

大数据发现非法传销网络

Uncovering the illegal pyramid networks by big data



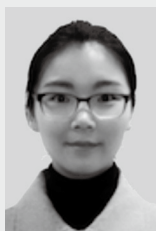
李艳丽 (1990-), 女, 电子科技大学大数据研究中心博士生, 主要研究方向为关键节点挖掘、网络结构分析、链路预测、推荐系统。



刘阳 (1981-), 男, 中国移动通信集团公司贵州分公司工程师、承载室经理, 具有较丰富的信息系统设计与管理经验。



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。



罗秀 (1985-), 女, 中国移动通信集团公司贵州分公司数据分析师, 主要研究方向为大数据分析与管理, 具有多年数据挖掘分析与数据管理工作经验。

中图分类号: TP391

文献标识码: A

doi: 10.11959/j.issn.2096-0271.2017056



徐腾 (1987-), 男, 中国移动通信集团公司贵州分公司中级通信工程师、网络部监控室传输主管, 主要研究方向为传送网运维管理, 具有多年传送网维护及项目管理工作经验。



翁先正 (1981-), 男, 中国移动通信集团公司贵州分公司工程师、网络运维主管, 主要研究方向为传送网运维管理, 具有多年传送网维护及项目管理工作经验。



马国彬 (1982-), 男, 中国移动通信集团公司贵州分公司中级通信工程师、注册信息安全专业人员、网络部网维中心传输主管, 主要研究方向为网络安全管理及传送网运维管理, 具有多年网络安全管理、传送网维护及项目管理工作经验。



尚杰 (1986-), 女, 中国移动通信集团公司黑龙江分公司人力资源部薪酬主管, 拥有国家法律职业资格证书以及处理网络安全法律相关问题的丰富经验。



许海泉 (1985-), 男, 就职于中国移动通信集团公司云南分公司, 主要工作方向为大数据应用场景分析探索, 具备丰富的数据分析、营销策划、项目管理专业知识及多年工作实践经验。



邓先晖 (1983-), 男, 中国移动通信集团公司贵州分公司中级通信工程师、网管支撑室骨干, 主要研究方向为支撑系统维护、实施、开发, 具有多年IT工作经验。



康丽 (1989-), 女, 中国移动通信集团公司黑龙江分公司市场部业务主管, 主要研究方向为市场运营, 具有多年市场营销经验。



侯丽霞 (1983-), 女, 广州咨元信息科技有限公司销售经理, 具有多年运营商大数据分析系统需求调研与分析经验。



陈端兵 (1971-), 男, 电子科技大学大数据研究中心副教授, 主要研究方向为数据挖掘、网络结构分析、信息传播与推荐。



周涛 (1982-), 男, 电子科技大学大数据研究中心教授、主任、博士生导师, 主要关注统计物理与信息科学、社会学、经济学等领域的交叉科学问题, 发表SCI论文200余篇, 引用16 000余次, H指数超过63。

非法传销早在20多年前由美国流入中国,在国内一直是一个屡禁不止的毒瘤。最近,东北大学毕业生李文星之死使得这个问题再次成为所有人关注的焦点。正如魏泽西之于虚假广告,徐玉玉之于电信诈骗,李文星的意外死亡产生的影响使得在国内多地明火执仗,存在多年,甚至和本地人民“水乳交融”的传销组织突然面临前所未有的打击。最近,国家工业和信息化部更是联合国家工商行政管理总局、教育部、公安部、人力资源和社会保障部4个部门印发了《关于开展以“招聘、介绍工作”为名从事传销活动专项整治工作的通知》,开展为期3个月(2017年8月15日—2017年11月15日)的传销活动专项整治行动。

传销危害人员之广,造成财产损失之大,历经时间之久远,骗术花样之繁多堪称触目惊心。然而再多的骗术、攻心术在数据面前立刻显现出最本质的单一模式。利用数据挖掘技术,结合网络分析理论,从传销组织的组织机制、行为模式角度识别传销,不仅可以以不变应万变,发现传销组织的团体涉案成员,还有助于发现传销组织中的终级背后操盘者,辅助公安部门根除传销组织,帮助各大运营商提醒客户最大力度地避免传销电话。此外,这还可以减少公安部门大量的物力、人力、财力,精准定位传销组织窝藏地点,减少人民不必要的财产和人员损失。

最近,笔者和中国移动通信集团公司合作探索了如何快速从通话行为中甄别传销网络,分析了一个子网络中4天的通话网络,并处理成无向网络。其中用户被分成4类:普通用户、某大型企业员工、服务账号、传销组织人员,分别从这些用户中选取一位作为种子节点,画出从这个节点出发按广度优先搜索获得的子网。这样的网络被称为自我中心网络(ego network)^[1,2]。图1~图4分别是

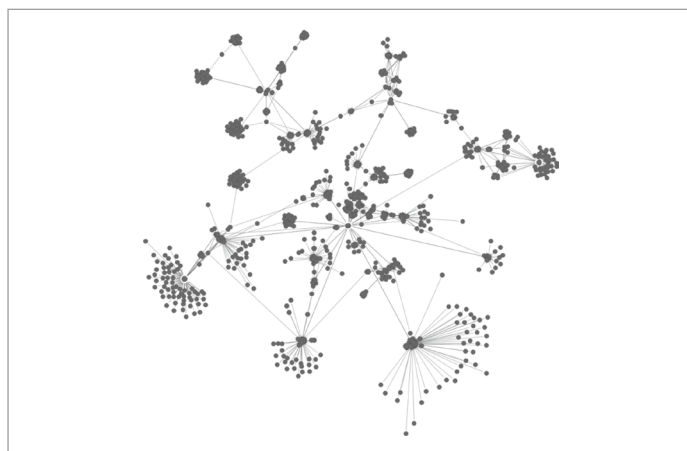


图1 某普通用户的自我中心网络

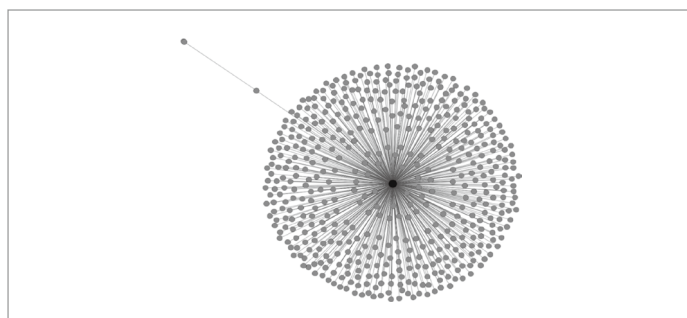


图2 某服务账号的自我中心网络

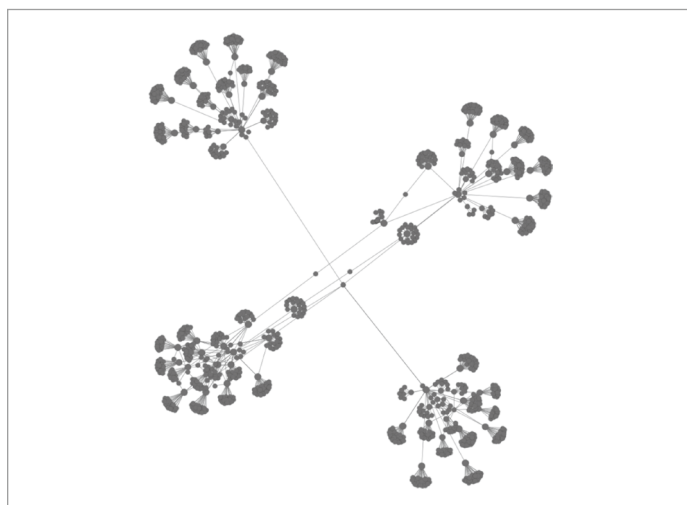


图3 某企业领导的自我中心网络(局限在本企业员工中)

可以看出,普通用户、服务账号自我中心网络与传销网络相去甚远,但是公司网络也有类似于传销网络的层次管理结构^[3-5]。通过4种常见用户群组通话网络的比较发

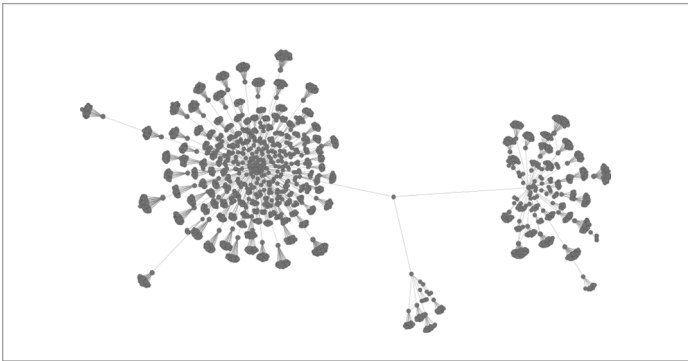


图 4 某非法传销组织头目的自我中心网络

现,传销网络呈现出与实际业务模式一致的类树状层级结构。而与传销组织有着相同层级模式的公司用户群网络层级模式也非常明显,但是非相邻层级之间、非同一组之间的节点也有通话(同一组指的是几

个节点有共同的上线节点,这里的上线和下线是以选择的根节点为参照)。因此,总结了传销网络3种明显的特征:节点之间不跨级联系、同级不抢占下线、同级非同上线节点不联系。基于总结的传销组织特征,定义了3类特征指标,见表1中的I1、I2、I3。I1指标中如果一个节点*i*有两个位于上一层的连边 j_1 和 j_2 ,则按照广度优先搜索的顺序,第一个扩展到*i*的节点被认为是*i*的父节点。

当然,传销网络还有很多特征是和一般网络不一样的,例如涉及传销网络的用户出度和入度分布都和其他网络大不一样^[6],如图5和图6所示。

本文用I4、I5、I6 3个指标量化不同网络出度和入度分布的差异。如果出度为*i*的

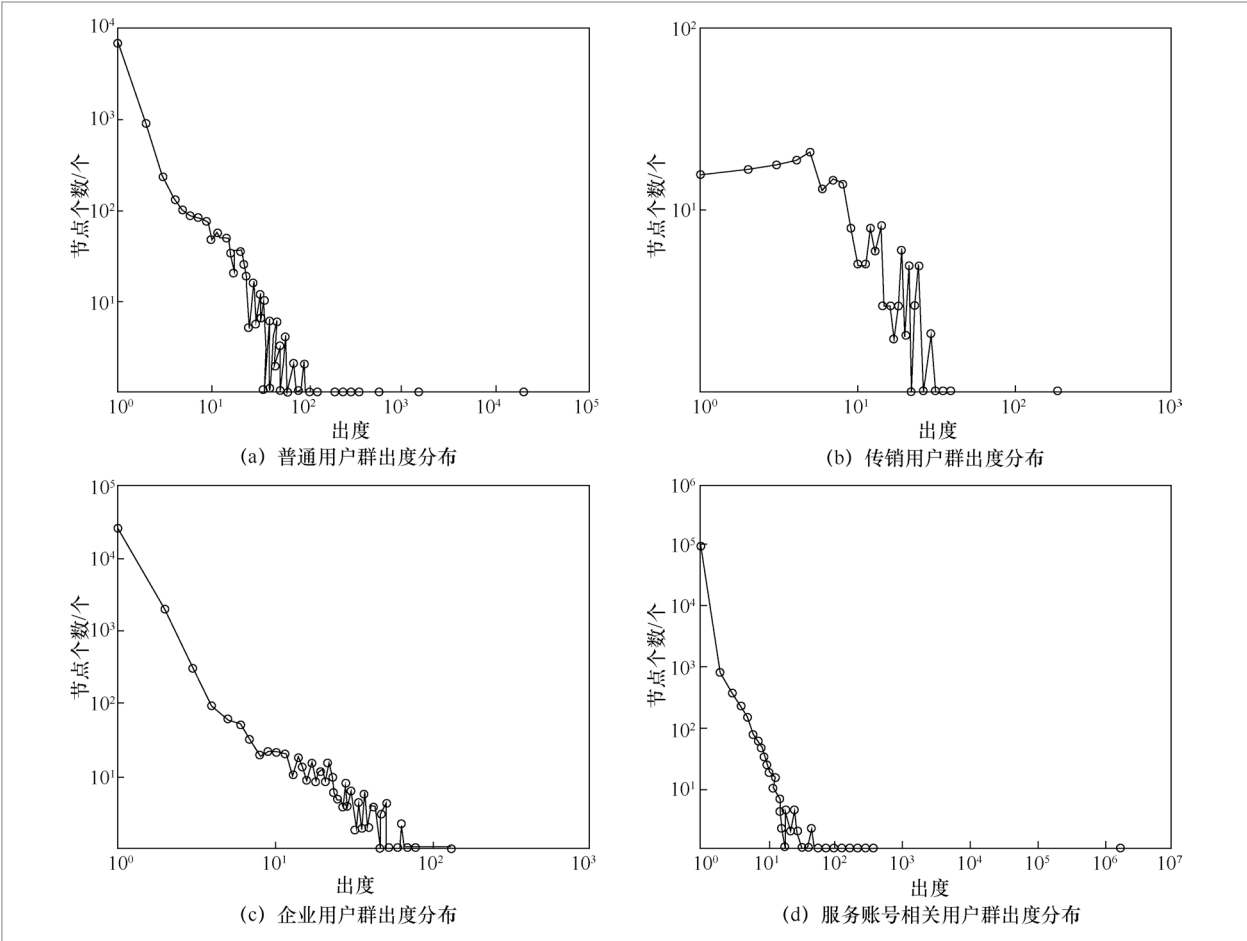


图 5 出度分布的比较

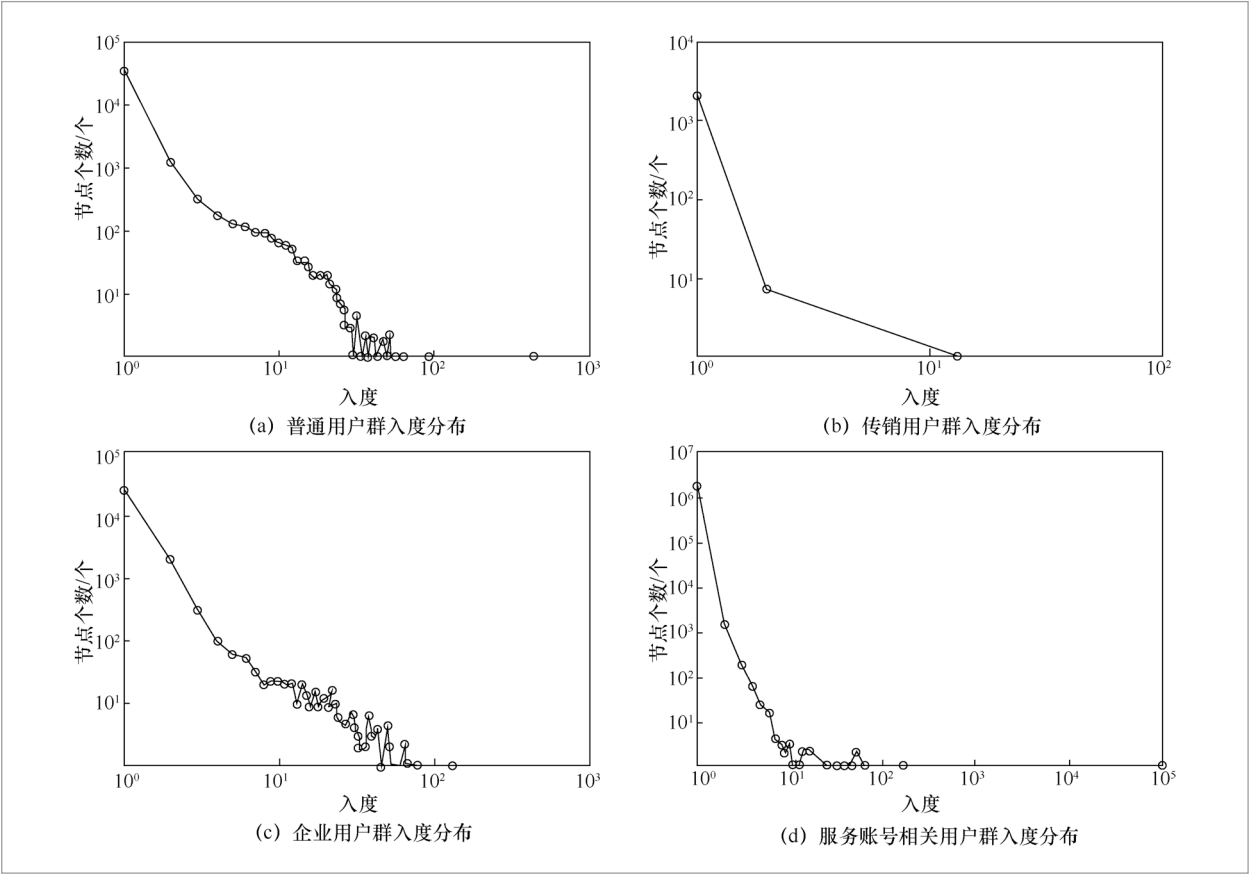


图6 入度分布的比较

节点个数是 n_i ,则 $I4=\frac{\sum_{i=2}^{10}n_i}{n_1}$ 。对于I5、I6,引

用信息熵的概念量化一个度序列的异质性^[7,8],其计算式为 $H=-\sum_{k=1}^{N-1}p(k)\log p(k)$,其中, N 为节点个数, $p(k)$ 是度为 k 的节点出现的概率,值越大,度序列分布越异质。从表2中可以看到,相比其他3类网络,传销用户群的出度分布和入度分布的信息熵值都是最小的,这也表示传销用户每个用户的接入上线数量和下线数量更加确定。

仅仅利用这6个指标就可以很好地分辨出不同自我中心网络是不是疑似传销网络。

根据以上分析,可以设计疑似传销网络的生长算法,从任何一个节点出发,生长出可能的最大疑似传销网络,然后根据生长情况判断

表1 网络属性量化指标

指标	含义
I1	非父子的连边数
I2	同层同一父节点的连边数
I3	同层不同父节点的连边数
I4	出度在区间[2,10]的节点总数与出度为1的节点总数之比
I5	出度分布的信息熵
I6	入度分布的信息熵

表2 基于4类有向无权网络计算的指标信息

指标	普通用户群	传销用户群	企业用户群	服务账号相关用户群
I1/条	3 151	15	1 816	1 642
I2/条	747	4	353	220
I3/条	1 316	0	801	263
I4	0.24	8.13	0.11	0.02
I5	4.0	2.8	4.2	3.2
I6	4.0	1.5	3.5	2.9

它是不是一个传销网络。见表3，笔者设计的算法识别的传销网络几乎能够覆盖节点数占比大约只有0.1%的疑似传销人群，但是对于其他已经甄别出来没有问题的人群完全不误判。原始数据的分类标签是由公安部门给出的，一些更细致的特征还可以进一步提高算法效果，但因为涉及敏感信息，本文此处不详细介绍。

如图7所示，笔者提出的算法基本覆盖了传销网络的所有核心，只有少量外围节点算法本身无法确认是不是传销分子。但是这完全不影响调查和抓获这个传销组织，特别是不会影响对传销头目和组织结构的识别。

个人隐私数据的泄露给了很多电信诈骗不法分子以可乘之机，但数据科学的正确使用也可以帮助揭示非法组织，保护人们的人身和财产安全。传销组织无法通过本文所描述的一些特征来躲避本算法的计算，因为这些特征更改意味着改变传销的组织和运营模式，这样的代价他们无法接受。并且本文描述的只是算法中最简单的一小部分，笔者并不仅仅只是通过这些特征量识别非法传

销。除了本文的研究，数据挖掘和网络分析技术在辅助挖掘犯罪分子方面的应用还很多，在拥有了移动轨迹、行为数据的基础上，可以更深入地分析犯罪分子异于常人的行为模式，达到挖掘潜在嫌疑人、快速定位嫌疑人的目的^[9]。

参考文献：

[1] LESKOVEC J, MCAULEY J J. Learning to discover social circles in ego networks[C]// Neural Information Processing Systems Conference, December 3–6, 2012, Nevada, USA. [S.l.:s.n.], 2012: 539–547.

[2] WANG Q, GAO J, ZHOU T, et al. Critical size of ego communication networks[J]. Europhysics Letters, 2016, 114(5): 58004.

[3] RAVASZ E, SOMERA A L, MONGRU D A, et al. Hierarchical organization of modularity in metabolic networks[J]. Science, 2002, 297(5586): 1551–1555.

[4] RAVASZ E, BARABASI A L. Hierarchical organization in complex networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2003, 67(2): 026112.

[5] CLAUSET A, MOORE C, NEWMAN M E J. Hierarchical structure and the prediction of missing links in networks[J]. Nature, 2008, 453(7191): 98–101.

[6] BARABASI A L, ALBERT R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439): 509–512.

[7] SHANNON C E. A note on the concept of entropy[J]. Bell System Technical Journal, 1948, 27(3): 379–423.

[8] WANG B, TANG H, GUO C, et al. Entropy optimization of scale-free networks' robustness to random failures[J]. Physica A: Statistical Mechanics and Its Applications, 2006, 363(2): 591–596.

[9] DU B, LIU C, ZHOU W, et al. Catch me if you can: detecting pickpocket suspects from large-scale transit records[C]// The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13–17, 2016, San Francisco, USA. New York: ACM Press, 2016: 87–96. □

表3 节点覆盖度

可疑种子来源	原有节点规模/个	覆盖节点规模/个
普通用户群	39 343	0
传销用户群	2 116	1 572
企业用户群	15 598	0
服务账号相关用户群	1 700 666	0

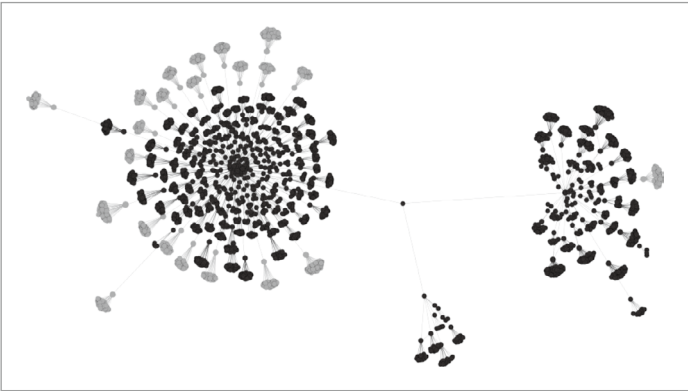


图7 算法识别出的可疑节点的覆盖区域
(黑色节点为算法识别出的可疑节点)