

(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
2016 年 6 月 16 日 (16.06.2016)



(10) 国际公布号
WO 2016/090877 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2015/081147
- (22) 国际申请日: 2015 年 6 月 10 日 (10.06.2015)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201410749244.X 2014 年 12 月 9 日 (09.12.2014) CN
- (71) 申请人: 深圳大学 (SHENZHEN UNIVERSITY)
[CN/CN]; 中国广东省深圳南山区南海大道 3688 号,
Guangdong 518000 (CN)。
- (72) 发明人: 李荣华 (LI, Ronghua); 中国广东省深圳南山
区南海大道 3688 号, Guangdong 518000 (CN)。 邱
宇轩 (QIU, Yuxuan); 中国广东省深圳南山区南海
大道 3688 号, Guangdong 518000 (CN)。 毛睿
(MAO, Rui); 中国广东省深圳南山区南海大道 3688
号, Guangdong 518000 (CN)。 秦璐 (QIN, Lu); 中国
广东省深圳南山区南海大道 3688 号, Guangdong
518000 (CN)。 金檀 (JIN, Tan); 中国广东省深圳南
山区南海大道 3688 号, Guangdong 518000 (CN)。 蔡
涛涛 (CAI, Taotao); 中国广东省深圳南山区南海大
道 3688 号, Guangdong 518000 (CN)。

(74) 代理人: 深圳市兴科达知识产权代理有限公司
(CHINA ZKD IP LAW FIRM); 中国广东省深圳南山
区科技园高新南一道 008 号创维大厦 A602, Guang-
dong 518057 (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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB,
GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS,
JP, KE, KG, KN, KP, KR, 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, PT, QA,
RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST,
SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ,
VC, VN, 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)。

本国际公布:

— 包括国际检索报告 (条约第 21 条(3))。

(54) Title: GENERALIZED MAXIMUM-DEGREE RANDOM WALK GRAPH SAMPLING ALGORITHM

(54) 发明名称: 一种广义最大度随机游走图抽样算法

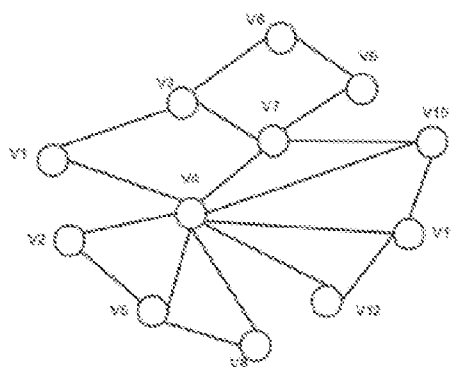


图 1 / Fig. 1

(57) Abstract: A generalized maximum-degree random walk graph sampling algorithm, comprising: random walking on a graph to acquire a sample; and structuring an unbiased estimation according to the acquired sample. The algorithm can effectively balance a large deviation problem of an RW algorithm and a sample repetition problem of an MD algorithm, thus improving the overall efficiency of acquiring sample points from a network.

(57) 摘要: 一种广义最大度随机游走图抽样算法, 其包括: 在图上随机游走采集样本; 根据采集得到的样本构造无偏估计。该算法能够有效地平衡 RW 算法的"大偏差问题"以及 MD 算法的"重复样本问题", 从而提升了从网络中采集样本点的整体效率。



WO 2016/090877 A1

一种广义最大度随机游走图抽样算法

技术领域

本发明属于大图数据挖掘技术领域，尤其涉及一种广义最大度随机游走图抽样算法。

背景技术

近年来，在线社交网络分析在学术界和工业界都引起了广泛关注。在所有在线社交网络分析的相关研究中，一个最为基本的研究问题是估计社交网络中的节点性质以及整个社交网络的拓扑特性。然而，由于很多在线的社交网络公司，例如腾讯、新浪微博、Facebook以及Twitter等，都没有向第三方发布其社交网络的图谱数据，并且整个社交图谱数据的大小对于第三方来说往往都是未知的。因此，广大从事社交网络分析的研究者和开发者都面临一个非常困难的数据采集问题。这里的主要难点在于，如何设计和开发出一种简便的方法来从一个“对于研究者不可见”的社交网络中提取出均匀的图节点样本。

为了解决这一问题，目前在学术界有很多基于爬虫技术的网络抽样方法被提出并广泛使用。可以把这些方法分为两大类：一类是基于图遍历的方法，另一类则是基于随机游走的方法。基于图遍历的方法主要是应用广度优先搜索(BFS, breadth-first search)或者深度优先搜索(DFS, depth-first search)采集节点。然而，这一类方法的主要缺点是在采集节点的过程中，算法会偏向于度比较高的节点，这显然与需要均匀的节点样本的目标不相符。并且，这一类算法对度

比较高的节点偏向多少无法从理论上刻画，因此很难纠正这一偏向，进而无法得到均匀的节点样本。目前，这一类算法逐渐被学术界和工业界弃用。基于随机游走的算法很好地解决了基于图遍历的算法的缺陷，它们可以直接生成无偏的节点样本，或者生成有偏但是偏向性已知的节点样本，故而这类算法在图采样中广受欢迎。目前有两种非常流行的基于随机游走的图抽样算法。第一种算法是重新加权的随机游走算法，称之为RW (re-weighted random walk) 算法；第二种算法是最大度随机游走算法，称之为MD (maximum-degree random walk) 算法。下面简要介绍这两种算法。

将网络抽象成一个图 $G=(V, E)$ ，其中 $n=|V|$ 代表节点的个数， $m=|E|$ 代表边的条数。令 $N(u)$ 为节点 $u \in V$ 的所有邻接节点的集合， $d_u=|N(u)|$ 表示节点 u 的度。令 $f:V \rightarrow R$ 是一个定义在节点集 V 上的实值函数，表示节点 u 的某种特性的值，例如节点的度，或者节点的某个属性值。在估计网络特性的问题中，目标是估计整个网络中所有节点的 $f(u)$ 值的平均值，记为 $E_{\pi^u}(f) = \sum_{u \in V} f(u)/n$ ，这里的 $\pi^u = [1/n, \dots, 1/n]$ 表示均匀分布。例如，如果定义 $f(u)=d_u$ ，那么 $E_{\pi^u}(f)$ 代表的是图 G 中节点度的平均值。如果定义 $f(u)=1_{\{d_u=d\}}$ ($1 \leq d \leq n-1$)，则 $E_{\pi^u}(f)$ 表示的是图 G 中节点的度分布，这里 $1_{\{d_u=d\}}$ 是一个指示函数，如果 $d_u=d$ ，则 $1_{\{d_u=d\}}=1$ ，否则 $1_{\{d_u=d\}}=0$ 。

在现有的文献中，RW和MD算法都能产生一个对 $E_{\pi^u}(f)$ 的无偏估计。RW算法是在图中执行一次随机游走来采集节点样本。众所周知，在一个非周期性的无向连通图中采用随机游走所采集到的节点样本并不

是一个均匀分布。根据随机游走的稳定分布理论，节点被选取的概率和节点的度成正比，也即对于 $u \in V$ ，有 $\pi^{rw}(u) = d_u / 2m$ ，这里的 π^{rw} 表示随机游走的稳定分布。因此，根据随机游走的采集样本策略，图中每个节点被采集到的概率是不一样的，度大的节点被采集到的概率比度小的节点被采集到的概率要大，也就是说随机游走的算法更偏向于度比较高的节点。为了纠正这种偏向性，RW算法采用了一种重新加权的策略。具体地，RW算法采用估计 $E_{\pi^{rw}}(f) = \frac{\sum_{u \in S} f(u) w^{rw}(u)}{\sum_{u \in S} w^{rw}(u)}$ （S表示采集到的样本节点的集合， $w^{rw}(u) \propto 1/d_u$ 代表节点u的权值，其中 $\propto$ 表示正比关系）来估计 $E_{\pi^u}(f)$ 。这一估计可以用重要性抽样（IS, important sampling）的框架来加以解释。具体地，IS框架采用的是相对比较容易实现的试验分布来代替目标分布采集样本节点，然后采用重要性加权来构造无偏估计。在RW算法中，目标分布是一个均匀分布 π^u ，试验分布是 π^{rw} 。根据IS框架，节点u的重要性权值为

$w^{rw}(u) \propto \pi^u(u) / \pi^{rw}(u) = 2m / nd_u \propto 1/d_u$ 。因此，根据IS框架，可以得到估计 $E_{\pi^{rw}}(f)$ ，并且在理论上可以证明 $E_{\pi^{rw}}(f)$ 是渐进无偏的。也即，当 $n \rightarrow \infty$ 时 $E_{\pi^{rw}}(f) \rightarrow E_{\pi^u}(f)$ ，而 $E_{\pi^{rw}}(f)$ 的方差取决于 $f(u) w^{rw}(u)$ 的方差。当 $f(u)$ 与 $w^{rw}(u) = \pi^u(u) / \pi^{rw}(u)$ 无关时， $E_{\pi^{rw}}(f)$ 的方差仅取决于 $\pi^u(u)$ 和 $\pi^{rw}(u)$ 的相近程度。根据“刘氏法则”，基于IS框架的抽样算法的估计精度依赖于试验分布与目标分布的卡方距离。二者的卡方距离越大，抽样算法的估计精度就越差。这里卡方距离的定义如下：令 p, q 分别为试验分布和目标分布，则 p 与 q 的卡方距离为 $\text{var}_p(q(X)/p(X))$ ，其中 var 表示方差。MD算法是一个无偏的图抽样算法，它是从一个动态构造的规则图上随

机游走采集节点，该算法能够直接得到均匀的节点样本。其原理是，通过在原始图的节点上加上自环，使得每个节点的度都等于图的最大度，生成一个规则图（节点度都相等的图称之为规则图）。当随机游走算法进行到节点 u 时，它以概率 $1/d_{\max}$ 从 u 节点的邻接节点集合 $N(u)$ 中随机选取一个节点，这里 $d_{\max}$ 表示图的最大度（度最大的节点的度）。根据这一过程，对于节点 u ，该算法将以 $(d_{\max}-d_u)/d_{\max}$ 的概率停留在原来的节点 u 上。使用重要性抽样（IS, important sampling）的框架，可知MD算法的试验分布 π^{md} 和目标分布 $\pi^u = [1/n, \dots, 1/n]$ 一致。因此，MD算法可以直接采用样本的均值来估计 $E_{\pi^u}(f)$ ，并且该估计也是渐进无偏的。

以上所述的算法中，根据IS框架，RW算法的试验分布 π^{rw} 与节点的度成正比，而目标分布是一个均匀分布 π^u 。在很多现实的社交网络中，网络的节点度往往并不均匀，而是呈现长尾现象。因此，在很多应用中，RW算法的试验分布 π^{rw} 和目标分布 π^u 有很大的偏离。根据“刘氏法则”，RW算法的有效性取决于 π^{rw} 和 π^u 的相近程度。所以在现实的网络中，RW算法往往会产生有很大的偏差，这个问题被称为“大偏差问题”（large deviation problem）。MD算法能够产生均匀的样本，因此它能够避免RW算法的“大偏差问题”。但它会产生自环

（self-loop），因而会产生很多重复的样本，并且这种情况在度比较小的节点上显得尤为严重。而过多的重复样本，通常会导致较大的估计方差，从而降低算法的估计精度，MD算法的这一缺陷被称为“重复样本问题”（repeated samples problem）。另外，在现实的很多

网络中，节点的最大度通常来说是未知的。为了解决这个问题，通常的做法是将最大度设为一个非常大的常数，从而保证该常数要大于真实的最大度。显然，这一方法会导致更多的自环，从而加重“重复样本问题”。

发明内容

本发明提供一种广义最大度随机游走图抽样算法，能够有效地平衡RW算法的“大偏差问题”以及MD算法的“重复样本问题”，从而提升了从网络中采集样本点的整体效率。

本发明通过以下技术手段实现：

一种广义最大度随机游走图抽样算法，包括以下步骤：

S1，在图上随机游走采集样本；采集到样本点集 S ；在图中随机选择节点 u 设为初始节点，并且将计数器 i 置为 1；使用 $d_u / \max\{d_u, C\}$ 作为参数生成一个几何随机变量 ξ_i 并加入集合 ξ ；将节点 u 作为 S_i ，并加入样本点集 S ；从节点 u 的邻接节点中等概率随机选取一个节点 v ；将节点 v 作为下一步的节点 u ，计数器 i 加 1，返回采集到的样本点集 S 和相应的几何随机变量集 ξ ；循环执行直至不满足条件；

S2，根据采集得到的样本构造无偏估计；构造无偏估计的公式为：

$$E_{\pi^{\text{gmd}}}(\mathbf{f}) = \frac{\sum_{i=1}^{|S|} f(S_i) \xi_i / \max\{d_{S_i}, C\}}{\sum_{i=1}^{|S|} \xi_i / \max\{d_{S_i}, C\}}$$

其中， S_i 表示算法收集到的第 i 个节点， ξ_i 指用来表示样本 S_i 的重复次数。

其中，在图上随机游走采集样本的概率转移方程如下：

$$P_{u,v}^{\text{GMD}} = \begin{cases} \frac{1}{\max\{d_u, C\}}, & \text{当 } v \text{ 为 } u \text{ 的邻接节点} \\ 0, & \text{其他情况} \end{cases}$$

其中 d_u 表示节点 u 的度， C 是一个非负整数。

以上的广义最大度随机游走算法，简称为 GMD 算法，能够有效地解决从一个“隐藏”的在线社交网络中提取均匀样本的问题，它很好地平衡了 RW 算法的“大偏差问题”，以及 MD 算法的“重复样本问题”。基于此，GMD 算法可以取代现有的广泛使用的 RW 以及 MD 算法来解决在线社交网络的抽样问题。

附图说明

图 1 为待进行随机游走算法样本采集的示意图。

具体实施方式

以下将结合具体的附图对本发明具体的实施方式进行详细说明。

本发明提供了一种新的广义最大度随机游走算法，以下简称 GMD 算法。

GMD 算法在 MD 算法之上引入一个参数 C (C 为一个非负整数) 来控制自环的数目，它的概率转移方程如下：

$$P_{u,v}^{\text{GMD}} = \begin{cases} \frac{1}{\max\{d_u, C\}}, & \text{当 } v \text{ 为 } u \text{ 的邻接节点} \\ 0, & \text{其他情况} \end{cases}$$

其中 C 是一个非负整数。

具体地，GMD 算法包括两个步骤：首先，通过上述的转移概率在图上随机游走采集样本；其次，根据采集得到的样本构造无偏估计。

其中，第一步的详细过程如下所示：

输入：图 $G = (V, E)$

输出：采集到的样本点集 S

1 在图中随机选择节点 u 设为初始节点，并且将计数器 i 置为 1

2 循环执行直至不满足条件

2.1 使用 $d_u / \max\{d_u, C\}$ 作为参数生成一个几何随机变量 ξ_i 并加入 ξ ；

2.2 将节点 u 作为 S_i ，并加入集合 S ；

2.3 从 u 的邻接节点中等概率随机选取一个节点 v ；

2.4 将节点 v 作为下一步的节点 u

2.5 计数器 i 加 1

3 返回采集到的样本点集 S 和相应的几何随机变量集 ξ

在这一步骤中，由于采用了一个几何随机变量来模拟随机游走算法在自环上停留的次数，使得随机游走算法不必真实地去游走自环，从而提升了算法的效率。换句话说，该随机游走算法中的几何随机变量 ξ_i 代表了样本 S_i 的重复次数。

采集完节点样本后，GMD算法通过以下公式构造无偏估计：

$$E_{\pi^{\text{gmd}}}(\mathbf{f}) = \frac{\sum_{i=1}^{|S|} f(S_i) \xi_i / \max\{d_{S_i}, C\}}{\sum_{i=1}^{|S|} \xi_i / \max\{d_{S_i}, C\}}$$

其中, S_i 表示算法收集到的第 i 个节点, ξ_i 指用来表示样本 S_i 的重复次数。

显然, 相对于 MD 算法, GMD 算法在每个图节点上添加的自环数要少于 MD 算法。因此, GMD 算法能够在一定程度上解决 MD 算法的“重复样本问题”。而且, GMD 算法还可以解决 MD 算法中的最大度未知的问题。此外, 还可以证明 GMD 算法的试验分布与目标分布 (均匀分布) 的卡方距离较 RW 算法的试验分布与目标分布的卡方距离要小。因此, GMD 算法也能够一定程度上解决 RW 算法的“大偏差问题”。

以下详细证明这个结论。

定理: $\text{var}_{\pi^{\text{gmd}}}(\frac{\pi(u)}{\pi^{\text{gmd}}(u)}) \leq \text{var}_{\pi^{\text{rw}}}(\frac{\pi(u)}{\pi^{\text{rw}}(u)})$, 其中的 $\pi(u)$ 为均匀分布, 即

$$\pi(u) = 1/n.$$

证明: 首先容易得到 $E_{\pi^{\text{gmd}}}[\pi(u)/\pi^{\text{gmd}}(u)] = \sum_{u \in V} \pi(u) = 1$ 。同样地, 有 $E_{\pi^{\text{rw}}}[\pi(u)/\pi^{\text{rw}}(u)] = 1$ 。因此, 要证明定理成立, 只需要证明 $E_{\pi^{\text{gmd}}}[(\pi(u)/\pi^{\text{gmd}}(u))^2] \leq E_{\pi^{\text{rw}}}[(\pi(u)/\pi^{\text{rw}}(u))^2]$ 成立即可。

具体地, 有

$$\begin{aligned} & E_{\pi^{\text{gmd}}}[(\pi(u)/\pi^{\text{gmd}}(u))^2] - E_{\pi^{\text{rw}}}[(\pi(u)/\pi^{\text{rw}}(u))^2] \\ &= \sum_{u \in V} \pi^2(u)/\pi^{\text{gmd}}(u) - \sum_{u \in V} \pi^2(u)/\pi^{\text{rw}}(u) \\ &= \sum_{u \in V} \pi^2(u)[1/\pi^{\text{gmd}}(u) - 1/\pi^{\text{rw}}(u)] \\ &= \sum_{u \in V} \pi^2(u) \left[\frac{\sum_{v \in V} \pi^{\text{gmd}}(u)}{\pi^{\text{gmd}}(u)} - \frac{\sum_{v \in V} \pi^{\text{rw}}(v)}{\pi^{\text{rw}}(u)} \right] \\ &= \sum_{u \in V} \sum_{v \in V} \pi^2(u) \left[\frac{\pi^{\text{gmd}}(v)}{\pi^{\text{gmd}}(u)} - \frac{\pi^{\text{rw}}(v)}{\pi^{\text{rw}}(u)} \right] \end{aligned}$$

根据定义, 有 $\pi^{\text{rw}}(v)/\pi^{\text{rw}}(u) = d_v/d_u$,

$$\pi^{\text{gmd}}(v) / \pi^{\text{gmd}}(u) = \max\{d_v, C\} / \max\{d_u, C\}。$$

$$\text{令 } g(u, v) = \pi^2(u) [\pi^{\text{gmd}}(v) / \pi^{\text{gmd}}(u) - \pi^{\text{rw}}(v) / \pi^{\text{rw}}(u)]。$$

对任意 $u, v \in V$, 令 $h(u, v) = g(u, v) + g(v, u)。$

为了证明 $E_{\pi^{\text{gmd}}}[(\pi(u) / \pi^{\text{gmd}}(u))^2] \leq E_{\pi^{\text{rw}}}[(\pi(u) / \pi^{\text{rw}}(u))^2]$, 只需证明 $h(u, v) \leq 0$

即可。显然, 当 $u=v$ 时, 有 $h(u, v)=0$ 。当 $u \neq v$ 时, 有:

$$h(u, v) = \frac{\max\{d_v, C\}}{\max\{d_u, C\}} - \frac{d_v}{d_u} + \frac{\max\{d_u, C\}}{\max\{d_v, C\}} - \frac{d_u}{d_v}。 \text{不失一般性, 令 } d_u \geq d_v。 \text{考虑}$$

以下三种情况:

(1) 如果 $d_u \geq d_v \geq C$, 有 $h(u, v)=0$;

(2) 如果 $d_u \geq C \geq d_v$, 有 $h(u, v) = \frac{C}{d_u} - \frac{d_v}{d_u} + \frac{d_u}{C} - \frac{d_u}{d_v} = \frac{(C-d_v)(Cd_v-d_u^2)}{Cd_u d_v} \leq 0$;

(3) 如果 $C \geq d_u \geq d_v$, 有 $h(u, v) = 1 - \frac{d_v}{d_u} + 1 - \frac{d_u}{d_v} \leq 0$ 。

综上所述, 有 $h(u, v) \leq 0$ 。

证明完毕。

以下进一步举例说明本发明。即通过介绍当 $C=0.5*d_{\max}=4$ 时, 广义最大度随机游走算法 (GMD算法) 从图1中抽取2个节点的具体实施过程, 以及通过抽取的样本节点估计图1中节点度的平均值的计算过程来说明GMD算法的算法流程。抽取更多的节点样本, 以及其它C值的情况与本例类似。

$$(1) \text{ 通过状态转移概率矩阵 } P_{u,v}^{\text{GMD}} = \begin{cases} \frac{1}{\max\{d_u, C\}}, & \text{当 } v \text{ 为 } u \text{ 的邻接节点} \\ 0, & \text{其他情况} \end{cases}$$

对图进行一次随机游走, 采集节点样本集合。

输入：图1

输出：采集得到包含2个样本点的集合S

- 1 在图中随机选择节点u设为初始节点。假设选择 v_1 作为初始节点，并且将计数器i置为1
- 2 使用 $d_u / \max\{d_u, C\} = d_{v_1} / \max\{d_{v_1}, C\} = 2 / \max\{2, 4\} = 0.5$ 生成一个几何随机变量 ξ_1 并加入 ξ ；不妨假设这里生成的几何随机变量 $\xi_1 = 2$ 。
- 3 将节点 v_1 作为 S_1 加入集合S;
- 4 从 v_1 的邻接节点中等概率随机选取一个节点v。假设选择的邻居节点为 v_4 。
- 5 将 v_4 作为下一步操作的初始节点
- 6 使用 $d_u / \max\{d_u, C\} = d_{v_4} / \max\{d_{v_4}, C\} = 8 / \max\{8, 4\} = 1$ 生成一个几何随机变量 ξ_2 并加入 ξ ；不妨假设这里生成的几何随机变量 $\xi_2 = 1$ 。
- 7 将节点 v_4 作为 S_2 加入集合S;
- 8 从 v_4 的邻接节点中等概率随机选取一个节点v。将其作为下一步操作的初始节点。
- 9 样本点采集完毕，采集过程结束。此时 $S = \{v_1, v_4\}$ ， $\xi = \{2, 1\}$

(2) 对采集到的样本点集通过 $E_{\pi^{\text{gmd}}}(\mathbf{f}) = \frac{\sum_{i=1}^{|\mathcal{S}|} f(S_i) \xi_i / \max\{d_{S_i}, C\}}{\sum_{i=1}^{|\mathcal{S}|} \xi_i / \max\{d_{S_i}, C\}}$ 来

估计图1中节点度的平均值。这里

$$E_{\pi^{\text{gmd}}}(\mathbf{f}) = \frac{\sum_{i=1}^{|\mathcal{S}|} f(S_i) \xi_i / \max\{d_{S_i}, C\}}{\sum_{i=1}^{|\mathcal{S}|} \xi_i / \max\{d_{S_i}, C\}} = \frac{2 \times 2 \times \frac{1}{4} + 8 \times 1 \times \frac{1}{8}}{2 \times \frac{1}{4} + 1 \times \frac{1}{8}} = 3.2, \text{ 说明由这个样本集}$$

估计图1中节点度的平均值为3.2。

由上可知，广义最大度随机游走算法，即 GMD 算法能够有效地解决从一个“隐藏”的在线社交网络中提取均匀样本的问题，它很好地平衡了 RW 算法的“大偏差问题”，以及 MD 算法的“重复样本问题”。基于此，GMD 算法可以取代现有的广泛使用的 RW 以及 MD 算法来解决在线社交网络的抽样问题。

1、一种广义最大度随机游走图抽样算法，包括以下步骤：

S1，在图上随机游走采集样本；采集到样本点集 S ；在图中随机选择节点 u 设为初始节点，并且将计数器 i 置为 1；使用 $d_u / \max\{d_u, C\}$ 作为参数生成一个几何随机变量 ξ_i 并加入集合 ξ ；将节点 u 作为 S_i ，并加入样本点集 S ；从节点 u 的邻接节点中等概率随机选取一个节点 v ；将节点 v 作为下一步的节点 u ，计数器 i 加 1，返回采集到的样本点集 S 和相应的几何随机变量集 ξ ；循环执行直至不满足条件；

S2，根据采集得到的样本构造无偏估计；构造无偏估计的公式为：

$$E_{\pi^{\text{gmd}}}(\mathbf{f}) = \frac{\sum_{i=1}^{|S|} f(S_i) \xi_i / \max\{d_{S_i}, C\}}{\sum_{i=1}^{|S|} \xi_i / \max\{d_{S_i}, C\}}$$

其中， S_i 表示算法收集到的第 i 个节点， ξ_i 指用来表示样本 S_i 的重复次数。

2、根据权利要求1所述的广义最大度随机游走图抽样算法，其特征在于：在图上随机游走采集样本的概率转移方程如下：

$$P_{u,v}^{\text{GMD}} = \begin{cases} \frac{1}{\max\{d_u, C\}}, & \text{当 } v \text{ 为 } u \text{ 的邻接节点} \\ 0, & \text{其他情况} \end{cases}$$

其中 d_u 表示节点 u 的度， C 是一个非负整数。

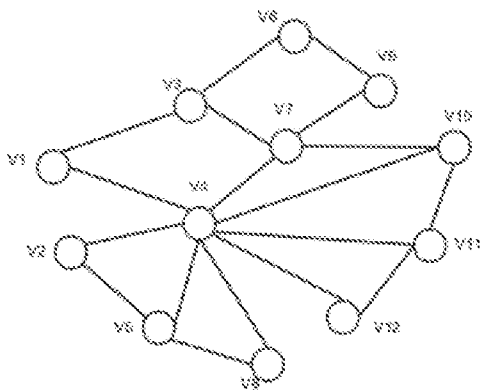


图 1

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2015/081147

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01) 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)

CNPAT, CNKI, WPI, EPODOC: random walk, graph, sample, large deviation, repeat, self loop, balance

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 104462374 A (SHENZHEN UNIVERSITY), 25 March 2015 (2015-013-25), description, paragraphs [0002]-[0071], claims 1-2, and figure 1	1-2
PX	LI, R.H., "On Random Walk Based Graph Sampling", ICDE CONFERENCE 2015, volume /, number /, 015) 0407 (015)-04-07), pages 931-933, and chapter IV	1-2
A	ZIV, B.Y. et al., "Approximating Aggregate Queries about Web Pages via Random Walks." PROCEEDINGS OF THE 26TH VLDB CONFERENCE, Cairo, Egypt, 2000., volume /, number /, 000) 0201 (000)-02-01), chapters 2-3	1-2
A	US 2014095616 A1 (7517700 CANADA INC. O/A GIRIH), 03 April 2014 (03.04.2014), the whole document	1-2
A	US 8396855 B2 (INTERNATIONAL BUSINESS MACHINES CORPORATION), 12 March 2013 (12.03.2013), the whole document	1-2

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	"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
"A" document defining the general state of the art which is not considered to be of particular relevance	"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
"E" earlier application or patent but published on or after the international filing date	"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
"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)	"&" document member of the same patent family
"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	

Date of the actual completion of the international search 17 August 2015 (17.08.2015)	Date of mailing of the international search report 09 September 2015 (09.09.2015)
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer WU, Huangfei Telephone No.: (86-10) 62413982

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2015/081147

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 104462374 A	25 March 2015	None	
US 2014095616 A1	03 April 2014	None	
US 8396855 B2	12 March 2013	US 2011295832 A1	01 December 2011
		KR 20110131094 A	06 December 2011

A. 主题的分类 G06F 17/30 (2006.01) 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)) CNPAT, CNKI, WPI, EPDOC: 随机游走, 随机行走, 图, 抽样, 取样, 样本, 大偏差, 重复, 自环, 平衡, random walk, graph, sample, large deviation, repeat, self loop, balance																				
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>PX</td> <td>CN 104462374 A (深圳大学) 2015年 3月 25日 (2015 - 03 - 25) 说明书第[0002]-[0071]段, 权利要求1-2, 图1</td> <td>1-2</td> </tr> <tr> <td>PX</td> <td>LI Rong-Hua 等, . "On Random Walk Based Graph Sampling." ICDE Conference 2015., 第/卷, 第/期, 015)0407 (015) - 04 - 07), 第931-933页, 第IV节</td> <td>1-2</td> </tr> <tr> <td>A</td> <td>ZIV Bar-Yossef 等, . "Approximating Aggregate Queries about Web Pages via Random Walks." Proceedings of the 26th VLDB Conference, Cairo, Egypt, 2000., 第/卷, 第/期, 000)0201 (000) - 02 - 01), 第2-3节</td> <td>1-2</td> </tr> <tr> <td>A</td> <td>US 2014095616 A1 (7517700 CANADA INC. O/A GIRIH) 2014年 4月 3日 (2014 - 04 - 03) 全文</td> <td>1-2</td> </tr> <tr> <td>A</td> <td>US 8396855 B2 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2013年 3月 12日 (2013 - 03 - 12) 全文</td> <td>1-2</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	PX	CN 104462374 A (深圳大学) 2015年 3月 25日 (2015 - 03 - 25) 说明书第[0002]-[0071]段, 权利要求1-2, 图1	1-2	PX	LI Rong-Hua 等, . "On Random Walk Based Graph Sampling." ICDE Conference 2015., 第/卷, 第/期, 015)0407 (015) - 04 - 07), 第931-933页, 第IV节	1-2	A	ZIV Bar-Yossef 等, . "Approximating Aggregate Queries about Web Pages via Random Walks." Proceedings of the 26th VLDB Conference, Cairo, Egypt, 2000., 第/卷, 第/期, 000)0201 (000) - 02 - 01), 第2-3节	1-2	A	US 2014095616 A1 (7517700 CANADA INC. O/A GIRIH) 2014年 4月 3日 (2014 - 04 - 03) 全文	1-2	A	US 8396855 B2 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2013年 3月 12日 (2013 - 03 - 12) 全文	1-2
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																		
PX	CN 104462374 A (深圳大学) 2015年 3月 25日 (2015 - 03 - 25) 说明书第[0002]-[0071]段, 权利要求1-2, 图1	1-2																		
PX	LI Rong-Hua 等, . "On Random Walk Based Graph Sampling." ICDE Conference 2015., 第/卷, 第/期, 015)0407 (015) - 04 - 07), 第931-933页, 第IV节	1-2																		
A	ZIV Bar-Yossef 等, . "Approximating Aggregate Queries about Web Pages via Random Walks." Proceedings of the 26th VLDB Conference, Cairo, Egypt, 2000., 第/卷, 第/期, 000)0201 (000) - 02 - 01), 第2-3节	1-2																		
A	US 2014095616 A1 (7517700 CANADA INC. O/A GIRIH) 2014年 4月 3日 (2014 - 04 - 03) 全文	1-2																		
A	US 8396855 B2 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2013年 3月 12日 (2013 - 03 - 12) 全文	1-2																		
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																				
* 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 2015年 8月 17日		国际检索报告邮寄日期 2015年 9月 9日																		
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局 (ISA/CN) 北京市海淀区蓟门桥西土城路6号 100088 中国 传真号 (86-10)62019451		授权官员 吴黄飞 电话号码 (86-10)62413982																		

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2015/081147

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	104462374	A	2015年 3月 25日	无	
US	2014095616	A1	2014年 4月 3日	无	
US	8396855	B2	2013年 3月 12日	US 2011295832 A1	2011年 12月 1日
				KR 20110131094 A	2011年 12月 6日