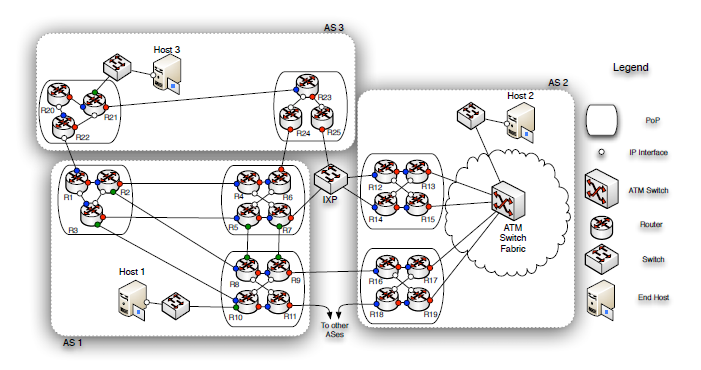
# Internet拓扑分类：

网络的物理拓扑，即网络中实体的互连关系和地理位置关系，在不同的粒度层级上有不同的意义，因而，可以将拓扑发现的技术按照粒度分成四个类别:接口级别，路由器级别，入网点(PoP)级别，自治域级别。从拓扑测量收集数据的方式来看，拓扑测量可以分为主动测量和被动测量。根据拓扑测量收集到的数据的性质，拓扑测量可以分为对控制平面的数据测量和对数据平面的数据测量。

ip级别的拓扑，节点是ip接口，边是ip链接。

路由器级别的拓扑，节点是路由器，边是路由器的邻接关系。

二者的关系：多对一，多0个ip对应一个路由器，多个ip链接对应一个路由器链接。



类traceroute技术得到接口级拓扑(完整性，准确性)，使用各种接口合并启发法聚合接口得到路由器级拓扑(对齐性)。

# 接口级拓扑测量，

traceroute是接口级拓扑测量中最常用的测量工具，也是Internet接口级别拓扑测量的仅有的原语(primitive)，其他的工具和技术都是traceroute-like的。

最早在1988年，van jacobson实现了ICMP traceroute。在之后的大部分UNIX-like系统发行版本中，traceroute都被实现成了UDP版本。

2001年，Michael Toren用TCP SYN探测包实现了traceroute，一定程度上克服了ICMP，UDP被防火墙过滤的缺陷，然而在网络边缘的主机仍有可能过滤未建立连接的TCP。

针对负载均衡造成的traceroute假路径问题，2006年，法国UPMC的Brice Augustin等发布了Paris traceroute工具，对traceroute进行了改进，通过操纵ICMP的首部，控制探测包在负载均衡存在的情况下只走一条路径。

针对大规模traceroute中观察点的选择问题，2009年，法国Brice Augustin使用了Targeted Probing(定向探测)技术，根据领域知识，精心选择TR的源和目的列表，并辅以LSSR选项，使得TR对IXP的peering的发现率大大提高。[35] 2011年，美国Boston University的B.Eriksson提出了将拓扑测量转化为统计学上的unseen species 问题，然后使用矩阵填充技术来推断未被大规模traceroute发现的设备和链接。[33]

针对大规模traceroute的负载问题，2006年，法国UPMC的Benoit Donnet在文章中提出了far probe技术，在监控点内部使用，并且提出top set技术，在监控点之间使用，降低了他raceroute的冗余和负载。[37]

traceroute大规模拓扑探测最早是从几个观测点出发进行的探测，例如1998年起，法国UoS Pansiot等人进行的研究。[39]

专用大规模拓扑探测平台有CAIDA的Skitter利用web服务器，以及Skitter的后继Arc。[40][41]

此外还有一般意义的分布式测量平台，iPlane[13]，和UW的RocketFuel[14]，都是基于2003年发起的PlanetLab项目的。

除此之外一些公共的Traceroute server，作为拓扑探测的窥镜(looking glass)，由于其测量点的分布覆盖范围广，能够保证较好的完全性，但是由于公共服务器对查询次数的限制，这种方法只适用于小尺度的拓扑测量。

此外还有crowd sourcing测量，不再适用专用的测量平台，而是利用Internet大量的用户基础，通过在浏览器安装插件，让终端用户执行测量工作，如Scriptroute[43]，Dimes[12]，Bitprobe[44]等。这种方法，需要额外的设计权衡，和安全性考虑。

针对traceroute正反向路径的不对称性，2010年，USC的E. Katz-Bassett实现了reverse-traceroute，利用了prespecified timestamp(TS)和 record route options(RR)来识别沿着反向路径的每一跳。[50]

子网发现作为traceroute的扩展，对ip路径上的子网拓扑进行了更近一步的探测。traceroute返回路径中转发路由接口的IP地址，2010年utaxes的M. Tozal 的tracenet可以访问到路由间的子网，并且标注IP地址之间的关系和通过测试获得的子网掩码。类似的工具还有同样是2011年M. Tozal发表的XNET工具。

traceroute本身的局限性：

首先，traceroute无法控制路由器对探测包的回复，对于不回复的路由器，无法得到ip地址，rtt等信息。

此外，由于一些ip地址的分配机制，icmp返回的ip地址可能不是合法的。

还有，由于往返路由的不对称性和拥塞的存在，icmp返回的rtt可能无法正确反映真实的传输延迟。

除了以上网络层的问题，数据链路层对于traceroute并不是完全透明的，有些数据链路层网络会故意向网络层隐藏自己，因而无法被traceroute发现并测量。

traceroute在大规模网络测量时的实际问题：首先观察点的选取问题，有的观察点很难发现一些路径，相关研究提出了统计和人工两种手段，提高路径发现率。其次，大规模traceroute测量平台要有合理的协作机制，如果不对冗余测量进行控制，和观察点拓扑距离较小的路由器将会有较大负载。针对该问题，通常在监控点内采用“远探针”手段，监控点间采用“协作监测”手段，去除冗余测量，做到高发现率的同时，降低负载。

traceroute改进方法：通过结合ip选项，控制探针包的路由走向，从源头控制traceroute路由，扩大测量范围。此外，利用同一子网段内，所有ip距观察点跳数差不超过一跳的性质，能够测量出接口层的一条路径上，所经过子网所有可达的ip地址，达到子网发现的目的，实现对接口级拓扑更详细的测量。

对于区域网络接口级拓扑结构的研究，大多利用了CAIDA的Skitter路由测量项目，以Skitter的接口级拓扑数据为基础，结合ICANN提供的地址分配信息，DNS，whois提供的域名信息，推断出区域拓扑结构。局限性是，需要限定问题规模，否则会影响性能，其次不能保证拓扑发现的完全性。

[8] R. Bush, O. Maennel, M. Roughan, and S. Uhlig, Internet optometry: assessing the broken glasses in Internet reachability, in Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. ACM, 2009, pp. 242-253.

Internet提供的最基础的服务是可达性，但是可达性的测量并不完善，所以可见性并不覆盖完整的可达性。

Randy Bush在2009年做的实验，在网络上广播了一个/25前缀(因为许多BGP路由器会过滤掉长度大于/24的前缀，所以/25并不会广播很远)，用BGP monitor的数据推断该前缀的理论可达性，用ping检验/25的真实可达性，发现两个测量平面上的数据不一致，真实的可达性要强于预期。不一致的原因，首先，可能因为BGP monitor无法发现一些被告知/25前缀的BGP路由器，其次，AS的真实可达性确实强于BGP路由信息提供的可达性，75%的stub BGP路由器使用默认路由，导致一些没有声明的前缀也可达。

Randy Bush的第二个实验是，使用AS路径下毒方法，来判断AS是否使用了默认路由。其原理是向上游AS发布下毒的路径(路径上带有被下毒AS本身的ASN，使前缀对其不可见)。然后用ping去探测实际可达性，如果仍可达，则判断AS使用了默认路由。实验结果证明，64%的被下毒AS都可达。

RandyBush的第三个实验是隐藏上游发现，对一个特定AS，对其上游所有AS下毒(并排除了默认路由的影响)，如果AS仍可达，则判断存在隐藏上游。这个实验证明了拓扑测量偏向性，AS间的P2P链接很难被发现，而C2P很容易被发现(只有很小比例的隐藏上游)。

[35] B. Augustin, B. Krishnamurthy, and W. Willinger, “IXPs: Mapped?” in Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. ACM, 2009, pp. 336–349.

IXP概念：由ISP外的第三方管理的链路层设施(交换机)。地理位置相近的ISP连接到IXP上，进行直接互联，而非通过上游互联网互联。IXP间的ISP通常采用peering交换策略。这种peering关系矩阵，现有的三种测量手段(TR，BGP，IRR)难以发现完整。

文章使用了Targeted Probing(定向探测)技术，也就是根据领域知识，精心选择TR的源和目的列表，并辅以LSSR选项，使得TR对peering的发现率大大提高。

# 别名解析技术和工具：

## 1.路由器级拓扑推断的挑战:

TR测量数据中，存在不可忽略的路由陷阱(routing trap)。例如，一般启发法推断时认为如果在一次TR中两条路径上有一跳不同，是因为上一跳发生负载均衡，然而也有可能正向路径唯一，反向路径发生了负载均衡。再比如，多次TR发现有一跳不同，一般会认为发生了路由改变，也有可能反向路径发生负载均衡，从不同接口进行了回复。

这两种路由陷阱都需要进行别名解析，把同属于一台路由器的ip地址(别名)映射到该路由器上。在实验中发现，在traceroute判断为负载均衡的情况中，有14%的真实情况是只有唯一的一条路径，在traceroute判断为路由改变的情况中，有32.1%的真实情况是路由并没有改变。[28]

## 2.基于指纹的方法：

### 2.1相同源地址：

法国Jean-Jacques PANSIOT，在1995年开始通过主动探测得到Internet的多播树拓扑，使用的探测手段是类traceroute，首先以自己主机作为源地址，向自己数据库(能够通信，支持源路由选项，能反向DNS查询)的5000台主机进行Traceroute，再从5000台主机采样11台，分别作为源地址再次进行Traceroute。得到的测量数据，在使用相同源地址方法进行别名解析，向可能是别名的ip地址对发送两个探测包，检测到返回包中的源地址发生变化，则认为两个地址是别名。[39]

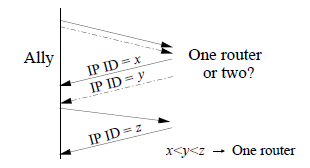
在2000年，CAIDA的Ken Keys 实现了iffinder别名解析工具，采用了相同源地址方法。

局限性，受制于路由器的应答策略。

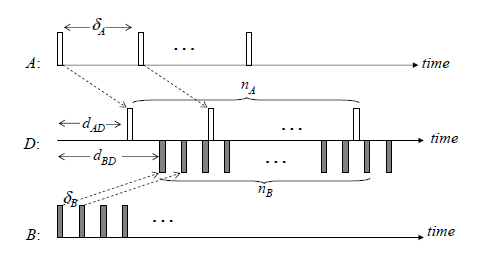
### 2.2 IPID计数器：

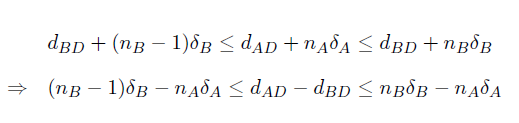
2002年华盛顿大学的Neil Spring提出了RocketFuel项目，作为ISP拓扑测量的引擎。使用了别名解析工具Ally(现在已经不再维护，而是作为Scriptroutede 一部分)。Ally所用到的原理就是IPID计数器。[14]

原理：向两个候选别名ip分别发探测包，根据回复包的IPID差值，若小于一定阈值，在向先回复的ip发送第三个包，如果三个包的回复IPID满足升序，且IPID差很小，则认为两个ip是别名。这样做给出了第二个回复包的IPID一个取值范围而不是一个确定的阈值，保证了不同IPID递增速率的路由器都适用，并且以很大的概率确定是别名。相反的，如果两个包返回得IPID距离较大，则认为是非别名，如果没有收到回复，则ip标记为未知。[14]



测量两个源点的单向延迟差，利用IPID推断包到达的先后顺序，两个B包夹一个A包(A包IPID在B的两个包之间)，根据不等关系可以确定A，B延迟的差的上下界。[57]





为了降低时间复杂度O(n^2)，Ally依据别名应当有相近TTL的原理，对TTL进行聚类，并使用逆向DNS查询进行分类，实现对候选别名集合的剪枝。

局限性，有些IPID计数器增加幅度不同，导致false negative。时间复杂度，不剪枝会很大O(n^2)，剪枝会不准确。[54]

2008年美国马里兰大学的Adam Bander，提出了RadarGun工具，改进了Ally工具，不再像Ally一样逐对探测别名，而是逐个多次探测ip列表，找到ID增速都一致的ip认为是别名。RadarGun，把复杂度降到了O(n)，并且解决了false negative问题。然而，如果对ip的探测间隔控制不好，ID会发生wrap around(折回)，导致新的错误判断。[58]

到2016年CAIDA使用的基于IPID的别名解析工具是MIDAR (Monotonic ID based Alias Resolution)，与Ally，RadarGun相比，MIDAR准确性更高，而且能够在大规模网络测量中实现产品化。在CAIDA的Macroscopic Internet Topology Data Kit (ITDK)项目中，MIDAR和iffinder共同使用作为别名解析工具。

主动探测的方法缺点，首先需要引入探测流量，其次主动探测不能控制中间路由器的行为，路由器可能故意不回复，ip实现可能不使用全局IPID计数器(为了隐藏自己的网络结构或者为了安全性考虑)。

### 2.3 IP选项：

IP选项别名解析的思路是，使用ip包首部中的ip选项，结合traceruote，获得路径的更多信息用来推断别名。

2008年UW的Neil Spring等提出了一种利用了RR选项(Record Route option)的别名解析工具DisCarte，使用析取逻辑规划的模型，将RR数据和TR数据对齐。2012年又将DisCarte整合到了Sidecar中。RR是ip首部的一个数组，路径上的路由器会在RR中插入自己的ip地址，协议设计最多能存储9个ip地址。RR选项可以克服TR的一些问题，发现不回复的路由器别名，隐藏路由器(不递减TTL)，匿名路由器(递减TTL但不回复icmp echo)。缺点是，最多支持9跳，RR包可能被过滤或丢弃，还有最重要的是，TR数据和RR数据对齐的复杂性，导致这种方法只适用于小范围的拓扑测量。[61]

此外，TS选项(pre-specified timestamp)也可以被用作别名解析。发包者在TS选项中指定4个ip，要求对应ip的设备给选项按照给出的ip顺序(即ip在前的写完后面的设备才能写)写入时间戳。将候选ip别名交叉写入TS如ABAB，BABA。回复中如果，将4个ip都写入了时间戳，并且时间戳数值相同(一个路由器只有单个时钟)，则认为AB互为别名。[51]

## 2.基于推断的方法：

### 2.1基于DNS-name：

相近的DNS-name认为是别名。单独使用缺陷很明显：需要ISP对路由器进行系统的命名，需要认为干预，在AS边界非常不准确。[30]

DNS-name在RocketFuel中，主要是用来路由器标识和地理位置映射，DNS标识相对于AS地址空间标识的优势是，AS地址空间对于不进行BGP广播的邻居而言，只用自己内部的地址编号，这样边界不容易被发现，此外同属于两个AS的路由器可能被编号成任意一个AS中的地址，还有DNS能区分出非路由器设备，避免AS造成的混淆。标识路由器时，利用DNS逆向查询得到的DNS名字，使用undns工具从中提取ISP标识路由器的字段。地理位置映射方面，利用ISP植入到DNS名字中的信息来推断路由器的地理位置，从而更好的识别POP和骨干，深入了解ISP的结构特征。[14]

这种标识路由器的局限性：准确度会受到DNS数据更新不及时的影响，完整度会由于ISP发布的DNS数据不完整而受到影响。[54]

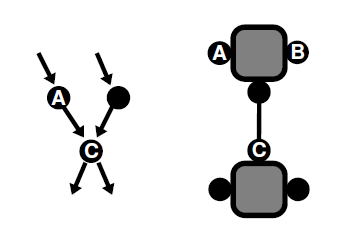
### 2.2基于图的启发法：

基于两条规则，对TR的有向图进行推断。

共同后继：适用于路由器点对点连接的情况(相对于多点接入网络和交换结构网络)，由于后继接口标识了唯一一个路由器，所以拥有共同后继的接口可以认为是别名。

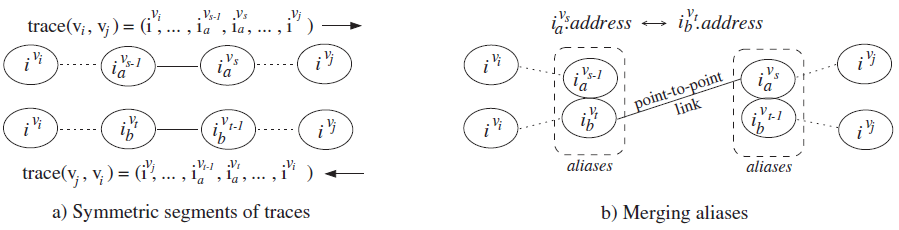
相同TR：相同TR中的所有接口都不是别名，因此一个具有n个节点的TR路径可以消除n^2个别名对可能。

基于图的方法只是用来做预处理用，减少下一步主动探测的数量。[54]



### 2.3基于子网的分析法：

2006年M.H.Gunes提出别名解析算法。思路是先用双向TR数据建立图，根据ip中子网划分的机制，找出分别在两个路径上的，且在同子网中的接口对，二者之间的连接认为是点对点连接，对于点对点连接之间的接口认为是别名。[59]



# 渐进路由器发现：

渐进路由器发现方法利用了支持MRINFO协议的路由器保存有邻居路由器列表，通过类似于OSPF的渐进路由器发现方法，逐步发现网络中所有路由器。

2004年，法国Jean-Jacques PANSIOT使用自主开发的基于mrinfo协议的mrinfo-rec工具，从单一路由器开始探测，直到08年结束。平均每日探测10万接口（不包括无法路由到的IP地址，特殊IP地址，隧道，无效接口等），发现1万左右路由器，超过850个AS。mrinfo-rec的局限性在于，适用范围只限制于支持多播的路由器，ICMP报文可能被过滤。[63]

2011年，PANSIOT针对对mrinfo-rec的问题做了改进，并发布了新的MERLIN工具。

[28] P. Marchetta, V. Persico, E. Katz-Bassett, and A. Pescap´e, “Dont trust traceroute (completely),” in ACM CoNEXT Student workshop, 2013

[39] J.-J. Pansiot and D. Grad, On routes and multicast trees in the Internet, ACM SIGCOMM Computer Communication Review, vol. 28, no. 1, pp. 41-50, 1998.

[14]N. Spring, R. Mahajan, and D. Wetherall Measuring ISP topologies with Rocketfuel, ACM SIGCOMM Computer Communication Review, vol. 32, no. 4, pp. 133-145, 2002.

[57] W. Chen, Y. Huang, B. F. Ribeiro, K. Suh, H. Zhang, E. d. S. e Silva, J. Kurose, and D. Towsley, Exploiting the IPID field to infer network path and end-system characteristics," in Passive and Active Network Measurement. Springer, 2005, pp. 108-120.

[58] A. Bender, R. Sherwood, and N. Spring, Fixing ally's growing pains with velocity modeling, in Proceedings of the 8th ACM SIGCOMM conference on Internet measurement. ACM, 2008, pp. 337-342.

[54] N. Spring, M. Dontcheva, M. Rodrig, and D. Wetherall, How to resolve IP aliases, Univ. Michigan, UW CSE Tech. Rep, pp. 04-05, 2004.

[59] M. H. Gunes and K. Sarac, Analytical IP alias resolution, in Communications, 2006. ICC'06. IEEE International Conference on, vol. 1. IEEE, 2006, pp. 459-464.

[30] R. Govindan and H. Tangmunarunkit, Heuristics for Internet Map Discovery, in INFOCOM 2000, vol. 3. IEEE, 2000, pp. 1371-1380.

[61] R. Sherwood, A. Bender, and N. Spring, Discarte: a disjunctive Internet cartographer," in ACM SIGCOMM Computer Communication Review, vol. 38, no. 4. ACM, 2008, pp. 303-314.

[51] W. de Donato, P. Marchetta, and A. Pescape, A hands-on look at active probing using the IP prespeci ed timestamp option, in Passive and Active Measurement. Springer, 2012, pp. 189-199.

[63] J.-J. Pansiot, P. M´ erindol, B. Donnet, and O. Bonaventure, “Extracting intra-domain topology from mrinfo probing,” in Passive and Active Measurement. Springer, 2010, pp. 81–90.

[5] M. C. Toren, tcptraceroute: an implementation of traceroute using TCP SYN packets

[7] B. Augustin, X. Cuvellier, B. Orgogozo, F. Viger, T. Friedman, M. Latapy, C. Magnien, and

R. Teixeira, Avoiding Traceroute Anomalies with Paris Traceroute, in Proceedings of the 6th ACM SIGCOMM conference on Internet measurement. ACM, 2006, pp. 153-158.

[35] B. Augustin, B. Krishnamurthy, and W. Willinger, “IXPs: Mapped?” in Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. ACM, 2009, pp. 336–349.

[33] B. Eriksson, P. Barford, J. Sommers, and R. Nowak, DomainImpute: Inferring unseen components in the Internet, in INFOCOM, 2011 Proceedings IEEE, pp, 171-175.

[37] B. Donnet, P. Raoult, T. Friedman, and M. Crovella, Deployment of an algorithm for large-scale topology discovery, Selected Areas in Communications, IEEE Journal on, vol. 24, no. 12, pp. 2210-2220, 2006.

[40] CAIDA, Macroscopic Topology Measurements Project and the Skitter infrastructure, http://www.caida.org/tools/measurement/skitter/.

[41] CAIDA, Macroscopic topology measurements project and the archipelago measurement

infrastructure, http://www.caida.org/projects/ark/, 2011.

[13] H. V. Madhyastha, T. Isdal, M. Piatek, C. Dixon, T. Anderson, A. Krishnamurthy, and

A. Venkataramani, iPlane: An information plane for distributed services, in Proceedings of the 7th symposium on Operating systems design and implementation. USENIX Association, 2006, pp. 367-380.

[42] Y. He, G. Siganos, M. Faloutsos, and S. Krishnamurthy, Lord of the links: a framework

for discovering missing links in the Internet topology, IEEE/ACM Transactions on Networking

(ToN), vol. 17, no. 2, pp. 391-404, 2009.

[43] N. T. Spring, D. Wetherall, and T. E. Anderson, Scriptroute: A Public Internet Measurement Facility, in USENIX Symposium on Internet Technologies and Systems, 2003.

[44] T. Isdal, M. Piatek, A. Krishnamurthy, and T. Anderson, Leveraging BitTorrent for end host measurements, in Passive and Active Network Measurement. Springer, 2007, pp. 32-41.

[12] Y. Shavitt and E. Shir, DIMES: Let the Internet measure itself, ACM SIGCOMM Computer

Communication Review, vol. 35, no. 5, pp. 71-74, 2005.

[50] E. Katz-Bassett, H. V. Madhyastha, V. K. Adhikari, C. Scott, J. Sherry, P. Van Wesep, T. E. Anderson, and A. Krishnamurthy, Reverse traceroute, in NSDI, vol. 10, 2010, pp. 219-234.

[53] M. Tozal and K. Sarac, Tracenet: an Internet topology data collector, in Proceedings of the 10th ACM SIGCOMM conference on Internet measurement. ACM, 2010, pp. 356-368.