FACULTAD DE INFORMÁTICA

ANÁLISIS DE REDES SOCIALES

PROYECTO FINAL

ANÁLISIS DE REDES DE MARVEL

GRUPO 09

AUTORES:

Lorena Jiménez Corta
Nerea Jiménez González
Bruno Torralbo Fernández
Rayner Tan Luc

Índice

1. Introducción

El objetivo general de este proyecto consiste en comparar las películas de Marvel con los cómics.

Para ello vamos a obtener los datos necesarios para posteriormente analizarlos e interpretarlos para obtener una buena conclusión o una solución aproximada al problema de los fans de la compañía: saber cuánto se asemeja la realidad de las películas a la ficción de los cómics.

2. Obtención de datos

En esta sección explicaremos como hemos conseguido recopilar los datos para el estudio de los grafos y las herramientas empleadas.

2.1 Películas

Para obtener la información de los personajes de las películas realizamos los siguientes pasos:

1. En primer lugar, seleccionamos las películas que queríamos analizar para realizar el estudio. Todas ellas se encuentran en el archivo *movies.pdf* adjunto con la entrega de la práctica.
2. Como idea inicial para obtener los personajes, pensamos en visualizar las películas y analizar los personajes de cada una de ellas. Concluimos con que no era buena práctica, ya que había un alto porcentaje de error a la hora de obtener los datos. Además, era necesario automatizar el proceso.
3. Con este motivo, decidimos hacer uso de la base de datos de "IMDb" para lograr nuestro objetivo. Una vez realizado el código, descubrimos que no era posible recoger la información necesaria.
4. Como alternativa, conseguimos obtener los datos de la base de datos de "theMoviedb".

Una vez hallada la manera de obtener la información, con ayuda de *python y java* generamos el código para obtener los resultados necesarios.

- *moviesIMDb.py*: Corresponde a nuestro primer intento de obtener datos con "IMDb". Se trata de un código que "no funciona", ya que mostramos como ejemplo que ni si quiera encuentra la película que buscamos. En caso de encontrarla, no incluye la información que precisamos. En la bibliografía incluimos un enlace en el que indica el mal estado de la base de datos.
- *searchMovies.py*: Se trata de un ejecutable de apoyo en el que introduces el título de la película que quieres buscar. El resultado es una lista de todas las películas relacionadas con nuestra búsqueda, mostrando el título, identificador y la fecha de estreno. Así, conseguimos encontrar los identificadores de todas las películas de nuestro estudio.
- *movies.py*: En este fichero, generamos los archivos *movies.csv* y *moviesNodes.csv.* El primero está formado por una cabecera "movie, characters", en el que, siguiendo la idea del fichero de *The Simpsons* de la práctica 1, incluimos la película seguida de todos los personajes que aparecen en ella separados por dos puntos. En el segundo, la cabecera contiene los campos "id, character", donde guardamos todos los personajes de las películas asignándoles un identificador.

NOTA: la película de Thor 2 solo alberga dos personajes (son los únicos almacenados en la base de datos).

- *RecogidaDatos.java:* Es el programa escrito en Java que recoge los datos del archivo mencionado anteriormente, el cual recoge la lista de personajes de cada película del archivo y va añadiendo una relación por cada pareja de personajes de la lista, y en el caso de que esa relación ya exista, aumenta un contador interno que se corresponde al peso de la arista de la relación en cuestión.

2.2 Cómics

Para conseguir la información que precisamos de los cómics, encontramos un fichero en el que se relacionaban los personajes con el cómic en el que aparecían. Una vez más, en la bibliografía indicamos el enlace donde lo encontramos. El archivo obtenido se encuentra al final de la página, en el enlace "*Cleaned source file*". De esta manera, partimos de la misma base de las películas.

Ante esto, nos topamos con un problema. Desconocíamos los cómics que se correspondían con las películas seleccionadas. Para solventarlo, decidimos apoyarnos en los ficheros de las películas que ya habíamos generado.

1. En primer lugar, con ayuda del fichero "*moviesNodes.csv*" filtramos todos los cómics en los que aparecían nuestros personajes.
2. Nos encontramos con otro problema. El formato en el que se guardan los personajes de las películas no es el mismo que el de los cómics. De esta manera, podemos guardar dos nodos que representen al mismo personaje. Por ejemplo, podemos guardar "*Tony Stark*" y "*Iron man*" en distintos nodos, cuando en realidad son el mismo personaje. Como solución, decidimos crear un diccionario en el que a partir de un único nombre, se recogieran todos los relacionados con él, de esta manera, nos aseguramos no repetir personajes.
3. Una vez recogidos todos los cómics, realizamos un segundo barrido. Esta vez, a partir de un cómic recopilamos todos sus personajes. De esta manera, generamos dos ficheros, uno que relaciona el cómic con los personajes que aparecen en él y otro que recoge todos los personajes de los cómics.
4. Al igual que en las películas, se utiliza el programa RecogidaDatos.java para recoger la lista de personajes que aparecen en cada cómic para relacionarlos entre sí.

2.3 Nodos y aristas

Los nodos representan a los personajes que aparecen en las películas y cómics respectivamente. En cuanto a las aristas, representan las relaciones entre los personajes. Dos personajes están relacionados si aparecen en la misma película o cómic.

3. Análisis de datos

En este apartado, se va a explicar todo el proceso del análisis de los datos obtenidos de las películas y de los cómics.

3.1 Objetivos del análisis

El objetivo principal es comprobar las redes obtenidas a raíz de procesar los archivos correspondientes a los cómics y a las películas, para determinar dónde aparecen más personajes, dónde aparecen más veces determinadas parejas de personajes o dónde aparecen más héroes que personajes secundarios.

3.2 Descripción de las métricas o algoritmos aplicados

Para poder llegar a analizar las redes, primero debemos obtener los archivos necesarios para tratarlos con Gephi, y para ello, se ha tenido que filtrar los nombres que aparecen en los cómics y en las películas, para intentar tener los mismos nodos principales en ambos grafos, sin importar los personajes secundarios de las películas o el resto de los héroes que aparezcan en los cómics.

Además, vamos a analizar las propiedades de cada red en base a las siguientes métricas:

- Número de aristas totales de la red.
- Densidad de la red.
- Tamaño de los hubs más grandes.
- Distancia media.
- Coeficiente de agrupamiento.

Analizaremos primero la red obtenida de las películas, todo con los datos obtenidos de Gephi.

En total, tiene 1479 nodos y 41970 aristas. Su densidad es de 0.038, lo cual significa que la red no está prácticamente conectada, excepto por varios nodos que se explicarán a continuación.

Los hubs más grandes son *"Helicopter Pilot"*, *"Tony Stark / Iron Man"* y *"Peter Parker / Spider-Man"*, con grado 369, 367 y 366 respectivamente, es decir, son los personajes que más personajes tienen relacionados, ya sean principales o secundarios.

La distancia media entre nodos es de 2.799, es decir, que cada pareja de nodos está separada entre sí por casi 3 nodos de media.

Para acabar, el coeficiente de agrupamiento de la red de películas es de 0.956, un valor bastante elevado, que significa que prácticamente todos los nodos están relacionados con sus nodos vecinos.

A continuación, vamos a analizar la red obtenida de los cómics, que igual que la red anterior, se analizan los datos obtenidos de Gephi.

Tiene un total de 6421 nodos y 171644 aristas, aproximadamente 4 veces la red de películas, con una densidad de 0.008, es decir, que la red no tiene prácticamente conexión alguna.

Los hubs más grandes son "*Captain America"* con grado 1919, "*Spider-Man / Peter Parker"* con 1754 y *"Iron Man / Tony Stark"*, con 1566.

La distancia media entre nodos es de 2.63, es decir, cada par de nodos están separados por casi 3 nodos de media.

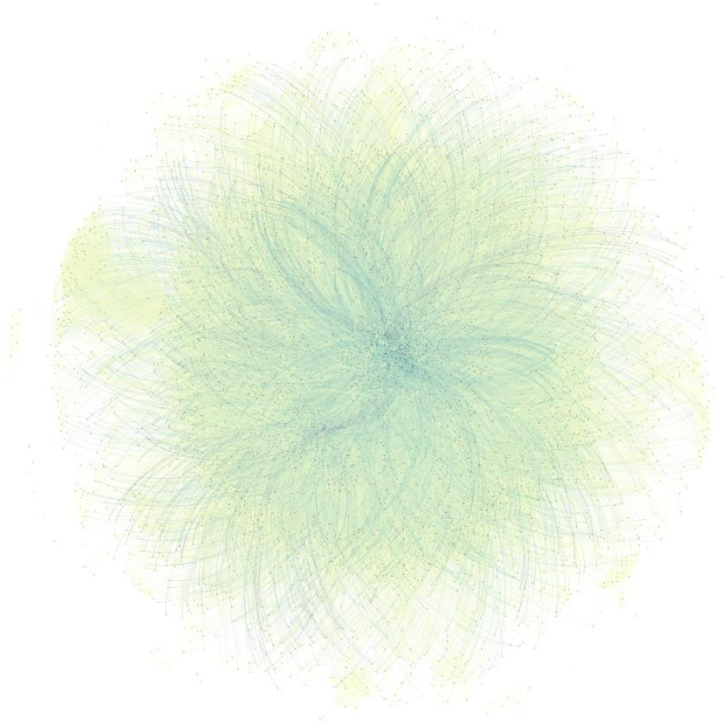
Y, por último, el coeficiente de agrupamiento de la red de cómics es de 0.78, que se traduce en que muchos de los nodos de la red están relacionados con sus nodos vecinos.

Como casos excepcionales, hay 18 nodos que no tienen grado, debido a que son cómics en los cuales sólo aparece un personaje.

3.3 Visualizaciones de los resultados obtenidos

La red anterior es la resultante del archivo de las películas, donde se pueden ver que los nodos están agrupados en pequeños conjuntos, correspondientes a cada película o saga de películas, uniéndose su mayoría por personajes comunes o por Stan-Lee, fundador y director de Marvel.

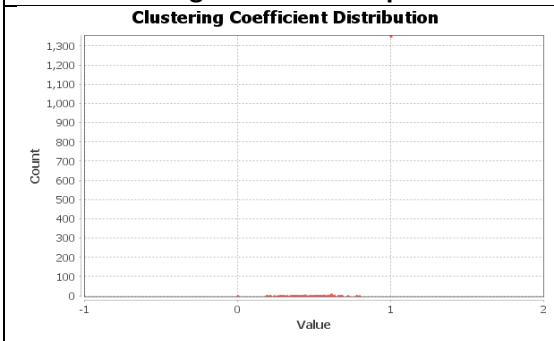
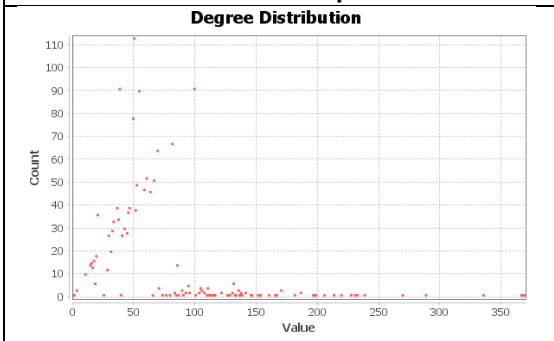
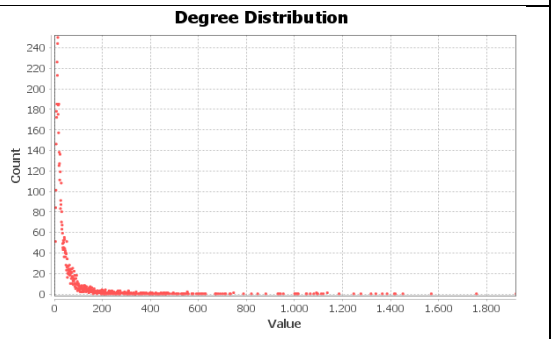
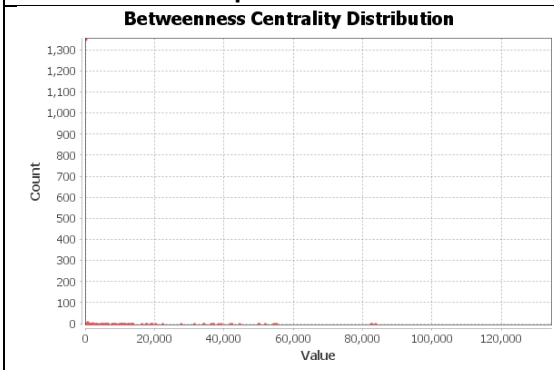
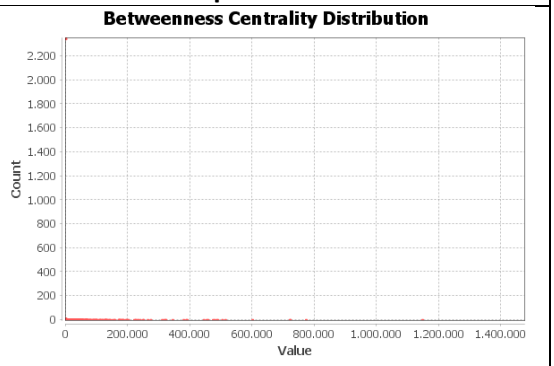
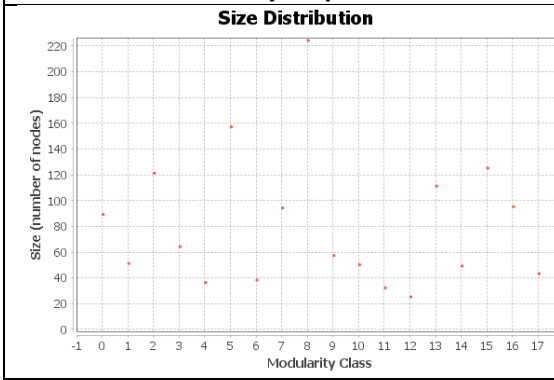
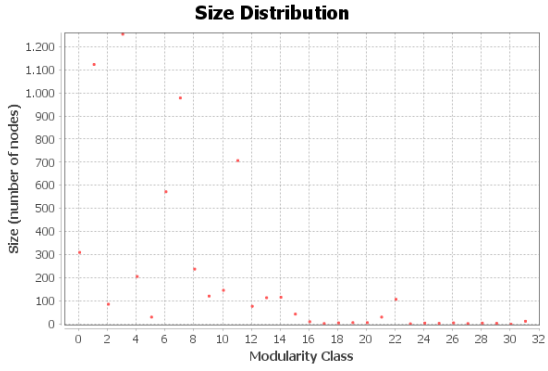
Las redes siguientes se corresponden al archivo de los cómics, donde se aprecia que la gran mayoría de personajes se encuentran alrededor de ciertos personajes centrales, aunque cada personaje tiene su propio conjunto de nodos (o *universo*, usando la terminología de Marvel).



Fruchterman Reingold

OpenOrd

A continuación, vamos a comparar varias de las gráficas obtenidas de cada red:

Clustering Coefficient de películas	Clustering Coefficient de cómics
	
Grado medio de películas	**Grado medio de cómics**
	
Betweenness Centrality Distribution de películas	**Betweenness Centrality Distribution de películas**
	
Modularity de películas	**Modularity de cómics**
	

A continuación, haremos una comparativa de los datos obtenidos de cada red:

	Nodos	Aristas	Densidad	Grado medio	Componente gigante	Distancia media	Clustering Coefficient
Películas	1479	41970	0.038	56.755	1377 (93.1%)y 40388(96.23%)	2.799	0.956
Comics	6439	171644	0.008	53.314	6403 (99.44%) y 171588 (99.97%)	2.63	0.78
Diferencia*	4.35	4.089	4.75	1.06	4.65 y 4.25	1.06	1.225

*Nota: El resultado de las diferencias se ha realizado mediante la división del valor más alto entre el valor más bajo de cada columna.

3.4 Descripción de las tareas realizadas durante el análisis de datos

En ambas redes, se ha utilizado la distribución Fruchterman Reingold, con la diferencia de que en la red de películas tiene por valores en Área = 10000.0, Gravedad = 10.0 y Velocidad = 10.0, mientras que en el caso de la red de cómics, al ser una red bastante mayor a la anterior, se han utilizado como valores en Área = 1000000.0, Gravedad = 4.0 y Velocidad = 10.0, para intentar obtener una visualización clara de los nodos, aunque no ha sido posible debido al volumen de nodos y aristas de la red. Debido a esto, hemos utilizado otra distribución, OpenOrd, usando los valores Liquid(%) = 50 y Expansion(%) = 50, y el resto de valores predeterminados.

Para obtener todos los datos necesarios para el análisis, se han utilizado los obtenidos de Gephi en vez de usar el modelo teórico.

4. Interpretación de los datos

En este apartado, vamos a proceder a interpretar y a comparar los datos obtenidos de ambas redes.

4.1 Interpretación de los resultados obtenidos

Nuestro objetivo principal es la comparación de los datos obtenidos de las películas con los datos obtenidos de los cómics.

En ambas redes se aprecian ciertos nodos con mayor grado, los cuales se corresponden con los personajes principales.

En el caso de la red de las películas, hay varios grupos separados de la red general, los cuales corresponden a películas co-producidas por Marvel, pero que no se incluyen en el universo cinematográfico de la compañía, tales como *El Motorista Fantasma, Los 4 Fantásticos* o *Blade*.

En el caso de la red de los cómics, los grupos que no se conectan con la componente gigante son los personajes que aparecen sin ningún otro personaje en su correspondiente cómic.

4.2 Limitaciones encontradas durante el análisis de los datos

La principal limitación ha sido a la hora de obtener los ficheros de los cómics, debido a que hemos tenido que filtrar muchos datos para poder obtener los personajes que necesitábamos. Aun así, después hemos tenido que volver a filtrar ciertos nodos de los obtenidos por tener algún tipo de carácter incorrecto, lo cual resultaba en nodos aparentemente repetidos, pero que el programa reconocía como nodos diferentes.

4.3 Conclusiones relevantes encontradas durante el análisis

Con los datos recabados en la tabla descrita anteriormente en el apartado de Análisis, podemos observar que, aunque la red de cómics sea 4 veces mayor que la red de películas, el resto de los datos son muy aproximados.

En ambas redes, las comunidades están bien diferenciadas, pudiendo distinguir a qué película o cómic pertenece cada una.

Como se ve en la gráfica del grado medio de cada red en el apartado anterior, observamos que la red de cómics se asemeja a una red libre de escala.

Observando que la red de películas contiene elementos concentradores o hubs, descartamos que pueda ser una red aleatoria y, por tanto, se asemeja más a una red libre de escala.

5. Bibliografía

- Url para el uso de IMDb para la obtención los personajes de las películas:
  https://imdbpy.sourceforge.io/
- Página donde descubrimos que IMDb no funcionaba bien:
  https://github.com/alberanid/imdbpy/issues/103
- Url para la obtención de datos de las películas definitiva con base de datos funcional:
  https://www.themoviedb.org/?language=es
- API para aprender a hacer la llamada rest para obtener los personajes de las películas:
  https://developers.themoviedb.org/3/movies/get-movie-credits
- Url para la obtención de datos de los cómics:
  http://syntagmatic.github.io/exposedata/marvel/