# 社会网络分析软件研究

刘 鹏 李先贤 王利娥
(广西师范大学广西多源信息挖掘与安全重点实验室 桂林 541004)
(广西师范大学计算机科学与信息工程学院 桂林 541004)

**摘 要** 随着社会网络数据规模的增长,人工处理方式已经不能满足社会网络分析的需求。介绍了4款常用的社会网络分析软件nodeXL、Pajeck、Gephi和networkX。从支持数据格式、可视化性能、统计分析功能、帮助文档、使用难度等多个方面对以上软件的特性进行对比分析,给出了客观的综合评价,并对选择和使用这些软件提出了相应的建议。

**关键词** 社会网络,数据分析,软件
**中图法分类号** TP317 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.12.037

## Study of Social Network Analysis Software

LIU Peng LI Xian-xian WANG Li-e
(Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, Guilin 541004, China)
(College of Computer Science and Information Technology, Guangxi Normal University, Guilin 541004, China)

**Abstract** With the growth in the size of social network data, manual processing methods cannot meet the needs of social network analysis. This paper reviewed 4 commonly used and well-documented social network analysis softwares, which are NodeXL, Pajeck, Gephi, and networkX. Through the comparison of the software's inputting data format, statistical performance, data visualization, help documentation, and difficulty of use, we gave an objective evaluation of their overall performance, and recommendations of selecting and use.

**Keywords** Social network, Data analysis, Software

## 1 简介

随着网络信息技术的发展,社会网络服务如新浪微博、QQ空间、人人网、淘宝网、LinkedIn和Google+等大量出现。社会网络服务给我们生活带来了极大的便利,使用社会网络服务能够方便地与朋友通讯、交流、发表自己的想法并及时获取信息。社会网络改变了信息及内容数据的产生和传播方式,我们在使用社会网络信息的同时也为社会网络贡献评论、照片、心得体会等数据。特别是随着手机、平板及可穿戴智能设备的发展,我们在线的时间越来越长,由此产生的数据越来越多,越来越详细,使得基于网络的虚拟数字世界将逐步同现实世界产生融合。社会网络服务所产生的数据是社会活动的真实反映,包含了社会生活各方面详细的信息,蕴含着无可估量的价值。这些数据吸引了科研工作者、政府、企业等各团体的注意,他们都希望能够获取并分析这些数据。

社会网络分析是采用图论等数学方法,通过研究大量个体(个人、组织等社会活动参与方)间的关系来寻找社会系统表面之下的内在规律。社会网络分析方法已经被广泛应用于推荐系统、电子商务系统、舆情监控系统等多种具体的应用中,并且在现代社会学、人类学、社会语言学、地理、社会心理学、资讯科学、经济学以及生物学等领域的研究中发挥了重要作用[10]。社会网络应用每天产生数以亿计的数据,传统的人工统计分析已经不能有效处理这些数据。挖掘和使用这些数据需要借助计算机及专门的分析软件进行处理。目前,社会网络数据分析使用的软件有:NetworkX、Pajek、NodeXL、NetDraw、NetLogo、UCIET、R、Gephi、SNAP等。这些软件的功能、用法和特长各不相同,因此有必要对社会网络的分析软件进行整理和介绍。本文选取了功能和特点具有代表性的NodeXL、Pajek、NetworkX和Gephi软件进行对比分析研究,对这些软件进行客观的评测并提出使用建议。这4款软件中,NodeXL、Pajek和Gephi属于独立软件,NetworkX属于工具类库。独立软件指能够独立运行使用,完成特定的工作的软件;工具类库通常只提供特定的功能函数接口,需要在特别的开发环境下调用这些功能完成特定的工作。

## 2 社会网络分析软件对比

社会网络是由许多节点构成的一种关系网络,节点通常是指个人或组织,边代表个人或组织间的交互关系[1]。社会

到稿日期:2015-02-08 返修日期:2015-03-29 本文受国家自然科学基金项目(61272535,11362003,61165009),广西科学研究与技术开发计划项目(14124004-4-11),广西区域多源信息集成与智能处理协同创新中心,"八桂学者"工程专项经费资助项目,广西高校科学技术研究项目(KY2015YB032),广西师范大学青年科研基金资助。

刘 鹏(1979—),男,硕士,讲师,主要研究领域为网络安全、数据分析,E-mail:liupeng@gxnu.edu.cn;李先贤(1969—),男,博士,教授,主要研究领域为网络、信息安全;王利娥(1981—),女,硕士,讲师,主要研究领域为对等网络、信息安全,E-mail:wanglie@gxnu.edu.cn(通信作者)。

网络需要进行建模才能在计算机中存储处理，一般采用图数据结构 $G(V,E,L,\phi)$ 表示社会网络，其中 $V$ 为节点的有限集合，表示社会网络中的个体；$E\subseteq V\times V$，表示社会网络中的关系；$L$ 为标签的集合，用于唯一标识社会网络中的个体；$\phi$：$V\rightarrow L$，为每个节点分配一个标签[11]。

图数据的存储主要有邻接矩阵和邻接链表两种数据结构。对于这两种模型不同分析软件有不同的实现方式，表现为不同格式的文件。常见的有格式有 Pajek(. net)、NetworkX (. yaml)、Gephi (. gexf)、GraphViz(. dot)、NodeXL (. csv,. txt,. xls,. xslt)、Tulip(. tlp,. dot) 等[2]。下文以应用较广泛的 Pajek 软件的. net 格式为例，对图数据的存储进行简要说明。如图 1 所示，. net 文件为一个纯文本文件，可以使用任何文本编辑器编辑。文件由两大部分组成，即点数据部分和关系部分。点数据部分以"* Vertices"开始，随后接空格，然后是所存储的点的个数。换行后开始存储点数据。点数据分为两部分，第一部分是整数代表的点的标识；第二部分是用双引号包围的标签，标签可以是任意的文本串。关系部分以"* Edges"开始，换行后，用数对表示关系，如"1 4"表示点 1 和点 4 间有一条边。图 1 所示记录内容为 1 个具有 7 个点和 7 条边的图，用 Pajek 软件图示化为图 3。此外. net 文件还支持弧表示法和矩阵表示法等，详见手册[3]。

```
1 *Vertices  7
2 1 "Actor 1"
3 2 "Actor 2"
4 3 "Actor 3"
5 4 "Event 1"
6 5 "Event 2"
7 6 "Event 3"
8 7 "Event 4"
9 *Edges
10 1 4
11 1 5
12 2 4
13 2 5
14 2 6
15 3 4
16 3 7
```

图 1 . net 文件

社会网络分析就是从多个不同角度(微观、宏观等)对社会网络进行分析，通过研究网络的局部模式对社会的整体行为进行推断，通过整体数据的聚类、分割对系统的行为做出合理的解释。主要的分析内容有：中心性，指分析人或组织在其社会网络中具有怎样的权力，或是处于怎样的中心地位；凝聚子群，分析联系紧密的社会成员组成的团体内部及团体间的关系。研究中需要用到的属性有：节点度、介数、平均距离、子图、聚集系数、最短路径等[12]。

对大量的社会网络数据进行分析研究，需要使用专业的数据分析工具。目前，社会网络分析工具有很多，需要解决的问题是怎样从众多的社会网络分析软件中找到适合使用者的工具软件。本文通过分析和测试主流的社会网络分析软件，选择其中一直在不断更新完善的 4 款免费软件 NodeXL、Pajeck、Gephi 和 networkX 进行详细的对比测试和评价，使读者能够了解这些软件的性能及特点，为选择、学习和使用这些软件提供一定帮助。各软件的版本信息及获取地址如表 1 所列。

表 1 社会网络分析软件版本及下载地址

| 软件 | 版本 | 发布时间 | 地址 |
|---|---|---|---|
| Pajek | 3.15 | 2014-03-04 | http://pajek.imfm.si/doku.php?id=pajek |
| NetworkX | 2.0 | 2014-07-21 | http://networkx.github.io/ |
| NodeXL | 1.0.1.331-Beta | 2014-01-23 | http://nodexl.codeplex.com/ |
| Gephi | 0.8.2-beta | 2013-01-03 | https://gephi.github.io/ |

## 2.1 Pajek

Pajek 是一款 Windows 平台的免费(非商业用途)软件，用于分析和显示复杂网络，也可以用于分析社会网络。Pajek 采用 Pascal 语言开发，只提供免费安装程序，未提供源代码。1996 年，Vladimir Batagelj 开发了 Pajek0.0.1 版，目前最新版本是 2014 年 3 月发布的 3.15 版本。Pajek 能够分析普通图(有向、无向、混合图)、多关系图、2-模图及动态图。Pajek 能挖掘输入数据的结构关系，根据节点的核心性、连通性等进行聚类分组，能够输出和显示划分的结果，并动态显示节点删减对划分结果的影响。Pajek 采用菜单驱动，图 2 为软件的主窗口。主窗口包含 17 个菜单和 6 个窗口部件。主控窗口的 17 个菜单包含软件所支持的全部功能操作，6 个窗口部件用于提供文件输入、保存和显示操作。File 菜单提供数据对象的输入、输出操作；Network 菜单提供输入数据的变换、移除、添加、更改等操作，软件的分析功能如求 $k$ 核、$p$ 团体、划分、聚类等操作都在此菜单下；Operation 菜单包括数据的收缩、提取、分区等操作选项；Draw 菜单提供显示操作，其它菜单功能详见使用手册[3]。图 3 是图数据图形化显示窗口，通过菜单能够对所显示图形的布局、颜色、标识等进行调整。

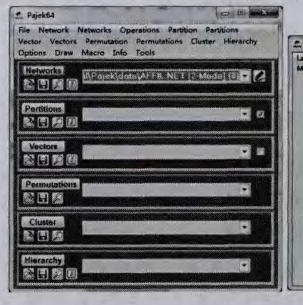


图 2 Pajek 主窗口

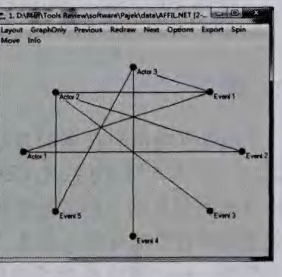


图 3 Pajek 数据显示

Pajek 使用的算法时间复杂度都低于 $O(n^2)$，相对于其它分析软件，在处理大型的网络分析时速度较快，但是付出的代价是统计功能不强。为了弥补统计功能的缺陷，Pajek 提供了 R 软件接口，通过 R 软件进行所需的统计分析。为了更好地体现 Pajek 在处理大规模数据时的优势，Vladimir Batagelj 还开发了一个特别版 PajekXXL，其能够完成 Pajek 同样的功能，内存需求却比原版少 2～3 倍。

## 2.2 NodeXL

NodeXL 是一款 Windows 平台下免费的开源(Ms-PL)社会网络可视化和分析软件工具，由微软研究院的 Marc Smith 领导的团队采用 C# 语言结合. net 平台技术完成，最新版本为 2014-01-23 发布的 1.0.1.331Beta 版。NodeXL 为非程序员提供了一个功能强大且易于使用的社会网络分析工具，它以模板的形式集成在 MS Excel(2007，2010 或 2013)中，利用熟悉的 Excel 表格作为数据展示和分析平台，用户能够快速掌握软件的使用方法。NodeXL 支持多种输入、输出格式，包

括 GraphML、Pajek、UCINet 和矩阵等[4]；能够直接从 Twitter、YouTube、Flickr 和电子邮件导入社交网络数据，或用多功能插件从 Facebook、Exchange 和万维网超链接直接抓取社交网络实时数据；能够对数据进行中心性分析(入/出度、中介中心性、接近中心性等)，能够对数据进行聚类分组，支持 Clauset-Newman-Moors，Wakita-Tsurumi，Girvan-Newman 3 种聚类方式[5]。NodeXL 数据可视化功能强大，图 4 为软件的主界面，图像区域的内容为 NodeXL 软件功能概要图，从图中可以直观地了解软件提供的功能。NodeXL 支持手动及多种自动布局并能实时缩放，能够对数据进行按指定条件动态筛选并显示；能够通过填写工作表的单元格来设置颜色、形状、大小、标签和透明度，或选择根据度中心性、中介中心性等度量自动填写属性并显示。

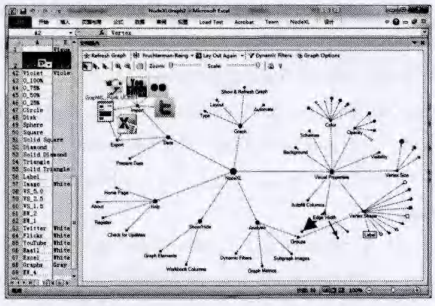

图 4　NodeXL 功能特性可视化图

### 2.3　Gephi

Gephi 是一款跨平台的免费开源(GPL3)网络分析和可视化软件，采用 Java 语言开发，OpenGL 为显示引擎[6]。目前 Gephi 中文版本为 2013-01-03 发布的 0.8.2-Beta 版，是 4 款软件中唯一支持中文菜单显示的。Gephi 软件的分析功能较弱，支持数据中心性分析和较少的聚类分析，但能很好地支持动态图数据分析。Gephi 支持的输入数据格式有“.dot”，“.gdf”，“.gml”，“.net”，“gml”，“.gexf”等，是 4 款软件中支持输入数据格式最多的。Gephi 的数据可视化效果是这 4 款软件中最好的，而且支持超过 5 万个点的复杂条件动态实时过滤。图 5 展示的是 Gephi 软件的主界面及示例数据“Les Miserables.gexf”可视化分析的效果。

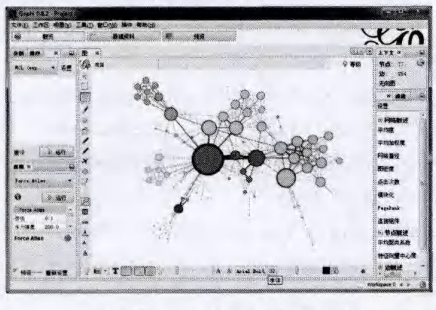

图 5　Gephi 主窗口及数据显示

### 2.4　NetworkX

NetworkX[7]是一套免费的图论与社会网络分析工具库，采用 Python 语言开发，遵循 BSD 开源协议，具有跨平台、易扩展、易使用等特点。NetworkX 支持无向图、有向图及多重图等多种图类型，并能够对不同图类型数据进行转换并保存输出。NetworkX 能够生成规则图、小世界图、ER 图、无标度图等多种经典数据。NetworkX 采用字典模式构建图的数据结构，实现了多种图的经典算法，能够进行度、边、密度、图半径、最短路径、聚集系数等基本分析特征的计算，通过组合这些算法能够按需进行中心性分析、凝聚子群分析等多种较复杂属性分析。NetworkX 是一套工具库，不能单独使用，需要 Python 作为宿主语言进行调用组合操作。Python 语言是一种很方便易学的脚本程序设计语言，用户只需要进行简单的学习就可以快速地调用 NetworkX 的功能函数完成要求的工作。NetworkX 采用 BSD 开源许可，是 4 款软件中开源程度最高的，后续开发的软件可以选择继续遵守 BSD 或是其它的自由软件条款，甚至是封闭软件[8]。NetworkX 的源代码结构清晰，风格简练，注释详尽，可以很方便地通过修改或扩展源码获得想要的功能。NetworkX 软件不具有数据可视化功能，但通过与 matplotlib 库相结合能够很方便美观地输出二维及三维图形。图 6 所示为一个使用 networkX 库生成小世界网络并显示的完整 Python 源代码，虽然只有 5 行，但完整地实现了数据的生成及显示功能。

```
1. import networkx as nx
2. import matplotlib.pyplot as plt
3. G=nx.random_graphs.watts_strogatz_graph(18,4,0.3)
4. nx.draw(G)
5. plt.show()
```

图 6　使用 NetworkX 生成小世界网络并显示的代码

上述 4 款软件的功能及详细参数对比如表 2 所列。

表 2　社会网络分析软件详细信息对比

| 软件 | Pajek | NodeXL | Gephi | NetworkX |
|---|---|---|---|---|
| 版本 | 3.15 | 1.0.1.331-Beta | 0.8.2-Beta | 2.0 |
| 发布时间 | 2014-03-04 | 2014-01-23 | 2013-01-03 | 2014-07-21 |
| 类型 | 独立软件 | 独立软件 | 独立软件 | 类库 |
| 支持系统 | Window | Windows | 跨平台 | 跨平台 |
| 开发语言 | Pascal | C#,.NET | Java | Python |
| 授权类型 | 免费(非商业) | Ms-PL | GPL3 | BSD |
| 可视化 | 是 | 是 | 是 | 否 |
| 支持输入文件格式 | Pajek(net,.mat),Vega(.vgr),GEDCM(.ged),UCINET(.dl),Ball and Stik(.bs),mac-MOL file(.mol) | Excel(.xls,.xslt),Pajek(.net),email,.csv,.dl(UCINet),GraphML | Gephi(.gexf),GraphViz(.dot),GUESS(.gdf),LEDA(.gml),Pajek(.net,.gml),.graphml,Tulip(.tlp,.dot),UCINET(.dl),Edge list(.csv) | NetworkX(.yaml,adjacency lists,and edge lists),GML,Graph6/Sparse6,GraphML,GraphViz(.dot),Pajek(.net),LEDA |
| 图类型 | | | | |
| 有/无向图 | 是 | 是 | 是 | 是 |
| 2-模图 | 是 | 否 | 否 | 是 |
| 多重图 | 是 | 否 | 否 | 是 |
| 动态图 | 是 | 否 | 是 | 否 |
| 分析功能 | | | | |
| 中心性分析 | 是 | 是 | 是 | 是 |
| 凝聚子群分析 | 是 | 是 | 部分 | 否 |
| PageRank | 否 | 是 | 是 | 否 |
| 过滤 | 是 | 是 | 是 | 否 |
| 数据抓取 | 否 | 是 | 否 | 否 |
| 生成数据 | 多种 | 否 | 随机图,动态图 | 多种 |
| 手册资料 | 有 | 有 | 有 | 有 |

对于表中可视化项目，NetworkX 软件包虽然可以结合

matplotlib 软件包完成可视化工作，但由于其不直接支持可视化输出，因此表中填写为“否”。凝聚子群分析项目，Gephi 只支持一种聚类分组模式，所以填写为“部分”。NetworkX 需要用户设计程序才能实现，所以为“否”。手册资料项目，NodeXL 的资料全面且集成在软件中比其它 3 个更方便，易于使用。

## 3 社会网分析软件综合评价

本节从软件功能、软件系统开放性、软件资料和学习难易程度 4 个方面对它们进行定量的综合评价。4 个评测大项总分 10 分，如表 3 所列。

软件功能项目包括支持数据格式(1 分)、支持图类型(1 分)、可视化(1 分)、分析功能(2 分)、数据抓取(1 分)、数据生成(1 分)。Gephi 支持输入数据格式种类最多，得 1 分，其它软件根据支持数据类型个数比给分；支持图类型，每支持 1 种得 0.25 分；其它项目，支持得 1 分，不支持记 0 分。软件资料(1 分)包括软件功能介绍、使用手册和在线帮助 3 部分。NodeXL 帮助资料包含详细的入门指导，且集成在软件系统中，易理解、使用，有在线帮助，得 1 分；Gephi 手册资料齐全，有在线帮助，得 0.8 分；其它两个软件没有在线帮助，得 0.5 分。系统开放性(1 分)，NetworkX 采用 BSD 开源协议，自由度最高，得 1 分；Gephi 和 NodeXL 提供源代码，得 0.8 分。学习难易程度(1 分)，学习难易程度关系到用户是否选择这款软件，直观易掌握的界面设计能减少使用者在学习过程中的障碍，提高用户体验。Gephi 和 NodeXL 都做得很好，得 1 分；Pajek 次之得 0.5 分；NetwokX 采用命令方式操作，界面友好性差，得分为 0。

表 3 综合评价表

| 软件 | Pajek | NodeXL | Gephi | NetworkX |
|---|---|---|---|---|
| 数据格式 | 0.8 | 0.9 | 1 | 0.7 |
| 图类型 | 1 | 0.25 | 0.5 | 0.75 |
| 可视化 | 1 | 1 | 1 | 0 |
| 分析 | 1.5 | 2 | 1.6 | 0.5 |
| 数据抓取 | 0 | 1 | 0 | 0 |
| 数据生成 | 1 | 0 | 0 | 1 |
| 资料 | 0.5 | 1 | 0.8 | 0.5 |
| 开放性 | 0 | 0.8 | 0.8 | 1 |
| 学习难度 | 0.5 | 1 | 1 | 0 |
| 总分 | 6.3 | 7.95 | 6.7 | 4.45 |

**结束语** 从上文分析比较可以得出这 4 款软件功能、特色各有不同。Gephi 支持的数据格式最多，数据可视化效果最好，且操作界面友好，但生成模拟数据却是弱项，适合做形象的数据分析探索和生成研究发布的效果图。NodeXL 由微软研发团队开发，延续了微软产品一贯的方便、友好、易用等特点，而且提供了数据抓取功能，是初学者不可多得的好工具，但它的速度却比较慢，只适合处理小规模的数据。Pajek 各项功能均衡，速度快，适合处理大规模数据，但却没有提供源代码，而且操作不是很方便。NetworkX 由于是软件类库，很多功能需要用户通过结合 Python 程序设计语言自行设计实现，易用性不如其它软件，但提供了最大的开发自由度，适合有特别分析需求的研发人员使用。

对于初学者，建议从 NodeXL 学起，NodeXL 的操作界面、方法与 Excel 相同，能够快速上手并进行简单的数据处理。掌握 NodeXL 后，如果觉得它不能满足增长的分析需求，可以根据使用者自身的情况选择下一个要学习的软件。如果使用者有计算机程序设计基础，建议深入学习 Python 类库 NetworkX，它功能强大且开放源代码，结合 Python 语言可以方便地扩充定制分析所需的流程和功能。如果使用者没有程序设计基础，最好选择学习和使用 Pajek、Geghi 或是其它的分析软件。多种软件组合使用，总能完成所需的解决方案。

## 参 考 文 献


[1] Aggarwal C C. An introduction to social network data analytics [M]. Springer, 2011

[2] Combe D, Largeron C, Egyed-Zsigmond E, et al. A comparative study of social network analysis tools[C]// Proceedings of the Web Intelligence & Virtual Enterprises. 2010

[3] Andrej M, Vladimir B. Pajek and Pajek-XXL Programs for Analysis and Visualization of Very Large Networks Reference Manual [EB/OL]. [2014-08-02]. http://mrvar.fdv.uni-lj.si/pajek/pajekman.pdf

[4] Smith M A, Shneiderman B, Milic-Frayling N, et al. Analyzing (social media) networks with NodeXL[C]// Proceedings of the Proceedings of the Fourth International Conference on Communities and Technologies. ACM, 2009

[5] Marin A, Wellman B. Social network analysis: An introduction [M]// The SAGE Handbook of Social Network Analysis. 2011: 11-25

[6] Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks [C]// ICWSM. San Jose, California, USA, 2009: 1-2

[7] Hagberg A, Daniel A, Sshult D, et al. Exploring network structure, dynamics, and function using NetworkX[C]// Proceedings of the 7th Python in Science Conference (SciPy2008). 2008

[8] Hagberg A, Schult D, Swart P. NetworkX Reference Release 1.9 [EB/OL]. [2014-08-02]. http://networkx.github.io/documentation/networkx-1.9/

[9] Huisman M, van Duijn M A. Software for social network analysis [M]// Scott, J Carrington P J, eds. The SAGE Handbook of Social Network Analysis. 2011: 578-600

[10] Hu Chan-gai, Zhu Li-jun. Complex network software analysis and evaluation [J]. Digital Library Forum, 2010(5): 33-39

[11] 丁兆云，贾焰，周斌，等. 社交网络影响力研究综述[J]. 计算机科学，2014，41(1)：48-53
Ding Zhao-yun, Jia Yan, Zhou Bin, et al. Survey of influence analysis for social networks [J]. Computer Science, 2014, 41(1): 48-53

[12] 傅颖斌，陈羽中. 基于链路预测的微博用户关系分析[J]. 计算机科学，2014，41(2)：201-205，244
Fu Ying-bin, Chen YU-zhong. Relationship analysis of microblogging user with link prediction [J]. Computer Science, 2014, 41(2): 201-205, 244