



Universidad
Carlos III de Madrid

TRABAJO FIN DE GRADO

SELECCIÓN ÓPTIMA DE GRUPOS EN UNA RED SOCIAL: APLICACIÓN AL MARKETING VIRAL

Autor: Elena Cerrato Hernández

Tutor: Elisenda Molina Ferragut

Grado: Ingeniería Informática y Administración de Empresas

Madrid, Junio de 2016



RESUMEN

Las redes sociales cada vez están más presentes en el día a día de más personas. Permiten comunicarse con usuarios de cualquier nacionalidad y estar informados de todo lo que ocurre en cualquier rincón del mundo. Pero no solo favorece a los usuarios de las redes, sino que permite a las grandes empresas realizar campañas publicitarias que accedan a un público objetivo inimaginable con los medios de comunicación tradicionales. A este tipo de campañas se las conoce como Marketing viral, y son el pilar fundamental de este documento.

Para poder aprovecharse de todas las ventajas que las redes sociales pueden llegar a ofrecer a las empresas, hay que descubrir qué individuos de la red pueden convencerse inicialmente para difundir la campaña, y conseguir que esta difusión sea la máxima posible.

El problema principal será la búsqueda en una red social de aquellos individuos o grupo de individuos que permitan que la campaña de marketing viral alcance la mayor difusión. Para solucionar este problema se va a modelar la red social como un grafo, y se estudiarán las medidas de selección óptima de grupos en función de la centralidad estructural de cada individuo y en función de su capacidad de difusión de los mensajes de forma dinámica a través de la red.

Se optará por el desarrollo del Modelo de Umbral Lineal debido a que es el que mejor se adapta al entorno del marketing viral. Una vez seleccionado se definirá la función a optimizar donde se tendrán en cuenta los nodos alcanzados y el tiempo de difusión.

Con el modelo elegido y la función definida, sólo quedará desarrollar una serie de funciones que se prueben sobre grafos tanto reales como aleatorios, de tal forma que teniendo en cuenta sus resultados, permitan elegir aquella función que mejores resultados dé en la búsqueda óptima de grupos en la red. El proyecto concluye con varias recomendaciones para su uso en aplicaciones reales.

Palabras clave: Marketing viral; Redes sociales; Grafo; Modelo de Difusión Dinámica; Modelo de Umbral Lineal; Grupos óptimos

SUMMARY

Social networks are continuously increasing their importance on the life of people. They allow us to communicate with user from any nation and be informed of everything going on around the globe. But not only do they favor the users, social networks also offer brand new channels for companies to invest in advertisement campaigns, reaching further than any traditional media. These campaigns are known as Viral Marketing, the basis of this project.

In order to benefit from all the advantages the social networks may present, the companies interested must discern which individuals should initially convince to spread the product/service, thus achieving maximum reach.

The scope of this project will be searching social networks for the group of users who will expand the campaign the furthest possible. Solving this problem will require social network modeling as a graph and studying optimum selection techniques related to both structural centrality and dynamic diffusion in networks.

After reviewing all the possibilities, the Lineal Threshold Model is selected for its better adaptation to the viral marketing environment. Once chosen, the optimization function is defined, taking into account the total number of reached nodes and the time invested for such.

With both model and function set, the functions are coded and a series of simulations are designed and carried out for real and randomly generated graphs. The results enable the selection of the best function for searching of optimal groups in networks. The project concludes with several recommendations for its use in real life scenarios.

Key Words: Viral Marketing; Social Networks; Graph; Dynamic Diffusion Model; Lineal Threshold Model; Optimal Groups

ÍNDICE

1	Introducción y Objetivos.....	1
1.1	¿Qué es el Marketing Viral?.....	1
1.2	Aplicación de Redes Sociales al Marketing Viral	4
2	Redes Sociales y Centralidad Individual	8
2.1	¿Qué es una Red Social?.....	8
2.2	La Red Social como un Grafo	9
2.3	Medidas estructurales de centralidad individual.....	15
2.3.1	Degree	15
2.3.2	Closeness.....	17
2.3.3	Betweenness.....	18
2.3.4	Eigenvector (Bonacich)	19
2.3.5	Page-Rank.....	19
3	Selección óptima de grupos de difusión.....	21
3.1	Modelos dinámicos basados en la Difusión de Innovaciones	21
3.1.1	Modelo General de Umbral (General Threshold Model).....	24
3.1.2	Modelo de Cascada Independiente de Kleinberg	26
3.1.3	Modelo del Votante (Ligget)	28
3.1.4	Selección del Modelo.....	28
3.2	Selección del grupo óptimo.....	29
3.2.1	Descripción de lo que ya hay hecho	31
3.2.2	Estimación de la difusión de un determinado grupo.....	32
3.2.3	Funciones de búsqueda del grupo óptimo	34
3.2.4	Heurísticos	38
3.2.5	Experiencia computacional.....	41
4	Conclusiones.....	52
5	Referencias.....	55
6	Glosario.....	57
7	Anexos.....	58
7.1	Código Fuente	58
7.2	Estudio piloto.....	58
7.3	Medidas de Centralidad Individual	59
7.4	Resultados de Métricas para los ocho grafos de depuración	59
7.5	Pruebas Completas.....	62

ÍNDICE DE ILUSTRACIONES

Ilustración 1. Grafo de Red de Amistades.....	10
Ilustración 2. Grafo de influencia de una red social	10
Ilustración 3. Matriz de Adyacencia de un Grafo	12
Ilustración 4. Matriz de Adyacencia ponderada de un grafo	12
Ilustración 5. Matriz de incidencia.....	13
Ilustración 6. Grafo de ejemplo para camino mínimo	14
Ilustración 7. Grafo con caminos mínimos entre nodos H y D.....	14
Ilustración 8. Grafo Kite para medidas de centralidad	15
Ilustración 9. Difusión en modelo progresivo.....	23
Ilustración 10. Difusión en modelo no progresivo. Ejemplo de no estabilidad...	23
Ilustración 11. Nomenclatura en modelo de Umbral Lineal.....	24
Ilustración 12. Difusión en modelo Progresivo con Umbral Lineal.....	25
Ilustración 13. Modelo de Cascada.	26
Ilustración 14. Difusión en modelo de Cascada.....	27
Ilustración 15. Modelo de Cascada de Kleinberg.....	27
Ilustración 16. Diagrama de flujo de procesos de la función threshold	34
Ilustración 17. Simplificación de función submodular a función convexa	35
Ilustración 18. Grafo Estrella de Estrellas	38
Ilustración 19. Grafo Conexo más Poco Conexo.....	40
Ilustración 20. Grafo Books con los nodos seleccionados en rojo formando el grupo óptimo	45
Ilustración 21. Nodos alcanzados en pruebas a Books	46
Ilustración 22. Tiempos de Difusión para pruebas de Books.....	47
Ilustración 23. Gráfica de Eficiencia de pruebas a Books	49
Ilustración 24. Clustering.....	53

ÍNDICE DE TABLAS

Tabla 1. Tipos de Grafos	11
Tabla 2. Resultados del Degree para el grafo Kite.....	15
Tabla 3. Resultados del Closeness en grafo Kite	17
Tabla 4. Resultados de Betweenness para Kite	18
Tabla 5. Resultados de Eigenvector para grafo Kite	19
Tabla 6. Resultados de Pagerank para Kite.....	20
Tabla 7. Espacio factible de soluciones y tipo de optimización para métricas de influencia	29
Tabla 8: Elección de modelos.....	31
Tabla 9. Pruebas para grafo Books	46
Tabla 10. Tiempos de difusión para pruebas de Books.....	47
Tabla 11. Medidas de Centralidad de Books.....	48
Tabla 12. Resultados de eficiencia de pruebas a Books.....	48
Tabla 13. Resultados de pruebas - Tiempos de Ejecución.....	50

Tabla 14. Resultados de pruebas - Nº de Nodos Alcanzados	50
Tabla 15. Nodos Alcanzados con medidas de Centralidad	51

ÍNDICE DE ECUACIONES

Ecuación 1. Métrica de Centralidad Degree sin adyacencia.....	16
Ecuación 2. Métrica de Centralidad Degree con Adyacencia.....	16
Ecuación 3. Métricas de Centralidad In-Degree y Out- Degree	16
Ecuación 4. Métrica de Centralidad Closeness	17
Ecuación 5. Medida de Centralidad Betweenness.....	18
Ecuación 6. Medida de Centralidad Eigenvector	19
Ecuación 7. Medida de Centralidad Pagerank.....	20
Ecuación 8. Función Biobjetivo a maximizar	30
Ecuación 9. Función Submodular	35
Ecuación 10. Estudios de nº de repeticiones para simulaciones.....	43

PSEUDOCÓDIGOS

Pseudocódigo 1. Threshold	34
Pseudocódigo 2. Threshold Greedy.....	36
Pseudocódigo 3. Threshold Greedy Stepwise	37





1 INTRODUCCIÓN Y OBJETIVOS

1.1 ¿Qué es el Marketing Viral?

Desde finales del pasado siglo hasta la actualidad, las tecnologías llevadas a cabo en los ámbitos de la informática y las telecomunicaciones han supuesto la llegada de la nueva Era de la Información.

Como consecuencia de esta revolución, las personas están expuestas a tal nivel de información que llegan a un estado de saturación en el cual no se puede captar, ni mucho menos retener, todos los datos a los que se tiene acceso. Los números en las redes sociales no engañan, y son un claro ejemplo del exceso de datos que se generan actualmente. Al día, en Instagram se comparten 40 mil millones de fotos [1], Facebook tiene a 1090 millones de usuarios en activo [2], y por Twitter se envían unos 500 millones de tweets [3]. Y éstos son sólo 3 ejemplos. Para poder ver las 10 horas de vídeo que se suben cada segundo a Youtube se necesitarían 36000 personas dedicadas exclusivamente a esa tarea [4]. Y va en aumento.

Uno de los sectores que se ven más afectados por esta saturación es el del marketing [5]. Los medios tradicionales donde se realizaban las campañas de publicidad, como la televisión o la prensa escrita, cada vez están más ligados a Internet (televisores inteligentes, visionado de los canales de televisión en streaming, prensa online), lo que ha provocado un cambio en la manera de realizar las campañas. En Internet no existen espacios especialmente definidos para la publicidad, los usuarios pueden instalar programas para bloquear cualquier anuncio, y, por si luchar contra esto fuera poco, las campañas además tienen que esforzarse por llamar la atención entre toda la información que se intercambia. [6]

Pero no todo son desventajas para este sector. Si se consigue llamar la atención de los usuarios hasta el punto de conseguir que sean ellos mismos los que difundan la campaña, Internet permite llegar a alcanzar una cantidad de público objetivo inimaginable para los medios tradicionales de difusión. Y lograr esto es el objetivo principal del marketing viral.

El **Marketing Viral** se define como un conjunto de técnicas de mercadotecnia, generalmente basadas en el uso del “word of mouth” (boca a oreja o boca a boca) a través de medios electrónicos (como redes sociales, mail, móviles), que buscan mejorar el reconocimiento de una marca mediante la difusión masiva de su campaña. Su nombre se debe a que su funcionamiento es similar a la expansión de los virus, de tal manera que son los consumidores “contagiados” los que se encargan de difundir la campaña. Esta propagación pandémica hace que la llegada al público objetivo sea lo más rápida posible.

Ventajas e inconvenientes del marketing viral

Existen una serie de ventajas e inconvenientes a la hora de utilizar una técnica viral en vez de una de otro tipo en una campaña de marketing [7]:

Ventajas

- ❖ No son necesarias grandes inversiones a la hora de llevar a cabo una campaña de marketing viral. Esto se debe, sobre todo, a que la difusión la realizan los propios usuarios, sin suponer un sobrecoste para la misma.
- ❖ Se puede llegar a muchas personas de diferentes áreas geográficas dentro del público objetivo, más que en otros medios de comunicación, ya que los usuarios saben a qué contactos les puede interesar y a cuáles no una determinada campaña.
- ❖ La información se transmite de unos usuarios a otros a gran velocidad, y de forma exponencial, lo que permite que mejore el conocimiento de marca, y con ello el posicionamiento de ésta en el mercado.
- ❖ El formato en el que se puede realizar la campaña, y los contenidos de la misma, pueden ser muy variados.
- ❖ Permite interactuar con los usuarios a través de correos o por medio de las redes sociales durante una campaña, y obtener así información de los mismos.

Inconvenientes

- ❖ Es difícil realizar una campaña que haga que los usuarios estén motivados a transmitirla a sus contactos y convertirla así en viral. Si se ve como una simple campaña publicitaria, el número de usuarios que la reenviarán a sus contactos es bajo.
- ❖ Existen problemas a la hora de reconocer a los usuarios clave en las redes sociales que permitan una mayor difusión y aceptación de la campaña entre sus conocidos. Ambos factores son muy importantes, ya que se puede dar el caso que un usuario transmita una campaña a todos sus contactos, pero estos no acepten sus recomendaciones.
- ❖ La información transmitida por la campaña puede ir variando a lo largo de la cadena de difusión y no hay ningún tipo de control, ni del mensaje ni de la distribución, por lo que es difícil planificar un seguimiento de la misma.
- ❖ Si la campaña es vía e-mail puede ser bloqueada por medio de los filtros para los virus y los correos de SPAM. Y si se realiza a través de redes sociales, puede llegar a ser vista como una intrusión a los usuarios.
- ❖ No es una técnica de marketing válida para todos los tipos de productos o marcas.
- ❖ Si existe insatisfacción por parte de los usuarios hacia la compañía para la que se realiza la campaña o se hace un uso inadecuado del marketing viral, se puede obtener el efecto contrario al deseado, dando paso a un word of mouth negativo.

Tipos de marketing viral

Las campañas de marketing viral pueden ser de diferentes clases, dependiendo del medio de difusión que utilicen y en qué se basen para llegar a ser virales. Algunos tipos de marketing viral son [8]:

- ❖ *Marketing encubierto*: Es el tipo de marketing viral más difícil de reconocer, ya que consiste en transmitir el mensaje sin que los usuarios lo identifiquen como una promoción. Al cabo del tiempo se revela la marca o producto a la que hacía referencia el mensaje, haciendo que parezca que los usuarios lo han descubierto por sus propios medios, y potenciando así el recuerdo de la campaña.
- ❖ *Viral incentivado*: Esta técnica consiste en ofrecer recompensas a los usuarios a cambio de que ayuden a la compañía que realiza la campaña. La ayuda puede ser de varios tipos, entre los que se encuentran el envío de los datos del propio usuario (correo electrónico, información del perfil), o el paso del mensaje a un número mínimo de contactos. La recompensa también puede ser de diferentes clases, siendo la más frecuente la realización de un concurso a través de alguna web o en una red social.
- ❖ *Mensajes de "pásalo"*: Es la técnica dentro del marketing viral que mejor representa el concepto del "boca a boca" electrónico. También se les conoce como mensajes o correos cadena. Consiste en el envío de un mensaje que, ya sea por su contenido cómico o emotivo, incita a los usuarios a reenviárselo a sus contactos.
- ❖ *Clubes de fans o asociaciones*: Se basan en la idea de los clubes de fans que se crean por algún personaje famoso. Para llevarlos a cabo se puede crear una página web o un grupo dentro de alguna red social donde se apoye la marca, compañía o producto sobre el que se está haciendo la campaña. Esta página debe ser desarrollada por un usuario que comente las noticias que vayan surgiendo y responda a las críticas. El usuario puede ser recompensado por la compañía de forma directa o por medio de promociones, pero no debe ser conocido por los demás usuarios, ya que afectaría a la credibilidad del usuario y, por consiguiente, a la eficacia de la campaña.
- ❖ *Bases de datos*: Son comunidades online consistentes en una base de datos ofrecida por una determinada compañía en la que los usuarios pueden inscribirse y crear listas de contactos con sus conocidos. Estos por su parte podrán añadir a sus respectivos allegados, y así sucesivamente, consiguiendo crear una base de datos de tipo viral, que podrá ser utilizada por la compañía para obtener los datos de los usuarios y enviar mensajes de forma masiva sin haber necesitado un gran esfuerzo para la recopilación de los mismos.

En cuanto a las vías en que el marketing viral se puede realizar a través de Internet, existen de varios tipos, entre las que se encuentran las páginas web, los blogs, el e-mail, las redes sociales, etc. De entre todas ellas, se van a estudiar las redes sociales y su aplicación al marketing viral, ya que son las que han sufrido una mayor expansión en cuanto al número de usuarios estos últimos años y tienen una gran capacidad de difusión de la información.

1.2 Aplicación de Redes Sociales al Marketing Viral

De acuerdo con las condiciones necesarias para que el desarrollo de una campaña se vuelva viral en Internet, el marketing puede encontrar en las redes sociales a un gran aliado para la divulgación masiva de los mensajes. Sobre todo teniendo en cuenta el número de usuarios en activo que tienen algunas redes sociales, como Facebook, Twitter, Youtube o Instagram, que llegan a contar con millones de usuarios en todo el mundo.

Las redes sociales tienen una serie de ventajas que pueden favorecer a que campañas de marketing lleguen a ser virales. Entre ellas se encuentran las siguientes [9]:

- ❖ La principal ventaja de las redes sociales es el elevado número de usuarios que participan activamente en ellas. Por ejemplo, Facebook tiene 1650 millones de usuarios en activo, Twitter 320 millones, Youtube 1000 millones, etc. De este número tan alto de usuarios interconectados se pueden aprovechar las campañas de marketing, ya que los usuarios pueden compartir los mensajes fácilmente y hacer recomendaciones a sus conocidos.
- ❖ No conllevan ningún coste para la compañía. Mientras que tener una web propia o publicar publicidad en otras páginas web supone algún tipo de coste para la empresa, el crear un perfil en una red social no tiene ningún coste adicional.
- ❖ Los usuarios pueden generar comentarios positivos que favorezcan a la campaña. Al ser realizados por personas ajenas a la empresa tienen mayor credibilidad que la propia campaña en sí. Por lo que si se consiguen generar comentarios positivos, los usuarios tendrán mejor imagen de la marca y será más probable que compartan el mensaje con sus respectivos contactos.
- ❖ Además de conseguir que la campaña se haga viral dentro de la propia red social, si se añade a la noticia un link a la página web de la compañía se puede conseguir “viralizar” también el número de visitas a la web, y con ello mejorar el reconocimiento de marca.
- ❖ Se puede ir conociendo en tiempo real la opinión de los clientes potenciales sobre la campaña. De esta manera se puede intentar rectificar si no recibe los resultados esperados o si recibe comentarios negativos por parte de los usuarios.
- ❖ A día de hoy existen navegadores web que permiten instalar extensiones para bloquear la publicidad dentro de las páginas web. Consiguiendo que los propios usuarios envíen la campaña de marketing desde sus cuentas a sus contactos se pueden evitar estos bloqueadores, ya que todo el contenido publicado y transmitido por los usuarios no se considera publicitario.

Para poder aprovechar estas ventajas y que la campaña tenga éxito en la red social, ésta no debe ser vista como un anuncio publicitario como tal. No se tiene que centrar en el producto o la marca que se esté tratando promocionar, sino que debe introducirse de forma sutil. Además, debe tener en cuenta el público objetivo al que

se está intentando acceder, ya que la forma de promocionar no es la misma para unos y otros segmentos poblacionales.

Otra de las características recomendables es que el mensaje sea cómico o emotivo. Es más sencillo que los usuarios compartan un mensaje por las emociones que les haya hecho sentir y que quieran que sus conocidos puedan vivir también, que un mensaje publicitario.

De esta forma, si se consigue distinguir correctamente el segmento de usuarios de la red social al que va dirigida la campaña y ésta es capaz de transmitir las emociones necesarias para convencer a los usuarios de su difusión, se podrá sacar el máximo partido a las ventajas que las redes sociales ofrecen. Como muestra de ello están los siguientes tres ejemplos de campañas publicitarias de marketing viral que supieron hacer un buen uso de las redes sociales:

- ❖ *Campaña Dove Real Beauty Sketches* [10]: Dove creó en 2013 una de las campañas de marketing que más repercusión ha tenido en las redes sociales a nivel internacional hasta la fecha. Es el claro ejemplo de cómo una campaña de marketing emotiva puede llegar a ser viral.

La campaña consistía en un artista forense que retrataba en dos ocasiones a una mujer sin haberla visto anteriormente. En primer lugar seguía las indicaciones de la propia mujer pintada, y en segundo lugar las de una persona que no conocía de nada a la retratada, demostrando la diferencia entre la percepción que alguien puede tener sobre sí mismo y lo que realmente ven los demás, y como las personas tienden a infravalorar su aspecto.

La campaña contó con un gran impacto en las redes: más de 114 millones de reproducciones en Youtube, fue traducida a 25 idiomas, se compartió en 660.000 ocasiones en Facebook durante los primeros 10 días después de su estreno, y se realizaron 1800 blogs escritos por usuarios ajenos a la empresa sobre la misma.

- ❖ *Campaña #comparteCocaCola* [11]: En esta campaña Coca Cola se centró en un segmento muy concreto de la población: los Millenials (generación más joven de consumidores de la compañía). Esta campaña no se basaba en un anuncio o un mensaje concreto, sino que la clave del mismo fue llamar la atención del segmento al que iba dirigido, y estos fueron los encargados de difundirlo por las redes sociales más conocidas.

Para llevar a cabo la campaña, Coca Cola cambió el diseño de sus latas y botellines, y añadió los nombres más comunes dentro de la generación objetivo. De esta manera, los jóvenes se acercaban a los establecimientos buscando la lata o botellín que tuviese su nombre, y se hacían una foto con ella, publicándola luego en las redes sociales (sobre todo Twitter, donde se añadía el hashtag #comparteCocaCola de la campaña).

Fue todo un éxito y un gran número de personas compartieron fotografías en las distintas redes sociales con sus coca-colas personalizadas, permitiendo a la compañía un aumento de sus ventas del 2%, cantidad nada despreciable teniendo en cuenta las dimensiones de Coca Cola.

- ❖ *Campaña Old Spice* [12]: El humor es otro de los factores que pueden ayudar a una campaña publicitaria a hacerse viral. Y ese es el caso de Old Spice. Old Spice es una marca, perteneciente a Procter & Gamble, de productos para ducha masculina. A comienzos de 2010 publicó el video “The Man Your Man Could Smell Like”, que aumentó en gran medida la notoriedad de la marca, por lo que a mediados de ese año lanzó la segunda parte, “The Return of The Man Your Man Could Smell Like”, en la Super Bowl.

Los resultados de ambas campañas fueron excelentes, obteniendo la segunda parte unas 6 millones de visitas en Youtube en las primeras 24 horas. Además de esta buena acogida en las redes sociales, consiguió que la marca incrementase sus ventas en un mes en un 107%, gracias en parte al número de visitas obtenidas por el video en Youtube, además de las parodias realizadas sobre el mismo.

Pero no siempre las campañas publicitarias se vuelven virales porque la compañía así lo buscase inicialmente. Hay veces que en las redes sociales se vuelven virales algunas campañas por motivos negativos.

Un ejemplo de esto es el caso de la campaña de Loewe en 2012 [13]. El principal error de esta campaña fue la segmentación y selección del público objetivo, ya que intentó buscar como posibles compradores en plena crisis económica al sector más castigado por la misma, los jóvenes. Otro de los errores fue la elección de los protagonistas del anuncio (en su mayoría hijos o familiares de famosos en la escena madrileña de los setenta y ochenta, además de bloggers, videoartistas y otros protagonistas de la actual noche madrileña), ya que el anuncio daba a entender que representaban a un amplio sector de la juventud española. Por todo esto, la campaña fue trending topic en Twitter en un solo día debido a la cantidad de comentarios críticos sobre la misma.

Y es que el uso del marketing viral en las redes sociales también puede tener inconvenientes, entre los que se encuentran los siguientes:

- ❖ Aunque el coste monetario de desarrollar una campaña de marketing en las redes sociales no es muy alto, sí que requiere de un tiempo y esfuerzo por parte de los desarrolladores, ya que tienen que ser capaces de contestar las dudas y críticas que vayan surgiendo por parte del público objetivo.
- ❖ Como se ha comentado en el caso de la campaña de Loewe, existe una falta de control por parte de la compañía de los comentarios que se van generando sobre la campaña, sobre todo si la viralidad se basa en los comentarios negativos sobre la misma. Esto puede llegar a producir un deterioro de la imagen de marca y una disminución de los beneficios.

Para intentar aprovecharse al máximo de las ventajas de las redes sociales, y evitar en la mayor medida posible los inconvenientes, una de las opciones es que la compañía no lance directamente la campaña, sino recompensar a usuarios de la red social para que la publiquen y hablen positivamente sobre ella. De esta manera se consigue que el resto de los usuarios no vean la campaña como un anuncio, siendo así más fácil la difusión y aceptación del producto, ya que suelen tener más en cuenta los gustos y sugerencias de aquellos a los que siguen que de las campañas publicitarias y lo que estas les ofrecen.

Pero hay que tener en cuenta que no todos los usuarios tienen el mismo número de seguidores en las redes sociales, ni ejercen el mismo impacto en ellos, ni exigen la misma recompensa para promocionar el producto o la marca. Todos estos factores tendrán que tenerse en consideración a la hora de realizar una campaña de marketing en las redes sociales. Se debe buscar la combinación de los tres que haga que se llegue y convenza de la compra del producto al mayor número de usuarios posible dentro de la red, ciñéndose al presupuesto máximo del que disponga la campaña.

El problema a abordar es la selección del grupo de usuarios que optimice estos tres aspectos; qué piezas de dominó elegir para conseguir la mejor reacción en cadena.

2 REDES SOCIALES Y CENTRALIDAD INDIVIDUAL

2.1 ¿Qué es una Red Social?

Para poder trabajar con una red social y ser capaces de localizar los individuos más importantes de la misma, primero hay que conocer qué es y cómo funciona.

Se define una **red social** como un conjunto de actores (entre los que se encuentran personas, grupos de personas u organizaciones) que están relacionados teniendo en cuenta algún tipo de característica. Gracias a ellas se pueden conocer las conexiones entre los actores de la red, el peso o poder de los individuos, si las relaciones entre los integrantes son lo suficientemente resistentes a largo plazo, etc. Factores que son clave para el estudio de los grupos de individuos. [14]

Aunque el conocimiento sobre las redes sociales en ámbitos como la sociología [15], la psicología o la economía puede remontarse hasta el siglo XIX, en el ámbito de la informática no surgieron hasta la segunda mitad del siglo XX. Aunque su concepción es relativamente reciente, su desarrollo a partir de entonces ha sido exponencial. Los hitos más importantes en su evolución son los siguientes [16]:

- ❖ En 1978 se crea el BBS (Bulletin Board Systems), que servía para compartir información entre amigos.
- ❖ En 1995 se crea lo que se considera el primer servicio de red social, Classmates, que sirve para poner en contacto a antiguos alumnos de entidades educativas.
- ❖ Surge en 1997 AOL Instant Messenger, primer servicio web que tiene chat para que los usuarios puedan comunicarse por escrito en tiempo real. En este mismo año aparece también Sixdegrees, primera red social donde se podía modificar el perfil y crear listados de amigos.
- ❖ En 2002 surge MySpace y en 2003 LinkedIn, la primera red social especializada en un sector específico, en este caso el mercado laboral.
- ❖ El 4 de febrero del 2004 nace Facebook, la red social con más usuarios activos del mundo a día de hoy: 1650 millones.
- ❖ En 2005 nace Youtube como sitio web de compartición de vídeos y en 2006 nace Twitter como servicio de microblogging.

A partir de estos años se crean un gran número de nuevos servicios de redes sociales (entre los que se encuentran Pinterest e Instagram en 2010). Las redes ya existentes comienzan a tener cada vez más popularidad, salvo algunas excepciones como MySpace, llegando a superar los cientos de millones de usuarios, con miles de millones de visitas diarias, y permitiendo la comunicación entre personas de todas las partes del mundo.

Todo esto ha supuesto un impacto enorme de las redes sociales en la sociedad. Han conseguido cambiar el estilo de vida de gran parte de la población, sobre todo facilitando la comunicación y permitiendo acceder a noticias e información de todo

el mundo en tiempo real. Y es que el nivel de usuarios al que pueden llegar las redes sociales era inimaginable hace tan solo 20 años. Por ejemplo, en enero de 2016, el dato sobre el número de usuarios que se conectaban al día a las diferentes redes sociales era de 2307 millones, es decir, el 31% de la población mundial. [17] Y la tendencia es que su uso siga en aumento (del 2015 al 2016 el aumento del uso de redes sociales se estima en un 10%, o lo que es lo mismo, 219 millones de personas).

Propiedades de las Redes Sociales

Dentro de las redes sociales se pueden identificar una serie de propiedades que son útiles a la hora de trabajar con ellas y extraer información. [18]

- ❖ *Efecto pequeño mundo*: Dentro de una red social los diferentes individuos están conectados unos a otros por un número de intermediarios relativamente pequeño. Esto permite que el paso de la información entre los individuos se realice rápidamente, una de las propiedades que más interesa en el marketing viral. Es en esta propiedad en la que se basa la teoría de que todas las personas del planeta están conectadas unas a otras por un número máximo de intermediarios de 6 personas.
- ❖ *Transitividad o formación de bloques*: En redes sociales es muy probable que si existe un individuo relacionado con otros dos, estos últimos también se encuentren relacionados entre sí.
- ❖ *Resistencia de las redes a romperse*: Aunque se elimine algún individuo dentro de la red, se tiende a aumentar las conexiones entre los individuos afectados para evitar la ruptura de la misma.
- ❖ *Diferentes tipos de individuos*: Es habitual encontrar diferentes categorías dentro de los individuos de una red social, lo que afecta también a las conexiones entre ellos. Los individuos semejantes estarán relacionados entre sí, lo que favorece a división de la red en grupos de características similares.

Todas estas propiedades surgen del estudio de las redes sociales utilizando la teoría de grafos, ya que toda red social puede representarse por medio de un grafo para su posterior estudio.

2.2 La Red Social como un Grafo

Toda red social puede representarse por medio de un grafo en el que los individuos que la forman son los nodos o vértices, y las relaciones existentes entre ellos son las aristas (si la relación es bidireccional) o arcos (si la relación es unidireccional). [19]

La teoría de grafos es una rama de las matemáticas y la computación encargada del estudio de los grafos. Permite estudiar, entre otras cosas, la importancia de los nodos de un grafo con respecto a sus vecinos, la centralidad de cada nodo en función de diferentes aspectos, el camino mínimo entre varios nodos del grafo, etc., información muy útil a la hora de realizar el estudio sobre una red social. Por este

motivo, el análisis de redes sociales para la obtención de grupos óptimos se va a basar en la teoría de grafos.

Los grafos se suelen representar utilizando puntos para simbolizar los vértices, y líneas para las aristas que relacionan unos vértices con otros. En cuanto a la notación utilizada para definir un grafo G , se va a utilizar la fórmula $G = (V, E)$, donde V representa al conjunto de vértices y E al conjunto de aristas o arcos entre los vértices.

Tipos de grafos

Existen muchos tipos de grafos diferentes: multígrafos, no dirigidos, pseudografos, vacíos, completos, perfectos, etc. Dentro del estudio de redes sociales interesa definir los dos grupos más importantes: los grafos no dirigidos y los grafos dirigidos.

- ❖ *Grafo no dirigido* $G = (V, E)$: El ejemplo más claro de las relaciones representadas por este tipo de grafos es el de la red social Facebook: si dos individuos dentro de la red social son considerados amigos es porque la relación siempre es recíproca, el usuario A es amigo de B y viceversa.

Este tipo de grafos contienen un conjunto de vértices (V) y un conjunto de relaciones (E) no vacíos, en el que las relaciones son bidireccionales, es decir, que los nodos están unidos por aristas, haciendo así que el grafo este formado por pares de nodos no ordenados.

El ejemplo que se puede ver a continuación representa las relaciones de amistad entre tres usuarios de una red. Los usuarios pueden conocerse, como en el caso de los nodos 1 y 2 o 2 y 3 del ejemplo, o no conocerse, como en el caso del 1 y 3, pero la relación es bidireccional: o existe para los dos o no existe.

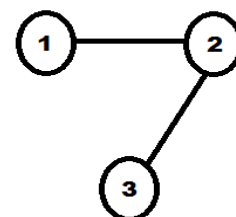


Ilustración 1. Grafo de Red de Amistades

- ❖ *Grafo dirigido* $G = (V, E)$: En este caso la red social que sirve como claro ejemplo de los grafos dirigidos es Twitter: un determinado usuario puede seguir a otro, pero no tiene por qué ser recíproco.

Este tipo de grafos contienen un conjunto de vértices (V) y un conjunto de relaciones (E) no vacíos, en el que las relaciones son unidireccionales, es decir, que los nodos están unidos por arcos, haciendo así que el grafo este formado por pares de nodos ordenados.

El ejemplo que se puede ver a continuación contempla la influencia que ejercen unos usuarios sobre otros en una red social. En este caso la relación es unidireccional, ya que un nodo puede ejercer influencia sobre el otro, y no al revés, como en el caso de los nodos 1 y 2, donde el 1 ejerce influencia sobre el 2, pero no es recíproco; pueden influirse el uno al otro pero no en el mismo grado, como es el caso de la relación entre el 2 y el 3, que al ejercer influencia cada

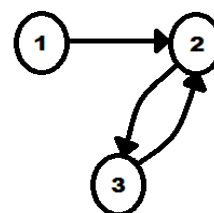


Ilustración 2. Grafo de influencia de una red social

uno de ellos sobre el otro deben representarse dos arcos diferentes, o que no haya ninguna relación, como entre los nodos 1 y 3.

Otro criterio de para catalogar los diferentes tipos de grafos es la conexión del mismo. En este caso se puede diferenciar si el grafo es conexo o no conexo.

Se dice que un grafo no dirigido es conexo si todos sus nodos están conectados entre sí. En caso contrario será un grafo no conexo.

En los grafos dirigidos se pueden distinguir dos tipos de grafos conexos: fuertemente conexos y débilmente conexos. Se dice que un grafo dirigido es fuertemente conexo si hay un camino entre dos nodos en ambas direcciones, es decir, que existe una conexión entre 1 y 2 y entre 2 y 1. Y se dice que un grafo dirigido es débilmente conexo cuando hay un camino entre cualquier par de nodos si no se tiene en cuenta la dirección de sus arcos.

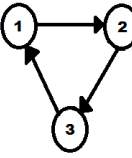
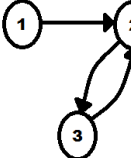
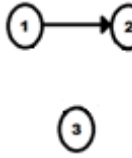
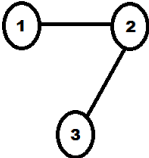
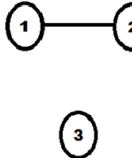
TIPOS DE GRAFOS	CONEXO		NO CONEXO
	FUERTE	DÉBIL	
DIRIGIDO			
NO DIRIGIDO			

Tabla 1. Tipos de Grafos

Representación y propiedades de los grafos

Además de los tipos de grafos con los que se puede desarrollar una determinada red, la teoría de grafos estudia los métodos de representación de los grafos. La más común es la representación gráfica, pero también se pueden utilizar las matrices de adyacencia e incidencia.

En un grafo no dirigido dos nodos son adyacentes, o son vecinos, si existe una arista que los une directamente, es decir, si están directamente relacionados. Siguiendo con el ejemplo utilizado para explicar los grafos no dirigidos, dos individuos en Facebook serían adyacentes si son amigos directos.

Por su parte, en un grafo dirigido la adyacencia entre nodos también es direccional. En el ejemplo de Twitter, si un individuo A tiene como seguidor a otro B, se dice que A es *adyacente a* B, y B es *adyacente desde* A. En estos casos se puede distinguir entre el vértice inicial, que en el ejemplo sería el individuo A, y el vértice final, que sería el B.

Teniendo en cuenta la definición de la adyacencia, se pueden construir matrices a partir de los nodos y la relación que existe entre ellos. Estas matrices, conocidas como matrices de adyacencia, son matrices compuestas por 0's y 1's, con tantas filas y columnas como nodos haya en el grafo. La matriz se irá rellenando recorriendo todas las posiciones, teniendo en cuenta que si el vértice que corresponde con una determinada fila i es adyacente al vértice que corresponde con una determinada columna j , la posición (i,j) de la matriz tendrá un 1, y un 0 en el caso contrario.

No existe una única matriz de adyacencia por cada grafo, ya que el orden de los vértices no es fijo, puede elegirse qué fila y columna corresponden a cada vértice en la matriz (existen $n!$ matrices de adyacencia posibles para un grafo, siendo n el número de nodos de la matriz).

Sin embargo, como en el caso de los grafos no dirigidos la relación es bidireccional, los valores correspondientes a la fila i y la columna j (posición ij) serán iguales a los valores de la fila j y la columna i (posición ji). Es decir, la matriz de adyacencia de este tipo de grafos será simétrica con respecto a la diagonal principal.

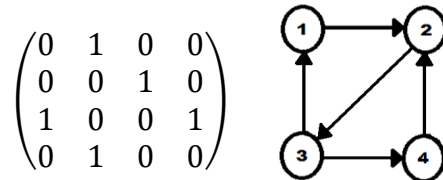


Ilustración 3. Matriz de Adyacencia de un Grafo

Las matrices de adyacencia permiten compactar los datos del grafo para que sea más sencillo trabajar con ellos y realizar cálculos sobre el grafo utilizando operaciones matriciales. Pero esta versión simple no es útil en todas las situaciones.

Por ejemplo, en el caso de un grafo que represente la influencia que ejercen unos usuarios sobre otros, se puede querer ponderar el nivel de esa influencia. Esos niveles se pueden representar con valores del 0 al 1, siendo 0 que no ejerce nada de influencia y 1 que ejerce una influencia total. En este caso la matriz de adyacencia simple no estaría teniendo en cuenta el valor de los pesos de las relaciones, sólo almacenaría si existe o no relación entre los nodos. Para poder solucionar esto, se utiliza la matriz de adyacencia ponderada.

La matriz de adyacencia ponderada se desarrolla igual que la matriz de adyacencia simple, sólo que en este caso no está formada por 0's y 1's, sino que

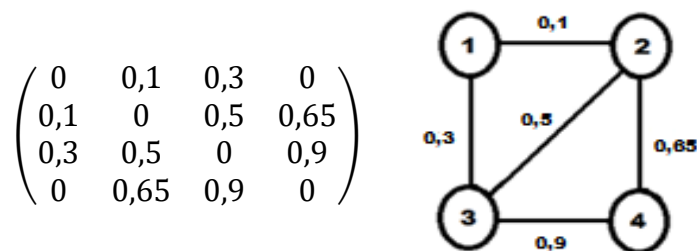


Ilustración 4. Matriz de Adyacencia ponderada de un grafo

dentro de cada posición (i,j) se utiliza el peso asignado a la relación entre el vértice i y el vértice j . En redes sociales también se conoce a esta matriz como la socio-matriz de la red.

La otra opción para la representación de un grafo es mediante una matriz de incidencia. Esta se puede utilizar, entre otras cosas, para el cálculo de alguna de las medidas de centralidad individual.

Se dice que una arista “e” es incidente con un par de nodos A y B cuando A y B son nodos adyacentes y “e” es la arista que los une. En el ejemplo de Facebook, si llamamos A y B a dos individuos que son amigos en la red social y e a la arista que representa esa relación de amistad en el grafo uniendo los nodos A y B, se diría que e es incidente tanto a A como a B.

La matriz de incidencia es, por tanto, la matriz donde se representan los vértices y aristas que son incidentes. Es una matriz de 0's y 1's y de tamaño $n \times m$, donde n equivale al número de vértices del grafo (filas de la matriz) y m al número de aristas (columnas de la matriz). Aparecerá un 1 en aquellas posiciones (i,j) en las que la arista e_j , correspondiente a la arista almacenada en la columna j , sea incidente con el vértice v_i , y 0 en el caso contrario.

El problema de las matrices de incidencia con respecto a las matrices de adyacencia es que sólo pueden representar los datos bidireccionales, los de los grafos no dirigidos. Por este motivo el uso de las matrices de adyacencia está mucho más extendido.

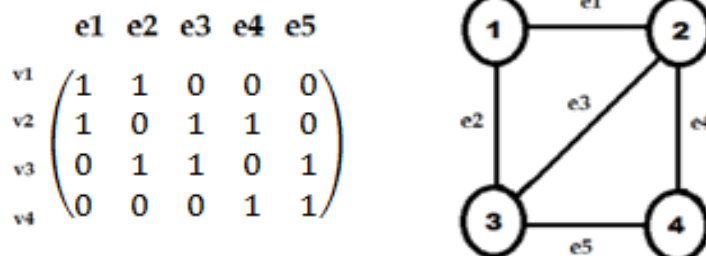


Ilustración 5. Matriz de incidencia

Las matrices de adyacencia e incidencia sirven para compactar los datos de las redes sociales, haciendo más fácil el representar el grafo de la red y trabajar con él. Pero para extraer conclusiones con respecto a la distribución de la misma y la importancia de los distintos nodos existen en teoría de grafos las medidas estructurales de centralidad individual y las medidas de centralidad de grupos.

Para poder comprender alguna de estas medidas de centralidad es necesario conocer que son el camino y el camino mínimo de un grafo.

Un camino entre dos nodos A y B de un grafo se corresponde con la secuencia de nodos que se deben atravesar para llegar desde el nodo inicial A al nodo final B. Puede existir más de un camino entre dos nodos, sobre todo si el grafo tiene un gran número de conexiones. En el caso de los grafos dirigidos, el camino sólo se podrá realizar siguiendo la direccionalidad de los arcos, así que no sólo será necesario que

exista una relación entre los
nodos que se van siguiendo,
sino que además ésta tiene
que ser en la dirección
correcta.

Por ejemplo, si tenemos en cuenta el siguiente grafo, se puede observar que existen un gran número de caminos para llegar desde A hasta G. Entre estos se pueden encontrar desde los más simples, como puede ser el camino {A, B, G}, a los más complejos, como el camino {A, C, F, H, B, D, E, G}.

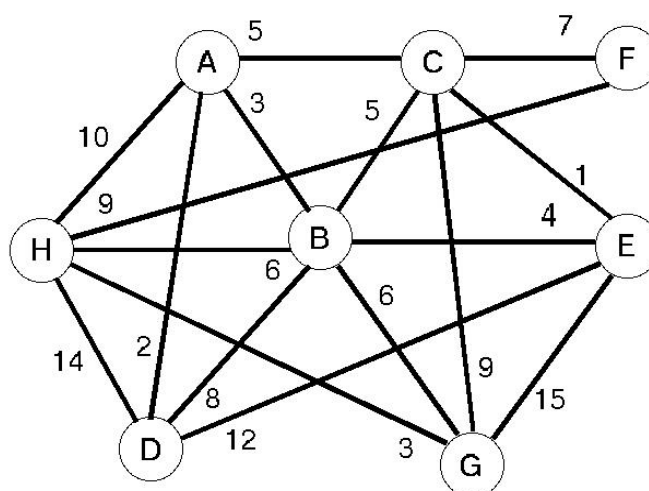


Ilustración 6. Grafo de ejemplo para camino mínimo

Como es lógico, en el caso de las medidas de centralidad basadas en los caminos el que interesa es aquel que hace que el recorrido sea el menos costoso de todos. Y es aquí donde entra el concepto de camino mínimo.

El camino mínimo entre dos nodos A y B es aquel camino que minimiza el coste que supone llegar del nodo de salida al nodo de llegada. En el caso de grafos cuyas aristas no estén ponderadas (sin pesos), el camino mínimo será aquel que haga mínimo el número de nodos que hay que pasar para llegar del nodo salida al nodo llegada. Por ejemplo, en el caso de que el grafo anterior no tuviese pesos en las aristas, para llegar del nodo H al nodo D el camino mínimo sería {H, D}, utilizaría la arista que une directamente a ambos nodos.

Sin embargo, si el grafo tiene pesos en las aristas, el camino mínimo será aquel que minimice la suma de éstos para llegar de un nodo a otro. Siguiendo con el

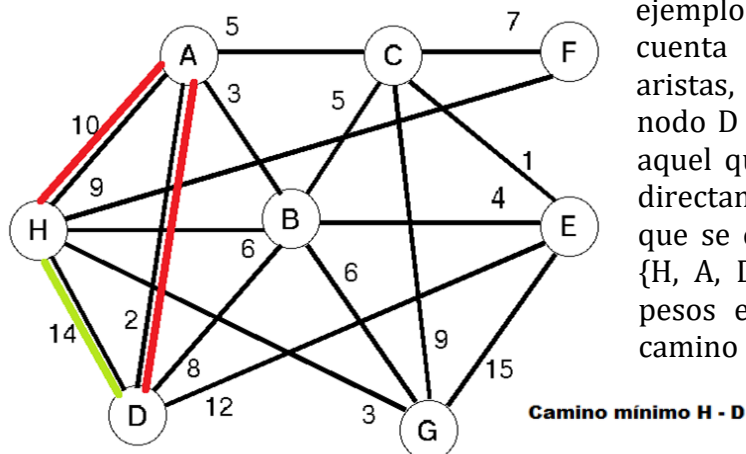


Ilustración 7. Grafo con caminos mínimos entre nodos H y D

ejemplo anterior, pero teniendo en cuenta esta vez los pesos de las aristas, para llegar del nodo H al nodo D el camino mínimo no sería aquel que utiliza la arista que une directamente a ambos nodos, sino que se corresponde con el camino $\{H, A, D\}$, ya que la suma de sus pesos es 12, frente a los 14 del camino $\{H, D\}$.

2.3 Medidas estructurales de centralidad individual

Uno de los usos principales que se le puede dar a la teoría de grafos cuando se está realizando el estudio de una red social es la capacidad de identificar cuál es el individuo o grupo de individuos más importantes dentro de la red.

Con esta idea, la de buscar a los actores más importantes de la red, la teoría de grafos estudia varias medidas que indican la centralidad de cada uno de los actores dentro de la misma. Es decir, estudia medidas que asignan a cada nodo un valor numérico asociado a la importancia que tiene dentro de la red. Como las diferentes medidas tienen un concepto distinto de lo que es ser el centro de un grafo, se puede elegir cuál utilizar en función del tipo de actor que se esté intentando encontrar.

Se van a estudiar 5 medidas de centralidad (**Degree, Closeness, Betweenness, Eigenvector y Page-Rank**) sobre el grafo *Krackhardt Kite*. Este grafo tiene la peculiaridad de que es el más pequeño en el que las medidas clásicas de centralidad, degree, closeness y betweenness, dan como nodo central 3 valores diferentes, facilitando la visualización de las diferencias existentes entre estas medidas. [20] [21]

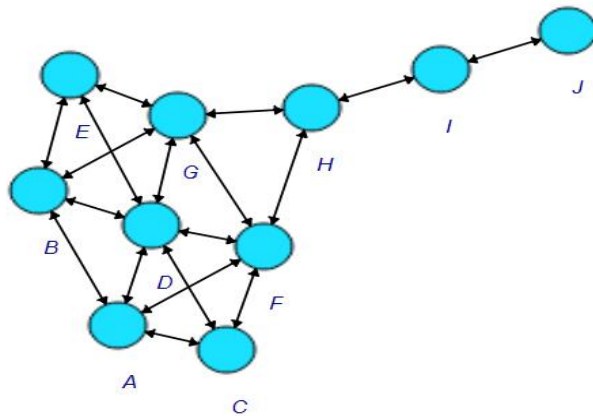


Ilustración 8. Grafo Kite para medidas de centralidad

2.3.1 Degree

La medida de centralidad *Degree* (o centralidad de grado), es la que se basa en la idea más simple de lo que tiene que ser un individuo central de la red: el actor central es aquel que más relaciones tiene con el resto de los actores en el grafo. Este valor también se conoce como grado del vértice en los grafos no dirigidos, y se define como el número de aristas incidentes con el nodo sobre el que se está calculando el valor, a excepción de los bucles, que suman el doble.

Teniendo en cuenta el grafo Kite, al calcular el degree sobre todos los nodos, se puede ver que el nodo D es el que tiene un mayor número de conexiones con el resto del grafo (es adyacente al A, B, C, E, F y G). El valor del degree de este nodo sería igual a 6, y como el resto de los nodos no llegan a alcanzar un valor tan alto (como se puede ver en la siguiente tabla), esta medida considera al nodo D como el nodo central del grafo.

A	B	C	D	E	F	G	H	I	J
4	4	3	6	3	5	5	3	2	1

Tabla 2. Resultados del Degree para el grafo Kite

Se puede escribir formalmente esta medida como sigue: Si tenemos un grafo G , que tiene un conjunto de vértices V y un conjunto de aristas E , podemos definir la centralidad de grado de un nodo v determinado como la suma de aristas que inciden en dicho nodo:

$$degree_v = \sum_i^N e_{iv}$$

Ecuación 1. Métrica de Centralidad Degree sin adyacencia

Donde llamamos e_{iv} a las aristas que unen al nodo i con el nodo v , y N al número total de nodos del grafo.

En el caso de disponer de la matriz de adyacencia, las aristas que unen a un nodo i con v vienen determinadas por la posición a_{iv} de la matriz. Por lo tanto la fórmula anterior se puede calcular como la suma de estas posiciones, ya que tendrán un valor de 1 en todas aquellas en las que exista un vértice entre el nodo i y el nodo v y un valor de 0 en caso contrario:

$$degree_v = \sum_i^N a_{iv}$$

Ecuación 2. Métrica de Centralidad Degree con Adyacencia

Todas estas medidas se han realizado teniendo en cuenta que el grafo fuese no dirigido. En el caso de los grafos dirigidos, esta medida es algo más compleja, ya que tiene en cuenta los arcos de entrada y de salida del nodo. Se pueden distinguir tres tipos de medidas de grado: el grado de entrada (in-degree), el grado de salida (out-degree) y el grado total, que se corresponde con la suma del grado de entrada más el grado de salida del nodo.

El grado de entrada se puede interpretar como el nivel de influencia que el resto de los nodos ejerce sobre el nodo v , y el grado de salida como el nivel de influencia que ejerce el nodo v sobre los demás nodos del grafo.

Si nos basamos en el uso de la matriz de adyacencia, la definición del cálculo del grado de entrada coincide con la dada para el cálculo del grado en los grafos no dirigidos. Sin embargo, como en los grafos dirigidos la matriz de adyacencia no tiene por qué ser simétrica, el grado de salida se calcula mediante las posiciones a_{vj} , en lugar de a_{iv} :

$$indegree_v = \sum_i^N a_{iv} \quad outdegree_v = \sum_j^N a_{vj}$$

Ecuación 3. Métricas de Centralidad In-Degree y Out-Degree

2.3.2 Closeness

La medida de centralidad *Closeness* (o centralidad de cercanía) es una medida algo más completa que el degree, ya que en el caso de que el grafo tenga pesos en sus aristas, ésta sí que los tiene en cuenta. Fue creada por Beauchamp, y consiste en calcular la distancia mínima desde el nodo que se está midiendo al resto de los nodos del grafo.

El cálculo de esta medida se puede utilizar para aproximar el tiempo que cada nodo tardaría en propagar la información de un nodo al resto de la red.

Para ello, el valor de centralidad que se asigna a cada nodo puede ser o bien la suma de todas las distancias del nodo a los demás, o bien el promedio de dicha distancia. En el caso de que el grafo no tenga pesos, la distancia mínima se corresponde con el número de aristas mínimo que haya entre un nodo y otro. Si el grafo tiene pesos, esta distancia se corresponderá con el cálculo del camino mínimo entre los nodos.

Volviendo a utilizar el grafo Kite, al ser éste un grafo sin pesos, la centralidad closeness tendrá en cuenta el camino mínimo calculado como el menor número de aristas entre un nodo y otro. La fórmula que se va a utilizar para el cálculo de esta medida en el grafo se puede representar de la siguiente forma:

$$closeness_v = \sum_{j=1}^N \frac{1}{distancia(v,j)}$$

Ecuación 4. Métrica de Centralidad Closeness

Donde v es el nodo sobre el que se está calculando la centralidad, $distancia(v,j)$ se corresponde con el camino mínimo del nodo v al j , y N es el número de nodos totales del grafo.

Si se realizan los cálculos sobre el grafo, se obtienen los siguientes resultados:

A	B	C	D	E	F	G	H	I	J
0,0588	0,0588	0,0556	0,0667	0,0556	0,0714	0,0714	0,0667	0,0476	0,0345

Tabla 3. Resultados del Closeness en grafo Kite

Es decir, que según la medida de centralidad closeness, los nodos F y G son los que están a una distancia menor del resto de los nodos del grafo. Esto tiene sentido observando el mismo, ya que F y G son los nodos más cercanos del conjunto de nodos conexos (A, B, C, D, E, F y G, que se encuentran muy bien comunicados entre sí a una distancia máxima de 2), y de la “cola de la cometa” (H, I y J).

En la fórmula se está dividiendo entre la longitud del camino mínimo, para que la medida closeness dé al nodo más central el valor más alto, y no al revés. Si se estuviesen sumando las distancias, el nodo más central sería aquel que tuviese un valor de closeness menor, y en ese caso la medida estaría devolviendo el valor de lejanía de cada nodo, no el de cercanía.

2.3.3 Betweenness

La medida de centralidad *Betweenness* (o centralidad de intermediación), se encarga de medir el número de veces que un nodo se encuentra dentro del camino mínimo que une a otros dos nodos. Esta medida se puede utilizar para conocer el papel que tiene cada nodo a la hora de conectar diferentes grupos dentro del grafo.

Los nodos que van a dar valores más altos con esta medida de centralidad son aquellos que si fuesen eliminados dividirían el grafo en dos grupos. Es decir, aquellos que al desaparecer hacen que el grafo sea inconexo.

Si se tiene en cuenta las redes sociales, los individuos que dan mayor valor en la centralidad *betweenness* son aquellos que controlan el flujo de información entre grupos de usuarios. Como se quiere maximizar la difusión de la información en la red, es interesante conocer esta medida para saber quiénes son las puertas de acceso de unos grupos a otros.

Se va a volver a calcular la medida de centralidad sobre el grafo Kite. La fórmula que se va a utilizar sobre los nodos del grafo para conocer el valor de intermediación es la siguiente:

$$betweenness_v = \sum_{i,j \neq v} \frac{b_{ivj}}{b_{ij}}$$

Ecuación 5. Medida de Centralidad Betweenness

Donde v es el nodo sobre el que se está calculando la centralidad, b_{ij} se corresponde con número de caminos mínimos entre el nodo i y el nodo j , y b_{ivj} se corresponde con el número de caminos mínimos del nodo i al nodo j que pasan por el nodo v .

Si se realizan los cálculos sobre el grafo, se obtienen los siguientes resultados:

A	B	C	D	E	F	G	H	I	J
0.833	0.833	0	3.667	0	8.333	8.333	14	8	0

Tabla 4. Resultados de Betweenness para Kite

Es decir, que según la medida de centralidad *betweenness*, el nodo H es el de mayor importancia en todo el grafo. Como se puede ver, existen dos grupos en la red claramente diferenciados, el conjunto de nodos conexos (A, B, C, D, E, F y G) y lo que se podría denominar como “cola de la cometa” (H, I y J), y el nexo de unión entre ambos grupos es el nodo H. Si éste se quita, el grafo quedaría dividido en dos grupos inconexos entre sí, de ahí que el valor de *betweenness* de H sea el mayor del grafo.

En los grafos no dirigidos, no se tiene en cuenta en el cálculo de la intermediación si el camino mínimo se realiza del nodo i al nodo j , o del j al i . Como se considera que es el mismo, no se suma dos veces. Sin embargo, en el caso de los grafos dirigidos sí que habrá que calcular los dos caminos mínimos y las veces que se encuentra el nodo v en ellos, ya que estos caminos no tienen por qué ser los mismos.

2.3.4 Eigenvector (Bonacich)

Phillip Bonacich propuso en 1972 la medida de centralidad Eigenvector (o centralidad del vector propio). Esta medida calcula el nivel de influencia que tiene un nodo en una determinada red. [22]

La centralidad en esta medida depende tanto del número de conexiones que tiene cada nodo como de la calidad de las mismas, ya que en este caso cada conexión tendrá un peso diferente en función del nivel de influencia de la red. Es decir, que cuanto más conectado esté un nodo, y cuanto mejores sean sus vecinos, mayor será su valor del eigenvector.

Se puede expresar la fórmula para el eigenvector de la siguiente forma:

$$evcent_v = \lambda \sum_{j \rightarrow v} evcent_j$$

Ecuación 6. Medida de Centralidad Eigenvector

Donde λ es un valor normalizador, y $\sum_{j \rightarrow v} evcent_j$ representa la suma de los valores de eigenvector de aquellos nodos que tienen una arista que termina en el nodo v .

Al estar normalizado, el valor que va a dar se puede encontrar entre el 0 y el 1, correspondiendo el valor 1 al nodo de mayor importancia.

En el caso del grafo kite, al no tener pesos las aristas, en el cálculo del vector propio la importancia de cada uno de los nodos vendrá determinada por su número de conexiones. Por este motivo, los cálculos van a dar resultados similares a los del degree:

A	B	C	D	E	F	G	H	I	J
0.732	0.732	0.594	1	0.594	0.827	0.827	0.407	0.010	0.023

Tabla 5. Resultados de Eigenvector para grafo Kite

Como se puede ver en la tabla, el nodo con mayor valor es el D, al igual que en el caso del degree, ya que es el que tiene un mayor número de conexiones. Sin embargo, mientras que el degree daba el mismo valor para los nodos C, E y H (3), en este caso el H vale un poco menos que los otros dos. Esto se debe a que los nodos a los que se encuentra conectado son menos importantes en la red que los conectados a C y a E.

2.3.5 Page-Rank

La centralidad Page-Rank es una medida creada por Google en el año 1999, basada en la centralidad del eigenvector. En esencia esta medida es la mejora del eigenvector para poder utilizarla en grafos dirigidos.

El page-rank se creó para ordenar las páginas web a mostrar por el motor de búsqueda para que las más relevantes apareciesen en las primeras posiciones. Para ello tiene en cuenta el número de enlaces entrantes de las demás páginas web hacia

la página de la que se está estudiando la centralidad, y la importancia relativa de la web de la que sale el enlace.

Cada enlace que apunta a un determinado nodo (es decir, los arcos entrantes) se pueden asemejar como un voto que da el nodo saliente a éste. Pero no es lo mismo que te de un voto un nodo con un número de enlaces muy pequeño, es decir, con poca importancia en la red, a que te lo dé el nodo con más enlaces de todo el grafo.

Para poder contemplar todo esto, la fórmula del page-rank tiene en cuenta los valores de page-rank de los nodos vecinos al nodo sobre el que se está calculando:

$$pagerank_v = (1 - d) + d \sum_i^n \frac{pagerank_i}{outdegree_i}$$

Ecuación 7. Medida de Centralidad Pagerank

Donde v es el nodo sobre el que se está calculando el pagerank, n es el número de nodos incidentes al nodo v, i va recorriendo los nodos incidentes a v y calculando la división entre su pagerank y su número de nodos salientes, y d es un factor de amortiguación con un valor de 0 a 1.

En el cálculo original, d representa la probabilidad de que un determinado usuario siga el link que enlaza las páginas v e i. Suele rondar el 0,85.

Como se puede ver en el grafo Kite, al ser no dirigido, el valor del nodo más central que se obtiene al utilizar la medida del page-rank es el mismo que en el eigenvector.

A	B	C	D	E	F	G	H	I	J
0.102	0.102	0.079	0.147	0.079	0.129	0.129	0.095	0.086	0.051

Tabla 6. Resultados de Pagerank para Kite

D es el nodo más central ya que al ser un grafo no dirigido y las aristas no tener pesos, prima el número de aristas incidentes al nodo, al igual que en el caso del cálculo del eigenvector.

3 SELECCIÓN ÓPTIMA DE GRUPOS DE DIFUSIÓN

Recapitulando lo dicho en los dos apartados anteriores, se está buscando encontrar en una red social al individuo o conjunto de individuos que permitan alcanzar la mayor difusión de una campaña en el ámbito del marketing viral. Para localizar a estos nodos se puede optar por buscar a los más centrales teniendo en cuenta las medidas estructurales de centralidad individual vistas en el apartado 2.3, o bien se puede optar por el desarrollo de un proceso dinámico de difusión del mensaje a través de la red. Se va a seguir por la segunda opción, la del desarrollo de un proceso dinámico para la selección del grupo óptimo, ya que es independiente de la estructura de la red, al contrario que las medidas de centralidad individual. En este desarrollo es en lo que va a consistir el apartado 3.

El problema que se va a intentar resolver en este caso es un **problema de difusión máxima**. Es decir, el problema consiste en buscar los nodos que permitan que la difusión de una campaña de marketing y la adquisición del producto ofertado por la campaña sea la máxima posible dentro de la red en la que se realice el estudio. Todo ello teniendo en cuenta ciertas condiciones, como por ejemplo, límites presupuestarios, el número de nodos que se quiere seleccionar inicialmente, etc. Para ello se tendrán en cuenta el nivel de influencia de cada uno de los integrantes de la red con respecto a sus conocidos, y se obtendrán los que sean más influyentes, incorporando el proceso dinámico de difusión de innovaciones a través de la red, que es de naturaleza estocástica.

Para medir la influencia de un nodo o grupo de nodos de un grafo se van a tener en cuenta dos criterios: el número esperado de nodos alcanzados y los periodos de tiempo esperados que tarda en conseguirlo. Antes de detallar los criterios empleados, se van a describir en la sección 3.1 los modelos dinámicos de difusión de innovaciones más comunes en la literatura. A continuación se va a describir en la sección 3.2 el problema de optimización que se va a resolver, analizando los procedimientos heurísticos implementados para su resolución. Por último, en el apartado 3.3 se va a describir la experiencia computacional, explicando los grafos seleccionados para las diferentes pruebas y los resultados más representativos.

3.1 Modelos dinámicos basados en la Difusión de Innovaciones

Los modelos dinámicos basados en la difusión de innovaciones son de los más utilizados dentro del sector del marketing a la hora de trabajar con redes sociales. El uso de los mismos permite buscar aquellos individuos más influyentes de una red no sólo teniendo en cuenta los datos estáticos del grafo que modela a la misma (las medidas estructurales de centralidad individual vistas en el apartado 2), sino que permite tener en cuenta la interacción entre los nodos de la red a lo largo del tiempo.

Los modelos dinámicos pueden ser de dos tipos:

- ❖ Los *modelos dinámicos progresivos* son aquellos en los que cuando se adquiere el producto, ya no se abandona su uso. Por ejemplo, en el caso de campañas de marketing, una vez que se consigue convencer al consumidor y adquiere el producto, ya no puede dejar de tenerlo.
- ❖ Los *modelos dinámicos no progresivos* son aquellos en los que, aunque se consiga convencer a un individuo de adquirir el producto, este puede abandonarlo después. El ejemplo más claro de los modelos no progresivos es el del modelo del votante. Un individuo puede cambiar su intención de voto en función de si le gusta más un determinado partido u otro. Aunque se le convenza de votar a uno de los partidos, esta decisión puede cambiar con el paso del tiempo.

Los modelos de difusión de innovaciones son aquellos en los que se busca dentro de una red social el grupo de nodos que al ser activados inicialmente tengan una difusión esperada mayor. Dado que se modela la red social como un grafo, los individuos como nodos y las relaciones entre ellos como aristas (o arcos en el caso de que la relación sea direccionada), los nodos más importantes en términos de difusión de innovaciones serán aquellos que sean capaces de alcanzar y convencer al mayor número de nodos posibles del grafo.

Antes de entrar en los tres modelos de difusión de innovaciones más conocidos se va a explicar con el ejemplo más simple como funciona un modelo de difusión. Para ello, definimos:

- ❖ *Nodo activado*: nodo que ha aceptado ya la innovación o ha adquirido un determinado producto de una campaña. Los nodos elegidos para iniciar la difusión se activarán desde el principio.
- ❖ *Nodo desactivado*: opuesto al activado, no ha aceptado la innovación.
- ❖ *Nodos vecinos*: nodos que se encuentran conectados por una arista. Un nodo desactivado solo podrá activarse cuando tenga al menos un nodo activado como vecino. Si se trata de un grafo dirigido para que un vecino i pueda influir sobre un nodo j debe de existir el arco (i,j) .

Teniendo esto en cuenta, el ejemplo más sencillo de modelo de difusión consistiría en un modelo **progresivo** de una red. Se parte de un conjunto de nodos activados inicialmente, denotado por S , que serán el grupo elegido para llevar a cabo la difusión. En cada periodo de tiempo, si existe en el grafo un nodo desactivado que tenga por vecino a un nodo activado, el estado del primero cambiará a activado en el siguiente periodo. Como es un modelo progresivo, ningún nodo que se haya conseguido activar volverá al estado de desactivado. Seguirán transcurriendo los periodos de tiempo hasta que el grafo llegue a un estado estable, es decir, que de un periodo al siguiente no se consiga convencer a más nodos para que pasen al estado de activado.

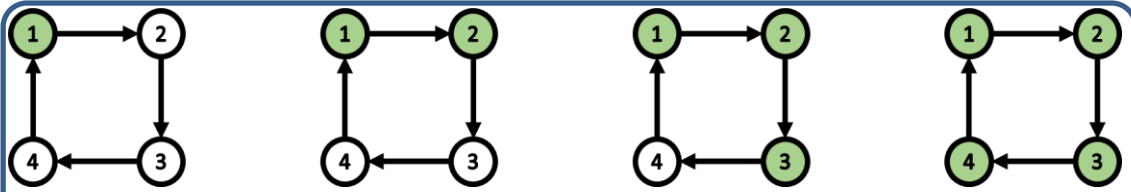


Ilustración 9. Difusión en modelo progresivo.

Descripción: se parte del nodo 1 como nodo inicial. En cada etapa se va influenciando/convenciendo a un nodo. Al final todos los nodos se activan (son convencidos).

El ejemplo más simple para un modelo **no progresivo** es similar al anterior, solo que en este caso un nodo activado con vecinos desactivados sí que podrá volver al estado de desactivación. En este caso, para llegar al estado estable se tiene que dar que ningún nodo del grafo active ni desactive a cualquiera de sus vecinos. Puede darse el caso de que no se alcance el estado estable, en un número finito de periodos, que explica que el tiempo de difusión en estas redes pueda ser infinito (Apartado 3.2). No obstante, bajo ciertas condiciones, se puede garantizar que alcanza un estado estacionario en el límite.

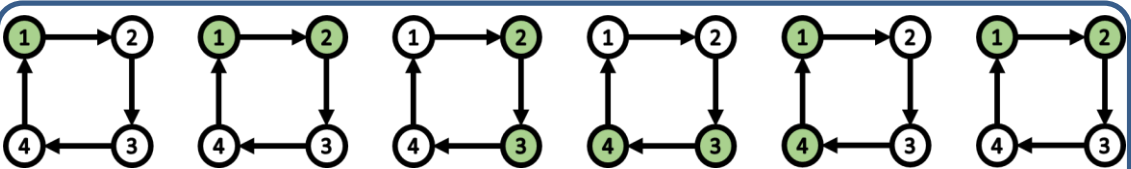


Ilustración 10. Difusión en modelo no progresivo. Ejemplo de no estabilidad

Descripción: se parte del nodo 1 como nodo inicial. En cada etapa se va influenciando/convenciendo a un nodo, pero a partir de la segunda etapa los nodos que están desactivados también desactivan a otros. No se alcanza la estabilidad (tiempo infinito de difusión).

Partiendo de estos modelos básicos se va a pasar a estudiar los tres más conocidos: el Modelo General de Umbral (General Threshold Model), el Modelo de Cascada Independiente de Kleinberg y el Modelo del Votante de Ligget.

3.1.1 Modelo General de Umbral (General Threshold Model)

Los modelos de umbral introducidos en Kempe, Kleinberg y Tardos (2003, 2005) [23] [24], y que tienen su base en el modelo de Granovetter (1978) [25], son el siguiente paso lógico a dar tras el modelo simple explicado. En el caso simple se está teniendo en cuenta que todos los nodos son iguales dentro de la red, y lo único que influye es el número de nodos a los que puede acceder. Sin embargo no es un ejemplo realista, ya que no todos los individuos tienen la misma importancia ni son capaces de “convencer” a sus vecinos con la misma facilidad. Además, no todos los individuos son igualmente fáciles de convencer, puede haber casos de individuos que sean propensos a adoptar innovaciones o a adquirir nuevos productos, y otros que sean reacios a ello. Para la mejora de todo esto se introducen los umbrales en el modelo de difusión.

El modelo de difusión de umbral más simple es el modelo de **umbral lineal**. En él cada nodo tiene un valor de umbral entre 0 y 1 (que se sortea aleatoriamente siguiendo una distribución $U(0,1)$ antes de iniciar el proceso de difusión), que identifica su predisposición a aceptar un producto o innovación. Es decir, es el umbral a partir del cual el nodo cambiaría de estado. Cuanto mayor sea este valor, más reacio será a aceptar el producto. Por su parte, las aristas también tendrán un peso, que tendrá que ver con el nivel de influencia ejercido entre los nodos que conectan.

Este modelo tiene más sentido aplicarlo sobre grafos dirigidos, ya que la influencia que ejerce un nodo sobre otro no tiene por qué ser la misma para ambos. Es más, no tiene ni por qué darse el caso de que la influencia sea recíproca, es decir, que si un nodo A influye sobre otro B, B puede que no influya sobre A. En este tipo de modelos, los arcos indican el peso de la influencia que ejerce el nodo del que sale el arco sobre el nodo al que llega.

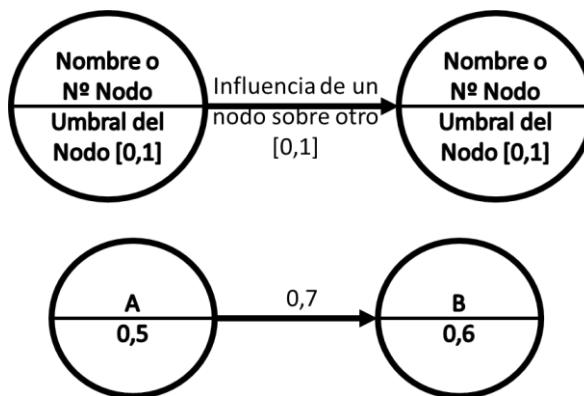
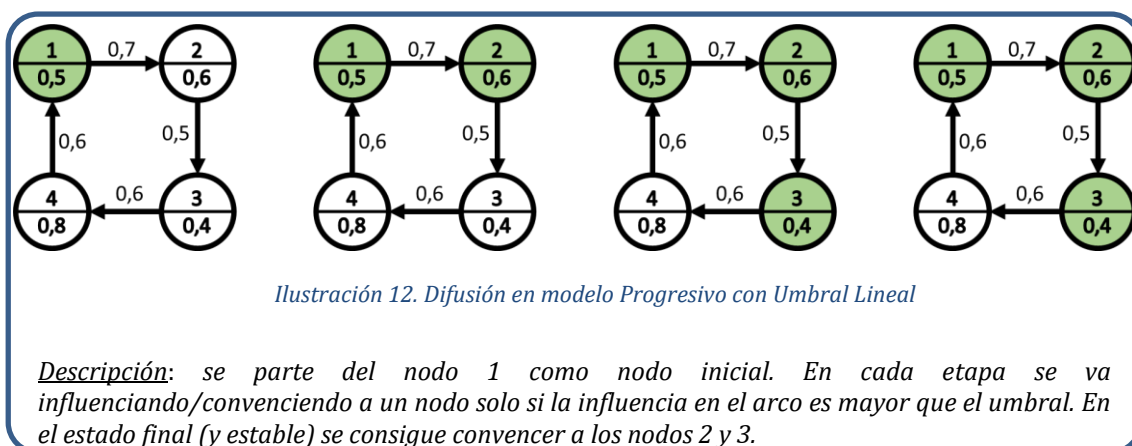


Ilustración 11. Nomenclatura en modelo de Umbral Lineal

El desarrollo de este modelo en su versión progresiva es como sigue. Se activan inicialmente un número concreto de nodos correspondiente al grupo del que se quiere conocer la difusión esperada. En cada periodo de tiempo, los nodos intentarán influir a sus vecinos para que cambien su estado al de activado. Si la suma de las influencias que los nodos activados ejercen sobre un determinado nodo es mayor o igual que el valor de su umbral, el nodo pasará a activado, y en caso contrario, permanecerá desactivado. Al encontrarnos en un modelo progresivo, no se podrá volver al estado inicial una vez activado. Se seguirán actualizando los estados de los nodos hasta que el grafo entre en un estado estable en el que ya ninguno de los activados consiga convencer a sus vecinos.



Pero este modelo es una de las posibles versiones que se engloban dentro del modelo de umbral de Granovetter, o Modelo General de Umbral. El modelo lineal de umbral asume que las influencias que ejercen los vecinos de un determinado nodo son estrictamente aditivas, sin embargo el modelo general de umbral va más allá.

Al igual que en el caso del umbral lineal, los nodos tienen un umbral asociado generado aleatoriamente siguiendo una distribución $U(0,1)$, y cada arista tiene un peso que se corresponde con el nivel de influencia que ejerce el nodo de salida del arco sobre el nodo de llegada. Sin embargo, los pesos que ejercen los nodos vecinos sobre un determinado nodo no tienen por qué sumarse para intentar alcanzar el umbral, sino que pueden seguir cualquier tipo de función g_v , siempre y cuando cumpla con las siguientes condiciones:

- ❖ Para cualquier conjunto de nodos A, debe existir un valor de $g_v(A)$ comprendido entre el 0 y el 1 que será el que se compare con el umbral del nodo v.
- ❖ La función g_v tendrá que ser monótona. Es decir, que la influencia de los nodos debe ir en aumento (o mantenerse constante) a medida que un nuevo nodo vecino adopte la innovación. De este modo si tenemos dos conjuntos de nodos vecinos, A y B, que cumplen que $A \subseteq B$, entonces la función $g_v(A)$ tendrá que ser siempre menor o igual que la función $g_v(B)$ ($g_v(A) \leq g_v(B)$).

Al utilizar funciones monótonas, este modelo no está contemplando el caso de que haya influencias negativas. Es decir, que un nodo esté menos interesado en adoptar una innovación en el caso de que uno o varios de sus vecinos la hayan adoptado anteriormente.

Tanto el modelo general de umbral como el modelo de umbral lineal explicados están centrados en los modelos de difusión progresivos, en los que si un nodo pasa de un estado de desactivado a activado, no puede volver atrás, y permanecerá activado el resto del tiempo en que se desarrolle el modelo. Pero se pueden llegar a calcular las versiones no progresivas a partir de estos. En Kempe, Kleinberg y Tardos (2003) proponen la siguiente aproximación para ello:

Dado un grafo no progresivo G que se va a ejecutar T veces, se puede crear un nuevo grafo G' compuesto por T copias del grafo G: $G_1, G_2, G_3, \dots, G_T$. Cada nodo v tendrá una copia v_i en cada grafo G_i , y se añadirán aristas desde los nodos vecinos

en G_{i-1} (u_{i-1}) hasta el nodo v_i . De esta manera los nuevos vecinos de v_i serán aquellos vecinos que se encuentran en el periodo de tiempo inmediatamente anterior. Con esto se conseguiría definir las reglas y niveles de influencia existentes en el grafo G para G' . De este modo, al estudiar el grafo G' de forma progresiva se podrán **extrapolar sus resultados como los datos del grafo G no progresivo**.

3.1.2 Modelo de Cascada Independiente de Kleinberg

El Modelo de Cascada es un modelo equivalente al Modelo General de Umbral, sólo que tiene también en cuenta el tiempo en el que cada vecino de un determinado nodo adopta la innovación. En este modelo se considera que el nivel de influencia de un nodo es inversamente proporcional al tiempo que lleva activado. Es decir, no tienen el mismo nivel de influencia los vecinos que acaban de adoptar un determinado producto o innovación, que aquellos que ya llevan varios periodos de tiempo activados. [26]

Al contrario que en el caso del modelo de umbral, los nodos no disponen de un umbral aleatorio a partir del cual aceptarían la innovación, ni las aristas tienen el peso correspondiente a la influencia que ejerce un nodo sobre otro.

El modelo de cascada, al igual que en todos los modelos de difusión de innovaciones, parte de un conjunto de nodos activados inicialmente. Cada nodo activado del grafo tendrá una única oportunidad para convencer a sus vecinos. Es decir, sólo tendrá una opción para conseguir cambiar el estado de sus vecinos desactivados a activados. Para saber si un nodo puede o no convencer a otro, existe una función de probabilidad que indica las posibilidades que tiene un nodo para convencer a su vecino, teniendo en cuenta todos los nodos que ya hayan intentado convencerle antes que él. Para ello se van almacenando todos los nodos que hayan intentado convencer al desactivado y no lo hayan conseguido. Si el nodo es capaz de activar al vecino desactivado, en el siguiente periodo de tiempo éste intentará activar a sus correspondientes vecinos que todavía no hayan sido alcanzados. En caso contrario, si no consigue activarlo, el nodo pasará a formar parte del conjunto de nodos que lo han intentado y no lo han conseguido para el cálculo de probabilidad del siguiente que lo intente.

El orden en el que los vecinos van intentando convencer a un determinado nodo y fallan no es influyente a la hora de calcular la probabilidad condicionada de un nuevo intento. Sin embargo, sí que influye que hayan participado unos nodos u otros. La receptividad de un nodo ante un cambio, una innovación o un producto, va a

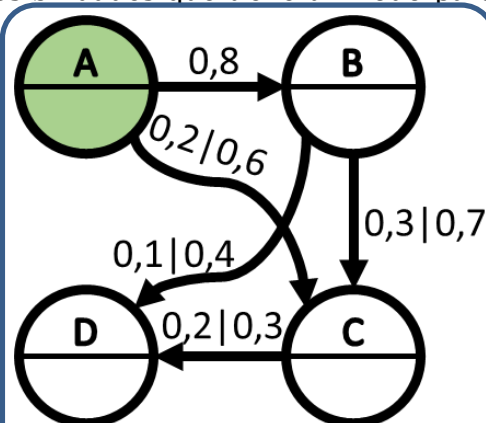


Ilustración 13. Modelo de Cascada.

Descripción: Los valores en los arcos son probabilidades de activar al nodo destino. En los arcos con dos valores, el primero se aplica cuando nadie ha intentado activar el nodo destino antes, y el otro cuando alguien ya lo ha intentado. La difusión se detalla en la ilustración 14

depender de los intentos pasados de sus nodos vecinos. En vez de tener una función g_v que tiene en cuenta los pesos de las aristas y los umbrales, se va a utilizar la función incremental de probabilidad $p_v(u, X)$, donde u es el vecino de v que está intentando convencerle de adoptar la innovación y X es el conjunto de vecinos que ya lo han intentado pero no lo han conseguido.

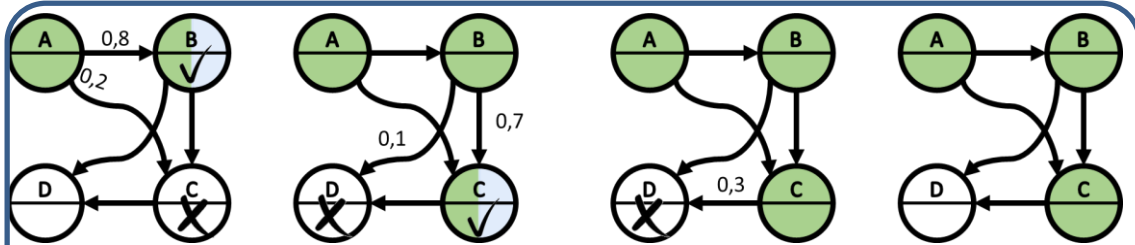


Ilustración 14. Difusión en modelo de Cascada

Descripción: se parte de la ilustración 13 como estado inicial (solo el nodo A activado). En cada etapa el nodo activado recientemente intenta activar a los que tiene conectados. En la primera etapa el nodo A consigue convencer al B (probabilidad de 0,8), pero falla al tratar de convencer al C. En la segunda etapa es el nodo B el que trata de convencer a C y D. Puesto que A ya intentó convencer a C, la probabilidad de B para convencer a C ha aumentado. B consigue convencer a C pero falla con D. En la tercera etapa C trata de convencer a D, sin éxito. El estado final del grafo son 3 nodos activados y uno desactivado. **IMPORTANTE:** esta es una de las muchas posibles difusiones, dado que es aleatorio.

Kempe, Kleinberg y Tardos (2005) demuestran que el modelo de cascada y el de umbral son equivalentes, ya que tanto la función g_v del modelo general de umbral se puede pasar a p_v del modelo de cascada independiente como viceversa.

El modelo de cascada también tiene casos especiales, entre los que se encuentra el Modelo de Cascada Independiente (Kempe, Kleinberg y Tardos, 2003). Este es un caso especial simple del modelo de cascada con rendimientos decrecientes, que tiene como característica que a medida que avanzan los periodos de tiempo, el nodo v es más difícil de convencer, ya que su probabilidad de aceptar una determinada innovación baja a medida que se le intenta convencer una vez más ($p_v(u, A) \geq p_v(u, B)$, donde $A \subseteq B$).

Otra de las características del modelo de cascada independiente es que la influencia de cada nodo vecino u sobre v es independiente de los nodos que hayan intentado influirle antes que él. La probabilidad de convencer a v en este caso sólo depende de la relación y la probabilidad asociada al nodo u , siendo esta probabilidad un parámetro que depende sólo de los dos nodos afectados. Por todo lo demás el modelo de cascada independiente es igual al modelo general de cascada.

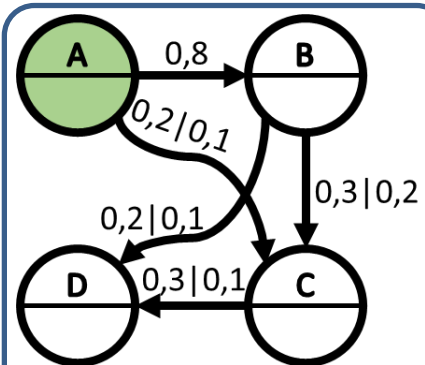


Ilustración 15. Modelo de Cascada de Kleinberg

Descripción: En comparación con la ilustración 13, en este caso las probabilidades en todos los arcos para cuando ya se ha intentado activar antes el nodo destino son menores que las iniciales

El estudio sobre las diferentes variantes de este modelo también se ha realizado sobre las versiones progresivas. Para poder realizar los cálculos en un modelo no progresivo se utiliza la misma técnica explicada para el modelo de umbral.

3.1.3 Modelo del Votante (Ligget)

Mientras que los dos modelos anteriores están relacionados entre sí, el modelo del votante es un modelo de difusión diferente. Los dos primeros modelos se centran más en el estudio de la difusión en entornos progresivos, donde no se suele volver al estado inicial. Sin embargo, como se explica más adelante, el modelo del votante tiene su base en las versiones no progresivas de los modelos de difusión, ya que estudia casos en los que se contempla el cambio de activo a no activo (es decir, contempla que se abandone la innovación).

El modelo del votante, introducido por Liggett, es el que mejor representa la difusión de opiniones que se producen dentro de las redes sociales. En cada uno de los pasos que se realizan en este modelo, cada nodo puede cambiar de opinión eligiendo aleatoriamente la opinión de cualquiera de sus vecinos, pasando a su estado en el siguiente periodo de tiempo. Si coincide que la opinión del nodo que selecciona es la misma que ya tiene, entonces se mantiene en el mismo estado. Como se puede deducir de aquí, este modelo se basa en las versiones no progresivas de los modelos de difusión y no tiene mucho sentido para modelos progresivos.

En la versión más general de este modelo, la probabilidad que un nodo tiene de cambiar de un estado a otro será igual a la probabilidad calculada mediante la regla de Laplace (número de casos favorables / número de casos posibles). Es decir, un nodo cambiará de estado con una probabilidad igual al número de nodos vecinos en el otro estado en el periodo de tiempo $t-1$ entre el número total de nodos, y permanecerá en su estado con una probabilidad igual al número de nodos vecinos en su mismo estado en el tiempo $t-1$ entre el número total de nodos.

3.1.4 Selección del Modelo

En el caso del marketing viral, este último modelo de difusión no tiene mucho sentido aplicarlo, ya que las campañas suelen consistir, en su mayoría, en que los usuarios adquieran un nuevo producto, y en estos casos no se suele volver del estado de activado (donde se ha adquirido el producto) al estado anterior de desactivado.

Entre los otros dos modelos, al elegir el Modelo de Cascada se estaría orientando el estudio sólo a aquellos casos en los que los nodos puedan convencer a sus vecinos una única vez. Esto es menos común en el ámbito del marketing viral, por eso no se va a seleccionar este modelo para su desarrollo.

Por todo esto, el modelo de difusión que se va a desarrollar para el estudio de la difusión máxima de campañas de marketing viral en las redes sociales será el Modelo de Umbral. En cuanto a la elección dentro de las diferentes versiones dentro del Modelo General de Umbral, se optará por el Modelo Lineal. Este modelo es el equivalente a que la función g_v sea igual a la suma de los pesos de las aristas, y se

realizará sobre ejemplos de modelos progresivos, en los que no se pueda volver al estado inicial una vez adoptado el nuevo estado.

3.2 Selección del grupo óptimo

En el apartado anterior se han definido diferentes procesos dinámicos de difusión de innovaciones y se ha seleccionado el Modelo Lineal de Umbral para nuestro estudio. A continuación se describe el problema que se va a resolver, que consiste en un problema de difusión máxima y se desarrolla la función objetivo que se ha seleccionado como más adecuada para el problema contemplado, tanto por su complejidad computacional, como por la adaptación al entorno del marketing viral.

Para medir la influencia de un nodo o grupo de nodos de un grafo se van a tener en cuenta dos criterios: el número esperado de nodos alcanzados y los periodos de tiempo esperados que tarda en conseguirlo:

- ❖ El **número esperado de nodos alcanzados** se medirá dejando que la red evolucione un tiempo indefinido hasta que no se alcancen nodos nuevos (el grafo llegue a un estado estable en el que la difusión se pare). Se buscará *maximizar* este criterio. Los valores que puede tomar están comprendidos entre los siguientes:
 - *Mínimo*: El número de nodos iniciales, es decir, los seleccionados en el grupo inicial. Este caso se dará si ninguno de ellos convence a nadie.
 - *Máximo*: El número de nodos total del grafo.
- ❖ El **tiempo esperado de difusión** no tiene por qué corresponderse con una medida de tiempo concreta (horas, días, semanas), sino que tienen en cuenta los cambios de estado del grafo. El tiempo de difusión se corresponderá con el instante en el que no se alcancen nuevos nodos (la difusión se pare). Se buscará *minimizar* este criterio. Los valores que puede tomar están comprendidos entre los siguientes:
 - *Mínimo*: 0, si ningún nodo inicial consigue convencer a nadie.
 - *Máximo*: Número de nodos del grafo menos el número de nodos iniciales en el caso de modelos de difusión progresivos, e ∞ en modelos de difusión no progresivos.

Criterios de Influencia	Mínimo	Máximo	Optimización
Nº esperado Nodos Alcanzados	Nº nodos iniciales	Nº nodos red	Maximizar
Tiempo esperado de difusión	0	(Nº nodos red – Nº nodos iniciales) ∞	Minimizar

Tabla 7. Espacio factible de soluciones y tipo de optimización para métricas de influencia

Teniendo en cuenta los dos criterios se tiene un problema **biojetivo**. Se resolverá uniendo las dos funciones objetivo en un único valor de influencia a través de una **función biobjetivo**. Esta va a ser la utilizada para encontrar la solución del problema de difusión máxima. La función biobjetivo puede escribirse de la siguiente forma:

$$\max Z^* = w_i * \sigma(S) + w_j * (-t(S))$$

Ecuación 8. Función Biobjetivo a maximizar

Donde w_i y w_j son los pesos con los que se va ponderar la importancia de una u otra métrica, $\sigma(S)$ se corresponde con el número esperado de nodos alcanzados por un grupo de nodos iniciales S determinado, y $t(S)$ se corresponde con el tiempo de difusión esperado de dicho grupo.

Como la optimización mediante el primer criterio era de maximización y la del segundo de minimización, se ha optado por cambiar el signo al tiempo de difusión, de tal forma que así la combinación de ambas deba maximizarse.

Por su parte, los pesos de cada uno de los criterios permiten dar prioridad a uno u otro en función de cuál se considere más relevante a la hora de calcular el grupo óptimo.

Pueden existir campañas de marketing que lo único que les interese sea el número de nodos alcanzados, y no les importe nada el tiempo de difusión, porque sea una campaña a largo plazo. En estos casos el peso w_i sería igual a 1 y el peso w_j igual a 0, y lo único que se estaría teniendo en cuenta es el grupo que maximice el número de nodos alcanzados en la red.

Por el contrario, puede darse el caso de una campaña a muy corto plazo (como puede ser una campaña de un producto en temporada de rebajas), en la que lo que interesa es que el tiempo de difusión sea lo más rápido posible. En este caso sería al revés, el peso w_i sería igual a 0 y el peso w_j igual a 1, y se optimizaría la función del tiempo. Eso sí, para los modelos de difusión en los que se minimiza el tiempo hay que especificar como una de las restricciones del problema una cobertura mínima, para evitar que se quede con aquellos que tardan un tiempo igual a 0, es decir, con los grupos que no llegan a difundirse.

El problema a desarrollar va a primar **el número de nodos alcanzados sobre el tiempo de difusión**. Es más, el tiempo de difusión se tendrá en cuenta sólo en el caso de que existan dos grupos que lleguen al mismo número de nodos, eligiendo aquel que lo haga en menor tiempo. Para ello, se asigna un valor a w_i igual al número de nodos del grafo y el w_j va a ser igual a 1. Como se verá más adelante, el modelo de difusión que se va a utilizar para el problema es un modelo *progresivo*, de forma que el tiempo en el peor de los casos será igual al número de nodos totales menos el número de nodos seleccionados. Por este motivo, al multiplicar el número de nodos alcanzados por el número total de nodos, siempre va primar el hecho de alcanzar un nodo más que el de mejorar en cualquier medida el tiempo.

El recorrido de la función biobjetivo (rango de valores que toma) depende de los pesos que se asignan a cada una de las métricas. Teniendo en cuenta los pesos que se han especificado para el problema de difusión propuesto, el espacio de soluciones será:

- ❖ *Mínimo*: El número de nodos iniciales multiplicado por el número de nodos totales, ya que el tiempo será igual a 0.
- ❖ *Máximo*: El número de nodos totales al cuadrado. (Sólo se dará este caso si se elige como grupo inicial a todos los nodos del grafo, siendo el tiempo como consecuencia igual a 0).

En la siguiente tabla se resume el problema que vamos a abordar. En cuanto al conjunto de grupos de búsqueda, que determinan el conjunto factible del problema de optimización, vamos a resolver el problema de la búsqueda de grupos k-óptimos.

Modelo de Difusión	Modelo de Umbral Lineal
Conjunto factible	Grupos de tamaño k
Búsqueda del óptimo	Difusión máxima, mínimo tiempo en empates

Tabla 8: Elección de modelos

El siguiente paso a realizar es desarrollar el programa que permita seleccionar el grupo óptimo para diferentes redes sociales teniendo en cuenta todo lo definido anteriormente. Para ello, partiendo de la base de que se trata de un problema combinatorio de complejidad NP-dura, se aplicarán **heurísticos**, reducciones de la magnitud del problema basadas en la relajación de las condiciones del problema inicial (ej. Probar solo en la mitad de los nodos del grafo en vez de en su totalidad). Aunque la solución obtenida mediante heurísticos pueda ser peor que la que ofrecen métodos más exhaustivos, el tiempo que se tarda en llegar a dicha solución será menor. De esta manera se podrá optar entre una solución costosa en tiempo pero con mayor precisión, y una solución menos precisa pero más rápida, comprobando los resultados de cada una en todas las redes sociales de prueba, estableciendo un **compromiso entre precisión y tiempo de ejecución**. A continuación se describe cada uno de los métodos desarrollados y se hace un estudio comparativo de los mismos aplicándolos a redes sociales reales y a redes sociales simuladas. Antes de ello se describe lo que ya hay hecho.

3.2.1 Descripción de lo que ya hay hecho

A día de hoy existen varios programas que realizan búsquedas de grupos óptimos. Las grandes compañías tendrán en cuenta sus propios algoritmos a la hora de elegir grupos en redes sociales. Google es el creador y utiliza una de las medidas de centralidad estructural individuales (el page-rank) para ponderar las distintas páginas web de cada búsqueda y ordenarlas y presentarlas en función de esta medida a los usuarios. Estos programas pueden ser privados de cada empresa, o ser públicos, como es el caso de algunos paquetes del lenguaje de programación R.

R va a ser el **lenguaje de programación seleccionado** para desarrollar el modelo de difusión. Por este motivo es interesante conocer qué medidas de difusión y centralidad se encuentran ya implementadas dentro del mismo. [27] [28]

Aunque no se ha encontrado ningún paquete dentro de R que realice directamente las medidas de difusión de innovaciones vistas en el punto anterior, sí que se han encontrado paquetes para el estudio de medidas estructurales de centralidad individual. Algunos de estos ejemplos serían los paquetes “*sna*” y “*keyplayer*”. Ambos tienen funciones que devuelven los valores para las diferentes medidas estructurales de centralidad individual de cada nodo del grafo, o el resultado del valor obtenido para un grupo de nodos dados.

Más concretamente, el paquete *keyplayer* permite buscar grupos óptimos de diferente tamaño dentro de un grafo teniendo en cuenta las medidas de centralidad individual. Esta función, llamada “*kpset*” va seleccionando los nodos en función de la medida de centralidad que se le indique, e irá comprobando, realizando un procedimiento de intercambios, si existen nodos que permitan mejorar la combinación que se ha ido eligiendo, en búsqueda de un óptimo local. El problema que tiene esta función es que es muy costosa en cuanto al tiempo de ejecución. [29]

Se utilizará el “*kpset*” para comparar el grupo óptimo que devuelve con los resultados obtenidos por el modelo de difusión desarrollado.

Otra de las razones por las que se ha elegido el lenguaje de programación R es debido a la calidad de generación de números pseudoaleatorios.

3.2.2 Estimación de la difusión de un determinado grupo

Aunque la estructura de la red sea conocida, como el modelo dinámico es estocástico, la función objetivo es una esperanza para la que no tenemos una expresión cerrada general. Para solucionar este problema, dado un grupo S de nodos inicialmente activos, se va a optar por realizar una simulación de Montecarlo para estimar el número medio de nodos alcanzados y el tiempo medio de difusión de dicho grupo S.

La **simulación de Montecarlo** consiste en asignar a aquellos datos desconocidos valores aleatorios y realizar los cálculos con ellos. Para que el estimador sea correcto, esta simulación se repetirá un número n de veces lo suficientemente grande (Análisis exhaustivo en Anexo 7.2) y se estimarán mediante la media de los resultados obtenidos.

Para desarrollar el modelo lineal de umbral (es decir, estimar $\sigma(S)$, para un grupo S) primero se va a mostrar la función que realiza esta estimación.

Threshold

En primer lugar, se va a desarrollar una función en la que se calculan los nodos alcanzados y el tiempo de difusión transcurrido hasta el estado estable para un grupo dado. Es el código que realmente implementa el Modelo Lineal de Umbral,

porque no busca el grupo que dé la mayor difusión, sino que sirve para estimar la difusión esperada de un grupo dado, que es la base del modelo. Es la función de partida para el resto de las funciones que se van a implementar.

Aunque no seleccione a un grupo de nodos óptimo, sirve para calcular otra serie de medidas que van a ser útiles a la hora de extraer conclusiones sobre el funcionamiento del modelo de difusión y su comparación con otros modelos:

- ❖ Se puede utilizar esta función para calcular las contribuciones marginales de un nodo a un determinado grupo. Para ello se puede comparar el valor esperado que se obtiene con el grupo S , y el valor esperado alcanzado añadiendo el nodo i al grupo, $S \cup \{i\}$.
- ❖ También se puede utilizar para realizar la comparación del modelo de difusión con respecto a los grupos obtenidos mediante las medidas de centralidad individual. Para ello se pasa por parámetro cada grupo a la función, y se contrasta el valor obtenido de este grupo con el obtenido mediante las funciones que implementan el modelo de difusión.
- ❖ Otro de los usos que se le puede dar, además de obtener la media (la estimación de $\sigma(S)$), es el de calcular la cuasivarianza muestral, que nos permite saber el número mínimo de veces que se tiene que repetir el cálculo aleatorio del umbral en la realización de la simulación de Montecarlo para no exceder un error máximo preestablecido.

Para mejor comprensión de la función que se va a desarrollar, se va a añadir un fragmento de pseudocódigo explicando los pasos más relevantes. En este caso las variables iniciales son:

- ❖ *Vector de nodos activados*: Vector de longitud n , siendo n igual al número de nodos totales, donde cada posición se corresponde con un nodo del grafo. Este vector se inicializará con todas las posiciones a 0. Cada posición almacenará un 1 si el nodo se encuentra en estado de activado, y un 0 en caso contrario.
- ❖ *Nodos alcanzados*: Variable donde se va almacenando el número de nodos activados en cada una de las iteraciones indicadas por loops. Para poder estimar el número esperado de nodos activados se hará la media de esta variable dividiéndola por el número de loops (simulación de Montecarlo).

- ❖ *Tiempo de difusión:* Variable donde se va almacenando el tiempo transcurrido en cada una de las iteraciones. Para poder estimar el tiempo esperado de difusión se hará la media de esta variable dividiéndola por el número de loops.

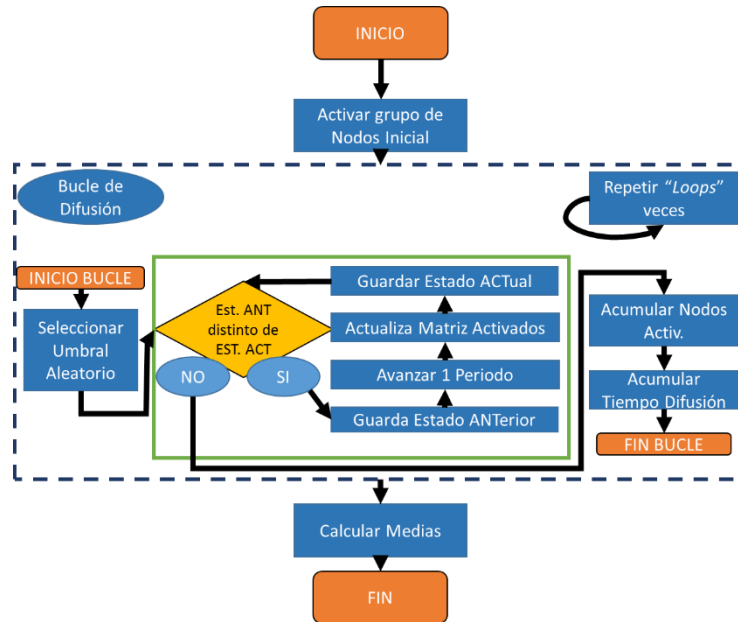


Ilustración 16. Diagrama de flujo de procesos de la función threshold

```

Threshold(grafo, grupo, loops = 10, ...)
    Activar nodos grupo en activados
    Hacer tantas veces como indique loops (MONTECARLO)
        Seleccionar umbral nodos aleatoriamente
        Mientras (estado anterior != estado actual)
            Guardar estado anterior
            Actualizar matriz activados
            Avanzar 1 periodo de tiempo transcurrido
            Guardar estado actual
        Sumar a nodos alcanzados los nodos activados
        Sumar a tiempo difusión el tiempo transcurrido
    Calcular óptimo con media de nodos alcanzados y tiempo difusión
    Devolver óptimo
    
```

Pseudocódigo 1. Threshold

El bucle de difusión será el mismo en el resto de las funciones y los heurísticos desarrollados para el modelo.

3.2.3 Funciones de búsqueda del grupo óptimo

Threshold Greedy

Esta función es la primera aproximación a la búsqueda de grupos k-óptimos utilizando el modelo lineal de umbral. Recibe por parámetro el tamaño que tiene que tener el grupo óptimo, y devuelve cuál es el grupo seleccionado, cuantos nodos se han activado en media, el tiempo de difusión medio hasta llegar a un estado estable y el tiempo de ejecución de la función. Este es el procedimiento propuesto por Kempe, Kleinberg y Tardos (2003, 2005). Basándose en un resultado de Nemhauser, Wolsey y Fisher (1978) [30], demuestran que dicho algoritmo garantiza

una difusión de al menos 0.63 veces la difusión generada por el conjunto de tamaño k -óptimo. Dicha demostración se basa en la submodularidad de la función $\sigma(S)$ [30].

$$\sigma(S \cup i) - \sigma(S) \geq \sigma(T \cup i) - \sigma(T), \quad \forall S \subseteq T$$

Ecuación 9. Función Submodular

Las funciones submodulares se pueden ver (simplificando e interpolando a un ejemplo sencillo) como una función convexa. Esto quiere decir que en la búsqueda del máximo, cada nuevo paso que se da hacia él (que en nuestro problema equivale a introducir un nuevo nodo en el grupo elegido o cambiar uno de ellos por uno mejor) obtiene una mejora más pequeña en cada periodo de difusión, tal y como se puede ver en la ilustración contigua. Siendo un poco más rigurosos, el modelo de umbral lineal y la función biobjetivo elegida conforman una función submodular porque la **contribución marginal de un nodo** (el número de nodos que se alcanza con él incluido en el grupo inicial menos el número de nodos que se alcanza sin él en el grupo inicial) **se reduce a medida que aumenta el número de nodos en el grupo inicial**. Se puede entender como que aunque un nodo ejerce una cierta influencia sobre sus nodos vecinos, ésta se torna inútil cuando nodos más lejanos consiguen convencer sin su ayuda a los vecinos del nodo de referencia.

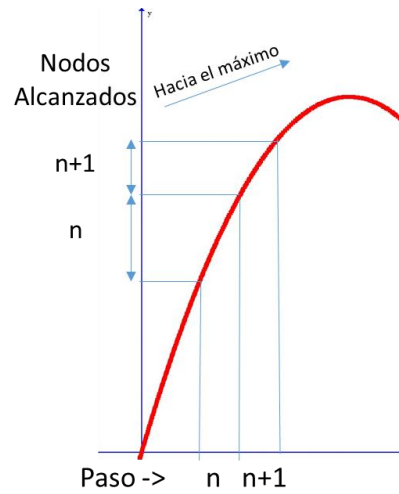


Ilustración 17. Simplificación de función submodular a función convexa

Con todo ello, el funcionamiento de *Threshold Greedy* (en adelante *Greedy*) se basa en ir eligiendo los mejores nodos uno a uno. Tras realizar los cálculos que realiza el threshold para la obtención del número esperado de nodos alcanzados en cada uno de los nodos del grafo, $\sigma(i)$, para todo nodo i , se contrasta cuáles son los resultados de cada uno de ellos y se seleccionará aquel que lo maximice. En caso de que haya empate entre varios nodos se seleccionará el que tenga un mayor número de enlaces salientes (mayor out-degree), y si el empate persiste, al que aparezca en primer lugar en el vector de nodos del grafo (ya que así es cómo lo realiza R). Una vez seleccionado el primer nodo, sea i_1 , se añadirá al grupo óptimo, y se volverá a empezar el proceso para buscar el segundo. Pero esta vez tendrá en cuenta que el nodo que ya ha entrado en el grupo óptimo también estará activado, ayudando a los demás a activar al resto de los nodos del grafo. De esta manera se incluirá al nodo i_2 que contribuya más a la difusión esperada de i_1 . Este proceso se repetirá sucesivamente hasta completar el número de nodos que tienen que formar parte del grupo óptimo.

Las variables iniciales de la función en este caso son las siguientes:

- ❖ *Nodos alcanzados y Tiempo de difusión*: Estas dos variables almacenarán la misma información que en el caso del threshold, solo que en este caso serán vectores de longitud n , donde cada posición se identifique con los resultados correspondientes a cada uno de los nodos.

- ❖ *Matriz de nodos activados*: A diferencia del threshold, los nodos activados se almacenarán en una matriz, donde cada fila se corresponderá con el vector de nodos activados de cada nodo. Esta matriz de tamaño $n \times n$, donde n es igual al número de nodos del grafo, se inicializará con todas las posiciones a 0, salvo las de la diagonal principal, que serán iguales a 1. En la matriz si la posición a_{ij} es igual a 0 significa que el nodo i no ha llegado a activar al nodo j . Y es 1 en caso contrario. De esta manera, cada fila i va a indicar el número de nodos que han llegado a ser activados por el nodo i .

Si se quiere que un nodo j aparezca activado en todo momento (como es el caso de los nodos activados inicialmente) lo que habrá que hacer es poner todos los valores de la columna j a 1 en la matriz.

- ❖ *Grupo*: Vector de longitud igual al número del grupo, con todas las posiciones inicialmente a 0 donde se van a ir almacenando los nodos seleccionados como óptimos.

```
ThresholdGreedy(grafo, n° grupo, loops = 10, ...)
    Hacer tantas veces como indique n° grupo
        Activar nodos seleccionados de grupo óptimo en activados
        Eliminar nodos seleccionados de filas en activados
        Realizar Bucle de Difusión de Threshold
        Seleccionar nodo óptimo
        Reiniciar matriz de activados
    Devolver grupo
```

Pseudocódigo 2. Threshold Greedy

Este algoritmo es muy simple, se encarga de seleccionar el mejor nodo partiendo de un grupo ya dado. Pero en ningún momento se plantea si alguno de los nodos que ya han entrado en el grupo podría ser sustituido por otro cuya contribución marginal al grupo sea mayor que la del que ya estaba elegido. Teniendo este problema en cuenta se va a implementar la función Greedy por pasos.

Threshold Greedy Stepwise

La función Greedy por pasos se basa en el procedimiento de regresión por pasos utilizado en la selección de variables para el análisis de regresión lineal múltiple. Esta función tiene en cuenta la posibilidad de que existan individuos fuera del grupo óptimo cuyas contribuciones marginales con un subgrupo del óptimo sean mayores que las que aporta el grupo óptimo ya seleccionado. Es decir, que al cambiar un solo nodo del grupo inicial se obtengan mejores resultados. Es la función más exhaustiva implementada a la hora de buscar el grupo de nodos óptimo.

En esta función lo primero que se realiza es una llamada a la del Greedy para obtener un grupo óptimo inicial sobre el que trabajar. A partir de ese grupo, se realizan las posibles combinaciones de sus k integrantes seleccionados en subgrupos de $k-1$ nodos y se almacenan en una matriz. Esta matriz se irá recorriendo fila a fila, y en cada iteración se realizará el mismo cálculo para encontrar el nodo que optimice la función objetivo utilizado en el Greedy, teniendo como grupo ya

seleccionado los valores almacenados en la fila a la que corresponda la iteración. Una vez se obtienen los resultados óptimos de todas las combinaciones, se elige aquella que maximice la función objetivo, teniendo en cuenta que los casos de empate se resolverán del mismo modo que en el greedy. Esta operación se repetirá hasta que no se pueda mejorar más el resultado obtenido por el grupo óptimo, y se devolverá tanto el grupo elegido, como el número de nodos alcanzados y el tiempo de difusión.

Como puede darse el caso de que la operación se siga repitiendo un número elevado de veces porque se vaya mejorando la función poco a poco, se va a establecer una condición de parada en la que se especifique que, una vez se haya realizado un número mínimo de veces la búsqueda del nuevo grupo óptimo, si no se mejora la función biobjetiva en al menos un 0,1%, la ejecución del programa finalizaría y se obtendría como grupo óptimo el último calculado.

Las variables iniciales de la función en este caso son las siguientes:

- ❖ *Nodos alcanzados, Tiempo de difusión y matriz de nodos activados*: Estas variables almacenarán la misma información que en el caso del Greedy
- ❖ *Grupo*: En este caso el grupo vendrá definido por la función Greedy
- ❖ *Óptimos*: Vector de tamaño igual al número del grupo menos 1, donde se van a ir almacenando los resultados de los óptimos obtenidos para cada una de las filas de la matriz de combinaciones. Se inicializan todos los valores a 0.

```
ThresholdGreedyStepwise(grafo, n° grupo, loops = 10, ...)
  GrupoInicial = ThresholdGreedy(grafo, n°grupo, loops,...)
  Mientras (GrupoInicial != GrupoFinal) y ((1 - optimoAnterior /
    optimoActual) > 0.001)
    Calcular matriz combinaciones de k-1 del grupo k
    Hacer tantas veces como filas tenga la matriz combinaciones
      Activar nodos en fila de combinaciones en activados
      Eliminar esos nodos de filas en activados
      Realizar Bucle de Difusión de Threshold
      Seleccionar nodo óptimo
      Almacenar el grupo en su fila de combinaciones
      Almacenar el óptimo en su posición del vector óptimos
      Reiniciar la matriz de activados
    Seleccionar grupo óptimo entre las filas de combinaciones
    Actualizar el GrupoInicial al grupo óptimo
    Actualizar el optimoActual al de la posición correspondiente
    en el vector de óptimos
    Reiniciar el vector de valores óptimos
  Devolver grupo
```

Pseudocódigo 3. Threshold Greedy Stepwise

3.2.4 Heurísticos

El problema existente tanto en las funciones ya programadas en R (las medidas estructurales de centralidad individual aplicadas a la selección de grupos) como en las dos funciones de búsqueda de grupos óptimos del apartado anterior (greedy y greedy por pasos) es su coste en tiempo de ejecución.

Si se está trabajando con redes no muy grandes, este problema no resulta importante, ya que la diferencia en tiempo de ejecución entre unas y otras medidas es despreciable, y, pudiendo elegir, es mejor seleccionar aquellos métodos que hagan la búsqueda de forma más exhaustiva. Sin embargo, si se trabaja con grafos más grandes, la complejidad de cálculo es cada vez mayor, y esto supone un aumento exponencial del tiempo de ejecución.

Es para estos últimos casos para los que se van a desarrollar 3 heurísticos que permiten aligerar las restricciones del problema inicial. De esta forma se consigue llegar a soluciones del problema en un tiempo menor que en los casos anteriores. Pero también hay que tener en cuenta que al prescindir de alguna de las restricciones del problema, el óptimo al que van a llegar los heurísticos puede ser peor que el obtenido en los métodos más exhaustivos.

Los 3 heurísticos tenidos en cuenta se pueden utilizar para mejorar tanto el tiempo de ejecución de la función Greedy como el del Greedy por pasos.

Heurístico 1 – Eliminar Amigos de Convencidos

Este heurístico consiste en eliminar de la posible elección de nuevos nodos para el grupo óptimo a todos aquellos que sean “amigos” de los que ya se encuentran seleccionados en dicho grupo. Es decir, elimina del cálculo a todos aquellos nodos que tengan un arco entrante de un nodo que ya haya sido alcanzado.

Este heurístico mejora el tiempo de ejecución ya que al quitar los nodos vecinos, se está ahorrando el tiempo que se tardase en calcular sus difusiones esperadas. Y cuanto más conexo sea el grafo o cuanto mayor sea el número de nodos que tenga el grupo óptimo, mayor será la ganancia en tiempo, ya que más nodos se estarán eliminando del cálculo.

Sin embargo, su funcionamiento no es tan bueno en todos los grafos. En aquellos en los que la distribución de sus nodos sea similar al tipo estrella de estrellas, como el ejemplo de la ilustración 18, no se podrá alcanzar de ningún modo el óptimo del grafo, ya que al eliminar los amigos del nodo central, se estarían eliminando los centros de las demás estrellas, que claramente son los que tendrían que entrar en el grupo óptimo.

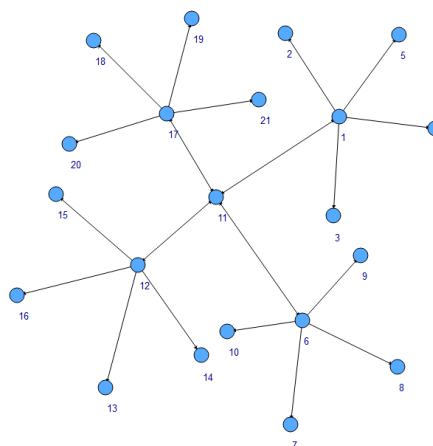


Ilustración 18. Grafo Estrella de Estrellas

El desarrollo de este heurístico para los modelos del greedy y el greedy por pasos es muy sencillo de implementar. Para el greedy basta con que por cada nuevo nodo que entre en el grupo óptimo, se eliminen del cálculo todos aquellos nodos hacia los que tenga un arco de salida. Para el greedy por pasos la única diferencia es que se tendrá que ir actualizando los vecinos que se eliminan en función del grupo que se vaya seleccionando.

Heurístico 2 – Fijar nodos del grupo óptimo del greedy

Este heurístico sólo puede utilizarse para optimizar la función del greedy por pasos, ya que se basa en el código implementado para esa función y se necesita un grupo inicial para poder realizarlo.

Este heurístico recibe por parámetro un valor numérico (`fixedGroup`), que se corresponde con el número de nodos que se van a mantener constantes dentro del grupo inicial obtenido en el cálculo del greedy. Consiste en fijar tantos nodos del final del grupo óptimo como se indique en dicho parámetro, de tal forma que al realizar las combinaciones del grupo, sólo tengan que calcularse las de los nodos “variables” del mismo.

Cuanto mayor sea el grupo que se deje fijo, menor será el tiempo de ejecución, pero también será menor la variación y optimización con respecto al greedy normal.

El problema de este heurístico es que al implementarse sólo sobre la función greedy por pasos, el tiempo de ejecución será siempre mayor que el del greedy. Es decir, que en el caso de grafos muy grandes para los que hasta el cálculo del greedy sea muy costoso, no podrá ser aplicado.

El desarrollo de este heurístico también es muy sencillo de implementar. Tras haber llamado a la función del greedy en el greedy por pasos, se seleccionan las p últimas posiciones del grupo (tantas como indique el valor pasado por parámetro) y se almacenan en otra variable. Esos nodos serán fijos para el resto de la ejecución de la función, de tal forma que a partir de aquí el código será el mismo que el del greedy por pasos, solo que la combinación de los valores del grupo se realizará para el número de nodos no fijo, y serán éstos los que se vayan actualizando.

Heurístico 3 – Reducción a nodos de mayor OutDegree

Este heurístico, al igual que el heurístico 1, va a optimizar el tiempo de ejecución eliminando nodos del cálculo de las funciones del greedy y el greedy por pasos. Sin embargo, la selección de los nodos a eliminar no va a depender de los nodos seleccionados en el grupo óptimo, sino de una medida de centralidad individual: el out-degree.

Para ello, se va a tener en cuenta la distribución del degree de todos los nodos del grafo, y se calcularán los cuartiles de la misma. El heurístico consistirá en seleccionar sólo aquellos nodos cuyo valor de out-degree (número de aristas salientes) sea mayor o igual al indicado por el cuartil que se pase por parámetro a la función. Y el resto de nodos se eliminarán del cálculo del grupo óptimo.

El parámetro que se pase a la función correspondiente al cuartil podrá tomar 5 valores, del 1 al 5:

- ❖ El 1 se corresponderá con el 0%, para el que no se eliminará ningún nodo del cálculo. Es lo mismo que calcular las funciones normales de greedy o greedy por pasos.
- ❖ El 2, 3 y 4 se corresponderán con los cuartiles del 25%, 50% y 75%, respectivamente.
- ❖ El 5 se corresponderá con el 100%. Sólo seleccionará los valores que tengan el valor máximo de out-degree del grafo.

Cuanto mayor sea el cuartil que se elija, menos nodos se podrán seleccionar para realizar el cálculo del grupo óptimo, por lo que el tiempo de ejecución será menor, pero también se podrá cometer un error mayor a la hora de la selección del grupo óptimo.

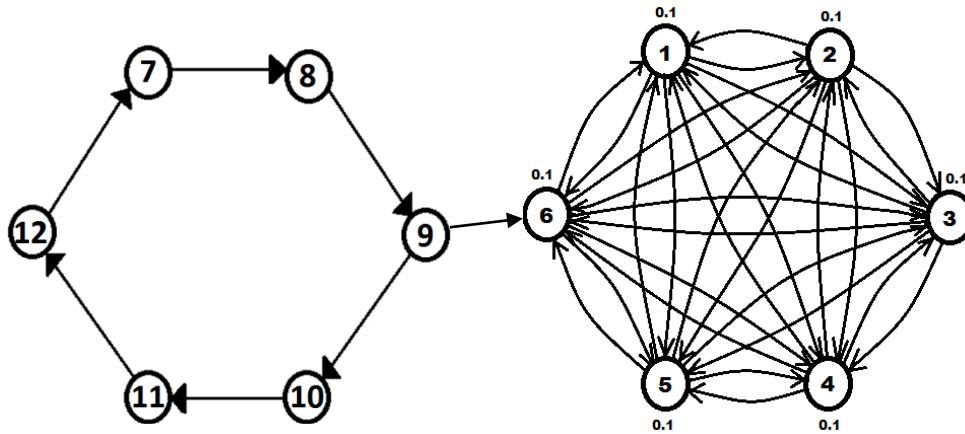


Ilustración 19. Grafo Conexo más Poco Conexo

Este heurístico funciona bien en grafos muy conexos, como ocurría con el heurístico 1. Sin embargo, si se tiene un grafo formado por dos subgrafos interconectados (uno de ellos muy conexo y el otro mucho menos, como se puede ver en la ilustración 18), este heurístico estaría ignorando el subgrafo menos conexo, no seleccionaría ningún nodo del mismo, y no sería capaz de llegar hasta él, lo que se alejaría del óptimo seleccionado por los métodos más exhaustivos.

El desarrollo de este heurístico para el modelo del greedy es muy sencillo de implementar. Antes de seleccionar ningún nodo, se eliminan de la lista de nodos candidatos para el grupo óptimo todos aquellos que tengan un valor de out-degree inferior al valor definido por el cuartil pasado por parámetro. A partir de ahí las operaciones a realizar son las mismas que las del greedy.

Por su parte, para el desarrollo del greedy por pasos hay que tener en cuenta que si el greedy no se ha hecho con el heurístico, puede devolver nodos cuyo valor de out-degree sea menor que el del cuartil especificado por parámetro. En este caso se optará por no eliminar ninguno de los nodos seleccionados por el greedy aunque se encuentren por debajo del cuartil.

3.2.5 Experiencia computacional

Como ya se ha dicho anteriormente, el programa seleccionado para realizar el cálculo del modelo de umbral lineal ha sido R. La elección de este programa se basa en que tiene un amplio número de paquetes para la realización de grafos y su interpretación. Dispone también de paquetes relacionados con las medidas de centralidad individual que permiten hacer comparaciones, sobre todo, entre sus tiempos de ejecución y los de las funciones creadas. Y, además, los números pseudoaleatorios generados por este programa son mejores que los que utilizan otros programas menos especializados en el sector estadístico. [31]

Pruebas realizadas

Una vez definido el problema a optimizar (difusión máxima), el modelo que se va a seguir para resolver este problema (modelo lineal de umbral), la función objetivo y los distintos programas creados para implementar el greedy, el greedy por pasos y los heurísticos, se debe probar cómo es el funcionamiento de cada uno de ellos.

Su finalidad es conocer la eficacia y eficiencia de las funciones desarrolladas en diferentes tipos de grafos, pudiendo decidir para qué tipos es mejor la utilización de una u otra función. Es decir, se podrá ver para qué grafos se aconseja el uso de **heurísticos**, ya que se aproximan al grupo óptimo mejorando cuantitativamente el tiempo de ejecución, para cuáles es mejor realizar los **métodos más exhaustivos** del greedy y greedy por pasos, o incluso si hay grafos en los que el mejor resultado viene dado por las **medidas** estructurales **de centralidad individual** y no por las funciones implementadas del modelo de difusión lineal de umbral.

Para ello se ha realizado una batería de pruebas sobre 13 grafos. Entre ellos se encuentran:

- ❖ *Grafo Friends*: Grafo de relaciones de amistad en un instituto. Está formado por 120 nodos que representan a los diferentes estudiantes dentro de la red, y los arcos son las relaciones de amistad entre ellos, siendo cada arco dirigido y ponderado en función del grado de amistad. Es una buena aproximación del funcionamiento de redes sociales y de cómo aprovecharse de ellas en el marketing viral, ya que es una simplificación del funcionamiento de redes más grandes, como puede ser Facebook, en las que los diferentes usuarios tienen distintos grados de amistad entre unos y otros. Ese grado de amistad puede interpretarse también como grado de influencia (las opiniones de los amigos suelen ser las que se tienen más en cuenta).
- ❖ *Grafo Books*: Grafo real en el que los 105 nodos que tiene se corresponden con libros sobre política, y la relación entre ellos con el nivel de influencia que tiene el haber comprado un determinado libro sobre la posible compra del otro en Amazon. Este grafo es **el más interesante desde el punto de vista del marketing viral**, ya que es un ejemplo de técnicas de mercado y de cómo funcionan las influencias en la adquisición de nuevos productos.

- ❖ *Grafo Dolphins*: Este grafo representa a una comunidad de 62 delfines y las agrupaciones que forman al relacionarse entre ellos. Es muy utilizado para modelar el comportamiento de las redes sociales cuando existen grupos claramente definidos en ella.
- ❖ *Grafo Faculty*: Grafo real de 81 nodos que tiene en cuenta las relaciones de amistad ponderadas y dirigidas de un grupo de amigos de una facultad de Reino Unido. Este grafo, similar al grafo Friends, puede modelar también el comportamiento de redes como Facebook, ya que tiene en cuenta el grado de amistad entre los estudiantes.
- ❖ *Grafos aleatorios 1, 2 y 3*: Estos grafos generados aleatoriamente en R tienen 200 nodos. La diferencia entre cada uno de ellos es que la densidad de conexiones entre los nodos varía, siendo para el grafo 1 igual al 1% , para el 2 igual al 10% y para el 3 igual al 0,8% (el grafo 2 es muy conexo, mientras que el 3 llega a tener hasta nodos aislados). Sirve para comprobar cómo influye la densidad de conexiones entre nodos en las diferencias de los valores óptimos alcanzados por las funciones más exhaustivas y los tres heurísticos.
- ❖ *Grafos aleatorios 4, 5 y 6*: Estos grafos generados aleatoriamente en R tienen 500 nodos. Sus densidades de conexión entre nodos son 0,7% para el 4, 5% para el 5 y 1,5% para el 6. Al igual que el caso anterior sirve para comparar los resultados de las funciones exhaustivas y los tres heurísticos.
- ❖ *Grafos aleatorios 7 y 8*: Estos grafos generados aleatoriamente en R tienen 1000 nodos. Si se mantuviesen las densidades de conexiones en los mismos porcentajes que para grafos más pequeños (como los de 200), los grafos serían muy conexos, no pudiendo probar las diferencias entre unas y otras funciones. Se daría el caso de que muchos de los grupos llegarían a alcanzar a todos los nodos del grafo. Por eso las densidades en este caso son 0.7% para el grafo 7 y 0.8% para el 8. Y aun así en las pruebas realizadas la mayoría de las funciones llegan a todos los nodos para estos dos casos
- ❖ *Grafo aleatorio 9*: Este grafo tienen 5000 nodos. Su densidad de conexiones es igual al 0,1%. Se han realizado pruebas sobre este grafo sólo para tener en cuenta como escala el tiempo de ejecución cuando se prueban las funciones con grafos más grandes.

Una vez definidos los grafos, se tiene que realizar sobre ellos un estudio piloto, para conocer el número de repeticiones que se van a tener que hacer en el cálculo de umbrales aleatorios para que el estimador del óptimo de la función biobjetivo (la media del número de nodos activados y el tiempo de difusión) por medio de Montecarlo pueda asegurarse que se encuentra por debajo de un error máximo del 2,5% con un intervalo de confianza del 95%.

Para realizar estos cálculos se ha tenido en cuenta también el número de nodos que se eligen en los grupos óptimos de cada grafo. En el caso del grafo Friends se seleccionan 10 nodos, para el grafo Books 8, y para el grafo Dolphins y el Faculty 4. En los aleatorios, los de 200 nodos seleccionan 10, los de 500 seleccionan 15, y los de 1000 y 5000 20.

Los estudios piloto realizados tendrán en cuenta la siguiente fórmula:

$$n \geq \left(\frac{z_{\alpha/2} \times s}{E} \right)^2$$

Ecuación 10. Estudios de n° de repeticiones para simulaciones

Donde n se corresponderá con el número de repeticiones mínimo que tendrá que tener el cálculo de umbrales aleatorios para que el estimador tenga un error menor al 2,5% en un intervalo de confianza del 95%. E se corresponderá con el error máximo permitido para el estimador, que en este caso será igual al 2,5% del valor que dé la media muestral. $z_{\alpha/2}$ es el valor crítico correspondiente al intervalo de confianza seleccionado, que al ser igual al 95% , hace que $z_{\alpha/2}$ sea igual a 1,96 . Y s se corresponde con la cuasi-desviación obtenida en el cálculo muestral.

Se han realizado los siguientes estudios para cada uno de los grafos (Anexo 7.2 – Estudio piloto):

- ❖ Para los grafos Friends y Books se han seleccionado 5 grupos aleatorios de longitud 1, 5 de longitud 2, 5 de longitud 3, etc., hasta llegar a 5 grupos de longitud 10. Para cada uno de ellos se ha buscado el valor óptimo y su cuasidesviación mediante la función de threshold. Con estos dos valores se ha obtenido el número mínimo de repeticiones del threshold, y teniendo en cuenta estos datos se ha decidido que para el grafo Friends se realicen 5000 repeticiones y para Books 1500.
- ❖ Por su parte, para los grafos Dolphins y Faculty, se han seleccionado también 5 grupos aleatorios para longitudes de 1 a 5. El resto del procedimiento ha sido el mismo, decidiendo que para ambos grafos se realizarán 1000 repeticiones.
- ❖ En cuanto a los grafos aleatorios, sólo se han seleccionado 5 grupos aleatorios del tamaño que se ha escogido para el grupo óptimo (10 para los de 200, 15 para los de 500 y 20 para los de 1000 y 5000), y se ha realizado el mismo proceso que para los grafos anteriores. La diferencia es que en este caso se elegirá como valor mínimo al peor dentro de cada grupo de grafos aleatorios, es decir, el peor dentro de todas las pruebas de los grafos de 200 para seleccionar el número de repeticiones de los de 200, el peor de los de 500 para las repeticiones de 500, y lo mismo con los dos de 1000. Teniendo esto en cuenta, se ha decidido que para los de 200 se realizarán 1000 repeticiones, para los de 500 se realizarán 100, y para los de 1000 se harán 10. Para el grafo de 5000 tan sólo se realizará 1 repetición.

Una vez definidos los grafos y el número de repeticiones necesarias de los umbrales aleatorios para que los estimadores de la función biobjetivo sean buenos estimadores, se realizan las pruebas necesarias sobre cada uno de los grafos.

De las 7 funciones encargadas de la búsqueda de grupos óptimos, se van a comprobar todas las combinaciones posibles entre las que no contemplan los

heurísticos y las que sí. Todas las funciones van a devolver como resultado el **grupo de nodos seleccionado, el tiempo de ejecución, el número de nodos activados y el tiempo de difusión hasta el estado estable**, de tal forma que las distintas pruebas sean comparables en esos términos.

Estas pruebas son idénticas para cada uno de los grafos, y consisten en las siguientes combinaciones de los tres heurísticos más las dos funciones del Greedy y el Greedy por pasos:

- ❖ *Threshold Greedy*: Prueba del modelo exhaustivo del greedy sobre el grafo.
- ❖ *Threshold Greedy Stepwise v1*: Prueba del modelo exhaustivo del greedy por pasos. Recibe como grupo inicial el obtenido del cálculo del greedy normal.
- ❖ *Threshold Greedy Stepwise Heurístico 1 v1*: Prueba correspondiente a la ejecución de la función greedy por pasos a la que se le aplica el heurístico 1 sobre el grupo inicial. Este grupo inicial se obtendrá del cálculo del greedy normal.
- ❖ *Threshold Greedy Heurístico 1*: Prueba correspondiente a la ejecución de la función greedy a la que se le aplica el heurístico 1.
- ❖ *Threshold Greedy Stepwise v2*: Prueba del modelo exhaustivo del greedy por pasos. Recibe como grupo inicial el obtenido del cálculo del greedy al que se le aplica el heurístico 1.
- ❖ *Threshold Greedy Stepwise Heurístico 1 v2*: Prueba correspondiente a la ejecución de la función greedy por pasos a la que se le aplica el heurístico 1 sobre el grupo inicial. Recibe como grupo inicial el obtenido en el cálculo del greedy al que también se le habrá aplicado el heurístico 1.
- ❖ *Threshold Greedy Stepwise Heurístico 2*: Prueba correspondiente a la ejecución de la función greedy por pasos a la que se le aplica el heurístico 2 sobre el grupo inicial. Este grupo inicial se obtendrá del cálculo del greedy normal.
- ❖ *Threshold Greedy Stepwise Heurístico 3 v1*: Prueba correspondiente a la ejecución de la función greedy por pasos a la que se le aplica el heurístico 3 sobre el grupo inicial. Este grupo inicial se obtendrá del cálculo del greedy normal.
- ❖ *Threshold Greedy Heurístico 3*: Prueba correspondiente a la ejecución de la función greedy a la que se le aplica el heurístico 3.
- ❖ *Threshold Greedy Stepwise Heurístico 3 v2*: Prueba correspondiente a la ejecución de la función greedy por pasos a la que se le aplica el heurístico 3 sobre el grupo inicial. Recibe como grupo inicial el obtenido en el cálculo del greedy al que también se le habrá aplicado el heurístico 3.

Además de estas métricas, se calcula para los grafos reales cuáles son los nodos seleccionados como más centrales teniendo en cuenta las siguientes medidas estructurales de centralidad individual (Anexo 7.3 - Centralidades Individuales):

- ❖ Closeness
- ❖ Betweenness

- ❖ Degree
- ❖ Eigenvector
- ❖ Page-Rank

Y se comprueba los resultados que obtienen estos grupos en el modelo de difusión lineal de umbral, mediante el uso de la función *threshold*, para comprobar cuanto se alejan del óptimo dado por el *greedy por pasos* (tomado como referencia por ser el de mayor carga computacional y el que mejores resultados debe generar) de cada uno de los grafos.

Antes de toda esta batería de pruebas se han realizado una serie de pruebas iniciales con grafos simples donde se ha comprobado que el funcionamiento de las diferentes funciones devuelve los resultados esperados mediante la comparación de los valores obtenidos para los grupos óptimos con un cálculo previo hecho a mano. (Anexo 7.4 – Resultados de Métricas para las 8 pruebas de pequeño tamaño)

Recopilación datos grafo Books

Como la cantidad de pruebas realizadas es relativamente alta, se va a optar por presentar los resultados obtenidos para las pruebas de sólo uno de los grafos. El resto se encuentra en el Anexo 7.5 – Pruebas Completas.

Se ha seleccionado el grafo Books porque es aquel que, como ya se dijo en su descripción, es el que más relación tiene con el entorno del marketing viral, ya que refleja cómo afectan las influencias del resto de los usuarios de una red sobre la decisión de compra de un individuo determinado.

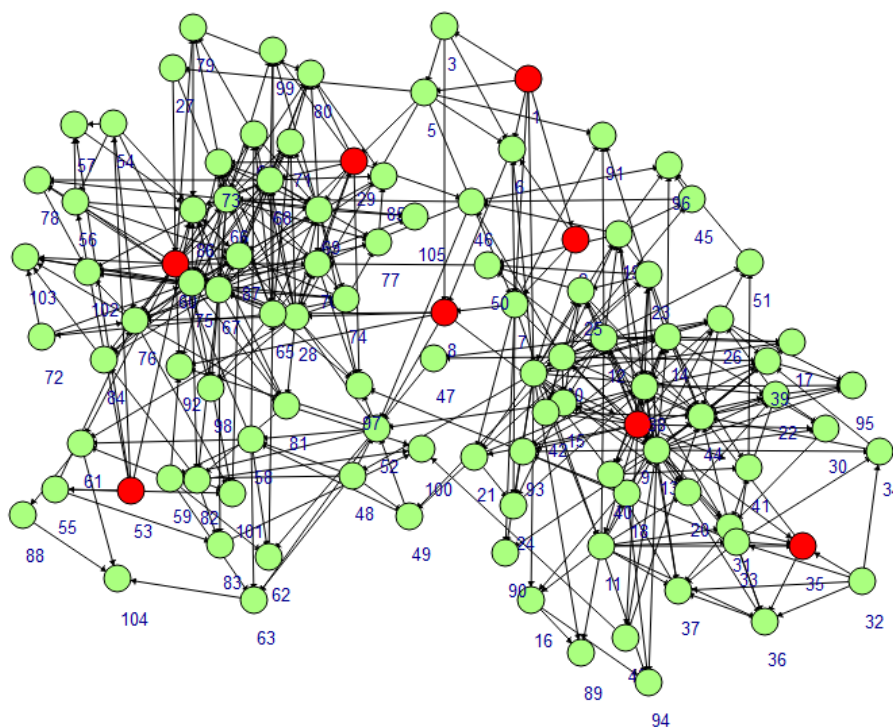


Ilustración 20. Grafo Books con los nodos seleccionados en rojo formando el grupo óptimo

Partimos del estudio para las diferentes funciones del número de nodos que cada una de ellas logra alcanzar. Como se puede comprobar tanto en la tabla como en el gráfico de la misma, las únicas variaciones en el número de nodos alcanzados vienen dadas por las funciones basadas enteramente en el heurístico 1. Esto ocurre porque como se puede ver en la imagen de la red, uno de los nodos seleccionados en el grupo óptimo (el 1) tiene un arco saliente hacia otro de los nodos seleccionados por el grupo óptimo del stepwise exhaustivo (el 2). Como el cálculo del heurístico 1 consiste en eliminar a los vecinos hacia los que tenga un arco saliente cada nodo que entra en el grupo óptimo, al entrar el nodo 1 se eliminará del cálculo al nodo 2. Y como éste ya no puede entrar, el modelo selecciona al nodo 52 en su lugar, con su correspondiente pérdida en el valor del número de nodos activados por el grupo óptimo.

No ocurre lo mismo para el stepwise que utiliza como grupo de entrada el del greedy basado en el heurístico 1, ya que introduce al nodo 2 cuando realiza la revisión de los nodos pasados como grupo inicial. Pero esta revisión hace que tiempo de ejecución del Stepwise-GHeur1 sea mayor que el del Stepwise-Greedy (17,01 minutos frente a 10,31). Esto se debe a que mientras que el greedy exhaustivo ya alcanza el grupo óptimo, el greedy basado en el heurístico 1 no, con la consiguiente pérdida de tiempo para corregir ese error.

TIPO DE PRUEBA	NODOS ALCANZADOS
Greedy	101,4647
Stepwise - Greedy	101,4647
GSHeur1 - Greedy	101,4647
GHeur1	100,126
Stepwise - GHeur1	101,4647
GSHeur1 - GHeur1	100,126
GSHeur2 (4) - Greedy	101,4647
GSHeur3 (50%) - Greedy	101,4647
GHeur3 (50%)	101,4647
GSHeur3 (50%) - GHeur3 (50%)	101,4647

Tabla 9. Pruebas para grafo Books

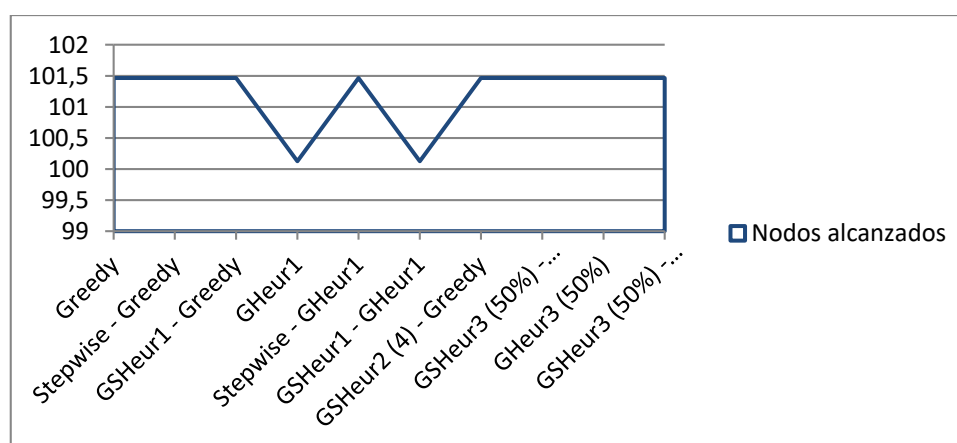


Ilustración 21. Nodos alcanzados en pruebas a Books

En cuanto a los tiempos de difusión calculados, se puede ver en los datos obtenidos que nos encontramos ante la misma situación que en el caso del número de nodos alcanzados. El heurístico 1 da los peores resultados también en el cálculo de esta métrica al eliminar al nodo 2 del grupo óptimo.

TIPO DE PRUEBA	TIEMPO DE DIFUSIÓN
Greedy	4,2627
Stepwise - Greedy	4,2627
GSHeur1 - Greedy	4,2627
GHeur1	4,2793
Stepwise - GHeur1	4,2627
GSHeur1 - GHeur1	4,2793
GSHeur2 (4) - Greedy	4,2627
GSHeur3 (50%) - Greedy	4,2627
GHeur3 (50%)	4,2627
GSHeur3 (50%) - GHeur3 (50%)	4,2627

Tabla 10. Tiempos de difusión para pruebas de Books

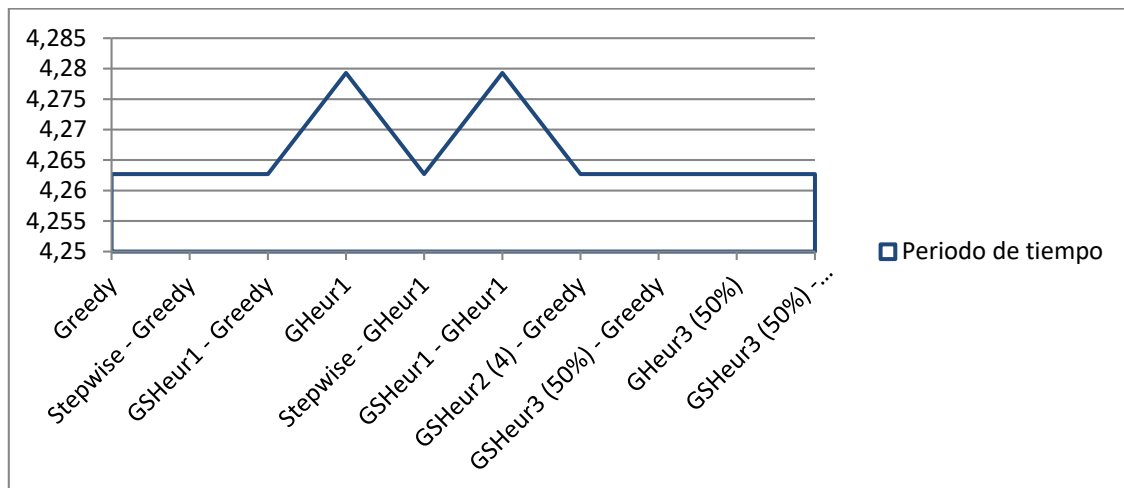


Ilustración 22. Tiempos de Difusión para pruebas de Books

Antes de explicar la última tabla de resultados obtenidos en el cálculo de las funciones del modelo de umbral lineal, se va a estudiar las medidas de centralidad individual sobre el grafo, y cómo de buenos son los grupos obtenidos en términos del modelo de umbral. Se va a hacer en este orden porque la información obtenida de las medidas de centralidad individual va a ayudar a explicar el resultado obtenido para la última de las medidas claves del grafo.

Se ha calculado cuáles van a ser los grupos óptimos seleccionados por cada una de las medidas de centralidad individual. Para poder compararlas con las medidas de difusión, se ha realizado el threshold con cada uno de esos grupos, comparando los resultados del óptimo con los obtenidos con la mejor de las medidas del umbral lineal: el stepwise exhaustivo.

Los datos que se han obtenido son los que se pueden ver en la tabla. Para este grafo, la medida betweenness es la que presenta el grupo óptimo que mejor resultado da al calcular con él la función objetivo del modelo del umbral lineal, mientras que el grupo seleccionado por el page-rank tiene una difusión muy pequeña teniendo en cuenta dicho modelo. Pero como se puede ver en el Anexo 7.3 – Centralidades Individuales esto no es representativo, ya que depende totalmente de la estructura del grafo.

MEDIDA CENTRALIDAD	VARIACIÓN CON RESPECTO AL ÓPTIMO DEL STEPWISE
Closeness	49,14%
Betweenness	90,19%
Degree	85,56%
Eigenvector	38,52%
PageRank	14,90%

Tabla 11. Medidas de Centralidad de Books

El último conjunto de valores a tener en cuenta en el cálculo de las diferentes funciones es el del tiempo de ejecución. Pero no sólo va a interesar lo que tarda cada función en ejecutarse, sino que también se va a querer saber el precio que se está pagando para lograr esa mejora del tiempo de ejecución, entendiendo como precio a la pérdida que supone el utilizar un método menos exhaustivo en el cálculo de la función objetivo. Para calcular esto se va a utilizar la siguiente medida: **$Z^*/\text{tiempoEjecución}$** .

Como se puede apreciar, sobre todo en la gráfica, el greedy basado en el heurístico 3 es el que mejor resultado da en función del tiempo que tarda en ejecutarse. Esto es así debido a que la medida de centralidad del degree (que es en la que se basa este heurístico), obtiene muy buenos resultados ya de por sí (85% del resultado óptimo del stepwise exhaustivo).

TIPO DE PRUEBA	$Z^*/\text{TIEMPO EJECUCIÓN}$
Greedy	1748,691429
Stepwise - Greedy	1032,932182
GSHeur1 - Greedy	1214,31366
GHeur1	2455,362313
Stepwise - GHeur1	626,074709
GSHeur1 - GHeur1	1507,740416
GSHeur2 (4) - Greedy	1242,652369
GSHeur3 (50%) - Greedy	1211,550717
GHeur3 (50%)	2787,835288
GSHeur3 (50%) - GHeur3 (50%)	1635,872627

Tabla 12. Resultados de eficiencia de pruebas a Books

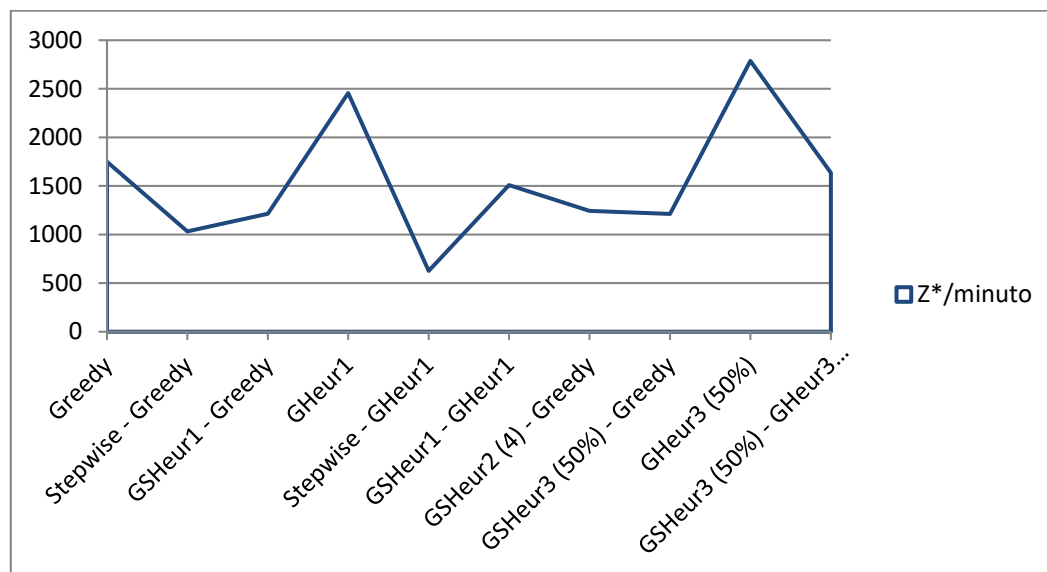


Ilustración 23. Gráfica de Eficiencia de pruebas a Books

Resultados de interés de las pruebas sobre los grafos

Tiempos de Difusión: no sufren grandes variaciones, y dado que se le ha otorgado mayor importancia al nº de nodos alcanzados, se deja para anexos el análisis del tiempo de difusión (así como los valores que toma la función biobjetivo). Dado que la función biobjetivo da prioridad al nº de nodos, ésta será la medida a usar en el análisis de resultados.

Tiempos de Ejecución: El mejor en todos los casos es el heurístico 3 aplicado al *greedy*, mientras que el peor oscila entre el *greedy por pasos* y el primer heurístico aplicado al *greedy por pasos*.

Se ha observado también que el aumento del número de loops a aplicar en la simulación (Montecarlo) eleva el tiempo de ejecución linealmente (el doble de loops implica aproximadamente el doble de tiempo), mientras que el tamaño del grafo hace que el tiempo aumente exponencialmente hasta el punto de hacer inviable una simulación adecuada para grafos del orden de 10.000 nodos en un ordenador de propósito general.

La tabla engloba los resultados relativos al tiempo de ejecución. Los colores en las celdas simbolizan la comparación con el resto (verde|mejor; rojo|peor). Los valores absolutos se resumen en máximo, mínimo y media para cada grafo.

Tiempos de Ejecución (%del Max)	gFriends	gBooks	gDolphins	gUKFaculty	GA Media
Greedy	32,0%	35,8%	59,3%	40,8%	49,7%
Stepwise - Greedy	100,0%	60,6%	100,0%	100,0%	98,3%
GSHeur1 - Greedy	73,6%	51,6%	90,5%	82,1%	68,2%
GHeur1	23,0%	25,2%	48,4%	32,9%	22,3%
Stepwise - GHeur1	88,2%	100,0%	89,6%	91,6%	75,9%
GSHeur1 - GHeur1	62,4%	41,0%	79,6%	74,5%	38,6%
GSHeur2 - Greedy	57,3%	50,4%	85,5%	84,3%	75,2%
GSHeur3 (50%) - Greedy	69,8%	51,7%	89,6%	81,5%	62,7%

GHeur3 (50%)	16,5%	22,5%	42,5%	27,3%	13,7%
GSHeur3 (50%) - GHeur3 (50%)	34,9%	38,3%	73,3%	63,9%	26,0%
MAX	100%	100%	100%	100%	100%
MEDIA	56%	48%	76%	68%	53%
MIN	16,5%	22,5%	42,5%	27,3%	13,7%
MAX (minutos)	147,2	17,0	2,2	5,0	39,6
MEDIA (minutos)	82,1	8,1	1,7	3,4	20,1
MIN (minutos)	24,4	3,8	0,9	1,4	5,0

Tabla 13. Resultados de pruebas - Tiempos de Ejecución

Nodos Alcanzados: En general se han obtenido resultados muy similares entre todos los métodos, ya que como se puede ver las diferencias de nodos alcanzados son mínimas en los heurísticos. Así mismo, se aprecia una enorme dependencia con la cantidad de conexiones del grafo (el grafo Dolphins es mucho menos denso en conexiones, y por tanto sus resultados se ven resentidos, como se puede ver en el % máximo de nodos alcanzados, que es tan sólo el 59 %, mientras que en el resto no se baja del 93% ni en los peores casos). También se aprecia como en los grafos basados en datos reales nunca se alcanza la totalidad del grafo sea cual sea el programa de optimización, mientras que en los grafos aleatorios sí que lo hace; quedando patente, por tanto, la importancia de usar grafos reales para las pruebas.

Nodos Alcanzados (% del MAX)	gFriends	gBooks	gDolphins	gUKFaculty	GA - Media
Greedy	99,9%	100,0%	100,0%	100,0%	99,6%
Stepwise - Greedy	100,0%	100,0%	100,0%	100,0%	100,0%
GSHeur1 - Greedy	100,0%	100,0%	100,0%	100,0%	100,0%
GHeur1	99,9%	98,7%	100,0%	100,0%	99,6%
Stepwise - GHeur1	100,0%	100,0%	100,0%	100,0%	100,0%
GSHeur1 - GHeur1	100,0%	98,7%	100,0%	100,0%	100,0%
GSHeur2 - Greedy	100,0%	100,0%	100,0%	100,0%	100,0%
GSHeur3 (50%) - Greedy	100,0%	100,0%	100,0%	100,0%	100,0%
GHeur3 (50%)	95,7%	100,0%	100,0%	97,9%	99,4%
GSHeur3 (50%) - GHeur3 (50%)	95,7%	100,0%	100,0%	97,9%	99,8%
MEDIA	99,1%	99,7%	100,0%	99,6%	99,8%
MIN	95,7%	98,7%	100,0%	97,9%	99,9%
MAX (% TOTAL)	98%	97%	59%	97%	100%
MEDIA (% TOTAL)	97%	96%	59%	96%	100%
MIN (% TOTAL)	93%	95%	59%	95%	100%
TOTAL (Nº Nodos)	120	105	62	81	(200,500,1K,5K)

Tabla 14. Resultados de pruebas - Nº de Nodos Alcanzados

Elección de Grupos Óptimos: como en todos los casos anteriores, presenta cierta correlación con la densidad de conexiones del grafo. En los grafos más dispersos (con menos conexiones) las elecciones suelen ser los nodos con mejores resultados para las medidas de centralidad (especialmente el *Degree*). En los grafos con mayor concentración los nodos elegidos no suelen tener medidas de centralidad especialmente altas, sino que son nodos relativamente aislados, a los que sería difícil acceder por otros caminos, y que tienen acceso a nodos con mayores capacidades de difusión.

Medidas de Centralidad: La mayoría de las medidas de centralidad resultan deficientes para la selección de grupos óptimos, al menos en comparación con los programas de difusión dinámica implementados en este trabajo. De entre ellas cabe destacar el Degree, que aun siendo la más sencilla, arroja los mejores resultados para esta familia de medidas (por lo que se refuerza su elección como base para el heurístico 3). De todas formas sí que se observan diferencias según la densidad de conexiones, viendo como en el grafo de Faculty (con gran densidad) todas las métricas funcionan bien, mientras que en el grafo de Books hay una gran dispersión de valores.

Nodos Alcanzados (% MAX)	gFriends	gBooks	gDolphins	gFaculty
Closeness	30,3%	47,5%	42,4%	92,0%
Betweenness	50,2%	87,2%	23,4%	92,4%
Degree	89,1%	82,7%	36,5%	92,0%
Eigenvector	56,2%	37,2%	19,2%	92,0%
PageRank	55,1%	14,4%	6,5%	92,0%

Tabla 15. Nodos Alcanzados con medidas de Centralidad

Función Submodular: durante la realización de las pruebas se observó que la mejora obtenida por el *Greedy por pasos* con respecto al *greedy* normal era muy pequeña (en el mejor caso de un 1,3% en nº de nodos alcanzados) para la diferencia de tiempos de ejecución entre las dos (entre dos y tres veces más). Esto se debe a que la función biobjetivo empleada para la optimización es de tipo submodular. (Página 13 de [26], explicación en apartado 3.2.3). Extrapolando, la mejora en nodos alcanzados por cada paso del *greedy por pasos* es minúscula (Tabla 14), y disminuye cuantos más nodos se alcanza.

Nº de conexiones del grafo: En las medidas se ha observado una fuerte correlación entre la calidad de los heurísticos 1 y 3 (nº nodos alcanzados con respecto al *greedy por pasos*, y tiempo de ejecución) y el número de conexiones. La causa se halla en que ambos se relacionan con medidas de centralidad: cuantas más conexiones tenga el grafo, más nodos eliminará el heurístico 1 en cada paso, y será menos probable que los nodos con mayor difusión que va a elegir el heurístico 3 no puedan llegar a casi todo el grafo. Por otro lado, un mayor número de conexiones afecta positivamente a todos los tiempos de ejecución y de difusión (reduciéndolos), debido a la mayor velocidad de convergencia por tener mayor capilaridad.

4 CONCLUSIONES

Todo el documento se ha centrado en la búsqueda en las redes sociales de aquel individuo o conjunto de individuos que van a permitir que se alcance una mayor difusión de una campaña en el ámbito del marketing viral.

Para esto se ha ido teniendo en cuenta cómo se modela una red social como un grafo, y se han estudiado las diferentes medidas que permiten seleccionar grupos de individuos según distintos criterios: o bien lo *central* que el individuo es en la red, utilizando medidas estructurales de centralidad individuales, o bien en función de lo *influyente* que es, teniendo en cuenta un **proceso dinámico de difusión** del mensaje a través de la red.

Entre las dos opciones se ha optado por el desarrollo de un modelo de difusión, y dentro de los distintos tipos que engloba esta categoría se ha seleccionado el **Modelo Lineal de Umbral**, ya que es el que mejor se adapta a las exigencias del marketing viral.

Una vez seleccionado el modelo de difusión, se establece la función biobjetivo donde se busca **maximizar el número esperado de nodos alcanzados y minimizar el tiempo esperado de difusión**, ponderando más el alcance de los nodos que el tiempo.

A partir de la función y el modelo de difusión se han implementado todas las pruebas y se han llevado a cabo sobre diferentes tipos de grafos tanto reales como generados de forma aleatoria. De los resultados de estas pruebas se puede obtener una idea clara de la función que mejor se adapta a la búsqueda óptima de grupos definida al principio. Por sus mejores prestaciones y que el óptimo local que alcanza no dista mucho del obtenido por la función implementada más precisa (el Stepwise), se selecciona la **función Greedy basada en el heurístico 3**. Es decir, en caso de llevar a cabo la búsqueda óptima de un grupo en una red social real, se emplearía esta función.

Aunque los modelos implementados sean sobre redes aleatorias, o redes reales pero de un tamaño reducido para lo que se suele ver en redes sociales (Facebook, Twitter, Youtube, etc.), existen diferentes estrategias que permiten que este modelo y las funciones desarrolladas puedan usarse en redes del mundo real. Algunas de estas estrategias vienen explicadas a continuación.

Clustering

La principal estrategia se basa en la filosofía del “divide y vencerás”: si la red es tan grande que no es posible calcular el grupo óptimo que se está buscando, se segmenta.

A esta estrategia también se la conoce como Clustering. Se llama cluster a un subgrupo de la red, y clustering a la técnica que consiste en realizar esas subdivisiones.

La red sobre la que se haya hecho clustering podrá ser tratada de dos formas distintas:

- ❖ La primera opción consistiría en identificar a cada uno de estos grupos o clusters como si se tratasen de un nodo, y hacer los cálculos sobre esta nueva red. Esta opción se podría hacer cuando se sepa que los grupos en los que se está dividiendo la red son lo suficientemente conexos como para que cualquier individuo del grupo sea capaz de difundir la información a todos los demás, y que además existen conexiones entre unos grupos y otros.
- ❖ En caso de que los grupos estén aislados unos de otros, o la difusión dentro de cada grupo no esté garantizada, se podrá realizar el cálculo de difusión dentro de cada cluster. La asignación de nodos del grupo óptimo se distribuirá entre los distintos clusters.

La división de la red puede hacerse mediante dos criterios: según su estructura (ilustración 18) o en función de criterios sobre los usuarios de la misma. Esta segunda opción es fácil de implementar con las redes sociales a día de hoy, porque la mayoría de ellas tienen almacenada una gran cantidad de información que permite segmentar a los usuarios en función de distintas características: gustos musicales, ubicación geográfica, edad, ideas políticas, canales de Youtube que siguen, etc.

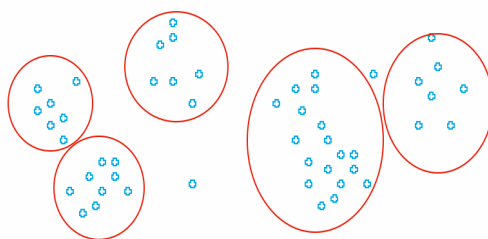


Ilustración 24. Clustering

En función de la campaña de marketing se podría seleccionar la división que más le interese de la red en términos de segmentación de mercado. Se podrá optar por aplicar el cálculo a un único cluster, en el que se encuentre su público objetivo, o intentar llegar a toda la red amoldando la campaña a cada uno de los clusters identificados.

Modificación del modelo

Otra de las posibles estrategias es amoldar el modelo de difusión seleccionado a las necesidades de la campaña. Para esto se puede optar por diferentes cambios:

- ❖ *Campañas centradas en difusión rápida:* Modificar los pesos en la función biobjetivo de tal forma que se pondere más el tiempo de difusión que el

número de nodos alcanzados. Para estos casos habrá que especificar una cobertura mínima.

- ❖ *Campañas a largo plazo*: Modificar los pesos en la función biobjetivo de tal forma que no se tenga en cuenta el tiempo esperado de difusión (peso del tiempo igual a 0).
- ❖ *Campañas con presupuesto fijo*: En este caso habría que modificar la función objetivo y el modelo de difusión para que se tenga en cuenta un presupuesto fijo de la campaña y el coste de convencer inicialmente a cada nodo.

Otra de las posibles estrategias de modificación sería cambiar la función de estimación de la difusión esperada para tener en cuenta los otros modelos de difusión de innovaciones vistos (el modelo de cascada y el modelo del votante). En el caso de campañas que se quiera evitar saturar al usuario desactivado con el acoso constante por parte de sus vecinos se utilizaría el modelo de cascada, y si la campaña se desarrolla para un modelo no progresivo, se podría utilizar el modelo del votante.

Éstos son solo algunos de los ejemplos que se puede desarrollar en campañas sobre redes sociales. Las posibilidades que suponen las redes sociales para el marketing quedan de sobra demostradas, porque aunque las redes sociales nacieron como plataformas para unir personas gratuitamente, sus aplicaciones para generar beneficio son tantas como nodos hay en la red.

*"If you're not paying for something, you're not the customer;
you're the product being sold" (Andrew Lewis)*

5 REFERENCIAS

- [1] «Datos de Instagram,» [En línea]. Available: <https://www.instagram.com/press>.
- [2] «Datos de Facebook,» [En línea]. Available: newsroom.fb.com/company-info/.
- [3] «Datos de Twitter,» [En línea]. Available: <https://about.twitter.com/es/company>.
- [4] «Datos de Youtube,» [En línea]. Available: www.reelseo.com/hours-minute-uploaded-youtube.
- [5] C. d. Pino, «Pensar en la Publicidad,» [En línea].
- [6] M. L. Raso, «La actividad Publicitaria en Internet,» Ra-Ma.
- [7] V. A. Arcos, S. S. M. Gutiérrez y R. P. Hernanz, «La aplicacion empresarial del marketing viral y el efecto boca-oreja electrónico. Opiniones de las empresas,» Universidad de Burgos, 2012.
- [8] «Tipos de Marketing,» [En línea]. Available: economiteca.com/todo-sobre-el-marketing-viral.
- [9] J. M. Maqueira y S. Bruque, «Marketing 2.0: El nuevo marketing en la Web de las Redes Sociales,» 2012.
- [10] «Campaña Dove Real Beauty Sketches,» [En línea]. Available: https://www.youtube.com/watch?v=XC-3g_NHQS4.
- [11] «Campaña Comparte CocaCola,» [En línea]. Available: <https://www.youtube.com/watch?v=TRGM9R-JuHk>.
- [12] «Campaña Old Spice,» [En línea]. Available: <https://www.youtube.com/watch?v=owGyVbfgUE>.
- [13] «Campaña Loewe ORO,» [En línea]. Available: <https://www.youtube.com/watch?v=UUWFWJ9fRoo>.
- [14] «Red Social: Definición,» [En línea]. Available: <http://definicion.de/red-social/>.
- [15] L. C. Freeman, The Development of Social Network Analysis. A study in the sociology of science (Capítulos 1 y 2).
- [16] «Los grandes hitos en la historia de las Redes Sociales,» [En línea]. Available: <http://www.marketingdirecto.com/digital-general/social-media-marketing/los-grandes-hitos-de-la-historia-de-las-redes-sociales/>.
- [17] [En línea]. Available: slideshare.net/wearesocialsg/digital-in-2016/7-wearesocialg_7GLOBAL.
- [18] M. Conrado y J. Tejada, «Redes Sociales: Artículos de Investigación Operativa».
- [19] K. H. Rosen, Matemática Discreta y sus aplicaciones (Capítulo 8).

- [20] S. Wasserman y K. Faust, *Social Network Analysis: Methods & Applications* (Chapter 5).
- [21] V. Umadevi, Estudio del Caso: Medidas de centralidad en el análisis de la red.
- [22] R. Criad, M. Romance y L. E. Solá, Teoría de Perron-Frobenius: importancia, poder y centralidad.
- [23] D. Kempe, J. Kleinberg y E. Tardos, «Maximizing the spread of influence in a social network.,» de *In Proc. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003, pp. 137-146.
- [24] D. Kempe, J. Kleinberg y E. Tardos., «Influential nodes in a diffusion model for social networks.,» de *In Proc. 32nd International Colloquium on Automata, Languages and Programming*, 2005, pp. 1127-1138.
- [25] M. Granovetter, «Threshold models of collective behavior.,» de *American Journal of Sociology*, 1978, pp. 83, 1420-1443.
- [26] J. Kleinberg, Cascading Behaviour in Networks: Algorithmic and Economic Issues.
- [27] «Librería igraph para grafos en R,» [En línea]. Available: <https://cran.r-project.org/web/packages/igraph/igraph.pdf>.
- [28] «Librería igraphdata,» [En línea]. Available: <https://cran.r-project.org/web/packages/igraphdata/igraphdata.pdf>.
- [29] «Librería Keyplayer,» [En línea]. Available: <https://cran.r-project.org/web/packages/keyplayer/keyplayer.pdf>.
- [30] G. Nemhauser, L. Wolsey y M. Fisher., «An analysis of the approximations for maximizing submodular set functions. Mathematical Programming,» 1978, pp. 14, 265-294.
- [31] D. y. S. R. Insúa, Simulación: Métodos y Aplicaciones, 2ª ed., Ra-Ma.
- [32] G. Nemhauser, L. Wolsey y M. Fisher., «An analysis of the approximations for maximizing submodular set functions,» de *Mathematical Programming*, 1978, pp. 14, 265-294.

6 GLOSARIO

Betweenness: Medida estructural de centralidad individual en la que el nodo más central será el que más veces aparezca como puente de un camino mínimo.

Closeness: Medida estructural de centralidad individual en la que el nodo más central será el que se encuentre a una distancia menor del resto de los nodos.

Degree: Medida estructural de centralidad individual en la que el nodo más central será aquel que tenga un mayor número de aristas incidentes.

Eigenvector: Medida estructural de centralidad individual para la que un nodo será central teniendo en cuenta la importancia de sus nodos vecinos.

Grafo: Se define el grafo $G=(V, E)$ a aquel formado por un conjunto V de nodos o actores y un conjunto E de aristas que representan las relaciones entre esos actores.

Heurístico: Función sobre la que se ha realizado una relajación del problema inicial, suponiendo una disminución de su complejidad.

Marketing Viral: Conjunto de técnicas de mercadotecnia basadas en el boca a boca electrónico. Se denomina viral porque su difusión se asemeja a la de los virus.

Modelo General de Umbral: Modelo dinámico de difusión de innovaciones que asigna a cada nodo un valor de umbral, a cada arista un peso, y los nodos activados serán capaces de cambiar de estado a sus vecinos si su nivel de influencia (peso de la arista) es mayor que el umbral.

Page-Rank: Medida estructural de centralidad individual basada en el Eigenvector, pero que se aplica para grafos dirigidos.

Red Social: Conjunto de individuos interconectados entre sí por relaciones de distinta índole (amistad, influencia, gustos comunes, etc.).

Threshold: Función desarrollada para implementar el modelo de umbral lineal. Se le pasa un grafo y un grupo y devuelve el número esperado de nodos alcanzados y el tiempo esperado de difusión.

Greedy: Función desarrollada para implementar la búsqueda de grupos óptimos teniendo en cuenta el modelo de umbral lineal.

Greedy por Pasos: Función desarrollada para optimizar la función del greedy, basándose en el modelo de regresión lineal múltiple.

7 ANEXOS

El conjunto de pruebas detalladas, así como el código empleado para ello se recogen en la siguiente carpeta de drive de libre acceso:

bit.ly/TFGElena-ANEXOS

7.1 Código Fuente

El código fuente de todo el proyecto, que incluye los archivos en R con la definición de los grafos, los estudios piloto y de convergencia, las funciones de selección del grupo óptimo, las pruebas y más, se recoge en un fichero en el siguiente enlace. La función más importante del fichero, la correspondiente al threshold, y las pruebas completas están correctamente comentadas, favoreciendo su lectura:

bit.ly/ElenaCerrato-TFG-CodigoFuente

7.2 Estudio piloto

Se realiza un análisis en base al número de elementos tomados en el grupo óptimo. Con los resultados de la simulación se definen los siguientes valores:

- ❖ Intervalo de confianza del 95%.
- ❖ Media.
- ❖ Cuasidesviación.
- ❖ Error Máximo (2,5% sobre la media).

A partir de ellos se obtiene el tamaño mínimo de la muestra, el número de loops o repeticiones del bucle de simulación (ver diagrama de flujo del Threshold) para cumplir con los parámetros definidos. Pudiendo elegir entre la media de los distintos estudios y el máximo, se va a optar por usar el máximo para mayor rigurosidad.

Estos estudios se realizan para todos los grafos empleados en las pruebas (los 9 aleatorios y los 4 reales), y sus resultados condicionarán los parámetros de las simulaciones finales, recogidas en anexos posteriores.

El estudio completo se puede encontrar en el siguiente enlace (se ha optado por conservar en formato Excel debido a su magnitud):

bit.ly/ElenaCerrato-TFG-EstudiosPiloto

Para asegurarse de que la realidad concuerda con los parámetros estimados en el estudio, se ha realizado un pormenorizado estudio de convergencia con múltiples simulaciones para los valores estimados en el piloto. Los resultados fueron

positivos, concluyendo que los valores estimados pueden usarse para las pruebas finales:

bit.ly/ElenaCerrato-TFG-EstudioConvergencia

7.3 Medidas de Centralidad Individual

Se han aplicado las 5 medidas detalladas en el apartado 2.3 a los 4 grafos reales (Delfines, Libros, Facultad y Amigos). Se recoge el grupo óptimo elegido, los nodos alcanzados al aplicar al grupo óptimo la función threshold, los periodos de tiempo de difusión del threshold, el valor para la función biobjetivo del máximo obtenido y su comparación con el stepwise (como óptimo de referencia).

Las medidas al completo se pueden encontrar en el siguiente enlace:

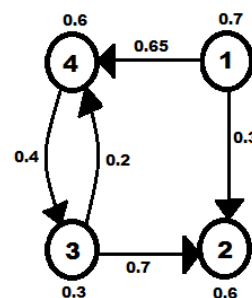
bit.ly/ElenaCerrato-TFG-MedidasCentralidad

7.4 Resultados de Métricas para los ocho grafos de depuración

Para probar las distintas funciones (todas las incluidas en Experiencia Computacional), depurarlas y asegurar su correcto funcionamiento se han empleado 8 grafos con estructuras sencillas y resultados calculables sin simular. Los valores que se incluyen aquí se obtuvieron y posteriormente se tomaron como referencia para comparaciones.

Grafo 1

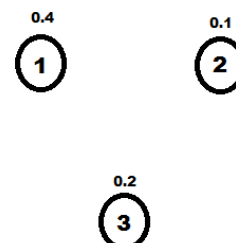
- ❖ Betweenness (2): Jugadores 1 y 4. 0.461832 segundos.
- ❖ Closeness (2): Jugadores 1 y 3. 0.3953369 segundos.
- ❖ Eigenvector (2): Jugadores 2 y 3. 0.3588209 segundos.
- ❖ Degree (2): Jugadores 1 y 3. 0.1090329 segundos.
- ❖ ThresholdGreedy (2): Jugadores 1 y 3. 0.7023389 segundos



Grafo 2

Da problemas al intentar obtener la matriz de adyacencia de un grafo, ya que no tiene aristas. Se debe pasar la matriz directamente a las funciones,

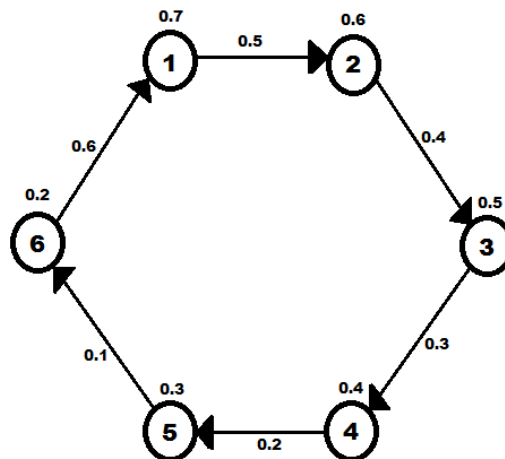
- ❖ Betweenness (2): Jugadores 1 y 2. 0.1979711 segundos
- ❖ Closeness (2): Jugadores 1 y 2. 0.1751001 segundos
- ❖ Eigenvector (2): Jugadores 1 y 2. 0.1170349 segundos
- ❖ Degree (2): Jugadores 1 y 2. 0.07148004 segundos



- ❖ ThresholdGreedy (2): Jugadores 1 y 2. 0.2224629 segundos

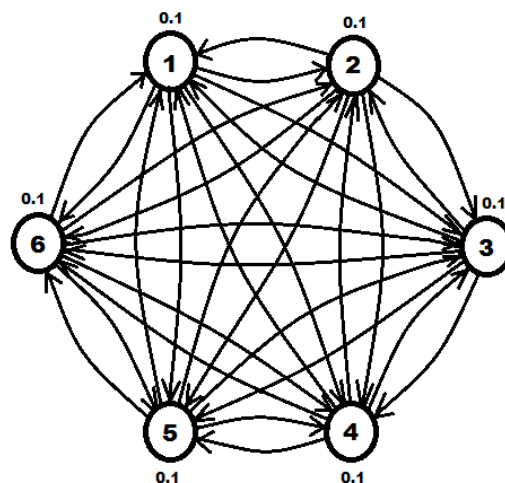
Grafo 3

- ❖ Betweenness (2): Jugadores 2 y 5. 0.9110541 segundos
- ❖ Closeness (2): Jugadores 2 y 4. 0.774019 segundos
- ❖ Eigenvector (2): Da problemas, el número de iteraciones es más alto del permitido y no da un resultado.
- ❖ Degree (2): Jugadores 2 y 6. 0.1889119 segundos
- ❖ ThresholdGreedy (2): Jugadores 1 y 2. 0.2317309 segundos



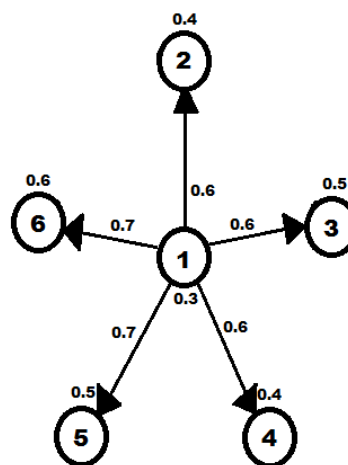
Grafo 4

- ❖ Betweenness (2): Jugadores 1 y 2. 0.9132409 segundos
- ❖ Closeness (2): Jugadores 1 y 2. 0.8019669 segundos
- ❖ Eigenvector (2): Jugadores 1 y 2. 0.686974 segundos
- ❖ Degree (2): Jugadores 1 y 2. 0.1910269 segundos
- ❖ ThresholdGreedy (2): Jugadores 1 y 2. 0.4051049 segundos

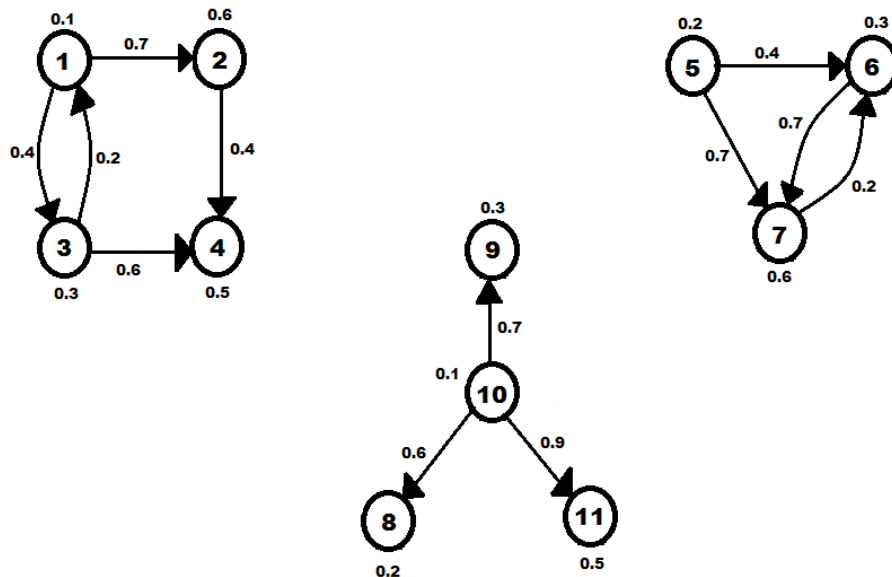


Grafo 5

- ❖ Betweenness (2): Jugadores 1 y 2. 0.909071 segundos
- ❖ Closeness (2): Jugadores 1 y 5. 1.016142 segundos
- ❖ Eigenvector (2): Jugadores 1 y 2. 0.5998549 segundos
- ❖ Degree (2): Jugadores 1 y 2. 0.191221 segundos
- ❖ ThresholdGreedy (2): Jugadores 1 y 2. 0.3829579 segundos



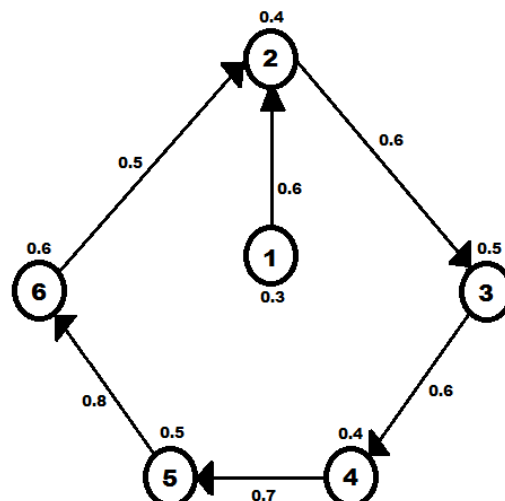
Grafo 6



- ❖ Betweenness (3): Jugadores 4, 6 y 10. 2.644615 segundos
- ❖ Closeness (3): Jugadores 3, 5 y 10. 2.279821 segundos
- ❖ Eigenvector (3): No sirve si el grafo no es conexo. Falla porque alcanza el número de iteraciones máximas.
- ❖ Degree (3): Jugadores 1, 5 y 10. 0.5995202 segundos
- ❖ ThresholdGreedy (3): Jugadores 10, 1 y 5. 0.8255439 segundos

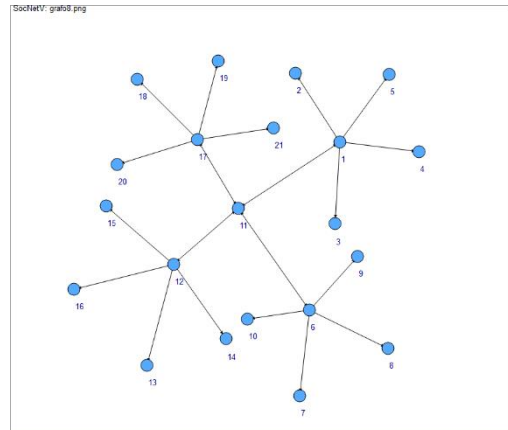
Grafo 7

- ❖ Betweenness (2): Jugadores 2 y 5. 0.907258 segundos
- ❖ Closeness (2): Jugadores 1 y 4. 0.80779 segundos
- ❖ Eigenvector (2): Jugadores 4 y 6. 0.756515 segundos
- ❖ Degree (2): Jugadores 1 y 5. 0.1911931 segundos
- ❖ ThresholdGreedy (2): Jugadores 1 y 4. 1.077722 segundos



Grafo 8

- ❖ Betweenness (4): Jugadores 2, 3, 11 y 18. 8.544698 segundos
- ❖ Closeness (4): Jugadores 1, 6, 12 y 17. 6.900178 segundos
- ❖ Eigenvector (4): Da error en la ejecución de este grafo
- ❖ Degree (4): Jugadores 1, 6, 12 y 17. 1.82563 segundos
- ❖ ThresholdGreedy (4): Jugadores 11, 1, 6 y 12. 1.225384 segundos
- ❖ ThresholdGreedyStepwise (4): Jugadores 1, 6, 17 y 12. 2.831893 segundos

**7.5 Pruebas Completas**

Para todos los grafos definidos (los 9 aleatorios y los 4 reales) se han ejecutado TODAS las funciones definidas en el apartado 3.2.:

- ❖ Greedy
- ❖ Stepwise - Greedy
- ❖ GSHeur1 - Greedy
- ❖ GHeur1
- ❖ Stepwise - GHeur1
- ❖ GSHeur1 - GHeur1
- ❖ GSHeur2 - Greedy
- ❖ GSHeur3 (50%) - Greedy
- ❖ GHeur3 (50%)
- ❖ GSHeur3 (50%) - GHeur3 (50%)

En todas ellas las medidas obtenidas son:

- ❖ Tipo de prueba
- ❖ Nodos seleccionados
- ❖ Tiempo de ejecución (m)
- ❖ Nodos totales del grafo
- ❖ Nodos alcanzados
- ❖ % Nodos alcanzados

- ❖ Periodos de tiempo
- ❖ Z^* - Valor de la función biobjetivo
- ❖ $Z^*/\text{Tiempo ejecución}$ – Métrica de comparación para evaluar el aprovechamiento del tiempo con respecto a los resultados de los distintos programas.

De forma adicional, y de cara a la extracción de conclusiones, se han generado gráficos comparativos de las más importantes (en el apartado 3.2.5 se recogen para el gBooks). El Excel con todos los resultados se encuentra en:

bit.ly/ElenaCerrato-TFG-PruebasCompletas