重要节点组影响力最大化实验总结

本文档主要总结上学期实验部分

song zhengqian

2021

目录

[一、研究内容 2](#_Toc67236794)

[二、对标论文 3](#_Toc67236795)

[三、研究内容 4](#_Toc67236796)

[1、信息熵 4](#_Toc67236797)

[2、重叠系数 5](#_Toc67236798)

[3、数据集介绍 5](#_Toc67236799)

[四、实验结果 5](#_Toc67236800)

[五、实验分析与展望 8](#_Toc67236801)

## 一、研究内容

**重要节点组挖掘：从社交网络中选出一组节点，使节点组的综合影响力最大化。**如图1所示，最左边为社交网络图G=(V,E)，每个节点i∈V代表一个社交用户，每条边eij∈E表示节点i与节点j之间存在社交联系。在社交网络G中根据某种规则选出一组种子节点，如图一中间部分，其中红色的三个节点即为选择出的一组种子节点S。将信息通过种子节点组S在网络G中进行信息传播，统计最终达到稳定状态后信息传播的范围，即为图1最右边所示的红色节点的数量就是信息所传播的范围。**而红色节点越大则代表信息传播范围越广，即挖掘到的节点组越重要**。

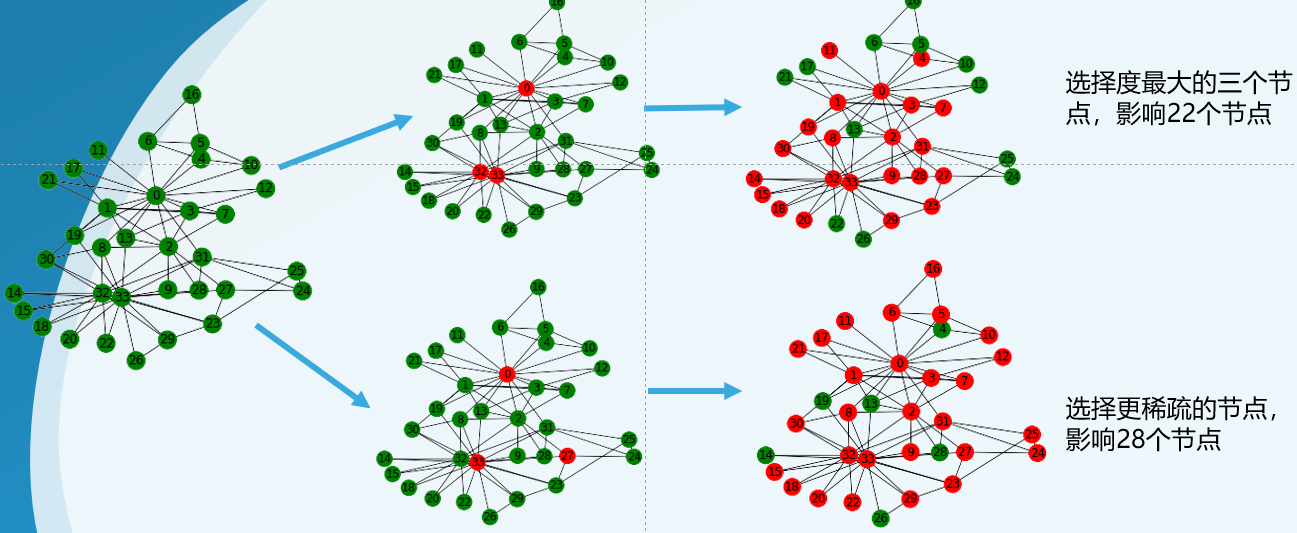


图1 研究内容示意图

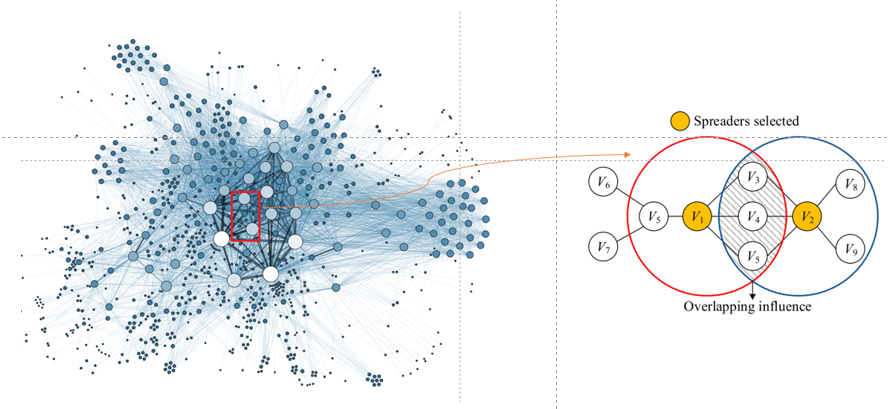


图2 影响力重叠示意图

由此可见，**重要节点组挖掘主要分为两部分：一是节点重要性度量，二是节点组综合影响力最大化。**目前大部分研究只针对于单个节点进行重要性度量，而将单个节点重要性度量方法直接套用到节点组（多个节点）重要性度量会产生**影响力重叠的问题**。其中图1上部分单纯使用单个节点重要性度量指标得到的节点组综合影响力不如下部分中选择更加稀疏的节点的综合影响力，这个问题在图2中得到了解释，即种子节点之间存在影响力重叠的问题，造成资源浪费，综合影响力无法最大化的问题。

基于以上的问题与思考，我目前的改进思路有以下三条：

1、使用信息熵与网络拓扑结构衡量节点重要性。（信息熵+边中心性）

其中**信息熵由节点的度以及介数中心性定义**，同时为了更充分利用网络的信息，**考虑对（无向无权图中的）边的权重进行度量**。节点的度包含了节点的局部信息，而节点的介数中心性则包含了节点的全局信息。网络中边的权重反应了两个节点之间的交互程度，而由于某些原因我们无法直接观测到边的权重，因此引入边中心性这个概念使得节点的重要性度量更加合理。

2、结合1与重叠系数实现削弱节点组影响力重叠的问题，实现综合影响力最大化。

使用单个节点重要性评价指标选择出的一组种子节点之间的影响力有重叠范围。针对这个问题，提出重叠系数这个指标，用于度量待选节点的邻居与种子节点的邻居重叠的比例。通过重叠系数削弱影响力重叠效应，使得种子节点的分布更加稀疏，达到影响力最大化的目标。

3、结合1、2与社团发现以及“桥节点”削弱节点组影响力重叠的问题，实现综合影响力最大化。

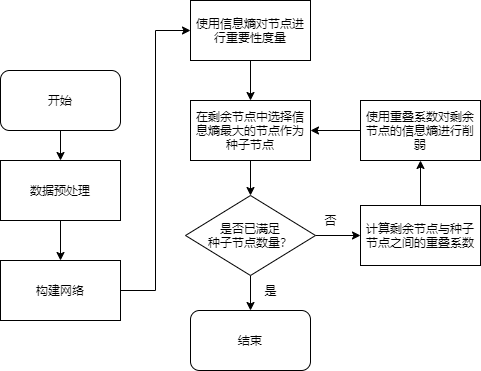
## 二、对标论文

1、Guo C , Yang L , Chen X , et al. Influential Nodes Identification in Complex Networks via Information Entropy[J]. Entropy, 2020, 22(2):242. 这篇使用信息熵衡量节点重要性

2、Beni H A, Bouyer A. TI-SC: top-k influential nodes selection based on community detection and scoring criteria in social networks[J]. Journal of Ambient Intelligence and Humanized Computing, 2020: 1-20. 这篇使用社团发现的观点进行重要节点组挖掘

## 三、研究内容

本部分主要总结上学期已完成的实验结果。**本实验的实验流程图如图3所示。即使用信息熵与网络拓扑结构衡量节点重要性。同时也加入了重叠系数进行优化。**



**图3 算法流程图**

### 1、信息熵

首先介绍节点信息熵的计算，节点v的信息熵的计算公式如公式1和公式2。其中公式1为信息熵的标准计算公式。而公式2则表示了信息由节点v传播到节点u的概率，其中表示节点u属于节点v的一阶邻居。公式2中的表示节点u的度（degree）。表示节点v的一阶邻居。

 (1)

 (2)

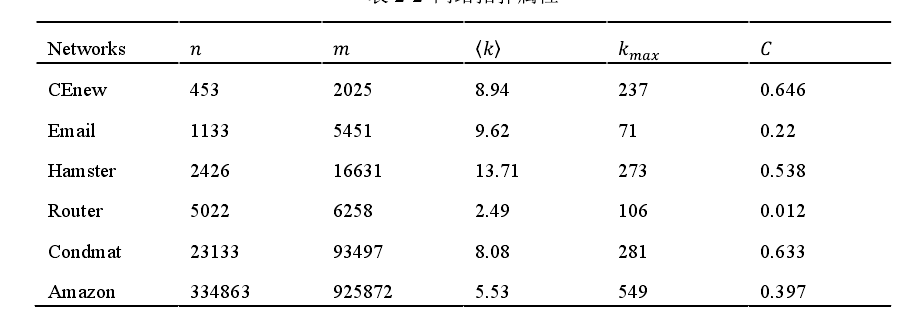
### 2、重叠系数

富人俱乐部现象，即具有较大度值的节点互为邻居，如果将这些节点选作种子节点，**传播过程中会出现大量共同邻居，使得种子节点的影响范围相互重叠，**示意图如图2所示。本方案提出重叠系数公式，用于**度量待选节点的邻居与种子节点的邻居重叠的比例**。在图2中，在我们已经选择节点V1作为种子节点后，判断是否选择节点V2作为种子节点。**重叠系数衡量了目标节点V2与种子节点V1之间的一阶邻居的重叠率。**如公式3所示，S1表示节点V1的一阶邻居集合，S2表示节点V2的一阶邻居集合。根据重叠系数公式可知，若重叠系数的值越大，则表示目标节点V2与种子节点V1有较大影响力重叠，因此我们可以利用**重叠系数对目标节点V2的重要性进行削弱，避免影响力重叠**。V2的信息熵更新公式如公式4所示。

 (3)

 (4)

### 3、数据集介绍



CEnew是线虫的新陈代谢网络；Email是一个电子邮件网络系统用户之间的收发网络；Hamster反映了网站http://www.hamsterster.com 的朋友和亲人关系网络；Router反映了因特网路由层面的拓扑结构关系；Condmat是arXiv上的论文作者合作关系网络，它反映了论文中各个作者之间的合作关系；Amazon反映了产品之间被同时购买的关系图。**所有网络均为无权无向网络**。

## 四、实验结果

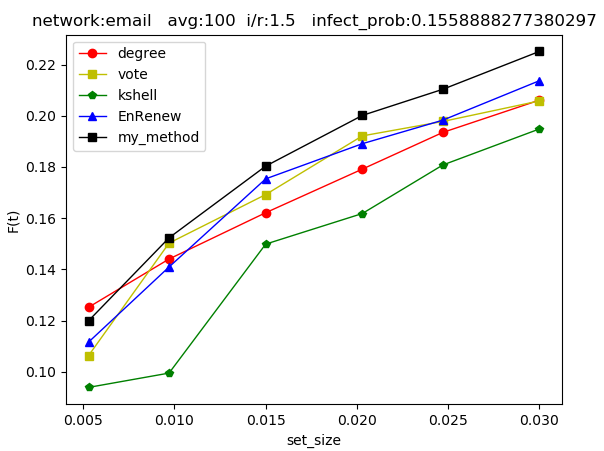
传播范围F(tc)实验结果

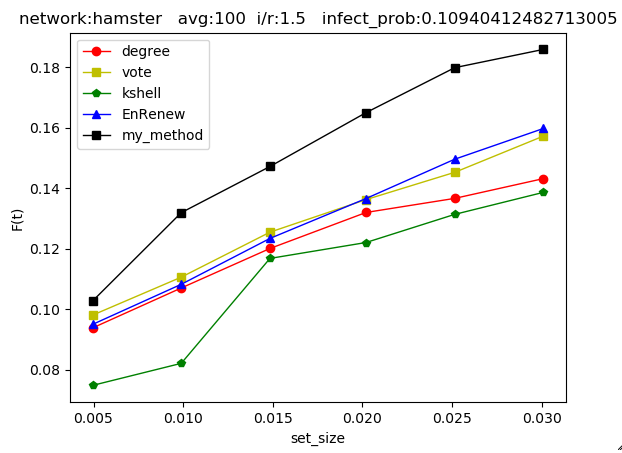
**传播范围是指经过算法选择的种子节点组在网络中进行信息传播，达到最终稳态时传播到信息的节点数量占总节点数量的比例。**是评判重要节点组挖掘效果的主要指标。其计算公式如公式5所示。

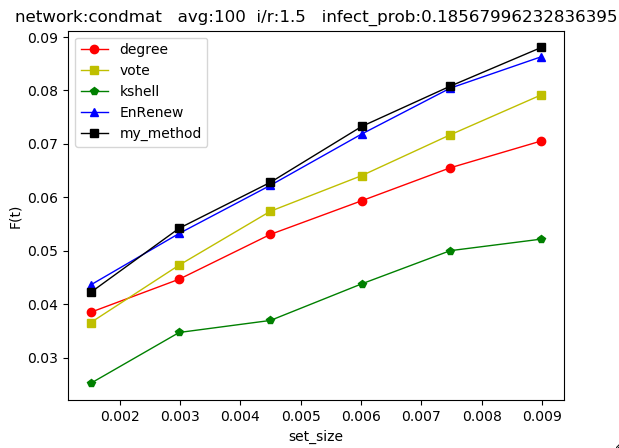
 (5)

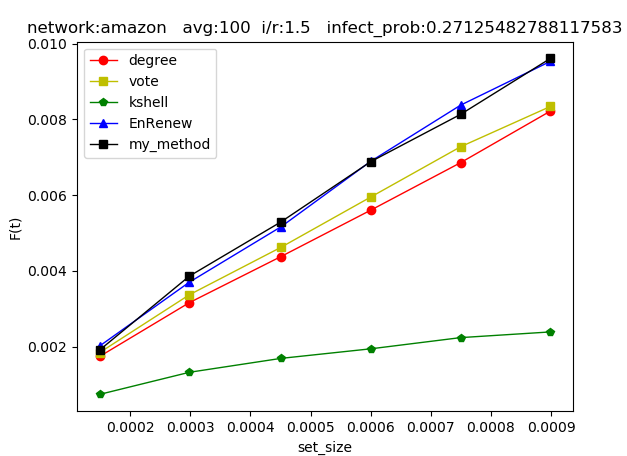
是指经过tc时刻，达到稳定状态时获得信息的节点数量，n为网络节点总数。

**其中黑色线条为我的实验方案的结果，蓝色线条为对标论文1的实验结果**。红色为单纯使用节点度作为指标的实验结果，黄色为2016年VoteRank算法的实验结果，绿色为kshell算法结果。可以看到**黑色线条在前两个小数据集（Email、hamster）上提升较大，而在后两个大数据集(condmat、Amazon)上提升较小。**









## 五、实验分析与展望

这个实验结果的信息熵定义是我最原始的方案，即只用节点的度这个指标来定义，**实验效果在小数据集上较好，在大数据集上提高不大**。分析原因可能是这种简单的信息熵定义方法**对于小数据集能不错的利用节点的局部以及全局信息，而对于大数据集，这种定义方式往往只能利用节点局部信息**，对于全局信息的利用几乎没有。因此下一步改进策略是对于定义**增加节点全局信息的利用，**如节点的介数中心性、边中心性等。