Tesis de grado de Ingeniería en Informática

Dinámicas del discurso político en una red de Twitter en la Argentina

Tesista: **Tomás Mussi Reyero**

Director: **Dr. Ing. Mariano G. Beiró**

Co-director: **Dr. Ing. J. Ignacio Alvarez-Hamelin**

Buenos Aires, septiembre de 2019

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Las ciencias sociales estudian el comportamiento humano tanto a nivel individual como colectivo y las interacciones entre las personas, con el objetivo de entender al hombre como ser social y mejorar su vida.

A lo largo de la historia, las formas de vincularse entre las personas y los grupos sociales han evolucionado de la mano de las innovaciones tecnológicas. De hecho, hoy en día las tecnologías de la información son una herramienta valiosa para las ciencias sociales, al permitir cuantificar y validar las teorías que ellas elaboran. Comencemos tomando como ejemplo a la *World Wide Web*: una red de distribución de documentos originalmente concebida para compartir trabajos entre investigadores. Desde su creación, la Web ha tenido un enorme crecimiento y su potencial no estuvo limitado solamente al ámbito académico. Cualquier persona que tuviera acceso a una computadora y conexión a dicha red podía producir y consumir información, por lo que desde ese entonces han surgido infinidad de páginas web con distintas finalidades. Estas aplicaciones creadas comenzaron a almacenar distintos datos sobre las personas que las consultaban, como quién accedía a qué recurso, qué producto compraba una persona, quién intercambiaba correo con quién, entre tantos otros ejemplos. En otras palabras, las páginas web comenzaron a registrar el comportamiento digital humano.

A medida que la tecnología evolucionó, comenzaron a hacerse frecuentes dispositivos con mayor capacidad de procesamiento y almacenamiento, de tamaños que caben en la palma de una mano y a un precio menor que sus antecesores, y se logró que una proporción cada vez más grande de la población pudiera acceder a los mismos. El avance tecnológico también permitió el crecimiento y expansión de los servicios sobre Internet, aumentando aún más la cantidad de datos e información que muchas empresas y organizaciones almacenaban sobre sus usuarios.

Este volumen de información sin precedentes ha brindado la posibilidad de estudiar el comportamiento humano a gran escala y también dar un enfoque cuantitativo al análisis

sociológico. Por ello es que ha surgido una nueva área de investigación conocida como Ciencias Sociales Computacionales, que estudia el comportamiento humano a gran escala utilizando recursos tecnológicos para tal fin y brinda un enfoque cuantitativo.

En este contexto se hizo común el uso del término *Big Data* para hablar de la revolución brindada por el análisis estadístico de enormes cantidades de datos, aprovechando la capacidad de procesamiento de las computadoras, para detectar patrones en el comportamiento de las personas. Las primeras investigaciones en esta área surgieron en el estudio de la movilidad de las personas utilizando datos obtenidos de las señales de celulares. Estos datos han permitido por ejemplo estudiar la movilidad humana, analizar fenómenos de segregación en ciudades, optimizar sistemas de transporte y mejorar la eficiencia de la vida urbana, entre muchas otras cosas. Otra área que se ha beneficiado de la abundancia de datos sobre movilidad es el estudio de la propagación de epidemias: hoy en día muchos modelos de propagación incluyen información sobre desplazamientos a nivel local, nacional e internacional con el objetivo de mejorar sus predicciones respecto a la probabilidad de un contagio masivo, y de ayudar a su prevención.

Otras investigaciones han puesto el foco en el análisis de comportamiento de los mercados, estudiando su dinámica a partir de datos sobre operaciones de compra y venta de instrumentos financieros, o fenómenos como las burbujas. Otros trabajos, en cambio, analizan las motivaciones de las personas para formar parte de redes colaborativas en Internet, y los motivos por los cuáles son más o menos propensas a hacerlo.

Por último, cabe mencionar que en los últimos 15 años ha habido una explosión en el uso de redes sociales como Facebook, Twitter e Instagram. A partir de ellos, podemos destacar algunas líneas de investigación orientadas a detectar en las mismas noticias falsas (*fake news*) o cuentas *spam*, y a diseñar algoritmos para recomendar productos o amistades a los usuarios. También existen estudios en redes sociales con el objetivo de detectar comportamientos criminales y actividades ilícitas o terroristas, o desarrollar campañas de *marketing* orientadas a la difusión de un producto o servicio.

En concreto, el objetivo que nos planteamos en la presente tesis es analizar tendencias entre grupos sociales, específicamente en el contexto político de las elecciones presidenciales argentinas de 2015, utilizando Twitter como herramienta de obtención de datos. Propondremos una forma de calcular qué tan símil es un usuario a otro usuarios a partir de su discurso, y en particular de los *hashtags* que utiliza. Propondremos una metodología para extraer tópicos de discusión a partir de los *hashtags*, y analizaremos cómo evoluciona el discurso de las personas que siguen a determinados candidatos políticos, focalizándonos en distintos momentos del período electoral: las primarias, la primera vuelta y el balotaje. Por último, mostraremos cómo distintos temas de discusión son utilizados en forma heterogénea por los seguidores de distintos candidatos.

Este capítulo ha servido de introducción al lector sobre el tema y objetivo de la presente tesis. El siguiente capítulo hace una revisión de trabajos previos del área, en algunos de los cuales nos basaremos para nuestro análisis. En el capítulo 3 se introducen las herramientas de análisis utilizadas en nuestro estudio y presentamos algunos de los algoritmos aplicados. El capítulo 4 detalla la metodología utilizada, desde la obtención de temas de debate por parte de los usuarios, pasando por la comparación estática de los mismos, luego haciendo una evolución de la similitud entre ellos y finalmente un experimento de predicción de la preferencia de usuarios por un candidato. El capítulo 5 extrae conclusiones del trabajo realizado y algunas mejoras propuestas para futuro. Finalmente se incluye la bibliografía consultada y la presentación en una conferencia que surgió del desarrollo del presente trabajo.

Capítulo 2

Estado del arte

En este capítulo presentaremos una reseña de la literatura existente relativa al estudio de redes sociales con un enfoque de sistemas complejos, cubriendo fenómenos como la influencia y el contagio, la formación de opinión y la homofilia, con un foco particular sobre la discusión política. Comenzaremos con algunos trabajos pioneros en el área, y luego presentaremos en orden histórico los principales aportes.

2.1. Influencia y contagio

La influencia social es el fenómeno por el cual un individuo cambia su opinión o comportamiento a partir de la opinión o comportamiento de los individuos con los que se encuentra conectado. Estos fenómenos son conocidos desde la antigüedad, aunque los sociólogos recién han tenido herramientas para estudiarlos en el siglo XX realizando trabajos de campo como los de Paul Lazarsfeld y Leon Festinger, o trabajos teóricos como los de Granovetter (1977). El éxito en influenciar a una persona se logra a través del contagio de ideas. El estudio del contagio en redes sociales permite modelar y entender las dinámicas por las cuales los usuarios adoptan pensamientos tomados de otros. Específicamente dentro de Twitter, los usuarios pueden tomar una variedad de acciones para demostrar la adopción de una idea: señalar el *tweet* (dándole un *like*), compartirlo (*retweetearlo*), o bien responderlo (*hacer una mención*), entre otras.

El trabajo de Granovetter (1977) sentó las bases para el estudio de influencia y contagio de ideas, donde estableció que existen dos tipos de vínculos, los fuertes (*strong ties*) que son los que las personas mantienen con seres queridos, y los débiles (*weak ties*), que son los vínculos que se mantienen con conocidos, contactos esporádicos. Dentro de las conclusiones, se determinó la importancia de los *weak ties* en la integración de individuos dentro de una comunidad. Granovetter afirmó que el comportamiento social y la adopción de ideas y costumbres siguen la regla del contagio simple, en el cual una

única exposición ante el agente transmisor alcanza para que el objetivo adopte esta idea. En otras palabras, el contagio simple funciona como el contagio de una enfermedad en la que una exposición a un virus o bacteria es suficiente para que una persona se enferme. De esta investigación de comportamiento humano ha surgido un abanico de aplicaciones y posteriores ramas de investigación. Más precisamente en el campo de las ciencias sociales computacionales, Centola y Macy (2007) limitaron el alcance de la afirmación que realizó Granovetter, haciendo una diferencia entre las reglas de contagio simple y contagio complejo. Establecieron que la adopción de ideas y costumbres no es igual al contagio de una enfermedad infecciosa. Se determinó entonces como contagio complejo aquél que requiere de exposición de un individuo a múltiples fuentes y se remarcó la diferencia con múltiples exposiciones a una misma fuente, para lograr adquirir una idea o costumbre.

Grabowicz *et al.* (2012) han estudiado y realizado comprobaciones empíricas en una red de usuarios de Twitter sobre la importancia de los *weak ties* en la difusión de información a través de *retweets* y menciones en la red. Para ello, aplicaron un método de búsqueda de comunidades en la red de usuarios, detectando la existencia de nodos que pertenecen a más de una comunidad y que hacen de intermediarios (*weak ties*) entre comunidades. Dichos nodos son los que mayor audiencia logran al adquirir información de un grupo y *retweetearla* hacia otros grupos. Por otra parte, encontraron que las menciones entre usuarios se concentran dentro de una misma comunidad o entre usuarios que pertenecen a distintas comunidades que son cercanas. Los autores asociaron este tipo de intercambio como *strong ties*.

La adopción de ideas de acuerdo a exposiciones provenientes de múltiples fuentes también da lugar al análisis del fenómeno de propagación de información en cascadas. La difusión de información en cascadas es un relevante objeto de estudio en las redes complejas. El principal interés es determinar las condiciones bajo las cuales ante el mismo evento pueden producirse resultados totalmente distintos: lograr la reproducción total del estímulo en la red o no lograr repercusión alguna. Watts (2002) ha estudiado cómo un pequeño evento en un conjunto inicial de nodos en la red puede desencadenar una cascada global. Analizó la difusión de información en cascada en redes aleatorias, utilizando un modelo de umbral (*threshold*) en el cual se considera la adopción de una idea como una decisión binaria (adoptar o no adoptar) y un nodo se contagia y propaga información si una fracción de sus vecinos es superior al *threshold* necesario para contagiarse. Este modelo responde a la teoría de contagio complejo, en el cual el contagio depende a la exposición de los vecinos. Evaluando este modelo con métodos de percolación exploró bajo qué condiciones la topología y los umbrales individuales pueden influir la propagación.

Hodas y Lerman (2014) realizaron un estudio sobre dos redes sociales: Twitter y

Digg. En el análisis de los datos disponibles determinaron que la exposición repetida de información inicialmente incrementa la probabilidad de infección, pero eventualmente la exposición comienza a ser inhibitoria. Han surgido varias explicaciones por las cuales sucede este fenómeno, incluyendo reglas de contagio complejo así como también un modelo de *threshold* lineal como el desarrollado por Watts (2002). Luego, los autores determinaron un factor que denominaron visibilidad, el cual consiste en considerar que la información es visible por parte del usuario dependiendo las particularidades que tiene la interfaz de usuario de la aplicación en uso. Tomando en consideración la visibilidad en las reglas de contagio, este se vuelve muy simple, cualquier exposición ante una fuente de información incrementa la probabilidad de difusión, lo cual puede ser explicado por reglas de contagio simple. Finalmente los autores propusieron un modelo que toma en cuenta el factor de visibilidad y logran predecir con un error menor al 1.5 % la respuesta de usuarios dentro de una ventana de tiempo de treinta segundos.

Independientemente de la teoría desarrollada por Watts (2002), Nekovee *et al.* (2007) también investigaron las reglas de difusión de rumores en redes sociales complejas, considerando que dentro de la red existen tres estados para los usuarios: ignorantes, difusores (*spreaders*) y detractores (*stiflers*). El modelo presentado tiene como innovación que utilizaron Cadenas Interactivas de Markov (*Interactive Markov Chains, IMCs*), en las que desarrollaron ecuaciones deterministas para el proceso de divulgación en redes complejas. Evaluaron dichas ecuaciones en grafos aleatorios y también en grafos con distribuciones de grado de nodos que siguen una ley de potencias (*power law*), emulando las mismas características que tienen las redes sociales de gran escala actuales. Descubrieron que existe un *threshold* para los nodos por debajo de cuyo valor el rumor no se logra divulgar en la red.

Más allá de estos modelos teóricos, Galuba *et al.* (2010) propusieron una variante: en vez de investigar la profundidad y características de las cascadas, tomaron un enfoque empírico y buscaron predecir en una red de Twitter qué probabilidad existe de que un usuario propague o no información. Propusieron dos modelos para estudiar la difusión de *URLs*, los cuales tienen en cuenta tres factores: la influencia que un usuario tiene sobre otro, la viralidad propia de la *URL* y el tiempo que un usuario tarda en propagar información dado que un usuario que él sigue ha hecho una publicación. El primer modelo (*At-Least-One model*) supone que un nodo puede ser influenciado principalmente por un único usuario, mientras que el otro modelo (*Linear Threshold model*) utiliza una función de *threshold* lineal para cada nodo, más parecido a la idea propuesta por Watts (2002). En conclusión, los autores lograron entrenar un modelo que predice correctamente qué *URLs* serán *tweeteadas* por un usuario con menos de un 15 % de falsos positivos.

Un análisis parecido realizaron Bakshy *et al.* (2011) que también analizaron cómo

usuarios comunes comparten *URLs* con el objetivo de identificar usuarios que generen los eventos de cascada en la red. Entrenaron un clasificador con información de los usuarios iniciales (*seed users*) (tales como cantidad de seguidores, de amigos, de *tweets*) así como la cantidad de *retweets* que generaron en el pasado, considerando la cantidad de *retweets* como una medida de influencia. Los autores concluyen que los usuarios que logran mayor difusión son aquellos que tienen mayor cantidad de seguidores. También afirman que existen efectos por fuera de las observaciones en Twitter que afectan a los fenómenos de cascada que están estudiando. Finalmente proponen funciones de costo para realizar campañas de *marketing* más efectivas.

La influencia en redes sociales es difícil de cuantificar. Ya de por sí el término influencia es difuso y para poder realizar un estudio es necesario eliminar cualquier ambigüedad. En su tesis, Rosenman (2012) definió influencia como la habilidad para, a través del comportamiento de una persona en Twitter, promover una actividad y transferir información. El autor hizo un profundo análisis sobre la influencia que ejercen celebridades, analizando métricas basadas en *retweets* y evaluando si una persona no hace un *retweet* pero utiliza el mismo contenido, vocabulario o *hashtags* que utilizó dicha celebridad en un *tweet*. Algunos parámetros que tomó en consideración fueron la cantidad de seguidores de un usuario y la cantidad de *retweets*. Su conclusión fue que existen varias formas de lograr influir a otras personas, pero que cualquier métrica de las anteriormente mencionadas no se puede aislar en su efecto.

Cha *et al.* (2010) analizaron la correlación entre tres medidas de influencia: el *in-degree* de un usuario en la red, la cantidad de *retweets* que logró y por último, la cantidad de menciones que se hacen hacia un usuario determinado. Comparando estas tres medidas, determinaron que el *in-degree* de un usuario por sí solo no es determinante en la influencia que puede ejercer. No así los *retweets*, que son considerados como la métrica más representativa de difusión en general, mientras que particularmente para celebridades, las menciones que realizan sus seguidores logran introducir temas de conversación dentro de la red, por ende se consideran también como medida de influencia.

Weng *et al.* (2010) también tuvieron como objetivo detectar cuáles eran los usuarios más influyentes en Twitter. Tomaron como base el algoritmo que utilizaron Page *et al.* (1999) en Google para revolucionar Internet, *PageRank*, una forma de clasificar páginas web de acuerdo a su relevancia, transformando las páginas web en nodos de un grafo y las aristas conectando nodos representadas por los hipervínculos entre ellas. Consideraron a los usuarios como los nodos del grafo y las aristas como las relaciones de seguimiento entre usuarios. Además, utilizaron el concepto de homofilia para definir las probabilidades de saltos del *random walker* que se utiliza dentro del algoritmo de *PageRank*, dando como resultado su modelo propuesto *TwitterRank*. Este algoritmo dio mejores resultados en

comparación a otros, pero los autores concluyeron que hay todavía mucho espacio para mejorar la predicción de los usuarios influyentes.

Finalmente, mencionamos algunos trabajos que estudiaron bajo distintos puntos de vista la relación entre los procesos de contagio y la topología de la red: Kitsak *et al.* (2010), Chen *et al.* (2012) y Jalili y Perc (2017) han explorado qué medidas de centralidad son más informativas de la capacidad de difusión de un nodo, mientras que Mehmood *et al.* (2013) y Barbieri *et al.* (2013) han vinculado la estructura comunitaria con los procesos en cascada.

2.2. Homofilia

La homofilia es uno de los fenómenos que se observan en las interacciones que mantienen las personas dentro de las redes sociales, definida como la tendencia entre personas a relacionarse con pares que resulten similares a uno mismo (McPherson *et al.*, 2001).

El trabajo de Kossinets y Watts (2009) sentó las bases para el estudio de homofilia. Este fenómeno predomina en las redes sociales, y trabajos como el que realizaron Conover *et al.* (2011), estudiaron cómo la homofilia entre usuarios logra polarizar la opinión política.

Cardoso *et al.* (2017) propusieron un análisis que se utilizará en la presente tesis. Su trabajo consistió en capturar usuarios y los *tweets* que publicaron, con sus respectivos *hashtags*. Construyeron una red de *hashtags*, conectándolos de acuerdo a su coocurrencia en *tweets*, y aplicaron un algoritmo de detección de comunidades sobre ella, agrupando los *hashtags* en temas. Luego, cada usuario se representó con un vector de temas sobre los cuáles publicó *tweets* y se comparó a los usuarios utilizando la similitud coseno como medida. Uno de los resultados más destacables se obtuvo al comparar la similitud entre usuarios que se siguen contra la similitud entre usuarios tomados aleatoriamente: se observó que la similitud es más alta entre usuarios que se siguen. Esta misma comparación la aplicaron entre usuarios que se mencionaron y midieron la similitud de usuarios en la red de seguidores calculando por un lado las relaciones recíprocas contra las relaciones no recíprocas. En este caso concluyeron que hay mayor homofilia entre usuarios que tuvieron menciones recíprocas.

An y Weber (2016) investigaron si hay homofilia dentro de grupos demográficos en Twitter, buscando identificar *hashtags* que fueran utilizados únicamente por cierto grupo. El desafío principal fue determinar el grupo demográfico de un usuario utilizando su foto de perfil. Para ello utilizaron una herramienta para procesar imágenes y obtener rasgos demográficos característicos. En su trabajo concluyeron que algunos *hashtags* son utilizados indistintamente mientras que otros son preferencialmente utilizados por ciertos

grupos demográficos.

Colleoni *et al.* (2014) utilizaron el concepto de homofilia aplicada al ámbito político, para observar si existe polarización de ideas. Una de las formas en que esta polarización se manifiesta es a través del fenómeno de cámaras de eco (*echo chambers*), efecto por el cual individuos eligen intercambiar ideas políticas con otros que piensan de una manera similar, reforzando así sus pensamientos y cerrando la posibilidad del debate de ideas. El objetivo de los autores fue determinar si Twitter favorece el debate político, o si se incrementa la exposición a personas de pensamientos similares y para ello analizaron una red de Twitter en Estados Unidos durante el año 2009. Realizaron análisis de sentimiento sobre *tweets* con un algoritmo semi supervisado, clasificaron a los usuarios y midieron la homofilia entre cada uno de los grupos resultantes. Luego generaron una red aleatoria de características similares a la red de usuarios bajo estudio y calcularon la homofilia existente dentro de esa red como parámetro de comparación. En conclusión, si se considera el discurso político, las personas de pensamiento demócrata exhibieron una mayor homofilia que personas de pensamiento republicano. En cambio, si se consideran usuarios que siguen cuentas de Twitter oficiales de políticos, se observó que existe mayor homofilia entre republicanos que entre demócratas. Los autores sugirieron que este fenómeno se debe a la dualidad de Twitter como red social y como medio de comunicación, por ende se observó el fenómeno de cámaras de eco entre usuarios que se siguen entre ellos, así como también una plataforma de debate abierta.

Del Vicario *et al.* (2016) estudiaron la evolución temporal de dos comunidades en Facebook, una sobre ciencia y la otra sobre conspiraciones, y realizaron análisis de sentimiento sobre los comentarios que escribieron los miembros de ambos grupos. Dieron por hecho que en ambas existen efectos de *echo chambers*, produciendo que los usuarios seleccionen fuentes de información a las cuales adhieren y formen grupos donde refuerzan sus pensamientos, excluyendo cualquier idea disidente. Utilizaron tres modelos distintos para predecir la evolución temporal de las comunidades y determinaron que el tamaño de los grupos tiene al principio un crecimiento exponencial y luego un período de estabilización hasta alcanzar un umbral máximo por el cual no podrá crecer más. Luego los autores procedieron a realizar análisis de sentimiento sobre la información y comentarios que compartieron usuarios dentro de los grupos, concluyendo que las personas de participación más activa son las que tienen más negatividad dentro del contenido que escriben.

2.3. Análisis político

En esta sección se desarrollan distintos trabajos que utilizan los conceptos explicados de influencia y homofilia específicamente en un contexto político dentro de las redes sociales.

Como se mencionó previamente, Conover *et al.* (2011) estudiaron una red de Twitter durante un período de seis semanas previas a las elecciones del 2010 en Estados Unidos. Dada la naturaleza política bipartita que existe en Estados Unidos, los autores eligieron un algoritmo de detección de comunidades para dividir la red en dos. Aplicando este algoritmo a las redes de *retweets* y de menciones, encontraron que la primera estaba altamente polarizada entre la izquierda y la derecha políticas. Además lograron extraer *hashtags* representativos de cada grupo

Por otro lado, Romero *et al.* (2011) investigaron la difusión de *hashtags* en una red de Twitter, clasificando manualmente en temas los *hashtags* más populares. Dentro de esa clasificación, existen *hashtags* que pertenecen a tópicos controversiales, como debates políticos. Del estudio de contagio realizado para los diferentes temas, encontraron que particularmente los temas controversiales necesitan de una mayor exposición que el promedio para lograr finalmente la adopción de una idea. Otra particularidad detectada es que dentro de la comunidad de los usuarios que inician la difusión (o *early adopters*) existe una mayor cantidad de triángulos (relaciones entre tres personas) y predominan los *weak ties*.

Barberá (2015) hizo un extenso trabajo del cual tomó como base la suposición de que dentro de las redes sociales existe el fenómeno de homofilia y construyó un modelo probabilístico Bayesiano. Aplicó dicho modelo en distintos conjuntos de datos de Twitter para inferir orientación política de individuos a partir de las conexiones dentro de la red de usuarios. El autor concluyó que el método aplicado logró clasificar exitosamente a la mayoría de los actores políticos y personas comunes de acuerdo a su orientación política.

Halberstam y Knight (2016) también dieron por sentada la existencia de homofilia y analizaron en una red de Twitter las diferencias de exposición a información que hay entre grupos mayoritarios y minoritarios. Construyeron un modelo y evaluaron exitosamente sus hipótesis: miembros del grupo mayoritario están expuestos a una mayor cantidad de información y de una forma más rápida, también que la información alcanza a individuos de pensamientos similares más rápido que a individuos de ideología opuesta.

Finalmente, Klašnja *et al.* (2016) realizaron un análisis de los desafíos técnicos que implica la utilización de datos extraídos de redes sociales para conocer la opinión pública de un determinado grupo de personas. Los autores destacaron el beneficio que tienen aplicaciones como Twitter para conocer la opinión pública y el bajo costo frente a otros

métodos sociológicos como las encuestas. Propusieron metodologías para sortear los problemas, sesgos y otros errores que se pueden cometer al utilizar datos proporcionados de la API pública de Twitter tales como la falta de estructura de los *tweets*, el muestreo de la población en estudio, la utilización de *hashtags* para marcar temas de interés, la utilización de la opinión y datos personales de los usuarios, entre otros. Entre las sugerencias que mencionaron para enfrentar estos desafíos, se incluyen:

- selección automática de palabras clave para mejorar la identificación de opinión política.
- utilizar modelos multinivel para corregir las conocidas diferencias entre la muestra y la población de votantes, y así mejorar la representatividad de la muestra.
- filtrar mensajes provenientes de cuentas *spam*, que en algunos casos representan una gran parte del conjunto de datos.

Capítulo 3

Análisis de redes complejas

En este capítulo explicaremos que la red social Twitter es un grafo, describiendo definiciones, propiedades y características de los grafos. Luego explicaremos que la red social de Twitter así como diversos sistemas del mundo real son grafos que tienen una estructura similar y han sido enmarcados dentro del área de estudio de **redes complejas** por su interés, tamaño y las características que comparten. Hablaremos del desafío que representa particionar en grupos un grafo, particularmente en dichas redes. Mencionaremos distintas técnicas de descubrimiento de comunidades en grafos y sus limitaciones. También trataremos con un tipo particular de grafos, denominados bipartitos, la forma de representarlos, el desarrollo TF-IDF (*Term Frequency - Inverse Document Frequency*), un método para realizar consultas de términos en documentos y textos, ponderando términos que sólo aparecen en ciertos documentos y restándole importancia a aquellos que están presentes en todos los documentos. Finalmente, hablaremos de la visualización de la información, de la importancia que tiene esta disciplina en general y de su aplicación en el área de estudio de redes complejas para poder extraer conclusiones.

3.1. Teoría de grafos

Muchas situaciones que ocurren en el mundo real pueden ser descritas con un diagrama consistiendo de un conjunto de puntos con líneas uniendo ciertos pares de puntos. Por ejemplo, estos puntos pueden representar personas, con las líneas uniendo pares de amigos. De la misma manera se puede representar la red de usuarios de Twitter: los puntos también representan personas y hay una línea que une a dos personas si una sigue a la otra. Así surge el concepto de grafo, siendo una abstracción matemática para modelar este tipo de situaciones.

Seguiremos la notación y definiciones del libro Bondy *et al.* (1976).

Un *grafo* G es una terna ordenada $(V(G), E(G), \Psi_G)$ consistiendo de un conjunto no

vacío $V(G)$ de *vértices*, un conjunto $E(G)$, disjunto de $V(G)$, de *aristas*, y una *función de incidencia* Ψ_G que asocia a cada arista de G un par no ordenado de vértices (no necesariamente distintos) de G. Si e es una arista y u, v vértices tales que $\Psi_G(e) = uv$, entonces se dice que e une a u y v; los vértices u, v se llaman extremos de e.

De aquí en adelante, nos referiremos a la terna que define a un grafo $(V(G), E(G), \Psi_G)$ simplemente como G, al conjunto de vértices $V(G)$ y al conjunto de aristas $E(G)$.

Algunas definiciones:

- Una arista se llama *lazo* si une un vértice consigo mismo.
- Dos vértices se llaman *adyacentes* si existe una arista entre ellos.
- Una arista *incide* sobre un vértice si el vértice es uno de sus extremos.
- Un grafo es *finito* si ambos conjuntos de vértices y aristas son finitos.

Grafo simple. Un grafo es *simple* si no tiene lazos y ningún par de aristas une los mismos pares de vértices.

Orden y tamaño del grafo. La cantidad de vértices y la cantidad de aristas del grafo están dadas por el tamaño del conjunto *V(G)* y *E(G)* respectivamente. Se asocia $n = |V(G)|$ denominado también *orden* y $m = |E(G)|$ denominado también *tamaño*.

Grafo completo. Un grafo simple es completo si cada par de vértices distintos está unido por una arista. Un corolario es que en un grafo completo, se cumple $m = \frac{n(n-1)}{2}$

Matriz de adyacencia de un grafo. Un grafo G tiene asociada una *matriz de adyacencia*, la cual es una matriz de tamaño $n \times n$ siendo n el orden del grafo. El elemento $a_{i,j}$ de la matriz es la cantidad de aristas que unen a los vértices v_i y v_j.

Grafo bipartito. Un grafo bipartito es aquel cuyo conjunto de vértices $V(G)$ se puede particionar en dos subconjuntos X e Y, de tal forma que cada arista tiene un extremo en X y el otro extremo en Y. La partición (X, Y) se llama *bipartición* del grafo. Un *grafo bipartito completo* es un grafo bipartito simple con la partición (X, Y) en la cual cada vértice de X está unido a cada vértice de Y; si $|X| = m$ y $|Y| = n$, tal grafo se lo denota como $K_{m,n}$.

Partición de un conjunto. Una partición de un conjunto $V(G)$ es una familia de subconjuntos no vacíos de $V(G)$, disjuntos dos a dos, cuya unión es $V(G)$:

- $i \in I, V_i(G) \subseteq V(G), V_i(G) \neq \emptyset$
- $\forall i, j \in I, i \neq j, V_i(G) \cap V_j(G) = \emptyset$
- $\bigcup_{i \in I} V_i(G) = V(G)$

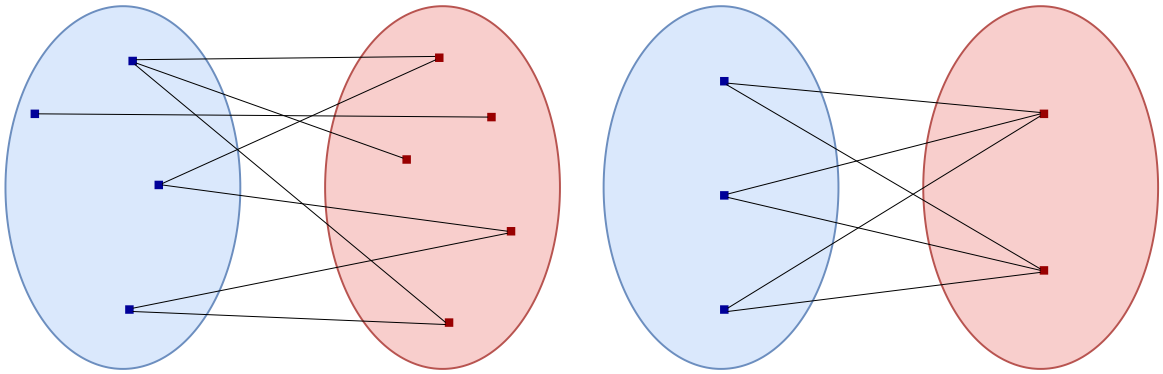


Figura 3.1: Grafo bipartito (izquierda). Grafo bipartito completo $K_{3,2}$(derecha)

Grado del vértice. Se define como *grado* $d_G(v)$ de un vértice v en G al número de aristas de G que inciden en v, cada lazo contando como dos aristas. Se denota $\delta(G)$ y $\Delta(G)$ al mínimo y máximo grado de vértices en G respectivamente.

Suma de los grados de un grafo. La sumatoria de los grados de los vértices de un grafo G es igual al doble de la cantidad de aristas:

$$\sum_{i=1}^{n} d_G(v_i) = 2 \cdot m$$

Grafo dirigido. Un *grafo dirigido* D es una terna ordenada $(V(D), E(D), \Psi_D)$ consistiendo de un conjunto no vacío $V(D)$ de *vértices*, un conjunto $E(D)$, disjunto de $V(D)$, de *arcos*, y una *función de incidencia* Ψ_D que asocia a cada arco de D un par **ordenado** de vértices (no necesariamente distintos) de D. Si e es una arco y u, v vértices tales que $\Psi_D(e) = uv$, entonces se dice que e une a u y v; los vértices u, v se llaman extremos de e. Si a es un arco que une desde u hacia v, v se denomina *cabeza* y u se denomina *cola* del arco.

La definición de grado se extiende, distinguiendo el *grado entrante* y el *grado saliente* del vértice siendo la cantidad de aristas que inciden y la cantidad de aristas que salen desde el vértice respectivamente. Utilizaremos la notación d_G^- para grado entrante y d_G^+ para el grado saliente de un vértice.

$$d_G^-(u) = |\{e = (x, y) : u = x\}|$$

$$d_G^+(v) = |\{e = (x, y) : v = y\}|$$

$$\sum_{i=1}^{n} d_G^-(v_i) = \sum_{i=1}^{n} d_G^+(v_i) = m$$

Camino. Un *camino* en G es una secuencia finita no vacía $W = v_0 e_1 v_1 e_2 v_2 ... e_k v_k$

cuyos términos son alternativamente vértices y aristas, de tal forma que para $1 \leq i \leq k$, los extremos de e_i son v_{i-1} y v_i. Se dice que W es un camino desde v_0 hasta v_k, o un (v_0, v_k)-camino.

En un grafo simple, un camino $v_0e_1v_1e_2v_2...e_kv_k$ está determinado por la secuencia $v_0v_1v_2...v_k$ de sus vértices, por lo tanto un camino en un grafo simple se puede especificar por la secuencia de vértices.

Dos vértices u y v en G están *conectados* si existe un (u, v)-camino en G. Si para cada par de vértices u y v en $V(G)$ existe un (u, v)-camino, se dice que el grafo es *conexo*. De lo contrario es *no conexo*.

Grafo pesado. Dado un grafo G, a cada arista e en $E(G)$ se le asocia un número real $w(e)$ llamado *peso*. Entonces G, en conjunto con estos pesos para cada una de sus aristas, se define como *grafo pesado*.

Problema del camino mínimo. Dado un grafo pesado G, el problema es encontrar un (u, v)-camino que minimice la suma de las aristas de dicho camino. Existe un algoritmo para resolver este problema llamado algoritmo de Dijkstra.

Subgrafo. Un grafo H es un *subgrafo* de G (denotado $H \subseteq G$) si $V(H) \subseteq V(G)$, $E(H) \subseteq E(G)$, y Ψ_H es la restricción de Ψ_G a $E(H)$.

Subgrafo inducido. Supongamos que $V'(G)$ es un subconjunto no vacío de $V(G)$. El subgrafo de G tal que su conjunto de vértices es $V'(G)$ y cuyo conjunto de aristas es el conjunto de aquellas aristas de G que tienen ambos extremos en $V'(G)$ se define como *subgrafo inducido* por $V'(G)$.

Clique. Un *clique* en un grafo no dirigido G es un conjunto de vértices $C \subseteq V(G)$ tal que todo par de vértices distintos en C son adyacentes. En otras palabras, se puede decir que un existe un clique C en un grafo G cuando el subgrafo inducido por $C \subseteq V(G)$ es un grafo completo.

Descomposición en *k*-núcleos. Sea $G = (V, E)$ un grafo no dirigido, y $H = (C, E(G)|C)$ el subgrafo inducido por $C \subset V(G)$ es un *k-núcleo* si y sólo si $\forall v \in C$: $d_H(v) \geq k$ y además H es el subgrafo máximo que cumple esta propiedad.

Descomposición en *k*-densos (*k-denses*). Se define $D_k(G)$ k-denso de un grafo G, para $k \geq 2$:

$$E(D_k(G)) = S \iff \{\forall e \in S : m_{G-\overline{S}}(e) \geq k-2\} \wedge \text{S es maximal con esta propiedad}$$

$$V(D_k(G)) = \{u \in V(G)/\exists v \in V(G) : u, v \in E(D_k(G))\}$$

3.2. Redes complejas

Ya establecimos que una red social como Twitter se puede representar con un grafo, de la misma forma que existen otros sistemas de diversas disciplinas que también se estructuran como elementos conectados de formas específicas: desde redes genéticas de proteínas, pasando por la red neuronal de una lombriz (nematodo Caenorhabditis elegans), hasta redes de transporte ya sea de energía eléctrica, de información, la red de computadoras conocida comúnmente como Internet, la red de páginas web o *World Wide Web*, entre tantas otras.

Estas redes representan sistemas muy diversos y sin embargo comparten ciertas características específicas, por lo que se las han denominado redes complejas. Estas son un área de estudio relativamente reciente dado que gracias a la capacidad de procesamiento y recopilación de datos que brindaron las computadoras se hizo posible realizar estudios en mayor detalle sobre ellas (Barabási y Albert, 1999). Mencionaremos algunas de sus características topológicas singulares que no están presentes en redes aleatorias.

3.2.1. Ley de potencia (*Power law*)

Dado un grafo $G = (V(G), E(G))$, se estudia la distribución de los grados de los vértices que tiene el grafo, es decir, se quiere observar cuantos vértices tienen grado $d(v) = 1$, $d(v) = 2$, ..., $d(v) = k$. La particularidad que tienen las redes complejas es que la función de probabilidad $P(k)$ que define la distribución de grados de los vértices del grafo es una ley de potencias:

$$P(k) \sim k^{-\gamma}$$

Esta función de probabilidad nos indica que existen muchos vértices que están poco conectados, por ejemplo personas comunes que son seguidos por sus amigos, mientras que existen unos pocos vértices, por ejemplo celebridades, que son seguidos por miles o incluso millones de personas. En el estudio realizado por Barabási y Albert (1999) indicaron que la distribución de la conexión de los vértices en forma de ley de potencias es posible gracias a que la red está en constante crecimiento (siempre existen nuevos usuarios en Twitter que crean una cuenta) y que esos nuevos vértices no eligen seguir a otras personas al azar sino que tienen cierta preferencia y/o sesgo a seguir a personas muy conectadas como por ejemplo celebridades. Este fenómeno de preferencia de los nuevos vértices a relacionarse con nodos ya conocidos se conoce como *preferential attachment*, y como consecuencia se obtiene que la probabilidad de que un vértice tenga más conexiones es directamente proporcional al grado del vértice (los ricos se hacen más ricos).

3.2.2. *Small world*

Watts y Strogatz (1998) denominaron **mundo pequeño** (*small world*) a esta propiedad que poseen las redes complejas, también conocido como *los seis grados de separación*. Este fenómeno se caracteriza por un elevado coeficiente de *clustering* y porque la distancia promedio que existe entre dos vértices cualesquiera de la red es bajo respecto a un grafo generado de forma aleatoria.

Los autores definieron un método para reconectar con cierta probabilidad p cada uno de los ejes de un grafo con una topología de anillo. Descubrieron que en el rango de valores que toma $p \in [0, 1]$, hay una transición donde la longitud promedio de camino entre dos vértices cualesquiera de la red decae rápidamente mientras que el coeficiente de *clustering* de la red se mantiene casi constante. Las redes que mantienen un camino promedio bajo entre todos sus vértices y un elevado coeficiente de *clustering* respecto a la red original las denominaron como redes *small world* y verificaron que las redes como la red neuronal de una lombriz, la red de transporte de energía eléctrica de Estados Unidos y la red de actores que trabajaron juntos en una película comparten este fenómeno.

3.2.3. *Assortative mixing*

Assortative mixing es una propiedad que tienen los tipos de conexiones entre usuarios, es decir, las aristas del grafo de la red social. En este tipo de redes, los vértices de mayor grado suelen conectarse con otros vértices de grado alto. El otro comportamiento observado es del tipo *disassortative mixing*, en el cual vértices bien conectados se asocian a vértices poco conectados. Internet y la Web son ejemplos de *disassortative mixing*.

3.2.4. Modularidad

Una de las particularidades de las redes complejas es la estructura comunitaria que presentan: existen vértices muy interconectados entre ellos, pero a su vez dichos vértices se encuentran conectados de forma dispersa con otros vértices de la red. Dicho conjunto altamente conectado se lo denomina comunidad.

Newman y Girvan (2004) definieron modularidad como una medida de la calidad de una partición del conjunto de vértices (ver sección 3.1) en comunidades de la red en cuestión. Los autores proponen construir una matriz cuadrada M de k filas por k columnas, siendo k la cantidad de comunidades en las cuales se particiona la red. Cada elemento e_{ij} se define como la cantidad de aristas que conectan vértices en la comunidad

c_i con vértices en la comunidad c_j :

$$e_{ij} = \frac{|\{(u,v) \in E(G) : u \in c_i, v \in c_j\}|}{|E(G)|}$$

La traza de esta matriz (suma de la diagonal principal) $Tr(M) = \sum_{i=1}^{k} e_{ii}$ es un indicador de como está interconectada cada una de las comunidades. Mientras mayor sea el valor, mayor es la cantidad de interconexiones dentro de las comunidades. Esta medida no impone restricción alguna sobre la cantidad de vértices dentro de cada partición. Daría lo mismo tener el conjunto entero de vértices en una única comunidad y el resto de las comunidades vacías y ello daría una modularidad de valor 1.

Por eso se obtiene la suma por filas (o columnas) de cada comunidad $a_i = \sum_{j=1}^{k} e_{ij}$. Dicho valor indica la fracción de vértices que se conectan con la comunidad i. Los autores finalmente determinan la modularidad de la matriz como:

$$Q = \sum_{i=1}^{k} e_{ii} - a_i^2 = Tr(M) - ||e^2||$$

Siendo $||e||$ la suma de los elementos de la matriz e. Dado el valor de la modularidad $Q \in [0,1]$, mientras más cercano a 1, mejor es la calidad de la partición. Los valores típicos para redes complejas suelen variar en el rango [0,3; 0,7]

3.3. Detección de comunidades

Dentro del análisis de redes complejas, un área de estudio relevante para la presente tesis es la estructura comunitaria que presentan dichas redes. Como mencionamos en la sección anterior, el grafo de usuarios que representa una red social tiene una estructura comunitaria donde los vértices suelen agruparse y tener muchas conexiones entre usuarios, mientras que hay una menor cantidad de aristas entre usuarios de distintas comunidades. Los distintos algoritmos de detección de comunidades pueden indicar que algunos vértices pertenecen a más de una comunidad, no hay necesidad de pertenencia única. El problema de detección de comunidades tiene distintos enfoques y de ellos depende la complejidad computacional y la efectividad/calidad de las comunidades detectadas.

Dentro de los métodos propuestos para detectar comunidades, existe el corte mínimo de un grafo. El objetivo del corte mínimo es minimizar la cantidad de aristas necesarias para desconectar un grafo y separar a los vértices en dos conjuntos. Esta solución puede ser únicamente útil si lo que se buscan son dos comunidades. En sí mismo, no es

un algoritmo de detección de comunidades, aunque sí se lo puede interpretar como un algoritmo que en base a la topología de la red logra separar el conjunto de vértices en dos y he ahí la detección de comunidades. La complejidad computacional es baja, dado que si las aristas tienen pesos enteros, el problema se resuelve aplicando el algoritmo de Ford-Fulkerson, que es un problema $\in P$.

Basado en la idea de corte mínimo, uno de los primeros fue el propuesto por Newman y Girvan (2004), que define el coeficiente de *betweenness* de una arista como la cantidad de caminos mínimos que pasan por una arista e. Se espera que las aristas que comunican comunidades tienen un alto coeficiente de *betweenness*, por lo que se realiza una poda en el grafo eliminando las aristas de mayor coeficiente hasta lograr separar los vértices en comunidades. Este algoritmo tiene complejidad computacional $O(E^2N)^3$.

Luego de que Newman introdujera el concepto de modularidad se desarrollaron algunos algoritmos que tienen como objetivo maximizarla: *Leading eigenvector* (Newman, 2006) se basa en el autovector asociado al autovalor de mayor valor de la matriz de modularidad del grafo para lograr el objetivo. *FastGreedy* (Clauset *et al.*, 2004) como su nombre lo indica es un algoritmo voraz. Inicia el procedimiento con todos los vértices representando una comunidad y empieza a unir nodos en comunidades en base a aquel par de comunidades que maximiza la modularidad (ver sección 3.2.4). El algoritmo finaliza cuando al unir comunidades no se incrementa la medida de modularidad. *Multilevel* (Blondel *et al.*, 2008): algoritmo de maximización de la modularidad. Es parecido a *FastGreedy*, con la variante que todos los vértices son una comunidad y se asocian vértices en una comunidad maximizando la modularidad de la red.

Hay un método en el cual un vértice es etiquetado de acuerdo a la etiqueta mayoritaria entre sus vecinos. El algoritmo se conoce como propagación de etiquetas (Xie y Szymanski, 2011) y comienza inicialmente con todos los vértices del grafo con su propia etiqueta. Se listan los vértices del grafo en orden aleatorio secuencial y para cada vértice dentro de la lista se lo etiqueta con la misma etiqueta que tiene la mayoría de sus vecinos. El algoritmo converge una vez que en todos los nodos coincide la etiqueta propia con la de sus vecinos.

Spinglass (Reichardt y Bornholdt, 2006) está basado en el modelo de Potts, un modelo extraído de la física y aplicado en la detección de comunidades.

Otros algoritmos se basan en *clustering* jerárquico, que consiste a partir de un conjunto inicial de elementos y una medida de distancia entre ellos, separarlos en distintos grupos (*clusters*) de acuerdo a dicha medida de distancia. Usualmente existen dos enfoques para este tipo de algoritmos: considerar al conjunto de vértices como una única comunidad e ir particionándola (*top-down*) o considerar a todos los vértices como una comunidad de un único vértice e ir agrupando comunidades pequeñas en más grandes hasta

el punto de convergencia (*bottom-up*). *Walktrap* (Pons y Latapy, 2005) es un ejemplo de *clustering* jerárquico. El algoritmo parte de la suposición de que caminos aleatorios (*random walks*) cortos deberían terminar en la misma comunidad, comenzando desde una partición no etiquetada.

Infomap, propuesto por (Rosvall *et al.*, 2009), descubre comunidades utilizando *random walks* en el grafo analizando el flujo de información que hay en la red. Los autores establecen una dualidad entre codificar un grafo con códigos de Huffman y buscar comunidades dentro del grafo. Se basan en la teoría de codificación de la información y la ecuación de entropía de Shannon como objetivo a minimizar y determinan que un *random walk* en el grafo permite codificar los vértices en grupo, y así buscan minimizar la cantidad de códigos de Huffman necesarios para representar la red. Utilizando esta idea, buscan el flujo de la red y los caminos que toma un caminante aleatorio (*random walker*) para segmentar los vértices en comunidades.

3.3.1. OSLOM

Lancichinetti *et al.* (2011) propusieron OSLOM como una mejora a los algoritmos de detección de comunidades hasta ese momento. El acrónimo representa *Order Statistics Local Optimization Method* y es un método estadístico que determina un *cluster* o comunidad en base a las conexiones locales de dicho *cluster*.

Se define la significación estadística de un *cluster* como la probabilidad de encontrar al mismo *cluster* en un modelo nulo de referencia. El modelo nulo utilizado fue *configuration model*, y permite generar un grafo aleatorio manteniendo la secuencia de grados de los vértices del grafo original.

El método mide la significación estadística de un *cluster* compuesto por vértices, toma a los vértices externos de la comunidad y calcula la probabilidad de que un vértice externo al *cluster* esté conectado con vértices del mismo en los grafos aleatorios generados. Mientras menor es la probabilidad de encontrar conexiones del vértice hacia el cluster en el modelo nulo, mejor es la posibilidad del vértice a pertenecer a la comunidad en el grafo original. Iterativamente se refinan las comunidades de vértices encontradas, hasta obtener varias comunidades solapadas. Luego se define una forma de poda determinando cuando dos comunidades comparten muchos vértices y eliminando una de ellas.

Una vez determinadas las comunidades, se inicia un proceso de jerarquización de las mismas, donde se genera un súper vértice representando a cada una de las comunidades encontradas, y conectando dichos súper vértices de acuerdo a las conexiones internas de los vértices y las comunidades. Nuevamente se vuelve a aplicar el algoritmo sobre este nuevo grafo generado. Este proceso jerárquico continúa hasta que en el proceso

de detectar el nuevo nivel de la red queda un único súper vértice que representa la comunidad de la red entera.

Finalmente la salida del algoritmo son cada una de las comunidades detectadas, el conjunto de vértices que pertenecen a ella y los niveles jerárquicos de las comunidades.

3.4. Redes bipartitas

Como ya hemos definido en la sección 3.1 algunos grafos tienen estructura bipartita. Una de las tantas interpretaciones que se le pueden dar a las redes bipartitas es que uno de los conjuntos de vértices son documentos que contienen palabras, y el otro conjunto de vértices son las palabras, existiendo conexiones entre un documento y una palabra si dicha palabra aparece en el documento. Esta interpretación permite construir la matriz documento-término, que es la matriz de adyacencia del grafo, considerando únicamente como filas los vértices que son documentos y como columnas los vértices que son términos (dado que no pueden existir conexiones documento-documento ni término-término por la naturaleza bipartita de la red). Con esta matriz, se desarrolló y definió una métrica llamada TF-IDF (*Term frequency - inverse document frequency*) que indica la relevancia de un término dentro de los documentos, considerando la frecuencia del término en los documentos (*term frequency*) así como también en que documentos aparece el término, dando un indicador de la especificidad del término dentro de los documentos.

A continuación veremos un ejemplo de cómo funciona la métrica de TF-IDF:

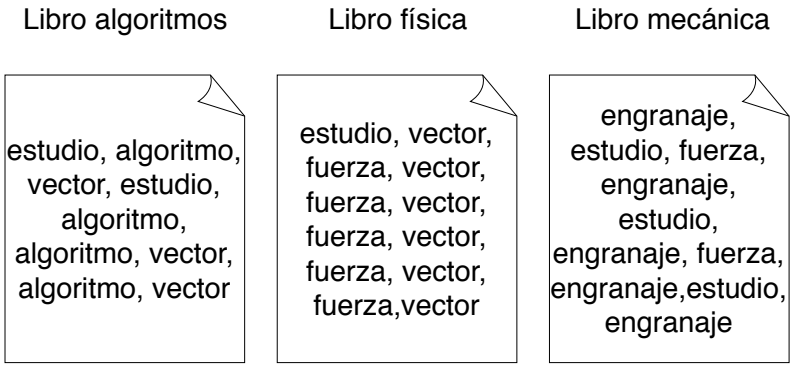


Figura 3.2: Comenzamos con tres documentos y las palabras que contienen

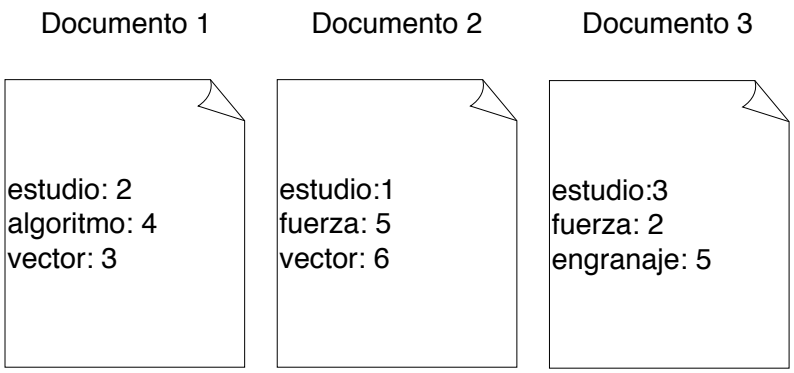


Figura 3.3: Para cada documento se cuenta cuántas veces aparece cada término

---

TF

	estudio	algoritmo	vector	fuerza	engranaje
D1	2	4	3	0	0
D2	1	0	6	5	0
D3	3	0	0	2	5

Figura 3.4: Se construye la matriz documento término, en filas tenemos todos los documentos y en columnas los términos. Si algún término no está en un documento, el valor es cero.

IDF	log(3/3)	log(3/1)	log(3/2)	log(3/2)	log(3/1)
	estudio	algoritmo	vector	fuerza	engranaje
D1	2 * log(3/3)	4 * log(3/1)	3 * log(3/2)	0 * log(3/2)	0 * log(3/2)
D2	1 * log(3/3)	0 * log(3/1)	6 * log(3/2)	5 * log(3/2)	0 * log(3/2)
D3	3 * log(3/3)	0 * log(3/1)	0 * log(3/2)	2 * log(3/2)	5 * log(3/2)

Figura 3.5: Para cada documento, se computa la proporción de cada término en el mismo.

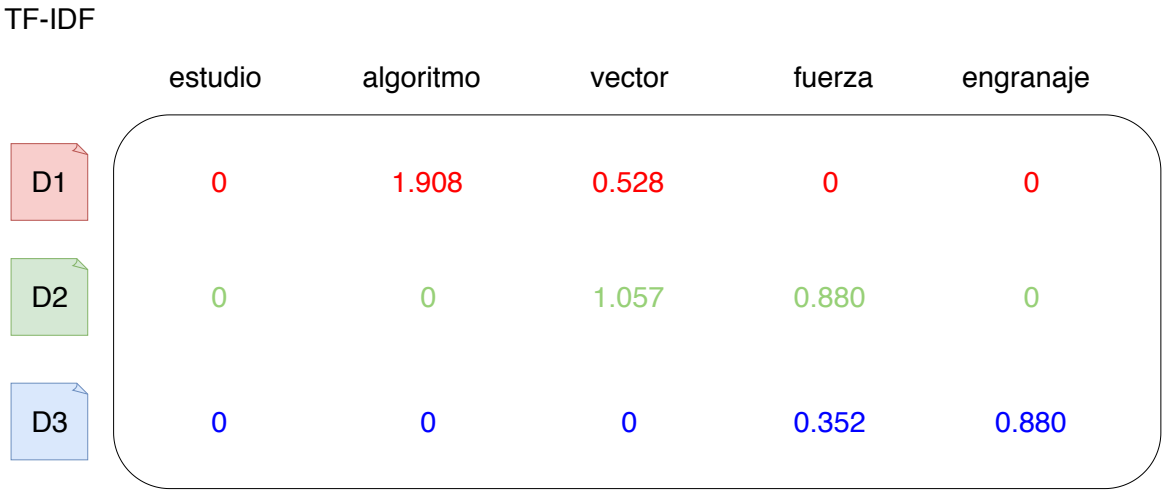


Figura 3.6: Finalmente, la métrica TF-IDF para cada término dentro de cada documento.

Considerando a los documentos por su representación vectorial de los términos que lo componen, se puede utilizar la similitud coseno para medir distancia entre ellos. La similitud coseno para dos vectores $u, v \in R^n$ se define como:

$$sim(u, v) = \cos(\theta) = \frac{\sum_{i=1}^{n} u_i \cdot v_i}{\sqrt{\sum_{i=1}^{n} u_i^2} \cdot \sqrt{\sum_{i=1}^{n} v_i^2}}$$

La función coseno está acotada en el intervalo $\cos(\theta) \in [-1, 1]$. Dados dos documentos, mientras más términos contengan en común, más próximos serán los vectores, por ende mayor será el valor del coseno del ángulo.

3.5. Visualización

Hemos determinado que la red de Twitter bajo estudio en la presente tesis es una red compleja, y para poder estudiarla necesitamos de herramientas que nos permitan comprender cómo está compuesta. La visualización emerge de la necesidad de interpretar conceptos abstractos, números y grandes cantidades de datos resumidos en imágenes que podemos comprender rápidamente, discriminando lo esencial de lo accesorio.

Visualizar significa formar en la mente una imagen visual de un concepto abstracto[1]. Es algo que va más allá de la percepción que tenemos las personas a través de los sentidos. En definitiva, la visualización de información es entonces la utilización de representaciones visuales e interactivas por computadora de datos abstractos, con el fin de aumentar su comprensión (Card, 1999). Nuestro interés es la visualización de grafos, particularmente la visualización de una red compleja como es el conjunto de datos bajo

[1]Diccionario de la Real Academia Española

análisis de Twitter.

Como primera instancia, el objetivo es lograr ver todos los vértices del grafo de forma dispersa en una imagen mientras se agrupan aquellos nodos que estén más interconectados y minimizar la cantidad de cruces de aristas para poder mostrar la mayor cantidad posible. Una de las herramientas más difundidas/potentes/utilizadas es Gephi (Bastian *et al.*, 2009) que permite ver grafos de una manera sencilla e interactiva. Permite variar los tamaños, colores, posición de los vértices, así como también filtrar aristas para poder obtener una imagen más legible de la interconexión de nodos. Además incluye cálculo de métricas de centralidad como la modularidad, distribución de grado de los vértices de forma de *power law* entre tantas otras funcionalidades. Para el análisis que queremos realizar en redes complejas Gephi presenta algunas limitaciones, principalmente porque no soporta el tamaño de la red que queremos visualizar.

La visualización de redes complejas es un importante objeto de estudio dentro del área porque no es una tarea trivial: en primer lugar algunas herramientas no tienen disponibles suficientes recursos computacionales como para mostrar un grafo entero, y segundo también porque en el caso de lograr dibujar en una imagen el grafo con todos los vértices y aristas que contiene, difícilmente se logre un resultado que pueda ser fácil de leer e interpretar a simple vista.

Por eso es que se han ido desarrollando a través de los años diversas técnicas para lograr ver, interpretar y poder obtener conclusiones en redes complejas.

Pajek (Batagelj y Mrvar, 1998) es un software que está pensado para tratar con redes complejas debido a las limitaciones mencionadas. Factoriza la red recursivamente en redes más pequeñas que pueden ser tratadas con algoritmos eficientes (subcuadráticos: $O(n), O(n \cdot log(n)), O(n\sqrt{n})$) y proveer herramientas de visualización.

LaNet-vi (Beiró, 2008), acrónimo de *Large Network visualization tool*, es una herramienta basada en la descomposición de un grafo en *k-núcleos* o también en k-densos (ver sección 3.1) para observar los vértices principales de la red y sus conexiones. Aplicando el algoritmo de descomposición de *k-núcleos*, clasifica a los vértices de mayor a menor y los dispone en un gráfico. Esta herramienta muestra una imagen de la red con los vértices pertenecientes al mayor núcleo, y a medida que se reduce el núcleo de los nodos, se empiezan a expandir los nodos hasta llegar hasta los extremos.

A continuación presentamos en la figura 3.7 la visualización de la red de coocurrencia de *hashtags*, que explicaremos en detalle en la sección 4.3. La escala de colores de la derecha representa el *k-denso* (ver sección 3.1) en el que se encuentra un vértice, utilizando los colores de un arcoíris. El color rojo representa a los vértices centrales, mientras que el violeta representa vértices alejados. A su vez, la escala de la izquierda representa el grado del vértice dentro del grafo. Mientras mayor el grado del vértice, mas grande será

el radio del círculo que lo representa.

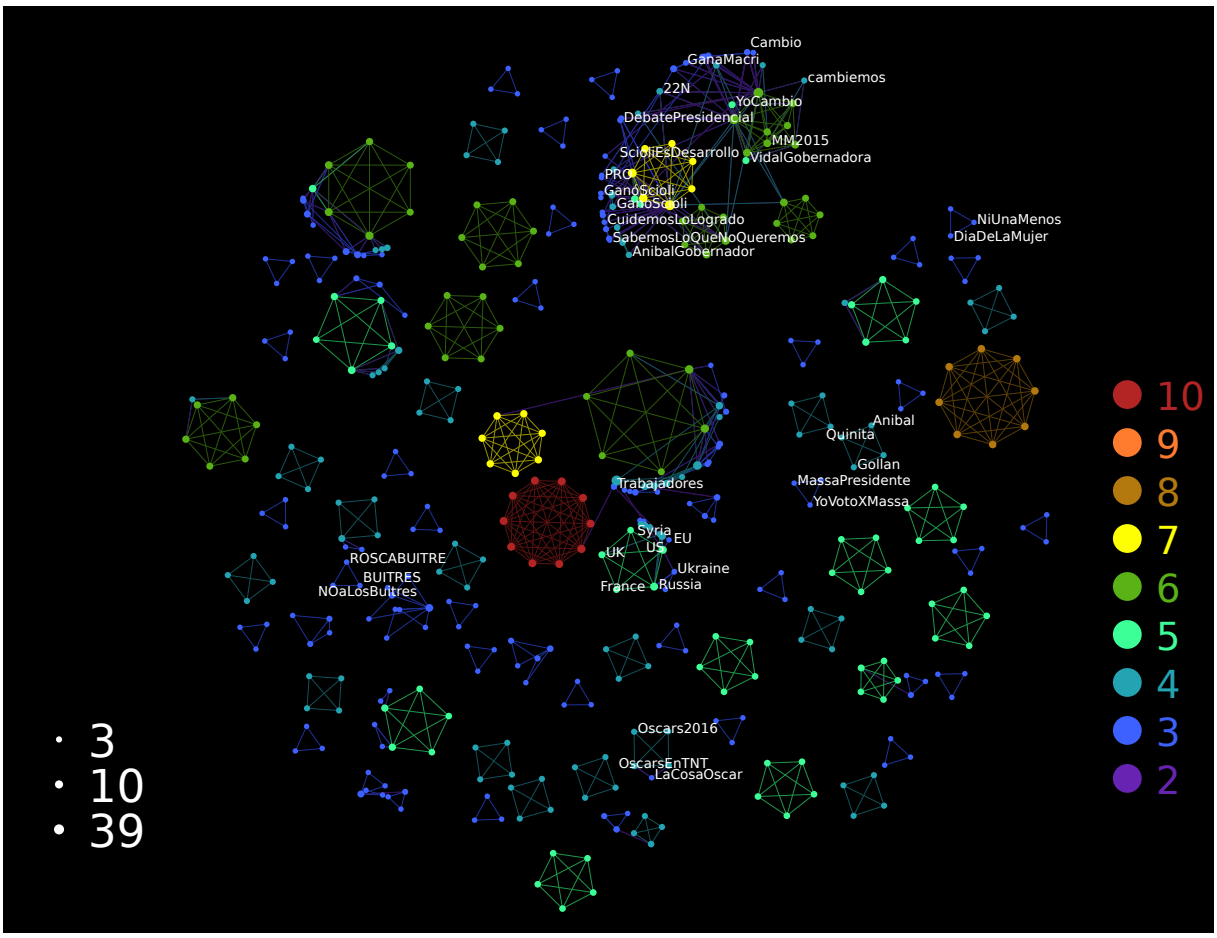


Figura 3.7: Descomposición en *k-densos* de la red de coocurrencia de *hashtags* para aquellos cuya coocurrencia es mayor o igual a 60. Se ven los vértices que se encuentran dentro del *k-denso* 3 en adelante.

Capítulo 4

Desarrollo

Luego de la introducción teórica y de herramientas, en este capítulo explicaremos todo el trabajo que realizamos en la presente tesis. El código del mismo se encuentra en el repositorio público de GitHub: `https://github.com/CoNexDat/twitter-topics`.

Disponemos de los *tweets* publicados por los usuarios, de ellos extrajimos los *hashtags* y los clasificamos dentro de temas de conversación, o como los llamaremos de aquí en adelante, tópicos, para luego determinar que los usuarios publicaron sobre tópicos en un determinado momento. Con ello primero hicimos un análisis dentro de todo el período, es decir, sin considerar la temporalidad de la utilización de *hashtags*, y medir la similitud entre usuarios de acuerdo a los tópicos que utilizaron. Luego, agregamos el factor temporal para hacer una progresión de la evolución de la similitud sobre grupos de usuarios que siguen a un único candidato.

También realizamos un análisis sobre la utilización temporal de los tópicos por parte de dichos usuarios y determinar cual de ellos utilizó más un tópico. Finalmente, experimentamos en la predicción de preferencias de usuarios por un candidato utilizando como base los tópicos que utilizaron.

4.1. Conjunto de datos

Panario (2016) realizó la extracción del conjunto de datos de Twitter utilizado en la presente tesis. El mismo se capturó a través de su interfaz pública expuesta para consultar usuarios, relaciones de seguimiento entre ellos y los *tweets* que los mismos publican.

Los datos fueron capturados dentro del período que abarcó las elecciones presidenciales de Argentina de 2015, desde el 15 de julio del mismo año hasta el 31 de marzo de 2016. Dentro de esa ventana de tiempo, se almacenaron aproximadamente 54 millones de *tweets* pertenecientes a 343 mil usuarios y 110 millones de relaciones seguidor-amigo entre los usuarios de la red.

Para capturar a los usuarios y sus *tweets* se eligieron siete cuentas oficiales correspondientes a políticos argentinos, a quienes de aquí en adelante denominaremos **supernodos**, y sus *tweets*:

- Cristina Fernández de Kirchner (@CFKArgentina): 532 *tweets*
- Aníbal Fernández (@FernandezAnibal): 322 *tweets*
- Ernesto Sanz (@SanzErnesto): 26 *tweets*
- Sergio Massa (@SegioMassa): 262 *tweets*
- Margarita Stolbizer (@Stolbizer): 465 *tweets*
- Daniel Scioli (@danielscioli): 845 *tweets*
- Mauricio Macri (@mauriciomacri): 724 *tweets*

Luego, se capturaron cuentas de usuarios que siguieran a alguno de los supernodos, sus *tweets* y relaciones de seguidor-amigo dentro de la red para continuar expandiendo el grafo de usuarios. Con este conjunto de datos podemos reconstruir el *grafo de usuarios de Twitter*, junto con los *tweets* publicados por ellos.

En la presente tesis utilizamos el grafo inducido de usuarios formado por las relaciones seguidor-amigo y obtenemos **7.616.836** de relaciones de quien sigue a quien como las *aristas* del grafo de usuarios. Aquí vemos reducido nuestro conjunto de usuarios (que son vértices del grafo) a **308.169** usuarios, de los cuales **218.353** de ellos escribieron los **50.093.913** de *tweets* que analizamos. Excluimos del análisis a los usuarios que no siguen o son seguidos por otros usuarios.

En la figura 4.1 observamos un ejemplo de los usuarios capturados y las relaciones de seguidor-amigo entre ellos. El color verde distingue a los nodos de los cuales se disponen

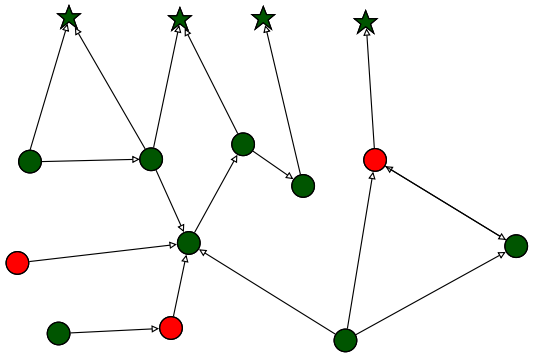


Figura 4.1: Red de Twitter capturada. Las estrellas representan a los supernodos y los círculos a usuarios. El color verde significa que ese usuario escribió al menos un *tweet*, en rojo si no se dispone de ningún *tweet*.

tweets publicados. En cambio, el color rojo denota la inexistencia de *tweets* en el conjunto de datos capturado.

4.1.1. Análisis previo de datos

En esta sección haremos un análisis de la cantidad de *tweets* publicados por usuario, cantidad de *tweets* por día durante el período capturado y visualización del subgrafo inducido de usuarios con un grado entrante mayor o igual a 5000. También veremos la distribución de utilización de *hashtags* por usuario.

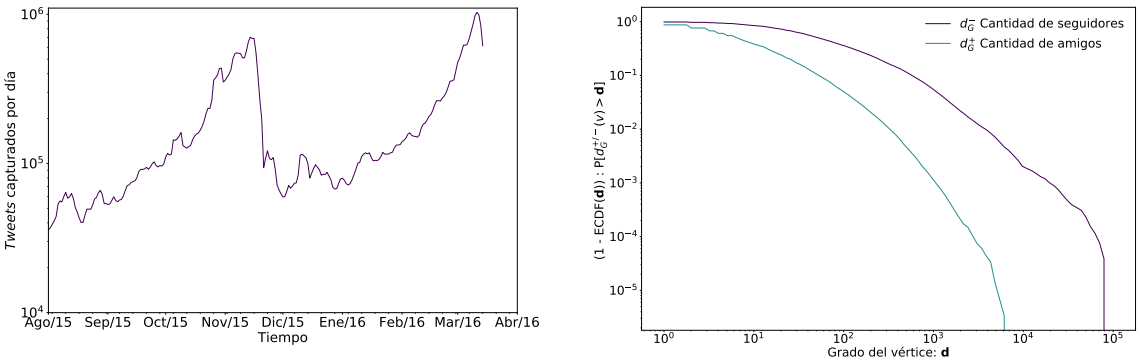


Figura 4.2: Cantidad de *Tweets* capturados por día (Izquierda). Distribución de grado entrante y saliente de cada usuario (Derecha)

A la izquierda de la figura 4.2 se puede observar la cantidad de *tweets* capturados en el período bajo análisis. La escala logarítmica marca la diferencia de *tweets* que hay en los meses de octubre y noviembre, pasando de cien mil a casi un millón de capturas diarias.

A la derecha de la figura 4.2 se puede apreciar la distribución de grados de los usuarios dentro de la red. No hay un ajuste perfecto de distribución de ley de potencias para los grados entrante y saliente, pero sí se observa que existe un reducido grupo de personas cuyo grado entrante es del orden de miles mientras que existen muchos usuarios cuyo grado entrante es del orden de cientos.

Dentro de la red de usuarios, presentamos en el cuadro 4.1 quienes son los treinta usuarios más seguidos de la red. Están presentes varias personas de la política, así como también se encuentran personas famosas y del periodismo.

Cuadro 4.1: Tabla de los usuarios con más seguidores en la red capturada de Twitter.

Usuario	Cantidad de seguidores en la red
Mauricio Macri	88138
Daniel Scioli	73777
Jorge Lanata PPT	70177
Clemente Cancela	62557
MARLEY - ALE WIEBE	51175
Sergio Massa	51130
H Rodríguez Larreta	48379
Radio Mitre	44886
Casa Rosada	43023
@lauritalonso	36585
Margarita Stolbizer	32385
Esteban Bullrich	31468
NiUnaMenos FETCHEVES	29154
Diego Santilli	28300
Daniel Tognetti	28243
Hernán Lombardi	26634
teleSUR TV	25620
Gabriela Cerruti	25417
Federico Pinedo	25315
Lizy Tagliani	24617
Juan Cabandié	24428
Gaby Levinas	22645
Chequeado	22517
Elisa Carrió.	20789
Sergio Bergman	20047
Natacha Jaitt	19678
PRO	19613
Fede Sturzenegger	19279
Equipo CFK	19142
Martín Insaurralde	19007

En la figura 4.3 se puede observar el subgrafo inducido por los usuarios con 5000 seguidores o más. La figura fue hecha con la herramienta Gephi (Bastian *et al.*, 2009), utilizando la disposición *Force Atlas 2*. Gephi utiliza una implementación del algoritmo voraz desarrollado por Blondel *et al.* (2008) que maximiza la modularidad. Se ejecutó dicho algoritmo y se detectaron tres comunidades distinguidas por los colores:

- Grupo violeta: aquí predominan personalidades políticas, entre las cuales se destacan los cuatro candidatos capturados a presidente. También se encuentran figuras periodísticas, tales como Pilar Rahola, Patricia Janiot, Clemente Cancela y Gabriel Levinas.

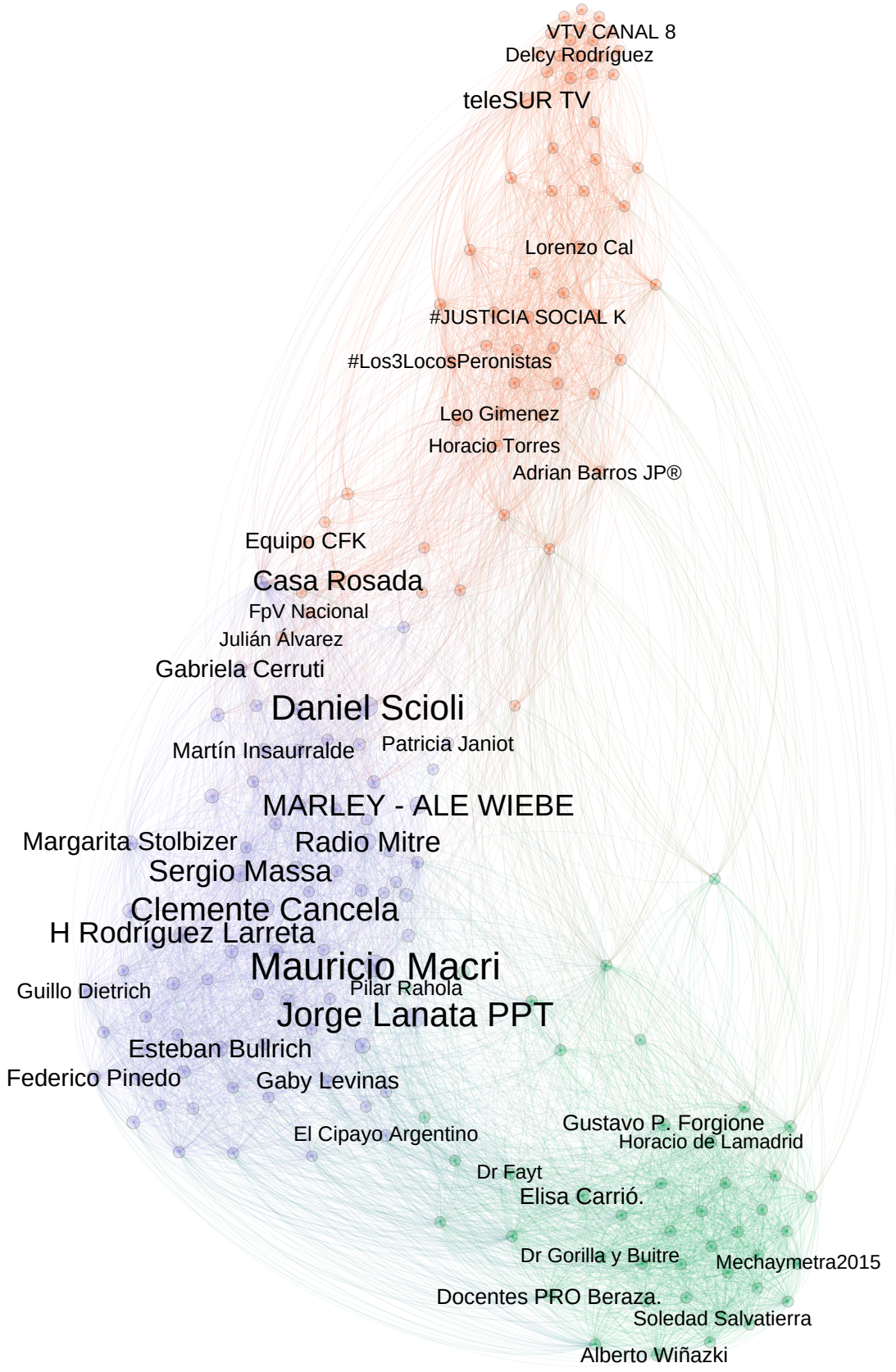


Figura 4.3: Grafo de usuarios de la red de Twitter para aquellos usuarios con 5000 seguidores o más.

- Grupo naranja: este grupo está caracterizado por cuentas de Twitter asociadas a la izquierda política tanto de Argentina con cuentas como Equipo CFK haciendo referencia a Cristina Fernández de Kirchner, y también de Venezuela con cuentas como teleSUR y Delcy Rodríguez.

- Grupo verde: aquí se encuentra la política Elisa Carrió, que comparte el mismo espacio político que el candidato Mauricio Macri.

Por último vemos la distribución de utilización de *hashtags* por parte de usuarios. La distribución de la izquierda es la cantidad de *hashtags* únicos utilizados mientras que la distribución de la derecha es la cantidad total de *hashtags* utilizados.

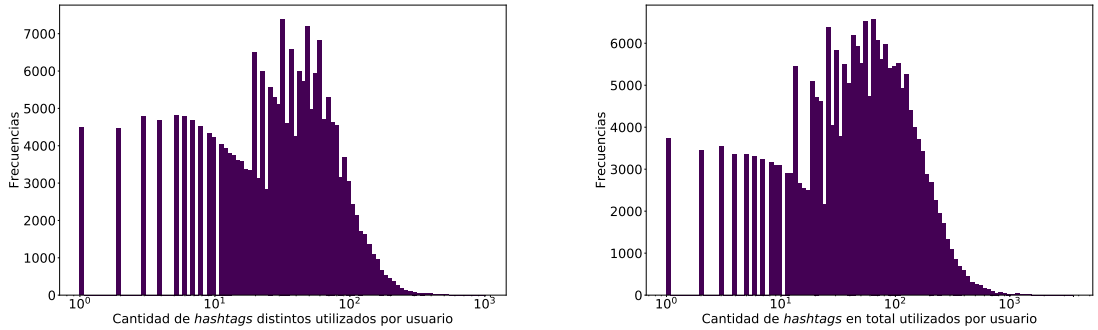


Figura 4.4: Cantidad de *hashtags* utilizados por usuario. *Hashtags* distintos (Izquierda). Total de *hashtags* (Derecha).

De la figura 4.4, se observa en el gráfico a la izquierda que hay pocos usuarios que hayan utilizado más de cien *hashtags* distintos en todo el período. Del gráfico a la derecha se concluye que son pocos los usuarios en proporción que no han utilizado ningún *hashtag* o que han utilizado pocas veces uno. A pesar de estas diferencias, los gráficos son muy similares por lo que podemos afirmar que un usuario no utiliza muchas veces un mismo *hashtag*. Si así fuera, habría una diferencia mayor entre ambos gráficos.

4.2. Contexto político de elecciones

Las elecciones primarias de Argentina en el año 2015 (PASO) se realizaron el 9 de agosto, donde se definieron los candidatos de cada uno de los partidos a participar en las elecciones generales. Luego, las elecciones presidenciales tuvieron lugar en primera instancia el 25 de octubre, y posteriormente el 22 de noviembre en balotaje. Dentro de los candidatos a presidente, enumeramos los capturados en nuestro conjunto de datos y el cargo ocupado al momento de las elecciones:

- Mauricio Macri: Jefe de Gobierno de la Ciudad Autónoma de Buenos Aires se presentó con la coalición política nacional opositora denominada Cambiemos.
- Daniel Scioli: Gobernador de la provincia de Buenos Aires. Proveniente del partido Frente para la Victoria, el mismo partido político de la presidente de la nación Cristina Fernández de Kirchner.
- Sergio Massa: Diputado Nacional de la coalición política del Frente Renovador. Se presentó a las elecciones dentro de una coalición llamada Unidos por una Nueva Alternativa.
- Margarita Stolbizer: Diputada Nacional de la coalición Frente Progresista, Cívico y Social.

Los resultados de la primera vuelta de las elecciones para los cuatro candidatos de nuestro interés fueron [1]:

1. Daniel Scioli: 37.08 %
2. Mauricio Macri: 34.15 %
3. Sergio Massa: 21.39 %
4. Margarita Stolbizer: 2.51 %

Dado que el primer candidato no obtuvo los votos necesarios para ser electo en primera vuelta, se programó el balotaje para el 22 de noviembre, cuyo resultado fue el siguiente [2]:

1. Mauricio Macri: 51.34 %
2. Daniel Scioli: 48.66 %

El presidente electo resultante fue Mauricio Macri por una pequeña ventaja sobre Daniel Scioli.

4.3. Detección de tópicos

En esta sección desarrollaremos la metodología aplicada para obtener los temas de debate que utilizaron los usuarios capturados en nuestro conjunto de datos. De los *tweets* publicados por los usuarios, extrajimos los *hashtags* para clasificarlos en tópicos. Finalmente realizamos una evaluación y visualización de los mismos.

[1] `https://www.argentina.gob.ar/sites/default/files/resultados_2015_nacionales_presidente_y_vicepresidente.xlsx`

[2] `https://www.argentina.gob.ar/sites/default/files/resultados_2015_segunda_vuelta_presidente_y_vicepresidente.xlsx`

4.3.1. Metodología aplicada

Para la detección, utilizamos como base los *hashtags* contenidos en *tweets* que publicaron los usuarios. Los *hashtags* son la clasificación de temas propia que tiene Twitter, y al disponer de una gran cantidad, agrupamos esos *hashtags* en temas de debate o **tópicos**.

También nos interesa puntualmente conocer los *hashtags* utilizados por los seguidores de políticos. En este análisis, identificamos a usuarios que siguen únicamente a uno de los cuatro candidatos a presidente, para marcarlos como seguidores unívocos de un político y estudiar *hashtags* posiblemente relacionados a un político.

Dentro del conjunto de datos, obtenemos aquellos *tweets* que no fueron *retweeteados*, es decir, nos quedamos con los contenidos originales, y no difusiones que hayan hecho terceros sobre el contenido original del *tweet*. De esta forma, tenemos la cantidad de veces que un *hashtag* fue utilizado en conjunto con otros. El primer filtro aplicado es eliminar aquellos *tweets* que tienen únicamente un solo *hashtags*.

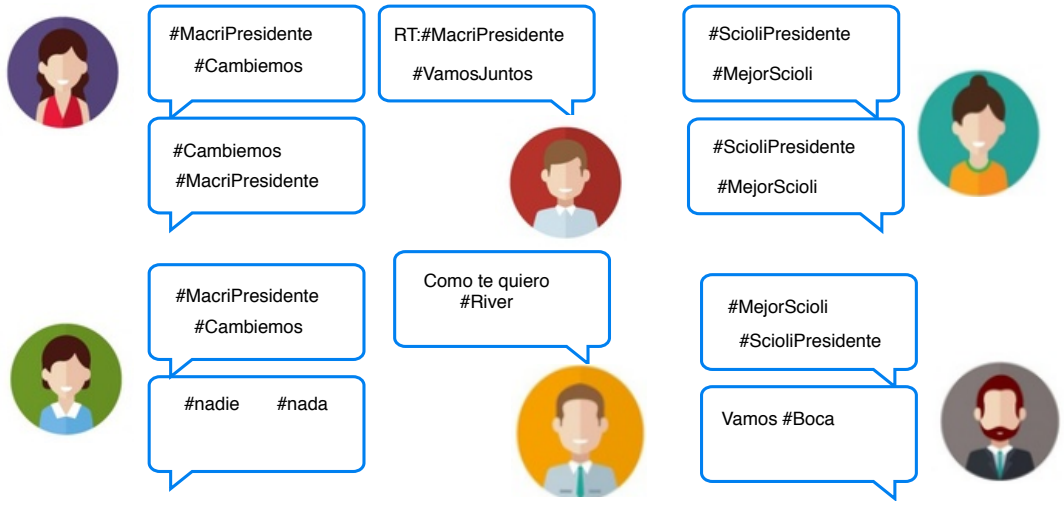


Figura 4.5: Ejemplo de publicación de *tweets*. `#ScioliPresidente` y `#MejorScioli` fueron utilizados en conjunto tres veces por dos personas distintas.

En el ejemplo presentado en la figura 4.5 no estarán presentes en la matriz de co-ocurrencia los *hashtags* `#nadie` y `#nada` porque fueron utilizados una única vez y no superan el umbral propuesto de usos. Tampoco estarán presentes `#River` ni `#Boca` dado que no fueron utilizados en conjunto con otros *hashtags*.

De los *hashtags* que superaron el umbral de utilización en conjunto, en la figura 4.6 observamos la matriz cuadrada y simétrica con los mismos, donde cada elemento de la matriz es la cantidad de veces que fueron utilizados en conjunto el *hashtag* fila con columna. La diagonal de la matriz no aporta información relevante al análisis, por eso se obvian los valores, que representarían la cantidad de usos total del *hashtag* en los *tweets*.

	#Macri Presidente	#Cambiemos	#Scioli Presidente	#MejorScioli
#Macri Presidente	-	3	1	0
#Cambiemos	3	-	1	0
#Scioli Presidente	1	1	-	3
#MejorScioli	0	0	3	-

Figura 4.6: En base a los *tweets*, generamos la matriz de coocurrencia (simétrica) de *hashtags*. #`ScioliPresidente` y #`MejorScioli` tienen una coocurrencia de tres.

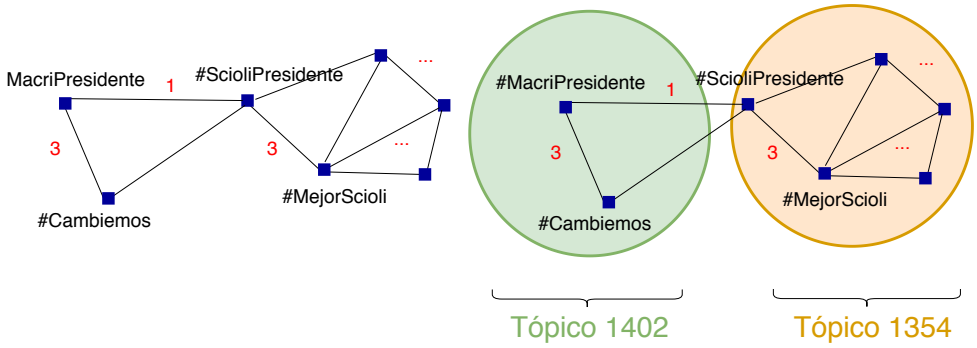


Figura 4.7: Grafo de *hashtags* a partir de la matriz de coocurrencia (izquierda). Detección de tópicos con algoritmo de detección de comunidades (derecha).

De la matriz de adyacencia presentada en la figura 4.6 se obtiene el grafo que representa en la figura 4.7 (izquierda). Luego se procede a la búsqueda de comunidades sobre el grafo (derecha).

Disponemos de una matriz que tiene como filas *tweets* y como columnas *hashtags*, al hacer el producto de la matriz traspuesta con ella misma, obtenemos la matriz de coocurrencia de *hashtags*. Esta matriz es cuadrada y simétrica, donde el elemento $a_{i,j}$ es la cantidad de veces que fueron utilizados en conjunto los *hashtags* i y j. En esta matriz eliminamos todos los elementos menores a tres, es decir, eliminamos conexiones entre *hashtags* que fueron utilizados en conjunto menos de tres veces.

Con los *hashtags* que quedaron en la matriz, aplicamos un filtro de acuerdo a qué tan comunes son los mismos dentro de los grupos de usuarios. Nuestra intención es filtrar contenido que es transversal a la contienda política, es decir que fue comentado en forma homogénea por los distintos grupos de seguidores. Un ejemplo de este tipo de

contenido son las temáticas relacionadas con deportes: no esperamos que los seguidores de un candidato sigan en su mayoría a un mismo equipo de fútbol, por ello nos interesa filtrar este tipo de contenido para segmentar mejor los tópicos.

Con el objetivo de filtrar *hashtags* de acuerdo a que tan relevantes son en las distintas comunidades de seguidores, utilizaremos la divergencia de Kullback y Leibler (1951). La divergencia de Kullback-Leibler mide la similitud entre dos distribuciones de probabilidad, una de ellas es la distribución que se quiere estimar y la otra distribución es la que se toma como referencia, calculando la entropía de la distribución a estimar y la entropía cruzada entre la distribución conocida y la estimada. Nuestra definición de entropía hace referencia a la entropía de la información que enunció en su trabajo Shannon (1948) para determinar cuál era la mínima cantidad de bits necesarios para codificar información en función de las probabilidades de ocurrencia.

En nuestro caso de estudio, partimos de la distribución de los grupos de usuarios que siguen únicamente a un candidato político, y la utilizamos como la distribución de probabilidad de referencia. La distribución de probabilidad a determinar es la distribución de usos de cada *hashtag* por parte de los usuarios que siguen a los candidatos. Mientras más parecida sea la distribución de la utilización de un *hashtag* h a la distribución de las comunidades de usuarios, menor será el valor de la divergencia de Kullback-Leibler del mismo. Se calcula de la siguiente forma:

$$D_{KL}(h) = \sum_{i=1}^{4} P_h(i) \cdot \log_2 \left[\frac{P_h(i)}{Q(i)} \right]$$

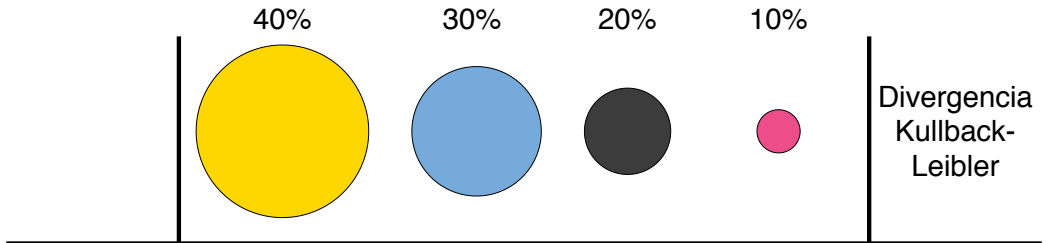
Siendo P_h la distribución de la utilización del *hashtag* h dentro de las cuatro comunidades de usuarios y Q la distribución de tamaños de dichas comunidades.

En la figura 4.8 presentamos ejemplos ficticios de como sería el cálculo para distintas distribuciones de probabilidad en usos de *hashtag* y una distribución ficticia y proporcionada de la probabilidad de referencia q_k que representa los tamaños de los grupos de los seguidores de candidatos.

En la figura 4.9 se presenta el cálculo de la divergencia para un *hashtag*. Si la probabilidad p_k es cero, como no está definido el logaritmo para números menores o iguales a cero, se descarta ese término y se continúa con las otras probabilidades.

Computamos los valores de la divergencia para todos los *hashtags*, y eliminamos el 5 % de los *hashtags* de menor valor de divergencia, dado que son las distribuciones más cercanas a la distribución de comunidades.

En el cuadro 4.2 se encuentran algunos de los *hashtags*, a la izquierda los que fueron eliminados de la matriz de coocurrencia y a la derecha aquellos que quedaron. Los eliminados corresponden a temas que no nos interesan dentro de nuestro análisis, como



	40%	30%	20%	10%	Divergencia Kullback-Leibler
#ArgentinaDebate	4	3	2	1	0.000
#MacriPresiente	40	2	3	1	0.485
#ScioliPresiente	4	60	1	1	0.826
#StolbizerPresidente	0	0	3	80	2.122
#YoVotoxMassa	0	0	40	0	1.609
#FutbolPorSiempre	5	5	3	2	0.012

Figura 4.8: Ejemplo divergencia de Kullback-Leibler de distintos *hashtags* y sus usos con una particular distribución de tamaños de comunidades.

$$D_{KL}(\#\texttt{SP}) = 0 \cdot \ln\left[\frac{0}{0{,}4}\right] + 0 \cdot \ln\left[\frac{0}{0{,}3}\right] + \frac{3}{83} \cdot \ln\left[\frac{\frac{3}{83}}{0{,}2}\right] + \frac{80}{83} \cdot \ln\left[\frac{\frac{80}{83}}{0{,}1}\right]$$

$$D_{KL}(\#\texttt{SP}) = 0{,}0361 \cdot (-1{,}7108) + 0{,}9639 \cdot 2{,}2658 = -0{,}0618 + 2{,}1840 = 2{,}122$$

Figura 4.9: Cálculo de la divergencia de Kullback-Leibler para el *hashtag* #`StolbizerPresidente`.

deportes y programas de TV. También quedan eliminados términos neutros como por ejemplo #`Eleccion2015`. En cambio, los *hashtags* remanentes tienen contenido político partidario.

Finalmente, la matriz obtenida es la matriz de coocurrencia de *hashtags*, la cual representa al grafo de coocurrencia, donde los vértices son *hashtags* y las conexiones entre ellos son la cantidad de veces que fueron utilizados en conjunto.

Sobre este grafo aplicamos el algoritmo de detección de comunidades OSLOM (Lancichinetti *et al.*, 2011) para clasificar los *hashtags* en tópicos.

El algoritmo OSLOM, desarrollado en la sección 3.3.1, determina la significación estadística local de un vértice de la red con respecto a sus vecinos, por lo que no necesariamente un vértice pertenece a una única comunidad. Es decir, un *hashtag* puede pertenecer a más de un tópico. Esto nos introduce algunos problemas, porque existen

Cuadro 4.2: Tabla de *hashtags* eliminados (izquierda) y *hashtags* remanentes (derecha).

Hashtags eliminados	$D_{Kullback-Leibler}$	*Hashtags* más importantes	$D_{Kullback-Leibler}$
SelecciónArgentina	0.0013	GanoStolbizer	2.5099
YoYaVote	0.0040	StolbizerPresidenta	2.4804
ShowMatch	0.0062	YoVotoAStolbizer	2.4686
BreakingBadEnAmerica	0.0068	StolbizerDebate	2.4190
Eleccion2015	0.0079	7MParoNacionalDeMujeres	2.4070
Rusia2018	0.0098	Progresistas	2.3683
EleccionesPresidenciales	0.0104	YovotoXMassa	2.2443
Argentina	0.0114	MassaAlBalotage	2.0947
ChapoGuzman	0.0116	MassaDebate	2.0337
Elecciones25deOctubre	0.0119	MacriNoPuede	2.0331
RiverEnJapon	0.0119	PresidenteDeMesa	1.7296
Monzón	0.0127	SanzPresidente	1.5347
Los8Escalones	0.0132	AbortoSeguroLegalyGratuito	1.4253
Economia	0.0151	EnergiasRenovablesYa	1.3671
DiputadosDeIzquierda	0.0154	MacriNuncaMas	0.9214
Piscis	0.0159	ScioliPresidenteEnPrimeraVuelta	0.8996
PASO	0.0159	GraciasPorEstos12AñosCFK	0.8867
CausaNisman	0.0167	MassaPresidente	0.8239
Balotaje2015	0.0178	YoVoteAScioli	0.7947
ArgentinaDebate	0.0181	ScioliPresidente	0.7435
PapaFrancisco	0.0212	MM2015	0.6321
FelizDiaDeLaBandera	0.0226	MacriPresidente	0.5240
ElFuturoPresidente	0.0236	IgualdadDeGénero	0.4412
Sarmiento	0.0243	*Refugees*	0.1928
VenezuelaDecide	0.0245	Oscars	0.0949
SodaStereo	0.0255	UnOscarParaLeo	0.0902
RelatosSalvajes	0.0266	NiUnaMenos	0.0868
CristinaKirchner	0.0299	Cuba	0.0816
LosSimpson	0.0303	AtentadoFrancia	0.0527
TheBeatles	0.0307	Merkel	0.0384

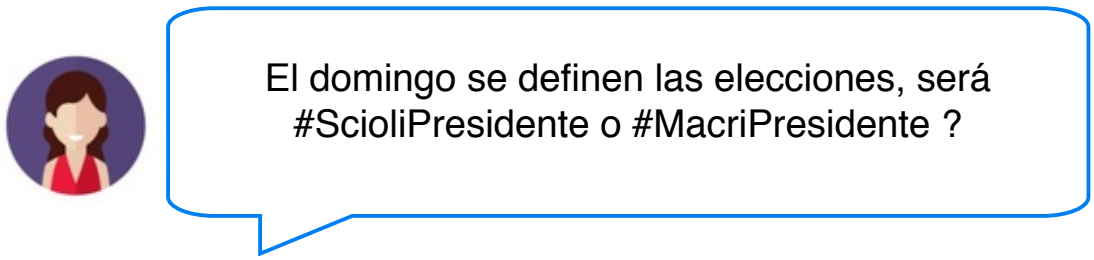


Figura 4.10: Ejemplo ficticio de un *tweet* que conecta dos *hashtags* que queremos en distintos tópicos.

---

tweets en los cuales se utilizan dos *hashtags* que queremos en tópicos distintos, y que por como fueron utilizados, quedan compartiendo algún tópico. La figura 4.10 ilustra un *tweet* inventado para ejemplificar esta situación.

Para aquellos *hashtags* que pertenecen a más de un tópico, observamos los tópicos de los *hashtags* vecinos en el grafo, y finalmente lo asignamos al tópico en el cual haya sido utilizado en conjunto mayor cantidad de veces con sus *hashtags* vecinos. Así, logramos clasificar `#MacriPresidente` en un tópico y `#ScioliPresidente` en otro. Asignamos a los 44.379 *hashtags* en 3131 tópicos.

4.3.2. Tópicos detectados

De los tópicos más utilizados en la red, se encuentran los que publicaron los usuarios de los cuatro candidatos presidenciales, que se destacan por contener *hashtags* que expresan sentimiento positivo hacia ellos. Dada la asignación realizada por el algoritmo de detección de comunidades, para un tópico en particular extraemos el subgrafo inducido por los *hashtags* pertenecientes a dicho tópico, y luego utilizamos LaNet-vi (Beiró, 2008) para realizar la descomposición en *k-densos* del subgrafo inducido y visualizarlo.

La escala en blanco a la izquierda representa el tamaño del vértice de acuerdo a su grado. La escala de colores a la derecha representa el *k-denso* al cual pertenece el vértice. El tamaño de los vértices y la cantidad de *k-densos* que hay dentro de la red, denotan una mayor comunicación entre los usuarios. Mientras mayor el grado del vértice, más fue utilizado un *hashtag*, y además, mientras mayor el *k-denso* al cual pertenece, significa que mayor fue la utilización en conjunto con otros *hashtags*. Los *hashtags* dentro del núcleo coloreado en rojo, son los de mayor conexión y los que más han sido utilizados en conjunto por parte de los usuarios.

La figura 4.11 es una visualización del tópico de Macri. Se observan en el centro del subgrafo *hashtags* fuertemente relacionados con Macri mientras que sólo en la periferia se encuentran *hashtags* en su contra.

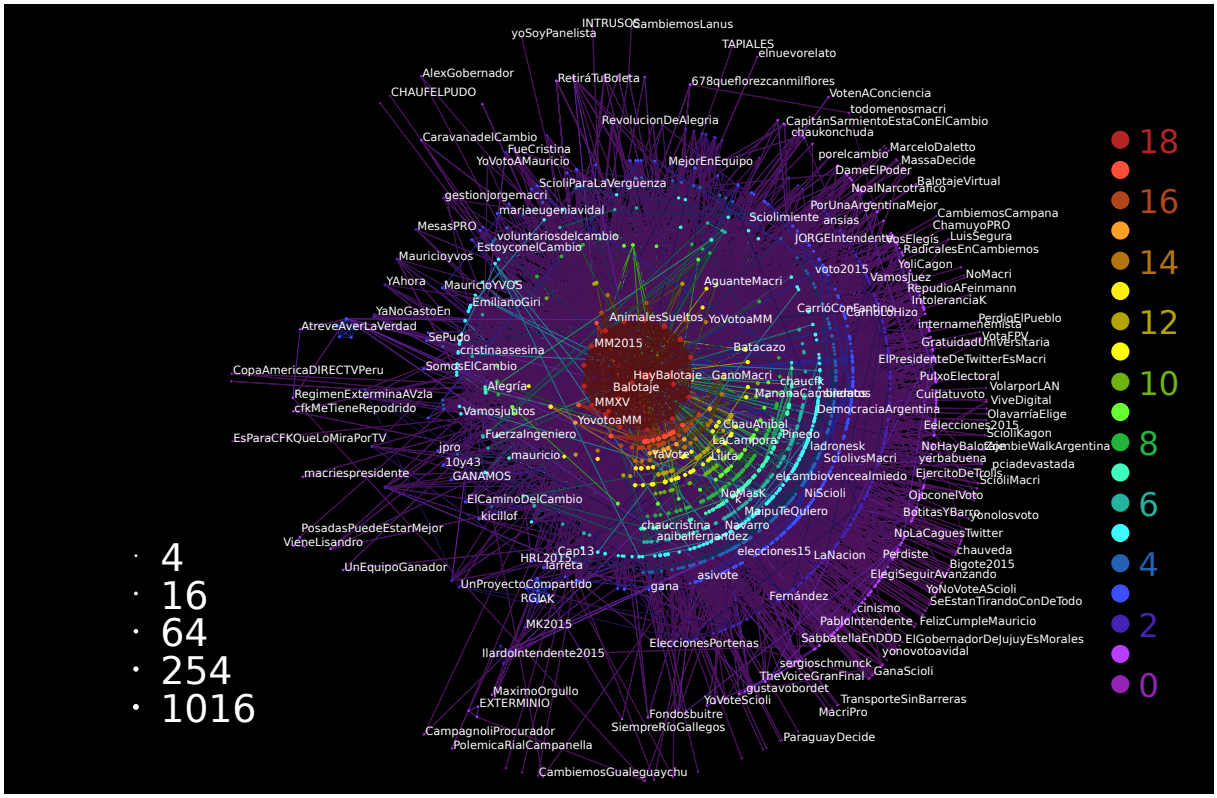


Figura 4.11: Visualización del tópico de Macri. Se destacan *hashtags* como #`YoVotoMM` y #`GanoMacri`, entre otros.

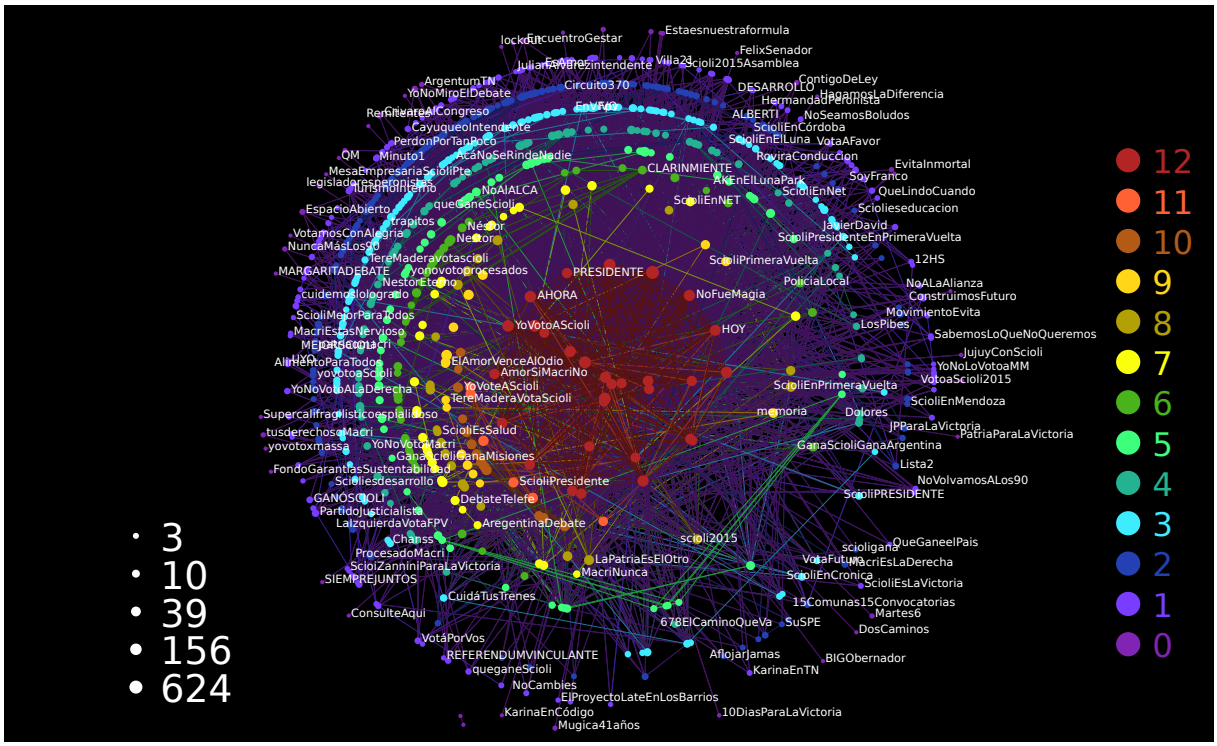


Figura 4.12: Visualización del tópico de Scioli. Se destacan *hashtags* como #`VotaScioli` y #`ScioliPresidente`, entre otros.

En la figura 4.12 se encuentra el tópico de Scioli. Al igual que el tópico de Macri, en el centro se concentran *hashtags* de apoyo al candidato y sólo en la periferia algunos que no están relacionados. Lo mismo ocurre en las figuras 4.13 y 4.14.

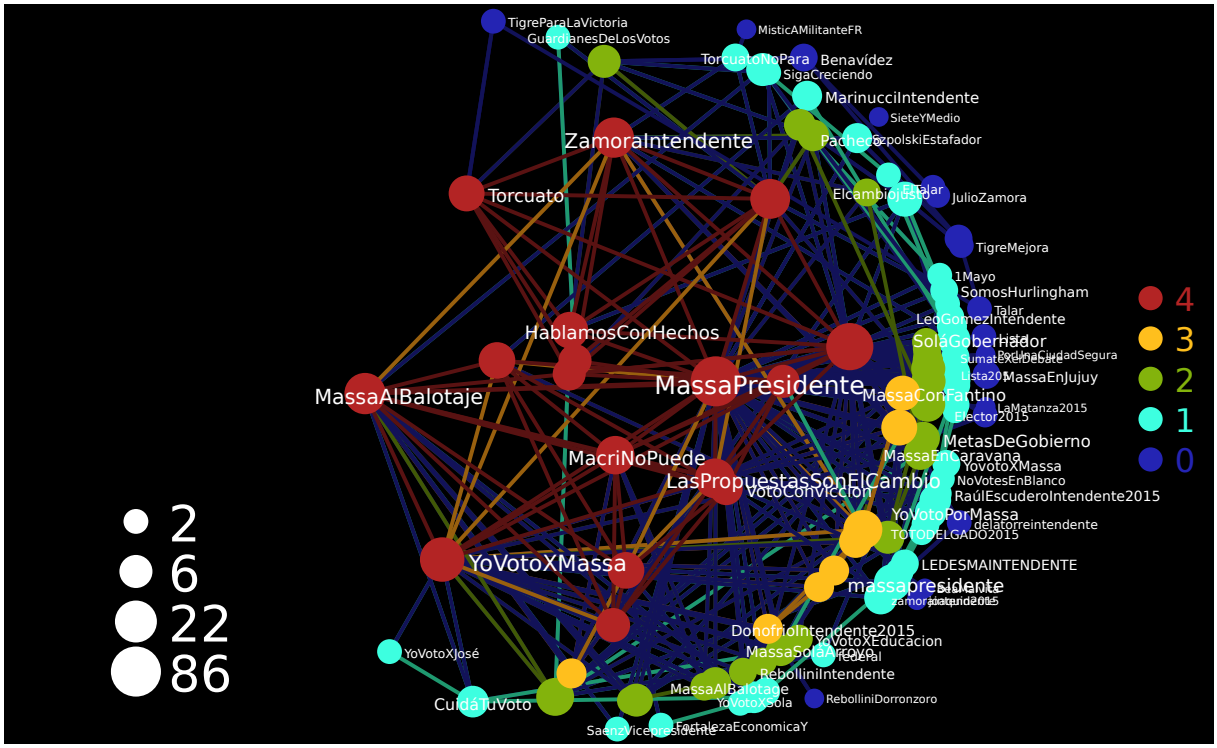


Figura 4.13: Visualización del tópico de Massa. Se destacan *hashtags* como #`MassaPresidente` y #`MassaAlBalotaje`, entre otros.

En las cuatro figuras presentadas 4.11 es el tópico de apoyo a Macri, 4.12 es el tópico de Scioli, 4.13 es el de Massa y finalmente el de Stolbizer 4.14. Se puede apreciar la popularidad de los tópicos con mirar la escala de los grados de los vértices, siendo el subgrafo inducido de mayor conexión el de Macri, y el último el de Stolbizer. Este efecto también está representado en la cantidad de *k-densos* que hay en los subgrafos inducidos. La popularidad de un tópico se puede apreciar al mirar la escala de grados de los vértices y también la cantidad de vértices que tiene el mismo. Los tópicos de Macri y Scioli contienen más vértices y hay vértices de mayor grado que en los tópicos de Massa y Stolbizer.

Además de los tópicos políticos, también logramos detectar otros temas sobre el contenido publicado por los usuarios. Por ejemplo, la figura 4.15 es un tópico sobre feminismo y cuestiones de género. Los tópicos no están restringidos al ámbito local de Argentina, sino también se encuentran tópicos como el de la figura 4.16 donde se observan *hashtags* que describen a Cuba.

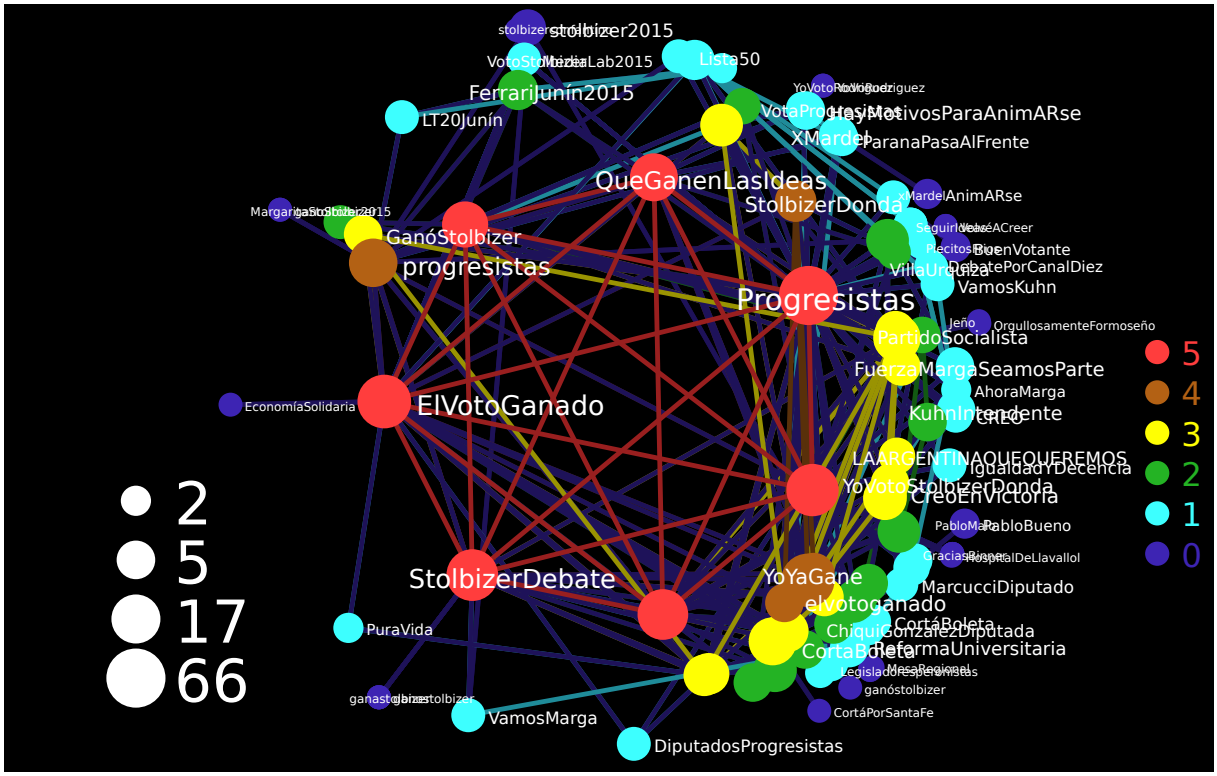


Figura 4.14: Visualización del tópico de Stolbizer. Se destacan *hashtags* como #StolbizerDebate y #Progresistas, entre otros.

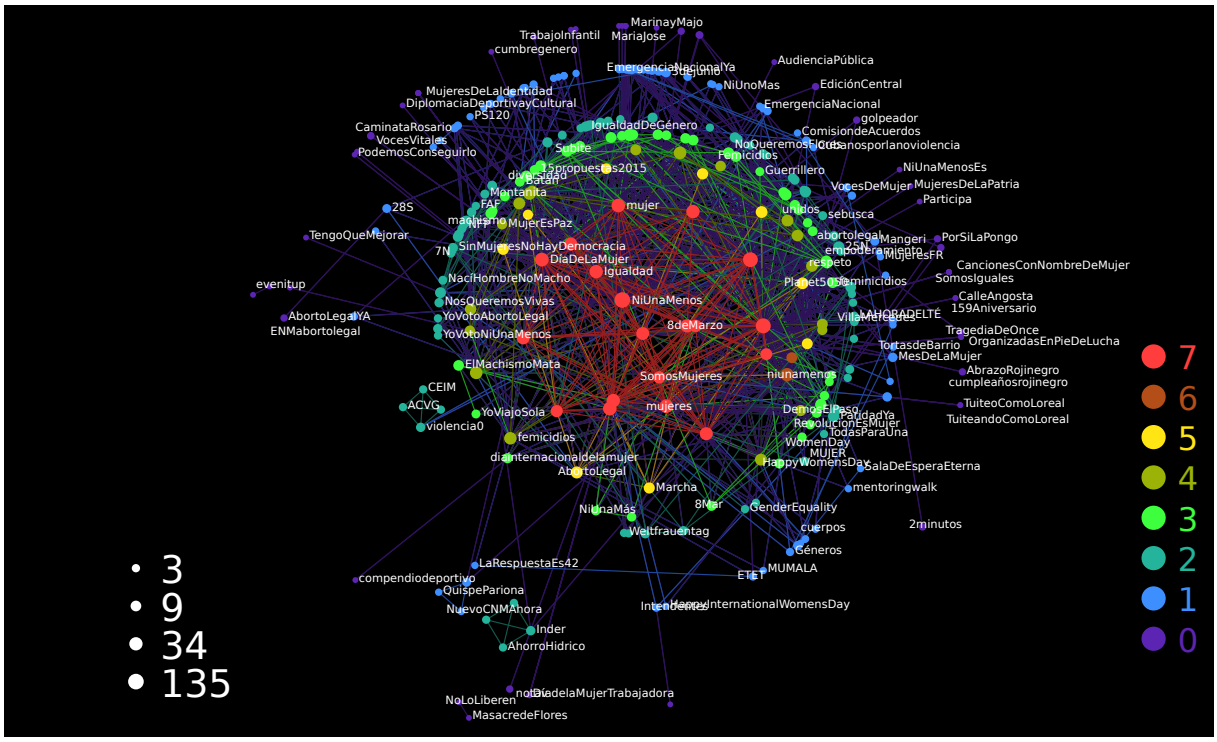


Figura 4.15: Tópico sobre feminismo. Se destaca el *hashtag* #NiUnaMenos.

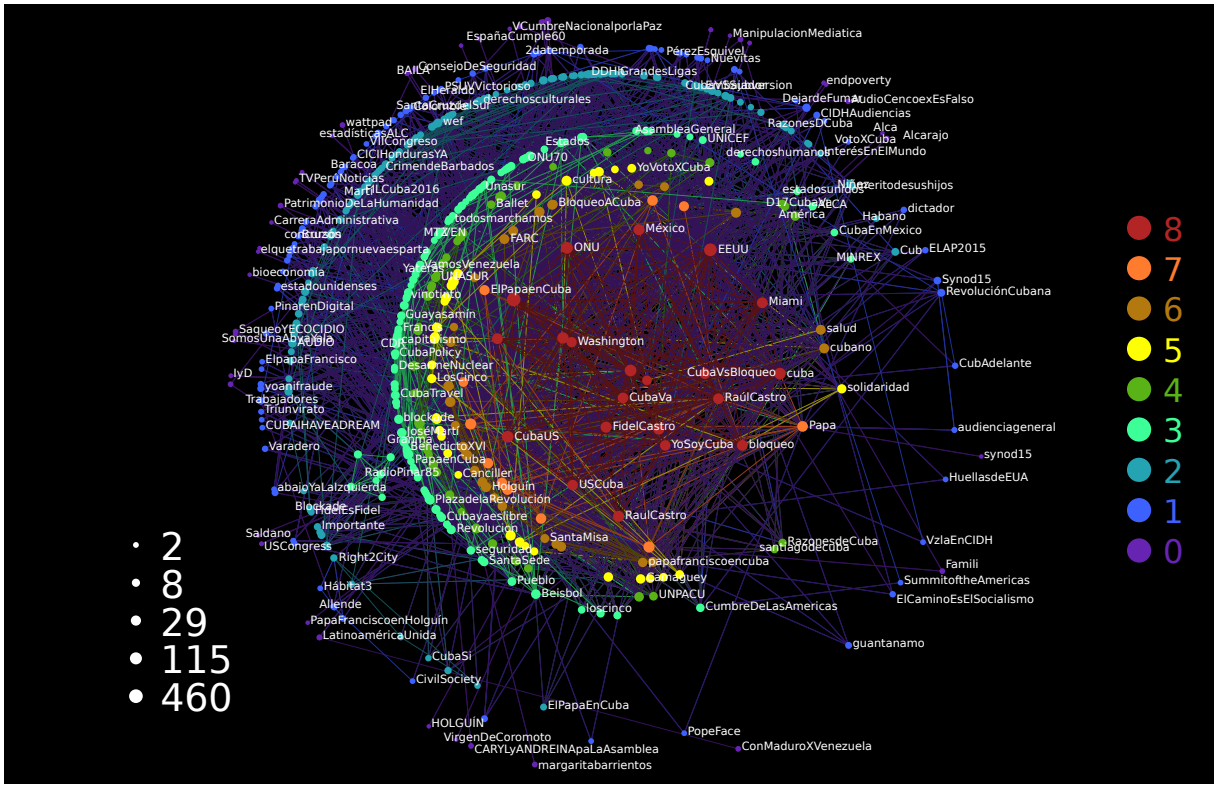


Figura 4.16: Tópico sobre Cuba.

En la figura 4.17 se encuentra un tópico sobre la premiación de los Óscar, un certamen

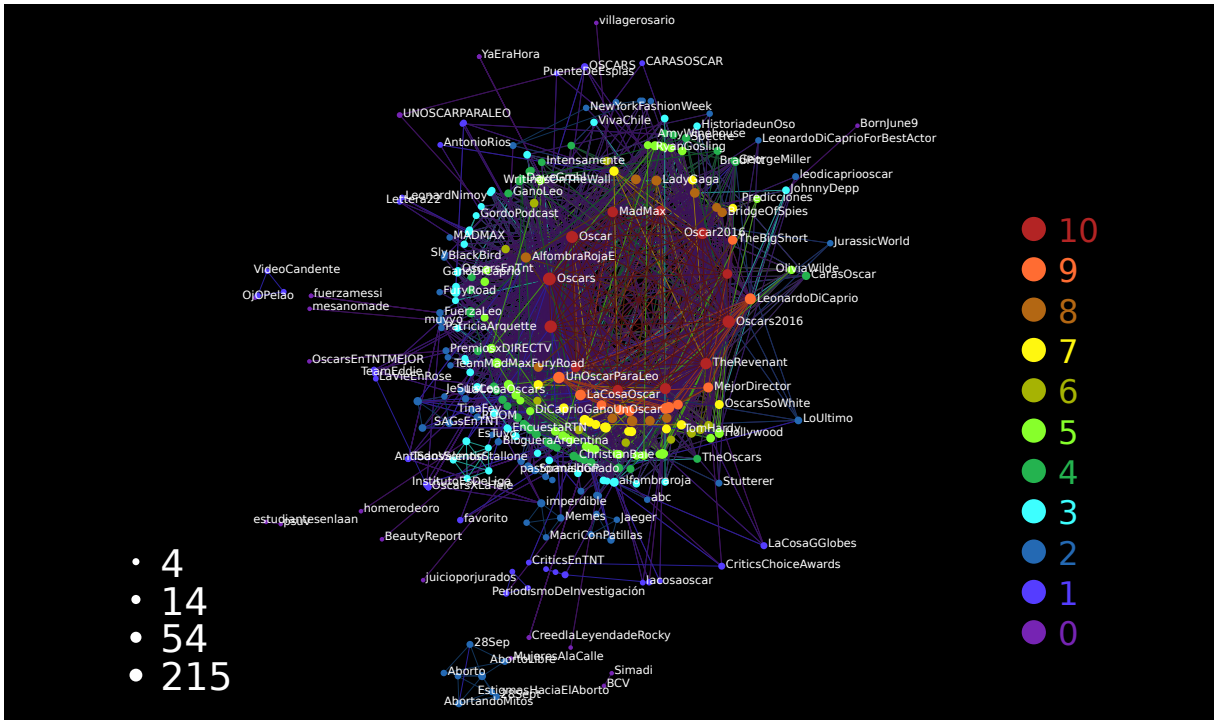


Figura 4.17: Tópico sobre celebridades y premios Óscar.

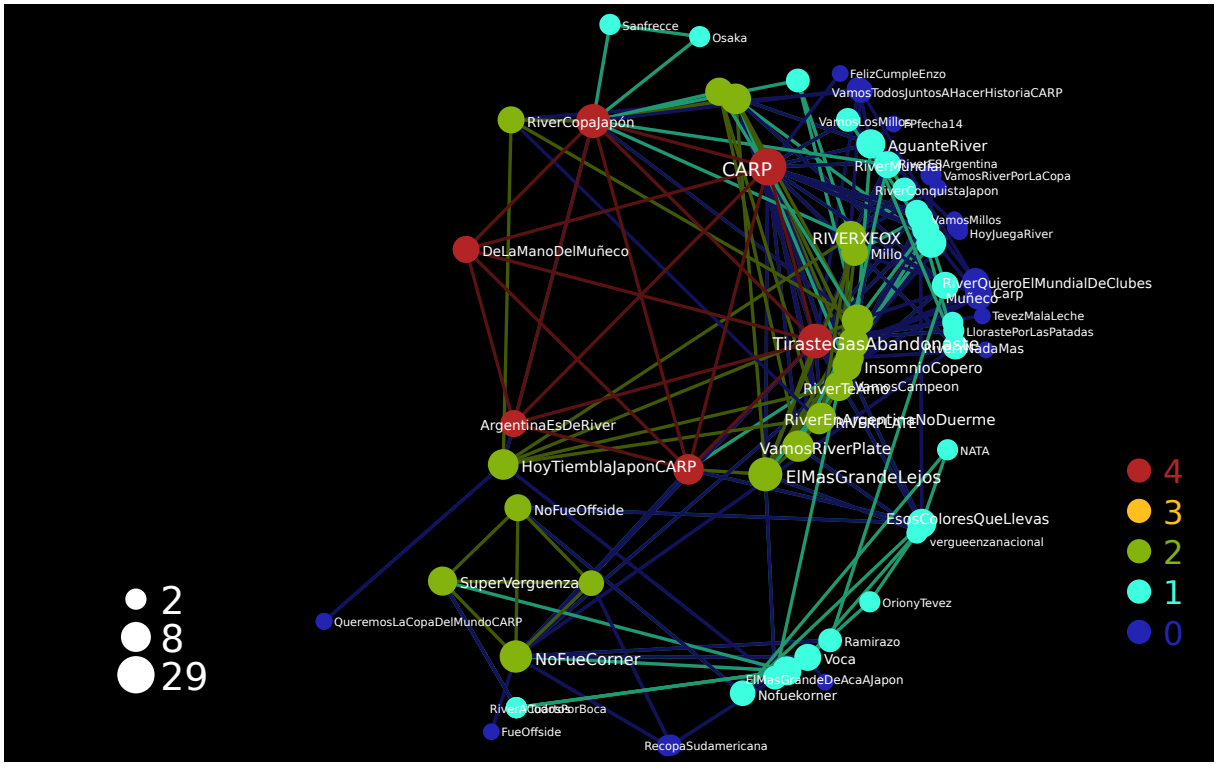


Figura 4.18: Tópico sobre el equipo de fútbol argentino River

anual que reconoce a las mejores películas. El actor Leonardo DiCaprio tiene un especial protagonismo en el tópico, dado luego de tantos años de actuar en películas importantes, ganó el premio Óscar a mejor actor.

En las figuras 4.18 y 4.19 se encuentran representados los dos clubes de fútbol más populares en Argentina: River Plate y Boca Juniors. El tópico de Boca Juniors presenta mayor cantidad de *hashtags* y conexiones entre los mismos, por lo que se puede deducir que los hinchas de este club tuvieron una mayor actividad en este conjunto de datos capturado que los hinchas de River.

Por último, hay dos tópicos fuera del contexto argentino. En la figura 4.20 observamos un tópico de debate internacional. En el centro se encuentran *hashtags* de países y de líderes mundiales. A medida que se aleja, se observan *hashtags* sobre temas humanitarios de interés mundial, por ejemplo #`Refugees`, #`BlackLivesMatter` y #`SyriaCrisis`. También están presentes *hashtags* que tratan sobre conflictos existentes en el mundo, principalmente en el Medio Oriente, entre ellos menciones a servicios de inteligencia y movimientos políticos.

En la figura 4.21 se puede observar un tópico sobre atentados terroristas sufridos en Francia. El centro del subgrafo hace referencia a los múltiples atentados sufridos el 13 de noviembre del 2015 en Francia. También está mencionado el atentado sufrido en el semanario satírico Charlie Hebdo el 12 de enero del mismo año.

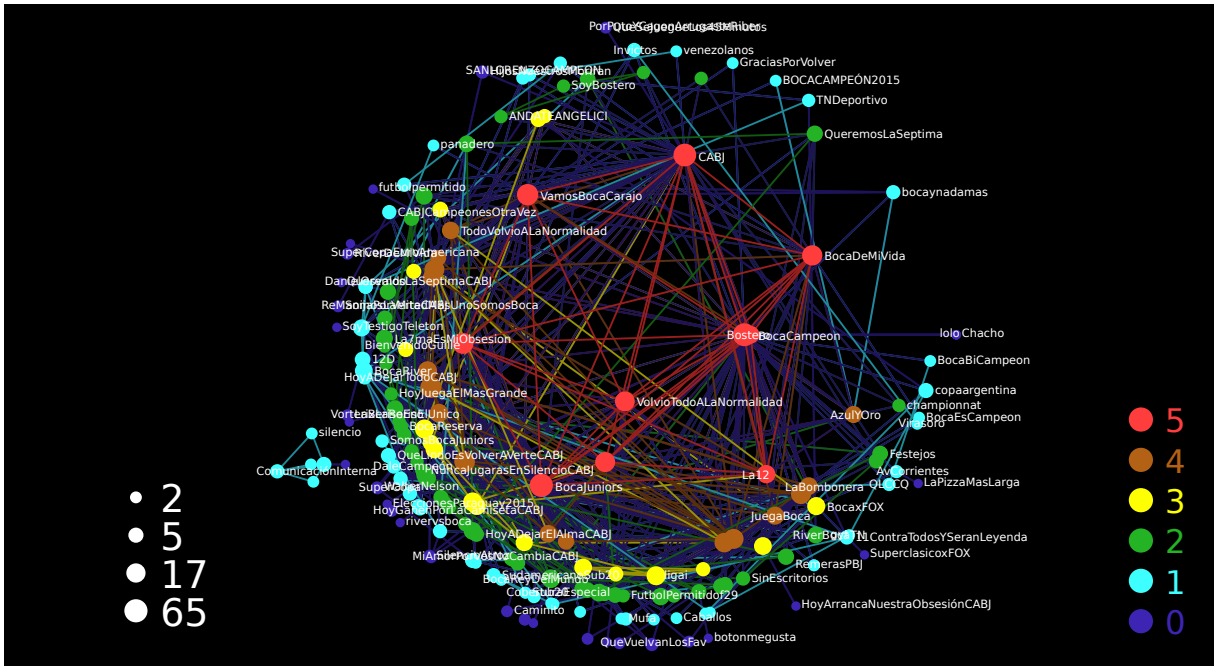


Figura 4.19: Tópico sobre el equipo de fútbol argentino Boca

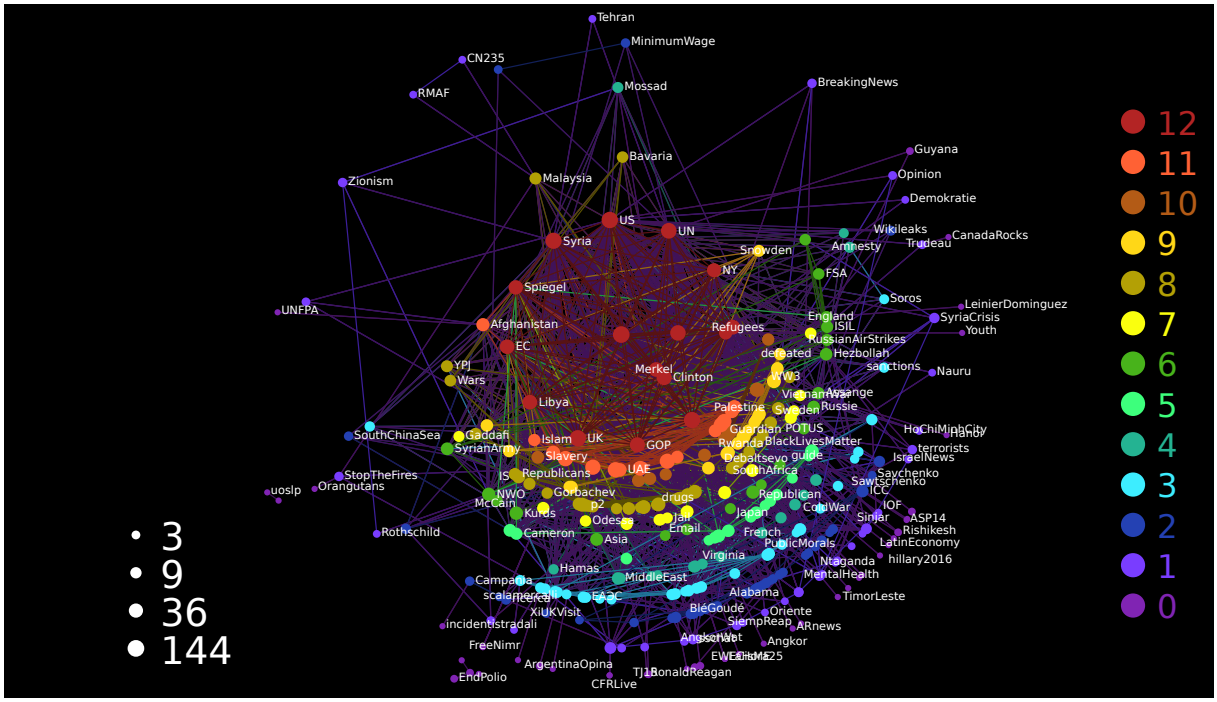


Figura 4.20: Tópico sobre política internacional, en el que se destacan temas como la guerra en Siria y las elecciones presidenciales en Estados Unidos, entre otros.

Los tópicos exhibidos muestran una gran cohesión entre los *hashtags* que los conforman. Al haber restringido a un único tópico los *hashtags* pertenecientes a múltiples comunidades, también es una contribución para que, por ejemplo, #`ScioliPresidente`

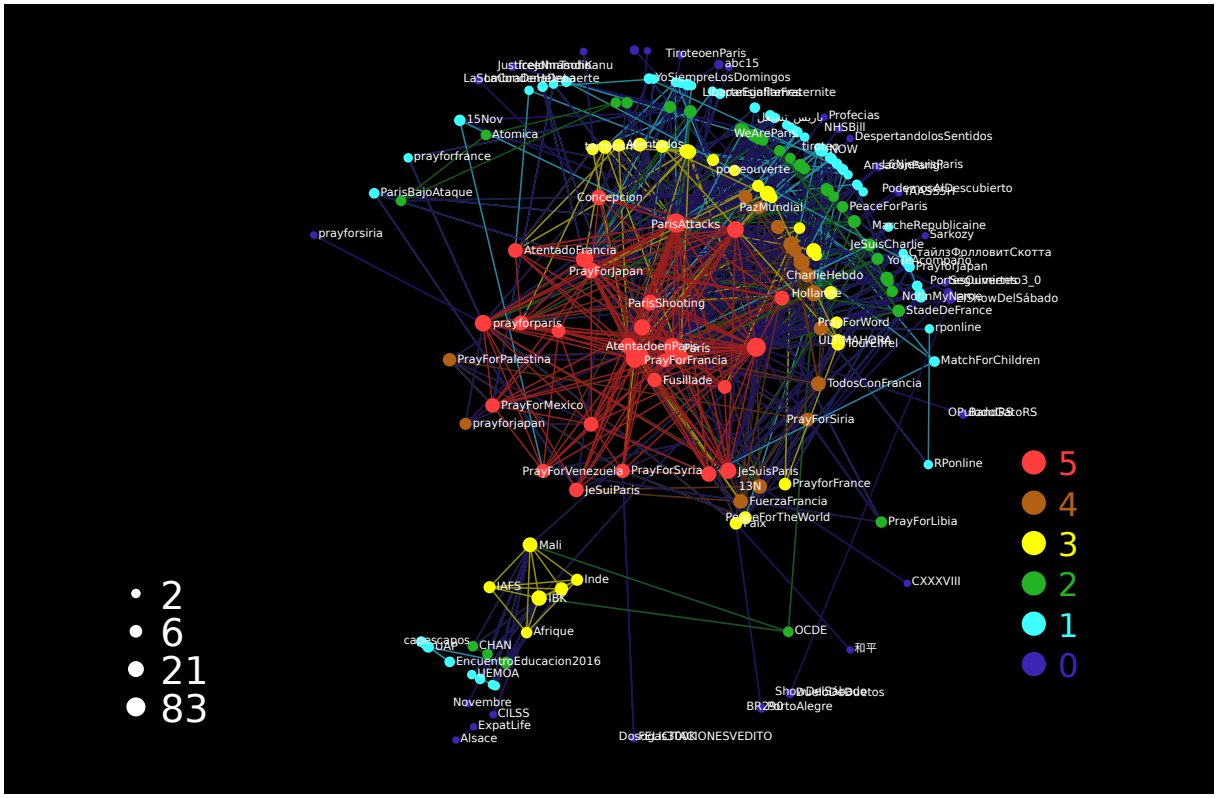


Figura 4.21: Tópico sobre atentados terroristas. En el núcleo principal del grafo se destacan los atentados sufridos el 13 de noviembre del 2015 en París.

---

no quede clasificado dentro del tópico de Macri.

4.4. Homofilia temática

Al haber clasificado los *hashtags* en tópicos, estamos en condiciones de analizar la cantidad de veces que habló un usuario sobre diversos temas y podemos representar a un usuario por su vector de tópicos. Con esta representación vectorial del usuario, analizamos si existe homofilia entre ellos (ver sección 2.2), utilizando como métrica la similitud coseno.

Un procesamiento previo que realizamos sobre la matriz de usuario-tópico fue aplicar TF-IDF para restar importancia a los tópicos que fueron tratados por todos los usuarios, y darle mayor relevancia a los tópicos que sólo fueron utilizados por ciertos usuarios. Utilizamos TF-IDF para discriminar términos relevantes en documentos, si consideramos a los usuarios como documentos y a los términos como tópicos, obtendremos una medida de que tan relevante es un tópico para cada usuario, de acuerdo también con la cantidad de veces que otros usuarios hayan utilizado el tópico.

El primer análisis que realizamos fue tomar la red de usuarios extraída de Twitter,

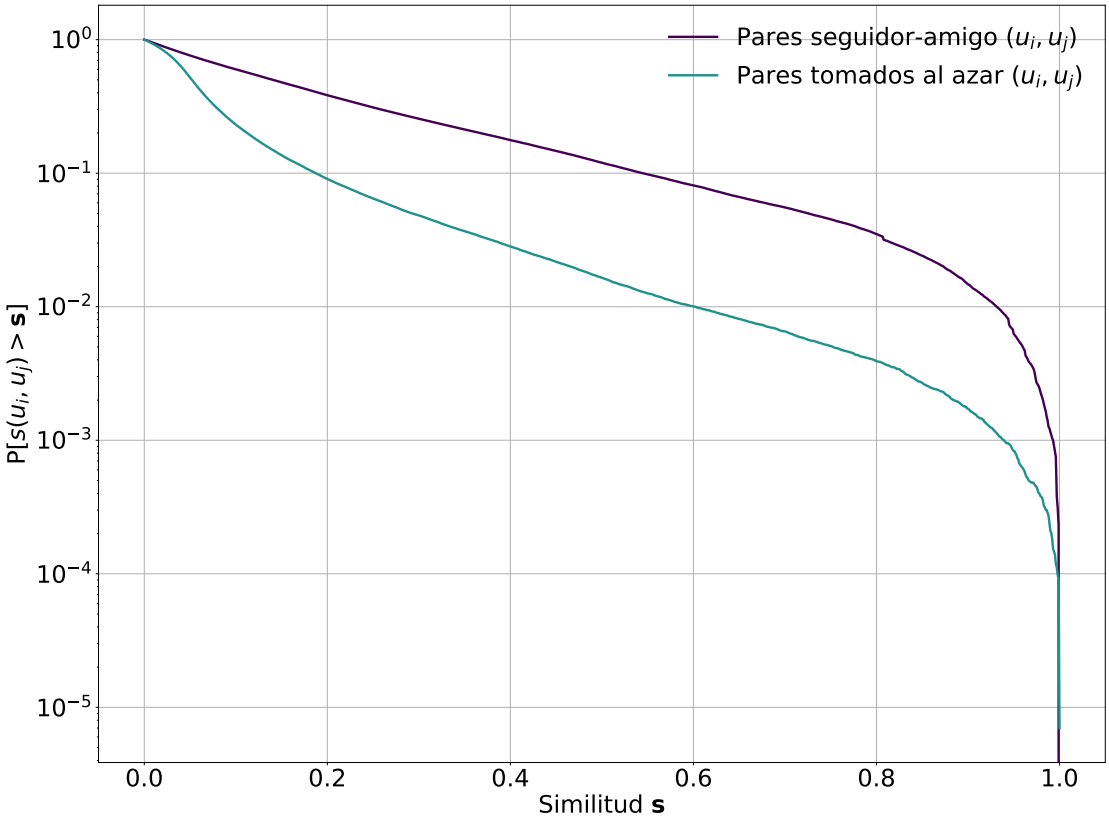


Figura 4.22: Distribución de probabilidad acumulada empírica inversa de la homofilia entre usuarios de la red de Twitter. Se comparan aquellos usuarios que tienen una relación de seguimiento contra usuarios tomados al azar.

---

y comparar la similitud entre usuarios que se siguen dentro de la red contra usuarios tomados al azar.

Para cada uno de los usuarios dentro de la red, computamos la similitud con cada una de las personas a las cuáles sigue, y obtenemos la mediana por cada uno de ellos. Posteriormente, para cada usuario dentro de la red, seleccionamos al azar la misma cantidad de personas a la que él sigue, computamos la similitud coseno y utilizamos el mismo estadístico sobre los valores. Así, para cada usuario tenemos la misma cantidad de valores tomados de la red de seguidores y tomados al azar.

En la figura 4.22 se puede observar que existe una homofilia marcada sobre los temas de debate que hay entre usuarios que se siguen entre ellos en comparación con usuarios tomados al azar y concluimos que efectivamente existe homofilia porque ambas curvas están claramente diferenciadas.

El siguiente análisis realizado fue determinar si existe homofilia entre los usuarios que siguen a un único candidato. De los pares de relación seguidor-amigo, restringimos a aquellos pares de seguimiento en los cuales ambos usuarios siguen únicamente a uno

de los cuatro candidatos. Luego, para cada uno de los pares computamos la similitud y distinguimos si siguen al mismo candidato o si siguen a candidatos distintos.

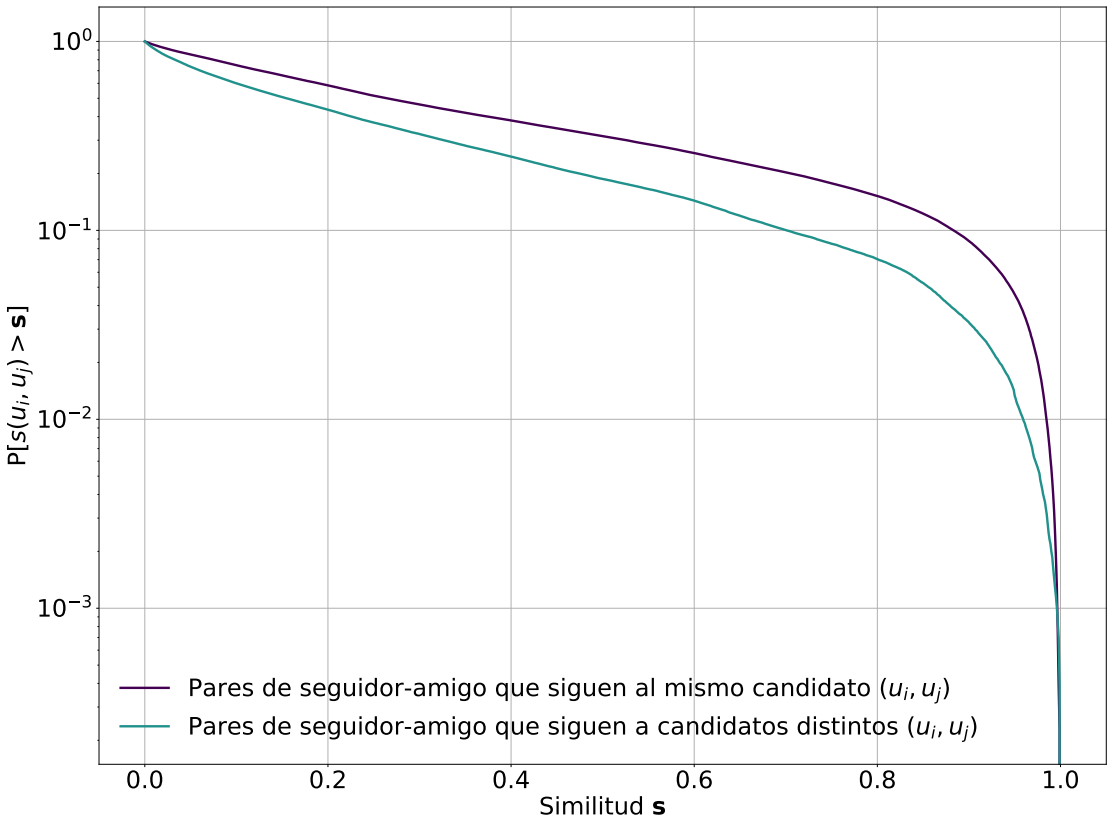


Figura 4.23: Distribución de probabilidad acumulada empírica inversa de la homofilia entre usuarios que siguen a un candidato. Se comparan aquellos usuarios que siguen al mismo candidato contra usuarios que siguen a candidatos distintos.

En la figura 4.23 observamos una mayor homofilia entre personas que siguen a un mismo candidato en comparación con personas que siguen a candidatos distintos. Si bien existe mayor homofilia para aquellas personas que siguen al mismo candidato, esta diferencia es menor a la observada en la figura 4.22.

4.5. Evolución política de los usuarios

Teniendo a disposición la información de que un usuario habló sobre cierto tema en un determinado momento, realizamos la evolución de la utilización de tópicos a través del tiempo. En la sección anterior hicimos un análisis de similitud en todo el período de captura de datos, ahora procedemos a construir matrices usuario-tópico para lapsos de 10 días, en la cual analizamos los temas que hablaron los usuarios en esa ventana y haremos la progresión en el tiempo. Al igual que en la sección anterior, para cada

una de estas matrices usuario tópico dentro de la ventana temporal aplicamos TF-IDF para disminuir el peso de los tópicos más comunes entre los usuarios, pero en este caso sólo estamos disminuyendo el peso de los *hashtags* dentro de la ventana de tiempo bajo análisis. Como ya hemos considerado antes los grupos de políticos, consideramos usuarios que siguen únicamente a una cuenta política de Twitter. Es decir, de los siete supernodos capturados, cuatro de ellos fueron candidatos a presidente en 2015: Macri, Scioli, Massa y Stolbizer, y analizamos aquellos usuarios que únicamente siguen a uno de estos cuatro políticos. Además de filtrar usuarios, también filtraremos tópicos, dado que hay algunos que fueron utilizados con una menor frecuencia. Para ello, buscamos la mayor cantidad de veces que fue utilizado un tópico, y a dicha cantidad la dividimos por un factor, para obtener un umbral de utilización mínima de tópicos. De los 3131 tópicos detectados, sólo 124 tienen una cantidad de utilización superior a dicho umbral, y son los que utilizaremos para medir similitud entre usuarios.

Para cada uno de los grupos de seguidores de políticos, medimos la similitud de un usuario con sus pares que siguen al mismo político, de esta forma tenemos una medida de homofilia para los seguidores de los candidatos y como cada uno de estos cuatro grupos fueron evolucionando a través del tiempo. Para compararlos, obtenemos una muestra de usuarios al azar y computamos la similitud que existe entre ellos para tener una referencia promedio dentro de la red. A continuación presentamos un listado de las acciones realizadas:

- Eliminamos los tópicos que fueron utilizados menos de tres veces. De los 3131 originales, quedan 1396 tópicos.
- Definimos un mínimo de utilización de tópicos con el cual eliminamos aquellos que hayan sido utilizados por debajo de dicho umbral. El valor del umbral es el cociente entre la cantidad de usos del tópico más utilizado dividido por un factor de 128: $umbral = \frac{\text{máx}\{\text{Uso tópico}\}}{128}$
- Generamos matrices usuario-tópico de acuerdo a las ventanas de tiempo de 10 días.
- Para cada una de las matrices, aplicamos TF-IDF sobre la matriz de 10 días.
- Para los cuatro candidatos, filtramos de cada una de las matrices aquellos usuarios que sólo siguen a un candidato, generando para cada una de ellas cuatro matrices con los usuarios filtrados. Adicionalmente, se calcula una quinta matriz para usuarios tomados al azar.
- Para cada matriz correspondiente a una ventana de tiempo, y para las cinco matrices derivadas de haber filtrado usuarios, se hace el producto matricial de la matriz

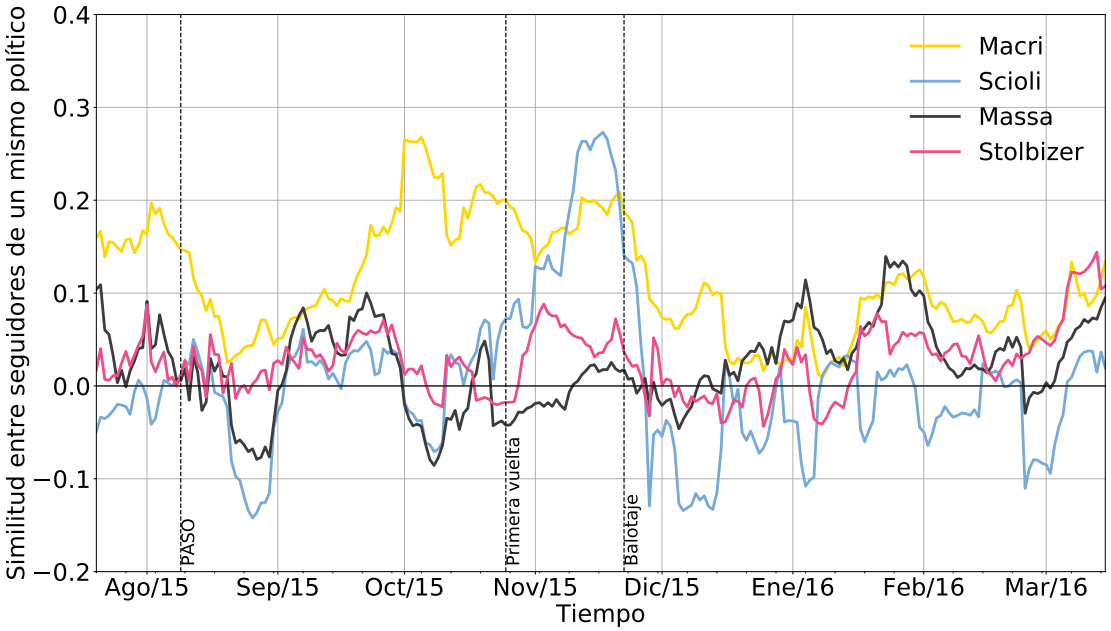


Figura 4.24: Media de similitud a través del tiempo de la diferencia para cada uno de los seguidores de candidatos dentro de su grupo y la similitud promedio en la red.

---

con su traspuesta para obtener una comparación intra-grupo. En cambio si se quieren comparar dos grupos de seguidores, se realiza el producto de la matriz de un candidato por la traspuesta de la matriz del otro candidato.

- Aplicamos la media sobre la porción triangular superior de la matriz (aquellos elementos $a_{i,j} : i < j$).
- Restamos esa media obtenida con la media de realizar el producto de la matriz de usuarios tomados al azar con su traspuesta.
- Graficamos.

Los valores de similitud están dentro del intervalo $[-1, 1]$ dado que se resta la similitud de un grupo con la similitud promedio dentro de la red. Los valores de similitud positivos indican que hablan de los mismos temas, por ende son símiles. En cambio, valores negativos indican que hablan de temas distintos, por ende son disímiles.

El siguiente gráfico muestra la similitud para cada uno de los grupos políticos menos la similitud promedio de la red.

En la figura 4.24 graficamos la similitud que hay de usuarios que siguen a un candidato, respecto a la similitud promedio que hay entre los usuarios de la red. Se puede observar que a medida que se aproxima la elección en la que participa un candidato, la similitud de los seguidores de un candidato tiende a crecer entre ellos.

Para realizar comparaciones entre grupos de seguidores, computamos la similitud entre dos grupos de usuarios, calculamos el promedio y luego restamos por la similitud promedio de la red.

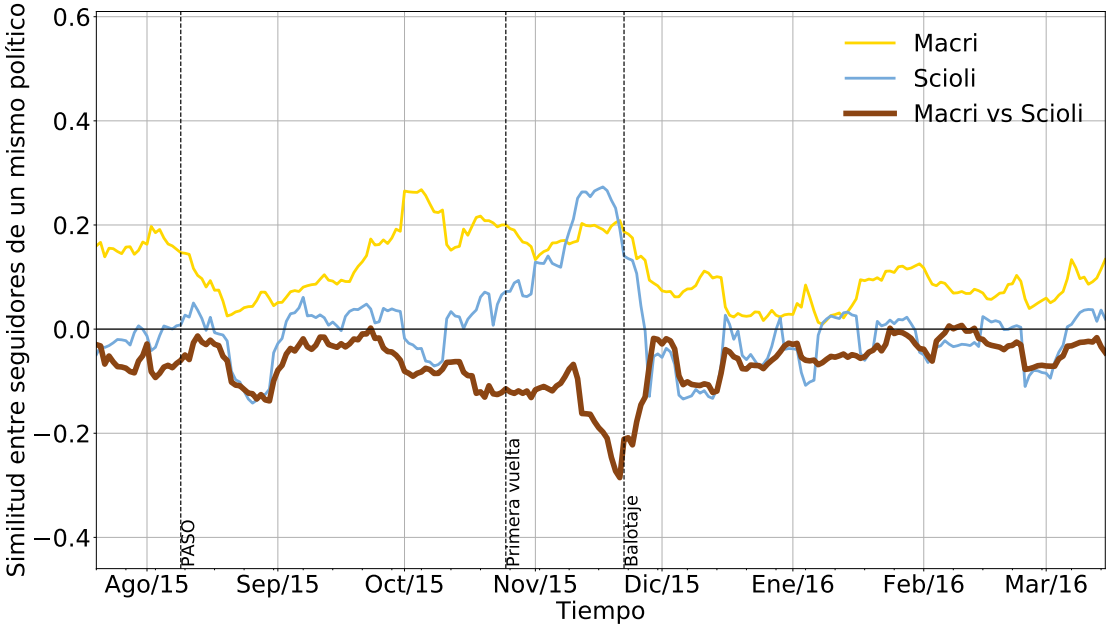


Figura 4.25: Media de similitud a través del tiempo de la diferencia para Macri (amarillo), Scioli (celeste), Macri versus Scioli (marrón) y la similitud promedio en la red.

En la figura 4.25 graficamos la similitud que existe entre seguidores de Macri respecto de la similitud promedio, así como también está graficado para Scioli. La tercer curva es la similitud promedio que hay entre seguidores de Macri contra seguidores de Scioli, respecto también de la similitud promedio. En la tercer curva, se observan valores por debajo de cero, llegando a un mínimo previo a la segunda vuelta de las elecciones. La curva de los seguidores de Macri muestra que la similitud de sus seguidores varía, con un máximo antes de la primer vuelta y luego se mantiene hasta el balotaje. En el caso de Scioli, hay partes con valores negativos (por debajo de la similitud al azar), y un fuerte crecimiento entre la primera vuelta y el balotaje.

La figura 4.26 muestra la similitud que existe entre los seguidores de Massa y Stolbizer contra los seguidores de los candidatos que pasaron a la segunda vuelta (Macri y Scioli). En particular, Massa tiene una similitud negativa con Macri antes de la primer vuelta, que luego se vuelve positiva y se incrementa hasta el balotaje. Esta tendencia explicaría en parte el triunfo de Macri en el balotaje, ya que Massa representó aproximadamente el 21 % del electorado. El caso de los seguidores de Stolbizer, siempre están alineados con Macri, con un máximo entre la primera vuelta y el balotaje. Ambos grupos de seguidores,

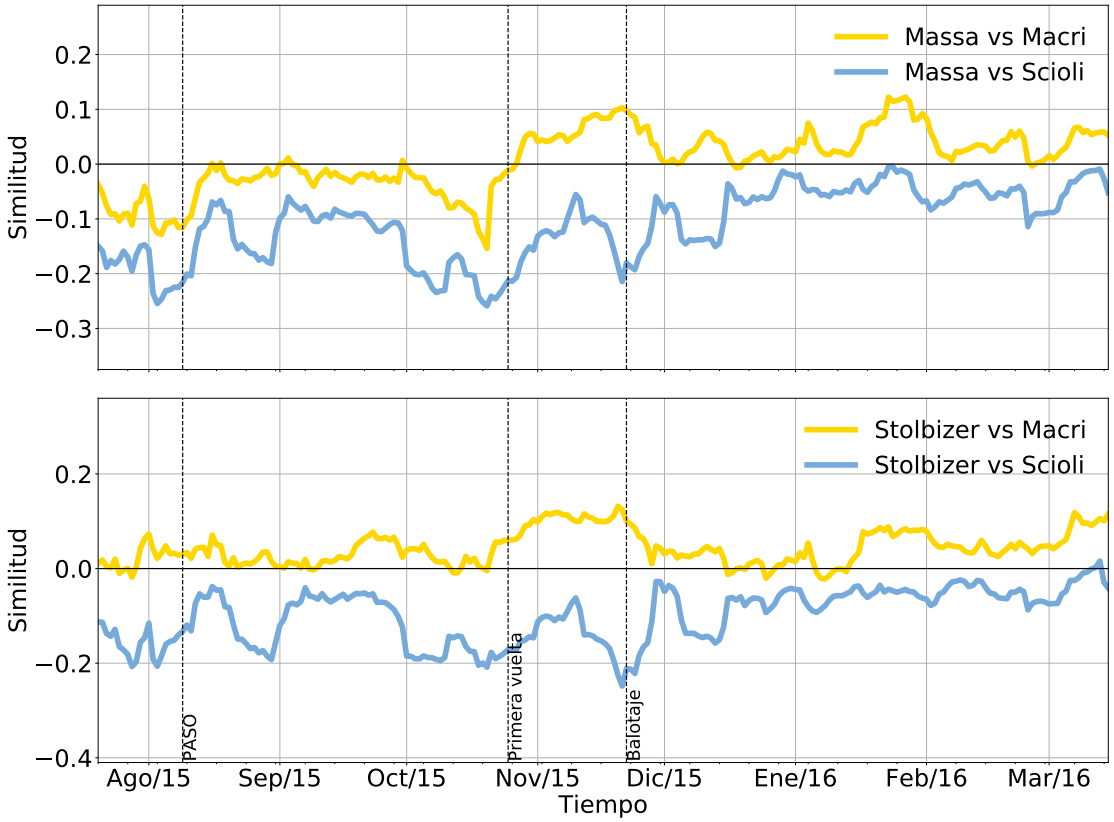


Figura 4.26: Media de similitud a través del tiempo de la diferencia entre los candidatos alternativos versus Macri y Scioli, con respecto a la similitud promedio de la red.

---

los de Massa y Stolbizer, mantienen una similitud negativa con Scioli para todo el período bajo análisis. Concluimos que los seguidores de candidatos eliminados se parecen mucho más a Macri que a Scioli en el período entero de captura, teniendo valores de similitud próximos al máximo cerca de la fecha del balotaje.

La comparación entre seguidores de Scioli contra seguidores de Massa que realizamos en la figura 4.27 es para observar en detalle la diferencia de discurso que existe entre ambos grupos. Observamos que la similitud intra grupos es mayor que la similitud entre los dos grupos. Hay varios picos negativos, tres de los cuales están próximos a los tres días de votación (PASO, primera y segunda vuelta).

La comparación de la figura 4.28 que realizamos entre los seguidores de Massa y Stolbizer es para confirmar el comportamiento observado en la figura 4.26 , dado que la comparación realizada no fue directa entre los candidatos. Aquí la tendencia es clara, los tópicos sobre los cuales hablan los seguidores de ambos candidatos son muy distintos previo a la primera vuelta de las elecciones. Luego, su discurso se alinea, convergiendo en la utilización de tópicos, y por ello el aumento de similitud entre ambos grupos.

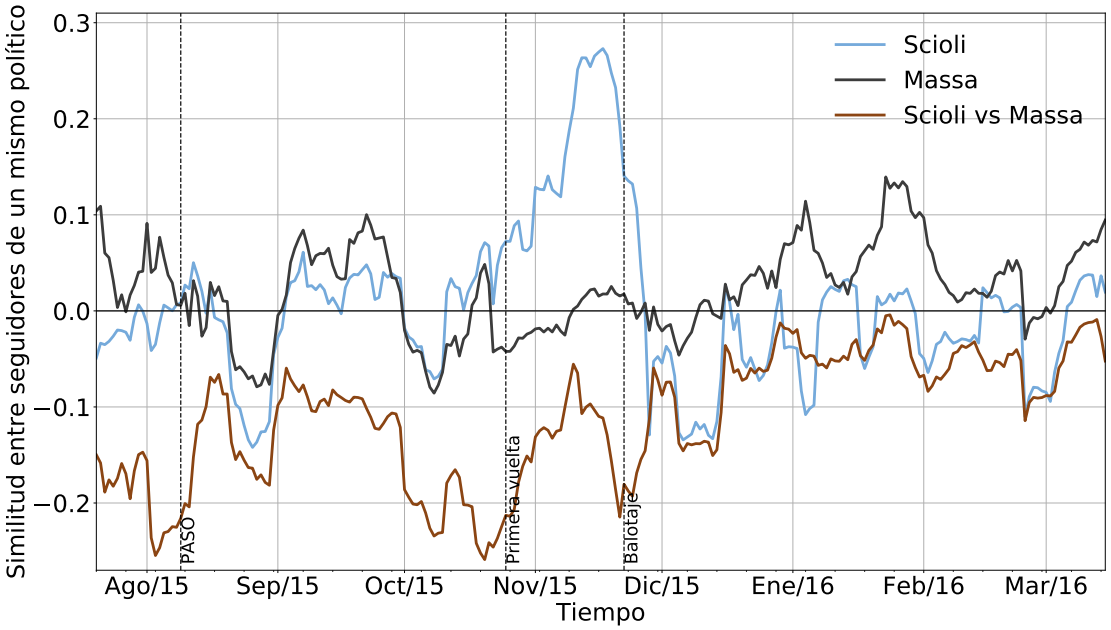


Figura 4.27: Media de similitud a través del tiempo de la diferencia para Scioli (celeste), Massa (negro), Scioli versus Massa (marrón) y la similitud promedio en la red.

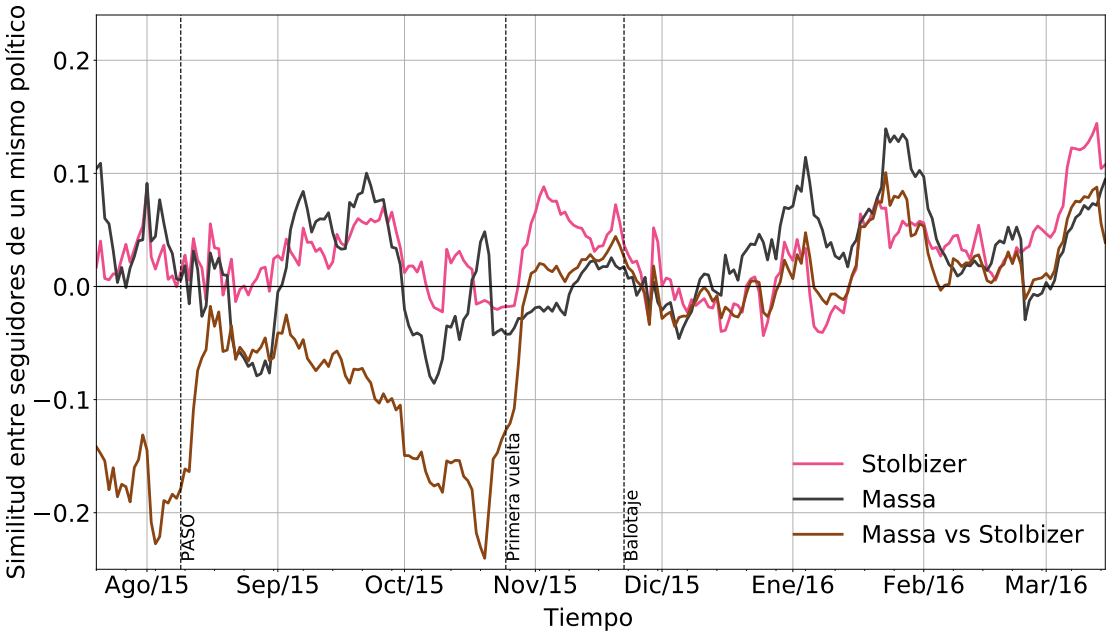


Figura 4.28: Media de similitud a través del tiempo de la diferencia para Massa (negro), Stolbizer (violeta), Massa versus Stolbizer (marrón) y la similitud promedio en la red.

4.5.1. Análisis de la evolución política

El objetivo de esta sección fue analizar la similitud de ciertos usuarios en base a la utilización por su parte de los tópicos detectados. Observamos que las fechas de las elecciones (PASO, primera y segunda vuelta) coinciden con los momentos de mayor amplitud en la fluctuación de la similitud de los usuarios, así como también, entre la primera y segunda vuelta se observa la alineación de tendencias entre los usuarios. El discurso entre los seguidores de Macri y Scioli es muy disímil, encontrando una máxima diferencia previo al enfrentamiento en las urnas en la segunda vuelta, como se observa en la figura 4.25. Además, entre la primera y la segunda vuelta se observa como los discursos de los seguidores de Massa y Stolbizer se asemejan al discurso de los seguidores de Macri, y no al discurso de los seguidores de Scioli.

4.6. Evolución temporal de tópicos

De los tópicos detectados por OSLOM, nos centramos en estudiar algunos en especial, entre ellos los más utilizados así como también aplicamos el mismo concepto de distancia entre distribuciones de probabilidad (divergencia de Kullback-Leibler) para seleccionar los tópicos más específicos dentro de las comunidades de usuarios. También veremos los tópicos donde predomina la utilización de uno de los cuatro grupos de seguidores políticos, lo que podría indicar interés por parte de los seguidores sobre un tema particular. Lo que se observa es la cantidad de veces que alguno de los usuarios que siguen exclusivamente a un candidato utilizaron algún *hashtag* de un tópico y realizamos un promedio.

En la figura 4.29 se observa la utilización del tópico de apoyo hacia Macri. Era esperable que predominen en la utilización los seguidores de Macri. También observamos como después de la primera vuelta, los seguidores de Massa y Stolbizer comienzan a utilizar más este tópico. Así como determinamos que existía una mayor afinidad entre los seguidores de Macri y Stolbizer, como se puede observar en la figura 4.26 de la sección anterior, aquí se ve reflejado en el incremento de usos del tópico por parte de los seguidores de Stolbizer.

En la figura 4.30 se observa la utilización del tópico de Scioli. Hay una clara preponderancia por parte de los seguidores de Scioli en el uso. A diferencia de lo que ocurre en el tópico de Macri, los seguidores de los otros tres candidatos casi no utilizan el tópico en todo el período.

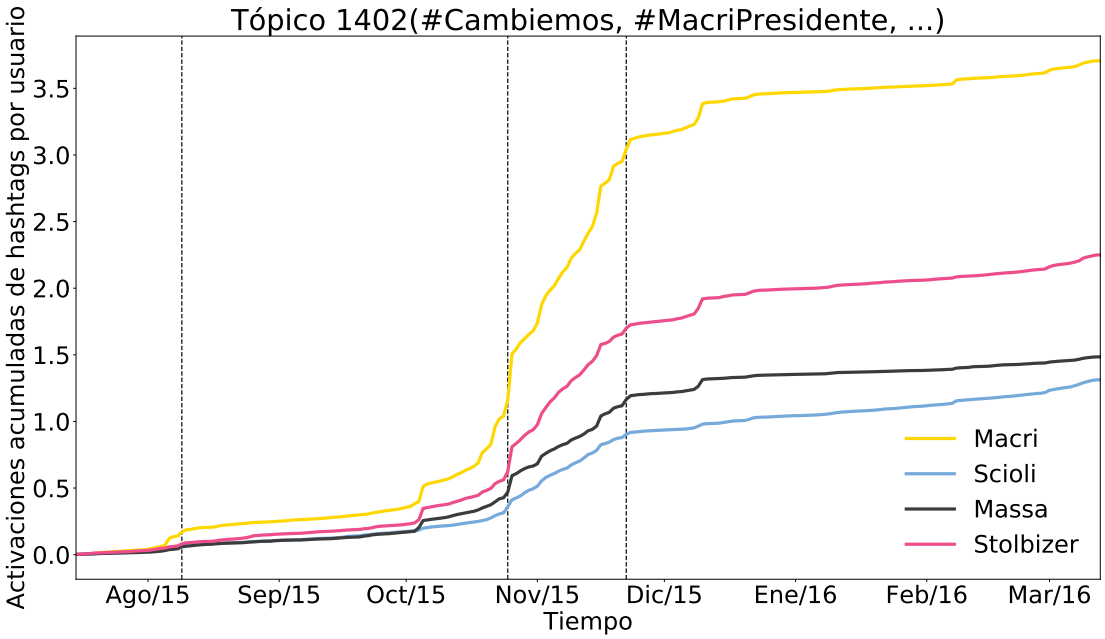


Figura 4.29: Evolución temporal promedio por usuario de la utilización del tópico de Macri.

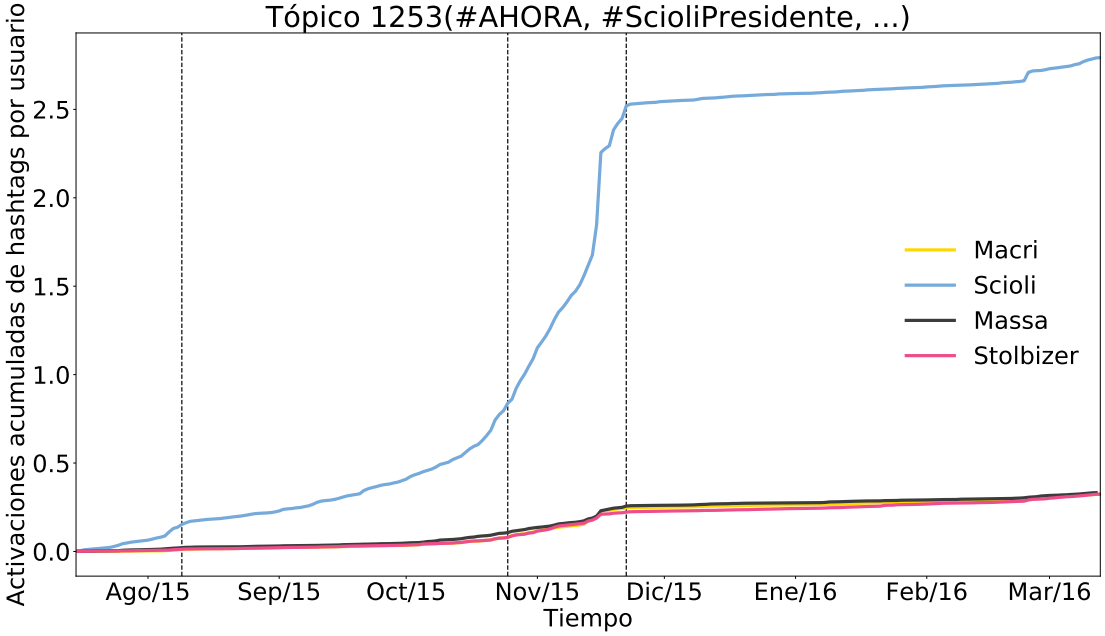


Figura 4.30: Evolución temporal promedio por usuario de la utilización del tópico de Scioli.

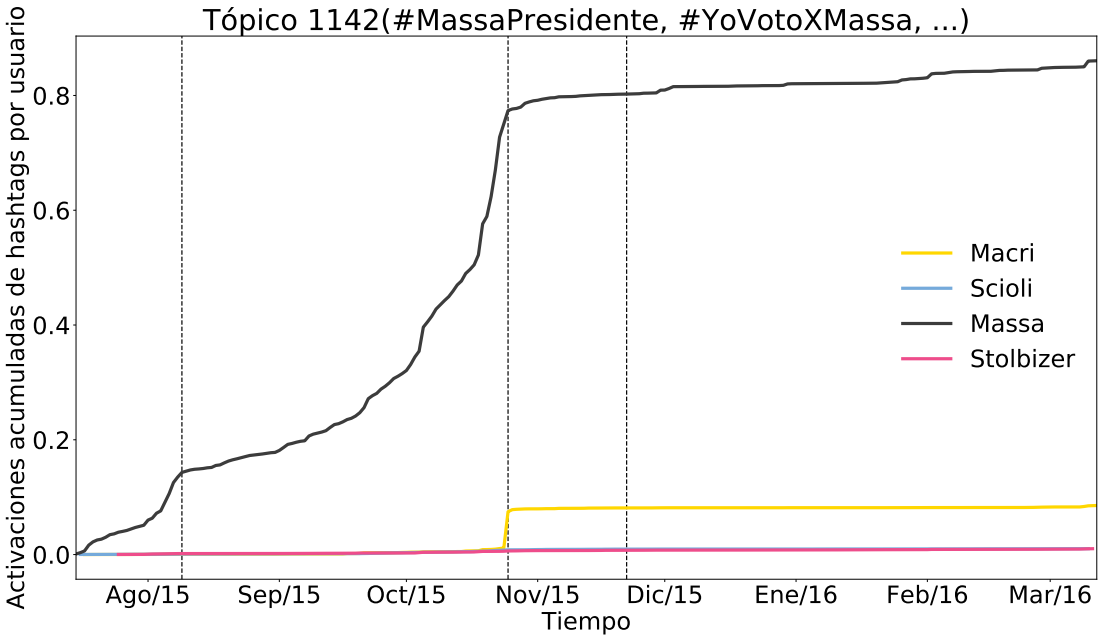


Figura 4.31: Evolución temporal promedio por usuario de la utilización del tópico de Massa.

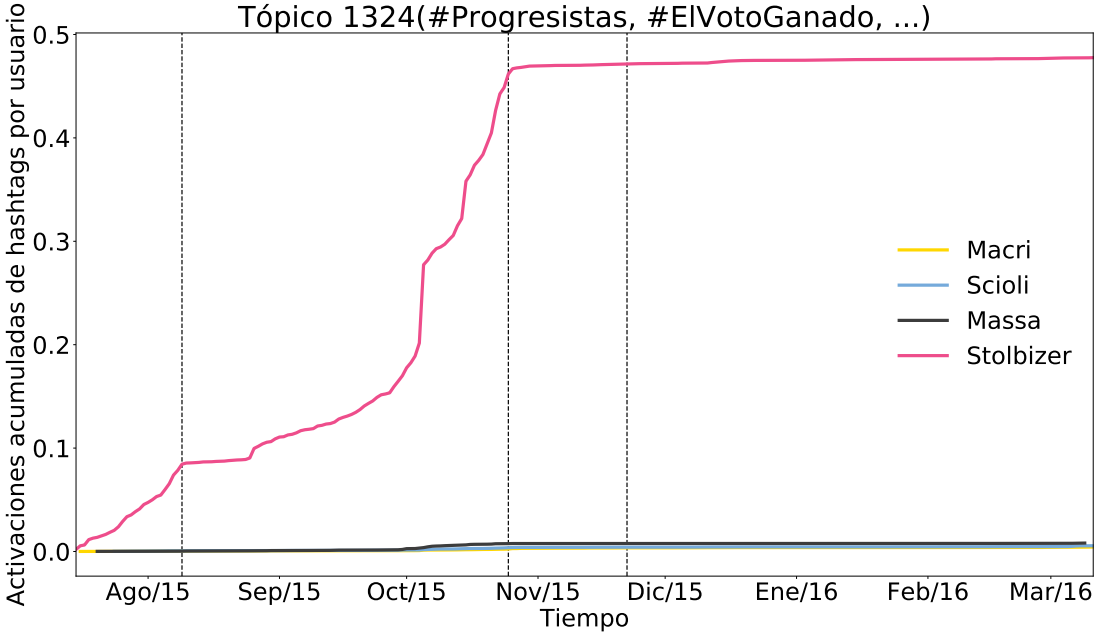


Figura 4.32: Evolución temporal promedio por usuario de la utilización del tópico de Stolbizer.

En las figuras 4.31 y 4.32 observamos el uso de los tópicos de Massa y Stolbizer respectivamente. Como ocurrió con los otros candidatos, el mayor uso lo realizan las personas que siguen al candidato. En estos casos, a diferencia de lo que ocurrió en el tópico de Scioli y sus seguidores en la figura 4.30, se observa que el crecimiento en el uso del tópico se da antes de la primera vuelta para luego estancarse. En cambio en el tópico de Scioli, el crecimiento se da entre la primera vuelta y el balotaje, con su posterior estancamiento.

La primera conclusión que se puede obtener de esta sección es una confirmación a lo que ya afirmamos en la sección 4.5.1 sobre la inclinación de votos de los seguidores de Massa y Stolbizer. La evidencia que hay en la figura 4.29 es confirmar dicha tendencia hacia Macri, dado que luego de la primera vuelta al quedar eliminados sus candidatos, se observa como los seguidores de Massa y Stolbizer comienzan a utilizar más intensivamente el tópico de Macri, efecto más atenuado en la figura 4.30 que corresponde al tópico de Scioli.

En la figura 4.33 se puede observar un tópico como se mencionó en la sección 4.3 relacionado al movimiento feminista, donde se destacan los seguidores de Stolbizer en la utilización del mismo y revelan los intereses que tienen en común con la agenda política de Stolbizer [3].

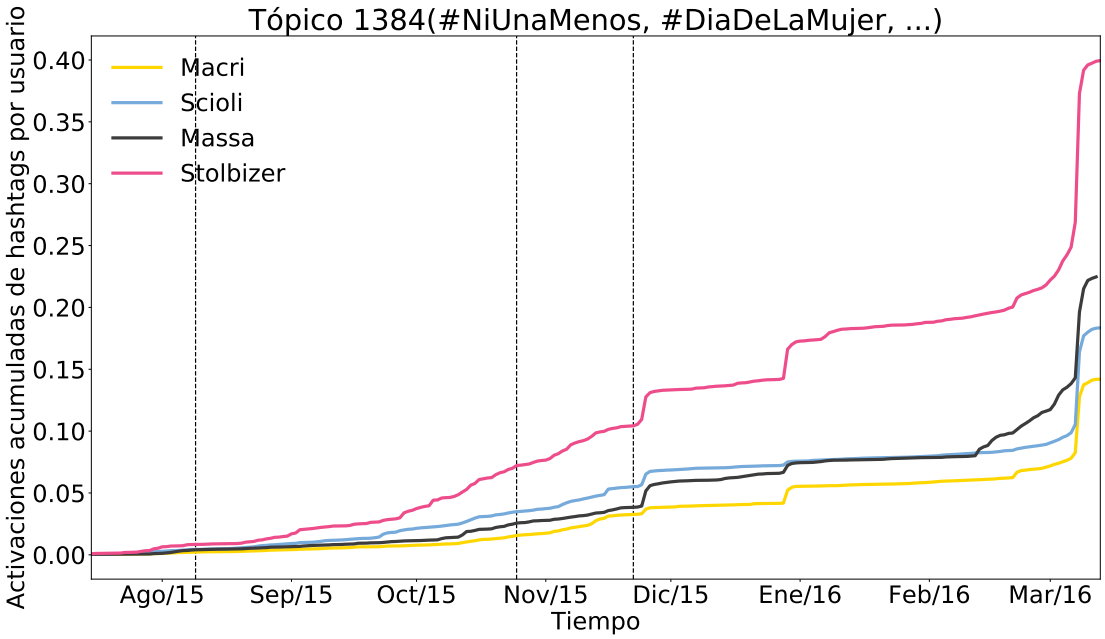


Figura 4.33: Evolución temporal promedio por usuario de la utilización del tópico sobre cuestiones de género.

[3] https://www.partidogen.org.ar/wp-content/uploads/2015/01/propuestas_para_laqq_vector.pdf

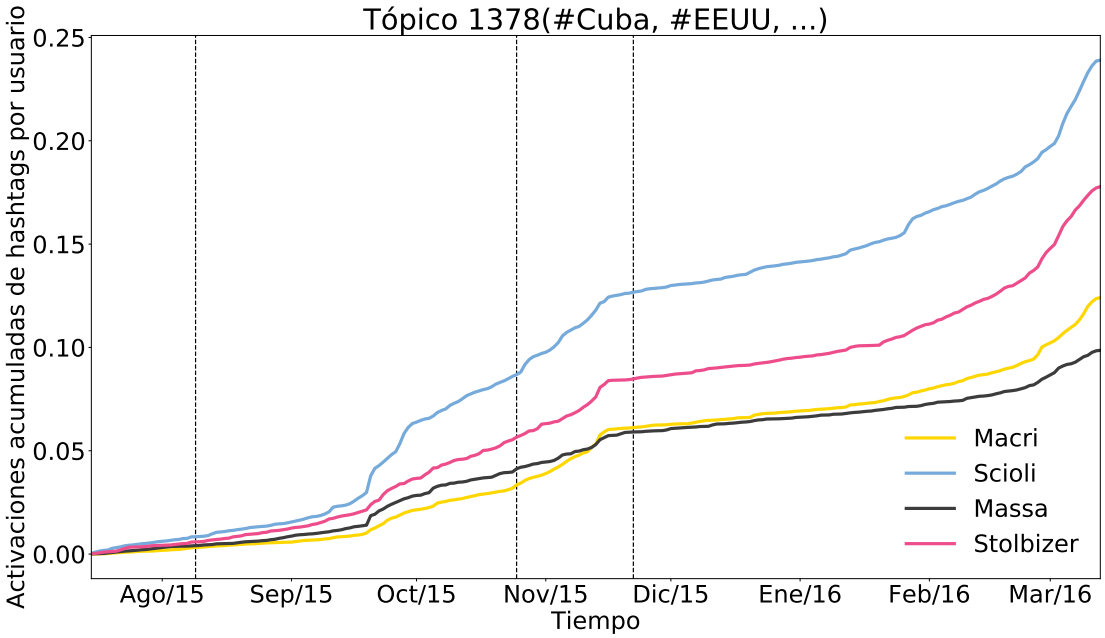


Figura 4.34: Evolución temporal promedio por usuario del tópico sobre Cuba.

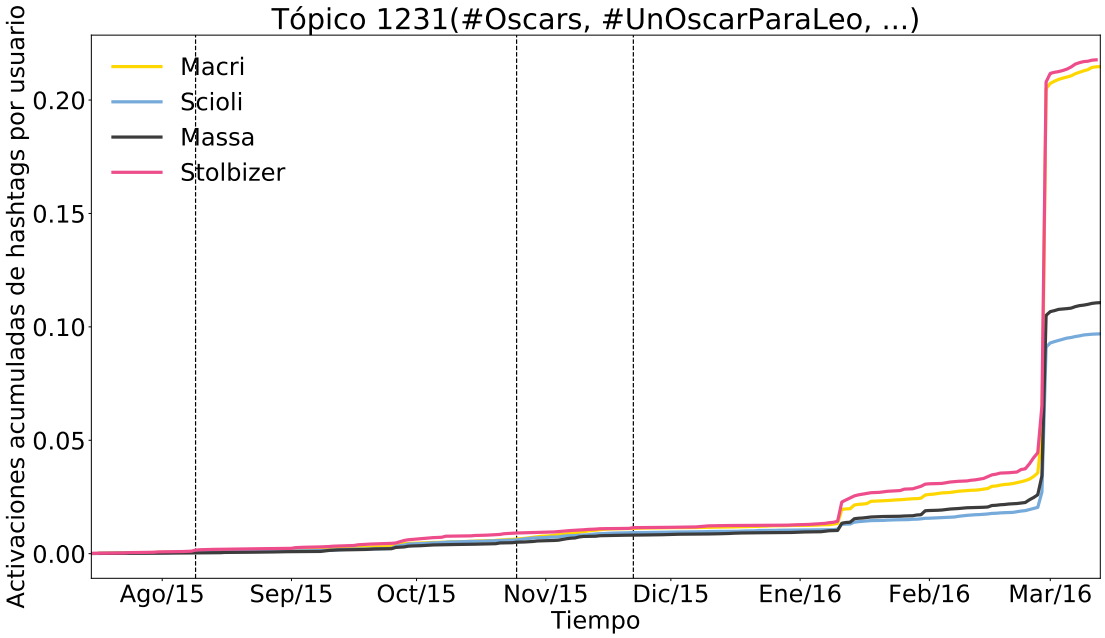


Figura 4.35: Evolución temporal promedio por usuario de la utilización del tópico de los premios Óscars.

En la figura 4.34 se observa que ciertos *hashtags* fueron utilizados siguiendo la misma dinámica (se activan en las mismas fechas); aunque también existe una preponderancia por parte de los seguidores de Scioli en su uso.

En la figura 4.35 se observa que hubo una participación baja en promedio por parte de los seguidores sobre los premios Óscar, por lo que no hay relevancia sobre analizar la tendencia de seguidores políticos. El uso mayoritario lo tienen los seguidores de Stolbizer, pero por una mínima diferencia. Es un tópico que no tiene impacto sobre la agenda política en Argentina.

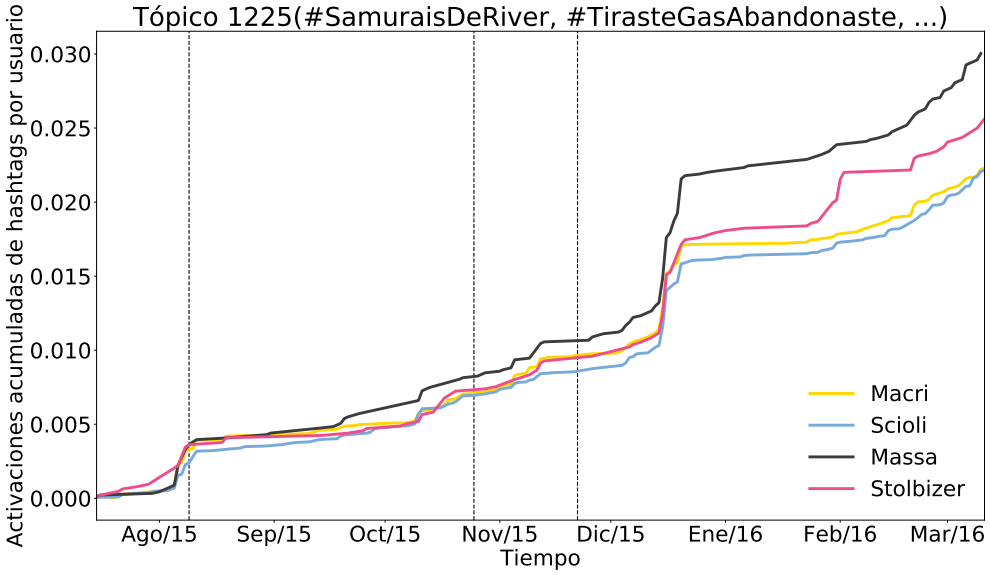


Figura 4.36: Evolución temporal promedio por usuario de la utilización del tópico del equipo de fútbol argentino River Plate.

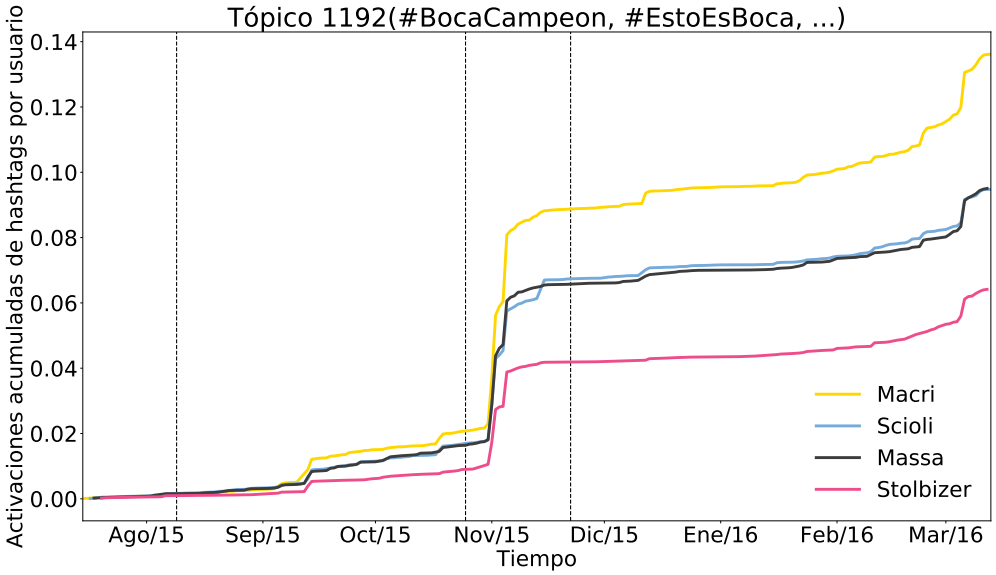


Figura 4.37: Evolución temporal promedio por usuario de la utilización del tópico del equipo de fútbol argentino Boca Juniors.

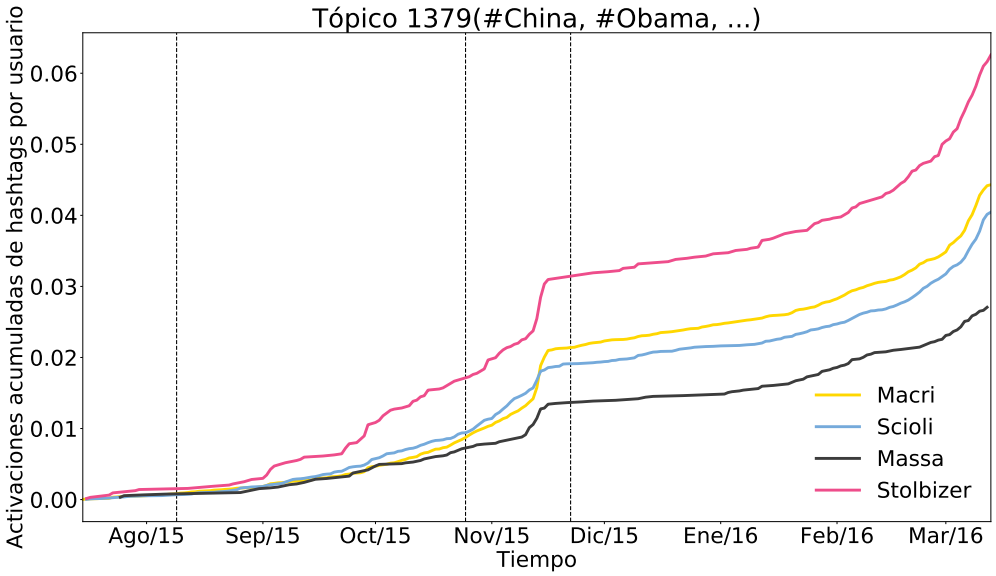


Figura 4.38: Evolución temporal promedio por usuario de la utilización del tópico sobre política internacional.

De las figuras 4.36 y 4.37 se puede observar que en comparación el tópico de Boca Juniors fue utilizado en promedio una mayor cantidad de veces, y esto se condice con la comparación de los *k-densos* de los gráficos 4.18 y 4.19, donde 4.19 tiene un mayor *k-denso*. También existe una concentración en el uso de ambos tópicos que probablemente tengan relación con eventos deportivos. Por ejemplo, en diciembre de 2015 el equipo River Plate viajó a Japón para jugar la copa del mundial de clubes FIFA, por lo que ese evento explicaría la cantidad de usos en 4.36 en el mes de diciembre. A diferencia, Boca Juniors sólo jugó el campeonato de fútbol local, por eso no tiene la misma cantidad de usos en diciembre y el mayor crecimiento en usos se registra en noviembre, mientras todavía se está disputando el torneo local.

En la figura 4.38 los seguidores de Stolbizer se destacan en la utilización de este tópico. Han utilizado en promedio mayor cantidad de veces el tópico, incluso siendo el grupo minoritario, por lo que agrega un interés especial en este tópico.

En la figura 4.39 se puede apreciar notablemente el incremento en la utilización del tópico los días posteriores al ataque sufrido el 13 de noviembre. También se exhibe un mayor interés por parte de los seguidores de Macri y Stolbizer, en menor medida por los seguidores de Massa y por último los seguidores de Scioli.

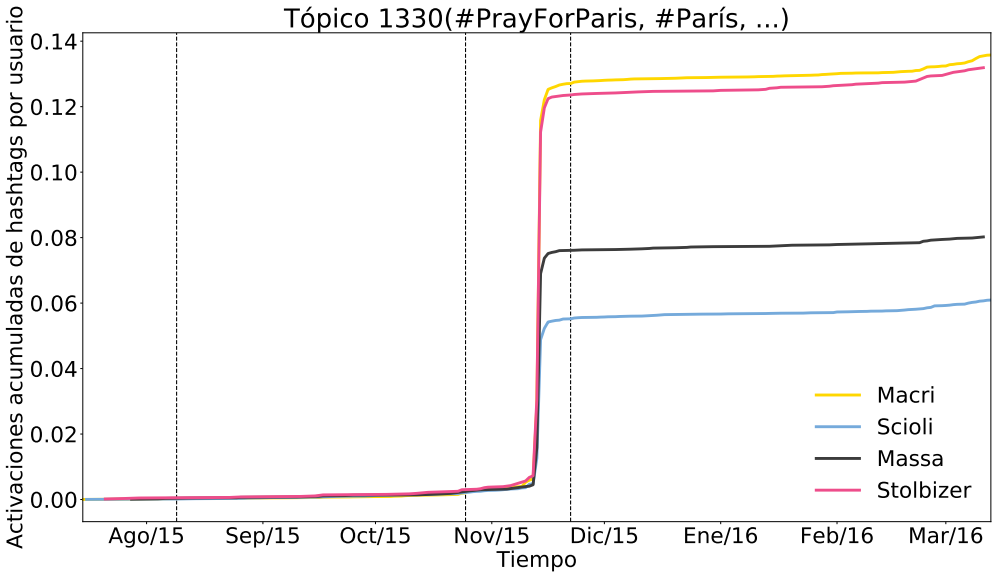


Figura 4.39: Evolución temporal promedio por usuario de la utilización del tópico sobre atentados terroristas.

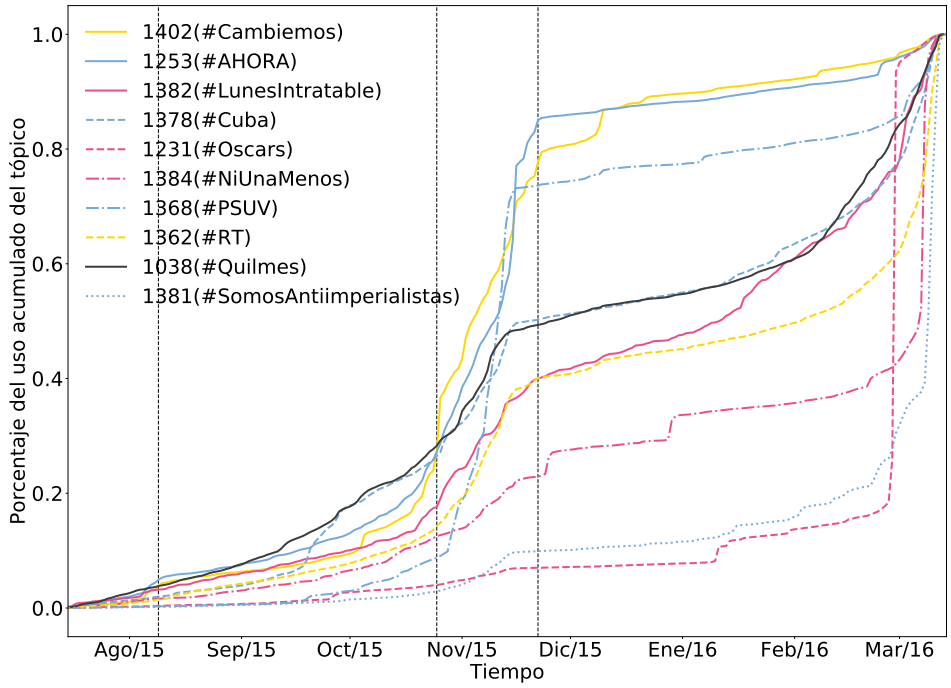


Figura 4.40: Probabilidad acumulada de la utilización de los diez tópicos más utilizados de la red. Cada tópico se encuentra coloreado de acuerdo a los seguidores de un candidato predominante.

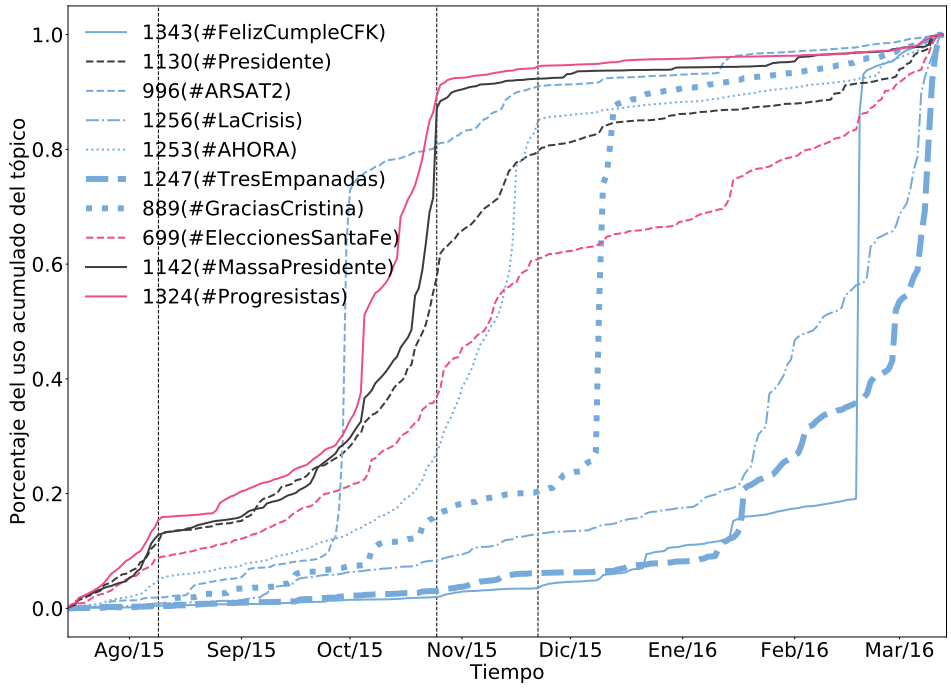


Figura 4.41: Probabilidad acumulada de la utilización de los diez tópicos más específicos por comunidad de la red. Cada tópico se encuentra coloreado de acuerdo a seguidores de un candidato predominante.

Finalmente, observamos el interés mayoritario de ciertos tópicos en los seguidores políticos, así como en la figura 4.40 como ya mencionamos se destaca la utilización de un tópico feminista los seguidores de Stolbizer, como el interés de los seguidores de Scioli en el tópico cuyo *hashtag* principal es #`SomosAntiImperialistas`. En la figura 4.41 están los tópicos de apoyo político hacia Massa y Stolbizer, así como también se encuentra un tópico de apoyo a Cristina Kirchner por parte de los seguidores de Scioli donde se destaca el *hashtag* #`GraciasCristina`.

4.7. Predicción de usuarios con intereses políticos

Por último, en esta sección aplicaremos herramientas de aprendizaje supervisado para clasificar usuarios identificados como seguidores de sólo un candidato. Utilizaremos dos representaciones vectoriales de usuarios para entrenar y clasificarlos, una de ellas es la representación vectorial de los tópicos utilizados por el usuario (ver secciones 4.3 y 4.4). La otra representación se obtiene del vector de *hashtags* que utilizó un usuario y posteriormente aplicar un algoritmo denominado LDA (Blei *et al.*, 2003). LDA es un modelo probabilístico generativo para colecciones de datos discretos. Describe a cada elemento de la colección como una combinación lineal de un conjunto de tópicos subyacente.

De esta forma, se representa a un usuario como la combinación lineal de los *hashtags* que utilizó, reduciendo la dimensión del vector del usuario a la misma dimensión que la representación de tópicos.

Para determinar la clase de los usuarios utilizaremos un clasificador denominado Bosque Aleatorio (*Random Forest*) desarrollado por Breiman (2001). El algoritmo consiste en generar un conjunto de árboles de decisión que crean distintas reglas para separa las muestras a clasificar. Los nodos se siguen dividiendo dentro del árbol hasta obtener nodos puros, es decir, las muestras que cumplen todas las condiciones dentro de una ramificación del árbol son de la misma clase.

Para validar y evaluar la clasificación realizada, utilizamos la técnica de validación cruzada (*cross validation*) en la cual se subdivide el conjunto de muestras que disponemos, una parte se utiliza para entrenar al modelo y la parte restante se utiliza para clasificar y evaluar la efectividad, dado que conocemos las clases de las muestras que estamos clasificando. Como las cantidad de muestras por clase es desequilibrada (hay una proporción mucho más grande de usuarios afines a Macri y Scioli en comparación con Massa y Stolbizer), utilizamos un método de partición denominado *Stratified K-fold*, el cual separa el conjunto de datos en entrenamiento y verificación asegurando que dentro de ambas particiones exista la misma proporción de muestras pertenecientes a cada una de las clases. Estos conceptos están desarrollados en detalle en Hastie *et al.* (2005).

Los parámetros utilizados para el clasificador fueron seleccionados utilizando el conjunto de entrenamiento y mediante la metodología de grilla de búsqueda (*grid search*), la cual consiste en probar todas las combinaciones de parámetros del algoritmo que se deseen, y el conjunto de datos a clasificar con sus clases. Este método informa la combinación de parámetros para la cual se obtuvo el mejor puntaje de clasificación en base a una métrica de referencia. En nuestro caso la métrica utilizada fue la exactitud ponderada (Brodersen *et al.*, 2010) sobre la clase de la muestra. A continuación listamos los parámetros que exploramos del algoritmo de Bosque Aleatorio:

- Cantidad de estimadores: cantidad de árboles de decisión que debe construir el bosque.
- Máxima profundidad: altura máxima que puede alcanzar el árbol al separar muestras dentro de sus nodos.
- Mínima cantidad de muestras de división: cantidad mínima de muestras en un nodo para realizar una división en nodos hijos.
- Mínima cantidad de muestras en nodo hoja: cantidad mínima de muestras requeridas para estar en un nodo hoja. Se realizará una división del nodo si deja al menos

la cantidad mínima de muestras en los nodos de las ramas izquierda y derecha.

- Máxima cantidad de características (*max features*): es la máxima cantidad de características a utilizar para crear una regla y realizar las división en los nodos.

- Máxima cantidad de nodos hoja: se determina una máxima cantidad de nodos hoja. Para la poda, se seleccionan los mejores nodos de acuerdo a una métrica de reducción relativa de impureza.

Cuadro 4.3: Parámetros del clasificador explorados. En negrita se destaca el valor seleccionado

Características	Parámetro	Valores explorados/ **seleccionados**
Tópicos	Cantidad de estimadores	[10, 100, **1000**]
	Máx profundidad	[**Ninguna**, 5, 8, 15, 25, 30]
	Mín muestras div	[**2**, 15, 100]
	Mín muestras hoja	[**1**, 10, 50]
	Máx *feat.*	[1, **10**, 100]
	Máx nodos hoja	[Ninguna, 100, **1000**]
LDA	Cantidad de estimadores	[10, 100, **1000**]
	Máx profundidad	[**Ninguna**, 5, 8, 15, 25, 30]
	Mín muestras div	[2, **15**, 100]
	Mín muestras hoja	[**1**, 10, 50]
	Máx *feat.*	[1, **10**, 100]
	Máx nodos hoja	[**Ninguna**, 100, 1000]

En el cuadro 4.3 se presentan los parámetros explorados para el clasificador. Los conjuntos de datos corresponden a las matrices de tópicos y reducción de dimensiones con LDA para datos dentro del período electoral completo. Luego de la exploración, para cada uno de los cuatro conjuntos de datos construimos sendos clasificadores y repetimos el método de validación cruzada.

Cuadro 4.4: Clasificación de usuarios que siguen a un único candidato.

Lapso	Características (*Features*)	Exactitud (*Accuracy*)	Exactitud ponderada (*Balanced accuracy*)
Hasta primera vuelta	Tópicos	84 %	77 %
Hasta primera vuelta	LDA	85 %	78 %
Periodo electoral completo	Tópicos	84 %	67 %
Periodo electoral completo	LDA	84 %	66 %

En el cuadro 4.4 se encuentran los resultados de la clasificación con validación cruzada. Al comparar los lapsos, observamos que la clasificación presenta mejores resultados limitando los datos a la primera vuelta, con una diferencia de aproximadamente diez puntos porcentuales en la exactitud ponderada. Esta diferencia es importante porque la exactitud ponderada nos indica que con datos tomados hasta la primera vuelta hay una mayor efectividad en la clasificación de usuarios de las clases minoritarias.

A pesar del decrecimiento de la exactitud ponderada en la comparación entre lapsos, la exactitud se mantiene constante. Esto puede suceder por una variación en la cantidad de usuarios entre la primera vuelta y el período completo. También pueden haber confusiones entre las clases minoritarias y mayoritarias luego de la primera vuelta, dado que los seguidores de Massa y Stolbizer comenzaron a utilizar los tópicos de Macri y Scioli.

Se observa que para un lapso determinado, al comparar las dos metodologías de extracción de características casi no hay diferencia. Para ambas métricas la diferencia es de un punto porcentual.

Por último, utilizamos matrices de confusión para ver la efectividad del Bosque Aleatorio en la clasificación. Una matriz de confusión es una disposición específica de una tabla, donde las distintas clases de la muestra aparecen como filas de la matriz. A su vez, para cada muestra el algoritmo realiza una predicción y la encasilla en una clase, que aparecen como columnas. La celda $c_{i,j}$ es una proporción del total de muestras que pertenecen a la clase i y fueron clasificadas como clase j. Si se suman los elementos de la matriz por cada una de las filas i, se obtiene el 100 % de la cantidad de muestras de la clase i (más menos un error de dígitos). Los verdaderos positivos se encuentran dentro de la diagonal principal, donde i es igual a j. En cambio, cuando i es distinto a j, la celda representa la cantidad de falsos positivos para la clase i generados por la clase j.

De las matrices de confusión del período electoral completo en la figura 4.43 podemos observar que los usuarios que *retweetearon* contenido de Massa y Stolbizer fueron confundidos por el algoritmo y clasificados como seguidores de Macri. Esta confusión es

consistente con lo desarrollado en secciones anteriores, los votos de Massa y Stolbizer se volcaron hacia Macri, por ello queda una proporción importante de seguidores que son clasificados como tales. En cambio, las matrices de confusión en la figura 4.42, con datos tomados hasta la primera vuelta electoral, exhiben una mejora importante en la clasificación de las clases minoritarias. Dado el lapso considerado, los usuarios que *retweetearon* contenido de Massa y Stolbizer estaban con sus candidatos vigentes, por lo tanto no habrían compartido tanto contenido en apoyo a Macri, por lo que menores proporciones de usuarios fueron confundidos.

Como mencionamos previamente, hay una mayor efectividad en la métrica de exactitud total, pero una peor efectividad en la de exactitud ponderada. Las matrices de confusión confirman que para el período electoral completo crece la efectividad en la clasificación de usuarios de Macri, pero también crece el error que se comete al clasificar usuarios que *retweetearon* contenido de los otros tres candidatos. Esta observación que se aprecia en el pase de la primera vuelta en la figura 4.42 hacia el período completo en la figura 4.43, confirmando los resultados presentados en el cuadro 4.4.

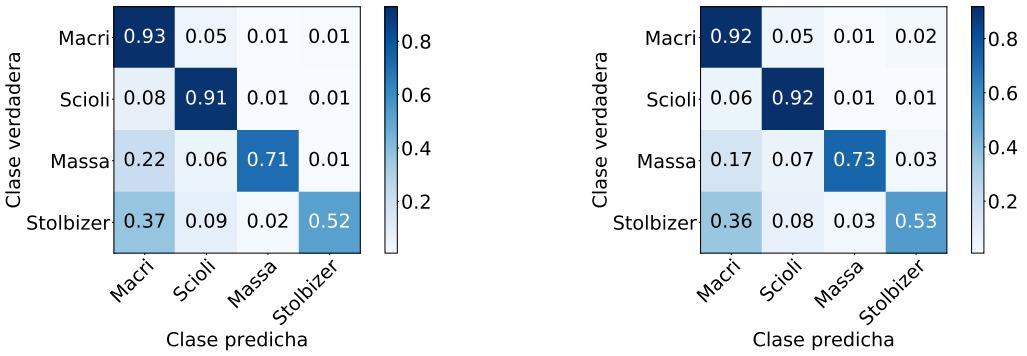


Figura 4.42: Matrices de confusión con datos tomados hasta la primera vuelta. Matriz de tópicos (Izquierda). Matriz con reducción de dimensiones LDA (Derecha).

Figura 4.43: Matrices de confusión con datos de todo el período electoral. Matriz de tópicos (Izquierda). Matriz con reducción de dimensiones LDA (Derecha).

Capítulo 5

Conclusiones

Hemos estudiado una red política de Twitter que abarcó el período de las elecciones presidenciales de Argentina de 2015. Analizamos algunas de las características del conjunto de datos, para luego desarrollar una metodología de extracción de tópicos de debate dentro de la red conformada por la coocurrencia en la utilización de los *hashtags* publicados por parte de los usuarios. Luego analizamos la similitud entre los usuarios de acuerdo a los tópicos detectados durante el período entero de captura de los datos y también la evolución temporal.

Exploramos trabajos previos y metodologías aplicadas para poder analizar características propias de una red social como el estudio de su topología para determinar que estímulos propagan una reacción en cadena o si existen reglas de contagio de información. En particular hemos aplicado algunas de estas metodologías al estudio de la homofilia entre personas y a la extracción de tópicos.

5.1. Contribuciones

El trabajo realizado amplió la visión y perspectiva sobre las elecciones presidenciales en Argentina en 2015. Detectamos temas de interés para personas, utilizando los *hashtags* como forma de representación de ideas dentro de un tema. En la evolución política de usuarios pudimos observar la discrepancia en los discursos de los dos candidatos que obtuvieron mayor cantidad de votos. También pudimos explicar el cambio de opinión de los seguidores de candidatos que quedaron excluidos de la segunda vuelta, y como su discurso cambió hacia el candidato que finalmente ganó las elecciones presidenciales. Finalmente, determinamos con una eficacia del 84 % a que candidato sigue una persona de acuerdo a los tópicos sobre los cuales publicó.

Las contribuciones de esta tesis dieron origen a un artículo para NetSci-X [1], una

[1] `https://netscisociety.net`

conferencia sobre redes complejas, y su posterior exposición el 3 de enero de 2019 en Santiago de Chile (Reyero *et al.*, 2019).

5.2. Trabajo futuro

Presentamos algunos puntos que serían de nuestro interés para continuar investigando:

- Selección de usuarios con intereses políticos: aportaría valor indagar en la selección de usuarios y utilizar otros indicadores de alineamiento político. En nuestro trabajo seleccionamos usuarios que siguieran a un único candidato, pero se podrían seleccionar usuarios que hayan *retweeteado* contenido de un único candidato, así como también analizar el texto de los *tweets* en busca de sentimiento positivo hacia alguno en particular.

- Detección de tópicos: daría un valor adicional y mejoraría la detección de tópicos el hecho de explorar otras alternativas para el filtro de contenido para la conformación de tópicos. Aumentar o disminuir el umbral de coocurrencia, utilizar un diccionario de palabras a excluir, analizar el texto dentro del *tweet* y realizar análisis de sentimiento, entre otras mejoras posibles.

- Representantes partidarios: sería interesante ampliar la selección de figuras políticas, yendo más allá de simplemente seleccionar al candidato a presidente. Existen personas dentro de los partidos políticos cuyo discurso es relevante, así como también personas fuera de afiliación política que son referentes para las personas y demuestran cierto apoyo hacia una fórmula presidencial en particular.

- Exploración en la predicción de afinidad política de usuarios: daría un valor adicional explorar metodologías en la determinación de las clases de los usuarios. Nuestra metodología consistió en seleccionar usuarios que hayan *retweeteado* contenido de un único candidato, pero esa acción no necesariamente significa apoyo. Se puede realizar análisis de sentimiento sobre los *tweets* publicados para determinar si efectivamente un usuario demuestra sentimiento positivo hacia un único candidato, y sentimiento neutral o negativo hacia los otros. Incluso se pueden hacer distinciones de usuarios, entre los que sólo demuestran sentimiento positivo hacia un candidato, y los que además demuestran sentimiento negativo hacia uno o más del resto de los candidatos.

Bibliografía

An, J., y Weber, I. (2016). # greysanatomy vs.# yankees: Demographics and hashtag use on twitter. *arXiv preprint arXiv:1603.01973*. 9

Bakshy, E., Hofman, J. M., Mason, W. A., y Watts, D. J. (2011). Everyone's an influencer: quantifying influence on twitter. En *Proceedings of the fourth acm international conference on web search and data mining* (pp. 65–74). 7

Barabási, A.-L., y Albert, R. (1999). Emergence of scaling in random networks. *science*, *286*(5439), 509–512. 17

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, *23*(1), 76–91. 11

Barbieri, N., Bonchi, F., y Manco, G. (2013). Cascade-based community detection. En *Proceedings of the sixth acm international conference on web search and data mining* (pp. 33–42). 9

Bastian, M., Heymann, S., Jacomy, M., *et al.* (2009). Gephi: an open source software for exploring and manipulating networks. *Icwsm*, *8*(2009), 361–362. 25, 30

Batagelj, V., y Mrvar, A. (1998). Pajek-program for large network analysis. *Connections*, *21*(2), 47–57. 25

Beiró, M. G. (2008). Visualización de redes complejas. *Tesis de grado en ingeniería informática. Fac. de Ingeniería (UBA).* (`http://cnet.fi.uba.ar/mariano.beiro/Tesis_Mariano_Beiro.pdf`) 25, 39

Blei, D. M., Ng, A. Y., y Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022. 62

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., y Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008. 20, 30

Bondy, J. A., Murty, U. S. R., *et al.* (1976). *Graph theory with applications* (Vol. 290). Macmillan London. 13

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32. 63

Brodersen, K. H., Ong, C. S., Stephan, K. E., y Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. En *2010 20th international conference on pattern recognition* (pp. 3121–3124). 63

Card, M. (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann. 24

Cardoso, F. M., Meloni, S., Santanche, A., y Moreno, Y. (2017). Topical homophily in online social systems. *arXiv preprint arXiv:1707.06525*. 9

Centola, D., y Macy, M. (2007). Complex contagions and the weakness of long ties. *American journal of Sociology*, *113*(3), 702–734. 6

Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P. K., *et al.* (2010). Measuring user influence in twitter: The million follower fallacy. *Icwsm*, *10*(10-17), 30. 8

Chen, D., Lü, L., Shang, M.-S., Zhang, Y.-C., y Zhou, T. (2012). Identifying influential nodes in complex networks. *Physica a: Statistical mechanics and its applications*, *391*(4), 1777–1787. 9

Clauset, A., Newman, M. E., y Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, *70*(6), 066111. 20

Colleoni, E., Rozza, A., y Arvidsson, A. (2014). Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication*, *64*(2), 317–332. 10

Conover, M., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., y Flammini, A. (2011). Political polarization on twitter. *Icwsm*, *133*, 89–96. 9, 11

Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., y Quattrociocchi, W. (2016). Echo chambers: Emotional contagion and group polarization on facebook. *Scientific reports*, *6*, 37825. 10

Galuba, W., Aberer, K., Chakraborty, D., Despotovic, Z., y Kellerer, W. (2010). Outtweeting the twitterers-predicting information cascades in microblogs. *WOSN*, *10*, 3–11. 7

Grabowicz, P. A., Ramasco, J. J., Moro, E., Pujol, J. M., y Eguiluz, V. M. (2012). Social features of online networks: The strength of intermediary ties in online social media. *PloS one*, *7*(1), e29358. 6

Granovetter, M. S. (1977). The strength of weak ties. En *Social networks* (pp. 347–367). Elsevier. 5

Halberstam, Y., y Knight, B. (2016). Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter. *Journal of Public Economics*, *143*, 73–88. 11

Hastie, T., Tibshirani, R., Friedman, J., y Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), 83–85. 63

Hodas, N. O., y Lerman, K. (2014). The simple rules of social contagion. *Scientific reports*, *4*, 4343. 6

Jalili, M., y Perc, M. (2017). Information cascades in complex networks. *Journal of Complex Networks*, *5*(5), 665–693. 9

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., y Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature physics*, *6*(11), 888. 9

Klašnja, M., Barberá, P., Beauchamp, N., Nagler, J., y Tucker, J. (2016). Measuring public opinion with social media data. *Handbook on polling and polling methods*. 11

Kossinets, G., y Watts, D. J. (2009). Origins of homophily in an evolving social network. *American journal of sociology*, *115*(2), 405–450. 9

Kullback, S., y Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79–86. 36

Lancichinetti, A., Radicchi, F., Ramasco, J. J., y Fortunato, S. (2011). Finding statistically significant communities in networks. *PloS one*, *6*(4), e18961. 21, 37

McPherson, M., Smith-Lovin, L., y Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, *27*(1), 415–444. 9

Mehmood, Y., Barbieri, N., Bonchi, F., y Ukkonen, A. (2013). Csi: Community-level social influence analysis. En *Joint european conference on machine learning and knowledge discovery in databases* (pp. 48–63). 9

Nekovee, M., Moreno, Y., Bianconi, G., y Marsili, M. (2007). Theory of rumour spreading in complex social networks. *Physica A: Statistical Mechanics and its Applications*, *374*(1), 457–470. 7

Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, *74*(3), 036104. 20

Newman, M. E., y Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, *69*(2), 026113. 18, 20

Page, L., Brin, S., Motwani, R., y Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web.* (Inf. Téc.). Stanford InfoLab. 8

Panario, M. (2016). Propagación de información en modelos sociales complejos. *Tesis de grado en ingeniería electrónica. Fac. de Ingeniería (UBA).* (`http://cnet.fi.uba.ar/matias_panario/Tesis_Matias_Panario.pdf`) 27

Pons, P., y Latapy, M. (2005). Computing communities in large networks using random walks. En *International symposium on computer and information sciences* (pp. 284–293). 21

Reichardt, J., y Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E*, *74*(1), 016110. 20

Reyero, T. M., Alvarez-Hamelin, J. I., y Beiró, M. G. (2019, Jan). *Topic-based study of a Twitter political network* (Inf. Téc.). NetSciX2019. Descargado de `http://netscix.net/program.html` 70

Romero, D. M., Meeder, B., y Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. En *Proceedings of the 20th international conference on world wide web* (pp. 695–704). 11

Rosenman, E. T. (2012). *Retweets-but not just retweets: Quantifying and predicting influence on twitter.* Cambridge. 8

Rosvall, M., Axelsson, D., y Bergstrom, C. T. (2009). The map equation. *The European Physical Journal Special Topics*, *178*(1), 13–23. 21

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*(3), 379–423. 36

Watts, D. J. (2002). A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, *99*(9), 5766–5771. 6, 7

Watts, D. J., y Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, *393*(6684), 440. 18

Weng, J., Lim, E.-P., Jiang, J., y He, Q. (2010). Twitterrank: finding topic-sensitive influential twitterers. En *Proceedings of the third acm international conference on web search and data mining* (pp. 261–270). 8

Xie, J., y Szymanski, B. K. (2011). Community detection using a neighborhood strength driven label propagation algorithm. En *2011 ieee network science workshop* (pp. 188–195). 20