基于海量数据的异常交易研究

摘　要

近年来，随着我国普惠金融的发展，贷款欺诈行为屡见不鲜，贷款归集现象较为严重，为商业银行风险管理迎来新的挑战。

本文通过对银行卡交易流水样本数据进行分析，综合运用社交网络、知识图谱的理论算法，运用大数据可视化工具，寻求对具有异常交易的资金归集群体进行捕捉，通过构建异常交易网络模型的方式对异常贷款行为进行分析，为信用风险管理的工作提供辅助。

通过构建策略模型工具的方式，建立识别金融交易属性中资金归集特性的欺诈行为，力求模型兼具实用性与创新型，对实际应用有一定的指导作用。

关键词：大数据；可疑交易；资金归集；社交网络；知识图谱

第 1 章 可疑交易分析的价值与意义

随着零售信贷业务的发展进步，线上自动化业务逐渐成为一种新的金融产品模式。这种由线下转为线上、由纸质人工变为数字自动、由服务渠道单一化转为多样化、由大众标准服务转为个性体验服务的模式转变，带来的不仅是客户体验的提升与效率的增加，随之而来的还有异常丰富的数据，逐步形成了多渠道多维度的海量数据。

因此，以新技术与海量数据为驱动的风险管理的模式已经得到快速的发展，互联网开放、分享、去中心化的特点可以提供更好的客户体验，但在实际的风险管理中，互联网客户的欺诈行为也较为严重，其中比较突出的现象之一，就是以贷款归集现象为表现的异常交易行为，为商业银行的风险管理带来了新的挑战。

如何从海量的数据中筛选出异常交易的数据，并对信息进行管理是一个难题。异常交易的突出特征是资金的异常归集行为，虚假的交易流水与

违规的贷款用途将增加风险管理的难度。通过大数据技术，结合风险管理经验，充分挖掘海量的数据中蕴含的特征信息，对人工分析难以捕捉的价值信息进行自动化识别，降低风险管理成本，提高精确识别能力，对商业银行的贷款管理具有重大的意义。

第 2 章 可疑交易网络构建

2.1 策略研究与数据准备

通过对某行四个地区的某年度共计 7495.78 万的数据进行分析挖掘，各地区交易流水数据量及占比如图 2.1 所示，可以看出交易流水数据量已达到千万级别，因此传统的数据分析及挖掘方法已难以解决，需要利用计算机大数据分析方法，基于海量数据的挖掘算法对数据进行进一步清洗、建模与挖掘分析。

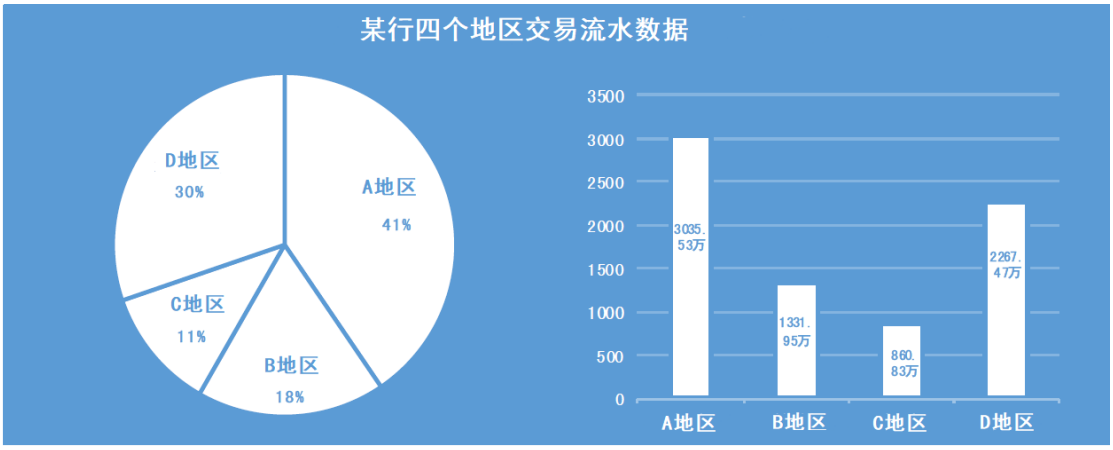

图 2.1　某行四个地区交易流水数据量及其占比

通过业务上对异常交易的特征定义，对交易流水数据分析，捕捉其中交易的关联性，从而作为构建复杂网络的基础。第二步为对海量数据的清洗，清洗原则包括将资金流入、资金流出、交易时间、交易对手、交易金额、交易频率、交易分类等十几个维度纳入考量，整理出三十五项清洗规则，从海量的资金交易流水中筛选出异常的交易，通过多维度的考量数据相关性、交易对手特征、交易时间范围、交易金额特征等，捕捉其中高风险的交易。最后，从高风险的异常交易流水中清洗出可疑的交易流水。

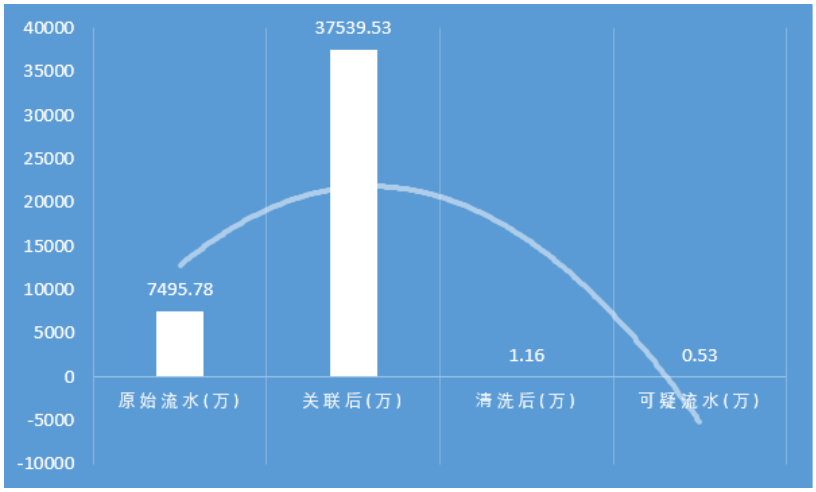

图 2.2　流水清洗数据量变化表

数据量变化范围如上图 2.2 所示，可以通过趋势线看出，经过复杂数据关联后，数据量由原始流水最初的 7495.78 万迅速增长了 5 倍，达到了 3.75 亿的数据量，通过清洗模型的清洗，异常交易的数据被筛选出来，最终获得的可疑流水约 0.53 万。流水的清洗工作是通过 SAS 工具完成的，得到了字段含义清晰、完整且规则的数据，为下一步构建可疑交易网络模型做准备。

2.2 欺诈网络分析模型构建

根据清洗模型的清洗结果，依据社交网络中的流-组算法（Stream-Group）进行建模和分析，并通过可视化的工具进行展示，以便更加直观的展示给风险管理人员，为其判断客户的风险提供数据上的支撑与依据。

2.2.1 建模原理

首先，需要做以下几点内容需要明确：（1）同一人名下有可能会对应多个卡号；（2）所有人的卡号视为一个集合，并包含在研究的数据集中；（3）多对一的归集和一对多的归集行为，对于我们研究的可疑交易，转出方与转入方是相反的。

Node(V)：节点。将每一个同一卡号的持卡人定义为一个节点。

Edge(E(t))：边。若两个节点之间有转账记录，则说明两者之间有关系，将两者之间的阶段用带时间的 t 的有向边进行标记，方向由转出方指向转入方。

Weight of Edge：边的权值。根据不同的情况，定义不同形式的边。如果两

个节点之间具有关系，可以通过定义边的权重的方式进行转账关系频率的表示。即两节点转账频繁，则边的权值大。若要获取转账金额与还款金额之间的关系，我们会定义两者的比率为边的权重，更加直观的展示出其贷款的还款金额与他人为其转账的金额之间的关系。

流-组(Stream—Group)算法在有向图挖掘上具有良好的效率。其流程大体为：首先，采用 S-Group 算法发现最新网络的社区结构；其次，计算最新网络的划分I^x与以当前网络图分割S^x的划分I^x的相似度；最后，根据划分的相似度和指定的阈值C_0判断是否出现变化点，如果时间片 t 不是变化的点，那么采用 Inc-Group 算法更新网络图分割S^x的划分I^x，否则开启一个新的网络图分割S^{x+1}。对于图分割矩阵S^x,假设有那个节点，那么图的矩阵表示如下：

$$S^{\mathrm{x}}=\begin{pmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \vdots \\ v_{m1} & \cdots & v_{mn} \end{pmatrix} \tag{2-1}$$

其中，

$$v_{ij}=\begin{cases} w_{ij}(i,j)\in E, w_{ij}=\dfrac{\sum w_{ij}(t)}{y} \\ 0,\text{其他} \end{cases} \tag{2-2}$$

给定图 G 的子图G',R 是 G 的关联矩阵，则G'的紧密度计算如下：

$$\mathrm{C}\big(\mathrm{V}(G')\big)=\frac{1}{B}\left(\sum_{i,j\in V(G')} r_{ij}-\frac{\sum_{i\in V(G')} r_i\times\sum_{j\in V(G')} r_j}{B}\right) \tag{2-3}$$

算法的详细过程、矩阵与图的计算，由于篇幅所限，暂不做详细介绍。

2.2.2 模型构建与可视化

基于以上原理，对数据进行建模，本课题的可视化构建工具采用的是 Geghi 0.8.2 beta 版对数据进行展示的，根据每一个账户之间的转账关系进行聚类，为了更加直观的对数据进行展示，再根据已经处理好的数据中的每一个节点的入度和出度，以及边的权值，对节点与边进行处理。

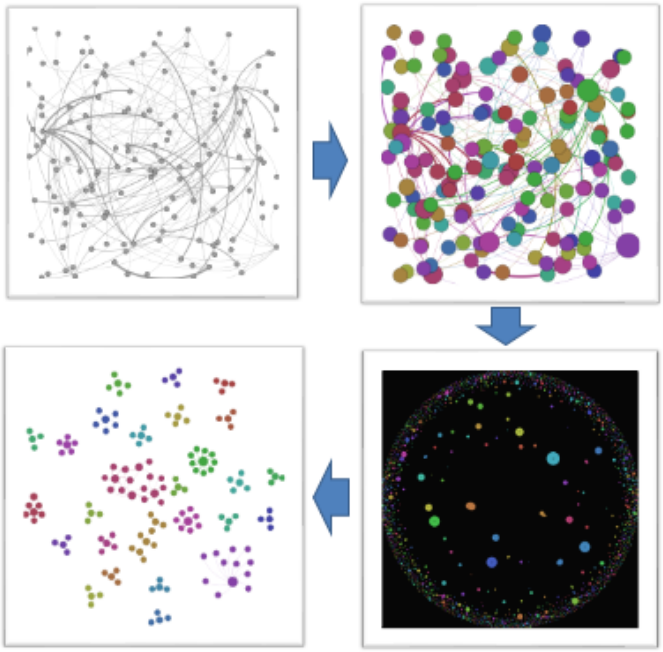

图 2.3　模型构建过程

如图 2.3 所示，为模型构建的流程，数据由最初的散点分布最终聚类成为各个聚簇，并根据各群体特征使用不同的颜色进行标记。图中左上图为初始阶段为进行聚类的数据点分布；右上图为根据节点的度与边的权值进行数据预处理，标记为不同的颜色；右下图为使用胡一凡算法进行聚类与数据布局；最终得到左下图所示的聚类簇，即通过算法与可视化工具获得了每一个进行资金归集的可疑群体，为了更加清晰的对捕捉的可疑群体进行展示，通过 Fruchter Atlas 算法对可疑数据进行重新布局，获得下图。

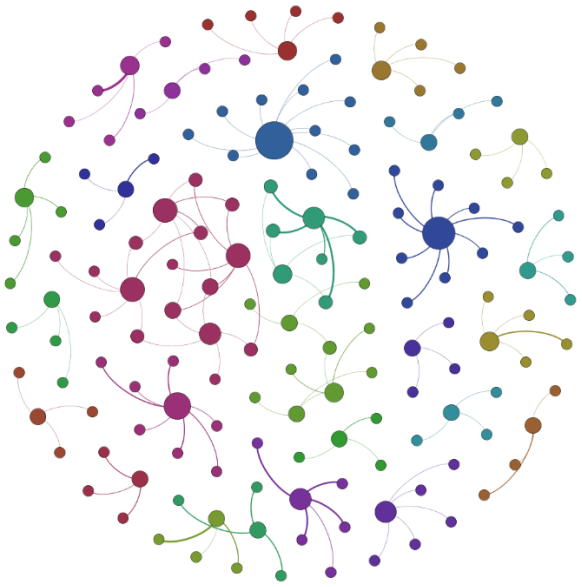

图 2.4　Fruchter Atlas 算法布局结果

如图 2.4 所示，采用 Fruchter Atlas 算法进行布局，获得了可直观展示资金流向的布局图。由上图，以展示的一对多的资金流水归集方式为例，每一种颜色的节点代表每一可疑的交易群体，在每个群体中，可疑交易的资金归集人为中心节点，其节点相对较大，边界点为可疑的交易客户。由于数据保密性要求，将客户信息进行隐匿，仅作结果展示。

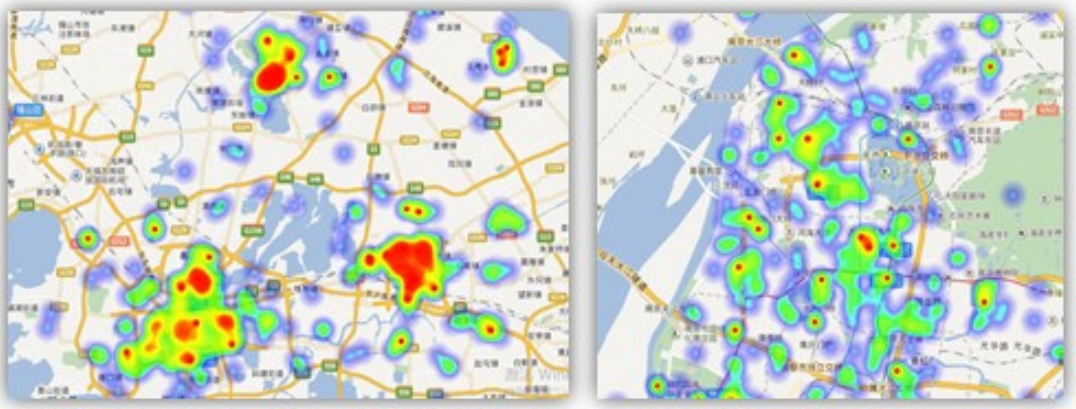

图 2.5　A 地区、C 地区可疑交易流水分布热图

另外，通过可疑交易数据可以获得可疑客户的地址范围信息，如图 2.5 所示，根据可疑交易客户的地址信息，我们可以可视化的展示出可以用户的主要地理位置，图中左半部分为 A 地区的可疑流水热图，右半部分为 C 地区的可疑流水热图，该图是根据地址获得经纬度，并通过可视化工具完成的，实现了以大数据的方式多维度的刻画客户风险。

第 3 章 可疑交易案例分析

可疑交易分为两种情况，其一为一对多的归集行为，其二为多对一的归集行为。如图 3.1 所示，为多对一归集资金的一个案例。此案例归集的资金量较为庞大，但是归集情况并不复杂。如图所示，以姚某为主等 11 人，在贷款每月还款日前均得到了姚某的转账用来归还其贷款本息，而且资金量较为庞大，该可疑归集圈共计归集资金约为 1513 万，若该可疑交易圈的资金归集确认为欺诈或者贷款挪用行为，其资金链一旦断裂，其可能在商业银行产生逾期或者造成不良，以至于造成损失。因此，对模型捕捉构建的类似该归集圈的交易行为需要进行重点的排查，对该归集圈涉及的贷款项目做进一步的排查和确认。

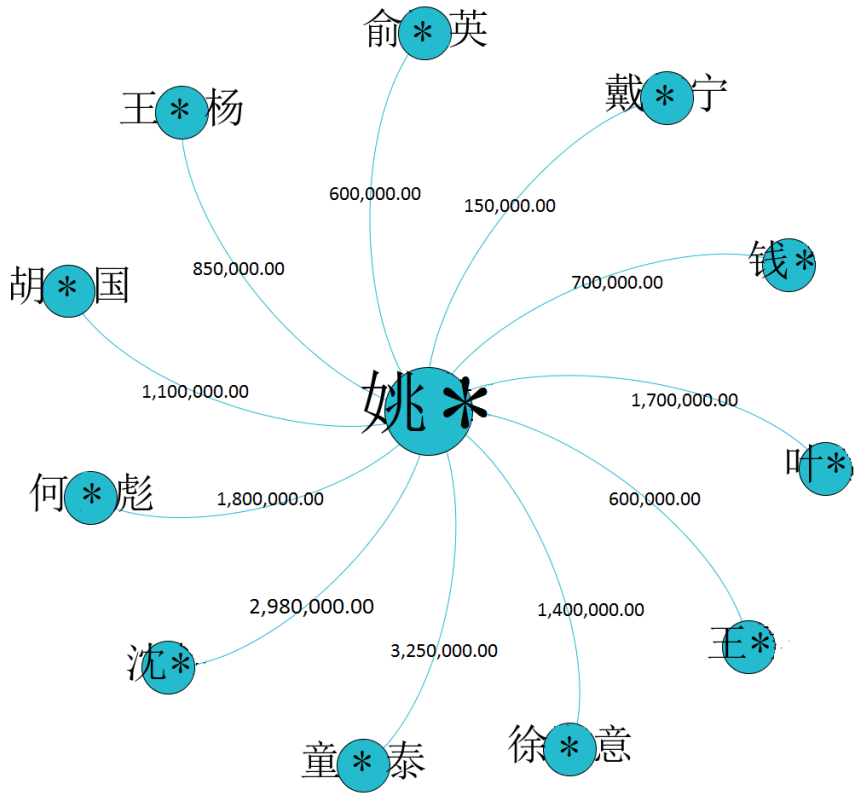

图 3.1　一对多归集案例

多对一进行资金归集的特征相对较好确认，因为其归集方式较为直接，通过交易直接进行归集，其特点是客户在获得贷款后，对放款金额进行转移，且转移的方式简单直接，可以通过数据直接建模获得。

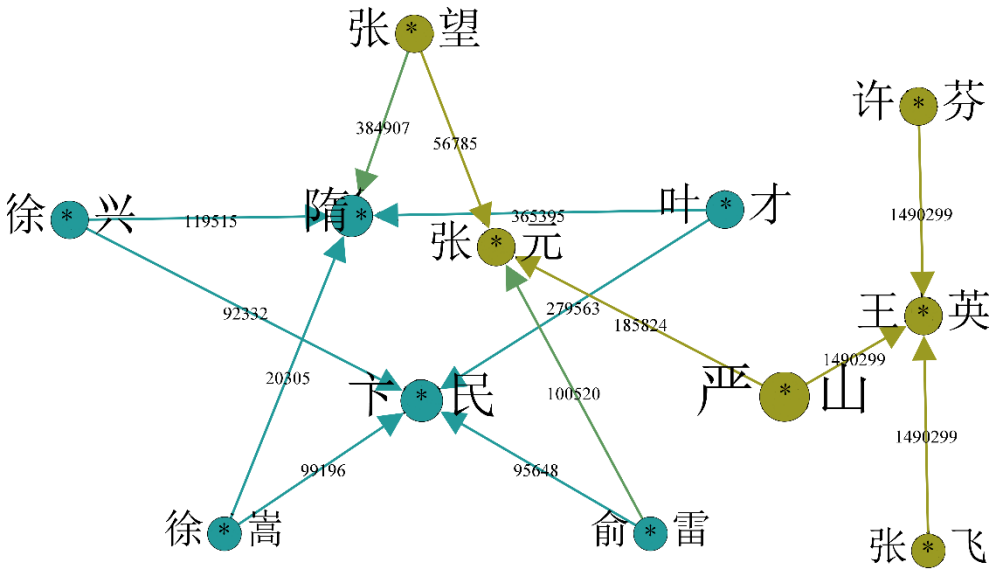

图 3.2　多对一归集案例图

如图 3.2 所示，为多对一进行转账归集的可疑群体，箭头所指向的节点即为归集方，我们可以看出主要的四个归集节点分别为卞某民、隋某、张某元和王某英，归集金额较为庞大，共计 627.1 万元。为了更位清晰的对其归集的情况进行掌握，我们可以继续对该网络进行放大，以卞某民为例，如图 3.3 所示，我们可以获得图 3.2 网络的第二层网络图。即为卞某民转账的四个人，卞某民接受转账为四张不同的账户卡，这四张卡均为卞某民名下的账户，但最终都归集到卞某民一人名下，总计金额为 31.57 万元。

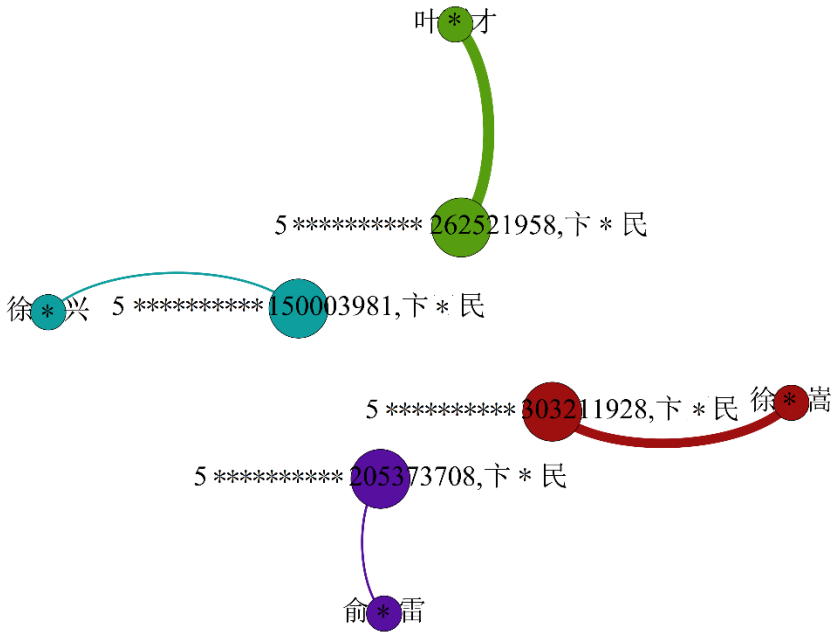

图 3.3　多对一归集案例第二层细化图

如有需要，模型仍然支持继续对该层网络进行放大，如下图 3.4 所示，展示出了为卞某民转账的用户的转账明细，四种颜色分别代表四个不同的账户，边上标有转账的金额，叶子节点标有转账人的姓名以及其转账时间，可以清晰的获得客户进行可疑的归集的时间、金额等信息数据。

由此，我们可以通过构建自动化模型的方式，以关联交易为线索，自动化的捕捉海量的数据中的线索资源，识别可疑的异常交易，大大降低手工工作量，解放人力、降低成本，而且模型的准确性也大大高于手工筛选，工作效率远高于人工处理。通过将经验模型化落地，也将大大的降低操作风险。

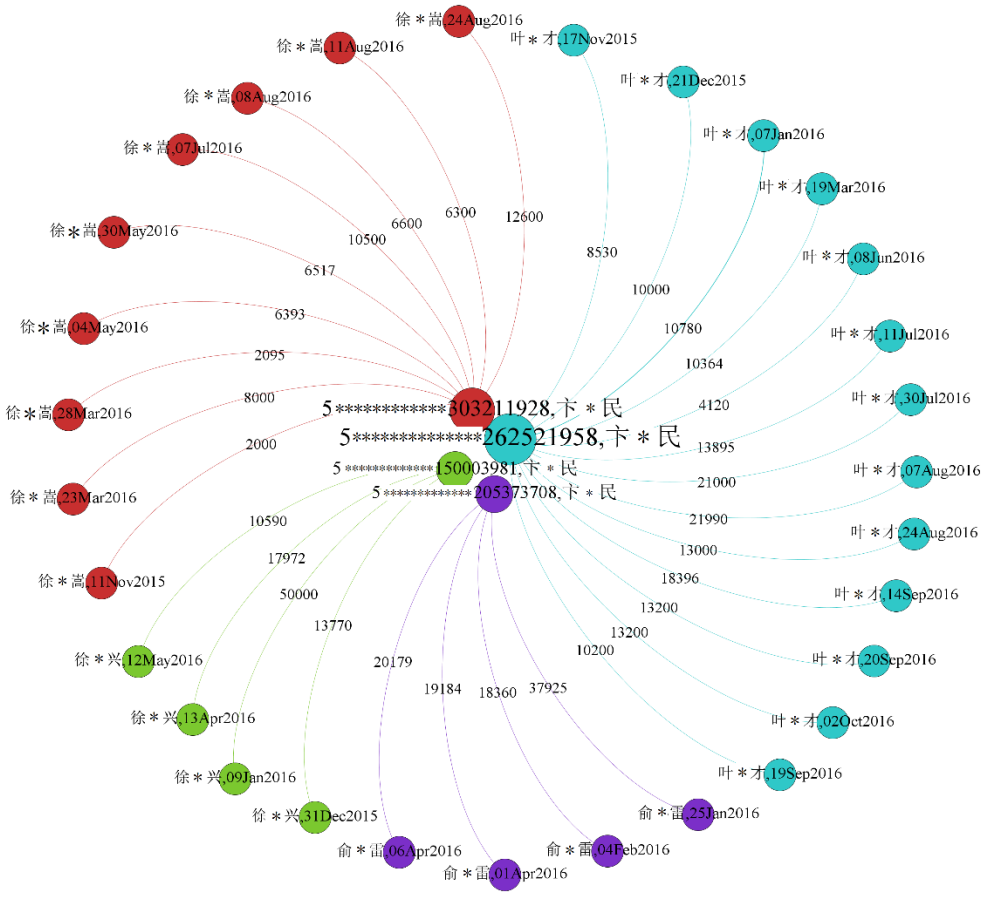

图 3.4　多对一归集案例第三层细化图

第 4 章 结论及展望

本文通过对海量交易数据的整理与分析，综合运用大数据工具与算法，结合社交网络、知识图谱等前沿理论，建立了一个具有实际应用价值的欺诈网络分析模型，以捕捉在银行信贷中的异常交易行为，通过构建社交网络的方式，识别欺诈性的资金归集的群体。同时，模型还支持网络的逐层拆解与放大，为商业银行信贷风险管理提供辅助工具，具有一定的实用价值。

由于篇幅所限，本文依然存在诸多不足之处，下一步拟对捕捉到的资金归集流水的特征做进一步提取与分析，以便下一步运用机器学习的算法进行优化，通过历史的交易行为捕捉其交易欺诈的行为特征，如交易的时间、地点以及行业特征、交易频率等，从而生成衍生变量，进一步的做欺诈分析。

随着欺诈手段的快速演变与提升，识别可疑异常交易的难度也在不断提升，通过构

建大数据机器学习模型并快速迭代应用于实际业务，将理论与实践将结合，动态、科学、精准的识别欺诈行为，进一步提升风险管理的有效性，以数据和技术驱动风险管理，助推我国普惠金融的发展。

参考文献

[1] 陈为. 大数据可视化与可视化分析. 技术应用， 2015, 11(2):1-7.

[2] 扬媛媛. 社交网络中的关系构建和亲密度分析. 计算机研究与发展, 2009(11):92-96.

[3] 刘峤, 李杨, 段宏, 等.知识图谱构建技术综述. 计算机研究与发展, 53(3):582-600 , 2016

[4] 先兴平, 吴涛. 知识图谱与网络表示学习. 产业与科技论坛, 15 (17):61-62 , 2016