FACULTAD DE CIENCIAS 2020/2021

DEPARTAMENTO DE FÍSICA TEÓRICA

GRADO EN FÍSICA

Trabajo Fin de Grado:

MODELOS DE DEFENSA ACTIVOS FRENTE A CIBERATAQUES EN REDES COMPLEJAS

Miguel Tarancón Cebrián

Dirigido por:
Alberto Aleta Casas
Yamir Moreno Vega

Resumen

Las redes de comunicación se han convertido en una pieza fundamental de los sistemas modernos de información. Sin embargo, esto ha hecho que los elementos de estas redes se conviertan en un objetivo de ataque. Para protegerlos, en una primera clasificación, podemos distinguir entre modelos de defensa pasivos, reactivos o activos. En estos últimos, tanto el atacante como el defensor pueden aprovechar las propiedades topológicas de la red para realizar su misión. En este contexto, se ha propuesto crear aplicaciones que utilicen las mismas técnicas que los gusanos emplean para propagarse a través de los dispositivos pero, en lugar de atacar el dispositivo, estos programas reforzarían la seguridad del sistema.

Cada gusano (el del atacante y el del defensor) puede utilizar vulnerabilidades distintas de los dispositivos y, por tanto, es posible que cada uno tenga características de propagación diferentes. Se propone explorar la dinámica de difusión de ambos gusanos y su interacción.

Índice

1. Introducción

Las epidemias llevan siendo un tema de estudio durante ya mucho tiempo debido a la importancia que pueden tener en las vidas de las personas. Al principio, debido a la limitación de los medios técnicos de los que se disponían para investigar, solo se estudiaban mediante modelos de ecuaciones que se intentaban resolver de forma analítica, para ver como influían los distintos parámetros en el desarrollo de los virus.

Conforme fueron avanzando los medios, se buscaron nuevas formas de abordar estos temas. Uno de los avances más importantes que se hicieron fue el uso de redes complejas a la hora de hacer simulaciones. Como rápidamente se vio, el uso de redes conformaba un punto de inflexión en este campo ya que permitía añadir muchos más grados de libertad a los modelos que hacían que los resultados que se obtenían cambiasen radicalmente con respecto a los obtenidos mediante los modelos anteriores. El inconveniente de estos avances era la necesidad de tener una ingente cantidad de datos reales para poder parametrizar los modelos. Pero los estudios no tardaron en realizarse debido al potencial que tenían estas nuevas formas de trabajar.

A la vez que se desarrollaban estas técnicas, había otro campo que empezaba a crecer a gran velocidad. Este era internet, el cual comenzaba a conectar más y más ordenadores de todo el mundo mediante una red. Esto propició la aparición de elementos malignos que viviesen en la red y se dedicasen a atacar los ordenadores que estaban conectados. Estos elementos son, obviamente, los virus informáticos, los cuales son programas informáticos que pueden "infectar" otros programas, modificándolos para incluir una copia suya.

El hecho de que los virus informáticos "viviesen" en la red de ordenadores, infectando algunos de ellos y propagándose por ella si había conexiones entre los programas, hizo que se viese una clara analogía con los modelos epidemiológicos que ya estaban muy estudiados. Por esta razón, se comenzaron a usar las herramientas que se habían desarrollado, tomando como datos reales las redes de ordenadores conectados que había. Los resultados que obtuvieron fueron bastante sorprendentes, al encontrar comportamientos extraños de los virus, que en una red aleatoria como la que consideraban sería muy extraño que ocurriesen. Como se acabó descubriendo posteriormente, esto era debido a que la topología de las redes que estudiaban no era como se imaginaban, si no que se trataban de redes libres de escala [1].

Como se vio rápidamente, los virus informáticos presentaban una clara amenaza, por lo que desde el primer momento se empezaron a desarrollar herramientas para contrarrestarlos. Las primeras que aparecieron fueron de tipo reactivas, es decir, una vez que se detectaba que uno de los ordenadores de la red había sido infectado, se aplicaban técnicas para eliminar al virus y dejar protegido al ordenador. Sin embargo, conforme la red de ordenadores conectados ha ido creciendo, este tipo de herramientas se han quedado algo obsoletas, en el sentido de que no aprovechan la conectividad de la red, a diferencia de los virus, los cuales van infectando ordenadores propagándose por ella. Esta es la razón por la cual el campo de la ciberdefensa activa está ganando importancia. Esta se encarga de desarrollar técnicas que sean capaces de proteger los ordenadores de amenazas aprovechándose de la conectividad de la red. De esta forma, no habría una asimetría entre los atacantes y los defensores [2].

Una de las técnicas con las que se ha comenzado a experimentar es la inclusión en la red de *white worms*.

Estos, al igual que los virus, se comportan como gusanos informáticos que viven en la red y se propagan de la misma forma que ellos. Sin embargo, su fin último es muy distinto, ya que se encargan de eliminar cualquier infección que pueda haber e intentan evitar cualquier otra posterior. Estos white worms están aún en las fases tempranas de desarrollo y quedan muchos problemas por resolver, sobre todo de carácter ético y legal. No obstante, se cree que pueden llegar a ser muy importantes para proteger en especial redes de aparatos IoT, las cuales cada vez crecen más y suponen una amenaza importante de ciberseguridad. Esto es debido a que este tipo de aparatos, por lo general, no tienen ni un *hardware* ni un *software* lo suficientemente potentes como para protegerse de ataques informáticos, por lo que resulta muy fácil infectarlos, provocando que se creen grandes redes de dispositivos infectados conectados a internet y capaces de hacer ataques muy poderosos. Estas redes son conocidas como *botnets*. La introducción en estas redes de un gusano que sea capaz de "curar" los dispositivos infectados y proteger al resto, permitiría evitar muchos problemas relacionados con las *botnets* [3].

Ya ha habido alguna propuesta de white worm capaz de realizar una tarea parecida a la descrita en el párrafo anterior, esta es AntibIoTic [4]. Más tarde se presentó la segunda versión de este gusano, AntibIoTic 2.0 [5]. Con esta nueva versión, los creadores trataron de solventar todos los problemas legales que tenía la primera versión, combinando el gusano con el nuevo paradigma *fog computing* de IoT. Este nuevo modelo trata de mover servicios que recogen los datos producidos por dispositivos IoT más cerca de estos, lo que reduciría el tráfico de datos en internet y mejoraría los servicios.

El objetivo del trabajo es desarrollar estrategias óptimas de defensa en función de las características de los gusanos y de la red de comunicación. Para ello, se estudiará la dinámica del sistema mediante un modelo compartimental sobre una red compleja.

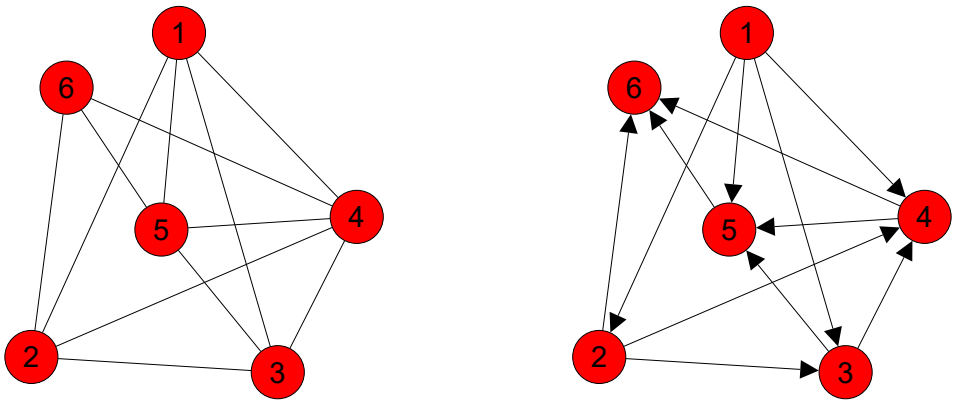
2. Teoría de redes

En esta sección se introducirán algunos de los aspectos más básicos en la ciencia de redes. Se empezará dando un pequeño resumen de los aspectos matemáticos más esenciales a la hora de estudiar redes, así como de las propiedades que sirven para caracterizarlas. Después se hará énfasis en dos tipos de distribuciones de grado concretas, las cuales serán esenciales en el transcurso de este trabajo, y, por último, se verán los modelos de redes que se han empleado.

Esta introducción a redes no es completa, ya que para eso ya hay buenas referencias como [6], si no más bien como una forma de establecer la terminología que se usará a lo largo del trabajo.

2.1. Definiciones y propiedades topológicas

Las redes complejas son estudiadas matemáticamente gracias a la teoría de grafos, ya que, formalmente, una red compleja puede ser representada mediante un grafo. Un grafo es una colección de nodos unidos por enlaces (*links*) [6]. Un grafo no dirigido $G = (\mathcal{N}, \mathcal{L})$ consiste en dos conjuntos $\mathcal{N}$ y $\mathcal{L}$, tales que $\mathcal{N} \neq \emptyset$ y $\mathcal{L}$ es un conjunto de pares desordenados de elementos de $\mathcal{N}$. Los elementos de $\mathcal{N}$ son los nodos, mientras que los de $\mathcal{L}$ son los links. El número de elementos en $\mathcal{N}$ y en $\mathcal{L}$ se denotan por N y K, respectivamente. Para identificar un grafo se usará la notación $G = (\mathcal{N}, \mathcal{L})$, o, simplemente, $G(N, K)$. En un grafo dirigido la única diferencia es que los pares que constituyen $\mathcal{L}$ son ordenados, ya que, como



(a) (b)

Figura 1: Representación gráfica de (a) un grafo no dirigido y (b) un grafo dirigido. Ambos grafos constan de $N = 6$ nodos y $K = 12$ enlaces. En el grafo dirigido los nodos vecinos son conectados mediante flechas indicando el sentido de cada link.

se verá ahora, el sentido en el que va el enlace es importante.

La manera de referirse a un nodo suele ser mediante su orden i en el conjunto $\mathcal{N}$. Después, en un grafo no dirigido los links se definen como una pareja de nodos distintos i y j y se denotan por l_{ij}. Dos nodos unidos por un enlace se dice que son vecinos. Por otro lado, en un grafo dirigido el orden de los nodos que definen un enlace es importante: l_{ij} se refiere a un enlace que va desde el nodo i al j, y $l_{ij} \neq l_{ji}$. Un grafo se suele dibujar mediante un punto para cada nodo y uniendo dos puntos por una línea si existe un link entre los correspondientes nodos [7]. En la Figura 1 se observan ejemplos de un grafo no dirigido y uno dirigido.

Hay ciertos casos donde puede haber más de un link entre los mismos nodos. La manera de referirse a estos vértices es como *multiedges*. También puede ocurrir que haya algún nodo conectado consigo mismo, lo que se conoce como *self-loop*. Notar que en ninguno de los grafos de la Figura 1 hay elementos de este tipo, ya que, según la definición que se ha dado de grafo, estos no están permitidos. Los grafos que contienen alguno de estos elementos se conocen como multigrafos [6]. En este trabajo el interés reside en los grafos más que en los multigrafos, más concretamente en los no dirigidos, es decir, como el que aparece en la Figura 1a. También existen grafos en los cuales cada enlace tiene un peso diferente, usualmente un número real, son los que se conocen como grafos ponderados. De todas formas, para este trabajo no se considerarán este tipo de grafos, si no que será como si todos los links tuvieran un peso unitario.

Otro de los conceptos claves a la hora de caracterizar redes es la capacidad de conexión entre dos nodos del grafo. De hecho, aunque dos nodos no sean vecinos, podrá ser accesible ir de uno a otro. De esta forma se define el camino del nodo i al nodo j como la secuencia de nodos vecinos que empieza en i y acaba en j. La longitud del camino se define como el número de enlaces en la secuencia. También existen los *paths*, que son caminos en los que no se pasa por ningún nodo más de una vez. Se dice que un grafo está conectado si para cada par de nodos distintos i y j hay, al menos, un *path* que los una. Si esto no se cumple, el grafo estará desconectado [7].

2.1.1. Matriz de adyacencia

La forma fundamental de representar matemáticamente una red es mediante la matriz de adyacencia. Considerando que se tiene un grafo $G = (\mathcal{N}, \mathcal{L})$, la matriz de adyacencia $\mathcal{A}$ se define como la matriz cuadrada $N \times N$ cuyos elementos A_{ij} $(i, j = 1, \ldots, N)$ cumplen

$$A_{ij} = \begin{cases} 1 \text{ si el link } l_{ij} \text{ existe,} \\ 0 \text{ de otra manera.} \end{cases} \tag{1}$$

En el caso de las redes que interesan en el trabajo (no dirigidas, simples, no pesadas), la matriz de adyacencia es simétrica, ya que si existe el link l_{ij} también existirá el l_{ji}, y cero en la diagonal principal, debido a que no tiene ni *multiedges* ni *self-loops* [6, 7]. Por tanto, considerando el grafo de la Figura 1a, su matriz de adyacencia sería

$$\mathcal{A} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \end{pmatrix}.$$

2.1.2. Grado y distribución de grado

En una red no dirigida $G = (\mathcal{N}, \mathcal{L})$ el grado de un nodo es el número de vértices conectados a él. A pesar de la simplicidad del concepto, es una de las herramientas más útiles y más usadas en redes. El grado de un nodo i se denota por k_i y se calcula en términos de la matriz de adyacencia como

$$k_i = \sum_{j \in \mathcal{N}} A_{ij}. \tag{2}$$

Cada vértice tiene dos finales y si hay K de ellos, en total habrá $2K$ finales de vértices. Pero el número de finales coincide también con la suma de los grados de todos los nodos, ya que

$$2K = \sum_{i,j \in \mathcal{N}} A_{ij} = \sum_{i \in \mathcal{N}} k_i. \tag{3}$$

Existen redes las cuales todos sus nodos tienen el mismo grado. En teoría de grafos estas se conocen como redes regulares [6].

La caracterización topológica más básica de una red se consigue gracias a su distribución de grado $P(k)$. Esta se define como la probabilidad de que al elegir un nodo de forma aleatoria, este tenga grado k o, equivalentemente, la fracción de nodos en la red con grado k. Para obtener información sobre cómo se distribuye el grado entre los nodos se calculan los momentos de la distribución. El momento n de $P(k)$ se define como

$$\langle k^n \rangle = \sum_k k^n P(k). \tag{4}$$

El primer momento $\langle k \rangle$ se corresponde con el grado medio de G [7].

En las redes no correlacionadas la distribución de grado caracteriza completamente las propiedades estadísticas. Sin embargo, en las redes reales suele ocurrir que existen correlaciones en el sentido de que

la probabilidad de que un nodo de grado k esté conectado a otro nodo de grado k' depende de k. En estos casos es necesario introducir la probabilidad condicionada $P(k' \mid k)$, que se define como la probabilidad de que un link desde un nodo de grado k apunte a un nodo de grado k'. $P(k' \mid k)$ cumple la condición de normalización $\sum_{k'} P(k' \mid k) = 1$ y la de equilibrio detallado $kP(k' \mid k) P(k) = k'P(k \mid k') P(k')$ [8].

En función de las correlaciones de grado que haya, se clasifica a las redes en dos grandes grupos, redes asortativas y redes disortativas. Las primeras se corresponden con aquellas en las que los nodos de grado alto tienden a conectarse con otros nodos de grado también alto. Por otro lado, en las disortativas los nodos de grado alto tienden a conectarse con nodos de grado bajo. Una forma de medir la asortatividad en una red es mediante el coeficiente de asortatividad r, que toma valores entre -1 y 1. Cuando $r < 0$ la red es disortativa y cuando $r > 0$ es asortativa. En el caso $r = 0$ la red es no correlacionada [9].

En redes no correlacionadas, en las que $P(k' \mid k)$ no depende de k, las condiciones de normalización y de balance detallado dan

$$\begin{aligned} &\sum_k kP\left(k' \mid k\right) P(k) = \sum_k k'P\left(k \mid k'\right) P\left(k'\right) \Rightarrow \\ &\Rightarrow P\left(k' \mid k\right) \sum_k kP(k) = k'P\left(k'\right) \sum_k P\left(k \mid k'\right) \Rightarrow \\ &\Rightarrow P\left(k' \mid k\right) \langle k \rangle = k'P\left(k'\right) \Rightarrow P\left(k' \mid k\right) = \frac{k'P(k')}{\langle k \rangle}. \end{aligned} \tag{5}$$

2.2. Modelos de redes

En la actualidad, gracias a la abundancia de datos y medidas de redes reales, se ha descubierto la existencia de diferentes tipos de redes, caracterizadas por una gran variabilidad de sus métricas básicas y propiedades estadísticas. Esto ha impulsado la investigación de diferentes modelos de generación de redes. La utilidad de estos modelos es que sirven como generadores de redes sintéticas con características parecidas a las de las redes reales y en las que se puede estudiar el comportamiento de procesos dinámicos.

El primer modelo que se propuso es la red aleatoria clásica de Erdös-Rényi [10]. Según este modelo, un grafo $G(N, K)$ es construido a partir de un conjunto de N nodos en el que cada uno de los $N(N-1)/2$ posibles enlaces está presente con probabilidad p. La distribución de grado de esta red está dada por una distribución binomial

$$P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k},$$

que en el límite de grado medio constante (específicamente $p = \langle k \rangle / (N-1)$) y N grande, tiende a una distribución de Poisson

$$P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}. \tag{6}$$

Por tanto, este modelo es adecuado en el caso de redes gobernadas únicamente por la estocástica, aunque $G(N, K)$ tienda a un grafo regular para N grande y p constante. La distribución de grado alcanza su máximo en torno al valor medio, denotando cierta homogeneidad estadística en los nodos [9]. En cuanto a las correlaciones de grado, para este modelo se encuentra que $r = 0$, por lo que se generan redes no

correlacionadas [11]. En la Figura 2a se observa el grafo de una red aleatoria, donde se ve la homogeneidad en los nodos de la que se ha hablado. Por otro lado, en la Figura 2b se observa la distribución de grado de otra red aleatoria, en la que se ve como se aproxima muy bien a una distribución de Poisson.

Sin embargo, la evidencia empírica ha demostrado que las redes reales no se comportan como redes aleatorias, sino que estas exhiben altos niveles de heterogeneidad. Las distribuciones estadísticas que caracterizan estas redes son generalmente desiguales y varían a lo largo de varios órdenes de magnitud. Suele ser esclarecedor representar las distribuciones de grado de redes reales para sacar más conclusiones. Esto es lo que hicieron en 1999 Barabási, Albert y Jeong cuando estudiaron la red WWW [12]. Lo que encontraron es que la gran mayoría de los nodos tenían un grado bajo, pero la distribución tenía una gran "cola" por la derecha, correspondiente con nodos de alto grado. A estos nodos tan bien conectados se les llama *hubs*.

Lo interesante de todo esto es que, estudios posteriores han llegado a la conclusión de que la mayoría de las redes reales tienen distribuciones de grado con una cola de *hubs* de alto grado como la descrita hace un momento. En el lenguaje de la estadística se conoce a este tipo de distribuciones por su término en inglés *heavy-tailed*, y suelen aproximarse por un comportamiento de una ley potencial de la forma $P(k) \sim k^{-\alpha}$, lo que implica una probabilidad no despreciable de encontrar nodos con alto grado.

A pesar de la simplicidad de este tipo de distribuciones, las cantidades que las describen se comportan de maneras sorprendentes [6]. Comenzando por la constante de normalización, tomando que se normaliza a partir de cierto grado $k_{min} > 0$, esta es

$$C = \frac{1}{\sum_{k=k_{min}}^{\infty} k^{-\alpha}} = \frac{1}{\zeta(\alpha, k_{min})},$$

donde $\zeta(\alpha, k_{min})$ es la función zeta generalizada. Así, la distribución completa es $P(k) = Ck^{-\alpha}$. Si se considera que en la cola la suma sobre k se aproxima bien con una integral, la constante queda $C \simeq (\alpha - 1)\, k_{min}^{\alpha-1}$.

Pasando ahora a los momentos de la distribución, usando la expresión (4) y separando la suma en dos partes, la del principio de la distribución y la de la cola (donde sigue una ley potencial), el momento n queda

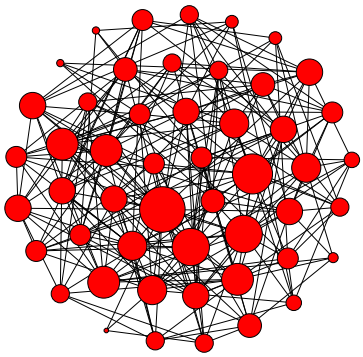
$$\langle k^n \rangle = \sum_{k=0}^{k_{min}-1} k^n P(k) + C \sum_{k=k_{min}}^{\infty} k^{n-\alpha}.$$

Al igual que para la constante de normalización, si se considera que para la cola, la suma se puede aproximar por una integral el momento queda

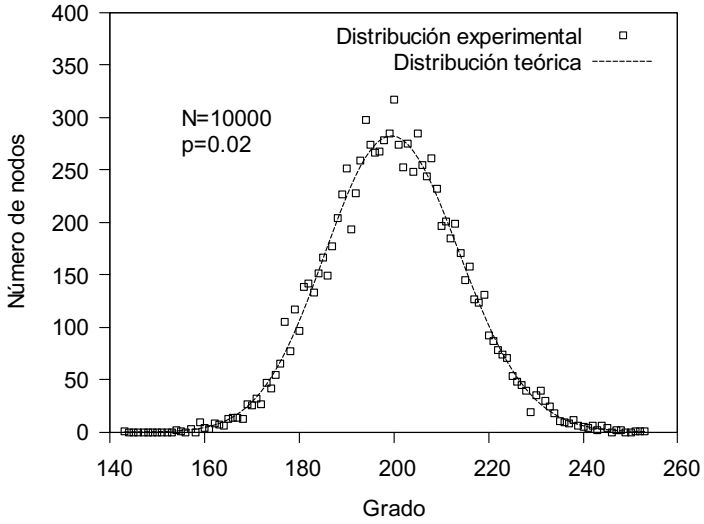
$$\langle k^n \rangle \simeq \sum_{k=0}^{k_{min}-1} k^n P(k) + \frac{C}{n-\alpha+1} \left[k^{n-\alpha+1} \right]_{k=k_{min}}^{\infty}.$$

El primer término es un número finito cuyo valor depende de la forma particular de la distribución no potencial para pequeños k. El segundo término depende de los valores de n y α. Si $n - \alpha + 1 < 0$ la integral tiene un valor finito, pero si $n - \alpha + 1 \geq 0$ la integral divergirá y con ella el momento $\langle k^n \rangle$. Por tanto, el momento n de la distribución será finito solo si $\alpha > n + 1$.

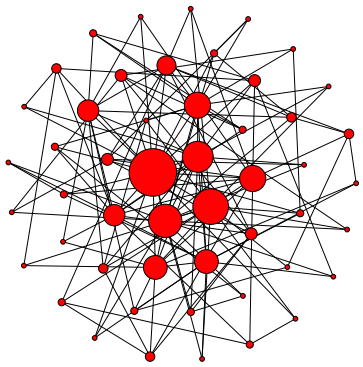
De especial interés es el segundo momento, el cual será finito si $\alpha > 3$. Lo que se encuentra es que, para muchas redes reales con leyes potenciales, el coeficiente α toma valores en el rango $2 \leq \alpha \leq 3$, por lo que



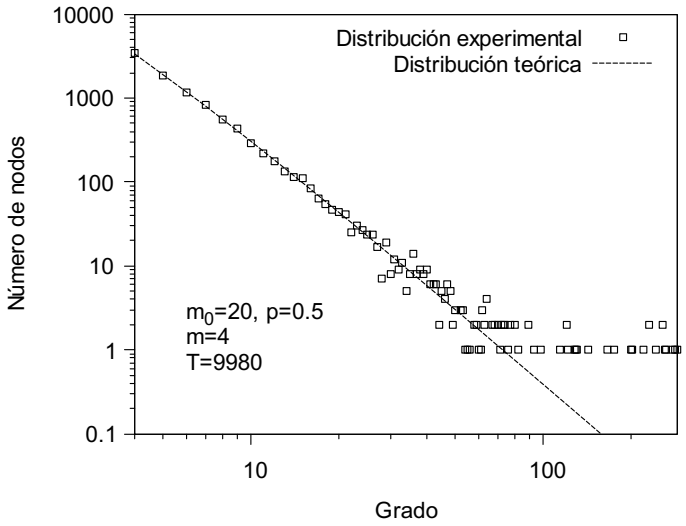
(a)



(b)



(c)



(d)

Figura 2: Diferentes ejemplos de modelos de redes. En (a) se representa el grafo y en (b) la distribución de grado para redes aleatorias. El grafo está construido a partir de una red con $N = 50$ nodos y una probabilidad de conexión $p = 0.2$, mientras que la distribución de grado se ha hecho a partir de una red con $N = 10000$ nodos y una probabilidad $p = 0.02$. Después, en (c) se representa el grafo y en (d) la distribución de grado de dos redes libres de escala. Estas redes han sido generadas gracias al modelo Barabási-Albert, tomando para el grafo los parámetros $m_0 = 5$, $m = 3$ y $T = 45$ y para la distribución de grado, $m_0 = 20$, $m = 4$ y $T = 9980$. En ambos grafos el tamaño de los nodos es proporcional a su grado. Las distribuciones de grado también se representan junto a sus curvas teóricas.

el segundo momento de estas redes tendría que divergir. En estos casos, las redes se denominan como libres de escala. Obviamente, los momentos solo divergirán cuando se esté en el límite de una red de tamaño infinito ($N \to \infty$). Para redes reales, debido al tamaño finito y a otras restricciones, los momentos no se harán infinitos, aunque sí que tendrán valores sorprendentemente grandes en comparación con el grado medio, reflejando las enormes fluctuaciones en la conectividad de los nodos [9].

Para crear este tipos de redes se han considerado diferentes paradigmas y modelos. Aquí se presenta el modelo Barabási-Albert (BA) [13], el cual es un modelo de red en crecimiento que considera que los nuevos nodos se conectarán a los nodos que ya estaban mediante una regla de conexión preferencial. En concreto, se considera que esta regla está basada en el grado de cada nodo, es decir, la probabilidad de añadir un enlace al nodo i es una función $F(k_i)$ de su grado. En su versión más simple el modelo funciona de la siguiente manera: (i) Se comienza con una pequeña red aleatoria de m_0 nodos, y a cada paso de tiempo se va añadiendo un nodo con m ($m < m_0$) links, que son conectados a los nodos antiguos de la red. (ii) Los nuevos enlaces se conectan al nodo i con una probabilidad $F(k_i) = k_i / \sum_j k_j$ [9].

La distribución de grado que genera este modelo es

$$P(k) = \frac{2m(m+1)}{k(k+1)(k+2)}, \tag{7}$$

que en el límite de k grande cumple $P(k) \sim k^{-3}$, por lo que se genera una red libre de escala con exponente $\alpha = 3$ [14]. Otra característica interesante es que para las redes que se generan con este modelo $r = 0$, es decir, son redes no correlacionadas [11]. En la Figura 2c se observa el grafo de una red libre de escala, donde se puede ver la heterogeneidad en el grado de cada nodo. Después, en la Figura 2d se representa la distribución de grado de otra red libre de escala, se ve como sigue una ley potencial, sobretodo para los grados más pequeños. Después, cuando ya crece el grado, debido a los efectos finitos el comportamiento se aleja del predicho.

Otro aspecto que no se ha tenido en cuenta sobre las redes es su carácter temporal. Hasta ahora se ha considerado que la topología de las redes era estática, ya que los conjuntos de nodos y links no cambiaban con el tiempo. Sin embargo, hay muchas redes reales que están lejos de ser estáticas. En algunas de estas redes, como la de Internet, la escala de tiempos característica a la que cambia la red es bastante pequeña. Es por esto, que en los casos donde las propiedades de los procesos dinámicos cambien mucho más rápido que la red, será una buena aproximación tomar redes estáticas [15]. Debido al ámbito que interesa para este trabajo, es decir, dispositivos conectados a internet que sufren ataques informáticos, tomar la aproximación anterior y usar redes estáticas es correcto.

Para finalizar esta sección se va a considerar una red real y se van a estudiar sus características. En concreto se considera la red Internet AS graph (2006) [16], la cual representa la estructura de internet al nivel de sistemas autónomos. Es una red de $N = 22963$ nodos y $K = 48436$ links. En la Figura 3 se observa su distribución de grado que, como se puede ver, sigue una ley potencial como la descrita. Comparándola con la distribución de la Figura 2d, para grados pequeños se comporta bastante bien, pero para grados grandes se desvía más, de igual forma que la otra. Se obtiene que el coeficiente de la ley potencial está en torno a $\alpha \simeq 2.3$, por lo que es una red libre de escala.

En cuanto a los momentos de la distribución, se obtienen $\langle k \rangle = 4.22$ y $\langle k^2 \rangle \simeq 1100$. Es decir, se ve como el segundo momento se hace muy grande, en comparación con el grado medio, por lo que cumple

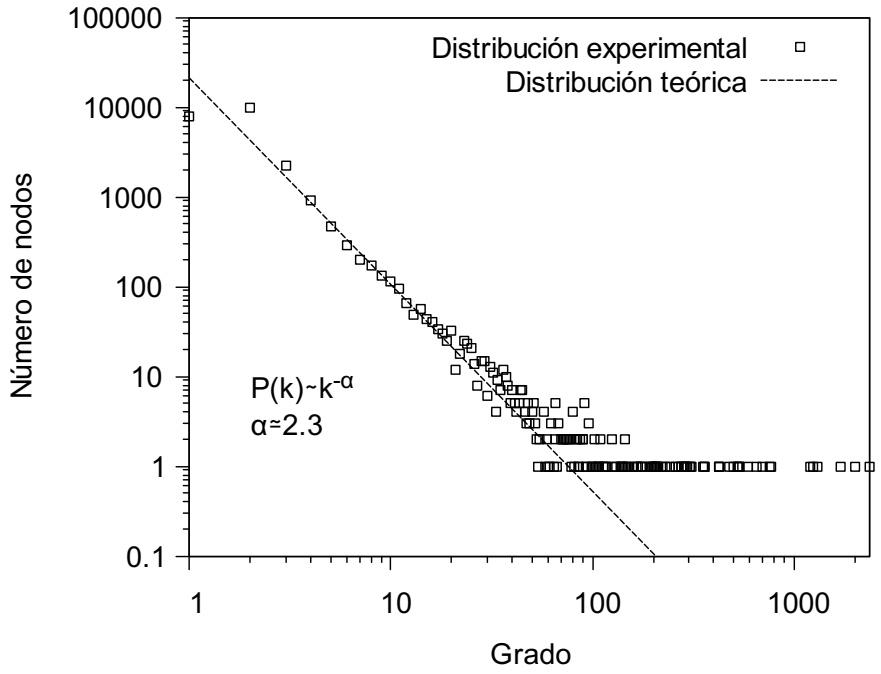


Figura 3: Distribución de grado de la red Internet AS graph (2006).

lo predicho para redes libre de escala. El coeficiente de asortatividad para esta red es $r = -0.2$, es decir, es una red ligeramente disortativa.

Debido a que esta red cumple las características descritas para una red real como las que interesan en este trabajo, será la que se use más adelante para realizar simulaciones y sacar resultados.

3. Procesos de propagación

El estudio de procesos de difusión lleva siendo un tema de gran interés desde hace tiempo. Esto se debe a la gran utilidad que tiene conocer cómo estos procesos se dan. Una de las aplicaciones más extendidas es la del estudio de difusión de epidemias. Gracias a los resultados que aportan estos estudios se pueden desarrollar estrategias para controlar y erradicar enfermedades. Pero no solo se puede modelar la difusión de enfermedades, sino que cualquier sistema que pueda verse como un proceso de contagio, también.

En esta sección se verán algunos de los modelos más básicos usados para modelar la difusión de enfermedades; después se estudiarán algunas aproximaciones que se pueden realizar a la hora de intentar abordar el problema de resolver este tipo de modelos; y, por último, se verá como se puede llevar todo esto a redes para ver como influye la topología de estas a los procesos de difusión.

3.1. Modelos de difusión de enfermedades

La asunción más básica realizada a la hora de modelar epidemias es la de considerar que la población se puede dividir en diferentes clases o compartimentos dependiendo del estado de la enfermedad. Esto es lo que se conoce como modelo compartimental [17]. En el caso más simple solo se consideran dos

estados: susceptible (S) e infectado (I). Los individuos en el estado susceptible son aquellos que no tienen la enfermedad pero si tienen contacto con alguien que sí la tenga, pueden infectarse. Por otro lado, los individuos del estado infectado son aquellos que tienen la enfermedad y pueden infectar a otros que no la tengan. A partir de aquí se pueden seguir añadiendo compartimentos para tener en cuenta más estados que pueda tener una enfermedad. Por ejemplo, suele ser útil el estado recuperado (R), donde estarían los individuos que ya han pasado la enfermedad y no pueden volver a contagiarse porque se han inmunizado [9].

El objetivo de estos modelos es conocer cómo evoluciona el número de individuos en cada estado a lo largo del tiempo. Para ello se tienen que definir los procesos básicos que tienen lugar al nivel de individuos y que gobiernan las transiciones de un estado a otro. Estos se caracterizan gracias a ciertos parámetros que describirían las probabilidades de que un individuo cambie de compartimento. Para obtener el valor de estos parámetros se recurre a métodos completamente experimentales.

El primer modelo compartimental que se considera es el SIS. Este consta de dos estados y solo puede haber dos transiciones

$$S + I \xrightarrow{\beta} 2I,$$
$$I \xrightarrow{\mu} S.$$

La primera describe el proceso de contagio y ocurre cuando un individuo infectado tiene contacto con uno susceptible. Está caracterizada por el parámetro β, que describe el ritmo de transmisión de la enfermedad. La segunda da cuenta de las recuperaciones de los individuos infectados, volviendo al estado susceptible, ya que pasar la enfermedad no otorga inmunidad. Esta está caracterizada por el parámetro μ, que representa el ritmo de recuperación. Como es obvio, la naturaleza de estos procesos es completamente distinta, ya que las recuperaciones ocurren de manera espontánea al cabo de cierto tiempo, mientras que los procesos de contagio dependen de los patrones de interacción entre individuos [6].

Otro modelo interesante es el SIR. Este consta de tres estados y la diferencia con el SIS es que los individuos que se recuperan de la enfermedad pasan al estado recuperado, ya que han adquirido inmunidad contra ella. De esta forma, las transiciones que puede haber son

$$S + I \xrightarrow{\beta} 2I,$$
$$I \xrightarrow{\mu} R.$$

Claramente, en este modelo la evolución no es infinita, sino que esta para en el momento en el que ya no hay individuos infectados. Esto quiere decir que todos los individuos están bien en el estado susceptible o bien en el protegido. A este tipo de estados se les conoce como estados absorbentes, ya que, cuando acaba la evolución, todos los nodos están en alguno de estos estados. Por el contrario, en el modelo SIS los individuos pueden infectarse una y otra vez, experimentando un ciclo $S \to I \to S$, que bajo ciertas condiciones se puede mantener para siempre, llegando a un estado estacionario en el que haya individuos infectados, conocido como estado endémico [9].

Existen muchos otros modelos compartimentales más complejos (ver Figura 4), pero para el interés de esta sección, con estos es suficiente. A partir de aquí, en lo que queda de sección se considerará el modelo SIS a la hora de hablar de los siguientes temas.

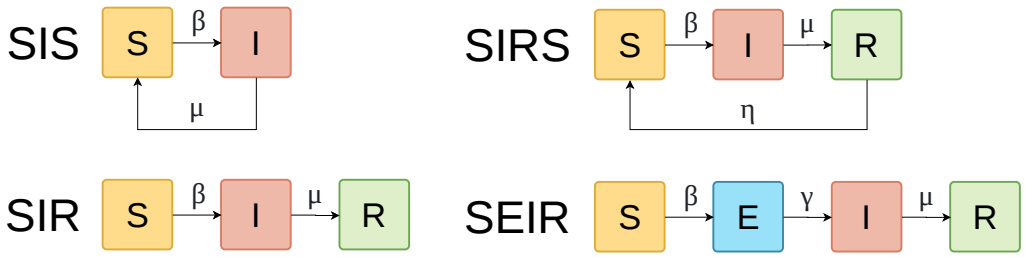


Figura 4: Representación con diagramas de flujo de diferentes modelos compartimentales usados en modelización de epidemias. Cada caja representa un compartimento, mientras que las flechas representan transiciones entre compartimentos, ocurriendo aleatoriamente de acuerdo a sus respectivos ritmos. El nuevo estado que aparece es latente (E), el cual representa a los individuos que están infectados y pueden contagiar, pero aún no tienen síntomas.

3.2. Aproximación *homogeneous mixing*

Los modelos descritos hasta ahora están basados en las propiedades de la enfermedad y de los individuos (para la recuperación), pero no se ha tenido en cuenta en ningún momento un elemento crucial en la difusión de enfermedades, la red de contactos.

La primera hipótesis clásica que se realiza sobre la red de contactos cuando se quiere estudiar la evolución de una enfermedad es la de *homogeneous mixing* [18, 19], también conocida como aproximación de campo medio, debido a las similitudes que tiene con este tipo de aproximaciones de la física estadística. Bajo este enfoque, se considera que la población está completamente mezclada y se asume que cada individuo tiene la misma probabilidad por unidad de tiempo de contactar con cualquier otro individuo. Esto, obviamente, no es una buena representación de cómo es el mundo en realidad, aun así, un estudio de los acercamientos clásicos resulta útil a la hora de estudiar la epidemiología en redes.

La gran ventaja que presenta esta aproximación es que permite escribir el modelo en la forma de un sistema de ecuaciones diferenciales ordinarias de las densidades de los individuos en cada estado. Este sistema puede ser resuelto mediante cualquier método de resolución numérica de ecuaciones diferenciales (como Runge-Kutta) y así obtener la evolución de la enfermedad a lo largo del tiempo. En teoría, como el proceso de difusión es completamente aleatorio, la evolución no está determinada unívocamente, ya que si la enfermedad volviese a propagarse por la misma población más de una vez, incluso bajo las mismas condiciones, cada vez se obtendría una evolución distinta. Sin embargo, resolviendo este sistema se alcanzaría siempre el mismo resultado. Esto se puede entender como que los resultados obtenidos con el sistema de ecuaciones son los que se obtendrían al hacer el promedio de muchas evoluciones bajo las mismas condiciones y son resultados completamente deterministas.

Se describe ahora como sería el modelo SIS bajo esta aproximación. Se supone que $S\,(t)$ es el número de individuos en el estado susceptible a tiempo t e $I\,(t)$ son los que están infectados. El número de infectados crece cuando los individuos susceptibles contraen la enfermedad de los infectados. Como este proceso está caracterizado por el ritmo de transición β, esto quiere decir que cada individuo infectado tiene, en promedio, $\langle k \rangle \beta$ contactos por unidad de tiempo con el resto de individuos de forma aleatoria, donde $\langle k \rangle$

es el número medio de contactos dentro de la red, i.e. el grado medio. Si la población total está formada por N individuos, la probabilidad de contactar con un susceptible es S/N, por lo que un infectado tiene de media $\langle k\rangle \beta S/N$ contactos con susceptibles por unidad de tiempo. Como hay I infectados en total, el ritmo global de nuevas infecciones será $\langle k\rangle \beta S I/N$. Por otro lado, como los infectados se recuperan a un ritmo μ, el ritmo de recuperaciones será μI [6].

Según lo explicado en el párrafo anterior, el sistema de ecuaciones que permite simular el modelo es el siguiente [9]

$$\frac{dS}{dt} = -\langle k\rangle \beta \frac{S I}{N} + \mu I, \tag{8a}$$

$$\frac{dI}{dt} = \langle k\rangle \beta \frac{S I}{N} - \mu I. \tag{8b}$$

Si ahora se definen las densidades de cada estado como $\rho^\alpha = N^\alpha/N$, siendo N^α el número de individuos en el estado α, el sistema de ecuaciones (8) queda

$$\frac{d\rho^S}{dt} = -\langle k\rangle \beta \rho^S \rho^I + \mu \rho^I, \tag{9a}$$

$$\frac{d\rho^I}{dt} = \langle k\rangle \beta \rho^S \rho^I - \mu \rho^I, \tag{9b}$$

con $\rho^S + \rho^I = 1$, por lo que una de las dos ecuaciones es redundante y con una es suficiente para describir el modelo.

3.3. Aproximación DBMF

Como ya se ha dicho, la aproximación *homogeneous mixing* no es del todo realista ya que no tiene en cuenta los diferentes patrones de contacto que puede haber en una red real. Pensar que todos los individuos tienen aproximadamente el mismo número de contactos y que pueden ser con cualquier otro individuo de forma aleatoria choca con la realidad. En la vida real cada individuo tendría su propio patrón de contacto social, por lo que resulta razonable pensar que estos diferentes patrones de contacto podrían representarse mediante una red compleja. Hay diferentes formas de introducir este aspecto dentro de un modelo. En concreto, en esta sección se describirá la aproximación *degree-based mean field* (DBMF) [20].

Bajo esta hipótesis de campo medio se supone que todos los nodos con el mismo grado k se comportan estadísticamente igual. Esta asunción implica que ya no es necesario considerar el estado específico de cada nodo, si no que las cantidades relevantes son las densidades $\rho_k^\alpha(t) = N_k^\alpha(t)/N_k$, que representan la densidad de nodos con grado k en el estado α (N_k^α es el número de individuos con grado k en el estado α, y N_k es la cantidad de nodos con grado k). La asunción también implica que cualquier nodo de grado k está conectado con la misma probabilidad $P(k' \mid k)$ a un nodo de grado k'. Esta aproximación es conveniente ya que reduce el número de grados de libertad del sistema enormemente.

En la teoría DBMF, las densidades ρ_k^α se pueden ver como la probabilidad de que un individuo de la población con grado k esté en el compartimento α. Aunque estas variables no son independientes, satisfacen la condición de normalización $\sum_\alpha \rho_k^\alpha = 1$. Después, la fracción total de individuos en el estado α se calcula como $\rho^\alpha = \sum_k P(k)\rho_k^\alpha$.

Esta teoría contiene, implícitamente, otra aproximación, ya que la equivalencia entre los nodos de un mismo grado considera la red en una perspectiva de campo medio en la que la matriz de adyacencia $\mathcal{A}$ se rompe completamente y solo se mantienen la distribución de grado y las correlaciones entre nodos de distinto grado. Esto es equivalente a pensar que la escala de tiempos de los procesos de difusión es mucho más lenta que la que caracteriza los cambios de los patrones de interacción en la red. Es como si la red estuviese constantemente reconectándose, pero preservando $P(k)$ y $P(k' \mid k)$. Aunque la teoría DBMF es obviamente una aproximación muy grande, es capaz de capturar el comportamiento de epidemias y de procesos dinámicos complejos [9].

Considerando el modelo SIS, las ecuaciones que gobiernan la evolución del sistema se pueden obtener siguiendo un razonamiento similar al hecho para el sistema de ecuaciones (8). El ritmo de recuperaciones tendrá una forma similar, pero ahora solo se contará la densidad de infectados con grado k, por lo que queda $\mu\rho_k^I$. En cuanto al ritmo de contagios, para un cierto grado k el número de contactos que realiza por unidad de tiempo es $k\beta$. La probabilidad de que un nodo de grado k sea susceptible es ρ_k^S, por lo que un infectado tiene de media $k\beta\rho_k^S$ contactos con susceptibles. Ahora, teniendo en cuenta la probabilidad de que el nodo esté conectado a otro de grado k' y este esté infectado: $P(k' \mid k)\,\rho_{k'}^I$, el ritmo de contagios quedaría $k\beta\rho_k^S P(k' \mid k)\,\rho_{k'}^I$. Este factor se debe sumar a todos los posibles valores de k'. Así, según lo explicado, el sistema queda

$$\frac{d\rho_k^S}{dt} = -k\beta\rho_k^S \sum_{k'} P\left(k' \mid k\right)\rho_{k'}^I + \mu\rho_k^I, \tag{10a}$$

$$\frac{d\rho_k^I}{dt} = k\beta\rho_k^S \sum_{k'} P\left(k' \mid k\right)\rho_{k'}^I - \mu\rho_k^I, \tag{10b}$$

donde se cumple que $\rho_k^S + \rho_k^I = 1$, por lo que con una de las ecuaciones es suficiente. Estas expresiones constituyen un sistema de dos ecuaciones para cada grado k, por lo que sería necesario resolverlo para cada uno de los grados simultáneamente para así obtener el resultado global. En el caso de tener un red no correlacionada, se cumple la ecuación (5) y el sistema queda

$$\frac{d\rho_k^S}{dt} = -k\beta\rho_k^S\Theta + \mu\rho_k^I, \tag{11a}$$

$$\frac{d\rho_k^I}{dt} = k\beta\rho_k^S\Theta - \mu\rho_k^I, \tag{11b}$$

donde

$$\Theta = \sum_{k'} \frac{k'P(k')}{\langle k\rangle}\rho_{k'}^I, \tag{12}$$

que sería la probabilidad de encontrar un nodo infectado tomando un vértice de forma aleatoria [9].

3.4. Propagación en redes: Gillespie

Hasta ahora se han empleado diferentes aproximaciones que permitían resolver estos modelos de forma completamente determinista. Sin embargo, este tipo de procesos son totalmente estocásticos, ya que los parámetros de los que dependen las transiciones son probabilidades. Para realizar estas simulaciones estocásticas se empleará el método de Monte Carlo y se controlará el estado de cada nodo, viendo todos

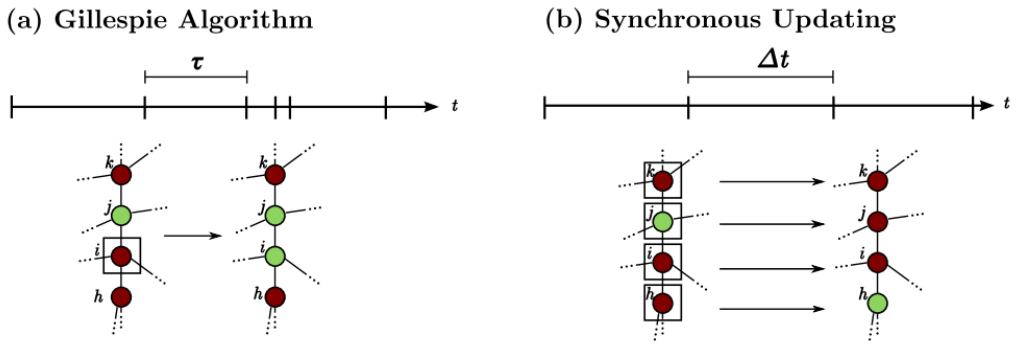


Figura 5: Esquemas de (a) el algoritmo de Gillespie y (b) la actualización síncrona en redes. Las marcas verticales en el eje del tiempo indican los momentos en los que se producen transiciones. En la actualización síncrona estos momentos ocurren cada un tiempo fijo Δt, mientras que en Gillespie el intervalo es una variable aleatoria τ. Los círculos verdes y rojos representan nodos de la red en los estados susceptible e infectado, respectivamente. Un cuadrado alrededor de un círculo indica que ha sido seleccionado para actualizarse en ese intervalo y podría cambiar su estado. En el algoritmo de Gillespie solo se elige un nodo y siempre cambia su estado. En la actualización síncrona todos los nodos tienen la opción de cambiar su estado, pero lo harán o no con una probabilidad que depende de su estado y del estado de sus vecinos. Imagen extraída de [23].

los posibles eventos que pueden ocurrir y la probabilidad con la que podría ocurrir cada uno. En la Figura 5 se pueden ver los esquemas de los dos métodos que se van a comentar a continuación.

Una primera forma de llevar a cabo este tipo de simulaciones es mediante la aproximación de tiempo discreto. En esta aproximación se divide el tiempo en pequeños intervalos Δt y en cada uno de ellos se calcula la probabilidad de que un nodo cambie su estado en el siguiente intervalo. Entonces, generando números aleatorios se puede escoger qué cambios son los que se producirán síncronamente y después pasar al siguiente intervalo. Una importante limitación que tiene este método es que, si los intervalos son muy largos, puede ocurrir que en un mismo intervalo ocurran varios eventos que podrían afectarse mutuamente. Por ejemplo, un nodo podría recuperarse y transmitir en el mismo intervalo, pero no se sabría qué ocurre antes. Esto podría arreglarse cogiendo intervalos muy pequeños, pero esto haría que la simulación fuese muy lenta [21].

Existen enfoques alternativos, como la aproximación de tiempo continuo. En este tipo de aproximaciones el estado de los nodos se actualiza de forma asíncrona, es decir, en cada intervalo de tiempo solo cambia el estado de un nodo. En concreto, en este trabajo se presenta el algoritmo de Gillespie [22]. Inicialmente, este método fue propuesto para hacer simulaciones estocásticas de reacciones químicas, pero posteriormente, debido al amplio rango de aplicabilidad que tenía, se empezó a usar para hacer simulaciones epidemiológicas en redes. En una simulación de Gillespie se calcula el tiempo que pasará hasta que ocurra el siguiente evento, sabiendo el ritmo combinado de todos los posibles eventos que podrían ocurrir en ese momento. El tiempo se escoge aleatoriamente de una distribución exponencial con ese ritmo. Después, usando otro número aleatorio se determina cuál de los posibles eventos es el que ocurre. Esto da una simulación estocástica exacta.

Según este método, dado el estado de la red, se pueden calcular las distribuciones de probabilidad que gobiernan tanto la longitud del intervalo hasta el próximo suceso como la decisión de qué nodo es el que se actualiza, gracias a los ritmos de transición individuales de cada nodo $\eta_i(t)$. Por ejemplo, para el modelo SIS estos ritmos de transición serían $\eta_i^S(t) = k_{inf}\beta$, donde k_{inf} es el número de nodos vecinos infectados, para los nodos susceptibles, y $\eta_i^I(t) = \mu$ para los nodos infectados. Así, si r es un número aleatorio uniforme en el intervalo $[0, 1)$ y se define el ritmo total de transición $\omega(t) = \sum_i \eta_i(t)$, el tiempo que pasa hasta el siguiente suceso es

$$\tau = \frac{1}{\omega} \ln\left(\frac{1}{r}\right). \tag{13}$$

Después, para elegir la transición que va a ocurrir se usa otro número aleatorio uniforme $u \in [0, 1)$ y se ve para qué nodo se cumple la condición

$$\sum_{j=1}^{k-1} \frac{\eta_j}{\omega} < u < \sum_{j=1}^{k} \frac{\eta_j}{\omega} \quad (k = 1, \ldots, N), \tag{14}$$

donde N es el número de nodos en la red. Así, el nodo k será el elegido para cambiar su estado en ese intervalo de tiempo. Este proceso se repetirá hasta que la población alcance un estado estacionario, bien porque todos los individuos han ido a un estado absorbente, como el estado recuperado del modelo SIR, o porque se ha llegado a un estado endémico [24].

4. Propagación de virus informáticos

Como ya se ha dicho anteriormente, todo lo explicado anteriormente tiene un rango de aplicabilidad muy amplio, y no se restringe únicamente al estudio de epidemias. Hay otros campos donde se usan métodos similares, como por ejemplo en biología, donde se usan redes biomecánicas para intentar entender los complejos procesos químicos que tienen lugar en las células e incluso descubrir nuevas terapias para tratar enfermedades. También se usan estos métodos para estudiar internet y así entender mejor cómo fluyen los datos por él o cómo se podría cambiar la red para que funcione mejor. Por nombrar un ejemplo más, también se estudian las redes de contactos sociales para comprender la naturaleza de las interacciones sociales y sus implicaciones en el comercio, la estructura de la sociedad, la difusión de información, etc [6].

En esta sección, en concreto, se estudiará un modelo completamente original que sirve para estudiar cómo se extienden un virus informático y un "antivirus" por una red tecnológica, como podría ser la red de internet. Se comenzará explicando el modelo y planteando las reacciones por las que se rige, y, finalmente, se pasará a ver como se comporta el modelo en diferentes tipos de redes.

4.1. Modelo

El modelo se basa en algunos trabajos hechos sobre ciberdefensa activa, donde se introduce en una red en la que se está propagando un virus informático, un white worm que irá protegiendo los dispositivos frente al virus [2], tal como ya se comentó en la introducción. En particular, para este modelo se considerará que tanto el virus como el white worm ven la misma red, por lo que, en principio, ninguno jugaría con ventaja respecto al otro.

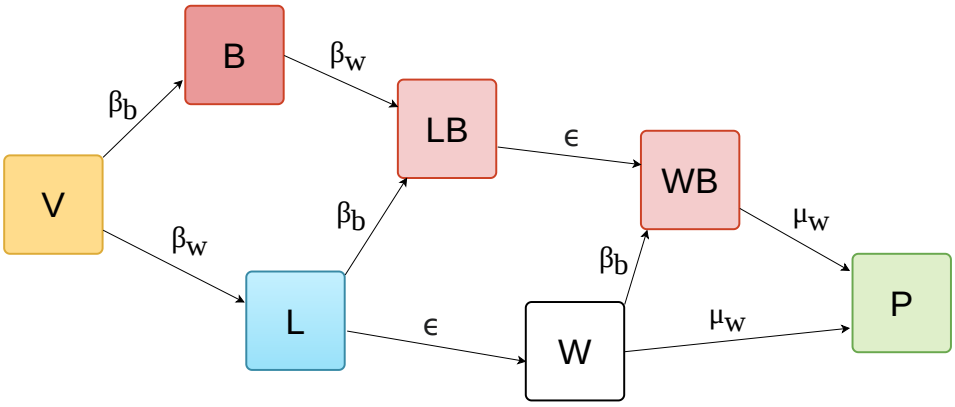


Figura 6: Representación del diagrama de flujo del modelo usado para simular la difusión de un virus informático y un white worm por una red tecnológica.

Las formas que tienen de trabajar los dos gusanos son sencillas. El virus se va propagando por la red a la vez que va intentando infectar al mayor número posible de dispositivos y así crear la *botnet* más grande posible. Una vez que el virus infecta un nodo, este puede infectar a todos los nodos con los que tiene contacto que no estén infectados aún. En contraposición, el white worm también se va propagando por la red y va "infectando" los dispositivos con un antivirus que arreglará las vulnerabilidades que pueda tener. Una vez que el antivirus está en un dispositivo, no actúa directamente, sino que espera un tiempo prudencial, avisando al usuario de que tiene que arreglar algo en su ordenador para protegerlo, por ejemplo. En función de cuánto tiempo pase en el dispositivo sin actuar, se dirá que el white worm es más o menos ético, ya que, a más tiempo sin actuar, más tiempo deja al usuario para arreglar el problema él mismo y no obliga al gusano a hacer modificaciones en el sistema para arreglarlo, y viceversa. Además, durante este periodo el white worm tampoco puede infectar a otros dispositivos. Una vez que el dispositivo queda protegido y ya se ha eliminado al virus, si lo hubiese, el white worm desaparece del sistema.

Se comienza describiendo el modelo compartimental. Este consta de 7 estados y 17 transiciones, cada una de ellas caracterizada por su respectivo ritmo de transición. Los estados son los siguientes:

- Vulnerable (V): los individuos en este estado son aquellos que no están infectados por ninguno de los dos gusanos, pero si tienen contacto con algún individuo que sí lo esté, pueden infectarse.
- Infectado negro (B): estos son los individuos que están infectados por el virus informático y pueden contagiar a otros que no lo tengan.
- Latente (L): estos individuos están infectados por el white worm, pero este aún no ha comenzado a actuar y está esperando a que el usuario arregle el problema él mismo. Los dispositivos en este estado no pueden infectar otros dispositivos que aún no tengan el white worm.
- Infectado blanco (W): cuando un individuo está infectado por el white worm y este está tratando de proteger el dispositivo, está en este estado. Los dispositivos en este estado sí pueden infectar otros dispositivos que aún no tengan el white worm.

- Infectado negro-latente (LB): estos son los individuos que han sido infectados por los dos gusanos, pero el white worm aún no ha comenzado actuar. Sería una mezcla de los estados B y L, por lo que el white worm no podría infectar otros individuos pero el virus sí.
- Infectado negro-blanco (WB): los individuos en este estado están infectados por los dos gusanos y el white worm está tratando de eliminar al virus del sistema para dejarlo protegido. Sería una mezcla de los estados B y W, por lo que tanto el white worm como el virus pueden infectar otros individuos.
- Protegido (P): estos son los individuos que ya no tienen ninguna vulnerabilidad en su sistema por lo que no tienen ninguno de los dos gusanos y no pueden ser infectados por ninguno de ellos.

Por otro lado, las transiciones que pueden ocurrir son

$$V + B \xrightarrow{\beta_b} 2B, \quad V + LB \xrightarrow{\beta_b} B + LB, \quad V + WB \xrightarrow{\beta_b} B + WB, \quad L + B \xrightarrow{\beta_b} LB + B, \quad L + LB \xrightarrow{\beta_b} 2LB, \quad L + WB \xrightarrow{\beta_b} LB + WB,$$

$$W + B \xrightarrow{\beta_b} WB + B, \quad W + LB \xrightarrow{\beta_b} WB + LB, \quad W + WB \xrightarrow{\beta_b} 2WB, \quad V + W \xrightarrow{\beta_w} L + W, \quad V + WB \xrightarrow{\beta_w} L + WB, \quad B + W \xrightarrow{\beta_w} LB + W,$$

$$B + WB \xrightarrow{\beta_w} LB + WB, \quad L \xrightarrow{\epsilon} W, \quad LB \xrightarrow{\epsilon} WB, \quad W \xrightarrow{\mu_w} P, \quad WB \xrightarrow{\mu_w} P.$$

Las nueve primeras se corresponden con la infección de un nodo por el virus informático. Como es obvio, solo pueden infectarse los nodos en los estados V, L y W, que son los que aún no lo tienen, y solo pueden infectar los nodos en los estados B, LB y WB, es decir, los que están infectados. Estas transiciones están controladas por el parámetro β_b, que es el ritmo de transmisión del virus. Después, las cuatro siguientes se corresponden con los procesos de contagio del white worm. Los únicos nodos que se podrán infectar son los que estén en los estados V y B, mientras que los que podrán infectar son los que estén en los estados W y WB. El parámetro de control en estas transiciones es β_w, que es el ritmo de transmisión del white worm. Las dos siguientes transiciones dan cuenta de los cambios entre que el white worm no actúe y se ponga a actuar. Solo se pueden dar desde estados en los que el white worm está latente en el individuo, es decir, L y LB. Están caracterizadas por el parámetro ϵ, que mediría la eticidad del white worm, de forma que cuanto más valga, menos ético será. Las dos últimas transiciones representan las protecciones de individuos debidas a la acción del white worm. Solo se pueden dar desde estados donde el white worm está de forma activa en el individuo, es decir, W y WB. Las controla el parámetro μ_w, que sería el ritmo de protección del white worm.

En la Figura 6 se puede ver el diagrama de flujo del modelo, con los diferentes estados y las transiciones entre estados que puede haber. Cabe destacar que teóricamente tendría que existir una transición más desde cada estado al estado protegido. Estas transiciones representarían la posibilidad de que un sistema arregla sus vulnerabilidades mediante una actualización de *software*, por ejemplo, y estarían caracterizadas por un parámetro μ_u. Sin embargo, se asume que la probabilidad de que ocurra algo así es muy pequeña, es decir, $\mu_u \ll 1$, por lo que estas transiciones son despreciadas desde el principio y no apa-

recen en ningún sitio. Otro aspecto interesante del modelo es que, debido a cómo son los estados y las transiciones que puede haber, existen tres estados absorbentes: V, B y P, por lo que cuando el sistema evolucione hasta un estado estacionario, solo se espera ver nodos en estos estados.

Por otro lado, siguiendo un procedimiento análogo al hecho para el modelo SIS en la Sección 3.3, se puede llegar al sistema de ecuaciones que gobierna la evolución del modelo en la aproximación DBMF para el caso no correlacionado

$$\frac{d\rho_k^V}{dt} = -\beta_b k \rho_k^V \Theta_b - \beta_w k \rho_k^V \Theta_w, \tag{15a}$$

$$\frac{d\rho_k^B}{dt} = \beta_b k \rho_k^V \Theta_b - \beta_w k \rho_k^B \Theta_w, \tag{15b}$$

$$\frac{d\rho_k^L}{dt} = \beta_w k \rho_k^V \Theta_w - \beta_b k \rho_k^L \Theta_b - \epsilon \rho_k^L, \tag{15c}$$

$$\frac{d\rho_k^{LB}}{dt} = \beta_w k \rho_k^B \Theta_w + \beta_b k \rho_k^L \Theta_b - \epsilon \rho_k^{LB}, \tag{15d}$$

$$\frac{d\rho_k^{WB}}{dt} = \epsilon \rho_k^{LB} + \beta_b k \rho_k^W \Theta_b - \mu_w \rho_k^{WB}, \tag{15e}$$

$$\frac{d\rho_k^W}{dt} = \epsilon \rho_k^L - \beta_b k \rho_k^W \Theta_b - \mu_w \rho_k^W, \tag{15f}$$

$$\frac{d\rho_k^P}{dt} = \mu_w \rho_k^W + \mu_w \rho_k^{WB}, \tag{15g}$$

donde

$$\Theta_w = \sum_{k'} \frac{k' P(k')}{\langle k \rangle} \left(\rho_{k'}^W + \rho_{k'}^{WB} \right) \quad \text{y} \quad \Theta_b = \sum_{k'} \frac{k' P(k')}{\langle k \rangle} \left(\rho_{k'}^B + \rho_{k'}^{LB} + \rho_{k'}^{WB} \right), \tag{16}$$

que son las probabilidades de encontrar un nodo infectado por el white worm o por el virus tomando un vértice aleatorio, respectivamente. El sistema se puede adimensionalizar definiendo las cantidades $\tau = \mu_w t$, $\lambda_w = \beta_w / \mu_w$, $\lambda_b = \beta_b / \mu_w$ y $\epsilon' = \epsilon / \mu_w$. Así queda

$$\frac{d\rho_k^V}{d\tau} = -\lambda_b k \rho_k^V \Theta_b - \lambda_w k \rho_k^V \Theta_w, \tag{17a}$$

$$\frac{d\rho_k^B}{d\tau} = \lambda_b k \rho_k^V \Theta_b - \lambda_w k \rho_k^B \Theta_w, \tag{17b}$$

$$\frac{d\rho_k^L}{d\tau} = \lambda_w k \rho_k^V \Theta_w - \lambda_b k \rho_k^L \Theta_b - \epsilon' \rho_k^L, \tag{17c}$$

$$\frac{d\rho_k^{LB}}{d\tau} = \lambda_w k \rho_k^B \Theta_w + \lambda_b k \rho_k^L \Theta_b - \epsilon' \rho_k^{LB}, \tag{17d}$$

$$\frac{d\rho_k^{WB}}{d\tau} = \epsilon' \rho_k^{LB} + \lambda_b k \rho_k^W \Theta_b - \rho_k^{WB}, \tag{17e}$$

$$\frac{d\rho_k^W}{d\tau} = \epsilon' \rho_k^L - \lambda_b k \rho_k^W \Theta_b - \rho_k^W, \tag{17f}$$

$$\frac{d\rho_k^P}{d\tau} = \rho_k^W + \rho_k^{WB}. \tag{17g}$$

Esto deja el sistema en función de solo tres parámetros y permite estudiarlo según el valor del cociente λ_w / λ_b, de forma que si es mayor que uno el white worm se propaga más fácilmente que el virus y si es menor al revés, y según el valor de ϵ', es decir, la eticidad del white worm.

4.2. Redes aleatorias

Ahora que ya se ha introducido el modelo y se han explicado todas las características necesarias sobre él, se pasa a ver cómo se comporta en diferentes tipos de redes. En esta sección se estudiará su comportamiento en redes aleatorias de Erdös-Rényi.

Los resultados que se sacarán son dos mapas de calor en función de las cantidades mencionadas en la sección anterior, uno de ellos representará la fracción total de infectados por el virus en el estado estacionario y el otro representará igualmente la fracción de infectados por el virus, pero será el valor máximo que ha alcanzado a lo largo de la evolución. El interés de este segundo mapa es debido a que es importante saber si en el algún momento de la evolución se ha creado una *botnet* tan grande que era capaz de realizar un ataque muy poderoso. De poco serviría que al final de la evolución quedasen muchos dispositivos protegidos si aun así el virus ha sido capaz de infectar muchos en algún momento. La fracción de infectados por el virus está representada por la cantidad $\rho^B + \rho^{LB} + \rho^{WB}$, ya que todos los individuos que pertenezcan a alguno de estos tres estados pueden infectar a otros individuos del virus.

Para realizar las simulaciones se ha tomado una red de grado medio $\langle k \rangle = 10$ con $N = 10000$ nodos y

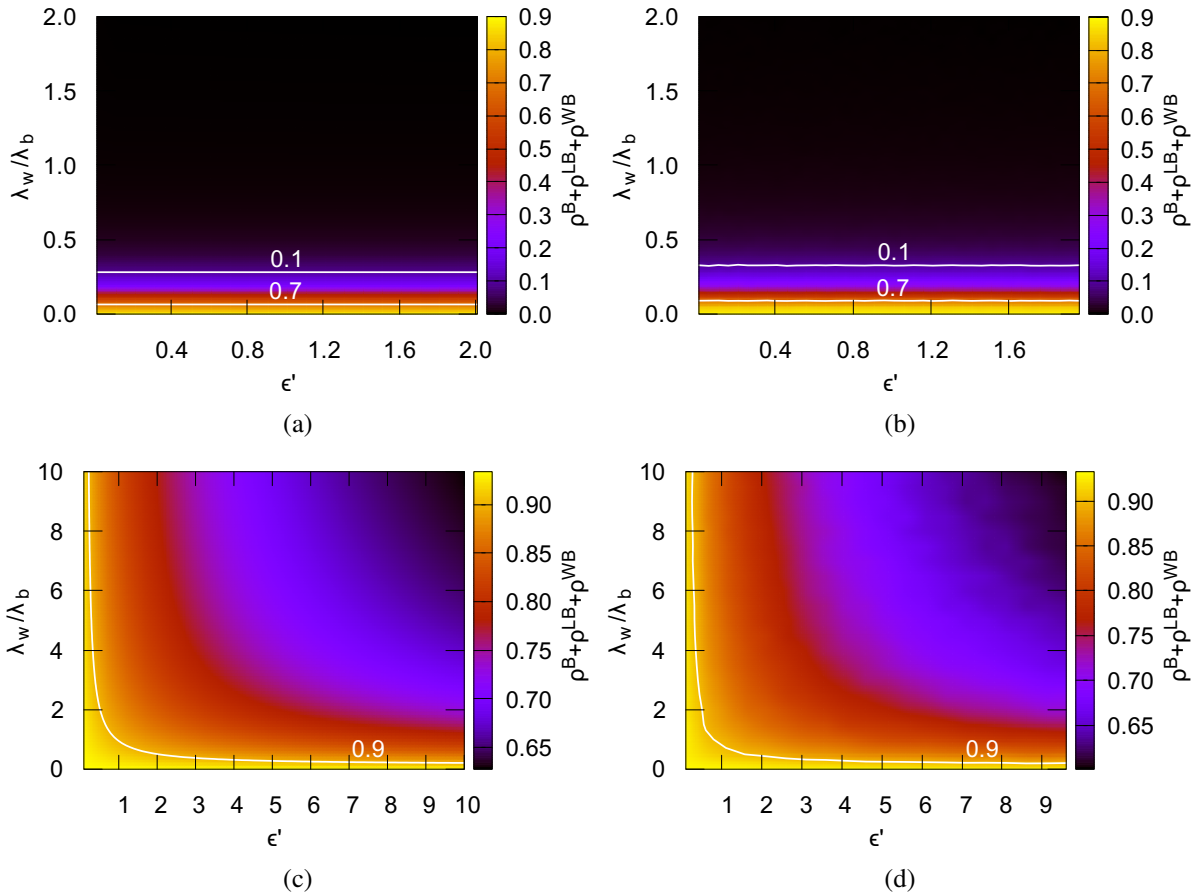


Figura 7: Mapas de calor obtenidos con la red aleatoria de $\langle k \rangle = 10$ para la fracción de infectados final con (a) Runge-Kutta y con (b) Gillespie, y los obtenidos para la fracción de infectados máxima con (c) Runge-Kutta y con (d) Gillespie. Las líneas representan una fracción de infectados constante.

$p = 0.001$. Las condiciones iniciales que se han tomado para las poblaciones de estados son $\rho^V = 0.8$, $\rho^B = 0.1$ y $\rho^W = 0.1$, es decir, al principio hay un 10 % de nodos infectados del white worm y otro 10 % del virus, el 80 % restante está en el estado vulnerable. Que haya un 10 % de nodos en el estado W al principio va a implicar que en el estado final nunca vaya a haber menos de un 10 % de nodos protegidos. Los resultados son obtenidos resolviendo el sistema de ecuaciones (17) con Runge-Kutta y, también, usando el algoritmo de Gillespie. Esto permitirá comparar ambos métodos.

En la Figura 7 se pueden ver los resultados obtenidos. Como se puede ver, hay mucha concordancia entre los resultados obtenidos para Runge-Kutta y los de Gillespie. Respecto a la fracción de infectados final, se ve como esta no depende de la eticidad del white worm y que se consigue erradicar el virus incluso aunque su ritmo de contagio sea mayor al del white worm. Se ve gracias a las isolíneas que la fracción final se anula muy rápidamente al aumentar el ritmo de contagio del white worm respecto al del virus. Después, en los mapas de la fracción máxima se observa un comportamiento distinto. Cuanto mayor sea la eticidad, menor es la fracción máxima, algo que era de esperar. Sin embargo, cuando la eticidad es muy baja, hay un pico de infectados para cualquier valor de λ_w/λ_b, ya que los dispositivos están mucho en el estado latente y dejan al virus propagarse sin problema.

La conclusión que se extrae de estos resultados es que, cuando el grado medio de la red empleada es pequeño, hay buena concordancia entre la aproximación DBMF y el algoritmo de Gillespie, y que los resultados obtenidos para una red aleatoria son los que cabría esperar desde un principio.

4.3. Redes libres de escala

En esta última sección se analizarán los resultados obtenidos al usar redes libres de escala para realizar las simulaciones. Se usarán una red Barabási-Albert y la red Internet AS graph mencionada en la Sección 2.2. La red Barabási-Albert se genera con los valores de parámetros $m_0 = 50$, $m = 2$ y $T = 9950$. Estos valores dan una red con un grado medio $\langle k \rangle \simeq 4.2$. Es decir, tanto la red BA como la red Internet AS graph tienen grados medios bastante bajos (como ya se dijo, $\langle k \rangle = 4.22$ para la red Internet AS

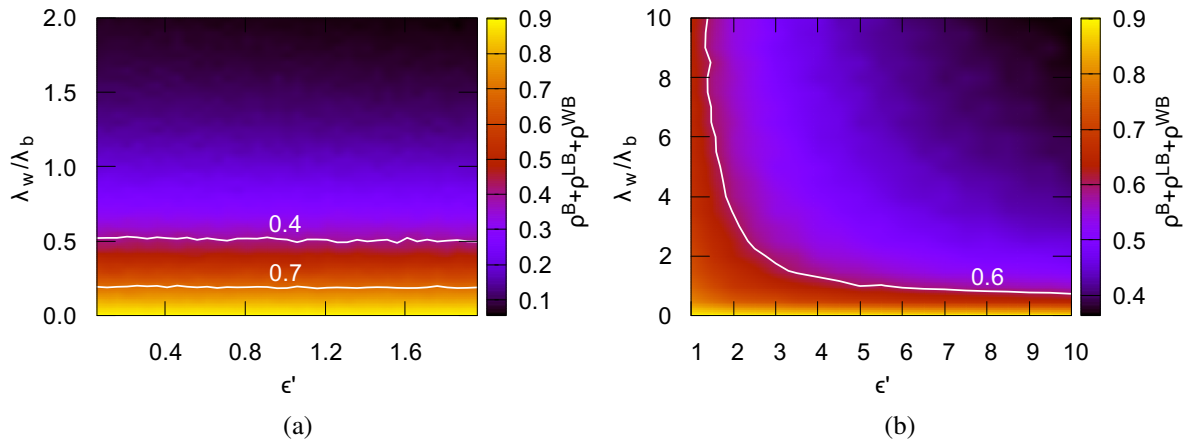


Figura 8: Mapas de calor obtenidos con la red Barabási-Albert para (a) la fracción de infectados final con Gillespie y para (b) la fracción de infectados máxima con Gillespie. Las líneas representan una fracción de infectados constante.

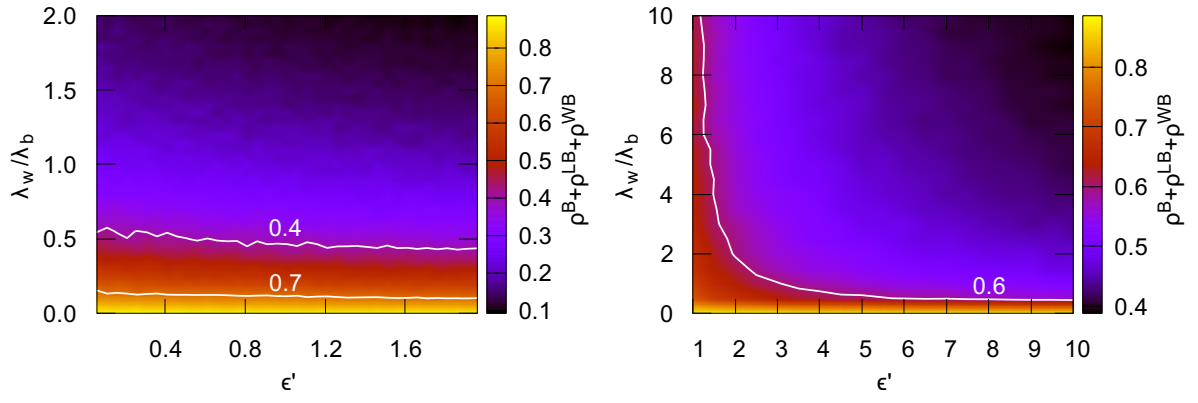
graph), por lo que debido a las conclusiones sacadas en el apartado anterior, se espera que los resultados obtenidos con la aproximación DBMF y con Gillespie sean muy parecidos. Por esta razón, solo se presentarán los resultados del algoritmo de Gillespie. En cuanto a las condiciones iniciales, son las mismas que en el apartado anterior.

Se comienza por la red Barabási-Albert. En la Figura 8 se observan los resultados y, como se puede ver, los comportamientos son bastante parecidos a los obtenidos para redes aleatorias, aunque hay diferencias notables en los resultados cuantitativos. Por tanto, comparando con los mapas de la Figura 7 se sacan las siguientes conclusiones.

En cuanto a la cantidad de infectados por el virus finales, que se puede ver en la Figura 8a, se obtiene que hay una cantidad no despreciable para un rango más amplio de valores de λ_w/λ_b, llegando incluso a no anularse cuando $\lambda_w/\lambda_b = 2$. A diferencia de la red aleatoria, que cuando $\lambda_w/\lambda_b = 0.5$ ya era prácticamente nula la densidad de infectados finales. Por otro lado, la dependencia de esta densidad con la eticidad ϵ' es igual en las dos redes, no existiendo ninguna influencia de este parámetro.

Pasando ahora a la fracción de infectados máxima, la cual se aprecia en la Figura 8b, se obtiene un comportamiento parecido, aunque ahora los valores se han reducido en comparación con los resultados de la red aleatoria. La densidad de infectados máxima ha disminuido en todo el rango de parámetros representado, por lo que, por ejemplo, no sería necesario que la eticidad del white worm fuese extremadamente baja para conseguir que la *botnet* no se hiciese muy grande en algún momento intermedio de la evolución.

Con la red Internet AS graph, cuyos resultados se pueden observar en la Figura 9, se obtienen resultados prácticamente análogos a los de la red BA, coincidiendo los valores de densidad de infectados, tanto finales como máximos, en los dos rangos de parámetros empleados. Por tanto, el análisis que se puede hacer es parecido para estas dos redes libres de escala. Como se puede ver, la leve disasortatividad que tiene la red Internet AS graph no ha influido para nada en los resultados.



(a) (b)

Figura 9: Mapas de calor obtenidos con la red Internet AS graph para (a) la fracción de infectados final con Gillespie y para (b) la fracción de infectados máxima con Gillespie. Las líneas representan una fracción de infectados constante.

En conclusión, las propiedades topológicas que tienen las redes libres de escala frente a las aleatorias posibilitan que la propagación del virus por la red no sea tan extendida, permitiendo al white worm proteger una gran cantidad de individuos sin que puede haber consecuencias fatales.

Como la mayoría de redes tecnológicas que existen actualmente son libres de escalas, muy parecidas a las analizadas en este trabajo, se deduce que para proteger una red real en la que hubiese un virus propagándose, podría usarse un white worm como el descrito. No sería necesario que la eticidad de este método estuviese muy comprometida, ya que podría hacerse un gusano muy ético pero con una capacidad de propagación por la red mucho mayor que el virus. Esto no es algo descabellado, ya que se supone que el white worm se introduce en la red con la ayuda de los encargados de administrarla, por lo que tienen un conocimiento mucho mayor de las propiedades de la red que los atacantes que han introducido el virus.

5. Conclusiones

A lo largo de este trabajo se han estudiado distintos aspectos de las redes, así como se ha presentado un modelo totalmente original para el estudio de protección de redes frente a ciberataques.

En la Sección 2 se presentaron las características más importantes que ayudan a determinar una red compleja. También se habló de diferentes modelos para producir redes sintéticas que comparten algunas características con las redes reales, como la distribución de grado en el modelo Barabási-Albert, y que permiten el estudio de ciertos comportamientos sin la necesidad de emplear redes reales, las cuales puede ser que no estén disponibles por diversas razones.

Después, a lo largo de la Sección 3 se presentaron las consideraciones principales que se realizan a la hora de estudiar la propagación de epidemias en poblaciones, debido a la estrecha relación que hay entre esto y la propagación de virus informáticos por redes tecnológicas. También se habló de algunas aproximaciones que se hacen para poder resolver los modelos epidemiológicos, tanto cuando se quiere abordar el problema resolviendo las ecuaciones diferenciales que describen el modelo (aproximación DBMF), como cuando se quiere estudiar la evolución directamente sobre una red compleja (algoritmo de Gillespie).

Finalmente, en la Sección 4 se describió el modelo empleado para simular la evolución de un virus informático por una red compleja cuando también hay presencia de un white worm que intenta acabar con él. Se vieron los resultados sobre redes aleatorias y libres de escalas y se analizaron sus diferencias. En concreto, se ha empleado una red real tecnológica que representa muy bien una situación típica en la que se podría aplicar esta estrategia de seguridad.

Todos los programas desarrollados durante la realización de este trabajo para hacer las simulaciones, las redes, etc. están disponibles en [25].

La principal conclusión que se extrae del estudio realizado con el modelo mencionado es que, debido a las propiedades que tienen las redes tecnológicas reales, que las hacen comportarse como redes libres de escala, es posible el desarrollo de un modelo de defensa activa. Este modelo usaría un white worm para proteger la red, sin que la eticidad se vea expuesta, algo muy importante debido a la gran cantidad de datos que hay moviéndose por cualquier red de internet actualmente.

Para concluir, sería interesante mencionar algún aspecto que ha quedado por estudiar sobre este modelo y que podría servir como punto de partida para trabajos posteriores. El más destacado es la inclusión de redes multicapa en el modelo. Esto permitiría representar la red tecnológica como una red de dos capas, en la que cada una de ellas sería la red que es capaz de ver cada gusano, ya que no tendrían porqué ver la misma red de interconexiones porque utilizan vulnerabilidades distintas de los dispositivos para propagarse.

Bibliografía

[1] Alberto Aleta. *Networks, Epidemics and Collective Behavior: from Physics to Data Science*. Tésis doct., Universidad de Zaragoza, 2020.

[2] Wenlian Lu, Shouhuai Xu y Xinlei Yi. Optimizing Active Cyber Defense. En *Decision and Game Theory for Security*, págs. 206–225. Springer International Publishing, 2013.

[3] Giovanni Ferronato. *IoT White Worms: design and application*. TFM, Universidad de Twente, 2020.

[4] Michele De Donno, Nicola Dragoni, Alberto Giaretta y Manuel Mazzara. AntibIoTic: Protecting IoT Devices Against DDoS Attacks. En *Proceedings of 5th International Conference in Software Engineering for Defence Applications*, págs. 59–72. Springer International Publishing, 2018.

[5] Michele De Donno y Nicola Dragoni. Combining AntibIoTic with Fog Computing: AntibIoTic 2.0. En *2019 IEEE 3rd International Conference on Fog and Edge Computing (ICFEC)*, págs. 1–6. IEEE, 2019.

[6] Mark Newman. *Networks: An Introduction*. Oxford Scholarship Online, 2018.

[7] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez y D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, 2006.

[8] Marián Boguñá y Romualdo Pastor-Satorras. Epidemic spreading in correlated complex networks. *Physical Review E*, 66(4):047104, 2002.

[9] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem y Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of Modern Physics*, 87(3):925–979, 2015.

[10] P. Erdös y A. Rényi. On Random Graphs I. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[11] Mark Newman. Assortative Mixing in Networks. *Physical Review Letters*, 89(20):208701, 2002.

[12] Réka Albert, Hawoong Jeong y Albert-László Barabási. Diameter of the World-Wide Web. *Nature*, 401(6749):130–131, 1999.

[13] Albert-László Barabási y Réka Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, 1999.

[14] S. N. Dorogovtsev, J. F. F. Mendes y A. N. Samukhin. Structure of Growing Networks with Preferential Linking. *Physical Review Letters*, 85(21):4633–4636, 2000.

[15] Romualdo Pastor-Satorras y Alessandro Vespignani. *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, 2004.

[16] Tiago de Paula Peixoto. Netzschleuder: the network catalogue, repository and centrifuge. Accedido: 03-03-2021.

[17] Matthew James Keeling y Pejman Rohani Princeton. Modeling Infectious Diseases in Humans and Animals. *Clinical Infectious Diseases*, 47(6):864–865, 2008.

[18] William Ogilvy Kermack y A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A*, 115(772):700–721, 1927.

[19] Roy M. Anderson y Robert M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1992.

[20] Alain Barrat, Marc Barthélemy y Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

[21] István Z. Kiss, Joel C. Miller y Péter L. Simon. *Mathematics of Epidemics on Networks: From Exact to Approximate Models*. Springer International Publishing, 2017.

[22] Daniel T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, 1976.

[23] Peter G. Fennell, Sergey Melnik y James P. Gleeson. Limitations of discrete-time approaches to continuous-time contagion dynamics. *Physical Review E*, 94(5):052125, 2016.

[24] Chao-Ran Cai, Zhi-Xi Wu, Michael Z. Q. Chen, Petter Holme y Jian-Yue Guan. Solving the Dynamic Correlation Problem of the Susceptible-Infected-Susceptible Model on Networks. *Physical Review Letters*, 116(25):258301, 2016.

[25] Miguel Tarancón. Repositorio con los scripts empleados en el TFG.