文章编号:1001 - 9081(2008)02 - 0367 - 04

基于 S-RTT 策略的 BitTorrent 文件共享模型

孙建华,王战国,陈 浩,石 林
(湖南大学 计算机与通信学院,长沙 410082)
(jhsun@ aimlab. org)

摘 要:分析了 BitTorrent 文件共享系统的工作流程,指出了在种子节点选取和连接方式上的缺点。介绍了网络定位技术 GNP,详细分析了 GNP 的工作流程,说明了它在网络中定位主机的优越性能。提出了一种 S-RTT 策略,进行种子节点选取和连接。将此策略引入 BitTorrent 文件共享系统中,使得 BitTorrent 系统能够灵活选择优质的种子节点、控制信息流量,有效改善了 BitTorrent 网络性能。

关键词:对等网络;BitTorrent;文件共享;网络定位;网络流量
中图分类号: TP393 **文献标志码**:A

BitTorrent file-sharing model based on S-RTT strategy

SUN Jian-hua, WANG Zhan-guo, CHEN Hao, SHI Lin
(*School of Computer and Communication, Hunan University, Changsha Hunan* 410082, *China*)

Abstract: In this paper, we analyzed the workflow of BitTorrent file-sharing system, and pointed out the shortcomings in ways of choosing and linking seed. The GNP technology was recommended, its workflow was analyzed in detail, and the superiority of it was shown in locating mainframe. This article described one kind of BitTorrent file-sharing model based on S-RTT strategy that led S-RTT strategy into BitTorrent system. It enables BitTorrent system to choose the good seeds flexibly and control network flow, and improves the network function of BitTorrent system effectively.

Key words: Peer-to-Peer; BitTorrent; file-sharing; network positioning; network traffic

0 引言

近年来,对等网络技术(Peer-to-Peer Networks,P2P)迅速发展,它最大的特点是网络客户之间直接共享信息,这使得网络客户之间真正实现了平等。在 P2P 模式中,Peer 之间可以直接相连,信息传送可以不必经过特定的服务器[1]。每一个 Peer 既可以是客户机,又可以作为服务器,文件信息真正实现了平等共享,这对网络客户具有很大的吸引力。

尽管 P2P 有很多优点,由于每个 Peer 都可以转发信息,致使网络信息流量急剧上升[2],拥塞现象经常发生。同时客户的信息存储和转发欲望是无限增长的,致使这种情况更加严峻。据统计 P2P 在网络中所占的流量已经超过一半,而且比例还有不断增大的趋势。P2P 网络通信会导致异常的流量峰值,所带来的网络拥塞、性能下降等问题,已影响到其他正常的网络应用,如 WWW、E-mail 等,缓慢的网页浏览和收发邮件速度引起普通客户的不满。如何减少 P2P 系统的网络流量,缓解拥塞状况就成为一个重要的课题。

本文从分析混合 P2P 网络典型代表 BitTorrent(以下简称 BT)系统的工作过程入手,分析 BT 文件共享系统[3]中网络流量、拥塞状况和资源占用的可改进性。介绍了网络定位技术,对典型的网络定位技术 GNP[4]进行了说明。在此基础之上提出了一种 S-RTT 策略,进行种子节点的选取和连接,并将此策略引入 BT 文件共享系统,此策略能够有效地优化 BT 系统的网络流量,解决网络拥塞及资源占用问题。

1 BitTorrent 系统分析

在 P2P 网络[1]中,每个网络节点 Peer 扮演了客户机和服务器的双重角色,既可以进行下载,也可以进行上传。共享信息分散在联网的客户机上,客户机之间可以不经过服务器,二者直接或间接相连。在混合式的 P2P 网络中,服务器为每一个联网的客户机保存一份注册清单,为请求的客户机进行搜索内容的查询处理。当某个客户机正常登录后,服务器检查该客户的清单信息。一方面服务器为客户机返回一个连接表,该表列出了包含该客户机所需资源信息端的名称和 IP 地址,客户机可以选择从哪些端进行下载。另一方面,服务器把该客户机提供的共享信息为其他客户机共享。

1.1 BitTorrent 系统简介

BT 协议是用于文件共享的混合 P2P 协议。BT 系统由 tracker 和 client 组成,拥有完整文件的 client 称为 seed,正在下载文件的 client 称为 downloader。tracker 作为服务器保存共享文件的相关信息和每个文件的共享用户信息。共享文件的信息保存在扩展名为 torrent 的文件中,包括 tracker 地址、文件块的大小和分块 hash。client 得到 torrent 文件后,与 tracker 通信,获得共享所需文件的 peer 列表,与这些 peer 建立连接,从下载速率最快的 N 个 peer 处下载数据。client 完成一文件分块的下载后,就可以开始上传,上传连接数为 N'。挑选上传 Peer 的原则是,下载速率排在前 N' 位的 Peer。完成下载后,client 由 downloader 变成 seed,只上传,不下载[3]。此时,挑选

收稿日期:2007 - 09 - 06;**修回日期**:2007 - 11 - 29。 **基金项目**:国家自然科学基金资助项目(60703096)。

作者简介:孙建华(1977 -),女,河南焦作人,博士,主要研究方向:网络安全、对等网络; 王战国(1978 -),男,陕西咸阳人,硕士研究生,主要研究方向:对等网络、网络安全; 陈浩(1977 -),男,湖南湘阴人,博士,主要研究方向:对等网络、流媒体; 石林(1980 -),男,湖南湘阴人,硕士研究生,主要研究方向:对等网络、图形学。

上传 Peer 的原则与前面相同。BT 下载结构如图 1 所示。

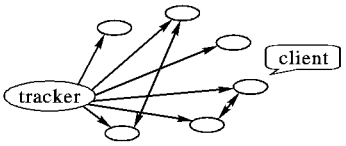

图 1 BT 下载结构

1.2 BitTorrent 系统存在的问题

BT 网络作为重叠网络，其通信路径选择有很高的灵活性。BT 系统中选择下载速率最快的 N 个 seed 节点进行连接，形成一个暂态的重叠网络。这种片面追求速率的策略造成了严重的流量瓶颈。

从覆盖网络的逻辑拓扑来看，片面追求速率的策略形成了大度数、高边介数、低簇集度的复杂网络[5]，追求高速率的选择机制满足了无尺度网络优先连接的机制，一些带宽丰富的节点成为优先连接的对象，使网络结构中存在少数度数极大的节点。其次，对于两个大度节点间的高速链路，边介数必然较大，成为瓶颈链路。第三，大度节点的邻居之间往往缺乏联系，即簇集度低。这种大度数、高边介数、低簇集度的网络更偏向树型结构而非网状结构，特点是流量集中在树干部分。

从实际的网络拓扑结构来看，这些大度节点往往是主干网上的核心节点，作用大、服务种类多、影响范围广，其链路的拥塞程度直接关系到网络的整体质量。另外由于实际网络拓扑的复杂性，一些看似不同的下载节点实际上通过相同的主干路径，加重了网络负担。再者，当小范围网络的多个节点同时向外连接时，会对该网的出口带宽形成巨大压力。

因此，应改变节点选择策略，综合考虑速率较快和距离较近两个标准，降低度数，增加簇集度，优先考虑小范围网络内的连接，尽量形成“本地”流，以减少对主干网的带宽占用，均衡 BT 流与其他类型流。

2 网络定位技术

准确测量网络中节点间距离，判断节点间簇集程度，与网络定位问题有关。由于节点之间路径通常很多，路径状况各不相同，无法直接用跳数来衡量节点间距离。这种情况下，可以用往返时间、传播延迟等作为“距离”的定义。

2.1 网络定位技术 GNP

GNP(Global Network Positioning)是一个新型网络定位系统，主要功能就是在给定网络中找出距离本机最近的 peer 节点。它代码简练，效率很高，很适合嵌入到其他应用系统中去。该系统采用了往返时间作为距离定义，往返时间是指传送一个简单的数据报到另一个主机并收到应答的时间。对于一条链路，往返时间随网络状况有变化，但在相对较长时间段内往返时间相对稳定，可以反映出某段时间内目标节点到测量节点之间的“距离”。网络正常时，较大往返时间意味着较长的物理距离。以下用距离指代往返时间。

GNP 中 Peer 节点分为两类：landmark 节点和 host 节点。系统根据一定策略选择一个 O 维空间，并在给定网络中选出 K 个节点作为 landmark，其余作为 host。K 个 landmark 节点之间通过发送 ICMP 报文确定彼此之间的距离，并确定自己在该空间 O 中的坐标。该坐标系统一旦建立起来，一段时间内将是比较稳定的。每个 host 节点依据系统提供的 landmark 节点距离信息，确定自己在 O 空间中的坐标，并计算出离自己最近的 landmark 节点。这样，某个 host 节点就很容易在给定网络中找出离自己最近的 host 节点。该系统计算出来的距离数值精确性较高，空间维数 O 和 landmark 节点数目 K 及位置对该值有一定的影响。其中，landmark 节点选取遵循最大分离、K 簇位数、最小簇内距离等标准[4]。下面以图 2 来说明，设 O 值为 3，K 值为 3 。

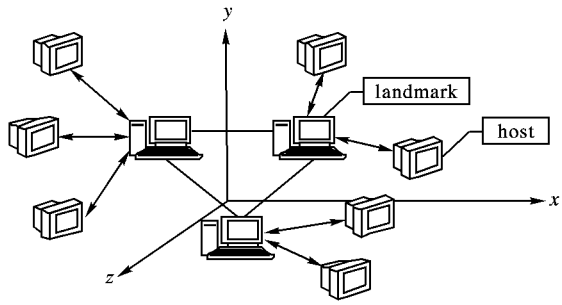

图 2 GNP 结构视图

2.2 距离误差公式

2.2.1 landmark 节点间距离误差

根据系统标准，要求选出的 landmark 节点间计算距离 $\hat{d}^O_{L_i,L_j}$ 和测量距离 d_{L_i,L_j} 的总误差最小，如下式：

$$f_{obj1}(c^O_{L_1},\cdots,c^O_{L_K}) = \sum_{L_i,L_j\in\{L_1,\cdots,L_K\}\,|\,i>j} \varepsilon(d_{L_i,L_j},\hat{d}^O_{L_i,L_j}) \tag{1}$$

其中，$c^O_{L_1}$ 为 landmark 节点 L_i 在空间 O 中的坐标，$\varepsilon(\cdot)$ 是一个误差测量函数。

$$\varepsilon(d_{L_i,L_j},\hat{d}^O_{L_i,L_j}) = \left(\frac{d_{L_i,L_j}-\hat{d}^O_{L_i,L_j}}{d_{L_i,L_j}}\right)^2 \tag{2}$$

2.2.2 host 节点与 landmark 节点间距离误差

host 节点测量自己到 K 个 landmark 节点间的距离，要求测出的这 K 个距离值与计算值之间的总误差也要达到最小，其中，c^O_H 为 host 节点 H 在空间 O 中的坐标。

$$f_{obj2}(c^O_H) = \sum_{L_i\in(L_1,\cdots,L_K)} \varepsilon(d_{L_iH},\hat{d}^O_{L_iH}) \tag{3}$$

2.3 GNP 定位网络的建立过程

1) GNP 定位系统根据特定策略从给定网络中挑选 K 个 Peer 作为 landmark 节点，并给出 O 维欧氏空间。

2) landmark 节点相互发送 ICMP 报文确定彼此间距离，根据系统策略在 O 空间中建立自己的坐标。

3) host 节点加入时，K 个 landmark 节点将它们的距离信息传递给该 host 节点，host 节点根据此信息计算自己在 O 空间中的坐标，并计算出离自己最近的 landmark 节点。

4) 依据该坐标系统提供的距离信息，某个 host 节点计算出距离自己最近的 host 节点。

这个系统中，landmark 节点是被动的，不主动向其他 host 发出查询，而是被动等待并响应 host 发出查询请求。系统一旦建立，landmark 节点坐标是稳定的，由于网络变化，host 位置可能会发生变化，它只是局部稳定，某段时间它搜索到的离自己最近的一些 host 节点可能与上一次的不同。

3 基于 S-RTT 策略的 BT 文件共享模型

3.1 S-RTT 策略

利用网络定位技术获得的信息，可选取往返时间较短的节点。在 BT 系统中依据下载速率和距离两个条件选取节点，可改变片面追求速率的简单节点选择策略。采用综合考虑下载速率和往返时间(Round Trip Time，RTT)的双重策略，

称之为 S-RTT 策略,如图 3 所示。

图 3　S-RTT 决策图

在 S-RTT 策略中,按照下载速率和往返时间的不同,待选择的节点可分为四类情况:

1)优选

选择下载速率 S 较高、往返时间 RTT 较短的节点进行连接,可以最大限度地缩短下载时间,避免形成"远程"数据流,此类节点应优先选择。

2)放弃

连接下载速率 S 较低、往返时间 RTT 较长的节点,形成的数据流往往是"远程"流,受网络环境影响较大,会使任务的下载时间大大增加,此类节点应予以放弃。

3)候补

对于下载速率 S 较低、往返时间 RTT 较短的节点,应视情况而定。一些情况下,在没有更好的选择时,可采用这些节点进行连接,称这些节点的状态为候补。

4)簇集

对于下载速率 S 很高,但往返时间 RTT 很长的节点,引入簇集机制。

簇集,即根据 GNP 返回的距离信息,将有共同资源需求的较小范围内的节点集中起来(一般在一个小范围网络内),选举一个速率最快的作为"代表"与远方资源节点建立连接,取得信息,再由"代表"在簇集内分发。

当小范围网络中出现第一个下载需求 Q_1 时,该节点就是自身的"代表",建立下载连接。若 Q_1 未下载完,在该小范围网络中出现第二个相同的下载需求 Q_2 时,两个节点之间进行选举,选出下载速率快的一个进行连接,继续下载,同时断掉原来的那个连接,后面的过程以此类推。在下载过程中,连向远端的连接始终保持一个。下载的同时,簇集内的节点之间相互传递自己所需数据。

在整个下载过程中,下载数据流在簇集内可能需要多次切换,设切换次数为 m,簇集内节点数为 n。设每次切换需要时间为 t,则总切换时间为 $T = mt$,平均每个节点的切换时间为 $\overline{T} = mt/n$。

利用 S-RTT 策略,可以将被选节点按照下载速率和往返时间分为上述四类,根据需要灵活选取下载节点和连接方式,这将显著改善网络性能。

3.2　基于 S-RTT 策略的 BitTorrent 文件共享模型

把 GNP 和 S-RTT 策略作为功能模块嵌入到 BT 系统中后,新的 BT 系统工作模型如图 4 所示。

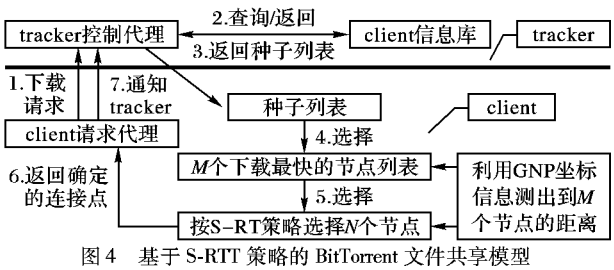

图 4　基于 S-RTT 策略的 BitTorrent 文件共享模型

整个工作过程要经历以下几个步骤:

1) BT 系统启动,运行 GNP,建立坐标系统;

2) client 端请求代理向 tracker 端控制代理发出下载请求;

3) 控制代理根据请求,到 tracker 端 client 信息库找到合适的节点,将它们的信息返回控制代理;

4) 控制代理将种子列表信息返回给 client 端;

5) client 选出下载速度最快的 M 个节点;

6) client 端定位模块 GNP 利用系统的坐标信息,测出到 M 个节点的距离,给出距离排序;

7) 按照 S-RTT 策略选出 N 个节点;

8) 将信息返回 client 请求代理;

9) client 代理通知 tracker 控制代理,自己选择下载的节点;

10) client 开始和选择的 N 个节点连接,开始下载信息。

如果 BT 网络中某个 client 节点下载速率越低,离本机距离越长,它就越不容易被选中。对于距离较长、速率较高的 client 节点,可采取合并下载流的方式。下载速率高、距离短的 client 节点优先被选择。这样,远距离高速链路流量大幅减少,短距离网络流量有所增大,网络整体状况得以优化。

3.3　模型分析

3.3.1　downloader 流量模型[3]

BT 系统中,每个 downloader 到达系统,就开始接受服务,这种自服务系统可以认为服务窗口是无穷的,可以用 M/G/∞ 排队模型来分析 BT 系统中的 downloader,downloader 的 M/G/∞ 模型和 seed 的 M/G/∞ 模型有紧密的对称性。考虑到二者的对称性,这里只考虑前者。

每个数据文件的共享子系统用一个 M/G/∞ 排队模型表示,共享数据文件 f 的 downloader 按照波松过程到达,到达间隔是独立同分布的负指数分布。每个 downloader 的下载时间,即接受服务的时间分布是 G。

下载时间由文件长度 Q 和平均下载速率 $\overline{b_d}$ 决定,下载时间 $T_d^{\Delta} = Q/\overline{b_d}$。当共享某个文件的客户较少,客户下载速率达不到客户最大下载速率时,下载时间与共享该文件的客户数有关;当共享某个文件的客户较多,客户下载速率达到客户最大下载速率时,下载时间与共享该文件的客户数无关。

根据 M/G/∞ 模型,可以得到 BT 系统中 downloader 的下列参量:

1) 平均下载时间

$$\overline{T_d^{\Delta}} = Q/\overline{b_d} + \overline{T} \tag{4}$$

2) 平均 downloader 数

平均 downloader 数就是 M/G/∞ 的平均队长,即:

$$\overline{N_d} = \lambda \times \overline{T_d^{\Delta}} = \lambda \times (Q/\overline{b_d} + \overline{T}) = \lambda \times Q/\overline{b_d} + \lambda \times \overline{T} \tag{5}$$

3) 流量

downloader 占用的平均下载带宽为:

$$\overline{B_{dl}} = \overline{N_d} \times \overline{b_d} \tag{6}$$

设 η 为 downloader 的上传率,即在下载一个文件的过程中,完成下载时上传数据量与下载数据量之比。

那么,downloader 占用的平均上传带宽为:

$$\overline{B_{up}} = \eta \times \overline{B_{dl}} \tag{7}$$

3.3.2　模型分析

以图 5 所示网络拓扑为例,对基于 S-RTT 策略的 BT 文件共享模型进行分析:黑色节点 5、6、7、8、9 为路由器,其余为终端节点。图中逻辑网络可视为两部分,节点 3、4、5、7 组成第一部分,节点 0、1、2、6、8、9 组成第二部分。链路 3 - 5 - 7 -

6-0、4-5、6-8、6-9为高速链路，其余为低速链路。节点0、1、2、3都开启了同一个BT软件，4号节点未参与BT应用，但与3号节点之间有大量数据往来。

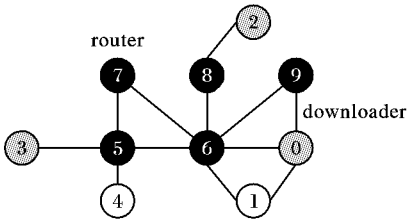

图5 一个典型的网络拓扑模型

第一种情况，设节点2和节点3都有节点0所需要的资源，按照速率第一的选择策略，节点0优先从节点3下载，路径为3-5-7-6-0。此时节点0与节点4共同竞争链路3-5，形成流量峰值。

第二种情况，节点0、1、2都从3处下载资源，按速率第一的选择策略，它们分别发起与节点3的连接，3-5-7-6成为瓶颈链路，这样对出口带宽和主干网络都造成压力。节点4的网络环境受到极大影响。

设节点0到节点2的RTT小于节点0到节点3的RTT。如果采用S-RTT策略，第一种情况下，节点0舍弃了距离较远的节点3，选择距离较近的节点2(候补)，避免了与节点4竞争带宽，尽管速率稍有降低，但没有对链路3-5-7-6形成冲击。根据式(4)、(6)、(7)，$\overline{T_d^{\Delta}}$、$\overline{B_{dl}}$、$\overline{B_{up}}$的值都有减小，由于兼顾了速率和距离两个因素，所以这些值降低的幅度不大；第二种情况下，节点0、1、2协同一致(簇集)，选出节点0作为“代表”仅建立一条从节点0到节点3的连接，再由节点0在簇集内分发，可以避免流量峰值。根据式(4)、(6)、(7)，$\overline{T_d^{\Delta}}$值增大了$\overline{T}$，和$\overline{T_d^{\Delta}}$相比，这个值很小可以忽略。链路3-5-7-6上$\overline{B_{dl}}$、$\overline{B_{up}}$的值都大幅减少。

可见，和现行的BT系统比较，新策略是速率与距离的折中，通过在高速率的前提下尽量选择离自己距离较近的节点连接，尽量形成“本地”下载流，对于长距离的下载流尽量进行合并。对于单个节点来说，下载的速率有可能小幅度下降，但整个网络的流量趋于均衡、合理，拥塞现象显著减少，平均每个数据载流将占用较少的网络资源，网络整体性能得到优化。

4 实验仿真

采用图5的网络拓扑，使用NS2仿真软件对新策略进行了仿真：

第一种情况下，流量分析结果如图6。

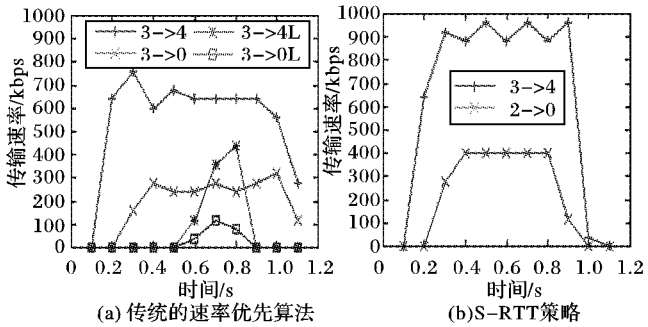

图6 第一种情况的仿真流量分析结果

图6(a)中，3->4、3->0分别为节点3发给节点4、0的有效数据流，3->4L、3->0L分别为节点3发给节点4、0的被丢弃的数据流。节点3到4的数据流刚开始时独占带宽，但0.2 s时，节点3到节点0的数据流抢占其带宽，两者的竞争使双方都出现丢包情况，链路吞吐量急剧下降。

图6(b)为采取S-RTT策略后的结果，节点0选择从节点2下载，避免了与节点4竞争带宽。节点3发给节点4的数据流保持较高的速率，节点2发给节点0的数据流也保持了较为理想的速率。两个数据流都没有出现丢包，都保持了较高的吞吐率。

第二种情况下，如果采用传统方法，节点0、1、2分别和节点3建立下载连接，这将使节点0、1、2、4都出现严重的丢包情况，链路3-5-7-6过度超载，和图6(a)情况类似。如果节点0、1、2采用簇集策略，选择节点0作为“代表”进行远程下载连接，再由“代表”0为节点1、2分发，将不会出现丢包现象。这里为了不让数据流3->4出现丢包，设定数据流3->0的下载速率略低于图6(a)中数据流3->0的下载速率。下载流量分析结果如图7。

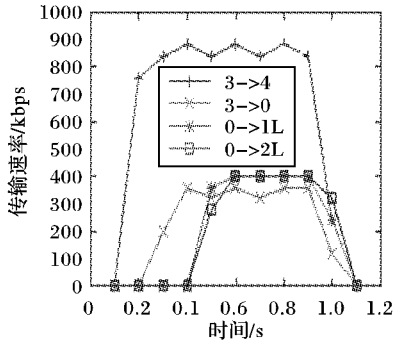

图7 第二种情况采用S-RTT簇集策的流量分析

从图7中可以看出，节点3到节点4的数据流没有出现丢包现象，维持较高的速率；节点3到“代表”节点0的下载流占用的带宽较小，没有丢包；节点0到节点1、2的分发流也正常进行。可见，和传统的连接策略相比，采用簇集策略，链路3-5-7-6中的流量约减少到原来的三分之一。

采用S-RTT策略，基本实现了BT网络中流量减小、均衡的目的。

5 结语

传统的BT节点选择策略只考虑速率因素，过于简单，没有考虑网络拓扑情况，全局流量极度不均衡。本文提出的S-RTT策略，使BT系统充分考虑距离因素，在选择下载节点时更加智能，在速率与距离之间取得了平衡，从而确保了BT流量的下降性、均衡性，缓解了其对全局网络的冲击。

参考文献：

[1] 李强，王宏，王乐春. 基于P2P的分布式网络管理模型研究[J]. 计算机工程，2006，32(13)：150-152.

[2] 李君，王攀，孙雁飞，等. P2P业务流量识别分析和控制研究[J]. 计算机工程，2006，32(11)：122-124.

[3] 周文莉，雷振明. BitTorrent文件共享系统的流量模型与文件评估方法[J]. 计算机工程，2006，32(13)：15-17.

[4] EUGENENG T S，ZHANG H. Towards global network positioning [C]// ACM SIGCOMM Internet Measurement Workshop 2001. New York：ACM，2001：25-29.

[5] NEWMAN M E J. The structure and function of complex networks [J]. SIAM Review，2003，45(2)：167-256.