通话社交网络社团结构实证研究

王 林,童昭维

(西安理工大学 自动化与信息工程学院,陕西 西安 710048)

摘 要:以社交网络中备受关注的通话社交网络为研究对象,对其社团结构进行分析。提出一种基于模糊综合评判分析通话社交网络权重的方法,并改进 CNM 算法进行社团划分。初步演示了通话社交网络的演化规律,为深入研究通话社交网络打下了坚实基础。

关键词:通话社交网络;加权网络;模糊综合评判方法;社团结构;改进的 CNM 算法

中图分类号:TP393 **文献标识码**:A **文章编号**:1674-7720(2013)04-0048-03

Empirical research on community structure of voice social networks

Wang Lin, Tong Zhaowei

(The Faculty of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China)

Abstract: The paper proposed a method of analyzing VSN weight based on fuzzy synthesize judgement method and the CNM algorithm is improved for the community partition. In the same time, the regularity of VSN evolution is demonstrated preliminarily in the paper, which lays a firm foundation for further studying of VSN.

Key words: voice social networks; weighted networks; fuzzy synthesize judgement method; community structure; improved CNM algorithm

虽然现实世界的复杂系统形式、功能各不相同,但其对应的网络结构却有很大的相似性。从个体层面的度、聚集系数,到整体层面的度分布、整体聚集系数等,处于中间的描述就是社团结构描述。借助复杂网络理论分析方法,科学家们成功地研究了社交网络。Onnela[1-2]从工作、家庭、休闲等方面分析了社交网络的结构,由此发现相互连接的强度和网络局域结构的关系。Szabo 和 Barabasi[2]研究表明由不同的通话网络发现两个不同地区的同一社交网络具有或强或弱的社团隔离效应的共性。Palla、Barasi 和 Vicsek[3-4]通过通话数据研究社会群体的演变发现。利用复杂网络理论分析社交网络拓扑性质并进行社团分解为进一步揭示通话社交网络的演化规律打下了基础。

社团结构划分算法的研究发展至今取得了很大的进展。早期的研究包括 Kernighan-Lin 算法、谱平分法和分级聚类方法等[5]。但在大规模及超大规模的网络社团结构分析中,算法时间复杂度和精确度的矛盾一直未能解决。目前,事先在不知道社团数目的情况下,最快算法的时间复杂度为 $O(n\log_2 n)$[6]。通话社交网络属于大规模加权网络,其中权重的分布是研究该网络的重要因素之一。本文提出一种基于模糊综合评判的方法来评价通话网络的权重,同时改进 CNM 算法使其适用于加权网络,并对现实加权通话社交网络进行实证研究。

1 通话社交网络建模

随机抽取某月某地发生通话行为的号码对。通话网络本身是一个有向网络,但若将此网络视为信息网络用作信息交流时,可把它当作无向网络。将通话加权网络中的用户抽象为图的节点,只要两个节点间有一次通话,就用一条边相连(即图的边)。考虑实验机器的运行速度,抽取某运营商 2011 年 1 月某地通话数据,其中包含的节点数为 344 522,边数为 697 489。通过计算通话数据中的通话时长、次数、频度等特性以建立加权通话网络模型。该加权通话网络反应某地区特定时间内基于各种社会关系进行过信息交流的状态。

构建的通话社交网络抽象为由点集 N 和边集 M 组成的图 $G=(V,E)$,其中节点数 $N=|V|=344\ 522$,边数 $M=|E|=697\ 489$。由于网络规模较大,考虑到相关工具的局限性,取部分节点与边建立模型,如图 1 所示。

由于从某运营商所获得的大量实验数据中包含外网的用户,故在建立移动通信关系网时不能较好地描述用户间通话关系。建立移动通信关系网络时,应滤掉外网联系记录,只保留网内通话联系记录,确保网络社团

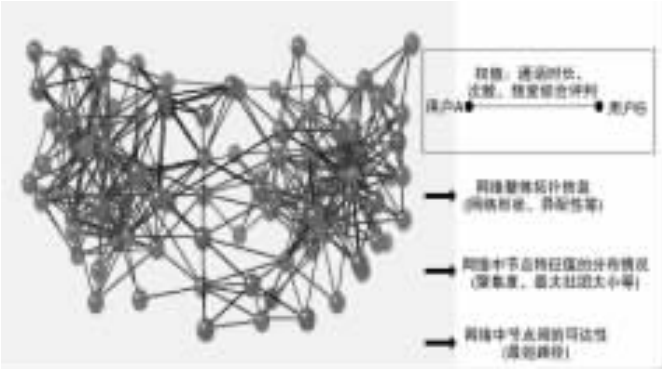

图 1 通话社交网络模型

的正确性，避免其偶然性。为反映用户真实的通话行为，排除电话销售和错误拨号的行为。同时去除每次通话时长小于 3 s 的记录和服务号码。当然这些做法会造成一些负面影响，但由于研究的时间跨度相对较长，所以造成的影响是有限的。

2 基于模糊综合评判的通话社交网络权重计算

2.1 加权网络权重定义

加权网络中每条边都具有度量连接强弱度的数值，为复杂网络节点间的关系和互相作用提供精确的描述方式。

目前加权网络有两种表示权重的方式：相似权和相异权[7]。相似权的权重表示权重越大，节点间的关系越紧密，两节点之间的距离就越小。相异权则相反，权重越大，两点间的关系越疏远；权重越小则关系越紧密。通常权重定义方式有以下 3 种[8-9]。

(1)常数权重

若加权采用常数权重分布，即网络中的每条边权重均相等且为常数。由此可知，二元网络是一种特殊的加权网络。

(2)服从指数分布的边权重

假设边的权重服从指数分布，即 $\varphi(\theta)=\theta e^{-\theta x}$，其中 $\theta>0$。服从指数分布的样本值均大于零，它与实际网络中边权重均大于零的情形一致。

(3)节点度乘积函数的边权重

设节点 i 与节点 j 的度分别为 k_i 和 k_j，则连接这两个节点的边权重定义为：$w_{ij}=(k_i k_j)^{\alpha}$，其中 α 可有效地调节节点的强度大小。

2.2 通话社交网络权重

本文研究的通话社交网络相似权 w_{ij} 范围为 $[0,\infty)$，也可归一化到 $[0,1]$。由于该网络的权值由通话频度、通话时长、通话次数等几方面共同决定，以上方法难以准确描述其权重，下面提出一种基于模糊综合评判的方法[10]对其权重进行定义。首先介绍模糊综合评判方法的思想。

模糊综合评判是对多种因素影响的事物做出全面评价的一种有效的多因素决策方法。设 $U=\{u_1,u_2,\cdots,u_n\}$ 为 n 种因素(或指标)，$V=\{v_1,v_2,\cdots,v_m\}$ 为 m 种评判(或等级)。

由于各种因素所处地位不同，作用也不一样，可用因素权重 $A=\{a_1,a_2,\cdots,a_n\}$ 来描述，它是因素集 U 的一个模糊子集。对于每一个因素 u_i，单独做出的一个评判 $f(u_i)$，可看作是 U 到 V 的一个模糊映射 f，由 f 可诱导出 U 到 V 的一个模糊关系 R_f，由 R_f 可诱导出 U 到 V 的一个模糊线性变换：

$$TR(A)=A\circ R=B \tag{1}$$

它是评判集 V 的一个模糊子集，即为综合评判。(U,V,R) 构成模糊综合评判决策模型。

对于通话社交网络，主要考虑通话频度 f、通话时长 t 以及通话次数 c 对网络中边权 w_{ij} 的影响。其中通话频度是衡量一个月某个号码的每个对端号码在该号码交往圈中出现的交往频率。因此其因素集定义为 $U=\{f,t,c\}$，然后将 U 与因素权重 A 进行模糊变化：

$$B=A\circ U=(B_{i1},B_{i2},\cdots,B_{iN})=(a_1,a_2,a_3)\begin{bmatrix} f_{i1},f_{i2},\cdots,f_{iN} \\ t_{i1},t_{i2},\cdots,t_{iN} \\ c_{i1},c_{i2},\cdots,c_{iN} \end{bmatrix} \tag{2}$$

其中 $i=1,2,\cdots,N$。

将合成的结果 B 进行归一化处理，得到通话网络的边权 w_{ij}，即：

$$w_{ij}=\frac{B_{ij}}{\sum_{i,j=1}^{N}B_{ij}}\quad(i,j=1,2,\cdots,N) \tag{3}$$

点权即节点的强度，记作 s_i，表示为连接该节点的所有边的权重之和。因此通话社交网络的点权为：

$$s_i=\sum_j w_{ij} \tag{4}$$

它描述了节点在整个通话社交网络中的重要程度。节点的点权越大，说明该节点的地位越重要，信息越易传播。

3 通话网络模型社团划分

3.1 算法描述

本算法是一种基于贪婪算法的凝聚算法，采用对 CNM 算法[5]引入网络权重的方法将其进行改进，算法的时间复杂度为 $O(n\log_2 n)$，接近线性复杂性。该算法步骤如下：

(1)初始化网络为 n 个社团，即每个节点就是一个独立的社团。初始的模块度 Q 满足：$Q=0$。初始的 $\boldsymbol{e}_{ij}$ 及辅助向量 $\boldsymbol{\alpha}_i$ 满足：

$$\boldsymbol{e}_{ij}=\begin{cases}\dfrac{\boldsymbol{w}_{ij}}{2m}, & \text{若节点 } i \text{ 与 } j \text{ 相连} \\ 0, & \text{若节点 } i \text{ 与 } j \text{ 不相连}\end{cases} \tag{5}$$

$$\boldsymbol{\alpha}_i=\frac{s_i}{2m} \tag{6}$$

其中 s_i 为节点 i 的点权，m 为网络总边权。

模块度增量矩阵 $\Delta\boldsymbol{Q}$ 与网络的连接矩阵 $\boldsymbol{A}$ 一样是一个稀疏矩阵。将它的每一行都存为一个平衡二叉树

(可在 O(logn)时间内找出所需某个元素)以及一个最大堆。初始化模块度增量矩阵元素满足:

$$\Delta Q_{ij}=\begin{cases}e_{ij}-\alpha_{ij}, & \text{若节点 } i \text{ 与 } j \text{ 相连}\\ 0, & \text{若节点 } i \text{ 与 } j \text{ 不相连}\end{cases} \tag{7}$$

由 ΔQ 中每一行的最大元素及该元素的相应两个社团的编号 i、j 构成最大堆 H。

(2)从最大堆 H 中选择最大的 ΔQ_{ij},合并相应的社团。更新模块度增量矩阵 ΔQ、最大堆 H 和辅助向量 α_i。通过 ΔQ_{ij} 值,更新模块度 Q。

(3)重复步骤(2),直到网络中所有的节点都归到一个社团。

以上存在的数据结构使得在步骤(2)中可快速更新其中的元素。标记符合要求的一部分 ΔQ_{ij} 元素,合并 i、j 社团并标记合并后的社团 j。更新第 j 行和第 j 列,同时删除第 i 行和第 i 列。更新方法如下。

首先判断社团 k 与社团 i、j 的连接状态。如果社团 k 同时与社团 i 和 j 相连,则:

$$\Delta Q'_{jk}=\Delta Q_{ik}+\Delta Q_{jk} \tag{8}$$

如果社团 k 仅与社团 i 相连而不与社团 j 相连,则:

$$\Delta Q'_{jk}=\Delta Q_{ik}-2\alpha_j\alpha_k \tag{9}$$

如果社团 k 仅与社团 j 相连而不与社团 i 相连,则:

$$\Delta Q'_{jk}=\Delta Q_{jk}-2\alpha_i\alpha_k \tag{10}$$

然后,更新最大堆 H。每次更新 ΔQ_{ij} 之后,更新 H 中相应的行与列的最大元素。

最后,更新辅助向量 α_i,同时记录合并之后的模块度值。

$$\alpha'_j=\alpha_i+\alpha_j;\alpha'_i=0 \tag{11}$$

3.2 划分结果

经计算得到,通话网络在 287 576 步模块度值 Q=0.786 148 达到极值,此时网络有最优的社团分解。此社团分解成 15 178 个社团,最大的社团由 18 016 个节点组成,占全部用户的 7.6%。模块度 Q(Q<1)是衡量算法分解好坏的标准,值越大分解的社团越精确。从实验中发现,最大模块度值 Q=0.786 148,与 1 较为接近,此时社团具备最优分解。

社团平均大小为 23,社团大小服从幂指数为 2.5 的幂律分布,如图 2 所示。出现这种分布的原因推测是由于通话社交网络的度呈幂律分布或者社团结构划分算法的动力学性质引起。

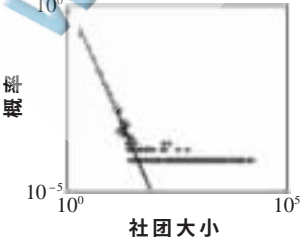

图 2 社团大小分布

3.3 算法复杂度比较

随着网络节点数的不断增加,考虑社团分解正确性、可靠性的前提下必须考虑算法复杂度。下图将改进的 CNM 算法与 GN 算法、Newman 快速算法在不同节点数的时间复杂度进行对比。

图 3 所示为 3 种算法在时间复杂度方面的对比。GN 算法相对比较耗时,仅适用于中等规模的网络;Newman 快速算法和 CNM 算法则可以分析节点数上万的大规模复杂网络,但由于改进的 CNM 算法同样采用了堆数据结构,与 Newman 快速算法相比计算速度有很大提高。

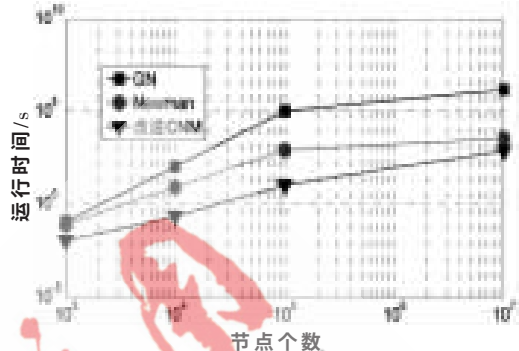

图 3 时间复杂度对比

本文基于某地的通话数据记录建立了一个大型的社交网络,并将模糊综合评判方法合理地应用于所构建的通话社交网络边权计算及评价中。为进一步分析社会网络的演化打下坚实的基础。

参考文献

[1] ONNELA J P, SARAMÄKI J, HYVÖNEN J, et al. Analysis of a large-scale weighted network of one-to-one human communication[J]. New Journal of Physics, 2007, 9(6): 179-184.

[2] ONNELA J P, SARAMÄKI J, HYVÖNEN J, et al. Structure and tie strengths in mobile communication networks[J]. PNAS, 2007, 104(18): 7332-7336.

[3] SZABÓG, BARABÁSI A. Preprint physics[S]. 2006.

[4] PALLA G, BARABASI A L, VICSEK T. Quantifying social groupe volution[J]. Nature, 2007(446): 664-667.

[5] 解冶, 汪小帆. 复杂网络中的社团结构分解算法研究综述[J]. 复杂系统与复杂性科学, 2005(3): 12.

[6] 骆志刚, 丁凡, 蒋小舟, 等. 复杂网络社团发现算法研究新进展[J]. 国防科技大学学报, 2011, 33(1): 47-52.

[7] 田柳, 迪增加, 姚虹. 权重分布对加权网络效率的影响[J]. 物理学报, 2011, 60(2): 1-5.

[8] 姚尊强, 尚可可, 许小可. 加权网络常用统计量[J]. 上海理工大学学报, 2012, 34(1): 18-26.

[9] 覃森, 戴冠中, 王林, 等. 不同权重定义下的静态与动态加权网络的比较分析[J]. 西北工业大学学报, 2007, 25(5): 672-676.

[10] 杨永萍, 李宝栋, 常文春. 基于模糊综合评判方法的研究及应用[J]. 兰州工业高等专业学报, 2006, 13(3): 49-52.

(收稿日期:2012-10-30)

作者简介:

王林,男,1963 年生,教授,主要研究方向:复杂系统与复杂网络,无线传感器网络以及计算机应用。

童昭维,女,1989 年生,研究生,主要研究方向:复杂网络。