利用网络数据预测企业失信行为

周涛[1]，李艳丽[1]，李倩[2]，陈端兵[1,3]，谢文波[1,3]，吴桐[2]，曾途[2]
1. 电子科技大学大数据研究中心，四川 成都 611731；2. 成都数联铭品科技有限公司，四川 成都 610041；
3. 成都数之联科技有限公司，四川 成都 610041

摘要
传统的企业信用水平分析方法多从企业规模、经营地、行业类别、注册与实缴资本等特征属性出发，缺少基于海量关联数据的深入分析。为了解决这个问题，采集、清洗了大量数据，建立了包含400多万家企业的有向投资网络，其中存在各类失信行为的企业有近26万家。研究结果显示，企业失信行为存在明显的“网络效应”，目标企业的股东或者投资企业若存在失信行为，则目标企业发生失信的风险远远大于平均值。基于此，提出了简单的预测企业失信行为的算法，其精确性远远超过了不考虑网络效应的回归方法。

关键词
失信行为预测；网络效应；企业征信

中图分类号：TN399 文献标识码：A doi: 10.11959/j.issn.2096-0271.2018049

Predicting the discredited behavior of enterprises via large-scale investment network

ZHOU Tao[1], LI Yanli[1], LI Qian[2], CHEN Duanbing[1,3], XIE Wenbo[1,3], WU Tong[2], ZENG Tu[2]
1. Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 611731, China
2. Business Big Data Inc., Chengdu 610041, China
3. Union Big Data Inc., Chengdu 610041, China

Abstract
Previous enterprise credit level analysis mainly focused on the features including enterprise size, place of operation, industry category, registration and paid-in capital, and lacked in-depth analysis based on massive data. A directed investment network consisted of more than 4 million enterprises was built up, among which nearly 260 000 enterprises have various discredited behaviors. The results show that there is an obvious "network effect" in the discredited behaviors of enterprises. If the target enterprise's shareholders or its invested enterprises have discredited behaviors, the risk of having discredited behaviors of the target enterprise is far greater than the average. Based on this, a simple generalized linear regression algorithm was proposed to predict the discredited behaviors of enterprises, which is far more accurate than the regression method without considering the network effect.

Key words
discredited behavior prediction, network effect, enterprise credit

1 引言

随着互联网、移动互联网、物联网和各种遥感探测技术的发展，一个“一切都被记录，一切都被分析”的数据化时代已经到来[1]。大数据的技术发展与应用实践已经为社会经济和人们的日常生活带来了显著的贡献[2]。在医疗领域，大数据和人工智能被用于精准识别医疗影像中早期的病灶，定位致病基因并开展相应的靶向治疗，实时监测评估健康状况，提前预警重大健康风险等[3]；在交通领域，海量数据和预测算法的结合能够帮助人们进行更高效的交通导航，尽可能减少拥堵时间[4]；在教育领域，对学生发展过程中学习和生活数据的分析可以帮助了解学生行为和学业表现之间的关系，从而设计更高效的个性化教育方案，提前对不利于学业发展的异常行为进行干预[5]等。

在各种大数据研究对象和大数据应用场景中，网络大数据是独具特色且受到广泛关注的方向[6-7]。网络科学是以网络为研究对象的一门有数百年历史的专业性很强的学科，又是众多学科中不同研究对象的统一抽象的表达方式[8]。目前万维网具有超过万亿的统一资源定位符（uniform resource locator，URL），Facebook有10亿个节点和千亿条连边，大脑神经元网络有数百亿节点……如何分析挖掘大规模网络中隐藏的信息，进一步应用于解决实际问题，已经成为学术界和企业界亟待解决的重要挑战。

金融网络分析是网络大数据在金融领域的应用，因为其可见的巨大价值，最近受到了广泛的关注[9-10]。真实的金融网络包括金融机构之间的业务关系网络[11]、投资机构和企业之间或投资机构之间因共同投资而形成的投资关系网络[12]等。这些网络往往度分布范围很广，网络连接总体上比较稀疏，网络连接呈现负相关性（度大的节点倾向于和度小的节点相连），但度很大的若干节点之间表现出“富人俱乐部效应”（即度最大的若干节点之间连接特别紧密）。金融网络分析可以帮助人们尽早发现金融风险，提升抗击重大风险的能力[13-14]。

本文拟将金融网络大数据分析技术应用于预测企业失信行为。笔者认为，定量刻画企业信用水平，进一步预测企业可能出现的各种失信行为，是建立信用社会、提升营商环境中不可或缺的重要环节，可以作为商务合作、股权投资、担保贷款、招商引资等业务的前置条件。传统的分析方法多从企业规模、经营地、行业类别、注册与实缴资本等特征属性出发，预测结果往往并不准确。最近大数据方法被广泛应用于企业征信领域，取得了大量、丰富且重要的成果[15-17]。但是，据笔者所知，目前尚没有直接利用大规模企业间投资关系网络来提升分析精确性的研究工作。本文采集清洗了大量数据，建立了包含400多万家企业的有向投资网络，其中存在各类失信行为的企业占比6%左右。研究结果显示，企业失信行为存在明显的“网络效应”，即目标企业的股东或者投资企业若存在失信行为，则目标企业发生失信的风险远远大于平均值。基于以上分析，笔者实现了一个简单的失信行为预测算法，其精确性远远超过了不考虑网络效应的回归方法。

2 数据描述和网络分析

本文分析的数据采自国家企业信用信息公示系统的公开数据，包括企业的基本

信息（如企业类型、企业注册资本、企业注册地、企业所属行业等）、工商变更等备案信息、各类行政处罚和其他失信行为公告等多项数据。**表1**给出了我国企业的行业分类信息。

本文分析的网络数据集包括4 020 504家企业，其中有259 760家企业存在至少一次失信行为（本文为了便于叙述，把各种行政处罚都归为失信行为），占比为6.46%。笔者建立了这些企业之间直接投资形成的有向网络，其中每一个节点代表一家企业，如果企业*i*投资了企业*j*（*i*是*j*的股东），则用一条有向边*i*->*j*表示。

图1显示了不同规模的4个有代表性的弱连通子图，其中灰色节点为没有失信行为的企业，黑色节点为有失信行为的企业。**图1**中连边的方向没有表现出来。

即便以有向网络弱连通作为判据，该网络也不是完全连通的网络。**图1**给出了所有弱连通子图规模（节点数）的Zipf分布

表1 企业行业信息字段和内容

字段	内容	字段	内容	字段	内容	字段	内容
A	农业、林业、牧业、渔业、服务业	G	交通运输、仓储和邮政业	M	科学研究和技术服务业	S	公共管理、社会保障和社会组织
B	采矿业	H	住宿和餐饮业	N	水利、环境和公共设施管理业	T	国际组织
C	制造业	I	信息传输、软件和信息技术服务业	O	居民服务、修理和其他服务业	Z	其他
D	电力、热力、燃气及水生产和供应业	J	金融业	P	教育		
E	建筑业	K	房地产业	Q	卫生和社会工作		
F	批发和零售业	L	租赁和商务服务业	R	文化、体育和娱乐业		

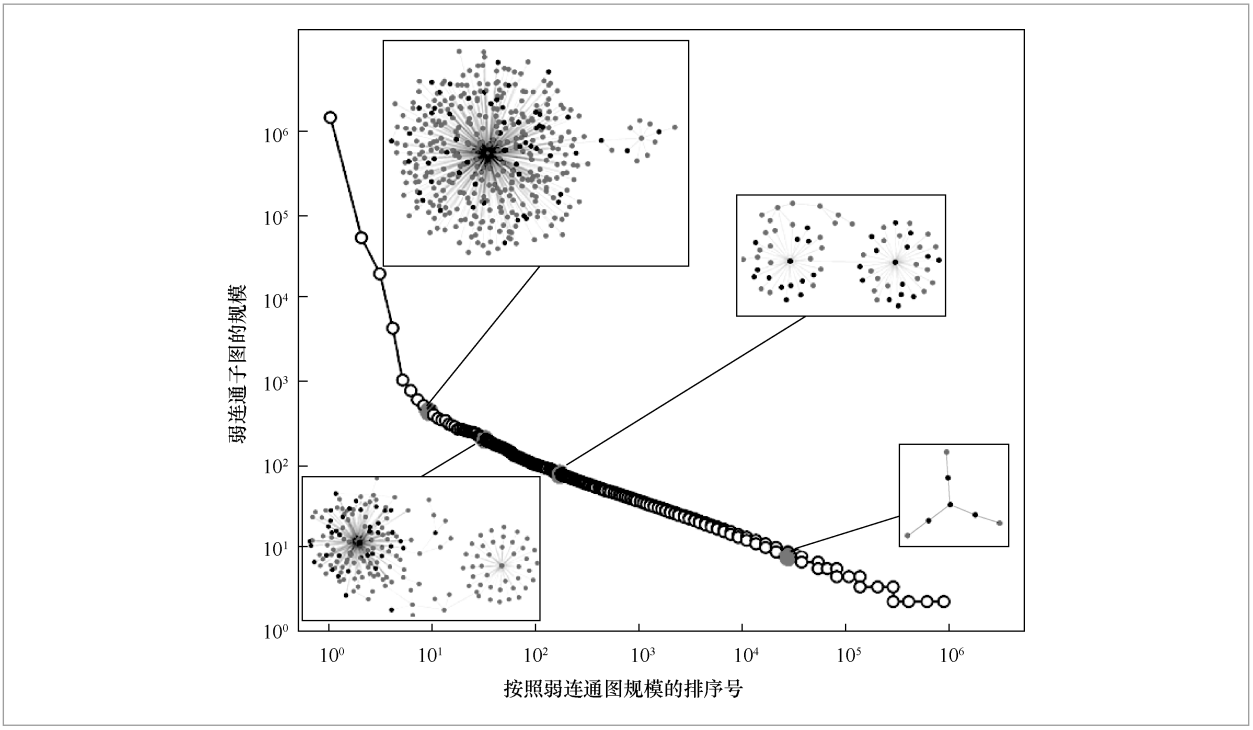

图1 按照弱连通图规模排序得到的 Zipf 分布

（关于Zipf分布的定义以及其与幂律分布和Heaps定律的关系，请参考文献[18]）。可以看到，仅有一个超级连通图规模在100万个节点以上，其余第二大连通图规模只有不到10万个节点。绝大部分连通图的规模都很小。笔者也在**图1**中选择了若干可以用可视化方法直接画出来的规模较小的连通图，方便读者直观地看到投资网络的结构（为了可视化效果，投资关系的方向在此处省略了）。后文为了便于叙述，将存在失信行为的企业简称为失信企业，读者要注意这并不是严格等同的概念。

笔者猜测，企业失信行为在企业之间的投资网络中具有明显的“网络效应”。也就是说，如果一个目标企业的投资者（法人股东）或其投资对象（子公司、控股公司、参股公司等）存在失信行为，那么这个目标企业存在失信行为的可能性也很大。为了方便、直观地观察上述网络效应，先考虑一种简单的情况，即将有向投资关系网络转化为无向网络加以考虑，暂时忽略连边的方向性。用$P(B|m)$表示所有邻居中失信企业数目大于或等于m的企业是一个失信企业的概率。显然，当$m=0$时，$P(B|0)=0.064\ 6$，就是整个数据集中失信企业的比例。**图2**给出了$P(B|m)$随m变化的曲线。该曲线上升的趋势非常明显，尤其在m比较小的时候（m很大的时候数据点很少，因此会出现一些波动）。即便只知道“目标企业的投资方和投资对象存在失信企业”这一信息（也就是$m=1$），该目标企业是失信企业的概率就从$P(B|0)=0.064\ 6$陡增到$P(B|1)=0.164\ 1$。而如果一家企业邻居中有3家或3家以上失信企业，它自己是失信企业的可能性会超过40%。笔者在**图2**中还用虚线强调了一个值（虚线所示），即如果已知目标企业的邻居中没有失信企业，则该目标企业自身是失信企业的可能性是$0.047\ 4$，比整个数据集中失信企业占比低了27%。

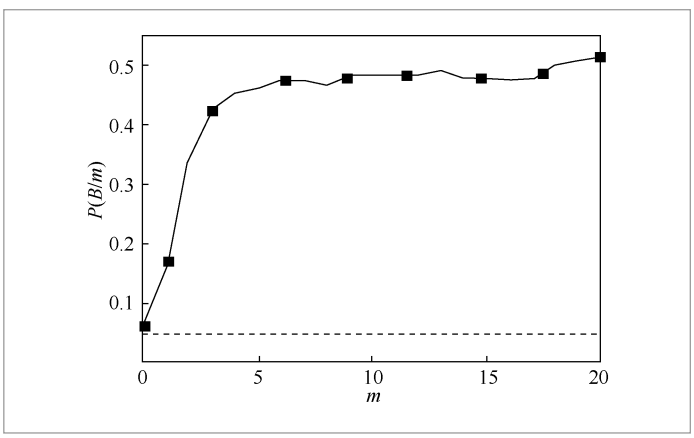

图2 $P(B/m)$随m的变化曲线

以上分析显示，投资网络模型对于分析企业失信行为而言是一个非常有效的工具。事实上，如同社交网络上吸烟－戒烟[19]、肥胖－减肥[20]的网络效应（如果你身边有很多烟鬼，那么你是烟鬼的可能性会大大增加；类似地，如果你的很多社交好友都在减肥，那么你减肥成功的可能性也会增大）一样，企业失信行为也有明显的网络效应，即目标企业的投资方或者投资对象若存在失信行为，则目标企业发生失信的风险会大大增加，而且随着失信邻居数目的增加，失信风险也快速增加。这种现象既可能来自投资关系本身的风险传递（如果目标企业的投资对象出现了信贷违约，说明该企业现金流出现了严重问题，那么目标企业可能已经为投资对象注入了现金，并且获得期望投资收益的可能性很小，因此目标企业自身也可能出现类似风险），也可能来自全行业的问题（例如环保标准上升后大量家具行业受到行政处罚，而这些同行业企业之间容易有产业链条上的投资关系）。

3 预测算法和预测结果

因为有一些企业基本信息数据不完

备，为了和只用基本特征信息进行分析的算法做对比，本文仅考虑基本信息完备的3 207 962家企业，其中有失信行为的企业有257 163家。因为本文的核心诉求是揭示企业失信行为的网络效应，所以不采用和对比非常复杂的算法。笔者选择了广义线性回归模型[21]，该模型拟合过程是并行的，计算速度非常快，适用于本文拟处理的数百万规模企业。因为预测企业失信行为是一个典型的二分类问题，选用伯努利分布作为广义线性模型的函数族，故一个企业是失信企业的似然值可由以下计算式拟合：

$$E(x,w)=[1+\exp(-w^Tx)]^{-1} \quad (1)$$

其中，x为特征向量，w为特征权重向量。**表2**给出了本文使用的企业基本特征和网络特征说明，其中二阶邻居是指与目标企业在无向图中距离恰好为2的企业集合。企业基本特征中的离散型已转化为多项2值特征——1代表“是”，0代表“否”。

本文实验中采用10次交叉验证，抽取9:1的数据作为训练集和测试集，通过10次实验求平均。每次实验根据训练集回归模型拟合的参数，对测试集中所有企业存在失信行为的期望值打分，把风险最高的企业排在前面。如果风险最高的Top N个企业中失信企业有r个，就定义精确度为：

$$\text{Precision}=r/N \quad (2)$$

图3给出了预测精确度随N的变化曲线。**图3**中，NOR表示使用全部的企业基本特征，S1表示使用网络特征中的一阶邻居特征，S2表示使用网络特征中的二阶邻居特征。ALL表示融合S1、S2和NOR的特征。

表2 预测模型中涉及的基础特征及网络特征

特征		说明
一阶邻居特征	失信被投资对象数	对这里的数量进行了处理，当值大于或等于4时，值取4
	失信被投资对象比例	失信被投资对象数/被投资对象的总数
	失信投资者数	对这里的数量进行了处理，当值大于或等于4时，值取4
	失信投资者比例	失信投资者数/投资者的总数
	失信一阶邻居数	对这里的数量进行了处理，当值大于或等于4时，值取4
	失信一阶邻居比例	失信一阶邻居数/一阶邻居的总数
二阶邻居特征	二阶邻居中失信被投资对象数	对这里的数量进行了处理，当值大于或等于4时，值取4
	二阶邻居中失信被投资对象比例	二阶邻居中失信被投资对象/二阶邻居中被投资对象总数
	二阶邻居中失信投资者数	对这里的数量进行了处理，当值大于或等于4时，值取4
	二阶邻居中失信投资者比例	二阶邻居中失信投资者数/二阶邻居中投资者总数
	失信二阶邻居数	对这里的数量进行了处理，当值大于或等于4时，值取4
	失信二阶邻居比例	失信二阶邻居数/二阶邻居总数
基本特征	企业类型	个体户、有限公司等类型，离散型
	企业资本	公司注册资本
	企业规模	根据工业和信息化部联企业[2011]300号文件《关于印发中小企业划型标准规定的通知》的标准，针对不同行业计算得到的企业规模被分为大型、中型、小型、微型4类，数值型
	地区	公司所在地理位置，离散型
	行业	公司所处行业细分领域（见表1），离散型

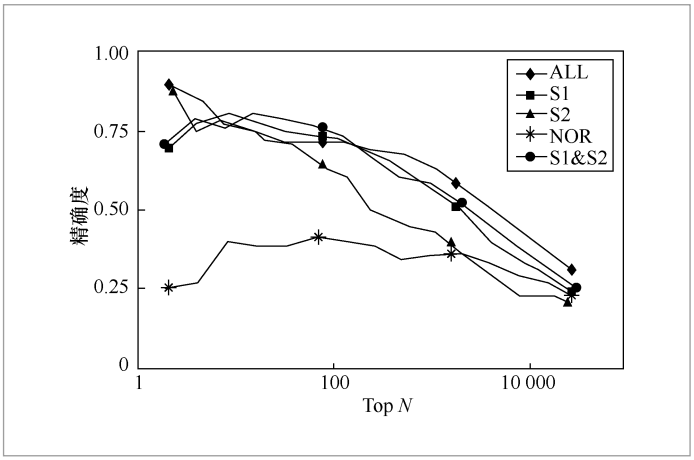

图 3 使用不同特征进行组合时预测模型的精确度对比

可以看到，网络特征中一阶邻居网络特征较二阶邻居网络特征预测效果更好，仅用企业基本特征不能很好地预测企业失信行为。如果结合了网络特征，预测的精确度能够被大幅度提升。其中预测出来失信风险最高的100家企业有70%以上有失信行为，前10 000家企业40%左右有失信行为。

4 结束语

本文采集了400多万家企业的真实数据，构建了大规模的企业间投资网络，揭示了企业失信行为显著的网络效应，即目标企业的投资方或者投资对象若存在失信行为，则目标企业发生失信的风险会大大增加。进一步的分析显示，随着失信邻居数目的增加，失信风险也快速增加。基于此，本文设计了一个简单的广义线性回归模型对企业失信行为进行预测。算法结果显示，网络特征的加入可以大幅度提高仅采用文本特征的算法的精确度。结合网络特征的算法可以在给出10 000家最高失信风险企业的情况下做到40%的预测精确度。

企业征信是金融生态体系，甚至经济生态体系建设中至关重要的一环。本文仅仅是开展了初步的尝试，就有效证明了大数据在金融征信领域的巨大应用价值。事实上，企业投资关系网络是应用潜力非常广泛的一类数据。举例而言，通过投资关系网络，可以发现一些异常的投资行为，例如通过多地、多次变更，形成企业A_1全资控股A_2，A_2全资控股A_3，…，A_{s-1}全资控股A_s，A_s全资控股A_1这样的长度为s的企业投资有向环，利用这样的有向环，企业A_1可以把1亿元注册资本给A_2，A_2再给A_3，以此类推，最终回到A_1。资金没有真正投入，但是每家企业的实缴注册资本都增加了1亿元。这些都是企业获取虚假资质、假造项目承接能力甚至非法集资常用的伎俩。这些重大金融经济风险问题转变成了在一个几千万个节点的企业投资关系网络中发现有向环的典型的图论问题。事实上，企业投资关系网络的应用还很多，本文的思路和方法还可以刻画金融担保圈的风险传播，提高识别金融担保圈中关键节点的准确度[22]，提升大数据在打击非法集资[23]、反洗钱[24]方面的效率和准确度。本文研究内容仅仅是其万千应用中的一角，希望能给金融管理工作者有益的启发，并激发数据科学家和数据工程师投身于网络大数据的研究和应用中。

参考文献:

[1] 维克托·迈尔-舍恩伯格，肯尼思·库克耶. 大数据时代：工作、生活与思维的大变革[M]. 盛杨燕，周涛，译. 杭州：浙江人民出版社，2013.
MAYER-SCHONBERGER V, CUKIER K. Big data: a revolution that will transform how we live, work, and think[M].

Translated by SHENG Y Y, ZHOU T. Hangzhou: Zhejiang People's Publishing House, 2013.

[2] 周涛. 为数据而生: 大数据创新实践[M]. 北京: 北京联合出版有限责任公司, 2016.
ZHOU T. Big data innovation[M]. Beijing: Beijing United Publishing Co. Ltd., 2016.

[3] 埃克里·托普. 颠覆医疗: 大数据时代的个人健康革命[M]. 张南, 等, 译. 北京: 电子工业出版社, 2014.
TOPOL E. The creative destruction of medicine[M]. Translated by ZHANG N, et al. Beijing: Publishing House of Electronics Industry, 2014.

[4] LV Y, DUAN Y, KANG W, et al. Traffic flow prediction with big data: a deep learning approach[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(2): 865-873.

[5] CAO Y, GAO J, LIAN D F, et al. Orderliness predicts academic performance: behavioral analysis on campus lifestyle[J]. Journal of the Royal Society Interface, 2018: 20180210.

[6] 王元卓, 靳小龙, 程学旗. 网络大数据: 现状与展望[J]. 计算机学报, 2013, 36(6): 1125-1138.
WANG Y Z, JIN X L, CHENG X Q. Network big data: present and future[J]. Chinese Journal of Computers, 2013, 36(6): 1125-1138.

[7] 周涛. 网络大数据——复杂网络的新挑战: 如何从海量数据获取信息[J]. 电子科技大学学报, 2013, 42(1): 7-8.
ZHOU T. Nework big data: a new challenge in complex network: how to dig out information from large-scale data set[J]. Journal of University of Electronic Science and Technology of China, 2013, 42(1): 7-8.

[8] 汪小帆, 李翔, 陈关荣. 网络科学导论[M]. 北京: 高等教育出版社, 2012.
WANG X F, LI X, CHEN H R. Network science an introduction[M]. Beijing: Higher Education Press, 2012.

[9] SCHWEITZER F, FAGIOLO G, SORNETTE D, et al. Economic networks: the new challenges[J]. Science, 2009, 325(5939): 422-425.

[10] 高见, 周涛. 大数据揭示经济发展状况[J]. 电子科技大学学报, 2016, 45(4): 625-633.
GAO J, ZHOU T. Big data reveal the status of economic development[J]. Journal of University of Electronic Science and Technology of China, 2016, 45(4): 625-633.

[11] SORAMÄKI K, BECH M L, ARNOLD J, et al. The topology of interbank payment flows[J]. Physica A, 2007, 379(1): 317-333.

[12] SONG D M, JIANG Z Q, ZHOU W X. Statistical properties of world investment networks[J]. Physica A, 2009, 388(12): 2450-2460.

[13] MAY R M, LEVIN S A, SUGIHARA G. Complex systems: ecology for bankers[J]. Nature, 2008, 451(7181): 893.

[14] BARDOSCIA M, BATTISTON S, CACCIOLI F, et al. Pathways towards instability in financial networks[J]. Nature Communications, 2017(8): 14416.

[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, 1(2): 2015024.
ZENG W, KONG X C, CHEN W, et al. Uncovering the risk in bank loans by big data[J]. Big Data Research, 2015, 1(2): 2015024.

[16] 石勇, 孟凡. 信用评分基本理论及其应用[J]. 大数据, 2017, 3(1): 2017003.
SHI Y, MENG F. Credit scoring:basic theory and applications[J]. Big Data Research, 2017, 3(1): 2017003.

[17] 李友元, 寇纲. 我国大数据征信的挑战及对策[J]. 大数据, 2017, 3(1): 2017004.
LI Y Y, KOU G. Challenge and countermeasure of big data credit reporting in China[J]. Big Data Research, 2017, 3(1): 2017004.

[18] LU L, ZHANG Z K, ZHOU T. Zipf's law leads to heaps' law: analyzing their relationin finite-size systems[J]. PLoS One, 2010, 5(12): e14139.

[19] CHRISTAKIS N A, FOWLER J H. The collective dynamics of smoking in a

large social network[J]. New England Journal of Medicine, 2008, 358(21): 2249-2258.

[20] BALLANTINE P W, STEPHENSON R J. Help me, I'm fat! Social support in online weight loss networks[J]. Journal of Consumer Behaviour, 2011, 10(6): 332-337.

[21] NELDER J, WEDDERBURN R. Generalized linear models[J]. Journal of the Royal Statistical Society(Series A), 1972, 135(3): 370-384.

[22] 陈国富, 刘新海. 我国担保圈大数据分析的初步发现[J]. 发展研究, 2015(5): 24-29.
CHEN G F, LIU X H. Preliminary findings of big data analysis in China's guarantee circles[J]. Development Research, 2015(5): 24-29.

[23] 张晓朴. 互联网金融监管的原则: 探索新金融监管范式[J]. 金融监管研究, 2014(2): 6-17.
ZHANG X P. Principles of Internet financial regulation: exploring new financial regulatory paradigm[J]. Financial Regulation Research, 2014(2): 6-17.

[24] 李新安. 大数据技术在反洗钱工作中的应用探析[J]. 时代金融, 2014(7): 259-260.
LI X A. Applying big data technology in anti-money[J]. Times Finance, 2014(7): 259-260.

作者简介

周涛(1982-),男,电子科技大学教授,主要从事统计物理与复杂性方面的研究。在《Physics Reports》《PNAS》《Nature Communications》等国际SCI期刊发表学术论文300余篇,引用20 000余次,H指数为68。2009年获教育部自然科学奖一等奖,2011年获第十二届中国青年科技奖,2014年起,历年入选Elesvier最具国际影响力中国科学家名单(物理天文类)。2015年当选第十二届中华全国青年联合会常务委员,并担任科学技术界别工作委员会副主任。2015年当选全国十大科技创新人物。2017年获全国创新争先奖。2018年起任四川省政协常务委员。

李艳丽(1990-),女,电子科技大学大数据研究中心博士生,主要研究方向为关键节点挖掘、链路预测、推荐系统。

作者简介

李倩(1987-),女,成都数联铭品科技有限公司数据管理部经理,主要研究方向为数据治理。

陈端兵(1971-),男,电子科技大学副教授、博士生导师,主要研究方向为数据挖掘、网络结构分析、信息传播与推荐。

谢文波(1990-),男,电子科技大学大数据研究中心博士生,主要研究方向为机器学习、推荐系统、数据挖掘。

吴桐(1984-),男,成都数联铭品科技有限公司首席财务官(CFO),主要研究方向为企业征信。

曾途(1984-),男,成都数联铭品科技有限公司CEO,主要研究方向为企业征信。

收稿日期:2018-08-09

基金项目:国家自然科学基金资助项目(No.61433014,No.61673085)

Foundation Items: The National Natural Science Foundation of China(No.61433014, No.61673085)