SGP: 一种基于图划分的社交网络采样方法

吴波

摘要

社交网络的代表性样本对于许多依赖于精确分析的互联网服务来说是必不可少的。一个好的社交网络抽样方法应该是能够生成与原始网络结构和分布相似的小样本网络。本文提出了一种基于图划分的采样算法——基于图划分的采样 (SGP)，用于对社交网络进行采样。SGP 算法首先将原始网络划分为几个子网络，然后在每个子网络中均匀采样。这一过程使 SGP 能有效地保持拓扑相似性，并能在原有网络中有效地保持对其他网络的控制。最后，我们在几个著名的数据集上评估 SGP，并与其他几种先进的图抽样算法的实验结果做了对比。

关键词：抽样算法; 社交网络; 图分区; 群落结构; 拓扑结构。

1 引言

在社交网络时代，社交网络 (Twitter、微博、MSN、Facebook、共引关系、信用网络) 随处可见。过去几年见证了在线社交网络的爆炸式增长，吸引了全世界的关注 (Wang et al.，2011)。社交网络的快速增长带来了新的挑战，因此许多研究关注在线社交网络。社交网络分析的一个瓶颈是网络数据量过大，难以处理。此外，由于隐私问题，一些网络数据无法访问。因此，我们必须发展抽样方法，从总体图中绘制有代表性的样本图。

在对社交网络的研究中，社交网络通常由不同类型的图表示。然后使用一些图挖掘技术 (图可视化技术、图结构分析技术等) 来辅助社会网络分析。然而，对于一个大规模的图，使用图挖掘方法直接处理整个图是非常困难的。寻找某种方法来加速大规模图挖掘过程是一个至关重要的问题。一个流行的解决方案是完成一个子图，它可以有效地表示原始图，以便我们能够使用这个子图进行模拟和分析[1]。在某些情况下，整个图是已知的，采样的目的是获得一个更小的图。子图的完成依赖于图采样过程。这个抽样过程的目的是选择一组顶点和边，使得到的子图遵从原始图的一些一般特征。图采样的常见目标是获取顶点的代表性子集、保留原图的某些性质。对于许多依赖于精确分析的互联网服务来说，社交网络的代表性样本是必不可少的。[2]一个好的社交网络抽样方法应该是能够生成与原始网络结构和分布相似的小样本网络。

一般来说，抽样大图会遇到 2 个问题。什么是好的抽样方法？什么是好的样本量？许多研究人员已经提出了他们的解决方案，以抽样社会网络。介绍了一些最新的抽样算法：随机节点 (VS) 抽样、随机边缘 (ES) 抽样、随机游走 (RW) 抽样、随机跳跃 (RJ) 抽样、森林火灾 (FF) 抽样、广度优先抽样 (BFS) 和其他抽样策略。在这些算法中，样本大小由用户预设，使用户可以得到自己理想的采样图。在具体的抽样过程中，保持抽样图与原图之间的相似性质是非常重要的。只有抽样图能很好地代表原始图，我们才能研究抽样图而不是原始图。如何评估采样图与原始图是否具有相似的性质? 现在有一些测量相似度的技术，我们将在后面中介绍。

本文提出了一种基于图划分 (SGP) 的社交网络采样算法。该算法首先将原始网络划分为几个子网

络，然后对每个子网络中的样本顶点进行分层。因此，该过程使 SGP 能够有效地保持采样网络与原始网络之间的拓扑相似性和群落结构相似性。

2 相关工作

在本节中，我们将分别介绍一些网络采样算法和性能评估。

2.1 图采样算法

在本节中，首先介绍最常讨论的抽样方法。顶点采样 (VS) 和边缘采样 (ES) 是两种经典的采样方法。它们也是更复杂方法的构建模块。带邻域的顶点采样操作与顶点采样类似，但邻域信息在一次探测中获得。我们还讨论了 VS 和 ES 的两种变体。基于遍历的采样 (TBS) 是一类很大的方法，因此我们在这里简要地提到它。

VS 和 ES 方法简单，适用于理论分析。它们可以有效地缩小原有社交网络的规模。然而，它们也导致了连接组件的规模如此之小，破坏了原有网络的拓扑结构。但是，在大多数实际应用中，由于各种各样的限制，例如不能枚举 ID 空间，不能直接执行 VS 和 ES。在这种情况下，TBS 变得更加实用，它只依赖于一个小型的初始拓扑 (例如几个种子节点)，并在探索过程中扩展它。注意 VS, ES 和 TBS 并不是完全不同的。某些 TBS 技术可用于生成 VS 或 ES。

基于遍历的采样有着悠久的历史，也是近年来的研究热点。基于遍历的采样器从一组初始顶点 (或边) 开始，并根据当前的观察结果展开样本。这类方法自然出现在网络爬行、隐藏人口调查等环境中。例如随机游走 (RW) 抽样，雪球抽样，应答驱动抽样，森林火灾等。RW 采样从随机选择一个顶点开始，然后在原始图上模拟 RW。RJ 采样与 RW 采样非常相似。唯一的区别是，在 RJ 采样下，我们以概率 c=0.15 随机跳转到图中的任何顶点。RW 抽样法在非二部无向图中自然地对边进行均匀抽样，而在断开图中则不能。另一方面，RW 方法及其变体倾向于高度顶点，使采样过程在局部区域进行。FF 采样是一个递归过程。首先，随机选择一个种子顶点，并开始“燃烧”传出的链接和相应的顶点。如果链接被烧毁，另一个端点上的顶点就有机会烧毁自己的链接，以此递归类推。

2.2 抽样评价方法

图采样算法使用户能够使用具有小范围影响的子图。然而，我们如何评估这些算法的性能？换句话说，我们如何评估采样图与其原始图之间的相似度？目前，已经提出了几种评价措施。测度法是通过计算抽样图的分布与原始图的相似度来表示两者的相似度。图具有一些性质，例如度分布、路径长度分布、聚类系数、k 核等等。这些属性反映了图的结构。通过比较采样图和原始图之间的这些性质，我们可以评价采样图是否具有与其原始图相似的结构，本文简要介绍了图的一些代表性性质。度分布：对于每个度 d，我们计算具有度 d 的顶点的数量。弱连接组件的大小分布：我们统计相同大小的弱连接组件的数量。跳线图：距离为 h 或小于 h 的可达节点对的个数 P(h)，其中 h 为跳数。图邻接矩阵奇异值随秩的分布：图的光谱特性通常遵循重尾分布。

对于关系图来说，拓扑结构和群体结构是两个重要的特征，它们可以反映网络的性质和社会关系。一个好的样本应该保持与原网络相似的拓扑结构和群落结构。本文重点研究了保持拓扑结构和群落结构相似的大规模图采样方法。本文提出了一种社会网络采样算法，该算法可以得到与原始网络拓扑结构和社区结构相似的样本网络。此外，我们在一些已有的评估技术上对一些知名的数据集进行了

评估。

3 本文方法

3.1 基于随机选择的抽样方法

3.1.1 随机节点抽样 (VS)

在经典的节点抽样[3]中，从 G 中独立地、均匀地随机选择节点包含在抽样图 G_s 中。我们首先直接选择 $V_s \subseteq V$ ，不包含拓扑信息 (例如；均匀地或根据 V 上的某种分布)。然后我们让 $E_s = \{(u,v) \in E | u \in V_s, v \in V_s\}$ ，即只保留采样顶点之间的边。分层抽样是调查研究中常用的一种方法, 通常只使用与顶点相关的信息，例如人口统计属性。我们也将其视为顶点采样。尽管节点抽样似乎能很好地捕获不同程度的节点，但由于它只包含了所选节点集的所有边，因此不太可能保留原有的连通性水平。

3.1.2 随机边缘抽样 (ES)

在经典的边采样中，从 G 中独立、均匀、随机地选取边，将其包含在抽样图 G_s 中。我们首先选择 $E_s \subseteq E$ 。然后令 $V_s^{(1)} = \{u, v | (u,v) \in E_s\}$ 。这个定义只出现在一些关于图的基本抽样方法的理论讨论中。不幸的是，ES 未能保留许多所需的图形属性[4]。由于边缘的独立采样，它不能保持聚类和连通性。然而，它更有可能捕获路径长度，因为它倾向于高度节点和包含所选边的两个端点。

3.2 基于图划分的抽样方法 (SGP)

3.2.1 边权计算

在社交网络图中，边可以扮演不同的角色。一些边连接同一群落的顶点，而相对较少的边连接不同群落的顶点。在本文中，我们提出了边权 E_W 来衡量边的作用。对于每条边 e ，计算 EW_e 的过程如下：

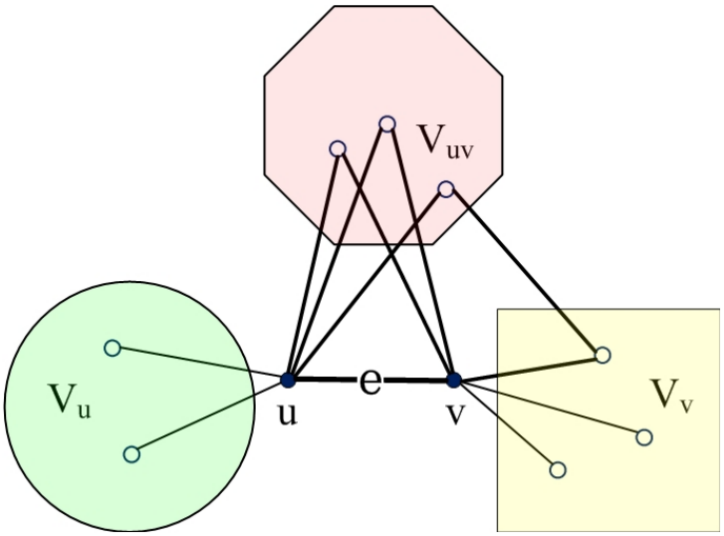

图 1: e(u,v) and its neighbours

首先，设顶点 u 和 v 是 e 的两个端点，那么如图 1所示，可以将 u 和 v 的邻居分为三组:

1. V_u 是一个顶点集合，它是 u 的邻居而不是 v 的邻居
2. V_v 是一个顶点集合，它是 v 的邻居而不是 u 的邻居
3. V_{uv} 是一个顶点集合，它是 u 和 v 的共同的邻居

那么 V_u 和 V_{uv} 或 V_v 和 V_{uv} 之间的每条边都是长度为 4 的循环的一部分，并且该循环包含边 e；V_{uv} 中的每个顶点都是包含 e 的长度为 3 的循环的一部分。所以 $(|V_{uv}|)/(|V_u|+|V_v|+|V_{uv}|)$ 可以表示长度为 3 的循环的比值。

然后，设 A 和 B 是 G 的任意两个顶点子集，并且 $N_e(A,B)$ 表示 A 和 B 之间的边数。设 $R(A,B)=N_e(A,B)/(|A|*|B|)$ 为 A 和 B 的实际边数与 A 和 B 的所有可能边数之比。特别地，我们让 $R(A)=2\left|N_e(A)\right|/(|A|*(|A|-1))$。

最后，我们可以通过以下表达式定义 G 中每个 e 的边权 EW：

$$EW_e=R(V_u,V_{uv})+R(V_u,V_v)+R(V_{uv},V_v)+R(V_{uv})+\frac{|V_{uv}|}{|V_u|+|V_v|+|V_{uv}|} \tag{1}$$

如图 1所示，V_{uv} 中的每个顶点都是长度为 3 的包含 e 的循环的一部分，所以 $(|V_{uv}|)/(|V_u|+|V_v|+|V_{uv}|)$ 可以表示包含 e 的长度为 3 的循环的数量之比。当边 e 的共同邻居集合 V_{uv} 趋于 0 时，EW_e 趋于 0。这意味着边 e 可能是两个不相交的子图的连接。相反，当 EW_e 越大时，边 e 的共同邻居集合 V_{uv} 趋于 1。那么边 e 可能被紧密地包含在某些群落中。

3.2.2 过滤边

计算完所有边权值后，根据 EW_e 对图 G 进行划分。当 EW_e 趋于 0 时，边 e 的共同邻居集合 V_{uv} 趋于 0，这意味着这条边是两个不相交的子图的连接。因此，EW_e 较小的边缘可能是两个群落之间的连接，EW_e 较大的边缘可能紧密包含在某个群落中。因此，我们通过阈值 τ 对边权值较小的边进行过滤，然后将原始图 G 划分为几个互不相交的子图。

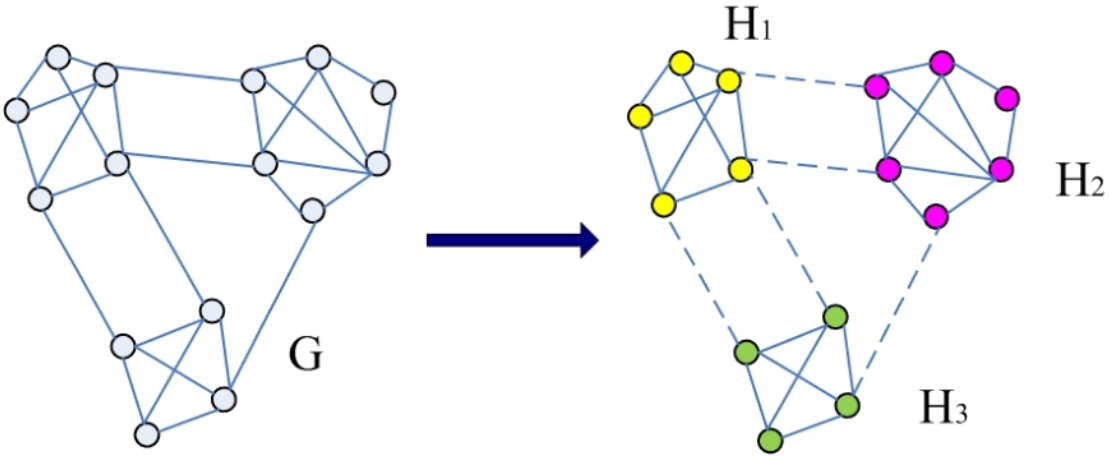

图 2: Filtering edges with smaller edge weight

如图 2所示，对用虚线表示的较小权重边进行过滤后，形成三个互不相交的子图 (G_1,G_2,G_3)。因此，通过过滤较小的权值边，可以将图 G 拆分为 q 个不相交子图 $\{G_1,G_2,\ldots,G_q\}$。

接下来我们介绍确定阈值 τ 的方法。设 $\max_{\text{EW}}$ 和 $\min_{\text{EW}}$ 分别表示 G 中的最大边权值和最小权值。然后我们让：

$$\tau=\min_{\text{EW}}+(\max_{\text{EW}}-\min_{\text{EW}})*r \tag{2}$$

其中 r 是一个参数，其经验值为 0.95。由式(2)可知，τ 的值趋于 $\max_{\text{EW}}$。这是因为对 EW_e 较大的边进行过滤，可以在很大程度上保证图划分的准确性。

然后我们对每个子图中的顶点进行采样。这样的设计可以保持样本网络与原始网络之间的共同结

构。在每个子图中采样也可以使被采样的顶点和边在 G 中全局分布，以保持拓扑结构。可以全局执行采样。也就是说，几乎在 G 的每一部分都有一些顶点。

3.2.3 图采样模型

在图论中，图中两个顶点之间的距离是连接它们的最短路径上的边的数量。这也被称为测地线距离，因为它是两个顶点之间的图测地线的长度。如果没有连接两个顶点的路径，也就是说，如果它们属于不同的连接组件，那么按照惯例，这个距离被定义为无穷大。图的直径是任意一对顶点之间的最大距离。要计算一个图的直径，首先要找出每一对顶点之间的最短路径。这些路径中最大的长度就是图的直径。图 3显示了一个直径为 7 的简单图的直径。图 3中的红色路径是任意两个顶点的所有最短路径中最长的路径之一。

图 4显示了分层抽样模型。对于原图 G 的子图 G_i，设 d 为初始图 G_i 的直径。我们在最大的路径上得到两个端点它们的长度等于直径 d 。随机选择一个端点 v_s 作为起始点后，我们将 v_s 添加到 V_S 。然后根据 v_i 到 v_s 的距离，将 V_{G_i} 划分为 d 个子集 $\{V_1, V_2, \ldots, V_d\}$, 其中 V_i 表示到 v_s 的距离为 i 的顶点集合。

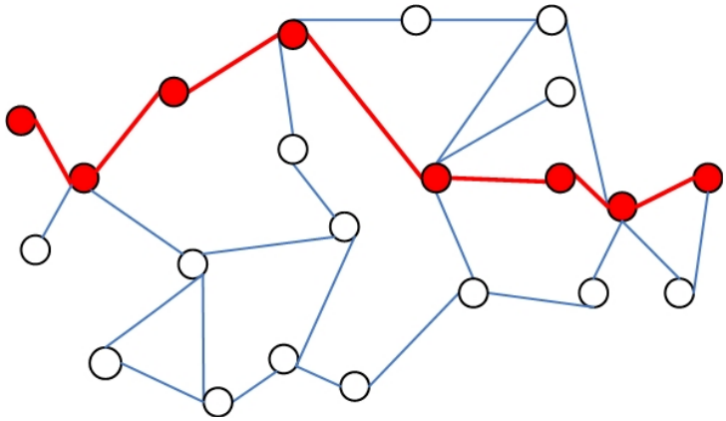

图 3: Diameter of the graph G

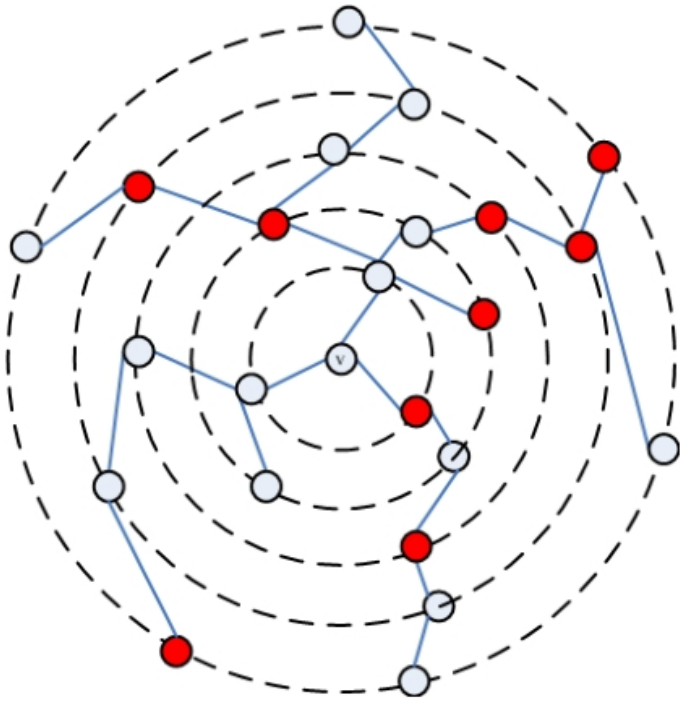

图 4: A stratified sampling model (diameter = 5)

我们将 S_i 表示为 V_i 的采样顶点集，所以 $S_0 = \{v_s\}$ 。那么 V_i 可以被拆分为两个子集 V_{i_o} 和 V_{in_o} 。V_{i_o} 中的顶点具有至少与 S_{i-1} 中的某个顶点有链接的特性，而 V_{in_o} 中的顶点具有与 S_{i-1} 中的顶点没有链接的特性。在 V_i 中采样时，我们在 V_{i_o} 中选取 k 个百分比的顶点，在 V_{in_o} 中选取剩余的顶点，参数 k 可以由用户决定，本文将 k 设为经验值 0.7。在选取顶点之后，我们将边添加到采样图中。我们之所以用分层的策略来抽样 $\{V_1, V_2, \ldots, V_d\}$ 中的顶点，是因为我们希望每个 V_i 中抽样顶点的比例几乎相同。

3.2.4 算法细节

为了清晰起见，我们将整个算法总结如下。提出的 SGP 采样算法包括两个步骤。

1. 首先，通过过滤边权值 EW_e 较小的边，将原始图划分为 q 个子图。

2. 其次，采用分层抽样模型对每个子图中的顶点进行抽样。

SGP 采样算法有一个参数：采样百分比 P (或采样大小 N)。给定一个抽样百分比 P ，我们的算法首先对原始图 G 进行分割，然后找到每个 G_i 直径的两个端点，随机选择一个端点作为起始顶点。最后，SGP 算法对每个子图中的样本顶点进行分层抽样。

3.2.5 算法拓展

我们的 SGP 模型的基本版本展示了原点图是连通图的情况。但是真正的网络并不是一直都是连接的，它可能有许多连接的组件。通过将该模型自然地扩展到实际网络中，我们提出了一种扩展方法：在每个连接组件中进行 SGP 过程。也就是说，我们必须增加一个额外的步骤，即得到原点图的连接分量。然后在每个连接组件上运行 SGP 算法。

3.3 基于遍历的抽样方法 (TBS)

3.3.1 雪球抽样 (SBS)

雪球抽样[5]在社会学研究中使用已久，可以用来对隐藏的群体（如吸毒者）进行调查。雪球使用来自随机选择的种子节点的 BFS 添加节点和边，但在雪球达到特定大小时停止。雪球抽样的过程如下：

1. 从顶点的初始集合 $V^{(0)}$ 开始。这个集合可以通过顶点上的随机样本或隐藏总体的侧面知识来获得。

2. 在阶段 i ，令 $\forall v \in V^{(i-1)}$ 命名 k 个相邻节点。如何命名根据具体的研究而有所不同。这等价于获得 $V^{(i-1)}$ 的入射边的样本。我们用 $E^{(i)}$ 来表示这些边。在这一阶段观察到的顶点为 $\tilde{V}^{(i)} = \{u, v | (u, v) \in E^{(i)}\}$ 。我们把新的顶点放到第 i 阶段，即：$V^{(i)} = \tilde{V}^{(i)} - \cup_{j=0}^{i-1} V^{(j)}$ 。

3. 这个过程持续了 t 个阶段。抽样图为 $G_s =< V_s, E_s >$ 其中 $V_s = \cup_{j=0}^{t} V^{(j)}$ 并且 $E_s = \cup_{j=1}^{t} E^{(j)}$ 。

有一种说法是 SBS 与 BFS 非常相似，BFS 完全扩展了当前顶点的邻域，而 SBS 只扩展了固定数量的邻域。滚雪球抽样准确地保持了雪球内部的网络连通性，但它存在边界偏差，因为许多外围节点（即上一轮抽样的那些节点）将丢失大量的邻居。

3.3.2 森林火灾抽样 (FFS)

Forest Fire 最初是作为一个图生成模型提出的，它捕捉了真实社会网络中一些重要的观察结果，如致密化规律、缩小直径和社区引导依恋。后来在许多后续的图采样作品中，它被简称为“森林火灾”。FFS 是 SBS 的概率版本[6]。在 SBS 中，每轮选择 k 个邻居；在 FFS 中，$K \sim \text{Geometric}(p)$ 个邻居被选择。SBS 和 FFS 通过设置 $p = \frac{1}{k}$ 关联起来，则我们有 $\text{E}[K] = k$。文章建议设置 $p = 0.7$，即每个节点平均烧毁 2.33 个边。FFS 采样算法的步骤如下：

1. 均匀随机地选择一个节点并将其添加到样本中，然后从它开始烧毁输出边，并将这些边与顶点一起添加到样本中。

2. 对每个烧毁的邻居递归地重复该过程。

3. 直到没有选择新的节点，然后选择一个新的随机节点继续该过程，直到获得所需的样本大小。

3.3.3 随机游走抽样 (RW)

随机游走算法的基本思想：从一个或一系列顶点开始遍历图，在任意一个顶点，遍历者以概率 α 游走到这个顶点的邻居。每次游走后得出一个概率分布，该概率分布刻画了图中每一个顶点被访问到的概率。用这个概率分布作为下一次游走的输入并反复迭代这一过程。当满足一定前提条件时，这个概率就会趋于收敛。收敛后，就可以得到一个平稳的概率分布。

随机游走抽样 (RW) 是无记忆的。在 SBS 中，前面阶段的参与者被排除在外。在 RW 中，可以重新访问一些顶点。RW 的无记忆性使得它更具有理论分析的吸引力，例如通过马尔可夫链分析得到平稳分布。随机游走算法的一步转移概率如下：

$$P_{vu} = \begin{cases} 1/d_v & u \in N(v) \\ 0 & otherwise \end{cases} \tag{3}$$

RW 算法对图进行随机游走，收集节点。在一个连通的、非周期的无向图中，随机游走的平稳分布与节点度成正比，由式 4可知，简单随机游走 RW 的平稳分布为 $\pi^{rw}(u) = d_u/2m$[7]。使用随机游走生成统计有效样本的前提是，在执行随机游走足够多的步数后，随机游走落在每个节点上的概率收敛到平稳分布 $\pi^{rw}(u)$ 。

$$\lim_{n\to\infty} P_{vu}^{(n)} = \begin{bmatrix} p_{11}^{(n)} & p_{12}^{(n)} & \cdots & p_{1u}^{(n)} & \cdots \\ p_{21}^{(n)} & p_{22}^{(n)} & \cdots & p_{2u}^{(n)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{v1}^{(n)} & p_{v2}^{(n)} & \cdots & p_{vu}^{(n)} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} = \begin{bmatrix} \frac{d_1}{2m} & \frac{d_2}{2m} & \cdots & \frac{d_u}{2m} & \cdots \\ \frac{d_1}{2m} & \frac{d_2}{2m} & \cdots & \frac{d_u}{2m} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{d_1}{2m} & \frac{d_2}{2m} & \cdots & \frac{d_u}{2m} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{4}$$

随机游走图抽样算法步骤如下：

1. 从一个初始顶点 $v^{(0)}$ 开始

2. 在第 i 步，选择一个相邻顶点 $u \in N(v^{(i-1)})$ (可能是均匀随机的，也可能是根据某种权重) 。设 $v^{(i)} \leftarrow u$ 为下一个顶点，并且包含 $\tilde{E}_s \leftarrow \tilde{E}_s + \{(v^{(i-1)}, v^{(i)})\}$ 这条边。

3. 重复 $t = \frac{B}{b}$ 步，返回采样图 $G_s =< V_s, E_s >$ 。根据不同的应用，有两种可能的顶点集和边集：

顶点的邻域未知：$V_s = \{v^{(0)}, v^{(1)}, \dots v^{(t)}\}$ 和 $E_s = \tilde{E}_s$ 顶点的邻域已知：设 $\tilde{V}_s = \{v^{(0)}, v^{(1)}, \dots v^{(t)}\}$ 。则 $E_s = \cup_{v \in \tilde{V}_s} \delta(v)$ 并且 $V_s = \{u, v | (u, v) \in E_s\}$。

然而，当我们想在一个有很多分量的图中随机游走抽样，所有的分量在图中都是不相连的，就需要跳跃。RJ 采样与 RW 采样非常相似。唯一的区别是，在 RJ 采样下，我们以概率 $\alpha = 0.15$ 随机跳转到图中的任何顶点，以概率 $1 - \alpha$ 随机游走当前顶点的邻居顶点。随机跳跃算法的一步转移概率如下：

$$P_{vu} = \begin{cases} \frac{\alpha}{N} + \frac{1-\alpha}{d_v} & u \in N(v) \\ \frac{\alpha}{N} & otherwise \end{cases} \tag{5}$$

3.3.4 Metropolis-Hastings 随机游走 (MHRW)

MH 算法是 Metropolis-Hastings 算法、无偏图采样的一个应用。Metropolis-Hastings 算法是马尔可夫链蒙特卡罗法的代表算法。在马尔科夫蒙特卡罗链中，Metropolis-Hastings 算法被广泛应用于从任意无向连通图中获得所需的顶点分布。它利用改进的随机游走从图中绘制节点。特别地，它修改了随机游走的跃迁概率，使其收敛为均匀分布。马尔可夫链蒙特卡罗法步骤如下：

1. 随机变量 x 的状态空间 E 上构造一个满足遍历定理的马尔可夫链，使其平稳分布为目标分布 $p(x)$

2. 从状态空间的某一点 x_0 出发，对构造的马尔可夫链进行随机游走

3. 行走时间足够长时 (时刻大于某个正整数 m)，即样本的分布收敛于平稳分布时，开始抽样得到样本 $\{x_{m+1}, x_{m+2}, \dots, x_n\}$

这里有几个重要的问题：1. 如何定义马尔可夫链，保证马尔可夫链蒙特卡罗法的条件成立。2. 如何确定收敛步数 m，保证样本抽样的无偏性。3. 如何确定迭代步数 n，保证遍历均值计算的精度。

假设要抽样的概率分布为 $\pi(x)$，MH 算法采用转移概率为 $p(x, x')$ 的马尔可夫链：

$$p(x, x') = q(x, x')\alpha(x, x') \tag{6}$$

其中 $q(x, x')$ 为建议分布，$\alpha(x, x')$ 为接受分布。建议分布 $q(x, x')$ 是另一个马尔可夫链的转移概率，并且 $q(x, x')$ 是不可约的。简而言之，一个不可约的马尔可夫链，从任意状态出发，经过充分长的时间后，可以达到任意状态。即存在一个时刻 t，$q(x, x')$ 满足 $P(X_t = x' | X_0 = x) > 0$ 。也就是说，时刻 0 从状态 x 出发，时刻 t 到达状态 x' 的概率大于 0，则称此马尔可夫链是不可约的，否则称此马尔可夫链是可约的。

转移概率为 $p(x, x')$ 的马尔可夫链上的随机游走以下方式进行：

1. 在时刻 $t-1$ 处于状态 x，即 $x_{t-1} = x$，则先按建议分布 $q(x, x')$ 抽样产生一个候选状态 x' 。

2. 按照接受分布 $\alpha(x, x')$ 抽样决定是否接受状态 x' 。以概率 $\alpha(x, x')$ 接受 x'，决定时刻 t 转移到状态 x'，以概率 $1 - \alpha(x, x')$ 拒绝 x'，决定时刻 t 停留在状态 x 。具体地，从区间（0，1）上的均匀分布中抽取一个随机数 u，决定时刻 t 的状态。

$$x_t = \begin{cases} x' & u \leqslant \alpha(x, x') \\ x & u > \alpha(x, x') \end{cases} \tag{7}$$

可以证明，转移概率为 $p(x, x')$ 的马尔可夫链是可逆马尔可夫链（满足遍历定理），其平稳分布就

是 $\pi(x)$，即要抽样的目标分布。

$$\pi(x)p(x,x') = \pi(x')p(x',x) \tag{8}$$

其中 $\pi(x)$ 是该马尔可夫链的平稳分布。若 $x = x'$，则上式显然成立。现在假设 $x \neq x'$，则：

$$\begin{aligned}
\pi(x)p(x,x') &= \pi(x)q(x,x')\min\left\{1, \frac{\pi(x')q(x',x)}{\pi(x)q(x,x')}\right\} \\
&= \min\{\pi(x)q(x,x'), \pi(x')q(x',x)\} \\
&= \pi(x')q(x',x)\min\left\{\frac{\pi(x)q(x,x')}{\pi(x')q(x',x)}, 1\right\} \\
&= \pi(x')p(x',x)
\end{aligned} \tag{9}$$

由式 9可知：

$$\begin{aligned}
\int \pi(x)p(x,x')dx &= \int \pi(x')p(x',x)dx \\
&= \pi(x')\int p(x',x)dx \\
&= \pi(x')
\end{aligned} \tag{10}$$

根据式 10以及平稳分布的定义，$\pi(x)$ 就是该马尔可夫链的平稳分布。

在 MHRW 图抽样中，建议分布 $q(v,u)$ 的转移概率为：

$$q(v,u) = \begin{cases} 1/d_v & u \in N(v) \\ 0 & otherwise \end{cases} \tag{11}$$

显然，在连通图中，$q(v,u)$ 是不可约的马尔可夫链。抽样的目标分布取 $\pi(v) = \pi(u) = 1/N$，则接受分布 $\alpha(v,u)$ 为：

$$\alpha(v,u) = \min\left\{1, \frac{\pi(u)q(u,v)}{\pi(v)q(v,u)}\right\} = \min\left\{1, \frac{d_v}{d_u}\right\} \quad ,u \in N(v) \tag{12}$$

综上所述，MHRW 的转移概率 P_{vu}^{MH} 为：

$$P_{vu}^{MH} = \begin{cases} 1/d_v \times \min\{1, d_v/d_u\} & u \in N(v) \\ 1 - \sum_{u \neq w} P_{vw}^{MH} & u = v \\ 0 & otherwise \end{cases} \tag{13}$$

Metropolis-Hastings RW 算法采样的步骤如下：

1. 随机选择一个初始顶点开始游走，直到样本的分布收敛于平稳分布。

2. 达到平稳分布后开始抽样，假设在第 k 步，遍历者处理顶点 v，根据建议分布 $q(v,u)$ 随机选取一个候选邻居顶点 u。

3. 根据接受分布 $\alpha(v,u)$ 接收或者拒绝顶点 u。

4. 重复上述过程，直到获得所需的样本大小。

最后，Metropolis-Hastings RW 抽样算法得到的样本分布为均匀分布：

$$\pi^{MH} = \left(\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N}\right) \tag{14}$$

当使用 MHRW 时，样本均值可以作为所有节点均值的无偏估计，即由 MH 得到的样本均值可以直接用于构造 $\mathbb{E}_{\pi^{MH}}(f)$ 的无偏估计量。有一点值得注意的是，MHRW 的应用场景比普通 RW 更有限。为了计算转移概率，必须知道相邻顶点的程度。这个信息通常是不可用的。然而，在某些情况下，这是

可能的。例如，在 P2P 网络中，对等体的数量通常是一个固定的参数。再比如，一些 osn 的 API 可能会返回除了 id 之外的丰富信息的好友列表（例如关注者和被关注者的数量）。

当然，和普通的随机游走一样，当我们想在一个有很多分量的图中进行随机游走时，所有的分量在图中都是不相连的，就需要跳跃。在 MHRJ 采样下，我们以概率 $d = 0.85$ 随机游走当前顶点的邻居顶点[8]，以概率 $1 - d$ 随机跳转到图中的任何顶点。所以建议分布 $q(v, u)$ 为：

$$q(v,u) = \begin{cases} \dfrac{1-d}{N} + \dfrac{d}{d_v} & u \in N(v) \\ \dfrac{1-d}{N} & otherwise \end{cases} \tag{15}$$

在带跳跃的均匀行走中，如果节点 v 和 u 之间有边，则接受分布 $\alpha(v, u)$ 为：

$$\alpha_{vu} = \min\left(\frac{\dfrac{1-d}{N} + \dfrac{d}{d_u}}{\dfrac{1-d}{N} + \dfrac{d}{d_v}}, 1 \right) \tag{16}$$

如果节点 v 和 u 之间没有边，则接受概率为 $\alpha_{vu} = 1 - d$ 。

3.4 Hansen-Hurwitz 估计量

在一个连通的、非周期的无向图中，普通随机游走（RW）的平稳分布与节点度成正比，即 $\pi^{rw}(u) = d_u/2m, u \in V$。基于这一事实，随机游走收集到的节点都倾向于高度节点。无论我们如何获得样本 (通过 VS, ES 或 TBS) 都无关紧要。只要我们知道样本概率，我们就可以使用偏差校正技术。为了纠正这种偏差，Re-weighted RW 算法使用了一种重加权策略，该策略可以使用著名的 HanseHurwitz 估计器。

假设我们有一个样本 $|V_s| = n_s$ ，其分布为 $\pi(v), \forall v \in V$ 。允许对单个顶点重复采样。现在我们要估计参数：

$$\theta = \frac{1}{n} \sum_{v \in V} g(v) \tag{17}$$

其中 $g(v)$ 是生成顶点标签的函数。（例如生成顶点的度的函数 $g(v) = d(v)$）。注意朴素估计量：

$$T_1 = \frac{1}{n_s} \sum_{v \in V_s} g(v), \quad n_s \to \infty, T_1 \to \sum_{v \in V} \pi(v) g(v) \tag{18}$$

在 T_1 表达式中，我们要把 $g(v)$ 换成 $h(v)$ ，我们期望：

$$n_s \to \infty \Rightarrow T_2 = \frac{1}{n_s} \sum_{v \in V_s} h(v) \to \theta = \frac{1}{n} \sum_{v \in V} g(v) \tag{19}$$

我们可以证明 $h(v)$ 有以下一个选择：

$$h(v) = \frac{g(v)}{n\pi(v)} \tag{20}$$

当 $n \to \infty$ 时：

$$T_2 = \frac{1}{n_s} \sum_{v \in V_s} h(v) = \mathrm{E}_\pi \left[\frac{g(v)}{n\pi(v)} \right] = \sum_{v \in V} \frac{g(v)}{n\pi(v)} \pi(v) = \frac{1}{n} \sum_{v \in V} g(v) \tag{21}$$

T_2 估计器仍然不能使用。有两个困难:

(1)n 在许多情况下可能是未知的。我们必须用 π 来估计 n 。令 $g(v) = 1, \forall v \in V$ ，我们知道 $\frac{1}{n_s} \sum_{v \in V_s} \frac{1}{n\pi(v)}$ 是 1 的一致估计量，即：

$$\frac{1}{n_s} \sum_{v \in V_s} \frac{1}{n\pi(v)} = 1 \Rightarrow \frac{1}{n_s} \sum_{v \in V_s} \frac{1}{\pi(v)} = n \tag{22}$$

所以我们找到了 n 的一致估计量。

$$\hat{n} = \frac{1}{n_s} \sum_{v \in V_s} \frac{1}{\pi(v)} \tag{23}$$

设参数 θ 的估计量为：

$$\hat{\theta} = \frac{1}{n_s} \sum_{v \in V_s} \frac{g(v)}{\hat{n}\pi(v)} \tag{24}$$

（2）由 RW 输出的平稳分布为 $\pi^{rw}(v) = d_v/2m$ ，$d(v)$ 可以在 TBS 过程中获得。然而，在大多数情况下，m 是未知的。幸运的是，用上面的估计量 $\hat{n}$ ，m 正好消掉，我们得到：

$$\hat{\theta} = \frac{1}{n_s} \sum_{v \in V_s} \frac{g(v)}{\frac{d(v)}{2m}} \frac{1}{\frac{1}{n_s} \sum_{v \in V_s} \frac{1}{\frac{d(v)}{2m}}} = \frac{1}{\sum_{v \in V_s} \frac{1}{d(v)}} \sum_{v \in V_s} \frac{g(v)}{d(v)} \tag{25}$$

在解决了这两个问题之后，HanseHurwitz 估计器与 RW 相结合是一种在不知道图的完整情况下估计图属性的实用方法。通过将 $g(v)$ 替换为特定于应用程序的函数。只要我们能从分布 π 中抽取样本，并且知道 $v, u \in V_s$ 中所有 $\pi(v)$ 与 $\pi(u)$ 的相对比值，就可以构造出类似的估计量。

3.5 抽样方法评价

网络采样方法的准确性通常通过比较结构网络统计量（例如，度）来衡量。我们首先定义一套常用的网络统计信息，然后讨论如何使用它们来定量比较抽样方法。

3.5.1 图的经典性质

通常考虑的网络统计可以从两个维度进行比较：局部统计与全局统计，点统计与分布。局部统计量用于描述局部图元素（例如，节点，边，子图）的特征；例如，节点度和节点聚类系数。另一方面，全局统计量用来描述整个图的一个特征；例如，全局聚类系数和图直径。类似地，点统计和分布之间也有区别。点统计量是单值统计量（例如直径），而分布是多值统计量（例如所有节点对的路径长度分布）。显然，一系列网络统计信息对于研究完整的图结构非常重要。我们首先定义一套常用的网络统计信息，然后讨论如何使用它们来定量比较抽样方法。

（1）度分布 (Degree distribution)：k 度节点的比例，对于所有 $k > 0$ 的节点：

$$p_k = \frac{|\{v \in V \mid \deg(v) = k\}|}{N} \tag{26}$$

为了理解图的连通性，度分布已经被许多研究者广泛研究。许多现实世界的网络被证明具有幂律度分布，例如在 Web 、引用图和在线社交网络。

（2）路径长度分布 (Path length distribution)：同时也被称为跳点分布 (hop plot distribution) ，最短路径距离 $dist(u, v)$ 为 h 的可达节点对所占的比例。对所有的 $h > 0$ 并且 $h \neq \infty$ ：

$$p_h = \frac{|\{(u, v) \in V \mid \text{dist}(u, v) = h\}|}{N^2} \tag{27}$$

路径长度分布对于了解节点之间的路径数量如何作为距离的函数 (即跳数) 扩展是非常重要的。

（3）聚类系数分布 (Clustering coefficient distribution)：聚类系数为 $(cc(v))$ c 的节点的所占比例，对所有的 $0 \leqslant c \leqslant 1$ ：

$$p_c = \frac{|\{v \in V' \mid cc(v) = c\}|}{|V'|}, \quad \text{where} \quad V' = \{v \in V \mid \deg(v) > 1\} \tag{28}$$

聚类系数是描述一个图中的某节点与其相连节点之间聚集成团的程度的一个系数。聚类系数的目标是比较群组的聚合紧密程度与其能够达到的聚合紧密程度。聚类系数的研究范围是无向图。聚类系数的计算公式如下：

$$cc(v) = \frac{2R_v}{N(v)(N(v)-1)} \tag{29}$$

其中，$cc(v)$ 为聚类系数，R_v 为 v 邻居的关系数（三角形计数），$N(v)$ 为 v 邻居的数量。聚类系数越大，则说明节点的邻居之间联系越紧密，即节点及其邻居聚集成团的可能性就越大。在社交网络和许多其他真实网络中，节点倾向于聚集。因此，聚类系数是捕获图的传递性的重要度量。

（4）K 核分布 (K-core distribution)：图中核心度为 k 的节点所占比例：

$$p_k = \frac{|\{v \in V \mid \text{K-core}(v) = k\}|}{N} \tag{30}$$

首先介绍一下 K 核子图的概念，K 核子图是指一张图的所有核心度不小于 k 的节点构成的子图。如果一个节点属于其所在图的 K 核子图，但不属于 K+1 核子图，则称该节点的核心度（或核数 Coreness）为 k。K 核常用来识别和提取图中的紧密连通群组。研究 K 核是社会网络分析的重要组成部分，因为它们展示了图的连通性和社区结构。

3.5.2 Kolmogorov-Smirnov 检验

Kolmogorov-Smirnov 检验[9]（K-S 检验）是对连续一维概率分布是否相等的一种非参数检验，可用于比较样本与参考概率分布（单样本 K-S 检验），或比较两个样本（双样本 K-S 检验）。KS 统计量是一个被广泛用于测量两个分布之间的一致性的统计量，它可以计算两个分布之间的最大垂直距离，其中 x 表示随机变量的范围，F_1 和 F_2 表示两个累积分布函数：

$$KS\left(F_1, F_2\right) = \max_x |F_1(x) - F_2(x)| \tag{31}$$

检验值越小，两个样本服从相同分布的概率越大。因此，我们在本文中使用 K-S 检验来度量两个分布的相似度，用 K-S 检验比较了原始图与抽样图之间的度分布、Hop plot 分布、聚类系数分布和 k 核分布。

3.5.3 斜散度 (Skew Divergence)

在介绍斜散度 (SD) 之前，先介绍一下 KullbackLeibler(KL) 散度的概念。KL 散度[5]，是一个用来衡量两个概率分布的相似性的一个度量指标。我们知道，现实世界里的任何观察都可以看成表示成信息和数据，一般来说，我们无法获取数据的总体，我们只能拿到数据的部分样本，根据数据的部分样本，我们会对数据的整体做一个近似的估计，而数据整体本身有一个真实的分布（我们可能永远无法知道）。那么近似估计的概率分布和数据整体真实的概率分布的相似度，或者说差异程度，可以用 KL 散度来表示。KL 散度又可称为相对熵，描述两个概率分布 P 和 Q 的差异或相似性，用 $D_{KL}(P\|Q)$ 表

示：

$$\begin{aligned} D_{KL}(P\|Q) &= H(P,Q) - H(P) \\ &= \sum_i P(x_i) \log \frac{1}{Q(x_i)} - \sum_i P(x_i) \log \frac{1}{P(x_i)} \\ &= \sum_i P(x_i) \log \frac{P(x_i)}{Q(x_i)} \end{aligned} \tag{32}$$

散度越小，说明概率 Q 与概率 P 之间越接近，那么估计的概率分布与真实的概率分布也就越接近。KL 散度有两个重要的性质：（1）非对此性：$D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$；（2）$D_{KL}(P\|Q) \geqslant 0$，仅在 $P = Q$ 时等于 0。性质 2 是十分重要的，可以用 Jensen 不等式证明。

$$\sum_i \lambda_i f(x_i) \geqslant f\left(\sum_i \lambda_i x_i\right) \tag{33}$$

斜散度 (SD)[10]用于评估两个概率密度函数之间的差异，即用于测量两个概率密度函数 P_1 和 P_2 之间的 KullbackLeibler(KL) 散度，在整个值范围内没有连续支持。当使用抽样分布时，KL 度量表示原始分布中的样本所需的额外比特的平均数。然而，由于 KL 散度没有为具有不同支持区域的分布定义，所以在计算 KL 散度之前，倾斜散度会使两个概率密度函数平滑：

$$SD(P_1, P_2, \alpha) = KL[\alpha P_1 + (1-\alpha)P_2 \| \alpha P_2 + (1-\alpha)P_1] \tag{34}$$

在非光滑分布上，使用 SD 近似 KL 散度的结果优于其他方法。在本文中，我们使用 $\alpha = 0.99$ 。

4 复现细节

4.1 与已有开源代码对比

郑重声明，没有参考任何相关源代码。

4.2 实验环境搭建

本次实验使用了 Pycharm，Python3.9.12 以及相应的库。

4.3 算法细节

关于 SGP 算法，论文中设置的图划分算法的过滤边的阈值过高，对群落属性很明显的网络划分得较为准确，对群落属性不明显的网络容易划分出很多孤立的顶点。因此，图划分算法在实际实现中，会根据网络的类型适当调整过滤边的阈值。分层抽样模型并未提及待采样集合 V_i 中的顶点与已采样集合 S_{i-1} 的顶点怎样才算有链接的特性，算法中认为两个顶点有边即为有链接的特性。分层抽样模型中，如果某个层次的顶点数量过少，程序中可能会计算出那一层抽取顶点的数量为 0，进而导致分层抽样断层。因此，算法中设置分层抽样模型对每一层顶点最少采样一个顶点，保证每一个层次都能抽样到一些顶点。此外，论文中并未提及采样到顶点以后，如何增加边，算法实现中认为拥有原始图全图的知识，抽取到顶点以后，根据原始图的信息补充所有原始图中存在的边。

关于随机游走算法，MHRW 的关键是如何定义马尔可夫链，保证马尔可夫链蒙特卡罗法的条件成立；如何判断马尔可夫链是否达到收敛状态，即确定收敛步数 m ，保证样本抽样的无偏性。如何定义满足条件的马尔可夫链以及证明已经在本文前面给出，判断马尔可夫链的收敛性通常是靠经验的。由之前的推导可知，MHRW 抽样的平稳分布为均匀分布，理论上收敛于平稳分布后，每个相等时间段内抽取到的样本均值可以作为所有节点均值的无偏估计，即每个相等时间段内抽到的样本均值相

等。因此，在实际的算法实现中，为了校验马尔可夫链是否收敛，先让遍历者行走足够大的步数 (原始网络顶点的十倍步数) 后，在每个相等长度的时间段内抽取一个顶点集 V_i ，最后获得一组顶点集 $\{V_1, V_2, \ldots, V_k\}$ ，如果 $E[g(v)], v \in V_i$ 稳定后，则可以认为马尔可夫链已经收敛，其中 $g(v)$ 设置为计算顶点 v 的度数的标签函数。算法实现中判断均值稳定的方法为百分数衡量法，百分数衡量法通过分析参数最大值和最小值差值占参数均值的百分值来判断稳定性，MHRW 算法中设置该百分值为 5%，即当样本最大均值和样本最小均值的差值不超过所有样本均值的 5%，就可以认为马尔可夫链达到了收敛状态。

4.4 创新点

关于 SGP 算法，划分群落之后，使用分层抽样模型对每个群落进行采样，不能很好的保持社区内部网络的连通性，容易抽样出孤立节点。然而，不跳跃的随机游走抽样算法，能够很好的保持单个社区网络的连通性，一定能够抽样出一个连通的子图，不会抽出孤立节点。因此，这里对 SGP 算法做出改进，在将原始网络划分为若干个群落以后，对每个群落执行随机游走抽样算法，这样既能保持原始网络的群落结构，又能保证每个群落的连通性。

接下来，在 5 个数据集（这 5 个数据集的详细描述将在第五章节给出）上运行原始的 SGP 算法以及改进后的算法 GPRW，并从 4 个方面对其抽样效果进行评价。图 5和图 6展示了采样图与原始图相应 4 个属性的概率分布，表 1和表 2分别给出了采样图与总体图详细的 KS、SD 偏差，对比实验结果如下所示。

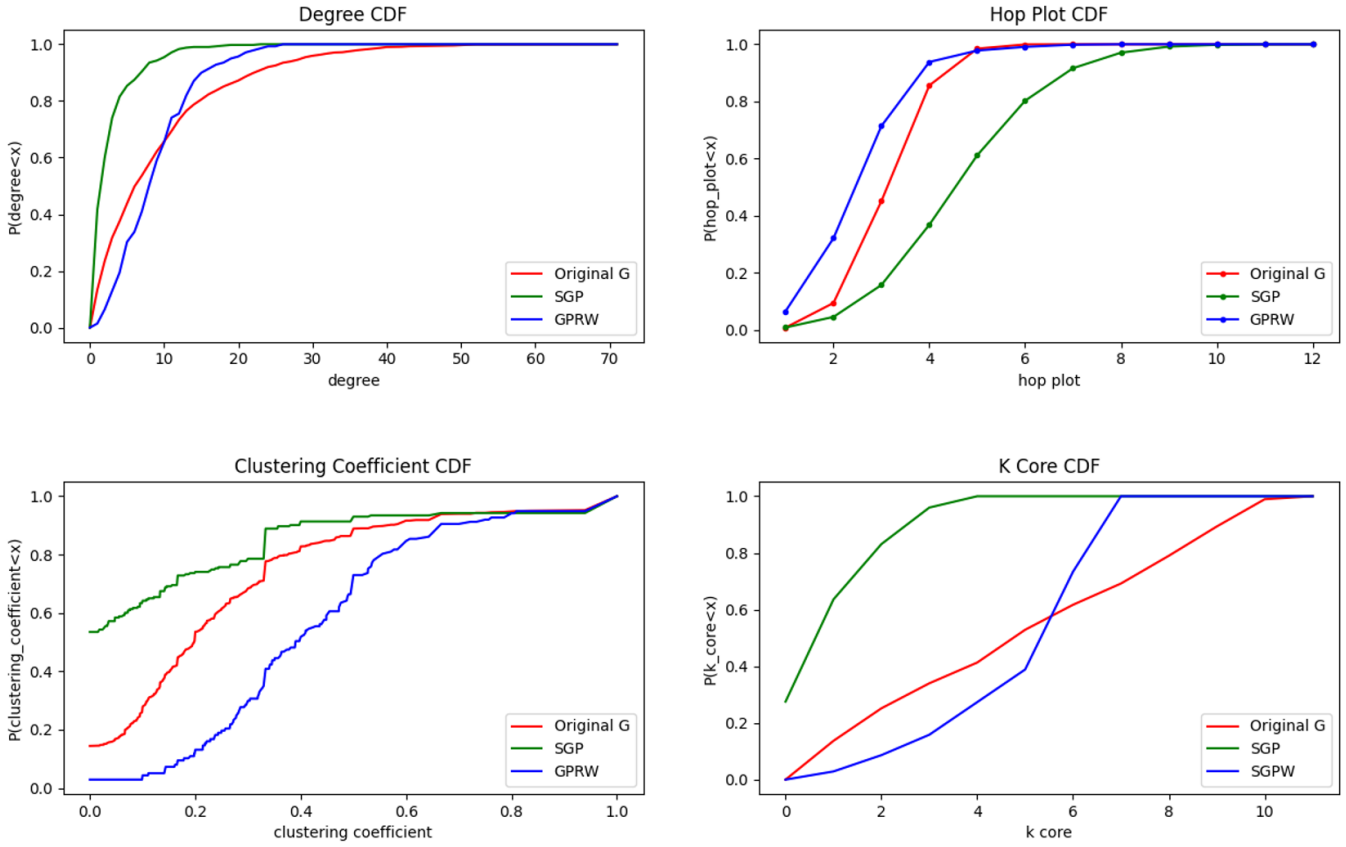

图 5: “email-univ”网络采样图与原始图的 degree、hop plot、clustering coefficient、k core 分布

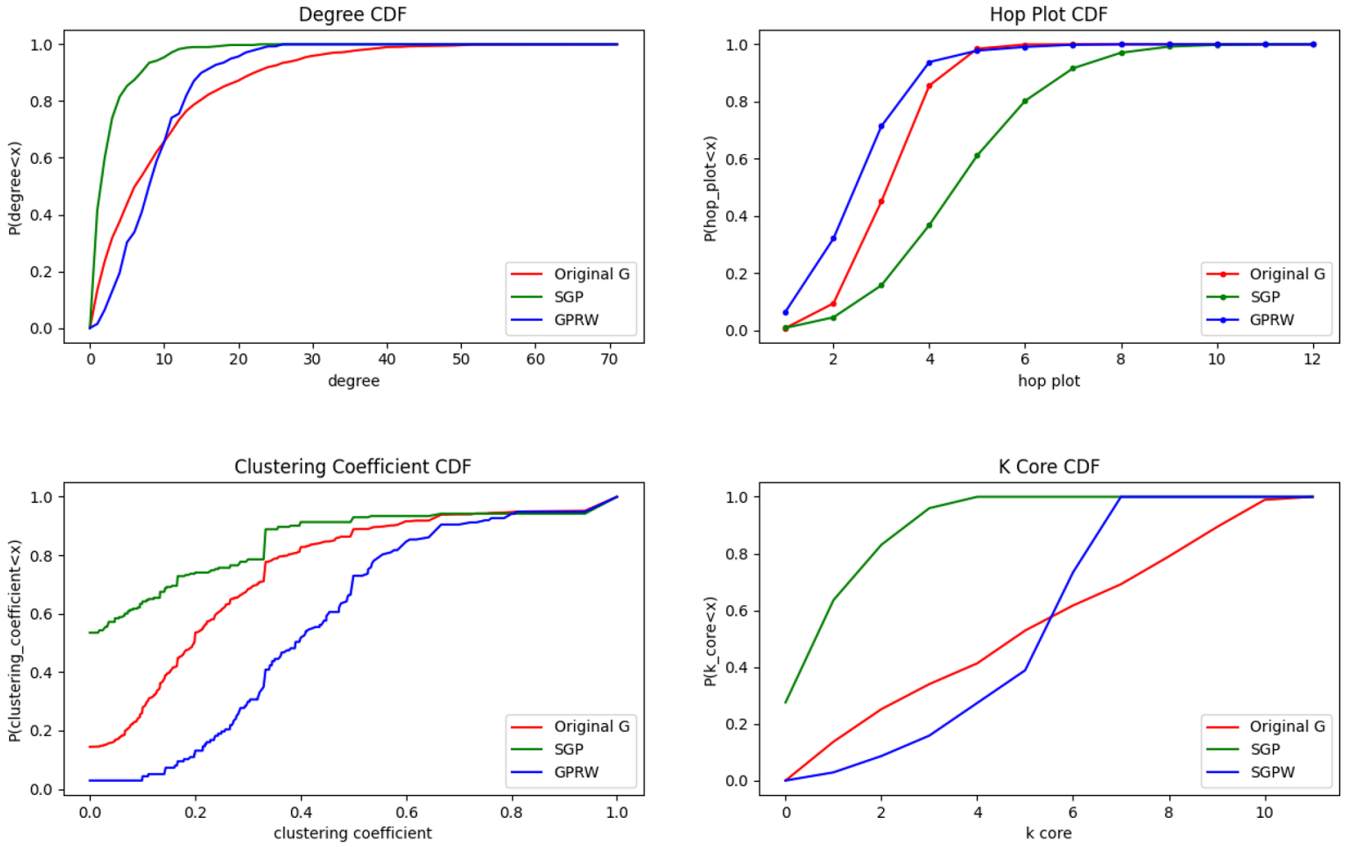

图 6: “lastfm-asia”网络采样图与原始图的 degree、hop plot、clustering coefficient、k core 分布

表 1: 抽样图与总体图的 KS 偏差 (P=0.2)

P=0.2	Dataset	ego-Facebook	email-univ	lastfm-asia	fb-pages-politician	fb-pages-company
SGP	degree	0.434554	0.396778	0.293296	0.432388	0.288813
	hop plot	0.294789	0.478100	0.570756	0.533208	0.397790
	cc	0.155262	0.380867	0.369621	0.332748	0.320204
	k-core	0.545030	0.597262	0.466487	0.588989	0.412683
GPRW	degree	0.113459	0.194446	0.305774	0.283534	0.058842
	hop plot	0.244538	0.188048	0.421711	0.355310	0.163180
	cc	0.153081	0.408645	0.476618	0.318805	0.406396
	k-core	0.173528	0.310618	0.347254	0.334933	0.088671

表 2: 抽样图与总体图的 SD 偏差 (P=0.2)

P=0.2	Dataset	ego-Facebook	email-univ	lastfm-asia	fb-pages-politician	fb-pages-company
SGP	degree	1.923542	1.605362	0.868362	1.549653	0.599759
	hop plot	0.504418	0.835312	1.193096	1.123788	0.647730
	cc	4.430056	2.771093	1.687156	2.560846	1.422390
	k-core	2.923095	3.217380	1.811019	2.368069	1.551825
GPRW	degree	0.705500	1.111029	0.600142	0.425736	0.129979
	hop plot	0.352709	0.202525	0.812018	0.615819	0.135688
	cc	4.524807	2.993184	2.719898	3.165356	1.865707
	k-core	1.708755	2.050722	0.662226	0.646880	0.188858

由上述实验结果可知，改进后的算法相较于原始的 SGP 算法，其度分布、hop plot 分布以及 k core 分布要更接近原始图的度分布、hop plot 分布以及 k core 分布，即改进后的算法能够更好的保留原始

图的度、hop plot 以及 k core 属性。因此，上述对 SGP 算法的改进是有效的。

5 实验结果分析

在本节中，我们将介绍几个真实社会网络的实验结果。它们分别是”ego-Facebook”、“email-univ ”、“lastfm-asia ”、“fb-pages-politician”、“fb-pages-company”，下表详细描述了五个数据集。

表 3: Datasets description

Datasets	Node	Edge	Diameter	Description
ego-Facebook	4039	88234	8	Facebook 社交圈 (匿名)
email-univ	1134	5452	8	Rovira i Virgili 大学 (Tarragona) 成员之间电子邮件交换网络的边缘列表。
lastfm-asia	7624	27806	15	亚洲 LastFM 用户的社交网络。
fb-pages-politician	5980	41706	14	有关 Facebook 页面的数据 (2017 年 11 月)。这些数据集代表了不同类别的蓝色验证 Facebook 页面网络。节点表示页面，边表示页面之间的相互喜欢。
fb-pages-company	14112	52126	15	有关 Facebook 页面的数据 (2017 年 11 月)。这些数据集代表了不同类别的蓝色验证 Facebook 页面网络。节点表示页面，边表示页面之间的相互喜欢。

首先，我们使用 Hansen-Hurwitz 估计量估计社交网络”ego-Facebook” 的平均度，社交网络”ego-Facebook” 中所有节点的平均度真值为 43.691，表 4展示了 Hansen-Hurwitz 估计量在社交网络”ego-Facebook” 上的的估计精度，表中的每个条目都是通过对社交网络”ego-Facebook” 运行 100 次 RW 抽样算法的平均估计值获得的。从估计结果可以看出，Hansen-Hurwitz 估计量能够比较准确地估计原始图的属性。

表 4: ego-Facebook 网络平均度估计结果

估计值	相对误差	估计值	相对误差	估计值	相对误差
43.682	0.019%	44.606	2.094%	45.543	4.240%
48.734	11.542 %	45.492	4.123%	44.311	1.419%
44.296	1.386%	44.347	1.502%	43.751	0.137%
43.865	0.399%	42.119	3.597%	42.028	3.805%
41.793	4.342%	44.482	1.811%	44.275	1.337%
45.549	4.254%	43.460	0.526%	42.614	2.463%
42.404	2.943%	43.436	0.582%		

接下来，我们评估我们的算法。图 7展示了使用 6 种抽样方法抽样示例网络“email-univ”得出的采样图以及原始图在 4 个图属性上的分布; 图 8展示了使用 6 种抽样方法抽样示例网络“lastfm-asia”得出的采样图以及原始图在 4 个图属性上的分布，其中抽样百分比为 20%。图 9给出了 6 个抽样方法所得出的采样图与原始图“email-univ”的平均 KS 距离；图 10给出了 6 个抽样方法所得出的采样图与原始图“lastfm-asia”的平均 KS 距离。此外，在表 5中给出了本文中所提到的抽样方法在五个数据集中采样得到的采样图与原始图详细的 KS 距离，表中的每个条目都是通过对每个数据集运行 20 次的 K-S 测试的平均值获得的。

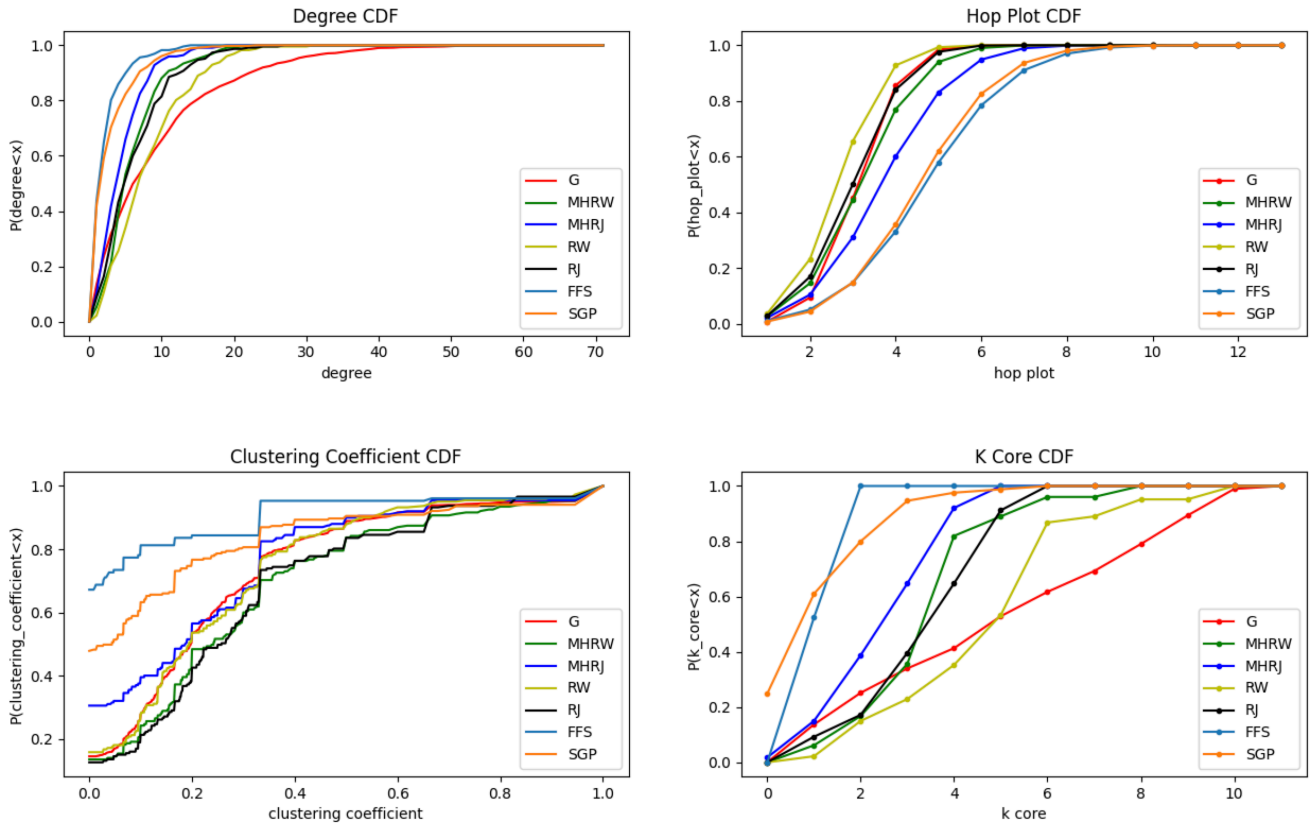

图 7: “email-univ”网络采样图与原始图的 degree、hop plot、clustering coefficient、k core 分布

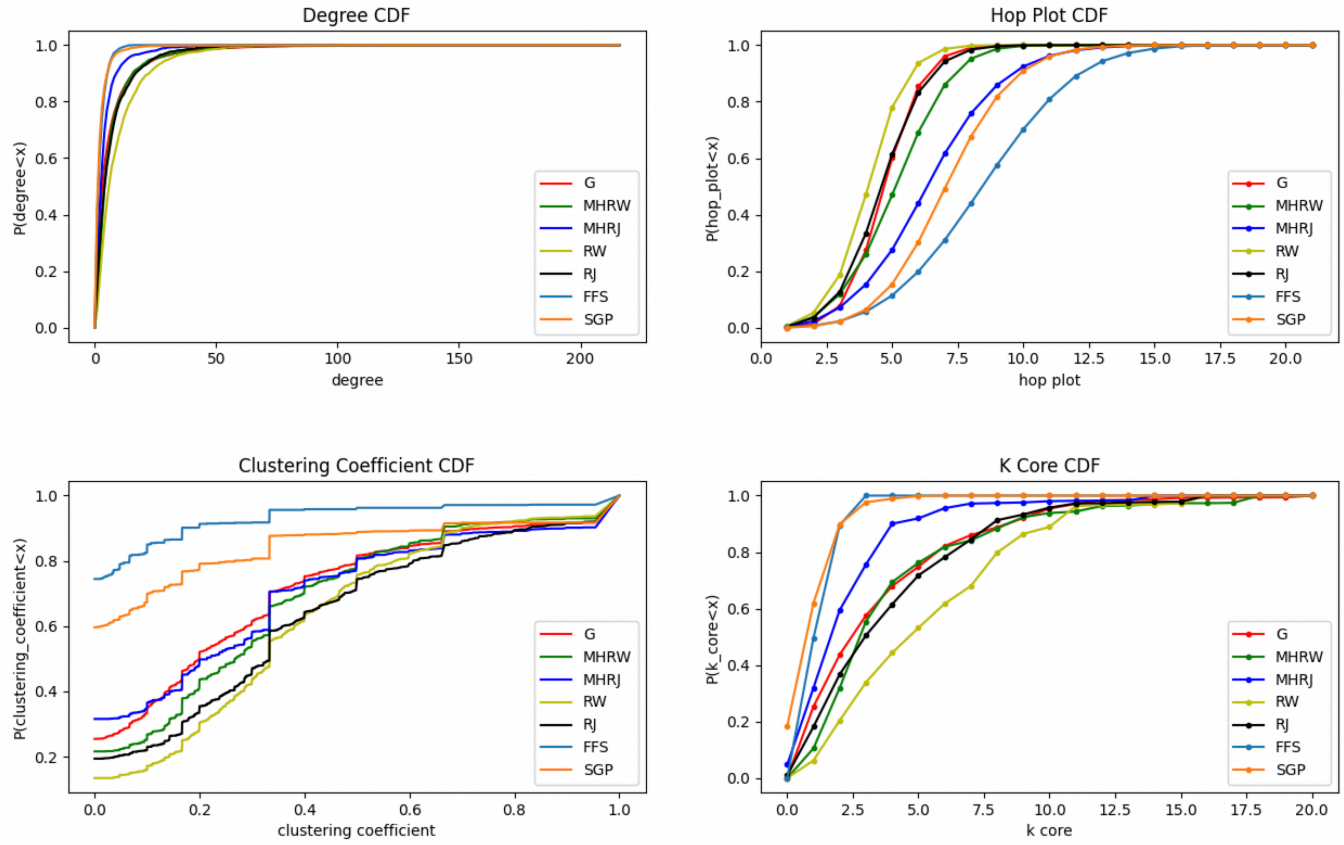

图 8: “lastfm-asia”网络采样图与原始图的 degree、hop plot、clustering coefficient、k core 分布

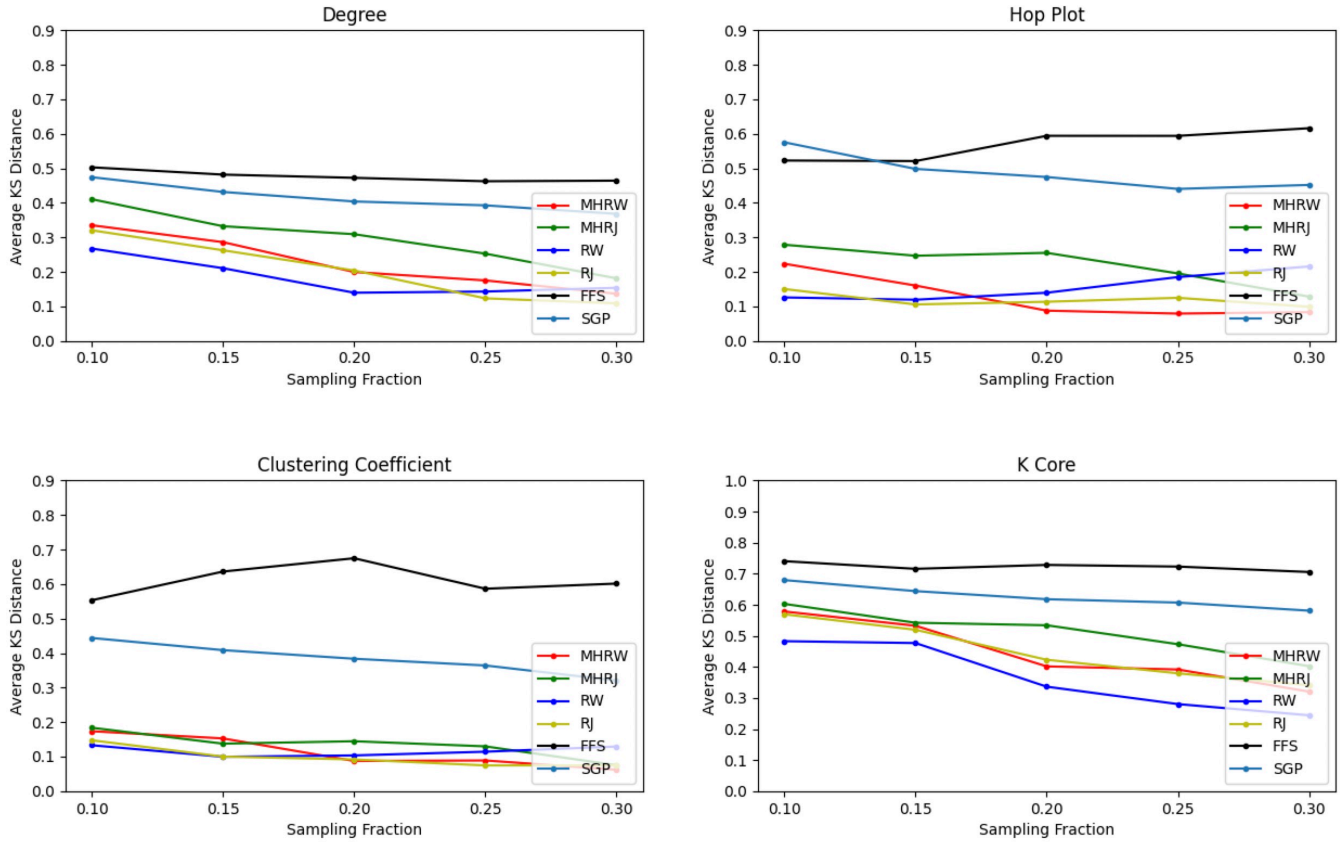

图 9: “email-univ”网络采样图与原始图的平均 KS 偏差

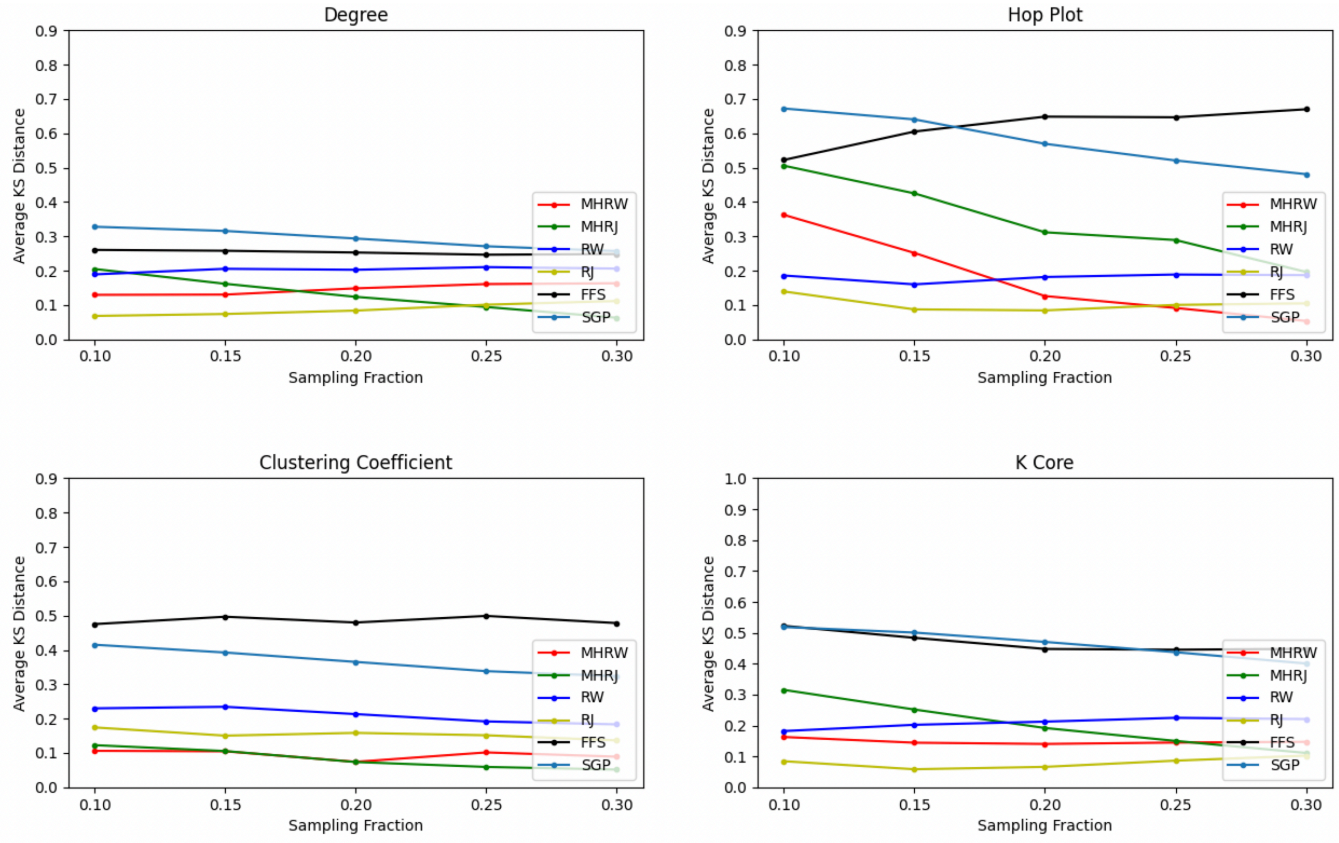

图 10: “lastfm-asia”网络采样图与原始图的平均 KS 偏差

表 5: 抽样图与总体图的 KS 偏差 (P=0.2)

P=0.2	Dataset	ego-Facebook	email-univ	lastfm-asia	fb-pages-politician	fb-pages-company
MHRW	degree	0.148468	0.214114	0.151788	0.131566	0.096371
	hop plot	0.219644	0.094430	0.170361	0.107242	0.123508
	cc	0.106837	0.078514	0.075809	0.074847	0.050826
	k-core	0.213048	0.445244	0.144033	0.148562	0.110335
MHRJ	degree	0.324144	0.273906	0.127604	0.188821	0.176141
	hop plot	0.363222	0.248594	0.288799	0.326535	0.256274
	cc	0.081047	0.129366	0.084440	0.070072	0.086020
	k-core	0.404948	0.511685	0.188021	0.242671	0.248692
RW	degree	0.131013	0.144726	0.210205	0.193686	0.151028
	hop plot	0.400165	0.132952	0.156668	0.239303	0.242597
	cc	0.203807	0.075890	0.204432	0.185394	0.064206
	k-core	0.186394	0.350748	0.215326	0.215915	0.163778
RJ	degree	0.261490	0.182609	0.074300	0.106336	0.043126
	hop plot	0.136851	0.097748	0.081774	0.061567	0.091444
	cc	0.195189	0.061971	0.144359	0.122343	0.024934
	k-core	0.358229	0.441894	0.057883	0.162857	0.081726
FFS	degree	0.736504	0.493775	0.251891	0.387471	0.320413
	hop plot	0.637284	0.606567	0.644228	0.599175	0.625904
	cc	0.862562	0.618551	0.470147	0.647158	0.505470
	k-core	0.903862	0.738176	0.440513	0.596707	0.572801
SGP	degree	0.430806	0.410745	0.293316	0.440582	0.282465
	hop plot	0.489817	0.499531	0.559119	0.663165	0.339861
	cc	0.128938	0.342999	0.363797	0.344285	0.318530
	k-core	0.545392	0.608822	0.466112	0.615806	0.403589

从图 9和图 10可以看出，随着采样百分比的增加，我们的方法在所有五个数据集中的测试值都有下降的趋势，因为原始图的样本越多越能更好地代表原始结构。同样，分布性质的第二个评价是 SD 散度，其评价结果如表 6所示，表中的每个条目都是通过对每个数据集运行 20 次的 SD 测试的平均值获得的。

表 6: 抽样图与总体图的 SD 偏差 (P=0.2)

P=0.2	Dataset	ego-Facebook	email-univ	lastfm-asia	fb-pages-politician	fb-pages-company
MHRW	degree	0.785126	0.966645	0.227933	0.354989	0.173849
	hop plot	0.311480	0.084241	0.167289	0.087732	0.083539
	cc	4.413959	2.233394	1.155719	2.147499	0.900125
	k-core	1.833156	2.329352	0.380515	0.678271	0.291061
MHRJ	degree	1.475913	1.110752	0.334109	0.574327	0.364061
	hop plot	0.672930	0.302205	0.369365	0.436675	0.290331
	cc	4.203741	2.394899	1.266521	2.180984	1.041083
	k-core	2.550173	2.653253	0.694425	1.229575	0.607791
RW	degree	0.600630	0.840148	0.301868	0.366635	0.200411
	hop plot	1.285643	0.084298	0.106078	0.276099	0.230055
	cc	4.652904	2.212016	1.235969	2.420260	0.865330
	k-core	1.806100	2.135521	0.374563	0.497115	0.361335
RJ	degree	1.250489	0.788866	0.131977	0.345088	0.105795
	hop plot	0.259606	0.065774	0.045386	0.047500	0.051835
	cc	4.190663	2.232257	1.225733	2.127499	0.905159
	k-core	2.315818	2.581791	0.203576	0.804555	0.301903
FFS	degree	4.014169	2.056808	0.727370	1.594898	0.886379
	hop plot	1.587436	1.394570	1.594332	1.360661	1.572710
	cc	5.738541	3.762471	2.146104	3.640641	2.131563
	k-core	5.598361	4.271301	2.260855	3.318076	2.701209
SGP	degree	1.861049	1.660388	0.867258	1.603093	0.587870
	hop plot	1.032378	0.900644	1.166940	1.732808	0.483671
	cc	4.480246	2.791167	1.675863	2.582243	1.419385
	k-core	2.914414	3.222441	1.816720	2.369019	1.576935

表 7显示了网络”lastfm-asia” 采样后直径的变化情况，由此可以看出，随机游走抽样算法能够较好的维护图的结构。

表 7: ”lastfm-asia” 网络采样图直径的变化

抽样方法	原始图直径	抽样图直径
MHRW	15	14
MHRJ	15	17
RW	15	11
RJ	15	16
FFS	15	18
SGP	15	21

由上述结果可知，抽样算法对不同类型的网络抽样有不同的抽样效果。随机游走算法在群落属性不明显的网络中抽样效果最好，在群落属性明显的网络中容易陷入局部区域并且偏向于收集高度节点。SGP 算法只适用于群落属性明显的网络，能较好地保持采样网络与原始网络的拓扑相似性和群落结构相似性。但是 SGP 算法的图划分算法计算每条边的权值需要整个网络的信息，局限性较大，并且对群落属性不明显的网络抽样，会出现很多孤立节点。

6 总结与展望

对于许多依赖于准确分析社会数据的互联网服务来说，推导一个具有代表性的抽样网络是非常重要的。本文的主要贡献是提出了一个划分网络抽样算法 SGP，并介绍了其他几种经典的图抽样算法。

我们进行了深入的分析和比较。在 5 个数据集上，使用 KS 检验法，SD 检验法，对 6 种图抽样算法的抽样效果从 4 个方面进行了评价与比较，并得出了每种抽样算法的优点与缺点以及适用的场景，并对 SGP 算法进行了一定的改进。

参考文献

[1] AHMED N K, BERCHMANS F, NEVILLE J, et al. Time-based sampling of social network activity graphs[C]//Proceedings of the eighth workshop on mining and learning with graphs. 2010: 1-9.

[2] AHMED N K, NEVILLE J, KOMPELLA R. Space-efficient sampling from social activity streams[C]//Proceedings of the 1st international workshop on big data, streams and heterogeneous source mining: algorithms, systems, programming models and applications. 2012: 53-60.

[3] NEWMAN M E. The structure of scientific collaboration networks[J]. Proceedings of the national academy of sciences, 2001, 98(2): 404-409.

[4] GJOKA M, KURANT M, BUTTS C T, et al. Walking in facebook: A case study of unbiased sampling of osns[C]//2010 Proceedings IEEE Infocom. 2010: 1-9.

[5] WATTS D J, STROGATZ S H. Collective dynamics of ‘small-world’ networks[J]. nature, 1998, 393(6684): 440-442.

[6] WANG S, HUANG L, HSU C H, et al. Collaboration reputation for trustworthy Web service selection in social networks[J]. Journal of Computer and System Sciences, 2016, 82(1): 130-143.

[7] LI R H, YU J X, QIN L, et al. On random walk based graph sampling[C]//2015 IEEE 31st international conference on data engineering. 2015: 927-938.

[8] QI X. Sampling Online Social Networks: Metropolis Hastings Random Walk and Random Walk[J]. arXiv preprint arXiv:2205.05885, 2022.

[9] HU P, LAU W C. A survey and taxonomy of graph sampling[J]. arXiv preprint arXiv:1308.5865, 2013.

[10] AHMED N K, NEVILLE J, KOMPELLA R. Network sampling: From static to streaming graphs[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2013, 8(2): 1-56.