Discovering interconnections between Uruguay and the world using popular internet traffic

Mateo Nogueira
mateo.nogueira@fing.edu.uy

Diego Kiedanski
dkiedanski@fing.edu.uy

Eduardo Grampín
grampin@fing.edu.uy

Facultad de Ingeniería,
Universidad de la República
Montevideo, Uruguay

Abstract

Understanding how Uruguay connects with the world has both practical and theoretical relevance. A lot of research has been done in this subject but none of them focused on Uruguay. In this investigation we intend to do analyze how Uruguay interconnects with the world using popular YouTube CDN traffic.

1 Introducción

Comprender cómo se interconecta Uruguay con el mundo a través de Internet tiene relevancia tanto práctica como teórica. El estudio de la topología de la red es una de las formas más intuitivas de comenzar dicho estudio. Este acercamiento, sin embargo, no contempla el tráfico sobre la red o el comportamiento de los usuarios.

Los servicios OTT (Over-The-Top) son uno de los mayores responsables del tráfico global, siendo video streaming 72% del tráfico de Internet en 2016 [1]. Es natural esperar que, al estudiar de dónde, cuándo y cómo proviene el contenido, estemos arrojando (mucha) luz sobre cómo nuestro país hace uso de Internet. Como ejemplo, se propone estudiar YouTube. Se tomó esta decisión debido a que el contenido es de acceso libre y bastante extenso.

In: Proceedings of the IV School of Systems and Networks (SSN 2018), Valdivia, Chile, October 29-31, 2018. Published at http://ceur-ws.org

El contenido de YouTube (y muchos OTT en general) es provisto desde CDNs (Content Delivery Networks), grandes redes distribuidas y opacas a usuarios externos. Esto hace que el problema de estudiar el origen y las rutas del tráfico sea mucho más complejo. Existen diversos trabajos que han intentado localizar los servidores y caches que proveen contenido a un determinado ISP, mayoritariamente combinando técnicas de análisis de tráfico con herramientas de diagnóstico (como ping y traceroute) [2][3][4].

La bibliografía existente ha explorado estas preguntas sobre todo para actores en Europa [2][5], pero poco o nada se ha hecho en América Latina y Uruguay. Un primer objetivo es por lo tanto reproducir estos resultados en nuestro país y ubicar geográficamente (en la medida de lo posible) los distintos servidores en los que se encuentra el contenido que consumimos, como cambian estos con el tiempo, que rutas utilizan para alcanzar nuestro país, etc. Lo que, es más, existen diversas mejoras posibles a dichos estudios que sería interesante poner en práctica.

Los estudios más recientes sobre CDN encontrados utilizan medidas pasivas, recolectadas desde un ISP [3][4]. Estos estudios, además, se basan en tráfico generado por personas en su uso cotidiano de estas plataformas, y, probablemente sin conocer el contenido en sí, debido al uso de TLS en la mayoría de los servicios. Además, el tráfico es analizado un tiempo después de generado. Nuestro trabajo utiliza tráfico generado por nosotros mismos, lo que nos permite tener un experimento más controlado, debido a que podemos decidir que parte del contenido de toda la plataforma nos interesa, y, nos da en un futuro, la posibilidad de volver a visitar el mismo contenido, para comparar si hubo cambios a nivel de tráfico. Se planea realizar una geolocalización de los servidores

de contenido lo más automatizada y rápido posible. Es decir, al encontrar un nuevo servidor de contenido desconocido, geolocalizarlo en el momento (o cuanto antes posible).

Aunque la metodología de trabajo está enfocada en utilizar YouTube, se puede generalizar a otros tipos de contenido. Es necesario, primero, conseguir algún servicio o plataforma que utilice contenido multimedia brindado por una CDN. Este paso esta por fuera del alcance del trabajo. Luego es necesario definir una forma de navegar por el contenido recolectando los servidores de los cuales es obtenido el contenido multimedia junto con otros datos específicos de ese contenido que, puedan resultar relevantes. Por ejemplo, en el caso de videos de YouTube la cantidad de visitas al video, fecha de carga e idioma del título consideramos que son ejemplos de esto. La forma de navegar por el contenido es dependiente sobre si se requiere realizar la recolección automáticamente o manualmente. Tomando lo primero como referencia, dado que, una recolección manual implicaría demasiado trabajo, se pueden utilizar en el caso de computadoras de escritorio herramientas de navegación automática como Selenium [6] junto con Browser-Mob Proxy[7] para capturar el tráfico (y asociarlo al contenido correspondiente). Finalmente, resta planificar un experimento para definir cuándo capturar los datos y la metodología a usar para analizarlos.

Una vez que se tiene la información de los servidores de origen, muchas preguntas surgen naturalmente: ¿Hay alguna correlación entre cercanía de los servidores y el contenido popular en Uruguay?, ¿Y entre los videos recomendados a usuarios?

2 Trabajo Previo

Investigaciones anteriores se han enfocado en descubrir el funcionamiento de YouTube [2][5]. Para reproducir un video, es necesario acceder a un frontend web para lo cual se obtiene una dirección IP consultando al servidor de DNS local. Luego, el frontend responde con un servidor de contenido inicial donde el video (que posee un identificador único) está alojado. Es posible que el servidor no contenga el video y el usuario sea redirigido hacia otro servidor. Pueden ocurrir múltiples redirecciones, y las mismas pueden ser de un servidor en un datacenter a otro en el mismo datacenter o a un datacenter diferente.

Los autores de [5], descubren namespaces de servidores de contenido junto con una jerarquía de redirecciones. 80% de las direcciones IP correspondían a Google/YouTube y las restantes a otros ISP como Comcast y Bell-Canada. Algunos hostnames de esas direcciones IP contenían códigos IATA de aeropuertos, lo que facilitaban la geolocalización de esos servidores. Además, investigan las posibles causas por las que el usuario es redireccionado.

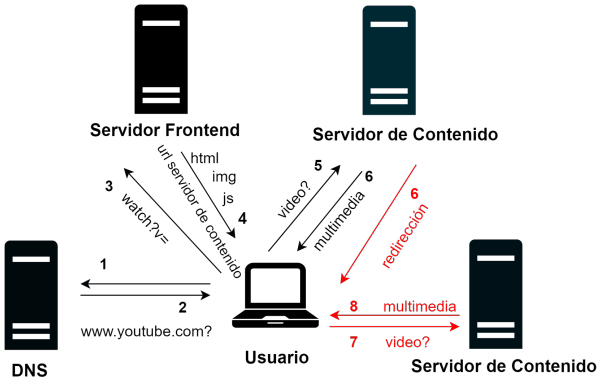

Figure 1: Pasos al mirar un video en YouTube.

3 Metodología

Para la investigación se intentó replicar los resultados obtenidos en los estudios previos. Sin embargo, se descubrió que los antiguos dominios utilizados por YouTube ya no son utilizados. En su lugar, se utilizan unos nuevos dominios de la forma ***.google-video.com**. Los nombres parecen estar agrupados por localización, pero no hay ninguna información que lo confirme. Las redirecciones a otro servidor cuando en el inicial no se encuentra el contenido siguen ocurriendo.

Se recolectaron aproximadamente 650 direcciones IP de servidores de donde se obtuvo contenido. Estos dominios fueron obtenidos utilizando un programa diseñado para navegar automáticamente por el sitio de YouTube, accediendo a un video inicial y accediendo a nuevos videos utilizando el video próximo indicado por la reproducción automática. Capturando paquetes, se guardaron los mensajes de DNS a los servidores ***.googlevideo.com**.

La mayoría de ellas están asignadas a Google y una minoría a nuestro ISP. Sin embargo, las direcciones IP asignadas a Google solo proveen contenido en caso de una redirección. Es decir, Google posee servidores de caché en nuestro ISP. Más aún, al hacer traceroutes hacia las direcciones de Google se descubre que el tráfico va de la red de nuestro ISP a la red de Google directamente sin otros ISP intermedios.

Se investigaron diversas técnicas de geolocalización y finalmente se decidió para cada servidor asignar como su localización la misma que un equipo cercano con bajo RTT. Suponiendo que el único retardo existente es el de propagación, se puede calcular la distancia entre el equipo y el objetivo. Siendo C la velocidad de la luz en el vacío, los bits viajan a aproximadamente $\frac{4}{9}C$ [8]. Por lo tanto, la distancia máxima posible entre el equipo y el objetivo es $\frac{4}{9}C * \frac{RTT}{2}$. Se eligió como bajo, un RTT menor o igual a $10ms$, lo que da un error de aproximadamente $670km$.

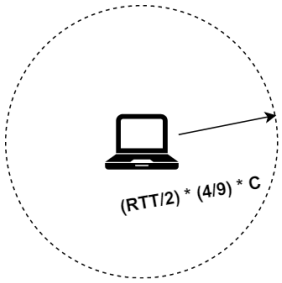

Figure 2: Distancia máxima posible a la que se encuentra un servidor.

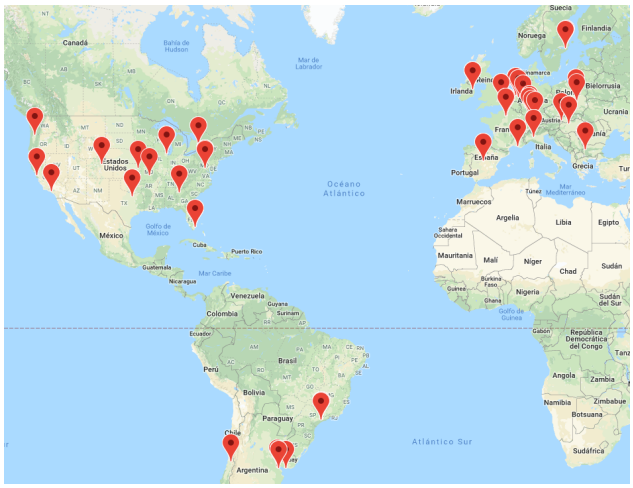

Figure 3: Servidores encontrados hasta el momento.

Los equipos utilizados para realizar las mediciones fueron probes del proyecto ATLAS de RIPE [9]. Los resultados arrojaron servidores en Montevideo, Buenos Aires, Santiago de Chile, San Pablo, Miami, Atlanta, Dallas, Kansas, Washington DC, Chicago, Toronto, Denver, Los Angeles, Palo Alto, Portland, Madrid, Marsella, Milán, Sofía, Budapest, Bratislava, Varsovia, Estocolmo, Paris, Amsterdam, Ede (Paises Bajos), Frankfurt, Hamburgo, Londres y Dublin.

Es probable que se siga intentando encontrar equipos con menor RTT hacia alguno de los servidores de forma de obtener un error más bajo. Por lo tanto, algunas ubicaciones podrían cambiar. Además, nuevos servidores pueden ser descubiertos al realizar el experimento final, luego de concurrida la investigación preliminar actual. Las localizaciones encontradas hasta ahora, pueden ser utilizadas para geolocalizar los nuevos servidores. Se utilizarían para concluir si, se encuentran en una localización conocida o, una distinta a las actuales, e intentar encontrar donde es.

4 Trabajo Futuro

Para el futuro, se planea continuar la investigación siguiendo los siguientes lineamientos.

- Integrar el sistema de geolocalización con el software de navegación de manera de geolocalizar los nuevos servidores al mismo tiempo en el que se descubren.
- Llevar a cabo un experimento de mayor porte, consumiendo videos desde varios equipos simultáneamente. Se estima que tenga una duración de aproximadamente una semana, un número no menor a cinco usuarios y a distintos horarios, por ejemplo, algunos por la mañana, otros en hora pico, etc. Antes de esto, es necesario terminar de definir que atributos de los videos podrían resultar útiles a la hora de analizar los resultados. Por ejemplo, cantidad de visitas.
- Aplicar técnicas de aprendizaje automático para analizar los datos recolectados, buscando relaciones entre el contenido y la localización del usuario, buscando resolver las dudas planteadas en la introducción.
- Investigar una posible relación entre los nombres de domino de los servidores de contenido y su ubicación geográfica.
- Obtener información sobre el funcionamiento del cacheo y redirecciones de YouTube luego de que un video no se encontró en un servidor.

References

[1] Cisco, "Cisco visual networking index: Forecast and methodology, 2016–2021," https://www.cisco.com/c/en/us/solutions/

collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html, 2017, [Online; accessed 17-June-2018].

[2] R. Torres, A. Finamore, J. R. Kim, M. Mellia, M. M. Munafo, and S. Rao, "Dissecting Video Server Selection Strategies in the YouTube CDN," in *2011 31st International Conference on Distributed Computing Systems*, Jun. 2011, pp. 248–257.

[3] P. Fiadino, M. Schiavone, and P. Casas, "Vivisecting WhatsApp in Cellular Networks: Servers, Flows, and Quality of Experience," in *7th Workshop on Traffic Monitoring and Analysis (TMA)*, ser. Traffic Monitoring and Analysis, M. Steiner, P. Barlet-Ros, and O. Bonaventure, Eds., vol. LNCS-9053, Barcelona, Spain, Apr. 2015, pp. 49–63. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01411179

[4] P. Fiadino, A. D'Alconzo, and P. Casas, "Characterizing web services provisioning via CDNs: The case of Facebook," in *2014 International Wireless Communications and Mobile Computing Conference (IWCMC)*, Aug. 2014, pp. 310–315.

[5] V. K. Adhikari, S. Jain, Y. Chen, and Z. L. Zhang, "Vivisecting YouTube: An active measurement study," in *2012 Proceedings IEEE INFOCOM*, Mar. 2012, pp. 2521–2525.

[6] Selenium Browser Automation, https://www.seleniumhq.org/, [Online; accessed 31-July-2018].

[7] Browsermob-Proxy, https://github.com/lightbody/browsermob-proxy, [Online; accessed 31-July-2018].

[8] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe, "Towards IP Geolocation Using Delay and Topology Measurements," in *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*, ser. IMC '06. New York, NY, USA: ACM, 2006, pp. 71–84. [Online]. Available: http://doi.acm.org/10.1145/1177080.1177090

[9] The RIPE Atlas measurement network, https://atlas.ripe.net/, [Online; accessed 31-July-2018].