Propagación de Información en Modelos Sociales Complejos

Tesis de grado presentada

por

Matías Panario

para el

Departamento de Electrónica,

Facultad de Ingeniería,

Universidad de Buenos Aires,

dirigida por

Dr. Ing. Sergio Lew,

co-dirigida por

Dr. Ing. José Ignacio Alvarez-Hamelin,

Buenos Aires, Argentina,

17 de mayo de 2016

Resumen

Diversos han sido los esfuerzos por caracterizar diferentes redes sociales tanto biológicas como artificiales.

Tomando como punto de partida modelos de comunicación asimétricos, hemos enfocado el análisis en el desarrollado de diversos algoritmos para el procesamiento de datos y extracción de características intrínsecas de grandes redes.

Gracias al avance de la tecnología, en la actualidad es posible obtener un elevado volumen de información y junto con un alto poder de procesamiento, se pueden aplicar técnicas y algoritmos con gran facilidad.

En lo particular, en el presente trabajo nos hemos centrado en la asimetría de una red social artificial partiendo de preguntas y estudios previos sobre otros redes sociales que aparecen en la naturaleza.

Nos hemos enfocado en caracterizar dichas redes con algoritmos actuales como la distribución de k-núcleos (*k-cores*) y coeficientes de agrupamiento (**clustering coefficient**) entre otros. El período elegido corresponde a las elecciones presidenciales de la República Argentina donde un gran volumen de información pudo ser recopilada. Este punto es de suma importancia en comparación con el análisis de otras redes sociales de animales donde el acceso a la información es limitado o el tiempo necesario para obtenerla puede afectar la identificación de dichas características.

Hemos identificado a través de dichos algoritmos y procedimientos una base de análisis para definir porque ***Twitter*** es utilizada ampliamente como red social de propagación de noticias.

Tags: *redes sociales artificiales, Twitter, tweet, k-núcleos, coeficiente de agrupamiento, grafo direccionado, supernodos, elecciones presidenciales en Argentina.*

Agradecimientos

Sección agradecimientos

Índice general

1. Introducción **1**

2. Materiales y Métodos **4**

2.1. Objetos 4

2.1.1. Usuario 4

2.1.2. Tweet 6

2.2. Modelo Social 6

2.3. Relaciones 8

2.4. Grafo 12

2.5. Arquitectura 12

2.5.1. Origen de los Datos 12

2.5.2. Servidores 14

2.5.3. Almacenamiento 15

2.6. Procesamiento 16

2.6.1. Python 16

2.6.2. Matlab 18

2.6.3. LaNet-Vi 18

2.7. Interfaces 18

2.7.1. Extracción de relaciones 20

2.7.2. Extracción de *tweets* 20

2.7.3. Flujo continuo de *tweets* 20

2.8. Set de datos 21

2.9. Tipos de Usuarios 23
2.10. K Núcleos 24

3. Resultados **25**

3.1. Topología 25
3.2. Distancia entre supernodos 29
3.3. Relación Retweet / Menciones 33
3.4. Clasificador de seguidores 38
3.5. Evolución Temporal 40
3.6. Retweets de seguidores 45
3.7. Alineación de Retransmisores 47
3.8. Flujo Continuo y *tweets* 50
3.9. Eventos 52
3.10. Análisis de k-núcleos 54
3.10.1. Remoción del núcleo principal 54
3.10.2. Evolución usuarios versus k-núcleos 61
3.10.3. Coeficiente de Agrupamiento 63

4. Conclusiones **64**

5. Apéndices **67**

5.1. Estructuras 67
5.1.1. Estructura de un *tweet* 68
5.1.2. Estructura de un usuario 72
5.2. Retransmisiones 73
5.3. Múltiples Aplicaciones 79

Índice de figuras

2.1. Usuario de ***Twitter*** 5
2.2. *tweet* 6
2.3. Modelo social asimétrico 8
2.4. Modelo social simétrico 9
2.5. Usuarios, *tweets* y relaciones 10
2.6. Seguidores y Amigos 11
2.7. Arquitectura 13
2.8. Ejemplo interface 19
2.9. Extracción de *tweets* 21
2.10. Supernodo versus usuario regular 23

3.1. Histograma de seguidores y amigos 26
3.2. Estimación α y D_{KS} 27
3.3. Divergencia de σ 28
3.4. Desvío estándar en función del set de datos 28
3.5. Número de seguidores por supernodo 29
3.7. Distancia de Jaccard 32
3.8. Dupla de supernodos: Menciones y retransmisiones 35
3.9. Menciones en función de retransmisiones 36
3.10. Performance del clasificador 39
3.11. CCDF retransmisiones y favoritos 41
3.12. Evolución temporal 43
3.13. *tweets* generados a lo largo del día 44

3.14. Propagación de contenido 46
3.15. Alineación de supernodos 48
3.16. Actividad en la red y *tweets* de los supernodos 51
3.17. Actividad de la red 53
3.18. Remoción del top core 59
3.19. Número de lazos y grado promedio 62
3.20. Coeficiente de Agrupamiento 63

5.1. *tweet* número 1 74
5.2. *tweet* número 2 75
5.3. *tweet* número 3 76
5.4. *tweet* número 4 77
5.5. *tweet* número 5 78
5.6. Extracción multi-aplicación 79

Capítulo 1

Introducción

Las sociedades siempre han posibilitado la interacción de sus individuos en diferentes formas. Animales, seres humanos y dispositivos electrónicos han establecido vínculos entre ellos para la obtención de un fin cualquiera sea. Desde el desarrollo social de un individuo, la administración de una colmena u hormiguero, la toma de decisiones o bien la comunicación digital de cualquier índole, siempre es posible definir los actores y la forma en que estos interactúan. Los diferentes tipos de entidades e interacciones definirán características intrínsecas de las redes subyacentes. Si pensamos a estos individuos como nodos y a las interacciones como lazos, cualquier tipo de acción y sociedad puede ser descripta de acuerdo a estas redes.

En el grupo de las sociedades de animales, siempre han surgido interrogantes acerca de la existencia de uno o múltiples líderes [King et al., 2015], si es posible identificarlos o en cuáles actividades estos participan o dejan de hacerlo. Aún más relevante es saber como ellos se comunican con el resto. En todos los casos, para responder y analizar este tipo cuestiones, es necesaria la recopilación y procesamiento de información en un tiempo razonable. Estudios realizados en chimpancés como [Nishida, 1983, Mitani and Amsler, 2003] se han prolongado por años a fin de recopilar eventos suficientes para responder a dichas preguntas. En otros estudios como [Lusseau and Newman, 2004] sólo se ha podido relevar cierto volumen de información acerca de las relaciones en la red debido a las limitaciones en la adquisición de

datos. Esto lleva a pensar en las limitaciones impuestas el muestreo en función de la dinámica que se desea observar.

Si analizamos sociedades de hormigas y abejas, cuestiones acerca de cómo la reina mantiene su *status* [Chen et al., 2010], cómo se establece el canal de comunicación entre individuos que cumplen diferentes funciones [Jeanson et al., 2003] o bien cómo optimizan procesos tales como la localización de alimentos [Morone and Makse, 2015, Plank and James, 2008, Reid et al., 2011], dan una idea acerca de cómo los individuos desarrollan las diversas actividades. Diversos estudios han propuesto modelos, similitudes y diferencias con otros tipos de sociedades.

En el caso de los mercados de valores, donde diferentes actores interactúan en la compra y venta de valores, el mismo es considerado un *Mercado Perfecto* ya que cumple con los requisitos para serlo: *ningún operador tiene la capacidad de influir en las cotizaciones; cualquier participante tiene igual acceso a la información relevante; y cada inversor conoce al mejor precio que puede comprar o vender* [Robert C. Merton, 1974]. Esto define características acerca de cómo la información es consumida por los diferentes actores, los principios [Soloviev and Saptsin, 2011] y la velocidad con la que las acciones son tomadas.

Por otro lado, existen diferentes redes sociales donde las personas interactúan entre sí tales como Internet [Wolfgang, 2008], blogs, Facebook o ***Twitter*** [Boyd and Ellison, 2007]. Cada una, con sus propias características, plantea interrogantes interesantes tales como qué rol cumple cada indivíduo [Conover et al., 2011], si es posible identificar líderes y cómo hacen estos para mantener su *status* [Benhardus and Kalita, 2013] o bien hasta dónde y cómo es propagada la información. Trabajos recientes con acceso global a información distribuida, han analizado las características de estos modelos particulares muy rápidamente y con un nivel de complejidad elevado respecto de otras redes [Bild et al., 2015a, Li et al., 2015]. Esto puntualmente es difícil de obtener en modelos sociales naturales como se mencionó en el estudio sobre los chimpancés.

En la presente tesis nos hemos enfocado en analizar si es posible fomentar o limitar la comunicación de un líder en un modelo social asimétrico.

Para el desarrollo de la presente tesis se ha elegido entonces la red social ***Twitter***, que en su naturaleza es un modelo social asimétrico de comunicación y su información es de libre acceso. Dentro de ella, nos hemos propuesto identificar sus características topológicas fundamentales y aquellas relacionadas con la propagación de la información. Nuestro principal objetivo es entender las características de aquellos actores que bien podrían responder a las cuestiones planteadas también para otras redes naturales.

Capítulo 2

Materiales y Métodos

En el presente capítulo se definirán los objetos, modelos, arquitectura y procedimientos utilizados para obtener, almacenar y extraer información útil de los datos utilizados durante el desarrollo de la tesis.

2.1. Objetos

En la red social ***Twitter*** se definen dos objetos: usuarios y *tweets*. Sobre ellos se centró el desarrollo de la tesis. Cada uno contiene características intrínsecas que los definen y que permiten establecer relaciones entre ellos. Se han definido relaciones entre usuarios, entre usuarios y *tweets* y entre *tweets*. Cada una de estas relaciones posee características fundamentales que permiten el funcionamiento de la red social como tal.

2.1.1. Usuario

Los usuarios representan las entidades que generan contenido y que interactúan con otras entidades a través de la propagación de información sobre en la red.

Estos usuarios están definidos con decenas de características, en el presente trabajo sólo nos hemos enfocamos en aquellas que permiten identificar la generación y propagación de información así como también la relación que une a las diferentes

entidades.

A continuación se detallan algunas de las propiedades que caracterizan a los usuarios:

- *id*: Cada usuario posee un identificador único que le permite incorporarse a la red.
- *nombre*: Representa el alias que todo usuario de la red puede adquirir.
- *subscriptores*: Listado de usuarios que están subscriptos a este usuario y los cuales consumirán la información que éste genere.
- *amigos*: Listado de otros usuarios a los que este usuario está subscripto y de los cuales consumirá información.
- *protegido*: Indica si el usuario es protegido o no.
- *estado*: Representa el último objeto *tweet* generado por el usuario.
- *cantidad de contenido generado*: Indica el número de objetos *tweet* generados por este usuario.
- *verificada*: Muestra si la cuenta esta verificada por la persona.

En la figura 2.1 se muestra un ejemplo de un usuario donde puede apreciarse la cantidad de personas que lo siguen, *Seguidores*, y la cantidad de personas a las que este usuario esta subscripto, *Amigos*.

Figura 2.1 – *Usuario de* ***Twitter****.*

Un ejemplo detallado con toda la información de un usuario se ha incluido en los apéndices.

2.1.2. Tweet

Los *tweets* son los objetos fundamentales en la comunicación dentro de ***Twitter***. Son las entidades que contienen el mensaje, siendo público para toda la red[1]. Este objeto tiene además como principal característica que el texto (cuerpo del mensaje) puede contener hasta 140 caracteres y además incluir imágenes y videos como contenido multimedia adicional. En la figura 2.2 se ejemplifica dicha entidad y se puede observar además la cantidad de usuarios que lo han retransmitido, *retweeters*, y la cantidad de usuarios que lo han identificado como favorito.

Figura 2.2 – *Este es el primer tweet escrito en la historia de* ***Twitter*** *y fue creado por el usuario* ***Jack****, creador de la red social.*

2.2. Modelo Social

El modelo social en ***Twitter*** se ha caracterizado como *asimétrico*. Aquí el establecimiento del vínculo entre dos entidades usuarios puede ser o no bidireccional pero no necesariamente tiene que existir en ambos sentidos para que se establezca. Un exhaustivo análisis acerca redes sociales asimétricas digitales ha sido desarrollado en [Boyd and Ellison, 2007].

Ejemplos de este tipo de red social se encuentran en la naturaleza, especialmente en hormigas y abejas, donde también se establecen lazos de comunicación unidireccionales sin esperar respuesta por parte del receptor. En estos casos el medio transmisor de la información no es digital sino químico, especialmente con el uso de feromonas. Existen diferentes tipos de feronomas que son secretadas por seres vivos con el objeto de provocar un comportamiento deseado en otros individuos. Dichas

[1] Es posible mandar mensajes privados, pero la presente tesis se centra en la información pública.

feromonas son consideradas como medios de trasmisión de señales tanto volátiles como no-volátiles. En este caso se han definido similitudes con los *tweets* dentro de la red social ***Twitter***, donde usuarios los crean esperando generar el impacto deseado en sus seguidores sin importar la opinión de estos.

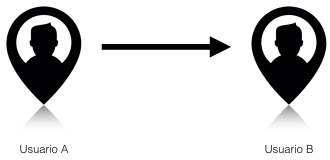
Otro ejemplo de red social asimétrica son los medios de comunicación gráficos y masivos, como los periódicos. El equipo editorial del diario es el encargado de recopilar las noticias de los diferentes autores y periodistas para realizar la publicación con el objetivo de llegar a la mayor cantidad de lectores. En estos casos, la opinión de la mayoría de los lectores no es recopilada y provista a los autores, planteándose la asimetría en el modelo.

Al ser un modelo asimétrico y direccional de comunicación, se establecen cuatro posibles vínculos entre cualquier dupla de usuarios A y B dentro de la red que se ilustran en la figura 2.3. En la figura 2.3a se observa que A sigue a B y consume todo el contenido que es publicado por B. En la figura 2.3b se observa que B sigue a A y consume todo el contenido que este último genera. En la figura 2.3c se observa que tanto A como B consumen el contenido generado por el otro. Finalmente, en la figura 2.3d se observa que ni A ni B se siguen, y a priori, ninguno se entera del contenido generado por el otro directamente[2].

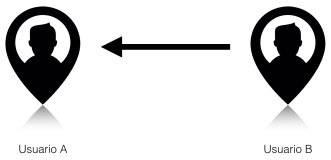
Estos cuatro posibles vínculos entre entidades de una red social asimétrica se verían reducidos a dos si la red social fuera simétrica. Esto implicaría que el vínculo se establece sólo si ambas partes concuerdan en establecerlo. Si no existe concordancia, entonces el vínculo no queda establecido. Estos dos estadios se pueden observar en la figura 2.4

Una de las características más importantes sobre la cual la presente tesis se enfoca, es la posibilidad de consumir la mayoría de la información generada en la red. ***Twitter*** es una red social que pone a disposición todo el contenido público generado por sus usuarios. Este punto es de suma importancia a la hora de recolectar los datos, almacenarlos y posteriormente analizarlos. Si toda la información se encuentra disponible, entonces un gran volumen de datos puede ser extraído y analizado. A

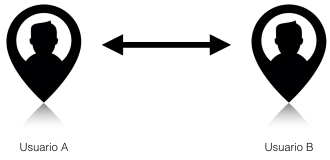
[2]Aunque pudiese haber posibles caminos indirectos para llegar de uno a otro.



(a) *A sigue a B.*



(b) *B sigue a A.*



(c) *Ambos se siguen mutuamente.*

(d) *Ningún usuario sigue al otro.*

Figura 2.3 – *Ejemplo de un modelo social asimétrico donde no es necesario establecer ambos sentidos de comunicación para establecer la misma.*

diferencia de la metodología empleada en estudios sociales tradicionales, como los que se han realizado sobre el comportamiento de chimpancés en la toma de liderazgo [Nishida, 1983] o sobre redes sociales en otras especies [Lusseau, 2003], las redes sociales artificiales minimizan la siguiente limitación: la imposibilidad de obtener gran volumen de información en un tiempo razonable que permitan obtener conclusiones significativas.

2.3. Relaciones

En la presente sección se describirán las relaciones y la interacción entre los usuarios (como entidades). Como se puede apreciar en la figura 2.5, múltiples relaciones entre diferentes usuarios pueden ser establecidas en la comunicación dentro de la red social.

(a) *Se establece el vínculo.* **(b)** *No se establece el vínculo.*

Figura 2.4 – *Ejemplo de un modelo social simétrico donde se establece o no la comunicación dependiendo de si ambas partes desean comunicarse.*

Escribir un *tweet* (POST)

Es la generación de contenido. Los usuarios son capaces de introducir nuevas entidades *tweet* a la red. Todos aquellos usuarios que se encuentran suscriptos al autor del *tweet*, serán automáticamente informados acerca del nuevo contenido generado.

Suscripción (FOLLOW)

Se define como la acción de un usuario de suscribirse a otro con el objeto de consumir el contenido generado por éste último. En esta red social asimétrica basada en la suscripción, este tipo de relación es de suma importancia y define la topología de la red con los lazos entrantes, *seguidores*, y los lazos salientes, *amigos*, a los usuarios.

La figura 2.6 muestra aquellos usuarios suscriptos a la entidad usuario A con flechas entrantes y resaltados en color naranja. Por el otro lado, el usuario A se encuentra suscripto a otras entidades usuarios, resaltadas en color verde. Bajo este esquema, las entidades *tweet* generadas por el usuario A serán consumidas por los usuarios naranja y a su vez el usuario A consumirá el contenido generado por los usuarios verdes.

Retransmitir contenido (RETWEET)

Es la acción de republicar contenido generado por otro usuario. Todos aquellos seguidores de los retransmisores también lo leerán. Este tipo de relación permite la

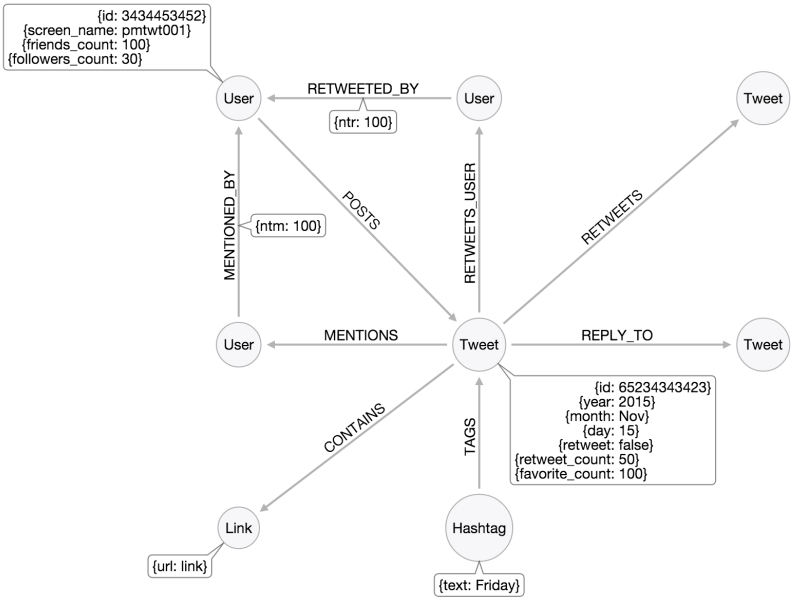


Figura 2.5 – *Relaciones entre usuarios y tweets. Modelo funcional centrado en el objeto tweet y su vinculación con otros tweets y usuarios a través de las diferentes opciones disponibles en* ***Twitter****.*

propagación de contenido a través de los suscriptores dando lugar al concepto de distancia entre el escritor y el usuario final, según la cantidad de saltos en el medio. Al retransmitir un *tweet*, el objeto original permanece inalterable en cuanto a su contenido. El usuario que retransmite no agrega ni elimina información durante el proceso de retransmisión.

Mención (@)

Al escribir un *tweet* un usuario puede mencionar a otro usuario. A este proceso se lo conoce como "mención". El usuario mencionado recibe una notificación donde se le informa que ha sido mencionado y recibe acceso al *tweet* donde se lo menciona. La forma para mencionar un usuario es mediante la inclusión del símbolo arroba (@) delante del nombre del usuario.

En el siguiente *tweet*, el usuario *MoralesGerardo1* ha sido mencionado dentro del texto.

"Acompaño y apoyo al Gobernador @MoralesGerardo1, que enfrenta con coraje a la máquina extorsiva y clientelar que maltrata a Jujuy."

Figura 2.6 – ***Seguidores y amigos***. *Se observa que múltiples usuarios son seguidores (color naranja) del Usuario A y al mismo tiempo este usuario es amigo de otros (color verde).*

Responder (REPLY)

Un usuario es capaz de responder a un *tweet* generado por otro usuario. El usuario que responde, genera un nuevo *tweet* identificado como respuesta al contenido original.

Marcar como favorito un Tweet (FAVORITE)

Los *tweet* pueden ser también marcados como favoritos. En este caso, el usuario consume el contenido y a su vez refuerza positivamente el mismo marcándolo como favorito.

Etiquetas (#)

Al generar un *tweet* es posible marcar ciertas palabras con el símbolo *numeral* delante. Esta acción refuerza la palabra marcada y permite rastrear la huella de la misma a lo largo de toda la red. Las palabras etiquetadas que son utilizadas por una gran cantidad de usuarios suelen convertirse en tendencia ya sea a nivel local o a nivel mundial. Ejemplo de esto último es la etiqueta *#earthquake* utilizada en *tweets* durante y luego de la ocurrencia de un terremoto.

2.4. Grafo

Se ha definido el grafo sobre los vínculos de *seguidores* entre los diferentes usuarios. Considerando la acción de subscripción, un usuario puede hacerlo a múltiples usuarios y a su vez múltiples usuarios pueden estar subscritos al primero. Definido así, el grafo es direccionado y muestra la asimetría que se presenta cuando los usuarios pueden seguirse mutuamente.

Dentro de la red, el rol de cada usuario ha sido definido de acuerdo a la interacción con el resto de las entidades así como con la cantidad de contenido generado. Un usuario puede ser partícipe activo en la generación de contenido, responder ante otro que lo menciona y utilizar muchas de las opciones provistas por la red social o bien ser sólo consumidor de información generada por otros.

Si el usuario consume información generada por otros, entonces es considerado como *consumidor*. Si a su vez se dedica a retransmitir contenido generado por terceros entonces es considerado como *retransmisor*.

2.5. Arquitectura

La figura 2.7 define las tecnologías y etapas utilizadas desde la obtención de los datos hasta el análisis de los resultados. El ecosistema de tecnologías se presenta de acuerdo a la posibilidad de consumir la información directamente desde la fuente, en este caso ***Twitter***.

2.5.1. Origen de los Datos

Han sido definidos dos elementos fundamentales para describir el origen de los datos obtenidos: el proveedor de los datos (la red social) y la puerta de acceso por la cual es posible acceder a esos datos.

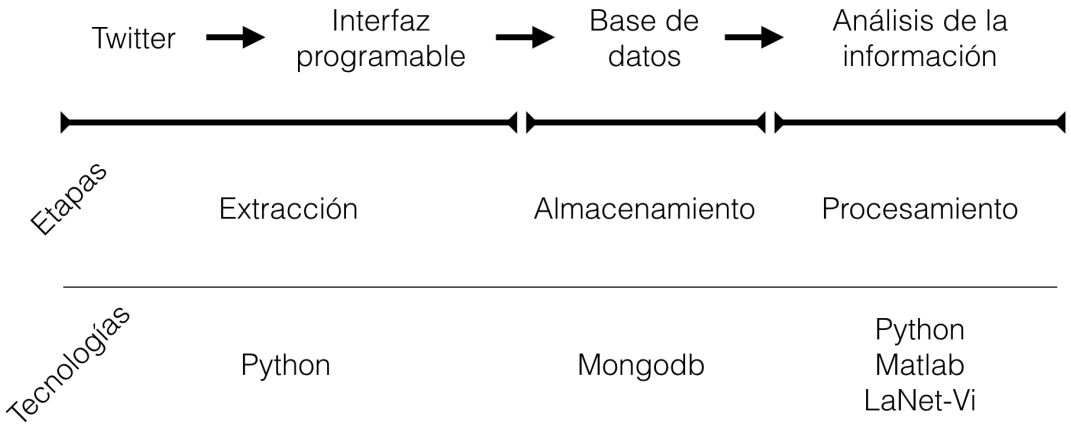


Figura 2.7 – ***Arquitectura de tecnologías y etapas***. *Se utilizaron diferentes tecnologías para poder extraer, almacenar y procesar los datos. Las tecnologías empleadas son en su mayoría de código abierto y fueron elegidas en función del tipo de objeto utilizado. JSON[ECMA International, 2011].*

Twitter como fuente de datos

La fuente de información de la presente tesis es la red social ***Twitter***. La misma permite, a través de su *interfaz de aplicación programable*[3], autenticar una aplicación y consumir la información deseada a través de las diferentes funciones provistas. La misma también impone límites en cuanto a la cantidad de información que puede ser extraída y respecto a la ventana de tiempo. Con el fin de incrementar la cantidad de información consumida se han creado 20 usuarios y aplicaciones para poder conectarse con múltiples instancias a la red, y así obtener más datos en menos tiempo.

En los apéndices se pueden encontrar ejemplos de cómo los objetos son provistos por ***Twitter***.

Se almacenó la información de los *tweets* y de los usuarios. Entre ellos: la relación existente entre usuarios y sus características así como el número de seguidores y amigos (usuarios), el número de retransmisiones y favoritos (*tweets*), marcas o tags

[3] API por sus siglas en inglés

utilizados (palabras), menciones (usuarios) y tiempo en el que ocurren los eventos (*tweets*).

API como puerta de entrada

Múltiples funciones han sido utilizadas para poder consumir los datos necesarios. Se enuncian a continuación con una breve descripción:

- GET friends/ids: Extrae la lista de *id* de los amigos de un usuario.
- GET followers/ids: Extrae la lista de *id* de los seguidores de un usuario.
- GET friends/list: Extrae la lista de amigos de un usuario incluyendo todos sus campos.
- GET followers/list: Extrae la lista de seguidores de un usuario incluyendo todos sus campos.
- GET users/show: Extrae todos los datos del usuario deseado incluyendo su más reciente *tweet*.
- GET statuses/retweeters/ids: Extrae los últimos 100 usuarios que han retransmitido un *tweet*.
- GET statuses/user_timeline: Extrae los últimos 3200 *tweets* de un usuario definido.
- GET application/rate_limit_status: Devuelve un objetivo con la cantidad de peticiones aún disponibles en la ventana de 15 minutos.

2.5.2. Servidores

Se han utilizado tres servidores con las siguientes especificaciones técnicas:

- Sistema Operativo (SO): 3.16.0-4-amd64 Debian 3.16.7-ckt11-1 (2015-05-24) x86 64 GNU Linux

- Procesadores: 1 x Intel(R) Xeon(R) CPU E5-2650 0 @ 2.00GHz
- Memoria RAM: 1GB
- Disco Rígido: Se utilizaron discos de estado sólido (SSD por sus siglas en inglés)
  - Partición del SO: 7.8GB
  - Partición de Datos: 21GB

En estos se han alojado las bases de datos y las interfaces de adquisición de datos.

Los tres servidores han sido provistos por *Amazon Elastic Cloud*[4], que permite la creación de máquinas virtuales con múltiples opciones de configuración y la posibilidad de modificar los recursos reservados de acuerdo a la demanda.

2.5.3. Almacenamiento

Hemos utilizado MongoDB[5] como motor de base de datos para dar soporte de almacenamiento de todos los datos consumidos desde ***Twitter***. El mismo es de código abierto, gratuito y orientado a objetos. En nuestro caso se ha elegido dicha tecnología ya que permite el almacenamiento de objetos JSON [ECMA International, 2011], consumidos desde el origen. La versión utilizada en los servidores es la 3.0.5.

Al comienzo del desarrollo hemos utilizado la versión 2.7, pero a lo largo del trabajo hemos migrado a la versión 3.0 con el fin de implementar el nuevo motor de almacenamiento **WiredTiger**[6] y posibilitar así:

- La compresión nativa de los datos. Mientras que MMAPv1 almacena los datos de forma binaria, WiredTiger utiliza librerías zLib[L. Peter Deutsch, 1996] para la compresión.

[4]https://aws.amazon.com/ec2/
[5]https://www.mongodb.org/
[6]http://www.wiredtiger.com/

- Incremento en la performance general del motor. Ambos motores permiten múltiples lecturas sobre una misma colección, pero WiredTiger permite también escrituras simultáneas a diferencia de MMAPv1 que sólo permite una.

2.6. Procesamiento

Se han utilizando principalmente las siguientes herramientas para analizar los datos obtenidos:

- Python v2.7: Intérprete de comandos orientado a objetos. Lenguaje de programación de alto nivel con semántica dinámica.
- Matlab 2015b Student Version: Ambiente numérico multi-paradigma y lenguaje de programación propietario.
- LaNetVi 3.0.1: Herramienta de visualización de grandes redes.

2.6.1. Python

En este lenguaje se han desarrollados todas las interfaces de acceso y extracción de datos y su posterior almacenamiento en las bases de datos correspondientes.

Múltiples módulos de diversas bibliotecas han sido utilizados con el fin de desarrollar tareas específicas. A continuación la lista de bibliotecas utilizadas:

- Medición de tiempos:
  - datetime
  - time
- Procesamiento matemático:
  - random
  - pylab
  - numpy

    - matplotlib

- Procesamiento de archivos y estructura de datos:
    - bisect
    - re
    - sets
    - json
    - csv
    - itertools
- Interacción con el sistema operativo:
    - sys
    - os
    - glob
    - argparse
- Procesamiento de redes:
    - networkx
    - graph_tool
- Obtención y almacenamiento de datos:
    - requests
    - urllib2
    - tweepy
    - pymongo

2.6.2. Matlab

Se ha utilizado Matlab para el manejo de vectores provenientes de agregaciones de valores de los datos extraídos de ***Twitter***. Por ejemplo, la clasificación de usuarios según afinidad política de acuerdo a la cantidad de *retweets* y menciones realizadas.

2.6.3. LaNet-Vi

Se ha utilizado esta herramienta para calcular y graficar la descomposición en *k-núcleos* de la red de seguidores generada con las relaciones extraidas de ***Twitter***.

Se obtuvieron y almacenaron los seguidores de diferentes usuarios. Estos usuarios se eligieron de una lista de retransmisores (retweeters) de los *tweets* de los supernodos. De esta lista de usuarios se almacenó la relación usuario - seguidor en una base de datos. Al finalizar el proceso de adquisición de datos se obtuvieron **32 millones** de usuarios (entre retransmisores y seguidores) con **111 millones** de relaciones.

Para la adquisición de datos se ha programado una interfaz que extrae de ***Twitter*** los *ids* de los seguidores de un usuario dado y almacena dicha relación en una base de datos MongoDB para su posterior procesamiento.

Luego, se han generado archivos de datos planos (csv) con los lazos del grafo inducido generado por el *tweet* y posteriormente han sido procesado con el programa LaNetVi. El resultado de este proceso es la obtención de una imagen con los nodos, el grado (sobre la izquierda) y el núcleo k (escala sobre la derecha) dibujados en un mismo gráfico. Dentro de la figura, los nodos se encuentran coloreados con el color del núcleo y los lazos con los colores correspondientes a los nodos con los que están conectados, facilitando una rápida identificación de los núcleos interconectados.

2.7. Interfaces

Como se ha mencionado anteriormente, diferentes interfaces fueron creadas con el fin de extraer los datos a través de las funciones provistas por la *API*. Cada una de estas interfaces utiliza una o más conexiones con el fin de reducir los tiempos de

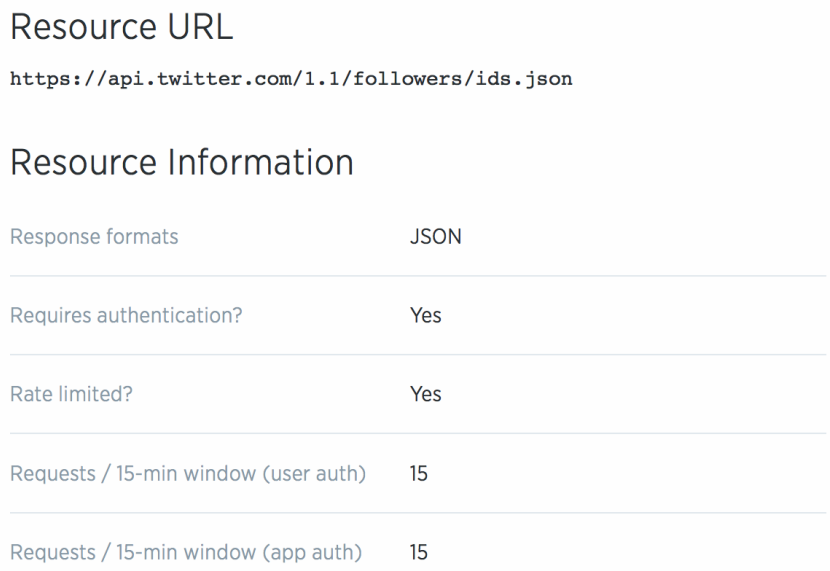


Figura 2.8 – ***Ejemplo interface para obtener id de lo seguidores.*** *Cada aplicación tiene sus propios parámetros, enlaces de acceso y número de extracciones por ventana de 15 minutos. En este caso, se pueden ejecutar como máximo 15 pedidos cada 15 minutos. Por cada pedido la aplicación devuelve 5000 ids lo que arroja unos 75 mil usuarios por ventana de tiempo.*

extracción. Para tal fin, 20 aplicaciones de ***Twitter*** fueron creadas y utilizadas de acuerdo a la demanda que la interfaz requiriese.

A modo de ejemplo, para la extracción de los seguidores de un usuario se utilizó la función **GET followers/ids**. Como se observa en la figura 2.8 esta función permite solicitar 15 pedidos cada 15 minutos o bien 75 mil usuarios (15 pedidos x 5000 ids por pedido). Esto resulta suficiente cuando los usuarios no poseen gran cantidad de seguidores, sin embargo, para unos pocos que tienen varios millones (como los supernodos), el tiempo de espera para completar la extracción se puede convertir en horas o incluso días. Para el caso de Macri con 4 millones de seguidores, el tiempo de espera para extraer todos los *ids* de sus seguidores sería de aproximadamente 13hs. Al haber creado 20 aplicaciones hemos reducido este tiempo a 30 minutos.

A continuación una descripción de la interfaces desarrolladas en Python junto con el propósito de las mismas.

2.7.1. Extracción de relaciones

Se han diseñado múltiples interfaces con el fin de extraer las relaciones que conforman el grafo. En la sección anterior hemos descripto el comportamiento de la función **GET followers/ids**, y la misma se ha utilizado para la extracción de los seguidores de los supernodos y para aquellos usuarios que retransmitieron un *tweet* de estos.

El uso de las múltiples aplicaciones fue fundamental ya que, en el primer caso, los supernodos tienen miles o millones de seguidores mientras que en el segundo caso, si bien es de esperar que decaiga el número de seguidores, se eleva la cantidad de pedidos que se hace a medida que el *tweet* obtiene más retransmisiones y por ende más usuarios.

Ambos procesos almacenan la información en diferentes bases de datos pero siempre lo hacen registrando el id del seguidor así como el del usuario que es seguido.

2.7.2. Extracción de *tweets*

En el caso de la extracción de *tweets* se ha utilizado la función **GET statuses/user_timeline**, la figura 2.9 muestra el número total de pedidos por ventanas de 15 minutos. Nuevamente fue importante el uso de las 20 aplicaciones a fin de poder extraer un volumen suficiente de contenido y almacenarlo en nuestras bases de datos para su posterior análisis.

2.7.3. Flujo continuo de *tweets*

Alternativamente a la extracción sistemática de *tweets* es posible utilizar la función **GET statuses/filter** que permite obtener cierta cantidad de objetos relacionados con el contenido definido en el filtro.

Este porcentaje es cercano al 1 % de la actividad total de la red. Si bien este número puede no parecer representativo, en momentos donde se registra una masiva cantidad de *tweets* relacionados al filtro en cuestión, la cantidad puede llegar a ser suficiente para diversos análisis donde se consolida la información. En nuestro

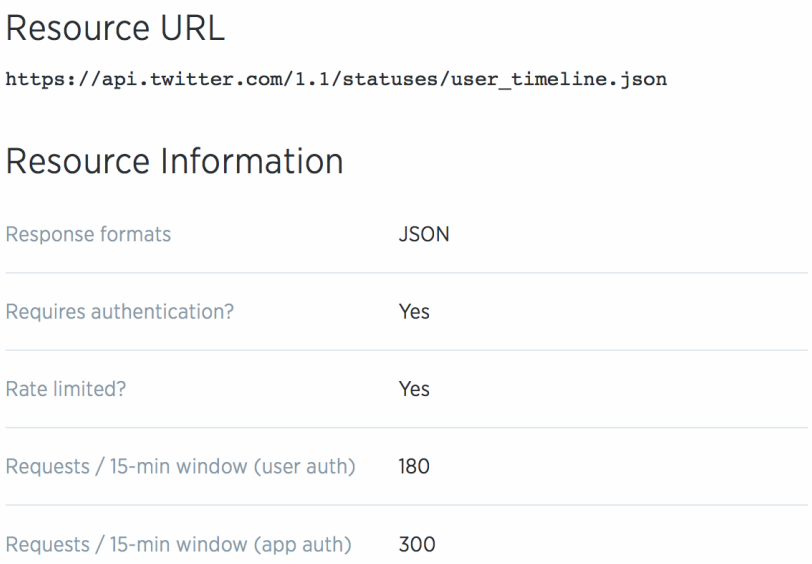


Figura 2.9 – ***Función utilizada para la extracción de tweets***. *En este caso, la función permite la extracción de los últimos 3200 tweets del usuario definido.*

análisis posterior hemos comparado resultados del flujo continuo obtenido de tres de los supernodos ante eventos puntuales junto con sus *tweets*.

El resultado obtenido de este tipo de interfaces es un objeto *tweet* que se almacena por completo en nuestras base de datos.

2.8. Set de datos

Durante el desarrollo de la tesis se almacenó una gran cantidad de *tweets* y relaciones entre usuarios de la red.

Los datos han sido extraidos en el período comprendido entre el *9 de Septiembre de 2015* y el *31 de Marzo de 2016*.

A continuación se detalla la cantidad de objetos almacenados, junto con una breve descripción acerca de la forma en que fueron adquiridos y el motivo.

tweets

Se almacenaron **54 millones** de *tweets* provenientes de **343 mil** usuarios que en algún momento han retransmitido contenido de alguno de los supernodos. Este set

de datos ha sido utilizado para analizar la cantidad de retransmisiones y menciones a los supernodos.

Se han almacenado los últimos **28 mil** *tweets* de los **8 supernodos** con el fin de estudiar la cantidad de retransmisiones y favoritos de estos.

Debido a la velocidad con la que el número de retransmisiones y favoritos crecen en el tiempo, fue necesario almacenar fotos de ciertos parámetros de los *tweets* cada 10 segundos a fin de poder analizar la dinámica de los mismos. A continuación se detalla el número de *tweets* por supernodo con información instantánea en este rango de tiempo.

1. Cristina Fernández de Kirchner (@CFKArgentina): 532
2. Aníbal Fernández (@FernandezAnibal): 322
3. Ernesto Sanz (@SanzErnesto): 26
4. Sergio Massa (@SergioMassa): 262
5. Margarita Stolbizer (@Stolbizer): 465
6. Daniel Scioli (@danielscioli): 845
7. Mauricio Macri (@mauriciomacri): 724

Se guardaron también **4 mil** *tweets* de los supernodos junto con los *id* de los retransmisores. Este tipo de dato es fundamental para el análisis de la robustez de la red de propagación del contenido generado.

Relaciones entre usuarios

Se seleccionaron supernodos y se extrajeron los seguidores de aquellos usuarios que habían retransmitido *tweets* de estos. El set de datos resultante se compone de **32 millones de usuarios** con **110 millones de relaciones** entre ellos.

Con este set de datos posteriormente se construye lo que denominaremos *grafo inducido del tweet*.

2.9. Tipos de Usuarios

En la presente tesis se han definido dos tipos de usuarios: *supernodos* y *consumidores*. El primer tipo corresponde a aquellos usuarios que generan contenido y que tienen en proporción más seguidores (lazo entrante) que amigos (lazo saliente). Con la presente definición como punto de partida para separar usuarios es posible inferir que aquellos denominados supernodos son los que generan contenido de interés para una gran cantidad de usuarios. Estos últimos (usuario-consumidor) deciden suscribirse a los supernodos.

Aleatoriamente se almacenaron 4 millones de usuarios junto con su número de *Seguidores* y *Amigos*.

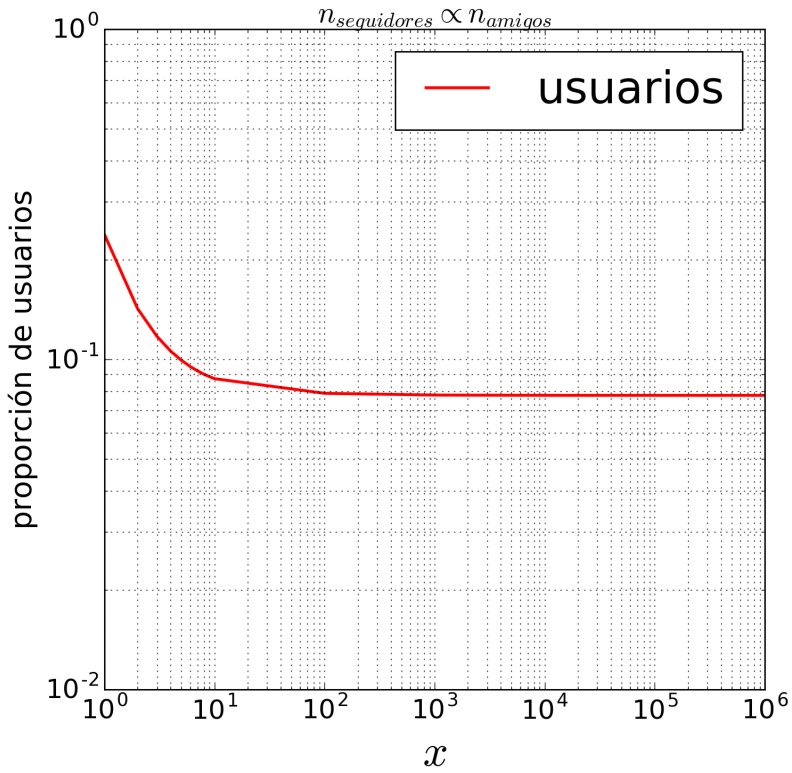


Figura 2.10 – ***Supernodo versus usuario regular.*** *Con 4 millones de usuarios se calculó la proporción de usuarios que tienen x veces más seguidores que amigos. Esta primera definición de supernodo discrimina entre aquellos potenciales generadores de contenido (muchos seguidores) de aquellos potenciales consumidores (muchos amigos).*

En la figura 2.10 se calculó la proporción de usuarios que tienen k veces más seguidores que amigos. Se puede ver que el 24 % de los usuarios poseen la misma cantidad de seguidores que de amigos, y luego desciende hasta alcanzar aproximadamente el 8 % que tienen 100 veces o más seguidores que amigos.

2.10. Descomposición en K Núcleos

El concepto de k capas en grafos ha sido introducido por Seidman en 1983[Seidman, 1983], cuya definición dice:

" 1 k-núcleo de un grafo G es el subgrafo maximal H tal que todos los vértices tengan grado al menos k. Esto implica que cada nodo del k-núcleo H tendrá al menos k vecinos dentro del grafo H. "

Para el presente trabajo se ha utilizado esta definición para obtener y graficar la descomposición de k-núcleos de lo que luego será definido como el grafo inducido por los *tweets*.

Capítulo 3

Resultados

3.1. Topología

La caracterización de la topología se basa principalmente en identificar la distribución de probabilidad del número de lazos dentro del grafo, en este caso dirigido. Los lazos entrantes a los nodos son denominados *Seguidores* y los salientes, *Amigos*. Este tipo de grandes redes poseen distribuciones de lazos según leyes de potencia [Barabási and Albert, 1999]. A continuación se enuncia la ley general detrás de las llamadas redes *scale-free*.

$$p(k) \sim k^{-\alpha} \tag{3.1}$$

Para la estimación de los difrentes parámetros se han utilizado las metodologías descriptas en [Clauset et al., 2009]. En la figura 3.1 hemos graficado la cantidad de usuarios en función de la cantidad de seguidores y amigos que se encontraron en la red. Ambas escalas son logarítmicas con el fin de mostrar las características del set de datos.

Para los diferentes cálculos se ha utilizado la librería *powerlaw* implementada en Python [Alstott et al., 2013].

El ajuste a una distribución de un set de datos es complejo y depende de múltiples variables y supuestos. Para el conjunto de datos en cuestión fue imposible

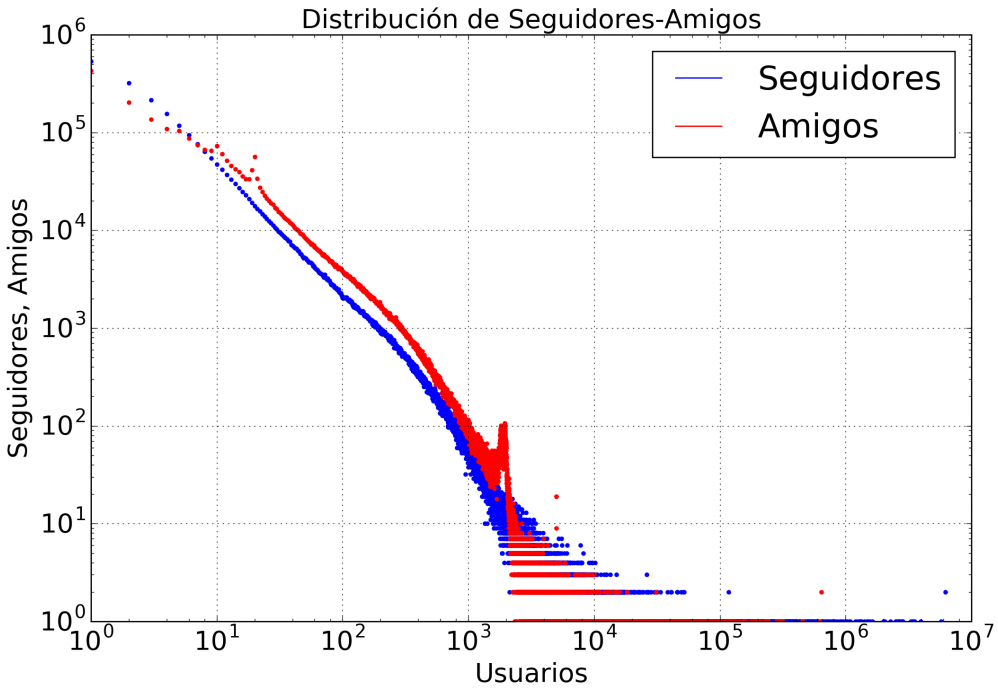
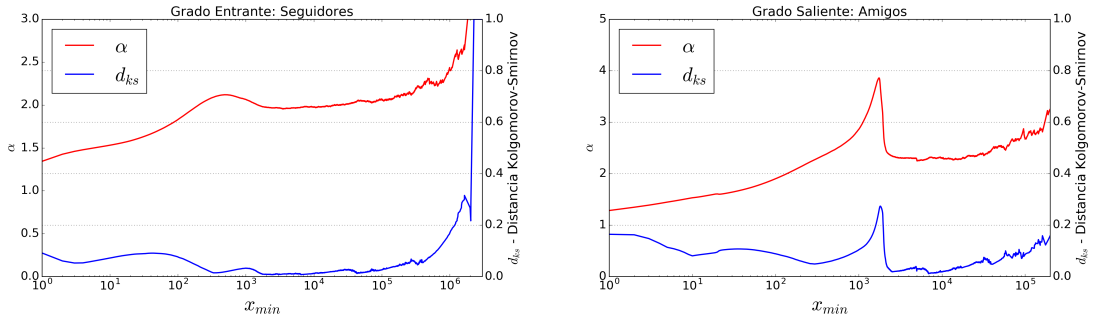


Figura 3.1 – ***Cantidad de usuarios según cantidad seguidores o amigos***. *Se relevó el número de seguidores y amigos de 3 millones de usuarios. Se graficó cantidad usuarios según cantidad de seguidores (grado entrante, azul) y cantidad de amigos (grado saliente, rojo).*

identificar la distribución subyacente y esto ocurre en general debido a las limitaciones impuestas por las técnicas del muestreo y/o características de los parámetros intrínsecos.

Para los casos enunciados, hemos sin embargo identificado el parámetro α de una distribución *Power Law* en función de la distancia de Kolgomorov-Smirnov (D_{KS}) tomando en consideración x_{min} y x_{max}. Se ha graficado el valor de α y D_{KS} en función del set de datos acotado al x_{min} fijo. La imagen 3.2 muestra los diferentes valores de los parámetros en función del x_{min}.

En el caso de los seguidores, hemos encontrado el mínimo absoluto de $D_{KS} = 0{,}0093$ en $x_{min} = 3{,}1k$ resultando un $\alpha = 1{,}96$. A su vez, se han analizado los puntos $x_{min} = 400$ que arroja $\alpha = 2{,}11$ y $D_{KS} = 0{,}0147$ y $x_{min} = 300k$ con $\alpha = 2{,}16$ y $D_{KS} = 0{,}0524$, siendo estos mínimos locales. El objetivo de buscar un mínimo local menor al absoluto es poder incorporar más usuarios al set de datos aunque



(a) **(b)**

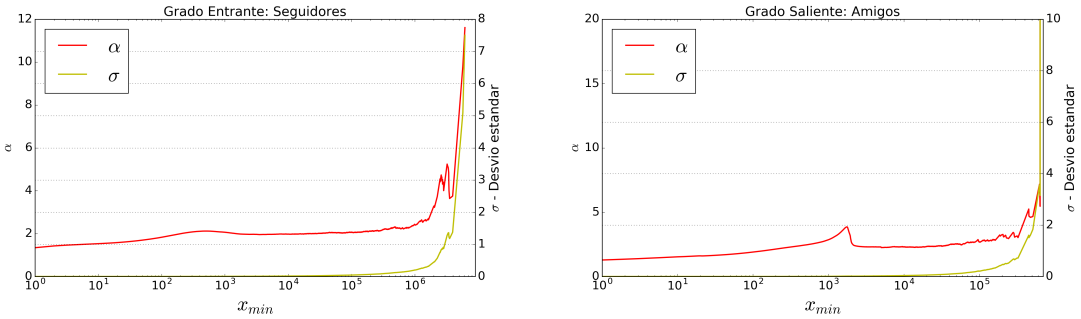
Figura 3.2 – ***Parámetro α y Distancia de Kolgomorov-Smirnov (KS).*** ***(a).*** *Con 2,7 millones de usuarios se calculó el parámetro α y la distancia de KS de la distribución estadísticas power law que mejor ajusta al dataset referido al grado entrante. Se utilizaron valores de x_{min} en el rango de 1 a $1M$ seguidores. El mínimo absoluto se encontró en ($x_{min} = 3{,}1k$, $\alpha = 1{,}96$, $D_{KS} = 0{,}0093$), el primer mínimo local en ($x_{min} = 400$, $\alpha = 2{,}11$, $D_{KS} = 0{,}0147$) y el último minimo local en ($x_{min} = 300k$, $\alpha = 2{,}16$, $D_{KS} = 0{,}0524$). Entre los mínimos locales, incluyendo el absoluto, los valores de α y D_{KS} se mantuvieron acotados.* ***(b).*** *Con $3{,}1M$ de usuarios se realizó el mismo procedimiento pero en este caso con el grado saliente. Se utilizaron valores de x_{min} en el rango de 1 a $100k$ amigos. El mínimo absoluto se encontró en ($x_{min} = 7{,}9k$, $\alpha = 2{,}30$, $D_{KS} = 0{,}0149$), el primer mínimo local se encontró en ($x_{min} = 300$, $\alpha = 2{,}28$, $D_{KS} = 0{,}0489$) y el segundo mínimo local en ($x_{min} = 2{,}5k$, $\alpha = 2{,}34$, $D_{KS} = 0{,}0187$). Con ambos grados la resolución de x_{min} es de 10 puntos por década.*

aumentado D_{KS}. En el caso de los amigos, hemos encontrado el mínimo absoluto de $D_{KS} = 0{,}0149$ en $x_{min} = 7{,}9k$ resultando un $\alpha = 2{,}30$. A su vez, se han analizado los puntos $x_{min} = 300$ que arroja $\alpha = 2{,}28$ y $D_{KS} = 0{,}0489$ y $x_{min} = 2{,}5k$ con $\alpha = 2{,}34$ y $D_{KS} = 0{,}0187$, siendo estos mínimos locales.

Más allá de que hayamos podido encontrar parámetros de distribuciones asociadas (ley de potencia), los grandes valores de x_{min} junto con D_{KS} produce que el ajuste no sea representativo del proceso subyacente.

Finalmente, se ha calculado el desvío estándar del grado entrante de los usuarios (seguidores). El cálculo lo hemos realizado en función del porcentaje de usuario incluidos y con una resolución del 1 % en cada paso.

Tanto en la figura 3.4, como en múltiples otras realizaciones, se ha podido apreciar una fluctuación enorme del desvío evidenciando que el mismo depende de los datos que fueron utilizados para el cálculo. Al no obtener una convergencia del mismo, y por consiguiente tampoco de la varianza, hemos podido definir la distribución de probabilidad asociada al proceso como de cola pesada. Este resultado comple-



(a) **(b)**

Figura 3.3 – ***Valores de*** α ***y*** σ ***en función de*** x_{min}

Luego de calcular y analizar los datos obtenidos de 3 millones de usuarios, se observó que la distribución no refleja una ley de potencia pura. ***Figura (a)****. Se observó que ambos parámetros divergen para el rango de Seguidores en el set de datos. Incluso se observó que* σ *es creciente para cualquier rango acotado de Seguidores que se tome.* ***Figura (b)****. En el caso de los Amigos (grado saliente), se observó también que ambos parámetros divergen para el rango definido. En este caso también* σ *es creciente para cualquier rango que se tome.*

menta lo evidenciado en la figura 3.1 con el histograma de los grados.

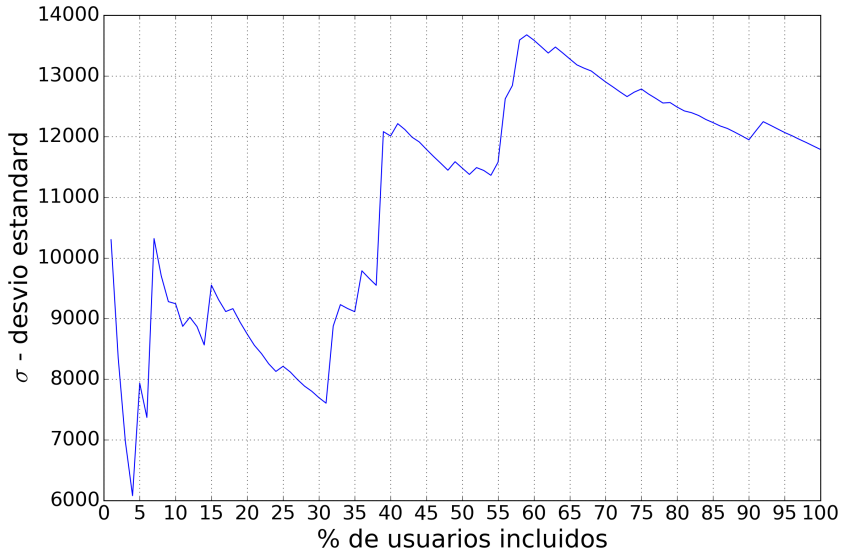


Figura 3.4 – ***Cálculo de*** σ ***en función del porcentaje de usuarios.*** *Se cálculo* σ *en función de la cantidad de usuarios incluidos en el set de datos. Los mismos fueron tomados al azar con una resolución de* 1 % *en cada cálculo. Se observó una variación total enorme entre los diferentes valores obtenidos. Un ejemplo de esto se evidenció entre los porcentajes* 4 %, 7 %, 31 % *y* 59 %. *Tanto el desvío, cómo la varianza, son dependientes del set de datos utilizado.*

3.2. Distancia entre supernodos

En trabajos previos tales como [Rosenman, 2012] se demostró que las métricas definidas para un grupo heterogéneo de usuarios pueden variar sus valores según la actividad que este grupo desarrolla. En el presente trabajo nos hemos enfocado en el grupo de *políticos de la República Argentina durante el año 2015* con el sólo propósito de aplicar diferentes técnicas de procesamiento al gran volumen de datos generado a lo largo del período. Se ha elegido un grupo característico donde todos los usuarios realizan la misma actividad. Para este análisis se incluyó al grupo y a todos su seguidores. Dentro del grupo de seguidores hay algunos que lo hacen a más de un supernodo a la vez. Hemos calculado la cantidad de usuarios según a cuantos supernodos se encuentran suscriptos. En la figura 3.5 se observa la distribución de usuarios en función de las suscripciones. El set de datos ha sido compuesto por 5,7 millones de usuarios únicos, donde 3,81 millones siguen solamente uno de los supernodos.

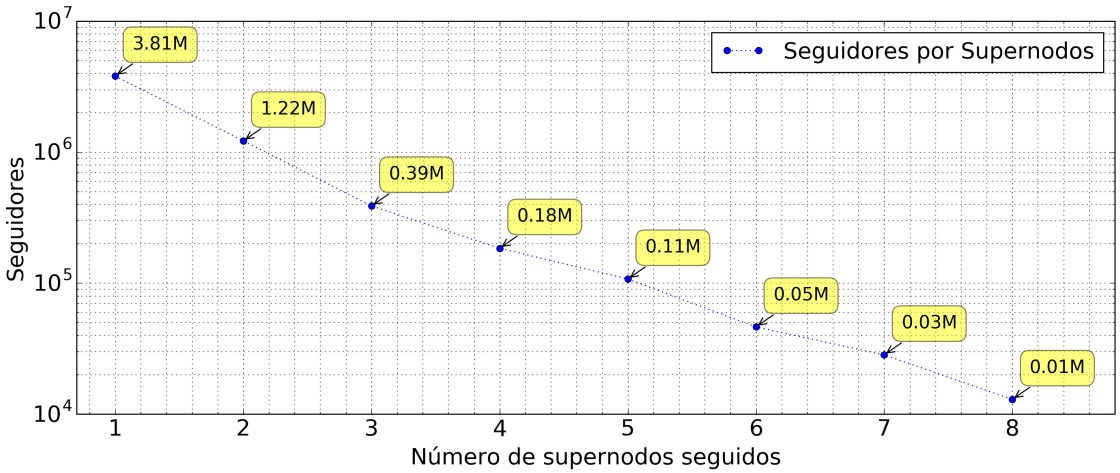


Figura 3.5 – ***Número de seguidores según su cantidad de suscripciones***. *Se definió al grupo de* ***supernodos*** *con los siguientes políticos: Massa, CFK, Stolbizer, Scioli, Fernandez, Carrio, Macri. A través de la API de* ***Twitter*** *se extrajeron los seguidores de cada uno, generándose un set de datos compuesto por* 5,7 *millones de usuarios únicos que siguen a uno o más supernodos. El* 87 % *de los usuarios siguen a 1 o 2 supernodos solamente. Escala logarítmica en el eje de ordenadas.*

Se ha planteado la posibilidad de medir una distancia entre los distintos supernodos, para ello nos hemos centrado en la distribución de seguidores propios y

compartidos calculando cuan similares son estos grupos entre todos ellos dando una idea de alineación política de los seguidores [Makazhanov and Rafiei, 2013, Conover et al., 2012].

En caso de que A y B sean grupos disjuntos, se define $J(A,B) = 0$ y $J(A,B) = 1$ cuando ambos grupos son idénticos. El índice es estrictamente acotado en el rango $[0,1]$. Cuando el índice es cercano a uno indica que los grupos comparten gran cantidad de usuarios y son grupos similares. Esto conlleva una distancia cercana a cero.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{3.2}$$

$$0 \leq J(A,B) \leq 1$$

$$D(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \tag{3.3}$$

Para el análisis se definió el siguiente grupo reducido de supernodos:

- Sergio Massa (SM)
- Mauricio Macri (MM)
- Daniel Scioli (DS)

Se tomaron los seguidores de cada uno de éstos y se han calculado el índice y la distancia de Jaccard según las ecuaciones 3.2 y 3.3.

La figura 3.6 muestra la cantidad de suscriptores que tienen los supernodos Macri (2,5 millones) y Scioli (1,14 millones) así como la cantidad que tienen en común (456 mil). Hemos incluido también los seguidores de Massa.

Se ha computado el índice y la distancia de Jaccard entre Macri y Scioli de acuerdo a las ecuaciones 3.4 y 3.5

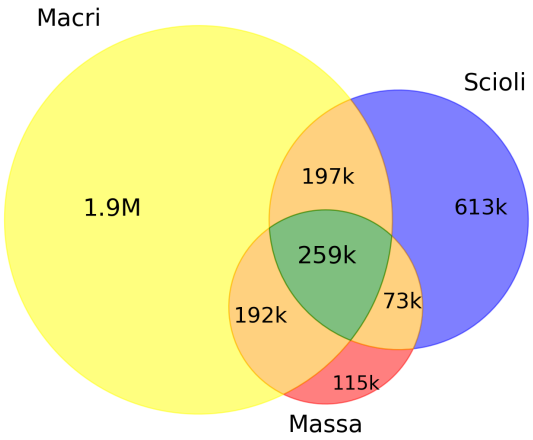


Figura 3.6 – ***Seguidores de Macri, Massa y Scioli.*** *Se observó la cantidad de usuarios propios y compartidos entre los candidatos del grupo de reducido de supernodos. El número de usuarios que siguen a los tres supernodos es mayor que la cantidad de seguidores que sólo siguen a dos: el área verde es mayor que cualquiera de las naranjas. La cantidad de seguidores del supernodo Macri triplica la cantidad de seguidores propios de Scioli.* 2,8 *millones de usuarios en el set de datos no siguen a ninguno de estos tres supernodos.*

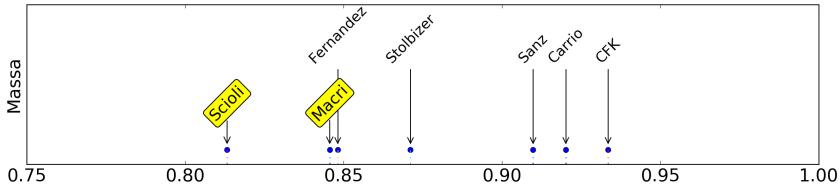
$$J(MM, DS) = \frac{|S_{MM} \cap S_{DS}|}{|S_{MM} \cup S_{DS}|} = \frac{197k + 259k}{2{,}5M + 1{,}14M - (197k + 259k)} = 0{,}14 \tag{3.4}$$

$$D(MM, DS) = 1 - J(MM, DS) = 1 - 0{,}14 = 0{,}85 \tag{3.5}$$

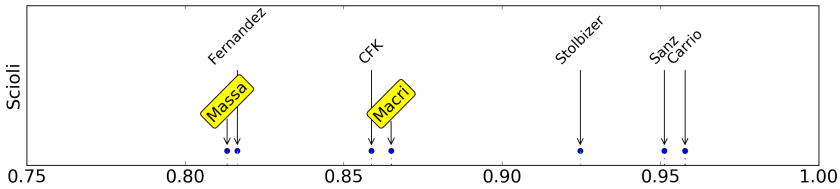
Utilizando el mismo procedimiento se ha calculado y graficado la distancia entre todos los supernodos. En la figura 3.7 se muestran las distancias desde los supernodos del grupo reducido al resto del grupo. En la 3.7a hemos observado que los seguidores de Massa se encuentran más cerca de Scioli que de Macri. En cambio, en la figura 3.7b se observa que los seguidores de Scioli se encuentran cercanos a Massa, Aníbal Fernandez y Cristina Fernandez de Kirschner antes que a Macri. Partiendo desde Macri, figura 3.7c, se observa que Massa y Cristina Fernandez de Kirchner, son los candidatos más cercanos en cuanto a seguidores compartidos, luego aparece Daniel Scioli y el resto de los supernodos.

A partir de lo descripto, se observa que los grupos de seguidores más distantes son los de Macri y Scioli mientras que los más cercanos son los de Massa y Scioli. Fuera del contexto de los seguidores de ***Twitter*** los dos primeros supernodos corresponden

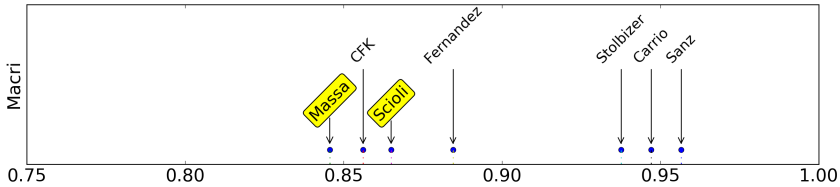
a agrupaciones políticas con orígenes distintos mientras que los dos segundos se encuentran en agrupaciones con un origen en común.



(a) $D_J(SM, \bullet)$ *hasta Scioli y Macri.*



(b) $D_J(DS, \bullet)$ *hasta Massa y Macri.*



(c) $D_J(MM, \bullet)$ *hasta Massa y Scioli.*

Figura 3.7 – ***Distancia de Jaccard***. *Se computó la distancia de Jaccard de 3 supernodos a los demás. El cómputo se realizó el día 9 de Septiembre de 2015.* $D(x, y)$ *es variable y depende de la cantidad de seguidores propios y compartidos con el resto del grupo. Solo se representó la distancia entre Scioli, Massa y Macri; el resto de los puntos corresponden a los demás supernodos.* $D_J(DS, SM) = 1 - J(DS, SM) = \frac{|S_{DS} \cup S_{SM}| - |S_{SM} \cap S_{DS}|}{|S_{SM} \cup S_{DS}|}$. S_{DS} *representa los seguidores de Scioli,* S_{SM} *representa los seguidores de Massa.* ***figura (a)***. *Computando* $D_J(SM, \bullet)$ *se apreció que se encuentra más cerca de Scioli que de Macri.* ***figura (b)***. *Computando* $D_J(DS, \bullet)$ *se apreció que se encuentra más cerca de Massa que de Macri.* ***figura (c)***. *Computando* $D_J(MM, \bullet)$ *se apreció que se encuentra más cerca de Massa que de Scioli. El término de cercanía comprende la masa de seguidores propia y en común con otro supernodo.*

3.3. Relación Retweet / Menciones

Los *tweets* (TW) son generados y propagados a través de la red por los diferentes usuarios. Existen múltiples formas de retransmitir (RT) un mensaje. En este trabajo nos hemos enfocado en la retransmisión a través de la función propia de la red[1]. El usuario que decide retransmitir un mensaje, al hacerlo, crea un nuevo *tweet*, pero no agrega información al mensaje original. Sobre este punto se ha desarrollado parte del análisis del presente capítulo: *retransmitir información sin alterar el contenido*. También, nos hemos enfocado en la información subyacente que se genera cuando un usuario menciona (@) a los supernodos. Ambos casos fueron utilizados para identificar alineación política de los usuarios con respecto a los supernodos.

El alcance del presente trabajo no es identificar el tipo de contenido ni el por qué de la retransmisión pero sí capitalizar la información desde el punto de vista de qué pretende realizar el usuario con dicha acción. En el caso de la retransmisión es directa la asociación de afinidad entre el retransmisor y el supernodo ya que el contenido no es modificado en el proceso (y suponemos que los *tweets* de los supernodos son positivos para ellos mismos).

En el caso de las menciones, el usuario tiene un grado de libertad mayor con respecto a la retransmisión ya que puede mencionar a un usuario dentro de un *tweet*, cualquiera sea el contenido del mismo.

Se han analizado 54 millones de *tweets* en busca de RT y menciones al grupo reducido de supernodos. Los usuarios pueden escribir miles de *tweets*, a su vez, pueden retransmitir otros o bien mencionar a otros usuarios. Con el fin de normalizar los datos de los diferentes usuarios, las ecuaciones 3.6 y 3.7 definen la normalización de RT y menciones en función de los TW escritos por cada usuario.

$$r_{RT} = \frac{RT_{usuario}}{TW_{usuario}} \tag{3.6}$$

[1]Los RT clásicos son aquellos generados por ***Twitter***. Al retransmitir, el nuevo *tweet* comienza con la sigla RT seguida por una mención (@) al usuario original.

$$r_{ME} = \frac{Menciones_{usuario}}{TW_{usuario}} \tag{3.7}$$

Esto normaliza las métricas en el rango de $[0, 1]$. En el caso cercano a cero indica que el usuario nunca ha RT o mencionado al supernodo. En contraposición, cercano a uno, indica que la mayoría de los *tweets* generados por el usuario son RT o que contienen menciones a alguno de los supernodos.

Nos hemos enfocado en 305 mil RT de los supernodos. Las figuras 3.8 (1a), (1b) y (1c) muestran como se distribuyen los usuarios según cuantos RT han retransmitido. Se ha observado que estos se distribuyen sobre ambos ejes dejando pocos usuarios que retransmiten contenido de más de un supernodo. En el caso de las figuras 3.8 (2a), (2b) y (2c) se observa una mayor dispersión alrededor de la coordenada $(0, 0)$. En este caso el usuario escritor del *tweet* tiene el grado de libertad de generar el contenido que más le interese y asociarlo al supernodo que quiera a través de las menciones (uso del @YYY). Al hacerlo, el contenido puede ser positivo, neutro o negativo respecto del usuario al que menciona. Se ha supuesto que el uso de las menciones dentro de la red es más amplio en el sentido de refuerzo o rechazo que la retransmisión del contenido de otro usuario. En este sentido se espera que un usuario alineado ideológicamente al supernodo YYY, retransmita mucho contenido generado por este, lo mencione positivamente y mencione negativamente al supernodo ZZZ (opositor a YYY). Como se describió al principio del capítulo la idea del presente trabajo no es analizar el contenido de los *tweets* sino extraer conclusiones acerca del comportamiento de los usuarios.

En conclusión, *Retransmitir contenido de un usuario refuerza el contenido original mientras que mencionar al supernodo escritor no da indicios acerca de la afinidad del propio usuario por ese supernodo.*

Para desarrollar el propuesto anterior, hemos analizado las menciones en función de la cantidad de RT que cada usuario transmitió. Solamente estamos enfocados en el análisis de las retransmisiones puras[2] de ***Twitter*** junto con las menciones.

[2]Un usuario podría escribir un *tweet* que comience con *'RT @YYY ...'* seguido del texto original pero esto no será considerado un retweet para la red.

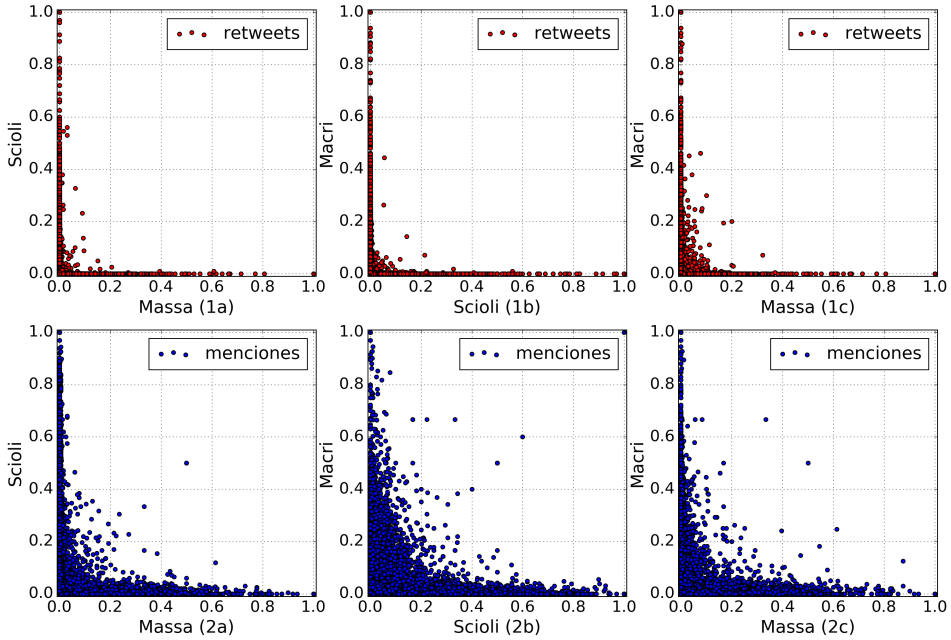


Figura 3.8 – ***Número de menciones y retrasmisiones de los supernodos tomados de a pares.*** *Para el siguiente análisis el dataset se conformó con* $54M$ *de tweets (TW) provenientes de* $239k$ *usuarios. Existen* $207k$ *menciones al supernodo @SergioMassa,* $550k$ *a @danielscioli,* $587k$ *a @mauriciomacri. Se mencionó a los supernodos* $1{,}3M$ *de veces,* $2{,}4\,\%$ *sobre* $55M$ *de menciones totales. En cuanto a las retrasmisiones, retweets (RT), se realizaron* $305k$ *a tweets de los supernodos. En la primer fila de gráficos, cada punto representa el cociente de RT a cada supernodo sobre el número total de tweets escritos por el usuario:* $RT_{usuario}/TW_{usuario}$.
Si el punto se encuentra cercano a 1 indica que el usuario retransmitió contenido en vez de generarlo, caso contrario, el usuario generó mucho contenido y retrasmitió poco. En todas las figuras un punto sobre cualquiera de los ejes significa que el usuario sólo retransmitió o mencionó a ese supernodo y nunca a otro. En la primer fila de gráficos los usuarios se concentran sobre los ejes, en cambio, en la segunda se observó una mayor dispersión alrededor de la coordenada (0,0). ***Figura (1c)****. Se observó un incremento en la cantidad de usuarios que RT contenidos generados por ambos supernodos en contraposición con las figuras* ***(1a)*** *y* ***(1b)****.*

Al momento de RT un *tweet* de otro usuario en la red se genera un nuevo *tweet* copia fiel del original pero que comienza con *'RT @YYY ....'*: por cada RT también se genera una mención al supernodo. A fin de eliminar el número de menciones debidas a los RT, este valor se sustrae del total de menciones relevadas. La ecuación 3.8 muestra como hemos calculado las coordenadas de los diferentes usuarios para luego ser graficados en la figura 3.9 respecto de la cantidad de RT y menciones puras a los supernodos.

$$(x, y) = \left(\frac{RT}{TW}, \frac{@ - RT}{TW} \right) \tag{3.8}$$

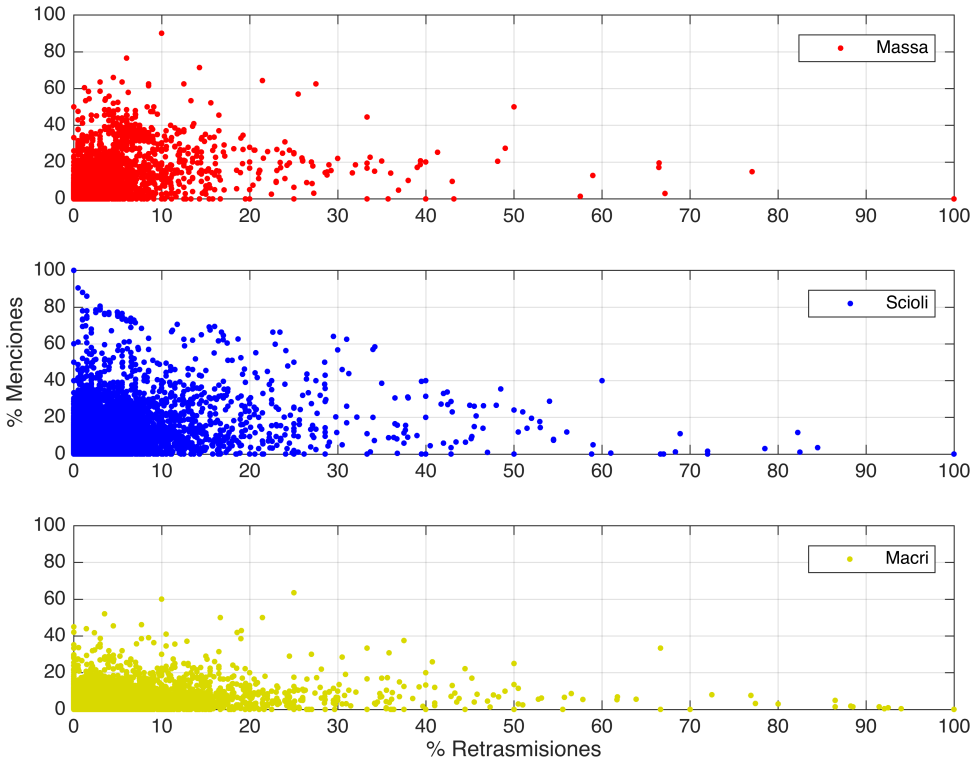


Figura 3.9 – ***Porcentaje de menciones en función del porcentaje de retrasmisiones para cada supernodo.*** *Se graficó el número de menciones en función del número de retransmisiones. En* ***Twitter****, por cada RT se genera también una mención. Se ajustó el valor quitando de las menciones el valor de los RT para mostrar las menciones puras.* ***Figuras (b)*** *y* ***(b)****. Se observó que los puntos azules se concentran en el rango* $(RT, @) = ([0, 10\,\%]; [0, 40\,\%])$ *mientras que los puntos amarillos lo hacen* $(RT, @) = ([0, 20\,\%]; [0, 20\,\%])$*. Los usuarios son más propensos a retrasmitir contenido de Macri y a mencionar a Scioli. Cada punto se definió de acuerdo a la siguiente normalización:* $(x, y)_{supernodo} = \left(\frac{RT}{TW}, \frac{@-RT}{TW} \right)_{supernodo}$.

En la figura 3.9 se observa que para el supernodo Scioli aparece una gran concentración de puntos azules en el rango $[0; 10\,\%]$ de RT y en el rango $[0; 40\,\%]$ de menciones, indicando que muchos usuarios mencionan al supernodo más que la retransmisión de su contenido. En el caso de Macri, observamos que el rango de RT y las menciones ocurren aproximadamente entre $[0; 20\,\%]$. En este caso los usuarios retransmiten contenido en la misma cantidad que lo mencionan.

Tomando en consideración que el RT no agrega información, es más fuerte la premisa de retransmisión que la mención. Por consiguiente se concluye que la información de Macri suele ser más aceptada dentro del set de datos analizado que aquella generada por Scioli.

3.4. Clasificador de seguidores

Una vez analizado cómo los usuarios mencionan y retransmiten contenido en la red, nos hemos enfocado en proponer una forma de clasificar su afinidad por alguno de los supernodos. Se han definido tres clases disjuntas donde cada una representa a los supernodos del grupo reducido (Macri, Scioli y Massa). El set de datos está compuesto por 56 mil usuarios que retransmitieron (r_{RT}) y/o mencionaron (r_{Me}) a uno o más de los supernodos.

Para diseñar el clasificador lineal hemos dividido al grupo de usuarios en dos sub-grupos: 50 mil y 6 mil usuarios pertenecientes al grupo de *entrenamiento* y *validación* respectivamente.

Se crearon dos clasificadores lineales, uno con información de los RT y el otro con información de las menciones.

Como regla general se ha definido la afinidad de un usuario a un supernodo tomando el mayor número de RT entre los tres supernodos. Esto define el grupo al que pertence este usuario. A continuación se han definido los vectores generados provistos al clasificador lineal para identificar el grupo de pertenencia de un usuario.

$$r_{RT}{}^{Macri} \mid r_{RT}{}^{Scioli} \mid r_{RT}{}^{Massa}$$

$$r_{ME}{}^{Macri} \mid r_{ME}{}^{Scioli} \mid r_{ME}{}^{Massa}$$

Para calcular la performance se ha repetido el proceso de selección de usuarios para los sub-grupos de entrenamiento y validación 500 veces. En cada ciclo los usuarios fueron elegidos aleatoriamente entre todo el set de datos[3]. Se ha graficado la mediana de las 500 realizaciones junto con el desvío estándar (barra de error) para los RT y para las menciones.

Hemos observado que al utilizar la retransmisión de *tweets*, el clasificador lineal tiene una performance $p = 0{,}9955$ con un error estándar del *bootstraping* de $\sigma =$

[3]En este caso es selección con remoción del set de datos, un mismo usuario no puede ser elegido para el grupo de entrenamiento y de prueba.

0,002 lo que da una elevada performance en el proceso de selección con un desvío pequeño.

Al comparar RT con menciones, el clasificador lineal disminuye notoriamente su performance a $p = 0{,}8701$ con un error estándar del bootstraping de más del doble respecto de los RT, $\sigma = 0{,}005$.

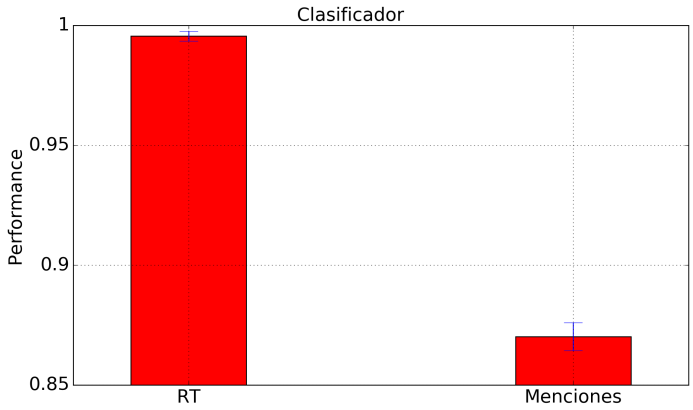


Figura 3.10 – ***Performance de un clasificador lineal con 3 clases*** *Los* 305 *mil retweets (RT) fueron escritos por* 56 *mil usuarios, las* 54 *millones menciones (@), por* 2,5 *millones de usuarios. Para clasificar la afinidad de un usuario a los supernodos se utilizó el máximo cociente* $RT_{usuario}{}^{supernodo}/TW_{usuario}$. *Se realizó el mismo procedimiento para las menciones:* $@_{usuario}{}^{supernodo}/TW_{usuario}$. *El set de datos se dividió en* 6 *mil usuarios aleatorios como grupo de muestreo y* 50 *mil usuarios como grupo de entrenamiento. Se repitió el proceso 500 veces regenerando los grupos aleatoriamente y almacenando el error de clasificación para cada realización. Se graficó la performance del clasificador para las retransmisiones con un valor de* $p = 0{,}9955$ *con un desvío estándar de* $\sigma = 0{,}002$. *En el caso del las menciones se observó* $p = 0{,}8701$ *y* $\sigma = 0{,}005$. *Se observó que la performance del clasificador disminuye y aumenta el desvío estándar utilizando las menciones en vez de las retransmisiones.*

La figura 3.10 muestra la performance de un clasificador lineal utilizando distancia de Mahalanobis [Mahalanobis, 1936].

Entonces, teniendo información acerca de qué retransmite un usuario podemos inferir con bajo error la alineación del mismo al grupo de supernodos. Sin embargo el uso de las menciones no resulta tan efectivo.

3.5. Evolución Temporal

Habiendo definido la topología de la red junto con la forma en que los usuarios propagan información, hemos analizado la evolución temporal del contenido en función de cuántos RT y favoritos (FAV) acumulan los *tweets* con el correr del tiempo.

Se ha definido como tiempo inicial al momento en el cual el contenido es generado y está disponible en la red[4].

Se ha construido el set de datos con 24 mil *tweets* escritos por los supernodos hasta el día 9 de Marzo de 2016. Para este análisis se ha tomado el grupo completo de supernodos a fin de poder obtener más objetos para analizar. Debido a limitaciones del API de ***Twitter*** sólo se permite la extracción de los últimos 3 mil *tweets* generados por cada usuario.

Al igual que con la topología, se ha caracterizado la dinámica con la que se propaga el contenido en la red y definido *estado estacionario* a aquel estado donde un *tweet* no obtiene más RT ni FAV habiendose estabilizado en un valor determinado.

En las figuras 3.11a y 3.11b se puede observar distribuciones de probabilidad de retransmisión de *tweets* complejas y con más de un proceso subyacente. Típico comportamiento de procesos de cola pesada. La figuras 3.11c y 3.11d muestran las distribución de probabilidad acumulada complementaria (CCDF por sus siglas en ingles) del set de datos y junto con la comparación de ajustes con diversos tipos de distribuciones. El ajuste de una distribución es complejo y depende del proceso y de las técnicas del muestreo, nuevamente la distribución que gobierna el proceso es de cola pesada [Bild et al., 2015b].

Del set de datos hemos seleccionado al azar tres *tweets* y graficado la evolución del número de RT y favoritos en el tiempo con el fin de mostrar la dinámica desde que nacen hasta que se estabilizan: figura 3.12a. Se ha observado que estas curvas tienden a ser crecientes, aún cuando existe la posibilidad de eliminar un RT o remover la marca de favorito.

[4]Cuando un usuario escribe un *tweet*, retransmite o marca como favorito contenido y/o menciona a otro usuario.

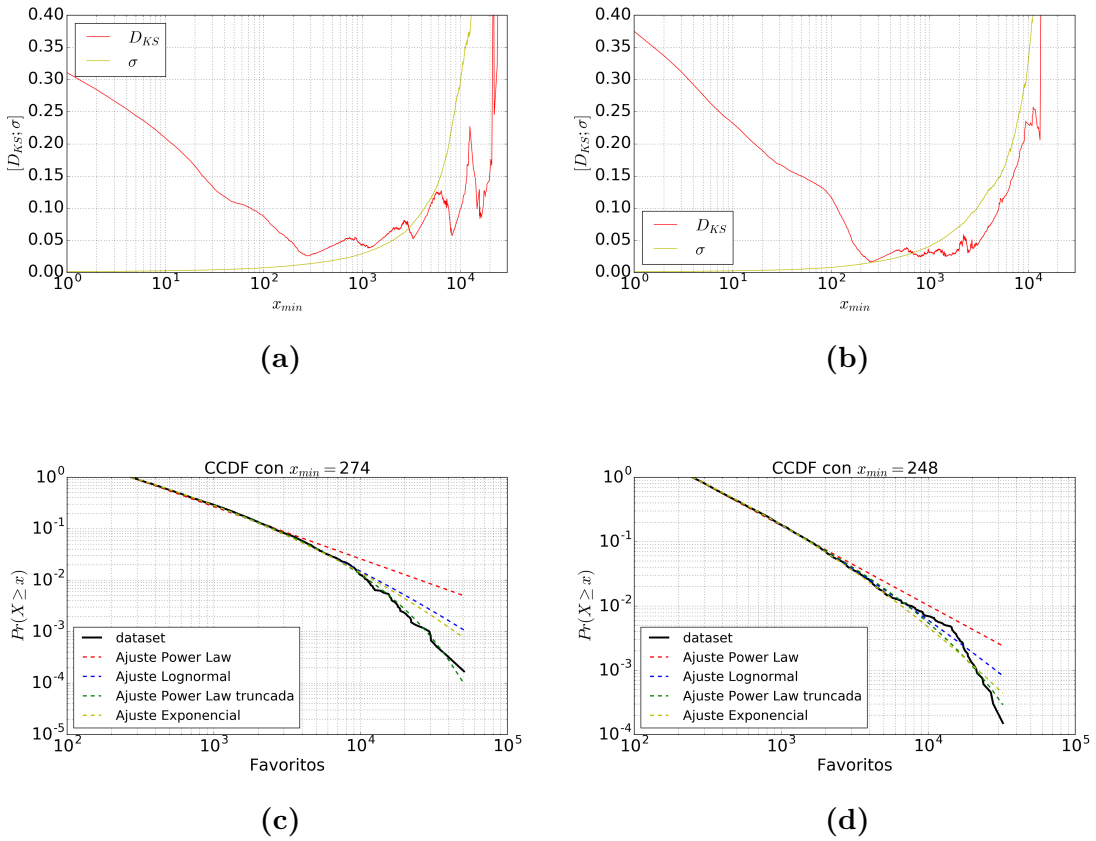
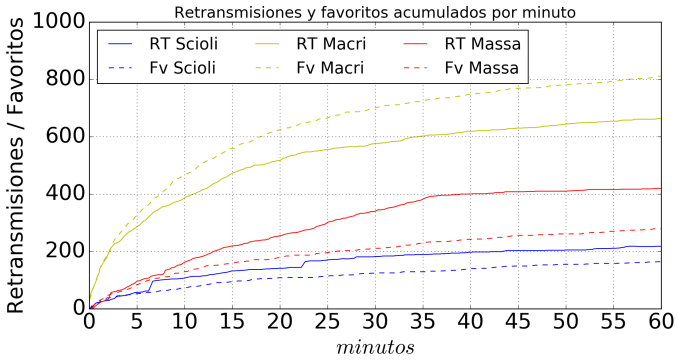


Figura 3.11 – ***CCDF de número de retransmisiones (RT) y favoritos (FAV) de los tweets***. *Se calculó la función de distribución de probabilidad acumulada complementaria (CCDF por sus siglas en inglés) para RT y FAV. Se conformó el set de datos con 24k tweets escritos por los supernodos hasta el día 9 de Marzo de 2016. De acuerdo a las limitaciones del API de* ***Twitter*** *sólo es posible extraer los últimos 3k tweets de cada usuario. Se calculó D_{KS}, σ y α de una distribución Power Law. Se graficaron distribuciones LogNormal, PowerLaw truncada y Exponencial para poder observar el mejor ajuste al set de datos propuesto con las consideraciones de x_{min} y σ.* ***Figuras (a) y (b)***. *Se observó que tanto en la retransmisión como en los favoritos el desvío estándar diverge llevando a un análisis nuevamente de una distribución de cola pesada.* ***Figura (c)***. *Se utilizó $x_{min} = 274$ (mínimo absoluto) y la distribución que mejor se ajusta es una power law truncada con $\alpha = 2{,}02$ para el modelo de retransmisiones de tweets.* ***Figura (d)***. *En el caso de las retransmisiones, Se observó que el mejor ajuste también con un modelo power law truncado con $x_{min} = 248$ y $\alpha = 2{,}24$.*

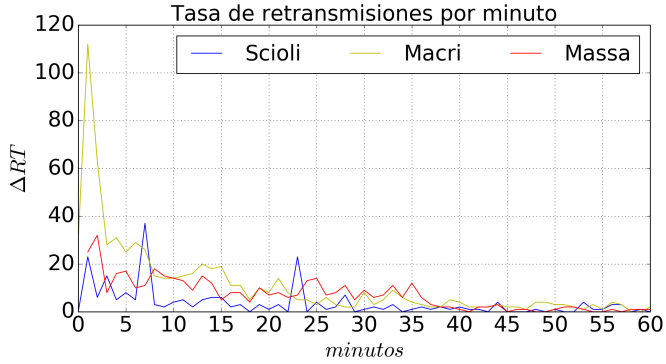
En la figura 3.12b se observa que la vida útil del *tweet* no es más que unos minutos desde el momento en que fue escrito[5]. Luego comienza a decrecer el impacto del mismo hasta caer en el olvido: cuando ningún otro usuario en la red lo retransmite ni lo marca lo como favorito.

[5]Definimos *tiempo de vida* como aquel tiempo transcurrido desde que el *tweet* es generado hasta que la tasa de crecimiento de RT o FAV alcanza su pico máximo y luego comienza a decrecer.

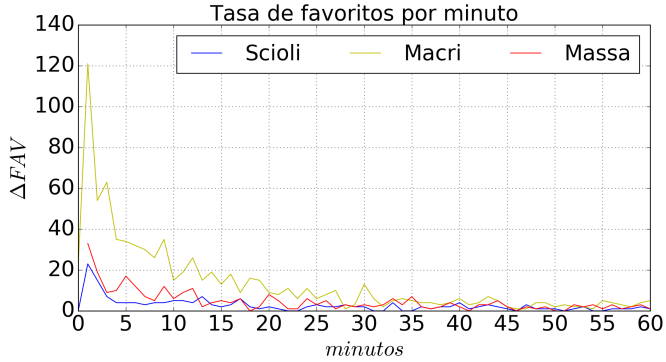
Al realizar este tipo de análisis y observar que el tiempo de vida es de apenas unos minutos, se ha analizado nuevamente el set de datos en búsqueda de aquellas horas donde mayoritariamente los supernodos escriben sus *tweets*. Se identificó que la mayoría de estos lo hacen alrededor de las 12pm y a las 8pm del día como se ilustra en la figura 3.13 que compara dicha actividad con la actividad de la red.



(a)



(b)



(c)

Figura 3.12 – ***Retransmisión de un tweet en el tiempo. Retransmisiones (RT) y Favoritos (FAV)***. *Se eligieron 3 tweets al azar, uno de cada supernodo, y se graficó la evolución del número de RT (línea continua), y la cantidad de FAV (línea punteada) en el tiempo.* ***Figura (a)***. *Estas funciones tienden a ser monótonas crecientes. En el caso de Scioli y Massa los tweet se observó FAV por encima de RT (línea contínua por encima de la punteada), y el caso opuesto en Macri.* ***Figuras (b) y (c)***. *En estas dos figuras se calculó la derivada temporal.* $f'[k] = f[k] - f[k-1]$ *dando así la velocidad de RT o FAV. Se observó también la rapidez con la que se extinguió un tweet. En el caso de Macri, Se hizo aproximadamente a los* 30 *minutos teniendo su pico máximo dentro de los primeros* 3*. En el caso de Scioli se observaron dos pico a los* 7 *y a los* 23 *minutos.*

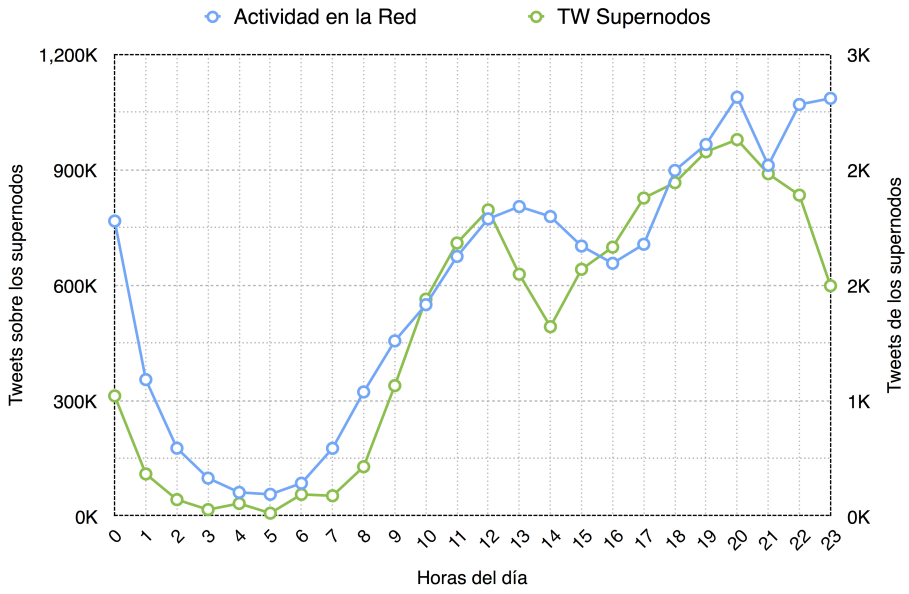


Figura 3.13 – ***Cantidad de tweets generados a lo largo del día***. *Se tomaron 24k tweets y se graficó la cantidad a lo largo de las horas del día. Se tomaron los contenidos generados hasta el día 9 de de Marzo de 2016.*

Como conclusión del presente análisis podemos decir que el tiempo de vida de los *tweets* es de algunos minutos, que su dinámica de RT y FAV es ser creciente y que el proceso que gobierna la distribución de probabilidad es de cola pesada. Por tales resultados es que los supernodos suelen escribir sus contenido en horarios donde la mayor audiencia se encuentra activa en la red.

3.6. Retweets de seguidores

Con el objetivo de cuantificar cuánto contenido es consumido por los seguidores y cuánto lo es por aquellos que hemos denominado *no-seguidores*, se han analizado los usuarios que retransmiten contenido junto con la relación que los une con los supernodos.

Partiendo del grafo de *seguidores* se ha definido S_{rx} como la cantidad de seguidores que retransmitieron el contenido sobre la cantidad del total de retransmisiones que ha obtenido un *tweet*.

$$S_{rx} = \frac{Seguidores}{Retransmisores} \tag{3.9}$$

Se ha conformado el set de datos con más de mil *tweets* donde se ha podido almacenar el *id* del usuario que ha retransmitido dicho contenido. A fin de obtener una muestra consistente, hemos definido como nivel mínimo de selección, a aquellos *tweets* donde se ha almacenado al menos el 75 % de los usuarios respecto del total del retransmisiones que obtiene dicho contenido[6].

En la figura 3.14 hemos graficado la distribución de probabilidad acumulada inversa para obtener un dado porcentaje de seguidores dentro del grupo de retransmisores un *tweet*.

Para los tres supernodos se ha podido apreciar con probabilidad cercana a uno que todos los *tweets* tienen más del 35 % de seguidores dentro del grupo de retransmisores.

Se ha observado que para el supernodo Macri, entre el segundo y tercer cuartil los mensajes obtienen entre un 90 % y un 93 % de seguidores del total de retransmisiones, indicando que con probabilidad mayor al 75 %, los *tweets* son propagados mayoritariamente por aquellos usuario a distancia uno del escritor, $D = 1$. El resto de los retransmisores se encuentran a una distancia mayor a uno de éste[7]. Se ha ob-

[6]Debido a una limitación de la API de ***Twitter*** solo es posible extraer los últimos 100 retransmisores de un *tweet*. Por tal motivo fue necesario obtener la lista de usuarios cada 10 segundos.

[7]Distancia 2 representa a un seguidor de un seguidor

servado este comportamiento ya que cuando un usuario retransmite contenido, todos sus seguidores son notificados, y estos pueden seguir propagándolo (aún cuando no son seguidores del autor del original). Para Scioli, hemos observado en el segundo y tercer cuartil un 82 % y un 87 %, por debajo del rango de Macri y finalmente se ha observado a Massa con un rango entre el 65 % y el 92 %, notoriamente por debajo de los otros dos candidatos.

En conclusión, La propagación o retransmisión de los *tweets* es mayoritariamente producida por los seguidores ($\sim 90\,\%$) de autor del mismo dejando un margen radicalmente menor ($\sim 10\,\%$) de retransmisores a una distancia mayor a uno ($D > 1$). Hemos podido apreciar entonces que el grafo inducido por el *tweet* posee un diámetro[8] pequeño alcanzando solamente un par de saltos.

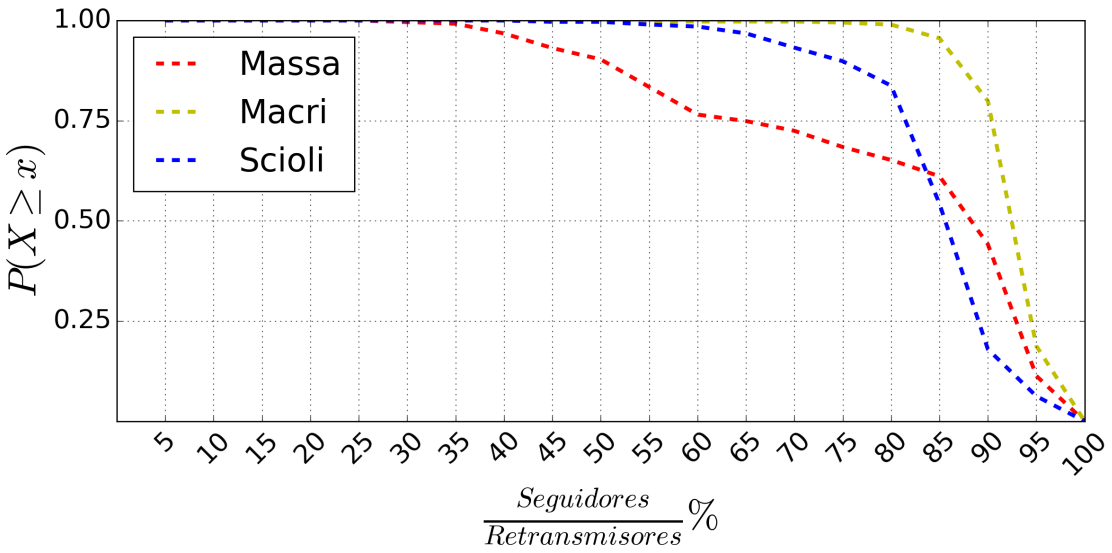


Figura 3.14 – ***Probabilidad Acumulada Inversa de la propagación en función de los seguidores*** *Se generó el set de datos con* 1222 *tweets de los supernodos en los cuales se almacenó al menos al* 75 % *de los retransmisores.* 592 *tweets corresponden a Scioli (*48 %*),* 383 *a Macri (*31 %*) y* 247 *a Massa (*21 %*). Se calculó el porcentaje de seguidores de los supernodos dentro de los retransmisores.* $S_{rx} = \frac{fo}{rx}$*. Se graficó la probabilidad acumulada inversa (P) de encontrar un tweet donde los seguidores fueron una cierta proporción de los retransmisores. En el caso de Macri, en el rango de* $25\,\% \leq P \leq 75\,\%$ *se encontró que los seguidores son entre* 90 % *y* 95 % *de los retransmisores. En el mismo rango, para Scioli, se encontró que la proporción de seguidores es entre* 82 % *y* 87 % *mientras que para Massa, entre* 65 % *y* 92 % .

[8] Distancia entre el origen y cualquier otro nodo dentro del grafo inducido

3.7. Alineación de Retransmisores

Hemos analizado el comportamiento de los retransmisores de los supernodos tomando a estos últimos de a pares. La figura 3.15 muestra en cada celda la cantidad de usuarios en común que han retransmitido contenido de los supernodos. La matriz resulta simétrica, $rts_{ij} = rts_{ji}$ y en la diagonal se ha utilizado el número de retransmisores propios del supernodo, rts_{ii}. Se ha encontrado que los valores son un orden de magnitud mayor que el resto de las celdas. Se han ordenado los supernodos de acuerdo a la inclinación política de los mismos y se han podido identificar dos grupos. Uno de ellos, compuesto por Macri, Sanz, Stolbizer y Massa (G1) y el segundo por Anibal Fernández, Cristina Fernández de Kirchner y Daniel Scioli (G2). Se ha podido identificar que el comportamiento de las retransmisiones de contenido es de acuerdo a la inclinación política de los usuarios. El supernodo Macri cuenta con el mayor número de retransmisiones propias donde hemos apreciado que supera en un orden de magnitud al resto de los supernodos y en dos ordenes de magnitud las retransmisiones cruzadas.

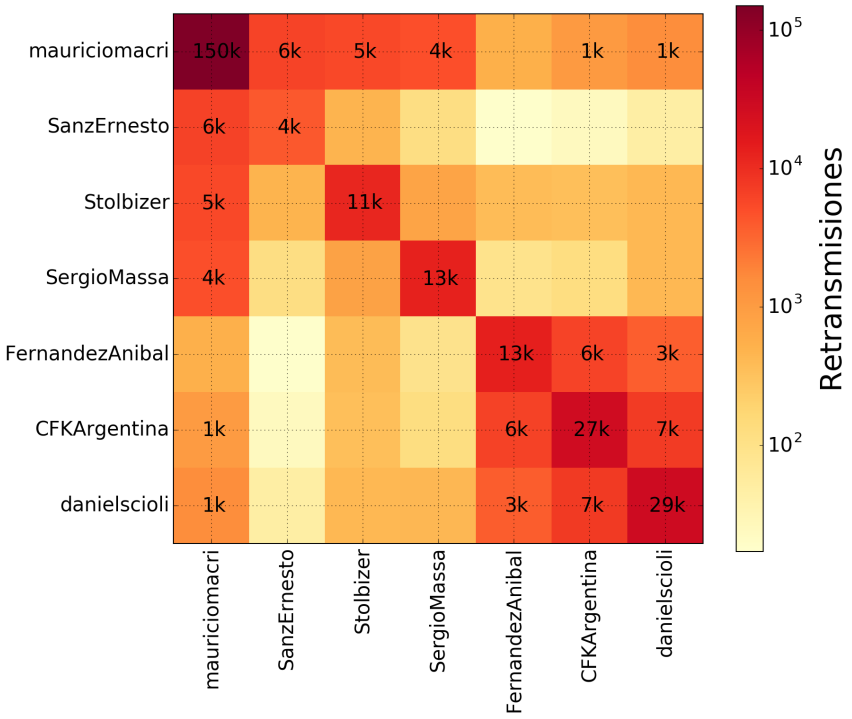


Figura 3.15 – ***Alineación por grupo.*** *El set de datos se generó con* 3,8*k tweets escritos por* 7 *supernodos incluyendo* 309 *mil retransmisiones. Se calculó la cantidad de usuarios que retransmitieron a cada par de supernodos y se agregaron las etiquetas cuando éste superaba índice superaba las mil unidades. La diagonal de la matriz contiene el número de usuarios que retransmitieron al supernodo. El orden de la matriz se ha especificado de acuerdo a la inclinación política de los supernodos.*

Es interesante notar que si agrupamos los supernodos de acuerdo a la cantidad de retransmisiones en común, observamos que mantienen similitud con el comportamiento político de los mismos

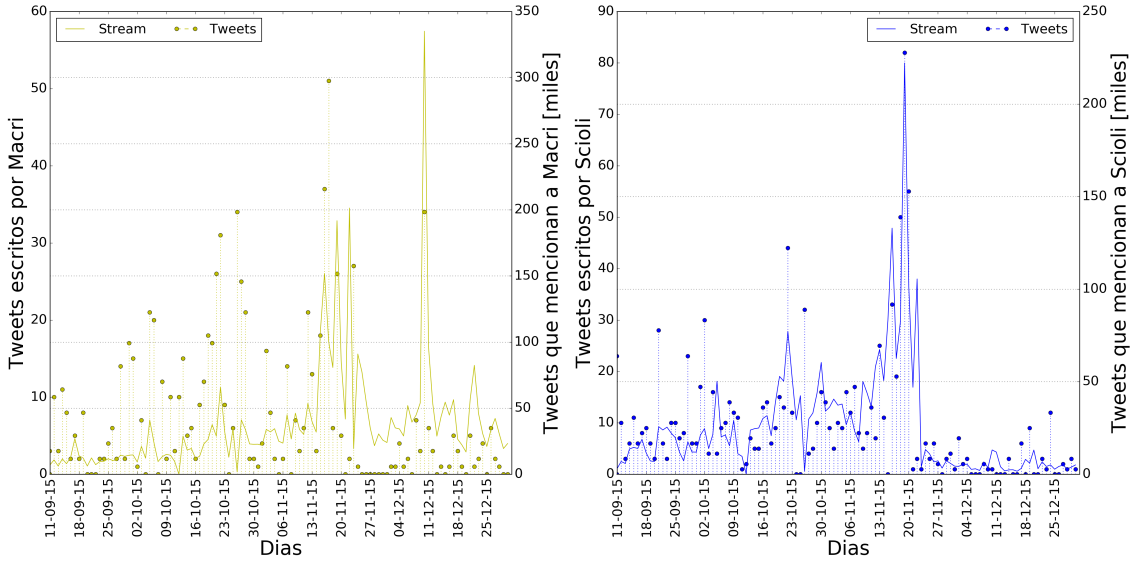
3.8. Flujo Continuo y *tweets*

En el período que va desde el día *11 de Septiembre* al *31 de Diciembre de 2015* hemos almacenado: la cantidad de *tweets* que los supernodos del grupo reducido han generado y la actividad de la red al mencionar a alguno de ellos.

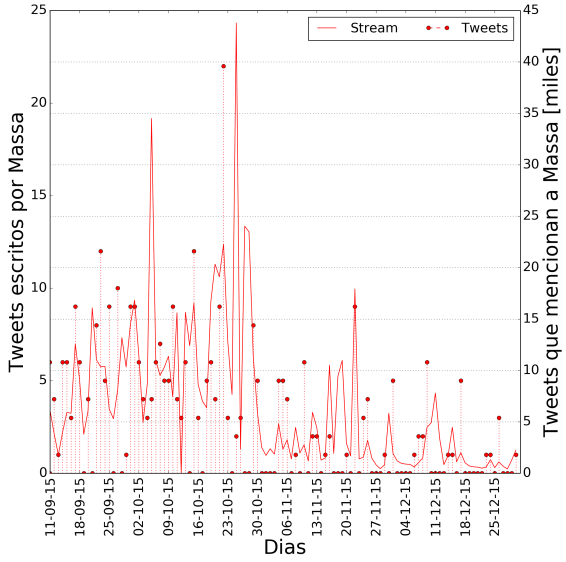
En la figura 3.16 se ha realizado un comparativo entre el nivel de actividad de los supernodos (*tweets* de Macri, Massa y Scioli) y la repercusión generada en la red por día. No se ha observado una correlación entre la actividad relevada y los *tweets* de los supernodos a nivel diario[9]. En cambio hemos observado que la actividad de la red depende de eventos fuera de la misma.

Ordenando los usuarios de acuerdo al nivel de actividad en ese período, hemos podido observar que Macri lidera el grupo mientras que Massa, es el que menor nivel de actividad registró. Tomando en contemplación que el tiempo de vida de un *tweet* es de algunos minutos, es de esperar que la actividad de la red responda en los minutos posteriores a la creación del contenido y su efecto se apacigüe con el correr del día (completo). A priori el análisis diario parece no tener suficiente granularidad como para inferir la dependencia de la actividad de la red en función del contenido generado por los supernodos.

[9]Es posible realizar un análisis similar a nivel hora.



(a) *Macri.* **(b)** *Scioli.*



(c) *Massa.*

Figura 3.16 – ***Actividad de la red y tweets escritos por los supernodos***. *El dataset se generó con los tweets escritos por los supernodos y aquellos en los que se los menciona en el período que va desde el 11 de Septiembre al 31 de Diciembre de 2015. De acuerdo a la limitación en la API de* ***Twitter*** *sólo es posible extraer alrededor del* 1 % *de la actividad de la red y hasta los últimos 3200 tweets de un usuario.* ***(a)***. *Se registraron para Macri* 809 *tweets propios y* 4,2M *menciones en la red. Se almacenaron cerca* 350k *menciones el día 10 de Diciembre siendo este el máximo del grupo de supernodos.* ***(b)***. *Para Scioli:* 1027 *y* 2,8M. *Se observó un decremento abrupto en la actividad de la red a partir de del día 23 de Noviembre.* ***(c)***. *Para Massa:* 327 *y* 748K *respectivamente. Se apreció que es el supernodo que menos actividad genera con alrededor del* 20 % *de actividad respecto de Macri durante el periodo analizado .*

3.9. Eventos

En la figura 3.17 hemos identificado cinco eventos cronológicos entre los días 9 de Septiembre y 17 de Diciembre de 2015: primer debate, primera vuelta electoral, segundo debate, balotaje y asunción del nuevo presidente electo. Se ha identificado un incremento en la actividad generada en la red durante el día posterior al *Primer Debate Presidencial* para los tres candidatos, figura 3.17a. El nivel se normaliza al segundo día. Se ha observado también que los tres supernodos mantienen niveles similares de actividad.

En la figura 3.17b se ha observado la actividad generada en los días previos y posteriores a la *Elección Presidencial*, donde hasta ese momento los tres supernodos continuaban generando un nivel similar.

En el día *5 de Octubre de 2015* se ha registrado el pico máximo de actividad con 50 mil *tweets*. A partir del día de la elección, se ha podido observar que la actividad de Massa y Macri disminuye mientras que la de Scioli comienza a ascender.

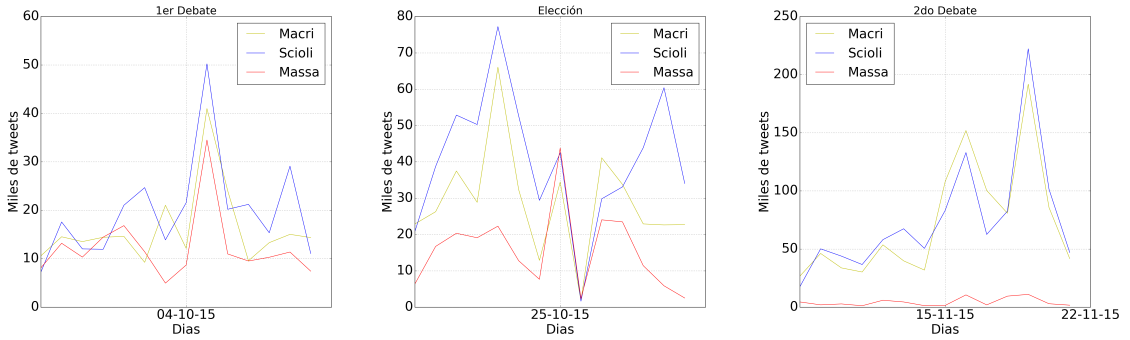
En la figura 3.17c hemos observado que Massa ya no genera actividad en la red mientras que Scioli y Macri siguen generando un volumen incluso mayor que en los eventos anteriores.

El máximo número de *tweets* ha sido registrado el día *19 de Noviembre de 2015* con más de 200 mil menciones a Scioli. En la figura 3.17d se ha observado que hasta el *22 de Octubre*, Scioli y Macri generaban la misma actividad en la red, sin embargo, en éste día Macri vuelve a generar 200 mil menciones mientras que Scioli sólo la mitad. Macri es elegido presidente.

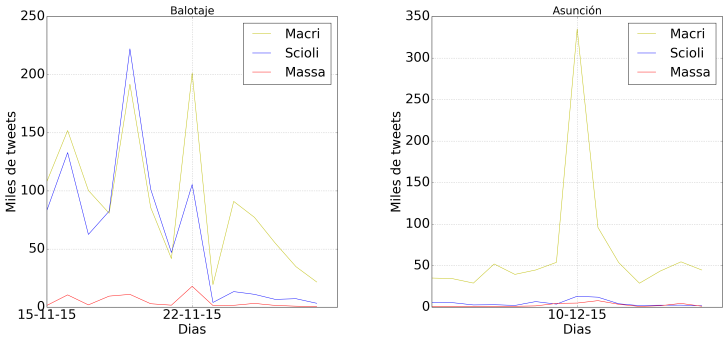
En el día *10 de Diciembre de 2015* se produce el cambio de gobierno y la gestión de Macri asume la presidencia. No se ha registrado actividad considerable de ninguno de los otros supernodos y Macri ha generado un pico máximo histórico del período analizado con alrededor de 350 mil menciones.

Hemos podido observar que la actividad en la red responde antes eventos externos de acuerdo a lo que ocurre con los supernodos analizados. Si un supernodo está relacionado con un evento puntual, entonces un incremento en el volumen de

actividad se verá registrado.



(a) *Primer debate.* **(b)** *Primera vuelta.* **(c)** *Segundo debate.*



(d) *Balotaje.* **(e)** *Asunción.*

Figura 3.17 – ***Actividad de la red ante eventos.*** *El dataset se conformó con* 6,6M *de tweets escritos entre el 9 de Septiembre y el 17 de Diciembre de 2015 que mencionaron a alguno de los supernodos. Macri obtuvo el* 3,5M *(*53 %*), Scioli* 2,5M *(*38 %*) y Massa* 600K *(*0,09 %*). La red responde ante eventos registrando incrementos o decrementos en la actividad.* ***(a)****. Se registró para los tres supernodos un nivel similar de actividad incluso cuando Scioli no ha participado del mismo. Los tres se encontraban en carrera presidencial.* ***(b)****. Luego de la primera elección se alzan Macri y Scioli como candidatos al balotaje, Massa queda fuera de la carrera presidencia. La actividad de este último disminuyó considerablemente en los días posteriores al evento.* ***(c)****. Durante la semana anterior y posterior al segundo debate, Macri y Scioli registraron similares niveles de actvidad. Massa generó poca actividad durante todo el período.* ***(d)****. Previamente al balotaje, Scioli y Macri siguieron generando actividad de forma similar. Durante el día del balotaje se observó un incremento de Macri y subsiguiente disminución de Scioli. Macri resultó electo presidente. Durante los días posteriores solo Macri continúo generando actividad en la red. Scioli y Massa casi no generaron actividad notoria durante este período.* ***(e)****. Macri siguió generando actividad en la red con un pico pronunciado en el dia de asunción presidencial (*350K *tweets). Tanto Scioli como Massa no generaron actividad notoria durante este período .*

3.10. Análisis de k-núcleos

3.10.1. Remoción del núcleo principal

Se seleccionado un *tweet* de Macri el cual ha sido retransmitido más de 12 mil veces. En pos de analizar la estructura del grafo inducido compuesta por las relaciones entre los retransmisores, se han obtenido los seguidores de al menos 8 mil de éstos. Finalmente hemos filtrado la red almacenada dejando solo el grafo inducido por el *tweet*: todos los lazos desde el supernodo hasta los retransmisores y todos aquellos lazos existentes entre retransmisores. Se ha utilizado la herramienta LaNet-Vi [Beiró et al., 2008] para dibujar los k-núcleos y lazos del grafo inducido completo como se puede apreciar en la figura 3.18a.

El núcleo máximo corresponde al maximal 57. En este caso se pudo apreciar que el escritor del *tweet* se encuentra dentro del núcleo principal el cual es representado en color rojo y se ubica dentro de la componente conexa mayor. No se han apreciado componentes conexas más allá de la principal. Se han observado que muchos de los usuarios de núcleos inferiores se conectan a usuarios en núcleos superiores, LaNet-Vi colorea la mitad del lazo con el color del nodo opuesto al conectado. Se ha podido apreciar que muchos nodos de color violeta se conectan a nodos en el núcleo rojo así como también a los de su mismo color (anillo exterior).

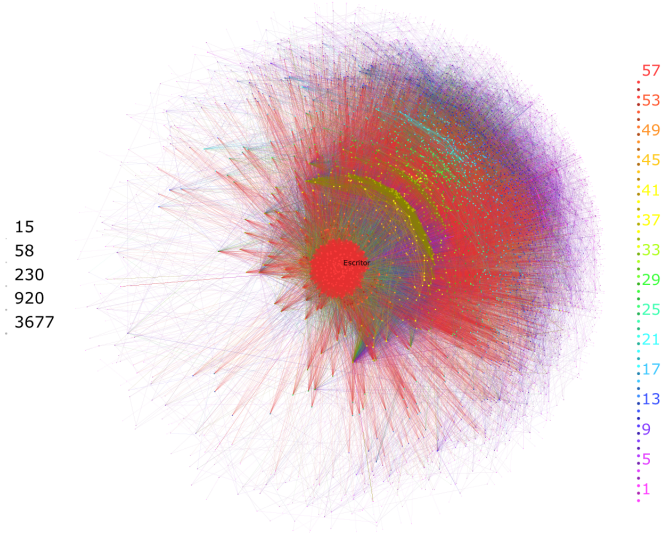
Luego se han quitado los usuarios y lazos de la capa 57 y se ha dibujado la nueva red. En la figura 3.18b hemos podido apreciar la máxima capa correspondiente al número 31. Nuevamente se ha apreciado una única componente conexa y que los usuarios de núcleos inferiores se conectan tanto a usuarios del mismo núcleo como a superiores. La notable disminución del número de la máxima capa es resultado de obtener una distribución de probabilidad de lazos de cola pesada.

En la figura 3.18c se han removido nuevamente todos los usuarios del máximo núcleo, en este caso el 31, siendo que la nueva capa máxima disminuye a 15. Hemos observado la existencia de una sola componente conexa mayoritaria y que los usuarios de núcleos inferiores se conectan a otros de su mismo núcleo así como a algunos pocos

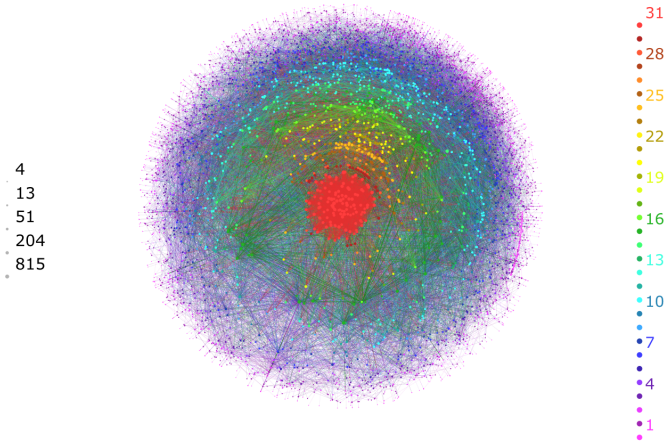
en superiores.

Se ha repetido el proceso cuatro veces más apreciando los resultados en las figuras 3.18d a 3.18g.

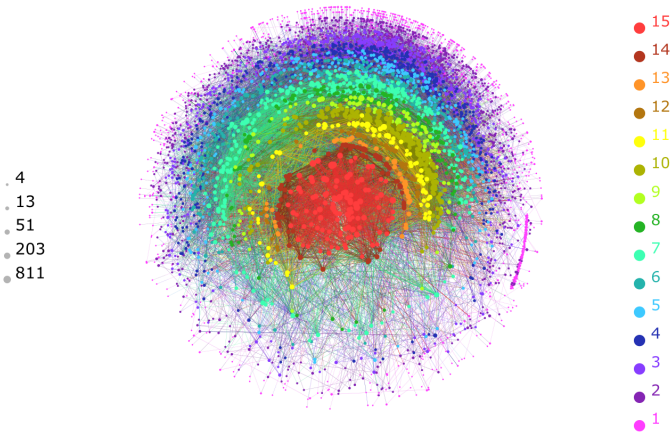
En todos los casos hemos observado que los usuarios siguen conectándose a un grupo reducido de usuarios en capas superiores. Al final de los seis pasos de remoción sigue existiendo una sola componente conexa mayoritaria centrada en aquellos usuarios del máximo núcleo. En los últimos tres pasos, al quitar el máximo núcleo, el nuevo es similar al anterior: de 10 se reduce a 7 y finalmente a 6. Esto ocurre en contraposición con los 26 removidos en el primer paso. En la figura 3.18g hemos observado a la mayoría de los usuarios conectarse a otros usuarios de la misma capa y algunos pocos conectarse a núcleos superiores. A su vez, se ha podido observar que el núcleo mayor comienza a dividirse en dos los núcleos no ocurriendo lo mismo en los inferiores los cuales siguen conectados en una sola componente.



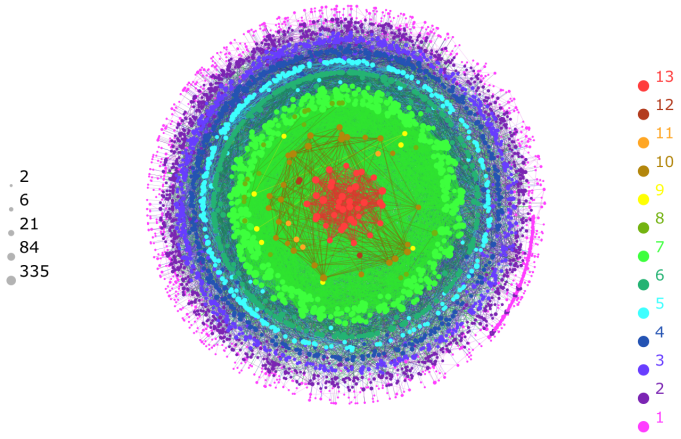
(a) 8*k retransmisiones.*



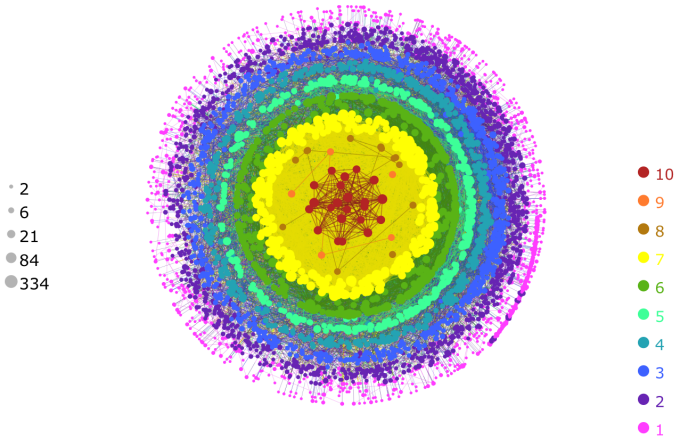
(b) *Sin los usuarios del core 57.*



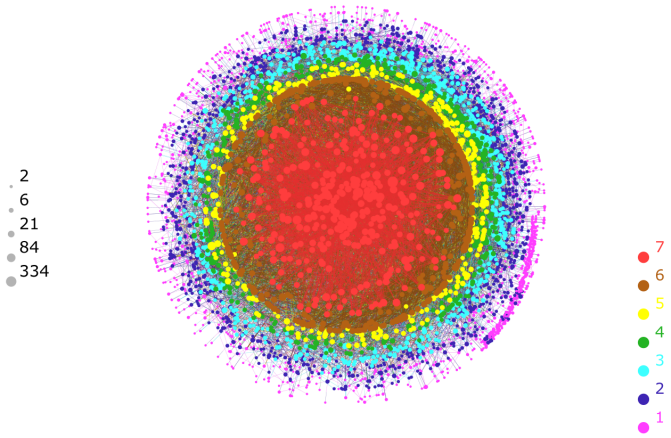
(c) *Sin los usuarios del core 31.*



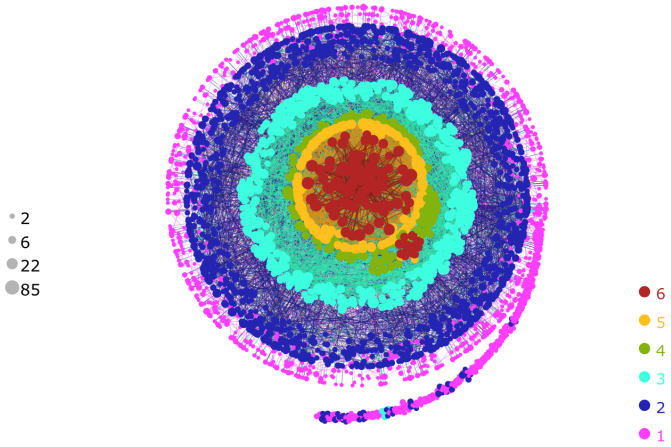
(d) *Sin los usuarios del core 15.*



(e) *Sin los usuarios del core 13.*



(f) *Sin los usuarios del core 10.*



(g) *Sin los usuarios del core 7.*

Figura 3.18 – ***Remoción del top core***. *Se eligió el tweet (id =* 668446892301418497*) que obtuvo* 12 *mil retransmisiones y de las cuales se almacenaron alrededor de* 8 *mil usuarios (*67 %*). De estos retransmisores se obtuvieron los seguidores y se conformó el grafo inducido partiendo desde el supernodo (escritor) hasta cada uno de estos retransmisores. Se definió como grafo inducido del tweet. Se generó el gráfico con la herramienta LaNetVi donde se visualiza la descomposición en k cores del grafo.* ***figura 3.18a****. Se observó que el máximo core encontrado es 57 (puntos rojo), que el escritor se encuentra dentro de este y la mayoría de las conexiones lo hacen hacia cores elevados (lazos rojos). Se relevaron* 111 *mil lazos. Grado promedio: 14.* ***figura 3.18b****. Se removieron los usuarios del core 57 y el nuevo top core paso a ser 31. Se observó una gran disminución de dicho top core. Se siguió apreciando la centralidad del grafo con una única componente conexa. El escritor fue removido en este paso.* ***figura 3.18c****. Se removieron los usuarios del core 31 dejando usuarios con 15 o menos conexiones. El número del top core se redujo al* 26 % *en solo dos remociones del top. Es coherente con la idea de que la topología obedece una distribución Power Law.* ***figura 3.18d****. Al cabo de la tercera remoción, el top core es de 13 y aún se observa una componente conexa solamente.* ***figura 3.18e*** *y* ***figura 3.18f****. Cuarto y quinto paso de remoción del top core. No se observan componente conexas además de la central.* ***figura 3.18g****. Al cabo de la remoción sistemática de cores por sexta vez, se observaron algunas pequeñas componentes conexas de bajo core sobre la parte inferior derecha de la figura. En este paso quedaron* 4603 *usuarios con* 9501 *lazos cada uno. Grado promedio: 2.*

A lo largo de la presente sección se ha graficado la descomposición en k-núcleos del grafo inducido por los 8 mil retransmisores de un *tweet* junto con la evolución de dicha descomposición a lo largo de múltiples pasos de remoción del núcleo superior. Otros *tweets* a modo de complementar con otros ejemplos se han agregado en los apéndices. En todos los casos se ha podido apreciar un comportamiento similar al arriba descripto donde:

- En todos los grafos inducidos se observa una sola componente conexa mayoritaria.
- Los nodos se conectan mayoritariamente a otros del mismo núcleo mientras unos pocos, a núcleos superiores.
- Los grafos inducidos comienzan en k-núcleos elevados y al cabo de seis pasos de remoción aún permanecen conexos.
- El k-núcleo siempre queda representado en el centro de la componente conexa.

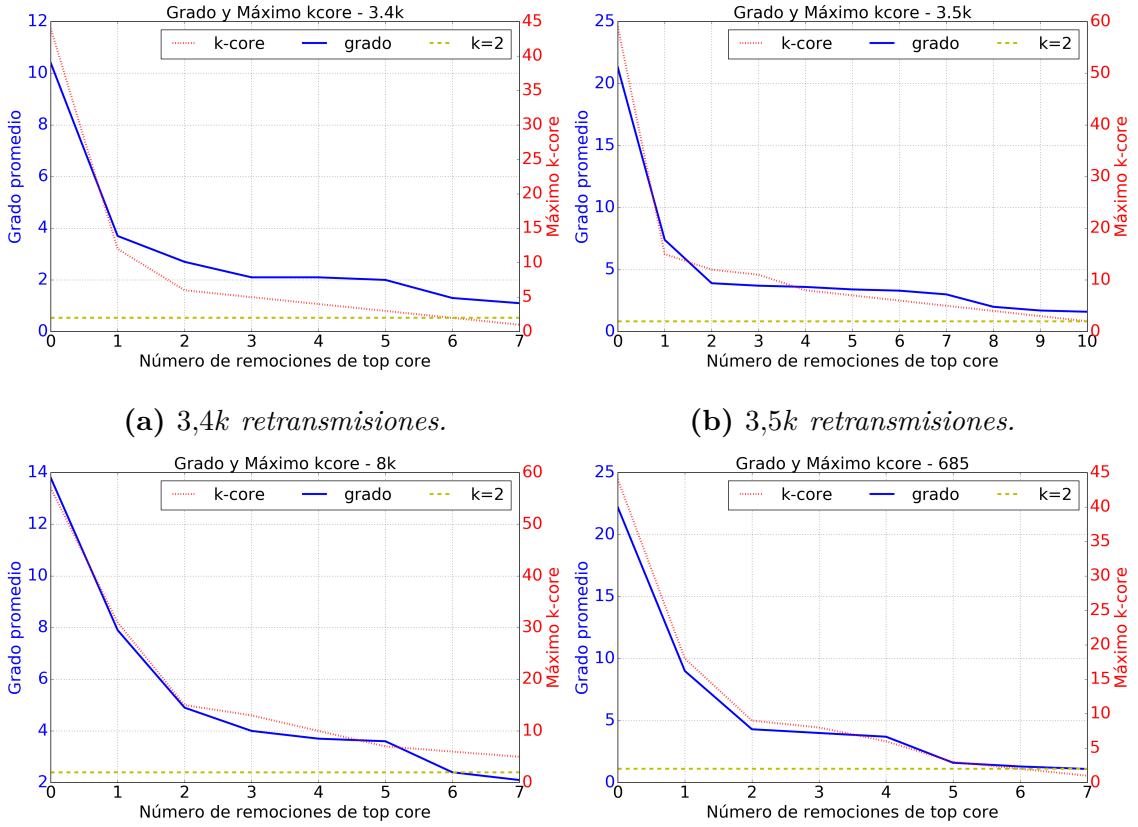
3.10.2. Evolución usuarios versus k-núcleos

Hemos tomado cuatro grafos inducidos por las retransmisiones y se ha graficado el grado promedio junto con el nuevo k-núcleo de acuerdo a la sucesiva remoción del núcleo superior.

En todos los casos se ha podido observar que la dinámica en la disminución del grado promedio acompaña la dinámica de la reducción del máximo número de capa encontrada en el grafo en cada paso. Además en tres de los cuatro casos al cabo del sexto paso de remoción no se había alcanzando aún a descomponer completamente el grafo quedando capas de grado dos o más. En el caso del *tweet* con 3,5 mil retransmisiones se ha observado que al cabo de nueve pasos la red seguía teniendo usuarios en capas mayores que uno.

Se ha agregado una línea recta en el *k-núcleo*= 2 para poder discriminar cuando ante las sucesivas remociones el máximo núcleo descendía por debajo este umbral.

De acuerdo a [Bhawalkar et al., 2012, Farzan et al., 2011] donde proponen anclar un número de vértices para optimizar la propagación de un mensaje en la red, con k-núcleos por encima de dos es posible encontrar una solución para evitar que la misma se desarme con sucesivas extracciones de nodos.



(a) 3,4k *retransmisiones.*

(b) 3,5k *retransmisiones.*

(c) 8k *retransmisiones.*

(d) 685 *retransmisiones.*

Figura 3.19 – ***Lazos y Grados***. *Para cada uno de los 4 tweets elegidos se graficaron los lazos por usuario o grado promedio.* ($gr = \frac{n_{lazos}}{n_{usuarios}}$) *y el máximo core dentro del grafo inducido. Se repitió el computo cada vez que se removian los usuarios del top core (core más alto). Se observó que en todos los casos tomados ambas curvas son similares a lo largo de los pasos de remoción.*

3.10.3. Coeficiente de Agrupamiento

A modo de calcular la conexidad de la componente mayoritaria encontrada en los sucesivos pasos de remoción de los k-núcleos superiores, se ha calculado el coeficiente de agrupamiento de acuerdo a [Newman, 2003].

En la figura 3.20 se ha podido apreciar que en todos los casos contemplados el coeficiente disminuye a medida que se remueven los k-núcleos superiores. Hemos podido observar a su vez que el coeficiente de agrupamiento definido en la ecuación 3.10 y su desvío estándar no indican que los sub-grafos comiencen a fraccionarse.

$$c = 3 * \frac{número\ de\ triángulos}{número\ de\ triplos} \tag{3.10}$$

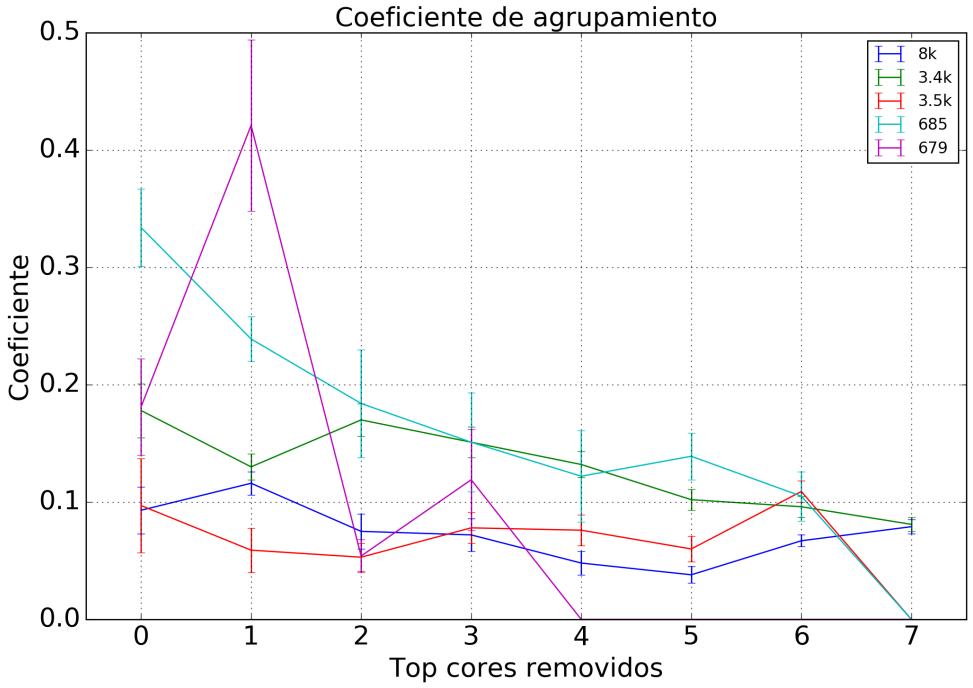


Figura 3.20 – ***Coeficiente de Agrupamiento***. *Se seleccionaron cinco tweets que fueron retransmitidos 8k, 3,5k, 3,4k, 685 y 679 veces. Se generó el grafo inducido para cada caso de acuerdo a los retransmisores de cada uno y a los vínculos existentes entre los ellos y el supernodo escritor del tweet. Se computó el coeficiente de agrupamiento para el grafo inducido. Se removieron todos los usuarios del máximo core encontrado y se recalculó el coeficiente. Se realizó este procedimiento seis veces más y se observó que el coeficiente de agrupamiento no presentó variaciones significativas para ninguno de los casos analizados.*

Capítulo 4

Conclusiones

A lo largo del presente trabajo se han planteado interrogantes acerca de diferentes tipos de redes sociales y cómo los vínculos se establecen en las mismas.

Nos hemos enfocado en el desarrollo de interfaces y procesos para la extracción, almacenamiento y consumo para el posterior análisis de información proveniente de la red social asimétrica Twitter.

Se ha presentado evidencia acerca de la topología de la misma y se han estudiado sus características para concluir en que la misma corresponde a un proceso de cola pesada. No fue posible identificar una sola distribución de probabilidad acorde a una distribución de ley de potencia que gobierne el proceso llevando a pensar que la realidad presenta características más complejas que una simple distribución.

Posteriormente, se ha definido un grupo de nodos al cual llamamos supernodos, los cuales generan contenido que otros usuarios, seguidores, consumen. En función del grupo de seguidores hemos definido y calculado la distancia que pudiese existir entre los diferentes supernodos.

Se han obtenido resultados que permiten asociar a la masa de seguidores con la alineación política de los supernodos. Aquellos que guardan una distancia menor entre sí, tendrán mas seguidores en común que otros más alejados. Se ha podido observar que los usuarios en la red son selectivos a la hora de consumir contenido y sólo se suscriben a aquellos supernodos que generan contenido interesante.

Luego de entender la topología y los diferentes actores, nos hemos enfocado en la forma en que éstos manipulan la información a través del uso de las menciones y retransmisiones. Como resultado se ha podido apreciar que los usuarios retransmiten contenido solo de aquellos supernodos que les interesa y poco o nada de los demás.

En cuanto a las menciones, aportan un grado de libertad extra respecto de las retransmisiones y por ende son menos precisas a la hora de buscar definir al los usuarios. Las menciones pueden ser positivas, negativas o neutras sin reforzar el contenido original como en el caso de las retransmisiones.

Se ha podido apreciar que los usuarios retransmiten y mencionan de diferentes formas a los supernodos dando la posibilidad de identificar cual de estos genera contenido de más alto impacto (y propagado).

A continuación, se ha analizado cuán efectivo es el uso de las menciones y retransmisiones para clasificar a los usuarios según su afinidad por los supernodos. Hemos encontrando que el clasificador lineal identifica la afiliación con muy buena performance al utilizar las retransmisiones y con una buena performance (menor que con las anteriores) al utilizar las menciones.

Se ha analizado la dinámica con la que evoluciona las retransmisiones y favoritos de los tweets donde nuevamente se han encontrado distribuciones de probabilidad de cola pesada.

Al estudiar el origen de la producción de las retransmisiones hemos encontrado que la mayoría de éstas son producidas por seguidores del autor del tweet ($\sim 90\,\%$).

Luego, se ha analizado si teniendo en cuenta la cantidad de usuarios que han retransmitido un tweet, podíamos encontrar cierta alineación partidaria entre los supernodos. Se ha encontrado que de acuerdo a la cantidad de retransmisiones que se obtenían al tomar dos supernodos, se podían agrupar en al menos dos grupos bien diferenciados. Reforzando el concepto del uso de las retransmisiones para propagar información afín. Con este análisis podemos concluir que twitter permite identificar afiliación política de usuarios y de supernodos.

Posteriormente, nos hemos enfocado en estudiar la correlación entre la actividad generada por los supernodos y la actividad en la red. No se ha podido encontrar una

correlación entre ambas a consolidando la información a nivel diario. Por otra parte se ha observado que la red refleja actividad acerca de los supernodos de acuerdo a eventos externos en los cuales éstos se ven involucrados.

Finalmente Se ha calculado y graficado la descomposición en k-núcleos del grafo inducido por la retransmisión de un tweet. Se ha propuesto quitar el núcleo superior a fin de estudiar si dicho grafo sufría algún grado de disociación lo cual no fue encontrado. Se ha repetido el proceso de remoción múltiples veces y a fin de verificar el grado de disociación hemos calculado el coeficiente de agrupamiento ante las sucesivas remociones. El mismo no ha incrementado su valor notoriamente llevando a pensar que la red permanece conectada y con una alta robustez luego de quitar los usuarios pertenecientes al k-núcleo superior.

Para futuros estudios se proponen los siguiente puntos de análisis: Identificar las distribuciones subyacentes a la topología de red presentada. Estudiar el diámetro total del grafo inducido por el retweet. Identificar si existe correlación lineal entre tweets de supernodos y actividad de la red. Calcular la descomposición en k-núcleos tomando el grafo inducido por dos o más tweets.

Como conclusión podemos decir que en este tipo de redes sociales asimétricas, donde se fomenta la propagación de información, es casi imposible limitar la comunicación entre líderes y seguidores.

Capítulo 5

Apéndices

5.1. Estructuras

En el siguiente apéndice se incluye un ejemplo con las propiedades de un *tweet* y de un usuario dentro de la red social ***Twitter*** con el fin de identificar todas las propiedades que las entidades poseen y que podría ser analizadas en posteriores desarrollos.

5.1.1. Estructura de un *tweet*


```
"id_str"  "593895901623496704"
"text"  "This is a #test tweet @LoveforTestingT with an image.
http://t.co/ZvgHovKZq4"
"source"  "<a href=\"http://twitter.com\" rel=\"nofollow\">Twitter Web
Client</a>"
"truncated"  false
"in_reply_to_status_id"  null
"in_reply_to_status_id_str"  null
"in_reply_to_user_id"  null
"in_reply_to_user_id_str"  null
"in_reply_to_screen_name"  null
"geo"  null
"coordinates"  null
"place"
    "id"  "fd70c22040963ac7"
    "url"  "https://api.twitter.com/1.1/geo/id/fd70c22040963ac7.json"
    "place_type"  "city"
    "name"  "Boulder"
    "full_name"  "Boulder, CO"
    "country_code"  "US"
    "country"  "United States"
    "bounding_box"
        "type"  "Polygon"
        "coordinates"


                    -105.3017759
                    39.953552


                    -105.3017759
                    40.094411


                    -105.183597
                    40.094411


                    -105.183597
                    39.953552




    "attributes"
```

```
"retweet_count"  0
"favorite_count"  0
"entities"
    "hashtags"

            "text"  "test"
            "indices"
                10
                15



    "trends"
    "urls"
    "user_mentions"

            "screen_name"  "LoveforTestingT"
            "name"  "Leah "
            "id"  2432909010
            "id_str"  "2432909010"
            "indices"
                22
                38



    "symbols"
    "media"

            "id"  593895901145346000
            "id_str"  "593895901145346048"
            "indices"
                54
                76

            "media_url"
"http://pbs.twimg.com/media/CD3xLO1UMAAvBgO.png"
            "media_url_https"
"https://pbs.twimg.com/media/CD3xLO1UMAAvBgO.png"
            "url"  "http://t.co/ZvgHovKZq4"
            "display_url"  "pic.twitter.com/ZvgHovKZq4"
            "expanded_url"
"http://twitter.com/johnd_test/status/593895901623496704/photo/1"
            "type"  "photo"
            "sizes"
```

```
                    "h"  340
                    "resize"  "fit"

                "thumb"
                    "w"  150
                    "h"  150
                    "resize"  "crop"

                "large"
                    "w"  512
                    "h"  512
                    "resize"  "fit"

                "medium"
                    "w"  512
                    "h"  512
                    "resize"  "fit"




"extended_entities"
    "media"

            "id"  593895901145346000
            "id_str"  "593895901145346048"
            "indices"
                54
                76

            "media_url"
"http://pbs.twimg.com/media/CD3xLO1UMAAvBgO.png"
            "media_url_https"
"https://pbs.twimg.com/media/CD3xLO1UMAAvBgO.png"
            "url"  "http://t.co/ZvgHovKZq4"
            "display_url"  "pic.twitter.com/ZvgHovKZq4"
            "expanded_url"
"http://twitter.com/johnd_test/status/593895901623496704/photo/1"
            "type"  "photo"
            "sizes"
                "small"
                    "w"  340
                    "h"  340
                    "resize"  "fit"
```

```
                    "w"  150
                    "h"  150
                    "resize"  "crop"

                "large"
                    "w"  512
                    "h"  512
                    "resize"  "fit"

                "medium"
                    "w"  512
                    "h"  512
                    "resize"  "fit"




"favorited"  false
"retweeted"  false
"possibly_sensitive"  false
"filter_level"  "low"
"lang"  "en"
"timestamp_ms"  "1430430791656"
```

5.1.2. Estructura de un usuario


```
    "id_str"  "2993982541"
    "name"  "Test Demo"
    "screen_name"  "jondee_test"
    "location"  "Denver, CO"
    "url"  null
    "description"  "this is a test account."
    "protected"  false
    "verified"  false
    "followers_count"  2
    "friends_count"  43
    "listed_count"  0
    "favourites_count"  0
    "statuses_count"  30
    "created_at"  "Sat Jan 24 00:12:53 +0000 2015"
    "utc_offset"  -21600
    "time_zone"  "Mountain Time (US & Canada)"
    "geo_enabled"  true
    "lang"  "en"
    "contributors_enabled"  false
    "is_translator"  false
    "profile_background_color"  "1A1B1F"
    "profile_background_image_url"
"http://abs.twimg.com/images/themes/theme9/bg.gif"
    "profile_background_image_url_https"
"https://abs.twimg.com/images/themes/theme9/bg.gif"
    "profile_background_tile"  false
    "profile_link_color"  "2FC2EF"
    "profile_sidebar_border_color"  "181A1E"
    "profile_sidebar_fill_color"  "252429"
    "profile_text_color"  "666666"
    "profile_use_background_image"  true
    "profile_image_url"
"http://pbs.twimg.com/profile_images/593893018911907840/YM9oLru3_normal.
png"
    "profile_image_url_https"
"https://pbs.twimg.com/profile_images/593893018911907840/YM9oLru3_normal.
png"
    "profile_banner_url"
"https://pbs.twimg.com/profile_banners/2993982541/1430430074"
    "default_profile"  false
    "default_profile_image"  false
    "following"  null
    "follow_request_sent"  null
    "notifications"  null

```

5.2. Retransmisiones

En el presente apéndice se ejemplifican *tweets* similares al utilizado en el capítulo de "resultados" donde hemos graficado los grafos inducidos y se ha removido la capa superior múltiples veces. El *tweet* de la figura 5.1 ha obtenido alrededor de mil cien retransmisiones y mil cuatrocientos favoritos aproximadamente. El mismo se encuentra en estado estacionario ya que no recibe más retransmisiones ni favoritos, habiendo sido escrito el Sábado 7 de Noviembre de 2015 a las 16:25hs. Se ha podido apreciar que ante los reiterados pasos de extracción del núcleo superior, se sigue identificando una única componente conexa en cuyo centro se ubican los nodos de la capa superior.

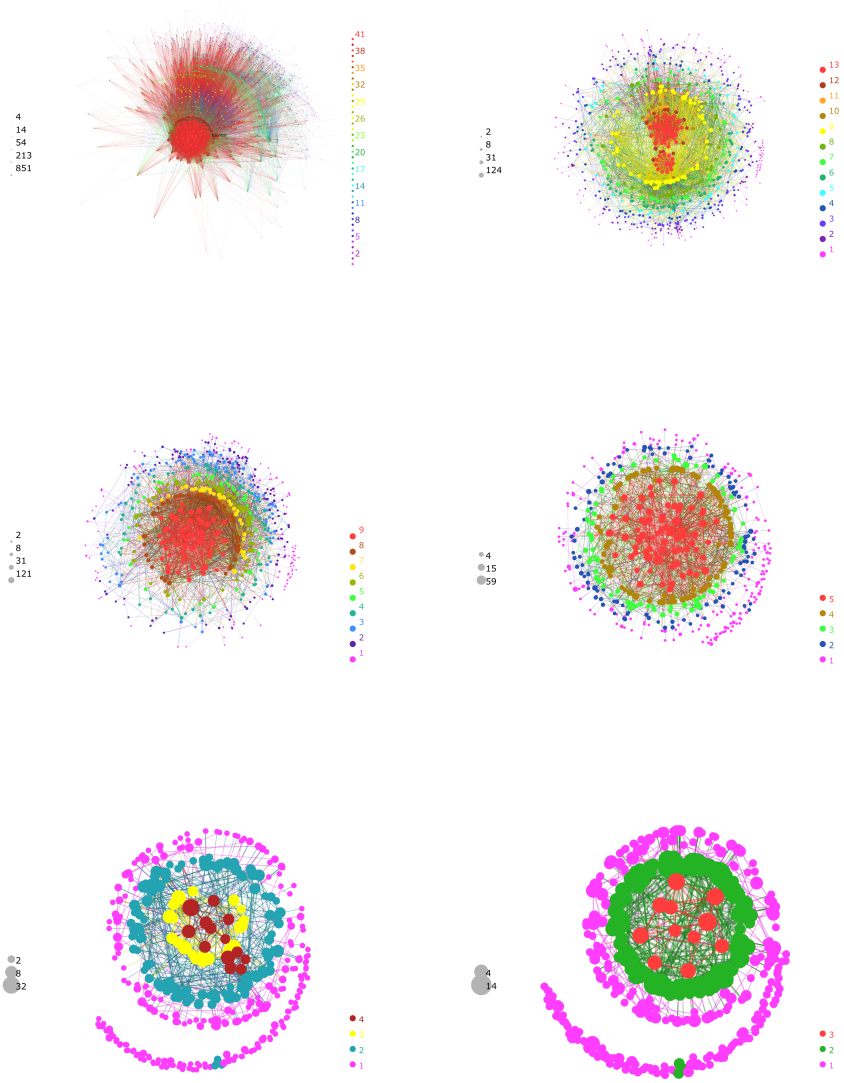


Figura 5.1 – *tweet: "Se cayeron los mitos de que los mismos van a gobernar siempre y la resignación. El 25 dijimos SE PUEDE. https:t.copcwQrBf2vc" .*

En el caso de la figura 5.2 el *tweet* ha obtenido más de 5 mil favoritos y arriba de 2 mil retransmisiones. El mismo procedimiento ha sido realizado en todos los ejemplos: remoción del núcleo superior observando la centralidad del grafo resultante. Fue escrito el 23 de Diciembre de 2015 a las 14:06hs.

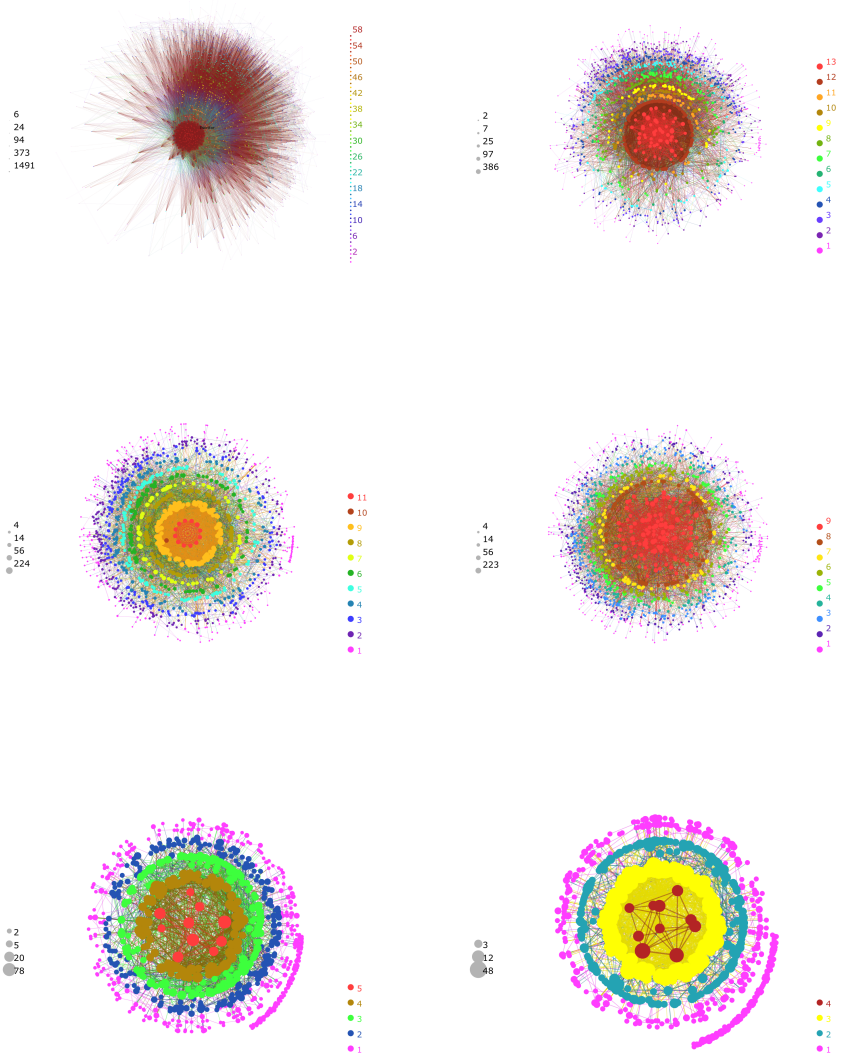


Figura 5.2 – *tweet: "Iniciamos las obras para el nuevo viaducto en Camino Negro que permitirá acceder directamente al Puente La Noria https:t.co/F7ZS4ASslO" .*

En el caso de la figura 5.3 el *tweet* ha obtenido más de 2500 favoritos y arriba de 1500 retransmisiones. Fue escrito el 9 de Noviembre de 2015 a las 15:03hs.

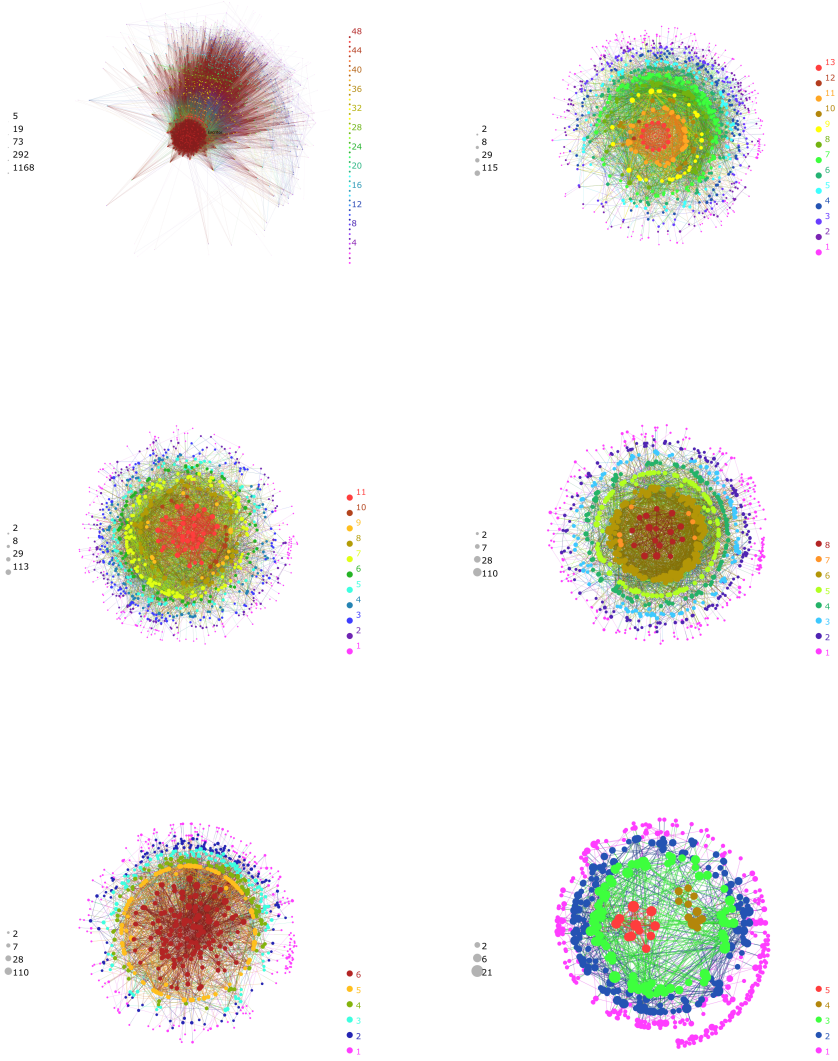


Figura 5.3 – *tweet: "Recorriendo #Paraná con la caravana del cambio https:t.colKHJ2FYOau" .*

El *tweet* de la figura 5.4 ha recibido mas de 18 mil favoritos y 12 mil retransmisiones. El mismo fue escrito el 22 de Noviembre de 2015 a las 12:12hs. El texto cita: "Presidente de la República Argentina. Casado. 4 hijos. Hincha de Boca.". Escrito por el supernodo Mauricio Macri, durante los comicios presidenciales.

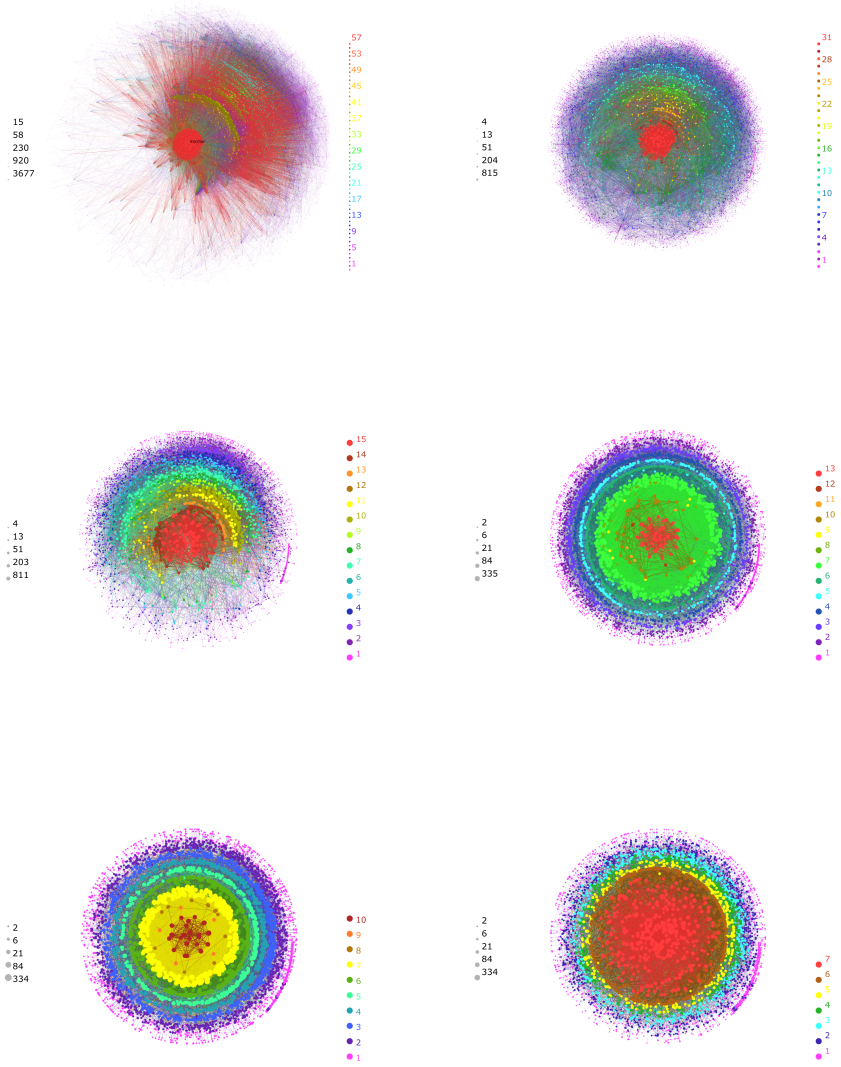


Figura 5.4 – *tweet: "Hoy es un día histórico. https:t.co72RlXoGfb3" .*

Este *tweet* ha obtenido 14 mil favoritos y más de 8 mil retransmisiones. Escrito por Mauricio Macri el día 23 de Noviembre de 2015 a las 11:08hs.

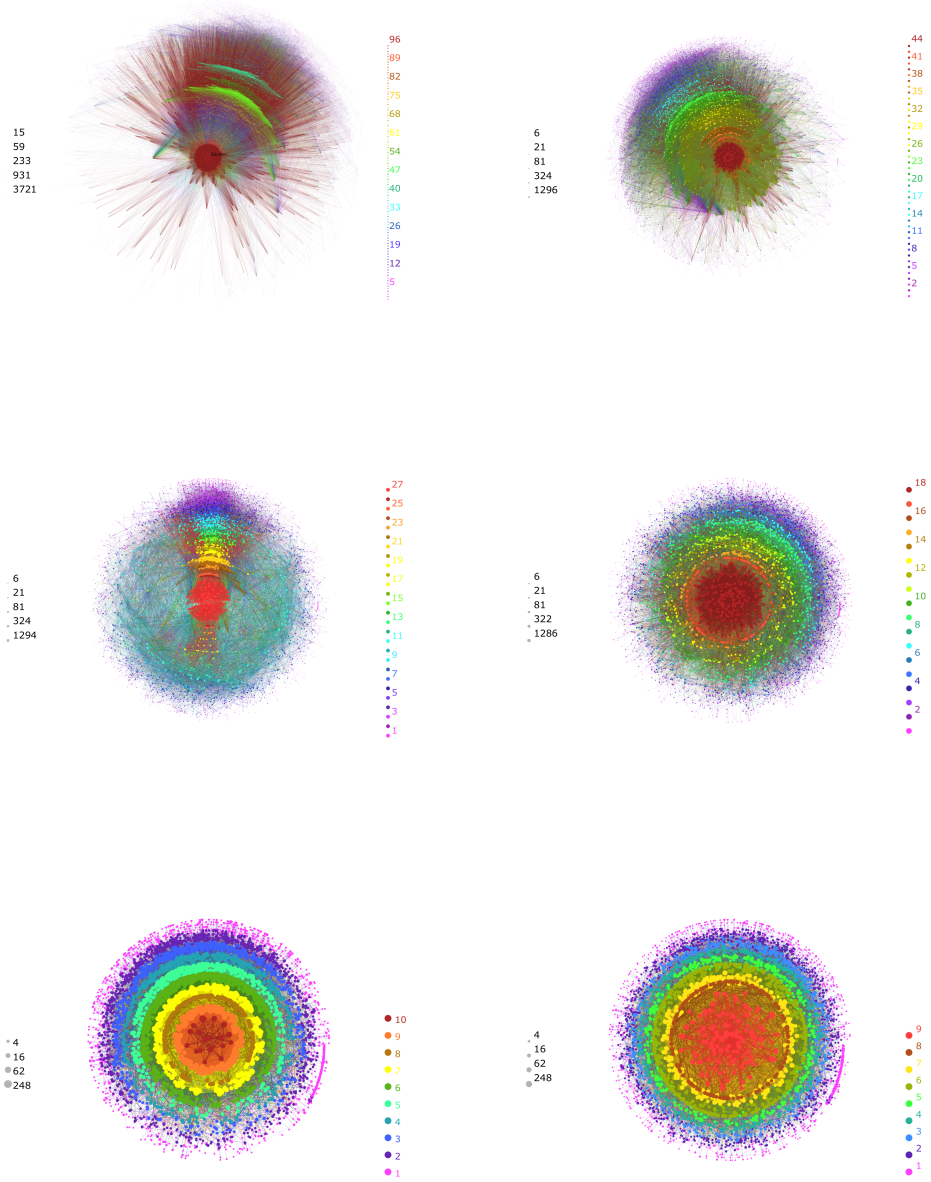
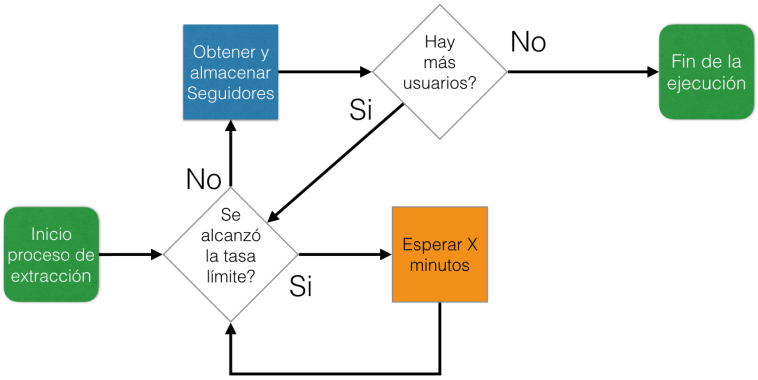


Figura 5.5 – *tweet: "Lo que pasó estos meses en la Argentina es inédito, todo tiene carácter de hazaña. Los argentinos han hecho posible lo imposible. #Cambiamos" .*

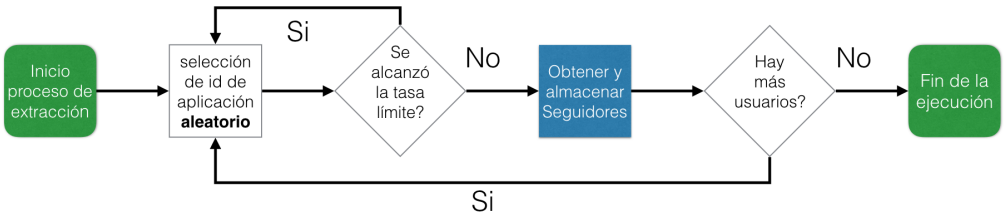
5.3. Múltiples Aplicaciones

Como se ha mencionado a lo largo de la tesis, múltiples aplicaciones de ***Twitter*** fueron creadas con el objeto de reducir los tiempos de extracción de las diferentes entidades y relaciones. En la figura 5.6a se observa el esquema inicial utilizado con una simple aplicación y el cual hemos programado en Python. Al pasar a la figura 5.6b se observa que los tiempos de espera han sido eliminados y se ha utilizado la selección aleatoria de aplicación para la extracción de los datos.

A fin de eliminar los tiempos de espera en cada interfaz se calculó el tiempo necesario para la extracción y almacenamiento de un ciclo y se calculó el total de aplicaciones necesarias para que no se alcance el límite impuesto en el período de 15 minutos.



(a) *Aplicación Única.*



(b) *Múltiples Aplicaciones.*

Figura 5.6 – ***Esquema de extracción de seguidores multi-aplicación.***

$$t_{ciclo} = t_{solicitud} + t_{twitter} + t_{respuesta} + t_{procesamiento} + t_{almacenamiento} \tag{5.1}$$

La ecuación 5.1 enuncia los tiempos contemplados por las interfaces para el cómputo de la cantidad de aplicaciones necesarias. Cada función de ***Twitter*** impone su propio límite a la cantidad de pedidos (o ciclos según hemos definido) que se pueden ejecutar por ventana (15 minutos).

Para el caso de la extracción de seguidores, el número de ciclos por ventana es de 15. Si $t_{ciclo} > \frac{15min}{n}$, siendo n el número de ciclos permitidos, entonces no es necesario utilizar más de una aplicación. En nuestro caso, los tiempos relevados rondaban los segundos. Al ser tan rápidos fue necesario introducir una espera o bien generar otra aplicación que pudiera seguir extrayendo datos mientras la primera se encontraba en modo espera.

Se desarrollaron en total 20 aplicaciones para evitar introducir retardos y optimizar los procesos de extracción. La reducción en los tiempos de ejecución de estos procesos se pueden apreciar en la siguiente tabla.

	Objetos	1 Aplicación	Múltiples Aplicaciones
Seguidores de Supernodos	10 millones	33 horas	1.6 horas
Seguidores de Seguidores	110 millones	15 días	18 horas
Tweets	54 millones	9.3 días	11 horas

Bibliografía

[Alstott et al., 2013] Alstott, J., Bullmore, E., and Plenz, D. (2013). powerlaw: a python package for analysis of heavy-tailed distributions. *Science*.

[Barabási and Albert, 1999] Barabási, A.-L. and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286:509–512.

[Beiró et al., 2008] Beiró, M. G., Alvarez-Hamelin, J. I., and Busch, J. R. (2008). A low complexity visualization tool that helps to perform complex systems analysis. *New Journal of Physics*, 10(12):125003.

[Benhardus and Kalita, 2013] Benhardus, J. and Kalita, J. (2013). Streaming trend detection in twitter. *International Journal of Web Based Communities*, 9(1):122–139.

[Bhawalkar et al., 2012] Bhawalkar, K., Kleinberg, J., Lewi, K., Roughgarden, T., and Sharma, A. (2012). Preventing unraveling in social networks: The anchored k-core problem. In *Proceedings of the 39th International Colloquium Conference on Automata, Languages, and Programming - Volume Part II*, ICALP'12, pages 440–451, Berlin, Heidelberg. Springer-Verlag.

[Bild et al., 2015a] Bild, D. R., Liu, Y., Dick, R. P., Mao, Z. M., and Wallach, D. S. (2015a). Aggregate Characterization of User Behavior in Twitter and Analysis of the Retweet Graph. *ACM Transactions on Internet Technology (TOIT) (2015): 4.*, 15(1):17.

[Bild et al., 2015b] Bild, D. R., Liu, Y., Dick, R. P., Mao, Z. M., and Wallach, D. S. (2015b). Aggregate characterization of user behavior in twitter and analysis of the retweet graph. *ACM Trans. Internet Technol.*, 15(1):4:1–4:24.

[Boyd and Ellison, 2007] Boyd, D. M. and Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1).

[Chen et al., 2010] Chen, H., Zhu, Y., Hu, K., and He, X. (2010). Hierarchical swarm model: A new approach to optimization. *Discrete Dynamics in Nature and Society*, 2010.

[Clauset et al., 2009] Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. In *ISSN 00361445. doi: 10.1137/070710111. URL http://dx.doi.org/10.1137/070710111.*

[Conover et al., 2011] Conover, M. D., Gonc, B., Ratkiewicz, J., Flammini, A., and Menczer, F. (2011). Predicting the Political Alignment of Twitter Users.

[Conover et al., 2012] Conover, M. D., Gonçalves, B., Flammini, A., and Menczer, F. (2012). Partisan asymmetries in online political activity. *EPJ Data Science*, 1(1):1–19.

[ECMA International, 2011] ECMA International (2011). *Standard ECMA-262 - ECMAScript Language Specification.* ECMA International, 5.1 edition.

[Farzan et al., 2011] Farzan, R., Dabbish, L., Kraut, R. E., and Postmes, T. (2011). Increasing Commitment to Online Communities by Designing for Social Presence. *Proceedings of the ACM 2011 conference on Computer supported cooperative work - CSCW '11*, (March 1923):321.

[Jeanson et al., 2003] Jeanson, R., Ratnieks, F. L. W., and Deneubourg, J.-L. (2003). Pheromone trail decay rates on different substrates in the Pharaoh's ant, \textit{Monomorium pharaonis}. *Physiol Entomol*, 28(3):192–198.

[King et al., 2015] King, A. J., Cowlishaw, G., King, A. J., and Cowlishaw, G. (2015). Leaders , followers , and group decision-making Leaders , followers and group decision-making. 0889(October):147–150.

[L. Peter Deutsch, 1996] L. Peter Deutsch, J.-L. G. (1996). ZLIB Compressed Data Format Specification. RFC 1560, RFC Editor.

[Li et al., 2015] Li, H., Mukherjee, A., Liu, B., Kornfield, R., and Emery, S. (2015). Detecting Campaign Promoters on Twitter Using Markov Random Fields. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2015-Janua(January):290–299.

[Lusseau, 2003] Lusseau, D. (2003). The emergent properties of a dolphin social network. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(Suppl 2):S186–S188.

[Lusseau and Newman, 2004] Lusseau, D. and Newman, M. E. J. (2004). Identifying the role that animals play in their social networks. *Proceedings of the Royal Society B: Biological Sciences*, 271(Suppl_6):S477–S481.

[Mahalanobis, 1936] Mahalanobis, P. C. (1936). On the generalised distance in statistics. In *Proceedings National Institute of Science, India*, volume 2, pages 49–55.

[Makazhanov and Rafiei, 2013] Makazhanov, A. and Rafiei, D. (2013). Predicting political preference of twitter users. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 298–305, New York, NY, USA. ACM.

[Mitani and Amsler, 2003] Mitani, J. and Amsler, S. (2003). Social and spatial aspects of male subgrouping in a community of wild chimpanzees. *Behaviour*, 140(7):869–884.

[Morone and Makse, 2015] Morone, F. and Makse, H. A. (2015). Influence maximization in complex networks through optimal percolation. *Nature*, 524(7563):65–U122.

[Newman, 2003] Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM REVIEW*, 45:167–256.

[Nishida, 1983] Nishida, T. (1983). Alpha status and agonistic alliance in wild chimpanzees (pan troglodytes schweinfurthii). *Primates*, 24(3):318–336.

[Plank and James, 2008] Plank, M. J. and James, A. (2008). Optimal foraging: Levy pattern or process? *Journal of the Royal Society, Interface / the Royal Society*, 5(26):1077–1086.

[Reid et al., 2011] Reid, C. R., Sumpter, D. J. T., and Beekman, M. (2011). Optimisation in a natural system: Argentine ants solve the Towers of Hanoi. *The Journal of experimental biology*, 214(Pt 1):50–58.

[Robert C. Merton, 1974] Robert C. Merton, M. G. S. (1974). The optimality of a competitive stock market. *The Bell Journal of Economics and Management Science*, 5(1):145–170.

[Rosenman, 2012] Rosenman, E. T. R. (2012). Retweets—but not just retweets: Quantifying and predicting influence on twitter.

[Seidman, 1983] Seidman, S. B. (1983). Network structure and minimum degree. *Social Networks*, 5(3):269 – 287.

[Soloviev and Saptsin, 2011] Soloviev, V. and Saptsin, V. (2011). Heisenberg uncertainty principle and economic analogues of basic physical quantities. *Working Paper*.

[Wolfgang, 2008] Wolfgang, J. (2008). On Measurement and Analysis of Internet Backbone Traffic. page 116.