文章编号:1672-3813(2015)02-0072-06;DOI:10.13306/j.1672-3813.2015.02.011

基于社交网络的社群生长模型

尤志强[1],管远盼[1],韩筱璞[1],邓小方[2],吕琳媛[1]
(1.杭州师范大学阿里巴巴复杂科学研究中心,杭州 311121;2.江西师范大学软件学院,南昌 330022)

摘要:基于腾讯 QQ 朋友网络数据,针对实际的用户结群行为和社会群组生长过程,提出一种基于共同兴趣的类渗流的扩散机制,并进行建模和分析。在腾讯 QQ 朋友关系网络上的数值模拟实验显示,模型得到的统计特征与真实的社群结构基本一致,表明这一机制是实际社群生长的重要驱动力。研究为进一步对社群生长趋势预测的研究提供了重要的理论支持。

关键词:社交网络;共同兴趣;渗流机制;兴趣扩散;结群行为;社群生长
中图分类号:N94 **文献标识码**:A

Modeling of Social Group Growth Based on Social Networks

YOU Zhiqiang, GUAN Yuanpan, HAN Xiaopu, DENG Xiaofang, LYU Linyuan
(1. Alibaba Research Center for Complexity Sciences, Hangzhou Normal University, Hangzhou 311121, China;
2. Software school, Jiangxi Normal University, Nanchang 330022, China)

Abstract: The structure of social group deeply influences the development and evolution of human society, but studies on this subject are relatively rare. Focusing on QQ friendship network, this paper proposes a percolation-like diffusion model which is based on users' common interest to simulate and analyze the clustering behaviors of users and the growing process of social groups. Numerical simulation on the real QQ friendship network of Tencent shows that the statistical features generated by our model accord with the real empirical properties of the group network. It indicates that this mechanism is an important driven-factor for the growth of real social group. This work provides vital theoretical evidence for the further studies on the prediction of social group growth.

Key words: social network; common interest; percolation; interest diffusion; clustering behavior; group growth

0 引言

社会网络结构特性研究一直是复杂网络学界的关注重点之一。在日常社交中,由于“物以类聚,人以群分”,往往存在结群行为,它对社会网络的结构特性和动力学效应有着深刻的影响。

一般而言,这种结群行为大致可以区分为两种类型。第一种类型为无特定形态的结群行为,它纯粹由社交过程中个体的社交偏好自发驱动,所形成的社群结构一般没有清晰的边界和明确的标签,是一种隐性结构,往往需要采用社区划分算法进行有效识别,例如在某个社会群体内部自发形成的一些亚群体等。另一种类型,则是在清晰的社群标签下进行,例如特定的社会组织与机构、共同喜好等,人群在这些标签指引之下进行社群化的交往,它

收稿日期:2014-10-16;修回日期:2015-01-16
基金项目:浙江省新苗人才计划项目(2013R421062),CCF-腾讯科研基金(CCF-Tecent AGR20130104);国家自然科学基金(11205040,11205042)
作者简介:尤志强(1990-),男,浙江金华人,硕士研究生,主要研究方向为复杂网络和数据挖掘。
通讯作者:韩筱璞(1981-),男,山东曹县人,博士,讲师,主要研究方向为复杂系统与人类动力学。

所形成的社群是显性的，一般有着明确的社群边界。本文把这类有着明确边界的社群称为“群结构”。

对于第一种类型的结群行为，当前已经有了相当深入的研究。社团结构[1-2]研究就是针对该类结群行为形成的社群结构。定性来说，社团结构表示的是团簇结构内节点之间的连边密度要远远大于这些节点与该团簇以外的节点的连边密度[3]，目前普遍被物理学界接受的定量表达则是基于 Newman 所提出的模块概念[4]。社团结构的理解以及性质挖掘，对于进一步推动社交网络的形成与演化、信息传播、重要节点识别、广告投放、舆论控制等研究及应用都具有相当重要的作用。相关学者通过实证分析，不仅发现万维网[5]、生物化学网络[6]等都存在社团结构，而且进一步发现社团结构存在自相似现象[7]等。一些学者则在实证基础上，提出一系列的社团结构识别算法，如 Kernighan-Lin 算法[8]、谱图分区算法[9]、模块最大化算法[10]等。另外，一些学者对团簇的形成机制做了相关研究，如 Andreas 等[11]利用脱离者模型研究社交网络的社团产生机制。Wang 等[12]则提出基于核心的算法来研究社团的演化。

然而，对于第二类结群行为，长期以来由于相关的社会网络数据的匮乏，此方面研究非常欠缺。目前，对于此种群结构的产生机制的认知的缺乏以及数据获取的难度，导致很多研究只能基于模拟数据进行，相关算法缺乏实际数据的有效检验。不过，随着在线社交网络的普及，数据的获取问题得到了缓解，如 MySpace、QQ、facebook 等社交网站拥有大量的活跃用户，这为研究真实社交网络的社群结构性质，探究网络演化、社群生长机制提供了机会。

本文针对第二类结群行为，获取了真实的 QQ 社交网络数据，对以 QQ 群为代表的社会群组结构的产生及生长机制进行了模型研究。基于对真实 QQ 群数据的实证研究[13]，提出了一种类渗流过程的兴趣扩散模型，并发现这种机制可以有效解释群组社交网络上群结构的产生以及生长过程，为理解社会群组结构的生长与演化提供了新思路。

1　模型

由于所研究的 QQ 社交网络数据是显性群结构，超图结构很自然地被用于分析该网络。文献[13]通过对该数据的实证分析发现群的产生主要是基于两种机制，一种是由线下社交关系的交互驱动产生，此类群的大小会在短时间内达到稳定；另一种则是基于共同爱好自发产生并逐渐演化，此类群中成员的社交关系与线下不一定对应。在本模型中，如果用户与好友处于同一地域、组织机构，或者具有相同的爱好，都视其拥有共同兴趣，若相邻朋友之间存在共同兴趣，他们就可能会加入同一个群。同时，还需要考虑相关结群限制因素：1）用户的有限精力，用户不可能同时与大量的直接邻居朋友产生结群行为；2）用户加群倾向，有些用户偏向于加入多个群；3）群规模的限制。

本模型中，每个用户都被赋予一个 N 维的二值随机兴趣向量 $\boldsymbol{H}=(h_1,h_2,\cdots,h_N)^{\mathrm{T}}$，$h_i=0$ 或 1，$i=1,2,\cdots N$。$\boldsymbol{H}$ 的每个维度表征用户在该方面的爱好，1 表示用户具有该维度兴趣，否则该维度值为 0。本文的计算中，固定 $N=10$。每一个用户具有一个兴趣构建概率 $P_v\in(0,\alpha]$，$0<\alpha\leqslant 1$，用于构建该用户的随机兴趣向量。考虑到不同用户在兴趣广度上存在差异，因此兴趣构建概率的大小对于不同用户存在一定差别。兴趣广泛的用户会有多维兴趣，这类用户的兴趣构建概率 P_v 值往往趋近于 α。另外，由于在极端情况下，有的用户的兴趣向量的所有维度值均为 0，此类情况下规定该用户将被随机赋予某一维兴趣，即随机将向量的某一维度赋值为 1，其余维度为 0。

此外，每个用户都有权利基于自身的兴趣点创建群，即 1 个用户有 n 个兴趣点则可建 n 个群。同时考虑到用户加群偏好的差异性，部分用户可能加多个群而其他用户倾向于加入少量群，因此引入加群偏好概率 $P_{\mathrm{join}}\in(0,\beta]$，$0<\beta\leqslant 1$，它代表相应用户加群的倾向性。$P_{\mathrm{join}}$的值越大表示用户自身具有更强的加群倾向。用户在模型初始阶段被赋予随机概率值 P_{join}。在模型中，社群生长依据朋友关系网络路径进行，但如图 2 所示，朋友关系网的度分布遵循 3 段式幂律分布，少部分用户有大量的邻居朋友（其数目可以超过 1 000）而大部分用户只有少量的邻居朋友（其数目甚至小于 10）。考虑到用户的精力有限，不可能同时与大量邻居保持结群行为互动，进一步基于 Dunbars 数[14]的考量，在社群生长过程中引入基于直接邻居数的衰减函数，以保证社群生长的平滑性，也进一步使模型贴近实际，否则会出现群大小分布的断层现象。Dunbars 数是指某个用户能维持紧密人际关系的人数上限：当拥有较小相邻朋友数时，用户可以与所有邻居维持良好的社交联系；然而，随着邻居数的增大，由于时间和精力的限制，用户仅能与一定数量的朋友保持较好的交往。该衰减函数定义为“漂移幂律”形式[15]：

$$f_d=c(x+K)^{-a} \tag{1}$$

其中，c 为常量，K 和 a 为可调参数，K 为对偏移幂律分布的测度，K 越小，函数越接近幂律分布，当 K 为 0 时，即为幂律分布。x 为相邻朋友数，当 x 较小时，此种情形下，用户几乎可以和其所有直接邻居朋友有结群交互行为。然而，当 x 超过某一阈值时，f_d 会按幂律衰减。在本文中，固定 $c=100$，$K=39$，$a=1.4$。

同时，在现实中，一些有相同偏好的用户往往同时会加入多个群，这可能是由于某些用户倾向于邀请其朋友加入若干群，这种情形通常发生在日常生活中有着紧密的线下关系的朋友之间。同一批用户同时加入多个群，会使得群与群之间的重叠用户数变大。因此，在模型中，为每一个用户定义了其协同加群概率 $P_c \in (0,\gamma]$，$0<\gamma\leqslant 1$，用以描述具有协同加群倾向的相应用户请求其他用户加入他所创建群的可能性。基于此，当需要判定用户 u_1 是否加入由用户 u_2 创建的群时，用户 u_1 的加群概率 P_{join} 的最终形式为

$$P'_{\text{join}} = \begin{cases} P_{\text{join}} + P_c, P_{\text{join}} + P_c \leqslant 1 \\ 1, P_{\text{join}} + P_c > 1 \end{cases} \tag{2}$$

其中，P_{join} 为用户 u_1 的原始加群概率，P_c 为用户 u_2 的协同加群概率，而 P'_{join} 为 u_1 的修正加群概率。P_c 越大，则表示 u_2 的邻居加入同样由 u_2 创建的群的概率越大。

2 模拟结果

基于以上讨论，本文在真实的 QQ 用户朋友关系网络进行该类渗流机制兴趣扩散模型的实验，模拟用户结群及社群生长行为，分析研究模型所得到的结果，以揭示社群生长的基本机制。该 QQ 朋友关系数据共包含1 052 129个用户节点和 8 022 535 条朋友关系边。该数据集抽取于同一个城市的用户数据，其平均簇系数和平均距离分别为 0.609 和 4.167。

图 1 描述了以用户 u_1 创建的群 G 在朋友关系网以及兴趣扩散规则基础上的生长过程。首先用户 u_1 基于第 i 维兴趣创建群 G，即表示该群中的成员用户必然具有第 i 维兴趣，该维度兴趣称为 G 的群兴趣，此时群 G 的用户集为 $S=\{u_1\}$。基于如前所述模型规则，群 G 的生长迭代步骤可以表述为：1)依据扩散邻居数衰减函数 f_d，该群首先向用户 u_1 的直接邻居进行扩张，对每一个邻居产生随机概率 $P_d \in [0,1)$，若 $P_d<f_d$，则将其加入初始用户集 S_1，本例中通过该处理得到初始用户集 $S_1=\{u_3,u_4,u_7,u_{11}\}$；2)在用户集 S_1 中，选取具有第 i 维兴趣的用户构建候选结群用户集 S_2，即 S_2 中的用户需要具有第 i 维兴趣；3)对 S_2 中用户逐个进行加群概率修正，然后对 S_2 中每个用户进行加群判定，即产生随机数 $P_r \in [0,1)$，如果满足 $P_r<P'_{\text{join}}$，则对应用户加入群 G，否则不加入，这里用户 u_{11} 虽然具有第 i 维兴趣，但仍以一定的概率选择不加入该群；4)当 u_1 的直接邻居处理完毕，此时该群当前用户集为 $S=\{u_1,u_3,u_4,u_7\}$，然后基于该批新加群的用户对群 G 进行进一步扩展。如图 1 中的用户 u_3，下一步需要考虑其直接邻居 u_2、u_5，同理考虑 u_4、u_7 的直接邻居。重复以上步骤，直至达到扩展边界为止。这里的扩展边界指：1)群规模达到上限 2 000；2)没有进一步可扩展的候选朋友集。另外，模型运行结果中，如果出现多个群同时具有相同的群兴趣以及同一批群成员，则进行去重处理，仅保留唯一的一个群，因为同一批用户基于同样的兴趣没有必要建立多个群。

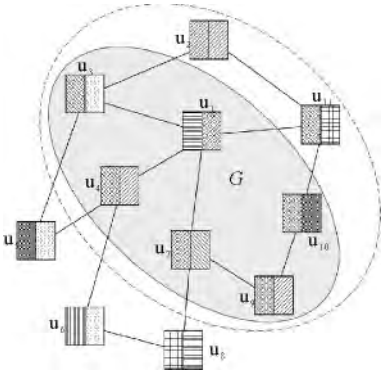


图 1 运用兴趣扩散模型基于朋友关系网的群生长过程的示意图

Fig. 1 Illustration of group growing process based on friend-network with the interest diffusion model

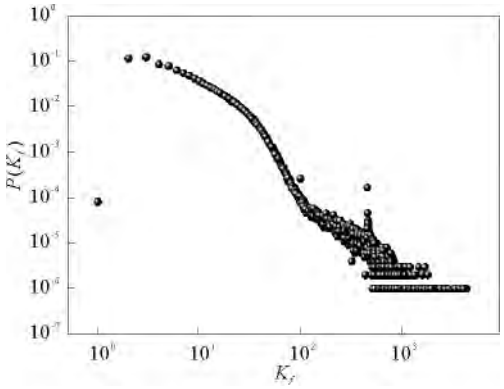
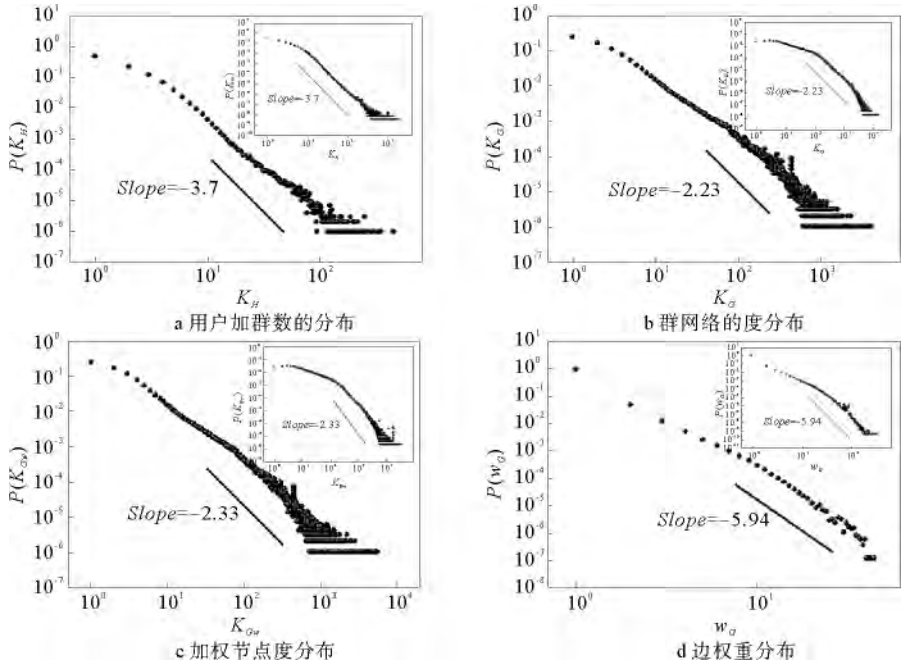


图 2 实际的 QQ 朋友关系网络的度分布

Fig. 2 The degree distribution of real QQ friend-network

图 1 中正方形小方块表示模型中 QQ 用户节点，节点之间的连边表示用户之间的朋友关系。该示意图中每个用户节点的兴趣向量中都只有两个维度的值为 1，其余为 0，使用不同图案的矩形框表示不同的兴趣。浅色底椭圆表示群 G，该群是用户 u_1 基于其右侧部分的第 i 维兴趣创建的群。图 1 中的虚线圈表示群 G 扩张的边界。

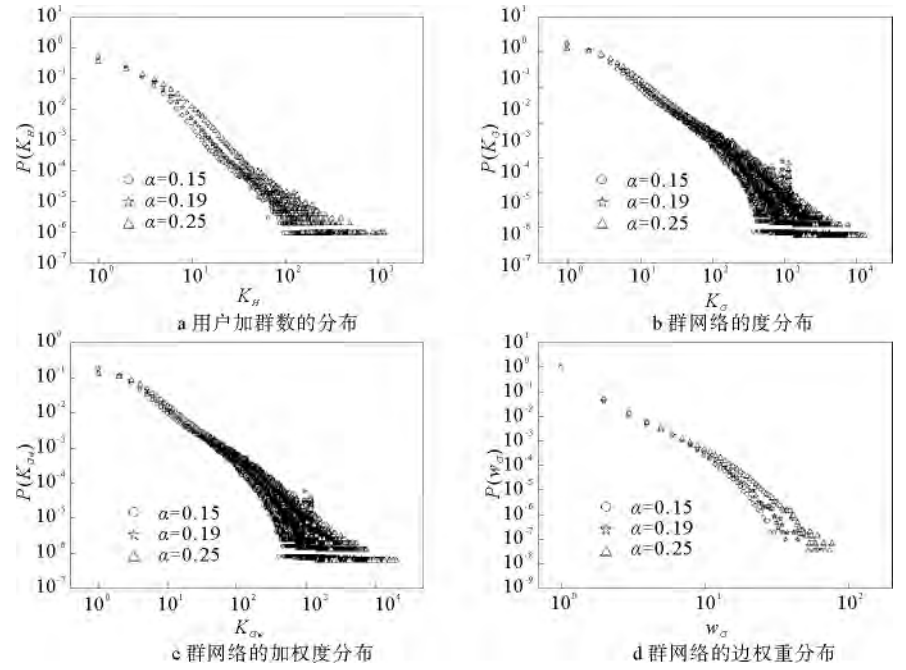
为了验证该模型的有效性，需要比较模型模拟得到的群关联网络与实际的群关联网络的统计特性差异[13]。在这里，首先考察了 4 种分布性质：1）用户加群数分布 $p(k_H)$，其中用户加群数使用符号 k_H 来表征，表示一个用户所加的群总数；2）节点度分布 $p(k_G)$，其中节点度使用符号 K_G 表示，也就是与该节点（即群）相连接的群数量；3）加权节点度分布 $p(k_{GW})$，其中加权节点度使用符号 k_{GW} 表示，反映一个群与其他所有群的共同成员总数；4）边权重分布 $p(w_G)$，其中边权重使用符号 w_G 表示，即两个群之间的共同成员数目。如图 3 所示，4 图分别展示了模型数据与实际数据在用户加群数、节点度、加权节点度以及边权重4类分布比较，可以看到模型得到的用户加群



大坐标系内散点表示模型结果，小坐标系内散点表示实证结果；模型参数 α,β,γ 分别为 0.19，0.195，1。

图 3　模型结果与实证结果在 4 种统计指标上的对比

Fig. 3　The comparison between model results and the empirical statistics on four metrics

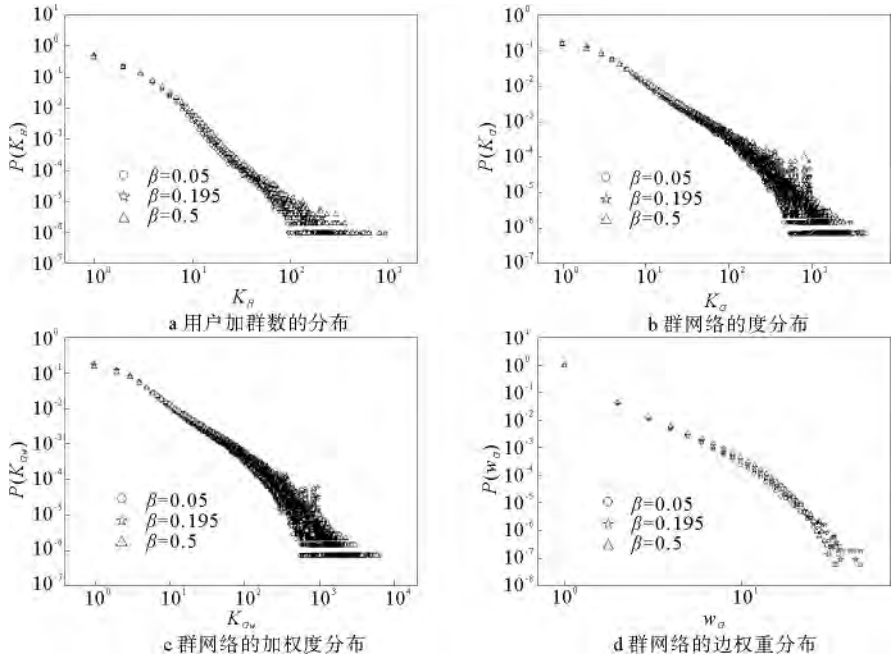


模型参数 β,γ 分别为 0.195，1。

图 4　不同兴趣构建概率上限下的模型结果

Fig. 4　Performance of different values of the limit of interest vector construction probability α

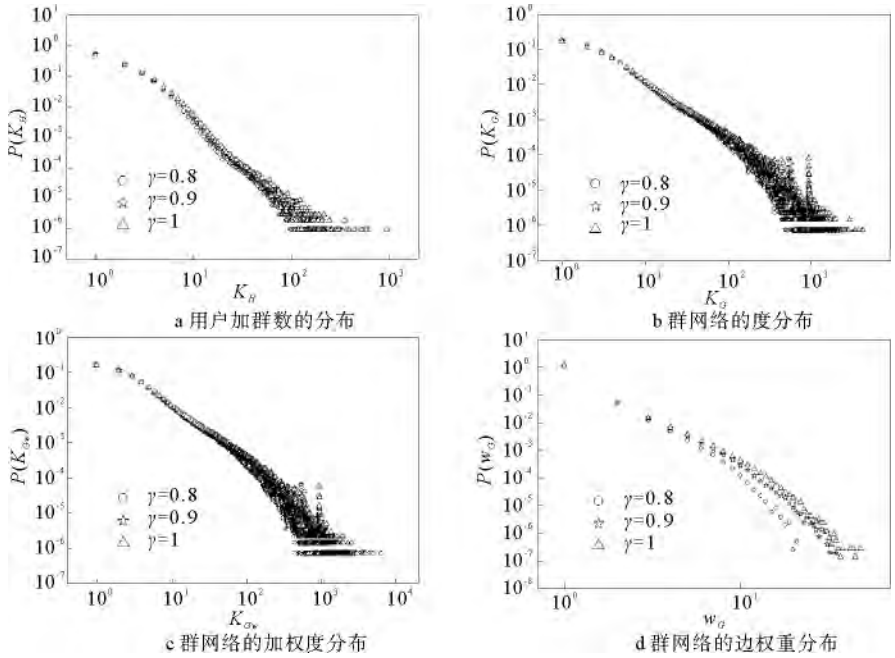
数分布以及群网络在节点度、加权节点度以及边权重方面呈现出的幂律分布与实际数据是一致的。在群网络度和加权度分布的头部，模型数据中小度群的比例与实证数据相比偏高，引起这种偏差的原因可能是模型未考虑活跃用户的退群行为以及群的人为解散行为，当群中活跃用户数较少时，维持一个群的必要性会降低，往往会产生大量用户退群或者群解散的行为。总体来说，模型所得到的结果与实证数据基本吻合，表明所提出的基于渗流机制的兴趣驱动的群生长机制可以有效描述实际的社群生长的基本机制。



模型参数 α,γ 分别为 0.19,1。

图 5 不同加群概率上限下的模型结果

Fig. 5 Performance of different values of the limit of Group-join probability β



模型参数 α,β 分别为 0.19,0.195。

图 6 不同协同加群概率上限下的模型结果

Fig. 6 Performance of different values of the limit of collaborative group-join probabilityγ

此外，对模型中主要参数的影响进行了研究，分别是：用户兴趣构建概率上限 α（如图 4）、用户加群概率上限 β（如图 5）以及用户协同加群概率上限 γ（如图 6）。我们发现，α,β,γ 的概率上限都会对用户加群数分布产生影响，

随着该上限值的提高，加群数分布头部出现降低趋势，而其他部分出现明显右偏，其中 α 造成的右偏现象最为明显，这主要是由于 α 的提高，使得用户平均兴趣点数量增加，直接造成用户可以对更多的群产生结群行为。此外，α、β、γ 对群权重网络的节点度分布与加群节点度分布产生一定影响，随着上限值的提高，大度节点以及大的加权度节点明显增多。另外，我们关注到 α 和 γ 对边权重分布结果的影响以及范围要强于 β 产生的影响。随着 α 和 γ 上限值的增加，网络中边权重大的边的比例明显增加。

3　结论

该模型是根据实证分析所发现的两类用户加群动机，基于好友共同兴趣的类渗流的扩散过程来揭示群组社交网络上社群的生长机制。在该种机制下，朋友关系网络上基于共同兴趣的连通簇，就成为了相应群组生长的边界。通过数值模拟，这一简单的模型可以产生出一系列与实际社群关系网络基本一致的统计特性，暗示这种基于共同兴趣的社群生长机制在实际的社群生长中扮演着重要角色。同时，这一机制的揭示，对于社群涌现和生长的预测，也有着相对重要的意义，例如可以通过对共同兴趣的判断来进行社群推荐或者预知潜在的社群关系等。总而言之，我们通过这样一个简单的演化模型，有效揭示出了社会群组生长过程的基本驱动力。

参考文献：

[1] Fortunato S. Community detection in graphs[J]. Physics Reports, 2010, 486(3): 75-174.

[2] Newman M E J. Detecting community structure in networks[J]. The European Physical Journal B, 2004, 38(2): 321-330.

[3] Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(9): 2658-2663.

[4] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical review E, 2004, 69(2): 026113.

[5] Eckmann J P, Moses E. Curvature of co-links uncovers hidden thematic layers in the world wide web[J]. Proceedings of the national academy of sciences, 2002, 99(9): 5825-5829.

[6] Holme P, Huss M, Jeong H. Subnetwork hierarchies of biochemical pathways[J]. Bioinformatics, 2003, 19(4): 532-538.

[7] Arenas A, Danon L, Diaz-Guilera A, et al. Community analysis in social networks[J]. The European Physical Journal B, 2004, 38(2): 373-380.

[8] Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs[J]. Bell system technical journal, 1970, 49(2): 291-307.

[9] Fiedler M. Algebraic connectivity of graphs[J]. Czechoslovak Mathematical Journal, 1973, 23(2): 298-305.

[10] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Physical review E, 2004, 69(6): 066133.

[11] Grönlund A, Holme P. Networking the seceder model: Group formation in social and economic systems[J]. Physical Review E, 2004,70(3): 036108.

[12] Wang Y, Wu B, Du N. Community evolution of social network: feature, algorithm and model[DB/OL]. [2008-04-28]. http://arxiv.org/pdf/0804.4356vl.pdf.

[13] You Z Q, Han X P, Lü L, et al. Empirical studies on the network of social groups: the case of Tencent QQ[DB/OL]. [2014-08-24]. http://arxiv.org/pdf/1408.5558vl.pdf.

[14] Dunbar R. How Many Friends Does One Person Need?Dunbar's Number and Other Evolutionary Quirks[M]. Cambridge, MA, USA: Harvard University press, 2010:21-34.

[15] Plate, Erich J. Methods of investigating urban wind fields-physical models[J]. Atmospheric Environment, 1999, 33(24): 3981-3989.

（责任编辑　耿金花）