知乎社交网络性质分析、聚类与预测

张悦嘉 16307110399 孟博宇 16307100012 林雅文 16307090185

摘　要 知乎是国内最大的知识问答平台，研究知乎用户网络的社交性对于互联网发展与社会学研究都具有重大意义。本文通过爬取知乎机构号下的所有粉丝及相互关注关系，研究以机构号为中心的知乎子网络，分析网络及其中心结点的基本属性，发展了基于相似度的边预测算法，生成了基于网络模型的社区画像标签，希冀为社交网络研究与知乎发展提供指导建议。

关键词 社交网络；知乎；社区画像；边预测

1 背景介绍

1.1 知乎——介于知识分享与社交网络之间的平台

知乎是中文互联网知名的可信赖问答社区，它致力于构建一个人人都可以便捷接入的知识分享网络，让人们便捷地与世界分享知识、经验和见解，发现更大的世界。作为一个问答社区，知乎连接着各行各业的用户。知乎用户们通过知识建立信任和连接，对热点事件或话题进行理性、深度、多维度的讨论，分享专业、有趣、多元的高质量内容，打造和提升个人品牌价值，发现并获得新机会[1]。知乎除了最基础的问答功能外，也有类似微信公众号发布“文章”与“专栏”的功能，类似微博发布“想法”的功能，类似视频平台举办“live”的功能； 在知乎，用户不仅可以围绕着某一感兴趣的话题进行相关的讨论，同时也可以和其他用户交流互动，对其他用户进行“关注”、“点赞”甚至“屏蔽”。因此，我们认为知乎是一个介于知识分享网络与社交网络之间的平台。那么，知乎的社交性究竟有多强？从社交网络的角度来看，知乎里的用户网络又有哪些独特的性质？

1.2 知乎官方机构号——集传播、互动、反馈为一体的媒体

知乎上，不仅有大量个人用户，还存在不少知名的官方机构号。这些机构号是集传播、互动、反馈为一体的媒体。对于官方机构号而言，在知乎上发布文章，有传播速度快、阅读量大的好处，更重要的是能够获取用户的实时反馈。此外，机构号能够借助知乎这一平台对粉丝的行为特征进行更为细致的描摹。

为了了解知乎社交网络的性质，我们从知乎上爬取相关用户信息进行分析。由于知乎用户数较大，我们仅从知乎上爬取两个具有代表性的子网络进行分析。这两个子网络分别是官方机构号“京师心理大学堂”和“36 氪”及它们的粉丝。

京师心理大学堂是北京师范大学心理学部在知乎上建立的官方机构号，其以打造最优质的心理科普平台为目标，以专业的视角，在知乎上生产了大量优质的文章，并对知乎上的提问从心理学角度进行解读，是知乎认可的心理学话题的优秀回答者。编辑推荐、知乎圆桌、知乎周刊和知乎日报共收录了京师心理大学堂的 82 个回答和 41 篇文章。京师心理大学堂拥有多达 27 万的粉丝，其中除了大量普通用户之外，还不乏心理学专业从业者、心理学专业学习者、以及其他心理服务平台。

36 氪是一个创业服务平台，致力于为创业者提供创业咨询、科技新闻、投融资对接、股

[1] 摘自知乎简介。

权投资、极速融资等创业服务，其在知乎上成立了同名官方机构号。36 氪同样也是知乎中非常活跃的官方机构号，其发表了 500 多篇文章和将近 2000 个想法，吸引了接近 25 万个粉丝。关注 36 氪的粉丝除了大量有志于创业的个人用户之外，还有腾讯、微软亚洲研究院等知名机构。

由于京师心理大学堂和 36 氪的受众不同，我们猜测 36 氪和京师心理大学堂这两个官方机构号所产生的社交网络也会具有不同的网络性质。两个机构号的粉丝数大体相当，通过对这两个不同并具有代表性的子网络的分析，我们可以从中窥探整个知乎社交网络的性质。

2 研究方法与数据来源

2.1 建立模型

我们将知乎用户视作结点，用户之间的关注关系视作有向边，构建有向图。记整个知乎网络的有向图为$G = G(V, E)$，则官方机构号 O 的粉丝群体$G_O = G\,(V_O, E_O)$由以下结点和边构成：

$$V_O = \Gamma_{in}(O) = \{v \in V | (v, O) \in E\}$$

$$E_O = \{(v_1, v_2) | v_1, v_2 \in V_O, (v_1, v_2) \in E\}$$

在下文中，用$\Gamma_{in}(v) = v' \in V|(v', v) \in E$表示 v 的入度邻居结点集合，用$\Gamma_{out}(v) = v' \in V|(v, v') \in E$表示 v 的出度邻居结点集合，而 v 的邻居结点$\Gamma(v) = \Gamma_{in}(v) \cup \Gamma_{out}(v)$是两者的并集。

定义 User 为网络中入度为 0 的用户，Leader 为网络中入度大于 0 的用户，即

$$U = \{v \in V | \Gamma_{in}(v) = \emptyset\}$$

$$L = \{v \in V | \Gamma_{in}(v) \neq \emptyset\}$$

2.2 研究方法

我们首先对结点的属性（即用户的个人信息）进行描述性分析，计算G_O的网络属性，包括度分布、PageRank 中心性、直径与平均路径长度等，并用可视化技术展现。通过了解网络的基本性质，可以大致判断知乎的社交性。

对网络与结点性质的进一步分析如下：

1. 针对用户的 PageRank 中心性，用随机森林模型分类。分析与用户影响力相关的属性，预测用户在机构号粉丝中的影响力，详见第 4 节。
2. 设计基于结点相似度的算法预测网络中边的形成，并用 AUC 对边预测性能进行评价，详见第 5 节。该算法可用于为普通用户推荐“知乎大 V”，即拥有众多粉丝的知乎用户。
3. 调用 Louvain 算法对用户进行聚类，挖掘粉丝群体中存在的社区。设计算法，用少数几个 Leader 用户的自我中心网络近似估计每一类别的子网络，并用 Leader 的标签给类别上标签，详见第 6 节。该算法可以为机构号制定有针对性的营销策略提供建议。

2.3 数据收集

2.3.1 爬虫

我们借助了 Python 的 Scrapy 框架，运用模拟登录进行数据爬取。

最初，我们从随机选取的种子结点出发，爬取他的粉丝列表与关注列表，迭代爬取他的粉丝与关注的粉丝列表与关注列表。然而，这样做的问题在于结点之间的边并不完整，

比如拥有 200 万粉的知乎用户张佳玮在网络中只有 2 条边。于是，我们每新加入一个结点，就遍历他的粉丝与关注列表，确保网络的完备性。但这一算法依旧有不足之处：从一个特定的人出发，不具有整体网络的代表性。

在知乎网络中，机构号用户下的粉丝为一个显式社区，所以我们决定爬取机构号的粉丝以及粉丝间的关注。但每次遍历新加入的节点的关注列表与粉丝列表，爬取速度过慢。

我们分析所有结点关注与粉丝的和大概是千万量级的，而且为了确认已经在我们数据中的结点哪些是大 v 结点粉丝，这样效率过低。但是我们如果已经知道有哪些结点，那么只需要遍历所有结点之间的关注就行，而不需要遍历粉丝。如果我们第一遍只遍历所有结点而不管边之间的关系，这个耗费大概是 10 万级别的，对于再重新遍历一遍已有结点的关注中是不是有在我们数据中的结点这种千万级别的操作是微不足道的，大大节省了我们爬虫的时间。

2.3.2 数据记录

用 MySQL 存储创建两张表，一张记录结点，一张记录边。在结点表中记录结点的知乎用户名以及一系列个人信息，在边表中记录边的起始节点和终止节点。另外给结点的用户名列加索引，加速查询。

3 数据简介与性质分析

3.1 数据简介

从 2019 年 5 月 28 日至 2019 年 5 月 31 日，我们从知乎上爬取官方机构号“京师心理大学堂”及“36 氪”的所有粉丝以及粉丝之间互相的关注关系。经清洗、去重处理后，得到统计数据如表 1 所示。

表 1 网络基本属性

	结点数	边数	互相关注边数	互相关注边比例	网络密度（$\times 10^{-4}$）	平均度
36 氪	231730	1903146	19777	1%	0.35	16.43
36 氪 Leader 网络	33778	908247	19777	2%	7.96	53.78
京师心理大学堂	277670	939435	4811	0.5%	0.12	6.77
京师心理大学堂 Leader 网络	10824	160087	4811	3%	13.67	29.58

从上述统计信息中，我们可以得到以下结论：

1. 从整体网络来看，36 氪网络密度更大，用户间的互动行为更强。同时，相互关注的边占总边数的比例也更高。
2. 从 Leader 网络来看，京师心理大学堂的网络密度更大，相互关注的边占总边数的比例更高。
3. 在 36 氪网络中，Leader 结点占比 14.6%，Leader 间的边数占比近五成；在京师心理大学堂网络中，Leader 结点仅占 3.9%，Leader 间的边数占比仅 17%。

除了网络信息外，我们还收集了用户的个人信息，包括回答数、文章数、是否认证用户、是否优秀回答者、专栏数、收藏数、被收藏数、粉丝数、关注专栏数、关注用户数、关注收

藏数、关注问题数、性别、举办的 Live 数、是否广告主、是否组织机构、参与公共编辑数、想法数、提问数、被感谢数、是否是 vip、被赞同数。在后续的分析中，可以用来刻画结点的属性。

3.2 描述性分析

3.2.1 基本信息分布

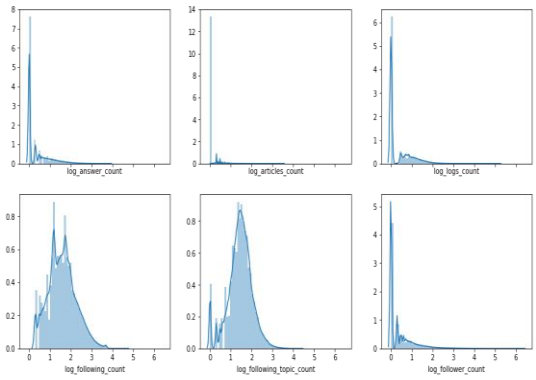


图 1：36 氪

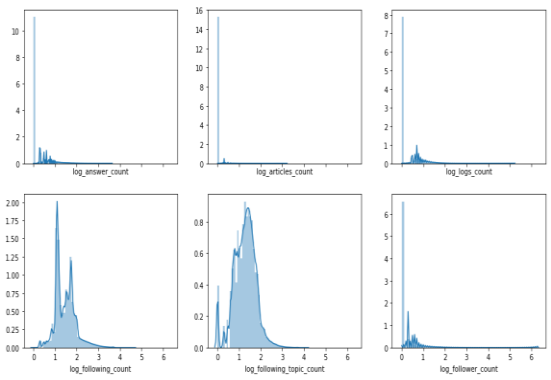


图 2：京师心理大学堂

我们发现，36 氪和京师心理大学堂这两个网络的对数回答数、对数文章数、对数参与公共编辑数、对数关注数、对数关注话题数、对数粉丝数分布相似。

上面三张图分别是对数回答数、对数文章数、对数参与公共编辑数的分布，可以发现，分布呈明显的右偏，两个网络大部分的粉丝，基本都不怎么回答问题，不怎么发文章，较少参与公共编辑；左下两张图分别是对数关注数和对数关注话题数的分布，可以发现，这些用户尽管不怎么回答问题，发表文章，但多多少少都会关注一些用户或话题。

从中可以一窥知乎的特点：大部分用户都是来“围观”，来知乎“寻找答案”，只有少数用户（很可能是某个领域的专业人士）才会踊跃发言。从右下的对数粉丝数分布也可以推测出，那些少数踊跃发言的用户拥有大量的粉丝，而大多数的用户粉丝数并不多。

3.2.2 性别比例

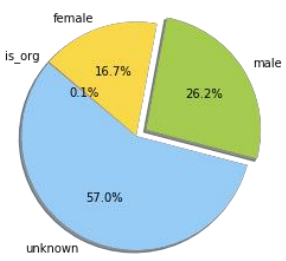


图 3：36 氪性别比例

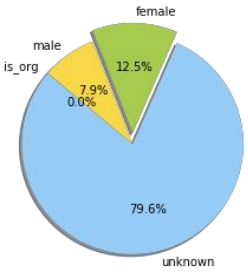


图 4：京师心理大学堂性别比例

两个网络中均有大部分的用户未填写自己的性别，已知性别的用户中，36 氪网络的关注者男性比例远高于女性；京师心理大学堂则是女性关注者比例更多。

3.3 网络性质分析

3.3.1 中心性

结点中心性是结点的属性，可以体现结点的影响力、权威性、中枢性，用来识别网络中最重要的结点。

图 9 和图 10 分别是两个网络中度中心性和特征向量中心性排名前十的结点。

	36氪		京师心理大学堂
丁香医生	59311	丁香医生	102345
王瑞恩	25717	刘看山	65413
楠爷	24554	即可运动	63087
周源	23813	中国科普博览	33638
Sean Ye	21638	简单心理	28844
以太创服	18344	木棉959	23007
汪惟	18191	叶壮	20552
王诺诺	16344	宏桑	20412
简浅	15702	动机在杭州	18791
万金油	15615	朵拉陈	18344

图 9 度中心性排名前 10 的结点

	36氪		京师心理大学堂
丁香医生	1	丁香医生	1
王瑞恩	0.446145	刘看山	0.637492
楠爷	0.424976	即可运动	0.614187
周源	0.421856	中国科普博览	0.330153
Sean Ye	0.372774	简单心理	0.282566
汪惟	0.317626	木棉959	0.224962
以太创服	0.315998	叶壮	0.203844
王诺诺	0.291675	宏桑	0.198883
山羊月	0.271893	动机在杭州	0.187587
简浅	0.266729	朵拉陈	0.18067

图 10 特征向量中心性排名前十的结点

PageRank 中心性的核心思想是改进没有出边的结点不会有中心性的问题，并且结点的出边不是获得结点的所有中心性，而是等额分配其中心性。图 11 是两个网络中 PageRank 中心性排名前十的结点。

	36氪		京师心理大学堂
周源	0.0221	丁香医生	0.044635
丁香医生	0.021304	中国科普博览	0.039199
机器之心	0.012275	动机在杭州	0.020232
楠爷	0.01075	刘看山	0.019475
王瑞恩	0.010576	叶壮	0.017437
王诺诺	0.010431	赵思家	0.016414
庄明（rosicky311）	0.009085	即刻运动Official	0.014323
以太创服	0.008779	Steve Shi	0.013943
汪惟	0.00867	刘柯	0.013455
Sean Ye	0.008067	王瑞恩	0.013083

图 11 PageRank 中心性排名前十的结点

3.3.2 直径与平均路径

我们需要穷举任意两个结点间的最短路径其中最长的为直径，平均值为网络平均路径。最短路径算法整体时间复杂度为 $O(n^3)$。但是数据量太大加上算法复杂度大，导致计算时间过长。我们想出的解决方法为只计算前文定义的 leader 结点之间的直径和平均路径，由于点与边大幅减少，所以计算速度变快。我们能根据 leader 节点的计算结果给出一个整个网络的上界，因为任意节点只需要连入 leader 网络（路径长度 1）通过 leader 网络的平均路径（记为 r）连出网络（路径长度 1）就能找到网络中的任意节点。注意，我们这里给出的上界是非常宽松的，因为很多网络外的节点共同属于一个 leader 节点的粉丝，他们并不需要通过 leader 网络找到彼此。计算结果如图 12 所示。可以看出两个网络的平均路径的上界都是 6 左右，符合六度分隔理论。

	Leader_diameter	All_diameter
36氪	13	<15
京师心理大学堂	13	<15

	Leader_avg_diameter	All_avg_diameter
36氪	4.36	<6.36
京师心理大学堂	4.61	<6.61

图 12 网络直径与最短路径

3.3.3 度分布

假设度数为 d 的结点在整个网络结点中所占的比例为 p，则根据幂律分布，有$\log p = k \log d + b$。分别用结点属性与网络中结点度数拟合斜率 k，结果如图 13 所示。可以看出曲线有明显的长尾效应，而左上方近似成为一条直线，符合幂律分布，具有无标度网络特征。

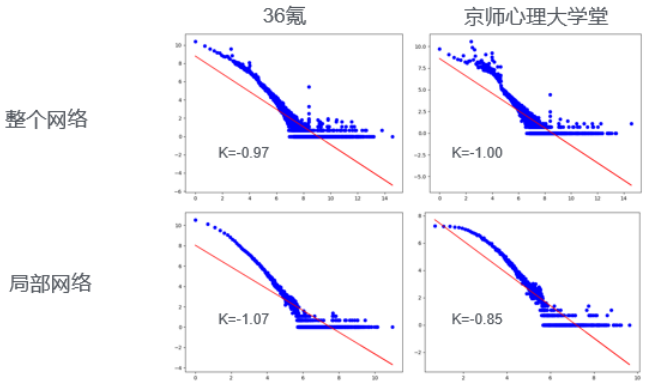


图 13 结点在知乎网络与子网络中的度分布

3.3.4 网络可视化

我们用 Gephi 软件进行可视化，选用数据集为两个网络中的 leader 节点，节点大小依据 PageRank 中心性，节点颜色依据基于 Louvain 算法算出的结点所属社区。

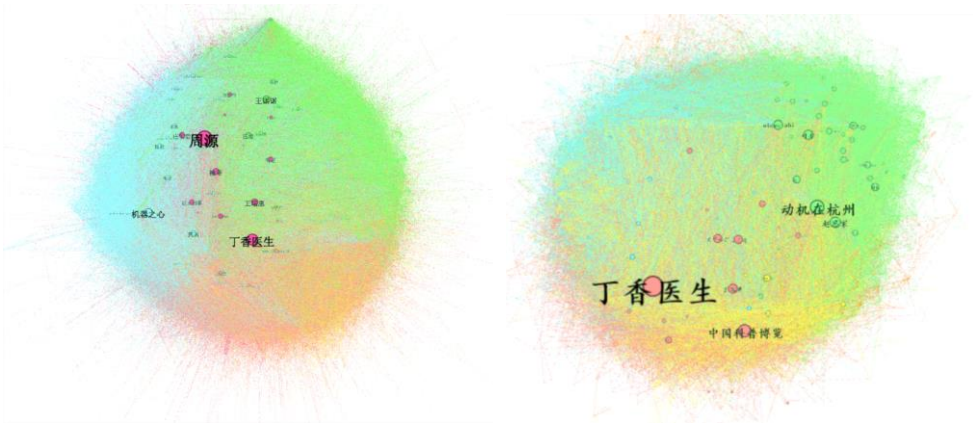


图 14 36 氪（左）与京师心理大学堂（右）网络可视化

4 数据挖掘——用户影响力因素分析

4.1 分类问题描述

前述网络分析中，我们根据用户之间的关注关系，分别计算了 36 氪和京师心理大学堂两个网络中每个用户的 PageRank 中心性。那么，中心性高的用户具有什么样的特征？什么样的用户在网络中更具影响力？我们分别将两个网络中 PageRank 中心性值>0.00001 的用户定义为高影响力用户，将其他用户定义为低影响力用户，通过随机森林模型寻找对用户影响力最为关键的因素。

4.2 数据处理与采样

我们分别将数据集按 60%、20%、20%的比例划分为训练集、验证集和测试集。36 氪和京师心理大学堂中分别有 5%和 1%的用户为高影响力用户，由于类别不均衡，我们采用 SMOTE 算法增加正样本数量使正负样本数量均衡。

4.3 结果与分析

最终随机森林模型的表现为：36 氪的 AUC 为 0.94，京师心理大学堂的 AUC 达到 0.97。根据随机森林模型的结果，对用户影响力最为关键的前十个因素如图 15 所示。

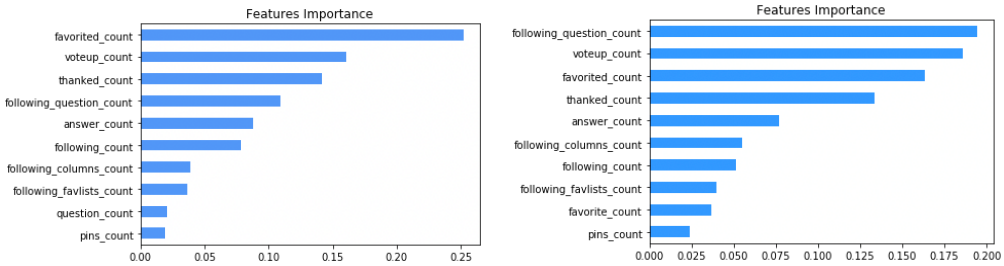


图 15　36 氪与京师心理大学堂-随机森林特征重要性

36 氪网络中，最能区分用户影响力的因素是被收藏数(favorited_count)，其次是被赞同数(voteup_count)和被感谢数(thanked_count)，说明一个有高影响力的用户往往会有高质量的回答和文章。收藏数的重要性最高，说明一个用户的回答或文章不仅需要受到认可和赞同，更要能够对其他用户产生启发，值得其他用户珍藏，这样的用户才会拥有更高的影响力。而京师心理大学堂中，最能区分用户影响力的因素则是关注问题数(following_question_count)，其次是被赞同数和被收藏数，这说明除了回答问题、发表文章以外，参与感对于影响力同样重要，用户需要多关注知乎中的问题并积极参与问题的讨论，才能拥有更高的影响力。

5 边预测

5.1 预测目标

Leader 之间本身边就很密集，预测这些边的产生并没有什么价值。所以我们要预测 Leader 结点与 User 结点是不是可能产生边，Leader 节点可以理解为内容的生产者，我们可以为 User 也就是普通用户推荐优质的内容。

5.2 基于相似度的边预测算法

5.2.1 Jaccard 指标

Jaccard 指标用于评价两个结点间的相似度。基于 Jaccard 指标为每条可能存在的边（不一定真实存在）打分，即

$$\sigma(u,l) = \frac{|\Gamma(u) \cap \Gamma(l)|}{|\Gamma(u) \cup \Gamma(l)|} = \frac{|\Gamma_{out}(u) \cap \Gamma(l)|}{|\Gamma_{out}(u) \cup \Gamma(l)|}$$

5.2.2 调整 Jaccard 指标

Jaccard 指标适用于无向图。当我们预测普通用户是否会关注大 V 时，考察他们的共同邻居是不合理的。受到协同过滤推荐算法的启发，我们考察普通用户所关注的 Leader 与他可能关注的 Leader 之间的相似度。这一指标合理的原因在于，Leader 网络相对稠密，连通性很好，可以近似看作无向图。调整后的 Jaccard 指标定义如下

$$\sigma^{adjusted}(u,l) = \frac{\left|\bigcup_{l' \in \Gamma_{out}(u)} \Gamma(l') \cap \Gamma(l)\right|}{\left|\bigcup_{l' \in \Gamma_{out}(u)} \Gamma(l') \cup \Gamma(l)\right|}$$

5.3 边预测效果评价

我们的训练集是选取所有 Leader 节点之间的 90 万边，以及 90%的 user 与 leader 之间的边；测试集是剩余的 10%的 user 与 leader 之间的边，以及随机生成的不存在的 user 与 leader 之间的边。评价指标是 AUC，可以理解为在测试集中的边的分数值比随机选择的一个不存在的边的分数值高的概率。每次随机从测试集中选取一条边与随机选择的不存在的边进行比较，如果测试集中的边分数值大于不存在的边的分数，那么就加一分，如果两个分数值相等就加 0.5 分，独立地比较 N 次。

$$AUC = \frac{\sum_{(i_1,j_1)\in E,(i_2,j_2)\notin E} I\big(\sigma(i_1,j_1) > \sigma(i_2,j_2)\big) + \frac{1}{2} I\big(\sigma(i_1,j_1) = \sigma(i_2,j_2)\big)}{N}$$

结果：普通 jaccard 指标：AUC=74%；调整后的 Jaccard 指标：AUC=95%。第二版边预测算法虽然准确率高，但存在不足之处，比如冷启动。

6 社区挖掘

6.1 聚类算法

我们调用了 Gephi 的模块化功能，将结点归入社区中，所用算法是基于模块度增益的 Louvain 迭代算法。

最终，将“36 氪”中的用户分为 56 类，其中超过 10 名用户的类别有 23 类，最大社区结点数为 33676；将“京师心理大学堂”中的用户分为 38 类，其中超过 10 名用户的类别只有六类，最大社区结点数为 69943。从上述统计中可以看出，“36 氪”的粉丝类别比较分散，来自各行各业各领域，网络同质性较低，而“京师心理大学堂”的粉丝类别比较集中，网络同质性较高。

6.2 社区画像

6.2.1 可解释网络

由于类别内用户所在领域太过集中，从领域的角度很难为每一类别上标签。因此，提出可解释结点 L'的概念。我们希望用尽可能少的结点的自我中心网络来近似代表整个网络，这些结点就被称为可解释结点，由可解释结点与结点间的边组成的网络被称为可解释网络。由于这些可解释结点在很大程度上解释了整个网络的组成，所以可以用这些结点的标签来解释类别标签。规定$L' \subseteq L$。严谨的数学表述如下

$$G \approx \bigcup_{l \in L' \subseteq L} G_l^{ego}\,(V_l,\, E_l)$$

where $V_l = \{l\} \cup \{\Gamma(l)\},\ E_l = \{(v_1,\, v_2) | v_1,\, v_2 \in V_l, (v_1, v_2) \in E\}$

从入度最大的结点开始，依次将可解释结点的自我中心网络加入可解释网络，计算可解释网络中结点与边占总网络的比例，直到比例达到某一阈值或曲线出现拐点为止。图#分别表现了“京师心理大学堂”可解释网络与“36 氪”可解释网络中，结点与边的比例随着结点数的增加而增加。

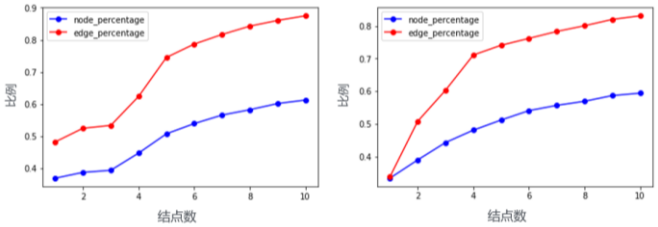


图 16 可解释网络解释性的变化

以左图“京师心理大学堂”的可解释网络为例，对网络的可解释性进行解读。可以看到，7 个大 V 的邻居网络包含了超过 80%的边和不到 60%的结点。这 7 位大 V 依次是：丁香医生（医疗健康服务平台）、刘看山（知乎吉祥物）、即刻运动（体育 APP）、简单心理（心理咨询平台）、木棉 959（京师心理大学堂主编、北师大心理专业在读硕士）、叶壮（作家、心理学培训师）、宏桑（百度运营经理）。由此，可以刻画京师心理大学堂粉丝的爱好——喜欢心理学，喜欢医学，需要心理咨询，爱好体育运动，对互联网感兴趣。

注意到从 1-3 结点数变化不大，说明刘看山和即刻运动的粉丝也关注了丁香医生。而后面几位大 V 带入了新的粉丝，尤其是简单心理和木棉 959 带入了大量新的结点和边，可以认为是京师心理大学堂从三个不同渠道获取的粉丝——医疗健康服务/心理咨询/心理知识科普。

6.2.2 社区特征

把讨论范围集中在结点数大于 10 的社区上。一部分社区呈现出两类典型的社区特征——拼凑型社区和紧密型社区，剩余社区处于二者之间，属于混合型社区。拼凑型社区的社交性较弱，而紧密型社区的社交性较强。

6.2.2.1 拼凑型社区

这类社区的特点是结点与边数同步增长，结点与边数在同一个数量级上，互粉边密度低——说明子网络由几位大 V 的自我中心网络合并形成，如图 17 所示。增长速度越快，说明这位大 V 引入的粉丝/边越多，是社区中的中心节点，如图 18 所示。

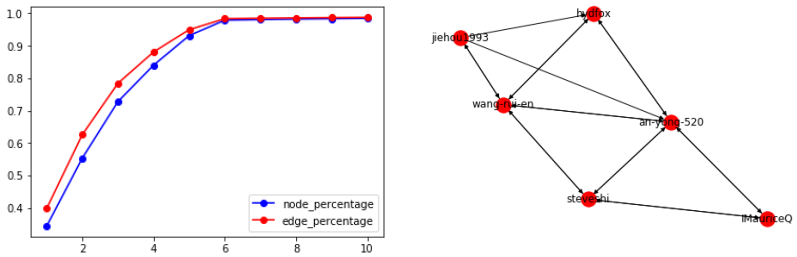


图 17 拼凑型社区示例 1

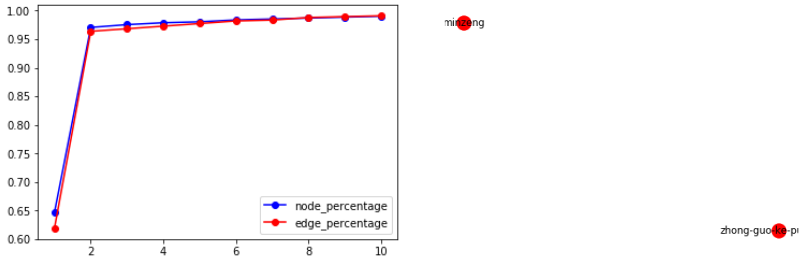


图 18 拼凑型社区示例 2

6.2.2.2 紧密型社区

这类社区的特征是其涵盖的边数大于点数，且互粉边密度高，说明大 V 之间形成了一个一个小圈子，如图 19 所示。

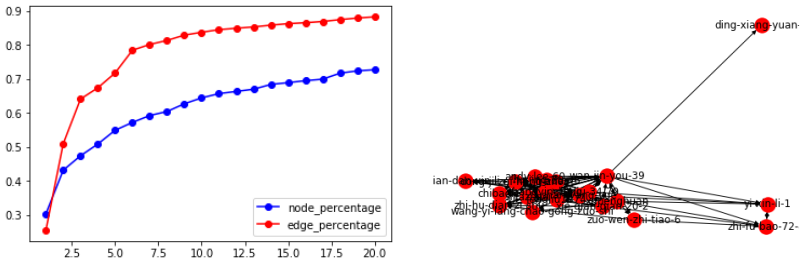
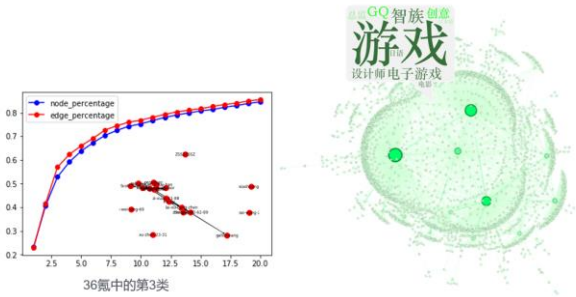


图 19 紧密型社区示例

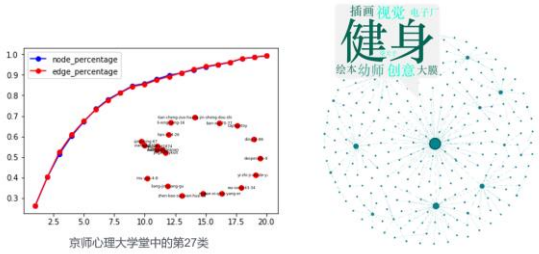
6.2.3 类别标签

获取了每一社区的可解释结点后，我们通过用结点的标签来为社区赋予标签，数据来源是相应用户的毕业院校、工作单位、一句话介绍、个人简介等信息，调用 jieba 包经过分词处理后，将这些词语作为类别标签，标签权重为词语的 tf-idf 值。现展示部分结果，词云中词语的大小表示标签的权重。

（1） 36 氪网络中一个典型的拼凑型社区，主要标签与游戏、设计有关。可以看到普通用户紧密地围绕在他们所关注的大 V 周围。

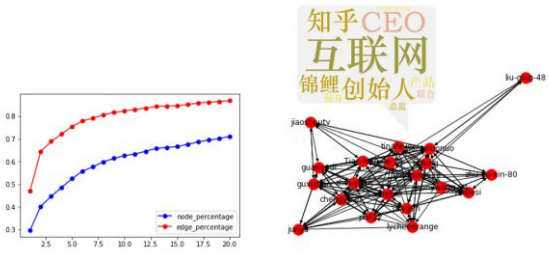


（2） 京师心理大学堂网络中一个典型的拼凑型社区，类别标签有“健身”、“插画”、“幼师”、“视觉”、绘本”等，五花八门什么都有。可以看到普通用户只与它所关注的大 V 相连。



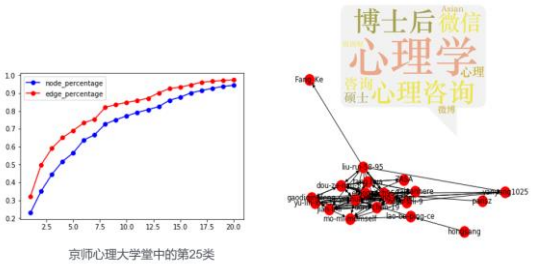
京师心理大学堂中的第27类

（3） 36 氪网络中一个典型的紧密型社区，类别标签有“互联网”、“创始人”、“知乎”、“CEO”等，大 V 之间关系紧密，而普通用户也同时关注了多位大 V。



36氪中的第2类

（4） 京师心理大学堂网络中一个典型的紧密型社区，主要标签为“心理学”、“心理咨询”、“博士后”、“微信”、“微博”等。从“微信”、“微博”等关键词可以看出这一社区的社交性非常强。



京师心理大学堂中的第25类

值得一提的是，紧密型社区的标签与机构号非常接近，比如 36 氪是一个创业服务平台，其紧密社区的标签为“创始人”、“互联网”，与 36 氪的主要业务息息相关；而京师心理大学堂是一个心理科普平台，其紧密社区的标签为“心理学”、“心理咨询”、“博士”，可见其中的大 V 都是这一领域的专业人士。由此我们可以断定紧密型社区中的用户与机构号的主要业务更加接近，他们是机构号粉丝中的核心成员。机构号可以根据不同的社区为用户制定不同的营销策略。

完整的社区挖掘与画像见附录。本算法后续的改进方向在于利用知识图谱，对原始标签进行概括和分类。

7 总结

7.1 结论与建议

本文通过爬取知乎网“36 氪”及“京师心理大学堂”机构号下的所有粉丝及相互关注关系，研究以机构号为中心的知乎子网络，获得主要结论如下：

1. 从网络结构来看，“36 氪”的网络密度更大，社交性比“京师心理大学堂”更强；但“京师心理大学堂”的 Leader 网络密度更大，社交性更强。京师心理大学堂有一个更小、更紧密的 Leader 网络，整个网络贡献者与参与者的比例比“36 氪”更小。两个网络估计出的网络直径类似，平均最短路径都在 6 左，符合六度分割理论。
2. 从“36 氪”和“京师心理大学堂”的粉丝分析我们可以发现，在知乎中，大部分用户都是来“围观”，来知乎“寻找答案”，只有少数用户（很可能是某个领域的专业人士）才会踊跃发言。那些少数踊跃发言的用户拥有大量的粉丝，而大多数的用户粉丝数并不多。
3. 36 氪网络中，最能区分用户影响力的因素是被收藏数，其次是被赞同数和被感谢数，说明一个有高影响力的用户往往会有高质量的回答和文章。而京师心理大学堂中，最能区分用户影响力的因素则是关注问题数，其次是被赞同数和被收藏数，这说明除了回答问题、发表文章以外，参与度对于影响力同样重要。
4. 通过比较两种边预测算法，我们发现以协同过滤的思想预测有向边更为合理，即评价 User 关注的 Leader 与他可能关注的 Leader 之间的相似度。这一算法的 AUC 可以达到 95%。
5. 我们利用基于 Louvain 的模块化聚类法挖掘社区，发现有两种类型的社区——拼凑型社区和紧密型社区。我们发现紧密型社区中的用户与机构号的主要业务更加接近，他们是机构号粉丝中的核心成员。

基于上述结论，我们为知乎网和知乎上的机构号提出以下合理建议：

1. 其实普通人之间的社交并不占主要地位，要多推荐优质内容给普通用户，并且希望普通用户成为内容生产者，这样才能使得整个知乎产生最大的价值。知乎用户并不关注与自身更相似的用户，而是关注与自己所关注的大 v 相似的用户，这一点可以用于我们的推荐中去。
2. 社区画像可以理解为业务层面的数据仓库，社区标签是多维分析的天然要素。通过社区画像，可以将用户群体切割成更细的粒度，为推荐系统、广告系统，精准营销提供基础。

7.2 展望与不足

我们希望去推测整个知乎网络长什么样子。如果以 36 氪社区为例，20 万结点之间产生 190 万条边，通过结点性质算出，他们与整个知乎网络产生了 7000 万条边，由比例推断，整个知乎有 700 万左右用户。但是我们查阅资料显示，用户群体远远超过这个数字。其实整个知乎网络远不如 36 氪社区这么密集，有很多很多游离在网络边缘的群体，他们只关注问题，而不关注优质内容的生产者。我们今后会对这些问题进行进一步研究。

囿于时间限制，我们在研究中只考虑了结点的相互关注关系与结点基本属性，而没有将用户与用户在问题、回答下的交互记录纳入考虑。而知乎作为一个问答型社区，用户与问题的交互也是一个很值得研究的方向。

我们只是对一小段时间进行数据采样，没有跟踪问题与用户随时间的变化。我们很感兴趣一个问题是如何变得热门的，我们也感兴趣一个人是如何成为大 v 的。这些东西我们会放在以后的研究中去做。

参考文献

1. Scrapy 模拟登录 2018 版知乎. https://www.cnblogs.com/zzzzzhangrui/p/8847724.html
2. 吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报, 2010, 39(5):651-661.
3. 宋学峰, 赵蔚, 高琳, et al. 社交问答网站知识共享的内容及社会网络分析——以知乎社区“在线教育”话题为例[J]. 现代教育技术, 2014, 24(6):70-77.
4. 李志宏, 吴煜山, 程裕. 基于从属关系的虚拟社区知识子群研究[J]. 科技管理研究, 2016, 36(5):104-110.
5. 知乎网用户知识共享研究[D]. 北京邮电大学, 2014.
6. Blondel V D , Guillaume J L , Lambiotte R , et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10):0-0.
7. Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

附录

附录 1：分工

张悦嘉：爬虫设计以及代码编写，数据库管理，社区挖掘及画像，上台汇报，报告撰写（任劳任怨的码农）
孟博宇：选题，文献搜集，爬虫算法设计，数据网络属性分析，边预测算法以及评价方法，上台汇报，报告撰写（天马行空的想法帝）
林雅文：数据描述性分析，数据挖掘，上台汇报，报告撰写（靠谱负责的数据分析师）

附录 2：社区挖掘与画像

36 氪社区画像（仅展示结点数大于 10 的社区）

类别编号	结点个数	主要可解释结点	标签
0	8834	['白诗诗', '叶倩倩', '尼克六六', 'Lachel', '梁悦', '童哲', '卫蓝', '左岸', 'nuttie', '欧阳畅', '杨大笨子', 'Chong', '万德尔 wonder', '杨又青', 'Scarborough']	['互联网', '公众', '白诗', '卫蓝', 'APP', '倩倩', '小猫', '方法论', '教育', '物理学']
1	4007	['山羊月', 'D.Han', '王郝', '卢诗翰', '刘不言', '李治林', '王藐', 'LLLBK', 'YoviaXU', '六号牙医', '阿尔托莉雅', 'Sir Pasco', 'Hermann', '江寒园', '莱特曼', '知己天涯 Samuel', '黄彬彬 Margaret', 'Mark', 'Paprika']	['博士', '在读', 'Assistant', 'Professor', '凝聚态', '心理学', '物理学', '临床', '公众', '硕士']
2	19063	['周源', '王诺诺', '周杰伦的小粉丝', '半十四', '赵世奇', '关雅荻', '田浩', '极乐', '郦橙锦妖', '王俊煜', '成远', '朱聿欣']	['互联网', '创始人', '知乎', '锦鲤', 'CEO', '联合', '电影', '化学', '001', '微信']
3	2599	['夏昊 BFA', '何瑶', '来须苍真', '余思', '王佳伦', '机智的 E 君', '奥斯卡', '窦月汐', '杨旭', 'Llia Yu 莉莉娅', '杜晓斑', '钱钱钱隆', '教日语的小慌先生', 'redhobor', '如云般飘过', 'Slade', '李威龙', 'Pseudoer', 'Sai WANG', '新店韩师傅']	['游戏', '电子游戏', '智族', 'GQ', '设计师', '创意', '总监', '日语', '电影', '日本语']
4	7552	['李靖', '黄有璨', '冯硕']	['互联网', '微信', '主页']
5	2935	['斌卡', 'Gabby 老师', '目目老师', '王彦翔', '奶黄包包包包', '胡海德']	['声乐', '健身', '互联网', '刷屏', '蝌普', '狂魔', '池人', '电音', '知乎', '萌物']
6	3391	['一苒', '张抗抗', '许良', 'Sean']	['汽车', '电动汽车', '3D', 'Hybrid', 'Powertrain', '工程系', '无人驾驶', '事业部', '汽车行业', '打印']
7	33676	['楠爷', '王瑞恩', '青锐吴斌', '空白白白白	['微信', '互联网', '数据分

		', '汪惟', '余青葭', '庄明浩（rosicky311）', '邹昕', '木棉 959', '沙丁鱼', '李石', 'Zhang Leslie', '翔宇情', '打不死的 little 强']	析', '创始人', '创业', '心理学', '并购', '投资', '创始', '合伙人']
8	9219	['车长']	['套路', '恋爱', '技巧', '互联网', '公众']
9	4544	['赖世雄', 'Emma', '陈芒果', '千里', '蔡韵 Iris', '七里', '豆浆 Boy 科技控', '唐文韬', '余知兮', '辩手李慕阳']	['创始人', '互联网', '大众传播', '芒果', '硕士', '法国', '赖世雄', '美少女', '常春藤']
10	5171	['万金油', '张大大', '牛姐', '猫叔聊地产', '社长有点野', '已注销', '张行', '水木知秋', '孙志超的新账号', '虞舜', '电商狗-老李', '许小喵', '陈凌轩', '勤奋的叶子君', '吴俊宇', '无可', 'Chongchong Zhang', '刘杨']	['公众', '电商', '互联网', '码字', '知乎', '微信', 'CEO', '区块', '咨询', '电子商务']
11	8684	['简浅', '吴清缘', '柳柳老师', '风墟', '阿正', '思桐', '靡靡之音', '逆鳞', '浅草先生', 'sMrZhao', '牛顿师兄', '旧文字', '观察君', '秋名山汽车驾驶员']	['互联网', '微信', '简族', '公众', '微博', 'ID', 'jianzu1126', '女流氓', '咨询', '私信']
12	1531	['Mr Diao', '诚言 SIR', '小宽', '武术研究所', '隔壁的六叔叔', '马克先生 Markk', '小西超人', '木南', 'H4 程靖', '大监制', 'Benny 是个文青']	['公众', '美食', '微信', '互联网', '形意', '空想家', '信号', 'JJJByes', '诚言', 'SIR']
13	17325	['老狼', '熊辰炎', '傅渥成', '王晋东不在家', '张明云']	['学习', 'Android', '博士后', '机器', '实习生', '迁移', '博士', '物理学', '互联网', '技术']
14	5505	['路过银河', '猫大叔', '刘大 1984', '纽约老闻', '何嘉文', '南区熊猫', '凌乐', '刘奶奶说牛奶奶的', '猎头 Will']	['TRUST', '财务', '持证人', '财报', '微信', '财会', '金融', '公众', '星空', '分析师']
15	1101	['木匠小强', '谢客官', '景晓萌', '鹏哥门窗科普', '一肥', '夭非妖育儿记', '假装在巴黎', 'kingconan', '木匠王双喜', '幻梦之始', '庞珞珞']	['amp', '家具', '宜家', '设计', '互联网', '产品', '公众', '私信', '主理', '良禽']
18	14464	['谢春霖', '利兄', '大梦 Power', '潇峰学长', '李云景', '看山不是山', '好姑娘老妖', '小木 Arvin', '讨喜 Hani', 'Pierre', 'Mia Li', 'ONE 字幕组', '运营喵芬妮', '水太深', '李老狼', '修身养性']	['公众', '互联网', 'PPT', '讨喜', 'Hani', '创始人', '富研社', '利兄', '治疗学', '观海']
24	206	['护法居士', '释兰迦叶', '李渊回', '水色天青叶子', 'tom', '能依', '乾知大始', '潇默城', '徐晓', '沐知', '一可轩', '青山如是', '陈方	['正法', '佛法', '如来', '互联网', '不堪其忧', '不改其乐', '受持', '速证', '寂

		岳']	空', '有涯']
25	814	['郭航初', '骆驼', 'bianlunnet', '张静远']	['CEO', '汉语言', '破过产', '过市', '辩论网', '官网', '官微', 'bianlunnet']
30	345	['游云', '土木林大锤', '圣权', 'Chenper', '子然', 'CHARVEN', '七叔成长工作室', '苏胖胖', 'Sheldon 的 Amy', 'biubiubiu', '一个没用的人', '马晨晨', '暮云山关', '悦苹', 'Llooooll', '孟尝梦不长']	['土木工程', '工程造价', '房地产', '项目管理', '慧眼识珠', '运营', '高工', '专业', '工程硕士']
37	2980	['Sean Ye', '无良 HR', '刀姐', '喧哗与低语', '田野', '章牧之', '王宇', '人力资源怪客', 'HeyJocelyn 酱', '陈焕', 'Steve SHEN', '三爷', 'europeduke', 'Waterwalker']	['人力资源', '互联网', '公众', '讲师', 'HR', 'UI', '总监', '职场', '招聘', '个人成长']
44	9762	['放心选·独孤评测', 'Edwin', '裘雪龙', '梨花', '翟羽', '超人测评', 'Junspr', '邹叔的任性', '康斯坦丁 FL', 'Nukeclear', 'KIKI', 'Martini', '唐一', '子由', '钢笔六六']	['互联网', '正畸', '种草', '付费', '评测', '美容', '牙齿', '广告学', '公众', '反派']
47	13738	['立党', '小约翰', '框框框子', '陈敏', '等风起的高达', 'AreYouKiddingMe', '兔撕鸡大老爷', '装睡的我', '可见迪', '妄明']	['毒言毒语', 'ID', 'cmdydy', '玻璃心', '公众', '半凉', '医学生', '朋克', '小约翰', '不杠']

京师心理大学堂社区画像（仅展示结点数大于 10 的社区）

类别编号	结点个数	主要可解释结点	标签
5	19015	['暗涌', '王瑞恩', '胡远东', 'Steve Shi', '覃宇辉', '小侯飞氘']	['心理学', '硕士', '心理咨询', 'Transdimensional', '宾夕法尼亚大学', '社会工作', '教育学', '知友', '法学专业', '氢气球']
6	68531	['动机在杭州', '山羊月', '赵思家', '万金油', '叶倩倩', '闻佳']	['神经科学', '博士', '不以己', '倩倩', '心理学', '小猫', '社会学', '脑科学', '知友', '神经学']
15	66259	['刘看山']	['知乎', '吉祥物', '北极狐', '全职', '市场推广', '有趣', '发现']
22	611	['曾旻', '雷云逸', '苏格拉没有底', '独行侠']	['心理学', '互联网', '应用', '慎独', '心理咨询', '在读', '硕士', '本科', '课程', '重建']
25	69943	['木棉 959', '叶壮', '在焉', '朵拉陈', '俞林鑫', '燕仰', '东华君', '宏桑', '简里里', '窦泽	['心理学', '心理咨询', '博士后', '微信', '硕士', '心

		南', '高地清风', '刘柯', '冯慎行']	理', '培训师', '微博', 'Asian', '教育']
27	274	['Stephanie', '珍宝碎片化', 'Todd', 'CaptainToy', '王爷不偏安', '小虫虫', '牧云', '婉兮清扬 er', '烟清清', '宁柠柠', '韩儒', '江城子', '狮子座黄金圣斗士', '布多', '丁一叉', '两斤香菇', '一只机智的羊', '樱桃 CC']	['健身', '创意', '视觉', '幼师', '插画', '电子厂', '大膜', '绘本', '柴犬手', '微博']