Tesis de Grado
de Ingeniería en Informática

Estudio de PeeringDB
para identificación y extracción de siblings

Director: Dr. Ing. Esteban Carisimo

Alumno: Augusto Arturi, *Padrón #97.498*
aarturi@fi.uba.ar

Facultad de Ingeniería, Universidad de Buenos Aires

27 de agosto de 2021

Índice general

Índice de figuras

Índice de cuadros

Índice de listados de código

Agradecimientos

Quiero agradecerle a mi director, Dr. Ing. Esteban Carísimo, por su dedicación y constancia invertida en mi para desarrollar esta Tesis. La forma de transmitir sus conocimientos y consejos son para mí, una virtud admirable que me han permitido estar motivado con este trabajo de principio a fin. Elaborar esta tesis con un profesional de su magnitud, fue una experiencia más que satisfactoria.

Dedicado a mis papás,
Rosana y Oscar.

Resumen

La red pública de Internet está compuesta por aproximadamente 70.000 entidades denominadas Sistemas Autónomos (del inglés Autonomous Systems, ASes), entre las cuales se encuentran ISPs, universidades, organismos públicos y proveedores de contenidos. En el sistema de ruteo de Internet, cada una de estas organizaciones se identifica al resto de la red por medio de un idenficador único denominado Número de Sistema Autónomo (del inglés Autonomous System Number, ASN). Aunque originalmente el protocolo BGP fue concebido para que cada organización se represente bajo un único ASN, dinámicas comerciales y de operaciones han llevado a que ciertas organizaciones controlen múltiples ASNs. En particular, el caso más frecuente es el de las empresas multinacionales, que operan en distintos países con distintos ASNs.

Esta practica de utilizar múltiples ASNs genera el problema de que los datos de ruteo no representan fehacientemente la totalidad del tamaño, y potencialmente la relevancia, de un conglomerado operando en Internet. Sin embargo, el mayor obstáculo se presenta a la hora de poder identificar el conjunto de ASNs operados por un mismo conglomerado. En esta Tesis se busca mejorar las técnicas de inferencia de siblings (conjunto de ASNs operados por una misma organización) a través de la exploración de la base de datos pública PeeringDB. Esta base de datos es diariamente usada por los operadores de Internet para compartir información que facilite el ruteo y acceso a sus redes. Frecuentemente los operadores comparten aquí sus siblings, convirtiendo esta plataforma en una posible fuente de datos para resolver la inferencia de siblings. El desafío de esta Tesis se centra en convertir la información presente en PeeringDB, carente de estructura ya que está pensada para ser leída por humanos, en una fuente de datos estructurada que reporte los siblings de un conglomerado.

Pre-Introducción

Esta Tesis se fragmenta en los siguientes capítulos:

1. **Contexto, motivación y postulación del problema**: El Capítulo 1 presenta el encuadre en el cual se desarrolló este trabajo de Tesis, comenzando con la motivación de desarrollar una nueva metodología de inferencia de *siblings*. Luego, a lo largo de este capítulo se explican conceptos fundamentales de la red de Sistemas Aútonomos para comprender el trabajo realizado, se presenta una revisión de la literatura y se postula la hipótesis alrededor de la cual se trabajará en el resto del trabajo. Finamlente, se concluye discutiendo el impacto que puede tener este trabajo en diferentes comunidades científicas, técnicas y profesionales.

2. **Metodología**: El Capítulo 2 comienza describiendo el estado del arte de las técnicas de inferencia de siblings y sus limitaciones. Luego, en este capítulo se presenta PeeringDB [35] y se argumenta la factibilidad de desarrollar una técnica de inferencias de siblings a partir de esta base de datos. Finalmente, este capítulo presenta detalladamente el desarrollo y la evaluación de una metodología de inferencia de siblings basada en PeeringDB.

3. **Resultados**: El Capítulo 3 presentará el análisis de los resultados obtenidos al aplicar la metodología propuesta en el Capítulo 2. En este capítulo se discutirá el impacto de esta metodología en la red de Sistemas Autónomos, y luego se hará especial énfasis en los grandes conglomerados y principales proveedores de tránsito a nivel global.

4. **Conclusiones**: Por último, el Capítulo 4 elabora las conclusiones obtenidas a lo largo de este trabajo de investigación.

Capítulo 1

Contexto, motivación y postulación del problema

En este capítulo abordaremos la motivación del trabajo (Sección 1.1) junto al marco teórico que explica los conceptos vinculados a Sistemas Autónomos y las relaciones entre ellos (Sección 1.2). Luego, presentaremos la postulación del problema a investigar con los objetivos que se proponen responder en esta Tesis de grado (Secciones 1.3 y 1.4). Por último, presentamos a las comunidades en las que este trabajo podría tener impacto junto a su beneficio en concreto (Sección 1.5)

1.1. Motivación

Esta Tesis propone mejorar la representación de los conglomerados en la red de Sistemas Autónomos. Existen múltiples indicios, además de evidencia, de que existen grandes conglomerados internacionales brindando servicios en Internet, que abarcan desde telefonía móvil, conectividad fija y provisión de contenidos. Sin embargo, las complejas estructuras de negocios de estas organizaciones, sumado a las caracterísitcas propias de los protocolos de Internet, generan barreras para determinar la dimensión global que tienen estos conglomerados dentro de Internet. Esto se debe a que muchos conglomerados segmentan sus redes replicando la estructutra comercial que tiene su organización (por ejemplo: una red por subsidiaria o una red por cada estructura del negocio). A su vez, los protocolo de Internet, y en particular BGP, no requieren de ninguna información comercial para su funcionamiento. Al poder capturar fehaciente la infraestrcutura de Internet controlada por una organización, se podrá, por ejemplo, analizar si existe una tendencia hacia la concentración de los recursos de la red.

1.2. Marco conceptual: Teoría de Redes de computadoras

Un Sistema Autónomo (del inglés Autonomous System, AS) es un conjunto de prefijos que son operados o gestionados por una misma organización de red que posee una clara y única política de ruteo [32]. Actualmente, en Junio de 2021, la red de Internet está compuesta por aproximadamente 72.000 ASes activos [13], incluyendo instituciones con distintas finalidades, tales como ISPs, compañías y universidades. Cada AS presente en la red de Internet es identificado bajo un único Número de Sistema Autónomo (del inglés Autonomous System Number, ASN) [42]. Los ASNs, al igual que las direcciones IP, son asignados por la Autoridad de Números Asignados en Internet (del inglés, Internet Assigned Numbers Authority, *IANA*) los cuales son los únicos habilitados a solicitar direcciones IPs al Registro Regional de Internet (del inglés Regional Internet Registry, *RIR*).

Hacia el interior de un AS, el ruteo está coordinado a través del uso de un mismo protocolo, el cual se ejecuta en cada uno de los routers del AS. De esta forma, para rutear un paquete donde el emisor y el receptor están dentro del mismo Sistema Autónomo, la ruta está enteramente determinada por este protocolo intra-AS.

Para que Internet se vuelva alcanzable, es necesario un protocolo inter-autónomo o de ruteo externo, en donde un paquete atravesará múltiples sistemas autónomos para llegar a su destino. El conjunto de ASes está interconectado vía enlaces dedicados y puntos de acceso público para el intercambio de información, haciendo uso del Border Gateway Protocol (*BGP*) [41], un protocolo que le permite a cada sistema autónomo propagar su información de ruteo sin revelar políticas o topologías internas a terceros [30]. Para que la comunicación sea efectiva, se necesita que todos los ASes utilicen el mismo protocolo de comunicación (BGP-4), bajo los principios de ser descentralizado, asincrónico y local. BGP es un protocolo que sirve para obtener rutas y destinos, así como tambíen divulgar los prefijos a los ASes vecinos, creando una estructura jerárquica y relación con intereses entre los distintos sistemas autónomos.

1.2.1. Relaciones entre Sistemas Autónomos

La interconexión de dos Sistemas Autónomos, es decir la disposnibilidad de enlaces que cursan tráfico entre dos ASes, puede ser clasificada en función de las relaciones comerciales que matienen ambas partes entre sí. Esto se ve principalmente reflejado en los prefijos que cada uno de los ASes en una relación comercial anuncia a través de BGP al resto de sus vecinos.

Gao y Rexford [32] propusieron el primer modelo identificando estas relaciones, proponiendo tres cateogrías posibles: *(i)* cliente-proveedor (del inglés, *customer-provider*, c2p), *(ii)* entre pares (del inglés, *peer-to-peer*, p2p), o *(iii)* de apoyo (del inglés, *backup*).

La relación cliente-proveedor es un vínculo comercial, con fines de lucro entre las partes. En este tipo de relaciones, una de las partes presenta una ubicación más privilegiada dentro de la esctructura de la red, entiendiéndose este privilegio como mayor alcance. Por ejemplo, este puede ser el caso de un proveedor de Internet municipal que se conecta con un proveedor de Internet de cobertura nacional, siendo este último, quien cuenta con mayor alcance dentro de la red. En este tipo de relaciones, el proveedor se ofrece ante el resto de la red, como una vía de acceso al cliente. A su vez, dada esta posición de privilegio, el proveedor le ofrece al cliente poder alcanzer cualquier destino de la red de Internet atravesando su infraestrcutura. Desde la perspectiva del tráfico, se supone que el intercambio de tráfico será asimétrico, ya que el proveedor es la vía de acceso para acceder al contenido disponible en el resto de la red. Por lo tanto, dado que el proveedor brinda su posición de privilegio al cliente, permitiendo que el tráfico desde el resto de la Internet fluya a través del proveedor para alcanzar el cliente, es el cliente quien debe compensarlo económicamente por estos servicios.

Las relaciones entre pares (p2p) no se celebran bajo la premisa de que una de las partes cuenta con una posición de privilegio respecto de la otra. Más aún, estas relaciones por lo general se basan en el principio de beneficio mutuo y se rigen bajo la cláusula de no tarifación en escenario de intercambio de tráfico simétrico. Sin embargo, a raíz de que este tipo de relaciones es no tarifada y para beneficio mutuo, este tipo de enlaces no les permite a las partes obtener una visibilidad total de la red por medio de sus contrapartes. El beneficio para los ASes de contar con estos enlaces evita que una fracción del tráfico sea enviada por los enlaces tarifados de los proveedores y se priorice utilizar enlances p2p para estos destinos. En ciertas ocasiones, como es el caso de los Puntos de Intercambio de Tráfico (del inglés Internet Exchange Point, IXPs) [12, 2], el uso de estos enlaces también tiene un impacto en la performance reduciendo la cantidad de intermediarios y la latencia para alcanzar el destino

La relación de backup funciona entre ASes vecinos para brindar conectividad ante fallas. Entre sistemas autónomos existe un enlace físico, que por factores externos o por un evento, se puede dañar, por ejemplo de un cliente hacia un proveedor. Un par de dominios administrativos proveen esta relación de apoyo para mantener la conectividad.

Sin embargo, se puede considerar que existe una categoría adicional, denominada "hermanos" (del inglés, *sibling-to-sibling*, s2s) , que es la relación a la que se le dará mayor énfasis en esta Tesis. Esta relación se establece al interconectar a dos ASes, con un límite administrativo en común, es decir, que ambos ASes pertenecen a la misma

organización (o conglomerado). Estos enlaces suelen aparecer como resultado de fusiones y adquisiciones, en determinados escenarios de gestión de red. Además, las políticas de ruteo e ingeniería de tráfico entre *siblings* son más flexibles que los anteriormente descriptos, por el hecho de pertenecer a la misma organización. Esta flexibilidad de ruteo genera que sea un desafío poder inferir *siblings* a través de información presente en las tablas de ruteo BGP.

1.2.2. Literatura relacionada

El estudio de la red de Sistemas Autónomos utiliza diferentes metodologías de recopilación de datos para analizar la estructura de la red. Sin embargo, una de las fuentes principales de datos son los anuncios de prefijos a través del protocolo BGP. A través de este mecanismo, y dado que BGP es un protocolo de vector de distancias, se construye un atributo denominado AS-PATH , que indica los ASes a atravesar para llegar al destino. La estructura de la red de ASes tiene un ordenamiento jerárquico, que puede ser inferida por medio de los AS-PATHS [21]. De igual manera, las relaciones comerciales entre los ASes también pueden ser inferidas por medio de los AS-PATHS [34].

Trabajos como el de Luckie *et al.* [34] han estudiando la relevancia de los ASes dentro de la estructura de Internet a partir de una métrica denominada *customer cones*. Esta métrica busca inferir las relaciones comerciales entre los Ases utilizando la información disponible en las tablas de ruteo BGP. Luego, a partir de esta información, se contruye un árbol jerárquico, donde el customer cone es el sub-árbol que dado un AS incluye los clientes, y recursivamente sus clientes hacia abajo. Sin embargo, los trabajo enfocados en inferir relaciones comerciales presentan limitaciones para identificar siblings. El principal argumento indica que no es posible distinguir cuando una organización opera múltiples ASes, ya que emplea políticas de exportación flexibles y diversas con proveedores, clientes y pares.

Trabajos previos enfocados en la identificación de siblings han basado sus metodologías en enriquecer la topología de Internet a partir de información externa a la de ruteo. Dimitropoulos *et al.* [16] propuso generar manualmente un diccionario de sinónimos de nombres basado en la información disponible en el Registro de Ruteo de Internet (del inglés Internet Routing Registry, IRR) [29]. Por ejemplo, este diccionario toma WorldNet Services y AT&T Israel como sinónimos de la empresea AT&T. Sin embargo, esta metodología requiere de trabajo manual lo que dificulta el mantenimiento de la base de datos. A su vez, requiere que los ASes cuenten con un perfil actualizado en IRR, la cual es una plataforma sin un alto grado de adopción.

Una propuesta alternativa ha sido la de Cai *et al.* [9], la cual genera una inferencia de

siblings a través de la información disponible en el servicio WHOIS [40]. Esta metodología propone asociar ASes a una misma organización a partir de los campos de contacto, direcciones e identificadores de la organización. Aunque esta metodología es ampliamente difundida, sus limitaciones se centran principalmente en las restricciones del servicio WHOIS, que serán explicadas en detalle en la Sección 2.1.

En 2004 se crea PeeringDB (PDB) [35], una iniciativa comunitaria de una organización sin fines de lucro dirigida y promovida por voluntarios, una herramienta pública para el crecimiento y entendimiento de Internet. PDB es una fuente de datos que puede complementar y validar inferencias realizadas por otras fuentes de datos que utilizan BGP. Esta fuente de datos es representativa de Internet, en términos de tránsito, contenido y proveedores, Lodhi *et al.* [33] determina en su estudio que la información disponible en PeeringDB es representativa, correcta y actual. Sin embargo, también se encuentran limitaciones de los AS representados como *stubs* [15] cuya actividad principal no está vinculada a Internet, sino en campos como educación/investigación, retail o redes con bajo volumen de tránsito, así como también aquellos que deciden no compartir ningún tipo de información por razones competitivas.

Otros estudios han utilizado PDB para estudiar organizaciones "hipergigantes". La información de PDB se utiliza para comprender a las organizaciones en su rol, caracterizarlas adecuadamente, y extraer aquellos atributos que permiten diferenciar organizaciones tradicionales de aquellas que no lo son. A partir de atributos como la capacidad de puertos, su alcance y perfil de tráfico saliente se puede inferir la magnitud de una organización hipergigante [7].

1.3. Postulación del problema

Existe evidencia de que múltiples ASes de Internet son operados por una misma compañía, sin embargo, en ocasiones se refleja que las técnicas actuales de inferencia de siblings no son capaces de capturarlos por completo. En esta Tesis, se propone utilizar la información presente en PeeringDB para aumentar la inferencia de siblings. En función del problema a explorar, se plantea la siguiente hipótesis .

Hipótesis. *Las metodologías actuales de inferencia de siblings generan una representación incompleta de la dimensión de conglomerados en la red de Sistemas Autónomos.*

A partir de los trabajos científicos previos que han demostrado el valor de PeeringDB como fuente de datos representativa de la red de ASes, esta Tesis propone desarrollar

una metodología de inferencia de siblings basada en esta fuente de información. Específicamente, el uso de PeeringDB para mejorar la detección de siblings luego de fusiones de compañías. Este motivo particular se debe a que suponemos que esta información puede llegar a ser reportada en esta plataforma como parte de las operaciones.

1.4. Objetivos específicos

Ésta Tesis propone crear una metodología que extraiga la información disponible en los registros de PeeringDB para identificar siblings. A través de la creación de esta metodología se busca responder las siguientes preguntas.

1. **Prevalencia en PDB**. Antes de construir una metodología basada en PDB, es importante determinar si PDB es una base de datos representativa de la red de ASes. Además, es importante investigar si esta base de datos no cuenta con sesgos en donde se sobre- o sub-representen ciertos países. En caso de encontrar desbalances, queremos investigar si la adopción de PDB se debe a algún motivo específico de ese país (ej: incentivos, política de grandes IXPs dométicos, etc). Esto lo respondemos en la Sección 2.3.2.

2. **Prevalencia de ASes activos en PeeringDB**. Los recursos de numeración, como es el caso de los ASN, pudieron haber sido delegados a una organización pero no estar siendo anunciados en el sistema de ruteo. Aunque el objetivo de esta Tesis es identificar siblings sin importar si se encuentran activos o no, suponemos que la metodología contará con mayor valor si captura los ASes que se encuentran en actividad. Esto lo respondemos en la Sección 2.3.2.

3. **Identificar los campos potenciales para el reporte de siblings**. Antes de comenzar a elaborar la metodología, es necesario comprender cuáles son los campos donde potencialmente se puede reportar la información correspondiente a los siblings.

4. **Identificar estructuras para reportar siblings en PeeringDB**. Al basar la metodología en PeeringDB es fundamental conocer si los operadores cuentan con alguna forma estandarizada o adoptada para reportar los siblings de su organización. Esto lo respondemos en la Sección 2.3.1.

5. **Generación de métodos de extracción**. Luego del aprendizaje producto de la exploración de PDB, se crearán mecanismos de extracción de los siblings reportados en los campos de PDB. Esto lo respondemos en la Sección 2.4.3.

6. **Validación** Luego de haber construído la metodología es importante evaluar si los números obtenidos corresponden fehacientemente a ASNs, si la inferencia de siblings es correcta y si han quedado ASNs sin identificar. Esto lo respondemos en la Sección 2.5.

7. **Comparación y contraste con metodologías del estado del arte**. La última etapa será la de determinar los beneficios de esta nueva metodología desarrollada, y combinar los esfuerzos de la metodología propuesta con las metodologías del estado del arte. Esto lo respondemos en la Sección 3.2.

8. **Aplicación**. Finalmente, como último paso, se aplicará la metodología creada y se analizará el impacto y las mejoras de esta nueva metodología en la detección de conglomerados. Esto lo respondemos en el Capítulo 3.

1.5. Transferencia a otras comunidades y al resto de la sociedad

La investigación presente en esta Tesis busca contribuir a diversas ramas de la ciencia interesadas en las dinámicas de la estructura de la red, incluyendo las ciencias informáticas, económicas, políticas y sociales. A continuación se presentan algunas comunidades en las que esta Tesis podría tener impacto, y además se brindan algunos ejemplos de cuál podría ser el beneficio concreto.

Comunidad de las ciencias informáticas dedicada al estudio de Internet. Mejorar las técnicas de identificación de siblings permitiría determinar con mayor precisión la dimensión de los conglomerados en la red de ASes. A través de esta nueva información se podrían llevar a cabo estudios que investigen si los conglomerados tienen una política de ruteo unificada. Además, al contar con estos datos más precisos se podría investigar cuál es el uso que le dan las grandes Redes de Distribución de Contenido (del inglés *Content Delivery Networks*, CDNs) [10, 23] a cada uno de los ASes que controlan.

Comunidad de las ciencias sociales dedicada al estudio de la tecnología. Áreas de las ciencias sociales dedicadas al estudio de Internet y su vínculo con la sociedad, incluyendo la economía, la sociología, las ciencias políticas y las ciencias de la comunicación, también podrían beneficiarse al contar con información más precisa sobre los conglomerados. Por ejemplo, muchas de estas disciplinas evalúan el impacto de las fusiones en el ecosistema de Internet. Sin embargo, hasta el momento no existe ninguna forma sistémica de detectar la fusión de compañías de Internet. Dado que producto de las fusiones se puede alterar la estructura de Internet, mejores técnicas de identificación

de siblings podrían generar una forma de detección sistemática de fusiones. A su vez, comprender el impacto que tiene una fusión en la estructura de la red podría ayudar a estudiar si existen posiciones monopólicas en términos de infraestructura.

Comunidad de operadores de Internet. El término operadores de Internet hace referencia a los técnicos e ingenieros encargados de la configuración, diseño y operación de las infraestructuras de comunicaciones pertenecientes a los Sistemas Autónomos. En un gran número de oportunidades los operadores de un Sistema Autónomo necesitan contactarse con sus pares de otro Sistema Autónomo. En estas circunstancias, el hecho de conocer que un AS forma parte de un conglomerado mayor, podría facilitar el contacto o la resolución del motivo que convoca al operador a generar la interacción.

Legisladores. A través de un mejor entendimiento de la fracción de la infraestructura de Internet en manos de un conglomerado, los legisladores podrán tomar decisiones respecto a la modificación de las políticas de telecomunicaciones . Estas políticas podrían incluir directrices para fomentar la competencia, limitar la expansión de grandes conglomerados o evaluar el rol de empresas multinacionales en el ecosistema doméstico de Internet.

Reguladores. Los reguladores, quienes auditan el cumplimiento del marco regulatorio y normativo, podrán beneficiarse de mejores metodologías de detección de conglomerados para llevar a cabo sus tareas. Dentro del amplio rango de entes reguladores con incumbencia en el mercados de telecomunicaciones (ENACOM, Defensa de la competencia, etc.) podrán utilizar estos datos para aprobar fusiones de compañías, otorgar licencias de telecomunicaciones o incluso llevar adelante el cumplimiento de leyes antimonopolio.

Capítulo 2

Metodología

En este capítulo abordamos la metodología desarrollada para inferir siblings, primero comenzaremos evaluando las metodologías del estado del arte (Sección 2.1). Luego, discutiremos la posibilidad de usar PDB para el desarrollo de una nueva metodología (ver Sección 2.2), junto a un análisis de los campos que presentan potencial para la inferencia (Sección 2.3). Propondremos una metodología y la evaluaremos (Secciones 2.4 y 2.5) y finalmente se explicarán las limitaciones encontradas (Sección 2.6)

2.1. Metodologías existentes y sus limitaciones

La metodología propuesta por Cai *et al.* [9] aplica un algoritmo de agrupamiento (del inglés, clustering) para descubrir las relaciones entre Sistemas Autónomos y Organizaciones a través de las operaciones [9]. El mapeo se realiza a partir de una heurística que utiliza datos de WHOIS para el año 2010. A partir de la extracción de datos crudos sobre ASes y su canonización, se descartan aquellos atributos considerados "genéricos" por ser comunes a muchas organizaciones. Por medio de uno o más atributos seleccionados, se forman los clusters que finalmente son etiquetados para tener un nombre amigable y representativo. Este agrupamiento se realiza a través de los campos OrgId, número telefónico, dominio email y posteriormente una combinación de estos. Para validar la metodología se realizaron dos observaciones, una interna a través de un ISP Tier 1 y otra externa, por medio de nueve organizaciones distintas. Los resultados que se obtienen permiten concluir que WHOIS es una base de datos con información valiosa para estudiar operaciones de ASes.

Por otro lado, la base de datos de WHOIS presenta una serie de limitaciones, tanto en la calidad del contenido como en su diseño. La primera de ellas es que la información está semiestructurada y no sigue un esquema consistente. Esto implica que debe aplicarse un parseo para leer los datos, que a escala produce incompletitudes y baja mantenibilidad

en el tiempo [31]. Además de que los datos tienen baja cohesión y se separan por RIR, existen distintas formas de estructurar la información acorde a cada uno. ARIN utiliza su propio formato a distinción de los demás registros. A su vez, cada RIR aplica distintas políticas para sus organizaciones, que como resultado conlleva a tener distintos campos y coberturas para cada Registro Regional. A pesar de iniciativas para la validación y verificación de datos, como *WHOIS Accuracy Program Specification (WAPS)* impulsada por ICANN [38], la información disponible aún puede ser obsoleta e imprecisa. Por último, el límite de consultas del protocolo es de 20 por hora y 200 por día desde la misma dirección IP. Aunque esta alternativa no está agotada y se pueden refinar técnicas del estado del arte para obtener una mejor representación de los datos, esto queda por fuera de los objetivos planteados en esta Tesis.

2.2. PeeringDB

PeeringDB (PDB) es una organización[1], que facilita el intercambio de información relacionada con interconexión, principalmente para Operadores de Red. PDB funciona como un punto de centralización de datos vinculados a peering que permite acelerar el proceso de búsqueda y conexión con otras redes. El peering es un intercambio de libre trafico para un beneficio mutuo entre las partes. Puede ser público, llevado a cabo a través de un IXP, donde una red puede conectarse con varias redes a través de esta conexión. También puede ser de carácter privado cuando la instalación de intercambio es de esta índole. Sin un IXP local, los proveedores de servicios de Internet deben utilizar la conectividad internacional para intercambiar y acceder al trafico global, generalmente a mayor costo [20]. En la actualidad, algunas organizaciones, como por ejemplo Microsoft [39], exigen a su contraparte contar con un registro en PeeringDB para establecer el peering. PDB está orientada hacia las operaciones. Permite que otras redes se conozcan entre ellas e intercambien información de como contactarse sin intermediarios. Funciona como un filtro para decidir, cuando, donde y con quien hacer peering en primera instancia.

El proceso de carga de una entrada a la base de datos es a través de un formato libre, en donde cada organización participante que lo genera es responsable de los datos que publica. Cada registro nuevo es sometido a un proceso de autenticación, en donde los administradores voluntarios validan los datos de entrada, de forma que no exista información replicada o falsificada. Una entrada será aceptada sólo si posee un registro de WHOIS actualizado y un email institucional de la persona que carga los datos [19].

CAIDA[2] posee un registro histórico de los datos de PDB a través de copias (del inglés,

[1]Sin fines de lucro, basada en miembros

[2]Center for Applied Internet Data Analysis based at the University of California [1]

dumps). Los datos son procesados, curados y estructurados en distintos formatos según año en que se capturaron. Para el desarrollo de la metodología se tomaron una serie de *snapshots* de los disponibles tanto para realizar como para validar lo implementado. Cada dump contiene los datos disponibles que se pueden observar desde la interfaz web, sin embargo, para el desarrollo de esta tesis, solamente se hará foco en aquellos documentos de la base de datos que son de interés para desarrollar una metodología por operaciones.

El listado de código 2.1 presenta un ejemplo de los datos estructurados en formato JSON para la empresa GTT Communications en una captura de PDB tomado el día primero de octubre del año 2020 [28]. El listado 2.1 se refiere a un ejemplo del documento `Net` donde observamos 16 campos , entre estos, existe un grupo específico referenciado a políticas como `policy_general`, un campo categórico en función de definir qué tan abierta o cerrada es la política general del ASN, `policy_ratio` booleano, `policy_contracts`, un categórico para definir si es requerido o no un contrato entre las partes para realizar peering y `policy_url` para especificar un sitio web con más información. Por otro lado se notifica la fecha de última actualización y creación a través de los campos `updated`, `created` respectivamente. El número de ASN y el nombre que lo vincula se describe en `asn` y `name`. Información de su alcance en `info_scope`, un campo categórico para especificar la región con la que interactúa el Sistema Autónomo. El campo notas en `notes` es de texto libre para notificar información de interés propia de la organización. Suponemos que este campo potencialmente puede brindar información referida a los siblings de la organización que opera este AS, se describirá en detalle en 2.3.1.3. `Aka` o "también conocido como", es un campo que hace mención de otra forma por la cual se puede conocer a la organización, del cual también se extraerá valor. Por último, el campo `org_id` es la clave primaria que provee la base de datos para vincular al conglomerado bajo un mismo identificador.

En el listado de código 2.2 obtenido de la misma captura, pero en otro documento se otorgan atributos de la organización. El listado se compone de 12 campos. Se puede verificar la dirección física que representa a la entidad a través de `address1`. Información de su nacionalidad y su sitio web oficial se obtienen por medio de los campos `country` y `website`. El atributo más significativo del documento, en función de distinguir el operador de cada ASNs es `id`. Este campo permite asignar de una manera inequívoca varios Sistemas Autónomos registrados en PDB a una misma organización. Se presentará en detalle en 2.3.1.1.

```
1 {"meta":
2   {"generated": 1601614591.736},
3     "data": [
```

```
4       {
5         "policy_ratio": true,
6         "info_unicast": true,
7         "policy_general":"Restrictive",
8         "website":"http://www.gtt.net",
9         "allow_ixp_update": false,
10        "updated":"2018-08-29T14:21:57Z",
11        "asn": 4436,
12        "policy_locations": "Required - International",
13        "name": "GTT Communications (AS4436)",
14        "info_scope": "Global",
15        "notes": "nLayer / AS4436 has been acquired by GTT
             Communications / AS3257 and is no longer directly
              peering.Please refer all peering related
             inquiries to peering [at] gtt [dot] net.",
16        "created": "2004-07-28T00:00:00Z",
17        "org_id": 8897,
18        "policy_url": "http://www.gtt.net/peering/",
19        "policy_contracts": "Required",
20        "aka": "Formerly known as nLayer Communications",
21        }}}
```

Listado de código 2.1: Listado de campos para estructura de datos JSON del documento Net en PeeringDB.

```
1 {"meta":
2     {"generated": 1601614574.828},
3         "data": [
4         {"website": "http://www.gtt.net/",
5         "city": "McLean",
6         "updated": "2018-07-11T19:57:34Z",
7         "name": "GTT Communications, Inc.",
8         "created": "2007-08-24T15:02:25Z",
9         "address1": "7900 Tysons One Place",
10        "notes": "",
11        "zipcode": "22102",
```

```
12          "id": 8897,
13          "state": "VA",
14          "status": "ok",
15          "country": "US",
16          }}}
```

Listado de código 2.2: Listado de campos para estructura de datos JSON del documento Org en PeeringDB.

Existen varias razones por las cuales se utiliza PeeringDB como recurso principal para la adquisición de datos. En primer lugar, según nuestros conocimientos, no se conoce ninguna otra alternativa que reúna tantos ASNs en un único lugar. La centralización simplifica la extracción para su posterior análisis y la prestación de la metodología. En segundo lugar, este recurso brinda datos en simultáneo, reporta la organización y los ASNs que la componen, y provee un mapeo indexado entre estos. La importancia de esto radica en que PDB permite vincular ASNs a organizaciones, donde cada entidad emisora del registro es la encargada de brindar información para su público conocimiento. En tercer lugar, existe evidencia de factibilidad para la inferencia de siblings como se verá en 2.3.1. A través de distintos campos se manifiestan indicios para inferir siblings, encontrando que los campos más interesantes para estudiar son `notas`, `aka` y `org_id`. Por último, esta base de datos es aceptada y utilizada por la comunidad científica para desarrollar investigaciones en distintos campos de investigación [33, 7]. Se concluye en estos estudios que, si bien es un recurso de participación voluntaria y no existe ningún mecanismo para verificar la precisión en lo reportado, la fuente de datos demuestra ser representativa, correcta y reciente.

2.3. Análsis de PeeringDB

En esta sección se presentará el análisis de los datos disponibles en PeeringDB. El análisis se compone de dos partes, uno preliminar (Sección 2.3.1) y otro de exploración de datos (Sección 2.3.2). El análisis preliminar consistirá en examinar los campos que por su contenido, presentan un potencial para inferir siblings. Los patrones identificados en esta exploración serán utilizados para el diseño de la metodología (ver Sección 2.4). Por otro lado, del análisis de exploración de datos se obtuvieron datos estadísticos del conjunto de datos, a fines de conocer su volumen y distribución. A su vez, se realizó un estudio por medio de n-gramas para comprender la semántica de los campos de inferencia.

2.3.1. Análsis Preliminar

En esta sección examinaremos los campos de PeeringDB `notas`, `aka` y `org_id` para determinar cuál es el uso que les dan los operadores a estos campos. Específicamente, nuestro interés se centra en investigar si estos campos pueden ser utilizados para reportar siblings. El análisis preliminar fue una herramienta para conocer la estructura de datos que compone PeeringDB. En esta estructura de documentos, claves y valores, se descartaron una serie de campos, que por su contenido, no aportan datos relevantes para inferir siblings. La examinación de esta estructura se ejecutó a partir de la búsqueda de ejemplos. Dentro de estos, se utilizaron organizaciones locales e internacionales de las cuales se conoce su estructura. A través de la variedad de casos reportados, se comprendió el valor semántico de cada campo. A continuación, se presentan los campos `notas`, `aka` y `org_id` del listado 2.1 por medio de ejemplos en la captura del 1 de Octubre del 2020.

2.3.1.1. Organización (`org_id`)

Las organizaciones son entidades de un nivel jerárquico superior las cuales aglomeran un conjunto de entidades (o también entradas) de ASes. En particular, este campo cuenta con un identificador para poder vincular cada una de las entidades ASes con la entidad Organización. En el Cuadro 2.1 se ilustran dos ejemplos de entidades de Internet, Akamai Technologies (AS20940) y Telefónica Brasil (AS18881). En la columna `ASN-ASname` se especifica el nombre de la organización, en primer lugar, Akamai y luego Telefónica Brasil. Del campo `Organización` (`org_id`), se extraen todas las entradas de PDB que se vinculan a una entidad Organización. Por ejemplo, para Akamai, a través de la estructura de datos, utilizando el identificador `org_id` se aglomeran tres ASNs (20189, 32787 y 20940). Asimismo, para Telefónica Brasil se reportan 9 siblings. En síntesis, esta estructura de datos permite identificar siblings por medio de este campo. En el Apéndice A.1 se presentan más ejemplos de este campo.

2.3.1.2. También conocido como (`aka`)

Éste es un campo de texto libre[3], cuya finalidad es referenciar otra forma de identificar al ASN a través de un seudónimo, por ejemplo con un nombre de marca. Sin embargo, en la práctica, los operadores no siempre lo utilizan para reportar seudónimos, generando variantes. En el Cuadro 2.2 se presentan ejemplos del campo `aka` para Claro Argentina (AS11664) y Microsoft (AS8075) donde no se reportan seudónimos. Para el primero, de Claro Argentina, se reportan tres compañias Techtel (AS11664), Ertach (AS17401) y

[3]El campo `aka` de las entradas de PDB tienen hasta 255 caracteres [37]

Organización	**ASname-ASN**
Akamai Technologies	Akamai Direct Connect - 20189
	Akamai Prolexic DDoS Mitigation - 32787
	Akamai Technologies - 20940
Telefónica Brasil S.A	TELEFÔNICA BRASIL - AS10429 - 10429
	TELEFÔNICA BRASIL - AS11419 - 11419
	TELEFÔNICA BRASIL - AS16885 - 16885
	TELEFÔNICA BRASIL - AS16911 - 16911
	TELEFÔNICA BRASIL - AS18881 - 18881
	TELEFÔNICA BRASIL - AS19182 - 19182
	TELEFÔNICA BRASIL - AS22092 - 22092
	TELEFÔNICA BRASIL - AS26599 - 26599
	TELEFÔNICA BRASIL - AS27699 - 27699

Cuadro 2.1: Ejemplos del campo `org_id` extraídos de PeeringDB para las entidades de Internet, Akamai (AS20940) y Telefónica Brasil (AS18881) donde se reportan los siblings observados en la columna ASname-ASN.

Telmex (AS19037). Por medio de una fuente externa, se comprobó que a la fecha de la captura, estas organizaciones efectivamente pertenecen al mismo conglomerado que Claro [4]. En el segundo ejemplo de Microsoft, la entrada en PDB reporta dos siblings de forma directa. Tanto 8068 como 8069 pertenecen a la misma entidad y se los referencia a través de su ASN, sin hacer uso de seudónimos o identificadores alfabéticos. Se concluye que a través de esta estructura de datos, se informan siblings en distintos formatos (nombres propios, ASNs o URLs). En el Apéndice A.3 se presentan más ejemplos de este campo.

2.3.1.3. Notas (`notes`)

La estructura de datos propuesta por PeeringDB le brinda al operador este campo de texto libre, en el cual puede manifestar y brindar información para el público que consulte su entrada. A diferencia de otros campos de texto libre, no existen restricciones de cantidades, permitiendo mayor detalle y libertad para expresar contenido que campos categóricos no permiten.

En el Cuadro 2.3 se presentan ejemplos del campo `notes` para las organizaciones Telin (AS7713) y Telecom Argentina (AS7303). En el primer ejemplo del cuadro, Telin, notifica un grupo de Sistemas Autónomos que son manejados por una entidad. En este grupo de 5 ASNs, se reportan siblings que hacen a un conglomerado. En el segundo ejemplo de Telecom Argentina se observa un ejemplo menos directo. En primer lugar, se introduce el texto con datos operacionales para después listar una serie de ASNs que están bajo su control. Se puede inferir a partir de esta nota que 7303, 10481 y 10318 pertenecen al mismo conglomerado.En el Apéndice A.2 se presentan más ejemplos de

ASN-ASname	También conocido como
11664-CLARO ARGENTINA	Techtel LMDS Comunicaciones Interactivas S.A. (ASN 11664,ASN 14535 - AR-TLCI-LACNIC); ERTACH S.A. (ASN 17401 - AR-MASA5-LACNIC); TELMEX; CTI Compania de Telefonas del Interior S.A. (ASN-19037 - AR-CCTI1-LACNIC)
AS8075-Microsoft	8068 8069

Cuadro 2.2: Ejemplos del campo `aka` extraídos de PeeringDB para las entidades de Internet, Claro Argentina (AS11664) y Microsoft (AS8075) donde se reportan los siblings observados en la columna "También Conocido Como".

este campo.

2.3.1.4. Combinación de campos

Los campos citados anteriormente no son excluyentes unos de otros, la estructura de datos de PDB permite su combinación. Con lo cual haciendo uso de una combinación de estos se puede extraer una mejor representación del conglomerado.

La Figura 2.1 presenta una estructura de datos jerárquica, donde en la parte superior se encuentra la macroentidad `org_id`, y por debajo, vinculada a esta, las entidades ASes, que incluyen los campos `aka` y `notes`. Para este caso particular, hay 8 ASes vinculados bajo el mismo `org_id`. Ahora centrando la atención en la entidad AS15169, cuando miramos `aka`, se reporta como seudónimo la marca Youtube, unidad de negocio de Google. Luego, si miramos el campo `notes` encontramos un reporte de siblings con los ASNs que también maneja la organización. Replicando el comportamiento para la entidad AS19527, por `aka`, sabemos que esta unidad se vincula al servicio Cloud de Google. A su vez, del campo `notes`, por la información operacional, inferimos que los ASNs reportados son siblings. Este comportamiento vale para los 8 registros agrupados bajo el mismo identificador. En el Apéndice A.4 se ilustran con detalle otros casos de usos combinados.

2.3.2. Análisis exploratorio de datos

Esta sección presenta la exploración de datos: en la primera parte se realiza un estudio estadístico para la estructura de datos obtenida en en 2.1 mientras que en la segunda parte un análisis semántico de los campos `notas` y `aka`. En el primer análisis se investiga la cobertura de PeeringDB de la red de ASes de Internet. La captura de 1 de Octubre

ASN-ASname	Notas
AS7713-TELIN	We manage international traffic of Group ASNs: 7713: Global Network 17974: Local Network ID 23693: Mobile Network ID 56308: Local Network SG 58731: Local Network TL
AS7303-TELECOM ARGENTINA	Telecom Argentina is the major broadband and mobile provider in Argentina, with more than 4.1 MM broadband subscribers, 4 MM fixed lines and 20 MM mobile lines. Other ASN under 7303 are 10481 and 10318.

Cuadro 2.3: Ejemplos del campo `notes` extraídos de PeeringDB para las entidades de Internet, Telin (AS7713) y Telecom Argentina (AS7303) donde se reportan los siblings observados en la columna Notas.

de 2020 muestra que PeeringDB cuenta con 20.245 ASes registrados, sobre un total de 89.246 delegados por los 5 RIRs, lo que equivale a un 22,68 %. La estructura de datos se compone de 18.369 identificadores para el campo `org_id`, los cuales vinculan ASNs a organizaciones.

PDB presenta una evolución a lo largo de los años, no solo por adoptar mayores características sino también por su adopción en operadores. En la Figura 2.2 se muestran las altas (`created`) y modificaciones (`updated`) de los registros de PDB del año 2010 en adelante. En el panel superior se muestra la función de distribución acumulada que manifiesta una tendencia alcista a partir del año 2016, donde se adoptó la versión 2.0 de PeeringDB. La mediana de `updated` es alcanzada para finales del 2019, mientras que `created` si bien manifiesta el mismo patrón a partir del nuevo versionado, su ascenso es más lineal. En el panel inferior se ilustra a través de un polígono de frecuencias los registros y actualizaciones de las entradas. Se observan dos picos de `updated` para Marzo de 2016 y Julio de 2020, con frecuencias de 1366 y 3192 respectivamente, para el campo `created` el máximo es alcanzado en Agosto de 2017 con 432 registros.

El Cuadro 2.4 representa el volumen de datos disponibles para estudiar en los campos `aka` y `notes`. Se puede observar que si bien el campo `aka` contiene al menos 3 veces más de registros no nulos en relación a `notes`, la presencia numérica de `aka` es un 45,49 % inferior. El contenido numérico disponible en cada campo será de interés al momento de inferir siblings de este.

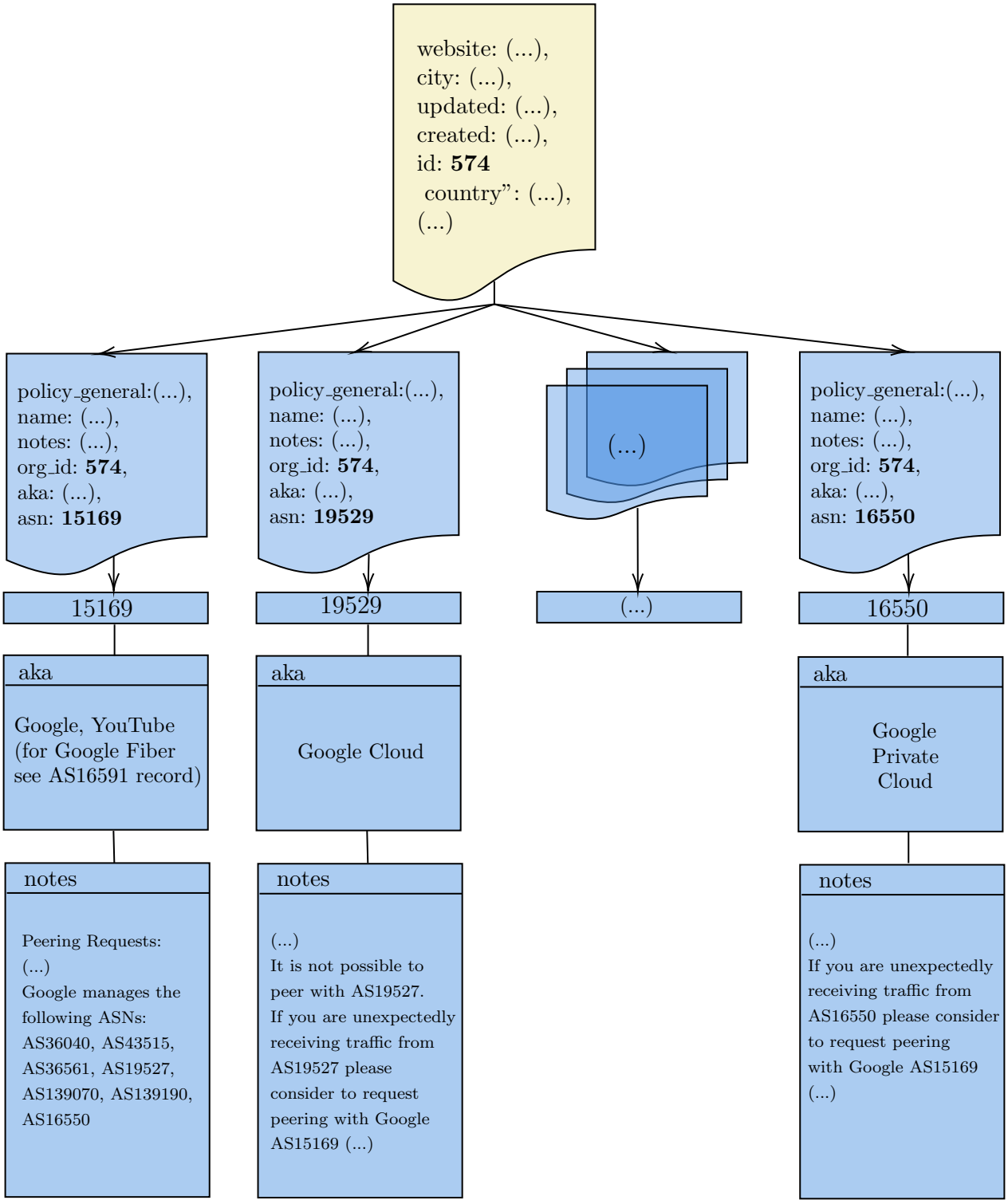


Figura 2.1: Ejemplo de la combinación de campos `org_id`, `aka` y `notes` extraídos de PeeringDB para Google LLC (AS15169). Los siblings se reportan a partir de una estructura de datos jerárquica, encabezada por la macroentidad `org_id` (indicada en amarillo) que vincula a un conjunto de 8 entradas (indicadas en azul). A su vez, de cada entrada se puede inferir información de ASNs por medio de los campos de texto libre `aka` y `notes`.

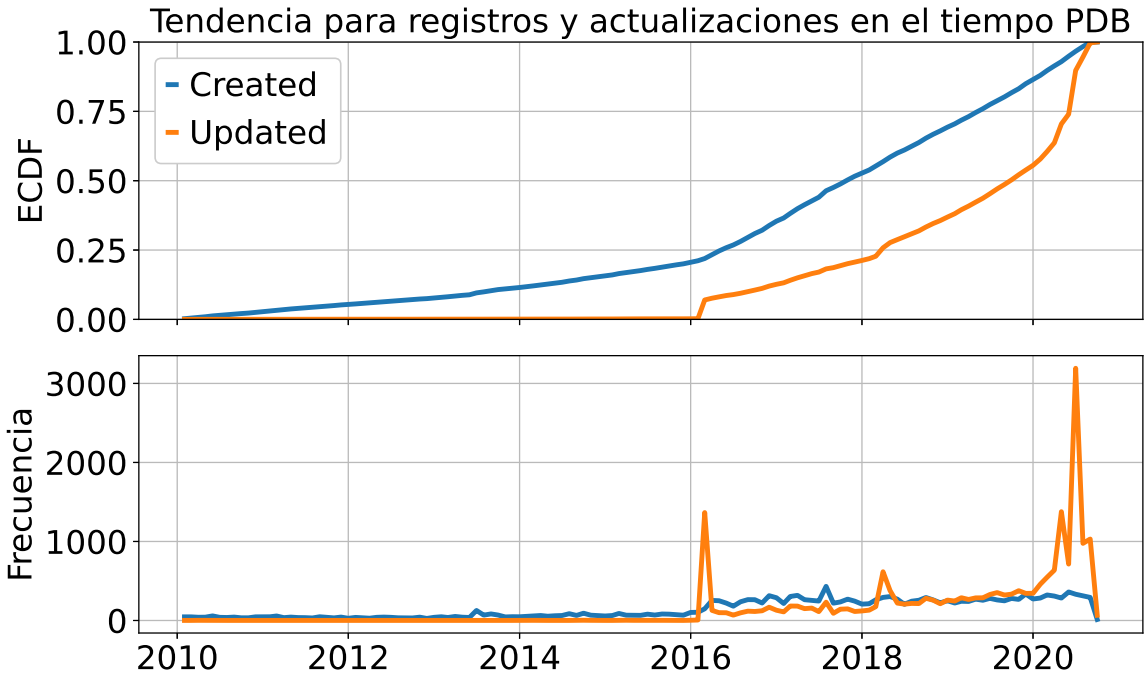


Figura 2.2: Tendencia para registros y modificaciones del año 2010 en adelante, a partir de los campos `created` y `updated` de PeeringDB. En el panel superior se muestra la función de distribución acumulada, mientras que en el panel inferior un polígono de frecuencias.

Volumen de datos			
Campo	*Registros Nulos*	*Registros No Nulos*	Registros con Presencia Númerica
Tambien Conocido Como	9512	10733	677
Notas	17559	2695	1488

Cuadro 2.4: Volumen de datos en campos `aka` y `notes` para la captura del 01/10/2020.

En la Figura 2.3 se observan los 20 países que más ASNs tienen registrados y cómo están representados en PDB. En primer lugar, para determinar la nacionalidad de un AS se recurrió a los archivos de delegación de los RIRs. Estos registros públicos documentan en qué país se encuentra registrada la organización a la cual fue delegado el recurso ASN. Luego, en función de esta definición de nacionalidad, se determinó la prevalencia de cada país en PDB. Dentro de los países mejor representados se encuentran Sudáfrica (61 %), Brasil (45 %) y Argentina (38 %). Se puede observar que dos de los países de más influencia dentro de Internet a nivel mundial están representados con menos del 12 %. Esto no quita que dentro de ese porcentaje se encuentren las organizaciones de mayor customer cone. Para examinar la tabla de países, sus respectivos ASNs y cobertura se puede consultar el Cuadro del Apéndice A.5 .

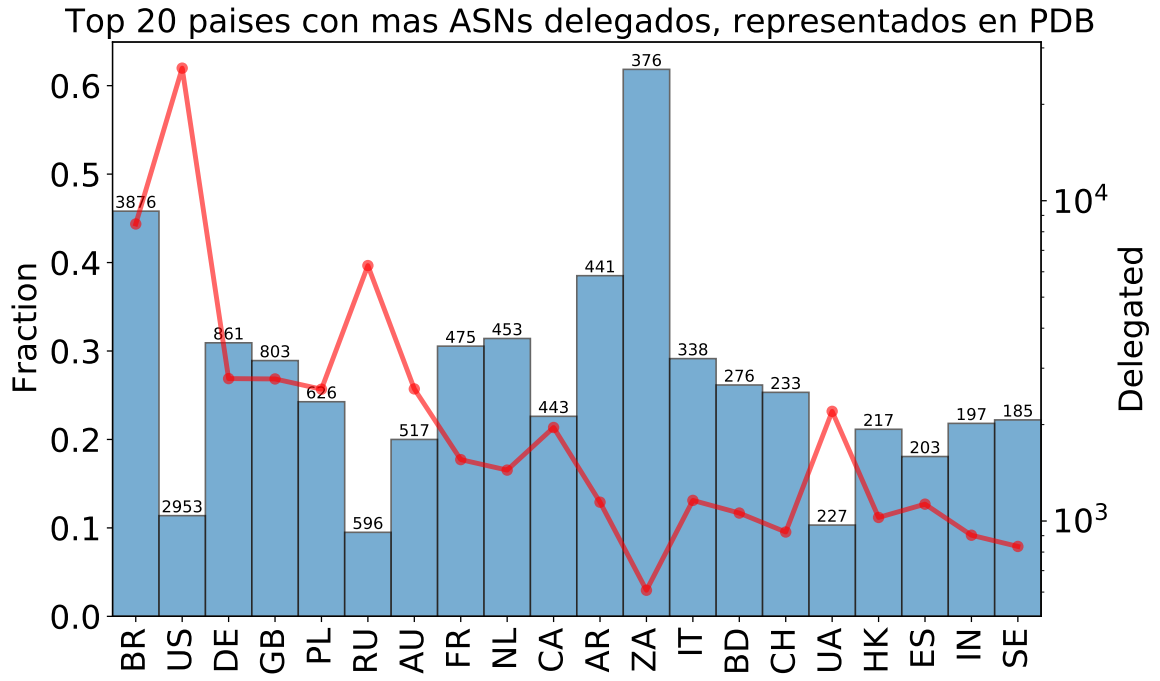


Figura 2.3: Ranking de los 20 países con mayor cantidad de ASNs delegados y su representación en PDB. A través del grafico de barras se ilustra la relación en forma fraccionaria con eje izquierdo. Sobre cada barra se indican la cantidad de entradas existentes en PDB para ese país. La curva roja indica la cantidad de ASNs delegados por los RIRs a cada pais con eje derecho.

Ahora enfocando nuestra atención en la semántica de los campos de texto libre haremos una descomposición del texto en n-gramas para clasificarlo a partir de su frecuencia utilizando TF-IDF [43] . El objetivo de este análisis es comprender, con mayor detalle, qué información transmiten los operadores en el campo `notes` evitando la lectura de los 2695 registros, con la intención de tener un análisis escalable para cualquier volumen de datos. Para un n-grama de tamaño 1 en el campo `notes`, se encontró a la palabra **peering** como la más frecuente del documento. También se encontraron las palabras *com, network, internet y asn* como sus siguientes como se ilustra en el Cuadro 2.5. En el estudio para n-gramas de tamaño 2, 3 y 4 se observó la misma tendencia, por lo que se puede concluir que el termino **peering** es un actor principal en la semántica del campo. En los Cuadros 2.6, 2.7 y 2.8 se detallan los 5 n-gramas de más frecuencia relativa para las distintas longitudes. Esto indica que en el campo se utilizan como expresiones más frecuentes, palabras vinculadas al dominio de Internet. Posteriormente se analizó el patrón *palabra+número*, con el objetivo de identificar las expresiones más usuales para introducir uno o más Sistemas Autónomos. A través de este análisis, se detectaron las expresiones por las cuales se hace uso de notación numérica, tanto para citar ASNs como para no hacerlo. Para el análisis del campo `notes`, el texto fue preprocesado, quitan-

Longitud 1	
N-grama	*TF-IDF*
peering	242.98
com	113.71
network	102.85
internet	83.37
peer	78.25

Cuadro 2.5: Ranking de los 5 n-gramas de tamaño 1 mas frecuentes por TF-IDF para el campo **notes** en PDB .

Longitud 2	
N-grama	*TF-IDF*
peering policy	69.47
peering requests	57.68
route servers	57.23
open peering	56.15
peeringdb com	48.21

Cuadro 2.6: Ranking de los 5 n-gramas de tamaño 2 mas frecuentes por TF-IDF para el campo **notes** en PDB .

Longitud 3	
N-grama	*TF-IDF*
open peering policy	68.28
internet service provider	36.68
www peeringdb com	22.46
asn used peering	20.78
welcome root asn	20.78

Cuadro 2.7: Ranking de los 5 n-gramas de tamaño 3 mas frecuentes por TF-IDF para el campo **notes** en PDB .

Longitud 4	
N-grama	*TF-IDF*
https www peeringdb com	24.88
peering root prefixes sourced	22.45
used peering root prefixes	22.45
asn used peering root	22.45
root prefixes sourced as3557	22.45

Cuadro 2.8: Ranking de los 5 n-gramas de tamaño 4 mas frecuentes por TF-IDF para el campo **notes** en PDB .

do caracteres especiales, números, espaciados múltiples y convirtiendo a minúsculas. Se utilizo la frecuencia de término (TF) para conocer cuantas veces una palabra ocurre en cada documento del campo `notes`. Se encuentran más significativas, aquellas que aparecen múltiples veces que las que no, al mismo tiempo si una palabra es frecuente en `notes` para los 2695 documentos, es porque la palabra si bien frecuente no necesariamente puede ser significativa (IDF). Por esta razón, fueron quitadas las *stop-words* del idioma inglés. A través de la combinación de estas dos métricas, se evaluó la relevancia de una palabra, en un documento, para una colección de documentos.

2.4. Metodología de inferencia de siblings

En función de lo presentado en las secciones anteriores, la metodología propuesta propone inferir siblings utilizando la información presente en los registros de PeeringDB. Este proceso se ilustra en la Figura 2.4 donde se inicia con el preprocesado de la captura seleccionada, para posteriormente extraer los siblings de los campos `org_id`, `aka` y `notes` a partir de técnicas distintas. Luego, se descartan números y ASNs que no corresponden a relaciones de siblings, a partir de procedimientos de filtrado específicos para cada campo de texto libre. Finalmente, se agrupan los listados de siblings extraidos para obtener

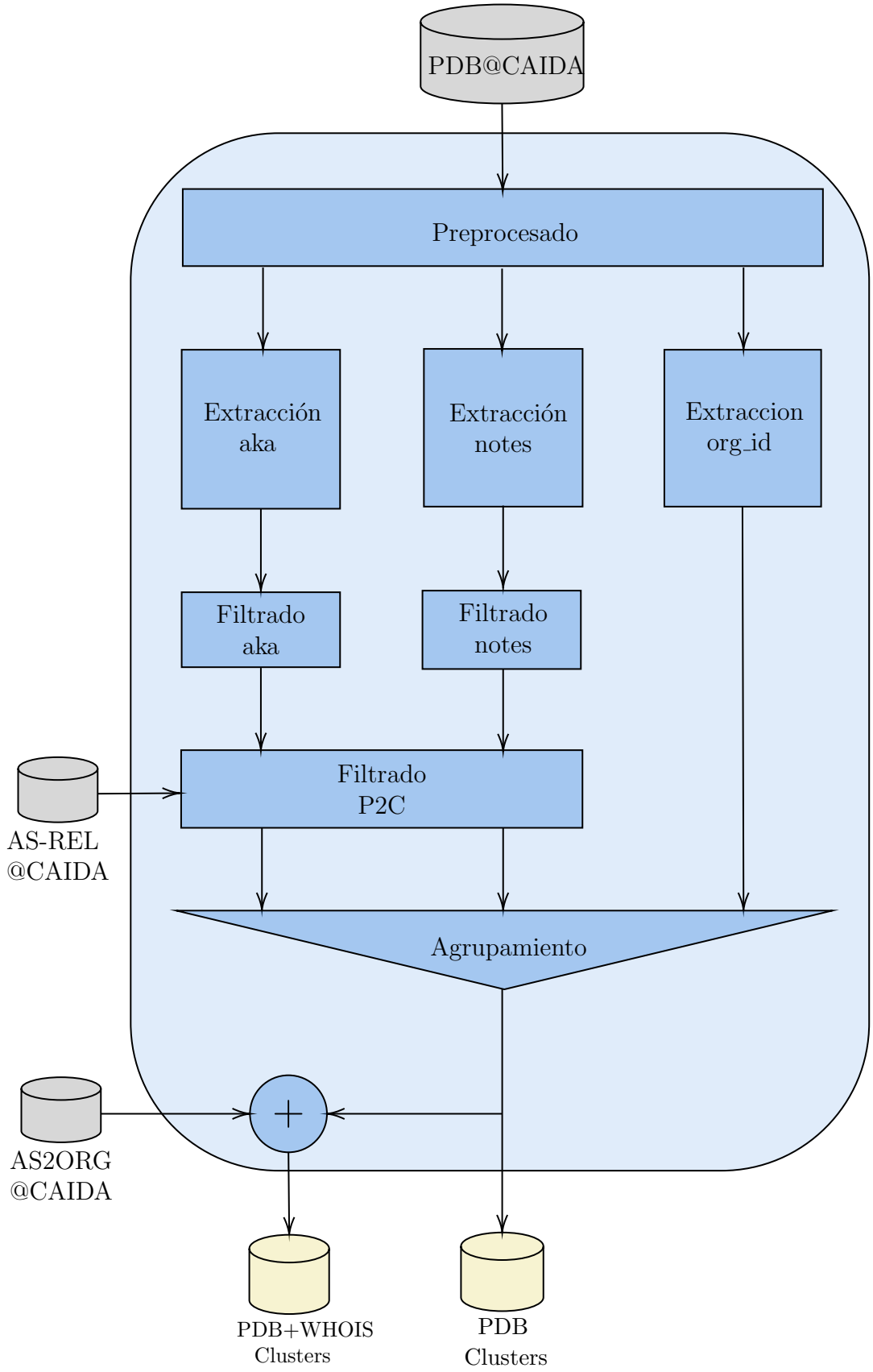


Figura 2.4: Estructura de la metodología propuesta para extraer siblings a partir de una captura de PeeringDB y generar clusters representativos de cada organización.

dos salidas diferentes. La primera, utilizando únicamente datos de PDB, mientras que la segunda, consiste en combinar el agrupamiento de PDB con los datos de WHOIS, presentes en el conjunto de datos AS2ORG [24] provisto por CAIDA. El objetivo de esta unión es representar el aporte que brinda la metodología propuesta al estado del arte.

2.4.1. Adquisición de datos

Las capturas diarias de PeeringDB son la fuente principal de datos de nuestra metodología. Para esto, nos basamos en la curaduría de datos de PeeringDB provista por CAIDA. El repositorio consta de dos partes, la versión 1 y la versión 2. Las capturas diarias de la versión 1 están disponibles como archivos `sql` y `sqlite` desde el 29 de julio de 2010 hasta el 13 de marzo de 2016. En 2016, PeeringDB cambió a un nuevo formato. Estos nuevos datos de la versión 2 están disponibles como archivos `sql` desde el 27 de mayo de 2016 hasta el 10 de marzo de 2018 y como archivos `json` desde el 11 de marzo de 2018 en adelante [36]. En el desarrollo de esta metodologia se utilizó, al igual que en el Análisis de 2.3, una captura del 1 de Octubre del 2020 [28].

2.4.2. Preprocesamiento de datos

En esta etapa se extrae de la estructura de datos, los campos `notes`, `aka` y `org_id`, que a partir del análisis previo, presentan características para inferir siblings. Las capturas de datos son provistas por medio de una estructura de datos en formato `JSON`, compuesto por multiples documentos, tales como `ixlan`, `api`, `fac` y `as_set`, entre otros. Como se ha explicado en la Sección 2.3.1, con los fines de extraer los siblings, el interés se reduce a los campos `notas`, `aka` y `org_id` de PeeringDB. Finalmente, se procesó el contenido alfanumérico de los campos de texto libre, de forma que sea uniforme en formato mayúscula.

2.4.3. Extracción de Sistemas Autónomos

El método de inferencia de siblings consiste en extraer de los campos `notas` y `aka` números que potencialmente sean ASNs, y a su vez tomar todos los ASNs reportados en la macroentidad `org_id`. A partir de las características identificadas para cada campo (ver Sección 2.3.1), se optó por generar una expresión regular (`regex`) de extracción específica para cada campo de texto libre, a fines de extraer la mayor cantidad de ASNs posibles.

Expresiones Directas		
Entrada	Regex	Salida
AS21202 IN THE NORDIC REGION	`AS[0-9]+`	[21202]
THIS AS IS BEING MIGRATED TO AS:5089	`AS:[0-9]+`	[5089]
OPERATING 2 ASNS (55818 AND 45147)	`ASNS.*`	[55818,45147]
THIS AS WILL BE MERGED SOON INTO AS 43646.	`AS [0-9]+`	[43646]
HAS 6 ORIGIN ASS: 10158, 45991, 38678, 9764, 7625, 38099	`ASS:.*`	[10158, 45991, 38678, 9764, 7625, 38099]
AUTONOMOUS SYSTEM (AS) 54113	`[(]AS[)] [0-9]+`	[54113]
IIG(ASN-58715) & ISP(ASN-63969)	`ASN-[0-9]+`	[58715, 63969]

Cuadro 2.9: Expresiones regulares directas en el campo notas.

2.4.3.1. Extracción de ASNs en el campo `Notes`

En función de la semántica estudiada en la Sección 2.3.2, se proponen 37 regexes cubriendo los patrones anteriormente identificados. El conjunto de expresiones regulares se puede clasificar en dos grupos, el primero hace énfasis sobre expresiones y combinaciones que se manifiestan de forma directa, en el Cuadro 2.9 se selecciona un subconjunto para ilustrar con ejemplos su funcionamiento. Mientras que el segundo grupo, indirecto, utiliza la semántica del texto para extraer ASNs. A través de ejemplos se muestra el resultado generado en el Cuadro 2.10.

El conjunto de expresiones regulares de extracción directa cuenta con 21 patrones que comparten entre sí el mismo patrón *Abreviatura de Sistema Autónomo + Número*. En la práctica, existe una variedad de formas de abreviar el termino **Sistemas Autónomos**, algunas de estas son *AS, ASN, ASNS, ASS y ASES*. Por lo tanto, de las expresiones directas se busca capturar el número inmediato posterior a una abreviatura. El resultado final de esta etapa es un listado de ASNs, obtenido a partir de todas las coincidencias

Expresiones Indirectas		
Entrada	*Regex*	*Salida*
ALSO MANAGES AS10987, 16486 AND 46498.	ALSO MANAGES.*	['10987', '16486', '46498']
ASN BEHIND 18200: 2198, 17480, 45345, 45461, 56055, 56089.	ASN BEHIND.*	['18200', '2198', '17480', '45345', '45461', '56055', '56089']
CRITEO ALSO MANAGES THE FOLLOWING ASNS: 44788, 53031, 55569	THE FOLLOWING ASNS.* ALSO MANAGES.*	['44788', '53031', '55569']
MERGING 8613 MERGING 31672	MERGING .*	['8613', '31672']
OTHER ASN'S WE CONTROL 62195, 133188, 133366	WE CONTROL .*	['62195', '133188', '133366']
WE ADMINISTERED ASN 28263, 262272, 53126 AND 265079.	ADMINISTERED ASN .*	['28669', '28263', '262272', '53126', '265079']
THIS ASN IS BEHIND 38195	IS BEHIND .*	['38195']
OTHER ASN UNDER 7303 ARE 10481 AND 10318.	ASN UNDER .*	['7303', '10481', '10318']

Cuadro 2.10: Expresiones regulares indirectas en el campo notas.

encontradas para cada patrón. El Cuadro 2.9 presenta algunos ejemplos de las expresiones regulares desarrolladas, junto con algunos casos reales y la `Salida` luego de aplicar estas expresiones. La columna `Entrada` de la tabla contiene un segmento de la nota y es el texto que se utiliza para comparar si la `regex` coincide en su patrón. En el primer ejemplo, *'AS21202 IN THENORDIC REGION'* se corresponde al patrón `AS[0-9]+`, de la entrada se obtiene *'AS21202'*, posteriormente se procesa la salida en un paso intermedio a fines de obtener solo contenido numérico. La entrada *'OPERATING 2 ASNS(55818 AND 45147)'* si bien aplica a un patrón directo, presenta una disposición de los ASNs que permite capturar al primero pero no al segundo. Se resolvió, utilizando la expresión regular `ASNS.*` que genera la salida intermedia *'ASNS(55818 AND 45147)'*, de la cual se extrae posteriormente solo los números.

El segundo grupo, de extracción indirecta, se compone de 15 expresiones regulares y comparten un comportamiento semántico que permite introducir a uno o más Sistemas Autónomos a través de n-gramas de longitud 1 a 3. Este grupo funciona como un complemento de todos aquellos números que efectivamente son ASNs, pero no preceden de una variante a la abreviatura de Sistema Autónomo. En la práctica, se observa que en reiterados casos se listan números que previamente fueron antecedidos por una expresión, es en ésta donde se centra la atención para ejecutar la extracción. El funcionamiento es análogo al descripto anteriormente, se presenta un subconjunto de estas expresiones en el Cuadro 2.10. En el primer ejemplo, la expresión *'ALSO MANAGES'* es la encargada de introducir el listado de siblings, se replica el comportamiento para las expresiones *' ASN BEHIND',' THE FOLLOWING ASNS',' MERGING','WE CONTROL','ADMINISTERED ASN', 'IS BEHIND','ASN UNDER'*, entre otras.

Las expresiones regulares indirectas proponen el operador `'.*'` al final de cada regex, con la intención de capturar todo el texto posterior e incluir ASNs en caso de existir. Este operador presupone un riesgo, que implica incluir números como los mencionados en el cuadro 2.11. Entre las expresiones, podemos encontrar la dirección física o IP de una entidad operadora, un número o prefijo telefónico para el contacto, un número involucrado en el contenido de una URL, RFCs, Normas ISO y datos numéricos vinculados prefijos. La extracción se ejecuta tomando esta estrategia de relajar las restricciones en las regex, para ponderar la inclusión de Sistemas Autónomos sobre dejarlos afuera.

2.4.3.2. Extracción de ASNs en el campo `Aka`

La extracción del campo `aka` se genera a partir de una única expresión regular que capture números de cuatro o más dígitos. A partir del estudio preliminar y el conocimiento adquirido en la exploración, se identificó que este campo contiene información

Expresiones que conducen a Falsos Positivos		
Expresion	*Valor*	*ASN*
Direcciones IP	LONDON: 193.239.117.31	20562
Números telefónicos	NOC CAN BE REACHED AT ALL TIMES AT 1-877-877-2120 OR +1-802-463-2111	5738
Normas ISO	ISO27001, IGT LEVEL 2, N3 AGGREGATOR.	48954
Comunidad 65535	WE ACCEPT & HONOR GRACEFUL_SHUTDOWN COMMUNITY 65535:0 (RFC 8326)	12779
RFCs	RFC-7454 IS ALSO SUPPORTED.	9201
Números a través de URLs.	HTTPS://WWW.PEERINGDB. COM/NET/1864	46038
Direcciones físicas.	578 LORIMER ST, PORT MELBOURNE VIC 3207	135639
Cantidad de prefijos	IPV4: 12000 IPV6: 2000	4826

Cuadro 2.11: Expresiones más frecuentes que conducen a falsos positivos en el campo `notes` de PDB.

Expresión Aka		
Entrada	Regex	Salida
Easynet (4589), MDNX (8190), NETCOM(5571), Solution1, CI-NET(8844), GRIFFIN(20500), Telecomplete/Fused(6320)	`\d{4,8}`	['4589', '8190', '5571', '8844', '20500', '6320']
K-NET, K-SYS (24904 60478)	`\d{4,8}`	['24904', '60478']
Mediaplex, Commission Junction, FastClick, Dotomi, ValueClick, SET.tv, 41041, 26762, 19834	`\d{4,8}`	['41041', '26762', '19834']
9722 18398 23741 23745 17999 9894 (IX Services)	`\d{4,8}`	['9722', '18398', '23741', '23745', '17999', '9894']
FKA AS29761	`\d{4,8}`	['29761']
Apple CDN AS6185	`\d{4,8}`	['6185']

Cuadro 2.12: Expresion regular para campo `aka` de PDB y las salidas obtenidas para cada entrada.

concisa y relevante. A su vez, el campo se limita a un largo máximo de 255 caracteres en el que concurrentemente se reportan ASNs en caso de haber contenido numérico. Al centrarse en una única expresión de cuatro o más dígitos, se dejan de lado 999 de 89.246 ASNs posibles, que representan el 1,11 % del espectro total delegado por los 5 RIRs a la fecha la captura. El Cuadro 2.12 ilustra algunos ejemplos de la extracción para el campo.

2.4.3.3. Extracción de ASNs en el campo `Org_id`

A través de los identificadores `org` del documento `org` y `org_id` del documento `net`, se genera una relación para unir a los documentosy luego agrupar el resultado por la clave primaria, obteniendo un listado de los ASNs involucrados. El procedimiento involucra a todos las entradas presentes en PDB, ya que cada una debe tener una organización asignada.

2.4.4. Filtrado de números espurios

Luego de la etapa de extracción es necesario remover ciertas cadenas que no son capaces de ser filtradas por los métodos de extracción de la etapa anterior. Para remover estos elementos que no corresponden a ASNs de siblings, la metodología incluye un filtro para los métodos de extracción en campos de texto libre.

2.4.4.1. Filtro de números espurios en campo `Notes`

El mecanismo planteado para remover los elementos espureos a la salida de las notas se basa en aplicar máscaras con 26 patrones diferentes, como se indica en alguno de los ejemplos del Cuadro 2.13 El filtro del campo notas tiene como objetivo remover números relacionados con:

(I) Números espurios frecuentemente presentes (e.g., 24x7x365)

(II) Números telefónicos y fechas

(III) Prefijos

Números espurios frecuentemente presentes

Los primeros 10 patrones obtenidos surgen de explorar el campo Notas y detectar los patrones numéricos más frecuentes. La Figura 2.5 presenta la frecuencia relativa de aparición de los primeros 10 números, indicando la existencia de un punto de inflexión al llegar al patrón 50. Entre estos números, podemos encontrar al 4 y 6 como los más frecuentes dado que se referencia en reiteradas ocasiones los protocolos *IPv4* e *IPv6*. Mientras que números como 50 y 100 establecen el límite de prefijos aceptados en la cadena para los protocolos IP.

Números telefónicos y fechas

Los prefijos telefónicos y las fechas de calendario son otros elementos numéricos que presentan inconvenientes a la salida de los métodos de extracción. Para resolver esta limitación se aplicaron, como indica el Cuadro 2.13, 9 patrones de dos dígitos. También se decidió incluir los años más recientes, dado que los operadores suelen referenciar la fecha al momento de reportar información, en especial los años 2019 y 2020.

Prefijos

A su vez, los prefijos IP expresados en notación decimal separada por puntos (Dotted-Decimal Notation) introducen otros números que no corresponden a ASNs y deben ser removidos. Asimismo, los operadores suelen compartir en los perfiles de PeeringDB la

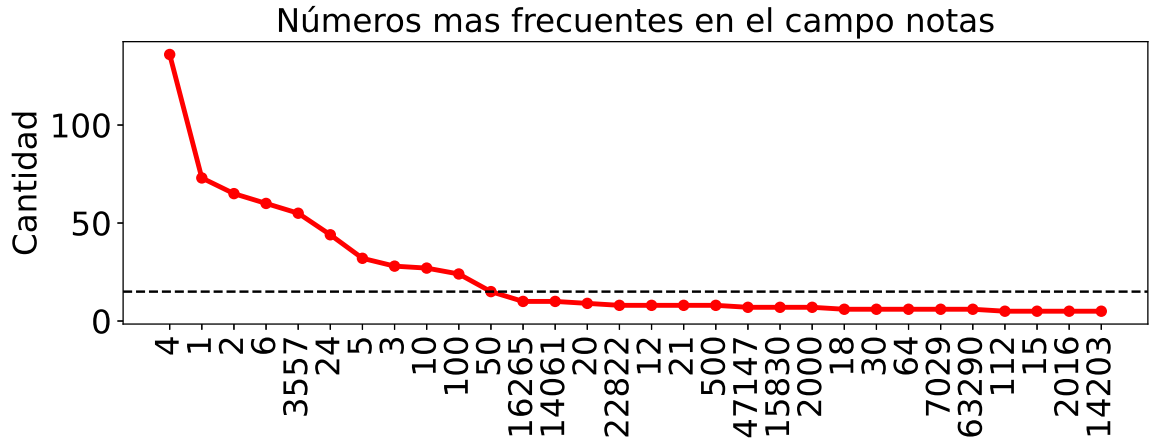


Figura 2.5: Números más frecuentes para el campo `notes`. Utilizando el método del codo se establece un *threshold* en línea punteada, para incluirlos como números espurios frecuentemente presentes.

Filtro de números espurios en `notes`	
Valores	*Categoría*
[4,1,2,6,5,3,10,100,50]	Números espurios frecuentemente presentes
[48,18,40,21,12,22,80,64,30,2019,2020]	Números telefónicos y fechas
[1000,400,252]	Prefijos

Cuadro 2.13: Valores filtrados como números espurios separados por distintas categorías para el campo `notes`

cantidad máxima de anuncios que admite su red. Esto se debe a que la mayor parte de los routers BGP limita por motivos de seguridad y performance la máxima cantidad de anuncios recibidos [3, 14]. Aunque esta es información sustancial para las operaciones, no involucra ASNs, por lo tanto, estos números deben ser descartados. La captura, en específico, presenta los patrones 252, 400 y 1000 del Cuadro 2.13 que fueron removidos.

2.4.4.2. Filtro de números espurios en campo `Aka`

Dado que la extracción considera únicamente números de al menos 4 dígitos, se propone filtrar únicamente fechas numéricas que remiten a años. La máscara de filtrado se compone de 51 números que comprenden la franja del año 1970 al 2021. La principal razón para eliminar este subconjunto se justifica con el uso que las organizaciones le dan al campo desde un punto de vista semántico. En reiteradas ocasiones se utiliza `aka` para informar acontecimientos que involucran años como ilustra el Cuadro 2.14. En

Filtro de números espurios en aka		
Entrada	*Categoría*	*ASN*
SWIPnet - TELE2/SWIPnet commercial Internet since 1991	Año desde que opera la organización	1257
City2020/ HAMCOM / LuenTEL	Año en nombre propio	12355
Kakao merged with Daum communications in 2014	Año de combinación	10158

Cuadro 2.14: Valores filtrados como números espurios separados por distintas categorías para el campo `aka`

los ejemplos se detallan años desde que opera la organización, años de combinaciones y adquisiciones con otras compañías y finalmente nombres propios que involucran un patrón de año+palabra.

2.4.5. Filtro proveedor a cliente (p2c)

Llegado a este punto del diagrama de la Figura 2.4, se cuenta con conjuntos de ASN inferidos de los campos `notes` y `aka`. Sin embargo, existe el riesgo de que, aunque estos números sean fehacientemente ASNs, estos no sean efectivamente siblings de la organización. Para descartar estos casos, se propone crear un filtro basado en aspectos topológicos de la red de ASes. Este filtro propone utilizar las relaciones comerciales del conjunto de datos de CAIDA [25].

Basándonos en las relaciones comerciales entre ASes definidas por Gao-Rexford, suponemos que los ASes inferidos como siblings serán en su mayoría clientes, pares o relaciones inexistentes[4]. Por lo tanto, si la mayoría de los ASes inferidos como siblings son proveedores del AS del que se generó la inferencia, probablemente el operador este reportando sus *upstream providers*. Suponemos que el hecho de reportar esta información, puede ser con el motivo de brindar visibilidad de la conectividad del AS, dando muestras de su confiabilidad, con el objetivo de atraer clientes. En base a estas suposiciones, el filtro p2c determina que se debe descartar el cluster si existe más de un proveedor entre los ASes inferidos. Definimos ASNx o ASN primario, como el Sistema Autónomo registrado por el operador en PDB, del cual se infieren siblings.

Las Figuras 2.6 y 2.7 ilustran el funcionamiento del filtro p2c para los dos escenarios posibles, el primero de rechazo y el otro de aceptación. Maxihost LTDA (AS262.287) es un datacenter que conecta una serie de ISPs, la Figura 2.6 presenta el funcionamiento.

[4]No es estrictamente necesario que los siblings cuenten con conectividad física entre sí.

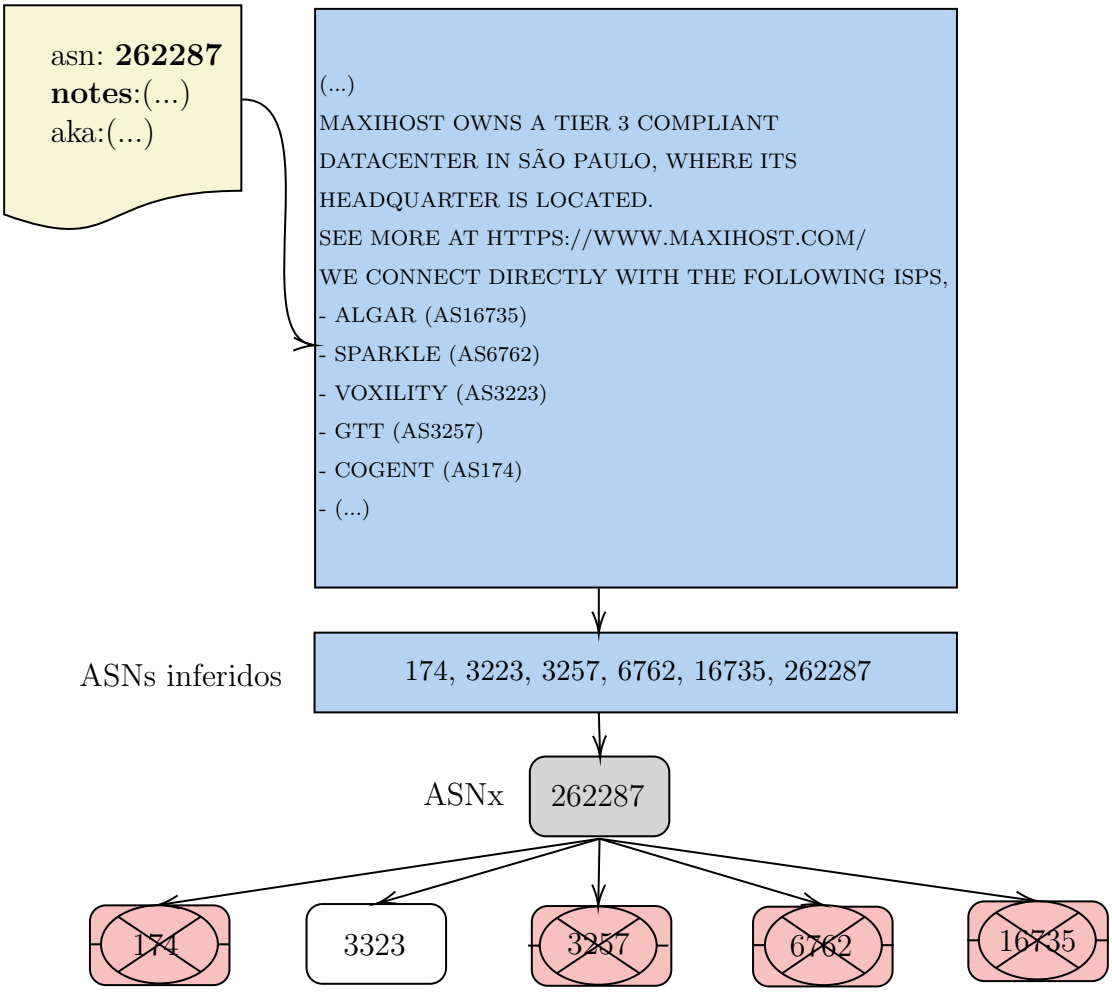


Figura 2.6: Filtrado p2c aplicado al ASNx 262.287 (Maxihost LTDA). Se rechaza la inferencia de siblings, ya que AS174, AS3257, AS6762 y AS16735 presentan una relación de proveedor frente al ASN primario.

En primer lugar, se extrae de una nota proveniente del ASNx 262.287 su listado de siblings. Luego, se contrasta para cada ASN inferido la relación frente al primario, en caso de existir más de una relación proveedor a cliente en el conjunto, se lo descarta. En este ejemplo en particular Cogent (AS174), GTT (AS3257), Telecom Italia Sparkle (AS6762) y Algar Telecom (AS16735) son proveedores de Maxihost LTDA (AS262.287). La Figura 2.7 demuestra un caso de aceptación para el AS7303 (Telecom Argentina), de este ASN principal se infiere el proveedor AS10481, previamente perteneciente a Fibertel. Es frecuente observar que las organizaciones que operan múltiples ASNs cuentan con un AS más expuesto. Es decir, una AS con mayor cantidad de vínculos. En caso de esta naturaleza, el AS más expuesto puede ser inferido como el proveedor de uno de sus siblings. Para no descartar clusters válidos como este, flexibilizamos la regla para que exista hasta un máximo de un proveedor.

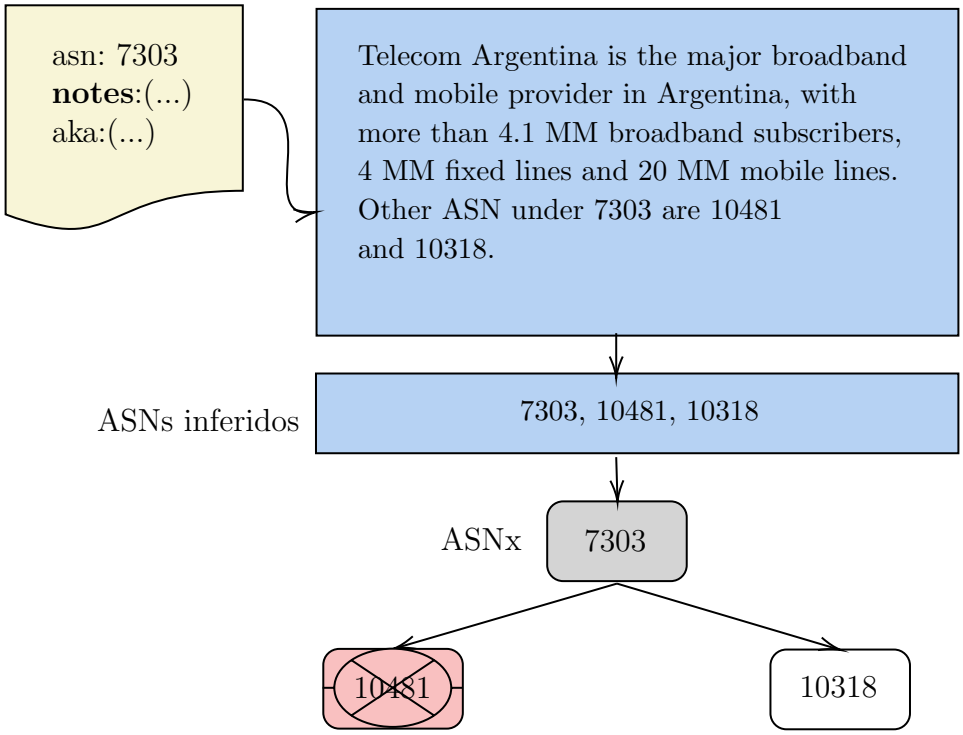


Figura 2.7: Filtrado p2c aplicado al ASNx 7303 (Telecom Argentina). Se acepta la inferencia de siblings, ya que la metodologia permite como máximo un proveedor al ASN principal, en este caso AS10481 (Fibertel).

2.4.6. Agrupamiento

Luego de concluir con las etapas de extracción y filtrado, es necesario combinar los conjuntos inferidos a través de cada uno de los campos. En la entrada se procesan todos los listados de siblings extraidos para los campos `notes`, `aka` y `org_id`. En esta etapa, se agrupan todos los clusters inferidos que cuenten con algún ASN en común. Es importante destacar que dados tres conjuntos de siblings inferidos A, B, y C pertenecientes al mismo cluster final, existe la posibilidad que $A \cap B = AS_0,\ B \cap C = AS_1\ y\ A \cap C\ = \emptyset$. La Figura 2.8 describe el funcionamiento con un ejemplo para la organización Anexia (AS47147). Centrando la atención en las columnas del grafico, se puede deducir el cluster final, mientras que de las filas, el campo del cual se infieren los siblings. A través de las celdas coloreadas distinguimos las intersecciones de conjuntos, en este ejemplo en particular el campo `org_id` reúne a 9 de los 11 ASNs que conforman el cluster, sin embargo, por medio del campo `notes` obtenemos los restantes. Tanto AS42360 como AS13902 son Sistemas Autónomos que se agrupan al cluster únicamente por `notes`.

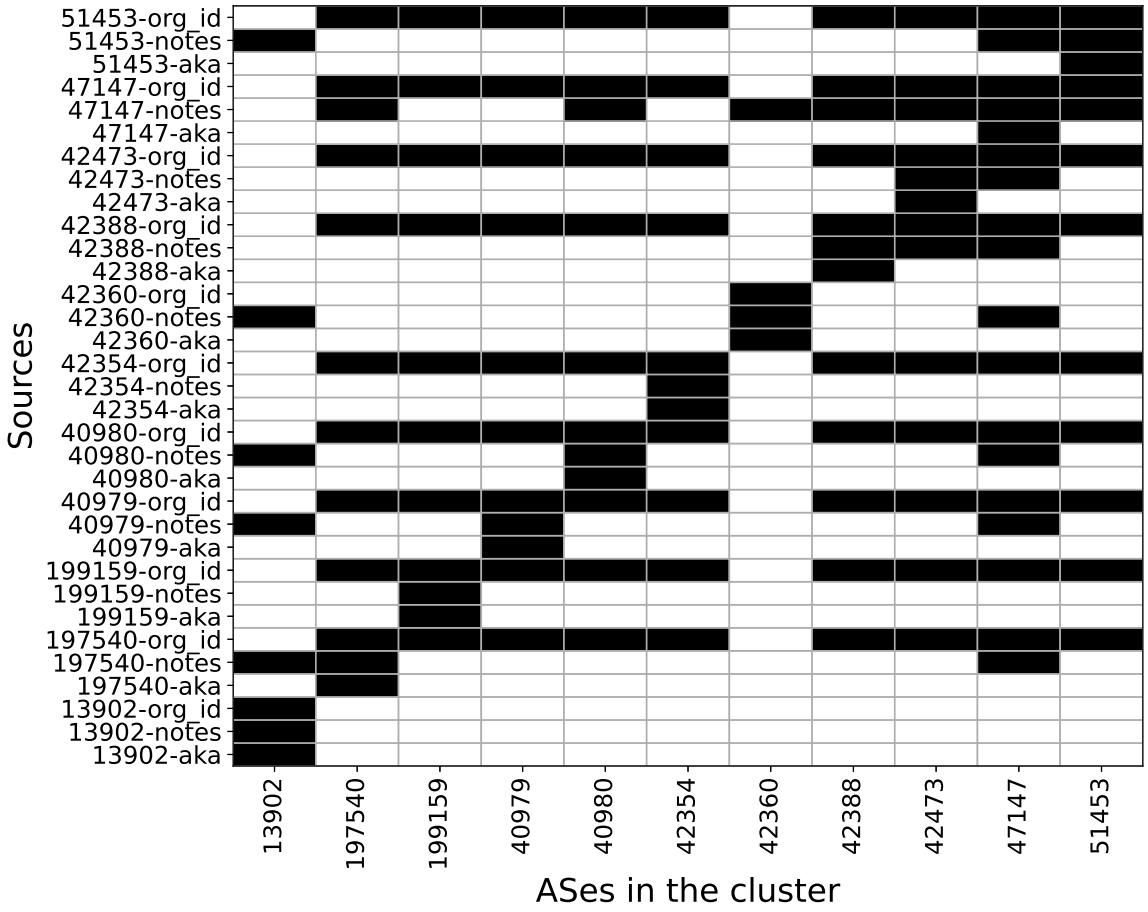


Figura 2.8: Intersección de conjuntos para agrupamiento con *Anexia*. A través de las columnas se deduce el cluster final obtenido utilizando PeeringDB, mientras que de las filas el campo del cual se infieren cada uno de los siblings.

2.4.7. Unión con siblings inferidos mediante WHOIS

El último paso del proceso (Fig. 2.4) consiste en combinar los siblings inferidos con PeeringDB con el conjunto de datos de siblings de AS2Org generado a partir de los datos de WHOIS [24]. El objetivo de esta combinación es aportar al estado del arte actual, a fines de representar el tamaño de las organizaciones en Internet con una nueva aproximación. Es importante destacar que la metodología de AS2Org a partir de los datos de WHOIS no cuenta con falsos positivos, sino que su limitación, dado los datos utilizados, son los falsos negativos. Por este motivo esta metodología aún tiene vigencia y es recomendable incorporarla a la metodología propuesta en esta tesis.

El procedimiento es análogo al descripto en la Seccion 2.4.6, donde partimos del agrupamiento realizado en la metodología para comparar uno a uno con los conjuntos de WHOIS, generando un nuevo cluster cuando existe un ASN en común. La Figura 2.9 ilustra como se obtiene una nueva presentación de Google en Internet utilizando la

metodología propuesta. En primer lugar, los Sistemas Autónomos se encuentran dispersos, sabiendo que fueron alocados por los distintos RIRs, a priori no existe una relación entre ellos que los vincule como representamos en la Figura 2.9a. Luego, aplicando la metodología propuesta por Cai *et al.* [9] utilizando datos de WHOIS, observamos 12 organizaciones en clusters diferenciadas por color, cada una con su respectiva etiqueta identificativa de la organización (Fig. 2.9b). Finalmente, ilustramos en la Figura 2.9c el agrupamiento obtenido luego de aplicar la metodología desarrollada con los datos de PeeringDB, en donde se unen distintas unidades de negocio que pertenecen a Google, como Youtube, Google Cloud y Google Fiber, entre otras.

2.5. Validación

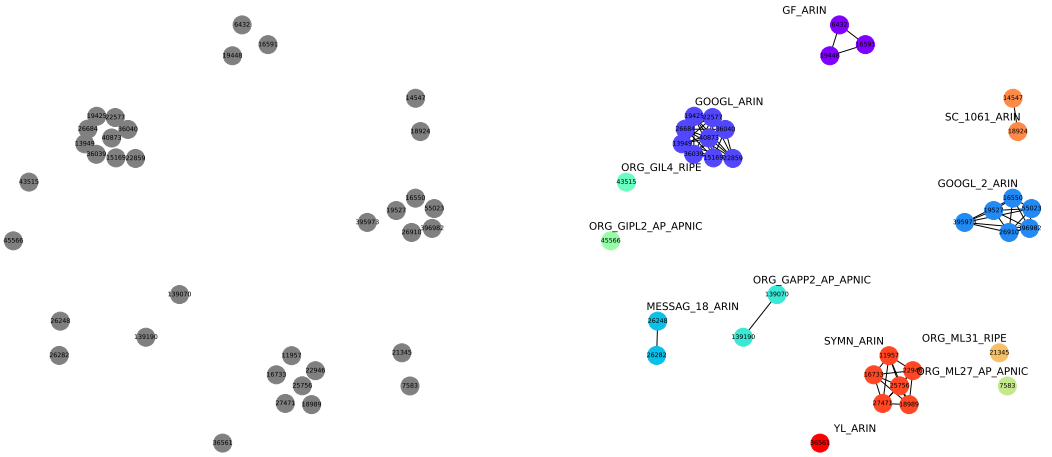
El objetivo de la validación es valorar las extracciones de la metodología a través de distintas evaluaciones y métricas. Los campos donde centramos el interés, son los de texto libre, ya que a diferencia del campo `org_id` existe un margen de error posible en la extracción. En cada una de las evaluaciones realizadas, se define un *verdadero positivo* (TP), *verdadero negativo* (TN), *falso positivo* (FP) y *falso negativo* (FN) de forma distinta, ya que cada evaluación propone un objetivo diferente. La etapa de validación se ha considerado luego del preprocesamiento, donde sólo se tienen en cuenta las entradas que cuentan con presencia numérica. Es importante destacar, que todas las validaciones fueron de inspección manual, es decir, que cada salida obtenida fue revisada y etiquetada por un humano.

Todas las evaluaciones son cuantificadas a través de las métricas de exactitud (del inglés, *Accuracy*), precisión exactitud (del inglés, *Precision*) y exahustividad (del inglés, *Recall*). La Ecuacion 2.1 describe la exactitud, que mide la proporción de aciertos, sin embargo si la clase objetivo es desbalanceada, está sola no es suficiente. Por esta razón, se complementó con la métrica de la Ecuación 2.2 ya que con la precisión podemos estudiar la calidad de la metodología a partir de los casos positivos, y por último de la Ecuacion 2.3 evaluar cómo responde la metodología en volumen, para identificar la clase objetivo.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{2.1}$$

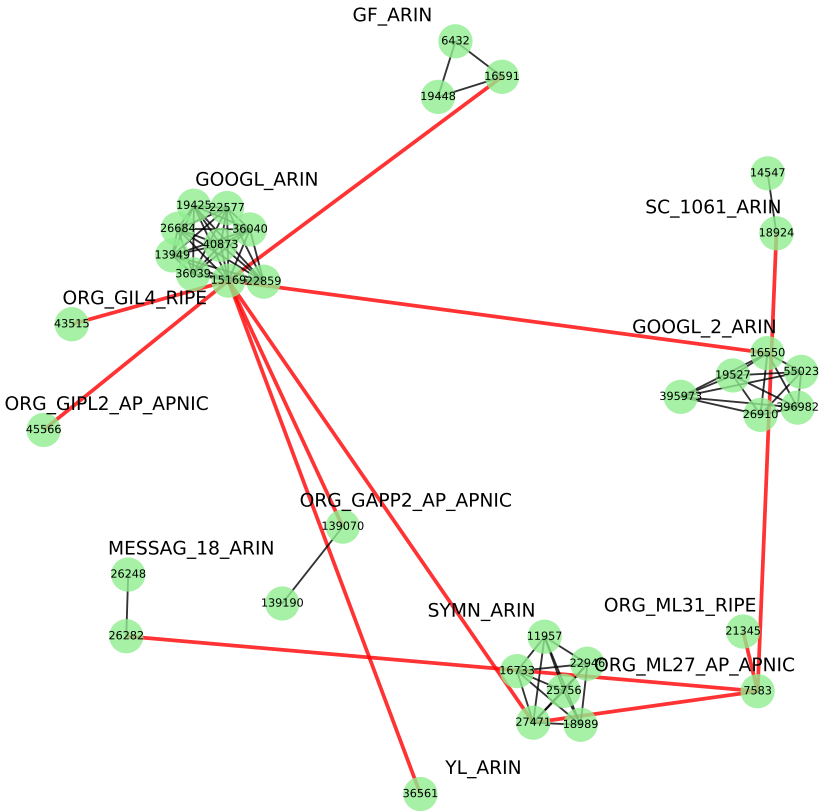
$$Precision = \frac{tp}{tp + fp} \tag{2.2}$$

$$Recall = \frac{tp}{tp + fn} \tag{2.3}$$



(a) ASNs crudos sin aplicar metodologías.

(b) Organizaciones representadas por metodología de AS2ORG.



(c) Organización final representada para Google por la metodología implementada en combinación con AS2ORG.

Figura 2.9: Evolución en la representación del conglomerado de Google a través de las metodologías.

Campo	Año	Evaluación	*n*	*fn*	*fp*	*tp*	tn	**Accuracy**	**Precision**	**Recall**
Notes	2020	A	1448	13	44	733	658	0.960	0.943	0.982
Notes	2020	B	790	13	8	752	17	0.973	0.989	0.983
Notes	2020	C	1394	17	17	1326	34	0.975	0.987	0.987
Notes	2016	A	753	13	34	273	433	0.937	0.889	0.954
Notes	2016	B	320	13	4	294	9	0.946	0.986	0.957
Notes	2016	C	639	41	14	561	23	0.913	0.975	0.931
Notes	2016	D	311	10	18	85	198	0.909	0.825	0.894
Notes	2016	D y B	113	10	11	90	2	0.814	0.891	0.900
Notes	2016	D y C	215	35	13	156	11	0.776	0.923	0.816
Notes	2010	A	247	5	4	80	158	0.963	0.952	0.941
Notes	2010	B	89	5	0	83	1	0.943	1.000	0.943
Notes	2010	D	151	5	2	41	103	0.953	0.953	0.891
Aka	2020	A	677	0	14	216	447	0.979	0.939	1.000
Aka	2020	B	230	0	3	218	9	0.986	0.986	1.000
Aka	2020	C	334	0	3	321	10	0.991	0.990	1.000
Aka	2016	D	110	0	0	103	7	1.000	1.000	1.000

Cuadro 2.15: Tabla de validación para campos de texto libre `notes` y `aka`.

Dado que el propósito es generar una metodología de evaluación sistemática para evaluar la extracción de siblings, abordamos la validación por medio de 4 instancias (*A*, *B*, *C* y *D*) con distintos objetivos. La validación *A* de la metodología, evalúa la clasificación de la nota en su completitud. Se considera un *verdadero positivo* aquella nota de la cual se lograron extraer todos los siblings posibles, mientras que un *verdadero negativo* es aquella nota que si bien tiene contenido numérico, no contiene ASNs y en defecto, es detectada como tal. Un *falso positivo*, es aquella salida que consta de al menos un registro numérico que no pertenece a un Sistema Autónomo. Una nota es clasificada como *falso negativo* cuando no se logró extraer el ASN, a pesar de estar disponible. La segunda validación de la metodología, tipo *B*, parte de un subconjunto del tipo *A*, centrando el objetivo en aquellas notas que contengan Sistemas Autónomos de forma estricta. La evaluación es considerada luego de aplicar el filtro de números espurios, para reducir la cantidad de falsos positivos, que luego de procesados transformarán a la salida en *verdaderos positivos*, *verdaderos negativos* o en su defecto, quedando igual. El criterio definido para *verdadero positivo*, *falso positivo* y *falso negativo* es análogo al inicial, con la salvedad de que los *verdaderos negativos* solo se obtienen de la reducción de *falsos positivos* de la instancia *A*. La validación tipo *C*, cambia la granularidad por una más específica, se clasifica por Sistema Autónomo en reemplazo de una nota en su totalidad como se hacía para las instancias *A* y *B*. El criterio definido para *verdadero positivo*, *verdadero negativo*, *falso positivo* y *falso negativo*, es análogo al de tipo *B*. La validación *D* evalúa clusters de salida condicionados a que la información del campo notas haya

sido alterada entre dos capturas. Esto lo podemos llamar validación de diferencia, y para poder identificar que este campo ha variado su contenido, utilizamos la distancia de Hamming [17]. El criterio definido para *verdadero positivo*, *verdadero negativo*, *falso positivo* y *falso negativo*, es análogo al de la instancia A. Todos los criterios explicados aplican de la misma forma para el campo `aka`.

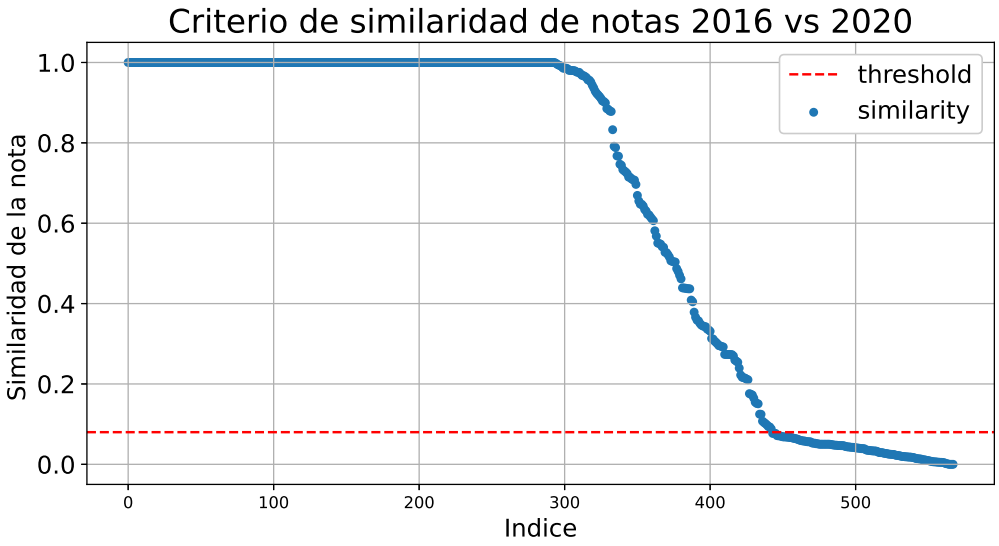


Figura 2.10: Criterio establecido para tomar una nota de la captura 2016 en la evaluación D. Se comparan todos los ASNs que presentan notas en 2020 y 2016 para medir la similaridad de su contenido, estableciendo un *threshold* en 0.09 para considerar la nota.

El Cuadro 2.15 ilustra los resultados obtenidos para las 16 validaciones realizadas sobre los campos `notes` y `aka`. La primera evaluación de tipo A, utiliza la captura del 1 de octubre del 2020 y los resultados obtenidos para 1448 notas son de 0.960 Accuracy, 0.943 Precision y 0.982 Recall. El tipo B, sobre el mismo conjunto de datos, hace énfasis en notas que contengan Sistemas Autónomos, dejando de lado las que no. Por esta razón, se parte de un subconjunto de 790 notas con un resultado de 0.973 Accuracy, 0.989 Precision y 0.983 Recall. La evaluación de instancia C, es equivalente a la B pero con un cambio de granularidad, que evalúa por número en vez de nota. El objetivo es evaluar desde otro punto de vista, ya que una nota puede contener una cantidad ilimitada de números. Por ejemplo, dentro de las 790 notas de B se encontraron 1394 números, de los cuales 1326 corresponden a Sistemas Autónomos bien clasificados, los resultados obtenidos de C fueron de 0.975 Accuracy, 0.987 Precision y 0.987 Recall. Luego, cambiamos los datos de entrada evaluando las mismas instancias con una captura del 27 de mayo de 2016 [27]. La evaluación D es una validación de diferencia entre las capturas de ambos años, quedando un subconjunto de 311 notas. El objetivo es garantizar que se evalúen entradas nuevas, para lograrlo, utilizamos la distancia de Hamming, que se computa superponiendo un

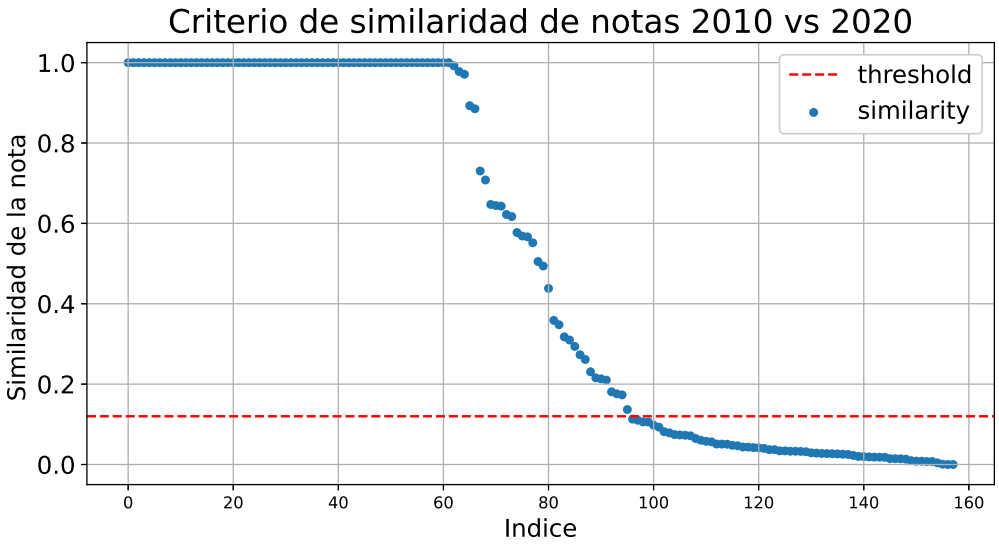


Figura 2.11: Criterio establecido para tomar una nota de la captura 2010 en la evaluación D. Se comparan todos los ASNs que presentan notas en 2020 y 2010 para medir la similaridad de su contenido, estableciendo un *threshold* en 0.12 para considerar la nota.

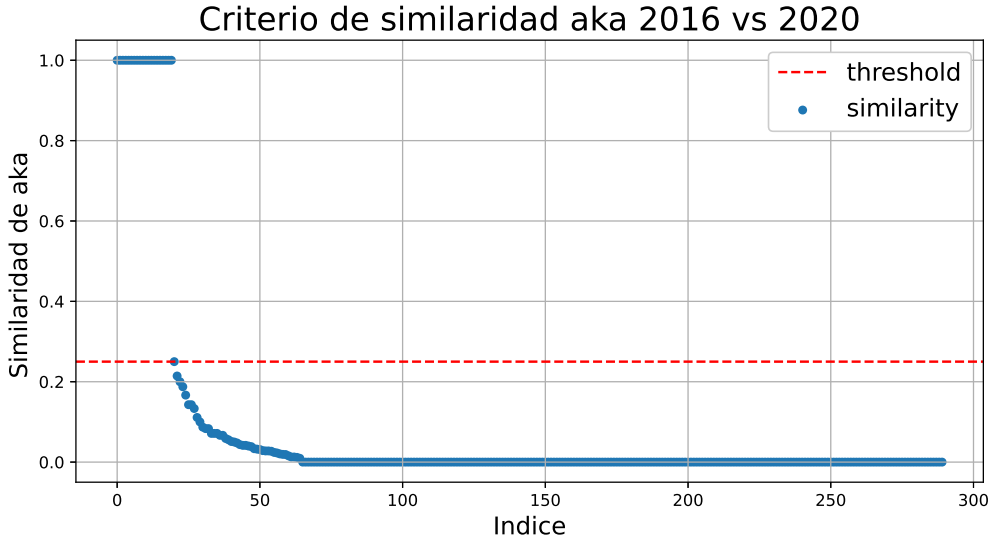


Figura 2.12: Criterio establecido para tomar una entrada de `aka` en la captura 2016 para la evaluación D. Se comparan los ASNs que presentan contenido en `aka` en 2020 y 2016 para medir la similaridad de su contenido, estableciendo un *threshold* en 0.25 para considerar la entrada.

---

string sobre el otro y encontrando los lugares donde existen variaciones, a índice bajo la relación entre strings es cada vez menor. El criterio establecido para considerar una nota suficientemente distinta se basó en el método del codo, estableciendo un *threshold* de 0,09 como se ilustra en la Figura 2.10. El resultado obtenido para D es de 0.909 Accuracy,

0.825 Precision y 0.894 Recall. Asimismo, se aplicaron las instancias B y C sobre la evaluación anterior, condicionando los datos para obtener un resultado en DyB de 0.814 Accuracy, 0.891 Precision, 0.900 Recall y para DyC 0.776 Accuracy, 0.923 Precision, 0.816 Recall. Finalmente, se repitieron las instancias A, B, C y D para una captura del 1 de diciembre del 2010 [26] cuyos los resultados se plasman en el Cuadro 2.15. El *threshold* utilizado en D para la similaridad de las notas de 2010 y 2020, fue de 0.12 como se ilustra en la Figura 2.11.

Por otro lado, realizamos las validaciones correspondientes para el campo `aka`. La instancia A, utiliza una captura del 1 de octubre del 2020, de la que se obtienen 677 registros con resultados de 0.979 Accuracy, 0.939 Precision y 1.000 Recall. El procedimiento para B es análogo al campo `notas`, de los 230 registros disponibles en `aka` obtuvimos 0.986 Accuracy, 0.986 Precision y 1.000 Recall. Dado que C es el equivalente a B pero con un cambio de granularidad los resultados son 0.991 Accuracy, 0.990 Precision y 1.000 Recall. La última validación para D, utilizando la captura del 2016 muestra la efectividad de la extracción, de las 110 entradas nuevas respecto de 2020, todas se clasifican correctamente. El *threshold* utilizado fue de 0.25, como se ilustra en la Figura 2.12.

2.6. Limitaciones

A continuación se listan limitaciones encontradas en la metodología propuesta.

1. No todos los ASNs inferidos a través de los campos `notes` y `aka` se pueden vincular a Organizaciones presentes en PDB. Sólo se vinculan los ASes y Organizaciones registrados a través de `org_id`.

2. No todos los ASNs inferidos a través de los campos `notes` y `aka` poseen una entrada en PDB.

3. La captura del 1 de Octubre del 2020, utilizada en la metodología equivale al 22,68 % de los ASNs delegados por los 5 RIRs.

4. El hecho de no implementar un preprocesado sobre el texto del campo `notes`, que incluya la eliminación de caracteres especiales, eliminación de stop words y aplicación de análisis morfológicos (Stemming and lemmatization) puede haber impactado en el hecho de usar una excesiva cantidad de regexes.

5. Utilizar los operadores `.*` en el regex, incrementa la inclusión de valores numéricos no deseados. Técnicas más refinadas quedan para trabajo futuro.

6. Se propone una definición de ASN primario en la sección 2.4.5 para aplicar el filtro p2c y establecer un marco teórico. En la practica no todas las organizaciones poseen un ASN de estas características.

7. El filtrado de números espurios tanto para el campo notas como aka, implica eliminar números que pueden ser potenciales ASN.

8. La validación de la metodología depende del factor humano a la hora de etiquetar los resultados.

9. Existe la posibilidad de que la entrada de un *ASX* esté reportando que los servicios de su compañía hayan migrado a los servicios de otro *ASY*. Por lo tanto, se podría inferir de forma incorrecta que el *ASX* y el *ASY* conforman el mismo conglomerado, cuando no es correcto.

Capítulo 3

Resultados

En este capítulo presentamos los resultados obtenidos de la metodología propuesta para la inferencia de siblings (Sección 3.1). A través de visualizaciones, plasmamos el aporte de los agrupamientos obtenidos en combinación con AS2ORG (Sección 3.2), junto a su impacto en los grandes proveedores de tránsito (Sección 3.3). A lo largo de este capítulo analizaremos frecuentemente clusters de diferentes tamaños, y los mencionaremos de la forma: `<N>`, para indicar un cluster de tamaño N. Por último, exploramos una alternativa al trabajo realizado, utilizando dendrogramas (Sección 3.4).

3.1. Resultados para metodología propuesta con PeeringDB

El primer paso es investigar los resultados de la metodología propuesta para la extracción de siblings a partir de los datos de PDB. El Cuadro 3.1 presenta los subconjuntos extraídos, los agrupamientos logrados y la cantidad total de ASNs involucrados en función de los campos utilizados para la extracción. El Cuadro 3.1 muestra que por medio del campo `notes` es posible extraer 660 subconjuntos que luego de ser procesados y agrupados forman 452 conjuntos representativos de una organización diferente totalizando la suma de 1048 ASes. Por medio del campo `aka`, la metodología infiere 226 subconjuntos que se pueden agrupar en 212 y contienen ASNs. A través del identificador de Organizaciones (`org_id`) ya presente en la esctructura de datos de PDB, se encuentran 18.369 organizaciones con un total de 20.245 ASes. Luego, combinando los campos `notes+aka` se obtienen 638 agrupamientos que reúnen a 1364 Sistemas Autonomos. Por último, la combinación más completa de la metodología, para los campos `notes+aka+org_id`, extrae 19.255 subconjuntos y 18.179 agrupamientos, totalizando 20.438 siblings.

Resultados de Metodologia con PeeringDB			
Campo	*Subconjuntos Extraidos*	*Agrupamientos*	*Cantidad de ASNs*
Notes	660	452	1048
Aka	226	212	361
Org_id	18.369	18.369	20.245
Notes+Aka	886	638	1364
Notes+Aka+Org_id	19.255	18.179	20.438

Cuadro 3.1: Cantidad de agrupamientos obtenidos a partir de las extracciones para los campos `notes`, `aka`, `org_id`, `notes+aka`, `notes+aka+org_id`

3.2. Resultados de metodología en combinación con AS2ORG

En esta sección analizamos los resultados obtenidos de la segunda salida de la metodología, que permite unir los siblings inferidos de AS2ORG [9], a fines de incrementar el conocimiento de las organizaciones en Internet. Una síntesis de los resultados se ilustra en la Figura 3.1, una función de distribución acumulada que representa la prevalencia que tiene cada tamaño de agrupamiento a partir de las distintas metodologías. La tendencia muestra que al menos un 90 % de los clusters de `<1>` y el agregado de información fomenta a que una parte de estos migre hacia agrupamientos de mayor tamaño. De acuerdo con los datos disponibles en el archivo de relaciones de CAIDA [25] del 1 de octubre del 2020, aproximadamente el 90 % de los ASNs existentes son *stubs*, es decir, que en una relación comerciales entre ASes, son únicamente clientes. Este concepto nos permite suponer que un gran número de estos *stubs* son en verdad el único ASN de pequeñas organizaciones. Por lo tanto, cualquier metodología propuesta para la inferencia de siblings se verá limitada por la estructura natural de Internet, por esta razón, se centra el foco en el impacto de cada sibling inferido, por encima de la cantidad.

Desde un punto de vista cuantitativo, podemos observar en la Figura 3.2 la tendencia para agrupamientos de `<1>`, a partir de las distintas metodologías de inferencia. El punto de partida es la metodología AS2Org que cuenta con 73.645 organizaciones de cluster atómicos. Luego, a medida que se incorporan siblings inferidos a través de los diferentes campos de PDB, el número se reduce hasta alcanzar 72.886 (una caida de %1.03), cuando se consideran todos los campos. La Figura 3.3 extendiente este análisis hasta los clusters de `<10>` ASes. A partir del cociente entre metodologías, vemos que no hay una tendencia clara de distrbucion en ASNs luego de incluir datos de PeeringDB. Esto se explica con

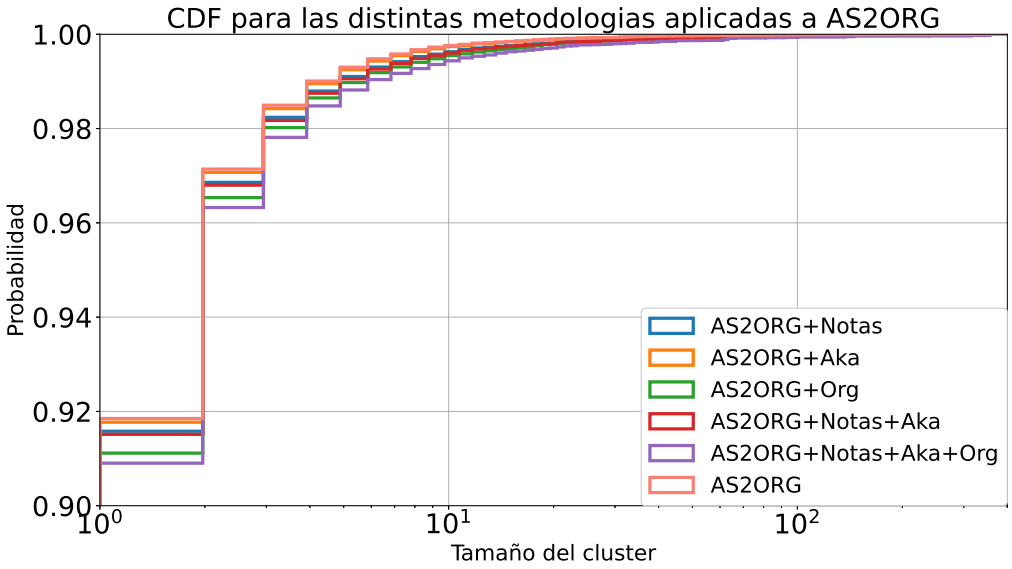


Figura 3.1: Función de distribución acumulada para todas las metodologías aplicadas en combinación con AS2ORG. En el eje Y distinguimos la probabilidad mientras que en el eje X el tamaño del cluster.

---

dos escenarios posibles, el primero es que luego de aplicada la metodología el tamaño de un cluster aumente y para que esto suceda, el tamaño de otro cluster tuvo que disminuir. Por ejemplo, para conglomerados de `<2>` existen 4242 por la metodología de AS2ORG, luego de aplicar `AS2ORG+notas+aka+org` se reduce su cantidad en 50, dando un total de 4192 organizaciones. Sin embargo, puede ocurrir el caso contrario en donde luego de aplicar la metodología la cantidad de conglomerados aumente, como es el caso de `<10>`. Como se muestra en el ejemplo de la Figura 3.4, Chief Telecom (AS17408) es una organización subrepresentada por otras 3 en AS2ORG, que luego de combinarse por la aplicación de la metodología, forman un único cluster de `<10>`.

La combinación de las metodologías PDB y AS2ORG genera la unión de conjuntos presentes en la metodología AS2ORG. La Figura 3.5 muestra para cada tamaño final $x \in (1, 20)$ de conglomerado (fila del gráfico), la distribución de la cantidad N de organizaciones que debieron unirse para formar un cluster `<X>`. Tomando como referencia organizaciones de tamaño `<10>` vemos que originalmente 21 conglomerados (%63.3) eran clusters de `<10>` y no se lograron unir a través de la nueva metodología. Sin embargo, se lograron juntar 4 veces, 2 organizaciones (%12.1) que conforman un conglomerado de 10 ASNs, también se logró juntar 2 veces, 4 organizaciones que formaban un cluster de 10 (%6). En último lugar, se logró juntar una sola vez a 8 organizaciones (%3), cuya suma de ASNs forman 10. La figura muestra el potencial de la metodología para fusionar clusters de AS2ORG, si bien el porcentaje de incidencia está vinculado a la

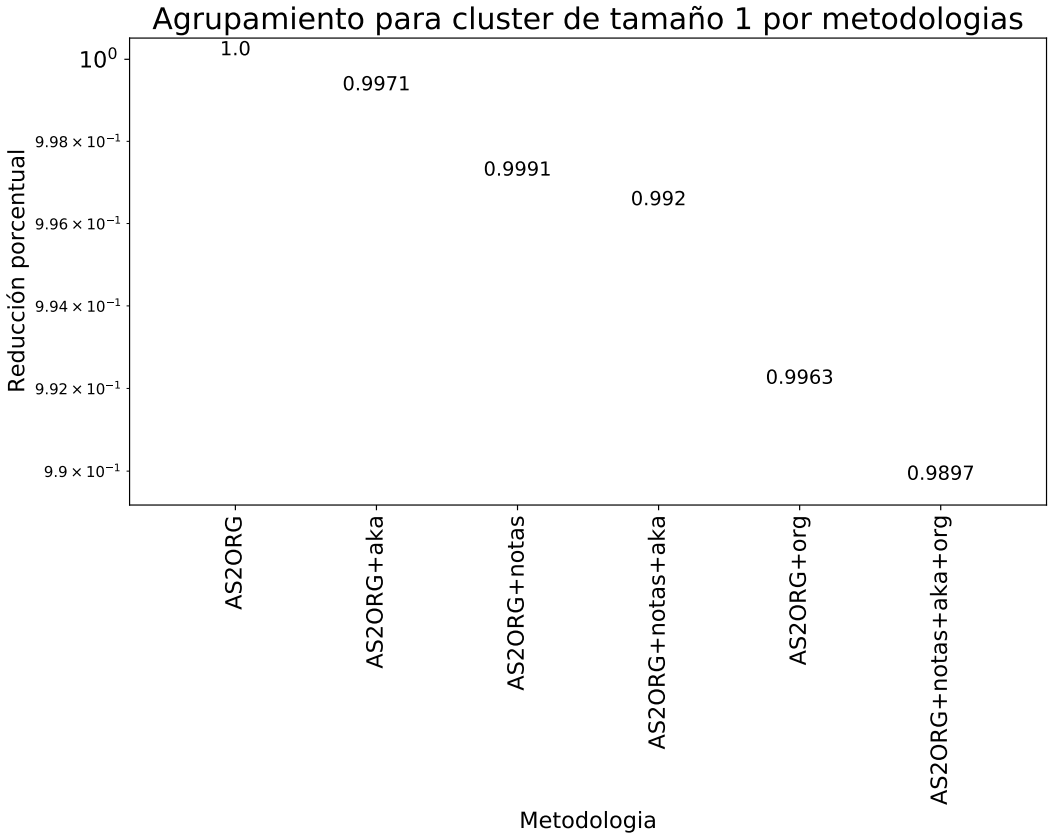


Figura 3.2: Reducción porcentual por metodologías para un cluster de tamaño 1. En el eje X se presenta la metodologia y en el eje Y la reduccion porcentual.

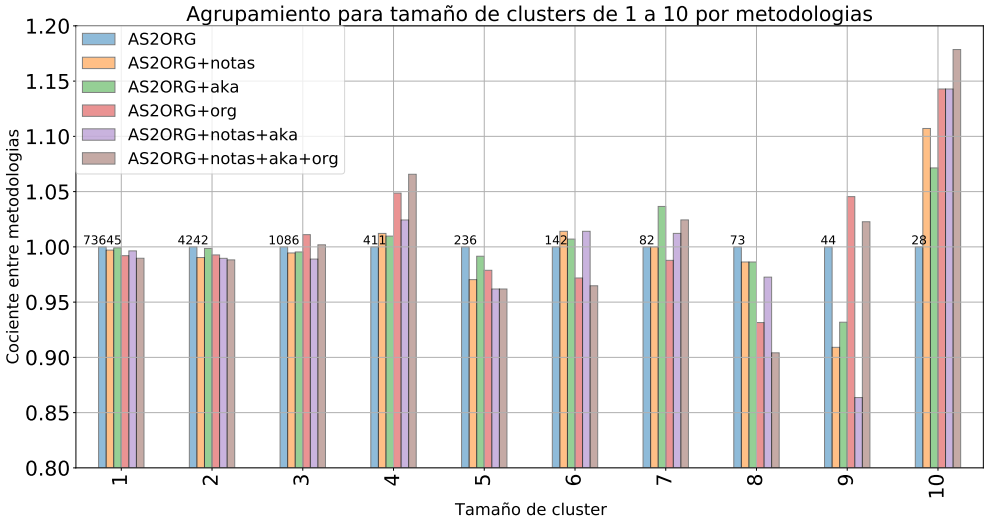


Figura 3.3: Agrupamiento para clusters de tamaño 1 a 10 para cada metodología propuesta. En el eje X se presenta el tamaño del cluster, mientras que en el eje Y el cociente entre metodologías tomando como denominador común a la metodologia AS2ORG.

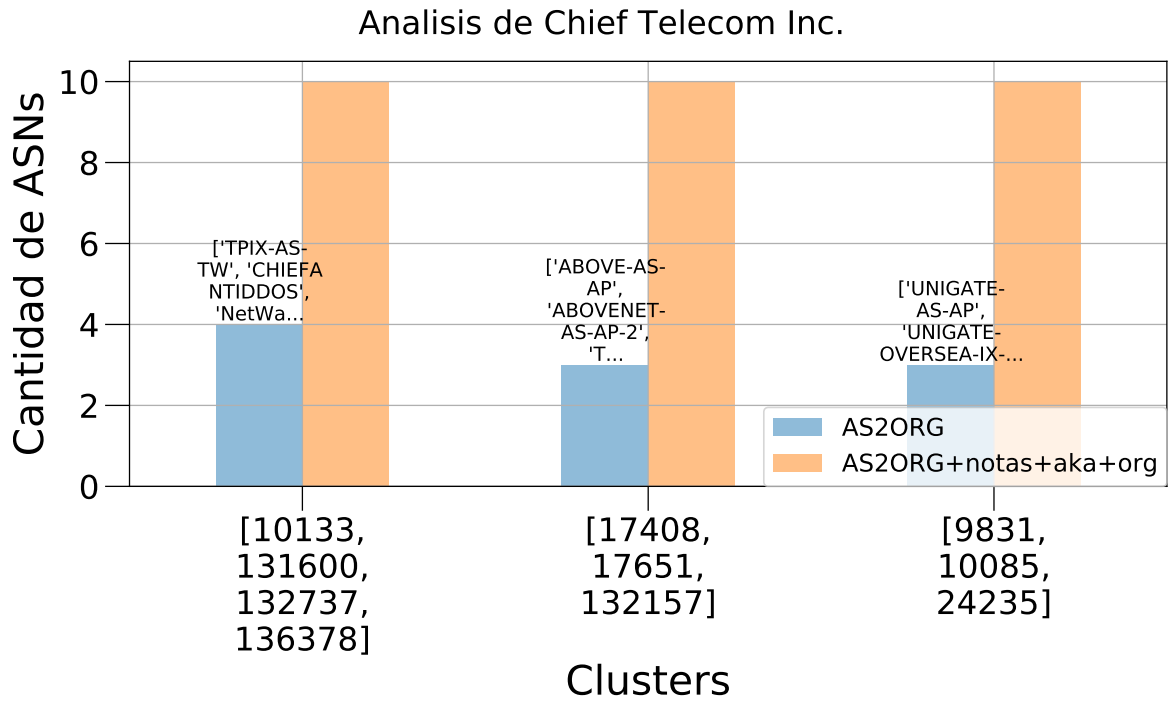


Figura 3.4: Ejemplo de nuevo cluster obtenido para Chief Telecom (AS17408) luego de agrupar utilizando la metodología `ASORG+notes+aka+org`. En azul se distinguen tres organizaciones que luego se convierten en una sola de tamaño 10.

cantidad de clusters de `<X>`, se observa un aporte en todos los tamaños. En el Apéndice A.6 se presenta un cuadro con los datos para un cluster de `<10>`.

Finalmente, retomamos los ejemplos presentados en la Sección 2.3.1, para ver el funcionamiento de la metodología en estos casos. La Figura 3.6 presenta el tamaño del cluster obtenido con la metodología PDB+AS2ORG para 16 Organizaciones, incluyendo grandes proveedores de tránsito, proveedores de contenidos, empresas de telefonía celular nacionales y de la región. En la figura se apilan en colores diferentes cada uno de los clusters de AS2Org que se unió por medio de PDB. Como vemos, en el caso del proveedor Level3+CenturyLink se alcanzaron los 136 ASes dentro del cluster. Esto es muy significativo, ya que Level3, mayor proveedor de tránsito global[1], fue recientemente adquirido por CenturyLink [11], y esto no es capturado por AS2Org. También en esta figura podemos observar que para el caso de GTT se han combinado un total de 18 clusters. En este caso también la metodología de PDB fue capaz de capturar la reciente fusion de GTT con KPN [22], que tampoco es capturada por AS2Org. Estos casos, como también el de Columbus Networks [8], demuestran que la metodología basada en PDB es capaz de reportar fusiones.

[1]Rank 1 según ASRANK [6]

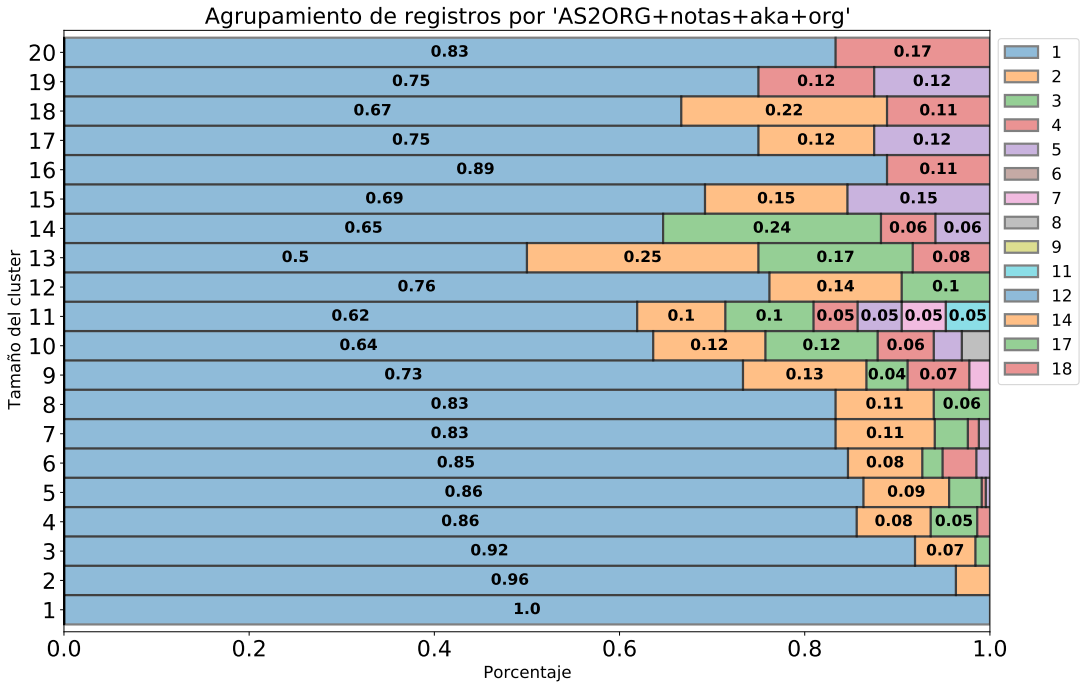


Figura 3.5: Número de veces que n organizaciones se unen para conformar clusters de tamaños de 1 a 20. A partir de la leyenda de cada color, se distingue la unión de n conglomerados, mientras que el número de la barra indica la distribución de veces que se unieron registros de AS2ORG para formar un conglomerado de tamaño `X`.

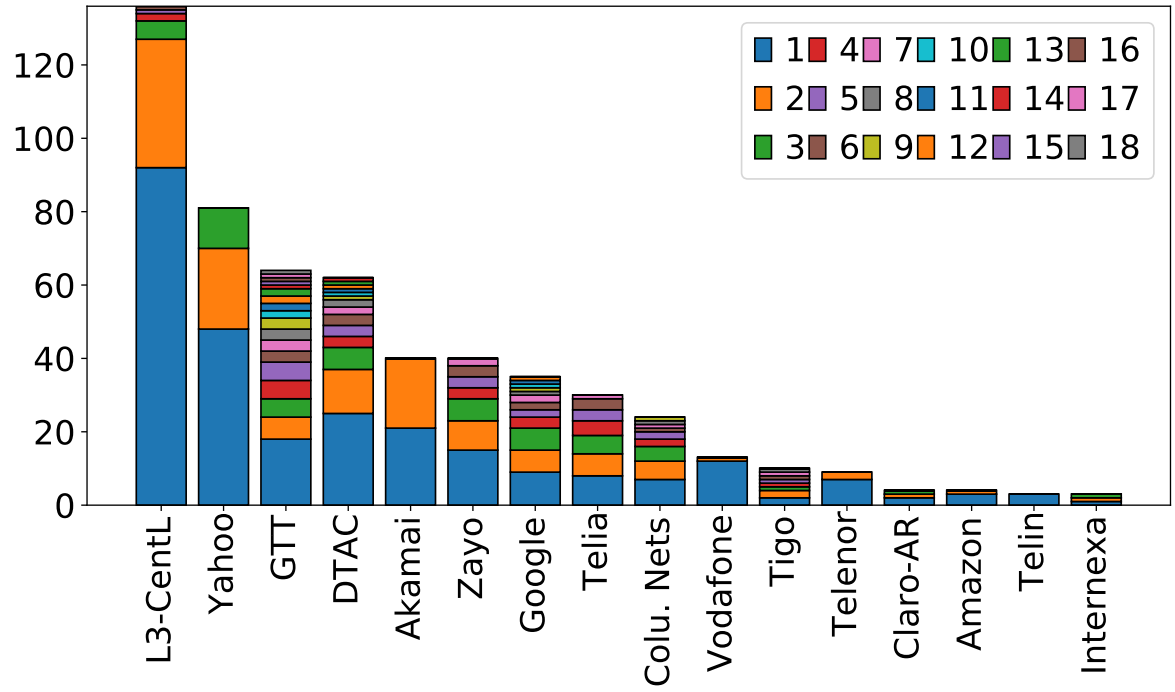


Figura 3.6: Contribución de la metodología propuesta sobre conglomerados de estructura conocida. En el eje X se ilustran 16 organizaciones, y para cada una se representa la unión de clusters obtenidos por AS2ORG que ahora conforman un conglomerado usando PDB. El eje Y representa en ASNs el tamaño del conglomerado obtenido.

3.3. Impacto de la nueva metodología en los grandes proveedores de tránsito.

A través de ASRANK [5], identificamos a los Sistemas Autónomos más preponderantes de Internet a fines de medir el impacto de la metodología en ellos. ASRANK es un ranking de Sistemas Autónomos creado por CAIDA que utiliza el *customer cone size* (ver Sec. 1.2.2) para posicionar a los ASNs por su número de clientes. En la Figura 3.7 comparamos el aporte de las metodologías `AS2ORG` y `AS2ORG+notes+aka+org` en cantidad de ASNs por medio del área cubierta para los mejores 10 posicionados del ranking. La Figura 3.8 ilustra la tendencia para los primeros 100 del ranking, donde en un 33 % de éste existe una contribucion de la metodología. Determinados ASes forman parten del mismo conglomerado como el rank 1 y rank 10, por lo que sus picos son alcanzados en la misma cantidad de ASNs.

La tendencia para los grandes proveedores de tránsito es distinta al común de las organizaciones de Internet. Esto se puede deducir de la CDF de la Figura 3.9, ya que desaparece la estructura atómica de las organizaciones vistas en la Figura 3.1. Si bien sigue siendo una función de cola larga, su transición es suave y la probabilidad de encon-

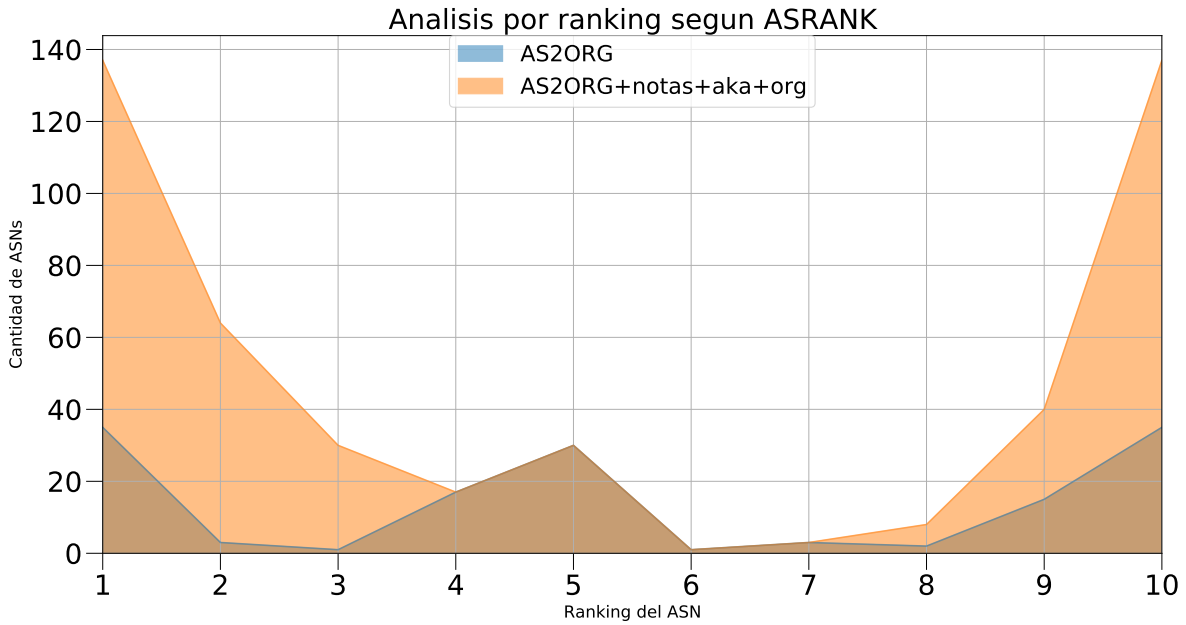


Figura 3.7: Impacto de la metodología sobre los primeros 10 proveedores de tránsito en Internet según *ASRANK*.

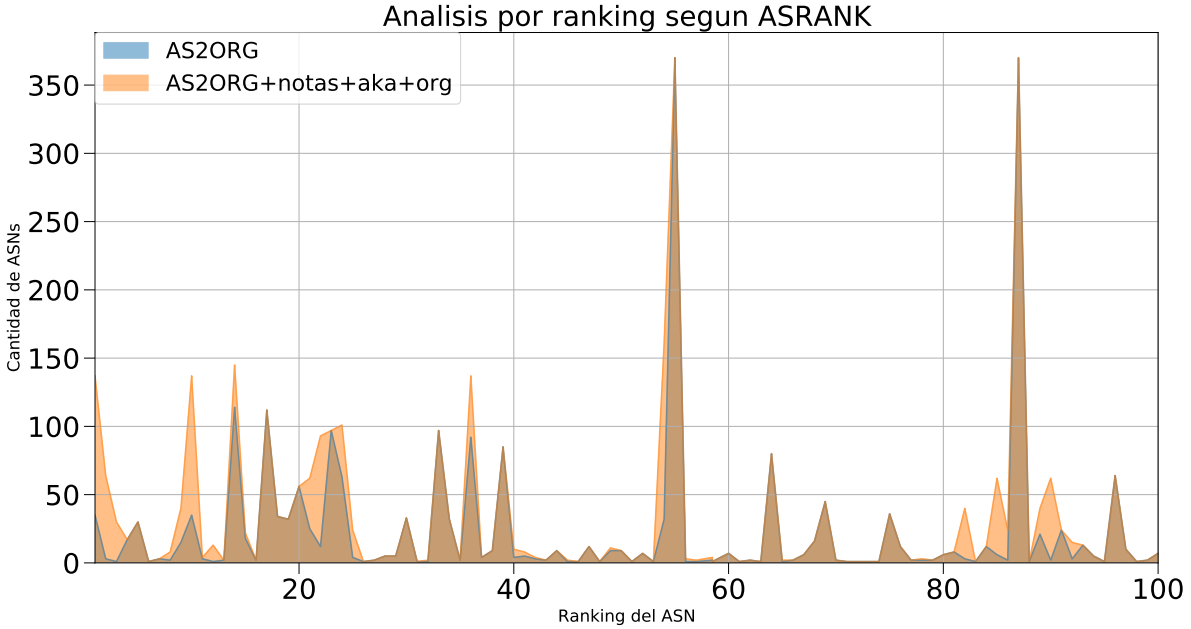


Figura 3.8: Impacto de la metodología sobre los primeros 100 proveedores de tránsito en Internet según *ASRANK*.

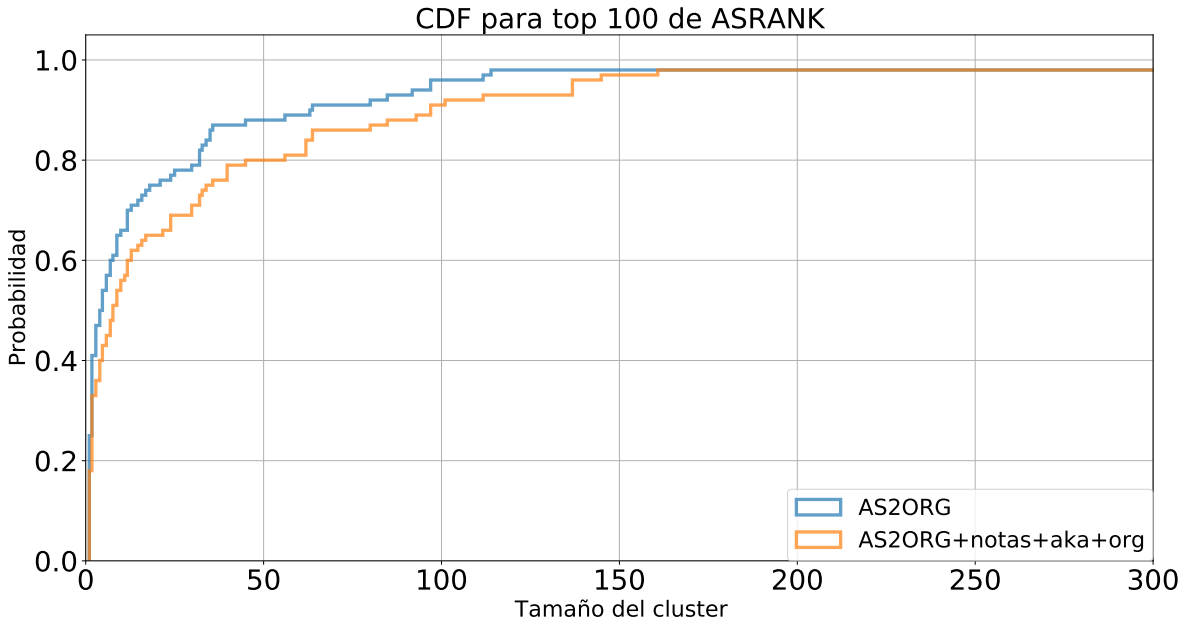


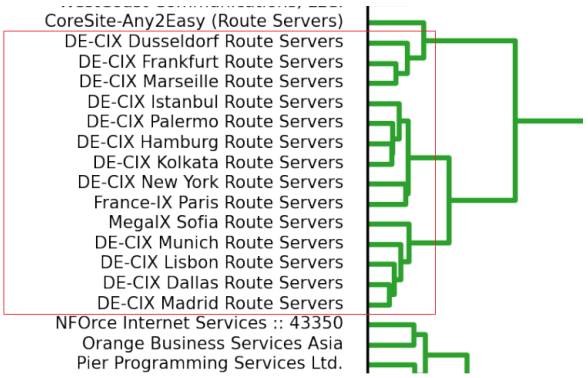
Figura 3.9: Función de distribución acumulada de los tamaños de la configuración inicial en comparación con la obtenida por la metodología propuesta, para los top 100 ASes de *ASRANK*.

trar clusters de tamaño mayor a 1 es inferior. También se puede observar, que la nueva metodología propuesta disminuye aún más la probabilidad de clusters entre `<1>` y `<150>`.

3.4. Exploración de trabajo con dendrogramas.

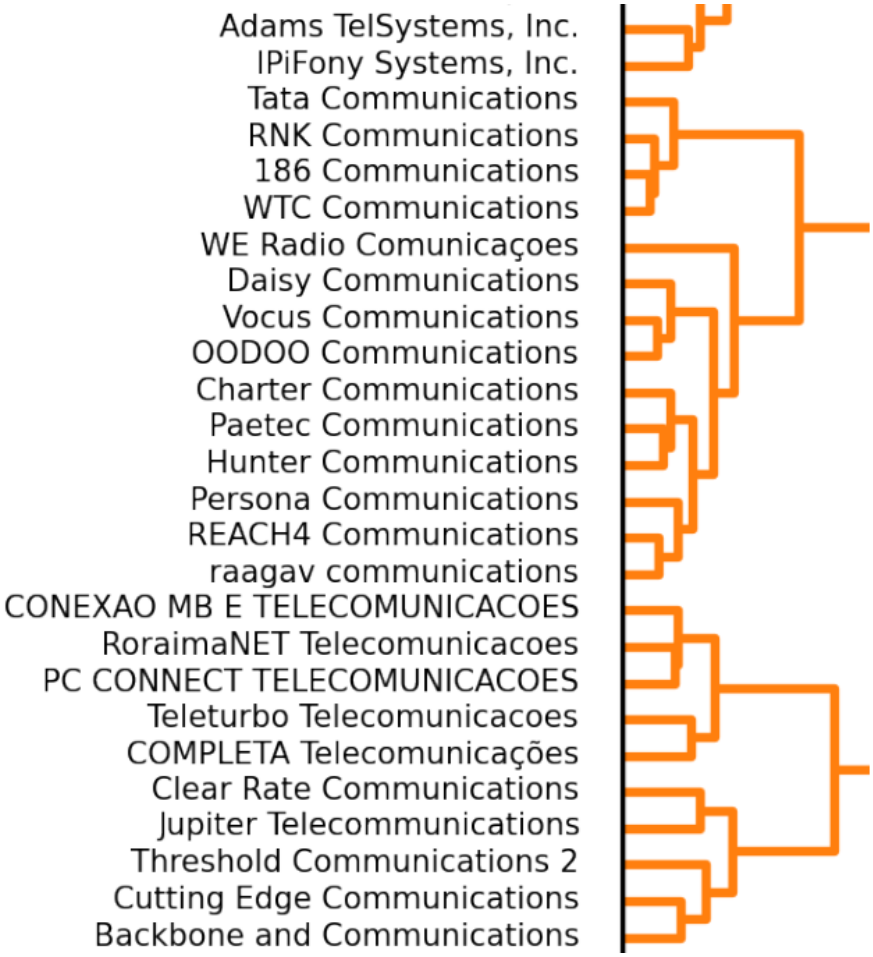
En esta sección realizamos un trabajo preliminar, utilizando dendrogramas a fines de contribuir con la metodología propuesta. El foco de esta exploración inicial está centrado en generar clusters tomando como distancia la cercanía entre organizaciones a través de los campos `name` y `aka`. A partir del conocimiento adquirido con la estructura de datos de PeeringDB y la semántica de sus campos, creemos que por medio de campos como `name` y `aka` se puede generar una aproximación que agrupe a los conglomerados. Las distancias utilizadas fueron las de Hamming [17] y Levenshtein [18], con el objetivo de crear una matriz cuadrada de *NxN* utilizando los campos seleccionados, para luego reagrupar aquellas filas o columnas por medio de un algoritmo de agrupamiento jerárquico aglomerativo. Un ejemplo de este tipo de algoritmo es el dendrograma, que se implementó sobre el subconjunto de 1488 Sistemas Autónomos utilizado para el campo `notas` (Sección 2.3.2) pero tomando como referencia las distancias del campo `name` para cada entrada. El Cuadro 3.2 ilustra una extracción de una matriz de 5x5 utilizando la distancia de Levenshtein, que mide el número mínimo de operaciones requeridas para

transformar una cadena de caracteres en otra. Notamos cómo Charter Communications (AS6543) y GTT Comunnications (AS4436) poseen una distancia menor entre ellos pese a ser conglomerados distintos. La Figura 3.10 muestra 3 casos de uso para la metodología, en 3.10a y 3.10b observamos dos casos de éxito para clasificar Decix y Limelight. Si bien Decix es capturada en la misma proporción que la metodología propuesta (Sección 2.4), el dendrograma también incluye organizaciones que no pertenecen al conglomerado como France-IX (AS57734) y Megaport (AS59899), esto se debe a que comparten los n-gramas *Route Servers*. El comportamiento se ve replicado en toda la exploración, siendo una de sus principales limitaciones. La Figura 3.10c demuestra un ejemplo de clasificación que incluye la palabra *Comunicaciones*, presente en muchas organizaciones que no se relacionan, siendo un término frecuente en el dominio de Internet. Luego utilizamos la distancia de Hamming, con la intención de estudiar el comportamiento de las organizaciones agrupadas bajo esta distancia. La tendencia resulta similar a la agrupada por Levenshtein, principalmente con desaciertos en n-gramas cortos como los ilustrados en la Figura 3.11a. También, presenta agrupamientos válidos para organizaciones como GTT (AS3257) y WTBTS (AS54235), ver Figuras 3.11b y 3.11c respectivamente. Un comportamiento interesante y que puede complementar a la metodología propuesta en esta Tesis, es el observado en la Figura 3.11d, si bien las organizaciones agrupadas no están relacionadas bajo los campos estudiados, podrían resultar ser parte del conglomerado de Telecom. Una parte de las adquisiones y fusiones de compañías únicamente se ve reflejado en campos como `aka` y `name`, sin el uso de contenido numérico. Este tipo de relaciones no se incluyen en la metodología desarrollada, pero pueden ser alcanzadas utilizando distancias entre palabras o caracteres.



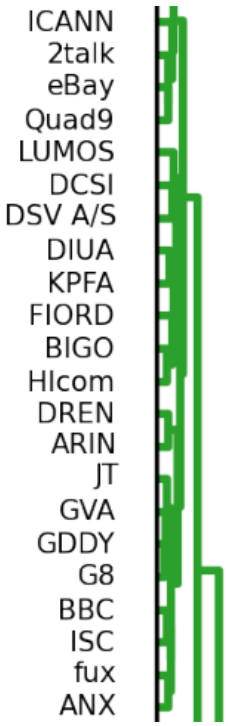
(a) Decix representado a través de dendrograma.

(b) Limelight representado a través de dendrograma

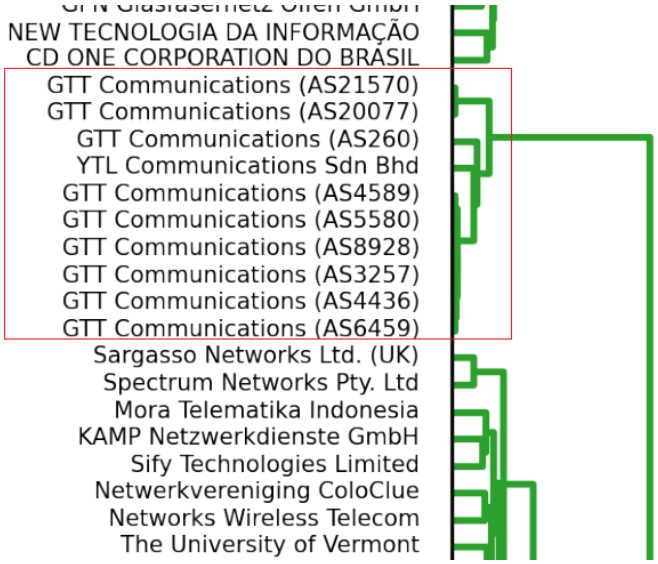


(c) Agrupamiento erróneo en el dendrograma tomando como referencia la palabra *Comunicaciones*.

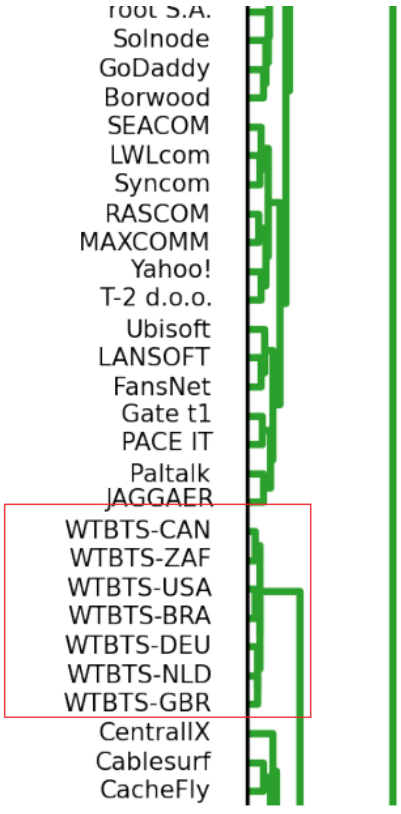
Figura 3.10: Ejemplos de agrupamientos logrados a partir de la distancia de Levenshtein para el campo `name` de PeeringDB.



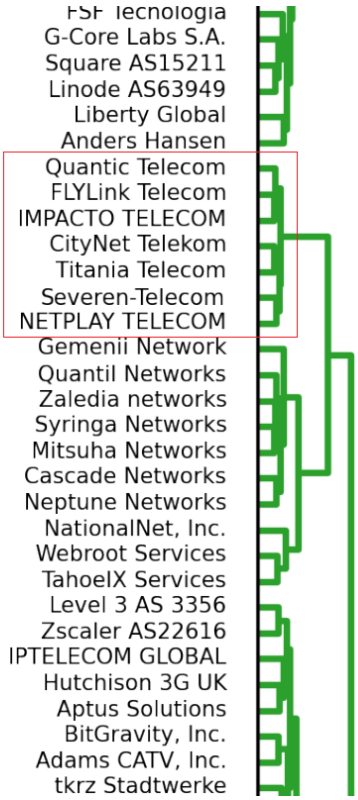
(a) Agrupamiento para n-grama de tamaño 1.



(b) GTT representado a través de dendrograma.



(c) WTBTS representado a través de dendrograma.



(d) Telecom representado a través de dendrograma.

Figura 3.11: Ejemplos de agrupamientos logrados a partir de la distancia de Hamming para el campo `name` de PeeringDB.

	GTT Communications (AS4436)	**Akamai Technologies**	**Limelight Networks Global**	**Charter Communications (7843)**	**Telia Carrier**	...
GTT Communications (AS4436)	0	23	23	10	22	...
Akamai Technologies	23	0	21	23	14	...
Limelight Networks Global	23	21	0	24	20	...
Charter Communications (7843)	10	23	24	0	24	...
Telia Carrier	22	14	20	24	0	...
...	...	...	...	...	...	...

Cuadro 3.2: Matriz de distancia de Levenshtein para campo `name` de PeeringDB.

Capítulo 4

Conclusiones

Luego del trabajo realizado, podemos concluir que PeeringDB es una fuente de datos significativa para la extracción e inferencia de siblings. Si bien cuenta con el 22.6 % de los ASNs delegados por los RIRs, su información es representativa de la red y en específico, de los operadores de mayor impacto de Internet. La metodología de extracción basada en `regexes` demuestra ser efectiva para obtener la información disponible, avalada por los resultados de validación. A través de la metodología `notes+aka+org` se extrajeron 19.255 subconjuntos y 18.179 agrupamientos, totalizando 20.438 siblings.

A partir de la metodología desarrollada y los nuevos agrupamientos generados en combinación con AS2ORG, se demuestra un aporte en la representación de conglomerados para la red de Sistemas Autónomos. Las nuevas representaciones son distribuidas en clusters de distintos tamaños, centrando su impacto en los grandes proveedores de tránsito etiquetados por ASRANK. La metodología `AS2ORG+notes+aka+org`, al ser la mas completa en información, propone los cambios más significativos.

Encontramos que el trabajo realizado puede ser complementado generando clusters a partir de la distancia de campos como `name` y `aka`. En ocasiones, los nombres de las organizaciones junto con su alias, hacen referencia a adquisiciones y fusiones entre compañías, sin referenciar Sistemas Autónomos en formato numérico.

Apéndice A

Apéndice

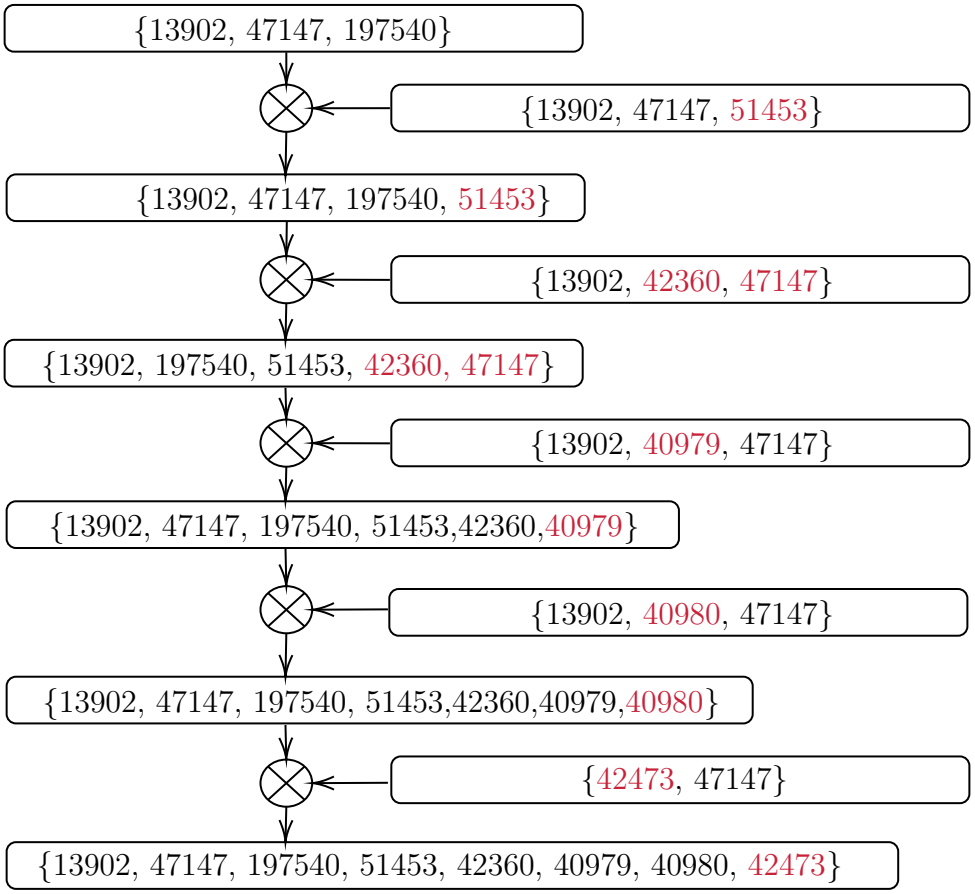


Figura A.1: Funcionamiento del agrupamiento para un ejemplo de extracciones filtradas del campo `notes`. A la izquierda se ilustra el agrupamiento parcial a medida que se incorporan nuevos subconjuntos de la derecha. Realizando la diferencia simétrica entre ambos subconjuntos se genera un listado final que contiene a todos los siblings inferidos para la organización.

Organización	**ASN-ASname**
Amazon.com	Amazon.com - 16509
	Amazon.com Tech Telecom -38895
Akamai Technologies	Akamai Direct Connect - 20189
	Akamai Prolexic DDoS Mitigation - 32787
	Akamai Technologies - 20940
Yahoo!	AOL - 1668
	Yahoo! - 10310
	Yahoo! ,Oath Global Network - 55517
	Yahoo! AS134706 - 134706
	Yahoo! AS17457 - 17457
	Yahoo! CHA - 394561
	Yahoo! UK - 204000
Telefonica Brasil S.A	TELEFÔNICA BRASIL - AS10429 - 10429
	TELEFÔNICA BRASIL - AS11419 - 11419
	TELEFÔNICA BRASIL - AS16885 - 16885
	TELEFÔNICA BRASIL - AS16911 - 16911
	TELEFÔNICA BRASIL - AS18881 - 18881
	TELEFÔNICA BRASIL - AS19182 - 19182
	TELEFÔNICA BRASIL - AS22092 - 22092
	TELEFÔNICA BRASIL - AS26599 - 26599
	TELEFÔNICA BRASIL - AS27699 - 27699

Cuadro A.1: Campo Organización.

ASN-ASname	**Notas**
AS3320-DTAC	AS3320 has no upstream - peering partners are encouraged to actively filter out all routes with AS3320 in the as-path received from any other peer or customer. Related networks: AS5483 Magyar Telekom AS5391 Croatian Telecom AS6855 Slovak Telecom AS12713 OTEGlobe AS9050 Telekom Romania AS8412 T-Mobile Austria AS13036 T-Mobile CZ AS50266 T-Mobile Thuis AS31615 T-Mobile Netherlands AS12912 T-Mobile Poland AS5588 GTS AS6821 Makedonski Telekom AS8585 Crnogorski Telekom
AS7713-TELIN	We manage international traffic of Group ASNs: 7713: Global Network 17974: Local Network ID 23693: Mobile Network ID 56308: Local Network SG 58731: Local Network TL
AS7303-TELECOM ARGENTINA	Telecom Argentina is the major broadband and mobile provider in Argentina, with more than 4.1 MM broadband subscribers, 4 MM fixed lines and 20 MM mobile lines. Other ASN under 7303 are 10481 and 10318.
AS2119-Telenor	Telenor in the Nordic: Norway and Sweden (Telenor Denmark/ASN9158 is connected to ASN2119). Fixed and Mobile customers - consumer and business.

Cuadro A.2: Campo Notas.

ASN-ASname	**También conocido como**
11664-CLARO ARGENTINA	Techtel LMDS Comunicaciones Interactivas S.A. (ASN 11664,ASN 14535 - AR-TLCI-LACNIC); ERTACH S.A. (ASN 17401 - AR-MASA5-LACNIC); TELMEX; CTI Compania de Telefonas del Interior S.A. (ASN-19037 - AR-CCTI1-LACNIC)
AS8075-Microsoft	8068 8069
AS262206-TigoMillicom	26617 Navega.com S.A. - 23243 Tigo Móvil Guatemala (Comcel Guatemala S.A.) - 27773 Millicom Cable El Salvador S.A. - 17079 Telemóvil El Salvador, S.A. - 52262 Telefónica Celular S.A. - 23383 Metrored S.A. - 20299 Newcom Limited - 262197

Cuadro A.3: Campo Tambíen conocido como.

ASN-ASname	**Campo**	**Valor**
AS15169-Google	También conocido como	Google, YouTube (for Google Fiber see AS16591 record)
	Notas	Peering Requests: https://isp.google.com/iwantpeering Peering Operational Issues: (...) Google manages the following ASNs: AS36040, AS43515, AS36561, AS19527, AS139070, AS139190, AS16550
	Organización	Google Cloud Indonesia - 139190 Google Cloud Korea - 139070 Google Corporate Network in APAC - 45566 Google LLC - 15169 Google LLC AS19527 - 19527 Google LLC AS36040 - 36040 Google LLC AS43515 - 43515 Google Private Cloud -16550
AS58453- China Mobile	Notas	Related ASN: AS9808 (CMNET) AS9231 (CMHK)
	Organización	China Mobile International - 58453 China Mobile International - Brazil - 268862 China Mobile International - NGN - 136750 China Mobile International - NII - 58807 China Mobile International - Oceania - 132389 China Mobile International - Russia - 209141 China Mobile International (Malaysia) Sdn.Bhd. - 139619 Guangdong Mobile Communication Co. Ltd. - 9808

Cuadro A.4: Campo Combinado

País	Cantidad de ASNs	Cobertura
BR	3876	45.8 %
US	2953	11.3 %
DE	861	30.9 %
GB	803	28.9 %
PL	626	24.2 %
RU	596	9.4 %
AU	517	20.0 %
FR	475	30.5 %
NL	453	31.4 %
CA	443	22.6 %
AR	441	38.5 %
ZA	376	61.8 %
IT	338	29.1 %
BD	277	26.1 %
CH	234	25.3 %
UA	227	10.3 %
HK	217	21.1 %
ES	204	18.0 %
IN	197	21.8 %
SE	185	22.2 %

Cuadro A.5: Ranking con los 20 países de más registros en PDB y su respectiva cobertura por los registrados en el RIR. La asignación de nacionalidades se tomó a partir de los archivos de delegación de los RIR.

Conglomerado de tamaño 10		
# Veces	Unión de n org.	Porc.
21	1	63.6 %
4	2	12.1 %
2	3	12.1 %
2	4	6 %
1	5	3 %
1	8	3 %

Cuadro A.6: Tabla explicativa de Fig. 3.5 para un conglomerado de tamaño 10. Se representan el numero de veces que n organizaciones se lograron unir para conformar un cluster de tamaño 10.

Bibliografía

[1] Mayo de 2021. URL: http://www.caida.org/.

[2] Bernhard Ager y col. "Anatomy of a large European IXP". En: *Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication*. 2012, págs. 163-174.

[3] *An empirical study of router response to large BGP routing table load — Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*. https://dl.acm.org/doi/abs/10.1145/637201.637233. (Accessed on 07/20/2021).

[4] *Aprobaron la fusión de Telmex y Claro - Diario 26*. https://www.diario26.com/119506--aprobaron-la-fusion-de-telmex-y-claro. (Accessed on 06/28/2021).

[5] *AS Rank: A ranking of the largest Autonomous Systems (AS) in the Internet*. https://asrank.caida.org/. (Accessed on 07/27/2021).

[6] *AS Rank: ASLevel 3 Parent, LLC (1)*. https://asrank.caida.org/orgs/589f9199b0. (Accessed on 08/01/2021).

[7] Timm Böttger, Félix Cuadrado y Steve Uhlig. "Looking for hypergiants in PeeringDB". En: *ACM SIGCOMM Computer Communication Review* 48 (sep. de 2018), págs. 13-19. DOI: 10.1145/3276799.3276801.

[8] *Cable & Wireless Communications — Liberty Global Completes Acquisition of Cable & Wireless Communications Plc*. https://www.cwc.com/live/news-and-media/press-releases/liberty-global-completes-acquisition-of-cable-wireless-communications-plc.html#:~:text=Liberty%20Global%20plc%20(%E2%80%9CLiberty%20Global,on%20an%20enterprise%20value%20basis.. (Accessed on 08/01/2021).

[9] Xue Cai y col. "Towards an AS-to-organization map". En: nov. de 2010, págs. 199-205. DOI: 10.1145/1879141.1879166.

[10] Matt Calder y col. "Mapping the expansion of Google's serving infrastructure". En: *Proceedings of the 2013 conference on Internet measurement conference*. 2013, págs. 313-326.

[11] *CenturyLink completes acquisition of Level 3 - Nov 1, 2017*. https://news.lumen.com/2017-11-01-CenturyLink-completes-acquisition-of-Level-3. (Accessed on 08/01/2021).

[12] Nikolaos Chatzis y col. "There is more to IXPs than meets the eye". En: *ACM SIGCOMM Computer Communication Review* 43.5 (2013), págs. 19-28.

[13] *CIDR report*. https://www.cidr-report.org/as2.0/. Accessed: 2021-07-25.

[14] *Configuring the BGP Maximum-Prefix Feature - Cisco*. https://www.cisco.com/c/en/us/support/docs/ip/border-gateway-protocol-bgp/25160-bgp-maximum-prefix.html. (Accessed on 07/20/2021).

[15] Amogh Dhamdhere y Constantine Dovrolis. "Twelve Years in the Evolution of the Internet Ecosystem". En: *Networking, IEEE/ACM Transactions on* 19 (nov. de 2011), págs. 1420-1433. DOI: 10.1109/TNET.2011.2119327.

[16] Xenofontas Dimitropoulos y col. "AS Relationships: Inference and Validation". En: *SIGCOMM Comput. Commun. Rev.* 37.1 (ene. de 2007), págs. 29-40. ISSN: 0146-4833. DOI: 10.1145/1198255.1198259. URL: https://doi.org/10.1145/1198255.1198259.

[17] *Distancia de Hamming - Wikipedia, la enciclopedia libre*. https://es.wikipedia.org/wiki/Distancia_de_Hamming. (Accessed on 07/23/2021).

[18] *Distancia de Levenshtein - Wikipedia, la enciclopedia libre*. https://es.wikipedia.org/wiki/Distancia_de_Levenshtein. (Accessed on 07/27/2021).

[19] Diego Dominguez. *Introducción a PeeringDB*. Accessed on 06/09/2021.

[20] *DrPeering White Paper - The Great Public vs Private Peering Debate*. https://drpeering.net/white-papers/Public-vs-Private-Peering-The-Great-Debate.html. (Accessed on 06/14/2021).

[21] Lixin Gao. "On Inferring Autonomous System Relationships in the Internet". En: *IEEE/ACM Trans. Netw.* 9.6 (dic. de 2001), págs. 733-745. ISSN: 1063-6692. DOI: 10.1109/90.974527. URL: https://doi.org/10.1109/90.974527.

[22] *GTT Completes Acquisition of KPN International*. https://www.gtt.net/se-en/media-centre/press-releases/gtt-completes-acquisition-of-kpn-international. (Accessed on 08/01/2021).

[23] Cheng Huang y col. "Measuring and evaluating large-scale CDNs". En: *ACM IMC*. Vol. 8. 2008, págs. 15-29.

[24] *Index of /datasets/as-organizations*. https://publicdata.caida.org/datasets/as-organizations/?C=N;O=D. (Accessed on 07/17/2021).

[25] *Index of /datasets/as-relationships/serial-1*. https://publicdata.caida.org/datasets/as-relationships/serial-1/. (Accessed on 07/12/2021).

[26] *Index of /datasets/peeringdb-v1/2010/12*. https://publicdata.caida.org/datasets/peeringdb-v1/2010/12/. (Accessed on 07/24/2021).

[27] *Index of /datasets/peeringdb-v2/2016/05*. https://publicdata.caida.org/datasets/peeringdb-v2/2016/05/. (Accessed on 07/24/2021).

[28] *Index of /datasets/peeringdb-v2/2020/10*. https://publicdata.caida.org/datasets/peeringdb-v2/2020/10/. (Accessed on 06/09/2021).

[29] *Internet Routing Registry (IRR)*. http://www.irr.net. Accessed: 2021-07-25.

[30] Keith W. Ross James F. Kurose. *Computer Networking: A Top-Down Approach*. Pearson, 2017. Cap. Cap 5.3.

[31] Suqi Liu y col. *Who is.com? Learning to Parse WHOIS Records*.

[32] Lixin Gao y J. Rexford. "Stable Internet routing without global coordination". En: *IEEE/ACM Transactions on Networking* 9.6 (2001), págs. 681-692. DOI: 10.1109/90.974523.

[33] Aemen Lodhi y col. "Using PeeringDB to Understand the Peering Ecosystem". En: *ACM SIGCOMM Computer Communication Review* 44 (abr. de 2014), págs. 20-27. DOI: 10.1145/2602204.2602208.

[34] Matthew Luckie y col. "AS Relationships, Customer Cones, and Validation". En: *Proceedings of the 2013 Conference on Internet Measurement Conference*. IMC '13. Barcelona, Spain: Association for Computing Machinery, 2013, págs. 243-256. ISBN: 9781450319539. DOI: 10.1145/2504730.2504735. URL: https://doi.org/10.1145/2504730.2504735.

[35] *PeeringDB*. https://www.peeringdb.com/. Accessed: 2020-11-10.

[36] *PeeringDB - CAIDA*. https://www.caida.org/catalog/datasets/peeringdb/. (Accessed on 07/05/2021).

[37] *PeeringDB API Documentation*. https://www.peeringdb.com/apidocs/#operation/create%20net. (Accessed on 06/24/2021).

[38] *Registrar Compliance Program*. https://www.icann.org/en/system/files/files/registrar-compliance-program-03nov16-en.pdf. (Accessed on 08/01/2021).

[39] *Requisitos previos para configurar el emparejamiento con Microsoft - Azure — Microsoft Docs*. https://docs.microsoft.com/es-es/azure/internet-peering/prerequisites. (Accessed on 06/09/2021).

[40] L. Daigle. *WHOIS Protocol Specification*. RFC 3912 (Draft Standard). RFC. Fremont, CA, USA: RFC Editor, sep. de 2004. DOI: 10.17487/RFC3912. URL: https://www.rfc-editor.org/rfc/rfc3912.txt.

[41] Y. Rekhter (Ed.), T. Li (Ed.) y S. Hares (Ed.) *A Border Gateway Protocol 4 (BGP-4)*. RFC 4271 (Draft Standard). RFC. Updated by RFCs 6286, 6608, 6793, 7606, 7607, 7705, 8212. Fremont, CA, USA: RFC Editor, ene. de 2006. DOI: 10.17487/RFC4271. URL: https://www.rfc-editor.org/rfc/rfc4271.txt.

[42] G. Huston y G. Michaelson. *Textual Representation of Autonomous System (AS) Numbers*. RFC 5396 (Proposed Standard). RFC. Fremont, CA, USA: RFC Editor, dic. de 2008. DOI: 10.17487/RFC5396. URL: https://www.rfc-editor.org/rfc/rfc5396.txt.

[43] *sklearn.feature_extraction.text.TfidfTransformer — scikit-learn 0.24.2 documentation*. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html. (Accessed on 07/04/2021).

Índice alfabético