

分类号: _____

U D C: _____

密级: _____

编号: _____

学 位 论 文

复杂网络中的随机游走研究

谢昀雅

指导教师姓名: 张旭 教授 河北工业大学

申请学位级别: 硕 士

学科、专业名称: 理论物理

论文提交日期: 2011 年 11 月

论文答辩日期: 2011 年 12 月

学位授予单位: 河北工业大学

答辩委员会主席: _____

评 阅 人: _____

2011 年 12 月

Dissertation Submitted to
Hebei University of Technology
for
The Master Degree of
Theoretical Physics

RANDOM WALKS ON COMPLEX NETWORKS

by
Xie Yunya

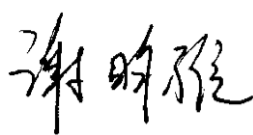
Supervisor: Prof. Zhang Xu

December 2011

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，进行研究工作所取得的成果。除文中已经注明引用的内容外，本学位论文不包含任何他人或集体已经发表的作品内容，也不包含本人为获得其他学位而使用过的材料。对本论文所涉及的研究工作做出贡献的其他个人或集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：



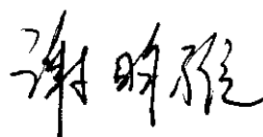
日期：2011.11.6

关于学位论文版权使用授权的说明

本人完全了解河北工业大学关于收集、保存、使用学位论文的以下规定：学校有权采用影印、缩印、扫描、数字化或其它手段保存论文；学校有权提供本学位论文全文或者部分内容的阅览服务；学校有权将学位论文的全部或部分内容编入有关数据库进行检索、交流；学校有权向国家有关部门或者机构送交论文的复印件和电子版。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：



日期：2011.11.6

导师签名：



日期：2011.11.6

复杂网络中的随机游走研究

摘 要

如今，我们对现实网络的研究日益广泛，如蛋白质相互作用网，社会联系网，英特网等。很多情况下我们只对这些网络中的部分内容有所了解和利用。即使我们可以得到完整的网络信息，只对整体网络中的一部分进行研究也更为“经济”。为了得到完整网络的子网，我们以随机节点抽样和随机连边抽样等方法对整体网络进行抽样。当随机节点和随机连边抽样存在一定的偏差时，寻找一种更为准确的抽样策略就成为了重要的研究课题。许多情况下，探索网络的过程类似于随机游走的过程，因此，研究、刻画不同的随机游走抽样策略对网络抽样效果的影响就十分重要。

本文中，我们在已有的无限制的随机游走（URW）的基础上发展出选择种子节点的随机游走（CSNRW），及不返回上一步的随机游走（NRRW），并探讨了这三种类型的随机游走在加权的美国航空网，ER，BA 和 WS 网络中的抽样情况。通过模拟研究了它们的抽样效率、度分布、抽样子网的平均度 $\langle k \rangle$ 和平均聚类系数 $\langle c \rangle$ 随抽样步数的改变，得到了一系列的相关结果：三类抽样策略在具有相同尺度和平均节点度的网络上具有不同的抽样效率和阈值；这些网络的子网度分布都只是略微的偏向更大的节点度；通过对 $\langle k \rangle$ 和 $\langle c \rangle$ 变化曲线的观察，发现子网的 $\langle k \rangle$ 和 $\langle c \rangle$ 可以在有限的抽样步数内回到原值，并且容易在加权网中被高估。我们的工作也揭示了，若一个节点具有大的节点度，则与它相邻的节点间更易具有较少的连接。

关键词：复杂网络，抽样，随机游走，抽样效率，度分布，平均度，平均聚类系数

RANDOM WALKS ON COMPLEX NETWORKS

ABSTRACT

Today we have done more and more works on real-world networks, such as protein interactions, social contacts, and the Internet. Most real world networks are only partially known and available to us. In another words, it is more “economical” for us to investigate a part of the full networks. In order to get the subnet from the original full network, we use the strategies of random node and random link samples. When random node and random link samples are biased, it becomes the key issue to find an accurate sampling strategy. In many cases, the process of exploring such networks resembles a random walk, so it is also interesting to investigate and characterize the effect of such networks which be covered by different strategies.

We discuss three types of random walks: unrestricted random walk (URW), choose seed node random walk (CSNRW), and no-retracing random walk (NRRW), sampled on weighted USAir, Erdős-Rényi (ER), Barabási-Albert (BA), and Watts-Strogatz (WS) networks respectively. The sample efficiency, degree distributions of subnets and the changes of average degree $\langle k \rangle$ and average clustering $\langle c \rangle$ have been investigated through simulation. Series of relevant results are obtained including the fact that three types of complex networks with the same size and average node degree have different sample efficiency and thresholds. And the degree distributions for four networks are similar to the real distributions, with a slight bias toward higher degrees. We furthermore show that $\langle k \rangle$ and $\langle c \rangle$ of subnet can come back to the original value within the limited steps, and more easily overestimated in the weighted network.

Our work display that when a node has large k , the other nodes linking with it have few links among themselves.

KEY WORDS: complex networks, sample, rand walks, sample efficiency, degree distribution, average degree, average clustering coefficient

目录

第一章 绪 论	1
§ 1-1 网络模型概述	2
1-1-1 ER 随机网络	2
1-1-2 BA 无标度网络	3
1-1-3 WS 小世界网络	4
§ 1-2 网络结构参数	4
§ 1-3 本文的研究内容	5
第二章 随机游走及网络抽样	6
§ 2-1 网络上的随机游走分析	6
2-1-1 模型和假设	6
2-1-2 混合率和传导率	7
§ 2-2 多路线随机游走	11
§ 2-3 网络抽样	12
§ 2-4 小结	14
第三章 随机游走抽样	15
§ 3-1 引言	15
§ 3-2 随机游走抽样的基本概念	16
§ 3-3 三类随机游走抽样方法	17
§ 3-4 抽样网络	17
§ 3-5 抽样效率	18
§ 3-6 抽样网络与原网络的度分布对比	19
§ 3-7 抽样子网平均度和平均聚类系数随步数的变化	20

§ 3-8 小结	23
第四章 总结与展望	24
参考文献	25
致 谢	29
攻读学位期间所取得的相关科研成果	30

第一章 绪 论

在过去的几十年中,一些关于现代社会中,人、物间关系的复杂联通性问题越来越吸引了人们的注意。而它的主要吸引点在于网络概念的描述,网络指一系列物体的相互作用模式,我们发现网络的概念在对众多研究话题的探讨评论中出现。事实上,在不同的语境中网络一词的使用往往具有多样性。

社会网络作为人类生存发展的外部环境,集中体现了个人与他人间的社会联系。在过去的半个世纪中,社会网络间的连接关系根本上由它的地理位置决定,由于科学技术的进步使得远距离旅行,全球通讯,数字交往变得更为容易,减弱了这些结构的传统局域性质,使得社会网络扩展到了更大的尺度规模上,从而社会网络的复杂性随人类历史的进步稳定增加。

我们产生的信息也具有相似网络结构,当高质量信息的供应者(出版商,新闻集团,学术机构)越来越少,有着不同可靠性、目的、意图的信息日益增加,信息网络也变得日趋复杂。在这种环境下,我们获取的每一条信息都依赖于此信息参考其它信息的路径及其它信息对此信息的授让权限,而这些其它信息往往会牵扯出一个大规模的网络连接。

我们的科学、经济系统也依赖于具有巨大复杂性的网络。现代的科学、经济系统由一些潜在的结构联系在一起,某一时刻局部的崩溃,就会导致连续故障和财政危机。这就使得对网络中单个节点行为的观测变得更为困难,修补这类网络也更具风险。网络的概念可以迅速的扩展到其它领域的讨论中:制造业现在具有它的全球供货网,销售公司有销售网,传媒公司有广告客户网。复杂网络的概念已经愈发的进入人们的视野,并吸引了众多科学工作者的关注。然而,在这些表述中,我们强调的并不是网络结构本身,而是它作为一个巨大传播个体在面对一些预料之外的事件时反应的复杂性,即网络上的动力学问题。如复杂网络中的控制、同步,传播机理与动力学分析,搜索策略及网络上的随机游走等。

当我们面对成千上百万的节点时,无论从人力还是物力的角度,对整个网络进行分析研究都显得很划算,如果我们能得到一个足够小又能基本代表原网络性质的子网络,则将大大节约我们的研究成本。例如,计算机通讯研究者在进行英特网路由协议的研究时,总是喜欢对边界网关协议(Border Gateway Protocol)进行逐条模拟,但在互联网上对数千节点进行模拟过为昂贵而不可行。针对这些问题,前人已经提出了一些复杂网络的抽样方法,如节点抽样,连边抽样,滚雪球抽样等,但都存在一些偏差。本文模拟了随机游走的抽样方法,并提出了两类改进的抽样方法,比较了抽样效率,及子网的诸多性质,发现随机游走的相关抽样方法能够很好的进行网络抽样,是一种便捷有效的抽样方法。

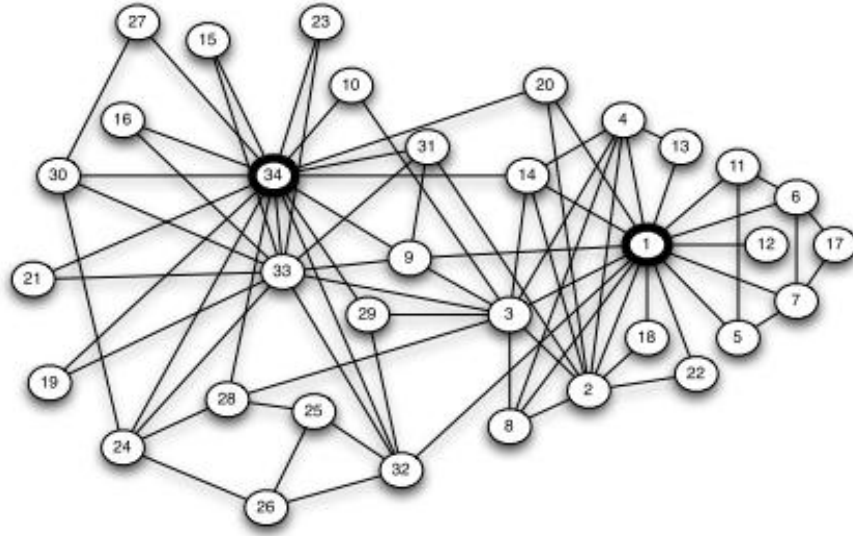


图 1.1 具有 34 个成员的空手道俱乐部的社会关系网络

Fig. 1.1: The social network of friendships within a 34-person karate club

§ 1-1 网络模型概述

前人在对大量现实网络进行了深入研究的基础上，构造了众多网络模型，这些模型不仅很好的刻画了现实网络的特征，也使得对现实网络的研究更为简便有效。以下，我们将对最有代表性的三类网络模型（ER 随机网络、BA 无标度网络、WS 小世界网络）加以介绍。

1-1-1 ER 随机网络

静态的随机图模型和拓扑生成元，例如 ER 模型^[1-3]和 Molloy、Reed 网络生成算法是以随机性为基本元素的最基本的网络模型。它们描述了我们知之甚少的网络性质，即如何从节点间建立连接。在缺少信息的情况下，最简单的假设即根据给定连接概率 p 任意的连接两个节点。在这种最初的假设中，ER 图 G 由 N 个伴有 E 条连边的不同节点构成，这些连边的终点由除始点外的 $N-1$ 个节点任意选择。为了使其更接近真实网络，我们可以多角度的对 ER 随机图进行扩展，简单的，我们可以将其扩展成为具有任意给定度分布的广义随机图。

1-1-2 BA 无标度网络

BA 模型可以作为网络动力学过程的一个示例^[4]，它体现出了当新的节点逐渐加入到网络的过程中，网络便呈现出了重尾度分布的特点。更深入的观察这类网络，我们发现新的连边并非随机的与原网络的节点相连，而是更倾向于连接那些度大的节点。例如，新加入的网页更喜欢连接那些非常流行的网站。相似的，一个具有庞大社会关系网的人相较于那些朋友圈较小或只有一些密友的人，更乐于结实新的朋友。在英特网中，新的服务供应者将与已经具有较好连接的供应者成为伙伴以实现集中性的优化作为目标，这类系统因此被基于有线连接机制的网络模型所描述，众所周知的如富者更富现象，Matthew 效应^[5]和 Gibrat 性质^[6]等。

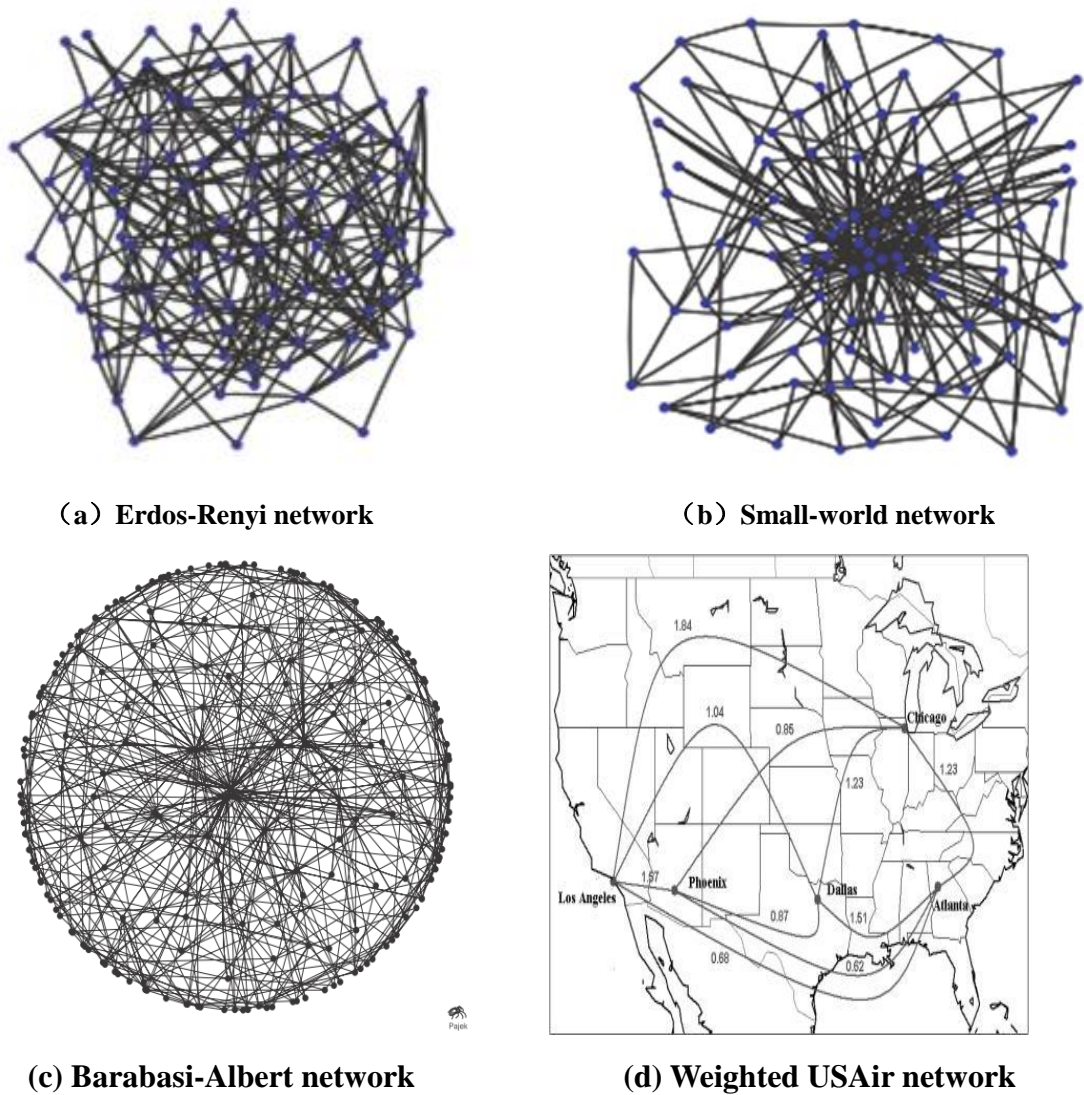


图 1.2 ER、WS、BA 网络模型及加权的美国航空网

Fig1.2 ER, WS, BA network models and weighted USAir network

1-1-3 WS 小世界网络

在随机图中，聚类系数由加入节点的度分布决定并在网络趋于极大值极限时消失。因此凭经验观察具有非常大聚类系数的真实网络，即便在概念上也是一个极大的挑战，这就激励人们调和平均聚类系数 $\langle C \rangle$ 至某特定值以定义模型。Watts 和 Strogatz 观察到社会网络具有当节点间存在高的聚集性的同时具有小的平均路径的事实后，便建立了一个新的模型，即在规则格子（具有大的聚类系数）抽取连边，进行随机化重连，使其更加趋近于纯随机网（具有小的平均路径长度）。便得到了具有大的聚类系数和小的平均路径长度的小世界网络。

§ 1-2 网络结构参数

节点 i 的度 (Degree) k 描述了 i 点与其它节点的连接关系，与 i 点连接的边的个数即此节点的度， k_i 。其平均值记为 $\langle k \rangle$ 。节点的度分布 $P(k)$ 表示随机选定节点的度恰好为 k 的概率。

平均聚类系数 (Clustering coefficient) C 刻画了，节点 i 的邻居间的连接情况。节点 i 的聚类系数可表示为：

$$C_i = 2E_i / (k_i(k_i - 1)) \quad (1.1)$$

其中 E_i 为 k_i 个节点间实际存在的边数。

最短路径 (Shortest path) 即两个顶点 i 、 j 之间边数最少的路径。 i 、 j 间的距离 d_{ij} 为经过 i 、 j 间最短路径的个数。网络的平均路径长度 L 为 d_{ij} 的平均值，记为： $L = \frac{1}{\frac{1}{2}N(N+1)} \sum_{i \geq j} d_{ij}$ ， N 为网络的节点个数。

介数 (Betweenness) 描述了经过节点 i 的最短路径的条数。节点的度数与介数间通常存在幂率关系。

§1-3 本文的研究内容

本文主要利用随机游走的相关方法进行复杂网络的抽样，并模拟研究了抽样子网的性质。全文的主要内容包括以下几个方面：

- 一. 由原始的限制随机游走抽样方法发展出根据节点度选择种子节点和不返回上一步的随机游走抽样方法。
- 二. 将三类抽样方法分别应用于能代表大多数网络特点的随机、无标度、小世界及加权的美国航空网络中，并对比了三类方法的抽样效率。
- 三. 通过模拟将三种方法在四类网络中的抽样子网的度分布与原始网络的度分布加以比较，以确定不同抽样方法对具有不同特点网络的度分布的影响。
- 四. 模拟抽样子网平均度及平均聚类系数随抽样步数的变化情况，动态分析抽样子网平均度及平均聚类系数的变化规律。

第二章 随机游走及网络抽样

§ 2-1 网络上的随机游走分析

在这一章中我们将对任意网络的随机游走行为加以分析。并确定随机游走过程中，一些重要参量的具体表达式。

2-1-1 模型和假设

以 $G = (V, E)$ 来代表网络，顶点 V 代表网络节点，边 $E \subseteq V \times V$ 代表节点间的连接。规定网络中没有节点与自身相连，也不存在两点间存在多条连边的情况。这并非只是简化了我们的模型，而且使得我们的网络模型更接近真实情况，如典型的 P2P 网络。我们定义网络的节点个数 $|V| = n$ ，节点度为 k 的节点个数为 n_k （若节点有 k 个邻居则 $\sum_k kn_k = 2|E|$ ）。对于所有节点它的节点度 k 一定小于网络大小 n ，在典型的实际网络中（如社会网和 P2P 网络）每一节点都仅与系统中其他顶点的子网相连。我们还以 p_k 代表在网络中随机均一选取度为 k 的节点的概率（ $p_k = n_k / n$ ）。网络中节点的平均度 $\bar{k} = \sum_k kp_k$ 。对于一个给定的网络，由概率 p_k 形成的分布被称为此网络的度分布。

原网络 G 上的随机游走过程实际为一类马尔科夫过程^[7] M_τ ，转移矩阵 $P = [P_{ij}]$ 被定义为：

$$p_{ij} = \begin{cases} \frac{1}{d(i)} & \text{if } (i, j) \in E \\ 0 & \text{otherwise} \end{cases} \quad (2.1)$$

其中， P_{ij} 是由 i 点移动到 j 点的概率， $d(i)$ 为 i 点的度。 P 研究每一步游走 l 至相应节点的概率。这个概率可以用状态概率矢量 $q^l = (q_1^l, q_2^l, \dots, q_n^l)$ 表示，其中 q_i^l 代表随机游走在第 l 次游走至 i 点的概率。进

而，这一概率可以写为 $q^l = q^{l-1} p$ 。

假设 G 有限且全连同，那么 M_G 就是不可约的：即每一节点都可由其他节点到达，两节点间的平均路径长度也有限。假设 G 也并非有向，则表明 M_G 为非周期的，所以我们可以应用马尔科夫链的基本理论^[8]。这一原理表明，在图 M_G 中存在一个各态历经的单一状态分布 π ， $\pi P = \pi$ ， $\pi = (\pi_1, \pi_2, \dots, \pi_n)$ ，其中 π_i 为：

$$\pi_i = \frac{d(i)}{2|E|} \quad (2.2)$$

直觉上， π 代表 M_G 中的稳态。上式中 π_i 代表当达到稳态分布时随机游走的某一步到达节点 i 的概率。这一概率正比于 i 点的度， $d(i)$ 。

基于 (2.2) 式，当 E 为常数， π_i 正比于 i 点的度 $d(i)$ 。所以我们更容易抽取出度大的节点，换句话说，子网的平均度一定大于等于最初的网络。这一点将在后文的数值模拟实验中得以验证。

2-1-2 混合率和传导率

我们感兴趣于随机游走以多快的速度趋近于 π ，我们将其大小称为混合率^[9]。为使方程(2.6)得以应用我们需要得到一个较快的收敛速度。

收敛速率与转移矩阵 P 的本征值有关。矢量 \vec{x} 是 P 的本征矢其本征值为 λ ，可表示为 $\vec{x}P = \lambda\vec{x}$ （因此，我们就可以假设 π 为 P 的本征矢，本征值为 1）。在文献[9]中我们知道 P 有 n 个真正的本征值 $\lambda_0 = 1 > \lambda_1 \geq \dots \geq \lambda_{n-1} \geq -1$ （事实上如果 G 非双向网络则 $\lambda_{n-1} > -1$ ）。通过文献[10]我们知道收敛于 π 的速率由 P 的第二大本征值的模控制。在很多现实网络中，我们可以方便的假设 $\lambda_1 > \lambda_{n-1}$ ^[28, 29, 35]。以下假设随机游走从 i 点开始^[9]：

$$|P_i^{(l)}(j) - \pi_j| \leq \sqrt{\frac{d(j)}{d(i)}} \lambda_1^l \quad (2.3)$$

上式中 $P_i^{(l)}$ 是当 i 为初态时随机游走在第 l 步的态分布。因此，我们可以期待在谱间隙 $1 - \lambda_1$ 处可以进行快速混合。

现在, λ_1 的值与网络 Φ_G 的传导率紧密相关。非正式的, 传导方法能多好的测量一个网络的连接情况, 由如下的定义可以看出。对于 $S \subseteq V$, 割集 S , $C(S)$ 是边的集合, 边的一边终点为 S , 另一终点为 \bar{S} , S 的体积, $vol(S)$, 被定义为 S 中节点度的集合, $vol(S) = \sum_{i \in S} d(i)$ 。进而, 我们可以对 G 的传导率进行如下计算:

$$\Phi_G = \min \frac{|C(S)|}{vol(S)} \quad (2.4)$$

传导和收敛的关系如下式 (Cheeger's 不等式) ^[9]:

$$\frac{\Phi_G^2}{2} \leq 1 - \lambda_1 \leq 2\Phi_G \quad (2.5)$$

所以我们可以由好的传导率获得高的混合速率, 即随机游走以很快的速率收敛于稳态分布 π 。由此我们可以推断, 当网络具有一个好的传导率时, 随机游走便可以更容易的游走到网络的其他区域, 快速的达到平衡态。我们知道许多真实网络 (尤其是通信网络) 和网络模型都具有高的连通性 ^[11-13]。

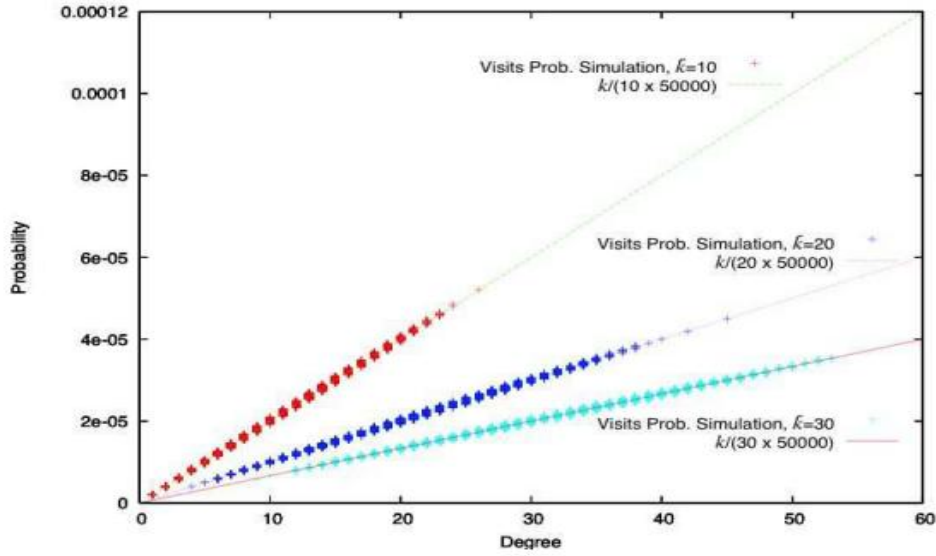
因此, 我们可以假设具有度 k 的节点在随机游走的某一步被访问到的概率, $P(k)$, 与 k 成比例, 并可按下式进行计算:

$$P(k) = \sum_{i \in V} \frac{d(i)}{2|E|} = n_k \frac{k}{\sum_j j n_j} = \frac{k p_k}{\bar{k}} \quad (2.6)$$

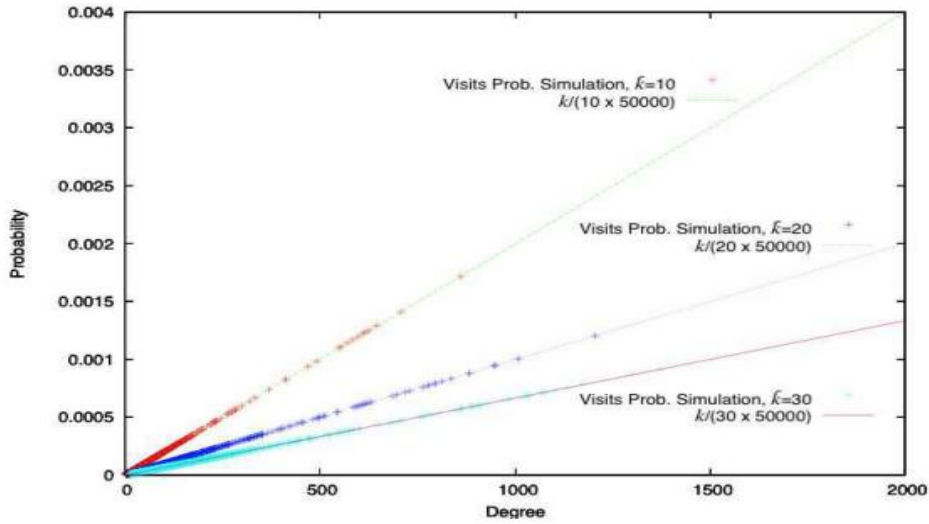
我们可以利用上式集中的对随机游走覆盖率进行计算。当然, 它的正确性依赖于从开始游走到稳态分布的距离, 或它可以多快的覆盖整个网络。另一个需要说明的是在随机游走在连续步中可能存在的连续性问题。我们估计了访问节点的平均个数和通过前步的估计值由确定步数的随机游走覆盖整个网络的情况。新的估计假设随机游走抽样与节点的随机抽样具有相似的统计性质, 其中, 尽管在连续的游走中存在依赖性, 但选择一个确定节点的概率正比于 k_i 。

在文献[14]中, 展示了独立抽样和随机游走抽样的相似性, 这也是我们分析方法的基础。如其作者所述, 在一个具有很好连通性和扩展性质 (这与 λ_1 紧密相关) 的网络中, 随机游走抽样和独立抽样具有相似的行为, 即选择某些节点的概率正比于节点度。

另外, 已经有一些实验用以证明这些假设的正确性。图 2.1 的结果证实上面的假设是正确的。另外, 我们注意到, 由方程(2.6)表述的性质在之前有关随机游走的工作[15]、[16]中已有假设, 更有[14]中的结论作为支持。



(a) Erdos-Renyi networks

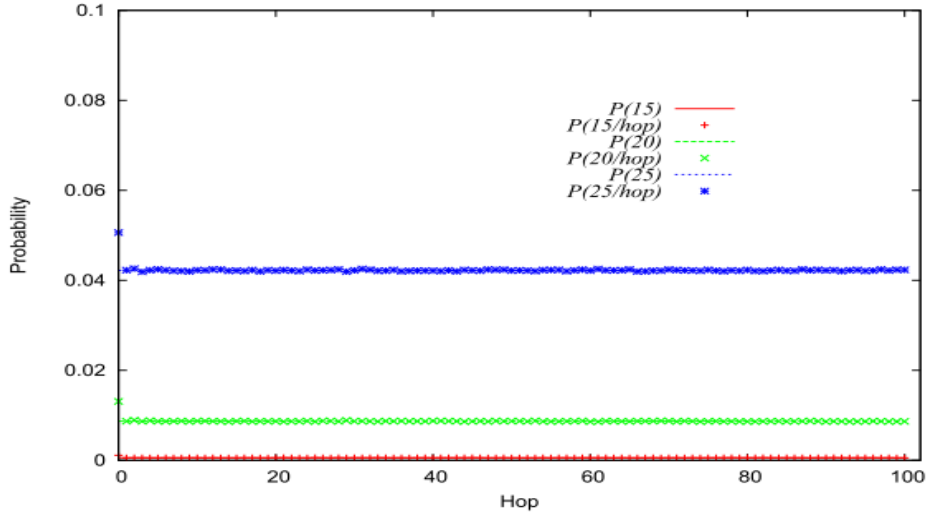


(b) Small-world networks

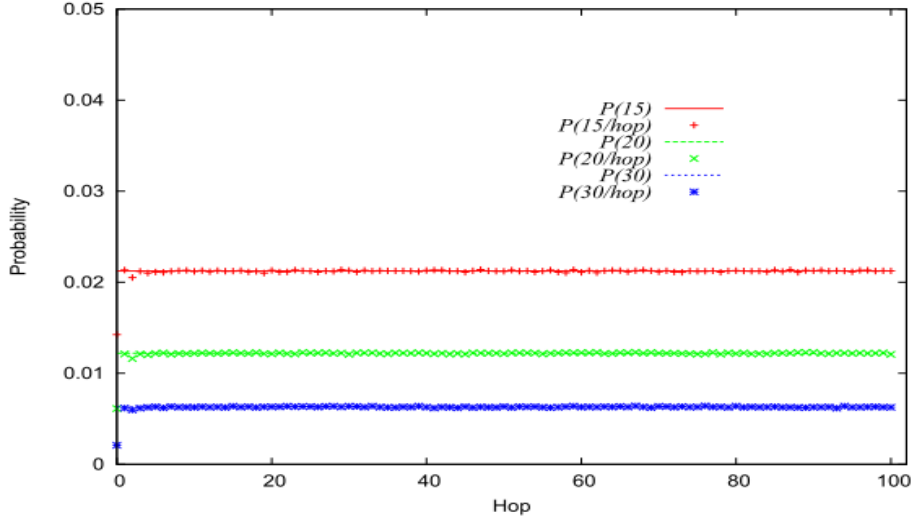
图 2.1 在此图中，我们展示了搜索信息达到特定节点的概率并以此作为此节点度的函数。图中针对含有 50000 节点的 Erdos-Renyi 和 small-world（幂率分布）网络，它们含有不同的平均节点度（10, 20 和 30）。相同的试验在含有 25000 和 100000 个节点的网络中完成，得到了相同的结果。从图中容易看出，搜索信息到达一特定节点的概率正比于节点度。

Fig.2.1 In these figures, we show the probability of a search message arriving at a particular node as a function of its degree. We have used both Erdos-Renyi and small-world (power-law) networks formed by 50, 000 nodes, with different average node degrees (10, 20 and 30). The same experiments have been performed with networks formed by 25, 000 and 100, 000 nodes, and we found similar results. As it can be readily seen, the probability of a search message arriving at a particular node is proportional to the degree of the node.

其它我们讨论的重要论点有，如方程(2.6)所述，随机游走抽样能以多快的速度达到稳定的状态分布。图 2.2 显示了游走的随机行为，从图中我们可以看到，在开始游走后，到达度为 k 节点的概率便几乎立即达到了 $P(k)$ 。



(a) Erdos-Renyi network, $\bar{k} = 30$.



(b) Small-world network, $\bar{k} = 10$.

图 2.2 上图对到达一个如模型所定义的度为 k 的节点的概率 $P(k)$ ，与随机游走在某一步到达具有节点度 k 的节点的精确概率进行了比较。试验针对 ER 和 small-world 网络中三类具有相同平均度和大小 ($n = 50 \cdot 10^4$) 的不同网络展开。

Fig.2.2. These figures compare the probability $P(k)$ of reaching a node of degree k as defined by the model,

with the measured probability of reaching a node of degree k at each hop of the random walk. Both for ER and small-world networks the experimental results are averaged over three different networks with the same average degree and size ($n = 50 \cdot 10^4$).

我们注意到，好的传导性质，在随机游走中表现为可以从网络上的一点在很少的步数内游走到网络中的任意其它节点，无论网络中是否存在如自环等的其他性质。

§ 2-2 多路线随机游走

我们可以把这个过程理解为 W 个游走者在具有 N 个节点的网络中随机扩散的行为。我们假设每一节点 i 根据其上游走者的个数具有占有数 W_i 。网络上游走者的总个数为 $W = \sum_i W_i$ ，每一个游走者根据转移比率沿连边进行扩散，转移比由节点度和其它因素共同决定。在简单的马尔科夫随机游走情况下，位于节点 i 上的游走者由 i 点沿连边 (i, j) 扩散的比率可表示为：

$$d_{ij} = \frac{r}{k_i} \quad (2.7)$$

其中 k_i 为节点 i 的度。这一关系简单的定义了从节点 i 出发沿任一条边扩散的统一比率，这对应于从任一点出发的总的比率 $\sum_{j \in v(i)} d_{ij}$ 等于 r 。

本文中，我们假设节点的性质只由节点的度表征，如果节点的度相等则其统计性质也就相等。这就允许我们将度的块变量定义为：

$$W_k = \frac{1}{N_k} \sum_{i|k_i=k} W_i \quad (2.8)$$

其中 N_k 为具有度 k 的节点的个数，变量 W_k 代表了位于度为 k 的节点上，游走者的平均个数。这也方便我们对网络中大范围的节点度进行考量。我们考虑在一个度分布为 $P(k)$ 的网络中，游走者一转移概率 r/k ，由度为 k 的节点转移至度为 k' 的节点。此游走过程可以由简单的平均场方程表示：

$$\partial_t W_k(t) = -r W_k(t) + k \sum_{k'} P(k'|k) \frac{r}{k'} W_{k'}(t) \quad (2.9)$$

上式中右边的第一项仅考虑了游走者以比率 r 游走出节点。第二项则考虑了从其他节点游走至节点 k 的情况。此项正比于节点连边 k 与其从每个邻居处过来的游走者的平均个数的乘积。

§ 2-3 网络抽样

目前,对于复杂网络的研究正进行的如火如荼,我们已经对诸多网络的各类性质进行了较全面的探讨。其中不乏包括英特网在内的技术型网络^[17],包括蛋白质相互作用网的生物网络^[18],和以科学合作网为例的社交网络^[19]。用以描述这些真实网络性质的各类模型也通过模拟和分析的方法得到了介绍、研究。近来,一些围绕关于在数据挖掘和认证现实网络实际意义时存在的可能错误和偏差的工作已经完成,这些工作大部分是针对社会网络和英特网展开的^[20-30]。

例如,当我们想通过参与者间的关系来构造一个社会网络时,由于调查只能针对所有人中的部分进行,所以我们所收集到的数据可能并不全面甚至是错误的^[22]!英特网的拓扑性质是通过聚集路径和跟踪路由推断的,这也仅仅能推断出了整体网络中的一小部分而已^[23-26]。在生物网络中,蛋白质相互作用网通过寻找在特定功能模型中的细胞功能来进行识别^[18]。通过实验来识别这类网络存在一些基本的自然极限。因此,我们所识别的所有这些网络都仅仅是完整结构中的抽样网络而已。另外,当对一些庞大的网络进行研究时,由于时间复杂性,对例如介数^[31,32]在内的一些性质的研究便不可避免的会用到抽样过程。

很多基于现实网络特征(如小世界效应^[33]和幂率度分布^[34,35])的网络模型已经被设计出来。若我们根据抽样网络所观察到的特点与现实网络原始结构的特点并不相同,情况又会如何?我们已经知道,在一些情况下,基于路由抽样方法所得到的抽样网络与原网络的拓扑性质有极大的不同^[23-26]。缺少数据在社会网络中所造成的影响也在文献[22]中得到讨论,其中显示出了一些问题,如在构思社会网时会导致数据的不完全性和对平均节点度、聚类系数和介数等网络参量的错误估计。基于此点,我们需要对这些网络参量进行更具一般意义的探讨。

从统计的意义上,当网络全体中的绝大部分被随机的抽出则我们就能通过抽样对网络进行公正合理的评估。然而,这一抽样法则不能被直接应用于进行网络抽样,因为,网络中包含节点和连边两类元素。例如,节点的度分布就是一个统计量,其中的节点度并不是一个节点的独立特征。节点通过网络的另一元素-连边逐一连接,我们通过连接节点的连边个数定义节点的度。相似的,网络的其他性质也紧密的

依赖于节点和连边间的相互关系。由于这两类元素的相互作用，就产生了几类不同的对网络进行抽样的方法，每种方法针对各类性质也给出了不同的特征。

在物理领域，我们已经针对作为抽样反过程的随机崩溃和有意连接做了大量的工作^[36-39]。因此，此类工作的研究方法也可以被应用到抽样过程中来，我们这里将对节点抽样、连边抽样和滚雪球抽样进行简要的介绍。

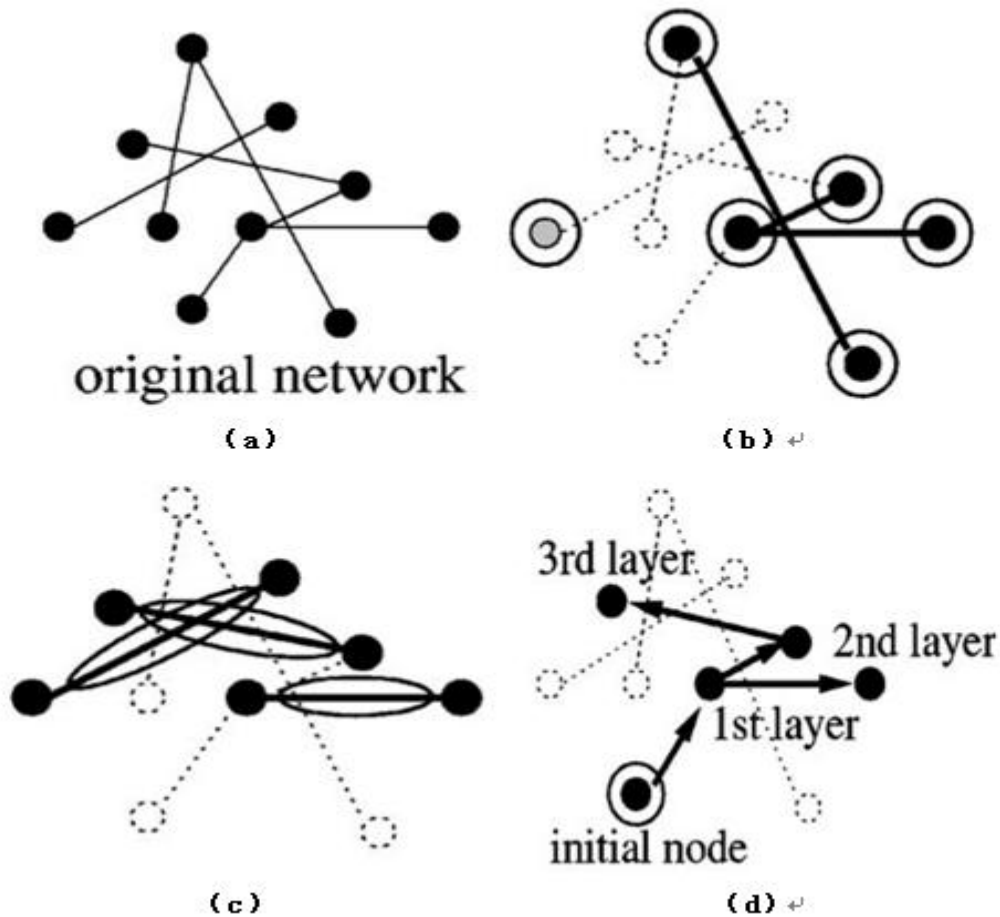


图 2.3 三类抽样方法 (a) 抽样原网络 (b) 节点抽样：选择被圈出的节点，保留节点间的三个连接，并去除孤立节点。(c) 连接抽样：选出 3 个被圈出的连接，及 6 个与它们相连的节点。(d) 滚雪球抽样：从被圈出的节点开始，通过连接路径选择取其相连的节点。

Fig.2.3 Three kinds of sampling method. (a) original network (b) Node sampling: Select the circled nodes, keep three links among them, and the isolated node is removed. (c) Link sampling: Select the three circled links and six nodes attached to them. (d) Snowball sampling: Starting from the circled node, select nodes and links attached to them by tracing links.

在节点抽样过程中，确定数目的节点被随机选出，节点间的连边也被保留下来。这种抽样方法的抽

样比例被定义为被选节点的个数与原网络总节点的个数比。在图 2.3(b)中，尽管可完全预言独立节点的存在，但为了方便起见，我们还是将其忽略，所以抽样网络中的节点个数要少于抽出节点的个数。因为关系到抽样网络的平均度和平均路径长度，我们需要对抽出节点及其连边的比例关系进行观察，发现若假设抽出节点的比例为 α ，它们之间的连接比例为 β ，则当我们随机的抽取节点时，满足关系 $\beta \sim \alpha^2$ 。

n 个抽取节点的最大连边可能为 $\binom{n}{2} = n(n-1)/2 \sim n^2$ 。

对于连边抽样，如图 2.3(c)所示，随机数目的连边及与它们相连的节点被随机选出。

对于滚雪球抽样^[40, 41]，我们首先选择一独立节点及其与它直接相连的连边，在下一步中，我们选出所有与上一步相连的节点，持续这个过程到理想数目的节点被抽样出来。第 n 步的被选节点集被定义为第 n 层。以图 2.3(d)为例，为了控制抽样节点的总数，我们从最后一层中选取任意数目的节点。这类似于计算渗透簇分数维的聚集增长方法^[42]。由于中心节点的高连接性，滚雪球方法更易于将其抽取出来。滚雪球抽样方法已经体现出了随机游走抽样的核心思想，对于不同类型的随机游走抽样，我们将在下文 中加以详细介绍。

§ 2-4 小结

本章中我们介绍了从原网络中抽取子网络的基本过程，并分析了用随机游走的策略在网络上进行抽样时，游走的混合率和传导率问题，发现用随机游走的策略对网络进行抽样时，达到稳态分布 π_i 时随机游走的某一步到达节点 i 的概率正比于 i 点的度， $d(i)$ 。收敛于稳态 π 的速率由转移矩阵 P 的第二大本征值的模控制。当网络具有一个好的传导率时，随机游走便可以更容易的游走到网络的其他区域，快速的达到平衡态。并给出了多路线随机游走的平均场方程。本章也介绍了节点、连边、滚雪球等抽样方法。

第三章 随机游走抽样

§ 3-1 引言

在过去的几年中，研究者在复杂网络领域做了大量的工作。这些工作涉及数学、统计物理、计算机科学、社会学、生物科学等交叉学科^[43-45]。已经有不同的网络模型被用来介绍和研究真实的网络。例如，WS 网络、BA 网络和 ER 网络等^[33, 34, 37]。但这些并不足以准确的探索大多数的网络。很多时候我们需要对整个现实网络加以研究以进行更为精确的分析。

但是，在实际应用中许多关于复杂网络的研究仅只是针对实际网络中的很小一部分。这一发现在蛋白质相互作用^[46-48]，基因规则^[49]、新陈代谢等网络^[50]中表现的十分明显。在这样的生物网络中，一个细胞只有很小一部分的子网被抽样出来。更重要的是，由于许多包括 Facebook 在内的在线社会网络^[51]，不愿公开他们的公司数据，所以作为研究者，特别难以得到完整的数据集。另一方面，无论从时间消耗和经济利益哪一方面考虑，对一完整而巨大的网络进行研究都是很不划算的。

这一观察也就随之提出了一个有趣而重要的问题——如何对一个具有代表性的完整网络进行网络抽样。当一个巨大的包含成千上万节点和连边的网络摆在我们眼前时，如何通过一个“好”的抽样方法得到一个足够“小”却足以代表原网络各项性质的子网络，研究结果尚不明确！根据相互影响的节点与连边间的关系，有几种方法（节点抽样、连边抽样、滚雪球抽样等）用以进行网络抽样，得到具有不同原网络特征的子网络。

研究者已经做了很多相关的工作，包括通过爬虫软件爬取的数据对 WWW 网络的抽样效果进行分析^[52]；通过路由探测器对因特网进行的调查^[53]；爬取一些公共的可以得到的社会网络，并首次得到了关于 Facebook 的无偏抽样网络^[54]。随机游走不仅被认为是研究社会问题的有效方法^[41, 55, 56]。探索复杂网络上的随机游走问题也具有较高的理论意义^[57-65]。另外，许多工作专门针对作为网络抽样方法的随机游走研究展开^[30, 71, 66-70]。

在参考文献[71]中提及了一系列相关的研究结果，他们研究并描述了，在不同的抽样策略下一些网络的节点及连边的覆盖效率。和这些网络的拓扑量度（节点的平均度和网络的平均聚类系数）如何得

到更准确的评估。

本文的研究直接针对更为一般的网络，包括 WS 小世界网络和加权的美国航空网等。我们更有兴趣的探索了不返回上一步节点和根据节点度分布的随机游走策略对网络抽样的影响。

以下，我们将首先概述有关随机游走网络抽样的基本概念进而通过数值分析，探索 ER、BA、WS 和加权的美国航空网的相关性质。

§ 3-2 随机游走抽样的基本概念

原网络 G 被定义为 V 和 E 的集合， $G = (V, E)$ ，其中 V 是元素的非空可数集，被称为定点或节点； E 是针对不同节点间连接的无序数组，被称作边或连接。

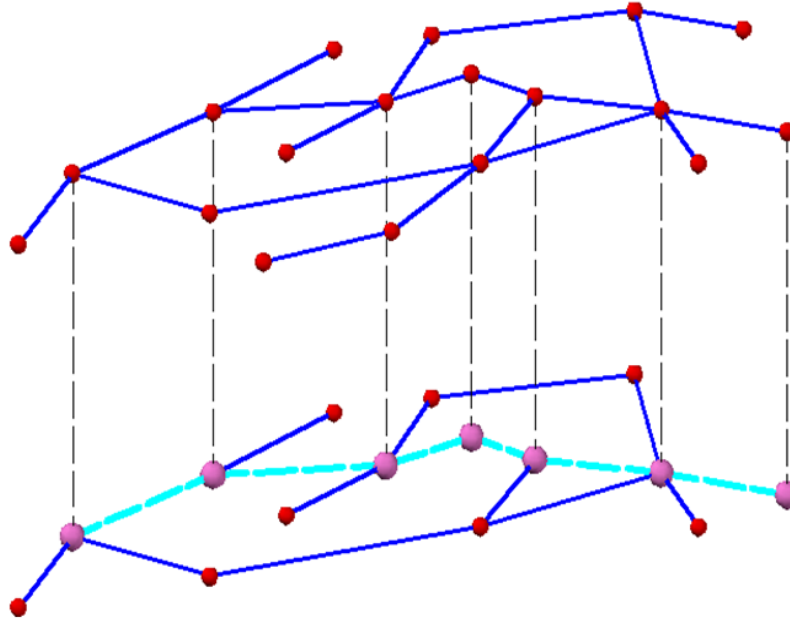


图 3.1 从原网络中进行抽样的游走路径图。其中，大球和虚线代表了具体的游走路径，下部分的其它节点为游走点的邻居。

Fig. 3.1 the way to sample a subnet from the original complex network. The bigger balls and the dashed line between them represent the random walk track. Other balls of under part are neighbors of the traced nodes.

在对原网络进行抽样时，我们可以得到一个不完全的特定子网 Γ ， $G \neq \Gamma$ 。它由游走者经过的节点、这些节点的邻居及其连接组成^[71]。我们用 $((i_1, A_1, E_1); (i_2, A_2, E_2), \dots, (i_M, A_M, E_M))$ 来代表抽样子网。

其中, i_p 是利用随机游走方法在 G 中抽取的节点, A_p 是节点的各自连接序列集, E_p 则包含了 A_p 中节点间的连边。我们发现根据以上定义必然有 $i_{p+1} \in A_p$, 而 (i_1, i_2, \dots, i_M) 则对应于沿 G 的游走路径。

图 3.1 表明了从原来的复杂网络中抽样出一个子网络的抽样路径。图的上半部分代表了最初的原网络; 下半部分代表了, 考虑抽样节点与邻居节点间连接的随机游走抽样得到的不完整子网络。

§ 3-3 三类随机游走抽样方法

1、无限制的随机游走 (URW)

这是一类最简单的随机游走, 节点 u 从它的邻居中等概率的任意选取一个节点 v 作为它下一步的要走的节点。

2、根据节点度选择种子节点的随机游走 (CSNRW)

所谓种子节点即游走者第一步选择的节点, 整个游走过程由这一点开始逐步向下进行, 所以显得格外重要。我们根据节点度选择种子节点, 节点的度越大被选择的概率就越大, 在某些网络中这也就限定随机游走的初始游走范围。这样的游走方式展现了初始的选择对后期抽样的影响。

3、不返回上一步的随机游走 (NRRW)

游走过程中, 除上一步已走过的节点外, 等概率选择其它邻居作为下一步游走节点。通过这一游走方式我们可以发现局部变化对整体游走产生的影响。

§ 3-4 抽样网络

我们选择 ER 随机、BA 无标度、WS 小世界和加权的美国航空网作为我们进行游走的初始网络。对于以上网络模型, 我们确定网络的总结点数 N 和每一节点对应的边数 m 。

在 ER 随即图中, 总结点数为 N , 我们假设两点间是否存在连接独立于其它边的连接情况, 所以每条边的存在有独立的概率 p 决定, $p = 2m / (N - 1)$ 。

对于 BA 网络, 总结点数为 N , 网络开始时具有 $m_0 = m$ 个节点, 每一次新节点被逐渐的加入到原网络中, 并与 m 个原网节点相连, 连接概率正比于所连节点的度的大小。

WS 小世界网络由 N 个节点组成, 每一节点具有 $m(m < N/2)$ 个邻居, 其后, 我们将连边以概率 $p=0.3$

随机的连接于其它节点。

美国航空网描述了美国部分机场的航线连接情况，它由 $N=332$ 个节点组成，每一连边的权重代表了两机场间航班的疏密情况。

§ 3-5 抽样效率

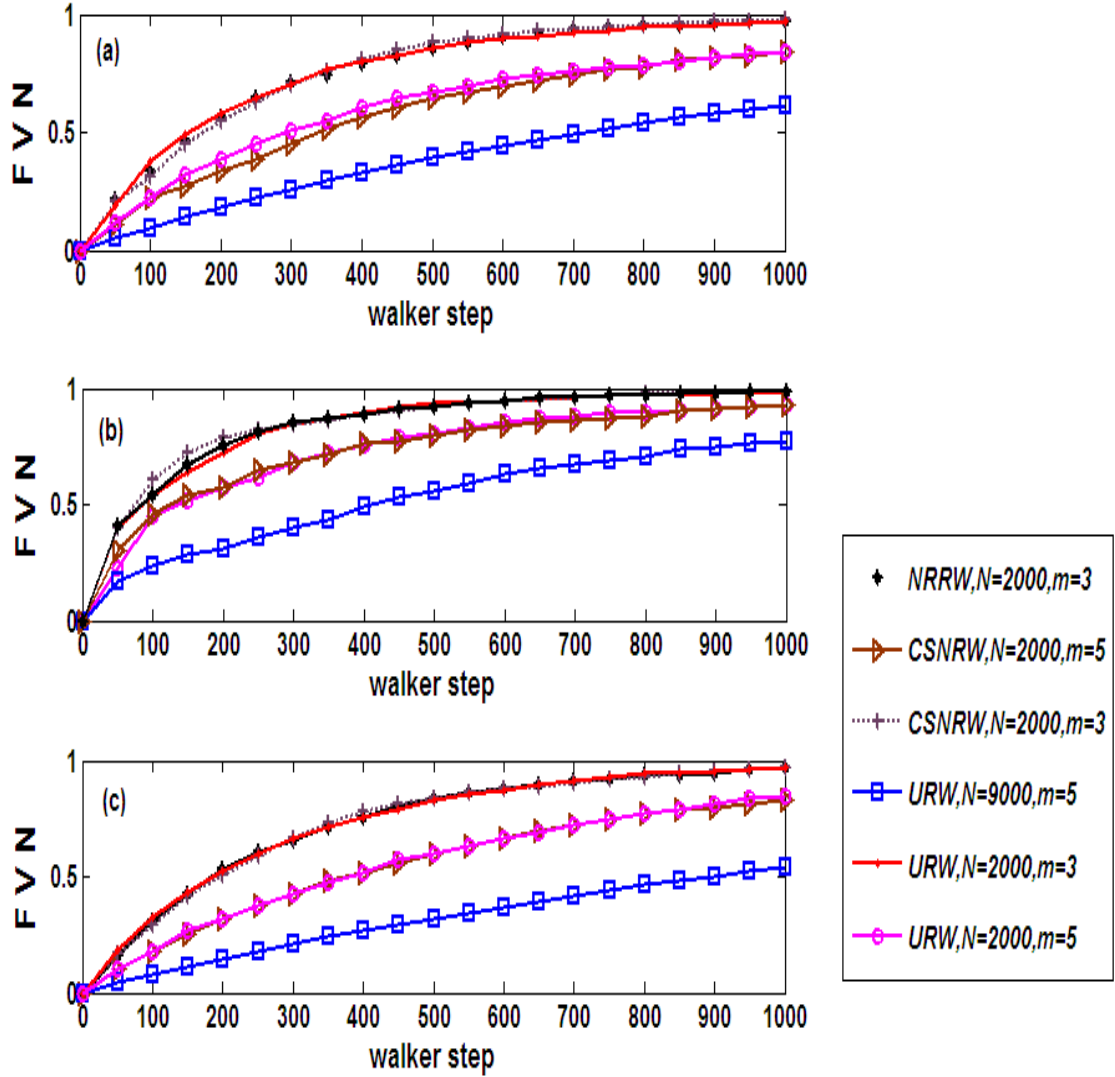


图 3.2 通过 URW、CSNRW 和 NRRW 抽样方法，在 ER (a)、BA (b)、WS (c) 网络上得到的抽样点的比例随游走步数 t 的变化情况，其中 $N=2000, 9000$ ， $m=3, 5$

Fig.3.2 The ratio of sampled nodes in terms of the steps t for $N = 2000, 9000$, and $m = 3, 5$, by the way of URW, CSNRW and NRRW, for ER (a), BA (b), WS (c) network models. FVN is the fraction of visited nodes.

图 3.2 解释了 N 、 m 是进行网络抽样的重要影响因素。我们固定 m ，而使 N 减小，抽样效率如预期般提高。当 $N=2000$ ， $m=3,5$ 时，我们发现减小 m 也可以使节点抽样更为有效。较小的 m 也会提高网络节点的抽样效率。这是因为，增大的 m 会使网络的连接更为复杂，游走者回到那些已被抽样的节点也更为容易。另外，当我们将随机游走的搜索策略 URW 改进为 $CSNRW$ 、 $NRRW$ 时，图 3.2 中的曲线即存在一个接近线性的初始区域，在游走足够的步数后，当访问到所有的节点时便出现了饱和状态。对于我们作为初始网络的三类复杂网络模型，它们代表了许多真实网络的基本特征，并具有不同的抽样特征和抽样阈值。我们发现，在 BA 网络上的随机游走，由于存在具有代表性的中心节点，所以比 WS 和 ER 网络具有更高的抽样效率。三种抽样策略在 ER 和 WS 网络中具有相同的表现，而 $CSNRW$ 和 $NRRW$ 抽样相较 URW 抽样在 BA 网络中具有更高的效率。这表明中心节点的存在不仅提高了网络的抽样效率，也使得每一游走步的改变对三种抽样策略更为敏感。

§ 3-6 抽样网络与原网络的度分布对比

在图 3.3 中我们可以轻易看出通过 URW 、 $CSNRW$ 和 $NRRW$ 方式抽样出的子网络的度分布与原网络的度分布存在些许差异。图 3.3 展示了 ER 、 BA 、 WS 和加权的美国航空网在用上述三类方法进行 100 步游走抽样后得到的抽样子网与原网络的度分布对比情况。有趣的是， ER 和 WS 网络的抽样子网与原网络具有相同的度分布形态，且更加偏向度大的一侧。另一方面，对于满足幂率分布的 BA 和美国航空网，具有类似的变化，即度分布的斜率变得更高。因为节点具有的邻居越多，这一点被走到的概率就越大。与此同时，略微的偏差也是对方程 2.2 的证明，说明抽样的确是从度大的节点开始的。很明显的，网络的加权对抽样子网度分布的影响并不大。所有随机游走的抽样方法都会导致对子网度分布的高估，但并不明显。

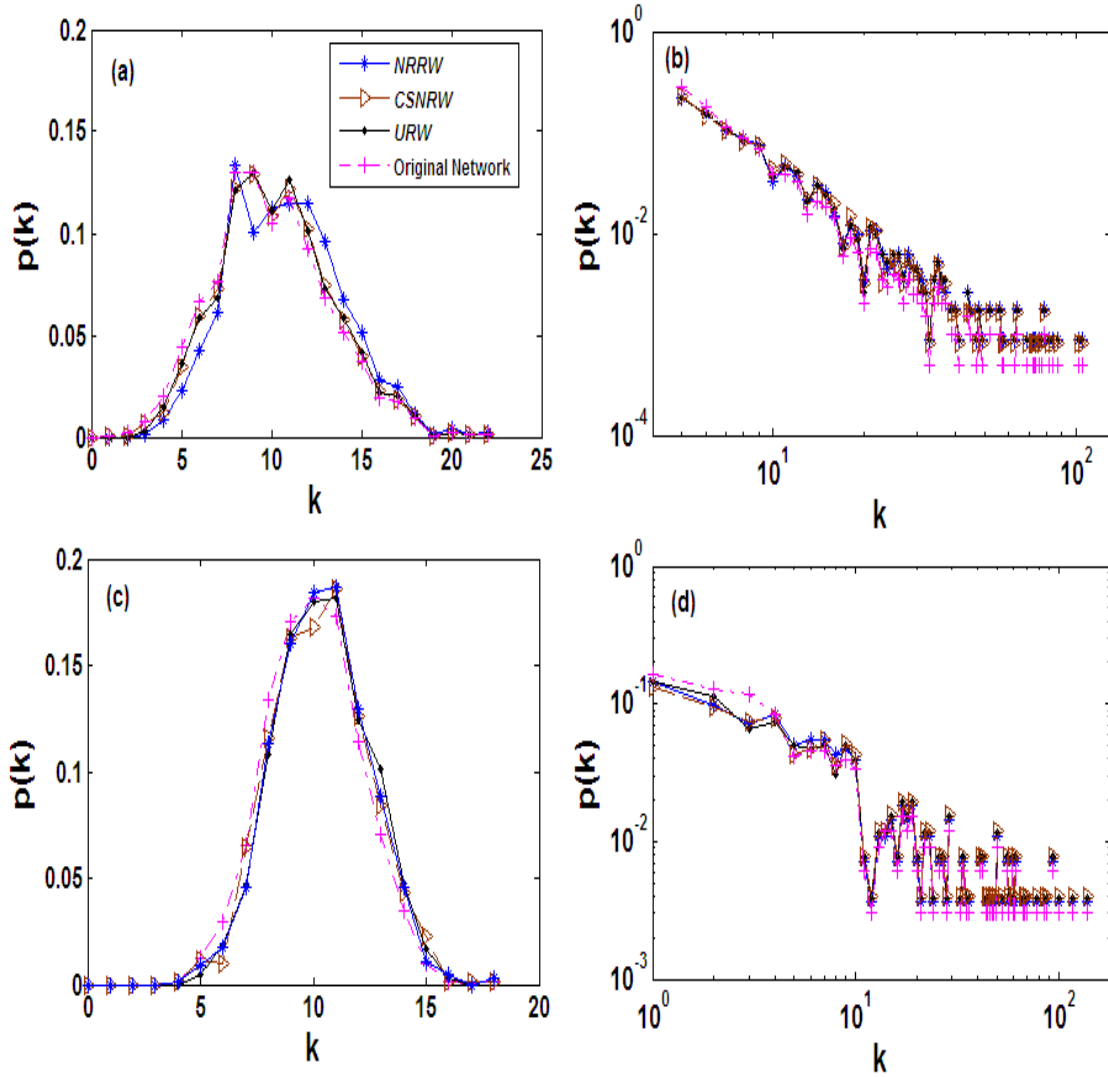


图 3.3 当 $N=2000$, $m=5$ 时 ER (a)、BA (b)、WS (c) 以及具有 332 个节点的美国航空网的度分布图。子网的度分布在 URW、CSNRW、NRRW 游走策略下游走 100 步后获得。

Fig.3.3 The degree distribution for the ER (a), BA (b), WS (c) networks with 2000 nodes and $m = 5$. And USAir (d) network with 332 nodes. The degree distributions in the subnets created by choosing each node through URW, CSNRW, NRRW strategy after 100 steps.

§ 3-7 抽样子网平均度和平均聚类系数随步数的变化

我们在图 3.4 中展示了抽样子网的平均度随游走步数的变化情况。我们以上述四类网络作为初始网络，在 ER, BA 和 WS 网络中， $N=2000$, $m=5$ 。具有 332 个节点的美国航空网，则通过飞机场间的班

次进行加权。

对于子网平均度的演化，图 3.4 展现出 $\langle k \rangle$ 逐渐降低。图 3.4 和方程 2.2 中的发现被进一步证实。在经过足够步数后，尽管略微高于原网平均度，子网的平均度基本可以回到初始 $\langle k \rangle$ 。另外，三种抽样策略在 ER 和 BA 网上具有相同的抽样效果，但是对于 WS 和美国航空网而言，CSNRW 的游走策略显示出了更高的效率。图 3.4 (d) 表征了 $\langle k \rangle$ 在美国航空网中梯形下降的趋势。

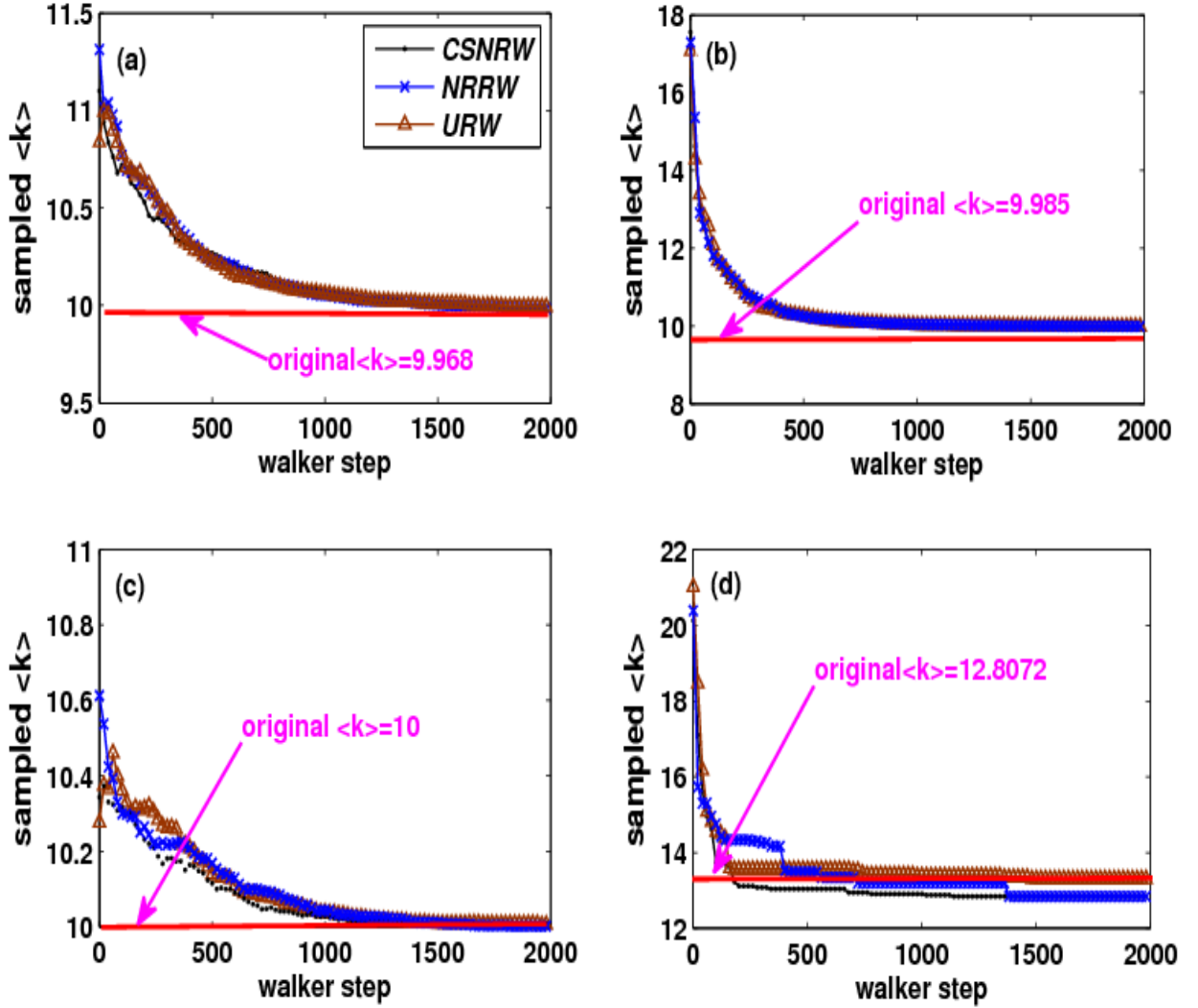


图 3.4 在 ER (a), BA (b), WS (c), USAir (d) 抽样子网中, $\langle k \rangle$ 在每一种随机游走抽样策略下随游走步数的变化情况

Fig.3.4 Walker step evolution of each random walk sample as a function of the $\langle k \rangle$ for sampled ER (a), BA (b), WS (c), USAir (d) subnets.

我们在图 3.5 中展现了平均聚类系数 $\langle c \rangle$ 在不同的抽样策略及四类网络中的改变情况。相较于 $\langle k \rangle$

在不同抽样策略下的改变曲线基本一致， $\langle c \rangle$ 在抽样的初始步中改变更为明显。URW 策略在初始时具有最大值。与此相对，CSNRW 策略所得到的值比其它策略得到的值更小。这就意味着，选择度大的种子节点会使网络的平均聚类系数降低。说明在网络的 $\langle k \rangle$ 反比于 $\langle c \rangle$ ，即当一个节点的度较大时与它相连的其它节点便更为陌生。WS 网络中，曲线总是低于初始值，进而逐步趋近原值。这都说明节点的邻居中 k 大的节点间更易互不相识。对于加权的美国航空网，我们发现无论子网中的 $\langle k \rangle$ 还是 $\langle c \rangle$ 都成梯形下降趋势。且美国航空子网的 $\langle c \rangle$ 也被高估。

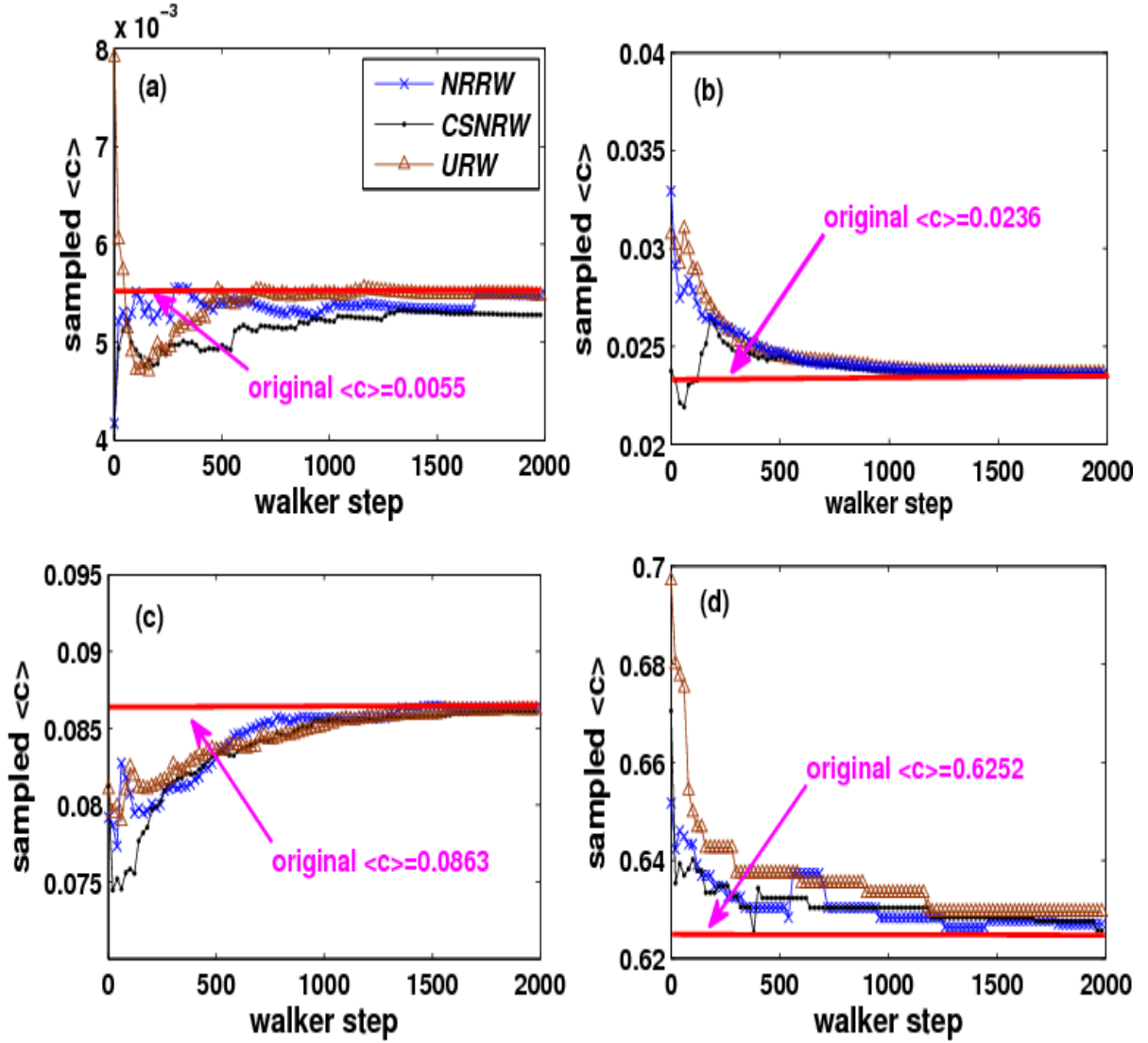


图 3.5 在 ER (a), BA (b), WS (c), USAir (d) 抽样子网中， $\langle c \rangle$ 在每一种随机游走抽样策略下随游走路数的变化情况

Fig.3.5 Walker step evolution of each random walk sample as a function of the $\langle c \rangle$ for sampled ER (a), BA (b), WS (c), USAir (d) subnets.

§ 3-8 小结

大多数现实网络只有部分为我们所知。并且对整体网络的一部分进行研究对我们而言显得更有意义。当节点和连接抽样存在偏差时，寻找一种更为准确的抽样策略就成为了一个重要课题。这里，我们探讨三种类型的随机游走：无限制的随机游走（URW），选择种子节点的随机游走（CSNRW），及不返回上一步的随机游走（NRRW），在加权的美国航空网，ER，BA，WS 网络中各自的抽样情况。我们通过模拟研究了诸如抽样效率，度分布，抽样子网的平均度 $\langle k \rangle$ 和平均聚类系数 $\langle c \rangle$ 等子网性质。得到了一系列的相关结果：当网络特征量足够小时我们可以从原网络中抽取出所有的节点。具有相同节点数目和节点度的三类不同复杂网络模型具有不同的抽样效率和阈值；无论在何种网络中，三种抽样策略下的子网度分布都偏向较高一侧。更重要的，我们展示出子网的 $\langle k \rangle$ 、 $\langle c \rangle$ 都可在有限步数内回到原值，并且在加权网络中更易被高估。这一变化过程也证明了这样一种观点：节点邻居的 k 越大则节点间便具有较少的连接。

第四章 总结与展望

现阶段我们只能对实际网络中很小的一部分加以研究。例如，在蛋白质组织中只有 10–80% 的蛋白质被加以分析研究。所以了解不同的随机游走抽样策略在一些具有代表性的网络中的抽样效果就显得格外重要。本文中，我们对三类游走的抽样效率和子网中的一些重要参量加以研究，这三类抽样策略（URW、CSNRW、NRRW），被分别应用于三类具有代表性的网络模型（ER, BA, WS）和加权的现实网络（USAir）中。并以此为原始网络进行模拟。

得到的一系列结果有力的证明了当 N 和 m 足够小时，就可以在较少的步数内从原网络中抽样出所有的节点。此外，三类抽样策略在无论何种网络中的抽样子网度分布都具有类似的表现（稍偏向度大的一侧）。我们也描述了抽样子网的平均度和平均聚类系数随抽样步数的演化情况。发现演化的过程曲线并不光滑，而且在加权网络中得到的结果也更易被高估。相对于平均度的改变，平均聚类系数的波动变化更为明显。通过对抽样子网 $\langle k \rangle$ 和 $\langle c \rangle$ 改变的比较，发现当某一节点的度较大时其邻居间便具有较少的连接。

虽然我们对实际的加权网络进行了研究，然而仍有大量的相关工作有待我们去完成。此外，研究抽样子网的诸如最短路径长度、网络直径等其它性质对于网络抽样的研究也很重要。发展新的抽样策略也是今后很重要的研究课题。

参考文献

- [1] P. Erdos, A. Renyi, On random graphs, Publ. Math. 6, 290–297, 1959.
- [2] P. Erdos, A. Renyi, On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci. 5, 17–60, 1960.
- [3] P. Erdos, A. Renyi, On the strength of connectedness of random graphs, Acta. Math. Sci. Hung. 12, 261–267, 1961.
- [4] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286, 509–512, 1999.
- [5] R. K. Merton, The Matthew effect in science, Science 159, 56–63, 1968.
- [6] H. A. Simon, On a class of skew distribution functions, Biometrika 42, 425–440, 1955.
- [7] L. Lovasz. Random walks on graphs. a survey. In Combinatorics, 1993
- [8] R. Motwani, P. Raghavan, Markov Chains and Random Walks, Cambridge University Press, Ch. 6, pp. 127–160, 1995.
- [9] L. Lovász, Random walks on graphs: A survey, 1993.
- [10] A. Sinclair, Improved bounds for mixing rates of marked chains and multicommodity flow, in: Lecture Notes in Computer Science, Proceedings of the 1st Latin American Symposium on Theoretical Informatics (LATIN 92), Vol. 583, Springer-Verlag, pp. 474–487, 1992.
- [11] A. Tahbaz-Salehi, A. Jadbabaie, Small world phenomenon, rapidly mixing markov chains, and average consensus algorithms, in: Proceedings of the 46th IEEE Conference on Decision and Control, IEEE Computer Society, pp. 276–281, 2007.
- [12] C. Gkantsidis, M. Mihail, A. Saberi, Conductance and congestion in power law graphs, ACM SIGMETRICS Performance Evaluation Review 31, 148–159, 2003.
- [13] A. Broder, R. Kumar, F. Maghoul, et al, Graph structure in the web, Computer Networks: The International Journal of Computer and Telecommunications Networking 33, 309–320, 2000.
- [14] C. Gkantsidis, M. Mihail, A. Saberi, Random walks in peer-to-peer networks, in: Proceedings of the Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies, INFOCOM 2004, Vol. 1, Hong Kong, pp. 120–130, , 2004.
- [15] L. A. Adamic, B. A. Huberman, R. M. Lukose, A. R. Puniyani, Search in power law networks, Physical Review E 64, 46135–46143, 2001.
- [16] M. E. J. Newman, S. H. Strogatz, D. J. Watts, Random graphs with arbitrary degree distributions and their applications, Physical Review E 64 (2) 026118–1, 026118–17, 2001.

- [17] M. Faloutsos, P. Faloutsos, C. Faloutsos, On power-law relationships of the Internet topology, *Comput. Commun. Rev.* 29, 251, 1999.
- [18] M. E. J. Newman, Scientific collaboration networks: I. Network construction and fundamental results, *Phys. Rev. E* 64, 016131, 2001.
- [19] M. E. J. Newman, Scientific collaboration networks: I. Network construction and fundamental results, *Phys. Rev. E* 64, 016131, 2001.
- [20] E. Costenbader, T. W. Valente, The stability of centrality measures when networks are sampled, *Soc. Networks* 25, 283, 2003.
- [21] G. Robins, P. Pattison, J. Woolcock, Models for social networks with missing data, *Soc. Networks* 26, 257, 2004.
- [22] G. Kossinets, Assortative model for social networks, eprint cond-mat/0306335, 2003.
- [23] T. Petermann, P. De Los Rios, Exploration of Scale-Free Networks - Do we measure the real exponents, *Eur. Phys. J. B* 38, 201, 2004.
- [24] A. Clauset, C. Moore, Accuracy and scaling phenomena in Internet mapping, *Phys. Rev. Lett.* 94, 018701, 2005.
- [25] D. Achlioptas, A. Clauset, D. Kempe, et al, Power-law degree distributions in regular graphs. eprint cond-mat/0503087, 2003.
- [26] L. Dall'Asta, I. Alvarez-Hamelin, A. Barrat, et al, Statistical theory of Internet exploration, *Phys. Rev. E* 71, 036135, 2005.
- [27] M. P. H. Stumpf, C. Wiuf, R. M. May, Subnets of scale-free networks are not scale-free, *Proc. Natl. Acad. Sci. U.S.A.* 102, 4221, 2005.
- [28] J. Scholz, M. DeJori, M. Stetter, et al, Statistical Mechanics and its Applications, *Physica A* 350, 622, 2005.
- [29] J-D. J. Han, D. Dupuy, N. Bertin, et al, Effect of sampling on topology predictions of protein-protein interaction networks, *Nat. Biotechnol.* 23, 839, 2005.
- [30] M.P.H. Stumpf, C. Wiuf, Sampling properties of random graphs: the degree distribution, *Phys. Rev. E* 72, 036118, 2005.
- [31] K-I. Goh, B. Kahng, D. Kim, Universal behavior of load distribution in scale-free networks, *Phys. Rev. Lett.* 87, 278701, 2001.
- [32] K-I. Goh, E. Oh, H. Jeong, et al, Classification of scale-free networks, *Proc. Natl. Acad. Sci. U.S.A.* 99, 12583, 2002.

- [33] D. J. Watts, S. H. Strogatz, Collective dynamics of ‘small-world networks, *Nature London* 393, 440, 1998.
- [34] A-L. Barabási, R. Albert, Emergence of scaling in random networks, *Science* 286, 509, 1999.
- [35] P. L. Krapivsky, S. Redner, Organization of Growing Random Networks, *Phys. Rev. E* 63, 066123, 2001.
- [36] R. Albert, H. Jeong, A-L. Barabási, Error and attack tolerance of complex networks, *Nature London* 406, 378, 2000.
- [37] R. Cohen, K. Erez, D. ben-Avraham, et al, Resilience of the Internet to Random Breakdowns, *Phys. Rev. Lett.* 85, 4626, 2000.
- [38] R. Cohen, K. Erez, D. ben-Avraham, et al, Breakdown of the internet under intentional attack, *Phys. Rev. Lett.* 86, 3682, 2001.
- [39] L. K. Gallos, R. Cohen, P. Argyrakis, et al, Stability and Topology of Scale-Free Networks under Attack and Defense Strategies, *Phys. Rev. Lett.* 94, 188701, 2005.
- [40] J. Wiley, Sons, The term snowball sampling is from Steven K. Thomson Sampling Inc., New York, 2002.
- [41] M. E. J. Newman, Ego-centered networks and the ripple effect, *Soc. Networks* 25, 83, 2003.
- [42] C. Song, S. Havlin, H. A. Makse, Self-similarity of complex networks, *Nature London* 433, 392, 2005.
- [43] M. E. J. Newman, The structure and function of complex networks, *SIAM Rev.* 45, 167, 2003.
- [44] R. Albert, A-L. Barabási, Statistical mechanics of complex networks, *Rev. Mod. Phys.* 74, 47, 2002.
- [45] S. N. Dorogovtsev and J. F. F. Mendes, Evolution of networks, *Adv. Phys.* 51, 1079, 2002.
- [46] P. Erdos, A. Renyi, *Publ. Math.*, et al, On the evolution of random graphs, *Acad. sci., Ser. A5*:17~60, 1960.
- [47] H. Qin, , H.S. Lu, , W. B. Wu, et al, Evolution of the yeast protein interaction networks, *Proc. Natl. Acad. Sci. USA* 100, 12820–12824, 2003.
- [48] S. Maslov, K. Sneppen, Specificity and stability in topology of protein networks, *Science* 296, 910–913, 2002.
- [49] V. Noort, B. Snel, M. Huynen, *EMBO Rep.* the yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model, 5 280–284, 2004.
- [50] H. Jeong, B. Tombor, R. Albert, et al, The large-scale organization of metabolic networks, *Nature* 407, 651–654, 2000.
- [51] M. J. E. Newman, J. Park, The origin of degree correlations in the Internet and other networks, *Phys. Rev. E*, 026112, 2003.
- [52] M. A. Serrano, A. Maguitman, M. Boguñá, S. Fortunato, and A. Vespignani, Decoding the structure of the WWW, eprint, cs.NI/0511035, 2005.

- [53] L. Dall'Asta, E. Alvarez-Hamelin, A. Barrat, A. Vázquez, et al, Exploring networks with traceroute-like probes: Theory and simulations, *Theor. Comput. Sci.* 355, 6, 2006.
- [54] M. Gjoka, M. Kuran, Carter T. Butts, et al, A Walk in Facebook: Uniform Sampling of Users in Online Social Networks, arXiv 0906.0060v3.[cs.SI], 2010.
- [55] M.J. Salganik, D.D. Heckathorn, Sampling and estimation in hidden populations using respondent-driven sampling, *Sociol. Methodol.* 34,193, 2004.
- [56] M.J. Salganik, P. S. Dodds, D. J. Watts, Experimental study of inequality and unpredictability in an artificial cultural market, *Science* 311, 854, 2006.
- [57] B. Tadic, Adaptive random walks on the class of web graphs, *Eur. Phys. J. B* 23, 221, 2001.
- [58] E.M. Boltt, D. ben Avraham, What is special about diffusion on scale-free nets?, *New J. Phys.* 7, 26, 2005.
- [59] J. D. Noh, H. Rieger, Constrained spin-dynamics description of random walks on hierarchical scale-free networks, *Phys. Rev. E* 69, 036111, 2004.
- [60] S. A. Pandit, R. E. Amritkar, Random spread on the family of small-world networks, *Phys. Rev. E* 63, 041104, 2001.
- [61] K. A. Eriksen, I. Simonsen, S. Maslov, et al, Modularity and extreme edges of the internet, *Phys. Rev. Lett.* 90, 148701, 2003.
- [62] J. D. Noh, H. Rieger, Random walks on complex networks, *Phys. Rev. Lett.* 92, 118701, 2004.
- [63] I. Simonsen, Diffusion and networks: A powerful combination!, *Physica A* 357, 317, 2005.
- [64] D. Volchenkov, P. Blanchard, Random walks along the streets and canals in compact cities: spectral analysis, dynamical modularity, information, and statistical mechanics, e-print physics/0608153. 2006.
- [65] S. Fortunato, A. Flammini, Random walks on directed networks: the case of pagerank, e-print physics/0604203, 2006.
- [66] M. P. H. Stumpf, C. Wiuf, R. M. May, From the cover: subnets of scale-free Networks are not scale-free: sampling properties of networks, *Proc. Natl. Acad. Sci. U.S.A.* 102, 4221. 2005.
- [67] S. H. Lee, P.-J. Kim, H. Jeong, Statistical properties of sampled networks, *Phys. Rev. E* 73, 016102, 2006.
- [68] S.-J. Yang, Exploring complex networks by walking on them, *Phys. Rev. E* 71, 016107, 2005.
- [69] A. Ramezanpour, Intermittent exploration on a scale-free network, e-print cond-mat/0607327, 2005.
- [70] D. Stauffer and M. Sahimi, Diffusion in scale-free networks with annealed disorder, *Phys. Rev. E* 72, 046128, 2001.
- [71] L.d.F. Costa, G. Travieso, Exploring complex networks through random walks, *Phys. Rev. E* 75, 016102, 2007

致 谢

洋洋洒洒几万字，不曾想能出于我手，两年半的时间，白驹过隙般一晃而过，昨日我熟悉的，早已越过学校的这扇窗，奔向自己的人生；今日熟悉我的，也来挥手送别。我们总是来不及怀念却早已各赴前程。如果说还留下些什么也绝非这一纸论文，几张照片，更多的是情——师生情、同窗情；是智——才智、理智。

不经历迷茫便总也不能寻找光明的方向，感谢张旭老师的言传身教，更感谢她跳出学海，站在人生的制高点为我指明方向；若说求学之路坎坷难行，科研之路就更是布满荆棘，它肆无忌惮的延伸出来遮蔽你的双眼，一不小心更是束缚住你的手脚让你动弹不得，衷心感谢李再东老师、赵培德老师、周国香老师、曹天光老师，是你们无私的给我“镰刀”，予我“拐杖”，让我摆脱束缚，拨开一条光明之路！也感谢一路陪伴的战友们——孟浩、姚淑芳、杨栋、路遥乾、李瑞金、孟香叶、赵永芳、赵鸿雁、朱素芳、赵飞；师姐——刘艳、孙亚周；师弟——王兵、李子阳、白建华、张鹏，是你们的鼓励让我一路向前，与你们的激辩使我获得智慧。感谢杨磊一路上默默地支持与鼓励，感谢上天让我们在人生最美好的年华相遇，执手并肩，开创属于我们的人生！

有些人，我不曾感谢；有些人，我一直想要感谢；对于有些人，感谢不足以表达心中之情，对于有些人的情，我将用一生来报！谨以此文献给最爱我以及我最爱的父母，作为我们人生中的一个小小纪念！

攻读学位期间所取得的相关科研成果

- [1] 谢昀雅, 孟浩, 张旭, 随机游走的网络抽样, 中国物理学会 2011 年秋季学术会议, 2011, 9.
- [2] 孟浩, 谢昀雅, 张旭, 大规模社交网络中基于结点度的链路预测, 中国物理学会 2011 年秋季学术会议, 2011, 9.
- [3] Yun-ya Xie, Xu Zhang, Hao Meng, Sampling from complex networks by random walks, Eur. Phys. J. B, contributed