

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2021 年 9 月 23 日 (23.09.2021)



(10) 国际公布号
WO 2021/184367 A1

(51) 国际专利分类号:
G06F 9/44 (2018.01)

(HUANG, Bingyang); 中国北京市海淀区清华
园一号, Beijing 100084 (CN)。

(21) 国际申请号: PCT/CN2020/080472

(74) 代理人: 北京中强智尚知识产权代理有限公司 (BEIJING STRONG WISDOM INTELLECTUAL PROPERTY AGENCY LTD.); 中国北京市朝阳区北苑东路 19 号院 2 号楼 2608 室, Beijing 100012 (CN)。

(22) 国际申请日: 2020 年 3 月 20 日 (20.03.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

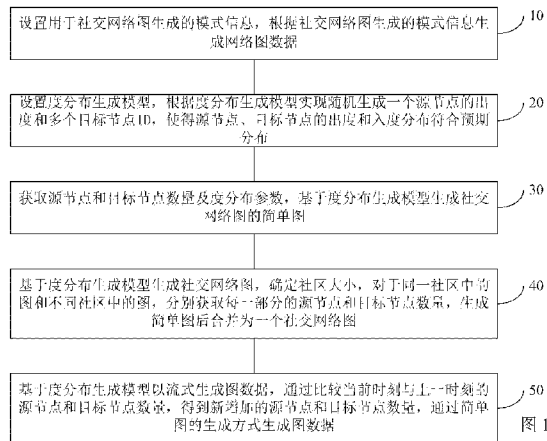
(71) 申请人: 清华大学(TSINGHUA UNIVERSITY) [CN/
CN]; 中国北京市海淀区清华园一号,
Beijing 100084 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家
保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG,
BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU,
CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB,
GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS,
JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,
LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX,
MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,

(72) 发明人: 王朝坤 (WANG, Chaokun); 中国北京
市海淀区清华园一号, Beijing 100084 (CN)。
王彬彬 (WANG, Binbin); 中国北京市海淀区
清华园一号, Beijing 100084 (CN)。 黄冰洋

(54) Title: SOCIAL NETWORK GRAPH GENERATION METHOD BASED ON DEGREE DISTRIBUTION GENERATION MODEL

(54) 发明名称: 基于度分布生成模型的社交网络图生成方法



- 10 Set mode information used for social network graph generation, and according to the mode information used for social network graph generation, generate network graph data
- 20 Set a degree distribution generation model, and according to the degree distribution generation model, randomly generate an out-degree of one source node and a plurality of target node IDs, such that the out-degree and in-degree distribution of the source node and target nodes is in accordance with an expected distribution
- 30 Acquire the number of source nodes and target nodes and degree distribution parameters, and on the basis of the degree distribution generation model, generate a simple graph of a social network graph
- 40 On the basis of the degree distribution generation model, generate the social network graph, determine the size of a community, and for graphs in the same community and graphs in different communities, respectively acquire the number of source nodes and target nodes of each part, and perform combination to form a social network graph after the simple graph is generated
- 50 On the basis of the degree distribution generation model, generate graph data in a streaming manner, obtain the number of newly added source nodes and newly added target nodes by comparing the number of source nodes and target nodes at the current moment with that at the previous moment, and generate the graph data in a simple graph generation manner

(57) Abstract: A social network graph generation method based on a degree distribution generation model. The method comprises: setting mode information used for social network graph generation, and according to the mode information used for social network graph generation, generating network graph data; setting a degree distribution generation model, and according to the degree distribution generation model, randomly generating an out-degree of one source node and a plurality of target node IDs, such that the out-degree and in-degree distribution of the source node and target nodes is in accordance with an expected distribution; on the basis of the degree

PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US,
UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

根据细则4.17的声明:

- 关于申请人有权申请并被授予专利(细则
4.17(ii))
- 发明人资格(细则4.17(iv))

本国际公布:

- 包括国际检索报告(条约第21条(3))。

distribution generation model, generating a simple graph of a social network graph; on the basis of the degree distribution generation model, generating the social network graph; and on the basis of the degree distribution generation model, generating graph data in a streaming manner. By means of the method, an out-degree can be efficiently determined for each node, and a corresponding number of target nodes can be determined; a generated graph node can represent a user in an actual network; an edge of a generated graph can represent the interaction relationship between users in a network; and the generated graph has the characteristics of a real social network and can serve as input data of various network analysis tasks.

(57) 摘要: 一种基于度分布生成模型的社交网络图生成方法, 包括: 设置用于社交网络图生成的模式信息, 根据社交网络图生成的模式信息生成网络图数据; 设置度分布生成模型, 根据度分布生成模型实现随机生成一个源节点的出度和多个目标节点ID, 使得源节点、目标节点的出度和入度分布符合预期分布; 基于度分布生成模型生成社交网络图的简单图; 基于度分布生成模型生成社交网络图; 基于度分布生成模型以流式生成图数据。上述方法可以高效地为每个节点确定出度, 并确定相应数量的目标节点, 生成的图节点可以表示实际网络中的用户, 生成图的边可以表示网络中用户的交互关系, 生成的图具有真实社交网络的特点, 能够作为各种网络分析任务的输入数据。

基于度分布生成模型的社交网络图生成方法

技术领域

本公开涉及计算机技术领域，尤其涉及一种基于度分布生成模型的社交网络图生成方法。

背景技术

社交网络图生成器旨在生成尽可能类似于真实社交网络的图数据。随着许多社交网络平台的普及，近年来，涌现了许多社交网络分析任务，如社区发现，社区搜索和网络表示学习。为了评估这些社交网络分析任务的算法的性能和可扩展性，真实网络数据和合成数据都是必要的。因此，社交网络图生成器变得越来越重要，特别是考虑到不同的算法关注于网络图的不同特征。

例如，使用层次聚类或矩阵分块技术的社区发现算法处理同构图，即网络图中只有一种类型的节点和边；而一些社区发现算法处理异构图，图中有多种标签的节点和多种类型的边。此外，真实网络中的社区可以分为重叠社区和非重叠社区，许多社交应用产生的图数据呈指数增长。

然而，现有的图数据生成器不能满足以上所有要求。模式驱动的方法能够使生成器有广泛应用场景，为不同的应用生成图数据。这类方法根据设计良好的模式生成数据，用户可以在模式信息中指定合成图的常见特征，例如节点标签和边的标签等。然而，这类方法不是专门为社交网络图设计的，不支持在合成图中包含社区结构，而且效率低，不适合生成大规模图数据。

LFR 是一种广泛使用的用于生成社交网络图的工具，根据节点与同一社区中节点之间的边与该节点的总度数的比例构建社区。节点入度分布符合幂律分布，而节点出度并不符合幂律分布。此外，LFR 在构建社区时的计算开销很高，因此不适合生成大规模图数据。

目前，已提出了一些用来生成大规模图数据的方法。其中应用最广泛的是 R-MAT 和 Kronecker。R-MAT 模型使用递归矩阵模型递归地选择邻接矩阵的一个象限，直到选择一个单元格，从而生成一条边。Kronecker 有随机 Kronecker 图 (Stochastic Kronecker Graph, 记为 SKG) 和确定的 Kronecker 图 (Deterministic Kronecker Graph, 记为 DKG) 两种模型，广泛应用的 SKG 模型是递归矩阵模型关于概率参数数量上的一个广义模型。R-MAT 的空间复杂度是 $O(|E|)$ ，Kronecker 的时间复杂度为 $O(|V|^2)$ 。TrillionG 提出了一种新的递归向量模型用于克服这两个模型很高的时空复杂度，可以高效生成万亿规模的图。然而，

TrillionG 生成的图没有社区结构，此外，如何设置噪声参数使得节点度数符合幂律分布并不明确。

因此，现有技术需要改进。

上述背景技术内容仅用于帮助理解本公开，而并不代表承认或认可所提及的任何内容

5 属于相对于本公开的公知常识的一部分。

发明内容

为解决上述技术问题，本公开提出了一种基于度分布生成模型的社交网络图生成方法。

基于本公开实施例的一个方面，公开一种基于度分布生成模型的社交网络图生成方法，

10 包括：

设置用于社交网络图生成的模式信息，根据社交网络图生成的模式信息生成网络图数据；

设置度分布生成模型，根据度分布生成模型实现随机生成一个源节点的出度和多个目标节点 ID，使得源节点、目标节点的出度和入度分布符合预期分布；

15 获取源节点和目标节点数量及度分布参数，基于度分布生成模型生成社交网络图的简单图；

基于度分布生成模型生成社交网络图，确定社区大小，对于同一社区中的图 and 不同社区中的图，分别获取每一部分的源节点和目标节点数量，生成简单图后合并为一个社交网络图；

20 基于度分布生成模型以流式生成图数据，通过比较当前时刻与上一时刻的源节点和目标节点数量，得到新增加的源节点和目标节点数量，通过简单图的生成方式生成图数据。

基于本公开的基于度分布生成模型的社交网络图生成方法的另一个实施例中，所述设置用于社交网络图生成的模式信息，根据社交网络图生成的模式信息生成网络图数据包括：

设置节点模式信息、设置边模式信息、设置社区模式信息、设置社交网络图模式信息、

25 设置用于生成社交网络图的符号表示；

所述节点模式信息包括： $VS = (lbl, amount, attr)$ ，其中 lbl 是节点标签，amount 是标签为 lbl 的节点数量，attr 是节点属性信息，是一个 key-value 结构的字典，key 为属性名，value 为属性取值；

30 所述边模式信息包括： $ES = (lbl, lbl_s, lbl_t, amount, distr_{in}, distr_{out}, attr)$ ，其中 lbl 是边的标签信息， lbl_s 和 lbl_t 分别为源节点和目标节点的标签，amount 为预期生成的边的数量，

distr_{in} 和 $\text{distr}_{\text{out}}$ 分别为目标节点的入度分布和源节点的出度分布, attr 为边的属性信息, 同样是一个 key-value 结构的字典;

所述社区模式信息包括: $\text{CS} = (\text{lbl}_e, \text{amount}, \lambda_s, \lambda_t, \rho)$, 表示在标签为 lbl_e 的边构成的图上构建社区, amount 为生成的社区数量, 社区大小服从幂律分布, λ_s 和 λ_t 分别为所有社区中源节点和目标节点数量服从幂律分布的参数, ρ 为社区分离参数;

所述社交网络图模式信息包括: $\text{SGS} = (\text{VSS}, \text{ESS}, \text{CSS})$, 其中 VSS 为节点模式的集合, ESS 为边的模式集合, CSS 为社区模式的集合;

所述设置用于生成的社交网络图, 其符号表示包括:

设置有向异构图 $G=(V, E)$, 其中 V 为节点集合, $E \subseteq V \times V$ 为边集;

一个节点 $v \in V$ 为一个三元组 $(\text{ID}, \text{lbl}, \text{attr})$, 其中 lbl 为节点标签, attr 为节点属性, ID 为一个节点的唯一标识;

一条边 $e \in E$ 以为一个四元组 $(v_s, v_t, \text{lbl}, \text{attr})$, v_s 和 v_t 分别为一条边的源节点和目标节点的 ID , lbl 为边的标签, attr 为边的属性, v_s 、 v_t 和 lbl 唯一确定一条边。

基于本公开的基于度分布生成模型的社交网络图生成方法的另一个实施例中, 所述社区分离参数 ρ 是取值范围为 $[0,1]$ 中的一个实数, ρ 的取值越大, 社区之间边的数量越多。

基于本公开的基于度分布生成模型的社交网络图生成方法的另一个实施例中, 所述设置度分布生成模型, 根据度分布生成模型实现随机生成一个源节点的出度和多个目标节点 ID , 使得源节点、目标节点的出度和入度分布符合预期分布包括:

计算度分布生成模型的概率质量函数:

$$p(x) = \begin{cases} \alpha P(D = x; \theta) & \text{if } x \in [d_{\min}, d_{\max}] \text{ and } x \in N^+; \\ 0 & \text{otherwise} \end{cases};$$

$d_{\min}$ 和 $d_{\max}$ 分别为节点的最小度数和最大度数, θ 为节点度数分布的参数, $P(D = x; \theta)$ 为度数为 x 的节点的存在概率, α 为归一化参数, 使得所有可能的度数存在概率之和为 1;

根据节点出度分布 $\text{distr}_{\text{out}}$ 、节点数量 n_s 和预期边的数量 n_e , 对一个源节点计算满足 $\text{distr}_{\text{out}}$ 分布的出度如下:

计算生成的边的数量:

$$n'_e = \sum_{x=\text{out}_{\min}}^{\text{out}_{\max}} x \cdot n_s \cdot \alpha \cdot P(D = x; \theta_{\text{out}});$$

其中, $\text{out}_{\min}$ 和 $\text{out}_{\max}$ 分别为模式信息中提供的源节点的最小出度和最大出度, θ_{out} 为节点出度分布 $\text{distr}_{\text{out}}$ 的参数;

调整节点最大出度 $\text{out}_{\max}$ 以使得 n'_e 与 n_e 相符;

计算累积分布函数: $F(x) = \sum_{i=out\_min}^x \alpha P(D = i; \theta_{out})$, 其中, $x \in [outd_{min}, outd_{max}]$;

计算生成一个满足累计分布函数为 $F(x)$ 的随机数, y 为一个服从均匀分布 $U(0,1)$ 的随机数, 则 $F^{-1}(y)$ 为一个源节点 u 生成的符合 $distr_{out}$ 分布的出度;

构造函数 $G(z) = \operatorname{argmax}_x F(x) \leq z, x \in [outd_{min}, outd_{max}]$, 其中, $z \in \{i \cdot step | i \in$

5 $N^+, step = \min P(D = x), i \cdot step \leq 1\}$;

生成一个源节点 u 的符合 $distr_{out}$ 分布的出度 $F^{-1}(y) = G(\lfloor \frac{y}{step} \rfloor \cdot step)$ 。

基于本公开的基于度分布生成模型的社交网络图生成方法的另一个实施例中, 所述调整节点最大出度 $outd_{max}$ 以使得 n'_e 与 n_e 相符包括:

如果 $n'_e < n_e$, 则增大 $outd_{max}$, 直到度数为 $outd_{max}$ 的节点数量小于 1, 或者 $n'_e \geq n_e$;

10 如果 $n'_e = n_e$, 则不需要调整 $outd_{max}$;

如果 $n'_e > n_e$, 则减小 $outd_{max}$, 直到 $n'_e \leq n_e$ 。

基于本公开的基于度分布生成模型的社交网络图生成方法的另一个实施例中, 还包括:

对于一个确定出度的节点, 生成一个符合预期分布的目标节点;

根据节点入度分布 $distr_{in}$, 目标节点数量 n_t , 预期边的数量 n_e , 得到符合预期节点入

15 度分布的一个目标节点 ID;

计算节点入度和的累积分布函数: $F_s(x) = \sum_{i=ind_{min}}^x \beta \cdot i \cdot \alpha \cdot P(D = i; \theta_{in})$,

式中, $x \in [ind_{min}, ind_{max}]$, β 为归一化参数, $\beta = \frac{1}{\sum_{i=ind_{min}}^{ind_{max}} i \cdot \alpha \cdot P(D = i; \theta_{in})}$;

构造辅助函数: $H_1(z) = F_s(\operatorname{argmax}_x F_s(x) \leq z)$, $H_2(z) = F_s(\operatorname{argmin}_x F_s(x) \geq z)$, 式

中, x 是节点入度, $z \in \{i \cdot step | i \in N^+\}$, $step = \min_{x \in [ind_{min}, ind_{max}]} (F_s(x+1) - F_s(x))$,

20 $i \cdot step \leq 1$;

计算两个累积分布函数值所对应的节点 ID 为: $G_1(z) = \sum_{i=ind_{min}}^{x_1} i \cdot \alpha \cdot P(D = i; \theta_{in})$,

$G_2(z) = \sum_{i=ind_{min}}^{x_2} i \cdot \alpha \cdot P(D = i; \theta_{in})$, 式中, $x_1 = \operatorname{argmax}_x F_s(x) \leq z$, $x_2 = \operatorname{argmin}_x F_s(x) \geq z$;

目标节点 ID 为: $G_1(z) + \left\lfloor (y - H_1(z)) \times \frac{G_2(z) - G_1(z)}{H_2(z) - H_1(z)} \right\rfloor$ 。

25 基于本公开的基于度分布生成模型的社交网络图生成方法的另一个实施例中, 所述获取源节点和目标节点数量及度分布参数, 基于度分布生成模型生成社交网络图的简单图包括:

通过输入参数源节点数量 n_s 、目标节点数量 n_t 、期望生成的边数量 n_e 、目标节点的入度分布 $distr_{in}$ 、源节点的出度分布 $distr_{out}$;

以 $n_s \times n_t$ 的矩阵 M 表示一个图, 以 $M_{ij} = 1$ 表示存在一条从源节点 v_i 指向目标节点 v_j 的有向边, $M_{ij} = 0$ 表示不存在有向边, 计算任一节点 u 的出度 $outd$;

确定 $outd$ 个不同的目标节点, 构建 $outd$ 条边。

5 基于本公开的基于度分布生成模型的社交网络图生成方法的另一个实施例中, 所述基于度分布生成模型生成社交网络图, 确定社区大小, 对于同一社区中的图和不同社区中的图, 分别获取每一部分的源节点和目标节点数量, 生成简单图后合并为一个社交网络图包括:

获取社交网络模式信息 S , 记 $d_{out}(u)$ 为节点 u 的出度, $d_{out}^i(u)$ 为节点 u 与同一社区中节点相关的出度, $d_{out}^e(u) = d_{out}(u) - d_{out}^i(u)$ 为节点 u 与其他社区中节点相关的出度, 计算
10 $d_{out}^e(u)$ 的概率密度函数为:

$$p(x) = \begin{cases} \alpha e^{-\frac{x}{1+\rho}} & \text{if } x \in [1, outd'_{max}] \\ 0 & \text{otherwise} \end{cases}, \text{ 式中, } \alpha \text{ 为归一化参数, } \rho \text{ 为社区分离参数, 其}$$

取值范围为 $[0,1]$ 之间的实数, $outd'_{max} = outd_{max} - d_{out}^i(u)$, $outd_{max}$ 为源节点最大出度, $p(x)$ 为单调递减函数;

将 $d_{out}^e(u)$ 视为连续随机变量, 根据概率密度函数的性质得到: $\int_1^{outd'_{max}} \alpha e^{-\frac{x}{1+\rho}} = 1$;

15 计算节点 u 与其他社区中结点相关的出度 $d_{out}^e(u)$ 为: $d_{out}^e(u) = -(1+\rho) \log(e^{-\frac{1}{1+\rho}} + y(e^{\frac{outd'_{max}}{1+\rho}} - e^{-\frac{1}{1+\rho}}))$, 式中, y 为一个符合 $U(0,1)$ 分布的随机数, 某一节点与其他社区中节点相关的出度 $d_{out}^e(u)$ 与 y 满足: $\int_1^{d_{out}^e(u)} \alpha e^{-\frac{x}{1+\rho}} = y$;

对于社交网络模式信息 S 中每个边模式 ES , 在社交网络模式信息 S 中找到相关的源节点模式 VS_s 、目标节点的模式 VS_t 、社区模式 CS , 生成的边的数量 $ES.amount$ 为 n_e , 源节点数量 $VS_s.amount$ 为 n_s , 目标节点数量 $VS_t.amount$ 为 n_t , 社区数量 $CS.amount$ 为 n_c , $CS.\lambda_s$ 和
20 $CS.\lambda_t$ 为两个幂律分布的参数;

根据社区数量 $CS.amount$, 社区大小分布参数 $CS.\lambda_s$ 和 $CS.\lambda_t$, 确定每个社区的矩阵表示的大小: $n_{s_1} \times n_{t_1}, \dots, n_{s_{n_c}} \times n_{t_{n_c}}$;

生成该社区与其他社区之间的边, 计算一个源节点 u 随机生成一个与其他社区之间节
25 点的出度 $d_{out}^e(u)$;

生成 $d_{out}^e(u)$ 个不同的目标节点, 从而为源节点 u 生成与其他社区中节点相连的边。

基于本公开的基于度分布生成模型的社交网络图生成方法的另一个实施例中, 所述基于度分布生成模型以流式生成图数据, 通过比较当前时刻与上一时刻的源节点和目标节点

数量, 得到需要新增加的源节点和目标节点数量, 通过简单图的生成方式生成图数据包括:

获取源节点数量 n_s 、目标节点数量 n_t 、期望生成的边数量 n_e 、目标节点的入度分布 distr_{in} 、源节点的出度分布 distr_{out} , 增长率参数 r_g , 所述增长率参数 r_g 为区间 $[0,1]$ 之间的实数;

5 初始化变量 pc_{it} 为0, pc_{tg} 为 r_g ;

对于每个ID小于 $n_s \cdot pc_{tg}$ 的源节点 u , 计算该节点的出度 $outd$;

生成 $outd$ 个不同的ID小于 $n_t \cdot pc_{tg}$ 的目标节点, 生成新的边。

与现有技术相比, 本公开具有如下优点:

10 采用本公开的基于度分布生成模型的社交网络图生成方法提出度分布生成模型以快速生成符合特定分布的随机数, 高效地为每个节点确定出度, 并确定相应数量的目标节点, 从而生成图, 生成图的边表示网络中用户的交互关系, 生成的图具有真实社交网络的特点, 能够作为各种网络分析任务的输入数据, 满足多种应用场景的需要, 生成方法高效, 可扩展性强, 适合生成大规模图数据。

附图说明

15 构成说明书的一部分的附图描述了本公开的实施例, 并且连同描述一起用于解释本公开的原理。

参照附图, 根据下面的详细描述, 可以更加清楚地理解本公开, 其中:

图1为本公开提出的基于度分布生成模型的社交网络图生成方法的流程。

20 具体实施方式

为使本公开实施例的目的、技术方案和优点更加清楚, 下面将结合本公开实施例中的附图, 对本公开实施例中的技术方案进行清楚、完整地描述, 显然, 所描述的实施例只是本公开一部分实施例, 而不是全部的实施例。基于本公开中的实施例, 本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例, 都属于本公开保护的范围。

25 随着互联网技术的不断发展, 社交网络平台蓬勃发展, 网络规模不断扩大, 网络平台中的用户数以亿计, 用户间的交互关系纷繁复杂。这样复杂的网络关系可以用图来表示, 用户表示为图中的节点, 用户间的交互关系, 例如, 关注、评论、点赞等行为表示为边。许多社交网络分析任务处理这样大规模的网络图用于辅助实际业务, 例如, 社区发现算法能够检测出网络中的社区结构, 这些结构信息能够辅助于风险控制任务和用户推荐任务等。

为了验证社交网络分析算法的有效性和可扩展性，往往需要使用合成数据集，因为抽取实际应用场景中的网络数据往往需要消耗大量资源，而快速合成大规模网络数据便十分必要。

本公开的目的在于提出一种基于度分布生成模型的社交网络图生成方法，该方法利用用户自定义的模式信息生成社交网络图，能够满足多种应用场景的需要，生成方法高效，可扩展性强，适合生成大规模图数据。生成图中的节点可以代表社交网络中的用户，边可以表示用户的行为。

下面结合附图和实施例对本公开提供的一种基于度分布生成模型的社交网络图生成方法进行更详细地说明。

基于度分布生成模型的社交网络图生成方法，在图的矩阵表示上生成数据，对于每个源节点，随机生成一个出度，再生成若干目标节点 ID，从而生成边。为了加速生成过程，我们提出了一种新的生成模型，称为度分布生成模型 ($D^2 G$)，使用这一模型，为一个源节点确定出度或生成一个目标节点的时间复杂度为 $O(1)$ 。因此，适合使用 $D^2 G$ 模型生成大规模图数据，此外， $D^2 G$ 是一个通用模型，只要提供度数分布的概率密度函数或概率质量函数，就可以使用该模型生成对应度数分布的图数据。

图 1 为本公开提出的基于度分布生成模型的社交网络图生成方法的流程图，如图 1 所示，所述基于度分布生成模型的社交网络图生成方法：

10，设置用于社交网络图生成的模式信息，根据社交网络图生成的模式信息生成网络图数据；

20，设置度分布生成模型，根据度分布生成模型实现随机生成一个源节点的出度和多个目标节点 ID，使得源节点、目标节点的出度和入度分布符合预期分布；

30，获取源节点和目标节点数量及度分布参数，基于度分布生成模型生成社交网络图的简单图；

40，基于度分布生成模型生成社交网络图，确定社区大小，对于同一社区中的图 and 不同社区中的图，分别获取每一部分的源节点和目标节点数量，生成简单图后合并为一个社交网络图；

50，基于度分布生成模型以流式生成图数据，通过比较当前时刻与上一时刻的源节点和目标节点数量，得到新增加的源节点和目标节点数量，通过简单图的生成方式生成图数据。

所述设置用于社交网络图生成的模式信息，根据社交网络图生成的模式信息生成网络图数据包括：

设置节点模式信息、设置边模式信息、设置社区模式信息、设置社交网络图模式信息、设置用于生成社交网络图的符号表示；

所述节点模式信息包括： $VS = (lbl, amount, attr)$ ，其中 lbl 是节点标签， $amount$ 是标签为 lbl 的节点数量， $attr$ 是节点属性信息， $attr$ 为可选值；

- 5 所述边模式信息包括： $ES = (lbl, lbl_s, lbl_t, amount, distr_{in}, distr_{out}, attr)$ ，其中 lbl 是边的标签信息， lbl_s 和 lbl_t 分别为源节点和目标节点的标签， $amount$ 为预期生成的边的数量， $distr_{in}$ 和 $distr_{out}$ 分别为目标节点的入度分布和源节点的出度分布， $attr$ 为边的属性信息， $attr$ 为可选值；

- 10 所述社区模式信息包括： $CS = (lbl_e, amount, \lambda_s, \lambda_t, \rho)$ ，表示在标签为 lbl_e 的边构成的图上构建社区， $amount$ 为生成的社区数量，社区大小服从幂律分布， λ_s 和 λ_t 分别为所有社区中源节点和目标节点数量服从幂律分布的参数， ρ 为社区分离参数，社区分离参数 ρ 是取值范围为 $[0,1]$ 中的一个实数， ρ 的取值越大，社区之间边的数量越多；

所述社交网络图模式信息包括： $SGS = (VSS, ESS, CSS)$ ，其中 VSS 为节点模式的集合， ESS 为边的模式集合， CSS 为社区模式的集合；

- 15 所述设置用于生成的社交网络图，其符号表示包括：

设置有向异构图 $G=(V,E)$ ，其中 V 为节点集合， $E \subseteq V \times V$ 为边集；

一个节点 $v \in V$ 为一个三元组 $(ID, lbl, attr)$ ，其中 lbl 为节点标签， $attr$ 为节点属性， ID 为一个节点的唯一标识；

- 20 一条边 $e \in E$ 以为一个四元组 $(v_s, v_t, lbl, attr)$ ， v_s 和 v_t 分别为一条边的源节点和目标节点的 ID ， lbl 为边的标签， $attr$ 为边的属性， v_s 、 v_t 和 lbl 唯一确定一条边。

节点和边的属性信息是可选的。

所述设置度分布生成模型，根据度分布生成模型实现一个源节点随机生成出度和目标节点 ID ，使得源节点、目标节点的出度和入度分布符合预期分布包括：

计算度分布生成模型的概率质量函数：

$$25 \quad p(x) = \begin{cases} \alpha P(D = x; \theta) & \text{if } x \in [d_{min}, d_{max}] \text{ and } x \in N^+; \\ 0 & \text{otherwise} \end{cases};$$

d_{min} 和 d_{max} 分别为节点的最小度数和最大度数， θ 为节点度数分布的参数， $P(D = x; \theta)$ 为度数为 x 的节点的存在概率， α 为归一化参数，使得所有可能的度数存在概率之和为1；

计算 α 的公式如下：

$$\alpha = \frac{1}{\sum_{x=d_{min}}^{d_{max}} P(D=x;\theta)};$$

累积分布函数 CDF 的公式如下:

$$F(x) = \sum_{i=d_{min}}^x \alpha P(D = i; \theta);$$

其中, $x \in [d_{min}, d_{max}]$;

- 5 为了生成一个满足累积分布函数为 $F(x)$ 的随机数, 首先生成一个服从均匀分布 $U(0,1)$ 的随机数 y , 然后 $F^{-1}(y)$ 就是生成的满足累积分布函数为 $F(x)$ 的随机数。为了高效地计算 $F^{-1}(y)$, 设计一个新的函数 G , 如下所示:

$$G(z) = \operatorname{argmax}_x F(x) \leq z, x \in [d_{min}, d_{max}] ;$$

其中, $z \in \{i \cdot \text{step} | i \in \mathbb{N}^+, \text{step} = \min P(D = x), i \cdot \text{step} \leq 1\}$ 。给定一个服从均匀分布

- 10 $U(0,1)$ 的随机数 y , 可以通过函数 $G(\cdot)$ 直接得到: $F^{-1}(y) = G(\lfloor \frac{y}{\text{step}} \rfloor \cdot \text{step})$ 。

使用度分布生成模型实现 $\text{GenOutDegree}(\text{distr}_{\text{out}}, n_s, n_e)$ 程序, 该程序输入的参数有: 节点出度分布 $\text{distr}_{\text{out}}$, 节点数量 n_s 和预期边的数量 n_e , 输出为一个满足 $\text{distr}_{\text{out}}$ 分布的出度。

- 15 根据节点出度分布 $\text{distr}_{\text{out}}$ 、节点数量 n_s 和预期边的数量 n_e , 对一个源节点计算满足 $\text{distr}_{\text{out}}$ 分布的出度如下:

计算生成的边的数量:

$$n'_e = \sum_{x=\text{outd}_{min}}^{\text{outd}_{max}} x \cdot n_s \cdot \alpha \cdot P(D = x; \theta_{\text{out}});$$

其中, outd_{min} 和 outd_{max} 分别为模式信息中提供的源节点的最小出度和最大出度, θ_{out} 为节点出度分布 $\text{distr}_{\text{out}}$ 的参数;

- 20 如果预期边的数量 $n_e = -1$, 则不需要调整参数, 生成的边的数量为 n'_e ; 否则, 调整节点最大出度 outd_{max} 以使得 n'_e 与 n_e 相符, 共有三种情况:

如果 $n'_e < n_e$, 则增大 outd_{max} , 直到度数为 outd_{max} 的节点数量小于 1, 或者 $n'_e \geq n_e$;

如果 $n'_e = n_e$, 则不需要调整 outd_{max} ;

如果 $n'_e > n_e$, 则减小 outd_{max} , 直到 $n'_e \leq n_e$ 。

- 25 计算累积分布函数: $F(x) = \sum_{i=\text{outd}_{min}}^x \alpha P(D = i; \theta_{\text{out}})$, 其中, $x \in [\text{outd}_{min}, \text{outd}_{max}]$;

计算生成一个满足累积分布函数为 $F(x)$ 的随机数, y 为一个服从均匀分布 $U(0,1)$ 的随机数, 则 $F^{-1}(y)$ 为一个源节点 u 生成的符合 $\text{distr}_{\text{out}}$ 分布的出度;

构造函数 $G(z) = \operatorname{argmax}_x F(x) \leq z, x \in [\text{outd}_{min}, \text{outd}_{max}]$, 其中, $z \in \{i \cdot \text{step} | i \in \mathbb{N}^+, \text{step} = \min P(D = x), i \cdot \text{step} \leq 1\}$;

生成一个源节点 u 的符合 $\text{distr}_{\text{out}}$ 分布的出度 $F^{-1}(y) = G(\lfloor \frac{y}{\text{step}} \rfloor \cdot \text{step})$ 。

对于一个确定出度的节点，生成一个符合预期分布的目标节点；

根据节点入度分布 distr_{in} ，目标节点数量 n_t ，预期边的数量 n_e ，得到符合预期节点入度分布的一个目标节点ID；

- 5 给出一个目标节点id与入度关系的约束条件。给定 n_t 个目标节点 $v_1, v_2, \dots, v_{n_t}$ ，按照id从小到大排序，这些节点入度非递增。这样的约束条件是合理的，因为如果希望节点id与入度之间没有明显的关系，可以随机生成一个 $[1, n_t]$ 上正整数的一个排列，然后将节点id映射为排列中对应的值。

计算节点入度和的累积分布函数： $F_s(x) = \sum_{i=\text{ind}_{\min}}^x \beta \cdot i \cdot \alpha \cdot P(D = i; \theta_{\text{in}})$,

- 10 式中， $x \in [\text{ind}_{\min}, \text{ind}_{\max}]$ ， β 为归一化参数， $\beta = \frac{1}{\sum_{i=\text{ind}_{\min}}^{\text{ind}_{\max}} i \cdot \alpha \cdot P(D = i; \theta_{\text{in}})}$ ；

给定一个服从 $U(0,1)$ 分布的随机数 y ，首先找到两个累积分布函数值 $F_s(x_1)$ 和 $F_s(x_2)$ ，满足 $F_s(x_1) \leq y \leq F_s(x_2)$ ， $x_1 + 1 \geq x_2$ ，以及这两个函数值对应的节点ID；

构造辅助函数： $H_1(z) = F_s(\arg\max_x F_s(x) \leq z)$ ， $H_2(z) = F_s(\arg\min_x F_s(x) \geq z)$ ，式中， x 是节点入度， $z \in \{i \cdot \text{step} | i \in N^+\}$ ， $\text{step} = \min_{x \in [\text{ind}_{\min}, \text{ind}_{\max}]} (F_s(x+1) - F_s(x))$ ，

- 15 $i \cdot \text{step} \leq 1$ ；

计算两个累积分布函数值所对应的节点ID为： $G_1(z) = \sum_{i=\text{ind}_{\min}}^{x_1} i \cdot \alpha \cdot P(D = i; \theta_{\text{in}})$ ， $G_2(z) = \sum_{i=\text{ind}_{\min}}^{x_2} i \cdot \alpha \cdot P(D = i; \theta_{\text{in}})$ ，式中， $x_1 = \arg\max_x F_s(x) \leq z$ ， $x_2 = \arg\min_x F_s(x) \geq z$ ；

- 20 生成一个服从 $U(0,1)$ 分布的随机数 y ，目标节点ID为： $G_1(z) + \lfloor (y - H_1(z)) \times \frac{G_2(z) - G_1(z)}{H_2(z) - H_1(z)} \rfloor$ 。

所述获取源节点和目标节点数量及度分布参数，基于度分布生成模型生成社交网络图的简单图包括：

通过输入参数源节点数量 n_s 、目标节点数量 n_t 、期望生成的边数量 n_e 、目标节点的入度分布 distr_{in} 、源节点的出度分布 $\text{distr}_{\text{out}}$ ；

- 25 以 $n_s \times n_t$ 的矩阵 M 表示一个图，以 $M_{ij} = 1$ 表示存在一条从源节点 v_i 指向目标节点 v_j 的有向边， $M_{ij} = 0$ 表示不存在有向边，计算任一节点 u 的出度 outd ；

确定 outd 个不同的目标节点，构建 outd 条边。

所述基于度分布生成模型生成社交网络图，确定社区大小，对于同一社区中的图和不同社区中的图，分别获取每一部分的源节点和目标节点数量，生成简单图后合并为一个社

交网络图包括:

获取社交网络模式信息 S , 记 $d_{out}(u)$ 为节点 u 的出度, $d_{out}^i(u)$ 为节点 u 与同一社区中节点相关的出度, $d_{out}^e(u) = d_{out}(u) - d_{out}^i(u)$ 为节点 u 与其他社区中节点相关的出度, 计算 $d_{out}^e(u)$ 的概率密度函数为:

$$5 \quad p(x) = \begin{cases} \alpha e^{-\frac{x}{1+\rho}} & \text{if } x \in [1, outd'_{max}] \\ 0 & \text{otherwise} \end{cases}, \text{ 式中, } \alpha \text{ 为归一化参数, } \rho \text{ 为社区分离参数, 其}$$

取值范围为 $[0,1]$ 之间的实数, $outd'_{max} = outd_{max} - d_{out}^i(u)$, $outd_{max}$ 为源节点最大出度, $p(x)$ 为单调递减函数, 表示一个源节点 u 有较小的 $d_{out}^e(u)$ 的概率更高, 即两个社区间节点连接稀疏的概率更高, 参数 ρ 越大, 一个源节点有较大的 $d_{out}^e(u)$ 的概率更高, 即不同社区之间边的数量更多;

$$10 \quad \text{将 } d_{out}^e(u) \text{ 视为连续随机变量, 根据概率密度函数的性质得到: } \int_1^{outd'_{max}} \alpha e^{-\frac{x}{1+\rho}} = 1;$$

计算节点 u 与其他社区中结点相关的出度 $d_{out}^e(u)$ 为: $d_{out}^e(u) = -(1+\rho) \log(e^{-\frac{1}{1+\rho}} + y(e^{\frac{outd'_{max}}{1+\rho}} - e^{-\frac{1}{1+\rho}}))$, 式中, y 为一个符合 $U(0,1)$ 分布的随机数, 某一节点与其他社区中节点相关的出度 $d_{out}^e(u)$ 与 y 满足: $\int_1^{d_{out}^e(u)} \alpha e^{-\frac{x}{1+\rho}} = y$;

对于社交网络模式信息 S 中每个边模式 ES , 在社交网络模式信息 S 中找到相关的源节点模式 VS_s 、目标节点的模式 VS_t 、社区模式 CS , 生成的边的数量 $ES.amount$ 为 n_e , 源节点数量 $VS_s.amount$ 为 n_s , 目标节点数量 $VS_t.amount$ 为 n_t , 社区数量 $CS.amount$ 为 n_c , $CS.\lambda_s$ 和 $CS.\lambda_t$ 为两个幂律分布的参数;

根据社区数量 $CS.amount$, 社区大小分布参数 $CS.\lambda_s$ 和 $CS.\lambda_t$, 确定每个社区的矩阵表示的大小: $n_{s_1} \times n_{t_1}, \dots, n_{s_{n_c}} \times n_{t_{n_c}}$;

20 生成该社区与其他社区之间的边, 计算一个源节点 u 随机生成一个与其他社区之间节点的出度 $d_{out}^e(u)$;

生成 $d_{out}^e(u)$ 个不同的目标节点, 从而为源节点 u 生成与其他社区中节点相连的边。

基于本公开的基于度分布生成模型的社交网络图生成方法的另一个实施例中, 所述基于度分布生成模型以流式生成图数据, 通过比较当前时刻与上一时刻的源节点和目标节点数量, 得到需要新增加的源节点和目标节点数量, 通过简单图的生成方式生成图数据包括:

25 获取源节点数量 n_s 、目标节点数量 n_t 、期望生成的边数量 n_e 、目标节点的入度分布 $distr_{in}$ 、源节点的出度分布 $distr_{out}$, 增长率参数 r_g , 所述增长率参数 r_g 为区间 $[0,1]$ 之间的实数;

初始化变量 pc_{lt} 为0, pc_{tg} 为 r_g ;

对于每个ID小于 $n_s \cdot pc_{tg}$ 的源节点 u , 计算该节点的出度 out_d ;

生成 out_d 个不同的ID小于 $n_t \cdot pc_{tg}$ 的目标节点, 生成新的边。

各种社交网络平台蓬勃发展, 网络规模不断扩大, 许多网络分析任务处理这些大规模网络用于实际业务中。获取实际网络图结构往往需要消耗大量资源, 因此有必要使用人工生成网络图技术高效地生成图, 用于评估网络分析算法的有效性和可扩展性。

本公开提出的方法利用用户自定义的模式信息生成社交网络图, 能够满足多种应用场景的需要。我们提出度分布生成模型以快速生成符合特定分布的随机数, 利用这一模型, 可以高效地为每个节点确定出度, 并确定相应数量的目标节点, 从而生成图, 生成的图节点可以表示实际网络中的用户; 生成图的边可以表示网络中用户的交互关系, 例如用户间关注关系、私信行为等。

生成的图具有真实社交网络的特点, 包括小世界、社区结构和幂律分布, 能够作为各种网络分析任务, 例如社区发现、社区搜索和网络表示学习等的输入数据。

对于本领域技术人员而言, 显然本公开实施例不限于上述示范性实施例的细节, 而且在不背离本公开实施例的精神或基本特征的情况下, 能够以其他的具体形式实现本公开实施例。因此, 无论从哪一点来看, 均应将实施例看作是示范性的, 而且是非限制性的, 本公开实施例的范围由所附权利要求而不是上述说明限定, 因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本公开实施例内。不应将权利要求中的任何附图标记视为限制所涉及的权利要求。此外, 显然“包括”一词不排除其他单元或步骤, 单数不排除复数。系统、装置或终端权利要求中陈述的多个单元、模块或装置也可以由同一个单元、模块或装置通过软件或者硬件来实现。第一, 第二等词语用来表示名称, 而并不表示任何特定的顺序。

最后应说明的是, 以上实施方式仅用以说明本公开实施例的技术方案而非限制, 尽管参照以上较佳实施方式对本公开实施例进行了详细说明, 本领域的普通技术人员应当理解, 可以对本公开实施例的技术方案进行修改或等同替换都不应脱离本公开实施例的技术方案的精神和范围。

权 利 要 求

1. 一种基于度分布生成模型的社交网络图生成方法，其特征在于，包括：

设置用于社交网络图生成的模式信息，根据社交网络图生成的模式信息生成网络图数据；

5 设置度分布生成模型，根据度分布生成模型实现随机生成一个源节点的出度和多个目标节点 ID，使得源节点、目标节点的出度和入度分布符合预期分布；

获取源节点和目标节点数量及度分布参数，基于度分布生成模型生成社交网络图的简单图；

10 基于度分布生成模型生成社交网络图，确定社区大小，对于同一社区中的图 and 不同社区中的图，分别获取每一部分的源节点和目标节点数量，生成简单图后合并为一个社交网络图；

基于度分布生成模型以流式生成图数据，通过比较当前时刻与上一时刻的源节点和目标节点数量，得到新增加的源节点和目标节点数量，通过简单图的生成方式生成图数据。

2. 根据权利要求 1 所述的基于度分布生成模型的社交网络图生成方法，其特征在于，
15 所述设置用于社交网络图生成的模式信息，根据社交网络图生成的模式信息生成网络图数据包括：

设置节点模式信息、设置边模式信息、设置社区模式信息、设置社交网络图模式信息、设置用于生成社交网络图的符号表示；

20 所述节点模式信息包括： $VS = (l_{bl}, amount, attr)$ ，其中 l_{bl} 是节点标签， $amount$ 是标签为 l_{bl} 的节点数量， $attr$ 是节点属性信息；

所述边模式信息包括： $ES = (l_{bl}, l_{bl_s}, l_{bl_t}, amount, distr_{in}, distr_{out}, attr)$ ，其中 l_{bl} 是边的标签信息， l_{bl_s} 和 l_{bl_t} 分别为源节点和目标节点的标签， $amount$ 为预期生成的边的数量， $distr_{in}$ 和 $distr_{out}$ 分别为目标节点的入度分布和源节点的出度分布， $attr$ 为边的属性信息，是一个 key-value 结构的字典，key 为属性名，value 为属性取值；

25 所述社区模式信息包括： $CS = (l_{bl_e}, amount, \lambda_s, \lambda_t, \rho)$ ，表示在标签为 l_{bl_e} 的边构成的图上构建社区， $amount$ 为生成的社区数量，社区大小服从幂律分布， λ_s 和 λ_t 分别为所有社区中源节点和目标节点数量服从幂律分布的参数， ρ 为社区分离参数；

所述社交网络图模式信息包括： $SGS = (VSS, ESS, CSS)$ ，其中 VSS 为节点模式的集合， ESS 为边的模式集合， CSS 为社区模式的集合；

30 所述设置用于生成的社交网络图，其符号表示包括：

设置有向异构图 $G=(V,E)$ ，其中 V 为节点集合， $E \subseteq V \times V$ 为边集；

一个节点 $v \in V$ 为一个三元组 $(ID, lbl, attr)$ ，其中 lbl 为节点标签， $attr$ 为节点属性， ID 为一个节点的唯一标识；

5 一条边 $e \in E$ 以为一个四元组 $(v_s, v_t, lbl, attr)$ ， v_s 和 v_t 分别为一条边的源节点和目标节点的 ID ， lbl 为边的标签， $attr$ 为边的属性， v_s 、 v_t 和 lbl 唯一确定一条边。

3.根据权利要求 2 所述的基于度分布生成模型的社交网络图生成方法，其特征在于，所述社区分离参数 ρ 是取值范围为 $[0,1]$ 中的一个实数， ρ 的取值越大，社区之间边的数量越多。

4.根据权利要求 1 所述的基于度分布生成模型的社交网络图生成方法，其特征在于，10 所述设置度分布生成模型，根据度分布生成模型实现随机生成一个源节点的出度和多个目标节点 ID ，使得源节点、目标节点的出度和入度分布符合预期分布包括：

计算度分布生成模型的概率质量函数：

$$p(x) = \begin{cases} \alpha P(D = x; \theta) & \text{if } x \in [d_{min}, d_{max}] \text{ and } x \in N^+; \\ 0 & \text{otherwise} \end{cases};$$

15 d_{min} 和 d_{max} 分别为节点的最小度数和最大度数， θ 为节点度数分布的参数， $P(D = x; \theta)$ 为度数为 x 的节点的存在概率， α 为归一化参数，使得所有可能的度数存在概率之和为 1；

根据节点出度分布 $distr_{out}$ 、节点数量 n_s 和预期边的数量 n_e ，对一个源节点计算满足 $distr_{out}$ 分布的出度如下：

计算生成的边的数量：

$$20 \quad n'_e = \sum_{x=outd_{min}}^{outd_{max}} x \cdot n_s \cdot \alpha \cdot P(D = x; \theta_{out});$$

其中， $outd_{min}$ 和 $outd_{max}$ 分别为模式信息中提供的源节点的最小出度和最大出度， θ_{out} 为节点出度分布 $distr_{out}$ 的参数；

调整节点最大出度 $outd_{max}$ 以使得 n'_e 与 n_e 相符；

25 计算节点出度的累积分布函数： $F(x) = \sum_{i=outd_{min}}^x \alpha P(D = i; \theta)$ ，其中， $x \in [outd_{min}, outd_{max}]$ ；

计算生成一个满足累计分布函数为 $F(x)$ 的随机数， y 为一个服从均匀分布 $U(0,1)$ 的随机数，则 $F^{-1}(y)$ 为一个源节点 u 生成的符合 $distr_{out}$ 分布的出度；

构造函数 $G(z) = \operatorname{argmax}_x F(x) \leq z, x \in [outd_{min}, outd_{max}]$ ，其中， $z \in \{i \cdot \text{step} | i \in N^+, \text{step} = \min P(D = x), i \cdot \text{step} \leq 1\}$ ；

生成一个源节点u的符合 $\text{distr}_{\text{out}}$ 分布的出度为: $F^{-1}(y) = G(\lfloor \frac{y}{\text{step}} \rfloor \cdot \text{step})$ 。

5. 根据权利要求 4 所述的基于度分布生成模型的社交网络图生成方法, 其特征在于, 所述调整节点最大出度 outd_{max} 以使得 n'_e 与 n_e 相符包括:

如果 $n'_e < n_e$, 则增大 outd_{max} , 直到度数为 outd_{max} 的节点数量小于 1, 或者 $n'_e \geq n_e$;

5 如果 $n'_e = n_e$, 则不需要调整 outd_{max} ;

如果 $n'_e > n_e$, 则减小 outd_{max} , 直到 $n'_e \leq n_e$ 。

6. 根据权利要求 4 所述的基于度分布生成模型的社交网络图生成方法, 其特征在于, 还包括:

对于一个确定出度的节点, 生成一个符合预期分布的目标节点;

10 根据节点入度分布 distr_{in} , 目标节点数量 n_t , 预期边的数量 n_e , 得到符合预期节点入度分布的一个目标节点 ID;

计算节点入度和的累积分布函数: $F_s(x) = \sum_{i=\text{ind}_{\text{min}}}^x \beta \cdot i \cdot \alpha \cdot P(D = i; \theta_{\text{in}})$,

式中, $x \in [\text{ind}_{\text{min}}, \text{ind}_{\text{max}}]$, β 为归一化参数, $\beta = \frac{1}{\sum_{i=\text{ind}_{\text{min}}}^{\text{ind}_{\text{max}}} i \cdot \alpha \cdot P(D = i; \theta_{\text{in}})}$;

构造辅助函数: $H_1(z) = F_s(\arg\max_x F_s(x) \leq z)$, $H_2(z) = F_s(\arg\min_x F_s(x) \geq z)$, 式

15 中, x 是节点入度, $z \in \{i \cdot \text{step} | i \in N^+\}$, $\text{step} = \min_{x \in [\text{ind}_{\text{min}}, \text{ind}_{\text{max}}]} (F_s(x+1) - F_s(x))$, $i \cdot \text{step} \leq 1$;

计算两个累积分布函数值所对应的节点ID为: $G_1(z) = \sum_{i=\text{ind}_{\text{min}}}^{x_1} i \cdot \alpha \cdot P(D = i; \theta_{\text{in}})$,

$G_2(z) = \sum_{i=\text{ind}_{\text{min}}}^{x_2} i \cdot \alpha \cdot P(D = i; \theta_{\text{in}})$, 式中, $x_1 = \arg\max_x F_s(x) \leq z$, $x_2 = \arg\min_x F_s(x) \geq z$;

20 目标节点ID为: $G_1(z) + \left\lfloor (y - H_1(z)) \times \frac{G_2(z) - G_1(z)}{H_2(z) - H_1(z)} \right\rfloor$ 。

7. 根据权利要求 6 所述的基于度分布生成模型的社交网络图生成方法, 其特征在于, 所述获取源节点和目标节点数量及度分布参数, 基于度分布生成模型生成社交网络图的简单图包括:

25 通过输入参数源节点数量 n_s 、目标节点数量 n_t 、期望生成的边数量 n_e 、目标节点的入度分布 distr_{in} 、源节点的出度分布 $\text{distr}_{\text{out}}$;

以 $n_s \times n_t$ 的矩阵 M 表示一个图, 以 $M_{ij} = 1$ 表示存在一条从源节点 v_i 指向目标节点 v_j 的有向边, $M_{ij} = 0$ 表示不存在有向边, 计算任一节点 u 的出度 outd ;

确定 outd 个不同的目标节点, 构建 outd 条边。

8. 根据权利要求 7 所述的基于度分布生成模型的社交网络图生成方法, 其特征在于,

所述基于度分布生成模型生成社交网络图，确定社区大小，对于同一社区中的图和不同社区中的图，分别获取每一部分的源节点和目标节点数量，生成简单图后合并为一个社交网络图包括：

获取社交网络模式信息 S ，记 $d_{out}(u)$ 为节点 u 的出度， $d_{out}^i(u)$ 为节点 u 与同一社区中节点相关的出度， $d_{out}^e(u) = d_{out}(u) - d_{out}^i(u)$ 为节点 u 与其他社区中节点相关的出度，计算 $d_{out}^e(u)$ 的概率密度函数为：

$$p(x) = \begin{cases} \alpha e^{-\frac{x}{1+\rho}} & \text{if } x \in [1, outd'_{max}] \\ 0 & \text{otherwise} \end{cases}$$

式中， α 为归一化参数， ρ 为社区分离参数，其

取值范围为 $[0,1]$ 之间的实数， $outd'_{max} = outd_{max} - d_{out}^i(u)$ ， $outd_{max}$ 为源节点最大出度， $p(x)$ 为单调递减函数；

10 将 $d_{out}^e(u)$ 视为连续随机变量，根据概率密度函数的性质得到： $\int_1^{outd'_{max}} \alpha e^{-\frac{x}{1+\rho}} = 1$ ；

计算节点 u 与其他社区中节点相关的出度 $d_{out}^e(u)$ 为： $d_{out}^e(u) = -(1+\rho) \log(e^{-\frac{1}{1+\rho}} + y(e^{\frac{outd'_{max}}{1+\rho}} - e^{-\frac{1}{1+\rho}}))$ ，式中， y 为一个符合 $U(0,1)$ 分布的随机数，某一节点与其他社区中节点相关的出度 $d_{out}^e(u)$ 与 y 满足： $\int_1^{d_{out}^e(u)} \alpha e^{-\frac{x}{1+\rho}} = y$ ；

15 对于社交网络模式信息 S 中每个边模式 ES ，在社交网络模式信息 S 中找到相关的源节点模式 VS_s 、目标节点的模式 VS_t 、社区模式 CS ，生成的边的数量 $ES.amount$ 为 n_e ，源节点数量 $VS_s.amount$ 为 n_s ，目标节点数量 $VS_t.amount$ 为 n_t ，社区数量 $CS.amount$ 为 n_c ， $CS.\lambda_s$ 和 $CS.\lambda_t$ 为两个幂律分布的参数；

根据社区数量 $CS.amount$ ，社区大小分布参数 $CS.\lambda_s$ 和 $CS.\lambda_t$ ，确定每个社区的矩阵表示的大小： $n_{s_1} \times n_{t_1}, \dots, n_{s_{n_c}} \times n_{t_{n_c}}$ ；

20 生成该社区与其他社区之间的边，计算一个源节点 u 随机生成一个与其他社区之间节点的出度 $d_{out}^e(u)$ ；

生成 $d_{out}^e(u)$ 个不同的目标节点，从而为源节点 u 生成与其他社区中节点相连的边。

9. 根据权利要求 8 所述的基于度分布生成模型的社交网络图生成方法，其特征在于，所述基于度分布生成模型以流式生成图数据，通过比较当前时刻与上一时刻的源节点和目标节点数量，得到需要新增加的源节点和目标节点数量，通过简单图的生成方式生成图数据包括：

获取源节点数量 n_s 、目标节点数量 n_t 、期望生成的边数量 n_e 、目标节点的入度分布 $distr_{in}$ 、源节点的出度分布 $distr_{out}$ ，增长率参数 r_g ，所述增长率参数 r_g 为区间 $[0,1]$ 之间的

实数;

初始化变量 pc_{lt} 为 0, pc_{tg} 为 r_g ;

对于每个ID小于 $n_s \cdot pc_{tg}$ 的源节点u, 计算该节点的出度 out_d ;

生成 out_d 个不同的ID小于 $n_t \cdot pc_{tg}$ 的目标节点, 生成新的边。

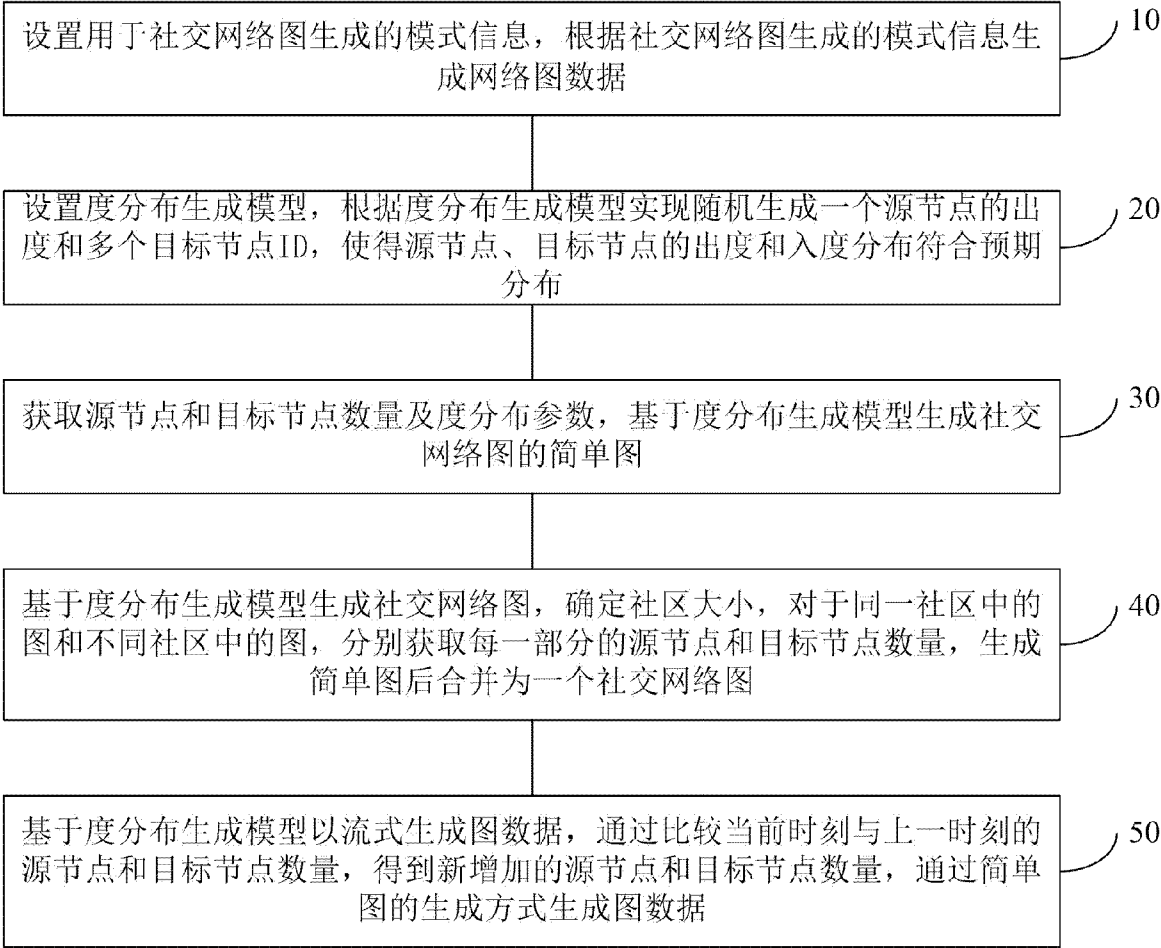


图 1

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/080472

A. CLASSIFICATION OF SUBJECT MATTER G06F 9/44(2018.01)i According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) G06F Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) CNKI, CNPAT, WPI, EPODOC, GOOGLE, IEEE: 网络图, 社交, 入度, 出度, 分布, 模型, 随机, 节点, network, construct+, random, establish, model, node		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 105550212 A (SHANGHAI DIANJI UNIVERSITY) 04 May 2016 (2016-05-04) description, paragraphs [0022]-[0048], and figures 1-5	1-9
A	CN 110659395 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 07 January 2020 (2020-01-07) entire document	1-9
A	CN 108510115 A (SHANDONG UNIVERSITY OF SCIENCE AND TECHNOLOGY) 07 September 2018 (2018-09-07) entire document	1-9
A	JP 2008052495 A (SONY CORP.) 06 March 2008 (2008-03-06) entire document	1-9
☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.		
* Special categories of cited documents: “A” document defining the general state of the art which is not considered to be of particular relevance “E” earlier application or patent but published on or after the international filing date “L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) “O” document referring to an oral disclosure, use, exhibition or other means “P” document published prior to the international filing date but later than the priority date claimed “T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention “X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone “Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art “&” document member of the same patent family		
Date of the actual completion of the international search 25 November 2020		Date of mailing of the international search report 21 December 2020
Name and mailing address of the ISA/CN China National Intellectual Property Administration (ISA/CN) No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088 China Facsimile No. (86-10)62019451		Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2020/080472

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	105550212	A	04 May 2016	None	
CN	110659395	A	07 January 2020	None	
CN	108510115	A	07 September 2018	None	
JP	2008052495	A	06 March 2008	None	

国际检索报告

国际申请号

PCT/CN2020/080472

A. 主题的分类 G06F 9/44 (2018.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类																	
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G06F 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) CNKI, CNPAT, WPI, EPDOC, GOOGLE, IEEE: 网络图, 社交, 入度, 出度, 分布, 模型, 随机, 节点, network, construct+, random, establish, model, node																	
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>A</td> <td>CN 105550212 A (上海电机学院) 2016年 5月 4日 (2016 - 05 - 04) 说明书第[0022]-[0048]段, 附图1-5</td> <td>1-9</td> </tr> <tr> <td>A</td> <td>CN 110659395 A (平安科技深圳有限公司) 2020年 1月 7日 (2020 - 01 - 07) 全文</td> <td>1-9</td> </tr> <tr> <td>A</td> <td>CN 108510115 A (山东科技大学) 2018年 9月 7日 (2018 - 09 - 07) 全文</td> <td>1-9</td> </tr> <tr> <td>A</td> <td>JP 2008052495 A (SONY CORP.) 2008年 3月 6日 (2008 - 03 - 06) 全文</td> <td>1-9</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	A	CN 105550212 A (上海电机学院) 2016年 5月 4日 (2016 - 05 - 04) 说明书第[0022]-[0048]段, 附图1-5	1-9	A	CN 110659395 A (平安科技深圳有限公司) 2020年 1月 7日 (2020 - 01 - 07) 全文	1-9	A	CN 108510115 A (山东科技大学) 2018年 9月 7日 (2018 - 09 - 07) 全文	1-9	A	JP 2008052495 A (SONY CORP.) 2008年 3月 6日 (2008 - 03 - 06) 全文	1-9
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
A	CN 105550212 A (上海电机学院) 2016年 5月 4日 (2016 - 05 - 04) 说明书第[0022]-[0048]段, 附图1-5	1-9															
A	CN 110659395 A (平安科技深圳有限公司) 2020年 1月 7日 (2020 - 01 - 07) 全文	1-9															
A	CN 108510115 A (山东科技大学) 2018年 9月 7日 (2018 - 09 - 07) 全文	1-9															
A	JP 2008052495 A (SONY CORP.) 2008年 3月 6日 (2008 - 03 - 06) 全文	1-9															
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																	
<table border="0"> <tr> <td> * 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 </td> <td> “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件 </td> </tr> </table>			* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件	“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件													
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件	“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																
国际检索实际完成的日期 2020年 11月 25日		国际检索报告邮寄日期 2020年 12月 21日															
ISA/CN的名称和邮寄地址 中国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		受权官员 吴瑶 电话号码 86-10-53961339															

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2020/080472

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	105550212	A	2016年 5月 4日	无	
CN	110659395	A	2020年 1月 7日	无	
CN	108510115	A	2018年 9月 7日	无	
JP	2008052495	A	2008年 3月 6日	无	