较并分析它们的优点;IFTIKHAR 等[3]基于监督机器学习算法,从非结构化的法律文本中提取所需的犯罪信息,自动从现有的文本中找到有用的和关键的犯罪团伙信息;MALEKAR[4]提出一种利用深度学习处理卷积神经网络的新方法,从监控视频中识别出抢劫、谋杀等可疑犯罪团伙活动;李万彪等[5]提出基于各类数据资源挖掘犯罪团伙信息,对数据资源建立关系数据模型,从而挖掘与嫌疑人关联的犯罪团伙及成员;李勇男[6]提出基于子图模式,对反恐情报数据集中的人员进行关联分析,得到涉恐人员犯罪的规律和特征,为打击恐怖活动的有效预测和防控提供依据.

基于图模式数据挖掘主要是从给定的图数据集中发现满足需求的拓扑结构,主要包括图的匹配、图数据中的关键字查询、频繁子图挖掘、聚类以及分类等.图数据挖掘是一个重要的研究课题,广泛应用在很多不同的学科领域,包括化学信息学、生物信息学和社会科学等.ACOSTA-MENDOZA 等[7]提出一种通过计算团来减少频繁子图挖掘数量的替代方法;JAYALAKSHMI 等[8]基于频率、熵和页面持续时间开发了 4 个满足需求的度量,用于检索满足需求的子图,提出一种满足用户度量的频繁子图挖掘算法.由于传统的方法仅基于单个阈值来挖掘图形模式,因此会造成重要模式信息的丢失.基于这个问题,LEE 等[9]提出一种新的图挖掘算法,该算法可以同时考虑图中每个元素的不同重要性和多个阈值;随后,LEE 等[10]又进一步提出一种基于长度递减支持度约束和加权最小有效扩展的加权频繁图模式挖掘方法.

随着犯罪情报数据的日益增长,大型犯罪数据集广泛使用图模式来表示情报数据结构.但以上提出的方法都不能避免子图同构带来的 NP 困难问题,因此,为了提高大型情报数据集的挖掘效率,本文提出基于简单图模型构建核心团伙集数据挖掘方法,有效避免子图同构问题.图模式数据结构对犯罪网络的拓扑结构有很好的表达方式,能够很清楚地揭示犯罪网络实体之间的关系模式.实体顶点以某种关系在图数据集中出现的次数被认为是核心团伙的关系强度或关系权重,出现次数越大则关系强度越大.通过对图数据集的分析和发现,挖掘出犯罪网络中频繁出现的核心团伙(即团伙构成的子图)及核心人员.

本文针对目前犯罪情报数据集的实际需求,利用图数据模型对犯罪网络数据进行分析,针对从大型图数据集中识别核心犯罪团伙和核心人物的问题,提出一种新的有效算法,该算法遵循一般图模式挖掘深度优先搜索的方法,结合新的思路和技术,避免传统的图模式挖掘中的子图同构问题,提高图模式数据挖掘效率.

所提方法的主要特点:(1)一个新的求解核心团伙挖掘问题描述;(2)通过连接和扩展这 2 个有效候选集产生操作,产生所有候选集且每个候选集唯一;(3)构建新的数据结构——团伙集列表,完全避免子图同构测试.

1　问题描述

考虑的犯罪实体是指嫌疑人个体,同时也包括嫌疑人相关的所有身份实体,如电话号码、银行账户、社交用户和邮箱账号等.用无向图表示犯罪团伙之间的关联信息,每个犯罪实体为图的顶点,2 个犯罪实体之间通过通信、通话、账户转账和住宿等联系形成交互行为,表示为图中相对应顶点之间的边.根据不同的联系类型建立相应的网络模型,如通话网络、转账网络和聊天等社交网络等,每个网络均对应一个具有相同顶点集的简单图.

定义简单图 $G=(\boldsymbol{V},\boldsymbol{E},\boldsymbol{L})$,其中 $\boldsymbol{V}$ 为对应犯罪实体的顶点集,$\boldsymbol{E}$ 为对应犯罪实体之间关联的边集,$\boldsymbol{L}$ 为对应顶点标签的集合且无重复,表示不同的犯罪实体.

一个简单图的每个连通子图可以看作是一个犯罪团伙.一般的团伙作案都有组织或领导头目,并且团伙内部有作案计划及分工,团伙中的成员在网络中担任的具体角色通常因时因事而异.但是,主要成员或核心成员之间的联系和配合在团伙犯罪中最为密切.犯罪团伙的这种密切关系体现在两个方面:一是连通子图包含的边数,即边数越多说明关系越密切,具有最多边数的连通子图是完全子图(或称为团);二是连

通子图在图集中出现的频度,次数越频繁说明犯罪团伙的关系越紧密.本文更关注连通子图的频繁性而忽略连通子图包含的边数,也就是说只要是连通子图,不论包含多少条边都视为一个犯罪团伙.如果连通子图出现的次数超过给定的最小频繁次数,称该犯罪团伙为核心团伙.

对于图数据集 $\boldsymbol{R}$,其中与图 G 同构的子图个数与图数据集 $\boldsymbol{R}$ 中所有图的数量的比值称为图 G 的支持度,记为 σ_G,表示为 $\sigma_G=\dfrac{|\{G'\in \boldsymbol{R}\mid G\subseteq G'\}|}{|\boldsymbol{R}|}$,其中 G' 为 $\boldsymbol{R}$ 中的一个子图.设定最小支持度阈值为 min_σ,如果图 G 的支持度大于或等于最小支持度($\sigma_G\geqslant$min_σ),那么图 G 称为频繁图或频繁子图.

如果 G' 是 G 的一个连通子图且$|\boldsymbol{V}'|=k$,其中 $\boldsymbol{V}'$是 G'的顶点集,则称 G'是 G 的 k-顶点连通子图,本文将 k-顶点连通子图称为 k-团伙,并简记为关于 k 个顶点的子集或 k-子集,也称连通子集.k-团伙在所有图网络中出现的频次称为支持度,这里 k-团伙的出现不同于子图同构,仅需要相同 k-子集是连通子集即可.如果 k-团伙的支持度大于或等于给定的最小支持度($\sigma_k\geqslant$min_σ),则称该团伙是 k-核心团伙.k-核心团伙包含 k 个核心人物,核心团伙至少有 2 人是为了多次实行犯罪而结合起来的,因此 $2\leqslant k\leqslant|\boldsymbol{V}|$.在实际应用中,一个核心团伙通常是 2~10 个核心成员的犯罪团伙,因此可设定核心成员上限为 p($2\leqslant p\leqslant 10$),根据 p 值挖掘出所有 k-核心团伙,其中 $2\leqslant k\leqslant p$.

综上所述,本文求解的问题描述如下:给定核心团伙成员数目上限为 p,所有犯罪嫌疑人集合 $\boldsymbol{V}=\{v_1, v_2, \cdots, v_N\}$,以及图数据集 $\boldsymbol{R}=\{G_1, G_2, \cdots, G_m\}$,其中 $G_i(G_i\in\boldsymbol{R}, i=1, 2, \cdots, m)$是以 $\boldsymbol{V}$ 中若干个犯罪嫌疑人员为顶点且通过某种交流方式形成的简单无向图,求出 $\boldsymbol{V}$ 中的所有 k-核心团伙,其中 $2\leqslant k\leqslant p$.

2 核心团伙挖掘

基于图数据集 $\boldsymbol{R}$,提出了 k-核心团伙挖掘算法(Core Gang Mining Algorithm, CGMA).首先对海量的情报数据进行预处理建立图模型,然后根据图模型建立核心团伙子集列表,定义连接和扩展 2 种基本操作构建新的候选核心团伙子集,并在构建新候选核心团伙子集的同时,对其出现的频度进行统计计算,提高算法效率.

2.1 建立图数据集

犯罪情报图集中的顶点表示嫌疑实体,边表示 2 个实体之间的联系,顶点的标签表示不同的实体,他们之间以通话、同住宿和同上网等联系途径设定,由若干顶点和连接顶点的边构成图数据结构 k-子图是包含 k 个顶点的子集,表示 k 个嫌疑实体的网络.用图模式表示情报数据网络是一种直观的表示方法,图模式不仅可以表示实体的拓扑信息,而且能分析实体的关联规律.

在进行核心团伙挖掘前,按照嫌疑人员身份相关的特征进行标识,如电话号码、证件标识、银行账户、社交用户和邮箱账号等.以金融网络为例,嫌疑人员各自的账户信息标识了不同的嫌疑人身份,形成的顶点如 A, B, C, D,结合账户之间的转账等交易行为,建立相应的简单图,如图 1 所示.

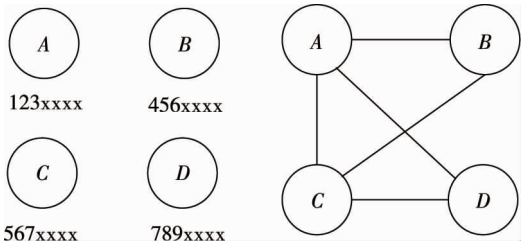

图 1 犯罪团伙简单图

犯罪团伙为实现特定的目标,在组织犯罪时,不同的团伙成员可能形成许多不同的网络拓扑结构.嫌疑人员之间通过通信、通话、账户转账和住宿等多种交互行为形成多个简单无向图,每个图的顶点都是同一个顶点集的子集.本文主要从顶点考虑核心团伙成员,忽略嫌疑人联系类别和形式,只关注嫌疑人员之间是否有联系出现.

对于给定的犯罪情报网络数据集,按照图 1 的方法对每个数据集进行预处理,图 2 给出了图数据集 $\boldsymbol{R}$ 的一个例子,其中涉及 7 嫌疑人的集合 $\boldsymbol{V}=\{A, B, C, D, E, F, G\}$,8 个代表不同联系方式的简单图 G_1,

G_2, …, G_8.

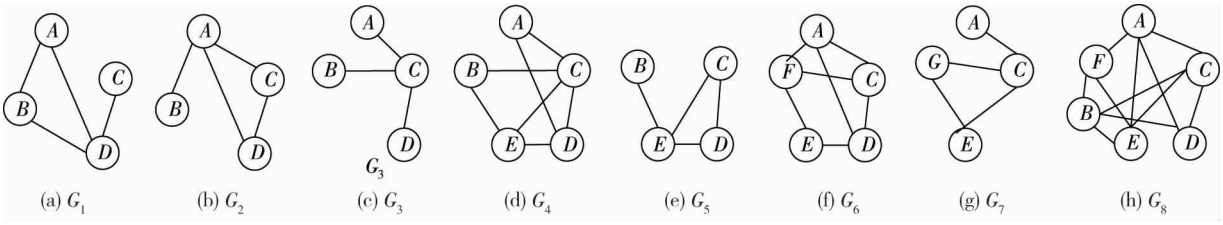

图 2　图数据集 $\boldsymbol{R}$

2.2　构建团伙集列表

将嫌疑人顶点集 $\boldsymbol{V}$ 中的每个元素按字典顺序从小到大排序,即 $\boldsymbol{V}=\{v_1, v_2, \cdots, v_n\}$,其中 $v_1<v_2<v_3<\cdots<v_n$. 设 $\boldsymbol{S}\subseteq\boldsymbol{V}$ 是 $\boldsymbol{V}$ 的顶点子集,且 $\boldsymbol{S}$ 中的顶点保持和 $\boldsymbol{V}$ 中相同的顺序,$\boldsymbol{V}$ 和 $\boldsymbol{S}$ 也分别称为有序集和有序子集.把 $\boldsymbol{V}$ 中的一个 k-团伙表示为包含 k 个顶点的有序子集 $\boldsymbol{S}$,且 $\boldsymbol{S}$ 在某个图 G_i 中的导出子图是连通子图.定义 k-团伙列表 LS_k 为

$\mathrm{LS}_k=\{<\boldsymbol{S}, f> | \boldsymbol{S}$ 是 G 的一个 k-团伙,即 $\boldsymbol{S}\subseteq\boldsymbol{V}$ 且 $|\boldsymbol{S}|=k$, f 是子集 $\boldsymbol{S}$ 出现的频次$\}$.

式中:$k=2, 3, \cdots, p$.

显然,当 $k=2$ 时,k-团伙 $\boldsymbol{S}$ 就是图 G_i 的边,即 LS_2 可由图 G_i 直接得到.当 $k>2$ 时,$(k-1)$-团伙采用连接和扩展 2 种操作可以求得 k-团伙.

2.2.1　连接

设 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 都是 k-团伙子集,若 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 有共同的$(k-1)$-子集,称 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 共核.2 个共核的 k-团伙子集可以连接成一个$(k+1)$-团伙子集 $\boldsymbol{Z}$,记为 $\boldsymbol{Z}=\mathrm{Join}(\boldsymbol{X},\boldsymbol{Y})$,且

$\boldsymbol{Z}=\mathrm{Join}(\boldsymbol{X}, \boldsymbol{Y})=\boldsymbol{X}\cup\boldsymbol{Y}$.

$\boldsymbol{Z}$ 中的元素保持和 $\boldsymbol{V}$ 中元素的顺序相同.例如,2 个共核的顶点子集$\{A, B, D\}$,$\{A, C, D\}$,通过连接操作得到$\{A, B, C, D\}$.

由于 $\boldsymbol{X}$, $\boldsymbol{Y}$ 是 2 个 k-团伙,即 $\boldsymbol{X}$,$\boldsymbol{Y}$ 在 G_i 中都是连通子集,而且它们有公共的$(k-1)$-子集,所以连接得到的子集 $\boldsymbol{Z}$ 在 G_i 中也一定是连通子集,保证了 $\boldsymbol{Z}$ 是$(k+1)$-团伙.

2.2.2　扩展

从已有候选 k-团伙 $\boldsymbol{X}$ 出发,根据 $\boldsymbol{X}$ 的有序子集的最后一个顶点 C,在 G_i(或 LS_2)中找到与 C 相邻接的顶点 $D(D>C)$,通过扩展成员得到新的$(k+1)$-团伙 $\boldsymbol{Y}$.记为 $\boldsymbol{Y}=\mathrm{Extension}(\boldsymbol{X}, D)$,且

$\boldsymbol{Y}=\mathrm{Extension}(\boldsymbol{X}, D)=\boldsymbol{X}\cup\{D\}$.

由于新引入的顶点 D 与 $\boldsymbol{X}$ 的最后一个顶点有边相连,所以 $\boldsymbol{Y}$ 是连通子集,即 $\boldsymbol{Y}$ 是一个$(k+1)$-团伙.如 3-团伙$\{A, B, C\}$,CD 是一条边,则扩展得到 4-团伙$\{A, B, C, D\}$.

上述 2 种运算总是以扩展优先,即只有当不能扩展时才考虑连接.以图 2 数据集 $\boldsymbol{R}$ 中的图 G_4为例,构建所有的 k-团伙列表 $\mathrm{LS_k}(k=2,3,4,5)$如下:

$\mathrm{LS}_2=\{<AC,1>, <AD,1>, <BC,1>, <BE,1>, <CD,1>, <CE,1>, <DE, 1>\}$;

$\mathrm{LS}_3=\{<ABC,1>, <ACD,1>, <ACE,1>, <ADE,1>, <BCD,1>, <BCE,1>, <BDE,1>, <CDE,1>\}$;

$\mathrm{LS}_4=\{<ABCD, 1>, <ABCE,1>, <ABDE,1>, <ACDE,1>, <BCDE,1>\}$;

$\mathrm{LS}_5=\{<ABCDE, 1>\}$.

因为只考虑一个图,所以每个 k-团伙出现的频次都是 1.

2.3　k-核心团伙挖掘算法 CGMA

算法的基本思想是对每个图建立候选 k-子集,同时计算每个 k-子集的频次,然后以此判定 k-核心团伙.下面给出挖掘 k-核心团伙挖掘算法 CGMA 的伪代码.

主算法:k-核心团伙挖掘算法 CGMA

输入:m 个简单图的图数据集 $\boldsymbol{R}$, n 个顶点的有序集 $\boldsymbol{V}$,团伙上限 p, 最小支持度阈值 min_σ

输出:k-团伙列表 LS_k, $k=2, 3, \cdots, p$

1.初始化 2 个团伙列表变量:LS[2:p], TempLS[2:p];

2.For $i=1$ to m do //对每个图 G_i 统计

3. $TempLS_2 \leftarrow \{s \mid s$ 是图 G_j 的边$\}$; // $TempLS_2$ 中的元素按边的大小顺序排列.

4. $LS_2 \leftarrow Add(LS_2, TempLS_2)$; //将 $TempLS_2$ 中的元素统计到 LS_2 中.

5. For $k=3$ to p do

6. $TempLS_k \leftarrow \varnothing$;

7. For each$X_{k-1} \in TempLS_{k-1}$ do

8. if X_{k-1}的最后一个顶点 C 和图 G_j 的顶点 $D(D>C)$ 相邻 then

9. $TempLS_k \leftarrow TempLS_k \cup \{Extension(X_{k-1}, D)\}$; //扩展

10. If $\exists Y_{k-1} \in TempLS_{k-1}$与 X_{k-1}共核 then

11. $TempLS_k \leftarrow TempLS_k \cup \{Join(X_{k-1}, Y_{k-1})\}$; //连接

12. $LS_k \leftarrow Add(LS_k, TempLS_k)$; //将 $TempLS_k$ 中的元素统计到 LS_k 中.

13.For $k=2$ to p do //根据 k-团伙的频次和最小支持度求出所有 k-核心团伙.

14. 删除 LS_k 中频次 $f<\min\_\sigma$ 的元素;

15.Return LS[2,$\cdots$, p].

其中调用了子程序 Add,描述如下:

子算法:$LS_k = Add(LS_k, TempLS_k)$; //将 $TempLS_k$ 中的元素统计到 LS_k 中.

输入:k-团伙及频度列表 LS_k,k-团伙临时列表 $TempLS_k$,且都已排序.

输出:更新了的 k-团伙及频度列表 LS_k.

1. $q \leftarrow LS_k$ 的第一个元素;

2. $s \leftarrow TempLS_k$ 的第一个元素;

3. While q, s 都不为空 do

4. If $q.S=s$ then //相同的 k-团伙出现,频度+1.

5. $q.f \leftarrow q.f+1$; $q \leftarrow q$ 的下一个元素; $s \leftarrow s$ 的下一个元素;

6. If $q.S<s$ then $q.f \leftarrow q.f+1$;

7. If $q.S>s$ then //列表中没有 k-团伙 s,需要添加.

8. 在 q 之前插入新元素<s, 1>; $s \leftarrow s$ 的下一个元素;

9.Return LS_k.

算法采用 2 个 k-团伙列表 LS 和 TempLS,分别用来保存最后的 k-团伙列表(包含频次)和每个图产生的临时 k-团伙列表(不含频次).

定理 1 算法 CGMA 能挖掘出所有的 k-团伙($k=2,\cdots, p$),且花费的时间不超过 $O(m \cdot n^p)$,其中 n 是顶点数,p 是常数,m 是关联图(无向简单图)的个数.

证明:首先,证明算法 CGMA 能挖掘出所有的 k-团伙($k=2,\cdots,p$).对任意的 k-团伙 $\boldsymbol{X}=\{x_1, x_2, \cdots, x_{k-1}, x_k\}$,其中顶点按顺序排列,对于最后 2 个顶点 x_{k-1}, x_k,分 2 种情况讨论:

1)顶点 x_{k-1}和 x_k 在 G_i 中相邻,且 $\boldsymbol{X}'=\{x_1, x_2,\cdots, x_{k-1}\}$是连通子集,即 $\boldsymbol{X}'$是$(k-1)$-团伙, 则 $\boldsymbol{X}$ 可以由 $\boldsymbol{X}'$通过扩展得到,即 $\boldsymbol{X}=Extension(\boldsymbol{X}', x_k)$.

2)顶点 x_{k-1}和 x_k 在 G_i 中相邻但 $\boldsymbol{X}'=\{x_1, x_2,\cdots, x_{k-1}\}$不是连通子集,或者顶点 x_{k-1}和 x_k 在 G_i 中不相邻.因为 $\boldsymbol{X}$ 是连通子集,必定存在生成树至少含有 2 个树叶结点 x, $y \in \boldsymbol{X}$,且去掉任何一个都不影响连通性.令 $\boldsymbol{M}=\boldsymbol{X}-\{x\}$, $\boldsymbol{N}=\boldsymbol{X}-\{y\}$,则 $\boldsymbol{M}$,$\boldsymbol{N}$ 都是连通子集,都是$(k-1)$-团伙, 且具有相同的核 $\boldsymbol{X}-\{x, y\}$,所以 $\boldsymbol{X}$ 可由 $\boldsymbol{M}$,$\boldsymbol{N}$ 通过连接操作得到,即 $\boldsymbol{X}=Join(\boldsymbol{M},\boldsymbol{N})$.

综上所知,所有的 k-团伙都可由$(k-1)$-团伙通过扩展或连接得到,而且 2-团伙是图集 $\boldsymbol{R}$ 中某子图 G_i 所有的边,可直接列出.因此根据归纳法可证得算法 CGMA 是正确的.

再次,证明算法 CGMA 的时间复杂度不超过 $O(m \cdot n^p)$. k-核心团伙是 k 个顶点的子集,顶点数 k 至少为 2 最多为 p,所以所有的 k-团伙($k=2,\cdots,p$)总数不超过 $C(n,2)+C(n,3)+C(n,4)+\cdots+C(n,p)=O(n^p)$.这些 k-团伙是针对 m 个图进行的,所以算法的时间复杂度为 $O(m \cdot n^p)$.

3　试验研究

通过试验对本文算法的有效性和正确性进行验证,主要从图数据集大小 m、顶点个数 n、团伙上限 p 对算法有效性的影响进行考察.算法在 Visual Studio 2013 开发环境支持下用 Visual C++实现,基于本地 Windows 调试器通过.采用的试验环境为单机 Intel(R) Core(TM) i7-5600 2.6 GHz CPU,8 G 内存,1 T 硬盘,Windows 10 Professional 操作系统.

由于一般图模式挖掘算法与本文解决的问题缺乏可比性,而且真实犯罪情报数据集存在诸多敏感数据,必须对真实数据集进行预处理后再用于试验评估.因此,本文试验首先在仿真模拟犯罪情报数据集上进行,然后在真实数据集上对算法的时间性能进行验证,最后利用文献[6]提出的关键成员挖掘分析方法 KMM(Key Member Mining)与本文算法进行比较.创建模拟数据集的主要参数:用户确定图数据集中顶点数目上限值 n 和本数据集中子图数目 m,其他数据如每个图的顶点、每个图的边数以及边均随机产生.本文主要是在图数据集大小 m、顶点个数 n、团伙上限 p 等不同参数条件下验证理论分析的结果及算法运行的效率.

3.1　模拟数据集

试验 1　考察 $m=8$, $n=6$, $p=6$ 时,以图 2 中数据集 $\boldsymbol{R}$ 为例,验证算法 CGMA 的正确性和有效性.

$k=2$, $k=3$, $k=4$, $k=5$ 的 k-团伙列表集如表 1 所示.得出的 6-团伙集 LS_6 只有(0 1 2 3 4 5) ----> Freq(1) In Graph(8),输出结果的含义为 6-团伙(0 1 2 3 4 5)在图 2 中出现的频次 Freq 为 1.为简单起见,用阿拉伯数字 0, 1, 2, 3, 4, 5 分别表示图 2 中的实体 A, B, C, D, E, F,由于 6-团伙只在 G_8 中出现,因此频次 Freq 为 1.如果规定用户最小支持度阈值 min_$\sigma=5$,则(0 2), (0 3), (2 3), (0 2 3), (0 1 2 3)均为核心团伙,且 02, 03, 23, 023, 0123 分别为 5 个核心团伙的核心成员,经人工检验证实了算法的正确性,并能得出有效结果.由表 1 可知:当 k 值不断增大时,候选团伙集的数量减少,且候选团伙集的频次急剧减少,满足阈值的核心团伙数量也随之减少.

试验 2　考察 $m=50$, $n=15$, $p=7$ 时,验证不同 k 值条件下数据集与团伙数量的关系.

团伙集数量与 k 的关系如表 2 所示.由表 2 可知:随着 k 值不断增大,k-团伙集的数量不断增加.由于 k 值是团伙人数,k-团伙的人数越多时,其出现的频次会减少,核心团伙数量也随之减少.当 min_σ 增大时,核心团伙数量急剧减少.表 2 中最小支持度阈值分别取 8, 12, 15 和 20 时,核心团伙的数量急剧减少,当 $k=5$ 时,min_$\sigma=20$ 核心团伙数量为 0,当 $k=7$ 时,min_$\sigma=15$ 和 min_$\sigma=20$ 核心团伙数量都为 0.

3.2　真实数据集

CGMA 算法的真实数据集来自于湖南神鹰实战平台数据集,该平台以湖南省公安厅大数据中心为基座,收集了大量的内网数据集(Intranet datasets)和外网数据集(Extranet datasets),其中内部公安专用网的数据集 1 008 大类,总计 1 104 亿条数据记录;外部公共网络数据集 37 大类,总计约 5 亿条数据记录.整个数据集涉及的领域主要有通信、金融、酒店服务、税务、市场监管、交通和公共资源交易等部门数据专线.本文从中提取了 9 大类犯罪网络模型,包含 53 个顶点数据标签和 320 个图的数据集 $\boldsymbol{R}$.

试验 3　考察 $m=270$, $n=53$, $p=10$ 时,在真实数据集中验证不同 k 值对算法 CGMA 执行时间的影响.

图 3 是 k 值与算法运行时间结果.由图 3 可知:在湖南神鹰实战平台数据集上,随着 k 值的不断增大,算法的运行时间不断增加,这也证实 CGMA 算法的主要时间复杂度 $O(n^p)$ 花费在寻找 k-团伙($k=2,\cdots,p$)的数量上.

设定最小支持度阈值 min_$\sigma=50\%$,$k=5$,取图数据集的规模为 $|\boldsymbol{R}|=320$.图 4 为在 2 个数据集中,不同输入图集数量对整个算法运行时间的影响.由图 4 可知:不同数据集大小的数据获得 k-频繁模式的时间非常接近,这也验证了前面算法理论分析的结果.

表 1 数据集 ***R*** 输出 k-团伙列表 LS_k

LS_2	LS_3	LS_4	LS_5
(0 1) ---->Freq(2) In Graph(1 2)	(0 1 2) ---->Freq(4) In Graph(2 3 4 8)	(0 1 2 3) ---->Freq(5) In Graph(1 2 3 4 8)	(0 1 2 3 4) ---->Freq(2) In Graph(4 8)
(0 2) ---->Freq(6) In Graph(2 3 4 6 7 8)	(0 1 3) ---->Freq(3) In Graph(1 2 8)	(0 1 2 4) ---->Freq(2) In Graph(4 8)	(0 1 2 3 5) ---->Freq(1) In Graph(8)
(0 3) ---->Freq(5) In Graph(1 2 4 6 8)	(0 1 4) ---->Freq(1) In Graph(8)	(0 1 2 5) ---->Freq(1) In Graph(8)	(0 1 2 4 5) ---->Freq(1) In Graph(8)
(0 4) ---->Freq(1) In Graph(8)	(0 1 5) ---->Freq(1) In Graph(8)	(0 1 3 4) ---->Freq(2) In Graph(4 8)	(0 1 3 4 5) ---->Freq(1) In Graph(8)
(0 5) ---->Freq(2) In Graph(6 8)	(0 2 3) ---->Freq(6) In Graph(1 2 3 4 6 8)	(0 1 3 5) ---->Freq(1) In Graph(8)	(0 2 3 4 5) ---->Freq(2) In Graph(6 8)
(1 2) ---->Freq(3) In Graph(3 4 8)	(0 2 4) ---->Freq(3) In Graph(4 7 8)	(0 1 4 5) ---->Freq(1) In Graph(8)	(1 2 3 4 5) ---->Freq(1) In Graph(8)
(1 3) ---->Freq(2) In Graph(1 8)	(0 2 5) ---->Freq(2) In Graph(6 8)	(0 2 3 4) ---->Freq(3) In Graph(4 6 8)	
(1 4) ---->Freq(3) In Graph(4 5 8)	(0 2 6) ---->Freq(1) In Graph(7)	(0 2 3 5) ---->Freq(2) In Graph(6 8)	
(1 5) ---->Freq(1) In Graph(8)	(0 3 4) ---->Freq(3) In Graph(4 6 8)	(0 2 4 5) ---->Freq(2) In Graph(6 8)	
(2 3) ---->Freq(7) In Graph(1 2 3 4 5 6 8)	(0 3 5) ---->Freq(2) In Graph(6 8)	(0 2 4 6) ---->Freq(1) In Graph(7)	
(2 4) ---->Freq(4) In Graph(4 5 7 8)	(0 4 5) ---->Freq(2) In Graph(6 8)	(0 3 4 5) ---->Freq(2) In Graph(6 8)	
(2 5) ---->Freq(1) In Graph(6)	(1 2 3) ---->Freq(4) In Graph(1 3 4 8)	(1 2 3 4) ---->Freq(3) In Graph(4 5 8)	
(2 6) ---->Freq(1) In Graph(7)	(1 2 4) ---->Freq(3) In Graph(4 5 8)	(1 2 3 5) ---->Freq(1) In Graph(8)	
(3 4) ---->Freq(3) In Graph(4 5 6)	(1 2 5) ---->Freq(1) In Graph(8)	(1 2 4 5) ---->Freq(1) In Graph(8)	
(4 5) ---->Freq(2) In Graph(6 8)	(1 3 4) ---->Freq(3) In Graph(4 5 8)	(1 3 4 5) ---->Freq(1) In Graph(8)	
(4 6) ---->Freq(1) In Graph(7)	(1 3 5) ---->Freq(1) In Graph(8)	(2 3 4 5) ---->Freq(2) In Graph(6 8)	
	(1 4 5) ---->Freq(1) In Graph(8)		
	(2 3 4) ---->Freq(4) In Graph(4 5 6 8)		
	(2 3 5) ---->Freq(1) In Graph(6)		
	(2 4 5) ---->Freq(2) In Graph(6 8)		
	(2 4 6) ---->Freq(1) In Graph(7)		
	(3 4 5) ---->Freq(1) In Graph(6)		

表 2 团伙集数量与 k 的关系

k 值	2	3	4	5	6	7
k-团伙数	105	455	1 365	3 003	5 005	6 435
min_σ=8	105	455	1 365	2 979	4 676	4 987
min_σ=12	105	455	1 199	1 218	563	140
min_σ=15	105	391	405	108	7	0
min_σ=20	79	59	4	0	0	0

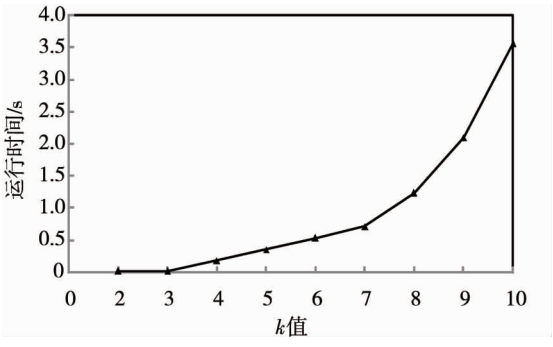

图 3 k 值与算法运行时间

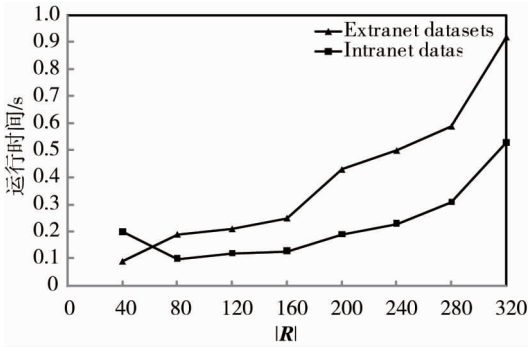

图 4 内外网数据集中的算法运行时间

为了进一步验证本文提出的 CGMA 算法的有效性,在实际数据集上取 $p=10$,比较 CGMA 算法和文献[6]中提出的 KMM 算法的时间性能.图 5 为在不同 k 值下 CGMA 算法和 KMM 算法的性能比较.由图 5 可知:在几乎所有情况下,尽管数据集的加速比不同,CGMA 算法都比 KMM 算法的运行时间更短.

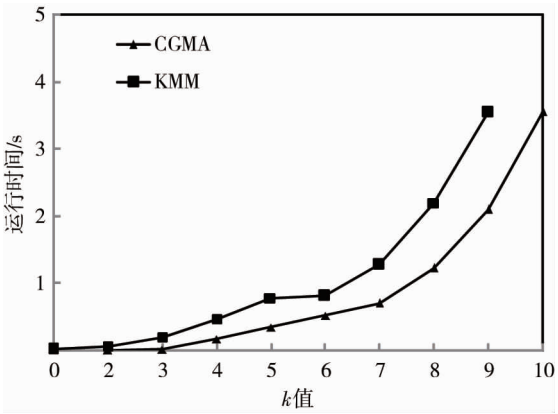

图 5　不同 k 值下 CGMA 算法和 KMM 算法的性能比较

最后根据情报数据的关联,从试验结果的 k-频繁模式中总共输出 2 360 条关联信息,满足用户指定阈值的有 337 条,每条信息都得出了不同数量的团伙和核心成员.这些信息可为监测预警、犯罪形势分析报告、行业风险评估报告和维稳形势分析报告等提供依据,同时也可为制定防范化解金融风险对策、打击经济犯罪提供参考.

4　结论

1)构建的候选核心团伙集数据结构可保证在图数据集挖掘,避免普遍的子图同构带来的 NP 困难问题.基于顶点集处理大型图数据集,实现数据采集、预处理和计算分析,为实现频繁子图结构在犯罪团伙挖掘中的应用提供理论支撑.

2)通过阈值完成核心团伙的特征分析,找出犯罪团伙顶点,认为犯罪团伙的顶点频度与核心成员的预测指标有较好的一致性.

3)提出连接和扩展两个操作算子,构建了产生完整候选集的模型,可以根据用户的需求调整团伙大小.所提出的犯罪团伙挖掘算法不需要执行任何子图同构测试,通过选择不同的 p 值,用户可以找到一组大小不同的 k-核心团伙.

参考文献:

[1] 胡公枢.集团性网络诈骗各参与人犯罪数额的认定[J].中国检察官,2021(18):52-55.

[2] DWIVEDI A, SINGH C, MISHRA A, et al. Comparative analysis of data mining in criminal and fraud detection[J]. International Journal of Psychosocial Rehabilitation, 2020, 24(6):1449-1460.

[3] IFTIKHAR A, JAFFRY S A, Malik M K. Information mining from criminal judgments of Lahore High Court[J]. IEEE Access, 2019, 7:59539-59547.

[4] MALEKAR M. Detecting criminal activities of surveillance videos using deep learning[J]. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 2021,7(1):188-193.

[5] 李万彪,余志,龚峻峰,等.基于关系数据模型的犯罪网络挖掘研究[J].中山大学学报(自然科学版), 2014,53(5):1-7.

[6] 李勇男.基于子图模式的反恐情报关联图集分析[J].现代情报,2019,39(7):37-43.

[7] ACOSTA-MENDOZA N, CARRASCO-OCHOA J A, MARTÍNEZ-TRINIDAD J F, et al. Mining clique frequent approximate subgraphs from multi-graph collections[J]. Applied Intelligence, 2020,50:878-892.

[8] JAYALAKSHMI N, PADMAJA P, SUMA G J. An approach for interestingsubgraph mining from web log data using w-gaston algorithm[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2019, 27(2):277-301.

[9] LEE G, YUN U. Mining frequent graph patterns considering both different importance and rarity of graph elements[M]//PARK J, STOJMENOVIC I, JEONG H Y, et al. Computer Science and its Applications. Heidelberg: Springer, 2015:179-184.

[10] LEE G, YUN U, KIM D. A weight-based approach: frequent graph pattern mining with length-decreasing support constraints using weighted smallest valid extension[J]. Journal of Computational and Theoretical Nanoscience, 2016, 22(9):2480-2484.