在现实生活中,人与人之间的关系和构成了社交网络。社交网络强调人际关系、关系内涵以及社交网络结构对社会现象的诠释。社交网络与人们的生活息息相关,其中随着移动通信技术的不断发展和移动义务的不断创新,这帮助了人们实现跨越时空、便捷沟通的梦想,实现了人们在任何时问、任何地点与任何人进行沟通与交流。利用手机通信获得大型社交网络数据是一种可行、经济、有效的手段。

随着移动网络的不断发展,用户对网络的质量也有越来越高的要求,中国的三大通信公司移动、联通、电信之间的竞争也越来越激烈。如何满足用户要求以提升用户感受成为运营商的一个重要课题。用户的通话数据具备了这方面得天独厚的优势,它真实反映了用户实际通话状况,用户通话数据也越来越受到各大运营商的重视,其在网络运营商中的用途也越来越广泛。

以移动通话社交网络作为复杂网络的一个实例,使用复杂网络理论方法对移动通话网络特性进行分析。包括对其个体层面的度、聚类系数等到整体层面的度分布、整体聚类系数的刻画,同时也描述了属于这之间的社团结构。社团结构具有属于同一社团的内部成员间联系紧密,不同社团间的成员连接松散的特点。其中设计出了一种基于各类图结构划分算法的方法,通过该算法对大规模的复杂通话网络进行实证研究。通话社交网络能够描述社交网络中人与人之间的信息交流及传播行为,研究其演化机制以及网络的拓扑性质,可从本质上揭示通话网络的发展与演化规律。通过对通话社交网络中的用户交互行为以及其对任意个体、整个网络的影响机制的研究,丰富并拓展了传统电话诈骗研究的缺陷. 这项研究对于移动运营商而言,具有高度战略参考性,比如:找出重点客户推送制订业务,预测与防止客户流失,客户群划分,找出诈骗或骚扰电话.等等.

从20世纪90年代以来,经过互联网的高速发展,人们已经生活在一个充满各种各样复杂网络的世界之中,且功能各不相同,但各种不同的复杂网络间也存在着许多的相同之处。其中相同的基本特征如下:

大规模的网络中节点数可达百万乃至千万,甚至更大规模.

存在于网络中的任一节点自身表现为非线性系统,且行为复杂性. 网络连接结构复杂.在现实情况中,整个网络的连接结构既不是完全属于理想情况的规则图也并非属于完全随机情况的网络,但却表现出了网络本身自组织的规律;

大型网络连接是非常稀疏的,在实际情况中,存在于大型复杂网络中的边数通常为O(N);

随着人们对复杂网络的深入研究,出现了一系列与小世界性、无标度性等基本统计特性并列的新性质,其中之一就是网络社团结构,它是复杂网络最普遍和最重要的拓扑结构属性之一。复杂网络研究进展集中于网络社团结构的研究。学者们试图揭示错综复杂的网络如何由相对独立而又相互交叉的社团组成。大量的实践研究表明,许多网络是异构网络,可将其分解为若干个社团,其中各个社团内部节点间连接较为紧密,而社团之间连接较为稀疏。对大规模复杂网络节点进行社团划分有助于更好地理解复杂网络不同的层次结构和其相应的功能特性,使其朝着对人类有益的方向发展。

借助复杂网络理论分析方法,利用通话数据网络对人与人之间的社会交往进行研究。目前,从复杂网络、社交网络角度进行的通话社交网络研究可总结为以下三类:

一、通过实证方法度量网络的统计特性。

二、对整体网络演化模型的深入研究。

三、研究用户行为之间所存在的规律,同时对网络特性以及用户行为进行讨论。

对以上问题的研究有助于对运营商对网络进行局部分析包括:VIP客户保障、客服支撑、网络优化、全网分析和评估以及解决用户流失的问题。

社交关系网络术语:

(一)规则网络

在一个全耦合网络中,任意两节点直接相连。在具有同等规模的网络中,全局耦合网络的平均路径长度最小:L=1。同时,具有最大的聚集系数:C=1。

然而,实际大规模网络大都都是很稀疏的,因此全局耦合网络模型作为实际网络的模型具有明显的局限性。

(二)随机网络

在复杂网络中,随机图与规则网络完全相反。其中,一个具有代表性典型随机图模型是ER模型。目前,随机网络已经成为数学中一个重要分支,在社会、交通、通信、计算科学技术等实际工程领域中都起着举足轻重的作用。

现实复杂网络一般都具有非常明显的聚类特性,而ER随机图的两节点间连接概率均为P,因此其聚集系数为C=p=<k>/N <<1。这就意味着,ER随机图几乎不具备聚类特性。所以,真实复杂网络要比相同规模的ER随机图的聚类系数高得多。

ER随机图不能很全面的刻画实际复杂网络的某些特性,但它描述了网络中的一些共性,因此在复杂网络模型及基本理论研究中依旧占据重要的地位。

(三)小世界网络模型

它是在一维规则的点阵中引入一定的随机性网络结构,毕竟完全规则网络和完全随机网络都无法准确描述现实生活中的网络结构状况。小世界网络模型的提出完善了网络模型刻画的理论基础。图中表示了通过随机化重连来实现从规则网络演化为随机网络的过程。

实践方法

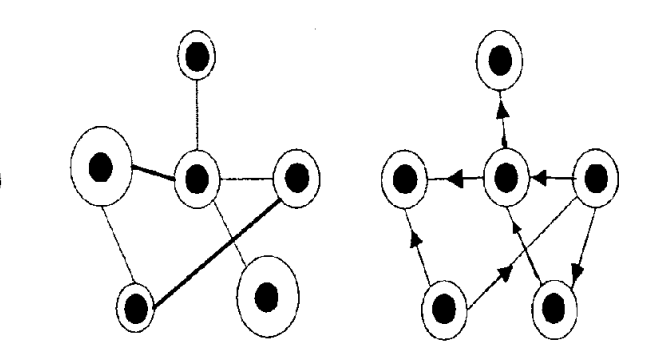
主要利用复杂网络方法对通话记录构建的通话社交网络进行实证研究。通过某运营商的某时间段的通话记录数据构建通话社交网络模型。该模型以单个通话用户作为网络中的独立节点,用户之间是否存在通话关系确定节点之间有无边相连。对通话社交网络的研究和分析有利于运营商的业务更好地扩展。随着通信技术的不断发展,从海量数据中挖掘社交网络中人类行为特征成为复杂网络的研究热点和重点。对通话社交网络的用户特征及整个网络的拓扑结构进行研究意义深远。

通过通话数据构建通话社交网络模型,包括对数据的预处理以及初步统计。分别统计了通话社交网络的度分布、平均路径、三角系数、连通图、聚集系数、同类性、PageRank系数、节点中心性等特征量。通过使用社团挖掘算法的对比,选取适合通话社交网络的算法进行社团分析。

复杂网络(Complex Network)是指具有复杂拓扑结构和复杂节点行为的网络系统,它是对现实世界中各种各样的大规模复杂系统的抽象。我国科学家钱学森对于复杂网络提出了一个相对全面的定义:在现实世界中,表现出“自组织、自相似、吸引子、小世界、无标度”的部分或者全部性质的网络称之为复杂网络。在复杂网络中,节点可表示为具备特定信息内涵的基本单位,边则是描述节点(基本单位)间的相互之间的关系或联系。鉴于图论以及统计物理等方法,进一步探索各种结构、功能等性质看似互不相同的网络之间的共性科学问题和处理它们的有效方法。

人们在刻画复杂网络结构的统计特性方面提出了许多概念以及方法,在自然科学领域,复杂网络研究的基本测度包括:度及度分布(Degree Distribution)特性、平均路径长度(Average Path Length)、聚类系数(Clustering Coefficient)及其分布特征、介数(Betwenness Centrality)及其分布特性、连通社群(Connected Components)的规模分布等。

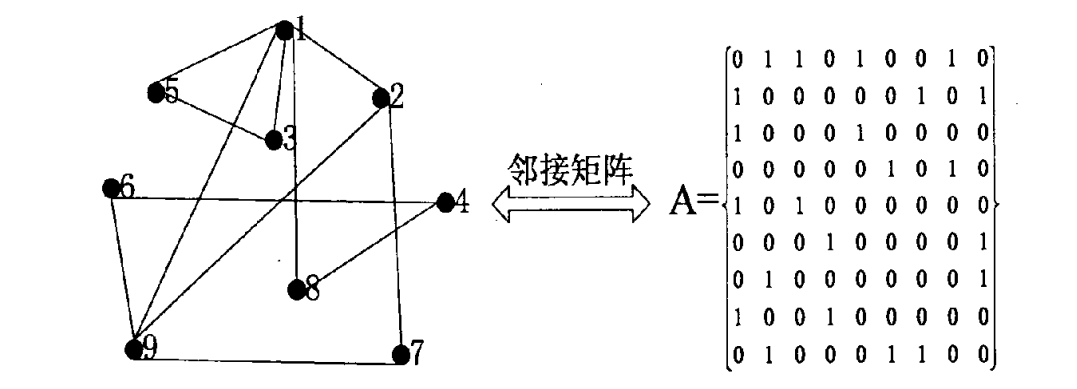
网络的图表示:G=(V,E)。



另外,复杂网络图结构中的信息通常可以通过邻接矩阵来存储。在这个矩阵中,如果

两个节点之间有变相连,则这两个节点对应的矩阵元素为1;如果两个节点之间没有边连

接,则相应位置上元素为0。



Spark GraphX

GraphX 是Apache Spark用于图和图并行计算的API。它扩展了Spark RDD，引入了一个新的图抽象： 有向多图（directed multigraph） ，每个节点和边都有自己的属性。

GraphX库提供了图算子（operator）来转换图数据，如subgraph、joinVertices和aggregateMessages。它提供了几种方法来从RDD或硬盘上的一堆节点和边中来构建一个图。它也提供了许多图算法和构造方法来进行图分析。我们将在后面讨论图算法。

图1展示了Apache Spark生态系统以及GraphX与其他库在整个框架中的关系。

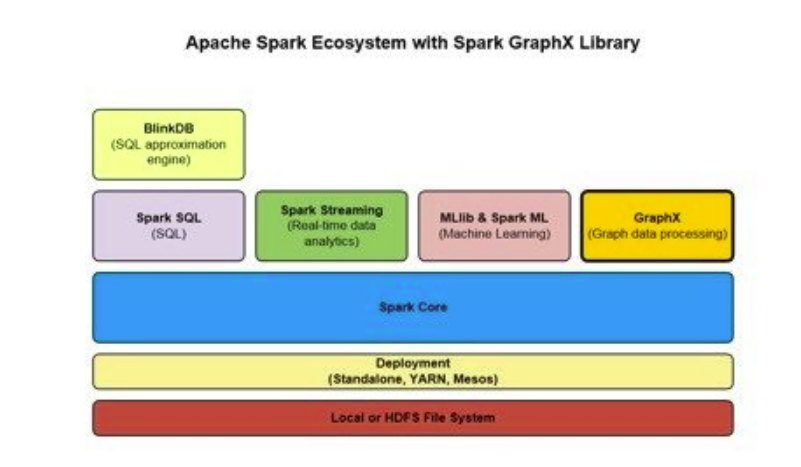
 通过内嵌的算子和算法，GraphX使得在图数据上运行分析变得更加容易。它还允许用户cache和uncache图数据，以在多次调用图的时候避免出现重复计算。

表1中列出了GraphX中的一些图算子。



GraphFrames

GraphFrames 是Spark图数据处理工具集的一个新工具，它将模式匹配和图算法等特征与Spark SQL整合在一起。节点和边被表示为DataFrames，而不是RDD对象。

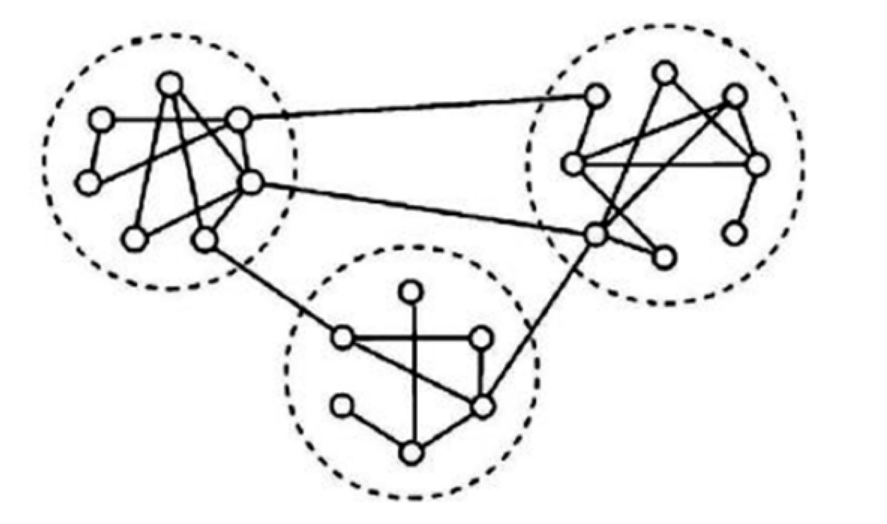
GraphFrames简化了图数据分析管道，优化了对图数据和关系数据的查询。与基于RDD的图处理相比，GraphFrames有下列优势：

* 在Scala API之外，还支持Python和Java。我们现在可以在这三门语言中使GraphX算法。
* 用Spark SQL和DataFrames获得更高级的查询能力。Graph-aware query planner使用物化视图来提高查询性能。我们也可以用Parquet、JSON和CSV等格式来存储和导入图。

#### 社区发现(CommunityDiscovery)

社交网络的核心是参与其中的用户以及用户之间的关系，“物以类聚，人以群分”，对一个大型网络调用社区发现算法，其实是对其按照某种标准进行了划分，在此基础上可对每个社区做进一步的发掘。而从计算的角度而言，社区分划相当于分解了任务，起到了降低计算复杂度的作用。因此，我们可以采用图模型来为其进行建模，其中的节点表示社交网络中一个个的用户，而边则表示用户与用户之间的关系，如果想对这些关系强度（或亲密度）进行区分的话，我们还可以为每条边赋予一个权重，权值越大表示关系强度越大（或者越亲密）。

如下图所示的社交网络，可将其分为三个社区，社区间关联关系弱，社区内关联关系强。



那些内部连接比较紧密的节点子集合对应的子图叫做社区（community），各社区节点集合彼此没有交集的称为非重叠型（disjoint）社区，有交集的称为重叠型（overlapping）社区。网络图中包含一个个社区的现象称为社区结构，社区结构是网络中的一个普遍特征。给定一个网络图，找出其社区结构的过程叫做社区发现（community detection）。

Community Detection算法很多、如：k-clique、label propagation 、fast unfolding、louvain等。

图分析算法介绍

PageRank

PageRank算法被用于确定图数据集中的一个对象的相关重要程度。它衡量图中每个节点的重要性，假设从其他节点到该节点的边代表着认可（endorsement）,那么边越多则认可度越高。

PageRank的一个经典例子就是Google的搜索引擎。基于有多少网页引用某个网页，Google使用PageRank来计算该网页的重要程度。

另一个例子是社交网络网站，如Twitter。如果一个Twitter用户被许多其他用户关注，那么该用户在网络中有较高的影响力。这种度量指标可被用于对关注者进行广告投放（100,000个用户关注一个厨师=> 很可能是食物爱好者）。

GraphX提供了两种Pageank的实现方法：静态的和动态的。

(1)静态PageRank：该算法迭代运行固定的次数，对图数据中的某给定节点集生成PageRank值。

(2)动态PageRank：该算法运行直至PageRank值收敛到一个预定义的误差容忍值。

Connected Components

图中的一个Connected Component就是一个连接的子图，其中，两个节点由边互相连接，并且子图中没有其他节点。也就是说，当两个节点之间存在关系时，这两个节点属于同一个Connected Component。子图中具有最低数值节点的ID被用于标记Connected Component。在社交网络例子中，可用Connected Component来创建图中的类（cluster）。

计算connected components时，有两种图遍历方法：

图数据处理中还有另外一个算法叫做 Strongly Connected Components (SCC)。如果图中每个节点都可到达所有的节点，那么这个图是强连接的。在有向图*G*中，如果两个顶点*vi,vj*间（*vi>vj*）有一条从*vi*到*vj*的有向路径，同时还有一条从*vj*到*vi*的有向路径，则称两个顶点强连通*(strongly connected)*。

Triangle Counting

Triangle counting是一种社区分析算法，它被用于确定经过图中每个节点的三角形的数量。如果一个节点有两个相邻节点而且这两个相邻节点之间有一条边，那么该节点是三角形的一部分。三角形是一个三节点的子图，其中每两个节点是相连的。Triangle counting算法返回一个图对象，我们可以从它上面提取节点。

Triangle counting被大量地用于社交网络分析中。它提供了衡量图数据聚类分析的方法，这对在社交网站（如LinkedIn或Facebook）中寻找社区和度量区域群落的粘度很有用。

Triangle Counting算法的其他用户案例有垃圾邮件检测和连接推荐。

与其他图算法相比，Triangle counting涉及大量的信息和复杂耗时的计算。因此，当你测试该算法时，确保在性能较好的机器上运行Spark程序。需要注意的是，PageRank衡量相关度，而Triangle counting衡量聚类结果。

Label propagation标签传播算法（LPA）

运行静态标签传播算法，用于检测网络中的社区。

算法基本理论：

根据LPA算法基本理论，每个节点的标签按相似度传播给相邻节点，在节点传播的每一步，每个节点根据相邻节点的标签来更新自己的标签，与该节点相似度越大，其相邻节点对其标注的影响权值越大，相似节点的标签越趋于一致，其标签就越容易传播。在标签传播过程中，保持已标注数据的标签不变，使其像一个源头把标签传向未标注数据。

最终，当迭代过程结束时，相似节点的概率分布也趋于相似，可以划分到同一个类别中，从而完成标签传播过程。

在LPA算法中节点的Label有同步更新与异步更新2种更新方法。同步更新方法在二分图中可能出现产生震荡情况。为了避免循环和保证收敛，LPA算法采用异步的策略更新节点的标签，并在每次迭代前对节点重新进行随机排序。

LPA算法的思路：

首先每个节点有一个自己特有的标签，节点会选择自己邻居中出现次数最多的标签，如果每个标签出现次数一样多，那么就随机选择一个标签替换自己原始的标签，如此往复，直到每个节点标签不再发生变化，那么持有相同标签的节点就归为一个社区。

算法优点：简单、高效、快速，适合大型复杂网络。

算法缺点：众所周知，划分结果不稳定，准确率不高，随机性强是这个算法致命的缺点。

标签传播算法，是所有基于标签的算法的基础，它是一种基于图的半监督学习方法。其基本思路是用已标记节点的标签信息去预测未标记节点的标签信息。

利用样本间的关系建立关系完全图模型，在完全图中，节点包括已标注和未标注数据，其边表示两个节点的相似度，节点的标签按相似度传递给其他节点。

标签数据就像是一个源头，可以对无标签数据进行标注，节点的相似度越大，标签越容易传播。

由于该算法简单易实现，算法执行时间短，复杂度低且分类效果好，其广泛地应用到多媒体信息分类、虚拟社区挖掘等领域中。

在电话诈骗业务中的算法应用：

* 在通话社交网络中，计算不同用户的社交频度如何（PageRank）
* 基于网络中的用户连接来对用户分群（Connected Components）
* 通话社区发现和对通话社交网络中的用户社区的粘度分析（Triangle Counting）
* Connected components
* Label propagation
* SVD++
* Strongly connected components
* Triangle count
* Single-Source-Shortest-Paths
* Community Detection

最短路径（Single-Source-Shortest-Paths） ：

最短距离和道路 在社交网络应用中也很有用。寻找从每一个结点到一个给定的终点t结点的最短路径。它们可被用于衡量网络中一个特定用户的相关度。最短路径越小，用户越相关。

平均路径长度

在网络拓扑结构中,网络的平均路径长度是网络的传输性能和效率的衡量标准量,是网络的全局特征,同时也是网络拓扑性质的重要参数之一。它反映了网络呈现的总体特征,研究最短路径在现实复杂网络中具有极其重要的意义。

在网络中,任意选取两个节点i和j,连接这两个节点的最短路径的边数定义为这两个节点间的距离d\_ij,网络的直径D定义为任意两节点距离的最大值。

网络的平均路径长度L也称作特征路径长度(Characteristic Path Length), 指网络中所有节点对的路径长度的平均值。

度和度分布

在复杂网络中,描述某一网络图的最基本属性是度(degree),同时也是网络节点属性中非常重要的概念。在图论中,某一节点的度是网络中的某一节点和节点之间关系(网络中的边)的数量表达。

对于有向图而言,节点的度分为入度(in.degree)和出度(out.degree)两种。其中,节点的入度通常指以某点作为有向图中的顶点,指向于该顶点的边数之和;节点的出度则指由有向图中某一顶点指出的边的数目。通过度可以很好的刻画网络中的全部节点的连接程度。节点度可以表示网络中个体的影响力和重要程度。一个节点的度越大则表示该节点从某种意义上来说其在网络中的影响力越大,地位亦“越重要”。

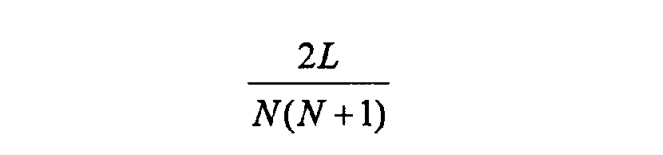
研究表明真实网络的度分布往往并非Poisson分布,而是遵循某种幂律形式,这种分布概率随着k值增大的衰减更为缓慢,即k值大的节点相比之下显著增多。将具备这样性质的网络称作“无标度网络”,如生物新陈代谢网络、科研合作网、国际互联网等,其度分布服从无标度性的幂律(power-law)分布,在双坐标下,这种分布表现为一条直线。

度在现实生活的不同领域中所表现的意义也各不相同。例如在通话网络中,度值越大的节点则表明对应的用户通话频度越高,在网络中越活跃,也就是说在这个人际关系网络中越重要,其对于移动通信运营商的贡献也越大;

密度

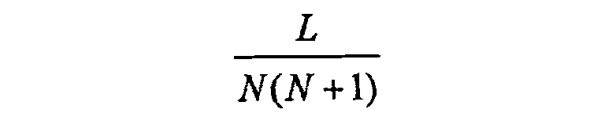
密度描述了一个图中各个节点间的关联程度。一个“完全图”是指一个所有节点间都相互邻接的图。即使在小网络中,这种完备性也是极其少见的。密度这个特征量试图对连线的总分布进行整体汇总,以测量图在多大程度上具备这种完整性。从图论的角度而言,密度指一个图凝聚力的总体水平,反映了图的“紧凑性(compactness)”。它依赖于网络结构中的两个参数:图中各个节点度数之和以及图的内含度(inclusiveness)。其中,图的内含度是指图中各类关联部分所包含的总的节点数,也就是图的总节点数减去孤立节点的个数。

图的密度定义为图中实际的连线数与最多可能拥有的连线数的比值,即



其中,L为图中实际连线数;N表示图中节点总数。N(N+1):为最多可能拥有的连线数。

在有向图中,密度的表达式则为:

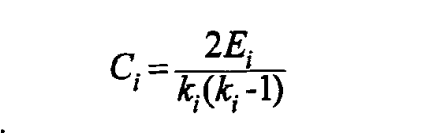


聚类系数

在现实生活的朋友圈子中,你的两个朋友彼此可相互认识是彼此的好朋友,这表明了两个人间朋友圈子的重合度。社交网络不同程度体现了“物以类聚,人以群分”的特性。

通过研究表明实际的复杂网络也并不是完全随机的,其中大部分网络在某种程度上也普遍存在类似于社交网络结构中的聚类特性。它描述了网络中节点聚集成团的程度,同时也是衡量网络结构的重要特征之一。

聚类系数衡量了连接在一起的集团与各自近邻中的共同近邻的数目,它是网络的静态特征。假设在一个网络图中,某节点i与其他节点之间存在k\_i条边。在理论上,所连接的k\_i条边的节点间最多可存在k(k\_i – 1)/2条边,则该节点的聚类系数q用实际存在的边数与最多可能存在边数之比来定义。其中,E\_i为实际存在的边数。



C\_i = 某节点实际存在的三角形个数 / 所有邻居节点两两组合的个数

很明显0<C<1。当网络中的全部节点都为孤立,节点之间相互无任何连接时,c=0;当网络为全局耦合时,也就是说网络中的任何两个节点均相互连接时,C=1。大规模复杂网络都具有聚类效应,虽然他们的聚类系数远小于l,但比O(N-1)大得多。

介数

在复杂网络中,还有一个与最短路径较近似的另一个衡量网络拓扑特性的重要指标即介数。介数通常有节点介数和边介数之分。节点介数是通过网络中经过该节点的最短路径的数目占最短路径总数目的比例来描述;边介数则指网络中通过该边的所有最短路径的数目与最短路径总数目之比。

介数反映了网络中相应节点或边在整个网络中的作用及影响力,同时也是一个很重要的全局几何量,具有很强的现实意义。例如,在社交网络中,节点介数或者边介数的分布特征反映了不同人在社交网络中的不同角色和地位,有助于研究人员发现关键节点,对制定相应的用户策略具有极其重要的意义。

[GN算法](http://arxiv.org/pdf/cond-mat/0112110.pdf)(社区发现):进行求解，首先求解每条边的介数（betweenness），然后将介数最大的边删去，再重新求解每条边新的介数，依此循环。对应图1，连接不同社区的边的介数最大，把它们删去后即可得若干个独立的社区。但是求解介数时间复杂度高，在大图上并不实用，这时候需要考虑对[图的抽样](https://cs.brown.edu/people/matteo/papers/RiondatoKornarop-BetweennessSampling.pdf)等问题。

区域社交

比如小区的小卖铺,菜鸟网络站点,

连通性

连通子图描述的是网络中的任意两个节点都相连的子图,其中子图中的节点数即该连通子图的规模,而该网络图的连通性则由整个网络图中连通子图的数量来表现。

同类性

网络的同类性反映了网络节点之间相互联系所表现出的某种趋势。同类性用”同配性系数”表示,网络的同类性又称为网络的异配、同配性问题,即网络的度相关性问题,往往由同配性系数(assortativity coefficient)来描述,记为r。当r>0时,网络中的节点倾向于同最类似与自身性质的节点相连;当r>0时,网络中的节点则倾向于同最不类似与自身性质的节点相连。

同配性系数(assortativity coefficient) :

同配系数定义为度的Person系数,是从节点度测度扩展而来的也称作度相关性（degree-degree Correlation）。

此系数刻画了统计意义中网络的度数较高的节点是倾向于与其他度数较高的节点相连还是倾向于与其他度数较低的节点相连的特征，实际上这两种情况都可能存在。

通过一个系数r来刻画网络的相配混合性。如果这个系数为正，则网络是同配的（即度值大的节点更倾向于连接度值大的节点）；为负，则网络为异配的（即 度值大的节点更倾向于连接度值小的节点）。很有趣的是几乎所有的社会网络都是同配的，而技术和生物网络～般为异配的。

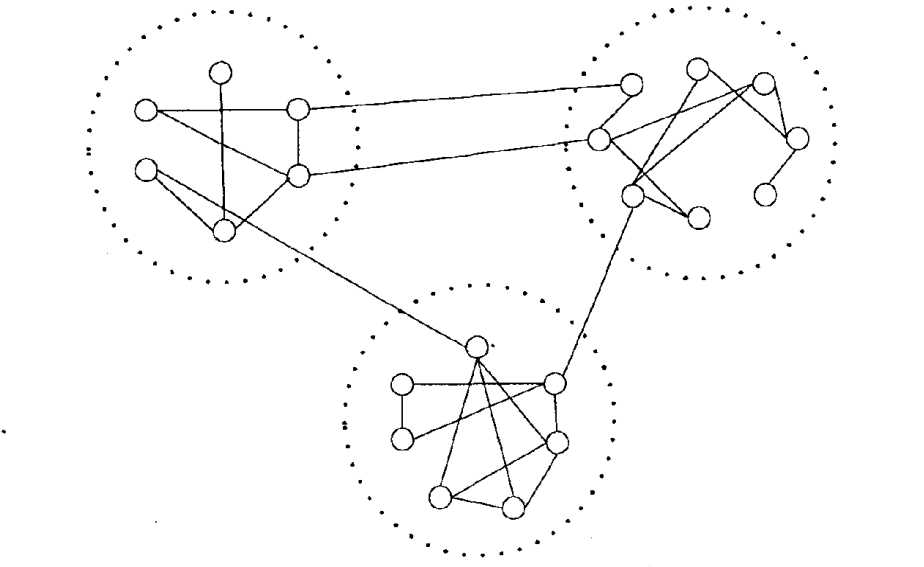
通过以上讨论，我们不难看出，真实网络一般都具有小世界特性、无标度特性、 高聚类系数和正负同配性系数。实际网络具有极高的复杂性，用非常简单的模型很难模拟所有的真实网络，所以需要通过不断的深入研究，构建合理的模型来反映实际网络的运行机制。

深入通话网络进行探究,针对90%非复杂小世界网络,如何区分网络是良性还是恶性的,这里通过大量分析发现,小世界网络中,均匀的强关联网络中的节点的出度与三角计数之间呈现正相关性, 而存在骚扰用户的网络中则不体现出节点的出度与三角计数有线性相关的关系.鉴于此结论,则可以通过计算小世界网络中成员的特征相关性来判断网络的好坏,即无相关性的网络则为骚扰用户藏匿的网络.

技术

在图分析代码示例中，我们将会用到下列技术：

|  |  |
| --- | --- |
| 技术 | 版本 |
| Apache Spark | 2.1.1 |
| Scala | 2.11 |
| JDK | 1.8 |
| Maven | 3.3 |



上图为社团划分做说明, 微信,人人网可以做社区划分,,而诈骗电话不行,,如上关系可以是一个菜鸟驿站,一个小卖铺,也可能是一个诈骗人在一个社区里

**Louvain**

Louvain算法其实是基于层次聚类的，图聚类。层次聚类依据是类间距更大，Louvain算法图聚类依据是模块度（公式在下方）。

Louvain算法包括两个阶段，迭代这两个阶段。

***1***不断地遍历网络中的结点，假设每个节点为1个社区，N个节点，初始化N个社区，尝试将单个结点加入能够使模块度提升最大的社区中，直到所有结点都不再变化。

模块度提升的定义如下：



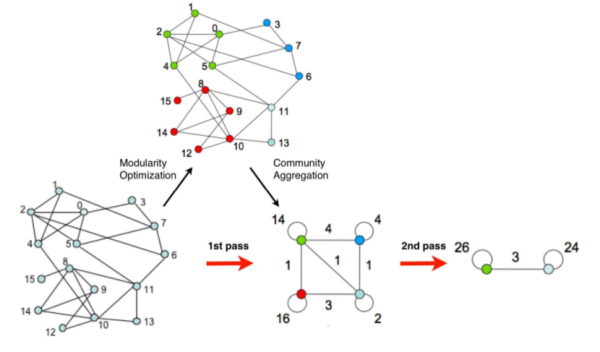
前项代表的意思为结点加入邻居社区后，社区内内所有边

后项代表着结点加入邻居社区后，社区所有边（包括内部边，以及连接外部的边）

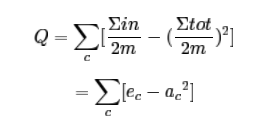
**目的是变化后希望子社区内边，外边少，即更聚集。**

**但是为什么后项要所有边呢？因为后项带了惩罚项的意思，没有惩罚项的话把所有都分成一个社区都分成一起就最大了。**

***2***它处理第一阶段的结果，将一个个小的社区归并为一个超结点来重新构造网络。它的执行流程如图所示：



迭代这两个步骤直至算法稳定（所有子社区模块度相加值不变）。



前项代表的意思为子社区内所有边，后项代表着子社区社区所有边（包括内部边，以及连接外部的边）。

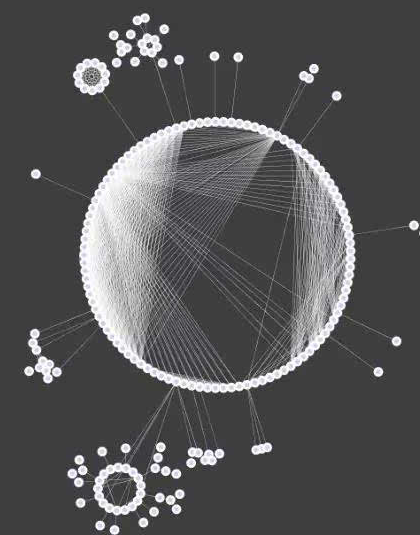
**模块度就是希望子社区内边，外边少，即更聚集。**

**但是为什么后项要所有边呢？因为后项带了惩罚项的意思，没有惩罚项的话把所有都分成一个社区都分成一起就最大了。**

## [**关系数据挖掘系列（1）公共安全行业团伙分析**](https://zhuanlan.zhihu.com/p/29411272)

作者：Zason.Zhang  
链接：https://zhuanlan.zhihu.com/p/29411272  
来源：知乎

本篇的关系指共同犯罪的关系，公共安全行业，会希望通过数据，对所有的打防控名单分成团伙。



分成团伙后可以知识发现，即做这件事的目的：

**1.寻找打防控团伙人员的关键人员**

**2.寻找与打防控团伙人员中有关联的正常人员**

因此，需求产生，开始进行数据挖掘：

*1.根据所有打防控人员的同案件关系数据形成图*

连接边为同案件的关系网络

*2.采用图聚类算法进行子社区识别出团伙*

可以选择两种方法聚类

第一种、

[图计算系列（1）图聚类Louvain算法](https://zhuanlan.zhihu.com/p/29380602)

louvain方法为层次聚类，可能聚集出链式团伙

第二种、

[图计算系列（2）图特征提取node2vec算法](https://zhuanlan.zhihu.com/p/29381100)

[机器学习系列（9）根据业务目的选择聚类算法](https://zhuanlan.zhihu.com/p/28811021)

将人在空间中转化成特征向量，然后再用划分聚类法进行聚类，聚集出簇行聚类

*3.通过度数计算确定每个团伙中心人员，计算连接不同社区的人员作为桥接人员*

*4.建立常驻人口的其他网络*

以同样的方式，建立其他网络并进行图聚类，例如根据住宿数据，建立同房间住宿的关系网络

*5.识别同住宿网络与团伙网络中的重叠，寻找与打防控人员中有关联的正常人员*

最终达到以下目的

**1.寻找打防控人员的关键人员**

**2.寻找与打防控人员中有关联的正常人员**

***不仅在公安领域，风控领域也适用于此方法***

***1.寻找欺诈人员的关键人员***

***2.寻找与欺诈人员中有关联的正常人员***

## [**实体数据挖掘系列（1）电信行业客户流失预测**](https://zhuanlan.zhihu.com/p/29400658)

作者：Zason.Zhang  
链接：https://zhuanlan.zhihu.com/p/29400658  
来源：知乎

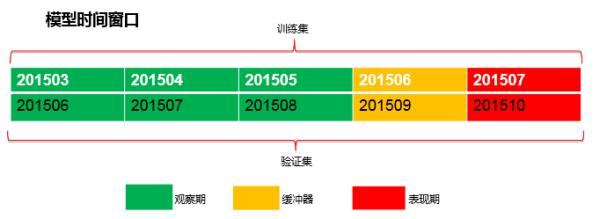
本篇的实体指客户，数据挖掘的文章会贴近应用，其中会涉及到一系列机器学习的算法，数据挖掘系列文章会从应用角度讲述如何使用机器学习算法实现业务目的。

某运营商，每个月会流失很大的用户量，但是由于用户量巨大，人力有限，不可能对所有用户进行策略性照顾。

如果有提前的一个月内，工作人员只能对10万个用户打电话挽留，就需要提前找出流失可能性top10万的用户，进行维系挽留。

ok，需求明确，现在开始通过行为属性与自然属性去挖掘用户！

**1.确立时间窗与用户**



用前三个月的行为，判断第五个月的行为，为什么这样呢？因为判断出第五个月这个人才会流失，那么刚好可以中间隔的一个月去打电话。

去除几乎没有行为的极低三无用户

**2.变量与自变量确定**

**变量的确定这方面我觉得经验丰富的重要性，这块如果是有经验踩过坑的人会省了很多时间，因为我在做其他领域项目的时候，没有任何领域经验的时候，觉得最耗时间的地方就是这个地方。**

自变量：表现期是否流失

变量：*基本信息*（基本信息在任何领域都非常有用，这里包括用户等级，是否打印发票，融合业务等等重要因素），*终端情况，账单，缴费，流量，短信，语音*

后4项为行为，会做出很多以三个月为统计的行为变量（基本上通过时间细分，和行为事件细分，例如三个月内上午通话时长，三个月内长途通话时长）

还可以以时间和事件为基准，制作衍生变量，均值（例，三个月通话均值），波动（例，第三个月通话相比三个月均值的变化），比例（三个月长途通话均值/三个月通话均值）

**3.变量选择**

根据文章

[机器学习系列（8）特征学习之有监督变量选择三步法](https://zhuanlan.zhihu.com/p/28796052)

进行选择

**4.分类算法**

根据

[机器学习系列（2）逻辑回归分类傻瓜式剖析](https://zhuanlan.zhihu.com/p/28641318)

进行建立模型

**5.模型测试评估**

根据

[机器学习系列（4）【面向业务】分类算法最普遍的两种评价方法](https://zhuanlan.zhihu.com/p/28705928)

第一种评价方法，看前20%的提升度

**6.模型sql固化系统中**

逻辑回归很简单，就两步完全可以用sql写出，因此可以用sql固化

每次业务人员跑出，最终的top10万人即可。

总结一下，自我感觉第二步的**变量与自变量确定**，如果是有经验的人会少浪费很多时间，在招聘行业经验人员的时候，这个地方会起到关键价值。

感性的说，客户流失就是客户活跃度忠诚度价值减少

最终实现模型的变量代表着客户

活跃度忠诚度价值（语音波动降低很多的会容易流失，没有融合业务等），这点很低的更容易流失。