Trabajo de Fin de Máster

Máster en Modelización e Investigación Matemática, Estadística y Computación

# Aplicación de modelos evolutivos para el análisis del cambio social y cultural en redes de comunicación de gran escala

**Ignacio Morer Zapata**

Director: Manuel González Bedia
Director: Miguel Aguilera Lizarraga

Junio 2015

Universidad Zaragoza

# Aplicación de modelos evolutivos para el análisis del cambio social y cultural en redes de comunicación de gran escala


## RESUMEN

La investigación dentro del marco de las redes complejas supone actualmente una línea importante en muchas disciplinas. Sus avances permiten abordar y comprender problemas que son inaccesibles para otras metodologías, ya que revelan propiedades de carácter contraintuitivo (emergentes). Más allá del estudio aislado de los componentes de un sistema, existe una gran cantidad de información que se puede extraer de la interacción entre éstos, y a la que es preciso acceder para comprender su funcionamiento. Para ello, la teoría de redes constituye una potente herramienta para el estudio de la emergencia, organización... de procesos de esta naturaleza.

Desde esta perspectiva, los movimientos sociales son un objeto interesante de análisis. El rápido desarrollo de las tecnologías de información está marcando indudablemente su evolución en los últimos años: Internet y su creciente disponibilidad en los dispositivos móviles, junto con la popularidad de las redes sociales online, aceleran y multiplican las posibilidades de comunicación ciudadana y reconfiguran constantemente sus modos de organizarse. Además, toda esta actividad genera una cantidad ingente de datos que son de gran interés para la investigación.

En mayo de 2011 se produjo una gran movilización ciudadana, conocida como "Movimiento 15M". La alta participación dio lugar a una extensa red de comunicación offline-online sin precedentes en este país. Las numerosas congregaciones y la aparición de acampadas urbanas ponían de manifiesto una organización proveniente de las redes y, aparentemente, distribuída y no centralizada.

El objetivo de este estudio es comprender aspectos del funcionamiento de una movilización de este tipo, mediante el acceso a una de sus capas de interacción digital (Twitter). Se parte de un extenso dataset que abarca desde el 13 de mayo (dos días antes de las principales manifestaciones) hasta el final del mes, que recoge mensajes relativos al movimiento. Las fases del análisis son dos:

1. Se realiza una descripción exhaustiva de la topología de las redes interactivas formadas cada día, partiendo de la teoría de grafos. Al mismo tiempo, se presentan fenómenos comunes de redes complejas reales y buscamos sus similitudes con nuestro caso de estudio.

2. Nos centramos en el análisis de los mecanismos de difusión de la información. Basándonos en otras investigaciones y en lo observado en la fase descriptiva, planteamos una hipótesis sobre el carácter evolutivo de estos procesos y construimos unos modelos que nos ayuden a corroborarla.

El análisis nos muestra un sistema de gran tamaño con propiedades topológicas muy buenas desde el punto de vista de la conectividad y la eficiencia, así como otras características poco evidentes, que nos ayuda a entender la interacción del proceso de comunicación. Además, concluimos que nuestra hipótesis evolutiva es adecuada ya que resulta útil para explicar cómo se propaga la información.

# Índice de contenidos

# Índice de figuras

# Índice de tablas

# Parte I

# Memoria

# Capítulo 1

# Introducción

## 1.1 TIC y movimientos sociales. Caso de estudio.

La movilización social constituye una importante herramienta de la ciudadanía para mostrar su descontento y defender sus derechos. Estos procesos dinámicos, y en concreto sus formas de comunicación, han estado tradicionalmente limitadas por la tecnología, que ha impuesto restricciones tanto en el acceso a la información como en el modo de interactuar. El gran desarrollo de las tecnologías de la información en los últimos años lleva a pensar que la forma en que se organizan, difunden y materializan estos movimientos evolucione al mismo tiempo.

En efecto, con el derribo de estas barreras, se habilita la creación de redes más horizontales, distribuidas, interactivas y accesibles, algo que modifica drásticamente el activismo popular. El ciudadano puede obtener información de forma más sencilla y selectiva, y puede intervenir activamente en el proceso gracias a una elevada (y creciente) conectividad.

Con el término "tecnopolítica" se ha definido el uso táctico y estratégico de las herramientas digitales para la organización, comunicación y acción colectiva [1]. Se han dado multitud de casos de autoorganización en red combinados con acciones visibles. Entre los más destacados se encuentran:

- Protestas anti-globalización en Seattle (noviembre, 1999): Internet y los medios digitales facilitaron la coordinación y la cobertura de las protestas. Se articuló una red compuesta por ONGs locales, ciudadanos y activistas grass-roots que se tradujo en un entramado global que proporcionaba canales de información, debate y, en última instancia, de acción [2].

- Reacción ante la supuesta manipulación mediática del gobierno de España (13 de marzo de 2004): tras los atentados del 11 de marzo en Madrid, se coordinó vía SMS, una protesta frente a la sede del partido gobernante. Se probó que esta movilización (y las posteriores) no fueron promovidas por actores sociales vinculados a partidos políticos, sino que fueron de naturaleza ciudadana. La repercusión de estas acciones supusieron importantes incrementos en las comunicaciones móviles, con un aumento del 30% en el tráfico de mensajes de texto respecto a otro sábado cualquiera [3]. Paralelamente, también la actividad en Internet se veía incrementada, especialmente en portales activistas y con una amplia participación de colectivos extranjeros.

- Wikileaks. Caso Cablegate (noviembre de 2010): una organización sin ánimo de lucro que, manteniendo el anonimato de las fuentes, filtra informes y documentos de interés público. Unas 250.000 comunicaciones del Departamento de Estado de Estados Unidos con sus embajadas se publicaron y, además, fueron enviadas a cinco importantes periódicos internacionales. La persecución que sufrió la organización por parte de gobiernos e instituciones privadas provocó multitud de reacciones de apoyo y protesta en la red.

- Primavera Árabe (2010): una serie de alzamientos populares en países del norte de África y Oriente Próximo surgen para reclamar libertades y democracia. El uso de blogs en Egipto tuvo un importante papel desde el 2004, y los enlaces en red se multiplicaron con las redes sociales online. A finales de 2010, la multitudinaria revolución en Túnez contagió movilizaciones similares en países vecinos.

- Anonymous: un pseudónimo bajo el que se agrupan activistas que realizan acciones a favor de la libertad de expresión y la independencia de la red. No se conoce su estructura (si la hay) y, a pesar de lo difuso de su organización, han realizado operaciones importantes. Un ejemplo es el ataque a sitios web de instituciones que actuaron contra Wikileaks, como Paypal y Mastercard.

- NoLesVotes: es un movimiento surgido de la ley Sinde, aprobada en febrero de 2011. El objetivo era desalentar el voto a todos aquellos partidos que apoyaron dicha aprobación. Muchos ven en este caso un precedente claro del movimiento 15M.

- Democracia Real Ya y 15M: este movimiento auto-organizado fue agrupando miles de personas en las redes que, implicta o explícitamente, se identificaron con el nombre Democracia Real Ya. Este colectivo proclamó su rechazo al bipartidismo y al dominio de las instituciones financieras, y a favor de una democracia más participativa, manteniendose ajeno a partidos políticos y sindicatos. El crecimiento en ciudadanos afines pasó desapercibido para los medios, mientras se construía un estado de ánimo colectivo. Todo esto condujo a una serie de protestas multitudinarias y pacíficas en 70 localidades el día 15 de mayo de 2011, lo que pasó a llamarse "Movimiento 15M".

En este trabajo nos centramos en este último. Su carácter novedoso, su cercanía geográfica y la relevancia de las TIC en su comunicación lo convierten en una buena oportunidad de estudio. Se dispone de un amplio dataset que recoge mensajes relacionados con el movimiento, desde el 13 hasta el 31 de mayo de 2011.

## 1.2 Twitter y los movimientos sociales

La aparición en 2006 de la red social Twitter supuso un gran impacto por su carácter novedoso. Ofrece un servicio de *microblogging*, a través del cual se envían mensajes cortos que son recibidos por una red de asociados. Los mensajes, llamados *tweets*, tienen una longitud máxima de 140 caracteres. Al ser publicados, son recibidos instantáneamente por los seguidores del autor. Al contrario de lo que sucede en otras plataformas, esas relaciones de seguimiento no son recíprocas ni impiden el acceso a la información: cualquiera (incluidos los no usuarios) pueden ver lo publicado en Twitter.

Con el uso de ciertos caracteres especiales se ha modificado la forma de interacción. Por ejemplo, se puede especificar el receptor del mensaje incluyendo el carácter '@' seguido del nombre del usuario objetivo. Esta práctica tiene tres posibles significados:

- Respuesta a un *tweet* (*reply*).

- Mención a un usuario (*mention*)

- Reenvío de un mensaje (*retweet*), si antes de '@' aparecen los caracteres 'RT'.

Otra modificación del funcionamiento de este servicio surge de la aparación de etiquetas, llamadas *hashtags*, que tienen la función de reunir mensajes relativos a un tema con la intención de facilitar la búsqueda y discusión sobre el mismo. Para etiquetar un tweet se antepone el carácter '#' seguido de una palabra o palabras sin espacios (por ejemplo, #15M). Cuando el uso de un *hashtag* se acelera notablemente se convierte en lo que en Twitter se conoce como *trending topic*. Se mide a distintas

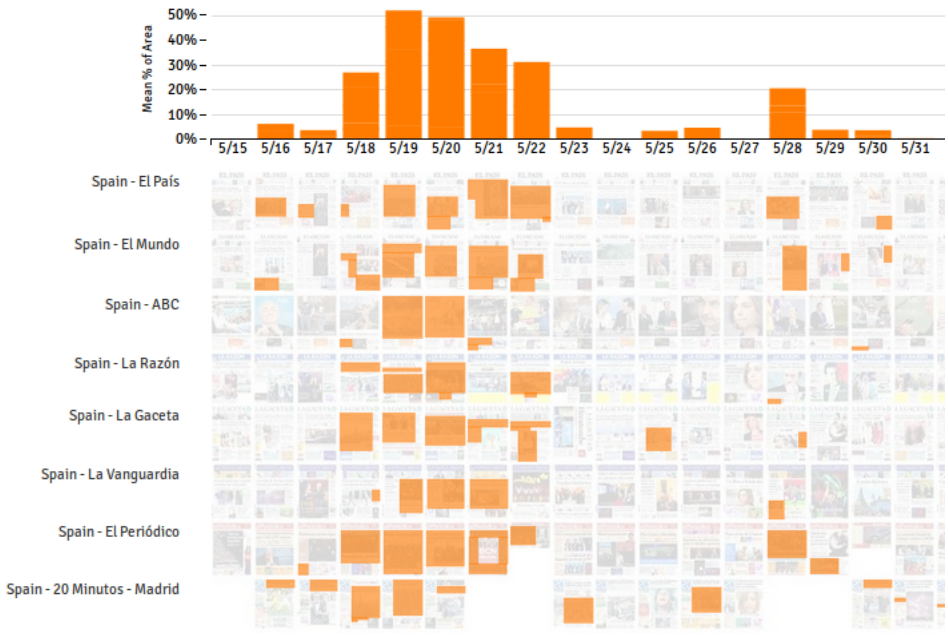



Figura 1.1: Repercusión mediática del Movimiento 15M, a través del análisis de portadas de prensa. Fuente: [4]

escalas: global, nacional y local. Llegar a ser *trending topic* implica haber logrado una enorme difusión y es considerado un éxito por aquellos que lo adoptan.

El *microblogging* mantiene una estrecha relación con los movimientos sociales: ofrece un medio de comunicación muy rápido y versátil, y con un potencial de propagación extraordinario. La posibilidad de englobar mensajes bajo un hashtag facilita el acceso directo a la información sobre un tema concreto. Sumado a un seguimiento activo, se puede construir un conjunto de datos representativo de un proceso de comunicación con identidad propia.

## 1.3 Motivacion y objetivo

Las acciones ciudadanas surgidas en momentos de agitación política suponen un fenómeno interesante para el estudio científico. La evolución de las tecnologías de la información y su evidente influencia en las formas de comunicación sugieren un cambio de paradigma científico que analice eficientemente estos procesos colectivos. Cuando la aproximación a estos fenómenos se realiza dentro de un marco más clásico es fácil interpretar errónea o sesgadamente la realidad. Sirva como ejemplo la escasa y desfasada repercusión en la prensa escrita de las movilizaciones multitudinarias del 15 de mayo de 2011 (Figura 1.1): hasta tres días después no hubo una presencia predominante de estos hechos en las portadas de prensa nacional [4].

Fenómenos como el 15M son de naturaleza muy llamativa. Originados por un estado de ánimo colectivo, se materializan en masivas acciones coordinadas cuyos participantes utilizan diversas formas de informarse y comunicarse. Nuestro objetivo global es conocer más acerca de la organización y el funcionamieto de estos movimientos auto-organizados en red. Como primer paso, pensamos que es clave analizar la estructura de sus interacciones. Esto se puede llevar a cabo gracias a los datos disponibles en las redes sociales *online* y con la teoría de grafos y redes complejas como principales

herramientas.

El segundo paso hacia nuestra meta es comprender los mecanismos de transmisión de información. Como todo proceso dinámico que tiene lugar en una red, está condicionado en primera instancia por la topología. No obstante, pensamos que también entran en juego otros factores que los hacen más complejos. Muchos procesos de propagación están restringidos por la naturaleza del medio en el que se encuentran: existen recursos limitados que son necesarios para que puedan desempeñar sus funciones. En el estudio de la transferencia de información se tiene en cuenta esto y por ello se proponen estrategias evolutivas para explicar el desarrollo cultural.

Creemos oportuno adoptar este enfoque para nuestro caso y pretendemos corroborar la hipótesis de que es un comportamiento evolutivo el que experimentan los mecanismos de difusión. Nuestra idea es identificar cómo las unidades de información se ven afectadas constantemente del entorno, por lo que su capacidad de reproducirse (y por tanto su permanencia) varía con el tiempo. Para ello, construiremos un modelo simple de tipo evolutivo cuyos mecanismos de reproducción se ajusten a nuestra hipótesis. Los datos simulados por este modelo se compararán con los reales y con otros dos modelos básicos para analizar qué parte de esa realidad somos capaces de capturar.

# Capítulo 2

# Bases teóricas

En este capítulo se presentan las características principales del enfoque adoptado para analizar un sistema, donde la clave se haya en la interacción entre sus partes. Consideramos que la primera etapa de un análisis de este tipo debe ser la explotación de la información estructural de la red, es decir, la disposición de sus componentes y las conexiones que los unen. Por ello, se presentan las medidas topológicas que se utilizarán para la descripción de nuestro caso de estudio, junto con algunos índices que sirvan para sustentar nuestra hipótesis evolutiva del capítulo 4.

## 2.1 Sistemas complejos: enfoque emergente

Frente a la perspectiva reduccionista de analizar un sistema a través de la suma de sus partes y sus relaciones lineales, surge como alternativa la teoría de la complejidad. El enfoque clásico tiende a hacer avanzar las disciplinas de forma independiente, provocando su divergencia y aislamiento progresivo [5]. No obstante, hay muchas características compartidas entre ellas y la aproximación científica a los sistemas particulares se puede realizar dentro de un marco global.

Para modelar eficientemente un sistema debemos comprender la relación del todo con sus partes, a través de las interacciones entre ellas. Partiendo de estas bases, el reto consiste en identificar las propiedades que emergen de la interacción. El acceso a estas se realiza a través de la información que contienen las uniones entre componentes, tanto desde un punto de vista estático (la estructura de red) como dinámico (la evolución de sus propiedades).

## 2.2 Redes complejas y teoría de grafos.

Recientemente, se están desarrollando multitud de técnicas y modelos que ayudan a entender e incluso predecir el comportamiento de los sistemas complejos. Frecuentemente, estos avances se inspiran en estudios empíricos de redes reales en distintos ámbitos:

- Tecnología: redes de suministro eléctrico[6], conexiones entre aeropuertos [7].
- Biología: redes de metabolismo [8], redes tróficas [9].
- Redes de información: enlaces en World Wide Web [15], co-autoría de artículos científicos [10].
- Redes sociales: red de intercambio de emails [11], relaciones de amistad [12].

Como herramienta de representación de sistemas complejos se usa la teoría de grafos, que supone una gran ayuda para detectar aquellas propiedades que emergen de las interacciones entre sus elementos. Estudia las propiedades de los grafos, $G$, estructuras formadas por un conjunto de nodos (vértices) $V$, unidos entre sí por enlaces (aristas) $E$:

$$G = (V, E)$$

$$E \subseteq (V \times V)$$

$$e_{12} = (v_1, v_2) \quad e_{12} \in E,\ v_1, v_2 \in V$$

A continuación se detallan una serie de propiedades topológicas básicas y fenómenos característicos de las redes complejas.

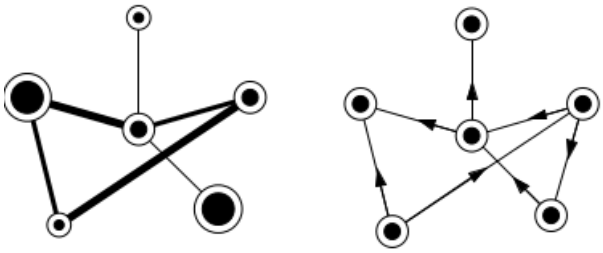
### 2.2.1 Topología de redes. Definiciones

#### 2.2.1.1 Direccionalidad y pesos

La presencia de atributos en nodos y enlaces aumentan la información del sistema y permite distinguir entre tipos básicos de redes. Típicamente, se realizan dos clasificaciones en función de la naturaleza de los enlaces. Si poseen una dirección, se distingue entre **redes no dirigidas**, cuyas aristas representan una conexión recíproca, y **redes dirigidas**, con enlaces que parten de un nodo origen y llegan a un nodo destino. Para estas últimas,

$$e_{12} = (v_1, v_2) \neq (v_2, v_1) = e_{21}$$

Asimismo, si los enlaces llevan asociado un peso, estamos ante **redes pesadas**. Es de vital importancia identificar la naturaleza de la red y valorar qué atributos debemos considerar, para estar seguros de incluir toda la información necesaria en el análisis (Figura 2.1).



(a) (b)

Figura 2.1: Direccionalidad: (a) Red no dirigida, con pesos en nodos y enlaces. (b) Red dirigida.

#### 2.2.1.2 Grado

La cantidad de enlaces adyacentes a un nodo se denomina **grado** y es denotado por $k_i$. Si se trata de una red dirigida, el grado es la suma de entrante y saliente:

$$k_i = k_i^+ + k_i^-$$

El grado pesado añade el valor de los pesos de las aristas. Este valor puede tener multitud de significados:

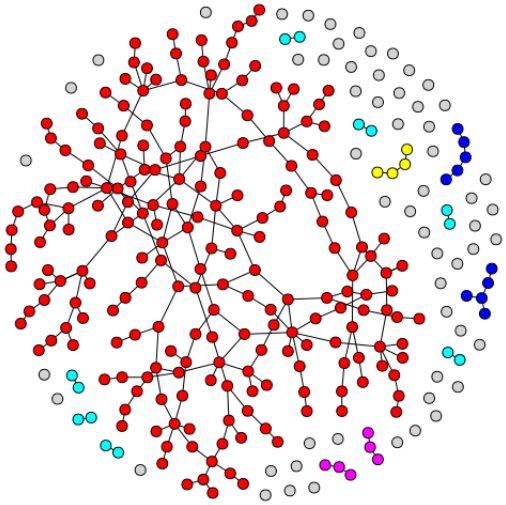


Figura 2.2: Componente gigante.

- El número de veces que se repite una interacción.
- La intensidad de la interacción.
- La distancia física.

Es habitual calcular el grado medio de los grafos para tener una idea de la conectividad global:

$$< k >= \frac{1}{n} \sum_i k_i$$

El caso de enlaces pesados se define análogamente:

$$< k_w >= \frac{1}{n} \sum_i k_{w,i}$$

#### 2.2.1.3 Componente conexo, débil y fuerte. Componente gigante

Los grafos pueden estar compuestos por subgrafos aislados unos de otros. Para grafos dirigidos, se definen dos tipos básicos de componentes conexos:

- Componente fuertemente conexo: para cada par de vértices $u$ y $v$ pertenecientes a este subgrafo existe un camino de $u$ a $v$ y viceversa.
- Componente débilmente conexo: la cohesión es independiente de la dirección de las aristas.

La existencia de un componente débilmente conexo que contenga una proporcion mayoritaria de los vértices, denominado **componente gigante**, es común en sistemas reales (Figura2.2). Constituye un síntoma del funcionamiento eficiente y coherente de un sistema [13].

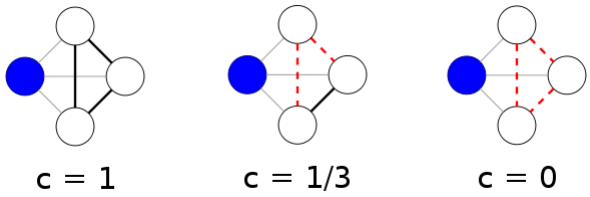



Figura 2.3: Coeficiente de clustering.

#### 2.2.1.4 Coeficiente de clustering

Es una propiedad local de los nodos, que mide el nivel de conectividad entre sus vecinos directos. Para un nodo $i$:

$$C_i = \frac{M_i}{k_i\,(k_i - 1)}$$
$$0 \leq C_i \leq 1$$

con $M_i$ el número de enlaces que se dan entre sus vecinos directos de $i$, y $k_i$ su grado (Figura 2.3) . A nivel global, se mide el valor medio del clustering para todos los nodos:

$$C = \frac{1}{n}\sum_i C_i$$

Con este valor se tiene un indicador de la robustez local de la red.

#### 2.2.1.5 Distancia promedio

La distancia entre dos nodos $i$ y $j$ (también llamado camino corto, o geodésico), $d\,(i,j)$, es el mínimo número de aristas que los separan. En un grafo $G = (V, E)$, con $n$ nodos, la distancia promedio $L$ es la media de longitud de todos los caminos existentes:

$$L = \frac{1}{n\,(n-1)}\sum_{i,j} d\,(i,j)\ \ \forall i,j \in V$$

Un valor bajo de $L$ es un signo de eficiencia de la red, puesto que sus componentes necesitan menos pasos para llegar de uno a otro.

### 2.2.2 El fenómeno del "mundo pequeño"

El origen de este concepto se encuentra en el experimento del psicólogo americano Stanley Milgram (1967). Se seleccionaron aleatoriamente habitantes del medio oeste americano y se les propuso hacer llegar una carta a un extraño de la costa este del país. Los únicos datos proporcionados sobre el destinatario eran el nombre, la ocupación y su localización geográfica aproximada. Cada persona enviaba el mensaje a alguien en su red directa de contactos que pudiese estar cerca del objetivo, basándose sólo en esos tres datos. El procedimiento continuaba hasta que el mensaje dejaba de viajar o llegaba a su destino. A pesar de que únicamente 64 de 296 mensajes alcanzaron el objetivo, lo hicieron en muy pocos pasos, concretamente, en una media de 5.2 conexiones.

Dos modelos clásicos de redes establecen las bases para la comprensión de este fenómeno:

- Redes regulares (del tipo *ring lattice*), en las que tanto coeficiente de clustering alto como distancia promedio son valores altos.

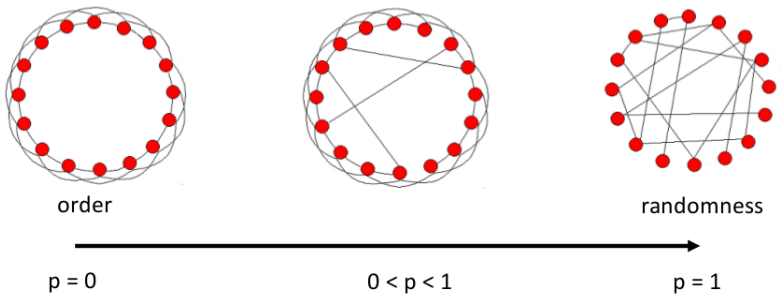



Figura 2.4: Aleatoriedad creciente: de red regular a red aleatoria.

- Redes aleatorias (modelo Erdos-Renyi), donde la distancia promedio es baja, pero la aleatoriedad dificulta agrupamientos locales ($C \downarrow$)

En muchas redes reales (como la de Milgram) se encuentran similitudes con ambos modelos. Con las redes aleatorias se comparte en una distancia promedio corta, mientras que el agrupamiento local es similar al de redes regulares.

Este fenómeno se conoce como **mundo pequeño**. Watts y Strogatz lo estudiaron y construyeron un modelo que replicase estas características [14]. Partiendo de la red regular, se recorren todas las aristas y con una probabilidad $p$ se cambia uno de sus extremos por otro individuo escogido aleatoriamente. Se va aumentando la aleatoriedad a través de $p$, obteniendo redes que cubren todo el espectro desde $p = 0$ (regular) hasta la aleatoriedad total, $p = 1$ (Figura 2.4). En los sistemas obtenidos para $p \in (0, 1)$ existe un rango intermedio que contiene las buenas propiedades de estructuración local del modelo regular y la eficacia de comunicación del modelo aleatorio. Matemáticamente, el comportamiento de $L$ y $C$ en función de $p$

$$p \to 0 \Rightarrow L \sim \frac{n}{2k}\,,\, C \sim \frac{3}{4}$$

$$p \to 1 \Rightarrow L \approx L_{random} \sim \frac{\ln(n)}{\ln(k)}\,,\, C \approx C_{random} \sim \frac{k}{n}$$

Cuando $p \to 0$ aparecen valores altos de $L$ (que crece linealmente con $n$) y de $C$ (en torno a $\frac{3}{4}$). Si $p \to 1$ las dos variables parecen alcanzar valores bajos al mismo tiempo ya que $L$ sólo crece logarítmicamente con $n$, y $C$ cae abruptamente. Sin embargo, la representación de la variación de los coeficientes con $p$ (normalizados por sus correspondientes valores en grafo regular, $L(0)$ y $C(0)$) revela un rango en el cual $L$ presenta valores bajos y $C$ mantiene valores elevados (Figura 2.5).

Para evaluar si se da este fenómeno, se comparan $C$ y $L$ con sus valores análogos del caso aleatorio y se comprueba si:

$$L \gtrsim L_{random}$$

$$C \gg C_{random}$$

En ocasiones se usa el índice $\sigma$ para contabilizar en qué medida la red presenta este fenómeno:

$$\sigma = \frac{C/C_{rand}}{L/L_{rand}}$$

Una red con $\sigma > 1$ se considera de mundo pequeño.

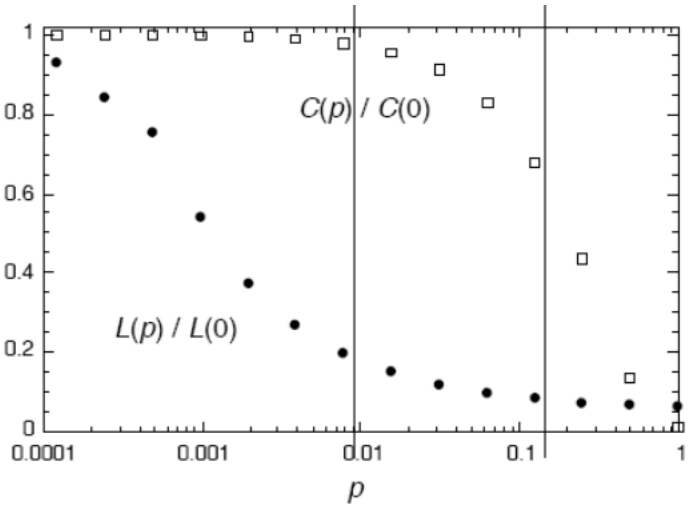



Figura 2.5: Evolución de $C$ y $L$ frente a $p$.Fuente:[14]

### 2.2.3 Distribución de grado. Redes libres de escala y conexión preferencial

Otra de las características presente asiduamente en sistemas reales deriva de la distribución del grado de sus nodos. La gran mayoría de los individuos presentan un grado bajo, mientras que sólo unos pocos, llamados *hubs*, están muy conectados. Esta distribución responde a una ley de potencias (*power-law*) caracterizada por una larga cola (Figura 2.6). La probabilidad de que un nodo escogido aleatoriamente

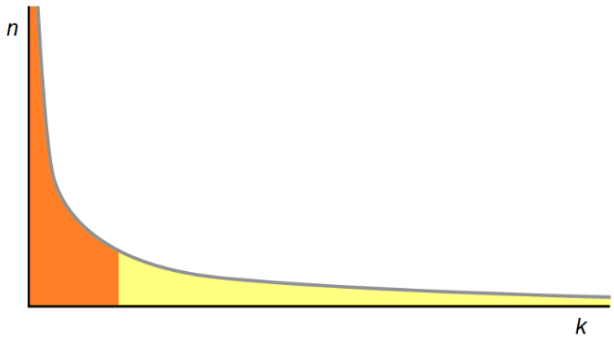



Figura 2.6: Distribución según ley de potencias.

tenga grado $k$ es:

$$p\left(k\right) = Ck^{-\alpha}$$

Su representación en escala logarítmica pone en evidencia este tipo de redes, conocido como **redes libres de escala**:

$$\log p\left(k\right) = \log C - \alpha \log k$$

donde $\alpha$ es el exponente de la ley de potencias. En la realidad, los valores del exponente recaen típicamente en el rango $\alpha \in (2, 3)$ [15, 16]. La forma más conveniente de representación es su función de distribución acumulada (complementaria), $P\left(X \geq x\right)$, que también sigue una ley de potencias de exponente $\beta = \alpha - 1$.

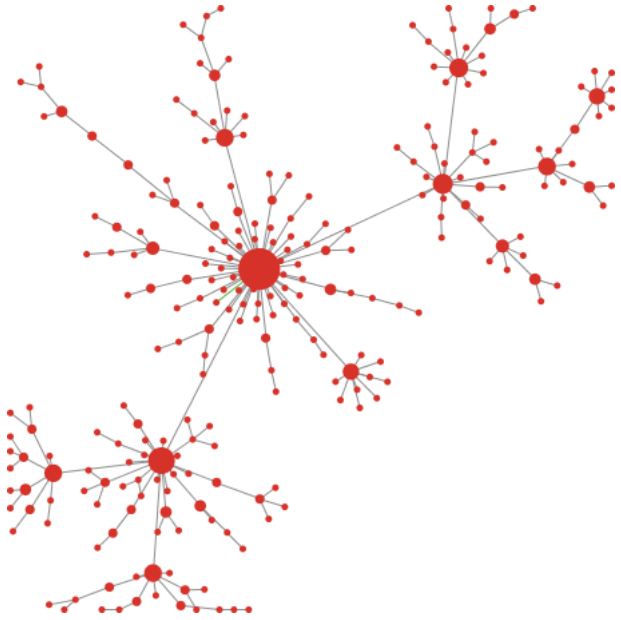


Figura 2.7: Red libre de escala.

Uno de los mecanismos clásicos que generan este tipo de redes es el denominado conexión preferencial (o *rich-get-richer*). Fue introducido por Price [10], y popularizado por Barabasi y Albert en un modelo de generación de redes de uso muy extendido [17], que se basa en dos principios básicos:

- Las redes crecen continuamente con la aparición de nuevas aristas.
- Las nuevas aristas se conectarán con mayor probabilidad a nodos que ya estén altamente conectados (Figura 2.7).

#### 2.2.3.1 Significado del exponente

Para evaluar el significado de $\alpha$ se utiliza el índice de Gini, usado tradicionalmente para medir la desigualdad en términos de ingresos económicos. Se calcula como una proporción de las áreas en el diagrama de la curva de Lorenz (Figura 2.8):

$$g = \frac{a}{a+b}$$

$$g \in [0, 1]$$

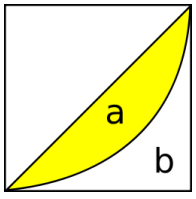



Figura 2.8: Curva de Lorenz e índice de Gini.

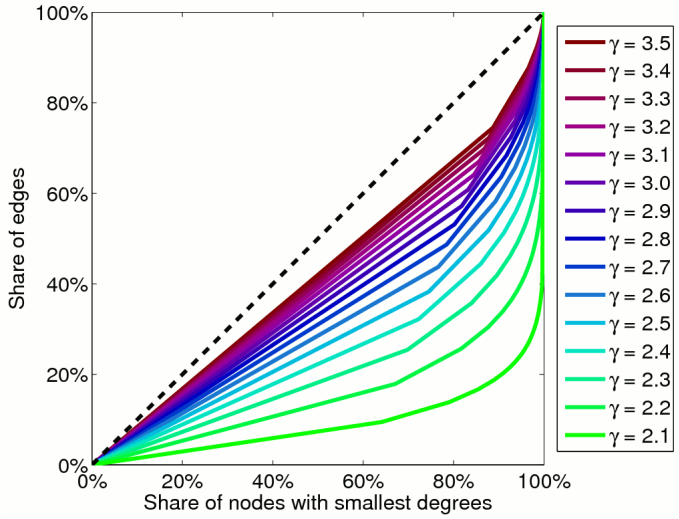



Figura 2.9: Curvas de Lorenz para ley de potencias. Fuente: [18]

Para ver qué relación guarda con las ley de potencias, se representa la curva de Lorenz para valores de $\alpha \in [2.1, 3.5]$ . Comprobamos que la distribución se hace más equitativa (el coeficiente de Gini disminuye) conforme aumenta el valor del exponente (Figura 2.9).

#### 2.2.3.2 Ajuste de power-law

A la hora de determinar si una serie de datos se ajusta a una distribución power-law es muy común incurrir en una serie de errores típicos. Con datos continuos, si se parte de un histograma, estamos incluyendo variables propias de esta representación que pueden influir en el ajuste obtenido (el número de bins y su anchura). Además, es habitual que los primeros puntos de la distribución no respondan bien al ajuste y que sólo a partir de un umbral, $x > x_{min}$, se siga esta ley. Por ello, para construir los modelos de ajuste se sigue el siguiente proceso [19]:

- Estimar los parámetros $x_{min}$ y $\alpha$ por el método de máxima verosimilitud.
- Comprobar la hipótesis de ajuste al modelo power-law con los parámetros anteriores a través de test estadísticos con un nivel de significación de 0.1.

### 2.2.4 Detección de comunidades

Además de los fenómenos presentados, otra propiedad típica de las redes complejas es una estructura dividida por comunidades [20]. Su presencia puede ayudar a identificar grupos más amplios cuya interacción los ha diferenciado del resto del sistema.

En definitiva, se pretende detectar grupos de nodos densamente conectados entre sí, en relación a lo conectados que están con el resto de la red. Existe un amplio conjunto de algoritmos que realizan esta tarea [21]. Recientemente, se han desarrollado técnicas más versátiles que mejoran los resultados [22][23]. Para evaluar la calidad de las particiones, se calcula la modularidad $Q$, que mide la densidad de enlaces dentro de las comunidades en comparación con los enlaces inter-comunitarios. En la práctica, se toma como válida una división con $Q > 0.3$ [24]. En el caso de aristas pesadas,

$$Q = \frac{1}{2m} \sum_{i,j} \left[ A_{ij} - \frac{k_i k_j}{2m} \right] \delta\left(c_i, c_j\right)$$

$$Q \in [-1, 1]$$

donde $A_{ij}$ representa el peso del enlace entre los nodos $i$ y $j$, $k_i = \sum_j A_{ij}$ es el grado pesado de $i$, $c_i$ es la comunidad a la que pertenece $i$, la función $\delta(u, v) = \begin{cases} 1 & u = v \\ 0 & u \neq v \end{cases}$, y $m = \frac{1}{2} \sum_{ij} A_{ij}$.

#### 2.2.4.1 El método de Louvain

Es un método basado en la optimización de la modularidad que admite redes dirigidas y pesadas [25]. El método despliega completamente las estructuras encontradas a distintos niveles de resolución. Así, se pueden buscar comunidades de mayor o menor tamaño. Se repiten iterativamente dos fases:

1. Se asigna a cada nodo su propia comunidad. Para cada nodo $i$ y sus vecinos $j$, se evalúa la ganancia de $Q$ que se obtiene situando a $i$ en la comunidad $C$ a la que pertenece $j$:

$$\triangle Q = \left[ \frac{\sum_{in} + k_{i,in}}{2m} - \left( \frac{\sum_{tot} + k_i}{2m} \right)^2 \right] - \left[ \frac{\sum_{in}}{2m} - \left( \frac{\sum_{tot}}{2m} \right)^2 - \left( \frac{k_i}{2m} \right)^2 \right]$$

   con $\sum_{in}$ la suma de pesos dentro de una comunidad $C$, $\sum_{tot}$ la suma de pesos de enlaces adyacentes a los nodos en $C$ y $k_{i,in}$ la suma de pesos de enlaces entre $i$ y nodos de $C$. Se da por concluida la fase cuando se obtiene un máximo local: ningún movimiento individual puede mejorar la modularidad.

2. Se construye una nueva red cuyos nodos son las comunidades encontradas en la fase 1, con enlaces pesados entre ellas. El peso de un enlace de esta red entre $i$ y $j$, se corresponden con la suma de los pesos de los componentes de $c_i$ que unidos a los de $c_j$. Se repite la fase 1 con esta nueva configuración.

De todas las divisiones realizadas se puede extraer información valiosa, como cuál es la actividad entre comunidades o comprobar si hay pequeños grupos dentro de otros mayores. El método proporciona por defecto la división con modularidad mayor.

## 2.3 Índices de diversidad

Añadimos esta sección que incluye medidas usadas típicamente en ecología para medir la diversidad de los ecosistemas. Las emplearemos en el capítulo 4 siguiendo la línea de la analogía de los mecanismos evolutivos presentada en la introducción.

La diversidad de un ecosistema es una medida cuantitativa que refleja, por un lado, el número de especies presentes en un conjunto de datos, y, simultáneamente, en qué medida están repartidos equitativamente. Existen varios índices de diversidad que permiten hacer ésto de forma sintética. Entre los de uso frecuente están el índice de diversidad de Simpson, el índice de Shannon-Wiener (o entropía de Shannon) y la *species evenness*, una normalización de la entropía de Shannon.

### 2.3.1 Índice de diversidad de Simpson

Mide la probabilidad de que dos individuos escogidos al azar sean de la misma especie:

$$l = \frac{\sum_{i=1}^{R} n_i (n_i - 1)}{N (N - 1)}$$

con $n_i$ número de individuos de especie $i$, $R$ el número total de especies y $N$ el número total de individuos del sistema. En conjuntos de datos grandes se puede aproximar por

$$\lambda = \sum_{i=1}^{R} p_i^2$$

$$\lambda \geq \frac{1}{R}$$

La característica principal de $\lambda$ es la ponderación mayor que asigna a las especies dominantes, puesto que, comparativamente, los grupos poco numerosos apenas contribuyen a la suma. Para que el índice sea creciente con la diversidad se suele emplear una transformación de éste. Aquí se usará la inversa:

$$D = \frac{1}{\lambda}$$

Nótese que para poder comparar entre diferentes sistemas conviene dividir por $D_{max}$, para que esté expresado en términos relativos a su máxima diversidad:

$$D' = \frac{D}{D_{max}}$$

$$D_{max} = \frac{1}{\lambda_{min}} = \frac{1}{(1/R)} = R$$

### 2.3.2 Entropía de Shannon y *species evenness*

Proviene del trabajo de Claude Shannon, y fue una de sus contribuciones a los orígenes de la teoría de la información. Aplicado a la ecología, mide la incertidumbre al predecir la especie a la que pertenecerá un individuo escogido aleatoriamente:

$$H = -\sum_{i=1}^{R} p_i \ln p_i$$

Igual que en el índice de Simpson, se suele hacer una transformación que permita la comparación entre sistemas de distintos tamaños. La denominada *species evenness,* $J$, es el índice $H$ normalizado por su valor máximo. Es útil cuando se quiere descartar la contribución del número de especies y medir únicamente el grado de similitud en la distribución de las poblaciones:

$$J = \frac{H}{H_{max}}$$

$$H_{max} = \log(R)$$

## 2.4 Herramientas

Existen muchas alternativas en cuanto a software para el análisis de redes. De entre todas ellas, se ha usado principalmente *networkx*, un paquete de software en lenguaje Python para el estudio de estructura, dinámica y funciones de las redes complejas. También se han empleado otros paquetes, siempre basados en Python, para la manipulación de datos y algún otro punto del análisis. Los detalles y el listado de paquetes usados se encuentra en el Apéndice A.

Adicionalmente, hemos usado *Gephi*, una plataforma interactiva de exploración y visualización de grafos de código abierto y gratuita. Con ella se han realizado cálculos complementarios y algunas visualizaciones.

## 2.5 Conclusión

En este capítulo se recogen las definiciones de conceptos y técnicas que emplearemos a lo largo del documento, agrupadas en dos bloques. En el primero (sección 2.2) explicamos el uso de la teoría de grafos como herramienta para estudiar la topología de las redes. Se comienza por presentar los distintos tipos de sistemas que hay, atendiendo a la naturaleza de sus nodos y enlaces. Después se definen las medidas para caracterizar la conectividad, tanto a nivel local como a nivel global. Con todo esto, repasamos tres fenómenos interesantes que se dan con frecuencia en redes complejas. Queremos comprobar si están presentes en nuestro caso de estudio, y para ello se explica cómo se generan, cómo se detectan y qué implicaciones tienen. Todo ello se aplicará en el Capítulo 3.

En el segundo bloque (sección 2.3) se definen tres índices usados habitualmente para medir la diversidad de especies en un ecosistema. El motivo de esto radica en la hipótesis propuesta en el Capítulo 4, que establece una analogía entre nuestros sistemas y un ecosistema. Creemos que una forma adecuada de evaluar similitudes entre ellos es comparar sus diversidades.

# Capítulo 3

# Análisis de topología y conectividad efectiva de la red de interacción

Procedemos a aplicar a nuestro caso lo presentado en la sección 2.2. En primer lugar, realizamos una descripción a grandes rasgos, empezando por las características de la base de datos y la extracción de la red de interacción. A partir de ahí, caracterizamos de forma más exhaustiva la topología de la red y evaluamos en qué medida el sistema contiene las propiedades de casos típicos de redes complejas.

Es especialmente importante analizar la topología del sistema por dos motivos. Por un lado, dado que se trata de una red formada por la interacción, el estudio de su estructura contiene la información sobre cómo se han establecido estas relaciones. Pero además, hay que tener en cuenta que es la topología de la red lo que sostiene los procesos dinámicos que se dan en ella, como los procesos de difusión de información, que analizaremos en el Capítulo 4. Así que las conclusiones que se obtengan aquí estarán directamente vinculadas a la siguiente parte del análisis y ayudarán a definir su planteamiento.

## 3.1 Descripción del dataset

Partimos de un extenso conjunto de datos [26] recogido a lo largo de 19 días, formado por 1438375 entradas que contienen en el cuerpo del mensaje alguna de estas palabras clave relacionadas con nuestro caso de estudio. Entre ellos están '#15M', así como otros *hashtags* relacionados con el movimiento: #nolesvotes, #democraciarealya, #spanishrevolution, #acampadasol... Se contabiliza el número de interacciones, así como su naturaleza (*retweet*, mención-respuesta). En la Tabla 3.1 y la Figura 3.1 se muestran los detalles por día y el volumen total de actividad, así como su división por categorías.

Más de la mitad de los mensajes recogidos son *retweets*, lo que denota la importancia de los procesos de difusión sobre el resto de interacciones. A pesar de que fue el día 15 el de las primeras manifestaciones, la actividad no explota hasta dos días despues. Ese momento recoge el efecto de la extensión del movimiento a gran parte del país. Posteriormente decae de forma abrupta para mantenerse en niveles bajos, similares a los iniciales. Como excepción se encuentra el día 27, en el que se produjo el desalojo de Plaça Catalunya, donde se encontraba la acampada de Barcelona.

## 3.2 Cohesión y conectividad

Construimos una red basada en la interacción del modo siguiente: una mención dentro del texto del mensaje se traduce en una pareja de nodos unida por un enlace dirigido, partiendo del autor y

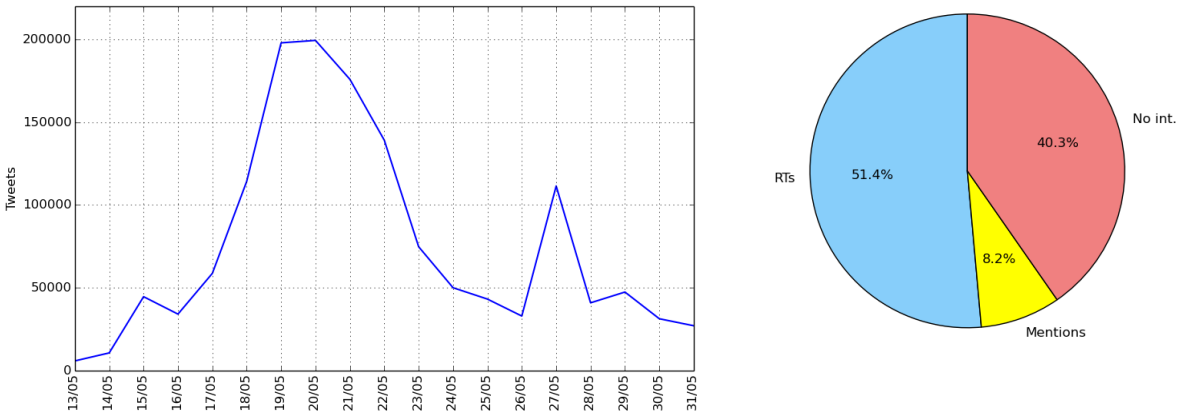



Figura 3.1: Actividad y tipología de los datos.

| Día | Total tweets | Interacciones | % Retweets | % Menciones | % No interacción |
|---|---|---|---|---|---|
| 13 | 5770 | 4471 | 51,73 | 6,38 | 41,89 |
| 14 | 10598 | 8219 | 56,05 | 4,62 | 39,33 |
| 15 | 44555 | 40439 | 59,34 | 4,92 | 35,74 |
| 16 | 34036 | 27017 | 53,75 | 6,70 | 39,56 |
| 17 | 58635 | 43579 | 53,74 | 6,38 | 39,88 |
| 18 | 114254 | 87225 | 53,94 | 6,48 | 39,58 |
| 19 | 197847 | 155517 | 50,03 | 8,14 | 41,82 |
| 20 | 199317 | 163147 | 49,83 | 9,03 | 41,15 |
| 21 | 175828 | 138136 | 49,25 | 9,25 | 41,50 |
| 22 | 139067 | 101145 | 48,24 | 7,65 | 44,11 |
| 23 | 74710 | 57777 | 51,27 | 8,53 | 40,20 |
| 24 | 49969 | 40496 | 46,63 | 10,48 | 42,89 |
| 25 | 43107 | 33454 | 46,12 | 10,01 | 43,87 |
| 26 | 32859 | 28334 | 46,85 | 10,30 | 42,86 |
| 27 | 111396 | 100403 | 58,63 | 8,25 | 33,12 |
| 28 | 40846 | 35752 | 54,23 | 9,18 | 36,59 |
| 29 | 47351 | 40777 | 57,73 | 7,10 | 35,17 |
| 30 | 31230 | 28132 | 52,77 | 8,96 | 38,27 |
| 31 | 27000 | 20168 | 46,98 | 9,82 | 43,20 |

Table 3.1: Resumen del conjunto de datos.

apuntando al usuario mencionado. El peso aumentará una unidad cada vez que se repita la interacción[1].

Como primera medida de la cohesión de la actividad de la red, se extrae el componente gigante. Casi la totalidad de enlaces (99.32%) y de los usuarios (95.27%) conforman el componente débilmente conexo más grande, lo que demuestra la cohesión del proceso (Tabla 3.2).

| Red global | | Componente gigante | |
|---|---|---|---|
| Nodos | 176480 | Nodos | 168137 (95.27%) |
| Enlaces | 884977 | Enlaces | 879009 (99.32%) |
| Total interacciones | 1154188 | Total interacciones | 1148122 (99.47%) |
| | | Grado medio | 10.4559 |

Tabla 3.2: Resumen de la red global.

El análisis a partir de este punto se realiza sobre las componentes gigantes, lo que resulta útil por dos razones:

1. Evita problemas a la hora de aplicar algoritmos. Por ejemplo, la distancia entre dos nodos de distintos componentes conexos sería infinita, y esto afectaría al cálculo de la distancia promedio.

2. Elimina ruido (mensajes que no pertenecen realmente al proceso). Por ejemplo, existen mensajes que con '15M' se refieran a otro asunto (15 millones, 15 minutos...), y por consiguiente, si en ellos se da una interacción difícilmente estará conectada a la componente gigante y será descartada.

Con el fin de identificar distintas fases en el proceso, se divide la red final en 19 redes diarias (datos detallados en el anexo B). El tamaño de la red diaria (Figura 3.2) sigue un comportamiento muy similar a la actividad total: comienza a crecer rápidamente a partir del día 16, alcanzando sus valores máximos en los días 19, 20 y 21. El número de participantes desciende hasta cerca de 10000 (día 26) para posteriormente mostrar otro pico de actividad el día 27. A lo largo del tiempo considerado, el componente gigante se mantiene siempre cerca del total.

Estudiamos ahora la conexión media de la red. En la Figura 3.3a se muestra la evolución del grado medio y grado medio pesado, $< k >$ y $< k_w >$. De forma promedio, se dan un mínimo de 4 interacciones distintas por usuario. El máximo se alcanza el día 15, momento en el que se materializan las manifestaciones y las acampadas. A partir del día 18 se aprecia otro aumento importante. Además, si atendemos a la distancia entre grado y grado pesado, vemos que también desde el día 18 aumenta del 10% para situarse en torno al 20% (Figura 3.3b). De estos datos interpretamos lo siguiente:

- El pico de conexiones el día 15 es un síntoma de red muy conectada. Se interactuó con una media 6.5 usuarios (7 interacciones en total si contamos enlaces recurrentes). En ese momento el sistema todavía no era muy grande ($\approx$ 11000 nodos, frente al máximo diario de casi 50000) pero los que lo componían se conectaron de manera muy amplia.

- La siguiente fase de conectividad alta coincide con los momentos de mayor participación, donde la repeteción en las interacciones aumenta. Estos datos sugieren que la red se ha familiarizado con las formas de comunicación, puesto que difusiones y respuestas recurrentes son más comunes, incluso en los momentos en los que $< k >$ es bajo.

[1] Nótese que un mensaje con varias menciones implica varias parejas de enlaces, mientras que uno sin menciones no se representa.

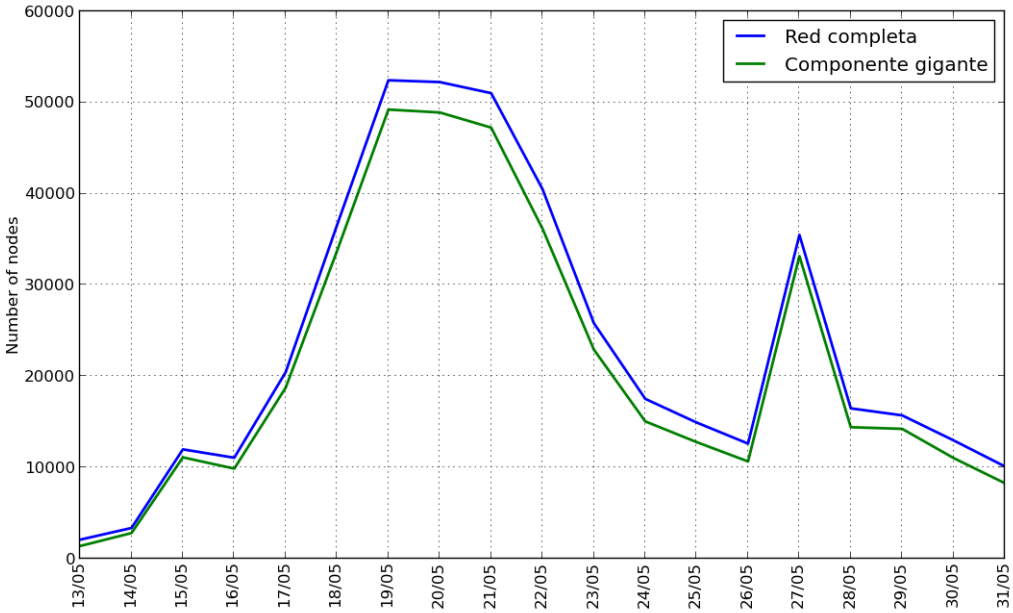



Figura 3.2: Tamaño de la red.

## 3.3 Mundo pequeño

Vamos a evaluar en qué medida nuestro sistema presenta características de mundo pequeño. Se calcula el índice $\sigma$ para las redes diarias (detalles en anexo B). La condición de mundo pequeño, $\sigma \gg 1$ , se cumple con creces en todos los casos (Figura 3.4). El valor de $\sigma$ crece con el tamaño de las redes, debido a que, al generar grafos aleatorios equivalentes, el clustering $C_{random}$ cae mucho más rápidamente que sus correspondientes valores reales $C$. Los niveles de interacción local estructurada se conservan, a pesar del aumento del tamaño de red. Al mismo tiempo, los valores de $L$ se mantienen por debajo de sus homólogos aleatorios, un indicador de la eficiencia en la transmisión de la información.

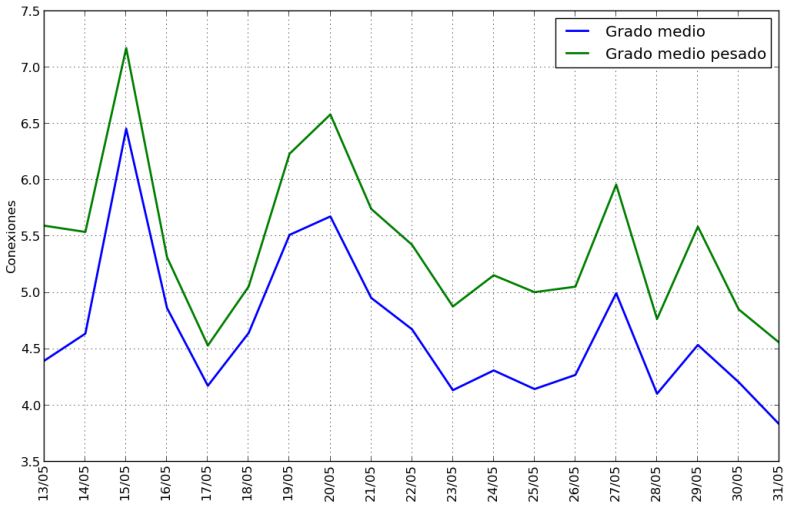
Se observan comportamientos interesantes:

- Del día 16 al 17 se produce un incremento notable del índice de mundo pequeño, cuadruplicando su valor (Figura 3.3). Esto sucede antes de la explosión de actividad a partir del día 18. Se podría pensar que cuando la red se convierte en un "mundo muy pequeño" se dan las condiciones necesarias para momentos de gran actividad.

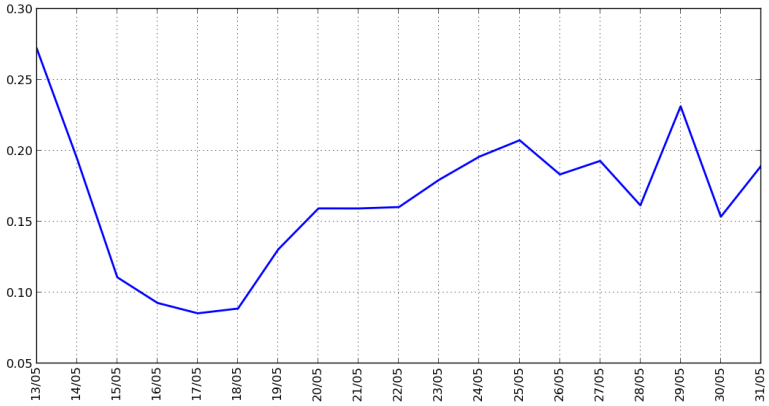
| Día | $n_i$ | $n_i/n_{i-1}$ | $\sigma_i$ | $\sigma_i/\sigma_{i-1}$ |
|---|---|---|---|---|
| 16/05 | 9861 | 0.889 | 419.5724 | 1.042 |
| 17/05 | 18723 | 1.898 | 1659.4422 | **3.955** |
| 18/05 | 33721 | 1.801 | 2282.0021 | 1.375 |

Tabla 3.3: Incremento del índice $\sigma$ de mundo pequeño.

- El valor de $\sigma$ no vuelve a los bajos valores de los 4 primeros días, a pesar de que el tamaño sí lo haga. Cuando la red alcanza su madurez es más difícil que pierda las buenas propiedades de




(a) Evolución grado medio y grado pesado medio.




(b) Distancia grado-grado pesado.

Figura 3.3: Conectividad media.

robustez local y eficiencia.

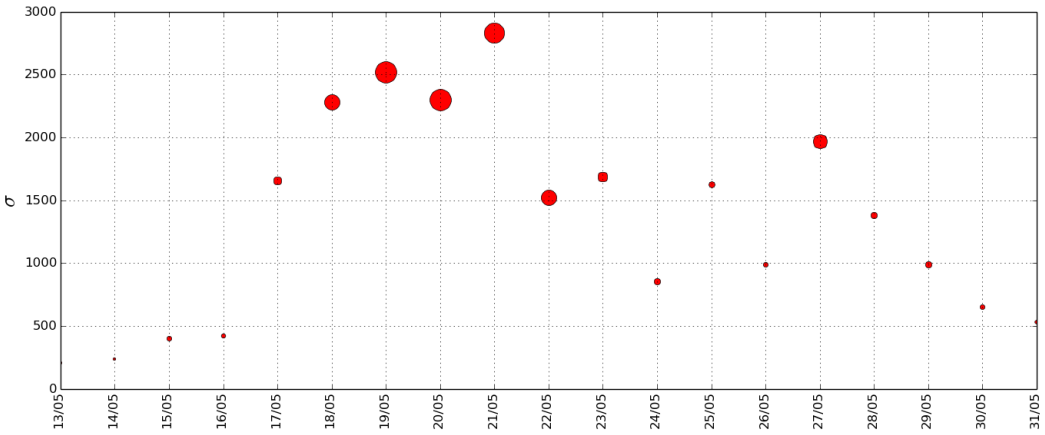



Figura 3.4: Índice $\sigma$ de mundo pequeño. Los puntos se representan con tamaño proporcional al número de nodos de su red.

## 3.4 Distribución de grados

Para obtener más detalles de la interacción, se estudia la distribución de grados entrantes y salientes de cada grafo diario. Como sucede habitualmente en sistemas de este tipo, las distribuciones parecen responder a leyes de potencia, presumiblemente alcanzadas a través de conexión preferencial. Como comprobación, se realiza un ajuste para cada serie de datos utilizando el método descrito en 2.2.3.2 y se comentan los resultados basándonos en los exponentes de las distribuciones power-law que se obtienen (Figura 3.5):

- Todos los exponentes (de entrada y salida, con y sin pesos) recaen dentro del rango típico de las redes libres de escala reales:
$$\alpha \in [2, 3]$$
- A partir del día 15, los exponentes de grados de salida son mayores que para los grados de entrada, para después estabilizarse ambos en torno a ciertos valores. La red experimenta un periodo de transición o aprendizaje en los primeros días, pero normaliza su funcionamiento.
- La distribución de los grados de salida es más equitativa ($\alpha$ mayor), según lo explicado en la sección 2.2.3.1. La razón de esta diferencia radica en la distinta naturaleza de los enlaces salientes y entrantes. Los salientes se hacen de forma intencionada, y están en cierto modo restringidos por la dedicación que un usuario puede dedicar a su actividad. En cambio, la entrada no supone ningún esfuerzo para el receptor. Por ello, el grado entrante está menos limitado y presenta una distribución más desigual.

Se muestra un ejemplo en la Figura 3.6. Para el mismo día, se ve cómo la pendiente de la recta es más elevada la distribución del grado de salida, $\alpha_{in} < \alpha_{out}$, lo que denota menos desigualdad.

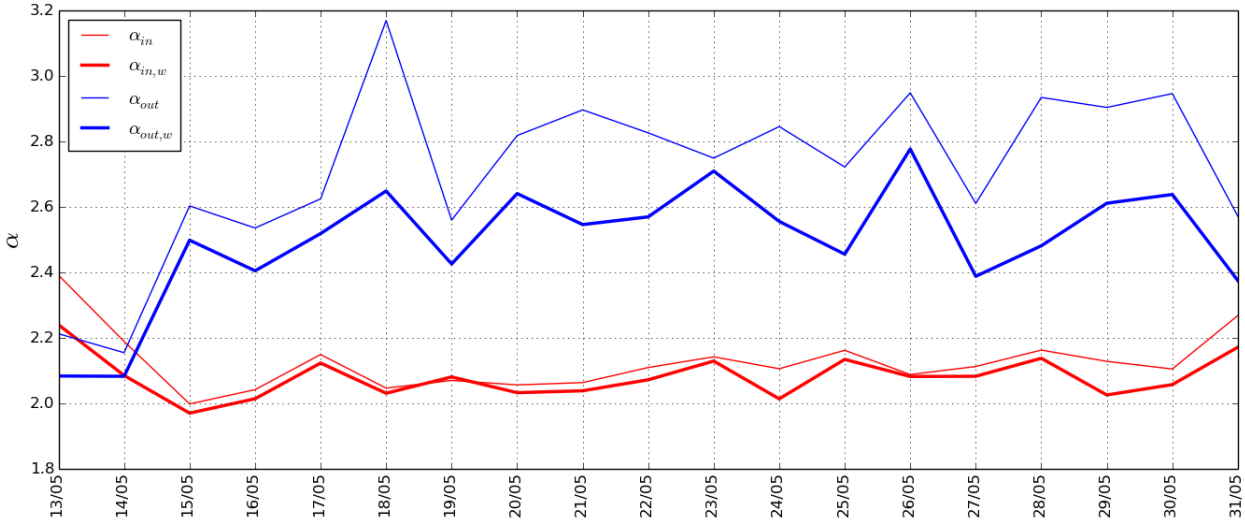



Figura 3.5: Exponentes de power-laws.

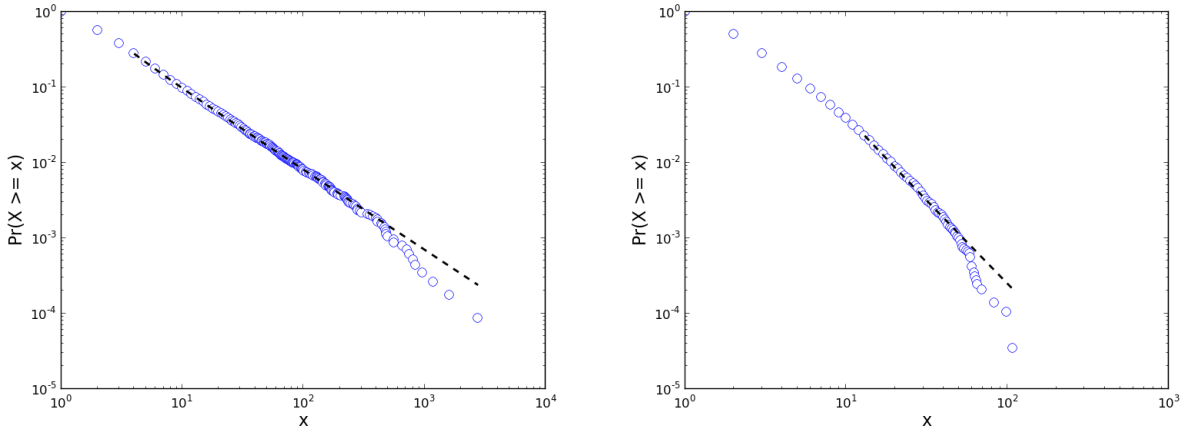



Figura 3.6: Ajuste power-law. Ejemplo: día 18 ($k_{in}$ y $k_{out}$).

## 3.5 Detección de comunidades

Se aplica el método descrito en 2.2.4 para detectar si en la red interactiva se forman comunidades de usuarios. El resultado es una división en 95 comunidades, con un valor de modularidad $Q = 0.361745$, que cumple la condición de validación típica $Q > 0.3$ [24]. La red resultante se puede ver en la Figura 3.7, en la que el tamaño de un nodo es proporcional al número de miembros de la comunidad, y su nombre pertenece al nodo de mayor grado.

A simple vista, se percibe una gran diferencia entre el tamaño relativo de cada comunidad. Representamos la proporción de nodos acumulada en función del número de comunidades, empezando por las de mayor tamaño (Figura 3.8). Los 13 grupos más numerosos ocupan más del 84% del tamaño total de la red [2] . Esto nos lleva a formular dos preguntas:

- ¿Se mantiene el protagonismo de las comunidades grandes a lo largo del proceso? Para responder a esto, se visualiza la proporción de interacción que llevan a cabo las comunidades en el apartado 3.5.1.
- Nos preguntamos cómo es el papel que juegan estas grandes comunidades, a pesar de su evidente superioridad en tamaño. Analizamos su forma de interactuar en el apartado 3.5.2.

### 3.5.1 Volumen de interacción por comunidades

Queremos evaluar el protagonismo de las comunidades en cada momento. Para ello, representamos la proporción de la interacción total para las 13 mayores comunidades en un *stream graph* (Figura 3.9) . Debajo, la evolución de dos índices que ayudan a explicar la primera gráfica:

- La diversidad de la interacción entre los grupos considerados, mediante el índice $J$ definido en 2.3.2. De esta forma, tenemos un valor que mide de forma global si la aportación de las comunidades consideradas es más o menos equitativa. Cuando la diversidad es máxima, $J_{max} = 1$.
- La variación media de la actividad de las $N = 13$ comunidades consideradas:

$$V_t = \frac{\sum_i |v_{i,t} - v_{i,t-1}|}{N}$$

  siendo $v_{i,t}$ la proporción de actividad de la comunidad $i$ en el día $t$.

Los resultados muestran las siguientes características:

- En los dos primeros días (en los que el número de usuarios todavía es reducido), gran parte del volumen de la actividad recae en dos comunidades, cuyos nodos representativos provienen de procesos previos afines al 15M[3]. De ahí que el día 15 se produzca un pico en la variación del volumen de actividad, puesto que una de las comunidades reduce drásticamente su peso y la otra lo incrementa.
- Posteriormente, el peso de cada grupo comienza a estabilizarse (la variación disminuye) y a repartirse de forma más equitativa (valores de $J$ altos, próximos a $J_{max} = 1$).
- Destacan otros dos picos en la variación:
  - Día 17 y 18, cuando la comunidad más grande se consolida[4] y empieza a generar actividad.

[2] Los datos detallados se encuentran en el Anexo (Tabla B.4).
[3] Los usuarios @democraciareal y @bufetalmeida fueron muy activos en la campaña NoLesVotes, mencionada en 1.1.
[4] La cuenta @acampadasol nace el día 16 alrededor de las 4:00.

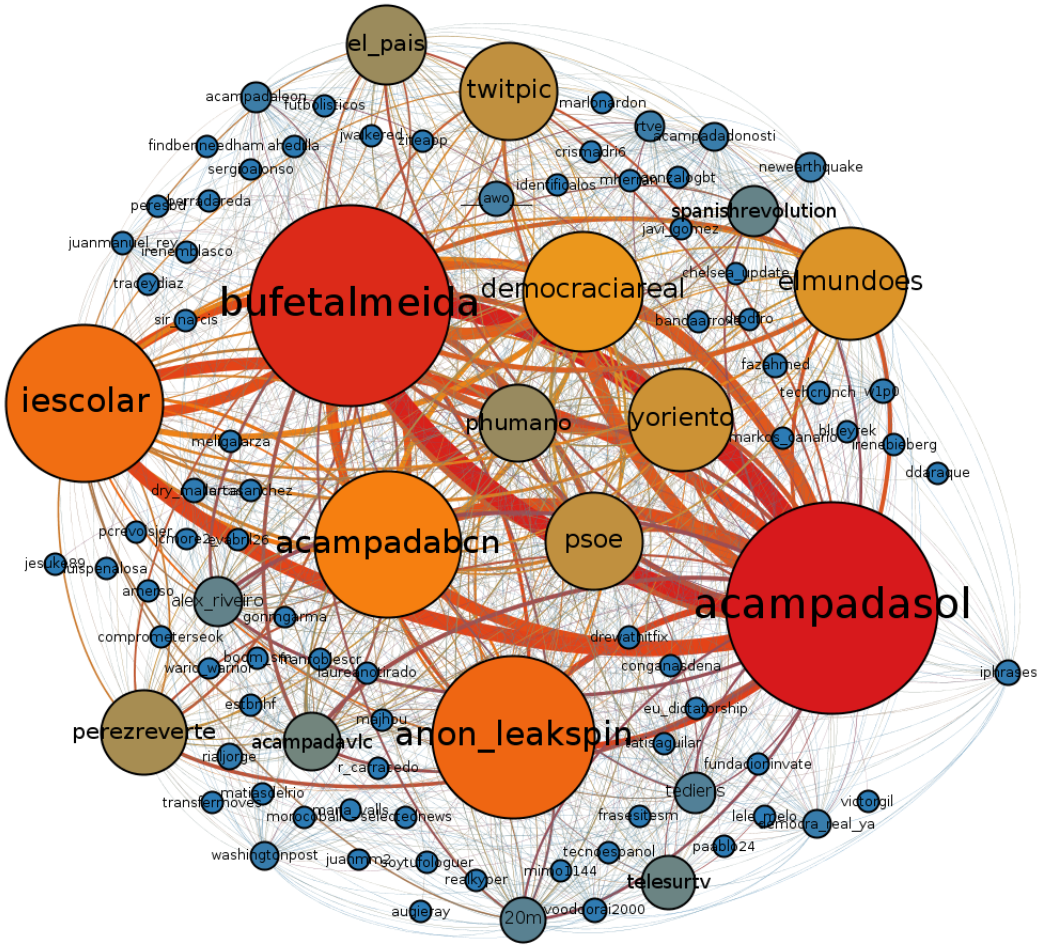



Figura 3.7: Red de comunidades. El nombre de cada grupo proviene del usuario con mayor grado. El tamaño de cada nodo es proporcional al número de miembros.

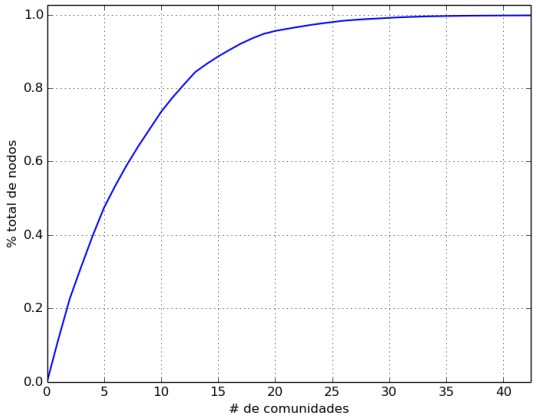



Figura 3.8: Tamaño relativo de las comunidades.

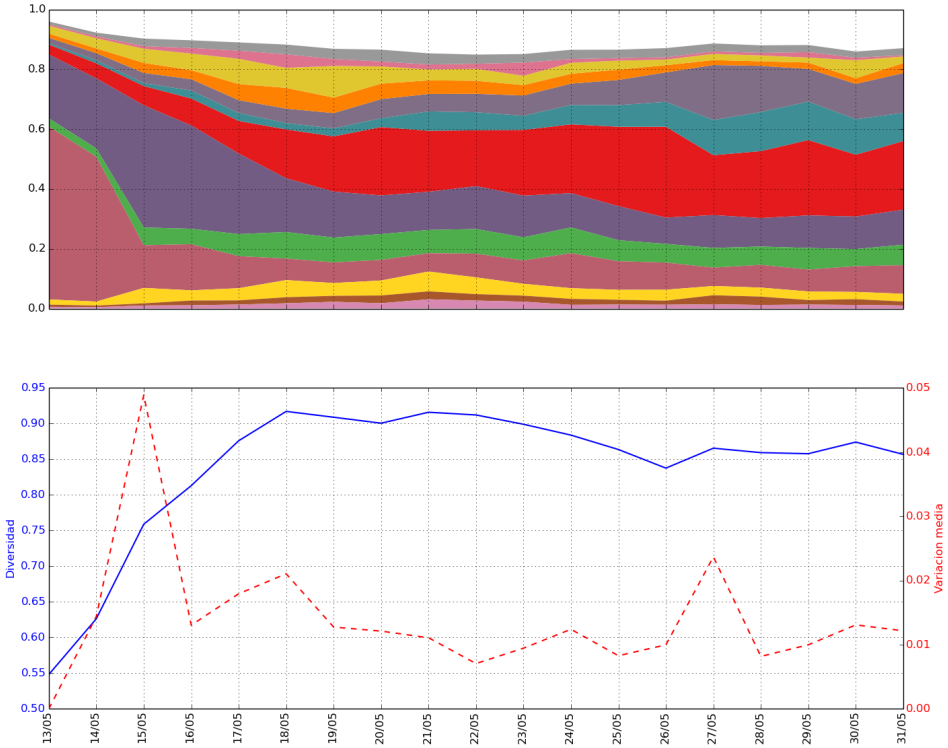



Figura 3.9: Actividad separada por comunidades.

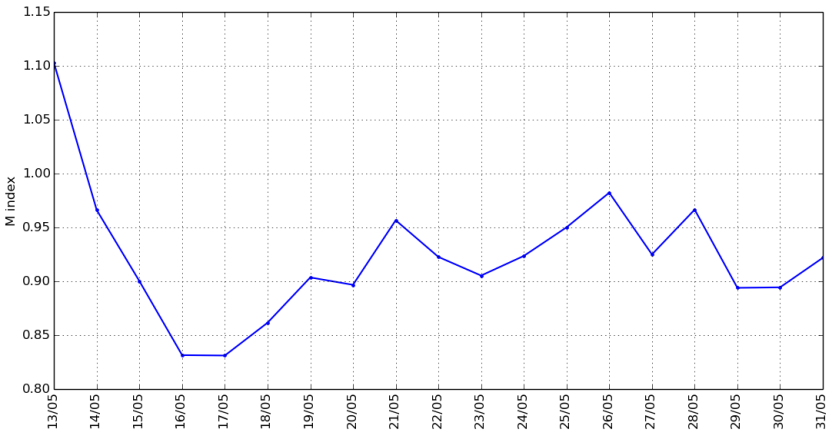



Figura 3.10: Índice $M$ global.

– Día 27: la comunidad centrada en 'acampadabcn' eleva su volumen de actividad notablemente. Corresponde con el día en el que fue desalojada la acampada de la ciudad de Barcelona.

### 3.5.2 El rol de las comunidades

Tras lo obtenido en el apartado anterior, pensamos en el papel que ha jugado cada comunidad en cada momento. Por ello, se proponen dos medidas:

- La relación entre el número de interacciones intra-comunidad e inter-comunidad:

$$M = m_{intra}/m_{inter}$$

  – $M > 1$ indica una presencia predominante de contactos dentro del grupo, mostrando un carácter más cerrado.

  – $M < 1$ caracteriza grupos de mayor apertura al resto de la red, no tan centrados en la actividad interna.

- La relación entre enlaces entrantes y salientes:

$$K_{IO} = k_{w,in}/k_{w,out}$$

  – $K_{IO} > 1$: la comunidad recibe más enlaces que los que salen de ella. En cierto modo, juega un papel de "referencia" mayor cuanto más se aleje de 1.

  – $K_{IO} < 1$: implica un rol más participativo. Sus usuarios realizan más contactos hacia otras comunidades que los que reciben.

Calculamos el índice $M$ para el total de la actividad por días (Figura 3.10). La actividad intracomunitaria destaca sobre la intercomunitaria ($M \uparrow$) en la primera etapa. Los usuarios son pocos y su interacción está más localizada dentro de sus comunidades. Posteriormente (días 16, 17 y 18), se detecta un periodo de apertura ($M \downarrow$) que coincide con incremento súbito de participantes. $M$ permanece en el intervalo [0.9, 1] hasta el final del proceso. A la vista de estos resultados, no se puede decir que haya

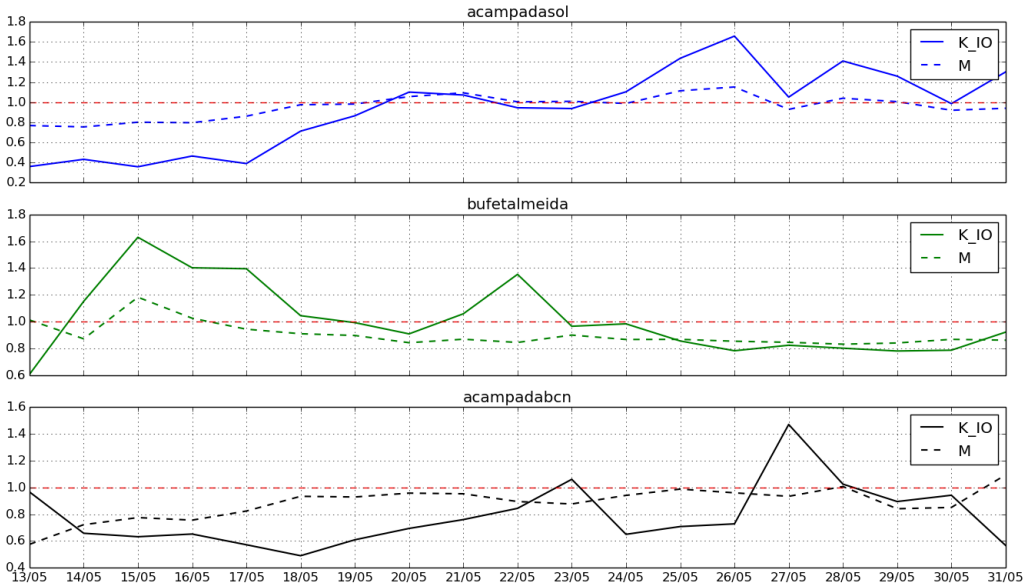



Figura 3.11: Evolución de $K_{IO}$ y $M$ en 3 comunidades.

momentos de actividad aislada por comunidades de forma generalizada. En gran parte del proceso, en concreto desde que alcanza cotas importantes de popularidad (día 18), las conexiones entre grupos superan las internas.

Para definir mejor el comportamiento de las redes, se ilustra la evolución de estas medidas en tres comunidades con diversas características (Figura 3.11). Los momentos en los que las conexiones entrantes superan a las salientes ($K_{IO} > 1$) se dan en puntos distintos para los tres ejemplos. Lo mismo sucede para $K_{IO} < 1$. Se puede decir que el papel de las grandes comunidades no es constante, sino que hay momentos en los que son una referencia en el proceso y en otros experimentan una apertura mayor en sus interacciones. El índice $M$ indica que estos grupos de interacción no son especialmente cerrados, sino que la actividad intra e inter-comunidad es bastante equilibrada. En el anexo B se muestran datos sobre el tamaño, los grados y los ratios comentados que describen la actividad de las comunidades más grandes ($n_i > 2000$).

## 3.6 Conclusiones

En términos generales, la estructura de la red presenta buenas propiedades en términos de conectividad, dando lugar a un proceso cohesionado en todo momento. La cercanía promedio entre los individuos es alta, algo importante tratándose de un sistema de este tamaño. Al mismo tiempo, el agrupamiento local y la estructura en comunidades identifica interacción a distintos niveles.

Encontramos ciertos aspectos relevantes que preceden la explosión de la actividad, como un aumento importante de la conectividad media y incremento destacado de propiedades de mundo pequeño. A pesar de identificar comunidades con un tamaño relativo muy grande, el papel que juegan no es fijo. El protagonismo se sucede a lo largo del proceso, y los distintos grupos pasan tanto por fases más participativas y abiertas como por otras en las que su actividad atrae una mayor actividad y son

consideradas una referencia.

La información obtenida de esta descripción topológica es interesante. Los resultados de todos los apartados muestran un periodo inicial de ajuste, en el que el proceso de comunicación va adquiriendo mayor protagonismo, aumentando el número de participantes, y ve como sus propiedades varían sustancialmente hasta que comienzan a ser más o menos estables. Esta estabilidad se mantiene independientemente de que el número de participantes se reduzca. Los usuarios de nueva incorporación se familiarizan con el proceso, presumiblemente porque se identifican con el movimiento, con cuentas colectivas, etc. y, en definitiva, porque comparten intereses comunes en la red y en el espacio físico.

# Capítulo 4

# Mecanismos de difusión

Una vez descrita la topología de una red, queremos centrar nuestra atención en los procesos que tienen lugar en ella. La propagación (también denominada difusión, transmisión, contagio) fue uno de los motivos originales por los que se empezó a estudiar las redes. La naturaleza del sistema especifica el tipo de propagación: transmisión de enfermedades, virus informáticos, rumores, etc. Constituye una extensa línea de investigación dentro de las redes complejas [27, 28].

En el presente capítulo se analizan aspectos de la difusión de información. Con lo estudiado hasta ahora e inspirándonos en otros trabajos, formulamos una hipótesis sobre el carácter evolutivo de estos mecanismos en nuestro caso de estudio. Tenemos como punto de partida el número de difusiones que se dan en el sistema real. Haremos uso de unos modelos que nos ayuden a entender porqué se alcanzan esos niveles de difusión y de qué forma se comporta este proceso.

## 4.1 Hipótesis: la difusión como proceso evolutivo

Un sistema puede experimentar procesos de contagio simultáneos que establezcan relaciones entre sí, determinadas por el medio en el que se encuentran. En este caso es importante estudiar las implicaciones que esto conlleva, y no tratar las propagaciones de manera independiente [29]. El origen de estas relaciones proviene a menudo de la presencia de recursos limitados en el medio, necesarios para que la propagación tenga éxito [30].

Las características descritas son similares a las percibidas en un ecosistema: un medio donde las especies persiguen la supervivencia a través de la reproducción. De ahí que la analogía con el ecosistema y el concepto de evolución darwiniana haya sido útil para entender el fenómeno de difusión. En concreto, fue propuesta como idea central del estudio de la transferencia de información cultural, en la denominada teoría memética [31].

Añadimos como ejemplo un estudio de la evolución de la tecnología a través de sus patentes [32]. En él, si una patente aparece como referencia en otra posterior se interpreta como una reproducción de la primera, ya que transmite parte de su naturaleza a un "descendiente". Una especie (patente) tiene mayor probabilidad de reproducirse cuanto mejor se adapte al medio. Esta capacidad de adaptación, denominada *fitness*, puede depender de multitutd de factores. Para las patentes, resultó ser muy importante para su prevalencia el hecho de abrir nuevos caminos tecnológicos (*door-opening innovations*).

Nuestro caso de estudio encaja dentro de estas características: en un ecosistema (una red social) conviven especies (mensajes) que si son capaces de adquirir recursos del medio (la atención de los usuarios y su voluntad de difundir ese contenido) se reproducirán. Por ello vemos adecuado formular la hipótesis de que la difusión de la información sigue un proceso evolutivo. Pensamos que:

- Los mensajes que van entrando en el proceso pueden propagarse simultáneamente y su capacidad de hacerlo se verá afectada por el resto de actividad.
- Existen recursos limitados en el medio. El tiempo y las características de cada usuario limitan la cantidad de información que pueden recibir y, más aún, qué parte de esta están dispuestos a difundir.
- Del mismo modo que las innovaciones tecnológicas, la aparición de mensajes innovadores, nuevos usuarios también puede implicar la apertura de nuevos nichos de interacción que atraigan la atención de la red.
- Hay ciertos indicios hallados en el capítulo 3 que encajan con este modelo. Por ejemplo, hemos visto en la sección 3.5 como puede haber distintas relaciones entre comunidades. Hay momentos en los que un grupo grande recibe mucha atención del exterior y otros en los que predominan enlaces salientes que salen de él. Pensamos que esto se puede percibir en la difusión: los usuarios que dediquen su atención a contenidos generados en su círculo pueden desviarla con facilidad a otros contenidos, internos o externos.
- A pesar de haber obtenido redes libres de escala (apartado 3.4), creemos que la conexión preferencial no es capaz de reproducir fielmente la naturaleza del proceso. Es cierto que puede jugar un papel relevante, pues un mensaje de amplio alcance multiplica sus posibilidades de repetición. Pero pensamos que la popularidad se verá continuamente influída por la aparición de nuevos individuos en el sistema.

La idea principal de este capítulo es construir un modelo con las características de un proceso evolutivo que nos ayude a confirmar esta hipótesis. Usaremos a modo de comparación otros dos modelos, cada uno con un objetivo distinto. Para ello, se analizarán las poblaciones reales y simuladas a través de las medidas propuestas en 2.3.

## 4.2 Modelos

Proponemos una serie de modelos que simulen procesos de propagación equivalentes a los reales. De acuerdo con la hipótesis evolutiva, el escenario se compone de una serie de especies que compiten por reproducirse y que, en caso de tener éxito, incrementan su población.

En consecuencia, trataremos cada *tweet* original $i$ como una especie susceptible de reproducirse a través de la difusión (*retweet*). Esas difusiones implican un aumento en su población asociada, $n_i$, cada vez que otro usuario difunda ese mensaje. En conjunto, habrá $R$ especies y la cantidad total de individuos será $N$.

| | |
|---|---|
| Especies | $i = 1, 2 \dots R$ |
| Poblaciones | $n = (n_1, n_2 \dots n_R)$ |
| Población total | $N = \sum_{i=1}^{R} n_i$ |

Tabla 4.1: Definición básica del sistema.

### 4.2.1 Características comunes: modelos sombra

La actividad en Twitter está fuertemente influida por las dinámicas propias de esta red social. Se ha demostrado que sigue ciertos patrones temporales de uso en función del momento del día [33].

Por otro lado, está la innegable influencia de factores externos. Se han realizado modelos dedicados exclusivamente a estudiar cómo repercute el exterior en la actividad interna de la red social [34].

El caso estudiado aquí está afectado por ambos factores. Mientras el primero es omnipresente en esta red social, la influencia externa es particular de cada proceso. Sirva como ejemplo el pico de actividad detectado en el capítulo 3, cuyo origen se localiza en un evento producido fuera de la propia red de comunicación[1].

Modelar estas influencias se escapa de nuestro objetivo. Necesitamos que los modelos no determinen por sí mismos los tiempos en los que nacen y se reproducen las especies, sino que lo hagan a modo de reflejo del caso real. Así, un periodo de alta actividad real, sea cual sea su origen, implicará alta actividad también en las simulaciones.

Planteamos un tipo de modelo que incorpore los datos temporales de la realidad. Se aplica un sistema de difusión "sombra" [32] que, por construcción, presenta las siguientes características de los sistemas reales:

- El momento del nacimiento real de una especie $i$, $t_i^{(n)}$, corresponde con un nacimiento de otra especie $j$ en el modelo. Hay un nacimiento sombra en el modelo por cada uno real. De esta forma, el número de mensajes $R(t)$ disponibles para la difusión en los modelos es el mismo que en la situación real.

- La reproducción real también implica una reproducción sombra. La unión de los tiempos de difusión reales $T_i = [t_{i,1}, t_{i,2}, \ldots, t_{i,n_i}]\ \forall i$ marcan el tiempo de reproducción en los modelos, de modo que el tamaño total $N(t)$ sea el mismo.

- Los modelos tomarán como *input* ambos conjuntos de datos temporales y la elección de la especie que se reproduce dependerá de las características de cada uno.

Los resultados de las simulaciones de cada modelo con estas características comunes nos servirá para comparar y decidir sobre la validez de éstos.

### 4.2.2 Modelo aleatorio

Utilizaremos en primer lugar un modelo que escoja aleatoriamente la especie que se va a reproducir. El motivo de este modelo "nulo" es establecer un punto de partida para la comparación con los datos reales y evidenciar, llegado el caso, la distancia entre estos y un comportamiento aleatorio.

La probabilidad de que el mensaje $i$ se difunda es

$$p_i = p = \frac{1}{R}$$

$$i = 1 \ldots R$$

### 4.2.3 Preferential attachment

En el apartado 3.4 se ha probado como estos sistemas muestran las características de las redes libres de escala. Como se explica en 2.2.3, uno de los mecanismos típicos que proporcionan redes libres de escala es la conexión preferencial (*rich-gets-richer*). Por tanto, se puede pensar que un comportamiento de este estilo guíe la transmisión de la información. La presencia de comunidades grandes, cuyos usuarios más conectados generan mucha actividad en su entorno, es otra de las razones para pensar en este mecanismo.

[1] El incremento mencionado corresponde al día 27 (Figura 3.2), desencadenado por el desalojo de Plaza Catalunya la mañana de ese mismo día.

Utilizaremos un modelo determinado por la conexión preferencial que nos permita evaluar si es un mecanismo válido o no. La probabilidad de reproducción de una especie es proporcional a su población:

$$p_i = \frac{n_i}{N}$$

con $N = \sum_i n_i$ el número total de individuos.

### 4.2.4 Modelo evolutivo

Se propone un modelo simple de naturaleza evolutiva. Las características principales son:

- Aparición de una especie: en el momento en el que se reproduce por primera vez el mensaje $i$ se inicializa su población, $n_i = 1$, y se le asigna un valor de fitness $f_i$ aleatorio.
- Modificación del entorno: la inclusión de una nueva especie supone la modificación de dos especies existentes. Se seleccionan dos poblaciones $j$, $k \neq i$ y se altera aleatoriamente su fitness.
- Selección: la probabilidad de reproducción es proporcional al fitness y a la población:

$$p_i \propto f_i n_i$$

- Muerte: se establece un número máximo de inviduos que pueden coexistir. Si con un nuevo nacimiento o reproducción se alcanza ese límite, el ecosistema se satura y un individuo desaparece por falta de recursos[2]. La probabilidad de muerte de la especie $i$ es función de su población y del fitness:

$$p_{\text{muerte},i} \propto n_i \, (1 - f_i)$$

## 4.3 Distribución de poblaciones

En este apartado, queremos analizar si las poblaciones de las especies al final de cada día se distribuyen de manera similar en el caso real y en las simulaciones. De esta forma podemos tener las primeras evidencias sobre lo adecuado de los modelos.

### 4.3.1 Poblaciones reales y simuladas

Al visualizar las poblaciones (mismo procedimiento que en 3.4) se intuye una ley de potencias (ejemplo en la Figura 4.1). Hay gran cantidad de especies que se reproducen muy poco, mientras que en la cola se encuentran unas pocas especies que han logrado acumular un número elevado de individuos.

En la tabla 4.2 se muestran los exponentes obtenidos en los ajustes a distribuciones power-law de los datos reales, de una simulación del preferential attachment y de una del evolutivo. El modelo aleatorio queda descartado porque no se puede ajustar de forma correcta[3]. Lo importante de esta sección es comprobar que ambos modelos dan el tipo adecuado de distribución, puesto que los tres sistemas tienen ajustes válidos en todo momento. Aún así, el preferential attachment presenta con frecuencia exponentes más cercanos al real que el evolutivo.

Hay que recalcar que, a pesar de que incluimos una única realización de los modelos, se ha probado previamente que éstos proporcionan valores con varianzas reducidas (Anexo C). Decidimos, para mayor claridad, considerar únicamente una simulación que tenga unos valores próximos a su media correspondiente.

[2] El efecto de la muerte tiene su efecto en $n_i$, y por tanto afecta a la probabilidad de selección y de muerte, pero no se ve reflejado a la hora de contar las poblaciones.

[3] Se incluirán sus resultados igualmente en el análisis más exhaustivo de las poblaciones (apartado 4.4)

| | Datos reales | | Pref. attachment | | Evolutivo | |
|---|---|---|---|---|---|---|
| Día | $\alpha$ | $x_{min}$ | $\alpha$ | $x_{min}$ | $\alpha$ | $x_{min}$ |
| 13/05 | 2.31 | 2 | 2.92 | 6 | **2.02** | 2 |
| 14/05 | 2.48 | 4 | **2.59** | 5 | 1.96 | 2 |
| 15/05 | 2.19 | 8 | 2.50 | 9 | **1.94** | 2 |
| 16/05 | 2.42 | 5 | **2.32** | 5 | 1.99 | 2 |
| 17/05 | 2.25 | 3 | **2.01** | 2 | 1.94 | 2 |
| 18/05 | 2.22 | 3 | **2.18** | 3 | 1.94 | 2 |
| 19/05 | 2.21 | 7 | **2.23** | 3 | 2.00 | 2 |
| 20/05 | 2.24 | 3 | **2.10** | 2 | 1.97 | 2 |
| 21/05 | 2.25 | 6 | 2.56 | 8 | **2.00** | 2 |
| 22/05 | 2.26 | 3 | **2.09** | 2 | 1.98 | 2 |
| 23/05 | 2.09 | 12 | 2.23 | 3 | **2.02** | 3 |
| 24/05 | 2.26 | 4 | **2.35** | 3 | 2.05 | 2 |
| 25/05 | 2.26 | 7 | **2.46** | 4 | **2.05** | 2 |
| 26/05 | 2.22 | 5 | 2.33 | 3 | **2.13** | 3 |
| 27/05 | 2.19 | 6 | **2.09** | 5 | 1.93 | 2 |
| 28/05 | 2.32 | 4 | **2.24** | 3 | 1.99 | 2 |
| 29/05 | 2.17 | 3 | 2.60 | 11 | **1.93** | 2 |
| 30/05 | 2.21 | 5 | **2.37** | 5 | 1.97 | 2 |
| 31/05 | 2.32 | 3 | **2.33** | 3 | 2.05 | 2 |

Tabla 4.2: Ajustes power-law: poblaciones reales.

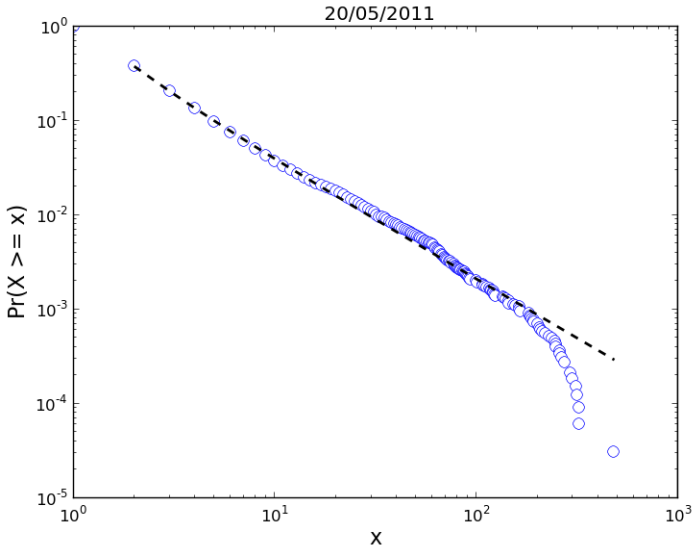



Figura 4.1: Distribución de poblaciones. Ejemplo: día 20.

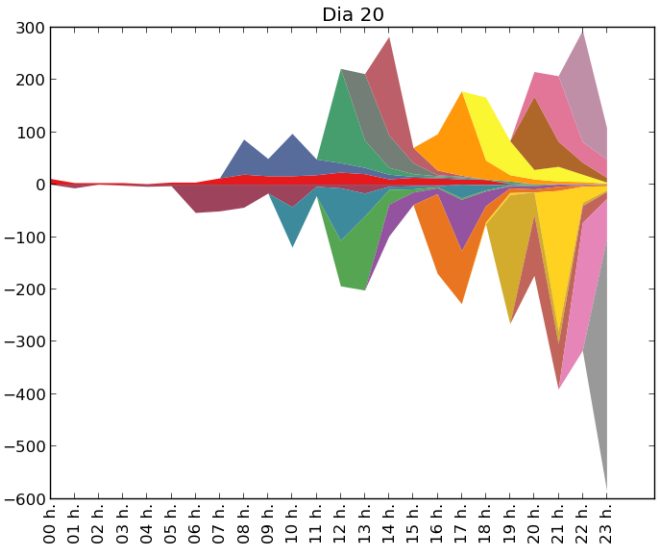



Figura 4.2: Reproducciones top20 RT (día 20)

### 4.3.2 Top 20

Nos centramos en el comportamiento de los 20 mensajes con más reproducciones, agrupado por horas. Esta actividad se puede visualizar en forma de *stream graph* (Figura 4.2). Se puede ver cómo crecen durante periodos de tiempo relativamente cortos, tras los cuales empiezan a disminuir el ratio de reproducción y son otros individuos los que se abren camino.

Se intuye que las aparaciones de nuevas especies influyen en el desarrollo de las demás. En este sentido, el modelo evolutivo podría ser una elección más acertada que el preferential attachment, ya que incluye la modificación del entorno. Esta característica tiene el potencial de modelar la generación de nuevos contenidos que atraigan súbitamente la atención de los usuarios.

## 4.4 Análisis de la diversidad

En el apartado anterior no se llega a ninguna conclusión sobre los modelos. Es necesario profundizar más en la descripción de las poblaciones para validar de forma clara un modelo u otro. Para ello se hace uso de las medidas de diversidad presentadas en 2.3. Los sistemas se caracterizan mediante unos índices que reflejan el número de especies y, al mismo tiempo, si el reparto de las poblaciones es equitativo. Para facilitar el análisis, se utilizan las versiones normalizadas por el valor máximo de ambos índices. De esta forma, se pueden comparar las diversidades de sistemas con tamaño distinto (los diferentes días considerados) y las diferencias entre las dos medidas empleadas.

Utilizaremos la distancia entre los índices del sistema real y los simulados para validar o rechazar los modelos. Se calcula de tres formas distintas:

- Comparando los índices cada jornada. Al final de cada día se tiene una distribución de poblaciones con toda la actividad acumulada en ese periodo. Se estudia en el apartado 4.4.1.
- Dentro de cada día, la diversidad va evolucionando hasta que llega a los valores finales, que son los calculados en el punto anterior. Así se puede identificar el comportamiento de los modelos de una forma más precisa (apartado 4.4.2).

- Por último, establecemos una ventana temporal que va recogiendo las poblaciones acumuladas dentro de su dominio. Poco a poco, se va desplazando la ventana, produciendo una señal por índice. Analizaremos la correlación de estas señales correspondientes al sistema real y a las simulaciones (apartado 4.4.3).

### 4.4.1 Diversidad total por día

El primer paso consiste en calcular los índices de diversidad totales para cada día en sistema real y simulaciones. Se representa la evolución de cada índice por separado (Figura 4.3). Se aprecian las distintas formas de evaluar la diversidad según el índice: mientras el valor más alto del índice $D$ ronda el 10% de su valor máximo (alcanzado por el modelo aleatorio), el índice $J$ alcanza el 96%. La mayor ponderación que otorga el índice de Simpson $D$ a las especies más numerosas tiene un efecto especialmente importante en distribuciones de ley de potencias. Mientras, el índice $J$ incorpora las contribuciones de muchas especies de poblaciones muy pequeñas, dando lugar a diversidades mucho más altas. En cualquier caso, comentamos los resultados centrándonos en el comportamiento de los modelos:

- El modelo aleatorio se mantiene alejado siempre de los valores objetivo, especialmente en el índice $J$, donde alcanza valores muy próximos a la diversidad máxima. Apenas se aprecian cambios relevantes cuando la variación de la diversidad real es mayor.

- El modelo de preferential attachment presenta diversidades más cercanas a las reales. Según el índice $D$, tiene la diversidad más cercana a la real en dos ocasiones (días 16/05 y 21/05). El índice $J$ coincide con el primero de los resultados, y además lo sitúa como más cercano en una tercera ocasión, el día 14/05.

- En la mayoría de casos, es el evolutivo el que mejor reproduce las condiciones de diversidad diaria.

Las medidas comparadas aquí no nos indican si los modelos se comportan de forma similar al sistema real. Para determinarlo, investigamos lo que sucede dentro de cada periodo en el que dividimos el proceso global.

### 4.4.2 Evolución de la diversidad acumulada

Queremos explorar la evolución que sufre la diversidad dentro de cada día y qué tipo de respuesta ofrecen los modelos. En primer lugar, medimos la diversidad total cada hora, teniendo en cuenta toda la actividad anterior. En la Tabla B.5 del Anexo se muestran los errores cuadráticos medios obtenidos entre la diversidad real y los tres modelos.

Observamos un ejemplo en la Figura 4.4. El índice $D$ expone comportamientos similares en la forma al real para los tres modelos, aunque el aleatorio aparece claramente distanciado. El preferential attachment se acerca cada vez más al real, mientras que el evolutivo apenas comete errores.

El índice $J$ resulta ser más exigente, pues muestra diferencias de mayor tamaño. El modelo evolutivo responde bien a las variaciones experimentadas por el sistema real, a lo que la conexión preferencial sólo reacciona de forma leve.

### 4.4.3 Diversidad: ventana móvil

El índice $J$ ha resultado ser el más útil a la hora de resaltar la precisión de los modelos. Por ello, estudiamos qué sucede cuando calculamos este índice a las poblaciones obtenidas dentro de una ventana móvil de longitud fija (3 horas), desplazándola cada 300 segundos. En la Figura 4.5 vemos dos ejemplos,

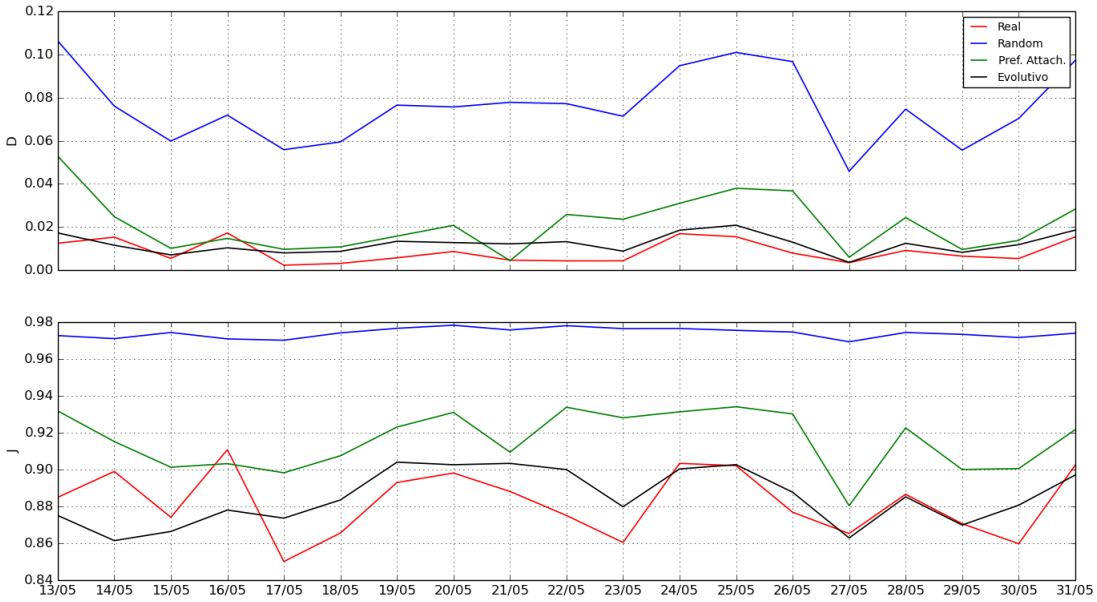



Figura 4.3: Índices de diversidad.

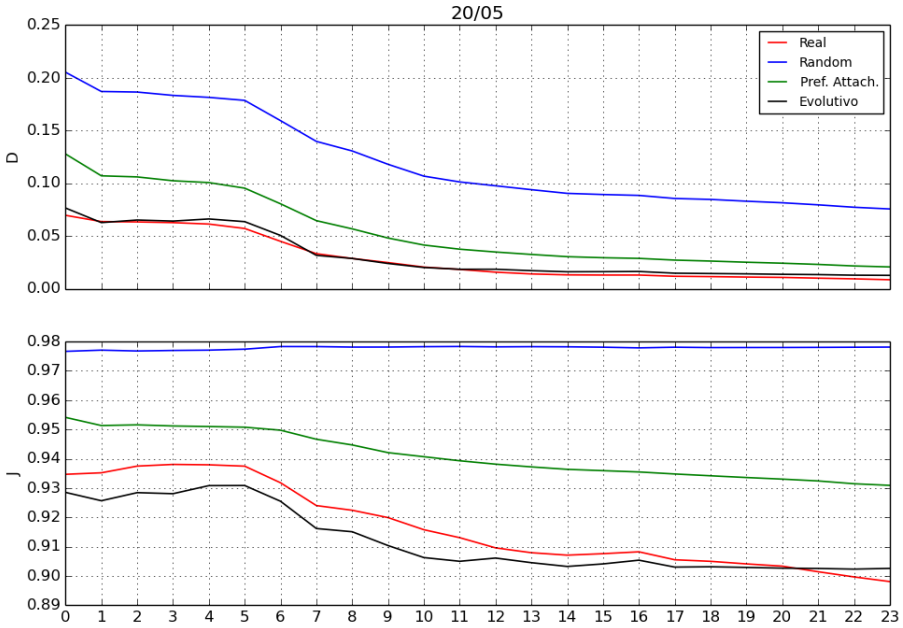



Figura 4.4: Evolución de la diversidad acumulada

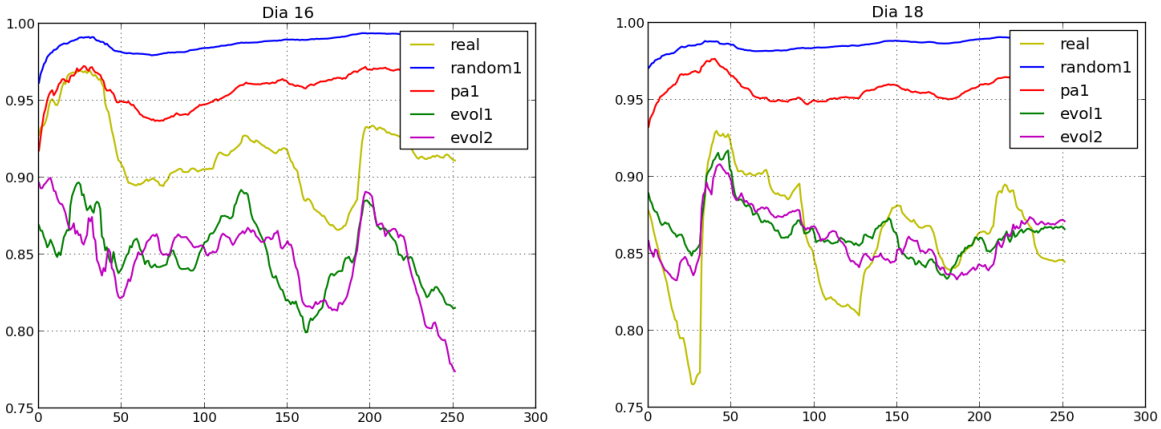



Figura 4.5: J en ventana de 3 horas.

que contienen una ejecución del modelo aleatorio y del preferential attachment y dos del evolutivo. En el primero de ellos, el resultado obtenido en el apartado 4.4.1 nos decía que el preferential attachment tenía una diversidad más cercana a la real. Se aprecia cómo el modelo evolutivo es capaz de reaccionar del mismo modo que los datos reales en ambos casos, aunque en el primero se mantiene distanciado del real. La conexión preferencial, por el contrario, no reproduce ni siquiera el cambio que se produce en $x = [150, 200]$.

Para obtener una medida de la proximidad entre modelo y datos reales calculamos la correlación entre las señales. Obtendremos un coeficiente de correlación para cada par de señales real-simulación y un conjunto de estos valores para cada uno de los 19 días. En la Figura 4.6 vemos cómo la correlación con los modelos evolutivos es más fuerte y menos variable que para el preferential attachment, y como el modelo aleatorio presenta de forma promedio una correlación casi nula.

## 4.5 Conclusiones

A pesar de que el modelo evolutivo y la conexión preferencial dan distribuciones de poblaciones hasta cierto punto similares a las reales, el análisis más exhaustivo a través de la diversidad muestra diferencias relevantes entre uno y otro. El modelo evolutivo es capaz de reaccionar a cambios tanto en las diversidades acumuladas como en el intervalo móvil, y proporciona al final de los periodos considerados valores más cercanos a los reales.

En este capítulo mostramos como la hipótesis evolutiva del comportamiento en los fenómenos de propagación es acertada. Las distintas unidades de información (tweets) entran en un escenario que se ve continuamente afectado por la aparición de nuevos contenidos. Su capacidad para reproducirse dependerá en cada momento del número de recursos disponibles en el medio.

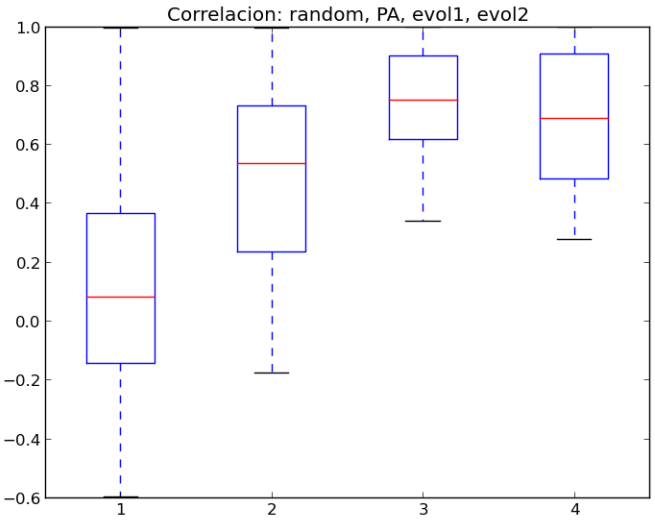



Figura 4.6: Coeficientes de correlación. Valores medios para cada modelo.

# Capítulo 5

# Conclusiones

Hemos presentado un análisis de un proceso de comunicación a gran escala asociado con un movimiento social. El estudio de la interacción se afronta con la teoría de grafos como herramienta principal, y revela a través de la topología propiedades a las que sería complicado acceder por otro camino. Los resultados muestran cómo la participación sigue una cierta estructura a distintos niveles y presenta características típicas de las redes complejas. A pesar del tamaño del sistema, las capacidades estructurales de las redes en lo que respecta a eficiencia son buenas. También es interesante cómo, a pesar de intuir un cierto nivel de jerarquía en las comunidades, el papel que juegan cambia constantemente. Es un proceso abierto que permite el cambio de protagonismo y de núcleos de generación de contenidos.

Por otro lado, hemos logrado identificar la forma en que se transmite la información. El análisis de la propagación presenta indicios evidentes de comportamiento evolutivo, tal y como sugería nuestra hipótesis. La atención de los usuarios que toman parte en el proceso es limitada y cambiante, y eso se traduce en un ecosistema que ofrece oportunidades de desarrollo a individuos nuevos. Por ello, el crecimiento total está restringido y su extensión en el tiempo es bastante reducida.

El caso de estudio proviene de un proceso complejo, que consta de muchas capas de interacción. Mediante el análisis realizado en una de ellas, y en concreto, de su estructura y de los procesos dinámicos que sustenta, obtenemos una visión completa del sistema y aplicable a lo que sucede en las demás capas. Se perciben constantes relaciones entre los hechos acontecidos y los datos observados, algo que favorece una aplicación directa del estudio.

El trabajo abre algunas líneas de investigación futuras que permitirían profundizar en este u otro proceso de comunicación. Entre ellas, las más interesantes se derivan del estudio de la transmisión de información. Es posible crear modelos que incluyan aspectos no recogidos aquí y que reflejen de una manera más eficiente y completa el comportamiento real. Se puede pensar en asignar el *fitness* de una manera más precisa teniendo en cuenta qué posición tiene en la red el nodo emisor o qué características tiene la información que genera. También es atractiva la idea de probar estos modelos sobre redes con estructuras conocidas y estudiar su relación con las redes reales.

# Referencias


[1] Toret, J., Calleja, A., Marín, O., Aragón,. P., Aguilera, M., Lumbreras, A. (2013). Tecnopolítica: la potencia de las multitudes conectadas. El sistema red 15M, un nuevo paradigma de la política distribuida. IN3 Working Paper Series - Universitat Oberta de Catalunya ISSN 2013-8644

[2] GlobaliseThis (2009). New media for peace, http://globalisethis.wordpress.com/2009/10/06/1999-new-media-revolution-in-seattle/

[3] Salido, N. (2006). Del 11M al 14M: Jornadas de móvil-ización social, ISBN 84-313-2374-4, págs. 271-284.

[4] Rey, P., Front page newspaper analysis, http://numeroteca.org/cat/frontpage-newspaper/ (2011-2012).

[5] Bar-Yam, Y. (2003). Dynamics of complex systems, The Advanced Book Studies in Nonlinearity series, Westview Press ISBN 0813341213.

[6] Rosas-Casals, M., Valverde, S., Solé, R. (2006). Topological Vulnerability of the European Power Grid under Errors and Attacks. Santa Fe Institute working paper.

[7] Amaral, L., Scala, A. y Barthélémy, M. (2000). Classes of small-world networks. PNAS October 10, 2000 vol. 97 no. 21 11149-11152.

[8] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A.-L. (2000). The large-scale organization of a metabolic networks, Nature 407, 651–654.

[9] Montoya, J. M. and Solé, R. V. (2002). Small world patterns in food webs, J. Theor. Bio. 214, 405–412.

[10] Price, D. J. de S., (1965). Networks of scientific papers, Science 149, 510–515.

[11] Guimerà, R., Danon, L., Díaz-Guilera, A., Giralt, F., and Arenas, A. (2002). Self-similar community structure in organisations, Physical Review E 68, 065103 (2003).

[12] Zachary, W. (1977) An information flow model for conflict and fission in small groups. Journal of antropological research, vol. 33

[13] Newman, M. (2001). Scientific collaboration networks I: Network construction and fundamental results. Physical Review E , 64(1):016131.

[14] Watts, D.J. y Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. Nature, vol. 393, June 1998, pp. 440–442.

[15] Albert, R., Jeong, H. y Barabási, A. (1999). Diameter of the World Wide Web. Nature, vol. 401, September 1999, pp. 130-131.

[16] Danon, L., House, T. a., Read, J. M. and Keeling, M. J. (2012). Social encounter networks: collective properties and disease transmission. Journal of the Royal Society Interface 9 , 2826.

[17] Barabási, A. y Albert, R. (1999) . Emergence of scaling in random networks. Science, vol. 286, pp. 509-512.

[18] Kunegis, J. (2012) The Power Law Paradox https://networkscience.wordpress.com/2012/04/19/

[19] Clauset, A., Shalizi, C. y Newman, M. (2009). Power-law distributions in empirical data. arXiv:0706.1062 [physics.data-an]

[20] Girvan, M. y Newman, M.E.J. (2002). Community structure in biological and social networks. Proc. Natl. Acad. Sci. USA 99 (12), 7821.

[21] Lancichinetti, A. y Fortunato, S. (2009). Community detection algorithms: a comparative analysis. Physical Review E 80, 056117

[22] Rosvall, M. y Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. USA 105, 1118

[23] Ronhovde, P. y Nussinov, Z. (2009). Multiresolution community detection for megascale networks by information-based replica correlations. Phys. Rev. E 80, 016109

[24] Newman, M., Clauset, Moore. Finding community structure in very large networks, arXiv:cond-mat/0408187 [cond-mat.stat-mech]

[25] Blondel, V., Guillaume, J.L., Lambiotte, R. y Lefebvre, E. (2008). Fast unfolding of communities in large networks. arXiv:0803.0476 [physics.soc-ph]

[26] Peña-López, I., Congosto, M., & Aragón, P. (2013). Spanish Indignados and the evolution of 15M: towards networked para-institutions. Big Data: Challenges and Opportunities, 359-386.

[27] Bearman, P. S., Moody, J., y Stovel, K., (2002). Chains of affection: The structure of adolescent romantic and sexual networks. Preprint, Department of Sociology, Columbia University .

[28] Pastor-Satorras, R. y Vespignani, A. (2001) Epidemic spreading in scale-free networks. Phys. Rev. Lett. 86, 3200

[29] Myers, S. y Leskovec, J. (2012). Clash of the Contagions: Cooperation and Competition in Information Diffusion. IEEE International Conference On Data Mining (ICDM).

[30] Weng, L., Flammini, A., Vespignani, A. y Wenczer, F. (2012). Competition among memes in a world of limited attention. Nature Scientific Reports 2, 335

[31] Dawkins, R. (1990). The Selfish Gene, Oxford University Press, ISBN 0-19-286092-5 (Chapter 11)

[32] Buchanan, A., Packard, N. y Bedau, M., (2010). Darwinian evolution of culture as reflected in patent records. ALIFE 2010: 831-837

[33] Raghavan, V., Ver Steeg, G., Galstyan, A. y Tartakovsky, A. (2013). Modeling Temporal Activity Patterns in Dynamic Social Networks. arXiv:1305.1980

[34] Myers, S., Zhu, C., Leskovec, J. (2012) Information Diffusion and External Influence in Networks. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2012.

# Parte II

# Anexos

# Anexo A

# Software

Este es el listado de todos los paquetes de Python que se han utilizado en el trabajo.

- networkx: creado para la generación, manipulación y el estudio de redes complejas. Se ha utilizado en primer lugar para generar los grafos a partir de la base de datos. Posteriormente, para el cálculo de propiedades topológicas y para la generación de grafos aleatorios con los que comparar los reales.

  https://networkx.github.io/

- community: es un módulo basado en NetworkX utilizado para la detección de comunidades. Usa el método de Louvain [25].

  http://perso.crans.org/aynaud/communities/

- powerlaw: un paquete de herramientas estadísticas para el manejo de distribuciones de ley de potencias con modelos de ajuste, tests de bondad de ajuste, visualización y validación.

  https://pypi.python.org/pypi/powerlaw

- numpy: paquete fundamental para la computación científica en Python.

  http://www.numpy.org/

- matplotlib.pyplot: herramientas para la respresentación de gráficas al estilo de Matlab.

  http://matplotlib.org/api/pyplot_api.html

- collections: un paquete de ayuda para el manejo de diccionarios de Python.

  https://docs.python.org/2/library/collections.html#module-collections

- stacked_graph: librería para realizar stream plots

  http://code.activestate.com/recipes/576633-stacked-graphs-using-matplotlib/

# Anexo B

# Tablas de datos

## B.1 Análisis de la topología

En esta sección se recogen los datos de todas las medidas realizadas en el Capítulo 3.

- En la Tabla B.1 se recogen, por días, los datos relativos al nḿero de nodos de la red global y cuántos de ellos pertenecen al componente gigante.
- La Tabla B.2 recoge las propiedades de mundo pequeño de cada sistema. Los valores $C_{random}$ y $L_{random}$ son valores medios de $C$ y $L$ que presentan 100 redes aleatorias equivalentes al sistema real.
- En la Tabla B.3 se listan los datos del ajuste a modelos de ley de potencias para las distribuciones de grado. Se incluyen los valores de los exponentes y el valor de $x$ mínimo desde el cual se realiza el mejor ajuste posible.
- Por último, la Tabla B.4 incluye los índices que describen de forma general la interacción de las 20 comunidades más grandes.

| Día | Nodos | Nodos en CG | % Conexión | Día | Nodos | Nodos en CG | % Conexión |
|---|---|---|---|---|---|---|---|
| 13/05/11 | 2071 | 1377 | 0.664896 | 23/05/11 | 25796 | 22900 | 0.887735 |
| 14/05/11 | 3356 | 2797 | 0.833433 | 24/05/11 | 17506 | 15037 | 0.858963 |
| 15/05/11 | 11970 | 11095 | 0.926901 | 25/05/11 | 14911 | 12765 | 0.856079 |
| 16/05/11 | 11055 | 9861 | 0.891995 | 26/05/11 | 12604 | 10649 | 0.844891 |
| 17/05/11 | 20430 | 18723 | 0.916446 | 27/05/11 | 35480 | 33159 | 0.934583 |
| 18/05/11 | 36546 | 33721 | 0.922700 | 28/05/11 | 16467 | 14400 | 0.874476 |
| 19/05/11 | 52427 | 49216 | 0.938753 | 29/05/11 | 15697 | 14220 | 0.905906 |
| 20/05/11 | 52230 | 48898 | 0.936205 | 30/05/11 | 12959 | 11024 | 0.850683 |
| 21/05/11 | 51015 | 47239 | 0.925983 | 31/05/11 | 10098 | 8260 | 0.817984 |
| 22/05/11 | 40493 | 36155 | 0.892870 | | | | |

Tabla B.1: Tamaño del componente gigante

| Día | $\sigma$ | $C$ | $C_{random}$ | $C/C_{random}$ | $L$ | $L_{random}$ | $L/L_{random}$ |
|---|---|---|---|---|---|---|---|
| 13/05/11 | **207.8947** | 0.2178 | 2.05E-03 | 106.0873 | 4.3883 | 8.5995 | 0.5103 |
| 14/05/11 | **239.3949** | 0.1681 | 1.44E-03 | 116.7638 | 4.3771 | 8.9741 | 0.4877 |
| 15/05/11 | **402.4974** | 0.1796 | 6.52E-04 | 275.7004 | 5.4313 | 7.9292 | 0.6850 |
| 16/05/11 | **419.5724** | 0.1640 | 6.54E-04 | 250.8019 | 6.0424 | 10.1085 | 0.5978 |
| 17/05/11 | **1659.4422** | 0.1309 | 1.32E-04 | 989.1556 | 7.4841 | 12.5556 | 0.5961 |
| 18/05/11 | **2282.0021** | 0.1441 | 1.01E-04 | 1420.0197 | 7.4216 | 11.9267 | 0.6223 |
| 19/05/11 | **2519.5050** | 0.1585 | 9.26E-05 | 1712.8660 | 7.1730 | 10.5509 | 0.6798 |
| 20/05/11 | **2301.0859** | 0.1918 | 1.31E-04 | 1469.5587 | 6.5325 | 10.2335 | 0.6386 |
| 21/05/11 | **2834.2016** | 0.1618 | 9.08E-05 | 1783.1936 | 7.2824 | 11.5747 | 0.6292 |
| 22/05/11 | **1524.9384** | 0.1571 | 1.59E-04 | 987.3332 | 7.7467 | 11.9649 | 0.6475 |
| 23/05/11 | **1686.3554** | 0.1641 | 1.45E-04 | 1134.6066 | 8.7866 | 13.0595 | 0.6728 |
| 24/05/11 | **855.516** | 0.1946 | 3.82E-04 | 509.3862 | 7.1035 | 11.9304 | 0.5954 |
| 25/05/11 | **1624.0139** | 0.1895 | 1.92E-04 | 987.4066 | 7.4247 | 12.2117 | 0.6080 |
| 26/05/11 | **988.1884** | 0.2141 | 3.63E-04 | 590.5037 | 6.9816 | 11.6834 | 0.5976 |
| 27/05/11 | **1966.8862** | 0.2056 | 1.75E-04 | 1177.3528 | 6.6388 | 11.0908 | 0.5986 |
| 28/05/11 | **1384.6093** | 0.1864 | 2.57E-04 | 725.9667 | 6.5452 | 12.4835 | 0.5243 |
| 29/05/11 | **991.4701** | 0.1852 | 2.71E-04 | 682.4866 | 7.6394 | 11.0980 | 0.6884 |
| 30/05/11 | **651.8203** | 0.1941 | 3.69E-04 | 526.5096 | 9.4626 | 11.7147 | 0.8078 |
| 31/05/11 | **532.8078** | 0.1788 | 4.73E-04 | 377.7278 | 8.9240 | 12.5878 | 0.7089 |

Tabla B.2: Coeficientes de clustering y distancias promedio.

| | Grado entrada | | Entrada (pesado) | | Grado salida | | Salida (pesado) | |
|---|---|---|---|---|---|---|---|---|
| **Dia** | $\alpha$ | $x_{min}$ | $\alpha$ | $x_{min}$ | $\alpha$ | $x_{min}$ | $\alpha$ | $x_{min}$ |
| **13** | 2.391 | 5 | 2.239 | 5 | 2.212 | 3 | 2.083 | 2 |
| **14** | 2.188 | 6 | 2.085 | 6 | 2.154 | 2 | 2.082 | 2 |
| **15** | 1.998 | 6 | 1.97 | 6 | 2.602 | 7 | 2.498 | 7 |
| **16** | 2.041 | 5 | 2.013 | 5 | 2.535 | 5 | 2.404 | 4 |
| **17** | 2.148 | 6 | 2.122 | 6 | 2.624 | 4 | 2.518 | 4 |
| **18** | 2.046 | 4 | 2.03 | 4 | 3.168 | 13 | 2.648 | 6 |
| **19** | 2.069 | 4 | 2.08 | 6 | 2.559 | 4 | 2.425 | 4 |
| **20** | 2.056 | 7 | 2.032 | 7 | 2.817 | 10 | 2.64 | 12 |
| **21** | 2.063 | 5 | 2.038 | 5 | 2.896 | 11 | 2.545 | 7 |
| **22.** | 2.109 | 4 | 2.071 | 4 | 2.825 | 8 | 2.569 | 8 |
| **23** | 2.141 | 4 | 2.128 | 5 | 2.748 | 5 | 2.709 | 9 |
| **24** | 2.105 | 4 | 2.013 | 3 | 2.845 | 5 | 2.555 | 5 |
| **25** | 2.161 | 4 | 2.134 | 5 | 2.721 | 4 | 2.455 | 4 |
| **26** | 2.087 | 3 | 2.081 | 4 | 2.948 | 7 | 2.776 | 10 |
| **27** | 2.112 | 9 | 2.082 | 10 | 2.61 | 4 | 2.387 | 4 |
| **28** | 2.162 | 8 | 2.137 | 9 | 2.934 | 6 | 2.48 | 3 |
| **29** | 2.128 | 6 | 2.025 | 4 | 2.903 | 13 | 2.611 | 9 |
| **30** | 2.104 | 4 | 2.057 | 4 | 2.945 | 5 | 2.637 | 5 |
| **31** | 2.268 | 7 | 2.171 | 6 | 2.571 | 3 | 2.374 | 3 |

Tabla B.3: Exponentes de power-law

| Comunidades detectadas | 95 |
|---|---|
| Modularidad | 0.361745 |

| Comunidad | $n_i$ | $k_{in}$ | $k_{out}$ | $k_{w,in}$ | $k_{w,out}$ | $K_{IO}$ | $M$ |
|---|---|---|---|---|---|---|---|
| acampadasol | 19674 | 56 | 51 | 109324 | 110616 | 0.9883 | 1.0077 |
| bufetalmeida | 18543 | 45 | 47 | 98376 | 93251 | 1.0550 | 0.9076 |
| anon_leakspin | 14592 | 39 | 37 | 25185 | 25790 | 0.9765 | **1.2810** |
| iescolar | 14069 | 38 | 39 | 63107 | 50448 | **1.2509** | **0.7693** |
| acampadabcn | 12919 | 37 | 36 | 43073 | 50297 | 0.8564 | 0.9355 |
| democraciareal | 10173 | 36 | 41 | 54127 | 58994 | 0.9175 | 0.8421 |
| elmundoes | 9388 | 33 | 32 | 22132 | 22501 | 0.9836 | 1.0749 |
| yoriento | 8514 | 36 | 35 | 26783 | 29965 | 0.8938 | 0.9189 |
| psoe | 7888 | 33 | 36 | 27886 | 33251 | 0.8387 | 0.9547 |
| twitpic | 7866 | 35 | 31 | 17073 | 12413 | **1.3754** | 0.8380 |
| perezreverte | 6571 | 32 | 31 | 17173 | 11379 | **1.5092** | **0.7882** |
| el_pais | 5935 | 31 | 32 | 13659 | 10876 | **1.2559** | 0.9172 |
| phumano | 5752 | 31 | 32 | 27544 | 21824 | **1.2621** | **0.7283** |
| acampadavlc | 3729 | 31 | 33 | 13789 | 20705 | **0.6660** | 0.9309 |
| telesurtv | 3366 | 31 | 30 | 3730 | 4900 | **0.7612** | **1.8006** |
| spanishrevolution | 3040 | 30 | 31 | 9839 | 11049 | 0.8905 | 0.8549 |
| alex_riveiro | 2870 | 32 | 32 | 14621 | 16788 | 0.8709 | **0.7726** |
| 20m | 2424 | 31 | 33 | 12660 | 12049 | 1.0507 | **0.7142** |
| tedieris | 2050 | 29 | 31 | 6976 | 10784 | **0.6469** | 0.8562 |

Tabla B.4: .

## B.2 Análisis de la difusión

Incluímos la tabla de errores cuadráticos medios obtenidos entre diversidades reales y las de los modelos.

| Día | Random | Pref. Attach. | Evolutivo | Random | Pref. Attach. | Evolutivo |
|---|---|---|---|---|---|---|
| 13/05 | 0.05841 | 0.02097 | 0.01177 | 0.03681 | 0.01508 | 0.00999 |
| 14/05 | 0.10534 | 0.045 | 0.04899 | 0.07048 | 0.03504 | 0.05 |
| 15/05 | 0.08648 | 0.03475 | 0.018 | 0.17402 | 0.13181 | 0.10368 |
| 16/05 | 0.05263 | 0.00586 | 0.02138 | 0.05642 | 0.00744 | 0.03905 |
| 17/05 | 0.06229 | 0.01467 | 0.00985 | 0.10429 | 0.05249 | 0.02028 |
| 18/05 | 0.08048 | 0.02759 | 0.01552 | 0.10045 | 0.04767 | 0.01966 |
| 19/05 | 0.08399 | 0.02042 | 0.0072 | 0.07558 | 0.02894 | 0.00694 |
| 20/05 | 0.09481 | 0.02804 | 0.0034 | 0.06257 | 0.02506 | 0.00614 |
| 21/05 | 0.08179 | 0.00207 | 0.0116 | 0.10073 | 0.03028 | 0.02763 |
| 22/05 | 0.11107 | 0.04964 | 0.01015 | 0.08088 | 0.04869 | 0.01617 |
| 23/05 | 0.11209 | 0.04959 | 0.01723 | 0.09013 | 0.05259 | 0.01429 |
| 24/05 | 0.1041 | 0.04113 | 0.00496 | 0.06229 | 0.02767 | 0.00548 |
| 25/05 | 0.10356 | 0.04172 | 0.01956 | 0.06516 | 0.03291 | 0.00641 |
| 26/05 | 0.10385 | 0.04218 | 0.01383 | 0.08462 | 0.0499 | 0.01458 |
| 27/05 | 0.06993 | 0.03441 | 0.01462 | 0.11399 | 0.05017 | 0.01585 |
| 28/05 | 0.09183 | 0.03988 | 0.01054 | 0.08573 | 0.04547 | 0.01241 |
| 29/05 | 0.06965 | 0.00518 | 0.00609 | 0.07784 | 0.02109 | 0.01091 |
| 30/05 | 0.07788 | 0.02368 | 0.01007 | 0.12411 | 0.06219 | 0.02997 |
| 31/05 | 0.11827 | 0.05112 | 0.03075 | 0.07151 | 0.03163 | 0.01372 |

Table B.5: Error cuadrático medio.

# Anexo C

# Variabilidad en los datos simulados

Realizamos una serie de simulaciones (50) para los tres modelos para comprobar la variabilidad de las medidas que se utilizan en el análisis. Para mayor claridad, se muestra el estudio realizado en uno de los días.

## C.1 Ajustes power-law

Recogemos en la Tabla C.1 los valores medios y la desviación estándar de las variables obtenidas en los ajustes para los tres modelos. Se incluyen el exponente de la ley de potencias $\alpha$, el valor de $x_{min}$ y el error estándar $SE$. Se puede ver cómo el exponente varía muy poco en los dos segundos casos (Figura C.1) y como el ajuste es bastante preciso, dados los valores encontrados en el error. El valor de $x_{min}$ confirma que el modelo evolutivo proporciona una distribución de poblaciones power-law desde un valor más bajo de $x$. Eso quiere decir que la ley de potencias recoge las especies menos frecuentes en mayor medida que el preferential attachment. Los ajustes del modelo aleatorio son bastante pobres y presentan errores muy altos.

## C.2 Diversidad

Procedemos de forma análoga con los índices de diversidad. Se muestra el comportamiento de las diversidades al final de cada día (apartado 4.4.1) en la Figura y el resumen en la Tabla C.2.

Los índices son bastante consistentes y podemos permitirnos el hacer las comparaciones en el apartado 4.4 con un caso típico o con los valores medios.

| | $\alpha$ | | $x_{min}$ | | $SE$ | |
|---|---|---|---|---|---|---|
| | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
| Modelo aleatorio | 7.186 | 1.663 | 10.0 | 1.743 | 0.693 | 0.569 |
| Preferential attachment | 2.298 | 0.071 | 4.18 | 1.244 | 0.043 | 0.011 |
| Evolutivo | 2.001 | 0.030 | 2.18 | 0.384 | 0.024 | 0.003 |

Tabla C.1: Variación exponente $\alpha$

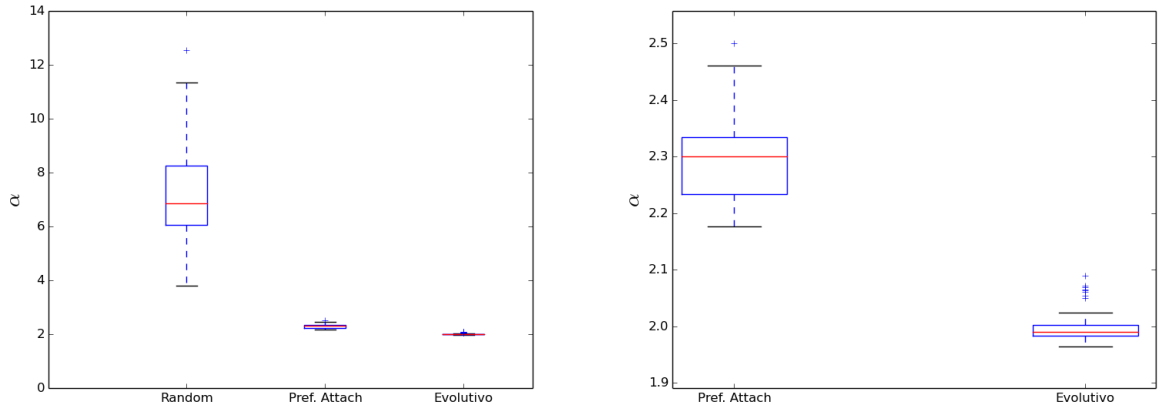



Figura C.1: Variación del exponente $\alpha$.

| | $D$ | | $J$ | |
|---|---|---|---|---|
| | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
| Modelo aleatorio | 0.0718 | 4E-4 | 0.971 | 3E-4 |
| Preferential attachment | 0.0142 | 1.6E-3 | 0.902 | 2.2E-3 |
| Evolutivo | 0.0121 | 1.7E-3 | 0.883 | 4.3E-3 |

Tabla C.2: Variación índices de diversidad.

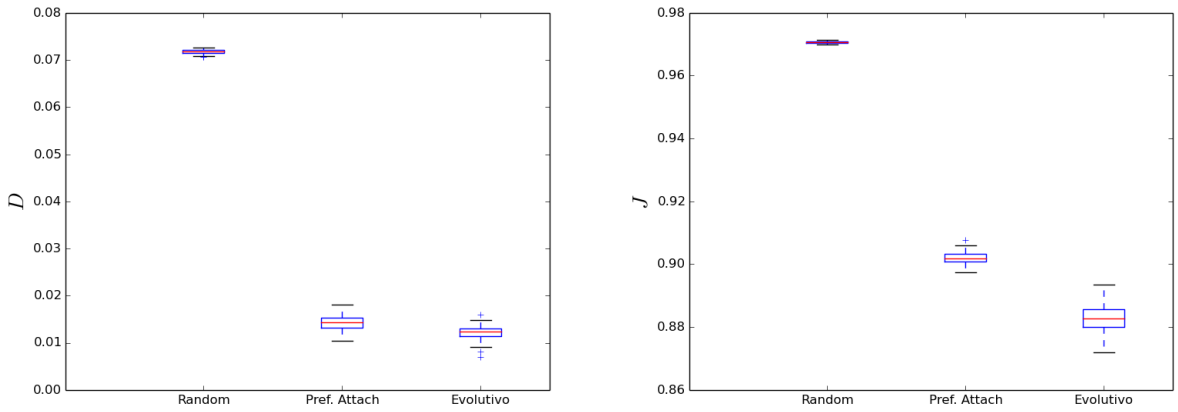



Figura C.2: Variación de índices $D$ y $J$ de diversidad.