基于社交网络（SNA）的数据挖掘和商业银行数字化营销

周学春[1] 王茜[2]
（1.中国民生银行，北京 100031;清华大学经济管理学院，北京 100084; 2.中国民生银行，北京 100031；）

摘要：数字化营销是指以数据挖掘技术为基础的客户关系管理、维护和提升。商业银行积累了海量的客户之间的转账交易数据，奠定了基于资金交易搭建客户关系网络的数据基础。采用数据挖掘（DM）和社交网络挖掘技术（SNA），论文首次基于商业银行零售客户之间转账交易数据，通过流水的清洗、有效边关系的界定、异常处理、圈子识别、圈子切割，有效实现客户关系网络图谱的构建和展示（三度人脉图和弱连通遍历图）。另一方面，基于客户社交网络关系图谱，论文创新探讨了以下几个方面的数字化营销应用。包括，圈子细分和聚类、客户网络影响力指数的测算、基于意见领袖的产品扩散研究、基于圈子的产品推荐指数分析、基于圈子的风险传染分析、基于圈子的客户提升指数分析。

关键词：数字化营销；社交网络；意见领袖识别；产品扩散；推荐指数；风险传染；

Abstract：Digital marketing is basing on data mining technology to perform customer relationship management. There is huge data in commercial bank, such as transaction data, this data is useful to build customer social network. Basing on data mining and social network analysis technology, this paper try to build customer relationship network through data cleaning, outlier processing, group detection, group cut. And this paper realizes the visualization of customer relationship network. On the other hand, basing on social network, this paper discuss some marketing application, such as group clustering, measurement of the index customer influence, product diffusion base on opinion leader, risk diffusion basing on group, and the lift index of customer asset.

Key word：Digital Marketing; Social Network Analysis; Opinion leader; product diffusion; recommendation index; risk diffusion;

基金项目：国家自然科学基金（71172212）、教育部人文社科青年基金（14yjc630197）

作者简介：
周学春（1986-，通讯作者），男，安徽滁州人，管理学博士、博士后，中国民生银行&清华大学经济管理学院；研究方向：商业银行&数据挖掘。电话：18511865725；邮箱：wdxiaochun@126.com。
王茜（1991-），女，北京人，中国民生银行总行小微金融部，硕士，研究方向：小微金融。

基于社交网络（SNA）的数据挖掘和商业银行数字化营销

摘要：数字化营销是指以数据挖掘技术为基础的客户关系管理、维护和提升。商业银行积累了海量的客户之间的转账交易数据，奠定了基于资金交易搭建客户关系网络的数据基础。采用数据挖掘（DM）和社交网络挖掘技术（SNA），论文首次基于商业银行零售客户之间转账交易数据，通过流水的清洗、有效边关系的界定、异常处理、圈子识别、圈子切割，有效实现客户关系网络图谱的构建和展示（三度人脉图和弱连通遍历图）。另一方面，基于客户社交网络关系图谱，论文创新探讨了以下几个方面的数字化营销应用。包括，圈子细分和聚类、客户网络影响力指数的测算、基于意见领袖的产品扩散研究、基于圈子的产品推荐指数分析、基于圈子的风险传染分析、基于圈子的客户提升指数分析。

关键词：数字化营销；社交网络；意见领袖识别；产品扩散；推荐指数；风险传染；

Abstract：Digital marketing is basing on data mining technology to perform customer relationship management. There is huge data in commercial bank, such as transaction data, this data is useful to build customer social network. Basing on data mining and social network analysis technology, this paper try to build customer relationship network through data cleaning, outlier processing, group detection, group cut. And this paper realizes the visualization of customer relationship network. On the other hand, basing on social network, this paper discuss some marketing application, such as group clustering, measurement of the index customer influence, product diffusion base on opinion leader, risk diffusion basing on group, and the lift index of customer asset.

Key word： Digital Marketing; Social Network Analysis; Opinion leader; product diffusion; recommendation index; risk diffusion;

1 数字化营销和基于交易的客户社交网络

（1）数字化营销的内涵

随着互联网的发展，以支付宝为代表的互联网金融产品不断加剧着对商业银行客户资源的争夺。为了应对互联网金融的冲击，商业银行数字化营销战略的启动有其现实意义。

数字化营销（Digital Marketing）是商业银行数据挖掘中常常会提到的一个概念。一般来说，数字化营销包含两个元素，即顾客关系管理（CRM）和数据挖掘（DM）。本质上，所谓数字化营销，它是指，基于数据挖掘技术，商业银行以客户为中心，开展和实施的客户获取、客户保持、客户维护和客户挽留等客户关系管理行为。换言之，就是把数据挖掘技术应用到传统的客户关系管理领域，借助于统计和机器学习的技术，对客户进行精细化管理。

换言之，商业银行的数字化营销，就是以数据仓库为基础，通过一定的技术和方法（统计算法和机器学习算法），从中提取出隐藏的、有价值的信息和知识，找出数据中呈现出的规律，从而能够解释已知的事实，预测未来的客户行为模式，有效辅助营销人员的客户关系管理。

（2）数字化营销和社交网络

“客户圈子营销”是当下较为前沿和热门的研究课题。以往商业银行在做客户维护和提升时，关注的仅仅是某个单一客户个体。但是，社会网络理论告诉我们，客户是嵌入在各种各样社会网络之中的，客户之间存在着关联网络。因此，客户圈子营销，就是把客户及其关联网络中的客户作为一个群体，进行整体营销，从而能够提高营销效率，提高客户的保持率。

商业银行沉淀了海量的客户之间转账交易数据，客户之间的转账实际上构成了客户之间的边关系，因此，基于客户之间的真实交易转账数据，可以构建以客户为节点的社会关系网络。基于该关系网络的数据挖掘，有助于识别意见领袖、基于圈子进行产品的扩散、测算产

品推荐指数、识别不同圈子类型、有效判断圈子的风险，从而有助于推动和加强基于数据挖掘的数字化营销落地应用。

2 关系网络的内涵和定义

（1）关系网络的内涵

关系网络（Relationship Network），又称社交网络（Social Network），它指的是，社会中个体与个体之间联系的集合，由点（个体）和各点之间的连线（个体之间的联系）组成的。从本质上来说，关系网络是个体为达到特定目的、在个体与个体之间进行信息交流和资源利用网络。社会网络体现着一种结构关系，它反映了行动者之间的社会关系。此外，通过这些边关系，个体之间传递物质、信息、观念、情感等资源。

（2）关系网络和六度分割理论

关于关系网络的一个最为典型的案例就是，六度分割理论。上世纪七十年代，社会心理学家想要通过陌生人之间的邮件转发来测定任意两个陌生人之间的关系距离。通过实验研究，学者们发现，平均通过六次转发，信件就可以从 A 陌生人到达 B 陌生人。这就是所谓的六度分割理论。它的推论是，任意两个陌生人之间想要发生联系，他们之间的间隔不会超过六个人。换言之，信息在一个关系网络中的扩散是比较迅速的，任意两个节点之间的最大链长远远小于群组的规模。

（3）基于关系网络进行营销的基本原理

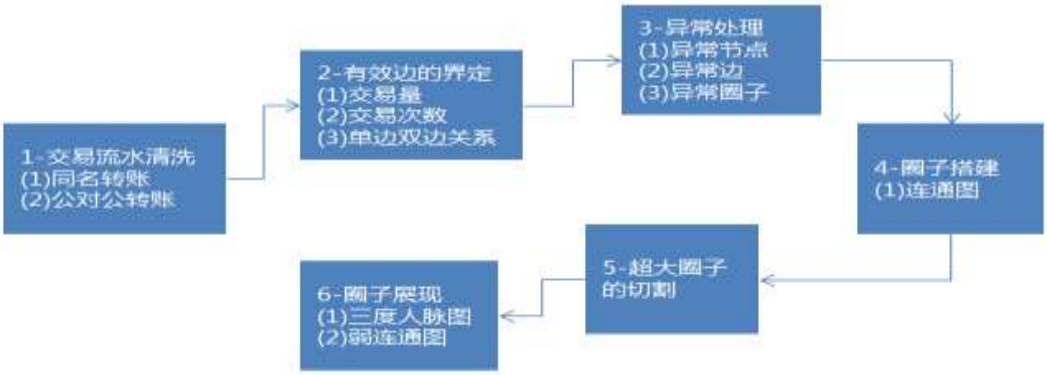
个体和个体之间形成关系网络的方式多种多样。例如，通过公司的邮件系统可以勾勒出员工之间信息流动网络。例如，基于网络的超链接，可以勾勒出网站之间的关系网络图谱（Relationship Network Map）。例如，通过人们之间的电话的呼入和呼出，能够勾勒出人际社交网络。例如，通过微博上的关注关系，也可以迅速勾勒出用户之间的关联网络。通过对社交网络的分析，可以从中抽离出很多的群组和社团（community）。根据“物以类聚人以群分的基本原理”，通常而言，处于同一个关系网络中的个体，他们之间的互动沟通较为频繁，行为特征上常常会呈现出一定的相似性。这个理论假设就是后续基于关系网络营销应用的基础。

（4）关系网络和客户之间的转账交易

传统商业银行中沉淀了大量的客户之间的转账交易关系，以某股份制银行为例，每个月客户之间的交易明细大约有一亿条。以转账关系为基础，当客户之间存在一笔转账交易记录时，我们就会认为客户之间形成一条边关系。并且，一般而言发生资金交易的双方，通常属于较强的关系，否则不会存在金钱上的往来。基于近半年的转账交易记录，我们就可以刻画出客户之间的资金交易关系网络。此外，基于圈子中成员行为的相似性和同质性，我们就可以进行相关的营销和应用。

3 基于资金交易的客户间关系网络的搭建

客户之间存在着转账交易关系，如何基于这样的数据，构建客户之间的资金交易网络？基于 SAS 和 R 软件进行海量数据的分析和挖掘，通过数据探索，我们认为关系网络的搭建包括如下几个步骤。即资金交易流水的清洗、有效边关系的界定、异常处理（异常点、异常边、异常圈子）、圈子识别、超大圈子的切割、圈子展现。



（1）资金交易流水清洗

第一，我们基于客户之间的交易流水构建客户资金交易关系网络，当客户和客户之间存在转账交易时，我们会认为两个客户之间存在一条边。但是，在界定有效边关系之前，我们需要进行交易流水的清洗。

第二，清洗的流水包括，剔除同名转账交易、剔除系统批量动账交易行为（非客户发起）、剔除公对公转账、剔除对公共户的转账（容易导致异常节点和异常圈子）、剔除交易对手为空的转账、剔除交易对手行为空的转账交易等。

（2）有效边的界定

第一，边是构建关系网络的基础，那么什么样的交易行为才算是有效的边关系？需要通过数据探索和数据分布界定一个标准。

第二，我们基于三个指标探索“有效边关系”的门槛，即某条边（客户和客户之间的近半年的交易汇总）的交易次数、交易金额、交易的单双边性（单向交易、有来有往的双向交易）。在这里我们引出“边关系稳定性规则”。例如，按照一定的通过上半年的数据搭建的变关系，通过设定一定的规则，有多少比例的边，仍然出现在下半年的边关系中。也就是说，通过探索一定的规则，使得边关系的保留率最高。边关系的保留率高，意味着后续搭建的圈子的稳定性也比较高，圈子成员之间的同质性和相似性也就越高，后续进行营销应用的结论也就相对可靠。

第三，基于数据探索，采用半年的时间窗口，我们发现近半年内的交易次数大于等于 2 次，或交易金额大于 2000，或双边关系，这样的边关系，稳定性最高，在下半年的边关系构建中保留率最高。因此作为有效边的门槛设定。

（3）异常处理

第一，基于流水清洗和有效边确认后，我们需要进行初步的圈子属性探索，包括节点属性、边关系属性、圈子属性。并且需要剔除较为异常的节点、边关系、圈子。

第二，异常节点。如果节点的点度中心度（即节点的出度和入度之和）较为异常时，可以考虑该节点的删除。例如，淘宝店主会更很多人发生转账关系，但是这种关系属于弱交易关系，不利于稳定圈子的搭建，对于这样的节点，应该进行剔除。在这里，我们采用均值标准差的思路识别异常节点。

第三，异常边。针对节点（客户）-节点（客户）之间的交易量、交易次数，呈现出异常特征时，我们也会考虑进行剔除。在这里，我们通过两种方式识别异常的边，即均值标准差思路和聚类思路。

第四，异常圈子。初步勾勒圈子时，对于圈子规模异常大的关系网络，需要分析这种超大异常圈子的形成原因，如果没有发现异常的要素，则需要考虑进行圈子的切割。

（4）圈子搭建

第一，基于连通图原理。当针对交易流水数据清洗完毕，并且处理了异常的节点、边关

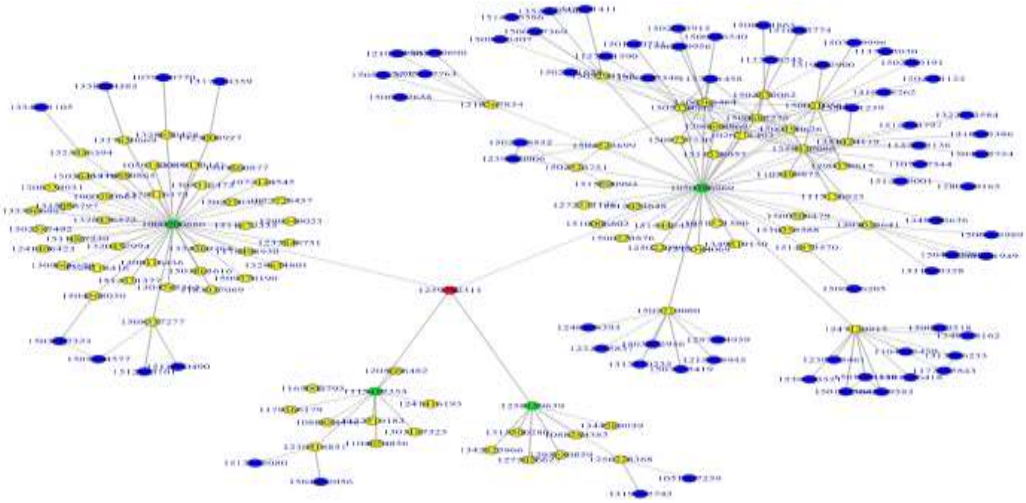
系和圈子，采用 R 软件，基于连通图和深度遍历原理，可以勾勒客户之间的基于资金交易的关系网络图谱。

（5）超大圈子切割

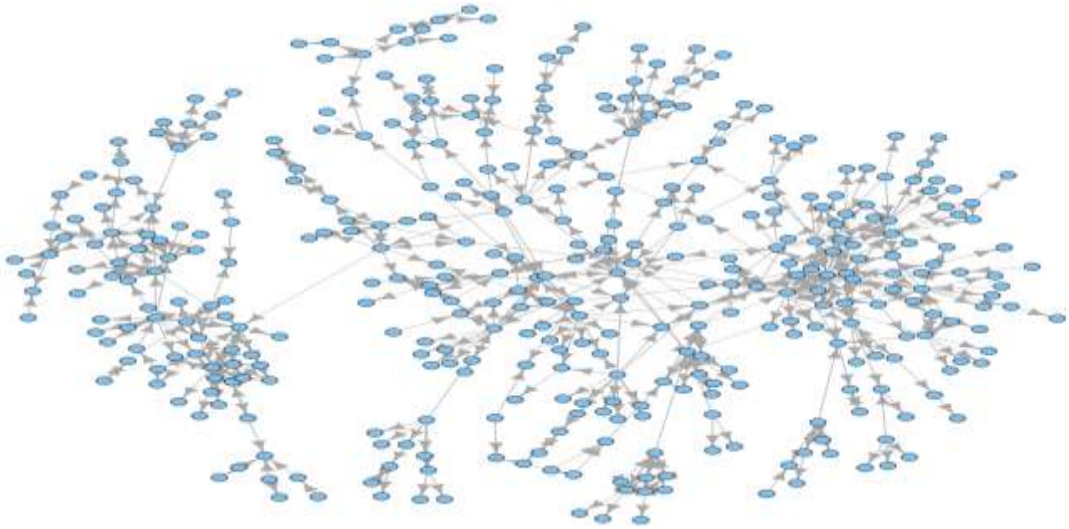
第一，针对圈子规模较大的网络，在排除了异常要素之后，我们会考虑采用均衡割（cut）原理进行圈子切割，使得切割后的两个圈子较为均匀，同时损失的边关系数量最少。

（6）圈子展现

第一，圈子的可视化展现包括两个方面，即三度人脉图和弱联通图展示。三度人脉图是指，刻画以某客户为中心节点的一度人脉、二度人脉和三度人脉。下图为某个客户的一度交易人脉、二度交易人脉和三度交易人脉。基于三度人脉图，我们可以构建客户的网络影响力指数，识别网络大 V 和意见领袖。

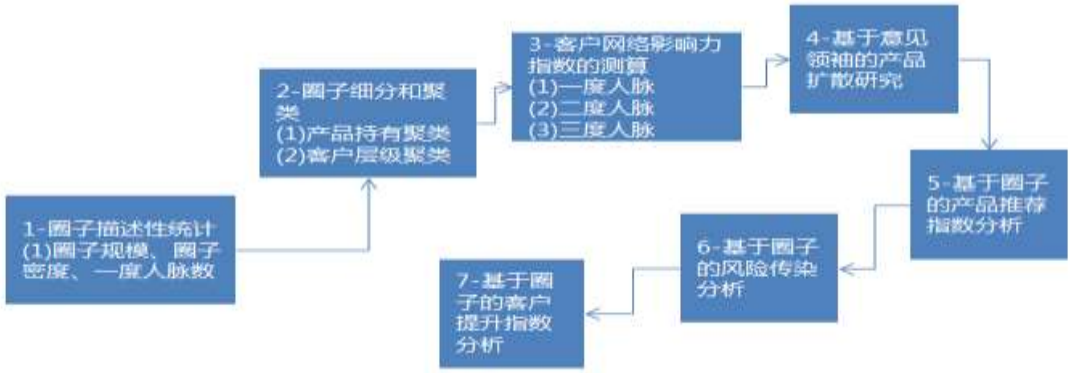


第二，弱连通图是基于节点的可达性，勾勒出节点顺着变关系能够触达的所有边和节点，所构成的关系网络。下图为某个群组规模为 298 的较大圈子。基于弱连通图和群组成员的同质性，我们可以进行产品推荐分析和风险传染分析等。



4 基于关系网络的数据化营销应用

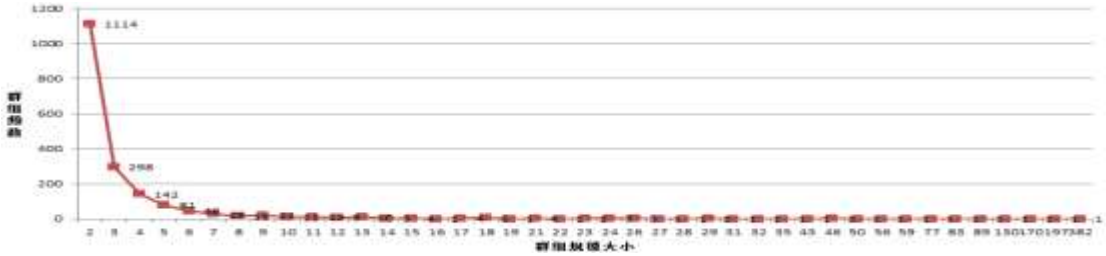
以上我们刻画了客户之间的资金交易网络图谱，基于该图谱，我们展开了如下几个方面的营销应用。即圈子描述性统计、圈子细分和聚类、客户网络影响力指数的测算、基于意见领袖的产品扩散研究、基于圈子的产品推荐指数分析、基于圈子的风险传染分析、基于圈子的客户提升指数分析。



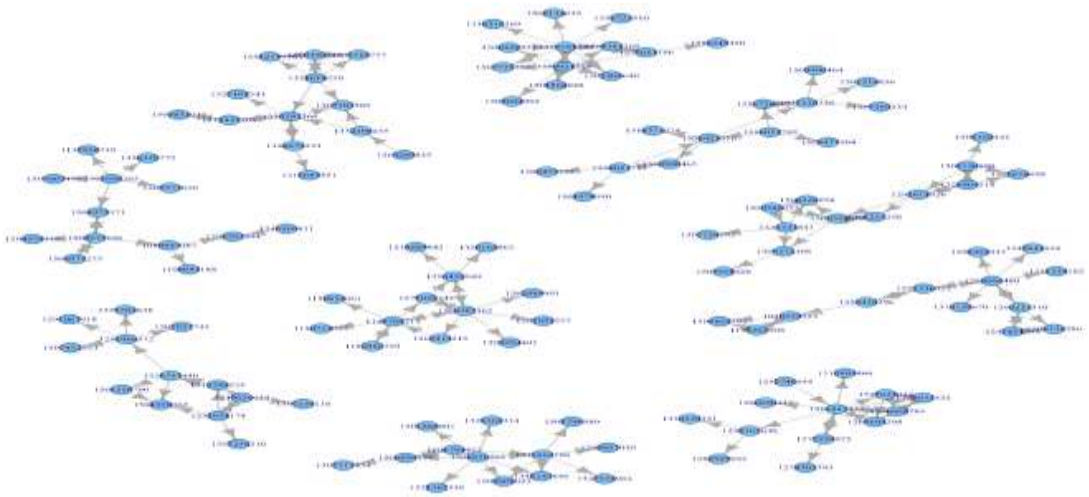
4.1 圈子描述性统计

（1）圈子规模分析

第一，圈子规模分布。下图为小微经济活动圈的群组规模-群组频数分布图。其中，最大的经济活动圈包括 382 名小微客户，最小的经济活动圈包含 2 名小微客户。数据表明，小微客户的经济活动群组规模以 2、3 和 4 为主，群组总数为 1554 个，在所有经济活动群组中占比为 84.4%。其中群组规模为 2 的经济活动圈子有 1114 个，规模为 3 的经济活动圈子有 298 个，规模为 4 的经济活动圈子有 142 个。



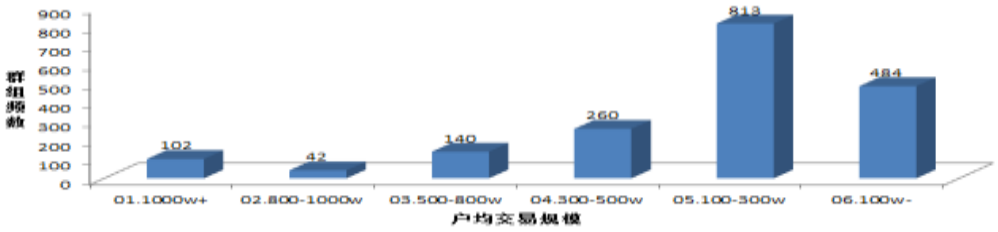
第二，可视化展示。下面两幅图为群组规模为 13 的 10 个资金交易关系网络图谱展示。其中，图中的箭头代表资金的流向。



（2）圈子中户均交易规模

第一，按照圈子中客户的户均交易规模对资金交易圈进行刻画。我们将户均交易规模划分为六个等级，即 1000w+，800-1000w，500-800w，300-500w，100-300w，100w-。

第二，户均交易规模*群组频数分布如下。其中户均交易额在 1000w 以上的圈子有 102 个，户均交易额在 100-300w 的圈子有 813 个。



（3）交易规模和群组密度

第一，群组密度刻画的是圈子中，客户之间互动强度、交易和关联密度。群组密度越大，该群组越属于紧密关联的群体，群体成员之间的互动频率越高，群组成员之间的同质性越高。

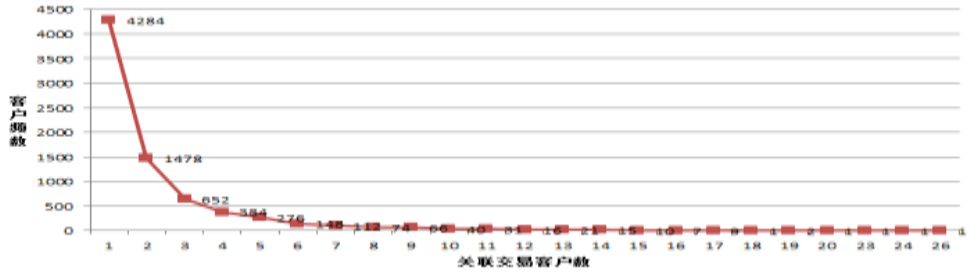
第二，我们通过户均交易规模和群组密度把圈子划分为四个群组，从而能够有效定位出高价值的交易圈。即户均交易规模高、群组关联密度高的经济活动圈有 465 个，属于最为优质的交易圈。

		01群组密度高	02群组密度低	合计
01户均交易高组	频数	465	455	920
	百分比	25.3%	24.7%	50.0%
02户均交易低组	频数	734	187	921
	百分比	39.9%	10.2%	50.0%
合计	频数	1199	642	1841
	百分比	65.1%	34.9%	100.0%

（4）客户一度人脉分布

第一，关联交易客户数描述的是小微客户在社会网络中的网络影响力，即一度人脉数量（直接发生交易和转账的客户）。一度人脉数的分布有较为类似幂分布特征。

第二，对某分行的 7630 名小微客户的关联交易客户数进行分布分析。可以发现，绝大多数小微客户的关联交易客户数集中在 1、2、3，在全部小微客户中占比为 84.06%。其中关联客户数为 1 的小微客户有 4284 户，关联客户数为 2 的小微客户有 1478 户，关联客户数为 3 的小微客户有 652 户。



4.2 圈子细分和聚类

基于以上步骤，我们可以勾勒出客户的关系网络圈子，结合圈子的属性和特征，我们可以做如下一些圈子层面的分析和聚类。包括产品偏好聚类和资产层级聚类。

（1）产品偏好聚类

第一，以圈子为样本单位，计算圈子中各种财富类产品占比（包括储蓄占比、理财占比、基金占比、保险占比），基于聚类 k-means 算法，可以得到基金圈、理财圈等非常典型的关系网络群体。

第二，实际结果表明，聚类算法生成六类群体，即理财股票圈、股票圈、高储蓄圈、低储蓄圈、理财圈、基金圈。并且，不同圈子群体在聚类指标上存在显著差异。

	圈1-理财股票圈	圈2-股票圈	圈3-高储蓄圈	圈4-低储蓄圈	圈5-理财圈	圈6-基金圈
储蓄占比	30.26%	18.04%	37.92%	13.42%	27.04%	27.08%
理财占比	11.09%	0.78%	0.01%	0.23%	14.88%	0.00%
基金占比	8.11%	0.42%	0.00%	0.14%	6.21%	13.25%
国债占比	0.30%	0.04%	0.05%	0.01%	0.29%	0.10%
股票占比	10.66%	10.45%	0.63%	0.00%	0.00%	3.62%
贵金属占比	0.29%	0.07%	0.03%	0.02%	0.04%	0.12%
保险占比	2.86%	0.73%	0.79%	0.45%	2.38%	1.14%

第三，基于产品偏好圈，可以存在以下的营销建议。例如，针对基金圈，群组成员对基金的偏好程度较高，可以针对没有购买基金的客户推荐基金产品。

（2）资产层级聚类

第一，一般而言，商业银行会根据客户的资产水平，将客户划分为不同的层级，例如欣然层级客户、悠然层级客户等。其中从资产水平来看，低层级<有效客户<欣然客户<悠然客户<悠然客户<卓然客户。

第二，考虑基于圈子中层级客户的占比进行聚类建模，分析建模方法同上，以圈子为单位，计算圈子的层级客户占比（包括有效客户占比、欣然客户占比、悠然客户占比、卓然客户占比、私银客户占比），基于聚类 k-means 算法，可以得到欣然圈、悠然圈等非常典型的关系网络群体。

第三，数据结果表明，聚类算法生成五类群体，即潜在有效圈子、潜在欣然圈子、潜在卓然圈子、低层级圈子、潜在悠然圈子。并且，不同圈子群体在聚类指标上存在显著差异。

	圈1-潜在有效客群	圈2-潜在欣然客群	圈3-潜在卓然客群	圈4-低层级客群	圈5-潜在悠然客群
低层级占比	66.51%	47.64%	37.16%	86.95%	35.93%
有效客户占比	21.13%	7.86%	4.70%	1.51%	21.88%
欣然客户占比	8.23%	36.35%	22.53%	5.93%	16.79%
悠然客户占比	0.98%	6.18%	14.66%	2.85%	18.80%
卓然客户占比	1.82%	0.29%	17.53%	1.03%	4.98%
私银客户占比	0.19%	0.36%	2.10%	0.16%	0.56%

第四，基于层级圈，可以存在以下的营销建议。例如，针对悠然圈，基于群组成员的同质性和相似性，我们有理由相信群组成员的资产水平都接近悠然级别，对于与悠然级别差距较大的客户，可以潜在的判断，这样的客户存在很大的资产提升潜质，从而为寻找高潜客户提供了很好的客户白名单线索。

4.3 客户网络影响力指数的测算

一般而言，网络大 v 的社会影响力较大，较容易通过口碑的传播，加快产品的市场渗透。那么如何基于关系网络和关系图谱识别意见领袖，判断客户的网络影响力？

第一，我们通过客户的“三度人脉”测算客户的网络影响力。所谓客户影响力指数，它表征的是客户的关系网络影响力，即客户能够触达和影响的人脉网络。在这里，以三度人脉理论为基础，我们通过三个指标表征客户的网络影响力指数，即圈子位置、关系数量、关系质量。

第二，指标含义。圈子位置刻画得是客户在圈子中的位置中心性基，图论中，常通过点度中心度、接近中心度和居间中心度三个指标测量。关系数量，从客户的一度人脉数量、二度人脉数量和三度人脉数量进行衡量。关系质量衡量的是客户和关联客户之间的关系强弱，这里我们通过过一度人脉人均关系强度、二度人脉人均关系强度、三度人脉人均关系强度，计算客户的关系质量。其中，人均关系强度，通过边关系的交易量和交易次数（0-1 标准化），基于经验权重计算得到（各占 50%的权重）。其次，对于所有涉及的指标，需要消除量纲的影响，因此都会进行 0-1 标准化。具体测算公式如下。

客户影响力指数=1/3(点度中心度+接近中心度+居间中心度)+ 1/3*(一度人脉数量+二度人脉数量+三度人脉数量)+ 1/3*(一度人脉强度+二度人脉强度+三度人脉强度)*

*其中，人脉强度=交易量*0.5+交易次数*0.5。*

第三，基于三个指标，我们可以计算出每个客户的影响力指数，影响力指数越高，客户越具备意见领袖的特征。如果客户的圈子影响力指数越高，该客户属于意见领袖的可能性越大，基于该客户进行口碑传播的价值性也就越大。

4.4 基于意见领袖的产品扩散研究

（1）意见领袖和产品扩散

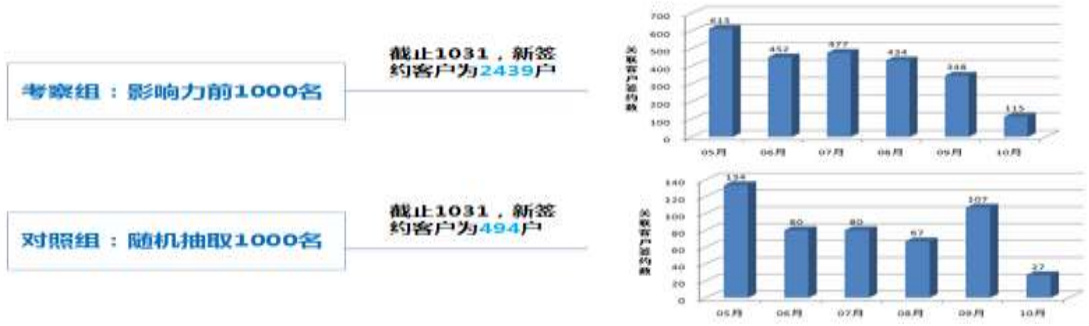
社会网络的分析告诉我们，意见领袖常常会带来更多的产品扩散，能够有效地加速产品的市场渗透。以贷款产品为例，如果高影响力客户签约了商贷通产品，其关联客户是否也会受到影响，从而签约成为我行的商贷通客户？

（2）实验研究

第一，基于某年 1-3 月份的交易流水构建关系网络，以该年 4 月份新签约商贷通的客户作为观察对象，考察一下新签约客户是否会在未来 6 个月内对其关联客户产生影响，从而带来更多的商贷通签约客户？

第二，四月份，新发放贷款客户为 380229 户。在这里，我们将签约客户拆分为两个群体。即考察组（网络影响指数排名前 1000 的新签约客户）、对照组（在新签约客户中随机抽取 1000 名客户）。对比分析 6 个月后，他们各自能够影响和带动多少关联客户。如下图所示。

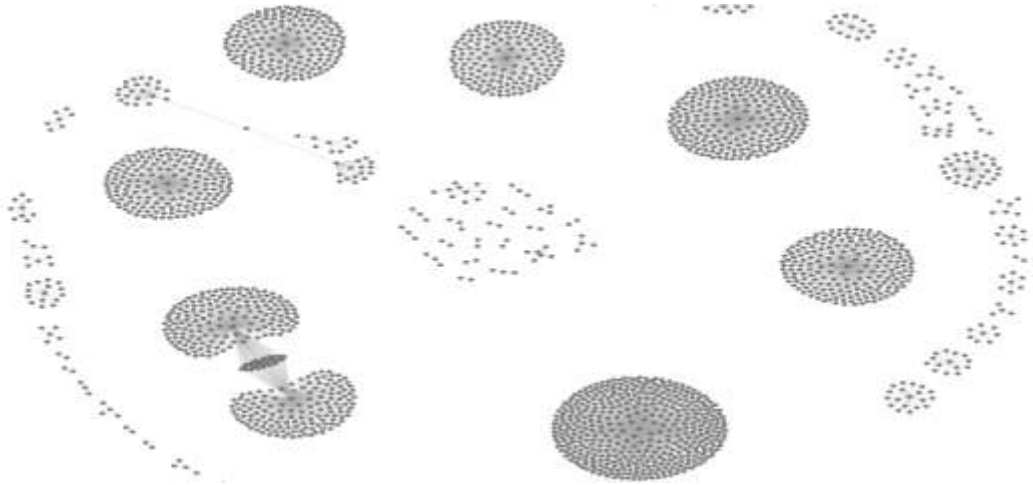
由下图可以看出，截至该年 10 月底，前 1000 名高影响力客户，其关联客户中有 2439 名客户签约商贷通。但是随机抽取的 1000 名客户中，其关联客户中仅仅有 494 名客户签约了商贷通。由此可以发现，高影响力客户的产品扩散和影响力是随机客户的 5 倍。



（3）原因分析

为什么会产生这样的差异呢？原因在于，意见领袖能够影响更多的关联客户。以上面为

例，前 1000 名高影响力客户的关联客户有 27434 名。而随机抽取的 1000 名客户，他们的关联客户仅有 1288 名。也就说，意见领袖能够影响更多的客户，更容易加速产品的扩散和传播。下图为高影响力客户的关联客户网络，如下图。由图可见，意见领袖有更多的关联客户。



（4）其他签约产品的检验

为了提高结论的一般性，我们对其他产品也进行了同样的检验。分析的对象包括：跨行通、基金、乐收银、资金归集、理财。结论再次得到验证，高影响力客户能够带来更多的产品扩散。具体而言，跨行通（提升度 2094%）、基金（提升度 647%）、乐收银（提升度 1455%）、资金归集（提升度 764%）、理财（提升度 1229%）。

产品类型	4月份新签约客户数	样本对照	关联客户数	1031关联客户签约	提升度
跨行通	106886	影响力前1000客户	10578	680	2094%
		随机抽取1000客户	526	31	
基金	151840	影响力前1000客户	23227	142	647%
		随机抽取1000客户	1502	19	
乐收银	22660	影响力前1000客户	10238	513	1455%
		随机抽取1000客户	757	33	
资金归集	15597	影响力前1000客户	6649	121	764%
		随机抽取1000客户	669	14	
理财	627107	影响力前1000客户	19408	186	1229%
		随机抽取1000客户	570	14	

综上，可以发现，通过意见领袖（高影响力客户）对关联客户进行影响和渗透，能够有效地加快商贷通产品的扩散，从而有助于有贷户客户数的快速增长，进而为贷款规模的稳定和持续增长奠定基础。

4.5 基于圈子的产品推荐指数分析

第一，如果一个交易圈或经营圈中，签约或者购买某种产品的客户数占比越高，基于圈子内客户行为和偏好的相似性，我们可以粗略估算出客户的产品购买倾向。

第二，产品推荐指数的测算包括两种方式，即占比推荐和权重推荐。前者的计算方式简单明了，通过计算购买客户数在圈子中的占比计算得到。后者需要考虑客户的三度人脉图中，购买客户和目标客户的距离和关系强度。

第三，以占比推荐为例，通过分析三方存管签约账户，数据分析表明，基于圈子计算得到的推荐指数得分较高的前 10%客群签约产品的概率，相对于随机筛选的客户，其签约率能够提高 2 倍左右。

推荐指数前百分比	交易圈中的行内客户			全行客户			提升度
	5月底未签约第三方存管客户数	10月底新签约第三方存管客户数	新签约占比	5月底未签约第三方存管客户数	10月底新签约第三方存管客户数	新签约占比	
5%	27840	291	1.05%	26819400	116748	0.44%	240.12%
5%-10%	39651	349	0.88%	26819400	116748	0.44%	202.20%
10%-20%	106523	618	0.58%	26819400	116748	0.44%	133.27%
20%-30%	118638	595	0.50%	26819400	116748	0.44%	115.21%
30%-40%	118638	603	0.51%	26819400	116748	0.44%	116.76%
40%-50%	118638	633	0.53%	26819400	116748	0.44%	122.57%
50%-60%	118637	603	0.51%	26819400	116748	0.44%	116.76%
60%-70%	118638	539	0.45%	26819400	116748	0.44%	104.37%
70%-80%	118638	553	0.47%	26819400	116748	0.44%	107.08%
80%-90%	118638	552	0.47%	26819400	116748	0.44%	106.88%
90%-100%	118637	577	0.49%	26819400	116748	0.44%	111.73%

4.6 基于圈子的风险传染分析

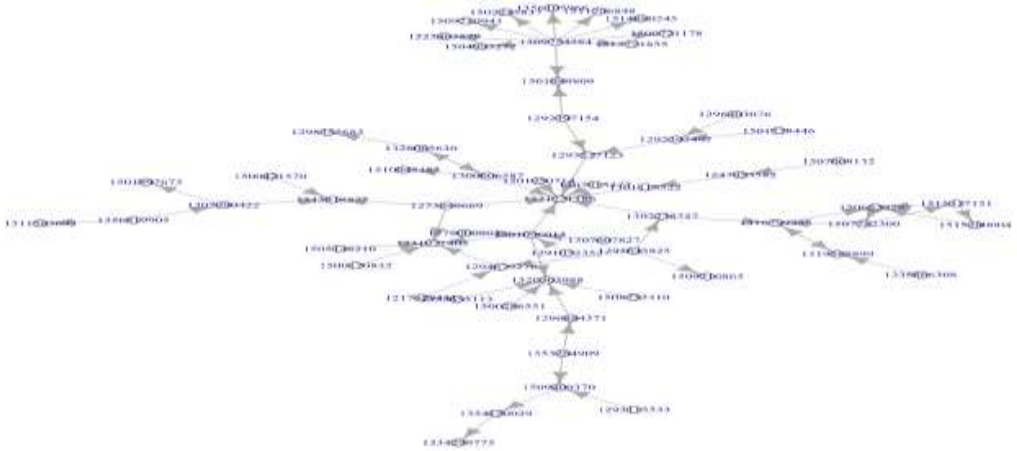
（1）客户圈子和风险传染

社会网络理论告诉我们，同一群组中的成员，因交易关系、互动关系、熟人关系，他们的行为常常表现出相似性和一致性。那么，我们会产生这样一个问题，如果某个小微客户发生过违约行为，其关联的交易客户是否也会有较高的违约倾向？

如果客户所在的交易圈中，历史上发生过违约行为的客户占比越高，我们认为该圈子中违约风险的传染可能性也就越高。原因在于，如果圈子中某个客户的资金周转出现问题（即该客户不具备履约能力），则可能会影响整个交易圈中的资金流动情况，甚至导致整个圈子的资金链断裂，使得该客户的违约风险扩散到整个圈子。

（2）基于资金交易网络的小微客户违约倾向分析

我们知道，小微客户深深扎根于社会网络和交易网络之中。客户之间因业务上下游关系、交易结算、资金拆解、转账关系等，形成错综复杂的资金交易网络。研究表明，在很多情况下，处于同一个资金网络中的客户，因为上下游关系、经营关系和交易关系，同一群组中的个体，他们的行为常常表现出一定的相似性。因为资金上的紧密关联，他们往往成为一个命运共同体，群组中资金流的健康性直接关系着整个群组是否能够稳健持续的发展。那么，这样就产生一个问题，如果群组中的个体，因为资金问题发生违约行为，这种违约风险是否更容易在群组成员内传染呢？下图是一个小微客户的交易社会网络图。



该图是一个群组规模为 62 连通图（成员数为 62），这 62 名成员之间因资金交易关系而连接成一个小网络。箭头的方向代表着资金的流向。从图中可以看出，不同的客户所处的位置有所差异，相互之间存在着资金的流入和流出。直观而言，客户号为 1321221186 处于较为中间的节点。

总体上，我们认为处于一个交易网络中的客户存在较为相似的属性和特征，行为倾向上

（尤其是违约行为）具有一定的传染性，原因有下。一般来说，违约行为的发生取决于两个要素，即履约的能力和履约的意愿。第一，从履约能力的角度来说，如果一个客户资金出现问题（即该客户的履约能力受到影响），可能会影响整个网络小组的资金流，从而导致传染效应，导致其他客户的违约。第二，如果群组中，有客户属于老赖，这种行为必然会产生示范效应，其他群组成员进行仿效，抱着法不责众的侥幸心理，必然会导致违约行为的扩散。

基于此，我们想要验证两个问题。第一，如果群组中有成员发生逾期或不良行为，这种违约行为是否会传染和扩散到其他群组成员？第二，如何计算关联客户的的违约概率？对于违约客户较多的群组，是否需要我们进行审慎的监督和管控？

（3）基于交易关系图谱的违约预测

我们根据某年 1 月-6 月的我行客户交易流水关系，构建客户之间的社会和资金网络。重点考察贷款客户之间的交易流水，识别贷款客户之间的关联关系。

同时，找出该年上半年期间，发生过逾期或不良行为的客户，并且从上面的交易关系中筛选出违约客户的交易图谱。统计表明，上半年出现过违约行为的客户数为 11040 户，从商贷通客户之间的交易关系中，一共能够识别出 1898 对交易关系（包含 2189 名小微客户）。在这 2189 名小微客户中，有 974 名违约客户，其余 1215 名为违约客户的关联客户。

我们考察该年 7-12 月份，存在逾期或不良行为的客户。发现，上半年违约客户的关联客户（1215 名关联客户）中，有 199 名发生逾期或不良行为。

这也就说明在违约客户的关联客户中，风险传染的比率高达 199/1215=16.38%。但是，以 12 月底的数据为例，随机的商贷通客户发生违约的概率为 13180/457392=2.88%。因此，通过关联客户识别违约客户的概率是随机识别的 5.68 倍。

违约预测效果对比

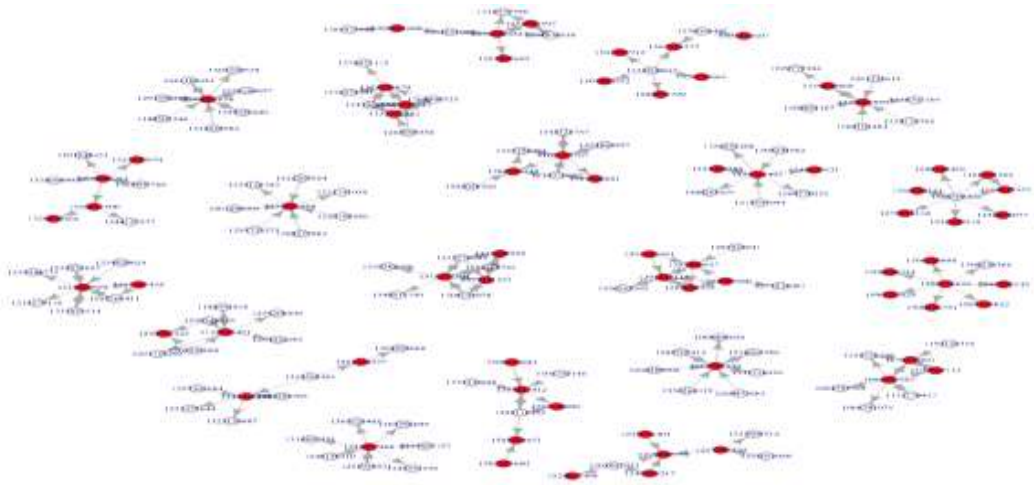
基于关联网络的违约预测		基于随机客户的违约预测	
上半年违约客户的关联客户数	1215	12 月底有贷户客户数	457392
下半年发生违约客户数	199	12 月底发生违约的客户数	13180
占比	16.38%	占比	2.88%
预测效果提升倍数	5.68		

（4）如何量化和计算关联客户的违约概率

由上文分析可以得知，基于违约客户的关联客户进行违约概率预测，其效果是随机效果的 5.68 倍。因此，通过对违约客户的关联客户进行监督和管控，能够很有效的降低违约风险，并且遏制风险的扩散。

第一，违约概率的测算。

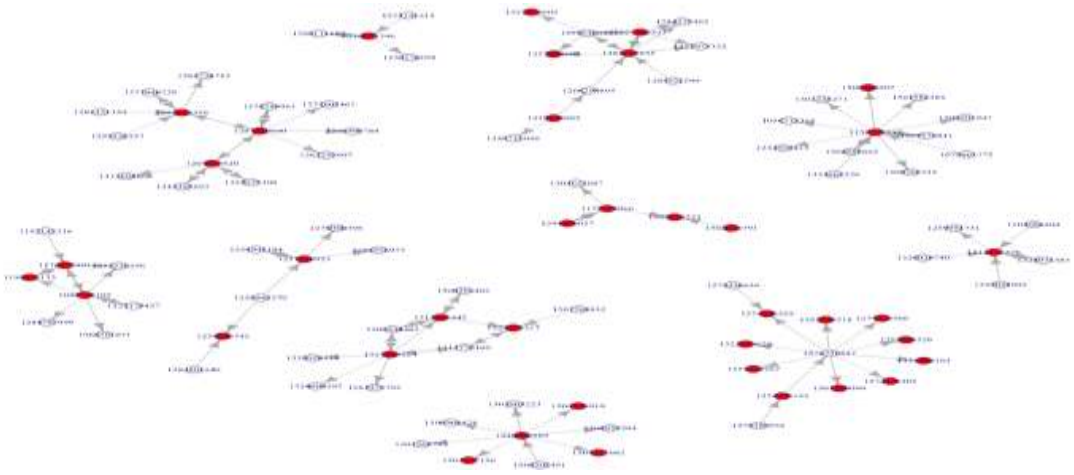
但是，对于关联客户，他们的违约概率是否存在差异，通过什么方法可以较为良好的预测和估计关联客户的违约概率？下图会给我们一些启发和思考。



上图是违约客户的关联网络，图中的红色节点为 2014 年下半年发生过逾期或不良行为的客户。上图选取的是群组规模=8 的所有群组情况。通过上图，我们可以非常直观的发现，在同等群组规模下（规模为 8），有些群组网络中违约客户占比很高（群组中有 7 名成员违约），有些占比很低（群组中只有 1 名成员违约）。基于群组成员行为的相似性，我们有理由相信，一个网络中违约客户越多，该群组中其余成员违约的概率也必然会很高。

第二，违约概率计算规则和步骤。

随机选取群组规模为 4、5、6、7、8、9、10、11、12、13、14 各一组。



数据挖掘中，一种较为常用和流行的样本分类和预测的技术是 KNN 算法（k-nearest neighbors），即 k 近邻。算法的基本原理为，未知客户的类别属性，常常和他最为相近的客户的分类属性相似。因此，常常可以通过关联客户的属性，来预测未知客户的属性。下面的关联客户违约概率预测，采用的就是这种方法。

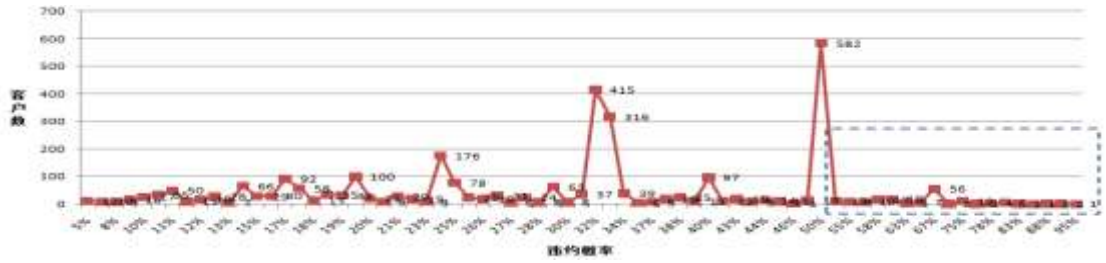
基于同群组成员行为的相似性，以及 KNN 算法的启发，我们采用这样一种方法计算关联客户的违约概率。即，群组成员的违约概率=群组中违约客户数/群组规模。如上图，左上角的群组，其群组规模=4，违约客户数=1，则群组中剩余的 3 名客户的潜在违约概率=1/4=25%。右下角的群组，其群组规模=13，违约客户数=10，则群组中剩余 3 名客户的潜在违约率为 10/13=76.92%。

基于上面的概率计算规则，我们处理加工步骤如下。首先，通过上半年的交易关系，构建商贷通客户之间的关联交易网络。其次，抓取出下半年发生过不良或逾期行为的客户。再次，把违约客户的所有交易关系抓取出来。然后，基于连通图原理，对交易关系进行群组划分。存在交易关系的客户会被划分到同一群组之中。最后，基于群组成员行为的相似性，计

算群组中，违约客户的占比作为其余客户未来违约概率的预测值。

第三，违约概率分布分析

基于此，我们可以计算出每位关联客户的潜在的违约概率。违约概率的分布图如下。



在 3008 名关联的、并且尚未结清的、商贷通客户中，违约概率大于 50%的客户数为 168 名，需要予以重要的监督和关注，防范风险在群组中的扩散和传染。同时，基于客户违约概率，我们也可以生成客户名单，下发分行，重点进行督促和执行。

第四，基于数据的验证

总体提升度。某年 7-12 月中，违约客户的关联客户（未结清）一共有 3008 名。此外，次年 1 月份，新出现的违约客户为 4503 名。在关联客户中，命中预测客户为 313 名，命中率为 10.41%。而 1 月份，客户随机违约的概率为 4503/252726=1.78%。命中效率提高约 5.84 倍。

1 月份数据验证

基于关联网络的违约预测		基于随机客户的违约预测	
某年 7-12 月违约客户的关联客户数	3008	1 月底未结清有贷户客户数	252726
次年 1 月底发生违约客户数	313	1 月底发生违约的客户数	4503
命中率	10.41%	占比	1.78%
预测效果提升倍数	5.84		

累积提升度。按照违约概率把预测名单划分为 10 等分，考察不同等分的命中率和累积命中率。数据表明，预测概率超过 50%的客户中，累积命中率为 23.81%，相较随机命中，提升度提高约 13.38 倍。

违约概率10等份	预测数	命中数	命中率	累积预测数	累积命中数	累积命中率	随机命中率	提升度
01.(90%,100%]	3	1	33.33%	3	1	33.33%	1.78%	18.73
02.(80%,90%]	5	3	60.00%	8	4	50.00%	1.78%	28.09
03.(70%,80%]	26	4	15.38%	34	8	23.53%	1.78%	13.22
04.(60%,70%]	68	11	16.18%	102	19	18.63%	1.78%	10.46
05.(50%,60%]	66	21	31.82%	168	40	23.81%	1.78%	13.38
06.(40%,50%]	663	77	11.61%	831	117	14.08%	1.78%	7.91
07.(30%,40%]	974	112	11.50%	1805	229	12.69%	1.78%	7.13
08.(20%,30%]	525	44	8.38%	2330	273	11.72%	1.78%	6.58
09.(10%,20%]	604	36	5.96%	2934	309	10.53%	1.78%	5.92
10.(0%,10%]	74	4	5.41%	3008	313	10.41%	1.78%	5.85

4.7 基于圈子的客户提升指数分析

第一，基于客户之间的交易转帐记录，我们可以勾勒出客户的经营圈（针对小微客户）或交易消费圈（针对个金客户）。通过计算两个圈子指标，即圈子密度和圈子户均金融资产。

第二，一方面，圈子密度越高，说明圈子内成员之间的交互越为频繁，客户之间的关系和联系越为紧密，那么圈子成员在行为特征和资产特征应该具有较高的相似性。另一方面，如果圈子成员 A 的金融资产显著小于圈子成员的户均金融资产，在圈子密度较高的情形下（成员之间相似性较高），那么成员 A 存在着较大的资产提升空间。

5 论文小结

论文创新点包括如下几个方面。

第一，探索了一套标准化的关系网络模型开发方案。在银行业内，首次基于客户之间的转账交易，构建客户的资金交易关系网络。相关的开发过程包括，资金交易流水清洗、有效边的界定、异常点线圈的处理、圈子搭建、圈子展现等。

第二，圈子聚类。首次以圈子为样本，采用 k-means 算法，展开圈子层面的聚类分析。基于圈子的特征属性（资产占比、层级客户占比），勾勒出产品偏好圈子（例如，基金圈、理财圈等）。其次，还基于圈子中客户的资产和层级属性，勾勒出层级圈子（贵宾圈、私银圈等）。

第三，网络影响力指数。首次构建了客户的网络影响力指数，基于影响力指数，可以快速有效识别意见领袖和核心客户。

第四，产品推荐指数。首次基于资金交易关系网络圈，开发了产品推荐指数。基于产品推荐指数，进行潜在产品的营销。

第五，风险传染指数。基于交易圈开发了风险传染指数（圈子中历史违约客户占比情况），相对于随机样本，基于圈子的违约客户捕获率能够提升 6 倍左右。

第六，资产提升指数。首次基于客户之间的资金交易网络，开发了资产提升指数（涉及两个测算指标，即圈子密度和圈子户均金融资产），有效辅助业务人员识别潜在的高价值客户。

6 参考文献

[1]扈中凯, 郑小林, 吴亚峰, 等. 基于用户评论挖掘的产品推荐算法[J]. 浙江大学学报: 工学版, 2013 (8): 1475-1485.

[2]李欣璐, 刘鲁. 基于协同过滤的银行产品推荐系统建模[D]. , 2007.

[3]那日萨, 刘影, 李媛. 消费者网络评论的情感模糊计算与产品推荐研究[J]. 廣西師範大學學報 (自然科學版), 2010, 28(1): 143-146.

[4]曹渝昆. 基于神经网络和模糊逻辑的智能推荐系统研究[D]. 重庆: 重庆大学, 2006.

[5]唐晓波, 樊静. 基于客户聚类的商品推荐[J]. 情报杂志, 2009 (6): 143-146.

[6]张彦超, 刘云, 张海峰, 等. 基于在线社交网络的信息传播模型[J]. 物理学报, 2011, 60(5): 50501-050501.

[7]张赛, 徐恪, 李海涛. 微博类社交网络中信息传播的测量与分析[J]. 西安交通大学学报, 2013, 47(2): 124-130.

[8]陈克寒, 韩盼盼, 吴健. 基于用户聚类的异构社交网络推荐算法[J]. 计算机学报, 2013, 36(2): 349-359.

[9]王辉, 韩江洪, 邓林, 等. 基于移动社交网络的谣言传播动力学研究[J]. 物理学报, 2013 (11): 96-107.

[10]吴信东, 李毅, 李磊. 在线社交网络影响力分析[J]. 计算机学报, 2014, 37(4): 735-752.

[11]邓夏玮. 基于社交网络的用户行为研究——用户行为分析与用户影响力建模[D]. 北京交通大学, 2012.

[12]康书龙. 基于用户行为及关系的社交网络节点影响力评价-以微博研究为例 [D][D]. 北京: 北京邮电大学, 2011.

[13]王莉. 基于数据挖掘技术的品牌客户画像管理初探[J]. 移动通信, 2008, 32(23): 77-82.

[14]崔琳. 基于客户画像的数据挖掘技术在 CRM 中的应用[D]. 东华大学, 2015.

[15]曾伟, 孔新川, 陈威, 等. 大数据发现银行贷款风险[J]. 大数据, 2015 (2): 112-115.

[16]黄鑫. 数字化时代商业银行 “智能化” 的思考[J]. 中国银行业, 2017, 3: 028.

[17]于立勇, 詹捷辉. 基于 Logistic 回归分析的违约概率预测研究[J]. 财经研究, 2004, 30(9): 15-23.

[18]高琪，辛乐．基于用户偏好度模型和情感计算的产品推荐算法[C]//第二十九届中国控制会议论文集. 2010.