基于 GitHub 用户行为的开发者社团发现

孙秋实 宋晏如

East China Normal University
School of Data Science and Engineering

2022 年 5 月 25 日

Outline

现实世界中的网络（图）

网络/图的特征

- 有向图
- Strong Local Clustering
- Power-Law（幂律分布）

以互联网理解 GitHub 数据集问题

- 网页链接/文章引用 → 仓库/人的相互依赖关系
- 互联网水军 → 仓库刷 Star/用户刷 Follow

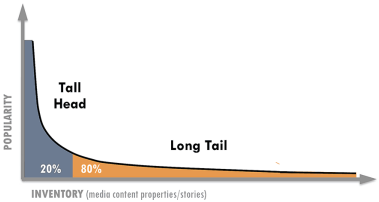


图: Power Law Distribution

链接排名算法的局限性

网页链接排名 PageRank

- 可以找出高权重用户，但并不能找到类似的人
- 可能会受极端值影响
- 无法考虑到用户的行为特征

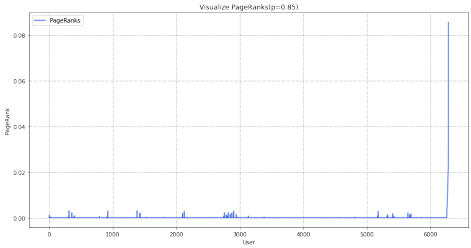


图: Page Rank

模块度算法初步

Louvain 模块度算法

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{1}$$

优化目标：模块度增益

$$\Delta Q = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m} \right)^2 \right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right] \tag{2}$$

*** 可采用分布式系统实现**

模块度算法初步

借助 Gephi 进行实验

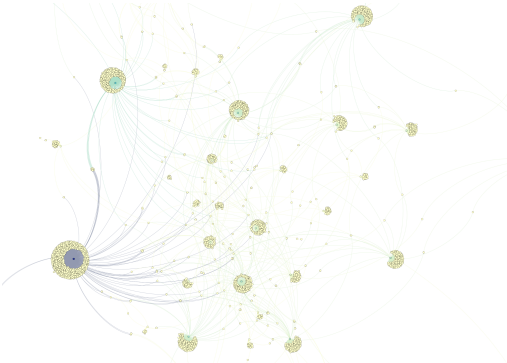
- 抽取 15k 用户数据（受到性能限制）
- Louvain 模块度算法与可视化

图: Gephi Modularity Alg & Visualization

模块度算法 Cont'd

社区划分的有效性

- Degree 的角度：显著的“长尾”性质
- 平均度数 1.035
- 社区的稀疏链接依靠极个别高入度公共节点



图：Degree

模块度算法 Cont'd

社区划分的有效性

- Modularity: 0.696, Num of Communities: 71
- **显著的“长尾”性质**
- *in practice it is found that a value above about 0.3 is a good indicator of significant community structure in a network*[1]

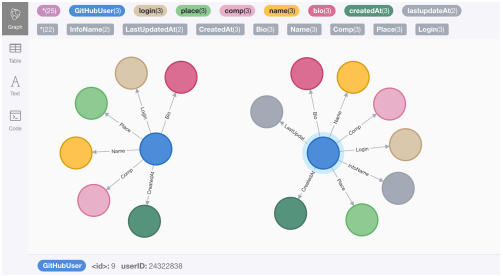
图: Community Size Distribution

[1] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. “Finding community structure in very large networks” .

模块度算法 Cont'd

模块度权重的设计策略

- 企业：54927 个（较难构成映射关系）
- 地点：28648 个（较难构成映射关系）[2]
- 用户活跃时间跨度：2021.11.8-2022.2.2
  - Min-Max 归一化 $\Delta Q = +0.101$
  - Log 指数转换 $\Delta Q = +0.087$



图：GitHub User's feature

[2] 需要进行共指消解

社区发现：标签传递

标签传递算法

- **Pros**：简单、快速、内存开销少
- **Cons**：
  - 迭代结果受初值影响
  - 随机性强，鲁棒性差
  - 易出现无意义的社区

案例分析 Cont'd

社区中的“离群”高权重节点[3]

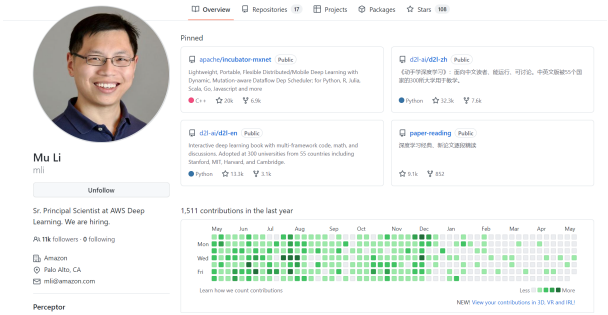


图: Muli's Profile

[3] **最极端例子**: Linus Torvalds 158k followers & 0 following

案例分析 Cont'd

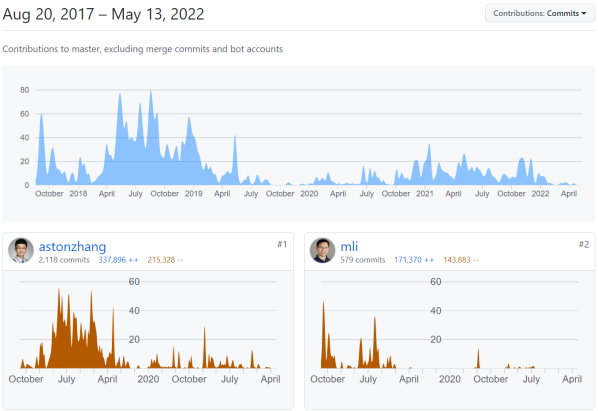
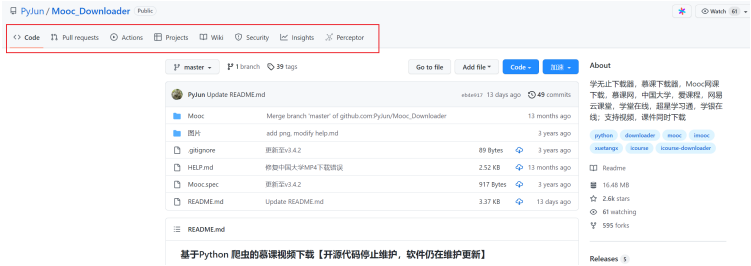


图: D2L Contribution

案例分析 Cont'd

问题用户

- 同样方式对涉及水军行为的用户进行探索
- 离群值用户，大量 followers 为仅关注该用户的低活跃度账号



图：问题用户

案例分析 Cont'd

问题用户

- .exe，而不是源代码
- 隐藏付费行为

学无止下载器-v2.0.1 Latest

学无止下载器最新稳定版本为 v2.0.1
使用说明:
选择下载 Setup-学无止下载器-v2.0.1-Win.exe
双击 exe文件进行解压安装即可，会自动在桌面创建一个快捷方式。
软件说明:
支持慕课网,网易云课堂,中国大学,爱课程,学堂在线,超星学习通,六大网站视频及课件下载!
新增支持哔哩哔哩，腾讯视频，优酷视频，爱奇艺，Youtube，Pornhub 等网站的视频下载!
普通用户:限速1M/s; VIP用户:无限速下载; SVIP用户支持登录下载已购买课程
GitHub：https://github.com/PyJun/Mooc_Downloader
帮助文档： https://github.com/PyJun/Mooc_Downloader/wiki

下载地址：

1. 蓝奏云下载： https://pyjun.lanzoui.com/b00n4ln4b
2. 百度云下载： https://pan.baidu.com/s/1rP5Q66j1xF3D5I5iAL335g
   觉得好用的话，就帮忙在右上角点个star哦

图：问题用户

社区发现的分布式实现

Spark 分布式实现[4]

Louvain 算法的计算瓶颈：第一轮迭代


```
['(80982860,{community:1201310,communitySigmaTot:475,internalWeight:0,nodeWeight:1})',
 '(8133487,{community:449224,communitySigmaTot:5575,internalWeight:0,nodeWeight:2})',
 '(30236891,{community:6313872,communitySigmaTot:499,internalWeight:0,nodeWeight:1})',
 '(4029488,{community:2138339,communitySigmaTot:983,internalWeight:0,nodeWeight:1})',
 '(12583921,{community:14338007,communitySigmaTot:809,internalWeight:0,nodeWeight:1})',
 '(10450243,{community:52195,communitySigmaTot:8215,internalWeight:0,nodeWeight:1})',
 '(93820582,{community:195327,communitySigmaTot:4283,internalWeight:0,nodeWeight:2})',
 '(30904526,{community:8583900,communitySigmaTot:147,internalWeight:0,nodeWeight:1})',
 '(73826061,{community:10350960,communitySigmaTot:510,internalWeight:0,nodeWeight:5})']
```


图：Louvain 算法的迭代，节点权重被合并

对 1.1M 用户进行社区发现

- 第一轮迭代：3771 个 Communities
- 第二轮迭代：1061 个 Communities
- ...

[4]可以借助 Spark 对接 MaxCompute 以 Pipeline 的形式进行

社区发现的拓展 Cont'd

不连通图的处理

- 紧密中心性算法的变体：Harmonic Centrality algorithm
- 用于处理数据量较少且社区不连通的情景

衡量社区的“宽度”

- KSP 算法
- 探索协作紧密程度

此部分在报告中展示

未来展望

对少部分数据进行标注

- 缩小算法搜索空间，缓解 LPA 等算法的不确定性
- 可以采用众包方式完成

对重叠社区的进一步探索

- 模块度算法的局限性
- 将社区重叠纳入考虑会更加符合现实世界规律

“实时”数据处理（Spark 分布式实现的拓展）

- 互相影响与推送的时效性
- 流数据处理方法

Acknowledgment

- 感谢倾听
- Q & A?

DaSE Data Science & Engineering