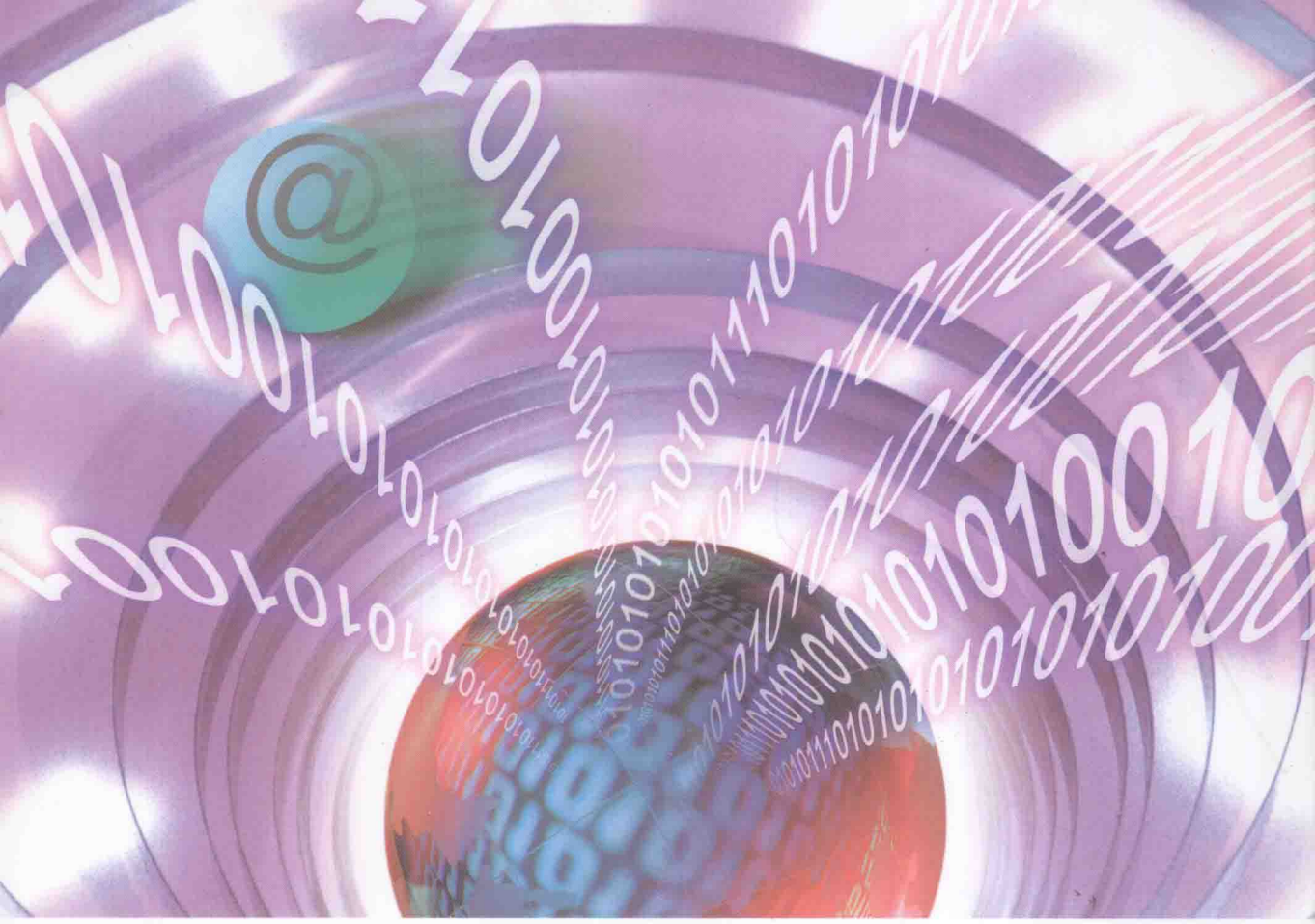

网络新技术系列丛书

HULIANWANG DASHUJU
WAJUE YU FENLEI

互联网大数据挖掘与分类

程光　周爱平　吴桦◎著

东南大学出版社
SOUTHEAST UNIVERSITY PRESS

网络新技术系列丛书

互联网大数据挖掘与分类

程 光 周爱平 吴 桦 著

东南大学出版社
·南京·

内 容 提 要

近年来，互联网的快速发展、新应用的不断出现、网络带宽的不断提高和网络数据流的急剧增加给互联网数据分析研究带来了技术挑战，互联网数据挖掘和分类对于网络计费、流量工程、网络安全等领域具有广泛应用价值. 本书主要针对互联网大数据挖掘与分类问题，系统介绍了作者在互联网数据分析处理方面的理论及实践的研究成果，主要介绍两个互联网大数据的挖掘和分类平台：基于 hadoop 集群网络被动测量数据分析平台和基于覆盖网的主动测量网络故障诊断平台，同时本书分别探讨了基于这两个平台的超点抽样检测方法、并行长持续时间流检测方法、面向 MapReduce 的大流识别方法、基于信息熵灵敏度的异常检测方法、HTTP 流量的页面关联、网络流的分类方法等六个方面的互联网大数据挖掘和分类问题。本书的内容对深入研究互联网数据测量和分析方法具有重要的借鉴意义，为网络安全和网络管理，特别是校园网的管理提供了参考。本书可供计算机科学、信息科学、网络工程及流量工程等学科的科研人员、大学教师和相关专业的研究生和本科生，以及从事计算机网络管理领域、网络工程及网络安全保护的技术人员阅读参考。

书在版编目(CIP)数据

互联网大数据挖掘与分类 / 程光，周爱平，吴桦著. —南京 ：东南大学出版社，2015.12
网络新技术系列丛书
ISBN 978-7-5641-6196-5

Ⅰ. ①互… Ⅱ. ①程… ②周… ③吴… Ⅲ. ①数据采集 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2015)第 301942 号

互联网大数据挖掘与分类

出版发行	东南大学出版社
出 版 人	江建中
社　　址	南京市四牌楼 2 号
邮　　编	210096
经　　销	全国各地新华书店
印　　刷	江苏兴化印刷有限公司
开　　本	787mm×1092mm　1/16
印　　张	15.25　彩插 8
字　　数	390 千字
书　　号	ISBN 978-7-5641-6196-5
版　　次	2015 年 12 月第 1 版
印　　次	2015 年 12 月第 1 次印刷
定　　价	40.00 元

前　言

近年来，互联网的快速发展、新应用的不断出现、网络带宽的不断提高和网络数据流的急剧增加，给互联网数据分析研究带来了技术挑战。互联网数据挖掘和分类在网络计费、流量工程、网络安全等领域具有广泛应用价值. 例如，在 CERNET 华东北地区网络的南京中心节点链路上总带宽超过 100 Gbps，中国国际出口带宽超过了 4 Tbps。同时由于有限系统资源与海量网络流量之间的矛盾存在，高速网络大数据环境下的网络流量突发性，传统网络流量数据处理算法自身的缺陷逐渐表现，多核处理器和云计算的并行分布式架构成为提高网络流量分析算法性能的有效途径，并开始得到学术界和工业界的关注。流量和网络应用的急剧增加，使流数、流持续时间、流长、信息熵、网页关联、流量分类等成为互联网大数据挖掘分类的关键问题，这些问题的有效解决可以减轻互联网大数据带来的影响，为网络运行和管理提供有力支撑。

本书针对互联网大数据挖掘和分类问题，主要介绍两个互联网大数据的挖掘和分类平台：基于 hadoop 集群网络被动测量数据分析平台和基于覆盖网的主动测量网络故障诊断分析平台，基于这两个平台本书分别探讨了基于抽样的超点检测方法、并行长持续时间流检测方法、基于 MapReduce 的大流识别方法、基于信息熵灵敏度的异常检测方法、HTTP 流量的页面关联、基于 DFI 的流量分类方法等六个方面的互联网大数据挖掘和分类问题。具体各章介绍内容如下：

第 1 章介绍在大数据背景下的网络测量相关技术及意义，介绍数据挖掘的基础理论，包括决策树算法、贝叶斯分类算法、支持向量机、聚类算法及 k-近邻算法；分析了互联网大数据处理分析所需要的分布式文件系统、MapReduce 的原理和处理流程等。

第 2 章利用云计算平台来解决海量互联网数据的分析问题，建立了一个基于 Hadoop 集群的网络数据分析平台，详细介绍了构成系统的四个模块的实现流程、实验环境的搭建、实验数据的来源等，评估了两个作业的正确性和实验分析系统的可扩展性，同时还分析了 MapReduce 作业运行过程中可能出现的性能瓶颈。

第 3 章针对基于流抽样的超点检测方法存在的检测精度低、实时性差、计算负荷重问题，提出了一种并行数据流方法，讨论可逆并行 Sketch 数据结构中参数对检

测精度的影响，从时间开销和存储开销分析了 PDS 方法的性能。实验表明 PDS 方法的链接度估计精度、超点检测精度及超点检测的处理时间均优于 CSE、JM 方法。

第 4 章研究了多核处理器硬件平台上的并行长持续时间流的检测问题，从共享数据结构和独立数据结构两方面设计了长持续时间流的并行检测算法，实验结果表明基于共享数据结构的长持续时间流的并行检测算法具有占用内存空间小的优点，基于独立数据结构的长持续时间流的并行检测算法具有时间效率高和检测精度高的优点，能够满足高速网络流量测量的应用需求。

第 5 章分析了 MapReduce 框架下 Reducer 之间产生负载不均衡的问题，提出了一种 MapReduce 框架下基于自适应抽样的大流识别方法，该方法由负责制定数据划分策略和在数据划分策略基础上进行大流识别两个作业构成。真实网络流量数据评价实验结果表明该方法具有较好的性能。

第 6 章通过理论和实验分析两种异常检测方法对信息熵的灵敏度影响，表明信息熵的灵敏度与流量在特征上分布的信息熵相关，借鉴 Kuai Xu 等人提出的网络流量描述方法，并将其应用于网络流量的异常检测。实验评估显示该方法对 Alpha、Scan、Probe 和 DDoS 等四类攻击具有较好的检测能力。

第 7 章主要研究解决网页浏览产生的 HTTP 流所属的页面关联的问题，在分析互联网广告流量以及 CDN 加速技术对 HTTP 流关联算法的影响后提出相应解决方案，设计并实现基于 DPI 和 DFI 的两种 HTTP 流关联算法，通过实验对两种不同算法正确性进行验证。

第 8 章主要解决基于统计特征机器学习的 DFI 流识别方法面临网络流数据的不平衡问题，从数据重采样、特征选择、分类算法展开流量分类方法的相关研究，实现对加密与隐私流量的分类，并在保证大类分类性能的情况下提高小类分类的准确率。

第 9 章主要解决在大规模互联网环境中采用主动探测的方法对网络故障进行推理问题，设计实现了一个基于社区的覆盖网监测与故障诊断系统，系统的客户端运行嵌有监测模块的覆盖网应用的节点并将监测结果上传给代理社区，代理社区将观测数据预处理后存入 chord，系统通过覆盖网实现对网络故障诊断和定位。

本书是笔者在互联网数据分析处理领域长期研究成果的总结，也包括了笔者培养的研究生参与的科研项目中的部分相关科研成果和论文。在本书撰写过程中，程志、王艳、梁一鑫、潘吴斌、王玉祥、马永、蔡雷、蔡振盛、戴震、戴冕、葛文锦、朱亚峰、陈新等研究生给予了支持，参与了本书部分章节的编写工作以及本书的整编、校验，全书由程光统稿。

本书的研究成果受国家高技术研究发展计划(863 计划)(2015AA015603)、江苏省未来网络创新研究院未来网络前瞻性研究项目(BY2013095－5－03)项目的资

助，在此表示感谢！同时感谢江苏省“六大人才高峰计划”的支持。在本书的撰写过程中，得到了东南大学计算机科学与工程学院、计算机网络和信息集成教育部重点实验室（东南大学）、东南大学出版社等单位领导和专家的大力支持，在此深表谢意！同时对笔者所引用的参考文献的作者及不慎疏漏的引文作者也一并致谢！

由于笔者水平有限，编写过程中难免存在很多不足及顾此失彼之处，敬请读者给予批评指正！

作 者

2015 年 10 月

目　录

1 绪 论

1.1 背景知识

1.1.1 研究背景

随着互联网技术的飞速发展，出现了越来越多的互联网应用，比如，音频视频、文件数据传输、网页信息浏览访问、基于 P2P 协议的业务等占用了大量的网络带宽。与此同时，网络接入设备的增多(主要是移动设备)，网络拓扑结构变得越来越复杂，骨干网路由器带宽在逐年的提升。很显然的是，我们目前已经生活在一个“大数据”的时代，数据的爆炸式增长出乎人们的想象，来自一份关于大数据时代数据量的预测分析报告指出，全球范围内以电子格式存储的数据在 2020 年前后，将达到 35ZB 的数据量，将会是 2009 年的 40 倍左右。来自第 36 次中国互联网络发展状况统计报告[1]，截至 2015 年 6 月，中国国际出口带宽为 4 717 761 Mbps，半年增长率为14.5%。图 1.1 是最近几年中国国际带宽及其增长率，与此同时，伴随着移动互联网、物联网，以及多种类型的移动智能终端的推广，网络设备，诸如路由器，交换机和用户移动设备将随着增长，网络数据流量增速也会变得更加迅猛，所以，如何对大数据时代网络的性能，安全，服务质量等指标进行统计分析，从而对网络服务质量进行监控和评估变得越来越困难。

图 1.1 中国国际带宽及其增长率图

目前的网络测量技术中，主要是使用集中处理平台对网络数据进行分析处理，数据源方面可以是由一个集中处理平台对单个测量点进行测量，或者由多个测量探针对多个测量点收集到的测量数据汇总到集中处理平台进行分析与处理。这使得当利用集中处理平台对海量测量数据进行处理时，平台的存储能力和计算能力会随着数据量的增长成为网络测量计算的瓶颈，如果对网络测量数据使用抽样技术减少数据量后再处理，又会明显降低测量结果的准确性，因此，

如何对海量互联网络数据进行测量分析是当前网络测量研究的一个重要点。

为了应对大数据时代的挑战，云计算的概念被提了出来，随着云计算技术的推广，利用云计算平台强大的数据处理能力来处理海量网络流量数据，已经成为当前网络测量的研究热点，另外，云平台的存储资源和计算资源的可扩展性也正是目前海量网络测量系统所需要的。

Apache 的 Hadoop[2] 是一个云计算开源项目，建立大数据的分布式的、可扩展的云计算框架是这个项目的目标。Hadoop 开源云计算平台实现了一个分布式文件系统 HDFS(相当于 Google 的 GFS)，一个并行数据处理算法框架 MapReduce(是 Google 的 MapReduce 算法的开源实现)，同时，提供了一个分布式数据库 HBase(相当于 Google 的 BigTable)。Hadoop 平台可以部署在低廉的硬件上，以减少成本；Hadoop 充分考虑到了数据的安全性，设计了副本保存机制；项目充分考虑到集群的可扩展性，Hadoop 集群可以通过对运算节点的增加以获得处理更大的数据量的能力；Hadoop 是一个开源免费的云计算项目，使得 Hadoop 技术得到了快速的发展。目前，Hadoop 广泛使用在雅虎，Facebook，IBM，百度，阿里巴巴等公司，成为了最著名的云计算项目之一。

我们根据网络数据量快速增长的网络测量环境的实际情况，设计了基于 Hadoop 云平台的网络流量测量系统，云平台的集群系统中的数据是通过测量探针对测量数据进行可靠的回收存储得到，然后利用云计算平台可扩展的数据处理能力来处理收集到的海量测量数据，实现了测量网络性能指标和用户行为分类两个作业，解决集中处理平台对海量数据处理困难的问题。

1.1.2 研究意义

大数据时代已经到来，云计算及其周边技术的迅猛发展充分证明了这个观点，各个行业都在部署自己的大数据战略。在网络测量领域，利用云计算平台强大的数据处理能力来处理网络测量中的海量数据，已经成为当前网络测量体系中数据统计分析的必然趋势，随着网络用户对互联网服务质量需求的增长，互联网服务提供商(ISP)凭借原有的网络测量技术已无法满足目前及以后的网络性能指标、安全信息等测量的监测需求。目前，国内三大运营商都在积极的部署自己的云计算中心，研究对海量的测量数据进行安全的存储管理和高效的并行计算，努力提高网络测量的处理效率，利用云计算平台的网络测量方法研究已经成为互联网服务提供商的一个重要研究点。传统的网络测量系统使用的单一的服务器，用集成管理的方法对网络数据进行分析处理，正如上文所述，不能保证数据存储和大量的测量数据存储的安全性，以及有限的计算能力，都成为集成测量系统的瓶颈，采样技术的使用会得到不准确的测量结果。为了对海量数据的网络测量进行高效的处理，我们设计的测量系统基于 Hadoop 云计算的网络测量系统，整个测量系统具有很强的灵活性和可拓展性，为网络测量大数据时代到来寻求到了一种解决方法。

1.2 网络测量技术

网络测量是按照一定的方法和技术，使用一系列软件程序或者硬件工具来测量网络的性能指标和网络行为的统称[3]。通常情况下可以将网络测量分成三个部分，分别是：测量对象、测量环境和测量方法。在这三个部分中，测量方法的选择比较重要，一般要满足稳健性、可重复性、准确性三个方面的要求；测量对象一般是指 Pcap 格式数据包数据、Netflow 格式流数据或者网络日志数据；测量环境一般是指测试运行其上的软件和硬件环境的描述，这里按平台可以是集中测量平台的或者分布测量平台。

1.2.1 网络测量常见对象

网络测量[4]的常见网络客体是网络数据报文，常见的格式是 Pcap 格式，通过分析统计报文的各种数据指标，可以获知网络的传输特性。对报文不同协议层的分析统计可以用于测量不同的性能测度指标，来反映网络表现[5]，因此在进行基于报文的网络测量时，往往是对报文根据协议进行分类研究。

网络测量的另外一个常见的客体是流，流的概念，指的是一系列的数据，以满足特定的流程规范和超时限制数据包组合，一般称为数据流。其中流规范的定义是指流中的报文必须具有相同属性组，一般使用的是五元组，分别是：源 IP 地址、宿 IP 地址、协议号、源端口号、宿端口号。当超过一定时间不活动时，就将一个流按照超时限制进行结束处理，从而使组流系统的内存和计算资源可以更充分的去处理新的流记录数据。在目前的实现情况中，实际的软件组流定义和数域将根据不同路由器厂家有所不同，甚至同一路由厂家的不同的协议版本也会有所不同，所以现在有各种不同的标准来实现流。在这些不同的流版本中，思科的 NetFlow 标准[6]由于在主流路由器中用的比较多，获得了比较广泛的应用。目前思科的 NetFlow 流记录格式广泛应用在 IP 流量统计、分析和计费等领域，并且成为了实事上的标准。目前，NetFlow 的协议版本已经发展到 Netflow V9。

当使用五元组进行组流时，NetFlow 的流结束机制一般包含表 1.1 所示的四条规则：

表 1.1 Netflow 流结束规则

规则编号	规则定义
编号 1	15 s 超时流断开规则
编号 2	30 min 内流未结束进行流断开规则
编号 2	TCP 流出现 FIN 或 RST 报文流断开规则
编号 4	流记录缓存空间已满规则

另外可以通过一些网络日志文件进行辅助的测量，网络日志文件也可以算是一种测量客体。

在本书实验中，主要使用两种研究客体进行研究工作，一个是 Pcap 格式网络报文数据包文件，另一个是根据五元组和结合表 1.1 设定的规则编写组流程序获得的 NetFlow 格式的流记录文件。

1.2.2 网络测量分析的问题

对于网络测量的研究领域，一般情况下包含三个部分：①准确地对一些网络性能参数数据的计算，这些性能参数主要包括往返时延、拥塞程度、站点的可达性、网络带宽、带宽的利用率、网络服务质量、数据包丢包率、服务器和客户端之间的响应时间、最大网络流量等；②通过计量模型的建立，一个合理有效的网络描述模型的建立，通过该模型的有效利用，可实现对网络行为分析与预测；③网络控制，使用从网络数据的性能参数的测量结果，并通过建立模型结果得到的结果反馈，实现合理配置网络资源和使用，网络的拓扑结构的调整，大型网络结构的动态描述，来监视网络，防止异常事件，防止大范围网络攻击，甚至为以后网络协议的设计提供研究参数，更或者是提供基础辅助依据参数实现对网络行为学的研究或者网络流量工程研究等。

1.2.3 单点测量常用方法

考虑到目前常用的测量方法基本上都是基于单机的测量方法，本节下面内容不做特指都是单机网络测量可以使用的方法。依照测量方法对网络运行的程度的影响程度，可以将网络测量可以分为两种测量方法：主动方式的网络测量技术和被动方式的网络测量技术。

主动方式网络测量技术和被动方式网络测量技术的区别在于，前者是主动地向网络中发出测试用的数据报文，网络的性能等信息是根据数据报文在网络上的传输情况来判断的，后者是被动地接收网络上的数据报文(一般使用镜像截获)，不会对网络运行造成任何影响。对不同的节点间的网络性能测量描述一般使用主动测量方式，而对单个节点的性能测量描述一般使用被动测量提方式。

主动方式的测量方法是通过注入测试流数据到网络(一般有接收的一方)，然后进行网络状态参数的数据的响应的测量，在此过程中，会对所测网络引入一部分附加的流量，网络的实际结果会因流量的引入产生一定的误差；根据引入的流量的多少，将进行适当的测量结果修订，然后反映网络的运行实际情况。如 Traceroute 程序，是一种主动方式的测量工具，用于反映路由信息，程序的方法是首先向目的地址发送探测性的数据报文，然后在此过程中会通过记录其返回数据报文信息来测量经过的路由器。

被动方式的测量方法主要是通过对网络流量的监视来完成网络测量任务，其测量机制是截获网络中正常的网络报文，达到尽可能地避免影响网络的正常运行的效果，所以对网络性能的影响很小。一般通过光分器在光纤上采集数据报文数据。对网络而言，没有影响网络的正常运行。

如果根据测量结果需不需要实时展示来对网络测量来分类的话，可以将测量可分为在线实时网络测量和离线网络测量两种类型。

在线实时网络测量方式，首先要求数据收集上，要做到可以实时收集网络流量数据，然后程序要能对收集的数据进行即刻分析处理，然后对结果进行输出展示，低速网络环境中，利用性能较好的硬件机器和优秀高效的程序可以做到，面试高速网络环境，一般只能对数据进行抽样处理，才能保证实时的处理，或者使用纯硬件工具。

离线测量方法是指事先采集网络数据报文保存起来，然后离线分析以存储的网络数据。使用离线方式相对而言比较简单，网络流量数据的收集存储工作和事后的测量分析工作是分开进行的，优势是结果分析比较灵活，但是当收集到海量数据时，也会出现机器存储瓶颈或者处理时间过长等问题。通常网络中的流量信息是通过开源工具软件 TcpDump 进行记录存储，然后编写相应的网络测量代码利用开源软件 Libpcap 提供的接口读取已经存储的数据进行相应的分析工作。

1.2.4 单点网络测量常用工具

Ping 工具是最基本的和常用的主动方式的网络测量工具，适应环境是在两个主机间，用于可达性、网络时延、丢包等测度的测量，是一种点对点方式的测量方法。另外，上文提到的 Traceroute 也是一种比较常用的路由信息获取工具。

对于其他主动的网络性能的计算测量，用得比较多的有 Pathload、Iperf、Netperf 等工具。Pathload[7]应用是一种主动方式的测量工具，程序用于确定在网络中两点节点之间的可用带宽；Iperf[8]也是一个主动方式测量工具，和 Pathload 一样是一种基于 Client/Server 方式的主动测量工具，能够提供网络吞吐率信息，以及振动、数据包丢包率、最大传输单元大小等统计信息，

利用这些信息，可以用于分析网络通信性能或者定位网络瓶颈；Netperf 是类似 Iperf 的网络性能测量工具。

被动方式的数据包网络性能监控工具，主要是通过 Tcpdump 和 Tcptrace[9] 这两个工具共同工作使用，作为离线数据的流截取和分析工具。Tcpdump 是由 Lawrence Brkely 国家实验室开发的，其目的是让开发者能更轻松地观察网络流量，它将主机的网络适配器接口设置成混杂模式，接收所有来自网络的数据包，并将它们以不同的格式显示在终端上或者保存在文件中供离线处理。Tcptrace 程序是由美国 Ohio 大学的 Shawn Ostermann 开发的工具，主要用于分析有 Tcpdump 应用程序捕获的 TCP 会话。它的作用主要是分析数据包回程时间、网络吞吐量、窗口大小、序列号信息以及网络会话的其他 TCP 特性等，与此同时，第三方插件 xplot 程序能将结果信息以图形的方式显示，便于观察和分析。

OpenDPI 是一个开源的深度报文检测开源软件，目前的版本是 1.3.0 版，该版本实现了对 118 种协议的分类支持，OpenDPI 允许第三方开发者编写协议解析代码，也可以根据需要修改其源代码，并部署在服务器上。另外，因为 OpenDPI 已经不在被维护和更新支持，目前，新的类似的开源软件 nDPI[10] 保持了高度受欢迎的 OpenDPI 的性质，它的扩展优化了原有的 OpenDPI 库，与此同时，增加了对新的协议的深度报文检测功能，为了支持不同平台的用户，它支持 Windows 和 Linux 各种版本，而且，可以改动 nDPI 的源码以适合自己的需要。

对数据流而言，一般指的是 Netflow 格式的数据，常用的软件工具有 Flowscan[11] 和 Flow-tools[12] 等。Flowscan 是一个遵照 GUN 开源协议的 IP 流量分析和报表工具，由一套 Perl 脚本和模块组成，FlowScan 包括了 Netflow 数据采集引擎 Cflowd，高性能数据库 RRD 以及一个可视化工具 RRDtool，通过组件之间的相互合作处理，Flowscan 能连续实时的监测由边界路由器发送的 Netflow 数据，根据预先定义的流量模式对流量进行分类，统计每类流量在一定时间粒度内的字节数、流数、报文数等信息，并且可以对相应的时间序列图进行绘制。Flowtools 是由美国俄亥俄大学开发的 Linux 系统环境下的流量收集和分析工具，它是一组实用工具的集合，包括 Flow-capturn、Flow-cat、Flow-print、Flow-stat 等工具，专门用于对 Netflow 数据进行捕获记录，合并文件，过滤输出，统计分析等。

1.2.5 分布式网络测量技术

对于并行网络测量分析，Cristian Morariu 等人提出了 DIPStorage[13]（Distributed Architecture for Storage of IP Flow Records），一个基于 P2P 的云平台的概念，每一个节点称为存储罐，以并行的方法处理网络数据流，每一个存储罐根据提前设定的规则存储相应的流数据，以后，无论是计算还是查询都可以根据规则在平台中找到相应数据；Chen 等人开发了一种使用 Hadoop 的 Snort 日志数据分析方法应用于大规模网络安全应用[14]；RIPE[15] 宣布了开发了一个用于 Hadoop 云平台的 Pcap 接口库用于对 Pcap 直接并行读处理；JongSuk R. Lee 等人为了能应对快速增长的海量电信数据[16]，提出了利用 Hadoop 云计算平台来检测电信网络异常方法，在论文中的并行分析阶段，使用的是一种随机自相似性的概念（Stochastic Self-similarity）来处理数据；在网络安全领域，卢森堡大学的几位研究人员提出了使用 Hadoop 的 Maprduce 并行算法去检测僵尸网络[17]，僵尸网络是目前网络安全的主要威胁之一，目前从集中服务主机的僵尸网络情况向分布式的、高度可扩展性、分散的 P2P 网络架构形式发展的趋势，检测和取证分析必须从网络的核心路由器取数据分析，然而核心路由的海量数据是一般方法很难处理的，因此，提出了一种使用 PageRank 算法并且基于可扩展的 Hadoop 云计算平台的方式，用于分析和检测 P2P 网络架构形式的僵尸网络主机。