UNIVERSIDAD PÚBLICA DE NAVARRA

TÉCNICAS DE RESOLUCIÓN DE ALIAS PARA LA OBTENCIÓN DE MAPAS DE INTERNET A NIVEL DE ROUTER

Tesis doctoral presentada porSantiago García Jiménez
dentro del Programa de Doctorado Computación distribuida y de altas prestaciones.
Dirigida por Dr. Eduardo Magaña Lizarrondo

UNIVERSIDAD PÚBLICA DE NAVARRA

TÉCNICAS DE RESOLUCIÓN DE ALIAS PARA LA OBTENCIÓN DE MAPAS DE INTERNET A NIVEL DE ROUTER

Tesis doctoral presentada porSantiago García Jiménez
dentro del Programa de Doctorado Computación distribuida y de altas prestaciones.

Dirigida por Dr. Eduardo Magaña Lizarrondo

El doctorando El director El director

Pamplona, Octubre de 2013

Índice general

Índice de figuras

ÍNDICE DE FIGURAS

Indice de tablas

Resumen

Mediante las técnicas de resolución de alias se pueden realizar estudios de topología que permiten inferir mapas de Internet a nivel de router. El descubrimiento de topologías a nivel IP se divide en dos fases, la fase de descubrimiento en la que se obtienen los datos referentes a las direcciones IP y los enlaces que existen entre ellas, y la fase de resolución de alias en la que se intenta agrupar aquellas direcciones IP que pertenecen al mismo router. Tras las dos fases se obtiene un grafo o mapa a nivel IP de la red en la que cada nodo del grafo representa un router y cada arco entre ellos representa un enlace entre dos routers.

A lo largo de este trabajo se estudian las técnicas de resolución de alias existentes y se proponen nuevas técnicas que mejoran en algún aspecto sus resultados.

La resolución de alias tiene un amplio estado del arte que se expone y se analiza en el trabajo. Se han evaluado las distintas técnicas de resolución tratando de identificar tanto sus bondades como las partes más problemáticas de cada una (Ally, Mercator, Midar, Radargun, TraceNet, PalmTree). En la mayoría de casos se ha tenido que reimplementar las técnicas debido a la falta de herramientas y medidas públicas. Este es el motivo por el que apenas existen estudios comparativos entre las técnicas de resolución de alias. A pesar de ser una temática en la que se han realizado muchas propuestas aún hay espacio para el desarollo de nuevas ideas que permitan una mejora en el área.

Se ha realizado un estudio de los comportamientos en la generación determinados campos de los paquetes de respuesta por parte de los routers. A los campos cuyos comportamientos se utilizan para la resolución de alias se les han denominado parámetros base. Para realizar las mediciones, se ha realizado el envío de distintos paquetes sonda con distintos tipos de medidas activas, variando los tipos

de paquetes así como la forma de realizar el envío. Los paquetes de tipo indirecto son los que mayores tasas de contestación proveen. En el caso específico del parámetro base IPID, a pesar de que los paquetes de tipo indirecto tienen mayor tasa de contestación, son los paquetes de tipo directo los que ofrecen mayores tasas de respuestas útiles. No obstante, para el parámetro base IPID las tasas de respuestas útiles tanto para las medidas de tipo directo como para las medidas de tipo indirecto son muy altas. A pesar del potencial que tienen, las medidas de tipo indirecto no se utilizan apenas en las técnicas de resolución.

El estudio de las diferentes técnicas de resolución muestra que las técnicas de resolución de alias encontradas en el estado del arte no superan tasas de identificación de un 25 % del total de parejas que se pueden formar con las direcciones IP de la red en estudio. La que mejores resultados ofrece es la técnica Radargun, que consigue tasas de alrededor de un 20 % de identificación. El estudio de las contestaciones a los paquetes sonda por parte de los routers muestra que las tasas bajas de identificación están ligadas en parte a la forma en la que se realizan los experimentos y no a la falta de potencial de identificación por parte de los comportamientos de los parámetros base que se utilizan a la hora de realizar la identificación.

A través del estudio de las técnicas del estado del arte se ha identificado un problema inherente a la forma de realizar la resolución de la técnica Radargun que hace decrecer de forma notable la completitud de la resolución obtenida. Mediante una regla de dimensionamiento se ha conseguido determinar el ancho de banda que se debe especificar en la herramienta para no perder completitud en las resoluciones. Por otro lado, se ha propuesto una estrategia de distribución temporal y espacial que permite que las técnicas de resolución con una agregación por parejas puedan realizar identificaciones de redes de gran tamaño.

Se ha realizado una propuesta de estrategia de reducción basada en el *IP-Offset* que consigue que con la utilización de sólo un 10 % de las parejas totales se obtengan resoluciones de un 75 % del total de alias, permitiendo reducir un 90 % el tráfico introducido en la red y permitiendo realizar las resoluciones en redes que debido a su tamaño no resultaría posible realizar de otra manera. El valor de *IP-Offset* en el que se basa la reducción consiste en la diferencia numérica de las dos direcciones IP que se desea identificar. La estrategia de reducción permite una preselección de

las parejas a las que se debe realizar las pruebas de resolución escogiendo aquellas que tienen más probabilidad de ser alias.

La técnica Ally-based es una técnica propuesta basada en el modo de funcionamiento de una de las técnicas del estado del arte. Identificando e intentando paliar los problemas que tenía la técnica original, se han conseguido mejoras basadas en el envío de tres tipos de paquetes sonda (ICMP, UDP y TCP) en lugar de sólo uno (UDP), el envío de un mayor número de paquetes realizando un envío de 20 paquetes sonda en lugar de 3 y por último un cambio en la forma de enviar los paquetes sonda, que en este caso se realizan cada 0,3 segundos para asegurar la contestación por parte de los routers a los que van dirigidos. Gracias a estos cambios se consigue aumentar la completitud desde valores alrededor del 7 % obtenidos en la técnica original, a valores alrededor de un 55 % de completitud permitiendo una mejor identificación de la red en estudio.

La técnica Pamplona-traceroute es una técnica propuesta en este trabajo con la que se alcanzan porcentajes de identificación de parejas de alrededor de un 65 %. Además esta técnica de resolución de alias permite disminuir de manera considerable el tráfico introducido en la red realizando en un mismo proceso la fase de descubrimiento y la fase de resolución. Esta técnica no utiliza medidas activas en la fase de resolución y no requiere de un incremento considerable de los paquetes sonda enviados en la fase de descubrimiento realizando la identificación mediante los paquetes de respuesta de tiempo excedido en tránsito. Esta técnica se puede combinar con la técnica Ally-based pudiendo llegar a conseguir resoluciones entre un 80 % y un 90 % del total de parejas del escenario.

Se han realizado contribuciones que obtienen una mejora notable en el proceso de resolución de alias permitiendo realizar el proceso de manera más completa e invirtiendo menor tráfico de sondeo en el proceso. Se ha realizado un estudio profundo de la herramientas disponibles hasta el momento y se ha evaluado su comportamiento en Internet. Además se han realizado propuestas que permiten el dimensionamiento de algunas técnicas existentes que permiten su ejecución sin perdida notable de completitud.

CAPÍTULO

1

Introducción

Internet está compuesta por miles de subredes interconectadas que son propiedad de distintas organizaciones. Estas organizaciones se encargan tanto del despliegue de equipos para permitir la conectividad con el resto de subredes, así como de ofrecer servicio de acceso a los usuarios finales.

Las distintas organizaciones siguen sus propios patrones de despliegue de los equipos que componen sus redes basándose en distintas motivaciones como por ejemplo motivaciones económicas, estratégicas o meramente tecnológicas. Por lo tanto, las distintas redes y conexiones que existen entre ellas, forman lo que conocemos como Internet que carece de una estructura claramente definida.

Es más, debido a la competitividad de las diferentes organizaciones así como para evitar posibles problemas de seguridad derivados del conocimiento de terceros de cómo es realmente el despliegue de dicha red, apenas hay redes de las que se tenga información pública de su estructura. Las únicas redes de las que se tiene una información completa sobre su estructura son generalmente redes académicas y de investigación nacionales como pueden ser las redes de GEANT [1], Internet2 [2] o Canet4 [3].

En los últimos años ha habido intentos de obtener información de las diferentes redes, así como de obtener información a nivel global de cómo es realmente Internet. Para ello existen diferentes formas de representación de la estructura de

Internet basadas en grafos que representan diferentes niveles de abstracción. Los grafos están compuestos por nodos y arcos que los unen y que pueden tener diferente significado. A estos grafos que representan la estructura de Internet se les denomina comúnmente mapas de Internet.

Un primer nivel de abstracción es aquel grafo en el que los nodos representan sistemas autónomos y los arcos identifican enlaces de interconexión entre sistemas autónomos [4]. Otra aproximación sería aquella en la que los nodos del grafo representan direcciones IP y los arcos que los unen representan enlaces físicos entre los routers a los que pertenecen dichas direcciones IP. Este tipo de grafos basados en direcciones IP y enlaces son los que se obtienen mediante la herramienta traceroute [5] y se utilizan en diversos proyectos de mapeo de Internet como OPTE [6] o el Internet Mapping Proyect [7].

La última aproximación es realizar un mapa de red a nivel de router, en el que cada nodo del grafo represente un router y cada arco represente los enlaces de dicho router con el resto. Los routers son los elementos funcionales básicos en las redes IP y se encargan de realizar la interconexión reenviando los paquetes por el camino adecuado para alcanzar el destino. Un router posee un número determinado de interfaces de red, conectadas a redes diferentes y por tanto con direcciones IP diferentes. La obtención de un mapa de Internet a nivel de router requiere de la realización de dos fases básicas: la primera es la llamada fase de descubrimiento y la segunda es la fase de resolución de alias.

La primera fase es la encargada de la obtención de las direcciones IP pertenecientes a la red en estudio y también de los enlaces existentes entre dichas direcciones IP (relación de adyacencia). La fase de resolución de alias es la encargada de determinar qué direcciones IP pertenecen al mismo router para agregarlas en el mismo nodo del grafo. Dos direcciones IP son *alias* si pertenecen al mismo router. En esta última fase es en la que se centra el trabajo de esta tesis.

1.1 Motivación

La principal motivación para la realización del trabajo es que la obtención de un mapa de Internet a nivel de router es un reto aun por solventar, sobre todo en la parte

que concierne a la fase de resolución de alias. Existen distintas estrategias que abordan la resolución de alias como Mercator[8] y Ally[4]. Sin embargo, sus resultados ofrecen bajos porcentajes de identificación que están lejos de una resolución completa de la red. El objetivo de la fase de resolución de alias es que para cada posible pareja que se pueda formar con las direcciones IP que se hayan obtenido en la fase de descubrimiento, se obtenga una respuesta de pertenencia o no al mismo router. Los métodos existentes consiguen una identificación de aproximadamente un 7,4 % de las posibles parejas según la red [9], por lo que hay posibilidad de mejora en las tasas de identificación.

Si nos centramos en las aplicaciones prácticas que se pueden ver beneficiadas por la obtención de un mapa de red a nivel de router, la primera que se podría citar es la de servir como entrada a simuladores. De esta forma, los estudios realizados mediante simulación, así como sus resultados, serán más acordes con la realidad. Hasta ahora las redes que servían de entrada para los simuladores para hacer estudios a nivel global son redes sintéticas basadas en algoritmos matemáticos que cumplen determinadas características como pueden ser las leyes de potencias [10]. El grado de imitación de dichas redes sintéticas comparadas con las de Internet no se ha podido medir de forma precisa debido a la dificultad de encontrar mapas de Internet a nivel de router fiables.

Las estrategias para la creación de un mapa de red pueden resultar muy útiles para la administración y gestión de redes. El uso de un mapa de red resulta una herramienta muchas veces indispensable para la detección de posibles fallos de configuración en la red, detección de cuellos de botella o para mejorar las latencias en la red. Puede existir ocasiones en las que la red a administrar sea muy grande y no se tenga información completa de cómo es físicamente. Herramientas comerciales como InterMapper [11] ayudan en la administración y gestión de las redes ofreciendo posibilidad de auto-mapeo de la red. Toda mejora en las estrategias de resolución de alias se verá reflejada en una mejora de esta inferencia automática dando como resultado un mapa mas acorde con la red real que tiene desplegada la organización bajo estudio.

Otra de las utilidades aplicables de la resolución de alias es en el campo de la geolocalización de direcciones IP [12]. Este tipo de herramientas se basan en

la geolocalización de los puntos intermedios (direcciones IP de routers) en los diferentes caminos que seguimos para llegar a la dirección IP que queremos geolocalizar. Cada geolocalización de una nueva dirección IP intermedia ofrece nuevas restricciones que permiten mejorar el resultado de geolocalización de la dirección IP en estudio. El saber que dos direcciones IP pertenecen al mismo router y que, por tanto, están situadas en el mismo punto geográfico, ofrece restricciones adicionales que permiten ser aun más preciso en dicha geolocalización.

Las aplicaciones de tipo P2P pueden verse favorecidos por estrategias que permitan un conocimiento mejor de la red de interconexión [13]. Normalmente la información con la que trabajan este tipo de programas esta relacionada con qué equipos están conectados a la red P2P, pero no con la información estructural de la red que interconecta los equipos. El modo en el que un nuevo equipo entra a formar parte de esa red P2P podría basarse en características más avanzadas con datos sobre la velocidad y cercanía física con otros equipos ya conectados.

Existen también motivaciones relacionadas con la mejora en el campo de la seguridad informática. En el trabajo de H. Burch et al. [14] se habla de un sistema para poder rastrear un ataque de denegación de servicio por inundación hasta la fuente que lo origina. Este tipo de ataques consisten en el envío de un número enorme de paquetes hacia un equipo víctima y que este último no pueda dar servicio a todas las peticiones que está recibiendo. La estrategia que se describe en el trabajo requiere tener previamente el mapa de Internet, o al menos el mapa de las redes hasta la fuente de dicho ataque. El proceso pasa por realizar peticiones hacia los distintos routers observando si asciende o no la tasa de pérdidas y de esta manera conseguir el camino que están siguiendo los paquetes hasta el punto de origen del ataque.

Como última utilidad a la obtención de un mapa de Internet, y que guarda cierta relación con la utilidad para los esquemas P2P, es el uso de éste para el posicionamiento de servidores como se puede ver en el trabajo de L. Qiu et al. [15]. Dependiendo de los objetivos que queramos darle a un servidor dado, la proximidad con los clientes pudiera ser algo esencial para ofrecer un mejor servicio a estos. La elección del proveedor de servicios se puede hacer atendiendo a un estudio de posiciones de donde están los clientes a los que deseamos dar un servicio determinado.

Con todo ello queda patente que, un mapa de Internet a nivel de router tiene un elevado interés como herramienta para la mejora de múltiples disciplinas.

1.2 Objetivos

El objetivo principal de la tesis es conseguir mejorar las tasas de identificación existentes en resolución de alias. La mejora en el porcentaje de identificación en resolución de alias supondrá la obtención de mapas de Internet más cercanos a la realidad. El objetivo no es tanto obtener una resolución de alias completa de Internet como el de ofrecer estrategias que permitan mejorar las tasas de identificación de la fase de resolución de alias teniendo en cuenta criterios de escalabilidad.

El objetivo anterior necesita de la elaboración de un completo estado del arte en los métodos de resolución de alias, pero también en las diferentes estrategias de descubrimiento de topologías. El estudio y comprensión de las estrategias existentes ayudan tanto a realizar pequeños cambios que permitan mejorar sus tasas de identificación en la fase de resolución de alias así como en la elaboración de nuevas estrategias de identificación.

El criterio de escalabilidad marca dos objetivos adicionales a cumplir. Uno de ellos consiste en proporcionar vías que permitan la unión de las dos fases del proceso de obtención de mapas de Internet a nivel de router. Las dos fases se encuentran separadas y sirve una de entrada de la siguiente. Si se consigue acoplar las dos tareas en una sola, esto permite ahorrar en volumen de tráfico enviado a la red y en tiempo de realización de las medidas.

Otro objetivo derivado del criterio de escalabilidad consiste en que las técnicas ideadas posean la propiedad de ser distribuibles. Si la carga, ya sea del proceso de sondeo o del procesado mismo de los datos para realizar la identificación de la fase de resolución de alias, es distribuible, significa que se puede disminuir el tiempo invertido en el proceso usando más equipos. Para ello, se utilizan técnicas de computación *grid* que permiten la distribución del proceso computacional.

Un objetivo adicional es el proceso de verificación y comparación de las diferentes técnicas de identificación de la fase de resolución de alias en Internet. Para la verificación inicial de las diferentes técnicas se hace uso de una maqueta completamente controlada. La maqueta es una pequeña red montada en el laboratorio en

la que se realizan las primeras pruebas y comparaciones de técnicas. Tras verificar que la técnica a estudio funciona de manera correcta en la maqueta se pasa al uso en redes publicas de Internet. Se realizan las medidas de identificación de la fase de resolución de alias sobre las redes de GEANT [1], Internet2 [2] y Canet4 [3] que forman parte de Internet, por lo que aparte de comprobar si las técnicas consiguen o no buenas tasas de identificación se puede observar como se comportan los routers y otros dispositivos (cortafuegos, balanceadores de carga, conformadores de tráfico, ...) situados en el camino a los routers que se pretende identificar.

Una vez terminado el proceso para redes con estructuras conocidas y verificando que no se cometen errores, se trasladan las medidas a Internet. Para dichas pruebas se hace uso de plataformas de medida como PlanteLab [16] y ETOMIC [17]. Este tipo de plataformas ofrecen al investigador una serie de máquinas distribuidas alrededor del mundo que le permiten realizar las medidas de red que desee. Se utilizan para realizar el descubrimiento de redes de interconexión y posteriormente la fase de resolución de alias. Con estas pruebas se obtienen tasas de identificación de los distintos métodos de resolución de alias y se realiza una comparación de su rendimiento.

1.3 Estructura del documento

El documento que sigue estará estructurado en las siguientes partes, la primera es la introducción. En ella se ha explicado la información de trasfondo sobre el que se sustenta el proyecto de tesis. El por que se ha elegido y la cuestión o problema que se pretende solucionar.

El punto siguiente titulado *Estrategias en la resolución de alias*, pretende hacer un repaso y mostrar al lector las diferentes técnicas existentes en el marco de investigación que se está tratando. Se dividirá en las dos fases ya mencionadas de descubrimiento y la fase de resolución de alias. En la primera parte se presentan las técnicas que permiten derivar la información de qué direcciones IP se encuentran en la red y, al menos en un principio, que enlaces existen entre estas direcciones IP.

La segunda parte hará referencia a las diferentes estrategias utilizadas para una vez obtenido un grupo de direcciones IP saber cuales de ellas pertenecen o no a un mismo router.

El los capítulos siguientes, se pretende mostrar las contribuciones realizadas a este campo tanto nuevas técnicas de identificación, mejoras de técnicas conocidas como medidas que se han hecho para determinar la estructura de Internet y el comportamiento de esta.

Para terminar el documento, se propondrán una serie de líneas futuras de investigación. A lo largo de este estudio se han podido ver mejoras en distintos puntos de las fases de descubrimiento de topologías, pero aun queda trabajo por realizar. Este punto pretende ser un punto de partida para investigaciones futuras que podrían dar pie a nuevos resultados o nuevas estrategias para obtener topologías más precisas y usando menos carga de red y computación.

Por último se anexarán los artículos publicados a lo largo del proceso de investigación con intención de que su consulta pueda ser más sencilla y rápida.

CAPÍTULO

2

Estrategias en la resolución de alias

2.1 Introducción

La generación de un mapa de red a nivel de router es un problema cuya resolución se puede descomponer en dos fases básicas llamadas fase de descubrimiento y fase de resolución de alias[18].

El objetivo de la fase de descubrimiento es recabar información sobre la red de interconexión de la que se desea realizar el mapa ya que normalmente no se disponen datos de la misma. La fase de descubrimiento consiste en recopilar un conjunto de direcciones IP que pertenecen a la red de interconexión y los enlaces de los routers a los que pertenecen dichas direcciones IP. Los métodos usados en esta fase se basan en sondear los routers intermedios en el camino a un destino. Si se realiza un grafo con la información exclusiva de esta fase los nodos representarían direcciones IP y no routers que es lo que se desea conseguir en el proceso de mapeo de red a nivel de router.

La fase de resolución de alias consiste en decidir si distintas direcciones IP pertenecen al mismo router. Un router puede tener varios interfaces conectados a distintas redes y por tanto con distintas direcciones IP. Esta fase se realiza con las

direcciones IP conseguidas en la fase de descubrimiento. El mapa de red resultante es un mapa que contiene, en caso de que la resolución sea completa, tantos nodos como routers hay en la red de interconexión. En comparación con el grafo resultante de la aplicación directa de la fase de descubrimiento, el número de nodos se ve reducido ya que todas las direcciones IP pertenecientes al mismo router se representan en el mismo nodo. Lo mismo pasa con el número de arcos que se reducen porque algunos de estos se ven duplicados al representar en el grafo un nodo por dirección IP en lugar de que el nodo corresponda a un router.

Esta tarea puede parecer sencilla ya que la intuición lleva a la falsa idea de que la solución pasa por un simple alineamiento de las direcciones IP pertenecientes a caminos opuestos entre cada par de máquinas finales. Por ejemplo, los paquetes de ida y vuelta pueden seguir caminos diferentes que pueden tener diferente número de saltos y hacer que resulte imposible realizar el alineamiento. Un segundo problema es que esa estrategia no puede asociar las direcciones IP pertenecientes a caminos diferentes que se cruzan en determinado router. El tercer problema es que, en numerosas ocasiones, no se dispone de la información del camino en ambos sentidos.

En la literatura existen distintas técnicas que pretenden solventar los problemas que aparecen a la hora de la realización de un mapa de red a nivel de router para ambas fases. En este capítulo se detallan las estrategias y técnicas que se encuentran en el estado del arte para el proceso de mapeado de una red a nivel de router.

2.2 Técnicas para la fase de descubrimiento

2.2.1 Registro de ruta

El protocolo IP permite almacenar la ruta que toman los distintos paquetes enviados a través de la red [19]. Esta opción se denomina registro de ruta (*Record Route*) y permite solicitar a los routers intermedios que adjunten su dirección IP dentro del paquete que se está reenviando. Conforme el paquete con la opción habilitada va atravesando los routers de camino a su destino, los routers añaden en el campo de opciones de dicho paquete la dirección IP de la interfaz por la que sale el paquete. En el caso de que el paquete sonda sea ICMP de tipo *Echo Request*, el equipo

final (el que está marcado como destino en el paquete sonda) incluirá también en su paquete de respuesta la opción habilitada y copia las direcciones IP que se encontraban en la opción de registro de ruta del paquete recibido. El paquete sonda es aquel que se envía desde un origen controlado y que se ha generado con unas características especiales (habilitando o modificando determinados campos). Esas características del paquete permiten que otro equipo devuelva un paquete de respuesta con el que se miden parámetros útiles para, por ejemplo, la resolución de alias.

El principal problema a la hora de utilizar este método de descubrimiento es el comportamiento de los routers ante el reenvío de paquetes con esta opción habilitada, ya que éstos pueden no insertar su dirección IP en el paquete según cómo se hayan configurado. También, debido a distintas implementaciones o configuraciones, existen routers que insertan la dirección IP de la interfaz de entrada en lugar de la de salida [20].

Por otro lado, debido al límite de tamaño de los campos de opciones del protocolo IP, el número máximo de direcciones IP que se pueden almacenar en esta opción es de 9. Esta capacidad resulta, en la mayoría de casos, insuficiente para reflejar el camino completo recorrido por los paquetes en Internet [21].

2.2.2 **Traceroute**

En el año 1989 Van Jacobson ideó otra forma de poder conocer las rutas que toman los paquetes a la hora de ser mandados desde un origen a un destino. Su herramienta, conocida como traceroute [5], se basa en el envío de paquetes sonda UDP a un destino y puerto en desuso (elegido de manera aleatoria), previa modificación del campo de la cabecera IP tiempo de vida (TTL). En este caso, los paquetes sonda permiten recibir contestaciones por parte de los routers del camino y la contestación por parte de la dirección IP destino a la que van dirigidos esos paquetes. El campo TTL se inicializa con un número y los routers lo van decrementando en una unidad hasta que llega a cero que es cuando estos lo descartan. Este campo TTL se usa para evitar que los paquetes enviados sean reenviados de forma continuada a través de la red debido a una configuración errónea.

En la herramienta traceroute, este campo TTL se modifica dándole valores desde 1 hasta un valor tal que los paquetes sonda lleguen al destino. Conforme los routers del camino vayan tirando los paquetes, irán informando de ello al origen mediante paquetes ICMP de error de tiempo excedido en transito. El destino al recibir un mensaje UDP a un puerto que no tiene abierto envía un ICMP de error de puerto inalcanzable a la dirección IP de origen de ese paquete sonda. Como resultado de esta técnica se obtienen las distintas direcciones IP de routers por las que pasan los paquetes desde un origen hasta un destino. También se obtiene una relación de adyacencia entre routers inferida de las direcciones IP consecutivas.

En implementaciones mas actuales del traceroute, también se utilizan paquetes sonda ICMP de tipo *Echo Request* en lugar de los paquetes de tipo UDP a un puerto en desuso. El procedimiento será el mismo, pero a la hora de detectar que se ha alcanzado el equipo final, se realizara mediante el paquete ICMP de tipo *Echo Reply* que es enviado por este al recibir el paquete sonda.

En la figura 2.1 se muestra el comportamiento del traceroute. En ella observamos como un paquete con valor de TTL igual a 1, devuelve un paquete ICMP de error de tiempo de vida excedido en tránsito que tendrá de origen la dirección IP del primer router por el que pasa el paquete. En la segunda parte de la figura se puede observar un paquete con mayor valor de TTL, en este caso el valor es 3. Este paquete llegará más lejos en la red y nos devolverá la dirección IP del router perteneciente al salto número 3 del camino.

La herramienta traceroute es ampliamente conocida y usada por prácticamente todos los gestores de red para, por ejemplo, la resolución de problemas de conectividad. Cuando alguien quiere saber en qué dirección IP del camino existe un problema que imposibilita la comunicación de un equipo concreto con un destino, el traceroute es una de las primeras herramientas que se utilizan. De esta manera se sabe hasta dónde se puede llegar y si el router que está generando el problema pertenece a nuestra red.

Esta herramienta tiene limitaciones a la hora de ser utilizada para la inferencia tanto de rutas como de direcciones IP pertenecientes a una red. Multitud de routers están configurados con reglas de filtrado que inhabilitan el envío de paquetes de tiempo excedido en transito. Este hecho ocasiona que en la inferencia del camino

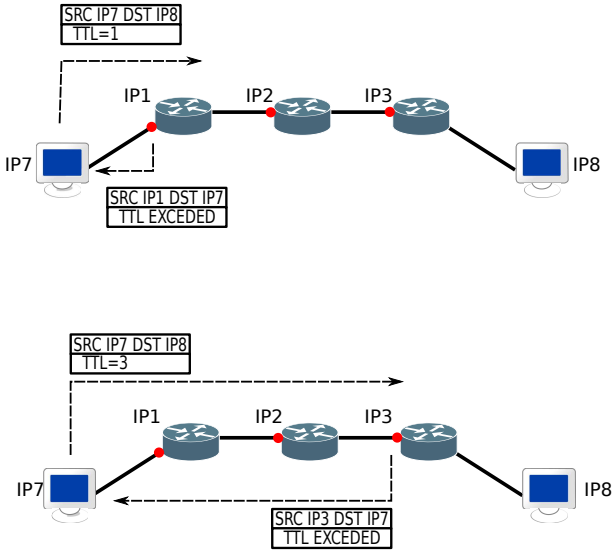


Figura 2.1: Ejemplo de funcionamiento de traceroute

pueda haber saltos de los que no se tiene información sobre su dirección IP (comúnmente aparecen identificados con el símbolo asterisco en las implementaciones del traceroute).

Por otro lado, existen routers que pueden realizar balanceo de carga. Este tipo de routers realizan redirecciones de los paquetes que pasan a través de ellos reenviándolos a distintos routers atendiendo a parámetros de congestión o distribución equitativa de la carga. Esto implica que pueden existir varios caminos entre un equipo origen y un equipo destino. Existen dos tipos principales de balanceo de carga, los que se realizan por paquete y los que realizan por flujo. En el primer caso, cada paquete puede tomar un camino diferente. En el otro caso, el balanceo de carga por flujo permite que los paquetes que compartan una misma tupla [protocolo, dirección IP origen, dirección IP destino, puerto origen y puerto destino], sigan el mismo camino.

Un problema que se deriva de la existencia de balanceos de carga en el camino que se estudia mediante traceroute es que algunos de los enlaces inferidos entre direcciones IP pueden ser erróneos [22]. Como los paquetes sonda que se envían pueden seguir caminos diferentes, ocasionan que dos direcciones IP recogidas en dos saltos consecutivos no tengan por qué estar conectadas entre si. En la figura 2.2 se puede observar un ejemplo de una inferencia errónea con motivo de un balanceo

de carga. En la figura existen dos caminos que pueden tomar los paquetes, el compuesto por las dirección IP1, IP2 y IP3 (se le denominará camino A) y el compuesto por las direcciones IP1, IP4 y IP5 (se le denominará camino B). El router que tiene la interfaz IP1 está realizando un balanceo de carga de manera que cada paquete recibido se envía al router con interfaz IP2 o al que tiene la interfaz IP4. Cuando el paquete enviado tiene TTL con valor de 2, el router que hace el balanceo decide enviarlo por el camino A con lo que el paquete ICMP de error de tiempo excedido en tránsito tiene la dirección IP2 como origen. En cambio cuando se realiza el envío con TTL valor 3, el balanceador decide enviarlo por el camino B, lo que ocasionará que el ICMP de tiempo excedido lo genere el router con dirección IP5. El resultado de inferir los enlaces en este caso particular es que se obtiene un enlace entre la dirección IP2 y la dirección IP5 que pertenecen en realidad a caminos diferentes y que por tanto no suponen un enlace real entre ambos routers.

2.2.3 **Paris-traceroute**

El Paris-traceroute [22] es una variación de la herramienta traceroute de Van Jacobson. Una mejora propuesta al sistema clásico de traceroute se basa en el aumento de los tipos de paquetes que se envían. Además del uso de los paquetes de tipo UDP, se amplia para que se puedan enviar también paquetes de tipo ICMP de *Echo Request* y paquetes de tipo TCP con el flag *SYN* habilitado a un puerto en desuso (el puerto se selecciona de manera aleatoria igual que en el traceroute).

El modo de detectar los routers del camino hasta el equipo final es igual que en el traceroute. Mediante la variación del valor del campo TTL se consigue que los equipos intermedios envíen paquetes de tiempo excedido en tránsito. En el caso de usar paquetes sonda de tipo ICMP de *Echo Request* se detecta que el equipo final ha recibido el paquete sonda porque envía un paquete ICMP de *Echo Reply* como respuesta. Si se utilizan paquetes sonda de tipo TCP, la detección de que el paquete ha llegado al equipo final será posible gracias a que éste deberá enviar un paquete de tipo TCP con el flag de *Reset* activado.

Mediante el uso de una mayor variedad de tipos de paquete, se consigue que existan más routers que contesten en el camino, ya que algunas redes o routers

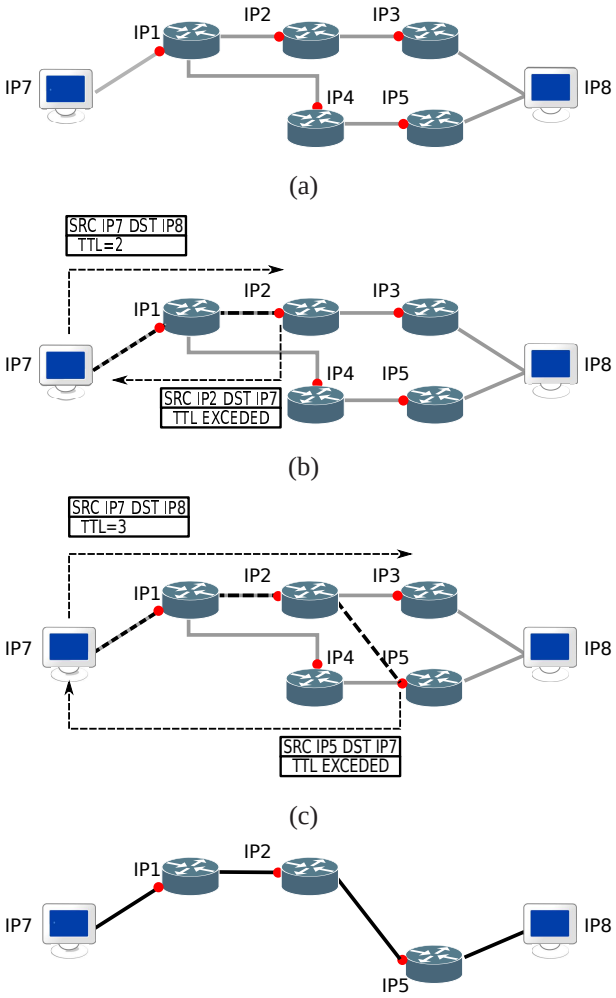


Figura 2.2: Ejemplo de inferencia de topología incorrecta por balanceos de carga

pueden estar configuradas para no contestar o filtrar determinados tipos de paquete. Con esta estrategia se consigue un mayor número de respuestas [22].

La herramienta Paris-traceroute envía los paquetes poniendo una serie de restricciones a los campos del protocolo para conseguir evitar el balanceo de carga por flujo. Esta característica es lo que da realmente el valor añadido a esta herramienta. Si se mantiene la tupla de todos los paquetes sonda, estos no se ven afectados por el balanceo de carga por flujo. Los campos a mantener en la capa de IP son el tipo de servicio (TOS), el campo de protocolo, el campo de dirección IP de origen y el campo de dirección IP destino. Por otra parte, en la cabecera del protocolo superior se deben mantener los primeros 4 bytes. Esto supone tanto en UDP como en TCP el no cambiar los puertos origen y destino. En el caso de utilizar paquetes de tipo ICMP, supone mantener el campo de tipo, de código y de checksum. Esta técnica sin embargo sigue sin solucionar el problema de balanceo de carga por paquete.

2.3 Estrategias para la fase de resolución de alias

A la hora de afrontar el problema de resolución de alias existen diferentes estrategias:

- En cuanto al tipo de estrategia según la necesidad de envío de tráfico sonda pueden usarse tanto estrategias activas cómo de inferencia.
- Según el destino de los paquetes sonda se encuentran estrategias que se basan en la medición directa y otras que se basan en la medición indirecta.
- Dependiendo del parámetro base utilizado para la resolución, existen estrategias basadas en IPID, marca temporal o registro de ruta.
- También existe la posibilidad de utilizar distintos tipos de paquetes sonda para provocar distintas contestaciones por parte de los routers.
- Hay estrategias basadas en la unificación de las dos fases de las que se compone el proceso de mapeado de redes a nivel de router.

- Las estrategias pueden variar según la distribución de las medidas, pueden hacer uso de un sólo punto de medida (estrategias centralizadas) o tener la posibilidad de distribuirse en varios puntos de medida (estrategias distribuidas).

- Existen estrategias que se diferencian por su forma de agregación. Unas se aplican por cada pareja de direcciones IP y otras que se pueden realizar por grupos de direcciones IP.

- Por último, existen estrategias basadas en reducción que permitirán la selección de determinados grupos de direcciones IP entre los que sea más probable encontrar alias.

2.3.1 Estrategias según la necesidad de envío de tráfico sonda

Esta estrategia tiene que ver con la generación del tráfico sobre la red. Existen estrategias de resolución de alias que no requieren de envío de tráfico [23]. Estas se sirven de las medidas ya realizadas por las técnicas para la fase de descubrimiento de topologías y deducen a partir de esos datos las direcciones IP que pertenecen al mismo router. A este tipo de estrategias se les denomina como basadas en inferencia. Por otro lado, las técnicas de resolución de alias que requieran del envío de tráfico extra sobre la red para poder realizar la resolución de alias son las denominadas estrategias activas [24]. Existe un tipo de estrategias que a pesar de obtener los alias a partir del estudio de las medidas de la fase de descubrimiento necesitan de una verificación mediante el envío de algo de tráfico a la red. Este tráfico resulta normalmente muy marginal por lo que a este tipo de estrategias se les denomina estrategias mixtas [18].

2.3.2 Estrategias según el destino de los paquetes sonda

Tiene que ver con la dirección IP destino del paquete sonda y la dirección IP de la que pretendemos obtener respuesta para la resolución de alias. Las técnicas de resolución de alias que utilizan estrategias directas son aquellas en las que el paquete sonda se envía directamente al equipo del que se quiere obtener información. El

paquete tendrá como destino la dirección IP del equipo del que se desea una respuesta. En las estrategias de carácter indirecto [25], el equipo que se quiere medir es diferente al marcado como destino en el paquete sonda. La técnica de traceroute, comentada con anterioridad, utiliza este tipo de estrategia para obtener las direcciones IP de los routers intermedios. Para obtenerlas, hace uso del campo TTL para forzar una respuesta de las direcciones IP intermedias y no de la dirección IP que utiliza como destino en el paquete sonda.

2.3.3 Estrategias según el parámetro base de la resolución

Las técnicas de resolución de alias pueden utilizar uno o varios campos de las cabeceras de protocolos de los paquetes de respuesta a la hora de realizar la inferencia para la obtención de alias pertenecientes al mismo router. Esta inferencia se realiza en base a las peculiaridades que tienen los distintos routers a la hora de rellenar los campos de los paquetes enviados como respuesta. Los campos que se utilizan para poder deducir qué direcciones IP pertenecen al mismo router se detallan a continuación [26].

IPID:

Este es el campo de identificación de la cabecera IP. Cuando se requiera de fragmentación de paquetes, este campo ayuda al reensamblado de los paquetes en el destino para conformar el paquete completo. Dicho identificador debe ser distinto para cada paquete generado por una misma máquina IP. Una implementación extendida para dar valores a este campo, y que sean diferentes entre los distintos paquetes enviados por parte del router, es mediante un contador. Éste contiene el valor del siguiente IPID a enviar, y cada nuevo paquete que se envía a la red incrementa en una unidad dicho contador [4]. Este contador es compartido en general por todas las interfaces del router. Otra forma de tratar el campo de IPID por parte de los equipos es mediante la generación de un número al azar. Existen otros tipos de comportamiento respecto a este campo que se han identificado en este trabajo y se presentan posteriormente.

Las inferencias de resolución que se pueden realizar mediante este tipo de comportamientos se basan en el estudio de crecimiento de los IPIDs de los paquetes

procedentes de dos direcciones IP dadas. Si las direcciones IP pertenecen al mismo equipo y éste tiene una implementación basada en contador, los IPIDs mezclados de ambas direcciones IP describen una secuencia de números creciente. En caso contrario se puede deducir que las direcciones IP a estudio no pertenecen al mismo router.

Marca temporal:

Hay diversos campos de distintas cabeceras de protocolos que permiten guardar un marca temporal:

- Marca temporal de la cabecera IP: este es el que aparece en el campo de opciones de la cabecera IP con 32 bits de tamaño. Si se encuentra habilitado representa los milisegundos transcurridos desde media noche (UTC).

- *Prespecified Timestamp* de la cabecera IP: se trata de un campo de opciones del protocolo IP que permite especificar hasta 4 direcciones IP de los que interese conocer el reloj. Esta marca temporal se mide en milisegundos desde media noche (UTC). La marca temporal se rellena cuando el paquete sonda atraviese el interfaz con la IP especificada y sólo se completa si los huecos reservados para los anteriores marcas temporales se han rellenado previamente [27].

- Marca temporal de la cabecera TCP: esta marca temporal se refresca aumentando en una unidad un contador en franjas desde 1 milisegundo hasta 1 segundo y tiene un tamaño de 32 bits. El contador se inicializa a un valor aleatorio a partir del cual se empieza a contar el tiempo [28].

- Marca temporal de la cabecera ICMP de tipo solicitud de marca temporal: los paquetes ICMP de tipo *Timestamp Request/Reply* permiten almacenar 3 marcas temporales diferentes (*origin*, *receive* y *transmit*) de 32 bits cada uno y representan milisegundos a partir de media noche (UTC). En el mismo paquete sonda se incluye la marca temporal del equipo desde el que se realiza la medida. La máquina destino genera un paquete de respuesta rellenando la marca temporal llamada *origin* con el valor de la marca temporal recibida en

el paquete sonda, la marca temporal *receive* la rellena con la marca temporal que marca su reloj cuando recibe el paquete sonda y por último la marca temporal *transmit* se rellena con la marca temporal en el momento del envío de la respuesta.

Mediante el estudio de este tipo de campos temporales se puede llegar a discernir si dos direcciones IP pertenecen o no a un mismo equipo debido a desviaciones temporales entre los relojes de los diferentes routers [28] o bien utilizando este campo de otros modos [27].

Registro de ruta:

Esta opción de la cabecera del protocolo IP, explicado ya con anterioridad, permite recopilar las direcciones IP de los routers que reenvían el paquete sonda. Las estrategias que permiten resolver alias usando este campo lo hacen mediante la alineación de los caminos derivados de este campo y de los derivados por las pruebas de traceroute, pero hay que recordar que existe un máximo de 9 direcciones IP almacenables mediante este tipo de paquetes [20].

Dirección IP origen:

La dirección IP origen de los paquetes de respuesta también es un campo que permite la identificación de aliases. Un paquete sonda puede ser enviado a una dirección IP destino y la máquina destino responder desde una dirección IP perteneciente a cualquiera de las otras interfaces. Si esto ocurre permite detectar que ambas direcciones IP pertenecen al mismo router [29].

2.3.4 Estrategias según el tipo de paquete sonda

La elección de un tipo adecuado de paquete sonda es una estrategia que permite obtener un mayor número de respuestas de los equipos que se desea medir [26]. Los principales tipos de paquetes sonda utilizados en el estado del arte son:

- ICMP de tipo *Echo Request*: este tipo de paquetes hacen que el equipo final devuelva un paquete de tipo ICMP de *Echo Reply* [18].

- ICMP de tipo *Timestamp Request*: este tipo de paquetes hacen que el equipo final conteste con un paquete ICMP de tipo *Timestamp Reply* [9].
- UDP: en caso de enviarse un paquete de tipo UDP a un puerto que no se esté usando por la maquina final, ésta devolverá un paquete de ICMP de error de tipo puerto inalcanzable [29] y [4].
- TCP: en caso de enviarse un paquete de este tipo a un puerto que no se esté usando en la maquina final, ésta devolverá un paquete de TCP marcado con el flag de *Reset* [30].
- Basado en TTL: se usa un valor de TTL limitado para provocar la respuesta de un router intermedio [25]. El equipo en el que el contador TTL alcance el valor de 0 descarta dicho paquete enviando un ICMP de error de tiempo excedido en tránsito. El protocolo del paquete sonda puede ser cualquiera de los anteriores, pero las contestaciones a este tipo de paquetes son muy diferentes, tanto en tasas de contestación como en el comportamiento de los routers a la hora de rellenar los distintos parámetros base de las diferentes cabeceras.

Debido a las características particulares de cada método de resolución de alias, algunos métodos usan varias de estas estrategias y otros solamente una de ellas. Cuantos más tipos de paquetes contemple una técnica dada más respuestas será capaz de obtener, permitiendo que un número mayor de direcciones IP le contesten y por ello tener un potencial de resolución de alias mayor.

2.3.5 Estrategias basadas en la unificación de fases

Existen técnicas que debido al tipo de paquetes sonda que necesitan para su inferencia permiten la unificación de las dos fases de descubrimiento y de resolución de alias usadas en la creación de mapas a nivel de router. Estas estrategias pueden necesitar de ciertas modificaciones o restricciones en la fase de descubrimiento para permitir que la técnica utilizada para la inferencia se realice con éxito. Por ejemplo, se puede pensar en realizar varios traceroutes para poder utilizar los campos de IPID de los paquetes de respuesta en la resolución de alias basándose en los valores

de los IPIDs. Estas técnicas resultan interesantes ya que consiguen reducir tanto el tiempo como el tráfico generado en la identificación.

2.3.6 Estrategias según la distribución de las medidas

Las estrategias de tipo centralizada son aquellas que no tienen la capacidad de poder ejecutarse en varios equipos sonda [30]. Debido a ello, la carga no es distribuible por lo que el tiempo que requiere la fase de resolución de alias no se puede reducir aun disponiendo de varios equipos sonda. Por otro lado, las estrategias distribuidas [25] permiten el reparto de la carga en varios equipos permitiendo reducir el tiempo invertido en la fase de resolución de alias y permiten a su vez que obtengan tasas de contestación mayores por parte de las direcciones IP en estudio. Sólo el hecho de que una estrategia sea distribuible no hace que sea mejor que otra. Por ejemplo, una estrategia de tipo centralizada puede requerir de tan sólo el envío de un paquete sonda por cada dirección IP, y por otro lado una estrategia distribuida podría requerir del envío de un paquete sonda por cada pareja posible de direcciones IP.

2.3.7 Estrategias según la forma de agregación

Existen técnicas que se tienen que realizar por cada pareja de direcciones IP por restricciones inherentes a la propia estrategia de identificación [4]. Este tipo de estrategias tienen un coste de orden cuadrático $O(N^2)$ ya que para un número dado N de direcciones IP se requiere de la realización de N^2 medidas. Esto tiene implicaciones de carácter computacional y de tráfico generado que hacen que según el número de direcciones IP a analizar resulte inviable el usar este tipo de estrategias. Por otro lado, las técnicas que permiten la resolución de direcciones IP en grupos [30], permiten normalmente el envío de paquetes sonda por grupo, de forma que el coste computacional y de tráfico generado es de orden lineal $O(N)$ por cada grupo.

2.3.8 Estrategias basadas en reducción

Las estrategias de reducción permiten identificar los grupos de direcciones IP con más probabilidades de ser alias previamente a las medidas de la fase de resolución

de alias [31]. Este tipo de métodos aparecen para solucionar problemas relacionados con la imposibilidad de realizar las pruebas de resolución de alias a todas las direcciones IP de la red. Después de aplicar la estrategia de reducción, el número de parejas candidatas al que se deben de realizar las medidas de la fase de resolución de alias es una parte del total. Al ser una estrategia probabilística, existe un porcentaje de error en todo este tipo de técnicas que deberá de añadirse al porcentaje de error ocasionado por la técnica de resolución utilizado. En la literatura se encuentran dos estrategias de este tipo, una basada en el campo TTL y otra basada en el campo de IPID.

Reducción basada en TTL

El primero de este tipo de estrategias está basado en el uso del campo TTL [24]. Se basa en que hay más probabilidad de encontrar una pareja de direcciones IP pertenecientes al mismo router entre aquellas direcciones IP que se encuentren más cercanas según su TTL. Por este motivo, este método divide las parejas de direcciones IP según su distancia usando el valor del campo TTL recibido desde un mismo punto de medida.

En el caso concreto en el que para llegar a las interfaces de un mismo router se recorriese el mismo camino y no existiesen asimetrías en la red, la distancia entre los valores de TTL de los paquetes recibidos desde todas ellas debería ser 0. El problema es que normalmente para llegar a dos interfaces IP de un mismo router se toman caminos diferentes ya sea por equipos que realizan balanceos de carga o por distintas políticas de enrutamiento.

Debido a las diferencias en los campos de TTL de paquetes provenientes de direcciones IP pertenecientes al mismo router, se selecciona un umbral para la comparación de los valores de TTL. Así, si paquetes provenientes de dos direcciones IP contienen valores de TTL que no superan dicho umbral, se aplica la fase de resolución de alias sobre ellas.

Esta estrategia requiere del uso de traceroutes desde un mismo equipo sonda para obtener las distancias en TTL desde este equipo hasta cada una de las direcciones IP. Debido a que todos los paquetes sonda han de enviarse desde el mismo nodo sonda la realización de las medidas no puede ser distribuida.

Reducción basado en IPID

Esta estrategia de reducción se basa en el campo IPID de la capa IP y en el uso particular de un contador incremental para rellenar dicho campo por parte de los routers. Debido a que el contador es compartido por todas las interfaces del router [29], si la distancia entre dos campos IPID de dos direcciones IP distintas es pequeño tendrán más probabilidades de pertenecer a un mismo router que dos direcciones IP cuya distancia en IPID sea mayor. En el caso de que el router no genere otro tráfico IP la distancia de IPID entre 2 respuestas consecutivas es la unidad. Sin embargo, dado que el router puede realizar envíos hacia otros equipos, se ha de determinar un umbral. Dicho umbral de diferencia entre IPIDs permite, como en el caso del uso del campo TTL, identificar a qué routers se les ha de realizar las pruebas de resolución de alias.

El problema del que adolece esta estrategia es que a pesar de que sea distribuible en varios equipos sonda, es muy dependiente del tiempo invertido en realizar la medida de los campos de IPID de las respuestas de las maquinas finales. Existen routers con incrementos de IPID cercanos a 200 unidades por segundo [9]. Esto implica que la medida se debe de realizar en menos de 330 segundos ya que el campo de IPID tiene un máximo de 65.535 números y en ese tiempo se provoca un reinicio del contador.

Otras estrategias de reducción

La técnica Iplane [32] explicada en la sección 2.5.3 utiliza en la propia técnica de resolución una serie de condiciones que se pueden considerar estrategia de reducción. En ella la determinación final para valorar si una pareja de direcciones IP pertenece o no al mismo router es mediante la cercanía del valor de los campos IPID pertenecientes a los paquetes de respuesta recibidos desde los routers. En este caso la comparación no se realiza entre cualquier pareja de direcciones IP, si no que se realiza primero un filtrado por sistema autónomo, después un filtrado por TTL para finalmente realizar la comparación de cercanía de IPID.

La estrategia de ventana deslizante utilizada por Midar [25] y explicada en la sección 2.5.11, también podría ser considerada una técnica de reducción. Esa estrategia consiste en realizar solo las medidas a aquellas direcciones IP que tienen

velocidades similares de tráfico dirigido hacia ellas. El tamaño de la ventana dictamina a qué direcciones IP hay que realizar las medidas, y al resto no se le harán.

2.4 Métricas de rendimiento de las técnicas de resolución de alias

Existen varias métricas que permiten comparar las distintas técnicas de resolución de alias: precisión, completitud, eficiencia y distribuibilidad [33]. A continuación se detallan cada una de estas métricas.

2.4.1 Precisión

Esta métrica está relacionada con la evaluación de la respuesta dada por una técnica de resolución de alias concreta para una pareja de direcciones IP. Las distintas técnicas de resolución de alias pueden dar 4 tipos de resultados para una pareja de direcciones IP (positivo, negativo, no concluyente y error) [9]. El resultado positivo se da cuando la técnica identifica una pareja de direcciones IP como pertenecientes al mismo router. Negativo cuando identifica que dos direcciones IP no pertenecen al mismo router. El caso de no concluyente es debido a un bajo número de contestaciones o a la ausencia de información suficiente suministrada por el parámetro base lo cual no hace posible emitir un veredicto. El caso de error se da cuando alguna de las direcciones IP de la pareja a las que se esta realizando la medida no responde a los paquetes sonda.

En caso de que la técnica devuelva un resultado positivo pero la pareja de direcciones IP no pertenezca al mismo router, se habla de falso positivo. Se habla de falso negativo cuando se obtiene un resultado negativo y realmente la pareja de direcciones IP pertenece al mismo router.

La precisión mide la relación entre parejas de direcciones IP identificadas correctamente (positivos y negativos) y las parejas identificadas incorrectamente (falsos positivos y falsos negativos) [34]. Dicha relación se expresa mediante el porcentaje de parejas con identificación fallida (la suma de falsos positivos y falsos negativos) respecto al número de parejas identificadas totales (la suma total de positivos, negativos, falsos positivos y falsos negativos).

Para evaluar la precisión es necesaria una topología conocida. Para ello se puede optar por el uso de una maqueta, que es una red artificial creada para realizar pruebas en un laboratorio, o por el uso de la información pública que ofrecen algunas redes académicas y de investigación nacionales como Geant [1], Canet4 [3] o GlobalNOC [2].

2.4.2 Completitud

La completitud mide la cantidad de parejas que consigue identificar una técnica de resolución de alias dada, es decir, la suma de resultados positivos y negativos. Una identificación completa significa que la suma de positivos y negativos es igual al total de posibles parejas que se pueden formar con las direcciones IP encontradas en la fase de descubrimiento [35]. Normalmente, las técnicas de la fase de resolución de alias no consiguen un 100 % de completitud debido a los filtrados de paquetes en routers y a la información insuficiente que provee el parámetro base de algunos paquetes de respuesta.

2.4.3 Eficiencia

La eficiencia se centra en el coste [35], tanto de tráfico como de computación, del proceso de identificación. Una técnica es mas eficiente cuanto menos tráfico introduce en la red y menos tiempo es necesario invertir en el cálculo de la misma. De este modo, una técnica de resolución de alias que no introduzca nada de tráfico adicional y su cálculo se realice en poco tiempo tiene la característica de ser eficiente. De las estrategias para la resolución de alias vistas con anterioridad, la que aporta mayor eficiencia es la que permite la agregación en grupo.

2.4.4 Distribuibilidad

Esta métrica evalúa la posibilidad de distribuir el tráfico sonda y el procesado de las respuestas en distintas máquinas. Se puede distribuir tanto la medida como el procesado de forma independiente. Normalmente, a igual complejidad computacional, la opción distribuible tenderá a escalar mejor además de que la distribución

de la parte de las medidas permite que filtros que se pueden encontrar para un equipo sonda dado, no se apliquen a otros equipos sonda.

2.5 Técnicas para la resolución de alias

En este apartado se detallan cada una de las técnicas del estado del arte para la fase de resolución de alias. Este conjunto de técnicas son las que permiten asociar distintas direcciones IP de un mismo router y son una pieza clave en la obtención de mapas a nivel de router. En las siguientes subsecciones se explicará con detalle el procedimiento a seguir para realizar una resolución de alias con cada una de ellas, el tipo de estrategias utilizadas y se realizará una valoración en función de las métricas vistas con anterioridad. Por regla general las distintas técnicas propuestas en el estado del arte no ofrecen una implementación pública de la herramienta para poder realizar una evaluación, por lo que muchas de ellas se han tenido que reimplementar. Las técnicas de las que se dispone una versión implementada de acceso público son Midar, Palmtree, TraceNet y Velocity modeling (Radargun) que, a día de hoy, ya no se encuentra disponible.

2.5.1 Mercator

Mercator [29] es la técnica de resolución de alias utilizada por el proyecto SKITTER [36] desarrollado por el grupo CAIDA [37]. El proyecto SKITTER realiza, a través de pruebas desde distintos nodos sonda, un mapa a nivel global de Internet obteniendo diferentes características de los equipos que se van descubriendo. Una de las características que se pretende obtener son las direcciones IP que pertenecen al mismo router y para ello se utiliza la técnica Mercator [29]. Mercator es una técnica activa basada en el campo de dirección origen del paquete de respuesta, que emplea paquetes sonda de tipo UDP directo al conjunto de direcciones IP que se desea identificar. Se trata de una técnica centralizada basada en grupos sin capacidad de unificación de fases.

Esta técnica de resolución de alias se aprovecha del comportamiento de ciertos routers a la hora de enviar los mensajes ICMP de error de puerto inalcanzable. Estos routers están configurados para que los paquetes de error sean enviados por

la interfaz de salida más cercana a la dirección IP origen del paquete sonda que lo ha ocasionado. La dirección IP de esa interfaz quedará reflejada en el campo IP origen de la cabecera IP del paquete recibido en el equipo sonda. En el caso de que tras el envío de dos paquetes sonda a diferentes direcciones IP se obtenga como respuesta paquetes ICMP de error con la misma dirección IP origen, se puede concluir que las 2 direcciones IP originales pertenecen al mismo router. En la figura 2.3 se muestra el comportamiento de un router con este tipo de configuración ante el envío de dos paquetes diferentes de tipo UDP a sus dos interfaces y cómo el envío del ICMP de error se realiza desde la misma interfaz de salida.

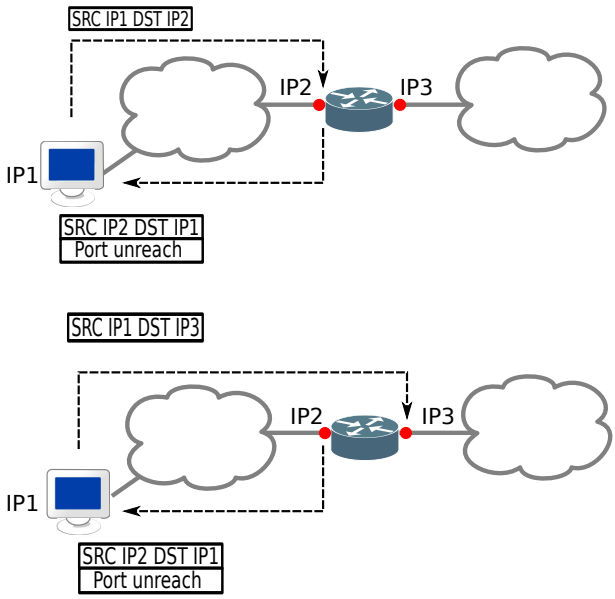


Figura 2.3: Ejemplo de funcionamiento de método Mercator

El hecho de que el router tenga una interfaz con camino más corto a la dirección IP origen del paquete sonda no quiere decir que los paquetes de todas las interfaces del router se envíen por ese mismo interfaz, por tanto esta técnica sólo permite detectar direcciones IP que pertenezcan al mismo router (identificación de positivos) y no permite deducir que dos direcciones IP pertenecen a routers diferentes (identificación de negativos).

Mercator ofrece una precisión y una eficiencia alta, pero tiene muchas deficiencias a nivel de completitud y distribución. La baja completitud es debida a las

distintas políticas de seguridad que hacen que los routers no contesten a los paquetes UDP enviados a un puerto aleatorio. La precisión es porque no caben falsos positivos ya que no se puede falsificar las direcciones IP origen de los paquetes. La eficiencia alta se justifica porque sólo hace falta el envío de un paquete sonda UDP a cada dirección IP que se quiere identificar. Por tanto el coste es totalmente lineal y la carga de red muy baja. La mala distribución se ocasiona porque si las medidas de esta técnica se distribuyen, no se cumplen los requisitos necesarios para ofrecer la identificación ya que se requiere enviar desde un mismo punto los paquetes sonda para que el camino más corto elegido por los distintos routers para la respuesta sea el mismo.

2.5.2 **Ally**

Ally [4] es una técnica de resolución de alias desarrollada por el equipo de Rocketfuel [38] en el año 2002. Tanto Ally como Mercator [30] se usan en el sistema de identificación Rocketfuel [38]. El equipo de Rocketfuel ha conseguido muy buenos mapas de Internet a nivel de sistema autónomo. La resolución de alias la usan para obtener medidas mejores de las capacidades de los links que interconectan los sistemas autónomos, ya que el resultado es más preciso si se realizan los estudios a nivel de router y no a nivel IP [4].

Ally es una técnica de resolución activa que basa sus medidas en el envío de paquetes UDP directos a un puerto en desuso de la pareja de direcciones IP. Los equipos destino contestan con un paquete ICMP de error de puerto inalcanzable. La resolución estudia el crecimiento del campo de IPID de dichos paquetes de respuesta. La técnica usa agregación basada en parejas a las que se realiza la identificación de forma independiente por lo que es distribuible mediante el reparto de todas las parejas existentes entre los equipos sonda que se dispongan.

La técnica envía 3 paquetes sonda hacia las 2 direcciones IP de las que se quiere saber si pertenecen al mismo router. Los dos primeros paquetes sonda se envían pegados a cada una de las direcciones IP y el tercero se envía a la interfaz cuya respuesta se obtenga en primer lugar con una separación de un segundo. Una vez obtenidos los tres paquetes de respuesta se estudia si los campos de IPID de los

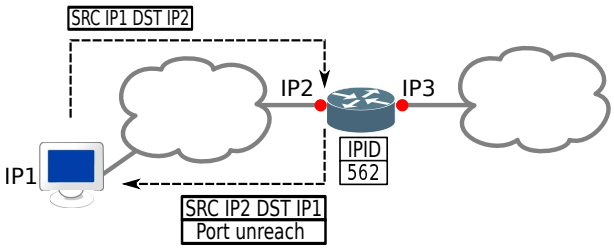
paquetes recibidos siguen un patrón creciente, en cuyo caso la pareja de direcciones IP se evalúa como perteneciente al mismo router (resultado positivo). En caso contrario la pareja de direcciones IP se evalúa como pertenecientes a distinto router (resultado negativo).

Con el objetivo de evitar fallos de reconocimiento ocasionados por direcciones IP que siguen un patrón aleatorio, se establece un umbral máximo para la diferencia de valores del IPID del primer paquete y el del último. Dicho umbral se fija en 200 unidades de diferencia para el campo IPID de la cabecera IP. Si en el estudio de los valores de los campos de IPID de los paquetes de respuesta se supera ese umbral, la técnica da el resultado de no concluyente para la pareja de direcciones IP.

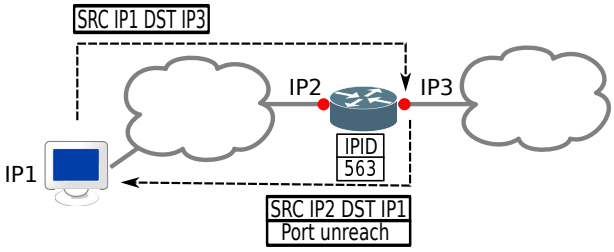
En la figura 2.4 se puede ver un ejemplo de cómo funciona la técnica Ally para la resolución de alias en dos direcciones IP pertenecientes al mismo router con contador común para dar valores al campo de IPID. El equipo sonda realiza el envío de los 3 paquetes (figura 2.4(a), 2.4(b) y 2.4(c)) con una espera de 1 segundo para el envío del último. Los valores que recibe en el campo IPID de los paquetes de respuesta son 562, 563 y por último 564 que permiten la formación de una secuencia creciente utilizando los valores en dicho campo. La ausencia de tráfico interferente por parte de otros equipos entre los paquetes enviados por el equipo sonda, hace que la secuencia de valores del IPID de los paquetes de respuesta aumenten de 1 en 1 (ver figura 2.4(d)). El veredicto final para este ejemplo es que las dos direcciones IP pertenecen al mismo equipo.

La técnica Ally tiene una precisión media dado que adolece tanto de falsos positivos como de falsos negativos [9]. Los routers, y las direcciones IP de los interfaces que poseen, pueden tener dos tipos de comportamientos para dar valores al campo de IPID: asignarlos mediante un contador incremental o mediante un contador aleatorio. Cuando se forman parejas de direcciones IP se pueden encontrar 4 combinaciones de estos comportamientos: incremental-incremental, incremental-aleatorio, aleatorio-incremental y aleatorio-aleatorio.

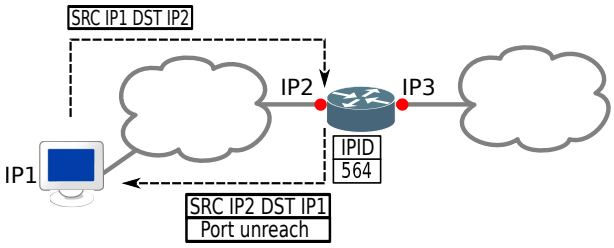
Para las parejas de tipo incremental-incremental e incremental-aleatorio se puede dar un falso positivo en el caso en que el valor del campo de IPID de la segunda dirección IP coincida casualmente con un valor entre el primer y el segundo IPID de la primera dirección IP (ver figuras 2.5(a) y 2.5(b)). Debido a las características del método, que utiliza tan sólo 3 paquetes sonda, es posible que los paquetes



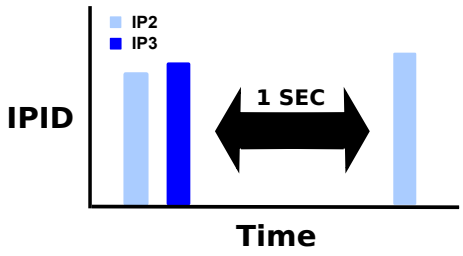
(a)



(b)



(c)



(d) Gráfica generada con los IPID de los paquetes respuesta

Figura 2.4: Ejemplo de funcionamiento de método Ally

de respuesta de dos direcciones IP que pertenezcan a distinto router generen una secuencia en los campos de IPID que sea creciente dando un falso positivo (ver figura 2.6). Se puede dar un falso negativo en el caso de una pareja incremental-incremental cuando el router envía mucho tráfico lo que ocasiona que exista un gran número de reseteos del contador en poco tiempo pudiendo dar como resultado que el primer y el ultimo IPID se dispongan de forma creciente y el segundo sea menor (ver figura 2.5(c)).

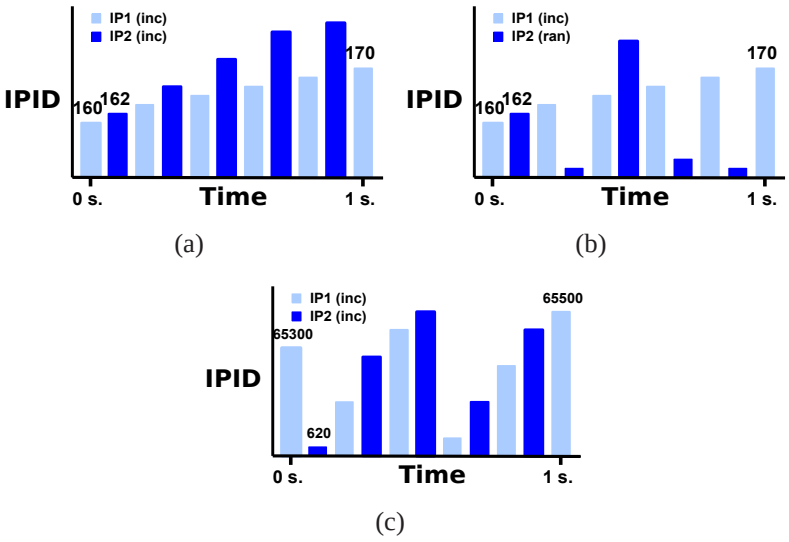
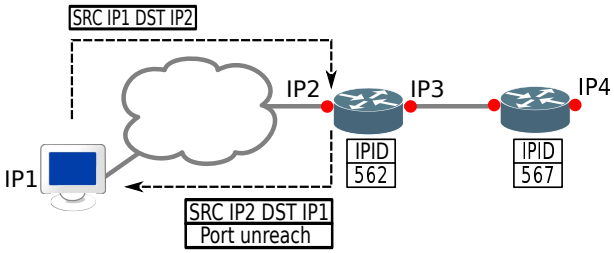
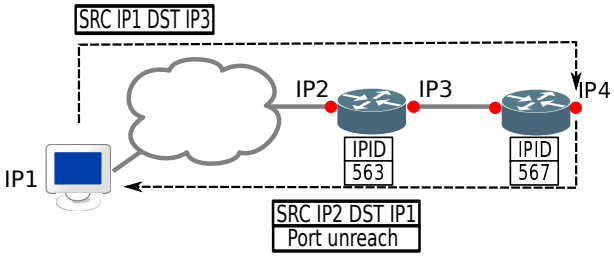


Figura 2.5: Diferentes errores de evaluación de parejas con falso positivo en 2.5(a) con dos direcciones IP incrementales y en 2.5(b) con una incremental y otra aleatoria. Y una pareja de la que se obtiene falso negativo en 2.5(c) siendo el router de tipo incremental.

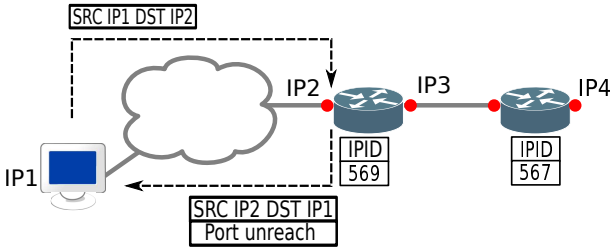
Para las parejas de tipo aleatorio-incremental y aleatorio-aleatorio puede haber un falso positivo en el caso de que, debido a los valores aleatorios que da el primer router, el primer y el tercer paquete de respuesta tengan dos campos de IPID que sigan un patrón creciente y no disten más de 200 unidades, y que además el segundo router provea una respuesta con un valor de IPID que se encuentre entre los dos anteriores (ver figuras 2.7(a) y 2.7(b)). Por último, el método puede dar un falso negativo en una pareja aleatorio-aleatorio si, como en el caso anterior, el primer y tercer paquetes de respuesta contienen un patrón creciente que no diste más de 200



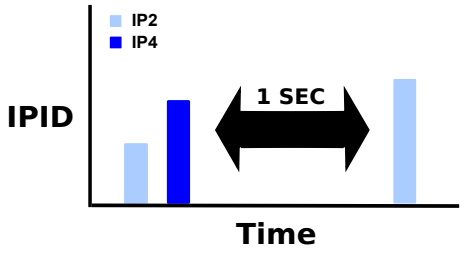
(a)



(b)



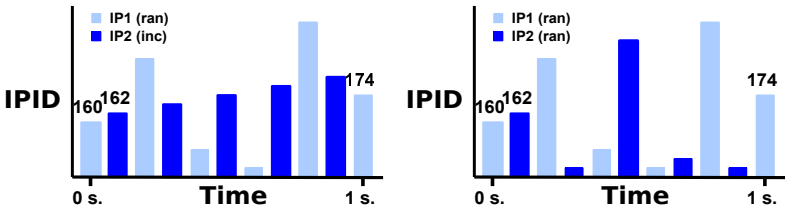
(c)



(d)

Figura 2.6: Ejemplo de funcionamiento de método Ally en dos routers que darán una respuesta de falso positivo

unidades y que el segundo tenga un IPID fuera del intervalo válido.



(a) (b)

Figura 2.7: Diferentes errores de evaluación de parejas tomando en 2.7(a) una de las direcciones IP aleatoria y la otra incremental y en 2.7(b) las dos direcciones IP son aleatorias.

La eficiencia de la técnica de Ally es baja debido a que se debe de realizar por parejas de direcciones IP haciendo que el coste de esta técnica sea de orden cuadrático $O(N^2)$ y además se debe de realizar el envío de 3 paquetes por cada pareja. Por otro lado, este nivel de agregación por parejas hace que se pueda realizar una distribución de las parejas entre equipos sonda haciendo que la distribuibilidad sea alta. Por último, la completitud de esta técnica es baja debido a que se basa únicamente en paquetes de tipo UDP y porque la restricción de que los dos primeros paquetes sonda se manden pegados hace que en muchos casos sólo se responda al primero de ellos.

2.5.3 Iplane

Un proyecto de referencia en el campo de las estrategias de creación de mapas de Internet es Iplane [32]. Dicho proyecto pretende la realización de un mapa de Internet a nivel mundial. Por este motivo no tiene tanta importancia la precisión ofrecida por la técnica de resolución de alias concreta como que la técnica sea eficiente para que se realice mediante pruebas que puedan extenderse a todo Internet.

La técnica que se utiliza en Iplane a la hora de hacer las pruebas de resolución de alias se basa en encontrar grupos con los posibles alias que pertenezcan al mismo sistema autónomo. Esa información de los alias se obtiene de las tablas de BGP de los routers. Una vez que se dispone de esa información, se realiza el envío

de paquetes sonda ICMP directo de tipo *Echo Request* consecutivamente a todas las direcciones IP del sistema autónomo para estudiar en el paquete de respuesta su TTL y su IPID. Se consideran pertenecientes al mismo router todas las direcciones IP que en el paquete respuesta tengan el mismo valor de campo TTL y que sus campos de IPID estén cercanos. La técnica es distribuible ya que cada nodo sonda puede realizar las pruebas a las direcciones IP pertenecientes a cada sistema autónomo y por tanto usa una forma de agregación grupal.

Mediante Iplane consigue realizar la comprobación de una parte muy amplia de Internet con datos de alias que se actualizan cada dos meses. Dicha información de alias se utiliza después para mediciones de anchos de banda disponibles entre routers y ofrece datos que se ajustan bastante bien al ancho de banda y capacidad real de los enlaces.

Esta técnica ofrece una tasa de precisión baja ya que debido a la forma en la que se realizan las medidas puede haber muchas direcciones IP que coincidan tanto en TTL como en IPID pero que no pertenezcan al mismo router. Por otro lado, la técnica no permite la detección de alias pertenecientes a routers frontera, ya que direcciones IP pertenecientes a distintos sistemas autónomos son catalogadas como no pertenecientes al mismo router (respuesta negativa), pero los routers frontera habitualmente tienen direcciones IP perteneciente a diferentes sistemas autónomos.

Por otro lado la completitud es baja debido a lo comentado ya sobre los routers frontera y a las restricciones que se exigen a las direcciones IP para ser parejas del mismo router: mismo valor en el campo de TTL y cercanía en el valor del campo IPID. En contraste con lo anterior, la eficiencia de la técnica es alta ya que no hay mucho gasto ni de ancho de banda ni de cálculo. Una consulta a las tablas BGP para saber el sistema autónomo al que pertenece cada dirección IP y un paquete sonda ICMP a cada una de las direcciones IP, es lo que se necesita para la resolución de alias. El gasto computacional pasa por la ordenación y filtrado de las direcciones por sistema autónomo y campo de TTL, con lo que las comparaciones finales por el campo IPID se ven muy reducidas. Por último, la técnica tiene una distribuibilidad alta ya que repartiendo las direcciones IP que pertenezcan al mismo sistema autónomo en el mismo nodo sonda, cada nodo sonda será capaz de realizar tanto las pruebas como el cálculo de las direcciones IP que son alias.

2.5.4 *Prespecified Timestamp*

La técnica de resolución de alias conocida como *Prespecified Timestamp* [27] es una técnica activa que utiliza paquetes sonda ICMP de tipo *Echo Request* directos con la opción de *Prespecified Timestamp* de la cabecera IP habilitada. Es una técnica fácilmente distribuible y que tiene un nivel de agregación por parejas. Se basa en el comportamiento de los routers cuando tienen que reenviar un paquete con dicha opción habilitada.

El campo de opciones *Prespecified Timestamp* permite la introducción de hasta 4 direcciones IP de las que se desea saber su marca temporal. Cuando el paquete se reenvía en el router que posea una interfaz especificada en el campo *Prespecified Timestamp* y sólo si los valores de marca temporal para las direcciones marcadas anteriormente se han rellenado ya, se introduce la marca temporal de dicho router. De esta forma, si se especifican 4 interfaces pertenecientes al mismo router y se envía un paquete directo a dicho router, se deben rellenar todas las marcas temporales. La forma de ordenar la tupla de 4 direcciones si se desea realizar la técnica a las dirección IP A y a la dirección IP B debe ser [IPA, IPB, IPA, IPB]. Al intercalar las direcciones IP, se evitan confusiones provocadas por la inserción de valores de marcas temporales por parte de los routers del camino.

Debido a que hay routers que no rellenan por completo el campo de *Prespecified Timestamp*, la técnica también contempla el envío de dos paquetes sonda por cada pareja de direcciones IP, definiendo las 4 direcciones aunque en el paquete de respuesta sólo se rellenen 2 o 3. Esto se hace para poder diferenciar el caso en que una de las direcciones IP esté añadiendo su marca temporal porque se encuentra en el camino a la máquina final que resulta indiferenciable de un alias cuando sólo se rellenan dos de las cuatro marcas temporales. Las direcciones de la opción se conforman de manera que la primera dirección IP a la que se pide la marca temporal sea la dirección destino y manteniendo que las dos direcciones IP sigan apareciendo de forma alternada en el campo de opciones. De esta forma, si un router sólo rellena 2 de las 4 marcas temporales, al enviar los 2 paquetes sonda, cada uno con un destino diferente, se obtienen respuestas que contendrán las marcas temporales de las dos direcciones IP. En el caso en que una de las direcciones IP pertenezca al camino, sólo en uno de los paquetes de respuesta se verán reflejadas las marcas

temporales de ambas direcciones IP. Esto es así porque cuando se envían paquetes directos hacia ambas direcciones, en una de ellas no se pasará por ambos routers porque uno de ellos esta antes que el otro, haciendo que no se rellene la marca temporal de uno de ellos.

Esta técnica permite detectar las parejas de direcciones IP que pertenecen al mismo router (resultado positivo) en el caso que un paquete sonda reciba una respuesta con las 4 marcas temporales rellenadas con el mismo valor y en el caso de recibir respuesta a 2 paquetes sonda y contengan 2 o 3 marcas temporales con el mismo valor.

Mediante esta técnica se puede detectar las parejas que no pertenecen al mismo router cuando los valores de las marcas temporales de dos direcciones IP difieran entre si. Esto sólo puede ocurrir si las dos direcciones IP comparten camino (resultado negativo). En caso de que no compartan camino, una de las direcciones IP no rellenará la marca temporal y no se podrá evaluar por lo que el resultado será no concluyente.

La técnica ofrece alta precisión ya que no hay manera de que dos routers diferentes den el mismo valor para las marcas de tiempo tal y como está pensada la petición de las 4 marcas. Al estar alternadas las marcas de tiempo, se imposibilita que un router intermedio pueda con motivo de una mala sincronización confundir a la técnica de resolución. Por otro lado, dos direcciones IP pertenecientes al mismo router nunca podrán diferir en sus marcas temporales por lo que tampoco hay confusión posible a la hora de dar las parejas de direcciones IP que no pertenecen al mismo router. La eficiencia de esta técnica es media ya que requiere del envío de sólo dos paquetes por cada pareja de direcciones IP, pero al tener que repartirse en parejas de direcciones IP el coste final es de orden cuadrático. La completitud es baja ya que existe un porcentaje reducido de routers que contestan a los paquetes rellenando las marcas temporales y porque la detección de parejas no pertenecientes al mismo router (resultados negativos) requiere que ambas direcciones IP se encuentren en el mismo camino. Para terminar, la distribuibilidad de esta técnica es alta ya que se puede repartir directamente el total de parejas que se desee identificar entre los equipos sonda de los que se dispone.

2.5.5 Velocity modeling

Velocity modeling es la técnica que se aplica en la herramienta Radargun [30]. Se trata de una técnica activa basada en paquetes sonda UDP y TCP directos. Utiliza el campo de IPID de los paquetes de respuesta de ICMP de error de tipo *Port Unreachable* y de paquetes de respuesta TCP con el flag de *Reset* habilitado. Se trata de una técnica centralizada con una agregación basada en grupos.

La técnica usa un número de medidas proporcional al número de direcciones IP que se quiere identificar (coste de orden lineal O(N)). La técnica se basa en el estudio del crecimiento de los campos de IPID de los paquetes de respuesta. En lugar de usar una estrategia basada en parejas como hace Ally [4], esta técnica realiza el envío de 200 paquetes sonda por dirección IP en rondas en las que se envía un paquete sonda a cada dirección IP por ronda.

Tras haber finalizado el proceso de medida, la estrategia de identificación pasa por la superposición de las distintas rectas que se pueden conformar para cada dirección IP con los valores de los campos de IPID y la marca temporal de recepción de cada paquete de respuesta. Se toman las rectas creadas a partir de las respuestas de las 2 direcciones IP que se desea identificar y se calcula el error cometido entre el valor del campo de IPID de una de las direcciones IP en un momento dado y el valor estimado que debería tener el campo de IPID de la dirección IP con la que se está comparando. El valor estimado que debería tener el campo de IPID para la segunda dirección se estima de la recta extrapolada entre los valores de IPID de la segunda dirección IP inmediatamente anterior y posterior del instante de tiempo del IPID de la primera dirección IP.

En la figura 2.8 se puede observar las rectas conformadas por los valores de IPID de 3 direcciones IP (direcciones IP1,IP2 y IP3). En ella se muestran los errores cometidos por la dirección IP1 respecto a las direcciones IP2 y IP3. Cada línea vertical que parte desde un valor de IPID de la dirección IP1 corta en el valor que tendría el campo de IPID de la otra dirección IP en ese instante de tiempo. El proceso se aplica a todos los valores de IPID de las direcciones IP que se desean identificar, pudiendo obtener una medida del error entre las rectas de dos direcciones IP aplicando la media a la suma de todos los errores.

La forma de identificar que dos direcciones IP pertenecen al mismo router (resultado positivo) es que la media del error cometido entre las rectas descritas con los valores de IPID de cada una sea inferior a 500. En caso de que el error cometido tenga un valor mayor a 1000 se considerará que las dos direcciones IP pertenecen a routers distintos (resultado negativo). Ambos rangos de error se han obtenido mediante ensayos experimentales calculándolos de forma que se minimizase el número de fallos de identificación (falsos positivos y falsos negativos) respecto de parejas obtenidas mediante el método Mercator [30].

La técnica ofrece una precisión media, ya que debido a restricciones temporales y de ancho de banda, a medida que se aumenta el número de direcciones IP a identificar aumenta el tiempo entre medidas del mismo equipo ocasionando que cada vez haya más tiempo y valores de IPID en los que se pueden confundir respuestas de direcciones IP pertenecientes a otros routers (falsos positivos). Los routers que tienen un crecimiento rápido de valores en el campo IPID ocasionan más reseteos del campo IPID haciendo que si hay mucho tiempo entre ellos, se identifiquen dirección IP como no pertenecientes al router que realmente sí lo son (falsos negativos). La eficiencia de esta técnica es alta ya que tiene un coste lineal con el número de direcciones IP. Por cada dirección IP se requiere el envío de 200 paquetes sonda que es una cantidad perfectamente asumible. La distribuibilidad es baja ya que la medida es centralizada. La completitud es también baja ya que hay rangos bastante amplios de error (desde 500 a 1000 valores de IPID) en los que el resultado es no concluyente. Además, debido a políticas relacionadas con seguridad en los routers, muchas direcciones no contestan a los paquetes sonda enviados con este proceso.

2.5.6 **Sidecar**

La técnica Sidecar [39] es una técnica activa que utiliza paquetes de tipo ICMP, UDP y TCP indirectos con la opción de registro de ruta activada. La técnica tiene una agregación por grupo y realiza las medidas de forma distribuida. Esta técnica permite la unificación de fases, ya que mientras se va haciendo el traceroute se recogen los paquetes de respuesta para examinar la opción de registro de ruta.

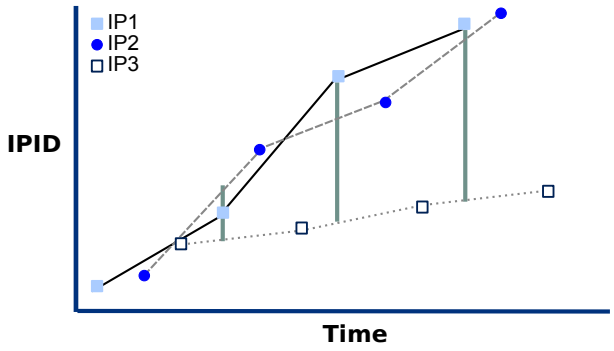


Figura 2.8: Ejemplo de proceso para cálculo de alias en la técnica *velocity modeling*

Esta técnica utiliza la información de la opción de registro de ruta que se puede obtener del paquete de respuesta de ICMP de error de tiempo de vida excedido. Dicho paquete contiene una copia del paquete sonda original de cuya opción registro de ruta se sacan las direcciones IP de los interfaces de salida de los routers del camino. La obtención de las direcciones IP de las interfaces de entrada de cada router se realiza mediante el campo IP origen del paquete de respuesta de cada uno de los paquetes sonda enviados utilizando una estrategia análoga a la realizada en la técnica de traceroute. La técnica alinea los datos de las direcciones IP de las interfaces de entrada con las direcciones IP de los interfaces de salida. En la figura 2.9 se puede observar un ejemplo de interfaces de entrada y salida de routers por los que pasa un paquete sonda enviado desde el equipo sonda con dirección IP7 hacia la maquina destino IP8.

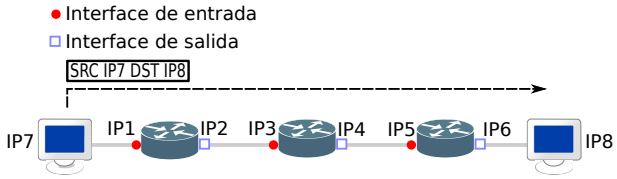


Figura 2.9: Interfaces de entrada y salida de los routers para un paquete enviado desde un equipo sonda.

La forma de corroborar que dos direcciones IP pertenecen al mismo router es examinar las direcciones que aparecen en el campo de registro de ruta de ambas respuestas con el objetivo de comprobar que los dos paquetes han seguido el mismo camino.

En esta técnica, el envío de paquetes sonda de tipo TCP se realiza usando la misma conexión. En numerosas ocasiones, debido a la configuración de los routers, los paquetes que forman parte de la misma conexión TCP no sufren balanceo de carga por lo que los caminos utilizando este tipo de paquetes son más estables. Esto permite cumplir el requisito de que los distintos paquetes sonda mantengan el camino y posibilita la detección de un mayor número de alias al emplear TCP como paquete sonda.

La inserción de la dirección IP de salida de cada router en el campo de registro de ruta no es una regla de obligado cumplimiento: se puede no introducir o introducir otra dirección IP de una interfaz perteneciente al mismo router. Por tanto, puede haber problemas a la hora de realizar el alineamiento de direcciones de entrada (obtenidas mediante la estrategia traceroute) y salida (obtenidas mediante la opción registro de ruta). El alineamiento requiere que cada uno de los paquetes enviados en el traceroute con la opción de registro de ruta activada sigan el mismo camino y que además, las direcciones IP obtenidas mediante el registro de ruta se correspondan con interfaces de salida de los routers y las obtenidas mediante el número de salto del traceroute sean las de entrada. Como este requisito no puede asegurarse en todos los casos, la técnica puede dar lugar a falsos positivos.

La precisión de esta técnica es de nivel medio, debido a que numerosas implementaciones de routers que no siguen la regla habitual ocasionan errores en la alineación de las direcciones IP y por tanto falsos positivos. Por otro lado la eficiencia conseguida es alta. De hecho, esta técnica puede hacerse directamente en la fase de descubrimiento habilitando la opción de registro de ruta y pos-procesando después los datos recogidos de los paquetes de respuesta. Por lo tanto, no se envía tráfico adicional a la red para realizar el proceso de identificación. La completitud de esta técnica es baja debido a que no permite la detección de routers situados en diferentes caminos, a filtrados o no contestaciones de los paquetes con el registro de ruta activado y al límite en el número de saltos que puede almacenar la opción de registro de ruta. La distribuibilidad de esta técnica es alta ya que los traceroutes se lanzan desde muchos nodos sonda para poder realizar un descubrimiento amplio de la red. También el procesado de las respuestas de esta técnica es distribuible, pudiéndose repartir los paquetes obtenidos como respuesta por cada traceroute entre los nodos sonda que se disponga para realizar la identificación de los alias.

2.5.7 Analitical Alias Resolver

La técnica *Analitical Alias Resolver* (AAR) [23], utiliza una estrategia de inferencia para el cálculo de los alias por lo que sólo utiliza la información obtenida a través de la fase de descubrimiento y por tanto no aplica en este caso el hablar de las estrategias específicas de la fase de resolución de alias.

La técnica se basa en el hecho de que existen redes, tanto de interconexión del núcleo de Internet como de acceso, que consisten en enlaces punto a punto entre routers. Como se trata de redes IP, y en los enlaces punto a punto sólo aparecen dos direcciones IP, los operadores de red suelen definirlas en subredes de máscara /30 o /31. Por esto, mediante la observación de las distintas rutas y direccionamiento de las interfaces de los distintos routers que se ven reflejados en los datos de la fase de descubrimiento, se pueden inferir los enlaces punto a punto y después se usa el dato de cada salto anterior al enlace encontrado para asociarlo con las direcciones pertenecientes al router del enlace.

En la figura 2.10 se puede observar un ejemplo de funcionamiento de la técnica. En ella se ve la información de direcciones IP y enlaces. En la figura existen dos direcciones IP que por el valor de sus direcciones IP pueden pertenecer al mismo enlace punto a punto, 10.0.0.1 y 10.0.0.2. Si ambas pertenecen al mismo enlace, las direcciones IP anteriores a dichas direcciones IP tienen que pertenecer también a los routers vecinos de éste por lo que la dirección IP2 pertenece al mismo router que 10.0.0.1 e IP6 al mismo router que 10.0.0.2.

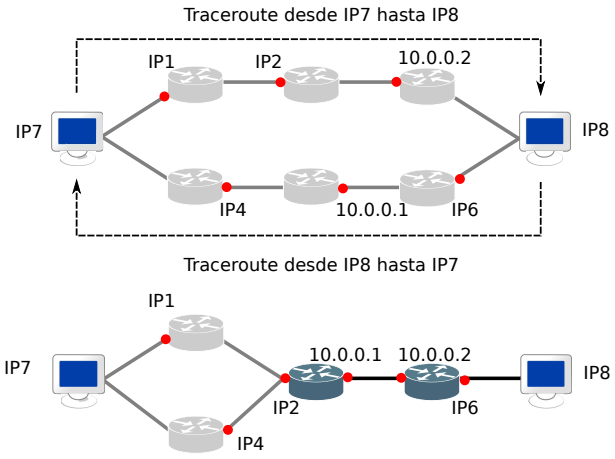


Figura 2.10: Ejemplo de método de inferencia AAR

El problema de esta técnica es que la máscara de la subred se está suponiendo como de punto a punto y no hay forma de verificarla. Además, se está suponiendo que los enlaces provenientes de la fase de descubrimiento son enlaces reales, pero puede haber errores por la existencia de balanceos de carga. Estos problemas hacen que la precisión del método sea media y no siempre los alias detectados formen parte del mismo router. La eficiencia del método es alta ya que no requiere de ningún tipo de medida adicional para realizarse. La distribuibilidad es baja ya que el proceso tal y como se define en el trabajo original no se distribuye [23], no obstante al ser un método de inferencia y no tener un gasto de ancho de banda no le afecta esta métrica de la misma forma que a las demás técnicas de resolución de alias. La completitud de esta técnica es baja también debido a que sólo detecta alias pertenecientes a enlaces de tipo punto a punto.

2.5.8 Analitical and Probe-based Alias Resolver

La técnica de resolución de alias *Analitical and Probe-based Alias Resolver* (APAR) [18] aparece como una mejora al método AAR [23], el cual utiliza solamente estrategias de inferencia sin el envío de ningún paquete sonda. APAR es una técnica mixta que se basa en el envío de paquetes sonda directos de ICMP de tipo *Echo Request*. Como parámetro base para la resolución utiliza el campo TTL y no permite la unificación de fases. La técnica es de carácter distribuible bajo unas condiciones específicas y posibilita la resolución en grupos de direcciones IP.

APAR realiza un análisis de las direcciones IP obtenidas a través de la fase de descubrimiento agrupando las direcciones IP que conformen redes punto a punto o multipunto con el fin de detectar los enlaces que posibiliten una identificación de los alias. Esto se consigue mediante la observación de las direcciones IP y agrupando aquellas que puedan formar parte de subredes con máscara desde /22 hasta /31.

Dado que pueden existir colisiones entre las pertenencias a las distintas redes, se utiliza una serie de condiciones para evaluar con que máscara se deben catalogar las redes a las que pertenecen los enlaces que se pretenden detectar. La primera es la condición de precisión, que invalida aquellas máscaras de red que hacen que aparezcan ciclos en los caminos obtenidos por el traceroute de las que

se han obtenido, es decir, que en un mismo camino no puede aparecer direcciones IP pertenecientes a la misma red más de una vez. La segunda es la llamada condición de completitud, que en este caso se define como el porcentaje total de las direcciones IP de esa red que se pueden observar en los caminos obtenidos por el traceroute. Hay que tener en cuenta que para el caso de las redes con máscara /30 o /31 la completitud siempre será 100 % ya que sólo se componen de una sola pareja de direcciones IP. La tercera condición se llama orden de procesado, y mediante ésta se elige la máscara que contenga mayor número de direcciones en el caso de que existan dos o más máscaras con el mismo porcentaje de completitud. La cuarta vuelve a tener relación con los ciclos e invalida cualquier máscara que produzca que en cualquiera de los caminos de los traceroutes obtenidos en la fase de descubrimiento se observen ciclos. En este caso hay que tener en cuenta que mediante las máscaras se infieren alias y dichos alias pueden ocasionar ciclos adicionales.

Una vez obtenidas las máscaras de las distintas subredes y con ellas los enlaces en las distintas trazas, se procede de una forma similar a la usada en la técnica AAR para la obtención de los alias mediante las direcciones IP de los routers predecesores y sucesores de los enlaces encontrados. En este caso se usan dos salvedades a la hora de catalogar un alias como válido. La primera, es que las dos direcciones IP que se quieren catalogar como alias tengan un predecesor o sucesor común, ya sea por tener en distintas rutas una misma dirección IP, por tener distintos alias de un mismo router o que las dos direcciones IP formen parte de dos subredes conectadas a un mismo router. La segunda característica que deben cumplir las direcciones IP a catalogar como alias es que se mantenga el valor TTL del paquete de respuesta a un paquete de ICMP de tipo *Echo Request* enviado como paquete sonda desde una misma maquina sonda.

En la figura 2.11 se puede observar la derivación de un alias a partir de haber encontrado los enlaces punto a punto que unen dos interfaces de un mismo router. Se supone en este ejemplo que las demás restricciones referentes a las máscaras se cumplen, y que se ha llegado a la conclusión de que existe un enlace entre las direcciones 10.0.0.1 y 10.0.0.2, y otro entre las direcciones 11.0.0.1 y 11.0.0.2. Como ambos enlaces se encuentran en sendos interfaces del mismo router, el alias de las direcciones 10.0.0.1 y 11.0.0.1 se calificará como válido siempre que se cumpla la condición adicional del TTL.

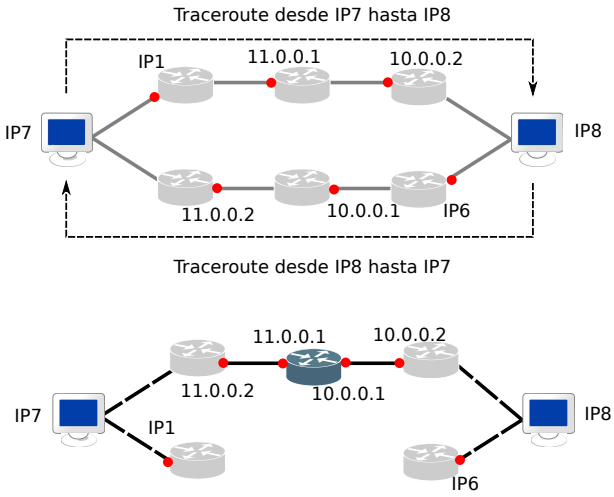


Figura 2.11: Ejemplo de obtención de posibles aliases en APAR

Esta técnica de resolución tiene una precisión media ya que mejora en cierta manera la precisión dada por la técnica AAR mediante el envío del paquete sonda a pesar de que las subredes seleccionadas sean mayores y esto en un principio pueda introducir un error mayor. Tiene una eficiencia alta dado que el único tráfico enviado se realiza solamente a las direcciones IP pre-identificadas como alias a modo de comprobación, y se requiere el envío de un sólo paquete. La completitud es media, ya que mejora la completitud dada por la técnica AAR pero no permite identificar direcciones IP pertenecientes al mismo router a las que se llegue a través de un camino diferente, y tampoco permite detectar aquellas direcciones IP de routers que formen parte de cruces entre caminos distintos. Por último, la distribuibilidad es media ya que el proceso inicial de inferencia no se realiza de forma distribuida, pero el envío de los paquetes sonda para la comprobación de los alias elegidos puede distribuirse entre varios equipos sonda. La única restricción que debe cumplir esta distribución de medidas es que todos los paquetes sonda que se envíen a las direcciones IP pertenecientes a un mismo router deben enviarse desde el mismo equipo sonda.

2.5.9 Tracenet

Tracenet [40] es una técnica mixta que utiliza paquetes sonda ICMP directos de tipo *Echo Request* y paquetes indirectos de tipo UDP a un puerto en desuso. Utiliza el

TTL y la dirección IP origen como parámetros base. Es una técnica con agregación a nivel de grupo que permite la unificación de fases y se utiliza de forma distribuida.

La técnica se basa en la inferencia de redes de una forma similar a la que se hace para derivar enlaces y alias en la técnica AAR, pero con la salvedad de que en este caso se realiza una búsqueda activa de los enlaces mediante el envío de tráfico con TTL limitado. A medida que se reciben respuestas desde cada dirección IP por parte de cada uno de los saltos del traceroute, se realizan envíos de tráfico ICMP hacia las direcciones IP que podrían formar parte de la misma subred. El proceso se inicia mediante la elección de una subred punto a punto /31 que se va haciendo más grande hasta que se corresponda con el tamaño de la red de la que realmente forma parte la dirección IP obtenida por el traceroute.

La máscara correcta se deduce a partir de envíos de *Echo Request* a todas las direcciones IP que componen la red. Si más de la mitad de las direcciones IP que forman parte de la subred definida contestan y cumplen una serie de heurísticos, se realiza el proceso de comprobar mediante heurísticos si realmente la red a la que pertenece es una red más grande (proceso de crecimiento). En cuanto se obtiene algún paquete de ICMP de error de tiempo excedido en tránsito se paraliza el proceso de crecimiento de la red y se vuelve a reducir el tamaño de ésta al tamaño que tenía con anterioridad, de manera que la dirección IP que ha devuelto el error no forme parte de la red. De manera adicional, también se para el proceso de crecimiento de las redes en cuanto el número de direcciones IP que contestan es menos que la mitad del número total de direcciones IP que abarca la red que se está intentado descubrir.

Hay una serie de heurísticos que permiten saber si una dirección pertenece o no a la red de la que se pretende obtener la mascara. El primero, verifica que una dirección IP pertenece a la subred limitando el campo de TTL del paquete sonda al número de salto en el que se encuentra la red en estudio y que no provoque un error de ICMP de tiempo excedido en tránsito. Es el heurístico que se utiliza para limitar el proceso de crecimiento de la red.

Para verificar que no se toman direcciones IP que estén a la misma distancia pero que pertenezcan a otras redes con dirección similar, se verifica que las direcciones IP que dan respuesta para el salto anterior son las direcciones conectadas punto a punto con los routers que no dan respuesta a éstas. Se puede ver en la

figura 2.12 un ejemplo en el que los paquetes sonda enviados con el valor N en el campo de TTL son contestados por todas las direcciones IP de la red 2.12(a), pero al reducir en una unidad dicho valor sólo lo harán las direcciones de un lado de los enlaces punto a punto 2.12(b). De esta manera se consigue desechar todas las direcciones IP cuyos dos extremos no se encuentren en la red definida.

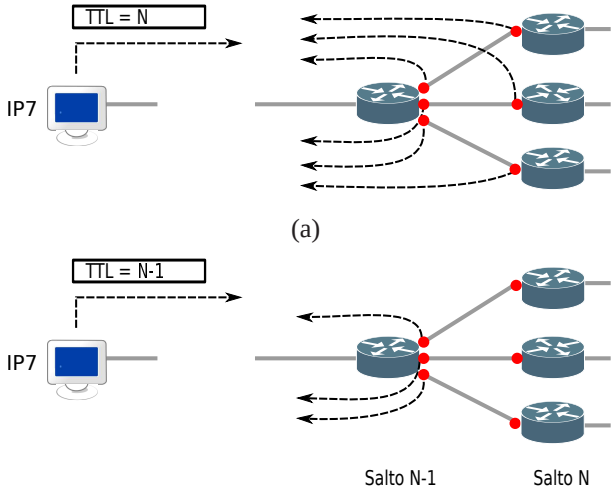


Figura 2.12: Ejemplo de comprobación de contestaciones de las interfaces punto a punto en el proceso de verificación de Tracenet

Se hace otra verificación de los extremos que quedan del proceso anterior y se les envía un paquete sonda con el valor de TTL limitado a dos saltos menos que la red en estudio. Las direcciones IP que pertenezcan a la red no deben contestar a estos.

La técnica restringe que la entrada a la red en estudio deba hacerse por los mismos interfaces, por este motivo se comprueba que el salto anterior para todas las direcciones IP pertenecientes a la red utiliza la misma dirección IP o alias detectados para ella.

Una vez obtenida toda la información de las redes bajo estudio de cada salto, mediante la dirección IP del siguiente salto y las direcciones IP de las interfaces de las subredes, se procede a un alineamiento de las direcciones IP que forman parte del mismo router. Con la máscara de red y la dirección IP de la interfaz de entrada de un router se deduce la dirección IP que debe tener la interfaz al que

está conectado, pudiendo asociarse éste último como alias de la respuesta del salto anterior. Este proceso es similar a la que se realizaba en el proceso de AAR, pero ahora se pueden detectar enlaces que forman parte de un enlace multipunto en lugar de limitarse a los enlaces punto a punto.

La técnica tiene una precisión alta, dado que los distintos heurísticos evitan la selección de enlaces erróneos haciendo que la evaluación final de los alias que derivan de ellos sea más precisa. La completitud es baja, ya que esta técnica permite detectar sólo los alias que pertenecen a un mismo camino. La eficiencia es media ya que requiere de un proceso de exploración de la red hacia todas las direcciones IP que pertenecen a las distintas redes del camino que se van infiriendo. Se trata de un proceso que requiere un paquete por cada una de las direcciones IP de la red y esto no debería suponer un coste importante. La técnica de Tracenet necesita realizarse desde varios equipos sonda ya que actúa como un traceroute modificado. El proceso que se encarga de las medidas en cada uno de los caminos acopla las tareas de descubrimiento y de resolución de alias. Por este motivo, las resoluciones de alias de todas las direcciones IP de un camino dado deben realizarse en el mismo equipo sonda por lo que la distribuibilidad de esta técnica es media.

2.5.10 **PalmTree**

La técnica Palmtree [41] usa una estrategia similar al Tracenet. Esta técnica utiliza paquetes ICMP de tipo *Echo Request* indirectos y paquetes de tipo UDP directos. Además la técnica es centralizada y realiza la resolución de alias por grupos.

La técnica se basa en encontrar los enlaces punto a punto entre dos direcciones IP para poder asociar las direcciones IP que pertenecen al mismo router. Por cada una de las direcciones IP de las que se quieren obtener los alias se realiza el envío de un paquete de tipo UDP a un puerto en desuso. En caso de que la dirección IP conteste, se obtiene el TTL del paquete UDP original guardado en la copia que hace el paquete ICMP de error enviado por el equipo destino. Mediante dicho valor de TTL se sabe a qué distancia está la dirección IP de la que queremos obtener el alias. Se procede al envío de un paquete sonda que contenga como valor de TTL el número de saltos que hay hasta la dirección IP en estudio, pero se envía a la pareja que tendría esa dirección si formase parte de una red punto a punto /30 o /31. Si la

contestación que se obtiene de ese paquete sonda es un paquete de ICMP de error de tiempo de vida excedido en tránsito, se toma la dirección origen de la cabecera IP de dicho paquete como alias. En caso contrario el resultado será no concluyente.

La precisión de esta técnica es de nivel medio, ya que cualquier cambio de ruta que suponga recorrer un camino con más routers intermedios provoca que Palmtree obtenga un falso positivo. En el mismo artículo en el que se describe la técnica se dan tasas de error de falsos positivos en el orden de un 0,01 % de los alias reconocidos [41]. La completitud de esta técnica es baja, ya que no permite la detección de parejas falsas y además requiere que para la detección del alias de una dirección IP, el equipo sonda esté situado más cerca que su pareja del enlace punto a punto. La eficiencia de esta técnica es alta ya que con un máximo de tres paquetes por dirección IP es capaz de realizar el proceso de identificación. Por último, la distribuibilidad de esta técnica es alta, ya que la única información que requiere la técnica para llevarse a cabo es la dirección IP de la que se quiere descubrir un alias, por lo que las diferentes direcciones IP tomadas en la fase de descubrimiento se pueden repartir entre los diferentes nodos sonda que se dispongan.

2.5.11 **Midar**

Midar [25] es una técnica de identificación activa que utiliza paquetes sonda de tipo ICMP, UDP y TCP directos e ICMP indirectos. El parámetro base utilizado para la resolución es el IPID. Es una técnica que tiene una agregación por grupos y que permite su distribución.

Esta técnica de identificación proviene de la técnica Radargun [42]. Radargun adolece de una tasa alta de falsos negativos y positivos. Midar trata de mejorar sus resultados.

Midar se basa en la monotonía y velocidad del incremento de los contadores de IPID de los distintos routers que siguen un patrón incremental. La principal premisa de esta técnica es que toda dirección IP que comparta contador de IPID, y que por tanto pertenezca al mismo router, debe describir siempre un patrón creciente en la generación de sus IPIDs.

Para comprobar que los valores de los campos de IPID de dos direcciones IP distintas son monótonos o no, en otras técnicas se utiliza la recta creciente formada

por los valores de IPID de la primera y se observa si al añadirle los valores de los campos de IPID de la segunda sigue describiendo dicha recta creciente. Midar utiliza una estrategia diferente que tiene en cuenta posibles retardos en el camino que pueden hacer que la respuesta de un paquete sonda enviado más tarde que otro pueda llegar antes. Para ello se marca tanto la salida de cada paquete sonda como la llegada de su respuesta. Con ese par de datos se puede acotar un segmento que determina entre qué tiempos se ha podido generar el IPID. El segmento asocia a un valor de IPID fijo el tiempo de envío del paquete sonda y el tiempo de recepción del paquete respuesta. Además, se forman áreas entre las medidas consecutivas a la misma dirección IP. El área se forma con el valor del tiempo de envío del primer paquete sonda enviado y el IPID contenido en el paquete de respuesta, y el tiempo de recepción de la respuesta del segundo paquete sonda enviado y el valor del campo de IPID contenido en dicha respuesta. Las distintas áreas que se pueden forman con los valores de IPID de una misma dirección IP se utilizan para verificar si los valores de IPID pertenecientes a otras direcciones IP pertenecen también al mismo router. Para ello se debe cumplir que los segmentos descritos con cada uno de los IPIDs de otra dirección IP intersequen con alguna de las áreas de la dirección IP original.

En la figura 2.13 se observan distintos valores de IPID recogidos en el tiempo. Aquellos que están unidos por rectángulos son los pertenecientes a la misma dirección IP, y como vemos cada IPID en lugar de ocupar un punto exacto en el eje x, se describe mediante un segmento que tiene como origen la marca temporal de envío del paquete sonda (Ti) y como punto final el tiempo de recepción de la respuesta a dicho paquete (Tf). El paquete de respuesta se ha podido generar en cualquier instante de tiempo dentro de dicho segmento.

En la figura, se pueden observar 3 casos de IPID, 2 corresponden con IPIDs que cumplen la condición de monotonía (a y b) y uno que no (c). En los 2 casos marcados como *possible alias* se puede observar como el segmento descrito por sus IPIDs interseca con el área que describe el rectángulo formado por dos de los valores de IPID de referencia. En el caso marcado como *not alias* se puede observar como el segmento no tiene ningún punto dentro del área descrita por los valores de IPID de la dirección IP con la que se quiere comparar, luego dicho IPID no cumple

la condición de monotonía exigida para que la dirección IP de la que se ha obtenido pueda ser considerada alias.

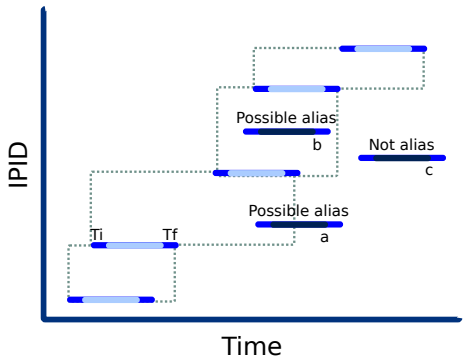


Figura 2.13: Ejemplo de IPIDs comparados con una serie de IPIDs para su verificación en Midar

La técnica no se basa únicamente en la condición de monotonía. Lo primero que evalúa esta técnica de resolución son las velocidades de los incrementos de IPID de cada dirección IP. Direcciones IP con velocidades muy dispares no se evalúan. Las velocidades se utilizan también para enviar paquetes sonda muy cercanos en aquellas direcciones IP que tienen velocidades altas y paquetes sonda menos cercanos para aquellas direcciones IP que no tienen velocidades tan altas. Para realizar esto se usa una estrategia que se denomina ventana deslizante. En la figura 2.14 se muestra un ejemplo en el que las líneas horizontales determinan a qué direcciones IP se va a realizar los envíos de las medidas. Las direcciones IP se ordenan por velocidades de mayor a menor (ver líneas verticales), y se empieza a realizar los envíos de paquetes sonda a las direcciones que requieren de mayor velocidad (IP1 y IP2). A cada ronda (ver eje vertical), la ventana se desplaza hacia la derecha, haciendo que ya no se envíen paquetes a las direcciones IP que requieren de más velocidad y añadiendo nuevas direcciones IP más lentas a las que realizar los envíos. Además, la ventana se aumenta permitiendo que abarque un mayor número de direcciones IP El proceso se itera hasta que se ha terminado el envío para todas las direcciones IP que tenemos en el subset.

La obtención de los valores de IPID se realiza mediante el uso de paquetes directos de tipo ICMP, UDP y TCP. Además usa paquetes ICMP de forma indirecta para obtener paquetes de error de tiempo excedido en tránsito. En este último tipo

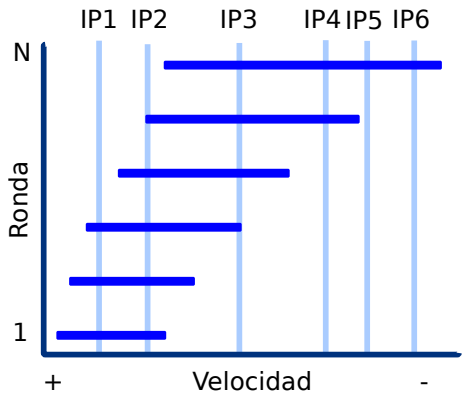


Figura 2.14: Ejemplo gráfico del método de ventana deslizante en Midar

los envíos se realizan a direcciones próximas a la IP de la que se desea realizar la medida con el fin de que ésta conteste.

Una vez obtenidos todos los datos, se somete a la prueba de monotonía a las direcciones IP cuyas medidas se puedan solapar y se declaran alias todas aquellas que la verifiquen.

La técnica Midar tiene una precisión alta, ya que las pruebas de monotonía con la inclusión de los offsets temporales hace que esta prueba no cometa apenas fallos. Además, la inclusión de la ventana deslizante permite que las direcciones IP que tienen crecimientos más rápidos se evalúen en grupos de direcciones IP más reducidos permitiendo enviar más paquetes a éstas en menor tiempo y de esta forma reducir los errores. La completitud de esta técnica es alta ya que permite la identificación de los alias con IPID incrementales y utiliza paquetes de varios tipos para permitir que un mayor número de direcciones IP contesten. La eficiencia es alta ya que permite descubrir los alias de todas las parejas de un conjunto de direcciones IP mediante algoritmos lineales. Y por último la técnica es distribuible ya que mediante el marcado temporal de los valores de IPID se permite hacer las medidas desde los nodos sonda de los que se disponga y después hacer el proceso de verificación de la monotonía en otro equipo.

2.5.12 **Discarte**

El método Discarte [20] se basa en el alineamiento de los valores que contienen los paquetes de respuesta en la opción de registro de ruta y los valores del campo

dirección origen utilizados en la estrategia de traceroute. Se trata de un método de inferencia ya que sólo usa los datos dados por los traceroutes previa activación de la opción de registro de ruta en los paquetes enviados. El método parte de que una vez conseguida la alineación de cada traceroute con su respectivo registro de ruta, se puede identificar fácilmente qué direcciones IP pertenecen al mismo router de manera parecida de la técnica Sidecar.

El alineamiento no resulta sencillo debido a la existencia de variedad de comportamientos de los routers a la hora de rellenar los campos de opción de registro de ruta en los paquetes que reenvían. En lo que respecta al campo de registro de ruta los comportamientos observados pueden dividirse en:

- Routers de tipo salida: que rellenan el campo de registro de ruta con la dirección IP de su interfaz de salida al reenviar el paquete.
- Routers de tipo no implementado: en los que no se rellena el campo de dirección IP en el registro de ruta.
- Routers de tipo MPLS: que son routers que se comportan como uno de tipo salida para las interfaces normales y como un no implementado cuando manda los paquetes por las interfaces conectadas vía MPLS.
- Routers de tipo llegada: que rellenan la opción de registro de ruta con la dirección IP de su interfaz de entrada cuando el paquete llega al router.
- Routers de tipo perezoso: que no decrementan el TTL en los paquetes con la opción registro de ruta habilitada y dejan pasar el paquete previamente rellenado con la opción de registro de ruta.
- Routers de tipo mixto: en los que para los paquetes que se reenvían actúan como un router de tipo salida pero si el TTL llega a cero en él, este se comporta como un router de tipo llegada.

Con todos estos tipos de comportamiento existen multitud de combinaciones posibles y sin tener información de qué tipo de router es cada uno, el realizar una alineación directa resulta prácticamente inviable. Esta técnica propone el uso de un sistema basado en reglas para solventar el problema. Este tipo de sistemas realiza

todo el árbol de posibles comportamientos con las reglas introducidas de cada uno de los routers para todos los traceroutes recogidos y realiza un proceso de evaluación para cada una de las topologías posibles en base a una serie de heurísticos. Cada topología obtiene una calificación en función de los heurísticos que cumpla y la que más puntuación obtenga será la utilizada para realizar el alineamiento de las direcciones IP obtenidas a través del campo de opciones de registro de ruta y las obtenidas a través de la estrategia de traceroute.

Los heurísticos utilizados para la calificación de las topologías son:

- El funcionamiento del registro de ruta debe ser consistente en todas las trazas. No puede aparecer un router que se comporte de distintas formas en distintos caminos.
- No deben aparecer bucles al mismo router.
- Para aquellos routers en los que se observe que disponen de enlaces multipunto, las direcciones IP de sus interfaces deben ser consecutivas.
- En el caso de que se disponga de información obtenida mediante la técnica Ally, se darán por ciertos los aliases obtenidos resultado de dicha técnica.
- Los routers de tipo perezoso no son muy frecuentes por lo que ante igual calificación, aquella topología que contenga menor número de routers de este tipo será la que se elija.

Para la calificación de cada topología generada mediante las reglas, el incumplimiento de un heurístico tendrá un peso en función del orden en el que se han citado los heurísticos con anterioridad, siendo el primer heurístico el más importante (por tanto el de mayor peso) y el último el que menor relevancia tiene (por tanto el que menor peso). La variación del valor de los pesos, mientras se mantenga el orden anterior, no aporta diferencias significativas en el resultado final de la topología obtenida ni, por tanto, de los aliases.

El proceso a seguir una vez realizadas las medidas es la de componer el sistema de reglas utilizando cada paquete de un salto dado con el siguiente salto. Se prepara una terna formada por el paquete de respuesta obtenido para el salto en la posición X, el paquete de respuesta del salto en la posición X+1 y la diferencia en el número

de direcciones IP que contienen sus campos de registro de ruta. A dicha diferencia se le denomina *alpha* y dependiendo de ésta se le añaden unas reglas u otras al sistema de reglas atendiendo a los posibles comportamientos de los routers para rellenar el registro de ruta comentados con anterioridad. De manera que si, por ejemplo, tenemos una terna [IP1, IP2, 1], implica que los posibles comportamientos que han tenido los routers en este caso particular han sido:

- IP1 y IP2 sean de tipo salida.
- IP1 y IP2 sean de tipo llegada.
- IP1 sea de tipo no implementado y IP2 sea de tipo llegada.
- IP1 sea de tipo salida y IP2 sea de tipo no implementado.
- IP1 sea de tipo no implementado, haya uno de tipo perezoso en medio y IP2 sea de tipo salida.
- IP1 sea de tipo no implementado, exista un router de tipo perezoso en medio y IP2 sea de tipo no implementado.

Hay que tener en cuenta que para distinto valor de *alpha* han de calcularse las distintas posibilidades de número y tipos de routers para la pareja de routers en estudio y así poder introducir estas posibilidades en el sistema de reglas.

Tras haber introducido todas las reglas se genera el árbol de posibles topologías y se les asigna una calificación según los heurísticos. La topología final se elige en función de la valoración dada por los heurísticos.

La técnica Discarte es una técnica de inferencia que ofrece una precisión media-baja ya que tiene una fiabilidad menor que la del Ally (recordemos que los autores eligen alias ofrecidos por Ally antes que los localizados mediante su técnica). Además todo el proceso se genera en función de meras suposiciones de cómo tendría que funcionar una red generada para ser óptima, pero no siempre tiene que ser así. La completitud de esta técnica es baja también ya que sólo permite alinear los alias que pertenezcan al mismo camino, y que contesten simultáneamente a paquetes con el registro de ruta habilitado y a los paquetes en los que se agote el TTL. La eficiencia de esta técnica es baja: a pesar de que únicamente requiere habilitar

la opción de registro de ruta en los paquetes sonda enviados para el descubrimiento a través del método traceroute, el proceso de cómputo de todas las reglas, el de todos los árboles posibles para las topologías y su calificación requiere de un coste computacional muy alto que resulta completamente inviable en cuanto se eleva el número de direcciones IP. Por último, la distribuibilidad es muy buena, ya que el sistema de reglas y de evaluación de topologías puede subdividirse, y de hecho se hace, en varios equipos de cómputo.

2.5.13 Técnica DNS

La técnica del DNS [24] se basa en la consulta de las bases de datos de DNS para obtener los nombres de host de cada una de las direcciones IP y en base a ellas poder identificar aquellas que pertenezcan al mismo router. Es una técnica activa y que utiliza paquetes de tipo UDP indirectos. El paquete UDP va dirigido hacia la base de datos del DNS y no hacia el equipo al que se desea realizar la prueba. Como parámetro base para la identificación se utiliza el payload del paquete DNS de respuesta.

La técnica se basa en que normalmente los operadores de red dan nombres descriptivos a los equipos que interconectan en la red, para permitir una fácil y rápida identificación de los equipos y saber qué equipos pueden o no estar interconectados entre sí. Un ejemplo de esto son los nombres de dos interfaces llamados sl-bb21-lon-14-0.sprintlink.net y sl-bb21-lon-8-0.sprintlink.net, en los cuales los códigos numéricos 14-0 y 8-0 parecen indicar el número de interface de una misma máquina. Por lo tanto, la técnica requiere primero de la realización de ingeniería inversa para descifrar la nomenclatura utilizada por un proveedor de servicios concreto. Cada proveedor tiene su propia forma de dar nombre a las distintas interfaces de sus equipos y ahí reside la complejidad. La técnica es totalmente dependiente de cómo de actualizados están las bases de datos de DNS respecto a la red en estudio. En algunos casos pueden darse situaciones en las que se cambien las direcciones IP de determinados equipos pero no se den de alta nuevos nombres que describan cual es su nueva situación.

La técnica ofrece niveles altos de precisión y eficiencia. Por ejemplo, en la comparación cruzada usando dos técnicas de resolución de alias en la red Planetlab no

devuelve ningún falso positivo ni falso negativo [24]. A nivel de eficiencia sólo exige el envío de un paquete por dirección IP hacia los servidores de DNS, de manera que es inmune a los filtrados por parte de los routers, y además no envía tráfico adicional hacia los routers de los que se quiere realizar una identificación. Por otro lado, la completitud es baja. Si se compara la técnica de DNS con la completitud del Ally, detecta solo entre un 45 % y un 70 % de las parejas de direcciones IP que detecta Ally. La distribuibilidad es alta, ya que en la fase de medida las peticiones a los DNS para obtener todos los nombres de dominio de cada dirección IP puede repartirse en varios nodos sonda. También la parte del procesado es fácilmente distribuible asignando a distintos equipos distintos subdominios. No obstante, el proceso de ingeniería inversa de los nombres asignados a las distintas interfaces de red no siempre es posible ya que un ISP no tiene por qué seguir una política de asignación de nombres concreta o que sea inferible para un usuario externo. Además, este proceso de ingeniería inversa es difícilmente automatizable.

2.6 Conclusiones

A lo largo de este capítulo se ha revisado las alternativas existentes para la fase de descubrimiento de direcciones IP y la fase de resolución de alias. Se ha identificado las estrategias utilizadas por las distintas técnicas de resolución de alias y las métricas para evaluar su rendimiento a la hora de afrontar la identificación de una red concreta. Además, se han asociado las técnicas con sus estrategias y valoración de métricas.

En las tablas 2.1 y 2.2 se presenta un resumen de las distintas estrategias y las valoraciones con las distintas métricas de las técnicas de resolución de alias.

No existe una técnica de resolución concreta ni un tipo de estrategia de las utilizadas en la fase de resolución de alias que sea claramente mejor. Las estrategias de carácter activo ofrecen normalmente una precisión y completitud mayores, pero requieren de introducir tráfico a la red y que las distintas direcciones IP en estudio respondan a los paquetes sonda que se les envían. Las estrategias de inferencia permiten realizar la identificación sin introducir tráfico adicional pero la precisión y la completitud son mejorables. Las estrategias basadas en estrategias de inferencia

Tipo de técnica	*Envío traf. sonda*	*Destino paq. sonda*	*Parámetro base*	*Tipo paquete sonda*	*Basada en unificación*	*Distribución medidas*	*Forma agregación*
Mercator	Activo	Directo	IP origen	UDP	No	Centralizada	Grupos
Ally	Activo	Directo	IPID	UDP	No	Distribuida	Parejas
Iplane	Activo	Directo	IPID, TTL	ICMP	No	Distribuida	Grupos
Prespec. Time.	Activo	Directo	Prespecified Time.	ICMP	No	Distribuida	Parejas
Radargun	Activo	Directo	IPID	UDP, TCP	No	Centralizada	Grupos
Sidecar	Activo	Indirecto	RR, IP origen	ICMP,UDP,TCP	Si	Distribuida	Grupos
AAR	Inferencia	-	-	-	-	-	-
APAR	Mixta	Directo	TTL	ICMP	No	Centralizada	Grupos
Tracenet	Mixta	Direct/indirect	IP,TTL	ICMP, UDP	Si	Distribuida	Grupos
Palmtree	Activa	Direct/indirect	IP, TTL	ICMP, UDP	No	Distribuida	Grupos
Midar	Activa	Direct/indirect	IPID	ICMP,UDP,TCP	No	Distribuida	Grupos
Discarte	Inferencia	-	-	-	-	-	-
DNS	Activa	Indirectos	Payload DNS	UDP	No	Distribuida	Grupos

Tabla 2.1: Estrategias utilizadas en las técnicas de resolución de alias

Tipo de técnica	*Precisión*	*Completitud*	*Eficiencia*	*Distribuibilidad*
Mercator	Alta	Baja	Alta	Baja
Ally	Media	Baja	Baja	Alta
Iplane	Baja	Baja	Alta	Alta
Prespecified Timestamps	Alta	Baja	Media	Alta
Radargun	Media	Baja	Alta	Baja
Sidecar	Media	Baja	Alta	Alta
AAR	Media	Baja	Alta	Baja
APAR	Media	Media	Alta	Media
Tracenet	Alta	Baja	Media	Media
Palmtree	Media	Baja	Alta	Alta
Midar	Alta	Alta	Alta	Alta
Discarte	Media-Baja	Baja	Baja	Alta
DNS	Alta	Baja	Alta	Alta

Tabla 2.2: Métricas para cada una de las técnicas de resolución de alias

pueden ser las idóneas en redes en las que por la razón que sea, no se pueda introducir tráfico adicional a la red o cuando los porcentajes de respuestas de la red en estudio sean muy bajos.

En lo que refiere a la cuestión de precisión y completitud, lo deseable es realizar una comparativa con porcentajes de errores que se cometen a la hora de realizar una técnica dada o, para el caso de la completitud, qué porcentaje del total de parejas se puede obtener. La evaluación de la precisión es dificultosa porque normalmente no se dispone de información de la red en estudio y de las que sí que se dispone de información suelen ser redes de características concretas que hacen que los resultados obtenidos no sean normalmente extrapolables al resto de redes.

En el caso concreto de la completitud existen dos problemas principales para no poder comparar las diferentes técnicas de resolución de alias. El primero es que no existe un estándar de medición de la completitud. En nuestro estudio definimos la completitud en función de que todas las parejas sean identificadas como pertenecientes o no a un router con la intención de saber cuando se ha realizado la resolución completa de la red, pero en otros trabajos se centran exclusivamente en los alias. En la comparativa realizada por CAIDA [34] la información de la red real se conoce ya que se han usado redes académicas y de investigación nacionales pero las únicas técnicas comparadas son Mercator y APAR. En otros trabajos, la completitud se mide mediante la comparación con las tasas de identificación obtenidas por la totalidad de las técnicas que estudian en él, por ejemplo en [24] se comparan el DNS, Ally y Mercator. Esto ocasiona que los porcentajes de completitud no se refieran al total de alias y no alias que tiene la red realmente. Dentro del tipo de estudios que utilizan un modelo comparativo, existe uno en que la completitud se basa en la contestación previa a otra técnica. En el trabajo de la técnica Radargun [30] la completitud se mide para el subset de direcciones para las que tenemos información de los alias gracias a la previa identificación mediante Mercator. El problema que conlleva esto es que se realizan las medidas de identificación al subgrupo de direcciones IP ya que dan una respuesta al Mercator lo que implica que el 100 % de estas direcciones va a responder a los paquetes sonda enviados por Radargun. Por último, el trabajo que presenta la técnica AAR [23] ofrece las tasas de identificación basándose en un subgrupo de parejas muy reducido (sólo 180

parejas) por lo que las medidas de completitud obtenidas en dicho trabajo son difícilmente extrapolables. Se tiene distintos datos que aportan los distintos trabajos pero no se puede hacer una valoración conjunta de cómo funcionan las distintas técnicas comparativamente, ni obtener los porcentajes exactos de completitud y de identificaciones fallidas de ellas debido a sus diferentes formatos. Por lo tanto es necesario un estudio comparativo de todas las técnicas de resolución basándose en las mismas métricas.

CAPÍTULO

3

Comportamientos de routers a medidas activas

3.1 Introducción

Se ha visto en capítulos anteriores cómo existen técnicas de resolución de alias que se basan en diferentes comportamientos de los routers a la hora de generar los paquetes de respuesta. La completitud de las diferentes técnicas está muy ligada tanto a las contestaciones de los distintos routers a los paquetes sonda como a los diferentes comportamientos de los routers a la hora de generar los distintos parámetros base. Se ha de tener en cuenta que a pesar de que para el tipo de paquete sonda utilizado por una técnica dada se obtenga una respuesta por parte del router, ésta puede no ser válida a la hora de realizar la resolución. Esto puede pasar por el incumplimiento de las características exigidas por la técnica para el parámetro base utilizado para la resolución. Las respuestas válidas son aquellas que por las características de su parámetro base y las exigencias requeridas por la técnica que se está empleando, puedan ser utilizadas para llevar a cabo la resolución de alias. Por ejemplo, en el caso de utilizar la técnica Midar sólo se pueden obtener los alias pertenecientes a los routers que contesten utilizando valores de IPID que sigan patrones crecientes. En Midar, aunque las tasas de contestación de las direcciones

IP sean un 100 %, si todas ellas se comportan dando valores aleatorios a los campos de IPID de los paquetes de respuesta (0 % de contestaciones válidas), la técnica no podrá dar ningún tipo de conclusión.

Los comportamientos de los routers a la hora de generar parámetros base son particulares de cada uno. A la hora de realizar una identificación completa con objeto de saber la completitud de una técnica dada se debe hacer medidas al conjunto de todas las posibles parejas de direcciones IP que se quieren identificar. El comportamiento de un router concreto no varía porque se le esté realizando un proceso de identificación contra una u otra dirección IP del mismo router. Por este motivo, un estudio de las contestaciones por parte de las direcciones IP a diferentes paquetes sonda y de los comportamientos de los routers a la hora de generar sus parámetros base ofrece información sobre los posibles niveles de completitud a la que se puede llegar. A partir de los distintos comportamientos de los routers, se puede saber qué técnicas son capaces de identificar parejas, los tipos de paquetes que es mejor utilizar para el proceso de identificación así como valorar el uso de una combinación de varias técnicas para realizar una mejor identificación.

Otra parte interesante que se obtiene de este estudio de las respuestas es el observar nuevos tipos de comportamientos que tienen los parámetros base y crear nuevas estrategias que permitan identificar alias a partir de ese tipo de comportamientos. Los porcentajes de respuestas válidas de un comportamiento determinado para un parámetro o parámetros base pueden incentivar la realización de un estudio más a fondo de dicho comportamiento para evaluar la viabilidad de una identificación basada en él.

Las técnicas que usan estrategias directas que ofrecen mayores tasas de identificación se basan en la utilización como parámetro base de 3 campos de la cabecera IP: el identificador de IP (IPID), la dirección IP origen y la marca temporal que se puede encontrar en los paquetes que tienen la opción de *Prespecified Timestamp* habilitada.

A lo largo de este capítulo se ha hecho un estudio experimental de los diferentes tipos de comportamiento de los routers a la hora de generar los parámetros base utilizados en las estrategias activas, así como los porcentajes de contestaciones totales y contestaciones válidas observadas para diferentes estrategias y paquetes sonda.

3.2 Escenario de medida

Para elaborar el estudio de los diferentes comportamientos de los routers en Internet se ha utilizado la plataforma de medida Planetlab [16]. Esta plataforma ofrece nodos en los que el usuario puede realizar experimentos accediendo a través de una terminal de comandos remota de tipo SSH. El total de nodos de los que dispone Planetlab es alrededor de 1030 nodos, pero debido a problemas ligados con equipos caídos, redes inaccesibles y problemas relacionados con la administración de los nodos a nivel local (los nodos sonda son administrados por las universidades y centros de investigación donde estos se emplazan), el número real de nodos accesibles para la ejecución de experimentos puede variar desde 100 a 500 normalmente. Esta plataforma se utiliza tanto para la fase de descubrimiento como para la fase de resolución de alias.

La fase de descubrimiento se ha basado en el uso de la herramienta Paris-traceroute. Esta herramienta se ejecuta desde 25 nodos de Planetlab que se corresponden con los nodos sonda. Se ha realizado el descubrimiento de las direcciones IP pertenecientes al camino desde cada uno de esos 25 nodos sonda hasta los 24 restantes. Cada camino se ha obtenido mediante Paris-traceroute utilizando paquetes de tipo ICMP, UDP y TCP lo que permite realizar una recogida de direcciones IP más amplia. Mediante este proceso se ha obtenido como resultado un subgrupo de direcciones IP perteneciente a la red de interconexión de Planetlab. En total, la red de interconexión obtenida se compone de 2037 direcciones IP. Dado que no hay información pública de la topología de la red, no se puede realizar una valoración exacta de la precisión de los métodos de resolución. El utilizar nodos sonda en redes académicas y de investigación nacionales, habría ofrecido la posibilidad de realizar una verificación en ese sentido, pero no se dispone en la plataforma Planetlab de suficiente número de nodos sonda alrededor de ninguna de las redes conocidas (GEANT, CaNet4 y GlobalNOC) para que sea posible realizar medidas de tipo indirecto.

El motivo de seleccionar sólo 25 de todos los nodos sonda disponibles para la realización de las pruebas de la fase de descubrimiento tiene como motivación el tener la posibilidad de poder ejecutar las medidas activas que se corresponden

con la fase de resolución de alias en un tiempo razonable. Tras la fase de descubrimiento se ha realizado la batería de medidas directas e indirectas para obtener los porcentajes de respuesta de las diferentes direcciones IP en estudio. Para las medidas directas se han utilizado los mismos nodos sonda que para la fase de descubrimiento. Para cada medida de este tipo solamente se requiere del envío de tráfico a una dirección IP, y con la intención de evitar problemas relacionados con filtrados a determinados nodos sonda, el envío se realiza desde varios nodos sonda simultáneamente. Para las pruebas de tipo indirecto el envío se realiza desde los mismos nodos sonda desde donde se obtuvieron las direcciones IP en la fase de descubrimiento. Las respuestas se obtienen con la herramienta Paris-traceroute utilizando los mismos destinos y repitiendo el envío 50 veces al mismo tiempo que se recogen los paquetes de respuesta enviados por los routers. Dichos paquetes de respuesta son los que se utilizan para el estudio de los parámetros base en medidas de tipo indirecto. En total, se han realizado pruebas directas con paquetes ICMP de tipo *Echo Request* y *Timestamp Request*, UDP y TCP con el flag de *SYN* activado, pruebas indirectas con paquetes ICMP de tipo *Echo Request*, UDP y TCP con el flag de *SYN* activado y por último pruebas directas e indirectas con paquetes sonda ICMP de tipo *Echo Request* con la opción de *Prespecified Timestamp* activada.

Por último, para la fase de resolución de alias se han utilizado un total de 200 nodos sonda del total que provee Planetlab para realizar el trabajo de cálculo. Las pruebas que tienen una agregación por grupos y no necesitan de su distribución, se ejecutaron desde una sola máquina sonda. Aquellas que tienen agregación por parejas, se distribuyeron entre el total de nodos sonda generando de esta forma una resolución más rápida.

En las siguientes secciones se realiza un estudio específico por cada parámetro base en el que identifica los diferentes comportamientos de los routers a la hora de generarlos, según el tipo y destino del paquete sonda. También se evalua la completitud que se puede conseguir con ellos. Para finalizar se realiza una comparativa de las diferentes técnicas de resolución.

3.3 Comportamientos del parámetro base IPID

El identificador IP (IPID) es el parámetro base utilizado por Ally [4], Radargun [30] y Midar [25]. Este tipo de técnicas se basan en los IPID de tipo incremental para realizar la resolución de alias, de manera que en el caso del Ally y Radargun pueden identificar parejas pertenecientes al mismo router mientras el router a identificar tenga comportamiento incremental a la hora de dar valores al campo de IPID y puede identificar parejas de direcciones IP que no pertenecen al mismo router mientras el comportamiento del IPID sea diferenciable. En el estado del arte se distinguen 2 tipos de comportamiento, incremental y aleatorio [25]. Por tanto, Ally y Radargun son capaces de identificar que dos direcciones IP pertenecen a distinto router (respuesta negativa) mientras los dos routers sigan un patrón incremental no sincronizado mientras exista uno de cada tipo (uno incremental y otro aleatorio). La técnica Midar sólo permite identificar direcciones IP pertenecientes al mismo router y no realiza la identificación de direcciones IP que no lo sean. Todas estas técnicas son tipo directo y la única técnica que ha introducido envío de paquetes sonda de tipo indirecto ha sido Midar, que ofrece la posibilidad de realizar la identificación mediante envíos de paquetes sonda ICMP de tipo indirecto.

El IPID es un parámetro base con gran importancia en la resolución de alias, que merece un estudio a fondo tanto de sus posibles comportamientos como del número de respuestas y validez de los mismos que ofrece para los distintos tipos de paquete sonda. Las técnicas basadas en este parámetro base son las que dan mejores resultados para la resolución de alias.

Se han hecho medidas a cada una de las direcciones IP del conjunto de 2037 direcciones obtenidas de la red de interconexión de Planetlab tomando el parámetro base IPID (usado por los métodos Ally,Radargun y Midar) se ha analizado los tipos de respuestas y el comportamiento de los routers que generan el parámetro base.

Después de la realización de las medidas se detectan 4 posibles comportamientos para el IPID:

- Cero: este tipo de comportamiento en los routers hace que el IPID para los paquetes de respuesta que envía tenga siempre el valor 0.

- Incremental: este tipo de comportamiento es el producido por los routers que utilizan un contador común a todos los interfaces del router para dar valor al IPID de los paquetes que envían. Cada vez que se envía un nuevo paquete el contador se incrementa en uno por lo que varios valores de IPID tomados de un mismo router siguen un patrón creciente.

- Aleatorio: este tipo de comportamiento es producido por aquellos routers que dan un valor al campo de IPID utilizando un generador de números aleatorios . Cada nuevo paquete que envía el router tiene un valor diferente en su campo de IPID.

- Copia: este tipo de comportamiento es producido por aquellos routers que realizan una copia del valor de IPID del paquete sonda que se le envía para dar el valor al campo IPID del paquete de respuesta.

Se ha de tener en cuenta que a pesar de haber detectado 4 tipos de comportamiento, estos sólo incluyen los paquetes que son respondidos por los routers. Existe también un porcentaje de routers que pueden no responder a los paquetes que se le envían y es un aspecto más a tener en cuenta para el estudio. Además, el mismo router para diferente estrategia de medida puede comportarse de forma diferente, de manera que se ha realizado el estudio diferenciando por tipo de estructura de protocolo usado (ICMP, UDP y TCP) y por tipo de estrategia de medida (directa o indirecta). En las tablas 3.1 y 3.2 se presentan los distintos comportamientos del campo IPID para paquetes sonda indirectos y directos respectivamente. En las tablas cada columna se corresponde con un comportamiento diferente de los routers para generar el parámetro base IPID menos las dos últimas que se corresponden con las direcciones IP que no han respondido (*unresponsive*) y con el total de contestaciones de las que se puede derivar una conclusión (*valid responses*). Esta última columna se ha generado por parejas a partir de la utilización del total de direcciones IP y sus diferentes comportamientos para el resto de columnas el resultado es por dirección IP. Para el cálculo de la última columna que tiene relación con la completitud que se puede llegar a obtener con las técnicas que utilicen este parámetro base, se han contabilizado todas las posibles combinaciones

de parejas de direcciones IP con un comportamiento del que se pueda dar un veredicto y se ha dividido entre el total de parejas posibles para obtener un porcentaje. Dicho porcentaje representa el número total de parejas a las que las técnicas que utilizan este parámetro base podrían identificar. Un estudio del crecimiento de los IPID encontrados en los paquetes de respuesta puede dar una valoración positiva (las dos direcciones IP pertenecen al mismo router) sólo si el comportamiento para el router al que pertenecen las dos direcciones IP es de tipo incremental. De forma análoga, el estudio del crecimiento de los IPID permite dar una valoración negativa (las dos direcciones IP pertenecen a distinto router) si los comportamientos de los routers a los que pertenece cada dirección IP es diferente (cero-incremental, cero-aleatorio, cero-copia, incremental-aleatorio, incremental-copia, aleatorio-copia o incremental-incremental no sincronizadas).

Type of probe packet	*Zero* (%)	*Incremental* (%)	*Random* (%)	*Copy* (%)	*Unresponsive* (%)	*Valid responsiveness(%)*
ICMP Echo	37.94	35.87	26.17	0	0	61.65
UDP	41.23	20.77	37.98	0	0	53.09
TCP	41.21	26.53	32.24	0	0	57.17

Tabla 3.1: Porcentajes de comportamientos de los routers a la hora de generar las respuestas según su identificador IP mediante prueba indirecta

Type of probe packet	*Zero* (%)	*Incremental* (%)	*Random* (%)	*Copy* (%)	*Unresponsive* (%)	*Valid responsiveness(%)*
ICMP Echo	0	48.40	13.59	35.00	2.99	70.51
ICMP Tstamp	0	25.92	6.67	16.54	50.85	18.74
UDP	0.78	0.04	0.29	6.23	92.63	0
TCP	3.04	33.08	63.81	0.04	0	55.26

Tabla 3.2: Porcentajes de comportamientos de los routers a la hora de generar las respuestas según su identificador IP mediante prueba directa

De los resultados se concluye que a pesar de que las medidas de tipo indirecto ofrecen una mayor tasa de respuesta, son las medidas de tipo directo las que ofrecen mayor tasa de contestaciones válidas. Se puede ver cómo el comportamiento de tipo cero es mucho mayor en los paquetes de tipo indirecto (con tasas cercanas a un 40 %), siendo prácticamente inexistente en el caso de utilizar paquetes sonda de tipo directo. Otro comportamiento específico de un tipo de medidas es el comportamiento de tipo copia, que sólo existe en el caso de utilizar medidas de tipo directo. Por último, en el caso de los paquetes sonda de tipo UDP se puede observar como la tasa de respuesta es muy reducida para los paquetes de tipo directo. Esto quiere decir que en muchos casos los routers ignoran el envío de respuestas ICMP de tipo puerto inalcanzable probablemente por razones ligadas con la seguridad.

3.4 Comportamientos del parámetro base *Prespecified Timestamp*

La técnica de resolución *Prespecified Timestamp* utiliza paquetes sonda ICMP directos de tipo *Echo Request* con la opción de *Prespecified Timestamp* habilitada. Para la completitud de dicha técnica, se ha evaluado sobre el conjunto de 2037 direcciones IP obtenidas en la fase de descubrimiento. Además, se han incorporado medidas realizadas mediante estrategias de tipo indirecto para evaluar que tal se comportarían las técnicas basadas en este parámetro sobre paquetes de contestación a paquetes sonda indirectos.

Para este parámetro base se han identificado 6 tipos de comportamiento diferentes:

- *n-tstamp*: Estos tipos de routers sólo rellenan de n igual 1 hasta n igual 4 (según el n que acompaña al nombre) de los huecos reservados para introducir la marca temporal a pesar de que las direcciones IP especificadas en el paquete sonda pertenezcan todas al mismo router.

- Siempre: Este tipo de routers rellenan todas las marcas temporales del paquete sonda, pertenezcan o no al mismo router.

- Ninguno: Este tipo de router no rellenan ninguna de las marcas temporales a pesar de que las direcciones especificadas en el campo de opciones pertenezcan al mismo.

Como ocurre en el caso anterior, se debe añadir el caso de que los paquetes sonda puedan ser no contestados por parte de los routers a los que se les envían las medidas. De todos los comportamientos, los únicos que pueden ofrecer una identificación en base a la estrategia utilizada en la técnica *Prespecified Timestamp* son los de los routers que introduzcan al menos 2 marcas temporales en el paquete de respuesta. En la tabla 3.3 se presentan los resultados de las medidas realizadas tanto con paquetes sonda directos como indirectos. Las columnas que se pueden ver en la tabla se corresponden con cada tipo de comportamiento de los routers a la hora de generar el parámetro base. Las 4 primeras se corresponden con routers que en su paquete de respuesta incluyen como máximo de 1 a 4 timestamps. Como en el caso de la tabla realizada para el caso del parámetro base IPID la columna de respuestas válidas se ha realizado por parejas aglutinando todas las combinaciones de parejas de direcciones IP que por el comportamiento del router al que pertenecen se puede derivar una conclusión.

Type of probe packet	*1-tstamp* (%)	*2-tstamp* (%)	*3-tstamp* (%)	*4-tstamp* (%)	*Always* (%)	*None* (%)	*Unresp.* (%)	*Valid resp.(%)*
ICMP Echo (direct)	9.51	6.08	0.09	36.70	1.47	0.09	45.94	42.87
ICMP Echo (indirect)	8.00	0.19	0.00	20.59	0.00	0.09	71.13	21.10

Tabla 3.3: Porcentajes de comportamientos de los routers a la hora de generar las respuestas a paquetes sonda con la opción de *Timestamp Prespecified* habilitada

Las técnicas indirectas ofrecen una tasa de no contestación (unresponsive) mayor que la ofrecida por el envío directo. Se puede observar un 71.13 % de no contestación por parte de los routers a los que se ha realizado el envío de paquetes indirectos frente a un 45.94 % en los que el envío de los paquetes sonda se ha realizado de forma directa. En este caso, las técnicas que utilizan paquetes sonda directos siguen dando una tasa de respuestas válidas mayor.

3.5 Comportamientos del parámetro base dirección IP origen

La técnica Mercator [29] utiliza como parámetro base para la identificación de alias el campo dirección IP origen de la cabecera IP. Esta técnica fue una de las primeras técnicas que permitieron la identificación de alias y sigue siendo ampliamente utilizada para este objetivo, pero a lo largo de los años tanto el comportamiento de los routers para la generación de este parámetro base como el filtrado por parte de los administradores de red ante los paquetes UDP enviados directamente a los routers ha reducido su efectividad. Para el estudio del comportamiento de los routers al rellenar la IP origen sólo se ha utilizado la información de las respuestas a los paquetes sonda directos de tipo UDP. Mediante la utilización de otros tipos de paquete no se puede obtener el tipo de respuestas que se necesitan para realizar la resolución. En el caso de realizar el envío de un paquete sonda ICMP ya sea de *Echo Request* como de *Timestamp Request*, o bien contesta la dirección destino del paquete sonda o el router simplemente no contesta. Para medidas directas de tipo TCP con el flag de *SYN* habilitado, los routers contestan con un paquete TCP de tipo *Reset* en lugar de con un paquete ICMP como en el caso de UDP por lo que tampoco es evaluado. Por otro lado, el empleo de estrategias indirectas para la resolución de alias mediante este parámetro base induce a demasiados errores ya que cualquier router que se encuentre a la misma distancia en saltos que otro, puede dar una respuesta a los paquetes sonda enviados de forma indirecta, por lo que se descarta la posibilidad de usar este tipo de estrategias. Las medidas realizadas para este parámetro base se presentan en la tabla 3.4 y se puede observar como el número de respuestas válidas es muy reducido. Esta tabla se compone de 4 columnas, la primera para las direcciones IP cuyos routers devuelven el paquete de respuesta desde la misma dirección a la que se le ha enviado el paquete, la segunda cuando la dirección desde la que se envía es diferente a la marcada en el paquete sonda, la tercera aglutina las direcciones IP que no dan una contestación a los paquetes sonda y las respuestas válidas son las parejas de direcciones IP a las que mediante técnicas basadas en la dirección IP origen se les puede dar un veredicto.

De los resultados se concluye que es una técnica que ofrece una completitud muy reducida. Dado que el porcentaje de routers que responden desde una direc-

Type of probe packet	*Same-interface* (%)	*Different-interface* (%)	*Unresponsive* (%)	*Valid responsiveness(%)*
UDP	0.09	7.26	92.63	0.5

Tabla 3.4: Porcentajes de comportamientos de los routers a la hora de generar las respuestas según su dirección origen

ción IP diferente es mayor que los porcentajes de routers que responden desde la misma dirección IP, se puede deducir que los routers con pocas interfaces (probablemente pertenecientes a routers del acceso de Internet) normalmente filtran este tipo de paquetes, mientras que los routers más grandes y por ello con mayor número de interfaces tienden a responder (probablemente routers pertenecientes al núcleo de Internet).

3.6 Completitud obtenida por las técnicas de resolución de alias

Se ha revisado en las secciones anteriores los diferentes porcentajes de respuestas válidas que es posible conseguir mediante la utilización de distintos parámetros base sometidos a diversos comportamientos de los routers. En dicho estudio se ha tenido en cuenta el tipo de paquete sonda utilizado (ICMP, UDP, TCP), la estrategia de envío (directo, indirecto) y los distintos comportamientos de generación de los parámetros base (IPID, IP origen y opción de *Prespecified Timestamp*). El porcentaje de respuestas válidas determina la máxima tasa de identificación alcanzable por las técnicas de resolución de alias que utilizan estrategias que permiten la identificación mediante determinados comportamientos de los routers a la hora de generar los parámetros base.

Una técnica puede no llegará a alcanzar los porcentajes de completitud obtenidos por las medidas de respuestas válidas en el estudio de un parámetro base debido a dos motivos principalmente. El primero está relacionado con la forma de realización de medidas si se está perdiendo información que resultaría útil para la resolución. El segundo esta relacionado con que la técnica no contemple la casuística de todos los posibles comportamientos de los routers respecto del parámetro base que utiliza y no pueda realizar una identificación o la haga de manera incorrecta. Un

ejemplo del primer motivo es el proceso de medida de la técnica Ally. La técnica tiene restricciones en cuanto a número de paquetes sonda (3 paquetes), la distancia temporal de los dos primeros paquetes (no se deja espacio entre ellos) y el offset máximo permitido para que los paquetes sean considerados alias (200 IPIDs). Las respuestas están limitadas por restricciones del router en el número de paquetes a responder a una dirección IP por segundo, porque el uso de sólo 3 IPIDs no es suficiente para mostrar la tendencia del router o porque por cuestiones de volumen de número de paquetes enviados por el router el offset máximo de 200 IPIDs pueda quedarse corto. El segundo motivo puede verse en el caso de la *Prespecified Timestamp* que no realiza una comprobación previa de los comportamientos de los diferentes routers esperando que los routers actúen siempre con un comportamiento de tipo *N-tstamp* o *unresponsive*. Para un router con tipo de comportamiento *always* una medida con dicha técnica implica la aparición de falsos positivos a la hora de realizar el proceso de resolución de alias, porque el router rellena todas las marcas temporales del campo de opciones aunque no sea su dirección IP la que aparezca para ser rellenada. En el paquete de respuesta se obtendrán marcas de tiempo de un mismo equipo por lo que la técnica catalogará las dos direcciones IP como pertenecientes al mismo router.

Por esta razón es necesario realizar la evaluación de los porcentajes de identificación para las técnicas de resolución de alias e identificar las técnicas que más se acercan a las cotas máximas de completitud.

La evaluación se ha realizado experimentalmente sobre el total de parejas de 2.037 direcciones IP obtenidas en la fase de descubrimiento. Se han realizado las pruebas de identificación desde 100 nodos sonda de la plataforma Planetlab mediante las técnicas Mercator, Ally, Palmtree y Ally-based. También se han utilizado los 25 nodos sonda utilizados en la fase de descubrimiento para las medidas de la técnica Tracenet y medidas desde un solo nodo para las medidas de Radargun y *Prespecified Timestamp*.

En la tabla 3.5 se presentan las distintas tasas de identificación que ofrecen las distintas técnicas de resolución de alias. Las columnas *Positives* y *Negatives* se corresponden con los porcentajes de identificación para las parejas de direcciones IP para las que cada técnica ha dado un veredicto de alias o no alias respectivamente. La columna *Completeness* contiene los porcentajes totales de parejas identificadas

como alias o no alias por cada técnica (suma de resultados positivos y negativos). Las columnas *Error* y *Unknown* tienen información del porcentaje de parejas para las que no se puede ofrecer un veredicto por la ausencia del paquete de respuesta o la ausencia de una contestación válida para realizar el proceso de identificación respectivamente. La columna *Resulting nodes* indica el número de nodos que tendría el mapa resultante a nivel de router tras el proceso de resolución. La columna *base-parameter* indica el tipo de parámetro base usado por la técnica de resolución de alias y la columna *Direct/Indirect* detalla el tipo de estrategia según el destino de los paquetes sonda. En la tabla aparece una técnica de resolución no comentada en el estado del arte llamada *Ally-based*. Dicha técnica es una propuesta propia que se explica en el capítulo 6.

Con intención de determinar a qué grado de completitud se puede llegar utilizando todas las técnicas de forma simultánea, en la tabla se muestra una fila adicional llamada *All*. Dicha fila muestra los diferentes porcentajes para una resolución de alias obtenida al unir los resultados de las diferentes técnicas de resolución.

Technique	*Positives* (%)	*Negatives* (%)	*Completeness* (%)	*Error* (%)	*Unknown* (%)	*Resulting* nodes	*Base-parameter*	*Direct/* Indirect
Mercator	0.00	0.00	0.00	0.00	99.99	2029	SourceIP	Direct
Palmtree	0.03	-	0.03	99.97	-	1343	SourceIP	Direct
Tracenet	0.10	-	0.10	99.9	-	857	SourceIP	Indirect
Ally	0.00	0.04	0.04	99.96	0.00	2025	IPID	Direct
Radargun	0.11	20.27	20.39	79.49	0.11	1625	IPID	Direct
Ally-based (6 packets)	0.07	19.72	19.80	7.65	72.55	1212	IPID	Direct
Ally-based (20 packets)	0.12	62.66	62.79	0.33	36.85	1129	IPID	Direct
Prespecified timestamps	0.06	0.24	0.31	99.68	-	1523	Timestamp	Indirect
All	0.34	73.85	74.19	0.03	25.77	492	All	Both

Tabla 3.5: Porcentajes de resolución de alias para las distintas técnicas

Se puede observar como en prácticamente ninguna técnica se llegan a las cotas máximas de identificación presentadas como respuestas válidas en las tablas 3.1 - 3.4, lo que supone que las técnicas aun pueden ser mejoradas para proveer una mayor completitud. Las cotas de completitud más altas se consiguen mediante la técnica *Ally-based* y mediante la unión de las distintas técnicas se llega a cotas de un 74.19 % de completitud.

3.7 Conclusiones

A partir del estudio realizado se ha podido determinar cómo la mayoría de las técnicas de resolución no obtienen todo el potencial de identificación ofrecido por las respuestas válidas de los distintos parámetros base, dejando posibilidades para mejora en cuestiones de completitud.

El estudio refleja que para el parámetro base IPID, las medidas de tipo indirecto ofrecen una tasa de respuesta mayor, pero las medidas directas ofrecen porcentajes mayores de respuestas válidas.

En el caso del parámetro base *Prespecified Timestamp* son las técnicas directas las que tienen un mayor porcentaje tanto de respuestas, como de respuestas válidas.

Los paquetes sonda de tipo UDP directos sufren numerosos filtrados que hacen que este tipo de paquetes no sean una buena opción a la hora de realizar medidas para resolución de alias. Por este motivo, las técnicas basadas en el parámetro base IP origen que utilizan este tipo de paquetes no ofrecen una buena completitud.

Los routers tienen diferente porcentaje de respuesta según si los paquetes sonda llevan habilitada o no la opción de *Prespecified Timestamp*. Los porcentajes de respuesta son notablemente mayores para los paquetes sonda sin dicha opción habilitada.

Por último, las medidas realizadas mediante las distintas técnicas de resolución de alias usando medidas activas muestran que los mejores porcentajes de identificación los ofrecen las técnicas Radargun y Ally-based con un 20.39 % y un 62.79 % respectivamente. Se ha de tener en cuenta que aunque Ally-based obtenga una mejor completitud, es una técnica con agregación por parejas, lo que supone que el coste de esta técnica tiene orden cuadrático. Radargun tiene un coste de orden lineal lo que la hace más ligera tanto desde el punto de vista del tráfico introducido en la red como del tiempo invertido en la generación de dicho tráfico. En el siguiente capítulo se realiza un estudio en detalle de ambos tipos de técnicas.

CAPÍTULO

4

Estrategias de coste lineal y de coste cuadrático

4.1 Introducción

A la hora de realizar el proceso de resolución de alias de grandes redes no se suele tener en cuenta las técnicas de resolución de alias que tienen coste de orden cuadrático. El motivo de no tenerlas en cuenta es que en este tipo de estrategias la forma de reconocer si dos direcciones IP pertenecen o no al mismo router requiere de una medida específica por cada par de direcciones IP, por lo que para un número dado de direcciones IP (N) hay que hacer medidas para todas las posibles parejas del conjunto de direcciones IP que se quiere identificar $(N*(N-1))$. Dependiendo de la topología, el número de routers puede llegar a ser muy grande con lo que no sea posible realizar las medidas requeridas en un tiempo razonable. Las distintas técnicas basadas en estrategias con coste de orden cuadrático se ven reemplazadas por estrategias con coste de orden lineal, a pesar de que éstas últimas ofrecen en general peores resultados de resolución de alias en métricas como completitud o precisión.

A pesar de que las técnicas con coste cuadrático no se tengan en cuenta en campañas de medida de gran envergadura (Iplane, Skitter, Ark), no existe ningún

estudio en el que se cuestione la viabilidad de realizar una identificación basándose en técnicas de coste de orden cuadrático.

Existen dos factores cruciales a la hora de evaluar la posibilidad de resolución de alias mediante la utilización de estrategias de coste cuadrático. El primero es el factor temporal y el segundo es la tasa de sondeo. El factor temporal tiene que ver con la estabilidad de las interfaces de los routers. Las direcciones IP que pertenecen a los distintos router pueden permanecer tiempos relativamente grandes sin cambiar (un router seguirá teniendo las mismas direcciones IP), por lo que la fase de resolución de alias puede extenderse en ese espacio de tiempo sin ocasionar ningún perjuicio a la resolución de alias.

Por otro lado, la tasa de sondeo tiene relación con la velocidad de envío de medidas activas por parte de los nodos sonda y con la limitación de los routers a la hora de generar respuestas a paquetes sonda al tratarse de una tarea de baja prioridad. Por ejemplo, el envío de paquetes sonda de tipo UDP a tasas de envío por debajo de los 0.3 segundos por paquete, hace que de forma general estos no sean respondidos por parte de los routers. Ambos factores están estrechamente relacionados ya que cumplir las exigencias temporales del primero puede suponer tener que realizar envíos a mayor velocidad, pero por otro lado se debe determinar si se tiene suficiente ancho de banda para poder cumplir las exigencias temporales, así como estudiar si los routers contestan a paquetes sonda enviados a las tasas de tráfico requeridas.

A lo largo de este capítulo se realiza un estudio de los dos factores (temporal y tasa de envío) y se realiza también una serie de consideraciones a tener en cuenta en el caso de utilizar estrategias lineales o cuadráticas para mejorar la resolución de alias. Para realizar este estudio se han utilizado dos técnicas de referencia. Para las lineales se ha elegido Radargun [30] y para las cuadráticas la técnica Ally-based [9] que se corresponden con las dos técnicas que ofrecen las mejores tasas de completitud.

4.2 Escenario de medida

Para poder realizar una valoración experimental para todas las métricas de las dos técnicas de resolución de alias (Ally-based y Radargun) se necesitan redes de referencia de las que se conozca la topología. Sólo de esa forma se puede realizar

una valoración de la precisión, ya que si se desconoce la estructura real de red la precisión sólo puede ser estimada en función de los resultados de las distintas técnicas mediante un proceso de comparación. Por este motivo se ha realizado la resolución de alias sobre las redes Canet4, GlobalNOC y Geant (ver tabla 4.1). De dichas redes se conoce sus direcciones IP y cuales pertenecen al mismo router. Por este motivo no hace falta realizar una fase de descubrimiento.

Para medir la completitud se ha optado por la utilización de 3 conjuntos diferentes de direcciones IP de la red de interconexión de Planetlab. Aunque las redes con información pública son muy útiles a la hora de realizar una medida de la precisión, el número y tipo de routers de éstas no es representativo. Por este motivo, y para tener una medida más cercana a la realidad, se han realizado traceroutes entre distintos nodos sonda de la red de Planetlab a nivel mundial y se han obtenido tres escenarios diferentes de 1.971, 1.282 y 4.844 direcciones IP.

La tabla 4.1 presenta el número de routers, número de direcciones IP, número de alias y el porcentaje sobre el total de parejas de direcciones IP que suponen dichos alias (*Aliases percentage*) conocidos a priori. Para las redes de interconexión de Planetlab sólo se sabe el número de direcciones IP por escenario porque no hay información pública de la topología de dicha red.

Network	*Routers*	*IP addresses*	Number of aliases	Aliases percentage
Canet4	6	103	1225	23.78
GlobalNOC	16	569	13832	8.58
Geant	19	493	7441	6.11
PlanetLab subset 1	-	1971	-	-
PlanetLab subset 2	-	1282	-	-
PlanetLab subset 3	-	4844	-	-

Tabla 4.1: Datos conocidos de los escenarios de medida seleccionados

Para la fase de descubrimiento del conjunto de direcciones IP de la red de interconexión de Planetlab se utiliza distinto número de nodos sonda pertenecientes a la misma plataforma. Para el primer escenario se utilizan 40 nodos sonda, para el segundo 20 y para el último se utiliza un total de 56 nodos sonda.

Para la fase de resolución de alias se utiliza otro conjunto de 100 nodos sonda, también pertenecientes a la plataforma Planetlab, para la realización de las medidas mediante la técnica Ally-based y se utiliza un nodo de ellos para la realización de

las medidas con la técnica Radargun al tratarse de una estrategia de resolución centralizada.

4.3 Estabilidad en los routers de Internet

Para evaluar si es posible realizar un mapa Internet a nivel de router mediante las técnicas de resolución de alias que tienen una agregación por parejas (coste de orden cuadrático), se debe identificar el tiempo máximo que se puede invertir en el proceso de medida de la fase de resolución de alias (factor temporal). Como se ha comentado con anterioridad, el factor temporal define el tiempo que se puede considerar estable la configuración IP de los interfaces de los routers a los que se quiere realizar la identificación.

No existen medidas de esta característica concreta de la red en el estado del arte. Los trabajos que se han encontrado tienen que ver con medidas sobre la estabilidad de los caminos y los balanceos de carga. En el trabajo de Labovitz et al. [43] se presentan medidas obtenidas a lo largo de 9 meses sobre la estabilidad de las rutas. Los resultados muestran cómo los caminos permanecen estables en un 60 % del total de tiempo y se llegan a porcentajes de estabilidad de un 90 % del tiempo de medida para el 90 % de los caminos observados. Siguiendo con el estudio de la estabilidad, en el trabajo de Butler y Mcdaniel [44] se realiza un estudio de la diversidad de caminos obtenida para parejas origen-destino a lo largo de un mes. Los resultados en dicho estudio muestran datos de 3 escenarios diferentes, uno catalogado como muy cambiante, otro que se cataloga como el más representativo y por último uno catalogado de muy estable. Se toma como medida de camino poco estable aquel que tiene 10 o más rutas diferentes a lo largo del mes de medición, y los porcentajes de parejas origen-destino de los escenarios que se consideran poco estables son entre un 15.3 % a un 0.06 %. Mediante la herramienta Dimes, una plataforma de medida basada en ordenadores de usuarios finales que dispone de miles de nodos sonda, en el trabajo de Weinsberg y Shavitt [45] se realizó una batida de medidas durante 9 días. En dicho trabajo se define la ruta dominante como la ruta tomada más veces entre una pareja origen-destino, y tan sólo el 10 % de los caminos utilizaban sólo esa ruta dominante. Para el resto existía algún instante de esos 9 días en los que la ruta tomaba un camino distinto. Por último, otro estudio interesante es el

mostrado en [46], en el que se utilizaron trazas de la base de datos de RIPE NCC. En ella se realizan medidas de traceroute entre todas las posibles combinaciones origen-destino de 50 nodos sonda entre 1998 y 2001. En este estudio concluyen que los caminos en Internet para los pares origen-destino ofrecen una estabilidad de 10 días.

Como de dichos estudios no se puede obtener la medida que se desea, se ha realizado un estudio que permite obtener la medida de cuanto es el tiempo máximo que debe invertirse en la fase de resolución de alias. Dicho tiempo viene ligado al tiempo en el que las interfaces IP de un router permanecen estables, que implica que las distintas direcciones IP pertenecientes al router no hayan sufrido cambios y se denominará como tiempo de estabilidad.

Para realizar la medida se han obtenido datos de dos proyectos diferentes: *Internet Mapping Project* (IMP [7]) y la plataforma ETOMIC [17]. El *Internet Mapping Project* es un proyecto que busca el mapeo de Internet desde un sólo nodo sonda a través de traceroutes a las distintas redes /24 de Internet. Las medidas públicas de este proyecto se prolongan a lo largo de 6 meses en el año 2001 y tiene granularidad temporal diferente según los diferentes destinos.

La plataforma ETOMIC dispone de una base de datos pública de medidas desde el 2007. Concretamente las medidas útiles para identificar el factor temporal que se quiere obtener son las pruebas de traceroute y Paris-traceroute. Estas se realizan de manera periódica durante 3 veces al día entre todos los nodos disponibles en la plataforma, pero debido a que se desea realizar el estudio desde el 2007 hasta el 2012 se han elegido las medidas entre los 18 nodos sonda que se corresponden con los que se disponían desde el inicio de la batería de medidas.

En las pruebas, se considera que si una dirección IP se encuentra a la distancia de N saltos a lo largo de mediciones sucesivas del camino entre dos nodos sonda, dicho router no ha variado en lo que concierne a su interfaz por lo que se considerará estable en el tiempo en el que hayan transcurrido dichas mediciones. Si un router por el que pasa el camino entre un nodo sonda origen y un nodo sonda destino cambia, entonces se observa una variación en el salto perteneciente al router. Del mismo modo, si una de las interfaces del router cambia, dicho cambio se ve reflejado en alguno de los caminos medidos.

En las figuras 4.1 y 4.2 pueden observarse las gráficas de distribución complementaria acumulada de probabilidad (CCDF) obtenidas a partir de los datos de ETOMIC e IMP respectivamente. En ellas se muestra el tiempo de estabilidad de los routers de las redes de interconexión. El número de medidas que se encuentran disponibles al público de estas dos plataformas esta en torno a las 32.000 medidas de traceroutes clásicos en el IMP y las 4.000.000 de medidas de distintos tipos de Paris-traceroute (mediante paquetes ICMP, UDP y TCP) en el caso de la plataforma ETOMIC. Cada punto de la gráfica muestra el porcentaje de routers que es estable durante al menos ese número de días.

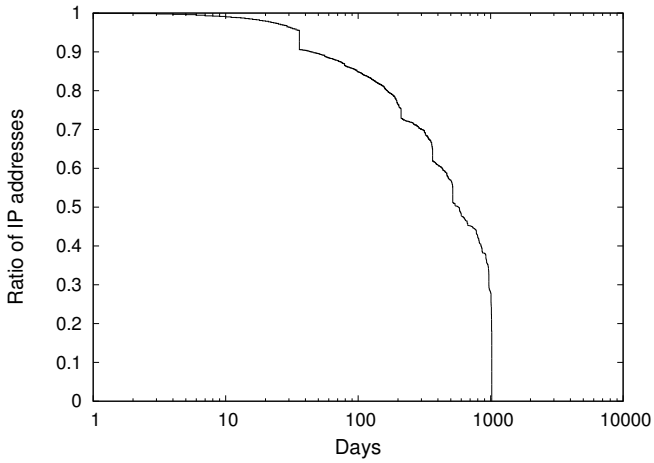


Figura 4.1: CCDF del tiempo de estabilidad en la red de interconexión de ETOMIC

En la figura 4.1 se presentan los datos para la red de interconexión de ETOMIC en la que se observa una estabilidad de al menos 200 días para el 80 % de los routers, de al menos 44 días para el 90 % de los routers y de al menos 11 días para el 99 % de los routers.

Por otro lado en la figura 4.2 se observan los datos de estabilidad de routers en la red de interconexión de *Internet Mapping Project*. En ella se observa una estabilidad de al menos 34 días para el 90 % de los routers. Las curvas de ambas gráficas tienen diferente por dos motivos principalmente, el hecho de que pertenezcan a redes diferentes ya que la red de interconexión ETOMIC utiliza sólo nodos de la Internet europea, pero la red IMP muestra equipos de toda la Internet mundial. Comparativamente, la red europea es más estable que la red Internet mundial. Otra

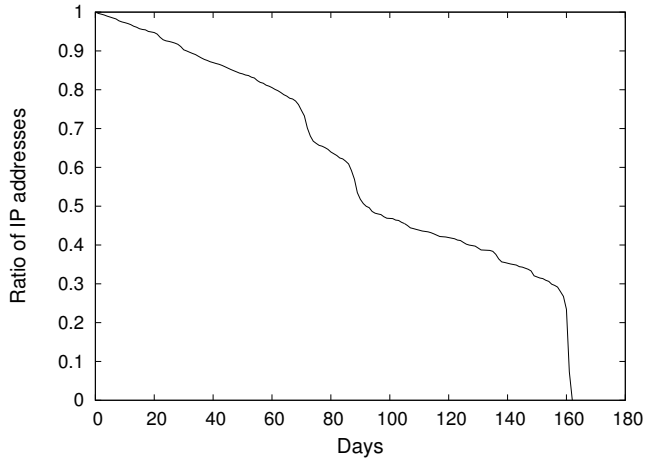


Figura 4.2: CCDF del tiempo de estabilidad en la red de interconexión del *Internet Mapping Project*

razón que hace que las gráficas difieran es que las medidas tienen distinta duración. Mientras que los datos de la red ETOMIC se prolongan durante más de 4 años, de la red de interconexión encontrada vía IMP sólo se disponen de datos de 6 meses.

Como resultado de estas dos medidas, se concluye que la configuración de un router a nivel de interfaces IP en Internet permanece estable a lo largo de un mes para el 90 % de los routers. Este dato es relevante ya que se utilizará para poner una cota máxima a la duración de la fase de resolución de alias.

4.4 Estrategias lineales

Las técnicas que ofrecen costes de orden lineal resultan muy interesantes ya que requieren introducir menos tráfico de sondeo en la red. Por regla general, esto también implica un menor coste computacional al tener que procesar un número mucho menor de respuestas. El menor coste tanto en tráfico como en carga computacional permiten que estas técnicas realicen la tarea de resolución en un tiempo menor. Por contra, y como se concluyó en el capítulo 3, este tipo de técnicas ofrecen porcentajes de resolución menores.

Se ha tomado la técnica Radargun como técnica de referencia y se ha realizado un estudio en detalle de la misma. Por defecto Radargun utiliza una tasa de envío de 80 Kbps y 30 rondas de medidas que se realizan enviando un paquete sonda a

cada una de las direcciones IP del conjunto al que se desea realizar la resolución. A pesar de que la herramienta permite hacer cambios manualmente en la tasa de envío no existe ningún tipo de directriz a la hora de dimensionar dicho parámetro.

Utilizar una tasa de envío fija para cualquier tamaño de escenario implica que el tiempo entre pruebas a la misma dirección IP cambia dependiendo de si el escenario tiene muchas o pocas direcciones IP. Con pocas direcciones IP el tiempo entre paquetes sonda a una misma dirección IP es pequeño, lo que puede ocasionar que ciertos routers no contesten a dichos paquetes sonda por tratarse de tareas no prioritarias. En el caso contrario, cuando el conjunto de direcciones IP es grande, el tiempo entre paquetes sonda a una misma dirección IP es grande, por lo que se pueden provocar reinicios de contador de IPID de los paquetes de respuesta por desbordamiento del contador o que en los crecimientos detectados exista mayor variabilidad, lo que deriva en la imposibilidad de realizar una resolución.

La figura 4.3 muestra los resultados de resolución para escenarios de distinto tamaño utilizando la técnica Radargun. Las direcciones IP para formar los distintos escenarios se han obtenido de la mezcla de las direcciones IP pertenecientes a las redes con información pública Geant, Canet4 y GlobalNOC. Utilizando dichas direcciones IP se han realizado las pruebas para escenarios de distintos tamaños haciendo subconjuntos aleatorios de todo el conjunto de direcciones disponibles. Los resultados muestran tanto el número de alias como el número de no alias.

Con tamaños de red menores a 70 direcciones IP se observa una peor resolución de alias. Tras un análisis de las respuestas, se puede llegar a la conclusión de que es debido a que los paquetes sonda no están siendo respondidos en escenarios con menor número de direcciones IP. El tiempo entre paquetes para 70 direcciones IP a la tasa de sondeo utilizada supone que el tiempo entre paquetes sea de 0,3 segundos. A partir de dicho umbral de tiempo entre paquetes los routers empiezan a no contestar los paquetes de respuesta por lo que se deben elegir escenarios y tasas de sondeo que permitan que el tiempo entre paquetes a una misma dirección IP sea superior a ese tiempo.

En la figura 4.3 se muestra como con escenarios con tamaño mayor a las 400 direcciones IP, los resultados empiezan a empeorar. Analizando los datos se puede concluir que es debido a que el tiempo entre paquetes se incrementa demasiado y esto provoca que los IPIDs tengan una alta variabilidad de su incremento y esto

hace que la técnica de Radargun no pueda asociar las secuencias de IPIDs pertenecientes al mismo router de la manera correcta. En la figura 4.3 se puede identificar otra cota de tiempo entre paquetes de 3,2 segundos para el que los valores de resolución bajan y esto sucede a partir de 520 direcciones IP. Estos mismos umbrales se verificarán si siguen apareciendo en diferentes escenarios para poder comprobar si es el tiempo entre paquetes lo que ocasiona todo el problema.

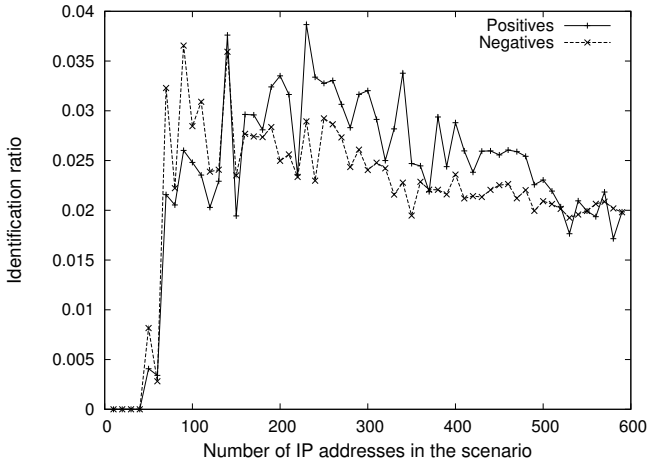


Figura 4.3: Resultados de identificación con RadarGun y diferente tamaño de escenario

Debido a que el escenario con topología real utilizado es muy reducido en cuanto al número de direcciones IP, se puede realizar una mejor observación del efecto del tiempo entre paquetes mediante la variación de la tasa de sondeo en lugar de variar el número de direcciones IP. En la figura 4.3 el efecto del tiempo entre paquetes en la resolución no se ve de una forma clara. Reducir la tasa de sondeo manteniendo el tamaño de escenario significa aumentar el tiempo entre paquetes. De la misma forma, aumentar la tasa de sondeo en el mismo tamaño de escenario implica que el tiempo entre paquetes sonda se verá reducido. En la figura 4.4, se muestran medidas experimentales realizadas a diversas tasas de sondeo en 3 escenarios de tamaño diferente (58, 116 y 944 direcciones IP). En Radargun el tamaño de los paquetes sonda es de 64 bytes y los paquetes sonda se envían hacia cada una de las direcciones IP del escenario. Se puede observar en la figura cómo para cada tasa de sondeo la variación del número de direcciones IP supone también una

variación en los porcentajes de identificación de la técnica Radargun ya deriva a la consecuente variación en el tiempo entre paquetes..

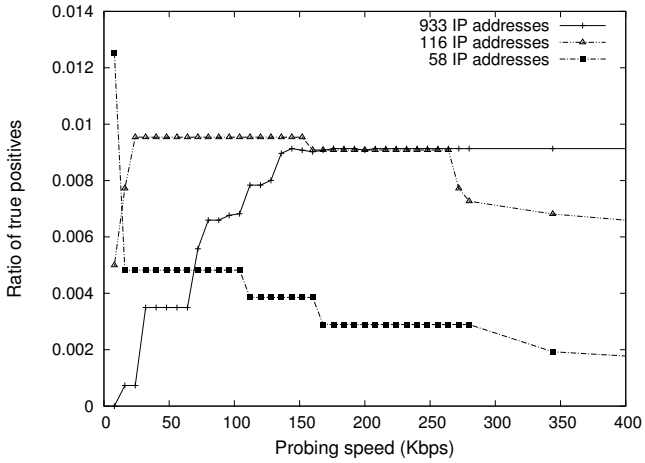


Figura 4.4: Resultados de identificación de Radargun utilizando distintos anchos de banda.

Se puede observar en la figura cómo para tasas de sondeo bajas los porcentajes de resolución son bajos debido al mencionado efecto de tener tiempos entre paquetes grandes. Por otro lado, para tasas de sondeo muy rápidas los porcentajes de resolución vuelven a ser bajos debido al efecto de tener un tiempo entre paquetes muy bajo. Este efecto no puede observarse en la curva que refleja el escenario de 933 direcciones IP porque no se ha utilizado una tasa de sondeo suficientemente rápida. La conclusión a la que se llega mediante la figura 4.3 y 4.4 es que existen dos límites en el tiempo entre paquetes necesarios para un buen comportamiento. La tasa de sondeo correcta debe estar entre 0,3 segundos y 3,2 segundos aproximadamente. Por lo tanto la técnica Radargun debe de ser utilizado con la tasa de envío correcta para permitir una resolución de alias de un determinado tamaño de red. La fórmula que determina que número de direcciones IP (N) puede tener un escenario que se mida utilizando una tasa de sondeo (B) viene dado por la ecuación 4.1. Esto implica que para los paquetes sonda enviados por defecto mediante la técnica Radargun y una tasa de sondeo de 10Mbps, esta técnica podrá operar en escenarios compuestos entre 4.000 y 60.000 direcciones IP.

$$N = (B * t)/S = (B * t)/(64 * 8) \text{ with t}\in[0.3,3.2] \tag{4.1}$$

A pesar de que esta ecuación tiene gran relevancia para la técnica, toda la problemática descrita anteriormente no se revisa en el trabajo original. Tampoco se identifica ninguna forma de dimensionar las tasas de envío para proporcionar una mejor identificación mediante la técnica Radargun.

Por otro lado, en términos de eficiencia, la estrategia Radargun ofrece características muy buenas. En la formula 4.2 se da el tiempo (T) que conlleva la realización de las pruebas para un número (N) de direcciones IP utilizando una tasa de envío específica (B). El tiempo se calcula dividiendo el número total de bits enviados (R) para el tamaño de escenario elegido y se divide entre la tasa de envío (B) a la que se realizan las medidas. El número total de bits enviados se calcula teniendo en cuenta que los paquetes sonda generados por la técnica Radargun son de 64 bytes y se realizan 30 rondas por cada dirección IP.

$$\begin{aligned} R &= N * 64 * 8 * 30 = 15,360N \\ T &= R/B \end{aligned} \tag{4.2}$$

Por ejemplo para un escenario con 5.000 direcciones IP y utilizando un ancho de banda de 10 Mbps, la fase de resolución de alias utilizando esta técnica supone aproximadamente 7,68 segundos.

4.5 Estrategias cuadráticas

Las técnicas de resolución de alias con coste de orden cuadrático son aquellas que tienen una agregación por parejas. En este tipo de técnicas las medidas se deben realizar, por necesidad de la propia técnica resolución, por parejas de direcciones IP utilizando uno o varios paquetes sonda por cada una de ellas. Este es el principal problema de este tipo de técnicas ya que a medida que se aumenta el número de direcciones IP del escenario, el número de medidas a realizar se incrementan de forma cuadrática.

Como técnica de referencia para este tipo de técnicas se ha utilizado la técnica Ally-based ya que es la técnica que mejores resultados ofrece respecto de las demás técnicas con agregación por parejas como se mostró en la sección 3.

Este tipo de técnicas al realizarse por parejas no tienen problemas de pérdida de precisión al aumentar el tamaño del escenario. Aumentar el tamaño del escenario implica un crecimiento del número de parejas al que realizar las medidas pero no implica una pérdida en la calidad de las respuestas ofrecidas por la técnica. Los problemas aparecen por el coste de la realización de esta identificación.

El hecho de tener que realizar las medidas por cada pareja y no para cada dirección IP, hace que el coste temporal (tiempo invertido en realización de medidas) se eleve por tener que realizar un número más elevado de medidas. El tiempo invertido no se debe de prolongar de forma indefinida para realizar una correcta resolución de alias ya que como se ha comprobado en la sección 4.3, en la que se ha analizado el tiempo de estabilidad de los routers, el proceso de medidas no podrá extenderse más allá de 30 días.

Para comprobar la viabilidad para redes grandes de una resolución realizada con una técnica de tipo cuadrática, se ha realizado una simulación del coste temporal del proceso de medidas para realizar dicha resolución. En el caso de la técnica Ally-based, se ha optado por la utilización de 16 paquetes sonda para realizar la identificación por cada pareja de direcciones IP [9] y se utiliza un ancho de banda de 10Mbps desde un único nodo sonda. Los resultados se pueden ver en la figura 4.5 en la que se muestran dos curvas: la primera utiliza la técnica de identificación sin realizar ningún tipo de proceso de reducción y la segunda hace uso del proceso de reducción IP-Offset que se describirá en el capitulo 5. La gráfica muestra como para 30 días, la técnica de resolución puede identificar hasta 58.986 direcciones IP y 180.776 en el caso de utilizar el proceso de reducción.

Incrementar el tamaño del escenario que se puede identificar sin extenderse más de los 30 días implica utilizar una tasa de envío mayor. La figura 4.6 muestra las duraciones de la fase de medida para distintas tasas de envío. En este caso, el incremento de la tasa de envío no supone un problema como sucede en el caso de Radargun ya que los paquetes sonda enviados a cada pareja mantienen el espaciado necesario de 0.3 segundos. La distribución se realiza de tal forma que no se realizan dos procesos de identificación diferentes utilizando la misma dirección IP, por lo que el tiempo entre paquetes para cada dirección IP permanece invariable respecto de realizar la técnica a una pareja determinada que está fijado en 0.3 segundos.

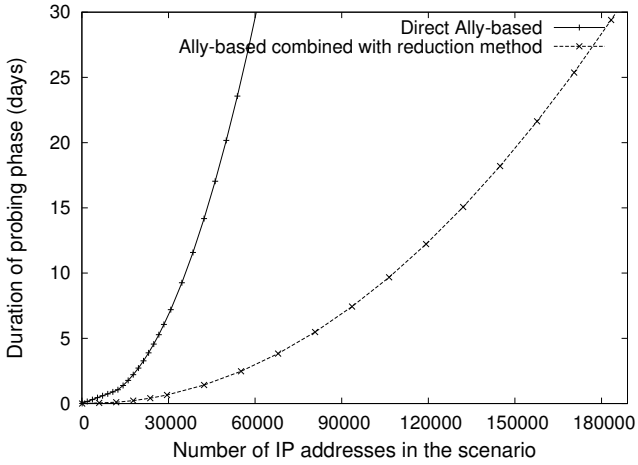


Figura 4.5: Duración del proceso de medidas en la técnica Ally-based

Incrementar la tasa de envío a 100 Mbps implica el poder realizar el proceso de resolución de alias a un tamaño de red de 591.212 direcciones IP.

El proceso de medidas de resolución al poder ser distribuido para realizarse desde diferentes sondas, permite que la tasa de envío se mantenga reducida por cada nodo sonda.

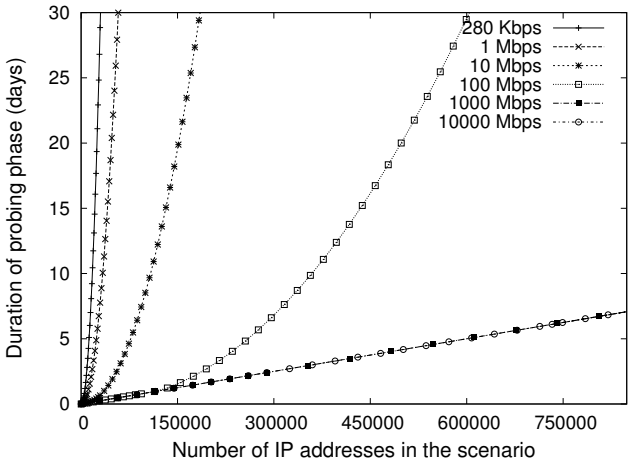


Figura 4.6: Duración de la fase de pruebas en la técnica Ally-based utilizando diferentes tasas de envío

De esta forma el proceso de resolución de alias para redes de gran tamaño puede extenderse tanto en tiempo (las medidas pueden realizarse a lo largo de un máximo de 30 días) como en el espacio (las medidas pueden distribuirse en diferente nodos

sonda). En la distribución espacial se tendrá que tener en cuenta la restricción de que en el mismo momento no se podrán realizar dos medidas a la misma dirección IP. La identificación realizada por cada uno de los nodos sonda puede copiarse a un nodo central desde el que se puede realizar el cálculo de la resolución completa. La distribución espacial permite tanto reducir el tiempo invertido en el proceso de medida como reducir la tasa de envío utilizada en cada nodo sonda.

La mala eficiencia de las técnicas de tipo cuadrático puede ser compensada por la distribución temporal y espacial. Por ejemplo, para la identificación de una red de 60.000 direcciones IP mediante la técnica Ally-based, empleando la misma tasa de sondeo que la utilizada por defecto en Radargun, se necesita 168 días. Si se utiliza una tasa de sondeo de 100Mbps la resolución podría realizarse en 53 días. Esta duración podría reducirse mediante el empleo de 10 nodos sonda a la misma tasa de sondeo, en cuyo caso la resolución consumirá sólo 16 días.

Desde el punto de vista del operador de red, un uso masivo del ancho de banda desde un único nodo sonda hacia varios destinos no supone un comportamiento sospechoso, pero un uso masivo de ancho de banda desde un nodo sonda hacia un sólo destino podría interpretarse como algún tipo de ataque de denegación de servicio. Esta es una razón adicional para justificar una distribución en el espacio de las medidas para el proceso de resolución.

En la figura 4.7 se visualiza el tiempo de prueba desde cada una de las direcciones IP para cada tamaño de escenario a identificar. En lugar de evaluar las tasas de sondeo desde el nodo sonda (*vantage point*) en este caso se realiza la evaluación para cada una de las direcciones IP objetivo de las que se quiere realizar las pruebas de resolución. Cada punto de la gráfica representa la duración en días de las medidas recibidas por los equipos asociados a esas direcciones IP objetivo, para un tamaño determinado de escenario. Cada línea muestra la duración del proceso de medida utilizando distinto número de nodos sonda. Se indica para cada línea el número de nodos sonda trabajando a una tasa de sondeo de 1.7 Kbps. El aumento del número de nodos sonda incrementa el número de paquetes sonda recibidos por cada dirección IP objetivo por unidad de tiempo (esa tasa de sondeo se representa entre paréntesis). Esta tasa de tráfico recibida es el agregado que se recibe de diferentes nodos sonda por lo que las restricciones de seguridad de la red destino se pueden sortear con más facilidad.

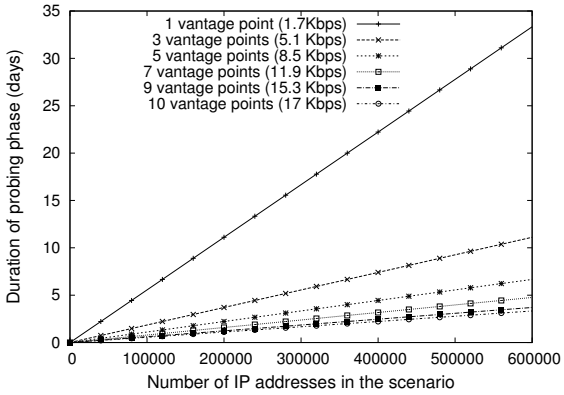


Figura 4.7: Duración del proceso de medidas de la técnica Ally-based para diferente número de nodos sonda y tasa de envío visto desde cada dirección IP destino

Por último queda verificar si es posible, con las restricciones mencionadas anteriormente, realizar una resolución de un número significativo de routers de Internet mediante la técnica Ally-based. Para realizar una evaluación del número de direcciones IP pertenecientes a routers que podría tener Internet se ha recogido la información de topología de dos fuentes, Scamper [47] y Dimes [48]. En ellos, el número de routers de Internet es de 290.000 en el caso de Scamper y de 579.236 en el caso de Dimes. Estos datos son compatibles con el número de redes anunciadas a través de BGP. Como el caso peor es el de Dimes, se realiza el estudio de coste para ese número de direcciones IP. En el caso de utilizar la técnica Ally-based desde 600 nodos sonda y extendiendo las pruebas durante un mes, se podría realizar una identificación utilizando una tasa de sondeo por nodo sonda de 1,7 Mbps que es una tasa de envío asumible y que se puede reducir mediante la ampliación del número de nodos sonda.

4.6 Completitudes ofrecidas por las técnicas lineales y cuadráticas

En este apartado se presenta el funcionamiento de las dos técnicas en estudio (Radargun y Ally-based) para los distintos escenarios descritos en la sección 4.2. Mediante la utilización de ambas técnicas para la resolución de alias en los escenarios de medida se obtienen datos sobre la completitud y precisión de estas.

Las tablas 4.2 y 4.3 presentan los resultados de resolución de alias para Radargun y Ally-based respectivamente en diferentes escenarios. Las 3 primeras redes (Geant, Canet4 y GlobalNOC) son redes de las que existe información sobre su topología pública disponible por lo que se conoce la información de alias y de las 3 siguientes (pertenecientes a la red de Planetlab) sólo se saben las direcciones IP pertenecientes a la red de interconexión que se han obtenido mediante estrategias de descubrimiento (Paris-traceroute). En ambas tablas, las columnas de positivos y negativos muestran el porcentaje de alias y no alias sobre el total de parejas. La columna *Completeness* muestra la completitud total obtenida (suma de positivos y negativos). Las columnas de *Error* y *Unknown* muestran el porcentaje de parejas que no ha respondido a los paquetes sonda o para las que la técnica no ha podido emitir una respuesta sobre el total de parejas. Las columnas de falsos positivos y falsos negativos muestran el porcentaje de parejas sobre los alias y los no alias que son erróneos, de manera que falsos positivos muestra el porcentaje de los positivos que son erróneos y falsos negativos muestra el porcentaje de los negativos que son erróneos. La columna de alias muestra el porcentaje de alias obtenidos mediante la técnica de resolución sobre el total de alias reales que existen en el escenario en estudio. Para las tres primeras redes la información de falsos positivos, falsos negativos y alias se obtiene de la información pública disponible para esas redes. En el caso de los resultados referentes a los escenarios de Planetlab, esa información no es pública por lo que los datos de falsos positivos y falsos negativos de la técnica Radargun se han calculado por comparación con los resultados obtenidos por la técnica Ally-based. La columna de alias no muestra datos para ninguna de las dos técnicas en las últimas redes ya que se desconoce el número real de alias para dichas redes.

Network	*Positives*	*Negatives*	*False positives*	*False negatives*	*Completeness*	*Aliases*	*Error*	*Unkwown*
Canet4	5.90	35.27	0	2.74	41.17	24.81	58.49	0.34
GlobalNOC	0	0.001	0	0.0006	0.001	0	99.99	0
Geant	0.12	1.50	0.004	0.08	1.62	1.96	98.21	0.17
Planetlab subset 1	0.055	28.30	0.008	0.002	28.355	-	71.56	0.08
Planetlab subset 2	0.013	46.76	0.001	0.016	46.773	-	53.18	0.04
Planetlab subset 3	0.00	15.60	18.11	0.48	15.61	-	84.29	0.00

Tabla 4.2: Porcentajes de identificación de alias obtenidos por Radargun

Network	*Positives*	*Negatives*	*False positives*	*False negatives*	*Completeness*	*Aliases*	*Error*	*Unkwown*
Canet4	9.26	48.73	0	0	57.99	38.94	8.12	33.89
GlobalNOC	4.41	42.94	0	0	47.35	51.39	10.15	42.50
Geant	5.11	85.67	0	0	90.78	83.63	0.16	9.06
Planetlab subset 1	0.11	47.68	-	-	47.79	-	3.07	49.14
Planetlab subset 2	0.09	44.50	-	-	44.59	-	4.04	51.37
Planetlab subset 3	0.40	50.10	-	-	50.51	-	15.05	34.44

Tabla 4.3: Porcentajes de resolución de alias mediante el uso de la técnica Ally-based

La tablas muestran que la técnica Ally-based es superior tanto en completitud (porcentaje identificado sobre el total de parejas existentes) como en precisión (ausencia de errores al dar un veredicto sobre una pareja de direcciones IP) para todos los escenarios menos para el subset 2 de Planetlab, en el que los resultados son parejos para ambas técnicas. En ese caso el porcentaje de completitud es un 46,7 % utilizando la técnica Radargun frente a un 44,59 % cuando se utiliza la técnica Ally-based. En el resto de escenarios, la técnica Ally-based ofrece tasas de mejora de la completitud desde 1.40 (en el caso de Canet4) a 47.350 (en el caso de GlobalNOC) veces las obtenidas por Radargun. Los resultados muestran además que en el caso del subset 1 de Planetlab el factor de mejora al utilizar la técnica Ally-based frente al uso de Radargun es de 1.6 veces mejor y en el caso del subset 3 de 3.2 veces.

En cuanto a precisión, la técnica Ally-based no comete ningún fallo para ninguna de las redes con información pública. Por el contrario, la técnica Radargun comete entre un 0.0006 % y un 2.74 % de respuestas fallidas en el proceso de resolución. Para las respuestas obtenidas con Radargun para los escenarios de Planetlab, este porcentaje de fallos asciende hasta un 18.11 % al realizar la comparación con los resultados obtenidas mediante la técnica Ally-based.

Con todo esto la técnica Ally-based supera claramente en las métricas de completitud y precisión a la técnica Radargun.

4.7 Conclusiones

En este capítulo se ha realizado el estudio comparativo entre las técnicas de resolución de alias con coste de orden lineal y las técnicas con coste de orden cuadráti-

co. Se han realizado medidas en escenarios reales y simulaciones que permiten mostrar el comportamiento de las técnicas ante diferentes escenarios poniendo de manifiesto sus características.

Se ha realizado un estudio del tiempo de estabilidad de los routers de Internet. Dicha medida refleja por cuanto tiempo un router permanece sin que la configuración de sus direcciones IP cambie. Mediante medidas realizadas desde la plataforma ETOMIC y desde el *Internet Mapping Project* se ha obtenido un tiempo de estabilidad de 30 días. Por tanto este es el tiempo máximo que se puede emplear para realizar la fase de resolución de alias.

A lo largo del capítulo se ha mostrado que la principal característica para elegir realizar la resolución de alias mediante técnicas con coste de orden lineal es la reducida tasa de sondeo y el menor tiempo que hay que invertir en la realización de las medidas. Por contra, este tipo de técnicas tiene una completitud y una precisión mejorables.

En la técnica Radargun se ha identificado una problemática relativa al tiempo entre paquetes que hace que la resolución de alias que ofrece se vea muy deteriorada. Se ha propuesto una regla de dimensionamiento de la tasa de sondeo en función del número de direcciones IP al que se quiere realizar la resolución.

Por último, se ha justificado que las técnicas con coste de orden cuadrático ofrecen tasas de completitud y precisión superiores a las técnicas con coste de orden lineal, pero tanto el número de paquetes sonda como el tiempo requerido para su realización son también muy superiores. Para permitir una resolución en redes de gran tamaño se propone la distribución espacial mediante el reparto de las medidas en distintos nodos sonda y la división temporal que permite espaciar las medidas a lo largo del tiempo de estabilidad de los routers (30 días). Mediante el uso de 600 nodos sonda y empleando un ancho de banda de 1,7 Mbps se puede realizar una identificación empleando técnicas con coste de orden cuadrático a un escenario con 579.236 direcciones IP, tamaño representativo del core de routers en Internet.

CAPÍTULO

5

Estrategia de reducción basada en IP-Ofsset

5.1 Introducción

Las estrategias de reducción se utilizan en procesos de resolución de alias de redes muy grandes como por ejemplo Internet. Este tipo de estrategias permiten la selección de aquellas direcciones IP que tienen más probabilidad de formar parte del mismo router para así realizar las técnicas de resolución sólo sobre aquellas direcciones IP con más probabilidad de ser alias y de esta forma reducir el tiempo y el tráfico de sondeo en la red.

Las técnicas de resolución de alias que utilizan una estrategia de agregación por parejas requieren de un uso intensivo de la red y de la realización de muchas más medidas que las técnicas de tipo lineal. Por lo tanto, es en este tipo de técnicas donde se suele utilizar las estrategias de reducción.

En el estado del arte, las principales estrategias de reducción son las basadas en TTL, en IPID y en sistemas autónomos (AS). La primera se basa en la premisa de que direcciones IP cercanas en número de saltos tienen más probabilidad de ser alias que direcciones IP lejanas. En principio, todas las interfaces pertenecientes a un mismo router son alcanzables por medio del camino más corto, lo que

supone que todas ellas podrían alcanzarse desde un mismo nodo sonda en un mismo número de saltos. Debido a balanceos de carga y los diferentes esquemas de selección de rutas, en multitud de ocasiones todas las interfaces del router no son alcanzables en el mismo número de saltos, pero la diferencia entre las distancias en saltos suele ser pequeña. En aquellas parejas de direcciones IP entre las que exista una diferencia en número de saltos no superior a un umbral se realizan las técnicas de resolución de alias. Si se opta por un umbral pequeño, se puede perder completitud pero el número total de parejas a analizar es menor. Conforme se utiliza un umbral mayor, aumenta el número de medidas a realizar y con ello la completitud.

La reducción basada en el parámetro base IPID se basa en la premisa de que direcciones IP que pertenezcan al mismo router rellenan las respuestas a los paquetes sonda enviados con IPIDs similares. Partiendo de la misma idea que la utilizada en las técnicas de resolución basadas en IPID, esta estrategia utiliza un envío de paquetes sonda a las direcciones IP que se desea identificar para obtener los IPIDs de los paquetes de respuesta. Si dos direcciones IP pertenecen al mismo router y éste genera los IPIDs de sus paquetes basándose en un contador creciente, estos IPIDs tienen más probabilidad de ser valores cercanos que si las medidas se han obtenido de direcciones IP pertenecientes a distinto router.

En la resolución basada en sistemas autónomos, el requisito para considerar una pareja de direcciones IP como posible alias es que ambas direcciones IP pertenezcan al mismo sistema autónomo. Los routers internos pertenecientes a un sistema autónomo tienen las direcciones IP de sus interfaces registradas en dicho sistema autónomo pero los routers frontera pertenecientes a un determinado sistema autónomo pueden tener interfaces con direcciones IP pertenecientes a distinto sistema autónomo. Por este motivo la estrategia no es muy precisa y se utiliza sólo en combinación con otras estrategias de reducción en los casos en que la reducción tenga más relevancia que la completitud.

A lo largo de este capítulo se propone una nueva estrategia de reducción basada en la distancia numérica entre valores de direcciones IP expresadas en formato entero sin signo de 32 bits que no requiere de medidas activas adicionales (sólo necesita conocer las direcciones IP). Para ello, se ha realizado un estudio de cómo se distribuyen los routers en Internet y se ha realizado una caracterización del espacio de direcciones IP que permite identificar las direcciones IP candidatas a

ser alias. En el capítulo también se detalla el funcionamiento de la estrategia y por último se realiza una valoración de la estrategia en distintos escenarios reales para verificar si la caracterización en la que se basa la estrategia de reducción es extrapolable a cualquier red.

5.2 Organización de Internet a nivel de router

Las redes y routers de Internet se organizan en sistemas autónomos administrados por diferentes entidades como proveedores de acceso a Internet, organizaciones de investigación o distintas empresas. Cada sistema autónomo contiene una o varias redes con rangos de direccionamiento que permiten que direcciones IP pertenecientes al mismo sistema autónomo puedan compartir prefijos.

Se ha realizado un estudio del direccionamiento de los diferentes sistemas autónomos de Internet para observar si la distribución de direcciones IP es uniforme con la intención de utilizar dicha distribución para realizar una estrategia de reducción.

La hipótesis de partida es que se pueden encontrar dos tipos de routers dentro de los sistemas autónomos en función de su conectividad: los routers internos y los routers frontera. Los routers internos poseen interfaces con direcciones IP cercanas en distancia numérica, dado que en un mismo sistema autónomo el rango de direcciones IP a utilizar es menor y es usual que el direccionamiento se realice con direcciones relativamente próximas dentro de una misma red. Existen también routers frontera que interconectan distintos sistemas autónomos. El enlace que conecta dos sistemas autónomos está dentro de una subred que forma parte del rango de direcciones de uno de los dos sistemas autónomos. Por tanto una de las direcciones IP de uno de los dos routers del enlace no pertenecerá al rango de direcciones del sistema autónomo al que pertenece cada router. Como los rangos de direccionamiento de los distintos sistemas autónomos no tienen por qué estar relacionados, las distancias numéricas de dos direcciones IP pertenecientes a distintos sistemas autónomos tendrán una distancia numérica IP diferente que la distancia numérica de las interfaces que pertenecen al rango de direcciones del mismo sistema autónomo. Diferentes sistemas autónomos se organizan de forma jerárquica

divididos en lo que se llaman *Tiers*. El grado del *Tier* marca la posición en la jerarquía de enrutamiento de Internet. Sistemas autónomos pertenecientes al *Tier-2* o *Tier-3* necesitan de un *Tier* superior para tener conectividad completa a toda Internet [49]. Los diferentes sistemas autónomos pueden realizar acuerdos de tipo proveedor-cliente entre *Tiers* de diferente nivel o acuerdos entre iguales cuando se realizan entre sistemas autónomos del mismo *Tier*.

La figura 5.1 muestra un ejemplo de jerarquía de conectividad con distintos niveles de *tier*. Como puede observarse, R2 situado en el sistema autónomo C es un router interno al sistema autónomo. Sus dos interfaces es probable que tengan direcciones IP con el mismo prefijo porque pertenecen al mismo sistema autónomo. Esto implica que las direcciones IP pertenecientes al router puedan ser direcciones IP muy cercanas y que sus distancias al expresar las direcciones IP en forma numérica (*IP-Offset*), no sea grande. Un router frontera puede interconectar distintos sistemas autónomos como el caso del router R1. Dicho router interconecta los sistemas autónomos A y C que poseen un espacio de direcciones diferente. El enlace que interconecta los dos sistemas autónomos se debe de formar con direcciones IP del espacio de direcciones asignado a uno de los dos sistemas autónomos. En este caso se asumirá que las dos direcciones IP de dicho enlace toman valores pertenecientes al sistema autónomo C luego el router R1 esta utilizando una dirección IP para dicho interfaz que no pertenece al rango de direcciones de su sistema autónomo. El resto de direcciones IP de sus interfaces, en cambio, sí pertenecen al sistema autónomo A. El *IP-Offset* del resto de sus interfaces respecto de la interfaz que ha tomado prestada la dirección IP del sistema autónomo C, son mayores que las encontradas por el router R2.

Los diferentes rangos de direccionamientos IP de los AS están relacionados con el grado del *Tier* al que pertenece. Para poder obtener la distribución de las direcciones IP se ha realizado un estudio a partir de la información disponible sobre las tablas BGP que se pueden encontrar en la web [50] tomando los 10 sistemas autónomos pertenecientes al *Tier1*, 52 sistemas autónomos seleccionados de forma aleatoria pertenecientes al *Tier2* y por último 2.500 sistemas autónomos también escogidos de forma aleatoria pertenecientes al *Tier3*. La información disponible cataloga distintos sistemas autónomos por *Tier* así como información de los rangos de direcciones que éstos tienen asignados. En la figura 5.2 se muestra la función

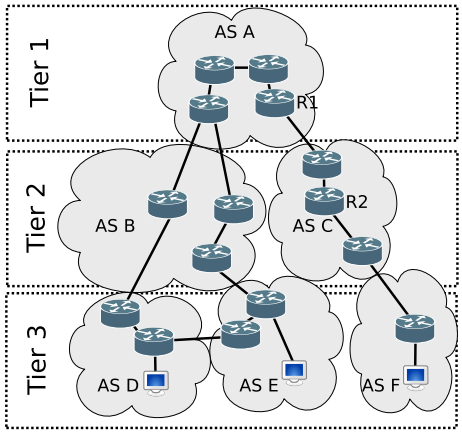


Figura 5.1: Ejemplo de ruta en Internet atravesando distintos routers y sistemas autónomos

de distribución acumulada complementaria (CCDF) de las subredes asignadas al *Tier1*, *Tier2* y *Tier3*. Cada punto de las curvas representa la probabilidad de que una dirección IP tenga ese valor numérico o uno mayor. Para obtener todas las direcciones IP se ha tomado el valor de cada subred representada numéricamente como un entero de 32 bits y se ha ponderado por el número de direcciones IP que contiene obteniendo la distribución numérica de las direcciones IP de los distintos *Tiers*.

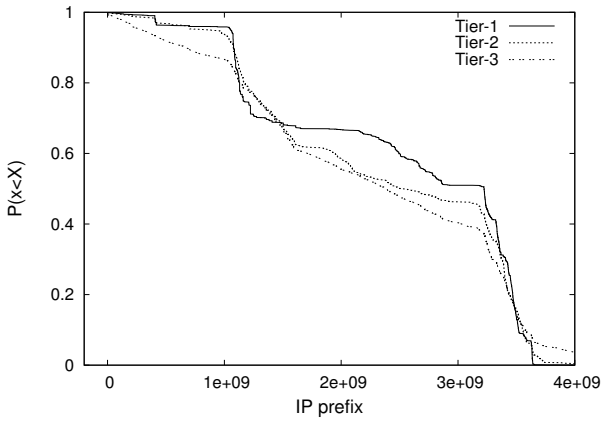


Figura 5.2: CCDF de los prefijos IP correspondientes a los sistemas autónomos Tier-1, Tier-2 y Tier-3

En la figura se observa como para los *Tier1* y *Tier2* existen dos escalones, el

primero en 1,2e+09 y el segundo en 3,4+e09 que indica que en esas dos zonas se concentran porcentajes entre el 70 % y el 80 % (alrededor de un 25 % en el salto cercano a 1,2e+09 y alrededor de un 50 % en el salto cercano a 3,4+e09) de las direcciones IP de los *tiers* 1 y 2. Con estos resultados se pueden estimar los valores que pueden tomar los *IP-Offsets* entre direcciones IP que pertenezcan a un mismo router interno (direcciones IP pertenecientes al mismo sistema autónomo) y aquellas que pertenecen a un router frontera (direcciones IP pertenecientes a distinto sistema autónomo). En el primer caso los *IP-Offsets* tienden a ser cercanos a 0 ya que las direcciones IP son del mismo rango de direcciones. En el segundo caso los *IP-Offsets* toman valores cercanos a 0 en los casos que estemos comparando direcciones IP pertenecientes a sistemas autónomos del mismo escalón o cercanos a 2,15+09 cuando estemos midiendo la distancia entre direcciones IP pertenecientes a sistemas autónomos que se sitúen en el primer escalón con sistemas autónomos pertenecientes al segundo.

Este tipo de organización con routers internos y frontera con características de direccionamiento específicas puede permitir realizar una preselección de aquellas direcciones IP con mayor probabilidad de pertenecer al mismo router. En las siguientes secciones se detallan los escenarios y las medidas realizadas en ellos mediante las cuales se puede observar que una estrategia de reducción basada en el *IP-Offset* es posible y que además ofrece unos índices de reducción equiparables o incluso mejores que las estrategias de reducción ya conocidas.

5.3 Escenario de medida

Para el estudio del *IP-Offset* se han utilizado distintos escenarios con diferente ubicación, número de nodos sonda y alcance. Los escenarios se han obtenido de una fase de descubrimiento realizada con Paris-traceroute mediante la utilización de las plataformas de medida Planetlab y ETOMIC. También se han utilizado para la validación las redes Geant, GlobalNOC y Canet4 de las que se conoce totalmente su despliegue de red incluyendo las direcciones IP de las interfaces de cada router. De esta manera se puede comprobar tanto la precisión cómo la completitud de la estrategia que se desee en las redes de interconexión europeas (ETOMIC), en la In-

ternet mundial (Planetlab) y en redes con despliegues más pequeños pertenecientes sólo al núcleo de Internet (Geant, GlobalNOC y Canet4).

Para el caso de las redes Planetlab y ETOMIC en los que no se conoce la información de su red la forma a proceder para comprobar el funcionamiento de la estrategia de reducción pasa por dos fases, la fase de descubrimiento y la fase de identificación que se realizan a toda la red desde un conjunto de nodos sonda.

Se desea verificar si la reducción obtenida mediante la estrategia está restringida solo a los alias obtenidos mediante técnicas basadas en Ally y Mercator o si la reducción funciona también para el resto de técnicas de identificación. Dado que la completitud obtenida de las redes sin información pública se ha obtenido mediante dichas técnicas de resolución y no ofrecen una completitud del 100 % puede que existan alias para los que la estrategia que se desea plantear no funcione. Por este motivo se ha realizado la evaluación de la estrategia de reducción en las redes Geant, Canet4 y GlobalNOC. De estas redes se dispone de información de todos los alias por lo que si la estrategia no funciona de manera efectiva podría estar ligado a que la estrategia solo funciona para reducir el número de parejas para determinadas técnicas de resolución.

En el escenario de Planetlab se han utilizado 18 nodos sonda, en el de Planetlab se han realizado dos medidas, la primera utilizando 18 nodos sonda y la segunda utilizando 50 nodos sonda. En todas ellas el proceso ha pasado por la realización de una fase de descubrimiento mediante medidas de Paris-traceroute entre todos los nodos sonda para obtener las direcciones IP de la red que los interconectan. Las medidas se han realizado utilizando paquetes de tipo ICMP, UDP y TCP y la utilización del Paris-traceroute ha permitido disminuir los balanceos de carga de algunas rutas. Tras la fase de descubrimiento se ha realizado una fase de resolución mediante las técnicas Mercator, Ally y Ally-based que han permitido identificar un subconjunto de las parejas alias y no alias de cada red de interconexión. Este proceso de descubrimiento y resolución no ha hecho falta en las redes Geant, Canet4 y GlobalNOC ya que se dispone información total de la red.

Los tamaños de las redes que se han utilizado han sido de 510 direcciones IP en el caso de la red ETOMIC, 369 en el caso de la red de Planetlab con 18 nodos sonda y 1.708 en el caso de Planetlab utilizando 50 nodos sonda. Las redes con información pública están formadas por 493 direcciones IP en el caso de Geant,

103 en el caso de Canet4 y 569 en el caso de GlobalNOC. Se puede ver un resumen de los tamaños de todas ellas en la tabla 5.1.

Network	*IP addresses*	*Total aliases*
Geant	493	7441
Canet4	103	1225
GlobalNOC	569	13832
Etomic	510	-
PlanetLab 18 vantage points	369	-
PlanetLab 50 vantage points	1.708	-

Tabla 5.1: Tamaño de los escenarios de medida seleccionados

5.4 Relación de alias según pertenencia a sistema autónomo

A lo largo de las siguientes secciones se van realizar distintos estudios de las distribuciones de las parejas de direcciones IP según lo que se ha denominado *IP-Offset* que representa el valor absoluto de la distancia numérica entre dos direcciones IP. En secciones anteriores se adelantaba que esta distancia podría ser utilizada para diseñar una nueva estrategia de reducción. A lo largo de las siguientes secciones se va a demostrar que es factible la utilización de dicha estrategia y que ofrece buenos resultados de reducción.

Las direcciones IP se concentran en rangos bastante concretos según su pertenencia al sistema autónomo. Al relacionar direcciones IP de alias pertenecientes al mismo sistema autónomo, el *IP-Offset* tenderá a centrarse en valores cercanos a 0, y en el caso de aquellas direcciones IP de alias pertenecientes a routers frontera cuyas interfaces puede pertenecer a otro sistema autónomo, tenderán a tener *IP-Offsets* con valores diferentes a 0. Se ha realizado un estudio de esta característica para el escenario de Planetlab-50 en el que se han calculado los *IP-Offsets* de todos los alias. También se ha obtenido el sistema autónomo de cada dirección IP para ver si es alias del mismo o diferente AS. En la figura 5.3 se muestra la función de distribución acumulada complementaria (CCDF) del *IP-Offset* para los alias de dirección IP que pertenecen al mismo o a distinto sistema autónomo.

Se observa que la distribución de la probabilidad de los alias para los routers pertenecientes al mismo sistema autónomo tiene diferente distribución que los que

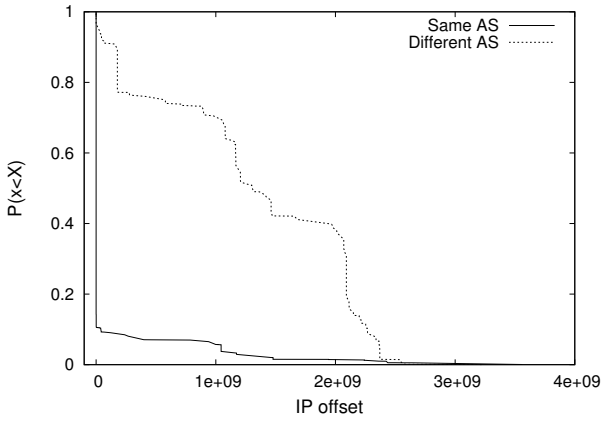


Figura 5.3: CCDF del *IP-Offset* para aliases en el mismo y en distinto AS

pertenecen a distintos sistema autónomo. Se puede observar como el perfil que muestra el *IP-Offset* de los alias pertenecientes al mismo sistema autónomo se concentra principalmente en rangos cercanos a 0. Por el contrario el perfil que describe el *IP-Offset* de los alias pertenecientes a distinto sistema autónomo se concentra en distintos tramos del espacio de posibles valores de *IP-Offset*. Se pueden observar 3 escalones principales en tres zonas diferentes, la zona 0 (valores en torno a 0), la zona 1 (valores en torno a 1,1e+09) y la zona 2 (valores en torno a 2,15e+09). La última zona tiene relación con las diferencias observadas en la distribución de los rangos de direcciones en los *tiers* 1 y 2.

De todas las parejas de direcciones IP del escenario que son alias, 768 corresponden al mismo sistema autónomo. Esto implica que cerca de un 75 % de los alias pertenecen al mismo sistema autónomo. Por lo tanto, el rango cercano de la zona 0 es el que ofrece mayor completitud muy útil en lo que respecta al diseño de una estrategia de reducción. La figura muestra claramente una disposición diferente para el *IP-Offset* de los routers internos que pertenecen al mismo sistema autónomo y los routers frontera que toman prestada una dirección IP del sistema autónomo con el que se van a conectar.

5.5 Relación de alias según distancia en saltos

Dado que existe una relación entre el *IP-Offset* y el tipo de router, podría existir una relación entre el *IP-Offset* y el número de salto en el que se encuentra la pareja

de direcciones IP a identificar respecto el nodo sonda. Si existe una relación entre estos dos parámetros significa que los routers frontera se encuentran a distancias concretas en lo que respecta al número de saltos, permitiendo implementar una estrategia de reducción utilizando ambos.

Para el estudio se ha utilizado el escenario de Planetlab-50 donde se ha analizado el *IP-Offset* de los alias según la distancia en número de saltos desde los extremos donde se han realizado los sondeos. Cada pareja de direcciones IP tiene varias distancias en saltos obtenidas desde diferentes nodos sonda, por lo que de cada pareja de direcciones IP se utiliza la menor distancia de todas ellas. En la figura 5.4 se muestran 3 perfiles pertenecientes a los alias de las zonas 0, 1 y 2 según su distancia en saltos hasta el nodo sonda más próximo (según su TTL). Para cada valor de TTL, se muestra el porcentaje de alias obtenido de cada zona respecto del total de alias obtenidos para dicho salto, de manera que la suma de los alias pertenecientes a las 3 zonas suman el 100 %.

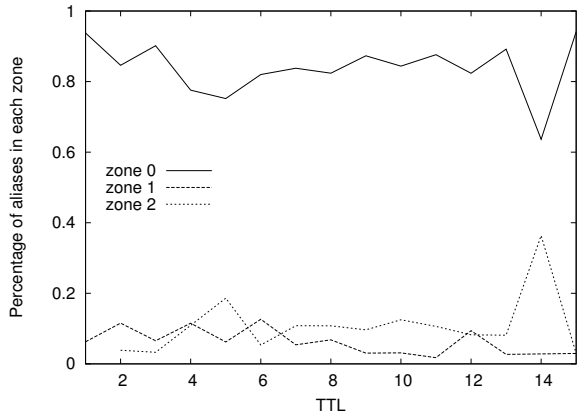


Figura 5.4: Porcentajes de aliases pertenecientes a las zonas según su distancia en saltos

Se puede observar en esta figura cómo la mayoría de alias se concentra en la zona 0, lo que significa que el *IP-offset* entre direcciones IP es cercano a 0 seguramente porque pertenecen al mismo sistema autónomo. Por otro lado, en el perfil de los alias pertenecientes a la zona 2 se observan dos picos principales para los saltos 5 y 14, lo que indica que en esos saltos es más probable un cambio de sistema autónomo con un valor de *IP-Offset* en el rango de 2,15e+09. A pesar de que

se puede observar un incremento de la probabilidad en dichos puntos, no parece suficiente para realizar un filtrado adicional de las parejas de direcciones IP según su TTL. En la zona 2 no se observan variaciones de comportamiento que dependan de la distancia.

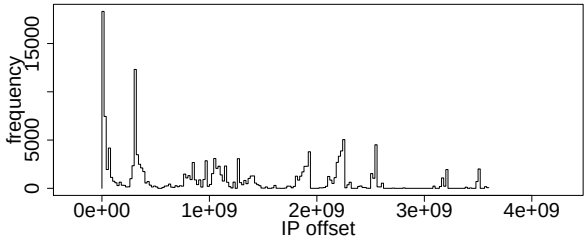
5.6 Relación de alias según el valor de *IP-Offset*

En esta sección se analiza la distribución de las direcciones IP que son alias. La intención final es la de obtener características en relación al *IP-Offset* que permitan discernir entre aquellas parejas de direcciones IP que son alias para poder elaborar una estrategia que permita una preselección de aquellas parejas que pertenecen al mismo router.

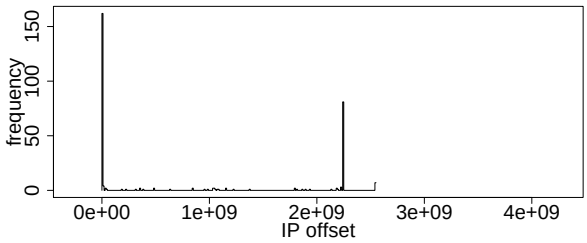
En la figura 5.5 se muestra el comportamiento del *IP-Offset* para la red de interconexión de ETOMIC. La figura esta compuesta por 3 subfiguras en las que se muestra en el eje horizontal los valores que puede tomar el *IP-Offset* y en el eje vertical medidas relacionadas con el número de veces que se observa cada *IP-Offset* concreto.

La figura 5.5.a muestra el histograma en el que cada punto representa el número de parejas que tienen el valor de *IP-Offset* que refleja el eje horizontal. Esta primera figura presenta los valores de *IP-Offset* para el total de parejas posibles de direcciones IP. En la figura 5.5.b se muestra lo mismo que en la figura anterior pero solamente para el espacio de valores de *IP-Offset* de las parejas de direcciones IP que son alias. Por último se muestra en 5.5.c la función de distribución acumulada complementaria de ambos casos para el mismo escenario. Cada punto de esta última gráfica muestra la probabilidad de que el *IP-Offset* sea mayor o igual al valor de *IP-Offset* de este, por lo que se puede observar mejor en qué rangos concretos de *IP-Offset* se concentran las diferentes parejas.

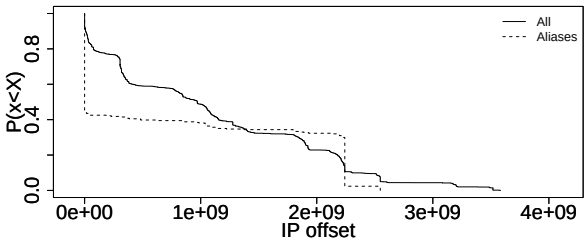
La figura 5.5.a muestra cómo el valor de *IP-Offset* para el conjunto de posibles parejas se distribuye a lo largo de todo el espacio de valores posibles del *IP-Offset*. Esto pone en evidencia que cada dirección IP de cada pareja pertenece a diferentes subredes. Por otro lado en la figura 5.5.b se puede observar como la distribución de *IP-Offset* se concentra sobre todo en dos valores principalmente, el 0 y el 2,15e+09 (pertenecientes a las zonas 0 y 2 anteriormente comentadas). Mediante la figura



(a)



(b)



(c)

Figura 5.5: Histograma del *IP-Offset* para todas las posibles parejas (a), para los alias (b), y el CCDF para ambos (c) en ETOMIC

5.5.c, que presenta el CCDF, se pueden identificar los *IP-Offsets* de mayor probabilidad de forma más sencilla ya que estos aparecen concentrados en dos escalones principales.

Otro escenario de análisis ha sido el correspondiente a Planetlab-18. La figura 5.6 muestra el CCDF del valor de *IP-Offset* para las parejas de direcciones del total de posibles parejas y del conjunto de parejas que son alias. Como se puede observar, el patrón de comportamiento en este escenario es muy similar al del primero. Se pueden observar también los dos escalones que se veían en el escenario anterior pero esta vez el escalón correspondiente al *IP-Offset* con valor 2,15e+09 (zona 3) tiene menor relevancia.

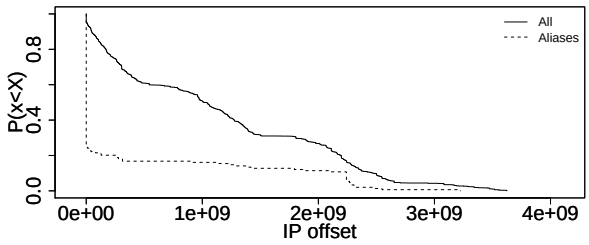
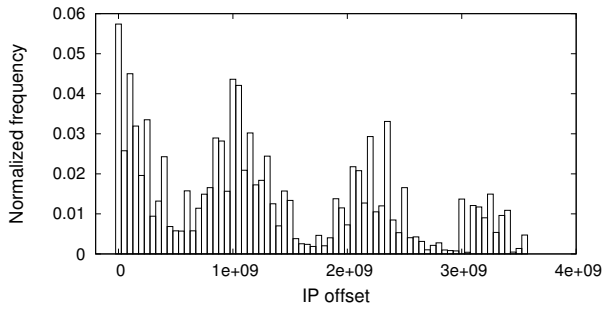


Figura 5.6: CCDF del *IP-Offset* para todas las posibles parejas y parejas sólo alias en el escenario Planetlab-18

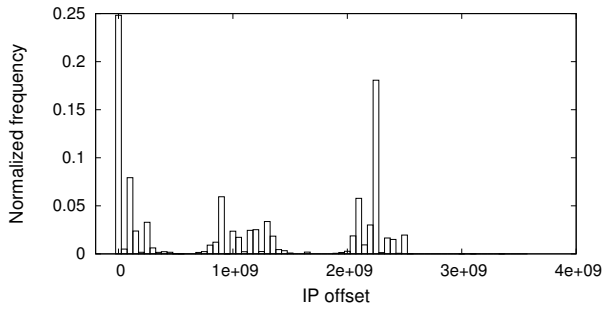
Finalmente, en el escenario más completo, la figura 5.7 muestra el resultado de las medidas realizadas para el escenario de Planetlab utilizando 50 nodos sonda. La figura muestra el histograma del valor de *IP-Offset* para todo el espacio de posibles parejas de direcciones IP (figura 5.7.a) y para el espacio de parejas de direcciones IP formados sólo por las parejas que son alias (figura 5.7.b).

En esta ocasión, con intención de dar mayor relevancia a los routers del núcleo de Internet, se han ponderado las parejas de direcciones tantas veces como aparecen en las trazas obtenidas en la fase de descubrimiento.

De nuevo la figura muestra que el *IP-Offset* se distribuye en todo el rango de *IP-Offset* cuando se aplica al total de parejas posibles mientras que cuando se aplica sólo a los aliases se pueden ver concentraciones en las zonas con valores alrededor de 0 (zona 0), 1,1e+09 (zona 1) y 2,15e+09 (zona 2). La identificación de esas tres zonas es relevante ya que aporta una forma de diferenciar las parejas de direcciones



(a)



(b)

Figura 5.7: Histograma del *IP-Offset* para todas las posibles parejas de IP (a) Y los para las parejas de direcciones IP que son alias (b) en el escenario de Planetlab-50

IP con mayor probabilidad de ser alias mediante el cálculo de su *IP-Offset*. Estas tres zonas se pueden observar mejor en la figura 5.8 en la que se presenta el CCDF del *IP-Offset* para el total de parejas posibles y para las parejas que son alias en el que se observan los tres escalones para los alias en las tres zonas indicadas y la distribución más uniforme para la curva del total de parejas. También existe la zona número 2 identificada en esta sección en torno a 1,1e+09. Se puede observar en las figuras 5.5 y 5.6 como en esa zona sí que hay cierto aumento de la probabilidad pero es mucho más pequeña que en este escenario y por ello no se había podido identificar visualmente. El aumento del tamaño de escenario ha permitido identificar esta nueva zona de alta probabilidad.

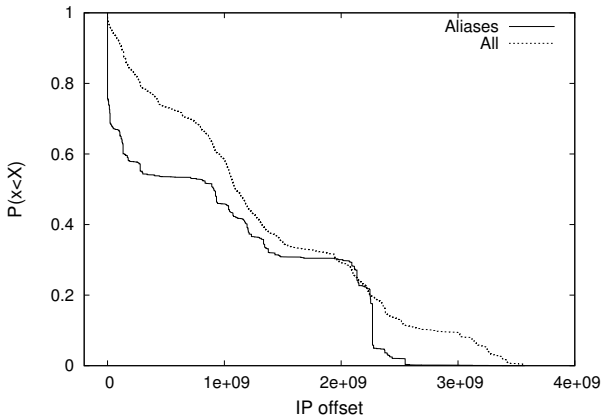
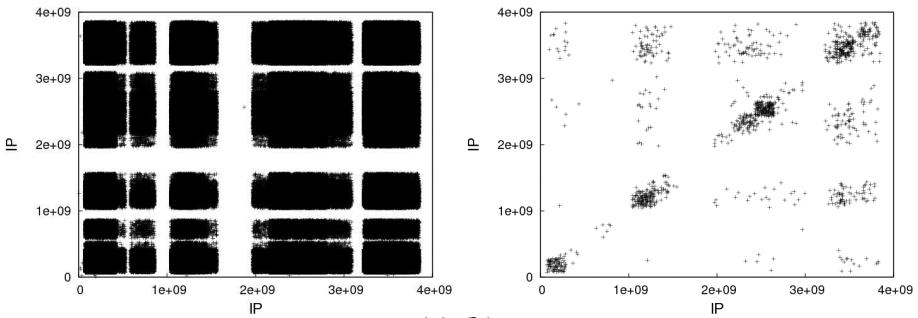


Figura 5.8: CCDF del *IP-Offset* para todas las posibles parejas y solo alias en Planetlab-50

Hasta ahora se han identificado las zonas de alta probabilidad del valor de *IP-Offset* para los alias, por lo que se sabe en qué valores de *IP-Offset* se concentran. Otra forma de representación es la de presentar los routers respecto del espacio total de direcciones IP existente para poder observar la distribución de los alias en el espacio real de direcciones y no según el espacio que representa su *IP-Offset*. En la figura 5.9 se presentan las distribuciones del total de parejas de direcciones IP y de los alias en el escenario Planetlab-50. El eje horizontal y el vertical marcan las distintas direcciones IP existentes en este escenario en formato numérico de 32 bits sin signo. En la figura 5.9.a se representa un punto por cada pareja de direcciones IP perteneciente al total de posibles parejas del escenario de medida y en la figura

5.9.b se representa un punto por cada pareja que sea alias. Se ha añadido un ruido aleatorio a cada punto para poder identificar el volumen de parejas de direcciones IP de cada zona. En la figura 5.9.a muestra como las direcciones IP se concentran dentro de los rangos IP que utilizan los distintos sistemas autónomos cubriendo gran parte del espacio total de direcciones IP. En la figura 5.9.b los puntos que representan alias se concentran en ciertas zonas. Los puntos cercanos a la diagonal representan las parejas de direcciones IP cuyo *IP-Offset* es cercano a 0 (cercanos a la recta Y=X identificada como zona 0). La segunda concentración de puntos relevante que se observa más lejana a la diagonal, se corresponde con la distancia en *IP-Offset* de 2,15e+09 definida previamente como zona 2 (puntos cercanos a la recta $Y = X \pm 2,15e+09$). Otra zona menos densa de puntos se sitúa más o menos a mitad de distancia de las dos zonas anteriores con mayor probabilidad. Esta zona de concentración de puntos corresponde con las parejas de direcciones IP pertenecientes a la zona 1 (puntos cercanos a la recta $Y = X \pm 1,1e+09$).



(a) (b)

Figura 5.9: Parejas posibles de direcciones IP (a) y solo alias (b) en Planetlab-50

En la figura 5.10 se puede ver el mismo estudio realizado esta vez en la red de interconexión de ETOMIC pudiendo identificar también una distribución de las direcciones IP similar y pudiendo observarse con mayor intensidad las parejas de direcciones pertenecientes a las zonas 0 y 2.

En la figura 5.10.a se puede observar diferencias en la distribución de direcciones totales con respecto a las obtenidas en Planetlab, ya que se está observando medidas de la Internet europea lo que supone un menor número de direcciones IP y rangos de direcciones IP limitados.

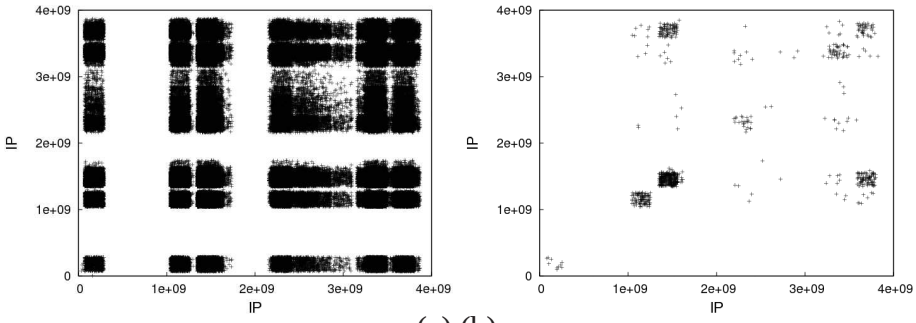


Figura 5.10: Parejas posibles de direcciones IP (a) y solo alias (b) en Etomic

Mediante el estudio realizado en esta sección se puede llegar a la conclusión de que el *IP-Offset* está estrechamente relacionado con las parejas de direcciones IP que son alias. La distribución particular de las parejas de direcciones IP que son alias puede ser utilizada para saber con anterioridad si una pareja de direcciones IP es buena candidata a ser alias. Las zonas 0, 1 y 2 permiten discernir aquellas parejas de direcciones IP con mayor probabilidad de ser alias por lo que puede definirse una estrategia de reducción basada en esta propiedad.

5.7 Estrategia de reducción basada en IP-Offset

En secciones anteriores se ha mostrado que el *IP-Offset* es una buena propiedad para utilizar en una estrategia de reducción. En esta sección se describe una nueva estrategia basada en dicha propiedad que permite reducir el número total de parejas a evaluar a la hora de realizar la fase de resolución permitiendo disminuir considerablemente el ancho de banda y el tiempo invertidos en la fase de resolución.

Un primer paso para la elaboración de la estrategia de reducción en *IP-Offset* se basa en la selección manual de los grupos de parejas de direcciones IP con gran probabilidad de ser alias. Para ello se seleccionan manualmente los tramos de *IP-Offset* con mayor probabilidad y se realizan las medidas sólo a esos conjuntos de parejas de direcciones IP. Se deben elegir unos rangos de valores de *IP-Offset* en los que se cumplan las condiciones de que el número de alias sea alto y el número de parejas a revisar sea pequeño. Con la intención de permitir realizar la selección de rangos para la estrategia de reducción *IP-Offset* de una manera automática y

no dependiente del observador, se propone la utilización de técnicas de *clustering* para automatizar el proceso de reducción. Para este proceso se han considerado dos técnicas de *clustering*, la técnica *K-means* [51] y la técnica *EM* (*expectation maximization*) [52].

Las técnicas de *clustering* permiten la división de un conjunto de datos en distintos clusters con propiedades similares. Este tipo de técnicas de clustering requieren de entrenamiento mediante un conjunto de datos de entrenamiento que permite que la técnica obtenga las propiedades de los diferentes clusters que debe formar. Una vez entrenada, la técnica se aplica a un conjunto de prueba del que se obtendrán clusters con propiedades similares al de entrenamiento. Por este motivo primero se debe utilizar un conjunto de entrenamiento que la técnica dividirá en clusters, y el aprendizaje del entrenamiento se aplica en otros conjuntos de prueba en el que se forman clusters con las características de los creados en el conjunto de entrenamiento.

La estrategia de reducción se basa en utilizar como datos de entrenamiento los datos de los valores de *IP-Offset* para el total de parejas de una red suficientemente genérica. Mediante ese entrenamiento se realiza el proceso de clustering a los alias de dicho escenario obteniendo de esta forma las medidas del número total de parejas que existen en cada cluster y el número de alias.

Mediante un proceso de ordenación se obtienen los clusters que ofrecen mayor número de alias resueltos por número de parejas requeridas para realizar la resolución. De esta manera se sabe qué clusters ofrecen mejores tasas de reducción. Se aplica entonces el proceso de clustering con el aprendizaje anterior al total de parejas posibles de direcciones IP que se pueden formar en una red nueva de la que se desee realizar la identificación. Con esto se obtienen las parejas de direcciones IP del nuevo escenario que pertenecen a cada cluster. Para terminar, se realiza la fase de resolución sólo a aquellas parejas que pertenezcan a los clusters con mayores tasas de identificación. Cuantos más clusters se tomen para realizar la resolución, mayor número de parejas deben pasar por el proceso de resolución pero a la vez se consigue mayores tasas de completitud.

Para poder realizar una valoración de cual de los dos métodos de clustering permite obtener unos grupos que ofrezcan una mejor reducción se ha realizado un estudio empleando el conjunto de direcciones IP obtenidas en los escenarios de

ETOMIC y de Planetlab-18. En este estudio los alias se han obtenido mediante las técnicas de resolución Mercator, Ally y Ally-based. En este primer estudio el conjunto de entrenamiento y el de pruebas es el mismo. El método de clustering que mejor funcione será aquel que ofrezca mayores tasas de identificación por número de parejas a utilizar. En la figura 5.11 se muestran las tasas de alias frente a las parejas necesarias sobre el total de parejas que se corresponden a los dos métodos de clustering utilizando dos criterios para realizar el proceso de ordenación. En un criterio se han utilizado los *IP-Offset* de las parejas que son alias y en el otro los *IP-Offsets* de las parejas que no lo son. También se han utilizado dos variantes diferentes del algoritmo *EM*, el primero utilizando el número óptimo de clusters (la técnica *EM* permite dicho cálculo) y en otro utilizando 15 grupos. En la figura se puede observar como entre los algoritmos que mejores porcentaje de identificación por parejas utilizadas se encuentra el *EM* utilizando los *IP-Offsets* de las parejas que son alias.

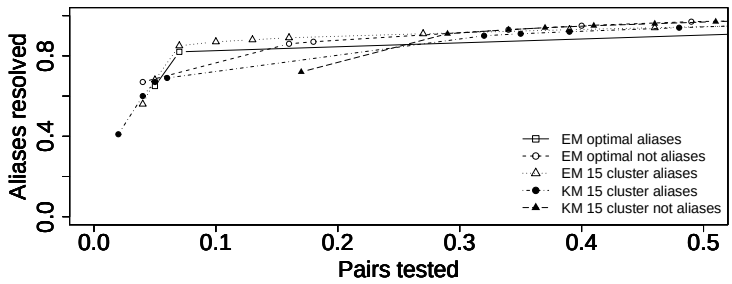


Figura 5.11: Comparación de tasas de reducción de las diferentes técnicas de clustering utilizando distintos criterios de ordenación

En el estudio se ha realizado el clustering del conjunto completo de direcciones IP con las diferentes técnicas para entrenar a los *IP-Offsets* del conjunto de entrenamiento. Estos clusters se han utilizado para identificar los conjuntos de los alias y los no alias (dependiendo de qué característica se ha utilizado para identificar los conjuntos a los que debemos de realizar la identificación). Puede que la aplicación de un segundo paso, que implique la utilización de diferente conjunto de datos para el entrenamiento y las pruebas derive en una disminución de las tasas de reducción. Por este motivo se han realizado varias medidas utilizando las distintas variantes

de los métodos de *clustering* aplicándolas en distintos escenarios. Se ha empleado el escenario de la red de ETOMIC para la fase de entrenamiento para realizar la estrategia de reducción tanto la red de ETOMIC como la red de Planetlab-18. También se ha empleado la red de Planetlab-18 para el proceso de entrenamiento para efectuar el proceso de reducción en la red ETOMIC y Planetlab-18. Los resultados pueden observarse en las figuras 5.12 - 5.16 que presentan cada una los resultados de diferentes estrategias de reducción. La figura 5.12 presenta la reducción mediante la utilización del algoritmo *EM* en la que el proceso de ordenación de clusters se ha realizado mediante los *IP-Offsets* de los alias y utilizando el número óptimo de clusters ofrecido por el algoritmo *EM*. La figura 5.13 muestra una reducción utilizando el mismo proceso que en el caso anterior, pero utilizando los *IP-Offsets* de las parejas que no son alias para el proceso de ordenación. La figura 5.14 utiliza el algoritmo *EM* con 15 clusters y con los alias para el proceso de ordenación de clusters. Por último, las figuras 5.15 y 5.16 se corresponden con la reducción ofrecida por el algoritmo *K-means* con 15 clusters utilizando los *IP-Offsets* de las parejas que son alias y las que no lo son respectivamente para efectuar la ordenación de los clusters.

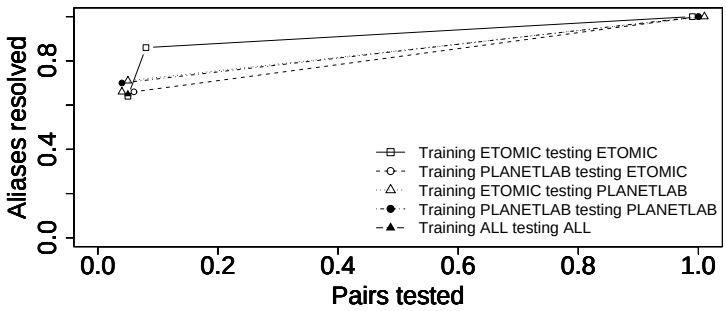


Figura 5.12: Técnica EM basada en alias y número de clusters óptimo utilizando diferentes escenarios de entrenamiento

Como se puede observar, es en la figura 5.14 en la que se observan mejores tasas de identificación en valores reducidos de número de parejas a utilizar para la resolución. Además, las tasas de identificación para porcentajes altos de parejas a

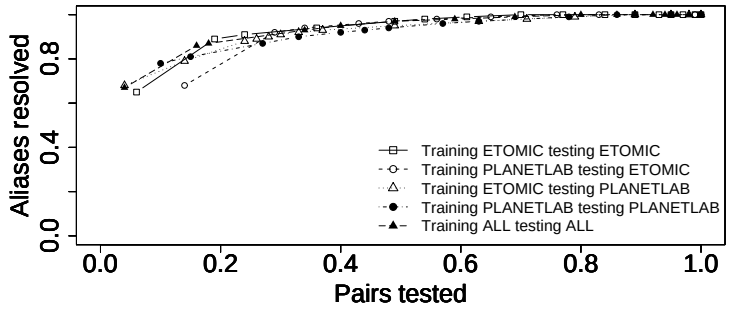


Figura 5.13: Técnica EM basada en parejas no alias y número de clusters óptimo utilizando diferentes escenarios de entrenamiento

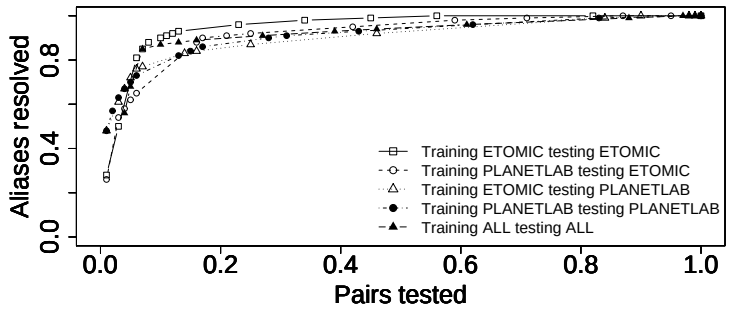


Figura 5.14: Técnica EM basada en alias y 15 clusters utilizando diferentes escenarios de entrenamiento

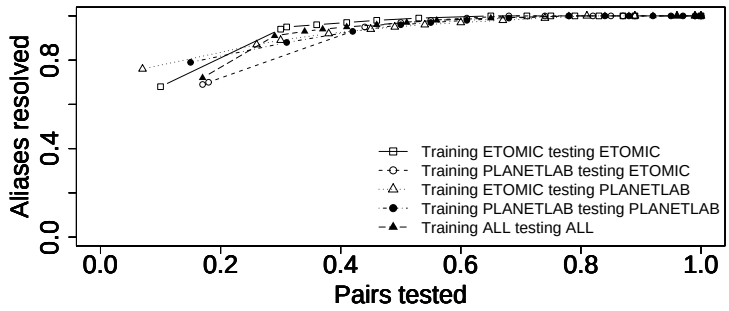


Figura 5.15: Técnica KM basada en parejas no alias y 15 clusters utilizando diferentes escenarios de entrenamiento

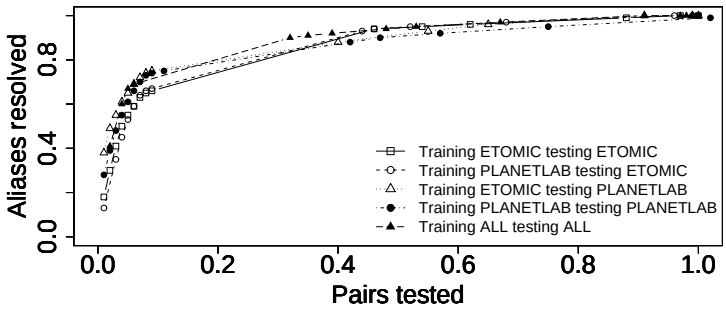


Figura 5.16: Técnica KM basada en alias y 15 clusters utilizando diferentes escenarios de entrenamiento

utilizar en la resolución se mantienen a niveles similares del resto de estrategias utilizadas. Mediante la utilización del método de clustering *EM* utilizando 15 grupos y realizando el proceso de ordenación mediante el número de alias que aparecen en cada cluster, se puede alcanzar tasas de identificación de alrededor del 90 % realizando la resolución a sólo el 10 % del total de las parejas de direcciones IP del escenario en el que se desea realizar la resolución.

También se ha realizado una evaluación de la reducción utilizando el escenario Planetlab-50. Se ha aplicado la reducción utilizando como conjunto de entrenamiento tres escenarios diferentes: ETOMIC, Planetlab-18 y Planetlab-50. La técnica utilizada para la realización de los grupos es el *EM* utilizando 15 grupos y para el proceso de ordenación de clusters el número de alias. Mediante esta medida se puede observar si el hecho de realizar el entrenamiento a subconjuntos más pequeños de la red de la que se desea realizar la identificación tiene o no una gran incidencia en la reducción y la identificación final.

Hay que tener en cuenta que se está realizando la identificación de una red de cerca de 1700 direcciones IP mediante el entrenamiento en escenarios del rango de 300-500 direcciones IP. En la figura 5.17 se pueden ver los resultados, y se observa que la utilización de un conjunto u otro de entrenamiento de *clustering* no tiene una gran incidencia en la identificación final ya que prácticamente las diferentes curvas coinciden entre sí.

En este caso se obtiene una identificación del 73 % cuando se utiliza alrededor

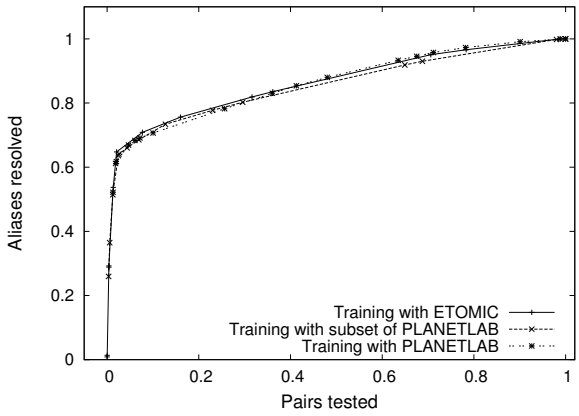


Figura 5.17: Resultados de resolución de alias utilizando la estrategia de reducción *IP-Offset* en el escenario de Planetlab de 50 nodos sonda

del 10 % de las parejas totales existentes en el escenario. Estas tasas resultan más realistas que las calculadas previamente ya que en el proceso de entrenamiento sí que se han utilizado escenarios más reducidos que el escenario final que se deseaba identificar, que es el proceso que se ha de seguir en caso de utilizar la estrategia de reducción en un escenario real de tamaño relevante. Por tanto, los diferentes clusters calculados en cualquiera de los 3 escenarios de esta medida se pueden utilizar para el proceso de reducción de otras redes de Internet.

En la figura 5.18 se pueden observar los diferentes clusters generados en el proceso de clustering en el escenario de Planetlab-50. En ella se observan diferentes segmentos que muestran el rango de *IP-Offset* que cubre cada cluster ordenados de mayor (cluster 1) a menor (cluster 13) porcentaje de alias por número de parejas del cluster. El cluster 1 es el que ofrece mayor tasa de identificación por menor número de parejas a utilizar y el cluster 13 el que menos después de quitar aquellos que no tenían ningún alias en su rango de *IP-Offset*. Los 2 clusters restantes (recordemos que la medida se ha realizado para 15 clusters), no contienen ningún alias.

Se puede observar como existen diferentes rangos centrados en las zonas identificadas en la sección anterior como zona 0, 1 y 2, siendo los rangos pertenecientes a las zonas 0 y 2 los que mayor tasa de identificación producen por número de parejas utilizadas. El detalle de la identificación por número de parejas utilizadas puede observarse en la tabla 5.2. La tabla muestra en las columnas *Alias in Cluster* y *Pairs*

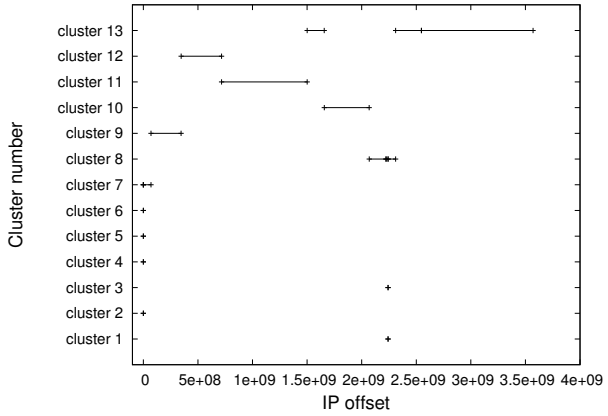


Figura 5.18: rangos de *IP-Offsets* de cada cluster obtenido en el escenario de Planetlab de 50 nodos sonda

Used los porcentajes de parejas que son alias y el porcentaje de parejas sobre el número de parejas totales respectivamente. La columna de *Alias Resolved* muestra el porcentaje sobre el total de alias que tiene el escenario que se encuentran en cada cluster. Se puede observar el resultado del proceso de ordenación y la existencia de rangos de *IP-Offset* para los que existe una concentración muy grande de parejas que son alias. Un ejemplo de ello es el grupo número 2 que ofrece un 27.49 % del total de los alias del escenario utilizando solamente un 0,3 % de las parejas totales del escenario.

Gracias a estas medidas se puede concluir que el proceso de reducción basado en *IP-Offset* se puede utilizar dando buenos resultados de identificación y sin necesidad de la identificación de los rangos mediante un proceso manual dependiente del observador. El proceso mediante algoritmos de clustering mediante el método *EM* utilizando 15 clusters y ordenación utilizando los *IP-Offsets* de las parejas que son alias, ofrece los mejores resultados de todos los que se han evaluado, llegando a alcanzar cotas de identificación de un 90 % con tan sólo el 10 % del total de parejas a identificar. En las siguientes secciones se evalúa la estrategia en redes con información pública de su estructura para verificar si la estrategia de reducción está ligada a la técnica de resolución o si por el contrario el proceso permite detectar alias independientemente de la técnica de resolución utilizada.

Cluster number	Alias in Cluster %	*Pairs Used %*	*Alias Resolved %*
1	6.923	0.0090	0.8712
2	5.362	0.3691	27.4927
3	4.528	0.0184	1.1616
4	1.918	0.8649	23.0396
5	1.211	0.0863	1.4520
6	1.086	0.7311	11.0358
7	0.078	5.6013	6.0987
8	0.038	8.3742	4.4530
9	0.028	15.293	6.2923
10	0.026	4.5432	1.6456
11	0.023	33.7891	11.2294
12	0.014	7.7265	1.5488
13	0.012	21.0741	3.6786

Tabla 5.2: Detalles de cada cluster

5.8 Evaluación de la estrategia de reducción basada en IP-Offset

En esta sección se evalúa la estrategia de reducción *IP-Offset* para los escenarios de medida de los que se dispone de información pública (Canet4, Geant y GlobalNOC). En las secciones anteriores se ha podido observar cómo para los alias obtenidos mediante ciertas técnicas de resolución (Mercator, Ally y Ally-based) esta estrategia de reducción ofrece tasas de reducción muy buenas.

La estrategia de reducción podría estar funcionando sólo para aquellos alias que son sólo identificables vía las técnicas de resolución utilizadas (Mercator, Ally y Ally-based) y no ofrecer una buena reducción para aquellas parejas que dichas técnicas de resolución no pueden obtener. En caso de que eso pase, las estrategias no se podrían utilizar para otras técnicas de resolución que obtengan alias no detectables por estas técnicas revisadas (Mercator, Ally y Ally-based). Para ello se realizan las estrategias de reducción a las redes Geant, Canet4 y GlobalNOC de las que se conocen todos los alias y de esta forma verificar que la estrategia se puede usar para cualquier técnica de resolución.

Las primeras medidas que se han realizado son para ver si la estrategia de reducción permite la reducción cuando se realiza una identificación de cada una de estas redes con información pública. Hay que resaltar que estas redes en estudio no son representativas de Internet sino de redes concretas del núcleo de Internet. Se

ha realizado un entrenamiento en el escenario ETOMIC y la ordenación de clusters en función de los alias de dicho escenario (obtenidos mediante técnicas de resolución). Esos clusters se han aplicado en cada una de las redes en estudio dando como resultado los que se pueden ver en la figura 5.19.

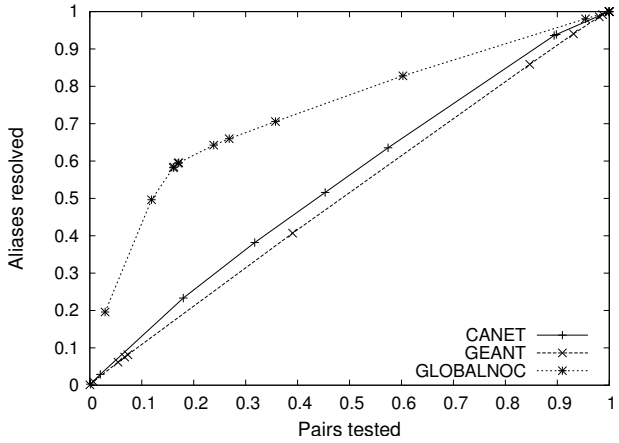


Figura 5.19: Resultados de aplicar la reducción basada en *IP-offset* utilizando clusters de ETOMIC sobre las redes de núcleo

En la figura se muestran de nuevo porcentajes del total de parejas utilizadas (eje horizontal) y porcentajes del total de alias identificados (eje vertical). Los puntos interesantes son los que ofrezcan un alto número de parejas identificadas con un número reducido de parejas totales a utilizar. Se puede observar en la figura que las tasas de reducción obtenidas son muy bajas, excepto para el caso de GlobalNOC donde las tasas aumentan ligeramente. Las curvas que representan las tasas de reducción son cercanas a la diagonal lo que indica que los clusters no ofrecen índices altos de alias por número de parejas a utilizar en el proceso de identificación.

Se ha realizado entonces un estudio del funcionamiento del proceso de entrenamiento en estas redes concretas y si existen clusters diferentes que podrían llegar a ofrecer buenas tasas de identificación utilizando una tasa reducida de número de parejas. Para ello se ha realizado el proceso de clustering directamente en las redes Canet4, GlobalNOC y Geant y se han comparado sus resultados con los obtenidos cuando el entrenamiento se realiza con el escenario de ETOMIC. Los resultados se pueden observar en la figura 5.20.

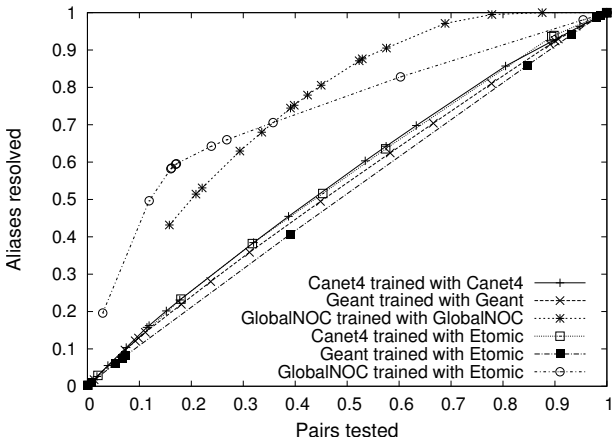


Figura 5.20: Resultados de aplicar la reducción basada en *IP-offset* utilizando distintas estrategias de clustering

Se puede observar como las tasas de reducción no mejoran notablemente respecto de las tasas de reducción ofrecidas utilizando el entrenamiento basado en el escenario de ETOMIC.

Los escenarios en estudio se componen de un número reducido de routers pertenecientes a una red muy concreta y con un espacio de direcciones muy reducido. En un escenario de identificación realista se disponen de distintos tipos de direcciones IP, algunas pertenecientes a redes de acceso, otras a distintas redes pertenecientes al núcleo y con gran variedad de direcciones IP diferentes pertenecientes a distintas redes. Esta variedad observada en las distintas redes es lo que ofrece el potencial a la estrategia de reducción basada en *IP-Offset*, lo que deriva en que si el conjunto total de direcciones IP para el que se está realizando la identificación no es suficientemente heterogéneo, la estrategia no ofrece buenos resultados.

Por este motivo se ha procedido a mezclar las direcciones IP de cada una de las redes en estudio con las obtenidas en el escenario de Planetlab de 50 nodos sonda. Este proceso permite verificar si se identificarían totalmente los alias pertenecientes a estas redes en el caso de que al realizar un estudio amplio de Internet se encuentren entre los routers a identificar.

En la figura 5.21 se pueden observar las tasas de reducción para este estudio. En él se ha empleado la técnica *EM* entrenada con el escenario ETOMIC mediante la ordenación de clusters utilizando los alias obtenidos mediante las técnicas de

identificación Mercator, Ally y Ally-based. Se ha aplicado los clusters sobre la red de Planetlab-50 añadiendo las direcciones IP de cada una de las redes públicas y se ha comprobado el número de alias de estas últimas que se han descubierto.

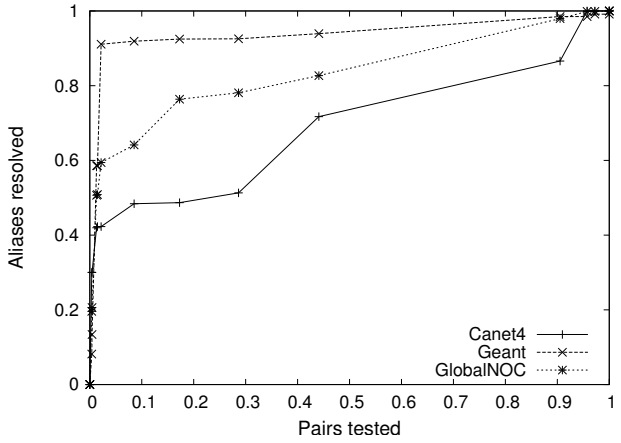


Figura 5.21: Resultados de aplicar la reducción basada en *IP-offset* utilizando el escenario de ETOMIC sobre Planetlab con 50 nodos+redes de núcleo

Se puede observar como mediante el uso de alrededor de un 20 % de las parejas totales se obtienen resultados que incluyen desde un 42 % a un 91 % de los alias totales dependiendo de la red pública.

El proceso mediante el cual se han obtenido los alias en el conjunto de entrenamiento ha sido a través de las técnicas Mercator, Ally y Ally-based, por lo que los alias detectados mediante los cuales se ha realizado el proceso de ordenación de clusters no contenía todos los alias. Puede que la reducción pudiese mejorarse si las técnicas de resolución fueran capaces de detectar todos los alias existentes en la red de entrenamiento pero aun no se dispone de una técnica o conjunto de técnicas que hayan verificado que ofrecen un 100 % de la identificación de alias.

Por tanto y mediante el estudio realizado se ha observado que el proceso de reducción basado en *IP-Offset* es un proceso que permite la reducción en grandes redes ofreciendo tasas de identificación de alias relativamente buenas utilizando un reducido número del total de parejas existentes en el escenario. Por otro lado, la estrategia no es dependiente de la técnica de resolución de alias utilizada por lo que se puede utilizar como proceso de reducción de alias con cualquier técnica de resolución.

5.9 Comparativa de *IP-Ofsset* con otras técnicas de reducción

Se desea comparar las tasas de reducción (porcentaje del total de parejas que es necesario utilizar para realizar la fase de resolución) y completitud obtenida (porcentaje del total de alias obtenidos) de esta nueva estrategias de reducción con las diferentes estrategias ya conocidas. Para la realización de este estudio se ha utilizado el escenario perteneciente a la plataforma ETOMIC, de la que se ha obtenido previamente la resolución de alias mediante las técnicas de resolución Mercator, Ally y Ally-based. De esta forma se pueden restringir las parejas que cumplen las condiciones que se imponen en cada una de las estrategias de reducción, observando cuantas parejas sería necesario utilizar en la identificación y qué porcentaje de los alias se obtendrían en caso de ser utilizadas. Para variar el número de porcentajes de parejas a utilizar en la resolución se han elegido 10 umbrales diferentes para cada una de las estrategias de reducción. En el caso de la estrategia de *IP-Offset*, se ha realizado mediante la selección inicial de unos rangos de *IP-Offset* seleccionados manualmente y después se ha ampliado el número de parejas de esos rangos. En el caso de las otras estrategias se han tomado distintos umbrales para las diferencias en TTL e IPID de los paquetes de respuesta de las direcciones IP del escenario que permiten distintas tasas de identificación.

En la figura 5.22 se puede observar una comparación del número de parejas que es necesario medir frente al número de alias obtenidos utilizando sólo las parejas preseleccionadas cuando se utilizan las estrategias de reducción basadas en TTL, IPID e *IP-Offset*.

Una buena estrategia de reducción es aquella que consigue minimizar el número de parejas a utilizar (presentadas en la figura en el eje horizontal) y maximizar el número de alias encontrados (representado en la figura mediante el eje vertical). Como se puede observar en la figura 5.22 la técnica que ofrece mejores resultados de alias para tasas bajas de número de parejas es la estrategia *IP-Offset*. Con alrededor de un 10 % de las parejas totales existentes, se consiguen identificaciones de alrededor de un 85 % de los alias. A medida que se incrementa el número de parejas a utilizar la estrategia *IP-Offset* se va equiparando al número de alias obtenidos por la estrategia basada en TTL. La estrategia de IPID es especial porque su rango de

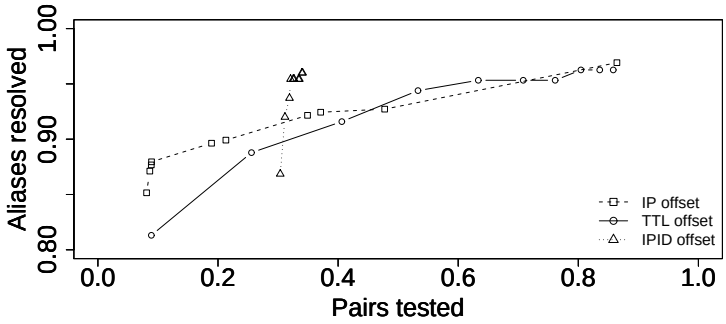


Figura 5.22: Comparación de las tasas de reducción obtenidas entre las distintas estrategias de reducción en el escenario ETOMIC

operación se centra en torno al 30 % de parejas, obteniendo los mejores resultados cuando se compara con las otras dos estrategias exclusivamente en ese rango.

La estrategia *IP-Offset* ofrece tasas de reducción parecidas o incluso mejores que los que se encuentran en la literatura, pero sin necesitar ningún tipo de medida activa adicional para realizar esta estrategia de reducción. Mientras que la estrategia de reducción basada en TTL requiere de la medida desde un mismo punto del número de salto de cada dirección IP y en el caso del IPID de una batida de medidas al mismo tiempo hacia todas las direcciones IP para obtener el IPID, en el caso de la estrategia *IP-Offset* la propiedad que se utiliza para poder evaluar la probabilidad de ser alias de cada pareja se ha obtenido exclusivamente mediante la fase de descubrimiento. Además la obtención de direcciones IP para esta estrategia de reducción no depende del tiempo, como ocurre en la estrategia de reducción basada en IPID, con lo que el proceso de medidas puede alargarse el tiempo que se desee. La estrategia *IP-Offset* tampoco depende de la localización, como ocurre en el caso de la estrategia basada en TTL, por lo que el proceso puede distribuirse sin que la estrategia de reducción se vea comprometida.

Estas dos ventajas se reflejan en la figura 5.23 en la que se muestra el volumen de tráfico de sondeo generado por cada una de las distintas estrategias para realizar el proceso de reducción para el escenario de medida de la red de interconexión de ETOMIC en función del número de nodos. La estrategia *IP-Offset* no necesita de tráfico adicional así que la recta que define su tráfico permanece en el valor

de 0 bytes transmitidos (en la gráfica aparece con valor 1 debido a que es una gráfica en formato logarítmico y es la única forma de representarlo). Para poder obtener valores que permitan la evaluación de todas las parejas del subset y obtener índices de reducción como los que se observan en la gráfica es necesario realizar las medidas para la estrategia TTL e IPID de la siguiente forma.

La estrategia basada en TTL necesita de una tasa de sondeo de un paquete por cada dirección IP de la red de interconexión. Por reducir el efecto de los filtrados se realiza la medida desde varios nodos sonda lo que supone un aumento proporcional del tráfico de sondeo. Si para una dirección IP determinada no se consigue TTL (debido al filtrado), el TTL de dicha dirección no se podrá comparar con el resto y no se podrá incluir en el proceso de reducción.

En el caso de utilizar la estrategia de reducción basada en IPID hay que recordar que el requisito temporal obliga a realizar las medidas por parejas. El crecimiento medio de los contadores de IPID está por debajo de los 200 IPIDs por segundo [9], lo que implica que como el número máximo que puede tener el contador es $2^{16} - 1$, si se realiza la medida en más de 5 minutos y medio el contador sufrirá un desbordamiento. Cuanto más se separen en tiempo los paquetes enviados a direcciones IP que pertenezcan al mismo equipo, más parejas cumplirán el requisito de tener que compararse para ofrecer las mismas cotas de identificación. La realización de las medidas de IPID por parejas permite el establecer un tiempo entre paquetes constante que permite medir de mejor manera las diferencias de IPID entre direcciones IP y deriva también en los altos resultados obtenidos por la estrategia de reducción. El número de paquetes que se introduce a la red es de $((n * (n-1)/2) * 2 * s)$ donde n es el número de direcciones IP de la red de interconexión del escenario de medida y s el tamaño en bytes de cada paquete enviado (64 bytes). El 2 que multiplica al número total de parejas se corresponde con los dos paquetes que se envían por pareja, una para cada dirección IP. Mediante la distribución de las pruebas en distintos nodos sonda se consigue reducir el tiempo y el tráfico generado por nodo sonda, pero como muestra la gráfica el tráfico total de la medida generado se mantiene.

Las condiciones impuestas en el estudio anterior hacen que las estrategias de reducción clásicas utilicen gran cantidad de tráfico para sus medidas previas por lo que en redes grandes resulta inadecuado utilizarlas. Este tipo de estrategias se

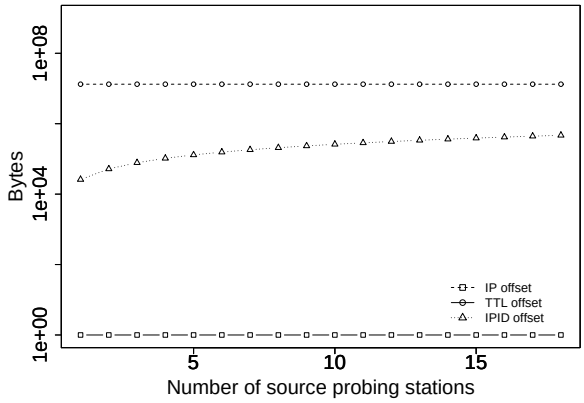


Figura 5.23: Comparación del tráfico de sondeo para las distintas estrategias de reducción en el escenario ETOMIC

utilizan precisamente para permitir la resolución de redes grandes por lo que tal y como se plantean las medidas previas no se podría utilizar. Se ha realizado un nuevo estudio comparativo de la estrategia de reducción *IP-Offset* en el escenario de red de Planetlab-50 frente a las estrategias de reducción basadas en TTL, IPID y AS. Esta vez se han obviado los problemas que pueden aparecer por la insuficiencia de medidas ocasionadas por la utilización de sólo un paquete para la obtención del TTL o el IPID. Se ha realizado una medida desde un mismo equipo sonda para obtener el IPID lo más rápido posible para evitar en la medida de lo posible los problemas derivados del desbordamiento en el campo de IPID. En el caso del TTL se han realizado medidas con el traceroute clásico para obtener el número de salto. Los resultados se pueden observar en la figura 5.24 en la que se observan tres curvas pertenecientes a las estrategias basadas en TTL, IPID y *IP-Offset* y un punto perteneciente a la estrategia basada en AS.

La estrategia basada en TTL muestra tasas de reducción cercanas a las obtenidas mediante la estrategia de *IP-Offset* cuando se utilizan más del 30 % de las parejas totales del escenario para realizar la resolución. Para porcentajes menores la estrategia que ofrece mejores resultados es la de *IP-Offset*. La estrategia que mayor variación ha tenido respecto el estudio anterior ha sido la de IPID, en la que se ve que debido a filtrados en los paquetes y por la forma de realizar la medida produce una reducción bastante baja y que además no llega a una resolución completa de la

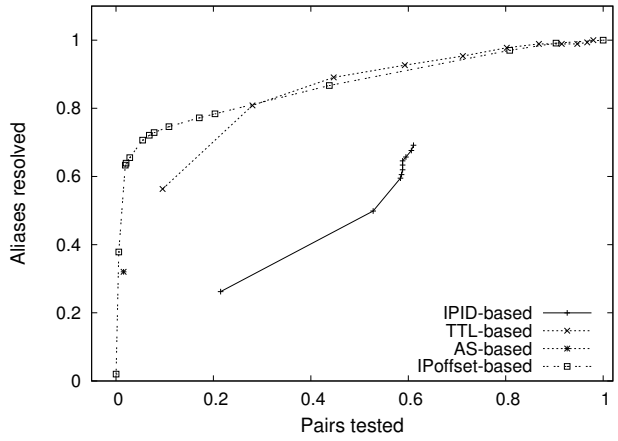


Figura 5.24: Comparacion de las distintas estrategias de reducción en el escenario de Planetlab de 50 nodos sonda

red. Por otro lado la técnica de reducción basada en AS permite obtener resultados cercanos al 32.04 % sobre el total de alias utilizando el 1.55 % de las parejas totales del escenario. Como esta técnica dictamina que se realice la resolución de aquellas direcciones IP que pertenezcan al mismo AS, esta estrategia no permite ampliar sus tasas de identificación mediante la relajación de ningún parámetro como ocurre en las demás estrategias donde se utilizan umbrales. En la estrategia AS sólo se obtiene un punto que representa la reducción obtenida.

Para concluir, se ha estudiado la aplicación de las estrategias de reducción clásicas también en las redes con información pública Geant, Canet4 y GlobalNOC. El proceso de obtención de TTLs y IPIDs se ha realizado mediante el envío de medidas desde un único nodo sonda. Los resultados para la estrategia de reducción basada en TTL se puede observar en la figura 5.25 y la reducción obtenida mediante la estrategia basada en IPID se puede observar en la figura 5.26.

Como puede observarse, las tasas de identificación y de reducción ofrecidas son limitadas no pudiendo ofrecer un 100 % de resolución de alias en ninguno de los casos. Esto está provocado sobre todo por los filtrados de paquetes que no permiten obtener información sobre el IPID o el TTL, no permitiendo que la estrategia de reducción pueda llevarse a cabo correctamente. Se pueden comparar con los resultados de la figura 5.21 donde se observa que *IP-Offset* da mejores resultados.

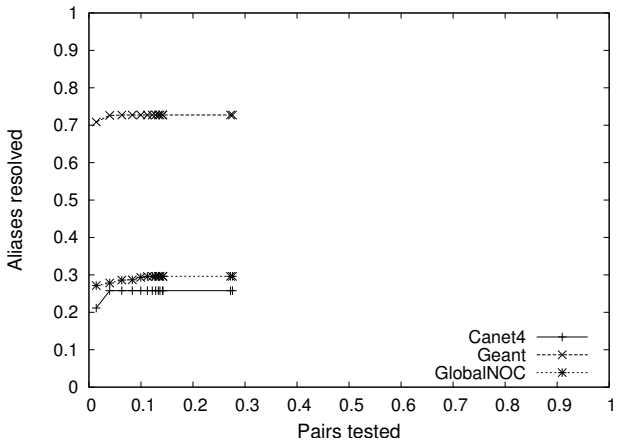


Figura 5.25: Reducción basada en TTL sobre las redes de núcleo

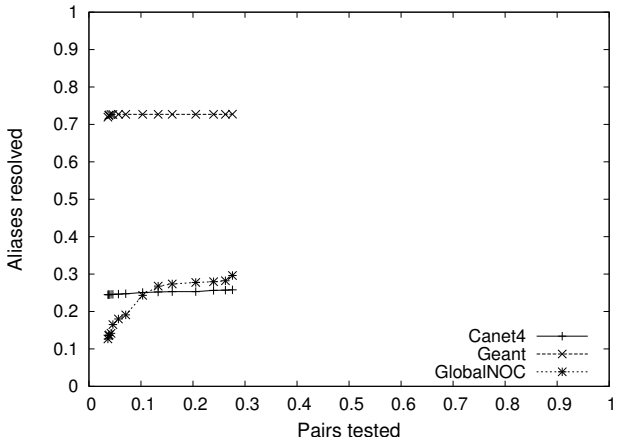


Figura 5.26: Reducción basada en IPID sobre las redes de núcleo

Cómo se ha estudiado la estrategia de reducción de *IP-Offset* ofrece buenos resultados de identificación y reducción.

5.10 Conclusiones

A lo largo de este capítulo se ha desarrollado una nueva estrategia de reducción que mejora las prestaciones de las que se pueden encontrar en el estado del arte. La estrategia se basa en el uso del parámetro *IP-Offset* y en su particular distribución para las parejas de direcciones IP que son alias.

Mediante técnicas de *clustering* como *EM* o *KM* se ha conseguido la automatización de la selección de los rangos de *IP-Offset* que ofrecen tasas de resolución altas para un número reducido de parejas a utilizar en la resolución. De esta manera no se utiliza un método manual de selección de clusters que sería lento y totalmente dependiente del observador.

La estrategia de reducción se ha verificado en 6 escenarios distintos obteniendo tasas de resolución altas mediante la utilización de un número reducido de parejas. Las tasas varían de un escenario a otro, pero una tasa de reducción conservadora cifra la identificación que se puede obtener en un 73 % de los alias utilizando tan sólo un 10 % del total de parejas de direcciones IP que se pueden formar con las direcciones IP del escenario en estudio.

Además, esta técnica de reducción no requiere de medidas adicionales para poder realizarse, ya que el *IP-Offset* puede obtenerse con los datos recogidos exclusivamente durante la fase de descubrimiento.

CAPÍTULO

6

Técnica de resolución de alias *Ally-based*

6.1 Introducción

Existen diferentes técnicas de resolución en el estado del arte pero sin duda una de las más extendidas y que mejores resultados ofrece es la técnica Ally como se ha visto en el capitulo 3. Esta técnica activa y directa se basa en la utilización de paquetes de tipo UDP hacia un puerto en desuso para obtener los valores de IPID de los paquetes de respuesta de ICMP de error por parte de los diferentes routers que se desea identificar. Esta técnica se utiliza en numerosas plataformas de medida como Ark [53] o Skitter [36] y es una de las técnicas de referencia más utilizadas del estado del arte.

A pesar de ser muy utilizada, esta técnica de resolución de alias posee múltiples defectos causados por su esquema de realización de medidas. Tal y como se observa en el capítulo que describe el estado del arte 2.5.2, esta técnica utiliza 3 paquetes de tipo UDP para realizar la resolución de cada pareja de alias. Esta técnica envía dos paquetes sonda sin tiempo de espera entre ellos hacia las dos direcciones IP de las que se desea saber si son alias. Se recibe entonces de cada una de ellas un paquetes de respuesta (uno desde cada una de las direcciones IP) ICMP de tipo

Port unreachable cuyos paquetes contienen el IPID1 (para la primera dirección IP a la que se le ha enviado el paquete sonda) y el IPID2 (para la segunda dirección IP). Un segundo más tarde se realiza el envío de un nuevo paquete sonda a la primera dirección IP obteniendo del ICMP de respuesta el IPID3. Con los 3 IPIDs obtenidos (IPID1, IPID2 y IPID3) se realiza una evaluación que resulta positiva si la distancia numérica entre IPID1 y IPID3 es menor a 200 y si se cumple que $IPID1 < IPID2 < IPID3$. Se considerarán no alias en caso contrario o error en caso de que no se hayan obtenido todos los paquetes de respuesta.

Esta forma de realizar los envíos y tratar los datos tiene dos problemas principales relacionados con el filtrado de paquetes en la red. El primero es que dos paquetes enviados a un mismo destino espaciados por un tiempo muy reducido normalmente no provocan una respuesta en los routers. Esto ocasiona que en multitud de ocasiones no se obtengan los paquetes mínimos necesarios para realizar la evaluación mediante esta técnica. En numerosas redes de Internet los paquetes de tipo UDP se someten a reglas de filtrado lo que es un problema adicional que ocasiona que los paquetes sonda no provoquen tampoco respuesta para poder realizar el proceso de resolución.

El uso de tan sólo 3 paquetes a la hora de realizar la resolución puede provocar que existan demasiados falsos positivos (parejas identificadas como alias que no lo son realmente) y falsos negativos (parejas identificadas como no alias que en realidad lo son) en el proceso de evaluación.

Por estos motivos se propone una variación de la técnica original de Ally a la que se le denomina Ally-based y que consiste en una serie de variaciones que solventen en mayor o menor medida los problemas expuestos anteriormente. A lo largo del capitulo se explica con detalle el proceso de resolución Ally-based, los escenarios donde se ha evaluado esta nueva técnica de resolución y el proceso de pruebas y comparaciones al que ha sido sometida.

6.2 Problemática en la técnica Ally

En la sección anterior se han enumerado algunos de los problemas de los que adolece la técnica Ally original.

6.2.1 Tipos de paquete

Los problemas derivados del filtrado de paquetes de tipo UDP se pueden revisar en los datos aportados en las tablas del capítulo 3. En la tabla 3.2 se mostraban las tasas de contestación para los distintos tipos de paquete de tipo directo, y se puede observar cómo los paquetes de tipo UDP son los que ofrecen una tasa de contestación más reducida con tasas de contestación de alrededor de un 8 % mientras que la utilización de otro tipo de paquetes directos puede llegar a ofrecer tasas de contestación cercanas al 100 % como en el caso de TCP. Por este motivo, para la nueva técnica de resolución a realizar se amplía el tipo de paquetes permitiendo recoger el IPID de los paquetes de respuesta a paquetes sonda de tipo TCP e ICMP.

6.2.2 Comportamientos de los routers

Los paquetes de respuesta directos en Ally ofrecen un abanico amplio de tipos de respuestas (*zero*, *incremental*, *copy* y *random*). Los paquetes de respuesta que esta técnica utiliza para identificar alias son los que se corresponden con un comportamiento incremental. La técnica original Ally no realiza un estudio de inferencia del tipo de router (desde el punto de vista de generación de los IPIDs). Por este motivo es posible que Ally interprete erróneamente los IPIDs los paquetes provenientes de determinados routers a la hora de realizar el estudio de los paquetes de respuesta. Por ello, direcciones IP del mismo router pueden ser identificadas como falsas (no pertenecientes al mismo router) en caso de tener comportamientos de tipo *random*, *zero* o *copy*.

Con la intención de realizar una estimación del número de falsos positivos (métrica relacionada con la precisión) se ha realizado un estudio de los comportamientos de los routers de la plataforma ETOMIC. En el conjunto de direcciones IP de ETOMIC se han identificado los comportamientos: *zero*, *incremental*, *copy*, *random* y *randomT*. En la tabla 6.1 se muestranlos resultados de dicho estudio en el que se pueden ver los porcentajes de contestación de los routers cuando se les envían distintos tipos de paquete sonda. Cada columna está etiquetada con el nombre del comportamiento del router que marca la forma de rellenar los campos de IPID. La última columna se corresponde con los paquetes sonda que por una u

Packet type	*Incremental*	*Random*	*RandomT*	*Zero*	*Unresponsive*
UDP	22.83	0.14	0	4.66	72.34
ICMP ECHO	18.53	0.37	0	50.13	30.95
ICMP TSTAMP	15.52	4.48	13.84	0.08	66.06
TCP	24.45	4.13	0	4.47	66.92

Tabla 6.1: Porcentajes de los diferentes comportamientos detectados para el IPID de los routers de la red de ETOMIC.

otra razón no han recibido respuesta por parte del router. Los porcentajes de contestación para los paquetes sonda utilizados en la técnica Ally se corresponden con la fila de paquetes UDP, que son el tipo de paquetes sonda que utiliza esta técnica. Los diferentes comportamientos del IPID son los ya explicados en secciones anteriores, pero en este caso aparece un nuevo comportamiento bastante marginal al que se le ha denominado *randomT*. Este comportamiento se corresponde con un comportamiento aleatorio para los diferentes IPIDs generados pero cuyo número aleatorio no varía para paquetes generados dentro del mismo segundo.

Dado que las medidas realizadas para la técnica Ally se deben realizar por parejas con los comportamientos observados en la tabla 6.1, se pueden realizar las siguientes combinaciones para parejas de direcciones IP que no pertenecen al mismo router: *Random-Random*, *Random-Incremental*, *Incremental-Incremental*, *Incremental-Random*, *Random-Zero* y *Incremental-Zero*. Las combinaciones de direcciones IP correspondientes a *Zero-Incremental* y *Zero-Random* no pueden dar pie a un falso positivo mediante este proceso de resolución. Dependiendo del comportamiento de los distintos routers a los que se realiza la resolución se pueden distinguir las siguientes probabilidades de falsos positivos a la hora de realizar la técnica Ally:

- *Random-Random:* Cuando el IPID se forma de manera aleatoria para las dos direcciones IP se debe calcular la probabilidad de que $IPID1 < IPID2 < IPID3$ con $|IPID3 - IPID1| \leq 200$ para que cumplan el umbral impuesto por la técnica Ally. El valor del primer IPID no es relevante y puede ser cualquiera. Si se tiene IPID1=0 y IPID2=1 entonces $IPID3 \in 2, 3, 4...,199$, si IPID2=2 entonces $IPID3 \in 3, 4, 5....,199$ y se puede iterar el proceso hasta que IPID2=198 y IPID3=199. La probabilidad de falso positivo en este caso es $P_{R,R} = \sum_{i=1}^{198} \frac{i}{65536^2} = 4{,}58\ 10^{-6}$.

- *Random-Incremental:* Los IPIDs del primer router ahora son aleatorios y los del segundo incrementales. A pesar de ello, del segundo router se obtiene un sólo IPID cuyo valor aislado es indiferenciable de un caso aleatorio. Por este motivo la probabilidad de falso positivo para este caso es la misma que la calculada para el caso anterior. $P_{R,I} = P_{R,R} = 4{,}58\ 10^{-6}$

- *Incremental-Random:* En este caso tenemos el caso particular en el que el primer router es de tipo incremental y el segundo es de tipo aleatorio. Se debe calcular la probabilidad de que $IPID1 < IPID2 < IPID3$ con IPID1 y IPID3 de tipo incremental. Se debe determinar la distancia entre IPID1 y IPID3 para poder entonces calcular la probabilidad de que $IPID2 \in (IPID1, IPID3)$. Esta distancia varía en función tanto del router cómo del instante de tiempo en el que se está midiendo, ya que la generación de paquetes puede variar en función de esos dos parámetros. Para evaluar la variabilidad de los incrementos de IPID en el tiempo se ha realizado una medición de los incrementos entre IPID consecutivos de respuestas desde todas las direcciones IP del escenario en estudio. En la figura 6.1 se puede observar la función de distribución acumulada de los incrementos de los valores de IPID obtenidos de paquetes consecutivos de los routers del escenario ETOMIC enviados con un espaciado de 0,4 segundos. Este espacio de tiempo se corresponde con el tiempo requerido para que se pueda obtener respuesta desde todas las direcciones IP del escenario. Si se desean obtener los incrementos por segundo de la técnica Ally se deben multiplicar las medidas por $1/0,4$. Cada punto presenta la probabilidad de que los incrementos de IPID sean igual o menores que su valor en el eje horizontal. Para realizar el cálculo teórico de la probabilidad de falso positivo en el caso las parejas incremental-aleatorio se utiliza la media de este cálculo experimental que sitúa dicho incremento en 22 IPIDs por segundo ($9 * (1/0,4) = 22$). Por este motivo la probabilidad de falso positivo para este caso es: $P_{I,R} = \frac{22}{65535} = 2{,}7\ 10^{-4}$

- *Incremental-Incremental:* Este es el caso en el que los dos routers son de tipo incremental. Como en el caso de *random-incremental* la segunda dirección IP actúa como si fuese un router de tipo aleatorio así que la probabilidad de falso positivo es igual que la del caso anterior. $P_{I,I} = P_{I,R} = 2{,}7\ 10^{-4}$

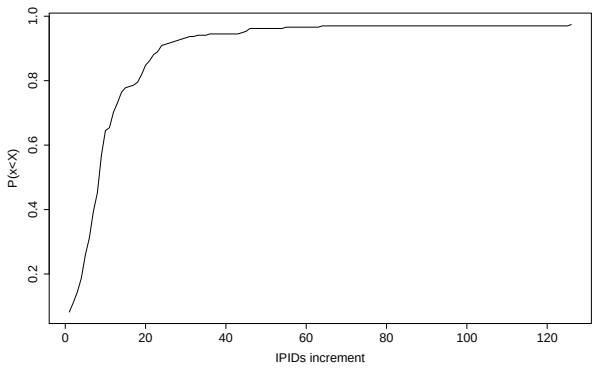


Figura 6.1: Distribución de las distancias de los IPID de dos paquetes de respuesta consecutivos para los routers de ETOMIC

- *Random-Zero:* Este caso es análogo al caso calculado para *random-random*. En este caso también se tiene un valor fijado (IPID2 tiene el valor 0) y los otros dos IPIDs deben de adquirir valores que hagan cumplir las reglas impuestas por la técnica Ally. Teniendo en cuenta que el valor de IPID sufre un desbordamiento y pasa de 65535 a 0, se debe cumplir que $IPID1 < 0 < IPID3$, con lo que si $IPID1 = 65535$ entonces $IPID3 \in \{1, 2, 3, 4, ..., 199\}$. En el caso de que $IPID1 = 65534$ entonces $IPID3 \in \{1, 2, 3, 4, ..., 198\}$. Al final la probabilidad queda de esta forma: $P_{R,0} = \sum_{i=1}^{198} \frac{i}{65536^2} = 4{,}58\ 10^{-6}$

- *Incremental-Zero:* La probabilidad en este caso será análoga a la del caso *Incremental-random*. Se dispone de una distancia de 22 IPIDs entre los IPIDs obtenidos en el primer y último paquete entre los que se tiene que situar el 0. Por tanto, la probabilidad es: $P_{I,0} = \frac{22}{65535} = 2{,}7\ 10^{-4}$

Mediante estas ecuaciones de probabilidad y los porcentajes de los diferentes tipos de routers se puede realizar un cálculo de la probabilidad total de obtener un falso positivo:

$$\begin{aligned} P &= P_{R,R} * 0{,}000025 + P_{R,I} * 0{,}00413100 + \\ &\quad + P_{I,R} * 0{,}004131 + P_{I,I} * 0{,}68260644 + \\ &\quad + P_{R,0} * 0{,}000843 + P_{I,0} * 0{,}13929732 = 2{,}23\ 10^{-4} \end{aligned} \tag{6.1}$$

A priori, esta probabilidad puede no resultar grande, pero en el proceso de resolución de alias de Internet las redes a identificar pueden tener miles de direcciones IP, lo que implica millones de parejas a evaluar y por tanto una aparición significativa de falsos positivos.

Una manera de reducir la probabilidad de una falsa resolución es el aumento del número de paquetes sonda utilizados. Mediante la utilización de los porcentajes de tipos de routers presentados en la tabla 6.1 y de los resultados de las pruebas para obtener los incrementos de los routers presentados en la figura 6.1, se ha realizado una simulación para poder observar el efecto del número de paquetes sonda utilizados en la probabilidad de obtener un falso positivo. Los resultados pueden observarse en la figura 6.2 donde se muestra el comportamiento de la probabilidad de falso positivo (eje vertical) frente al número de paquetes sonda utilizado desde 3 hasta 60 (eje horizontal).

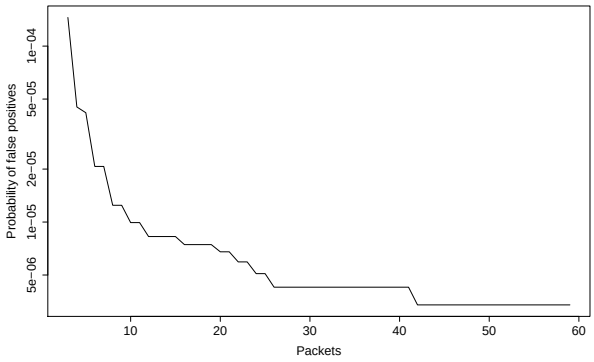


Figura 6.2: Probabilidad de falsos positivos respecto el número de paquetes sonda utilizados.

Se puede observar como la curva tiene un descenso bastante pronunciado desde 3 hasta 10 paquetes, momento en el que el descenso se vuelve menos abrupto. Finalmente para más de 25 paquetes el descenso de la probabilidad de error se vuelve mucho más moderado. Una cifra en torno a los 20 paquetes puede ser una decisión acertada para ofrecer unas tasas de error pequeñas sin elevar demasiado el número de paquetes sonda a utilizar.

6.3 Especificación de la técnica de resolución de alias Ally-based

Los problemas analizados en la sección anterior motivan la introducción de ciertas variaciones que permitan mejorar las tasas tanto de completitud como de precisión de la técnica Ally. Los cambios introducidos no son muy grandes pero permiten mejorar notablemente las tasas de resolución de la técnica Ally original.

El primer cambio a utilizar en esta nueva técnica tiene relación con la utilización de un abanico más amplio de tipos de paquete para realizar las medidas. Como se podía observar en la sección anterior, debido a los filtrados de paquetes, respuestas a medidas utilizando diferente paquete sonda obtienen distintas tasas de contestación. Los tipos de paquetes a utilizar se corresponderán con los protocolos TCP, UDP e ICMP. La obtención de las respuestas con cada tipo de paquete se realiza de forma diferente, utilizando distintos comportamientos de los routers ante paquetes con características diferentes.

En el caso de utilizar paquetes TCP se utilizan paquetes de dicho tipo con el flag de *SYN* habilitado y se envían a un puerto aleatorio. La dirección IP destino no dispondrá de un servicio en escucha en dicho puerto por lo que devuelve un paquete TCP con el flag de *RESET* activado.

En el caso de realizar la medida en base a paquetes de tipo UDP se utiliza la misma estrategia que la utilizada en la técnica original.

Por último, en el caso de realizar el envío utilizando paquetes de tipo ICMP se utilizan dos estrategias diferentes, la primera es la utilización de paquetes de tipo *Echo Request* a la dirección IP destino que contesta mediante un paquete ICMP de tipo *Echo Reply*. La segunda estrategia consiste en emplear paquetes ICMP de tipo *Timestamp Request* que se envían a la dirección IP destino y ésta responde con paquetes de respuesta de *Timestamp Reply*.

Mediante estos nuevos tipos de paquetes sonda utilizados para obtener respuesta por parte de los routers se consigue esquivar algunas de las reglas de filtrado más estrictas que se pueden encontrar hoy en día en los routers.

El segundo cambio que se ha realizado respecto de la técnica Ally original tiene que ver con el espaciado entre paquetes. Algunos routers desechan paquetes enviados hacia ellos cuando los paquetes se reciben muy juntos. El motivo de hacerlo

es el evitar ataques de inundación o de denegación de servicio. Esto ha motivado el realizar un cambio en la forma en que se envían los paquetes sonda. En esta ocasión los paquetes enviados se espacian el mismo tiempo y a una tasa de envío de 0,3 segundos por paquete. Esta tasa de envío se ha obtenido experimentalmente y permite que la mayoría de los routers contesten de manera satisfactoria.

Otro cambio realizado de la nueva técnica tiene que ver con la cantidad de paquetes sonda enviados hacia la pareja de direcciones a la que se desea realizar la resolución. Tal y como se observa en la figura 6.2 existe una probabilidad de obtener un falso positivo de 10^{-4} en la técnica Ally original. A medida que se aumenta el número de paquetes sonda la probabilidad de un falso positivo decrementa. Se ha optado por la utilización de 20 paquetes sonda para la realización de esta medida.

Por tanto la técnica Ally-based utiliza un total de 4 tipos diferentes de paquetes (ICMP *Echo Request*, ICMP *Timestamp Request*, UDP y TCP) para sus medidas. Se realiza por cada tipo de paquete el envío de 20 paquetes sonda por cada dirección IP de la pareja que se desea realizar. El tiempo entre paquetes utilizado es de 0,3 segundos.

Como se puede observar la cantidad de datos introducidos en la red por pareja de direcciones IP es muy superior al de la técnica original. En el caso de la nueva técnica se realiza el envío de $4 * 20 * 2 * 64 = 10240$ bytes, son 4 tipos de paquete por los 20 paquetes utilizados para cada una de las 2 direcciones IP que se quieren analizar y 64 bytes que ocupa cada paquete sonda enviado. Esta cifra es 50 veces superior que la utilizada en la técnica original donde se introducen en la red tan sólo $3 * 64 = 192$ bytes, 3 paquetes por el tamaño mínimo de paquete.

En la técnica también se añade la realización de un estudio previo del crecimiento de los valores de IPID obtenidos de cada una de las direcciones IP que permite saber si cada una de ellas viene de un router de tipo incremental o no. Ally-based etiqueta como falsas aquellas direcciones IP cuyos paquetes de respuesta no tengan el mismo comportamiento para sus IPIDs.

6.4 Escenario de medida

En esta sección se describen los escenarios en los que se han realizado las medidas. Como en capítulos anteriores se han utilizado dos tipos de escenarios diferentes, la

primera usa redes de Internet genéricas siendo necesaria una fase previa de descubrimiento para poder obtenerlas y no teniendo una referencia de la precisión que se consigue en ellas debido a la falta de información sobre la estructura de la red. Estas medidas se realizan utilizando la plataforma ETOMIC y Planetlab. El segundo tipo de escenario contiene direcciones IP pertenecientes a redes concretas de las que se dispone de información completa sobre su estructura de red. Las redes con información pública disponible son las redes Geant, Canet4 y GlobalNOC.

En el escenario de medida ETOMIC se ha realizado un proceso de descubrimiento utilizando la herramienta Paris-traceroute desde 18 nodos sonda del total de la plataforma. Los traceroutes se realizan utilizando como origen y destino todas las posibles parejas de nodos sonda de manera que se obtienen las direcciones IP pertenecientes a todas las rutas entre los nodos sonda utilizados. Un proceso análogo se ha utilizado en la red de Planetlab, para la que se han realizado un mayor número de medidas. Se han obtenido primero tres medidas utilizando 3 grupos de 15 nodos sonda diferentes, después se ha realizado otra medida utilizando 50 nodos sonda y por último utilizando 55 nodos sonda. El total de escenarios obtenidos a partir de la red de Planetlab han sido 5, con 370, 306, 141, 1123 y 4844 direcciones IP. Estas redes permiten conocer cómo se comporta la técnica de resolución en redes pertenecientes a Internet, permitiendo conocer los porcentajes de resolución para futuras medidas en diferentes redes que atraviesen distintas redes pertenecientes a esta.

En el caso de las redes con información pública la fase de descubrimiento no ha sido necesaria porque ya se conocen las diferentes direcciones IP de cada escenario. Estas redes se utilizan para conocer la precisión obtenida por la técnica de resolución. Estas redes son muy específicas y los porcentajes de completitud pueden no ser genéricos pero al disponerse de información pública permiten obtener una medida de la precisión que ofrecen las técnicas de resolución. Estas redes son más reducidas en número de direcciones IP que las obtenidas en los escenarios de Planetlab y ETOMIC. La red Geant dispone de 309 direcciones IP, la red Canet4 de 103 direcciones IP y por último la red GlobalNOC de 593.

6.5 Completitud obtenida por las técnicas de resolución

Cómo se ha observado en secciones anteriores la técnica de resolución Ally-based se basa en dos cambios principales de la técnica Ally original. Los cambios tienen relación con el tipo de paquete sonda enviado así como en la forma de envío. En las primeras medidas se ha realizado comparaciones entre los resultados de completitud obtenidos por la técnica Ally-based y por las técnicas Ally y Mercator. Los resultados de resolución de la técnica Ally-based se presentan separados por tipo de paquete, de manera que se pueda evaluar mediante qué tipo de paquete sonda se obtienen mejores resultados de resolución. En la tabla 6.2 se puede observar las tasas de resolución para las técnicas Mercator, Ally y Ally-based utilizando cada tipo de paquete sonda para el escenario de ETOMIC. Las columnas presentan los resultados positivos, negativos, no concluyentes y errores de cada una de las técnicas. La columna de nodos acumulados muestra el número de routers que se obtiene en total tras utilizar el proceso de resolución de cada fila y todos los anteriores. La columna enlaces acumulados presenta el análogo a la columna anterior pero presenta el número de enlaces en lugar del número de routers. Por último la columna de identificación total presenta el porcentaje total de resolución obtenido tras aplicar el proceso de resolución de cada fila y el de las filas anteriores.

Method	*Positive*	*Negative*	*Not conclusive*	*Error*	*Nodes acumulated*	*Links acumulated*	*Total identified*
Mercator	.02	0	9.35	90.63	545	710	0.02
Ally	.03	7.35	0	92.62	520	692	7.40
Ally-based(UDP)	.06	7.77	0	92.17	506	685	11.79
Ally-based(ECHO)	.21	54.81	19.12	25.86	440	588	62.03
Ally-based(TCP)	.01	3.27	.31	96.41	434	580	63.08
Ally-based(TIME)	.06	12.52	7.91	79.51	434	580	63.17
Total	0.30	62.87	11.58	25.25	434	580	63.17

Tabla 6.2: Resultados de resolución de alias en un escenario real (en % de parejas)

Los resultados de la resolución muestran una notable mejoría de la técnica Ally-based frente a las técnicas Ally y Mercator en todos los casos menos en el caso particular de la utilización de paquetes de tipo TCP. En el caso particular de paquetes UDP, el mismo tipo de paquete utilizado por la técnica Ally original, se puede

observar como tanto la estrategia utilizada para el tratamiento de los datos como el aumento del número de paquetes sonda permiten una mejora de la resolución de los alias (se pasa de un 0,3 a un 0,6 % de resolución) y una mejora más leve en el número de parejas identificadas como falsas (se mejora de un 7,35 % a un 7,77 %). Por otro lado se puede observar que el mejor tipo de paquete sonda a utilizar, en lo que se refiere a completitud, es el ICMP de tipo *echo request*. Si se realiza una resolución utilizando todas las técnicas medidas se obtienen tasas de completitud de un 63,17 % de las parejas identificadas. Si sólo se utilizasen las técnicas del estado del arte las tasas de resolución quedarían reducidas a un 7,40 % de las parejas. Tal y como se puede ver la mejora que los cambios ofrecen a la resolución es notable.

Con intención de poder observar la precisión de la técnica Ally-based se ha realizado las medidas de resolución a las redes con información publica disponible Geant, GlobalNOC y Canet4. Los resultados de la resolución se muestran en la tablas 6.3, 6.4 y 6.5 por cada una de las redes. Dado que se tiene información pública de las redes se ha añadido una columna más a las mostradas en el caso anterior. La columna de falsos negativos ofrece el porcentaje de parejas que se han identificado erróneamente como no pertenecientes al mismo router y que gracias al conocimiento total de la estructura de red se pueden detectar.

Method	Positive %	Negative %	False Negatives %	Not conclusive %	Error %	Number of nodes	Total identified %
Mercator	0	0	0	100	0	318	0
Ally	0.089	1.944	0.000065	0	97.966	288	2.033
IPID_UDP	0.104	2.003	0	0	97.892	287	2.217
IPID_TCP	0.111	1.939	0	0	97.948	286	2.348
IPID_ECHO	4.947	82.745	0	0	12.306	48	87.869
IPID_TIME	4.532	75.940	0	0	19.526	47	90.790

Tabla 6.3: Detalles de resolución de alias sobre la red Geant

En las distintas tablas se puede observar como mientras que con la técnica Ally original siempre se obtiene un porcentaje dado de falsos negativos, en la técnica mejorada Ally-based esta tasa se mantiene siempre en 0. Por otro lado se puede ver en todas las redes, que mediante la técnica Ally-based utilizando los distintos tipos de paquetes se consigue una notable mejora de las tasas de resolución. Se pueden observar incrementos en las tasas de resolución menos notables como en el caso

Method	Positive %	Negative %	False Negatives %	Unknown %	Error %	Number of nodes	Total identified %
Mercator	0	0	0	99.952	0.047	104	0
Ally	6.566	40.186	0.0002	0.0477	53.199	34	46.752
IPID_UDP	7.211	39.708	0	0	53.080	31	47.540
IPID_TCP	7.521	39.947	0	0	52.531	30	50.214
IPID_ECHO	8.213	42.335	0	0	49.450	30	51.289
IPID_TIME	6.661	35.792	0	0	57.545	30	51.289

Tabla 6.4: Detalles de resolución de alias sobre la red Canet4

Method	Positive %	Negative %	False Negatives %	Unknown %	Error %	Number of nodes	Total identified %
Mercator	0	0	0	100	0	574	0
Ally	0.030	0.038	0.0007	0	99.930	559	0.069
IPID_UDP	0.031	0.036	0	0	99.931	559	0.069
IPID_TCP	0.033	0.048	0	0	99.918	557	0.084
IPID_ECHO	4.659	48.321	0	0	47.018	103	52.985
IPID_TIME	0.060	0.498	0	0	99.440	103	52.985

Tabla 6.5: Detalles de resolución de alias sobre la red GlobalNOC

de la red Canet4, en el que se mejora desde unas tasas de un 46.752 % mediante las técnicas Mercator y Ally a tasas de un 51.289 % en la red Canet4, y mejoras que van desde un 2.033 % cuando se utilizan Mercator y Ally original a tasas de un 90.790 % cuando además se utiliza también la técnica Ally-based en el caso de la red Geant. Los datos mostrados ponen de manifiesto que la mejora en precisión y completitud respecto de la técnica Ally es importante.

Por último se ofrece una última medida de las tasas de resolución que se pueden obtener mediante la técnica Ally-based en los diferentes escenarios obtenidos para la red de Planetlab. Los resultados se pueden observar en la tabla 6.6 que muestra las tasas de resolución ofrecidas por la técnica Ally-based. En esta tabla se presentan directamente los resultados de la técnica utilizando todos los tipos de paquetes sonda.

Se puede observar cómo las tasas de resolución se concentran en torno al 50 % de completitud. Estas tasas de resolución son las tasas esperables cuando esta técnica se utilice en otras redes genéricas de Internet.

La comparación de los resultados de esta técnica con otras técnicas de resolución además de Ally y Mercator revisadas en el estado del arte, se muestran en

Network	*Identification %*
Planetlab 1 (15 nodos sonda)	45.45
Planetlab 2 (15 nodos sonda)	31.17
Planetlab 3 (15 nodos sonda)	58.74
Planetlab 4 (50 nodos sonda)	51.39
Planetlab 5 (55 nodos sonda)	50.41

Tabla 6.6: Tasas de identificación mediante la técnica Ally-based

la tabla 3.5 del capítulo 3. En ella se observaba como la técnica Ally-based que se está proponiendo es la que mejores tasas de resolución ofrece con porcentajes cercanos a 62.79 % de completitud seguida de la técnica Radargun con un 20.39 % de completitud.

A pesar de que esta técnica ofrezca buenas tasas de resolución, se realiza por parejas y utiliza un gran número de paquetes sonda. En el estudio mostrado en el capítulo 4 se puede observar como el gasto en número de paquetes y en tiempo en comparación de las estrategias con agregación por parejas con otras técnicas con agregación por grupos es mucho mayor. No obstante en la misma sección se presenta una vía mediante la distribución temporal y espacial de las medidas que permite la resolución de redes de routers de tamaño comparable al de Internet (tamaños cercanos a 600.000 direcciones IP) utilizando técnicas de agregación por parejas como Ally-based. El gasto total en tráfico sigue siendo elevado pero se distribuye en distintos nodos sonda y el tiempo invertido en la resolución se ve reducido el suficiente tiempo como para conseguir realizarla antes de que más de un 10 % de los routers hayan podido cambiar su configuración de red (este tiempo se tabulaba en un mes).

6.6 Conclusiones

En este capítulo se han evaluado los cambios realizados a la técnica Ally para permitir una mejora tanto en la completitud como en la precisión de la técnica original. A esta técnica derivada de la técnica Ally original se le ha denominado técnica Ally-based.

La técnica Ally-based varía, respecto de la técnica original, la cantidad de paquetes sonda utilizados, el tipo de paquetes utilizados y la forma de envío de los

paquetes. Además, la técnica realiza un estudio previo del crecimiento de los IPIDs provenientes de cada una de las direcciones IP para prevenir la resolución de alias errónea que pueden ocasionar routers con diferentes comportamientos en la generación del campo IPID.

Gracias a esta forma de realizar tanto las medidas como su procesado se consiguen tasas de resolución superiores a las obtenidas por la técnica Ally original. Las tasas de completitud se incrementan de tasas cercanas al 7.40 %, obtenidas por la técnica original, a tasas de cerca del 63.17 % de resolución, para la técnica modificada, lo que supone obtener resoluciones en un rango 9 veces mayor a las ofrecidas por Ally. La técnica Ally-based ofrece también mayor precisión reduciendo a 0 el porcentaje de parejas resueltas de manera errónea como puede verse en el estudio realizado hacia las redes Geant, Canet4 y GlobalNOC.

Se han realizado pruebas en distintos escenarios obteniendo tasas de resolución similares lo que confirmaría que la técnica se puede utilizar en distintas redes de Internet obteniendo similares resultados.

CAPÍTULO

7

Técnica de resolución de alias Pamplona-traceroute

7.1 Introducción

La mayoría de técnicas de resolución de alias activas usan estrategias de tipo directo en el envío de paquetes sonda. Las únicas técnicas del estado del arte que ofrecen una resolución de alias basada en medidas indirectas son Midar y Tracenet. Estas dos técnicas utilizan el envío de paquetes sonda a la dirección IP perteneciente al otro lado del enlace que conecta un router con el siguiente. Dicha dirección se obtiene vía una inferencia basada en las máscaras, que usa la premisa de que en general los routers de Internet pertenecientes al mismo enlace compartirán la misma máscara /30 o /31. La forma de obtener un paquete de respuesta de tipo indirecto en estas técnicas es la de dirigir el paquete hacia la dirección IP comentada pero inicializando el TTL del paquete sonda con el valor de número de salto en el que se encuentra el router que se desea medir.

Como se pudo observar en el estudio del capítulo 3, los paquetes de tipo indirecto ofrecen tasas de respuesta altas y comportamientos de los routers en la generación de IPIDs para este tipo de paquetes que son válidos para realizar una resolución de alias.

En este capítulo se propone una técnica de resolución de alias basada en medidas indirectas denominada Pamplona-traceroute. La técnica permite la agregación de fases haciendo que no sean necesarias medidas adicionales a las que se realizan en la fase de descubrimiento. La resolución que puede obtener es alta ya que se basa en la utilización de las respuestas a los paquetes indirectos y además el uso de distinto tipo de paquetes para realizar la resolución facilita que las parejas resueltas no se hayan identificado mediante estrategias de tipo directo. Pamplona-traceroute utiliza los IPIDs de los paquetes de respuesta obtenidos en la fase de descubrimiento para realizar la resolución de alias de las direcciones IP obtenidas en ella.

A continuación se presenta la técnica Pamplona-traceroute, una descripción de los distintos escenarios donde se han realizado las pruebas de evaluación y un análisis de la técnica para verificar su comportamiento en dichos escenarios.

7.2 Especificación de la técnica de resolución de alias Pamplona-traceroute

La técnica de resolución Pamplona-traceroute se compone de 3 fases. Una primera parte que permite la recolección de datos mediante medidas indirectas, otra que realiza un pre-procesamiento de las medidas recogidas durante la fase anterior, y una última fase que realiza la resolución de alias.

7.2.1 Fase de recolección de datos

En esta primera fase se realiza la recolección de datos con los que después se realiza la resolución de alias. El proceso consiste en lanzar instancias del Paris-traceroute en cada uno de los nodos sonda disponibles alrededor de la red que se desea identificar. Para cada nodo sonda, el resto de routers se marcarán como destino de cada medida de Paris-traceroute que se realizan utilizando un sólo paquete sonda por cada TTL. En caso de que el nodo sonda destino no conteste, se define como TTL máximo 30 tal y como está definido por defecto en la mayoría de implementaciones de traceroute. Cada dirección IP perteneciente a los routers del camino enviarán como respuesta a los diferentes paquetes sonda paquetes ICMP de error de *tiempo excedido en tránsito*. A pesar de que los paquetes de tipo ICMP por regla general

están filtrados en numerosas redes, el uso de medidas mediante herramientas de tipo traceroute está muy extendido por lo que este tipo de paquetes concreto no suele estar filtrado.

Los paquetes de respuesta de tiempo excedido se recolectan en cada una de las sondas ya que contienen el campo de IPID con el que se efectúa después la resolución de alias. Este proceso de resolución de alias utilizando este campo IPID necesita de varios valores para realizarse dado que se necesita inferir la evolución que siguen los IPIDs de una determinada dirección IP. Con la intención de obtener varias medidas de IPID para cada una de las direcciones IP pertenecientes a los distintos caminos se realizan I rondas de las medidas de Paris-traceroute en cada nodo sonda para cada uno de los distintos destinos.

De cada paquete de respuesta se recogen los siguientes datos:

- La dirección IP desde la que se envía el ICMP de respuesta.
- El TTL en el que se encuentra dicha dirección IP.
- El IPID recibido en el paquete de respuesta enviado por la dirección IP.
- El timestamp asociado al tiempo de recepción de dicho paquete de respuesta.

Para realizar la resolución mediante la técnica Pamplona-traceroute se hace uso de distintos tipos de paquete sonda para evitar problemas derivados con los filtrados y obtener una mejor completitud en la resolución de la red en estudio. Las medidas de Paris-traceroute se realizan mediante paquetes ICMP de *Echo request*, paquetes UDP dirigidos a un puerto aleatorio y paquetes TCP con el flag de *ACK* activado dirigidos a un puerto aleatorio. Los paquetes utilizados para la identificación son los paquetes de ICMP de *tiempo excedido en tránsito* recibidos desde cada uno de los routers de los distintos saltos debido a que el TTL del paquete sonda ha llegado a 0 sin alcanzar su destino final.

La utilización de la estrategia de Paris-traceroute a la hora de realizar las medidas está motivada porque evita la mayoría de balanceos de carga que son por flujo [22], de modo que todos los paquetes enviados en una ronda a un nodo sonda destino utilizan la misma ruta. Las distintas rondas necesarias para la realización del Pamplona-traceroute permiten ampliar el conocimiento de la red dado que cada

una puede tomar rutas diferentes obteniendo nuevas direcciones IP para la misma red.

En cada ronda se ejecutan los Paris-traceroutes hacia todos los nodos destino al mismo tiempo. Al terminar la primera ronda, que supone la terminación del Paris-traceroute más lento, se ejecuta la siguiente hasta completar un total de 50 rondas de Paris-traceroute. Una vez terminado este proceso para un tipo de paquete sonda, se inician las rondas para el siguiente tipo de paquete hasta completar las medidas para todos los tipos de paquete (ICMP, UDP y TCP).

Todos los nodos sonda utilizados en el proceso deben estar sincronizados para permitir la agregación y posterior comparación de las diferentes medidas tomadas desde estos.

Las siguientes fases se realizan de forma separada por tipo de paquete sonda, ya que el comportamiento de los IPIDs para distintos tipos de paquetes sonda pueden cambiar en el mismo router.

7.2.2 **Fase de pre-procesamiento**

Una vez que la fase de recolección ha terminado se procede a realizar una fase de pre-procesamiento. Esta fase permite la distinción de los diferentes comportamientos de los routers así como la ordenación temporal de los IPIDs. Tras este proceso se procede al uso de estrategias de resolución mediante la verificación de las secuencias de IPIDs que pertenecen al mismo router.

Esta fase de pre-procesamiento se puede realizar de dos formas diferentes, mediante un procesado centralizado o mediante un procesado distribuido. En cualquiera de los casos se necesita de una copia o acceso a los datos completos de las trazas recogidas por cada nodo sonda utilizado en la fase anterior.

Antes de entrar en la fase de resolución se realiza una verificación del comportamiento de los IPIDs de cada una de las direcciones IP. Estos comportamientos dependen de cada router concreto y se han observado los siguiente tipos de comportamiento: *incremental*, *random*, *randomT* y *reset*. Para la resolución de alias la tipología deseada es la incremental que es la que permite la identificación de parejas pertenecientes al mismo router por el alineamiento de sus IPIDs. Las demás

tipologías a pesar de no ofrecer la posibilidad de identificar las direcciones IP pertenecientes al mismo router, permiten diferenciar entre paquetes pertenecientes a distinto router cuando los valores de los IPIDs provenientes de dos direcciones IP tienen distinto comportamiento.

Además de la verificación del comportamiento de los IPIDs se realiza una ordenación temporal de los paquetes recibidos por los distintos nodos sonda. Se debe tener en cuenta que los IPIDs provenientes de routers de tipo incremental tienen relación con el instante de tiempo en el que han generado por lo que la ordenación temporal de los distintos paquetes antes de la fase de resolución final resulta importante.

Tras esta fase de pre-procesamiento se obtienen los IPIDs pertenecientes a las distintas direcciones IP ordenados temporalmente y divididos por comportamiento permitiendo que la fase de resolución se realice con mayor rapidez teniendo tan sólo que verificar la linealidad de las secuencias numéricas de aquellos IPIDs con un comportamiento incremental y no de todos ellos.

7.2.3 **Fase de resolución**

En esta fase se finaliza el proceso de resolución en el que se realiza el veredicto de si cada pareja de direcciones IP pertenece al mismo router. El proceso de alineamiento debe de tener en cuenta que los distintos IPIDs han sido recogidos en instantes de tiempos diferentes.

El procedimiento para decidir que dos direcciones IP pertenecen al mismo router es el siguiente:

1. La secuencia de IPIDs siguen un patrón creciente cuando se comparan conjuntamente IPIDs de las dos direcciones IP.

2. IPIDs en el mismo segundo no son tenidos en cuenta.

3. IPIDs distanciados más de 3 segundos no son tenidos en cuenta.

4. Deben existir al menos dos mezclas de incrementos en los IPIDs de la secuencia.

5. En caso de tener un desbordamiento del IPID se permite como máximo un offset entre IPIDs de 200 valores.

La regla número 1 es la consideración básica utilizada en casi todo proceso de resolución de alias mediante IPID. Los decrementos en la secuencia de IPIDs sólo son permitidos cuando se cumple la regla número 5, que es la encargada de resolver los problemas de desbordamiento en el campo de IPID.

Las reglas 2 y 3 tienen relación con la forma en la que se están obteniendo los IPIDs. La regla 2 evita tener problemas con IPIDs que se hayan recibido a la vez haciendo imposible la distinción de cual de ellos fue generado en origen primero. La regla 3 tiene relación con la distancia que pueden tener 2 IPIDs pertenecientes al mismo router cuando ha pasado mucho tiempo entre la generación de ambos. Se marca un límite temporal máximo para evitar que por la comparación de este tipo de IPIDs se derive en una conclusión falsa.

Por último la regla número 4 viene motivada porque debido al tiempo transcurrido entre los diferentes IPIDs puede ocasionar que las comparaciones entre los IPIDs se realicen siempre de IPIDs obtenidos por una dirección IP (IP_1) con IPIDs de otra (IP_2) y nunca se comparen los IPIDs obtenidos de IP_2 con los de IP_1. Esto puede pasar si el tiempo transcurrido entre la obtención de cada IPID de IP_1 y el siguiente IPID obtenido de IP_2 está dentro de los límites especificados, pero el tiempo transcurrido entre cada uno de los IPIDs de IP_2 y el siguiente de IP_1 está fuera de dicho rango ocasionando que se deseche la diferencia de valores que haya entre ellos. Para evitar esta situación, la regla 4 exige al menos el cumplimiento de dos mezclas de incrementos en la secuencia de IPIDs formada por ambas direcciones IP. Por ejemplo, en la siguiente secuencia de IPIDs para una pareja de direcciones IP $(IP_1\,5, IP_2\,10, IP_2\,12, IP_1\,14, IP_2\,16, IP_1\,18)$ se pueden obtener las siguientes parejas de IPIDs: $(IP_1\,5, IP_2\,10)$, $(IP_2\,12, IP_1\,14)$, $(IP_1\,14, IP_2\,16)$ y $(IP_2\,16, IP_1\,18)$. En esta secuencia de IPIDs se observan 3 mezclas de incrementos ya que la dirección IP que aparece en primer lugar en las secuencias cambia 3 veces. Debido a las reglas 2 y 3 no todos los incrementos son utilizados por lo que es necesario este requisito.

Tras esta descripción técnica del procedimiento a utilizar, en la siguiente sección se describe una serie de escenarios en los que se realizan las pruebas de resolución utilizando esta técnica.

7.3 Escenario de medida

La técnica de resolución Pamplona-traceroute se ha evaluado en escenarios que forman parte de las redes de interconexión de las plataformas Planetlab y ETOMIC. Para el caso de ETOMIC se han escogido un total de 6 nodos desde los que se realizan las medidas de Pamplona-traceroute obteniendo un total de 91 direcciones IP. En el caso de Planetlab se han elegido 4 escenarios diferentes, 3 de ellos obtenidos desde 15 nodos sonda y el último obtenido desde 50 nodos sonda de la misma plataforma. De los escenarios de Planetlab se han obtenido 370 (Planetlab1), 306 (Planetlab2) y 141 (Planetlab3) direcciones IP desde 15 nodos sonda y un último escenario mayor compuesto por 1.123 direcciones (Planetlab4) IP obtenido del compuesto por 50 nodos sonda.

Estos escenarios no disponen de información pública que permita verificar la precisión obtenida pero permiten evaluar los porcentajes de completitud obtenidos por la técnica. Una forma de poder realizar una verificación de la precisión hubiese podido ser la evaluación de Pamplona-traceroute en las redes Geant, Canet4 y GlobalNOC pero no se disponen de suficiente número de nodos sonda cuyas medidas atraviesen las direcciones pertenecientes al núcleo de estas tres redes. Por este motivo no se ha podido obtener mediante medidas indirectas un porcentaje suficientemente representativo de direcciones IP.

Para proveer de verificación al estudio realizado se han utilizado dos vías, la primera ha sido la utilización de otra técnica de resolución para verificar la resolución obtenida por el Pamplona-traceroute. La técnica Ally-based ofrece porcentajes de resolución altos y su precisión ha sido comprobada en capítulos anteriores. De modo que los resultados de resolución obtenidos por la técnica Pamplona-traceroute serán comparados con los obtenidos por la técnica Ally-based para observar si se produce alguna identificación errónea. La segunda vía para tener una medida de la precisión ha sido la realización de medidas en una maqueta completamente controlada. Esta maqueta se compone de 9 routers cisco y una máquina

Linux que actúa como router. Es un escenario reducido compuesto por 25 direcciones IP diferentes que implica la obtención de la resolución de alrededor de 300 parejas de direcciones IP. La red puede ser totalmente revisada mediante los paquetes sonda indirectos generados desde 4 nodos sonda localizados estratégicamente dentro de esta. La estructura completa de la maqueta puede verse en la figura 7.1.

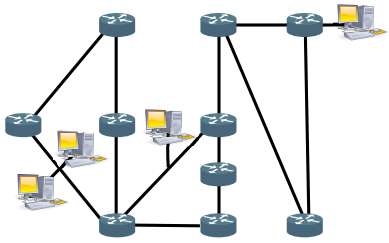


Figura 7.1: Escenario maqueta del que se conoce su estructura de red utilizado en las medidas de verificación

En la siguiente sección la técnica Pamplona-traceroute se evalúa según su completitud, precisión, eficiencia y distribuibilidad en los escenarios presentados.

7.4 Evaluación de la técnica Pamplona-traceroute

En esta sección se realiza una evaluación de la técnica Pamplona-traceroute. Para ello se emplean las 4 métricas conocidas: completitud, precisión, eficiencia y distribuibilidad. A continuación se detalla la evaluación de cada métrica.

7.4.1 Completitud y precisión

La completitud de la técnica Pamplona-traceroute viene limitada por la cantidad de medidas útiles obtenidas en la fase de recolección. Un primer aspecto a analizar son las tasas de respuesta por tipo de paquete sonda. En el escenario más grande, catalogado como Planetlab4, los porcentajes de paquetes sonda para los que no se obtiene respuesta son 3.78 % para paquetes sonda de tipo ICMP, 6.04 % para paquetes sonda de tipo UDP y 3.28 % para paquetes sonda de tipo TCP.

A pesar de todo, la contestación de un router no es condición suficiente para poder realizar la resolución de alias ya que en ella entra en juego el tipo de comportamiento que tienen los routers que se intentan identificar a la hora de generar

los IPIDs de los paquetes de respuesta. En la tabla 7.1 se muestran los comportamientos de tipo incremental (que permiten la resolución de parejas que son alias) y aquellos que no lo son (que no permitan la inferencia de alias). Las columnas presentan los porcentajes de cada comportamiento según el tipo de paquete y el escenario. La etiqueta no incremental se corresponde con la unión de los comportamientos *random*, *randomT*, *zero* y *copy* estudiados en el capítulo 3. Como se puede observar las tasas de direcciones IP que provienen de routers con un comportamiento incremental es bastante alto para el caso particular de ICMP en todos los escenarios. El resto de tipos de paquetes sonda ofrecen tasas distintas para cada escenario particular.

Network	*ICMP: % incremental*	*ICMP: % not incremental*	*UDP: % incremental*	*UDP: % not incremental*	*TCP: % incremental*	*TCP: % not incremental*
Etomic	40.00	60.00	42.04	57.95	40.24	59.75
Planetlab subset 1	32.86	67.13	18.32	81.67	38.67	61.32
Planetlab subset 2	47.76	52.23	31.57	68.42	52.23	47.76
Planetlab subset 3	29.77	70.22	19.14	80.85	47.45	52.54
Planetlab subset 4	32.19	67.80	30.27	69.72	35.97	64.02

Tabla 7.1: Porcentajes de comportamientos para el campo de IPID de las distintas direcciones IP al utilizar los paquetes sonda de Pamplona-traceroute en los distintos escenarios

Dado que se conoce el TTL de cada dirección IP se ha realizado una medida del número de salto donde se concentran los comportamientos que se corresponden con las direcciones IP pertenecientes a routers no incrementales. En la figura 7.2 se presenta para cada número de salto (eje horizontal), qué porcentaje de direcciones IP de ese salto (eje vertical) no se corresponden con un comportamiento incremental. Se observa cómo para valores de TTL que se corresponden con routers de acceso (TTL situados a los extremos del escenario) el número de routers con comportamiento no incremental tiene valores bajos, mientras que para TTLs correspondientes al núcleo de Internet los valores no incrementales ascienden de manera considerable.

Esta figura muestra que para el reconocimiento de alias, esta técnica centra su capacidad de resolución en los routers que forman parte de los routers de acceso. Esto limita la completitud de esta técnica en los routers pertenecientes al núcleo de Internet.

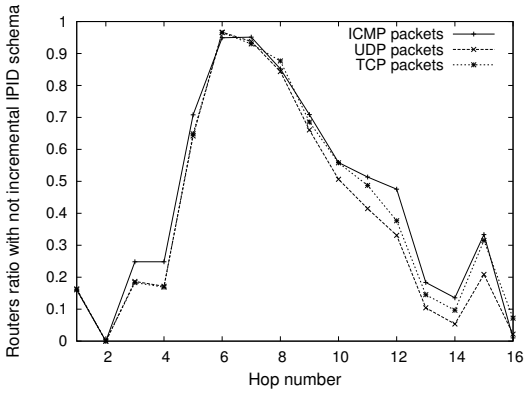


Figura 7.2: Histograma normalizado del número de routers con comportamiento no incremental según el TTL utilizando los paquetes sonda de Pamplona-traceroute

Otra forma de representación es la que se muestra en la figura 7.3 donde se presenta la función de probabilidad acumulada para los diferentes TTLs (eje horizontal) de la probabilidad de encontrar un comportamiento no incremental.

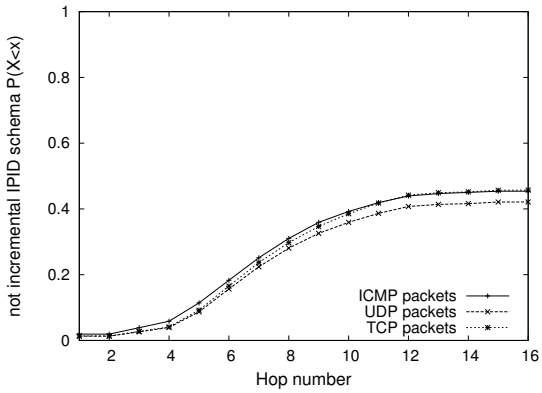


Figura 7.3: CDF del número de routers con comportamiento no incremental según el TTL para los paquetes sonda de Pamplona-traceroute

De nuevo se vuelve a observar que el crecimiento de la probabilidad de encontrar un comportamiento no incremental se acumula en el centro de la figura donde se sitúan los routers pertenecientes al núcleo de Internet. No obstante, la capacidad de identificación de esta técnica puede combinarse con otras para obtener una identificación total mayor. Si observamos la figura 7.4 que presenta la función de probabilidad acumulada análoga a la anterior pero para la técnica Ally-based se puede observar como las dos técnicas son complementarias situando los tramos

donde Pamplona-traceroute ofrece buenas tasas de resolución en los routers de acceso y Ally-based que ofrece buenas tasas de identificación en todos los TTLs.

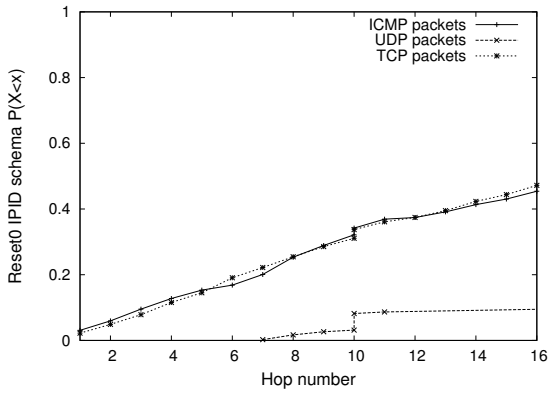


Figura 7.4: CDF del número de routers con comportamiento no incremental según el TTL para los paquetes sonda de Ally-based

Una resolución completa ofrecida por una técnica combinada se puede idear en base a una fase de descubrimiento realizada mediante la técnica Pamplona-traceroute y una fase de resolución posterior complementada mediante Ally-based. Gracias a la combinación de la resolución ofrecida por las dos técnicas combinadas pueden conseguirse altas tasas de completitud que se muestran mas adelante en esta misma sección.

El único escenario de medida con información de la estructura de red real donde se han realizado las medidas ha sido en el escenario maqueta. Los resultados para la resolución de alias para Pamplona-traceroute y para una selección de las demás técnicas de resolución se presenta en la tabla 7.2. Como en otros estudios realizados con anterioridad en la tabla se muestran columnas con los alias (positivos), no alias (negativos) y total de parejas identificadas (positivos más negativos). Al conocer totalmente la estructura de red también se han incluido columnas que muestran el porcentaje de falsos positivos (parejas catalogadas como alias pero que no lo son) y falsos negativos (parejas catalogadas como no alias pero que realmente lo son) así como el porcentaje total de alias resueltos. Se puede observar como para este caso particular, la técnica Pamplona-traceroute ofrece la mejor identificación en comparación con el resto de técnicas que se han medido. Mediante esta

técnica se obtiene un 89.85 % de completitud y se identifican un 94.44 % del total de alias existentes. Al contrario de lo que puede verse en otras técnicas como Palmtree, Prespecified-timestamp y TraceNet, esta técnica no adolece de resoluciones erróneas en esta red.

Method	*Positives %*	*Negatives %*	*False positives %*	*False negatives %*	*Completeness %*	*Aliases %*
Pamplona-traceroute	6.15	83.69	0	0	89.85	94.44
Ally-based methods	4.71	44.56	0	0	49.27	72.22
Palmtree	3.62	-	00.72	-	3.62	55.55
Prespecified-timestamps	0.36	5.07	0	00.72	5.43	5.55
Radargun	0.36	9.78	0	0	10.14	5.55
TraceNET	3.26	-	00.72	-	3.26	50.00

Tabla 7.2: Comparativas de completitud y precisión en el escenario maqueta utilizado

A pesar de que los resultados de completitud en esta red concreta son muy buenos, los resultados para redes grandes de Internet también resultan interesantes.

Se ha realizado una evaluación de Pamplona-traceroute en los escenarios ETOMIC y Planetlab 1-4 en los que se ha realizado una verificación de los alias y no alias identificados por la técnica Pamplona-traceroute por comparación con la técnica Ally-based. Los resultados se pueden ver por tipo de paquete en las tablas 7.3-7.5. En las tablas se muestran medidas de identificación total (suma de positivos más negativos) obtenidos por la técnica de resolución, el número de direcciones IP de cada escenario (con cada tipo de paquete el mismo escenario puede tener un número diferente de direcciones IP) y una medida del porcentaje de falsos positivos y falsos negativos.

Network	*IP addresses*	*Identification %*	*False positives %*	*False negatives %*
Etomic	85	66.32	0.00	0.00
Planetlab 1	286	59.31	0.00	0.18
Planetlab 2	224	62.31	0.00	0.10
Planetlab 3	131	50.31	0.00	0.00
Planetlab 4	963	55.31	0.00	0.04

Tabla 7.3: Tasas de completitud y precisión para paquetes ICMP en Pamplona-traceroute

Tal y como se esperaba las tasas de completitud obtenidas son más reducidas que las obtenidas en el escenario maqueta. Se debe tener en cuenta que en este caso

Network	*IP addresses*	*Identification %*	*False positives %*	*False negatives %*
Etomic	88	66.02	0.00	0.00
Planetlab 1	322	60.38	0.00	0.00
Planetlab 2	266	62.07	0.00	0.05
Planetlab 3	141	56.19	0.00	0.00
Planetlab 4	1004	53.38	0.00	0.01

Tabla 7.4: Tasas de completitud y precisión para paquetes UDP en Pamplona-traceroute

Network	*IP addresses*	*Identification %*	*False positives %*	*False negatives %*
Etomic	82	65.83	0.00	0.00
Planetlab 1	287	56.25	0.00	0.03
Planetlab 2	268	50.82	0.00	0.02
Planetlab 3	118	59.82	0.00	0.37
Planetlab 4	970	43.86	0.00	0.09

Tabla 7.5: Tasas de completitud y precisión para paquetes TCP en Pamplona-traceroute

se han separado las identificaciones por tipo de paquete, pero las tasa de completitud han bajado de un 89,85 % a tasas de completitud de alrededor de un 55 % (la tasa mínima es de un 43,86 % y la tasa máxima alcanza un 66,32 %). No se observan grandes diferencias entre las tasas de completitud obtenidas aunque las más altas, como ha estado observándose en estudios anteriores, son las ofrecidas por los paquetes sonda de tipo ICMP. Por otro lado, en lo que concierne la precisión, en todas ellas la tasa de falsos positivos es nula pero las tasas de falsos negativos alcanza tasas de un 0,37 % como sucede en el escenario Planetlab 3 de la tabla 7.5.

La técnica Pamplona-traceroute utiliza un agregado de las resoluciones obtenidas mediante todos los tipos de paquetes, por lo que en la tabla 7.6 se presenta un agregado de la completitud ofrecida por las tres resoluciones correspondientes a cada tipo de paquete sonda enviado. La tabla se divide en 5 columnas que muestran el número total de direcciones IP que resultan del agregado en cada escenario, la resolución ofrecida mediante el uso de cada tipo de paquete sonda y por último la resolución final que ofrece la técnica Pamplona-traceroute. Como se puede observar, los resultados de identificación final ascienden a un 60-80 %.

En comparación con otras técnicas del estado del arte, como por ejemplo Radargun,

Network	*Number of vantage points*	*Total IP addresses*	*ICMP completeness over total %*	*UDP completeness over total %*	*TCP completeness over total %*	*Total completeness %*
Etomic	6	91	57.0	60.3	52.0	70.0
Planetlab 1	15	370	35.35	45.67	33.76	66.83
Planetlab 2	15	306	33.27	46.83	38.93	72.86
Planetlab 3	15	141	44.56	57.84	38.12	81.85
Planetlab 4	50	1123	40.65	42.65	38.92	62.83

Tabla 7.6: Tasas de completitud para el agregado total de tipos de paquetes sonda utilizados por Pamplona-traceroute

la técnica Pamplona-traceroute ofrece unas tasas altas de resolución y unas tasas de errores en la identificación muy reducidos. En la tabla 7.7 se presentan los resultados de precisión equivalentes para la técnica Radargun en los mismos escenarios. La tabla muestra los porcentajes de identificación, falsos positivos y falsos negativos obtenidos por la técnica. Observando las tasas de resolución, la técnica Pamplona-traceroute ofrece mayores tasas de completitud ofreciendo al mismo tiempo tasas notablemente inferiores de errores en la resolución. Por ejemplo para el caso de Planetlab4 las tasas de completitud obtenidas por Pamplona-traceroute son de un 62.83 % mientras que las tasas obtenidas mediante Radargun sólo alcanzan un 26.74 % del total.

Network	*Completeness %*	*False positives %*	*False negatives %*
Etomic	13.74	0.00	0.16
Planetlab 1	25.80	7.89	0.01
Planetlab 2	23.19	44.76	0.04
Planetlab 3	18.33	0.00	0.00
Planetlab 4	26.74	25.85	0.00

Tabla 7.7: Tasas de completitud y precisión obtenidas con Radargun

Adicionalmente a las medidas de completitud obtenidas solamente mediante Pamplona-traceroute se ha realizado una medida que combine la utilización de esta con la técnica Ally-based. En las figuras 7.3 y 7.4 se observa que las técnicas Pamplona-traceroute y Ally-based parecen poder combinarse por lo que se ha realizado un estudio de las tasas de completitud obtenidas mediante la utilización de estas dos técnicas combinadas en los mismo escenarios utilizados para los estudios anteriores. La tabla 7.8 muestra la completitud obtenida por la técnica Pamplona-traceroute, las parejas que necesitan de realizar el proceso mediante Ally-based

porque no han sido identificadas y por último la completitud ofrecida por la combinación de ambas para los escenarios ETOMIC, y Planetlab 1-4.

Network	*Pamplona-traceroute identif. %*	*% of pairs to check with Ally-based methods*	*Combined identification %*
Etomic	70.08	29.91	93.06
Planetlab 1	66.83	33.17	81.26
Planetlab 2	72.86	27.14	72.96
Planetlab 3	81.85	18.15	88.11
Planetlab 4	62.83	37.17	81.27

Tabla 7.8: Tasas de identificación combinando las resoluciones obtenidas por las técnicas Pamplona-traceroute y Ally-based

La tabla muestra que las tasas de completitud obtenidas de la combinación de Pamplona-traceroute y Ally-based alcanzan el 93,06 % en la red de interconexión de ETOMIC. Las tasas varían desde porcentajes de un 72,96 % a porcentajes de un 93,06 % pero en media la completitud obtenida tiende hacia un 82 %. Los resultados son muy interesantes ya que permiten obtener una completitud muy alta realizando las medidas correspondientes a la técnica Ally-based solamente en las parejas que no han sido previamente identificadas, reduciendo el uso de esta a solamente un 18.15 %-37.17 % de las parejas totales.

7.4.2 **Eficiencia**

La eficiencia muestra si una técnica dada consume mucho tiempo o introduce mucho tráfico en la red en la realización de sus medidas. En esta sección se realiza un estudio de la eficiencia que ofrece la técnica Pamplona-traceroute.

Los paquetes sonda utilizados por Pamplona-traceroute son paquetes de tamaño mínimo (64 bytes) y se envían hasta un máximo de N paquetes por cada Paris-traceroute, este es el TTL máximo con el que se enviarán los paquetes desde cada nodo sonda antes de considerar que el nodo destino no responde a los paquetes sonda enviados. Por cada salto se envían un total de H paquetes sonda, de manera que se obtienen H paquetes de respuesta por cada dirección IP del camino del que se estén obteniendo las respuestas. El planteamiento de Pamplona-traceroute se basa en su ejecución en una serie de nodos sonda entre los que se realizan las

medidas de Paris-traceroute, por lo que si se dispone de M nodos sonda, desde cada uno se realizan M-1 medidas. Por último, esta técnica utiliza 3 tipos diferentes de paquetes sonda para los que se repite el proceso y se realizan I rondas, por lo que la ecuación final a partir de la que se obtiene el total de bytes introducidos en la red es: $I*(M-1)*N*H*64*3$. Si se tiene en cuenta que H se fija en 1 porque en principio las rutas van a ser estables por la utilización del Pamplona-traceroute, N se ha fijado en 30 tal y como se hace en la mayoría de implementaciones de traceroute y el I utilizado es 50 el gasto total introducido dependiente del número de nodos sonda disponibles (M) la ecuación toma la siguiente forma $50*(M-1)*30*1*64*3 = (M-1)*288{,}000$

El número de rondas (I) puede ser ajustado dependiendo de la precisión que se desea obtener. Con un valor mayor de I se obtiene mayor número de IPIDs de los paquetes de respuesta debido al mayor número de rondas y por tanto mayor número de paquetes sonda enviados lo que deriva en una mejor identificación. La figura 7.5 muestra tanto los porcentajes de completitud obtenidos, falsos positivos y negativos respecto al número total de parejas (eje vertical). Se puede observar como a medida que se aumenta el número de iteraciones (eje horizontal) se incrementa la completitud ofrecida por la técnica y se reduce el número de errores en la identificación. Se puede ver que un valor de aproximadamente 20 iteraciones provee un compromiso, permitiendo obtener valores de completitud cercanos a 0.45 que es el valor donde parece estabilizarse la completitud según lo observado en la figura 7.5.

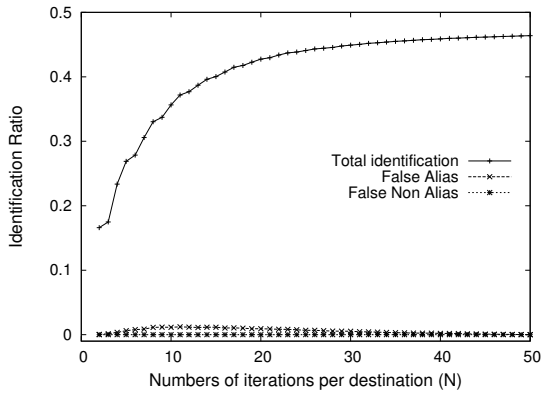


Figura 7.5: Resultados de resolución de alias número de iteraciones mediante Pamplona-traceroute utilizando paquetes sonda ICMP

En la figura 7.6 se muestra el CCFD de la duración de las medidas de Pamplona-traceroute para los 3 tipos de paquetes utilizados. La principal diferencia observada es que las curvas de TCP e ICMP siguen patrones diferentes a la obtenida para UDP. Esto sucede porque normalmente las direcciones IP escogidas como destino de cada Pamplona-traceroute contestan a los paquetes UDP con paquetes ICMP de error de tipo *Port Unreachable*, pero en el caso de los ICMP o los TCP los paquetes no llegan al destino final siendo filtrados por algún equipo de la red de acceso a los nodos destino o los propios nodos destino simplemente no responden a los paquetes enviados. Esto causa que el número de paquetes sonda enviados para el Paris-traceroute se prolongue hasta el máximo de saltos a comprobar (fijado en 30) y por tanto crezca el tiempo invertido en la ejecución de las medidas para esos tipos de paquete concretos. Observando los datos temporales obtenidos y dependiendo del número de iteraciones seleccionado, la fase de recolección de datos de la técnica Pamplona-traceroute se puede prolongar por minutos.

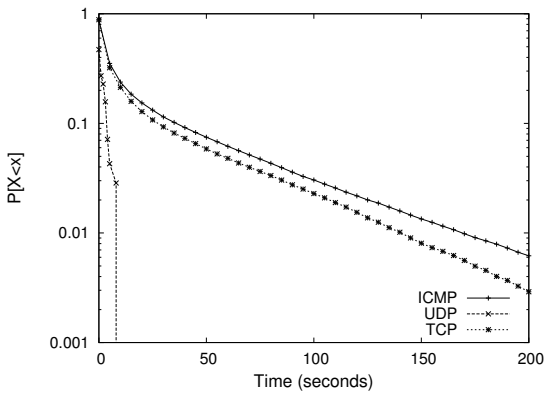


Figura 7.6: Funcion de probabilidad acumulada complementario del tiempo necesario para realizar cada traceorute en Pamplona-traceroute

7.4.3 **Distribuibilidad**

La técnica Pamplona-traceroute puede ser distribuida en varios nodos sonda. En la fase de recolección de datos, la distribución es necesaria para obtener los datos mediante los que la técnica realiza la resolución. De esta fase distribuida se obtienen

las direcciones IP de la red que se desea resolver y los datos necesarios para realizar la resolución de las direcciones IP obtenidas (medidas necesarias previas para la fase de resolución).

La fases de recolección de datos y la de resolución mediante Pamplona-traceroute pueden ser distribuidas o centralizadas. Todas las posibles combinaciones de las direcciones IP de la red deben de ser verificadas, y ese proceso puede realizarse directamente en una sola máquina o distribuirse y que un número de maquinas se reparta el procesado de las diferentes parejas. Dado que para la realización de la técnica se requiere de la utilización de varios nodos sonda, en estos mismos nodos sonda pueden también ser ejecutadas las fases de recolección de datos y de resolución. A medida que se añaden nodos al proceso de resolución el tiempo que invierten en el procesado se reduce linealmente, lo que permite un escalado sencillo del problema.

Todo nodo utilizado para la fase de resolución debe tener disponible los datos pertinentes de las direcciones IP de las parejas que se deseen identificar en dicho nodo. El proceso más simple para que los nodos tengan disponibles dichos datos es el de copiar el total de los datos de medidas a todos los nodos que se utilicen para el procesado. Por tanto esta técnica permite su distribución en todas sus fases.

7.5 Conclusiones

La técnica Pamplona-traceroute posibilita la agregación de las fases de descubrimiento y resolución lo que supone una gran ventaja tanto en tiempo invertido como en tráfico introducido en la red.

Las tasas de completitud obtenidas son altas obteniendo tasas en los escenarios utilizados entre un 62,83 y un 81,85 % por la utilización de medidas de tipo indirecto aprovechando tres tipos diferentes de paquetes ICMP, UDP y TCP.

La técnica adolece de falsos negativos verificados a través de la técnica Ally-based, pero estos porcentajes de error están muy por debajo de las tasas de error producidas por otras técnicas utilizadas en el estado del arte como Radargun.

La resolución de las parejas que son alias en esta técnica funciona mejor para redes de acceso a Internet que para los routers pertenecientes al núcleo de Inter-

net debido a los comportamientos no incrementales que estos últimos tienen en la generación del campo de IPID.

Se ha realizado una combinación de la resolución obtenida por la técnica Ally-based y Pamplona-traceroute, y se han obtenido tasas de identificación entre un 72.96 y un 93.06 %. Estas tasas de completitud con porcentajes muy altos lo que supone que se esta identificando la red casi por completo.

CAPÍTULO

8

Conclusiones y líneas futuras

8.1 Conclusiones

El descubrimiento de topologías y la resolución de alias son útiles para multitud de aplicaciones y han sido ampliamente estudiados en el estado del arte. Este trabajo se centra en la parte de resolución de alias.

Tras analizar las técnicas de resolución de alias que aparecen en el estado del arte, se puede llegar a la conclusión de que sigue siendo un problema abierto y no existe una técnica concreta de resolución de alias que sea claramente mejor. Las técnicas que utilizan estrategias activas ofrecen tasas mayores de completitud y de precisión pero requieren de la introducción de tráfico en la red. Las técnicas que utilizan estrategias de inferencia no requieren de la introducción de tráfico en la red por lo que en redes en las que las tasas de contestación de los routers sean bajas o en las que no sea posible realizar medidas activas son las que se pueden utilizar.

La problemática a la hora de valorar las métricas para las distintas técnicas de resolución de alias es la de que no existe un estándar para evaluar las distintas métricas, ni siquiera la métrica de completitud que permite medir el número total de parejas identificadas. En nuestro caso la completitud se mide con el porcentaje de parejas para las que las técnicas emiten un veredicto de positivo o negativo respecto al total de parejas. Esta medida de completitud permite conocer cuando

se ha terminado de identificar por completo una red. La falta de uniformidad de las métricas ocasiona que no se pueda realizar una comparación objetiva de las distintas técnicas publicadas en distintos estudios. Otro problema que se encuentra a la hora de realizar una valoración de las distintas métricas es que la mayor parte de las herramientas del estado del arte que implementan técnicas de resolución de alias y sus trazas asociadas no están disponibles públicamente lo que ocasiona que se haya tenido que realizar una reimplementación de las técnicas para poder evaluarlas.

Es importante que las nuevas técnicas puedan identificar tanto las parejas de direcciones IP que pertenecen al mismo router como las que no con el objetivo de saber cuando se ha conseguido una resolución de alias completa de la red que se desea estudiar.

La mayoría de las técnicas de resolución no obtienen todo el potencial de identificación que ofrecen los distintos parámetros base de las respuestas válidas obtenidas. De los distintos tipos de medidas activas realizadas, las que mejores tasas de contestación ofrecen son las que utilizan paquetes sonda indirectos, pero a pesar de ello son las técnicas directas las que ofrecen mejores tasas de contestaciones válidas para su utilización en una resolución de alias. Los paquetes ICMP son los que son sometidos a menos filtrados y por tanto los que mejores tasas de contestación proporcionan. El parámetro base IPID tanto en su versión directa como indirecta proporciona buenos valores de contestaciones válidas siendo el parámetro base que mejores tasas de completitud permite obtener.

Se ha identificado una problemática relativa al tiempo entre paquetes de la técnica de resolución de alias Radargun que hace que la resolución de alias que ofrece se vea muy deteriorada y se ha propuesto una regla de dimensionamiento que permite mejorar sus resultados. Esta regla de dimensionamiento permite saber qué ancho de banda utilizar a la hora de realizar las medidas de Radargun en una red con un número de direcciones IP. Los resultados de las resoluciones basadas en medidas realizadas fuera de las franjas temporales marcadas por la regla de dimensionamiento calculada se ven deterioradas y ofrecen peores resultados de completitud.

Se ha identificado el tiempo máximo que permanecen los routers sin ningún cambio en la configuración de sus direcciones IP fijando este valor en un periodo

medio de 30 días. Gracias a ello se ha realizado una propuesta que permite la realización de las técnicas de resolución de alias con agregación por parejas mediante la distribución en el tiempo y el espacio de las medidas dentro de dicho periodo de tiempo mejorando de esta forma la completitud obtenida por parte de las técnicas de resolución de alias para redes de gran tamaño al posibilitar la utilización de este tipo de técnicas con agregación por parejas.

Se ha realizado una propuesta de estrategia de reducción llamada *IP-Offset* que se basa en el valor absoluto de la diferencia entre las direcciones IP de una pareja de direcciones expresadas en forma numérica. Esta estrategia de reducción permite mediante técnicas de clustering la preselección de las parejas con más probabilidad de ser alias permitiendo reducir el número de medidas a realizar. Se han conseguido que utilizando sólo un 10 % del total de parejas posibles que se pueden formar con las direcciones IP de la red se obtengan tasas de completitud cercanas a un 73 %. Esta estrategia de reducción además de estar entre las que ofrecen mejores tasas de reducción dentro de las analizadas en el estado del arte no requiere de la introducción de tráfico adicional al utilizado durante la fase de descubrimiento, lo que la diferencia del resto de propuestas.

Otra propuesta realizada es la técnica de resolución de alias Ally-based. Esta técnica de resolución de alias es una variación de la técnica original Ally en la que se cambia el número total de paquetes sonda utilizados, el patrón de envío utilizado y se amplía el tipo de paquetes a utilizar. Las tasas de completitud se ven incrementadas de tasas cercanas al 7,40 % cuando se utiliza la técnica original a tasas cercanas al 63,17 % cuando se aplica la técnica propuesta.

Por último se ha propuesto una técnica de resolución llamada Pamplona-traceroute que no necesita de medidas activas en la fase de resolución e incrementa la completitud obtenida sin aumentar de manera considerable el tráfico introducido en la red en la fase de descubrimiento. Las tasas de completitud obtenidas mediante esta técnica oscilan entre un 62,83 % y un 81,85 % del total de parejas. Además la técnica Pamplona-traceroute puede ser combinada con la técnica Ally-based pudiendo mejorar sustancialmente la identificación consiguiendo tasas de completitud entre el 72.96 % y el 93,06 %.

Tras el trabajo realizado se han obtenido mejoras notables en el proceso de resolución de alias que permiten que las redes inferidas a través de las medidas

activas tengan una estructura más realista. A pesar de los avances conseguidos en el área estos años, el problema aun no se ha solucionado por completo y no existe técnica a día de hoy que permita una completitud del 100 % con la que se pueda asegurar que la red inferida es igual a la real. Esto quiere decir que se puede seguir trabajando en las técnicas de resolución para poder llegar a inferir la estructura de las redes completamente.

8.2 Líneas futuras

Como se ha comentado previamente el problema de la resolución de alias no ha sido resuelto por completo, y las redes, routers, protocolos y fabricantes van cambiando a medida que pasa el tiempo.

Todos los comportamientos de los parámetros base que se han observado en los paquetes de respuesta no se utilizan en las distintas técnicas. Sería interesante identificar los factores que hacen que no se aproveche todo el potencial de identificación ofrecido por los comportamientos de los parámetros base en las técnicas de resolución de alias que se utilizan actualmente así como idear nuevas técnicas de resolución de alias que aprovechen otros comportamientos de los parámetros base que ahora no se utilizan. A modo de ejemplo, centrándose en el caso particular del parámetro base timestamp, se puede estudiar la deriva de reloj en base a los paquetes de respuesta obtenidos de los distintos routers e identificar como pertenecientes al mismo router a aquellos cuya deriva sea similar.

Por otro lado, una posible vía de estudio puede ser el obtener nuevos parámetros base cuyos comportamientos particulares permitan la resolución de nuevas parejas de direcciones IP o la reducción del número de parejas a utilizar para realizar una resolución.

Una importante parte que no se ha estudiado apenas en este trabajo es la fase de descubrimiento. Esta fase se compone de la identificación de enlaces y el descubrimiento de direcciones IP de los routers de la red. Las técnicas de descubrimiento no han evolucionado notablemente desde la aparición del traceroute de Van Jacobson. En el trabajo de Paris-traceroute se propuso una serie de cambios que permiten al traceroute seguir una ruta sin sufrir balanceos de ruta por flujo, pero aun sigue

teniendo los problemas derivados de los balanceos de ruta por paquete y los problemas derivados de routers que no contestan a los paquetes sonda.

El actual despliegue y utilización de las redes IPv6 y las características particulares tanto de este nuevo protocolo de red así como de los routers con soporte IPv6, abren un campo de investigación importante en el que trabajar para la identificación de topologías de red. La mayoría de técnicas deberán reimplementarse o simplemente no podrán utilizarse dado que este nuevo protocolo contiene variaciones de los parámetros base que se utilizan para la resolución de alias en redes IPv4.

Existen otras líneas de investigación que se pueden derivar de los mapas de red a nivel de router obtenidos mediante las técnicas de resolución de alias. Las áreas en las que pueden utilizarse dichos mapas varían desde temáticas sociológicas, ya que dichos mapas de red representan relaciones corporativas entre distintas compañías y varían a medida que los avances tecnológicos permiten la expansión y mejora de la comunicación con lugares donde antes no se tenía acceso, a temáticas más técnicas como la geolocalización de los equipos de una red determinada.

CAPÍTULO

A

Articulos publicados

A.1 Resolución de alias para el cálculo de topologías

- Autores: S. García, E. Magaña, M. Izal y D. Morató.

- Publicado: VI Jornadas de ingeniería telemática, Jitel 2007, Málaga del 17 al 19 de Septiembre del 2007.

- Abstract: The network topology is a fundamental parameter for managers and researchers. The traditional methodology for discovering the topology of a network is based on the tool traceroute, used from several vantage points in different subnetworks. The result is a set of sink trees where the nodes are the discovered IP addreses from the routers. However, few tools have faced the problem of identifying the nodes in different sink trees as interfaces in the same router. This paper shows a new methodology for this problem of alias resolution. It has been used in the european research network using the ETOMIC platform. It shows that the traditional methodologies are not effective in today's networking scenario but can be easily improved at least in a factor of 3 in the number of successes.

A.2 Techniques for better alias resolution in Internet topology discovery

- Autores: S. García-Jiménez, E. Magaña, D. Morató and M. Izal.
- Publicado: 11th IFIP/IEEE International Symposium on Integrated Network Management, IM2009, 1 - 5 June 2009, Long Island, New York, USA.
- Abstract: One of the challenging problems related with network topology discovery in Internet is the process of IP address alias identification. Topology information is usually obtained from a set of traceroutes that provide IP addresses of routers in the path from a source to a destination. If these traceroutes are repeated between several source/destination pairs we can get a sampling of all IP addresses for crossed routers. In order to generate the topology graph in which each router is a node, it is needed to identify all IP addresses that belong to the same router. In this work we propose improvements over existing methods to obtain alias identification related mainly with the types and options in probing packets.

A.3 Improving Efficiency of IP Alias Resolution based on Offsets between IP Addresses

- Autores: S. García-Jiménez, E. Magaña, D. Morató and M. Izal.
- Publicado: 21st International Teletraffic Congress, ITC21, 15-17 September 2009 in Paris, France.
- Abstract: In order to get a router-level topology in Internet, IP address alias resolution techniques allow to identify IP addresses that belong to the same router. There are several proposals to make this identification, some based on active measurements and others based on inference studies. The former provides more accuracy and completeness, however efficiency is very low because of the high number of probes needed. These methods probe IP addresses in pairs. With thousands or even more IP addresses to check for aliases,

the number of tests gets too high. In order to reduce the number of probes, we propose to select the pairs of IP addresses to test for aliasing using information available a priori. This selection will be based on the offset (numerical distance) between the IP addresses to test. We will show that we can improve efficiency of active alias identification with almost no loss on completeness and without generating probing traffic. The technique is also adaptable to a distributed measurement scenario.

A.4 IP addresses distribution in Internet and its application on reduction methods for IP alias resolution

- Autores: S. García-Jiménez, E. Magaña, M. Izal and D. Morató
- Publicado: 4th IEEE LCN Workshop on Network Measurements, WNM 2009, Oct. 23rd 2009 Zürich, Switzerland.
- Abstract: Discovery of Internet topology is an important and open task. It is difficulted by the high number of networks and internetworking equipments, and even by the dynamic of those interconnections. Mapping Internet at router-level needs to identify IP addresses that belong to the same router. This is called IP address alias resolution and classical methods in the state of the art like Ally need to test IP addresses in pairs. This means a very high cost in traffic generated and time consumption, specially with an increasing topology size. Some methods have been proposed to reduce the number of pairs of IP addresses to compare based on the TTL or IP identifier fields from the IP header. However both need extra traffic and they have problems with the probing distribution between several probing nodes. This paper proposes to use the distribution of IP addresses in Internet Autonomous Systems in order to reduce the number of IP addresses to compare. The difference between pairs of IP addresses is used to know a priori if they are candidates to be alias with certain probability. Performance evaluation has been made using Planetlab and Etomic measurement platforms. The paper justifies the reduction method, obtaining high reduction ratios without injecting extra traffic in

the network and with the possibility to distribute the process for alias resolution.

A.5 On the performance and improvement of alias resolution methods for Internet core networks

- Autores: S. Garcia-Jimenez, E.magaña, D. Morató and M. Izal.
- Publicado: Annals of Telecommunications 2011, vol. 66, no1-2, pp. 31-43, ISSN 1958-9395.
- Abstract: Internet is a huge interconnection of thousands of networks with different technology, equipment, configuration and administrative owner. This, added to the lack of public information about those individual infrastructures, makes a difficult task to provide a so called Internet map: a topological map with information of routers, in- terconnections between routers and IP addressing configuration. Traditional topology discovery methods based on traceroutes only provide IP addresses in the path between end-nodes. Some of those IP addresses can belong to the same router, and this iden- tification is made by alias resolution methods. Therefore, alias resolution allows to provide router-level map of the Internet with important applications in network simu- lation, protocol design, network management, network security, network service design and geolocation. In this paper, alias resolution methods are analyzed in Internet core networks (GlobalNOC, Canet4 and Geant). This allows to identify peculiar behaviors in these core networks, improving alias resolution methods. Simultaneously, reduction methods are used to decrease the number of probing packets in alias resolution meth- ods.

A.6 Probing distribution in time and space for IP alias resolution

- Autores: S. Garcia-Jimenez, E. Magaña, D. Morató and M. Izal.
- Publicado: Enviado a Journal of Internet Technology JIT.

- Abstract: The Internet is composed of thousands of networks, interconnected to provide end-to-end IP (Internet Protocol) connectivity. However, Very little public information is provided about these networks and their interconnections. The information needed to create an Internet map of the routers and the links between those routers must be derived from techniques for discover- ing IP addresses (traceroute) and for associating IP addresses that belong to the same router (IP aliases). Both processes IP address discovery and IP alias resolution require a large measurement infrastructure, and they introduce variable amounts of traffic into the network. Although systematic proposals have been made for creating a scalable IP address discovery system, the equivalent system for resolving IP aliases is far from being determined. In this paper, new proposals for obtaining a scalable IP aliasing system are evaluated and compared with existing solutions. Distributing measurements along multiple vantage points (spatial distribution) and extending probing tasks over time (temporal distribution) have been identified as the key methods for reducing the overhead of IP alias resolution.

A.7 Pamplona-traceroute: topology discovery and alias resolution to build router level Internet maps

- Autores: S. Garcia-Jimenez, E. Magaña, D. Morató and M. Izal.
- Publicado: Enviado a Computer Communications.
- Abstract: An Internet topology at the router level not only needs to discover IP addresses in Internet paths (traceroute) but also needs to identify IP addresses belonging to the same router (IP aliases). Both processes, discovery and IP alias resolution, have traditionally been independent tasks. In this paper, a new tool called Pamplona-traceroute is proposed to improve uppon current results in a state of the art for Internet topology construction at the router level. Indirect probing using TTL-scoped UDP packets, usually in the discovery phases, are reused in IP alias resolution phases, providing high identification rates, especially in access routers.

Bibliografía

[1] Geant official site. http://www.geant.net/pages/home.aspx. 5, 10, 30

[2] Globalnoc official site. http://globalnoc.iu.edu/. 5, 10, 30

[3] Canet4 looking glass web tool. http://dooka.canet4.net/lg/lg.php. 5, 10, 30

[4] N. Spring, R. Mahajan, and D. Wetherall. Measuring ISP topologies with Rocketfuel. In *In Proc. ACM SIGCOMM*, pages 133–145, Pittsburgh, August 2002. 6, 7, 22, 25, 26, 33, 42, 69

[5] V. Jacobson. ftp://ftp.ee.lbl.gov/traceroute.tar.gz, October 1989. 6, 15

[6] The opte project. http://opte.org/, November 2003. 6

[7] Internet mapping project raw internet mapping data page. http://cheswick.com/ches/map/dbs/index.html. 6, 83

[8] Jean Jacques Pansiot and Dominique Grad. On Routes and Multicast Trees in the Internet. *ACM SIGCOMM Comput. Commun. Rev.*, 28:41–50, January 1998. 7

[9] Santiago Garcia-Jimenez, Eduardo Magaña, Daniel Morato, and Mikel Izal. Techniques for better alias resolution in Internet topology discovery. In *Published in 11th IFIP/IEEE International Symposium on Integrated Network*

Managemen miniconference, pages 513–520, New York, USA, June 2009. 7, 25, 28, 29, 34, 80, 90, 127

[10] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proc. ACM SIGCOMM*, 1999. 7

[11] Intermapper web page. `http://www.intermapper.com/`. 7

[12] E. Katz-Bassett, J.P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe. Towards IP Geolocation Using Delay and Topology Measurements. In *Proc. USENIX Internet Measurement Conference*, pages 71–84, Rio de Janeiro, Brazil, 2006. 7

[13] Mirrezaei S.I., J Shahparian, and M Ghodsi. A Topology-Aware Load Balancing Algorithm for P2P systems. pages 1–6, Michigan, USA, November 2009. 8

[14] Hal Burch and Bill Cheswick. Tracing Anonymous Packets to their Approximate Source. In *Proceedings of the 14th USENIX conference on System administration*, pages 319–328, New Orleans, Louisiana, USA, December 2000. 8

[15] Lili Qiu, Venkata N. Padmanabhan, and Geoffrey M. Voelker. On the placement of web server replicas. *In Proceedings of IEEE INFOCOM*, 2001. 8

[16] B. Chun, D. Culler, T. Roscoe, A. Bavier, L. Peterson, M. Wawrzoniak, and M. Bowman. Planetlab: An overlay testbed for broad-coverage services. *ACM SIGCOMM Computer Communications Review*, 33:3–12, July 2003. 10, 67

[17] D. Morato, E. Magaña, M. Izal, J. Aracil, F. Naranjo, F. Astiz, U. Alonso, I. Csabai, P. Haga, G. Somin, J. Seger, and G. Vattay. The European Traffic Observatory InfraestruCture (ETOMIC): A testbed for universal active and passive measurements. In *Proc. TRIDENTCOM*, pages 283–289, 2005. 10, 83

[18] Mehmet H. Gunes and Kamil Sarac. Resolving IP aliases in building traceroute-based Internet maps. *IEEE/ACM Transactions on Networking*, 17:1738–1751, December 2009. 13, 21, 24, 47

[19] Jon Postel. Rfc 791 - internet protocol. http://www.ietf.org/rfc/rfc791.txt. 14

[20] Rob Sherwood, Adam Bender, and Neil Spring. Discarte: a disjunctive internet cartographer. In *Proceedings of the ACM SIGCOMM 2008 conference on Data communication*, SIGCOMM '08, pages 303–314, New York, NY, USA, August 2008. ACM. 15, 24, 56

[21] Aiguo Fei, Guangyu Pei, Roy Liu, and Lixia Zhang. Measurements on delay and hop-count of the internet. In *in IEEE GLOBECOM'98 - Internet Mini-Conference*, 1998. 15

[22] Brice Augustin, Xavier Cuvellier, Benjamin Orgogozo, Fabien Viget, Matthieu Latapy Timur Friedman, Clemence Magnien, and Renata Teixeira. Avoiding traceroute anomalies with paris traceroute. In *6th ACM SIGCOMM*, pages 153–158, Rio de Janeiro, Brazil, October 2006. 17, 18, 20, 151

[23] Mehmet Gunes and Kamil Sarac. Analytical IP alias resolution. In *ICC '06. IEEE International Conference on Communications*, pages 459–464, Istanbul, June 2006. 21, 46, 47, 63

[24] N. Spring, M. Dontcheva, M. Rodrig, and D. Wetherall. How to resolve IP aliases. Technical Report UW-CSE-TR 04-05-04, Washington Univ. Computer Science, 2004. 21, 27, 60, 61, 63

[25] K. Keys, Y. Hyun, M. Luckie, and K. Claffy. Internet-scale ipv4 alias resolution with midar. *Networking, IEEE/ACM Transactions on*, PP(99):1, 2012. 22, 25, 26, 28, 53, 69

[26] Santiago Garcia-Jimenez, Eduardo Magaña, Mikel Izal, and Daniel Morató. Validity of router responses for ip aliases resolution. In Robert Bestak, Lukas Kencl, Li Li, Joerg Widmer, and Hao Yin, editors, *NETWORKING 2012*, volume 7289 of *Lecture Notes in Computer Science*, pages 358–369. Springer Berlin / Heidelberg, 2012. 22, 24

[27] Justine Sherry, Ethan Katz-Bassett, Mary Pimenova, Harsha V. Madhyastha, Thomas Anderson, and Arvind Krishnamurthy. Resolving IP aliases with prespecified timestamps. In *Proceedings of the 10th annual conference on Internet measurement*, IMC '10, pages 172–178, New York, NY, USA, November 2010. ACM. 23, 24, 40

[28] Sebastian Zander. An improved clock-skew measurement technique for revealing hidden services. 23, 24

[29] R. Govindan and H. Tangmunarunkit. Heuristics for internet map discovery. In *Proc. IEEE INFOCOM*, March 2000. 24, 25, 28, 31, 74

[30] Adam Bender, Rod Sherwood, and Neil Spring. Fixing Ally's Growing Pains with Velocity Modeling. In *(IMC 08) 8th ACM SIGCOMM conference on Internet measurement*, pages 337–342, New York, NY, USA, October 2008. ACM. 25, 26, 33, 42, 43, 63, 69, 80

[31] Hal Burch. Measuring an IP Network in situ. Carnegie Mellon University, PhD thesis, ISBN 0-542-01549-8, 2005. 27

[32] Harsha V. Madhyastha, Tomas Isdal, Michael Piatek, Colin Dixon, Thomas Anderson, Arvind Krishnamurthy, and Arun Venkataramani. iPlane: An information Plane for Distributed Services. In *7th USENIX Symposium on Operating Systems Desing and Implementation*, pages 367–380, Seattle, WA, November 2006. 28, 38

[33] Jean-Jacques Pansiot and Dominique Grad. On routes and multicast trees in the Internet. *Computer Communication Review*, 28:41–50, 1998. 29

[34] K. Keys. Internet-Scale IP Alias Resolution Techniques. *ACM SIGCOMM Computer Communication Review (CCR)*, 40(1):50–55, Jan 2010. 29, 63

[35] Santiago Garcia-Jimenez, Eduardo Magaña, Daniel Morato, and Mikel Izal. Improving efficiency of IP alias resolution based on offsets between IP addresses. In *Published in 21st International Teletraffic Congress (ITC 21)*, pages 1 – 8, Paris, France, September 2009. 30

[36] D. McRobb, K. Claffy, and T. Monk. Skitter: CAIDA's macroscopic Internet topology discovery and tracking tool. Available from http://www.caida.org/tools/measurement/skitter/, 1999. 31, 133

[37] Caida research group web page. http://www.caida.org/. 31

[38] Rocketfuel. http://www.cs.washington.edu/research/networking/rocket
33

[39] Rob Sherwood and Neil Spring. Touring the internet in a tcp sidecar. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, IMC '06, pages 339–344, New York, NY, USA, 2006. ACM. 43

[40] Kamil Sarac M. Engin Tozal. Tracenet: An internet topology data collector. *Internet Measurement Conference IMC*, pages 356–368, November 2010. 49

[41] Kamil Sarac M. Engin Tozal. Palmtree: An ip alias resolution algorithm with linear probing complexity. *Computer Communications*, 34(5):658–669, April 2011. 52, 53

[42] Radargun's web page. http://www.cs.umd.edu/~bender/radargun/. 53

[43] Craig Labovitz, Abha Ahuja, and Farnam Jahanian. Experimental study of internet stability and backbone failures. In *In FTCS99*, pages 278–285, 1999. 82

[44] Kevin Butler and Patrick Mcdaniel. Optimizing BGP Security by Exploiting Path Stability. In *In ACM CCS*, pages 298–310, 2006. 82

[45] Udi Weinsberg, Yuval Shavitt, and Yaron Schwartz. Stability and Symmetry of Internet Routing. In *Proceedings of the 28th IEEE international conference on Computer Communications Workshops*, pages 407–408, April 2009. 82

[46] Milena Janic, Fernando Kuipers, Xiaoming Zhou, and Piet Van Mieghem. Implications for qos provisioning based on traceroute measurements. In *QofIS'02*, pages 3–14, 2002. 83

[47] Scamper caida web. http://www.caida.org/tools/measurement/scamper/. 93

[48] Y. Shavitt and E. Shir. DIMES: Let the Internet Measure Itself. *ACM SIGCOMM Computer Communication Review*, 35(5):71–74, October 2005. 93

[49] Lixin Gao and Feng Wang. The extent of as path inflation by routing policies. *7th IEEE Global Internet Symposium (Taipei, Taiwan)*, November 2002. 100

[50] A launch pad for network & internet management related resources. http://www.netconfigs.com. 100

[51] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967. 114

[52] N. M. Laird D. B. Rubin A. P. Dempster. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977. 114

[53] CAIDA. ARK, Archipelago Measurement Infrastructure. http://www.caida.org/projects/ark/, 2002. 133