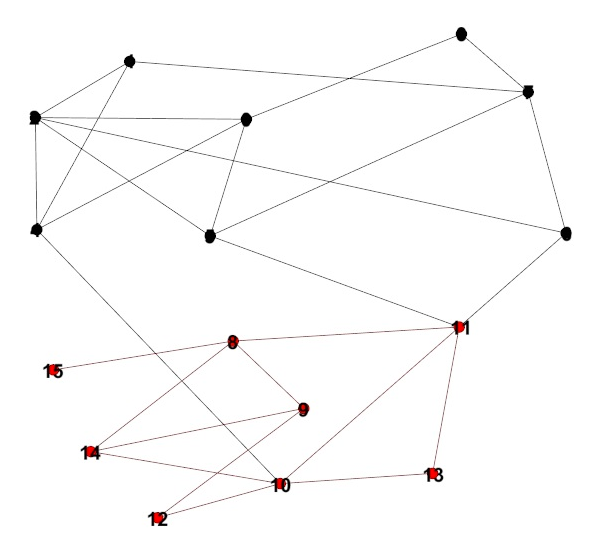
社区发现算法调研报告

调研人:李亚彬

工号:x8149

**一 社区发现（Community detection）算法定义**

在社交及其他群体性活动中，用户相当于一个节点，节点通过彼此互动行为构成了一个网络的结构，在这样的网络中，有的用户之间的连接较为紧密，有的连接关系较为稀疏。连接较为紧密的部分可以看成一个社区，其内部的节点之间有较为紧密的连接，而在两个社区间则相对连接较为稀疏，这便称为**社团结构。**



上图用红色点和黑色点对其进行标注，整个网络被划分成了两个部分，其中，这两个部分的内部连接较为紧密，而这两个社区之间的连接则较为稀疏。

**社区发现与聚类**

社区与聚类是两个很容易混淆的概念，让我们花费大量篇幅解释这二者的异同点。

通过社区发现算法定义可知，社团检测侧重于找到网络中联系紧密的部分，而经常忽略节点本身的属性（attributes）。

聚类，顾名思义是将属于同一类的目标聚在一起，通常在聚类之前我们是不知道目标有哪些类型，这也是一种典型的无监督学习方法。聚类算法最核心的部分是计算两个目标之间的**距离**（或者称为相似度），距离近则它俩是一类，距离远，那就自成一派，或者去找其它距离近的。距离计算方式有欧式距离，曼哈顿距离，余弦相似度等，都是直接用目标特征构成的向量来计算的，所以，聚类侧重于找到一堆属性相似的目标，从而忽略了目标与目标之间的联系。

聚类算法也可以应用于复杂网络，但是无法保证每个集群都具有良好的社区结构。

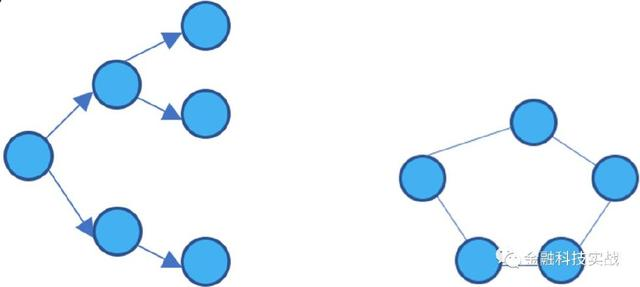
社团检测和聚类存在区别，但是两者又是可以结合起来的。比如，我们现在有一个网络，只知道顶点和边的情况，顶点的属性是未知的。那么在做社团检测的时候，可以将顶点与顶点之间的关系构成一个邻接矩阵，通过一系列变化或者就这个邻接矩阵而言，将每个行看作一个属性，每个列看作目标，就可以很轻松的转为聚类，用聚类的方法求解。所以，两者并没有完全独立，只是考虑的角度不同，可以结合使用。现在社交网络方向有一个很热门的就是用attributes来辅助进行社团检测，是对传统的社团检测和聚类方法的一种改进，两者优势互补。

**相同点：**二者都属于无监督学习。聚类算法本质是“簇内距离最小化，簇间距离最大化”。社区发现使得“同一社区内部的节点链接较为紧密，社区之间的节点链接较为稀疏”。

**不同点：**聚类能够适应多种数据(点数据,图数据,轨迹数据等等),处理多种属性类型。 而社区发只适用于图数据，它依赖于edge这种单一的属性类型。

聚类是更general的方法，社区发现本质上就是图中结构紧密的节点的聚类。

**二 社区发现算法在风控领域的应用**



有向图 无向图

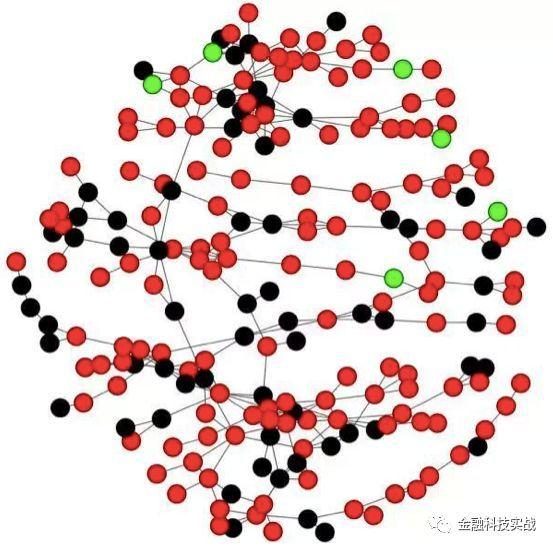
无向图仅表示节点和节点之间是否有关系，如同学、朋友等社交关系都是无方向的社交网络。例如：在P2P行业反欺诈建模中，可以通过申请者通讯录去获取其社会关系，假设张三和李四的通讯录都有老赖王五，那么，张三和李四的贷款申请违约风险就会比较高。

有向图相比于无向图会携带方向信息，如电话、邮件、短信等都是有方向的社交网络，一个最简单的例子就是传销图。传销有非常成熟的上下线制度，是发展团队十分迅速有效的手法，也被互联网公司广泛用于发展用户——好友邀请制度，此外，保险销售公司也有类似的提成机制。如被不法分子利用规则，对互联网公司，产生的后果就是大规模虚假注册；对保险销售公司，产生的后果就是内外勾结骗取额外提成。

近年来，消费金融行业快速发展，相比于传统商业银行，形成了自己独特的优势：填写字段少、在线操作、审核速度快、放贷及时。这类申请人群通常因缺乏征信信息（一是客户年轻化，而是一些消费金融公司不具有查询征信的资格）而给消费金融企业带来了巨大的信用和欺诈风险。

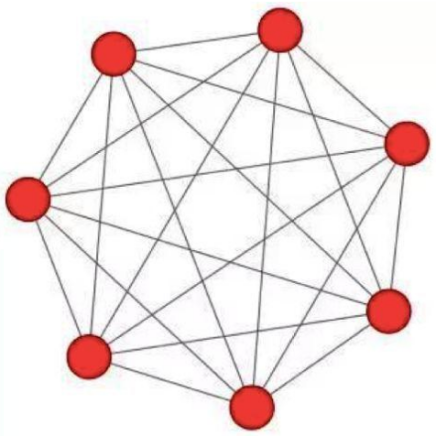
如何在有限信用记录甚至是“零”信用记录下进行更准确的风险控制和欺诈识别是消费金融公司降本增效的关键问题。

下图是一个诈骗团伙构成的社区结构。



图中，红色的点代表被拒绝的用户，黑色的点代表穿越用户（通过申请但有逾期表现的用户），绿色的点代表通过且表现良好的用户。总结一下，该团伙的拒绝率达到66.8%，说明该团伙的平均用户信用值较低；穿越用户占所有通过用户的91.4%，进一步验证了该团伙的欺诈性。

派系是指任意两个点都相连的节点的集合，又称为完全子图。

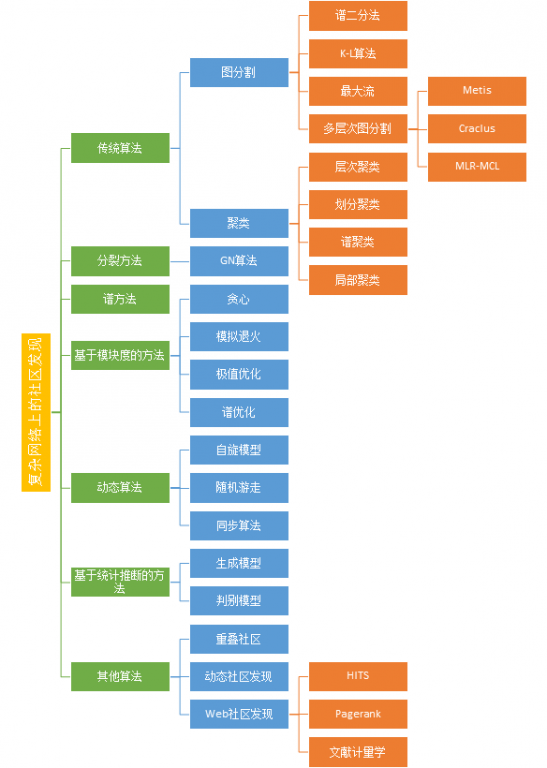


特别地，在利用社交网络分析（SNA）进行社区分析时，派系图具有更高的风险性。这种图的背后通常是多人协作的团伙作案，其两两互通表示两两认识，背后的目的多为相互勾结，伪造信息以达到消费金融借贷审核要求，且这类社群多有内外勾结的情况，需要重点关注。

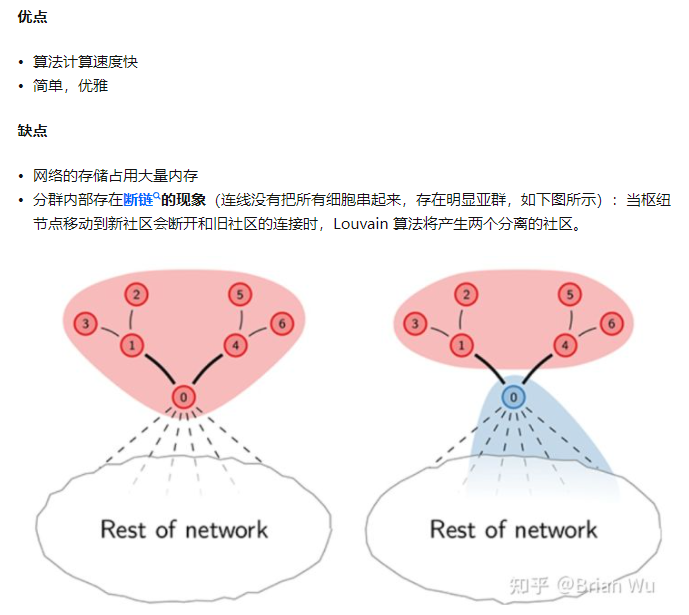
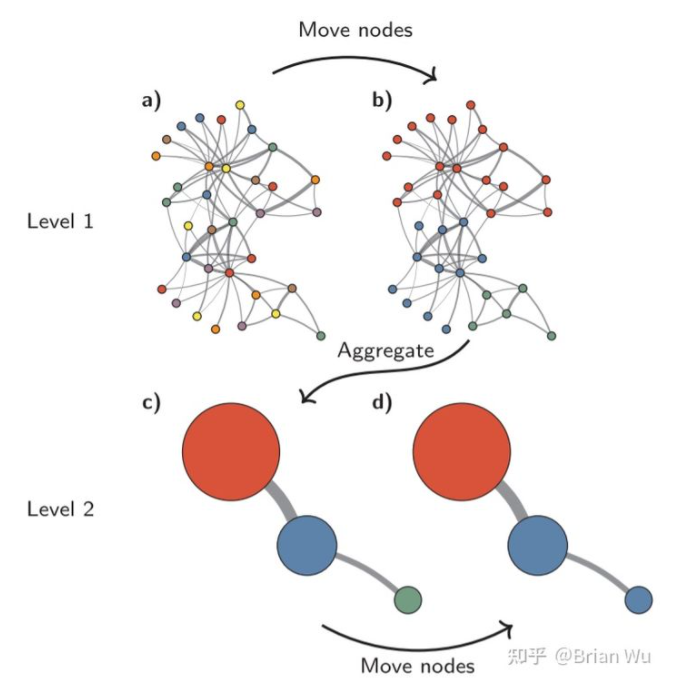
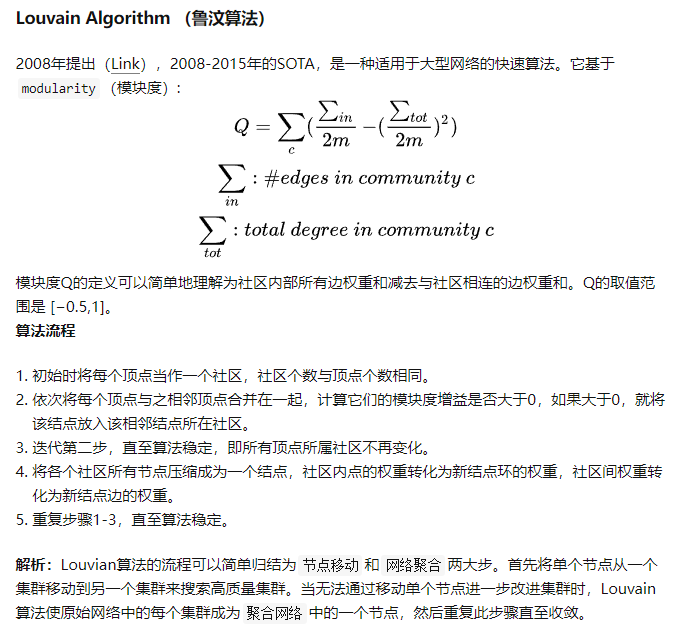
三 社区发现算法分类及简介（待续...）

由社区结构在风控领域应用可知，好的社区网络划分对结果至关重要，社区发现算法（不全）总览如下：





效率比较高的算法有Louvain，Label Propagation，Infomap等，其中以infomap综合优势最好，因为infomap通吃所有类型的网络(有向无向有权无权)，且是线性时间，发现的社区质量也比较好。在实际应用中，也有人测试出Louvain下效果更好些。说明社区发现类算法并不存在一个最好的算法，因为在现实数据中对于社区或者说团伙的定义千差万别，不一定都跟算法的假设匹配。为获得更好的预测效果，通常需要多种社区算法联用。



参考资料：

1. https://blog.csdn.net/xiaoxiaoniaoer1/article/details/103669269?spm=1001.2101.3001.6650.7&utm\_medium=distribute.pc\_relevant.none-task-blog-2%7Edefault%7EBlogCommendFromBaidu%7ERate-7.pc\_relevant\_default&depth\_1-utm\_source=distribute.pc\_relevant.none-task-blog-2%7Edefault%7EBlogCommendFromBaidu%7ERate-7.pc\_relevant\_default&utm\_relevant\_index=10
2. <https://blog.csdn.net/qq_43584847/article/details/102561438?utm_medium=distribute.pc_relevant.none-task-blog-2~default~baidujs_title~default-0.pc_relevant_default&spm=1001.2101.3001.4242.1&utm_relevant_index=3>
3. <https://zhuanlan.zhihu.com/p/428821516>
4. [https://www.infoq.cn/article/C99whYfeGILp1W\*M75cl](https://www.infoq.cn/article/C99whYfeGILp1W*M75cl)
5. <https://www.zhihu.com/question/29042018>