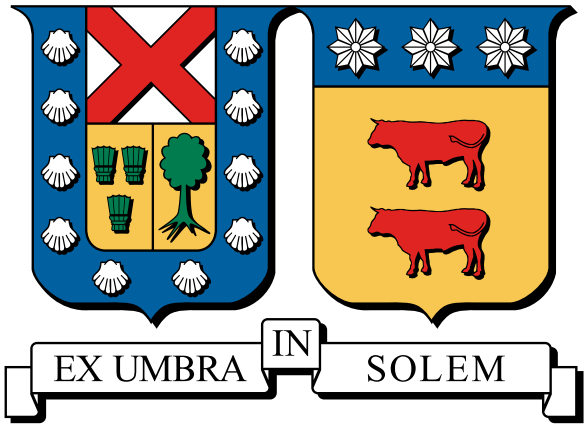
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE INFORMÁTICA

SANTIAGO – CHILE



"PREDICCIÓN DE AFINIDAD E INTERACCIÓN ENTRE USUARIOS EN REDES SOCIALES EN LÍNEA"

MATÍAS ESTRADA IRRIBARRA

TESIS PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: MARCELO MENDOZA
PROFESOR CORREFERENTE: ANDRÉS MOREIRA

Octubre – 2016

TITULO DE LA TESIS:

PREDICCIÓN DE AFINIDAD E INTERACCIÓN ENTRE USUARIOS EN REDES SOCIALES EN LÍNEA

AUTOR:

MATÍAS IGNACIO ESTRADA IRRIBARRA

TRABAJO DE TESIS, presentado en cumplimiento parcial de los requisitos para el Grado de Magíster en Ciencias de la ingeniería Informática de la Universidad Técnica Federico Santa María.

Dr. Marcelo Mendoza
_____
Director de Tesis

Dr. Dr. Andrés Moreira
_____
Co-referente Interno

Dr. Sebastián Ríos
_____
Co-referente Externo

Dr. J. Ricardo Ñanculef
_____
Presidente Comisión Examen

Santiago, Chile. Octubre de 2016.

Resumen

La predicción de enlaces es el problema de inferir la existencia de un enlance entre un par de nodos en un grafo en el futuro cercano. Para llevar a cabo esta tarea es común usar información provista por los nodos y aristas observados, ante lo cual se pueden crear distintos métodos de puntuación de enlaces. Usualmente, estos métodos evalúan estructuras locales del grafo observado, asumiendo que nodos que están más cercanos en el período de observación, tendrán mayor posibilidad de generar un enlace en el futuro.

En esta tesis se diseñan y evalúan algoritmos de recomendación de usuarios que puedan ser de interés a un usuario dado, lo que se denomina recomendación *P2P* en redes sociales en línea. Los algoritmos se basarán en la caracterización de las diferencias entre niveles de actividad de usuarios. El supuesto de esta tesis es que el nivel de interacción real entre dos usuarios se relaciona con los patrones de actividad que ambos usuarios han tenido en el pasado y por tanto, a partir de estos patrones es posible generalizar y pronosticar el éxito o fracaso de una recomendación P2P.

En esta tesis se analizan diferentes características topológicas en redes reales con el objetivo de explicar la creación de nuevos enlaces. En particular se estudia la red Skout, una red social para conocer personas, donde en una representación como grafo cada usuario corresponde a un nodo y la existencia de una relación de amistad entre dos usuarios corresponde a una arista.

Luego de determinar las características que mejor se adaptan al problema de predicción de enlaces con los datos reales, se realizan evaluaciones de modelos generativos como Watts-Strogatz y Barabasi-Albert, evaluando la efectividad del uso de éstas características topológicas: autoridad, grado y transitividad. Además, se analiza el uso de la segmentación de grafos, aplicando las características antes mencionadas en cada sub-grafo generado. Finalmente, los resultados muestran que la red real (Skout) se comporta de manera similar a un modelo generativo (Watts-Strogatz).

Publicaciones relacionadas

- M. Estrada y M. Mendoza, *Affinity Prediction in Online Social Networks* [13], 2014 , *VI Chilean conference of pattern recognition 2014*, Jornadas Chilenas de la Computación Talca, Chile. Noviembre 2014. DOI: 10.1049/14.2014.0012. Institution of Engineering and Technology.
- Predicción de afinidad entre usuarios en redes sociales en línea, M. Estrada y M. Mendoza, Encuentro de Tesistas, Jornadas chilenas de la computación 2014, Talca, Chile.
- *Revisiting link prediction: evolving models and real data findings*, M. Mendoza y M. Estrada, *Network Science Cambridge university press*, Abri 2015 en revisión.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1. Motivación

Existen diversos tipos de redes tales como redes tecnológicas, de información, biológicas y sociales. En el caso particular de esta propuesta, se trabajará con una red social en particular llamada Skout [1] la cual consiste en una red social para conocer personas alrededor del mundo.

En la tabla 1.1 se pueden observar distintos tipos de redes sociales con distintas características para cada red. Las características pueden estar dadas por la naturaleza de las redes, por lo que pueden existir limitaciones físicas o necesidades específicas que hacen que una red presente cierto tipo de propiedades. En el caso de las redes sociales en línea la cantidad de conexiones existentes es mucho menor a la cantidad de conexiones posibles y, por tanto, el grado promedio en la red es bajo, por lo que este tipo de redes son dispersas.

Debido a la disponibilidad de datos y a los diversos fenómenos que ocurren en entornos donde los datos generados obedecen el formato de datos en red, es que se ha desarrollado un creciente interés por el análisis de redes complejas y su uso para diversos propósitos, siendo el ámbito de las redes sociales en línea uno de los más llamativos en la última década. En este trabajo de tesis se explorará el uso de datos extraídos desde redes sociales en línea para proveer recomendaciones P2P. En particular, el interés se centra en explorar

[1] http://www.skout.com

	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r
social	film actors	undirected	449913	25516482	113.43	3.48	2.3	0.20	0.78	0.208
	company directors	undirected	7673	55392	14.44	4.60	–	0.59	0.88	0.276
	math coauthorship	undirected	253339	496489	3.92	7.57	–	0.15	0.34	0.120
	physics coauthorship	undirected	52909	245300	9.27	6.19	–	0.45	0.56	0.363
	biology coauthorship	undirected	1520251	11803064	15.53	4.92	–	0.088	0.60	0.127
	telephone call graph	undirected	47000000	80000000	3.16		2.1			
	email messages	directed	59912	86300	1.44	4.95	1.5/2.0		0.16	
	email address books	directed	16881	57029	3.38	5.22	–	0.17	0.13	0.092
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001	−0.029
	sexual contacts	undirected	2810				3.2			
information	WWW `nd.edu`	directed	269504	1497135	5.55	11.27	2.1/2.4	0.11	0.29	−0.067
	WWW Altavista	directed	203549046	2130000000	10.46	16.18	2.1/2.7			
	citation network	directed	783339	6716198	8.57		3.0/–			
	Roget's Thesaurus	directed	1022	5103	4.99	4.87	–	0.13	0.15	0.157
	word co-occurrence	undirected	460902	17000000	70.13		2.7		0.44	
technological	Internet	undirected	10697	31992	5.98	3.31	2.5	0.035	0.39	−0.189
	power grid	undirected	4941	6594	2.67	18.99	–	0.10	0.080	−0.003
	train routes	undirected	587	19603	66.79	2.16	–		0.69	−0.033
	software packages	directed	1439	1723	1.20	2.42	1.6/1.4	0.070	0.082	−0.016
	software classes	directed	1377	2213	1.61	1.51	–	0.033	0.012	−0.119
	electronic circuits	undirected	24097	53248	4.34	11.05	3.0	0.010	0.030	−0.154
	peer-to-peer network	undirected	880	1296	1.47	4.28	2.1	0.012	0.011	−0.366
biological	metabolic network	undirected	765	3686	9.64	2.56	2.2	0.090	0.67	−0.240
	protein interactions	undirected	2115	2240	2.12	6.80	2.4	0.072	0.071	−0.156
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23	−0.263
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087	−0.326
	neural network	directed	307	2359	7.68	3.97	–	0.18	0.28	−0.226

Cuadro 1.1: Estadísticas básicas para redes publicadas.

recomendaciones de afinidad entre usuarios, esto es, cuán interesado puede estar un usuario en conocer y establecer una relación con otro. Esta propuesta se relaciona también con sistemas de recomendación, brevemente descrito en la sección 1.2.

1.2. Sistemas de recomendación

Los sistemas de recomendación fueron adquiriendo una vital relevancia durante la última década. Es así que a finales de los años 90 emergen sistemas de recomendación en el entonces creciente mercado electrónico, dando origen al área *Collaborative Filltering* (CF) [6]. El objetivo principal de CF es predecir la relevancia de un artículo para un usuario, esto generalmente realizado a través de la proximidad de dos usuarios. Esta proximidad puede ser obtenida mediante tres enfoques: 1) similitud entre el historial de interacción de

los usuarios con los artículos, b) similitud de características de los usuarios, y c) similitud de los artículos de interés de los usuarios. La proximidad en el caso de la similitud entre el historial de interacción de usuarios se logra en base al registro histórico de actividad de usuarios relacionada con artículos, como por ejemplo, comparando la cantidad de artículos en común que compraron dos usuarios, dando paso a la recomendación de productos sobre los cuales los usuarios no comparten preferencias. Distinto es el caso de la similitud basada en usuarios, en la que se utilizan características de éstos, para luego recomendar artículos entre usuarios con propiedades similares, tal como se muestra en la figura 1.1. Finalmente, el caso de recomendación basada en artículos es análogo al basado en usuarios, pero en sentido contrario, es decir, se recomiendan artículos similares a los que son de interés del usuario objetivo.

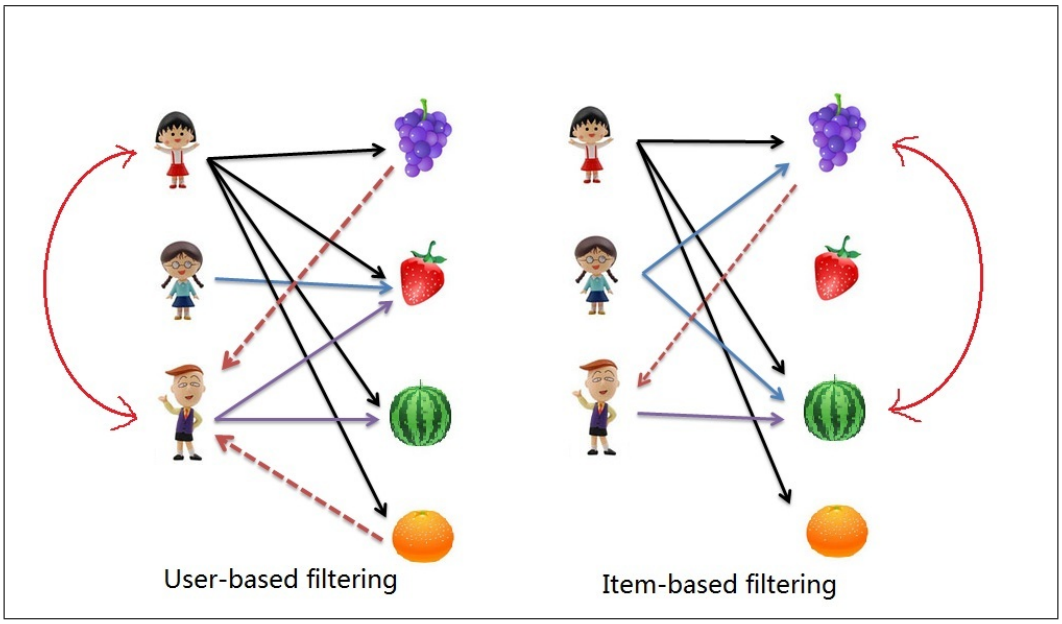


Figura 1.1: Diagrama explicativo de *Collaborative Filtering*

En algunas ocasiones se ha abordado este problema como un problema de predicción de enlaces [7], tomando las ventajas que este enfoque trae para este tipo de problemas, donde agrupar usuarios se realiza de manera más natural y eficiente. En este trabajo de tesis se

seguirá esta aproximación, modelando una recomendación P2P y una red social como un problema de predicción de enlaces.

1.3. Predicción de enlaces

En el ámbito de las redes sociales en línea, la recomendación de usuarios se realiza principalmente a través del grafo completo existente [43] o a través de la transitividad [38]. En el primer caso, una recomendación se establece en base a funciones de similitud usando características topológicas [5]. En el segundo caso se generalizan las relaciones basándose en transitividad en triángulos, como por ejemplo, si A es amigo de B y B es amigo de C, entonces A y C pueden ser amigos [22]. Estas últimas recomendaciones se describen como relaciones basadas en amigos de los amigos (*FOAF*). En general, las redes sociales en línea son dispersas [10], es decir, el total de posibles enlaces que puede generar un usuario es mucho mayor que el número de enlaces que realmente el usuario tiene, tal como se describe en la siguiente expresión de densidad en redes sociales: $E \approx kV$, donde E es el número de enlaces, V es el número de nodos o usuarios y k es un valor mucho mayor al número total de usuarios. Esto provoca que la predicción de enlaces en redes sociales sea una tarea muy difícil, con resultados bastante pobres (dependiendo del tipo de algoritmo o técnica utilizada y de la red en sí misma). Estos resultados varían sustancialmente, encontrándose resultados desde un 3 % hasta un 54.8 % de mejora por sobre predictor aleatorio [25]. En general, los algoritmos de predicción de enlaces en redes sociales se clasifican según el tipo de características que usan para realizar sus recomendaciones. Se pueden encontrar: 1) Algoritmos basados en similitud de perfiles de usuarios, lo que está basado en el principio de homofilia, 2) Algoritmos basados en la descripción de información transaccional de los usuarios, como intercambio de mensajes, participación en foros, y en general *text mining* de sus posts, y 3) algoritmos basados en la caracterización topológica de cada usuario. Estos diferentes enfoques se describirán en las secciones 1.3.1, 1.3.2 y 1.3.3 respectivamente.

1.3.1. Homofilia

La homofilia es la tendencia de individuos a relacionarse con otros individuos de similares características en distintos aspectos [33]. En base a esto se espera que una persona tienda a generar vínculos con otras con las cuales comparten características tales como religión, situación socio-económica, ubicación geográfica, idioma entre otras. Además, se ha estudiado que las relaciones entre individuos no similares tiende a disolverse a una tasa mayor que las relaciones basadas en homofilia. Según el principio de homofilia, la afinidad entre dos usuarios desconocidos puede inferirse a partir de la similitud de sus perfiles.

1.3.2. *Text mining* sobre información transaccional

Las redes sociales proveen mucha información de lo que los usuarios hacen y comparten, tal como mensajes directos, menciones, etiquetado y envío de fotos o archivos entre otras acciones. Esta información puede ser muy útil para determinar la fuerza de conexión entre enlaces, como ha sido propuesto y estudiado por Kahanda y Neville [19]. Según el principio de análisis de texto, la afinidad entre dos usuarios desconocidos puede inferirse a partir de la similitud del contenido recuperable desde sus mensajes o publicaciones.

1.3.3. Caracterización topológica

Las características topológicas de una red se relacionan con la formación de futuros enlaces [38] debido a que las redes se ajustan a determinados modelos de evolución. Esto quiere decir que las características topológicas pueden ser buenos descriptores de la evolución esperada de una red y que ellas están condicionadas al modelo dinámico que explica el crecimiento de una red. Luego, es de interés estudiar si estas características pueden tener

buenas propiedades para pronósticos realizados a nivel de nodos, lo cual no es evidente ya que estas características representan síntesis a nivel estructural.

Algunas características topológicas de interés:

- Diámetro: Longitud del camino más largo de las rutas más cortas entre los pares de nodos.
- Distancia: Cantidad de aristas para la ruta más corta entre dos nodos.
- Grado: Número de vecinos a distancia unitaria. Si la red es dirigida puede distinguirse entre grado de entrada y grado de salida.
- Número de ciclos: Número máximo de ciclos independientes en el grafo.
- *Betweeness centrality*: Número de caminos cortos que pasan por un nodo.

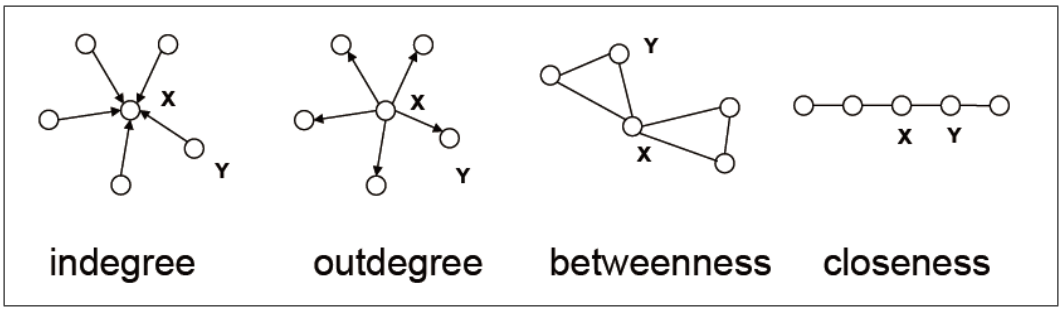


Figura 1.2: Algunas características topológicas de redes.

Las características topológicas de una red pueden dar ciertas pistas acerca del comportamiento ante ciertos eventos en la red. Por ejemplo, una red que contenga nodos con un valor alto en el índice de *Betweeness centrality*, serían susceptibles a ataques intencionados. No se ha explorado aún cuán relevantes pueden ser estas características en el problema de recomendación P2P de usuarios. Dentro de estas características, es de especial interés explorar la utilidad de la asortatividad en este problema, la cual se explica en la siguiente sección.

1.3.4. Coeficiente de afinidad selectiva

El término *assortativity* corresponde a un coeficiente que denota la preferencia de los nodos para vincularse con otros nodos de similares características. En general, las redes sociales tienen mezclas de afinidad selectiva entre los nodos [35].

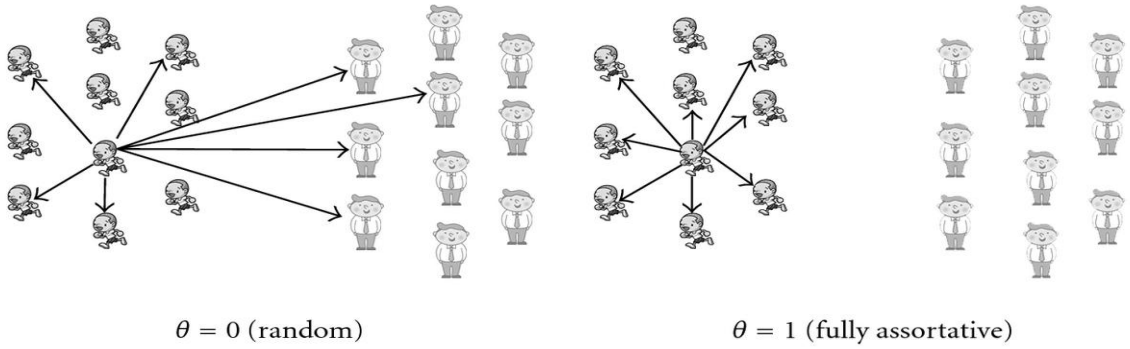


Figura 1.3: Diagrama explicativo de coeficiente de afinidad preferencial

En el caso de la red social a estudiar (Skout) puede ser útil el coeficiente de afinidad preferencial, el cual representa la relación entre los grados de los extremos de las aristas, o sea, puede extraer información acerca de la jerarquía existente de la red, por ejemplo, los usuarios de grado bajo interactúan con usuarios de grado alto o con usuarios que tienen una cantidad similar de amigos. Según esta característica, un usuario A podría ser afín a un usuario dado si es que su grados de afinidad selectiva son coincidentes.

1.3.5. Hipótesis y verificación

Hipótesis 1 *La naturaleza de la red determina el método que mejor se adecua para el problema de Link Prediction.*

Para probar esta hipótesis de trabajo se lleva a cabo una batería de pruebas sintéticas en base a modelos generativos de redes con el fin de mostrar la efectividad en la predicción de enlaces respecto de los parámetros utilizados para las mejores características encontradas en pruebas reales, comparando el comportamiento entre ambas pruebas. Con esto, se puede determinar el modelo generativo que más se acerca a la red social, con lo cual se determinan ciertos comportamientos o patrones esperados y se puede tomar ventaja de esto en el problema de predicción de enlaces.

1.3.6. Discusión

En la propuesta original de tesis se propuso trabajar con características particulares de la red, en específico con el principio de afinidad selectiva o *assortativity*, postulando que el nivel de interacción entre dos usuarios se relaciona con patrones de actividad que han tenido en el pasado. En pruebas preliminares se observó la diversa gama de métodos y enfoques para resolver el problema de *Link Prediction*. Varios de estos métodos recogen el principio del nivel de interacción entre usuarios de diversas maneras, ya sea mediante relaciones de amistad, envío de mensajes y perfiles de usuario. Ante esta situación se evaluaron diversos métodos obteniendo resultados dispares, lo cual motivó el trabajo de tesis final, debido a que además de que el problema de *Link Prediction* está bien estudiado, es un problema difícil y su aplicación entornos reales implica explorar redes muy dispersas.

Capítulo 2

Trabajo relacionado

La recomendación de usuarios es lograda principalmente mediante propiedades de transitividad (Newman *et al* [38]) o *Friend of a Friends* [46]. Si un usuario A es amigo de un usuario B y el usuario B es a la vez amigo del usuario C, entonces A y C probablemente serán amigos [22]. En general, las redes sociales en línea son dispersas [10], lo que significa que la cantidad total de potenciales enlaces a ser creado es mucho mayor que los enlaces observados. La baja densidad de enlaces convierte el problema de predicción de enlaces en un problema muy difícil entregando mejoras sobre un predictor aleatorio que van del 3 % al 54 % [25]. En los últimos años un número de diferentes algoritmos ha sido desarrollado para predecir nuevos enlaces en varios tipos de redes tales como redes de colaboración científica [25], redes de interacción proteína-proteína [30], redes de energía eléctrica [30] y redes sociales [34], entre otras. En la actualidad hay varios índices de similitud disponibles para el problema de predicción de enlaces, entre ellos se encuentra los que se basan en el usuario o en los nodos [27]. Otros índices de similitud que son ampliamente usados son los que se basan en características topológicas locales o globales de una red. Las características locales son construidas a partir del sub-grafo al cual pertenece el usuario y no se requiere del grafo completo. Uno de los índices más simples es *Common Neighbors* [38]. Éste índice considera la cantidad de vecinos en común que tienen dos nodos. Cuando éste índice es mayor, entonces es más probable la creación de un enlace en el futuro entre el par de nodos [22]. El índice de Jaccard [15] es el cuociente entre la cardinalidad de los vecinos en común con la cardinalidad de la unión de ambos vecindarios, obteniendo un valor que representa la proporción entre la cardinalidad de ambos conjuntos. Otros índices usualmente

utilizados son el índice de Salton [44], también conocido como similitud coseno, y el índice de Sorensen [47] que es principalmente utilizado en redes ecológicas [31]. Otro índice local conocido es el índice de Adamic y Adar [1]. Los índices globales requieren de la red entera para ser calculados. Uno de esos índices es conocido como el índice Katz [20]. Éste índice suma la cantidad de caminos existentes entre dos nodos multiplicándolo por un factor de amortiguación. Si este factor es muy pequeño, entonces el índice se comportará de manera similar al índice *Common Neighbors*. El índice de camino local [30] es muy similar al propuesto por Katz, pero usa un largo de camino limitado, todo esto asociado a una complejidad computacional menor. Finalmente se encuentra el índice recursivo conocido como SimRank [18] que es calculado utilizando procesos de *random walk*, propagándose por el grafo con un factor de decaimiento.

2.1. *Link Prediction* Estático

El problema de *Link Prediction* contempla distintas etapas que pueden afectar el rendimiento de una posible solución, como es el caso de la ventana de observación utilizada, uno de los factores que influye en la calidad de la solución [26].
Una de las variantes más utilizadas, ya sea por simpleza o menor esfuerzo computacional requerido, es la observación estática, es decir, utilizar una ventana de observación estática y previamente definida, sobre la cual se ejecutan los distintos experimentos.

2.1.1. Enfoques precursores

En la literatura se pueden encontrar distintos trabajos que se basan en períodos de observación estática, como es el caso del trabajo de Silva *et al* [46], en el que desarrollan un

algortimo que utiliza características topológicas locales, como la densidad de enlaces en los conjuntos resultantes de la unión o intersección de los vecindarios de dos nodos. Hay otros trabajos en los que la localidad del algoritmo depende de un parámetro que acota o limita la exploración, como es el caso de Lu *et al* [30]. A los índices topológicos usados anteriormente se le pueden agregar pesos con información relacionada a la red, como es el caso de Murata *et al* [34], en el cual se agregó información a los índices comúnmente utilizados. Existen otros dos trabajos similares en los cuales se usa la información transaccional de una red: uno es el de Kahanda *et al* [19] que utilizan los mensajes, *posts* e imágenes enviadas y el otro es de Roth *et al* [43] en el cual se basan en el grafo implícito generado a través de las interacciones entre los usuarios en un servicio de correos electrónicos. Bliss *et al* [5] combinaron distintos índices de similitud para luego ajustar los parámetros para darles pesos a cada uno de estos índices utilizando un algoritmo evolutivo, evaluando y evolucionando los parámetros con el pasar del tiempo. Cabe destacar, que en este trabajo de tesis no se varía la ventana de observación, sin embargo, se utilizan los datos en distintas fechas como entradas para ajustar los parámetros necesarios. Finalmente, en el trabajo de Jacobs *et al* [16] se realiza una encuesta para poder etiquetar los enlaces para una posterior evaluación utilizando la periodicidad de las interacciones entre usuarios. Por otro lado, Zhao *et al* [53] utilizan vectores de características basados en la información de los usuarios con una componente opcional basada en propiedades locales.

Referencia	Foco	Conjunto de pruebas
Liben-Nowell *et al* [25]	Variados índices de similitud	Red de colaboración científica
Murata *et al* [34]	Índices de similitud con pesos	Red social
Kahanda *et al* [19]	Información transaccional	Facebook
Lu *et al* [30]	Índice de similitud semi-global (acotado)	Interacción proteína-proteína Colaboración científica Energía eléctrica
Silva *et al* [46]	Densidad de vencindarios	Red Social
Roth *et al* [43]	Grafo implícito de interacciones	Correos electrónicos (Gmail)
Bliss *et al* [5]	Algoritmo evolutivo para combinación lineal de índices de similitud	Twitter
Zhao *et al* [53]	Vectores de características de nodos	interacción proteína-proteína
Jacobs *et al* [16]	Periodicidad entre interacciones	Interacción en un juego en línea

Cuadro 2.1: Trabajos realizados con observación estática

2.2. Enfoques alternativos

Otra gama de trabajos en el área de *Link Prediction* apuntan a diversificar o explorar otros algoritmos y metodologías para llevar a cabo la tarea de predecir enlaces. Es así como Lichtenwalter *et al* [26] generan un algoritmo de flujo y además evalúan distintos tamaños de ventanas de observación.

El trabajo de Chiluka *et al* [7] modifica implementaciones previas para utilizar los algoritmos en las redes basadas en el contenido generado por los usuarios. En el caso de Backstrom [2] utiliza *Supervised Random Walks* combinando la información estructural con la información a nivel de nodos.

Luego, tenemos a Wang *et al* [11] que se basan en tensores y matrices. Un trabajo más reciente (Shin *et al* [45]) plantea el uso de la técnica de aproximación *Low Rank* para mejorar la eficiencia y la escalabilidad del algoritmo.

Una manera diferente de mejorar la eficiencia y poder escalar a medida de que el tamaño de la red aumenta, es aprovechar características globales mediante métodos que se aproximen a métodos globales. Es así como Symeonidis *et al* [48] proponen el uso de *Spectral Clustering* para el problema de *Link Prediction*. Este enfoque considera los k primeros valores propios de la matriz Laplaciana para generar k clusters. Posteriormente se calcula la distancia Bray-Curtis entre nodos en el espacio k-dimensional formado por los valores propios seleccionados.
Finalmente, tenemos el caso de la aplicación de lógica difusa con Bastani *et al* [4] y la teoría de juegos con Zappella *et al* [52].

Referencia	Foco	Conjunto de pruebas
Lichtenwalter *et al* [26]	Algoritmo de flujo	Llamadas telefónicas, colaboración científica
Backstrom *et al* [2]	*Supervised Random Walks*	Facebook, red de colaboración científica
Chiluka *et al* [7]	Índices de similitud adaptados	Flickr
Shin *et al* [45]	Aproximación mediante *Low Rank*	Red club de karate
Wang *et al* [50]	Modelos probabilísticos locales	DBLP, colaboración científica
Dunlavy *et al* [11]	Modelos basado en tensores y matrices	DBLP
Bastani *et al* [4]	Lógica difusa	Red de colaboración científica
Zappella *et al* [52]	Teoría de juegos	Red social comercial
Symeonidis *et al* [48]	*Spectral Clustering*	Red social opiniones

Cuadro 2.2: Trabajos con enfoques alternativos

2.3. Segmentación de redes

En la comunidad de *Link Prediction* existen varios trabajos en los que se realiza un proceso de segmentación de la redes combinado con alguna otra técnica o característica para poder realizar la tarea de predecir enlaces, como en el caso Symeonidis *et al* [48]. Por esto, es importante revisar algunas de las técnicas más conocidas para la detección de comunidades o para segmentar redes.

2.3.1. *Clustering* espectral

Ésta técnica de segmentación considera los *k* primeros valores propios de la matriz Laplaciana para generar *k* clusters mediante los *k* vectores propios que representan un espacio k dimensional. Dentro de este espacio k-dimensional, se pueden clasificar los miembros de cada uno de los k cluster usando alguna medida de distancia.

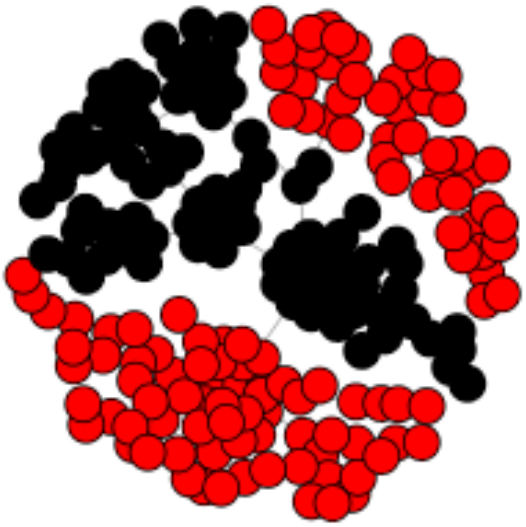


Figura 2.1: Análisis espectral para Red Barabási-Albert de 200 nodos

2.3.2. Método Walktrap para detección de comunidades

El método planteado por Pons *et al* [40] se elabora en base a la intuición de que una caminata aleatoria tiende a quedar atrapada en partes densamente conexas, correspondientes a comunidades, por lo que se puede generar un dendrograma representando la jerarquía del grafo en base a la distancia entre vértices y uniones de vértices.

2.3.3. Método *Edge Betweenness* para detección de comunidades

El método *Edge Betweenness* para detección de comunidades planteado por Newman *et al* [37] calcula el valor de *edge betweenness* de cada nodo y elimina el nodo más alto, para luego iterar sobre la misma operación. Esto genera un dendrograma representando la jerar-

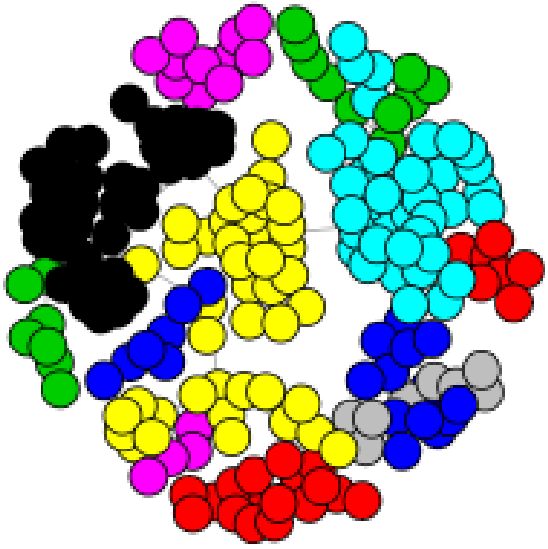


Figura 2.2: Método Walktrap

quía de los nodos y segmentando el grafo a medida que se eliminan nodos en cada iteración.

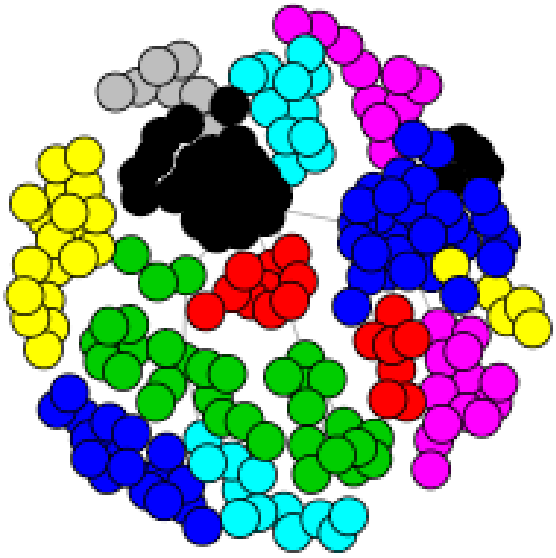


Figura 2.3: Método *Edge Betweenness*

2.3.4. *Fast Greedy Modularity* para detección de comunidades

Fast Greedy Modularity es el método propuesto por Clauset *et al* [9], el cuál es un algoritmo aglomerativo jerárquico que se basa en un índice de modularidad Q. Para determinar las comunidades se calcula el valor de cambio del índice de modularidad ΔQ en cada iteración y luego se itera hasta que la comunidad persista.

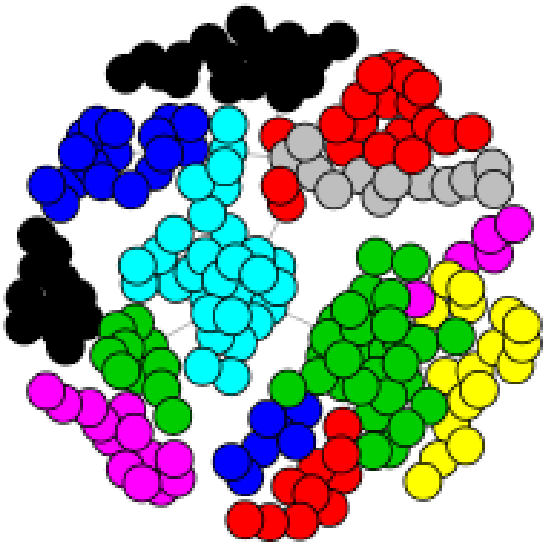


Figura 2.4: Método *Fast Greedy Modularity*

2.3.5. *Leading eigenvector*

El método *Leading eigenvector* postulado por Newman [36] usa el primer valor propio positivo para particionar la matriz de modularidad $Q = A - P$, donde A es la matriz de adyacencia y P es la matriz de de probabilidad de existencia de enlaces. Una vez particionada la red en dos comunidades, se puede repetir este proceso para particionar en más de dos comunidades la red.

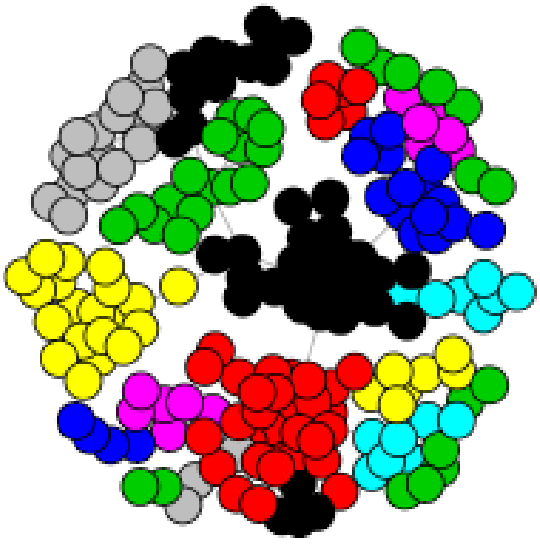


Figura 2.5: Método *Leading Eigenvector*

2.4. Modelos generativos

2.4.1. Watts-Strogatz

El modelo Watts y Strogatz en teoría de redes se emplea para la construcción de algunas redes de mundo pequeño. Genéricamente se trata de un modelo de generación de grafos aleatorios con distancias medias pequeñas y valores altos del coeficiente de agrupamiento. El modelo matemático toma el nombre de la investigación realizada por los matemáticos Duncan Watts y Steven Strogatz en el año 1998 en la revista Nature [51]. El modelo se genera a partir de un enrejado tipo anillo, donde cada nodo está conectado directamente a k vecinos. Posteriormente, se procede al proceso de recableado, en donde con probabilidad p se determina si un enlace es recableado o no. Para recablear un enlace entre los nodos n_i y n_j se escoge con probabilidad p un nodo n_k de todos los posibles, evitando bucles y enlaces repetidos, formando un enlace entre n_i y n_k, eliminando el enlace original con el

nodo n_j. El parámetro p asociado a la probabilidad de recableo determina cómo será la red resultante. Cuando p es 0, es una red regular tipo malla o grilla; cuando p es 1 el resultado es una red similar a una red aleatoria tipo Erdós-Renyi [12]. En la imagen 2.6 se observa el efecto del parámetro p sobre el resultado generado.

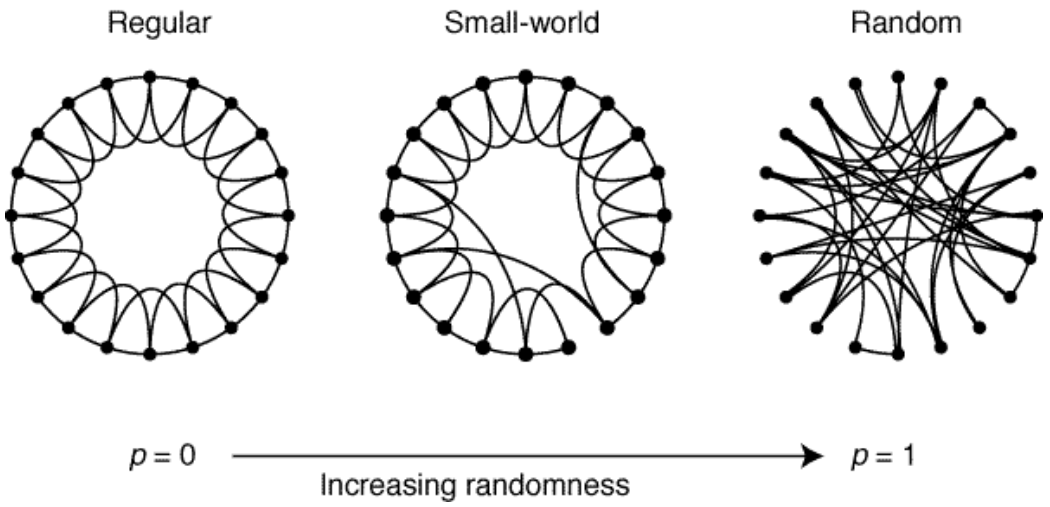


Figura 2.6: Modelo generativo de Watts-Strogatz [51].

2.4.2. Barabási-Albert

El Modelo de Barabási–Albert [3] es un algoritmo empleado para generar redes aleatorias complejas libres de escala empleando una regla o mecanismo denominado conexión preferencial. Las redes generadas por este algoritmo poseen una distribución de grado de tipo Ley de Potencia y que se denominan redes libres de escalas. La red se genera a partir de la incorporación de nuevos nodos, dado un nodo v, la probabilidad de generar el arco hacia v_i, indicada como p_i es proporcional al número de enlaces ya existentes de v_i. La ecuación 2.1 muestra la probabilidad p_i asociada al grado k_i del nodo objetivo respecto de la suma total de grados existentes en la red al momento de agregar un nuevo nodo. Las redes de

este tipo son muy frecuentes en los sistemas elaborados por el ser humano así como en la naturaleza. Ejemplos de sistemas de este tipo son Internet, redes de citas, redes eléctricas y algunas redes sociales. El modelo toma el nombre de Albert-László Barabási y Réka Albert autores que lo popularizaron en 1999.

$$p_i = \frac{k_i}{\sum_j k_j} \tag{2.1}$$

2.4.3. Erdös-Rényi

El modelo Erdös–Rényi, nombrado así por ser un estudio que realizaron los matemáticos Paul Erdős y Alfréd Rényi [12], es uno de los métodos empleados en la generación de grafos aleatorios. En este modelo se tiene que un nuevo nodo se enlaza con igual probabilidad p con el resto de la red, es decir posee una independencia estadística con el resto de los nodos de la red.

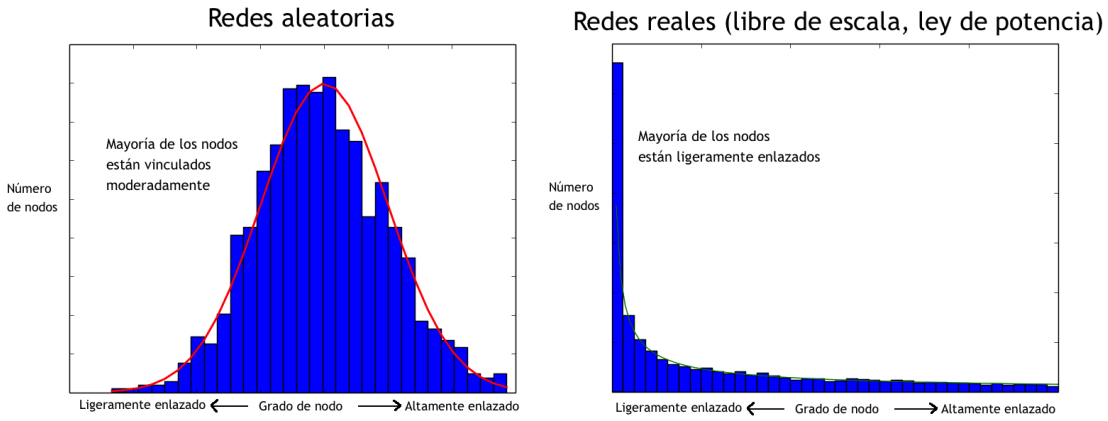


Figura 2.7: Comparación de distribución de grados entre redes aleatorias y libres de escala

En la imagen 2.7 se puede observar la diferencia entre la distribución de grados entre

redes aleatorias y libres de escala. En el caso de las redes aleatorias (tipo Erdós-Renyi), las distribución de grados sigue una distribución normal, es decir, la mayoría de los nodos presentan grados cercanos al promedio, mientras que en el caso de las redes libres de escala (tipo Barabási-Albert) se observa una ley de Zipf [41], donde muchos nodos están conectados a muy pocos, y muy pocos nodos están conectados a muchos nodos. En el caso particular de las redes generativas Watts-Strogatz, la distribución de grados dependerá del parámetro p utilizado, es decir, cuando p es cero, todos los nodos tienen el mismo grado; cuando p es 1 se comporta como una red aleatoria y cuando es cercano a 0.5 muestra una distribución normal desplazada.

Capítulo 3

Análisis descriptivo

En este capítulo se exploran dos aproximaciones al problema de predicción de enlaces. La primera es modelar el problema como un problema de clasificación de enlaces entre dos clases (verdadero y falso). El segundo enfoque es usar umbrales de localidad para definir una colección de semillas desde la cual se puede generar una lista de candidatos.

Este capítulo fue publicado parcialmente en Estrada *et. al* [13]

3.1. Estrategia para predicción de afinidad

Sea $G = (V, E)$ un grafo no dirigido. En un punto dado del tiempo t_0 asumimos que todos los vértices $v \in V$ son conocidos pero solo un subconjunto de enlaces E es conocido. Sea E^{obs} el subconjunto de enlaces $e \in E$ conocidos en t_0, y sea E^{miss} el subconjunto de enlaces no observados en t_0, donde $E^{miss} = E \setminus E^{obs}$. Dado el problema de predecir enlaces en E^{miss} dado $G = (V, E^{obs})$, una manera natural de abordarlo es creando un modelo a partir de $G = (V, E^{obs})$ capaz de predecir enlaces potenciales en E^{miss} tal como se señala a continuación:

$$\mathcal{P}\left(E^{miss} \mid E^{obs}, \mathcal{X}_v\right),$$

donde $\mathcal{X}_v$ es una colección de características de los vértices extraídas desde G. Es usual estimar un modelo $\hat{\mathcal{P}}$ usando instancias de datos etiquetadas sobre $\mathcal{X}_v$, modelando la tarea de predicción de enlaces como un problema de clasificación de enlaces entre dos clases,

correspondientes a, enlaces reales y falsos. Para realizar esto, un conjunto de enlaces E_0 es observado en un periodo posterior a t_0, asumiendo que estos enlaces son buenos descriptores de los enlaces en E^{miss}. Entonces, estos enlaces son caracterizados sobre $\mathcal{X}_v$ y etiquetados como enlaces reales de ejemplo. Desde el conjunto $E \times E \setminus \{E_0 \cup E^{obs}\}$ se extrae una muestra aleatoria de enlaces, caracterizada sobre $\mathcal{X}_v$ y etiquetada como enlaces falsos de ejemplo. En consecuencia, un modelo es ajustado para estimar $\hat{\mathcal{P}}$ de acuerdo a una función de criterio dada. La figura 3.1 muestra la aproximación descrita anteriormente.

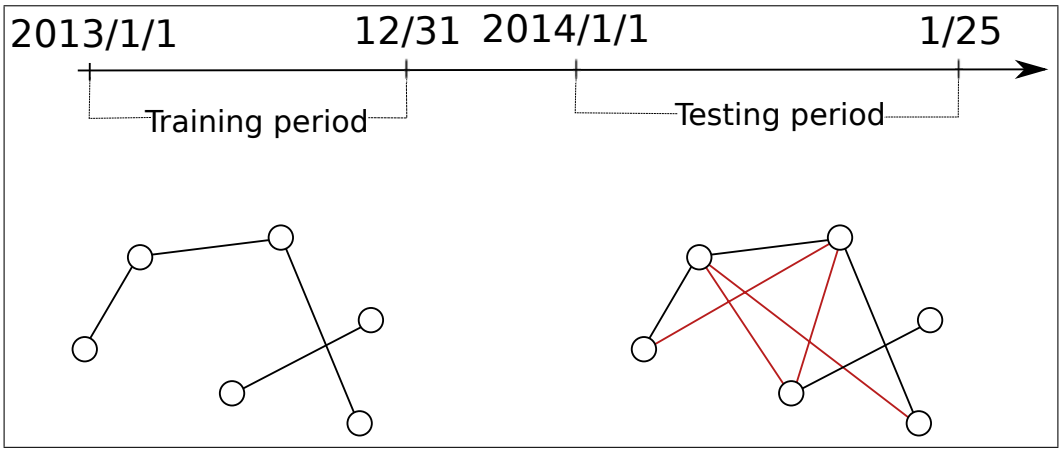


Figura 3.1: *Link prediction* como un problema de clasificación.

Es común considerar un período largo de observación respecto a la red y calcular $\mathcal{X}_v$. En la comunidad de máquinas de aprendizaje este período es conocido como período de entrenamiento. En la figura 3.1 el período de entrenamiento abarca un año completo y t_0 es el último día de ese año (31 de Diciembre). Después de t_0 y durante 25 días, nuevos enlaces (marcados por líneas rojas) son registrados y almacenados para crear una colección de enlaces reales. Este período de tiempo es llamado período de pruebas. Cada enlace no observado durante el período de entrenamiento y de pruebas es considerado como un enlace falso. Luego, enlaces falsos (cada enlace no incluido en la figura) y enlaces reales (arcos rojos) son usados para crear un conjunto de datos con ejemplos de ocurrencias verdaderas y falsas caracterizado sobre $\mathcal{X}_v$. Un número de factores puede ayudar a explicar el éxito de tal estrategia. Primero, el resultado depende de la calidad del *solver* de máquinas de

aprendizaje. Es habitual tratar con modelos sobredimensionados limitando la capacidad de $\hat{\mathcal{P}}$ de generalizar a nuevas instancias. La creación de un modelo con buenas propiedades de generalización está lejos de ser una tarea fácil. Además, el éxito depende de la elección de t_0 y de la existencia de un período de entrenamiento y de prueba considerable. En esta sección se compara la estrategia descrita anteriormente respecto a una estrategia basada en *ranking*. Una estrategia basada en *ranking* considera dos pasos para ordenar enlaces candidatos. Para un vértice dado u, se realiza un paso de recuperación de vértices, usando restricciones de localidad para recuperar una lista de vértices cercanos a u. Luego, esta lista es ordenada de acuerdo a criterios de relevancia dados. El siguiente pseudocódigo ilustra la estrategia basada en *ranking*.

Predicción de enlaces basada en *ranking*

```
1:  procedure LINK RANKING(u, th)
2:  retrieval:
3:      seeds ← ()
4:      Γ(u) ← fetch neighbor list of u
5:      for each v in Γ(u) do
6:          if locality(u, v) > th then
7:              seeds ← v
8:  ranking:
9:      scores ← ()
10:     for each v in seeds do
11:         for each c in Γ(v) do
12:             scores[c] ← X_v(c)
13:     return top K elements of scores
```

Como se muestra en el pseudocódigo, se requieren dos pasos para abordar el problema de predicción de enlaces como un problema basado en *ranking*. El primer paso desarrolla un paso de recuperación sobre el conjunto de vecinos de u. Para hacer esto, una restricción asociada a un umbral de localidad es verificada, colocando los vértices en la lista de semillas. Luego, en el paso de ordenamiento o *ranking* del algoritmo, por cada vecino c de cada semilla recuperada en el paso anterior, un puntaje es calculado sobre $\mathcal{X}_v(c)$. Finalmente, los mejores K elementos de la lista de candidatos son retornados.

3.2. Características para predicción de afinidad

3.3. Características locales

En esta sección se explorará el uso de una medida de localidad: Jaccard.

Coeficiente de Jaccard Sea $G = (V, E^{obs})$ el grafo observado en el tiempo t_0. Sea u y v un par de vértices en V, y sea $\Gamma(u)$ y $\Gamma(v)$ el vecindario de u y v, respectivamente. El coeficiente de similitud de Jaccard de u y v es la proporción entre los vecinos en común y la unión de los dos vecindarios. Formalmente:

$$\texttt{Jaccard}(\texttt{u},\texttt{v}) = \frac{|\,\Gamma(\texttt{u}) \cap \Gamma(\texttt{v})\,|}{|\,\Gamma(\texttt{u}) \cup \Gamma(\texttt{v})\,|}.$$

El dominio de valores del coeficiente de *Jaccard* se mueve entre 0 y 1, siendo 0 dos nodos que no tienen nungún vecino en común y 1 cuando dos nodos comparten todos sus vecinos.

3.4. Características globales

En esta sección se explora el uso de medidas globales para la predicción de enlaces: Grado normalizado, medidas basadas en *HITS*, y transitividad. Estas características son independientes de las consultas, es decir, estas medidas definen una colección de estimaciones a nivel de vértices. El valor de una medida de esta colección es el mismo para todo el grafo, razón por la que estas características serán llamadas características globales. Las definiciones están dadas a continuación.

Coeficiente de grado Sea $\Gamma(u)$ el vecindario del vértice u y sea $\mid \Gamma(u) \mid$ la cardinalidad de este conjunto, también conocido como grado de u. Se define coeficiente de grado normalizando $\mid \Gamma(u) \mid$ con el máximo grado de G. Formalmente:

$$\texttt{Degree(u)} = \frac{\mid \Gamma(\texttt{u}) \mid}{\texttt{Max}_{\texttt{v}\in\texttt{G}} \mid \Gamma(\texttt{v}) \mid}.$$

Coeficiente de transitividad Sea $\Gamma(u)$ el vecindario del vértice u. El coeficiente de transitividad $\texttt{Transitivity(u)}$ (también conocido como coeficiente de *clustering*) es la razón entre el número de enlaces en $\Gamma(u)$ y el número máximo de esos enlaces. Si $\Gamma(u)$ tiene e_u enlaces, entonces se tiene:

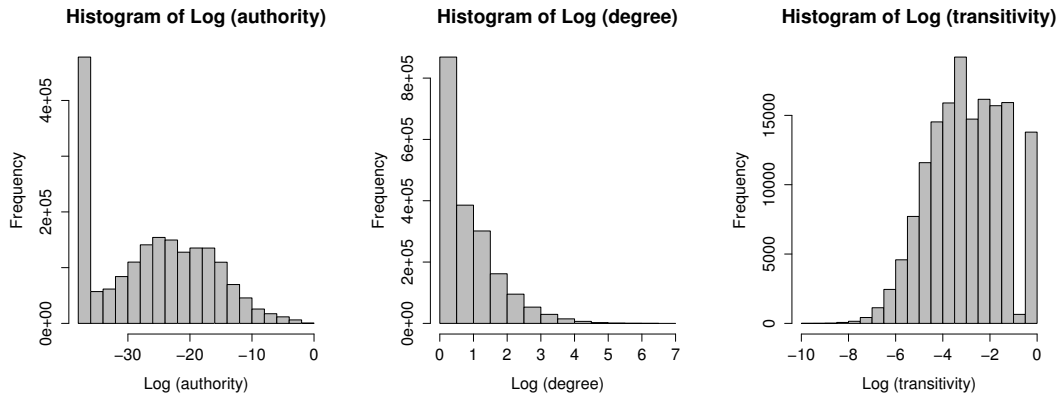
$$\texttt{Transitivity(u)} = \frac{\texttt{e}_\texttt{u}}{\frac{|\Gamma(\texttt{u})|\cdot(|\Gamma(\texttt{u})|-1)}{2}}.$$

Coeficiente HITS Los coeficientes HITS (Búsqueda inducida de tópicos en hipertexto) provienen de la comunidad de recuperación de la información, propuesto por Kleinberg [21] y usado originalmente para *ranking* de páginas web. La idea es que las páginas que tienen muchos enlaces apuntando a ellas son llamadas autoridades y las páginas que tienen muchos enlaces de salida son llamadas *hubs*. Buenos *hubs* apuntan a buenas autoridades, y viceversa. Sean $\texttt{hub(u)}$ y $\texttt{auth(u)}$ los coeficiente de *hub* y autoridad para un vértice u. La siguiente ecuación puede ser resuelta a través de un algoritmo iterativo que aborda el problema definido por:

$$\texttt{hub(u)} = \sum_{\texttt{v}\in\texttt{G}|\texttt{u}\to\texttt{v}} \texttt{auth(v)},$$

$$\texttt{auth}(u) = \sum_{v \in G | v \to u} \texttt{hub}(v).$$

En grafos no dirigidos ambos coeficientes tienen el mismo valor. Se obtuvo el grafo de enlaces de Skout que corresponde a un año completo (2013). Luego, para una colección de 542.010 usuarios que crearon enlaces al menos una vez durante el período de observación, se calcularon estas tres medidas descritas anteriormente. Estos histogramas aparecen en las figuras 3.2a, 3.2b y 3.2c.



(a) Autoridad (b) Grado (c) Transitividad

Figura 3.2: Histogramas de características globales

Las figuras 3.2a, 3.2b y 3.2c muestran histogramas log(a+1) de las medidas de autoridad, grado y transitividad. Autoridad muestra una porción significativa de usuarios en la primera sección. Sin embargo, un grupo de usuarios está concentrado en el rango entre -30 y -10, indicando que esta medida es relevante para esta tarea. Algo similar ocurre en el histograma de log transitividad, donde los coeficientes están concentrados alrededor de -4 y un gran grupo en la sección cero, indicando la presencia de cliqués. Finalmente, se puede observar que la característica de grado normalizado está concentrada en valores bajos.

3.5. Evaluación experimental

3.5.1. Resultados

En esta sección se explora la tarea de predicción de enlaces con datos reales obtenidos desde Skout. Se registró cada enlace creado durante el año 2013 generando un grafo no dirigido con 3.855.389 enlaces entre 1.920.015 usuarios.

Algunas estadísticas básicas de este grafo están detalladas en la tabla 3.1.

Feature	Value
E	3,855,389
V	1,920,015
Diameter	7

Cuadro 3.1: Estadísticas básicas para el grafo analizado.

Se registraron los enlaces creados durante los primeros 25 días de Enero de 2014. Esta colección de enlaces registrados comprende 582.199 nuevos enlaces durante este período entre usuarios cuyas cuentas fueron creadas antes del 1 de Enero (usuarios antiguos desde ahora en adelante). Un total de 428.341 usuarios antiguos adicionaron un nuevo amigo durante el período de observación. La distribución de nuevos amigos por usuario es mostrada en la figura 3.3.

Como muestra la figura 3.1, la distribución de nuevos amigos por usuario sigue la ley de los ricos obtienen más riquezas: Solo unos pocos usuarios concentran muchos nuevos amigos mientras que la mayoría de los usuarios solo forma uno o dos relaciones de amistad durante el período. El número de enlaces creados por día durante la observación es mostrada en la figura 3.4. Se puede observar que solo dos días exhiben alzas significativas en términos de creación de enlaces (el día tercero y el duodécimo día) y los otros están

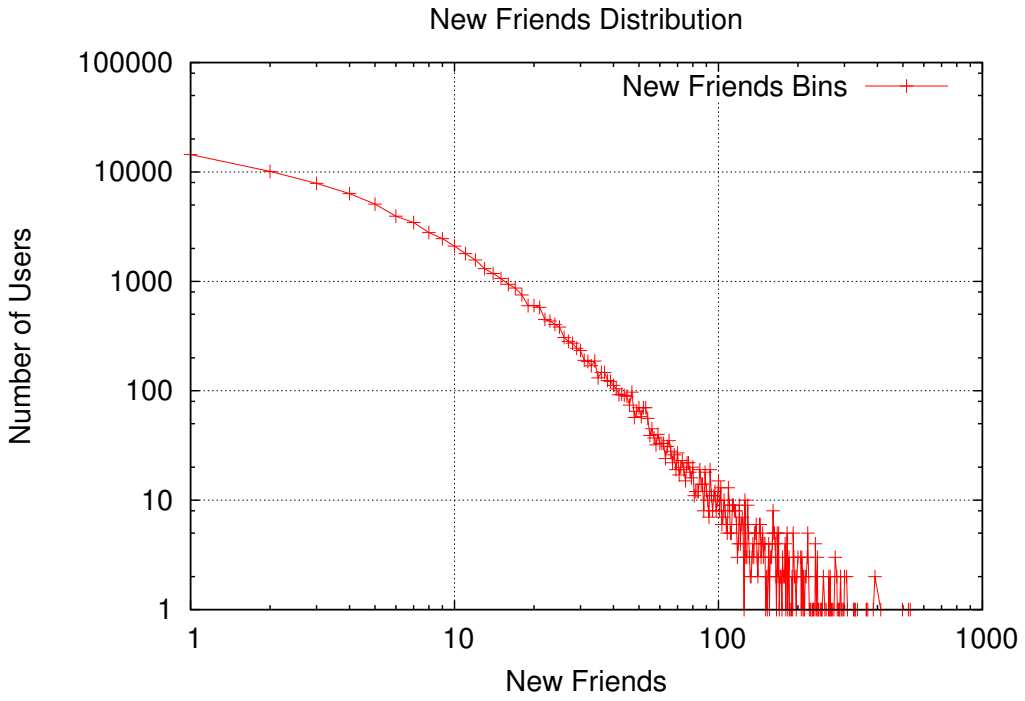


Figura 3.3: Distribución del número de enlaces creados por usuarios durante el período de observación.

concentrados alrededor de 30.000 enlaces por día.

Un total de 76.848 usuarios aceptaron al menos una solicitud de amistad durante el período de observación. A esos usuarios se les denomina como “usuarios activos” del período. La colección de enlaces potenciales usadas para crear el conjunto de datos de máquinas de aprendizaje es el conjunto de enlaces no observado que se inicia a partir de usuarios activos y que apuntan a usuarios con al menos un nuevo enlace durante el período de observación, descartando de este conjunto los enlaces potenciales que pudieron ser creados entre usuarios inactivos. Luego se reduce el problema de clasificación al de separación entre enlaces no observados y reales que se inician en usuarios activos. Esta metodología es descrita en la figura 3.5.

Este escenario favorece a las máquinas de aprendizaje. A medida que se utiliza informa-

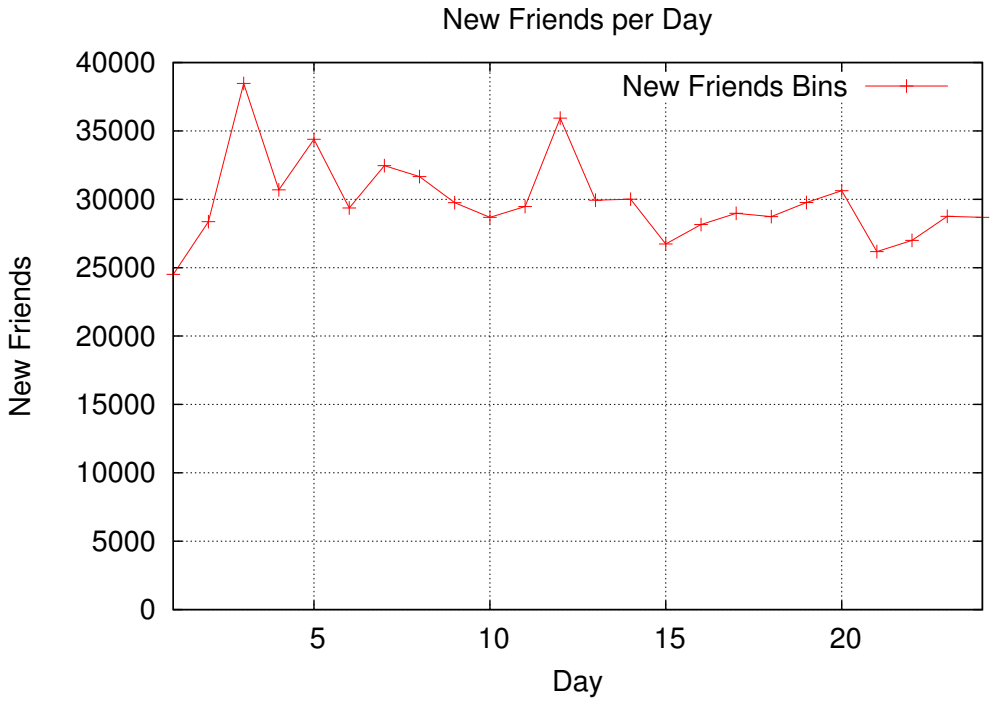


Figura 3.4: Número de enlaces creados por día durante el período de observación

ción sobre cuáles usuarios crearon enlaces durante el período de observación, el conjunto de datos puede tomar ventaja de esta información, reduciendo el problema de predicción de enlaces a un problema de clasificación entre enlaces reales y falsos entre usuarios activos. De hecho, el problema real es mucho más difícil, debido a que no se tiene apriori la información sobre los usuarios activos. Por lo tanto, el clasificador tiene que lidiar con el conjunto completo de enlaces potenciales, que crece de manera cuadrática a medida que crece V. A pesar de estas consideraciones, este enfoque es útil para ilustrar cómo las medidas globales se comportan en este problema, descubriendo algunas propiedades de los datos. En este trabajo preliminar, se balancean las instancias de datos reales y falsas para evitar resultados sesgados o parciales. Esto se realiza mediante el muestreo aleatorio uniforme de todo el conjunto de instancias etiquetadas. Como resultado se obtuvo un conjunto de 343.887 enlaces etiquetados como reales y 345.000 enlaces etiquetados como falsos.

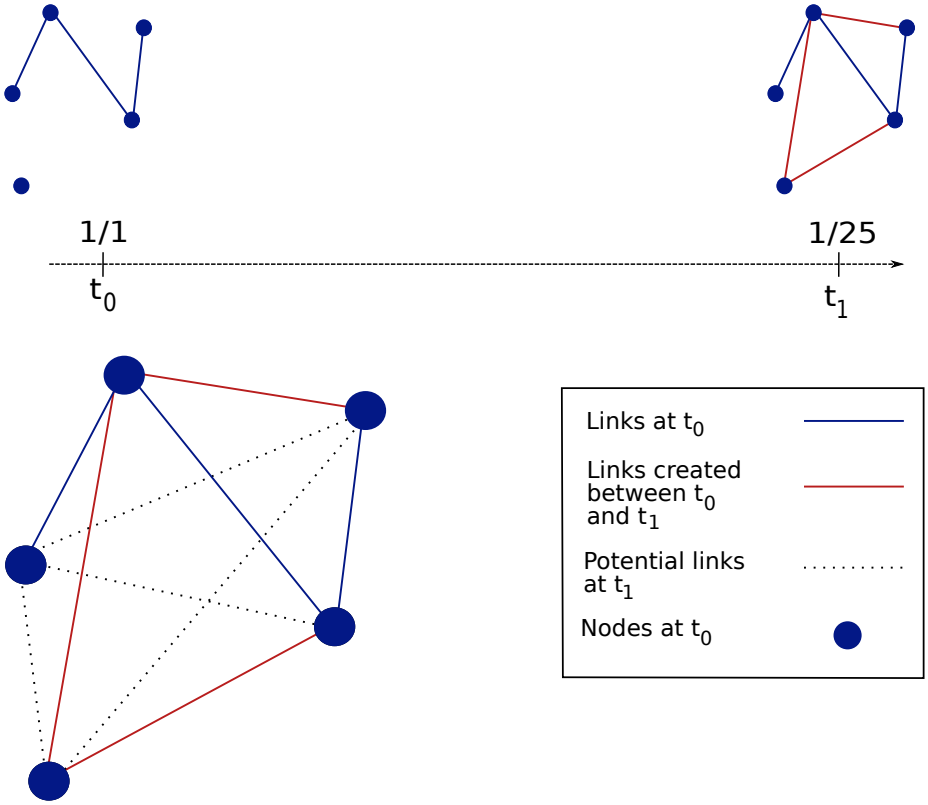


Figura 3.5: Metodología de creación del conjunto de datos. Enlaces y nodos azules representan el grafo en t_0. Enlaces rojos representan enlaces creados durante el período de observación. Enlaces mostrados con línea punteada indican enlaces potenciales no observados. En la estrategia a utilizar, enlaces de línea punteada son etiquetados como instancias de enlaces falsos y los enlaces rojos son etiquetados como instancias reales.

Para cada instancia de datos considerada en este conjunto y para cada vértice de cada enlace etiquetado, se calcularon las medidas de autoridad, grado normalizado y transitividad. En la tabla 3.2 se muestra los valores de ganancia de información para cada característica considerada en el conjunto de datos.

Como muestra la tabla 3.2, la característica más relevante para este problema es el puntaje de autoridad del segundo vértice. Como el primer vértice corresponde al usuario activo (el que acepta o rechaza la solicitud de amistad), la autoridad del segundo usuario es una medida de la visibilidad del usuario candidato para el resto del grafo (qué tan conectado

Feature	Information Gain
Authority 2	0.9155
Authority 1	0.5367
Transitivity 2	0.1192
Degree 2	0.1054
Transitivity 1	0.0392
Degree 1	0.0227

Cuadro 3.2: Valores de ganancia de información para cada característica considerada en el conjunto de datos. Características de usuarios activos están representadas con un subíndice 1.

está el candidato al grafo). Por otro lado, las características de los usuarios activos que están relacionadas con su localidad son marginalmente relevantes al problema, indicando que la actual conectividad del vecindario del usuario activo no se relaciona con la creación de enlaces nuevos. El gráfico de matriz del conjunto de datos está representado en la figura 3.6. Se aplica una función logarítmica a cada característica para facilitar la visualización de los gráficos.

Tal como muestra la figura 3.6, cada comparación entre características muestran la ausencia de correlación. En consecuencia, se descarta el uso de una selección de un conjunto de características. En la tabla 3.3 se muestra los resultados en rendimiento de los clasificadores creados para resolver el problema de clasificación. Se utilizó una estrategia de validación cruzada para evaluar cada clasificador. Los *solvers* usados fueron *Naive Bayes*, J48 (árboles de decisión) y regresión logística.

Como se muestra en la Tabla 3.3, el mejor resultado es logrado usando un árbol de decisión J48, con un balance de rendimiento casi perfecto entre ambas clases. Los otros *solvers* exhiben resultados sesgados, a pesar del hecho de que el conjunto de datos etiquetado es balanceado. De hecho, *Naive Bayes* y la regresión logística sesgan sus resultados a la detección de enlaces falsos, con un alto valor para la tasa de Verdaderos Positivos para esta clase. Respecto a los resultados para enlaces verdaderos, son muy pobres, con tasas

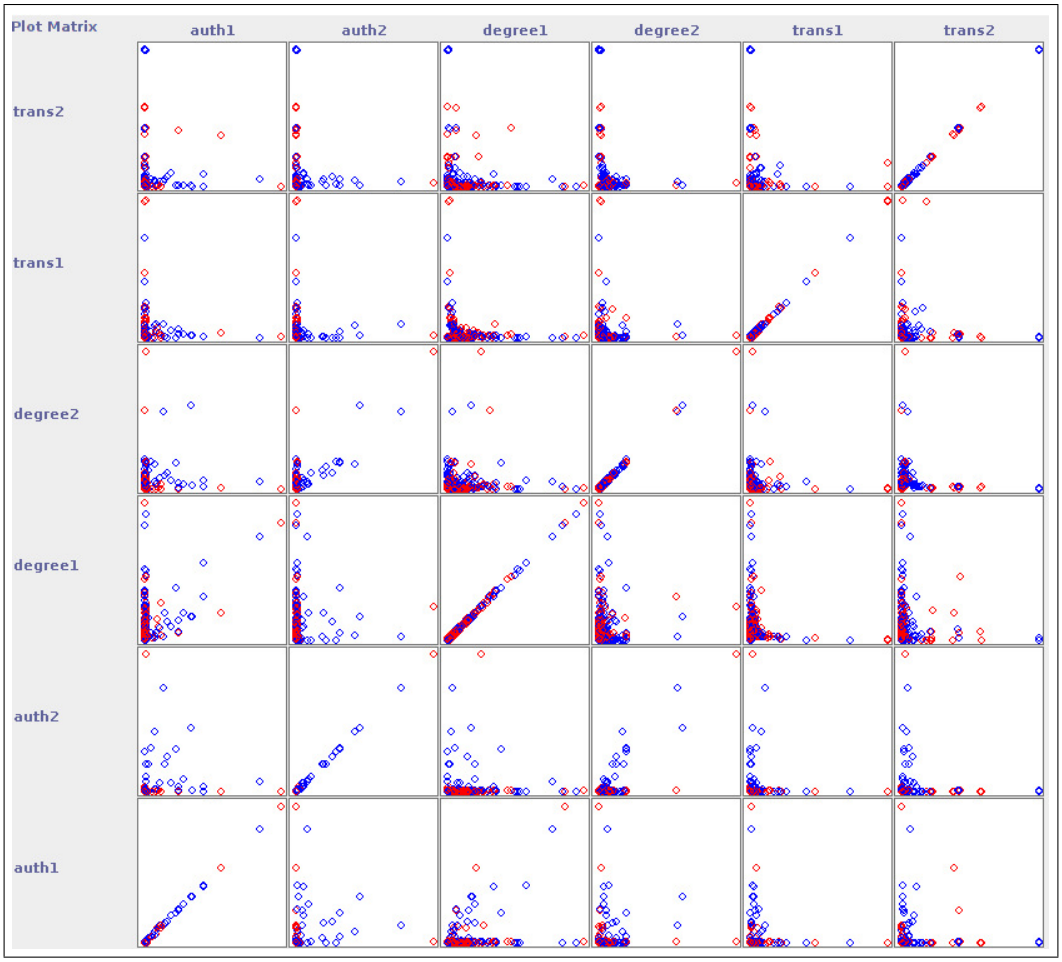


Figura 3.6: Matriz de correlación de las características consideradas en el conjunto de datos. Enlaces reales son representados por los puntos rojos y enlaces falsos por los puntos azules.

de Verdaderos Positivos iguales a 0.138 y 0.316 para Bayes y regresión logística, respectivamente. Estos resultados implican bajas tasas de *recall* y, en consecuencia, bajas tasas de *F-measure*. Además, en valores de *TP* y *FP* se percibe notoriamente el sobreajuste producido por el desequilibrio entre clases determinado por la naturaleza de la red, es decir, la proporción de $\frac{FP+TN}{FN+TP}$ es alta, produciendo el mencionado desequilibrio con *Naive Bayes* y con regresión logística. Debido a lo anterior, se analizará los resultados obtenidos usando J48, descartando Bayes y Regresión logística para la siguiente etapa de análisis.

En la figura 3.7 se muestra el árbol de decisión J48 creado usando el conjunto de datos.

	Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Bayes	real	0.138	**0.035**	0.796	0.138	0.235	0.183	0.605	0.635
	false	0.965	0.862	0.529	0.965	0.683	0.183	0.605	0.572
	W Avg.	0.552	0.449	0.662	0.552	0.460	0.183	0.605	0.604
J48	real	0.642	0.232	0.734	0.642	0.685	0.414	0.789	**0.813**
	false	0.768	0.358	0.683	0.768	0.723	0.414	0.789	0.778
	W Avg.	**0.705**	0.295	**0.708**	**0.705**	**0.704**	0.414	**0.789**	0.795
Log	real	0.316	**0.132**	0.705	0.316	0.437	0.221	0.662	0.666
	false	0.868	0.684	0.560	0.868	0.681	0.221	0.662	0.625
	W Avg.	0.593	0.408	0.633	0.593	0.559	0.221	0.662	0.646

Cuadro 3.3: Medidas de rendimiento del problema de clasificación entre enlaces verdaderos y falsos.

Se introdujo una restricción de podado para un mejor entendimiento de su estructura. La restricción de poda fue introducida para hacerse cargo del balance entre la descripción de datos y el rendimiento, limitando el efecto de la restricción en 5 puntos porcentuales sobre *F-measure*.

La figura 3.7 muestra que la característica más relevante para esta tarea es la autoridad del segundo vértice, resultado que es consistente con el análisis de ganancia de información. De hecho, un valor alto de autoridad para el segundo vértice (el vértice candidato) es suficiente evidencia para la predicción de enlaces reales para 35.296 enlaces sobre 2.993 falsos positivos. Candidatos con bajos puntajes de autoridad necesitan ser descritos usando grado y transitividad. Finalmente, la caracterización de los usuarios activos es marginal para el problema, siendo considerada para esta tarea el grado. De hecho, un bajo grado para usuarios activos permite detectar enlaces reales para 80.346 casos sobre 33.628 falsos positivos.

Ahora se explorará el uso de umbrales de localidad para la selección de candidatos. Se calculó el coeficiente de Jaccard para cada vértice en el grafo almacenado hasta el primero de Enero de 2014, es decir, para la colección de 3.855.389 enlaces creados durante el 2013. El histograma para estos coeficientes se muestra en la figura 3.8.

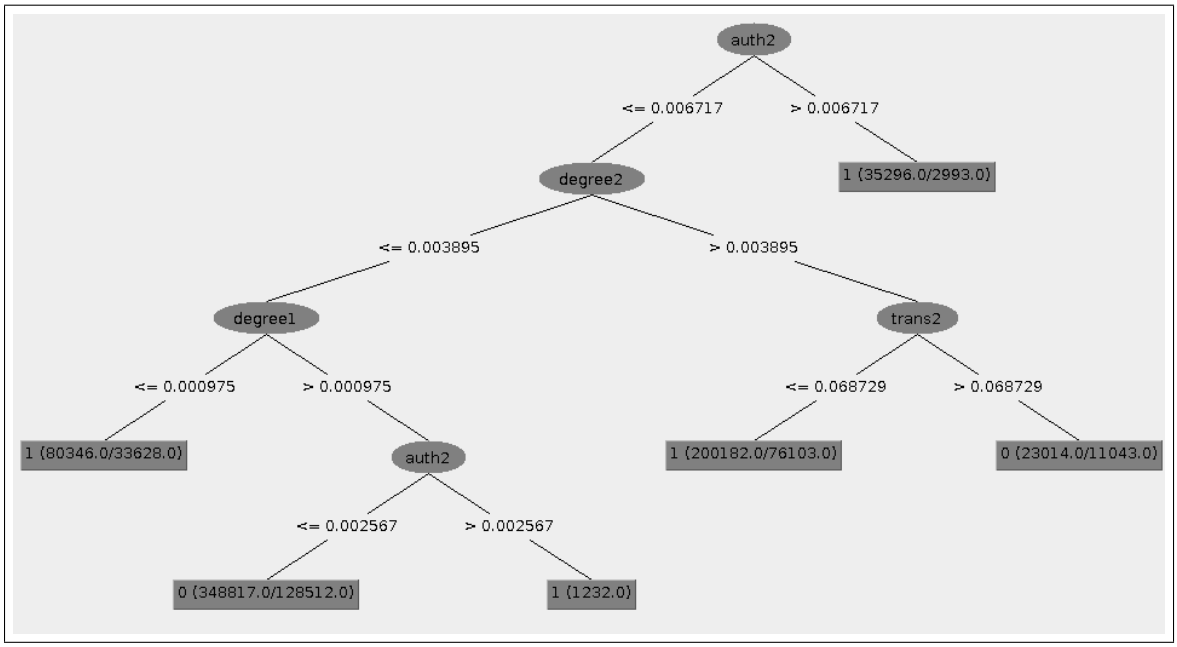


Figura 3.7: Árbol de decisión para el problema de clasificación verdadero/falso.

Tal como muestra el histograma de la figura 3.8, el principal casillero de agrupación está en el origen, pero un número significativo de vecinos logran altos valores para esta medida. Se utilizarán 3 umbrales de localidad usando Jaccard: 0.1, 0.2 y 0.3. Estos valores permitirán recuperar un gran número de semillas para el algoritmo de *ranking* de vértices.

Se iniciará el análisis desde la colección de usuarios que crearon al menos un nuevo enlace de amistad durante el período de observación. Este conjunto contiene 76.848 vértices, y por cada vértice de este conjunto se obtiene la colección de vecinos con un valor del índice de Jaccard por sobre el valor del umbral.

Se obtuvieron 70.498, 61.056, y 34.929 semillas para los umbrales de Jaccard iguales a 0.1, 0.2 y 0.3, respectivamente. Luego, para cada semilla fue recuperado su conjunto de vecinos, los que son considerados como amigos candidatos para cada usuario activo. Posteriormente, la lista de enlaces entre usuarios activos y sus candidatos fue evaluada en el grafo. Si un enlace fue encontrado en el grafo almacenado hasta el 1 de Enero de 2014,

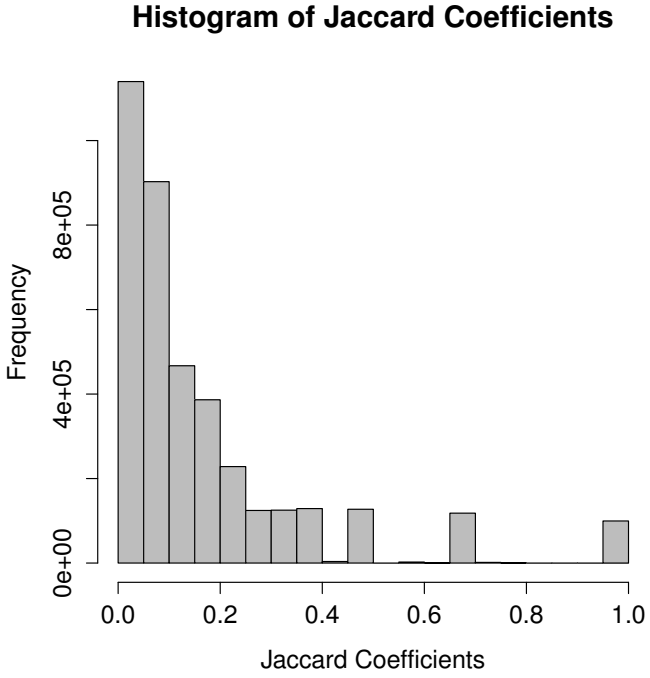


Figura 3.8: Histograma para coeficientes de Jaccard.

se elimina de la colección.

Si un enlace fue encontrado en el período de observación (del 1 al 25 de Enero), entonces se etiqueta como enlace real. Finalmente, si un enlace no fue encontrado, se etiqueta como enlace falso. 5.138, 1.875 y 810 enlaces reales fueron encontrados usando la estrategia postulada, y un total de 235.348, 135.022 y 61.027 casos fueron etiquetados para valores de umbrales iguales a 0.1, 0.2 y 0.3, respectivamente. Se balanceó el conjunto de datos usando re-muestreo para evitar el análisis de resultados sesgados.

En esta sección se exploró la factibilidad de una estrategia basada en *ranking* construyendo clasificadores para cada conjunto de datos etiquetado. Los rendimientos resultantes son mostrados en la tabla 3.4.

th = 0.1	Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Bayes	real	0.392	0.064	0.861	0.392	0.538	0.391	0.800	0.799
	false	0.936	0.608	0.606	0.936	0.736	0.391	0.800	0.742
	W Avg.	0.664	0.336	0.733	0.664	0.637	0.391	0.800	0.770
J48	real	**0.920**	**0.055**	0.943	0.920	0.931	0.864	0.975	0.965
	false	**0.945**	**0.080**	0.921	0.945	0.933	0.864	0.975	0.978
	W Avg.	0.932	0.068	0.932	0.932	0.932	0.864	0.975	0.971
Log	real	0.683	0.168	0.803	0.683	0.738	0.521	0.819	0.813
	false	0.832	0.317	0.724	0.832	0.774	0.521	0.819	0.782
	W Avg.	0.758	0.242	0.764	0.758	0.756	0.521	0.819	0.798
th = 0.2	Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Bayes	real	0.559	0.073	0.884	0.559	0.685	0.523	0.837	0.833
	false	0.927	0.441	0.678	0.927	0.783	0.523	0.837	0.787
	W Avg.	0.743	0.257	0.781	0.743	0.734	0.523	0.837	0.810
J48	real	**0.899**	**0.064**	0.933	0.899	0.916	0.836	0.972	0.967
	false	**0.936**	**0.101**	0.903	0.936	0.919	0.836	0.972	0.973
	W Avg.	0.917	0.083	0.918	0.917	0.917	0.836	0.972	0.970
Log	real	0.742	0.137	0.844	0.742	0.790	0.609	0.844	0.833
	false	0.863	0.258	0.770	0.863	0.813	0.609	0.844	0.802
	W Avg.	0.802	0.198	0.807	0.802	0.801	0.609	0.844	0.818
th = 0.3	Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Bayes	real	0.756	0.097	0.887	0.756	0.816	0.666	0.862	0.854
	false	0.903	0.244	0.787	0.903	0.841	0.666	0.862	0.815
	W Avg.	0.829	0.170	0.837	0.829	0.829	0.666	0.862	0.834
J48	real	**0.919**	**0.062**	0.937	0.919	0.928	0.857	0.975	0.972
	false	**0.938**	**0.081**	0.920	0.938	0.929	0.857	0.975	0.975
	W Avg.	0.928	0.072	0.929	0.928	0.928	0.857	0.975	0.973
Log	real	0.823	0.102	0.890	0.823	0.855	0.723	0.871	0.864
	false	0.898	0.177	0.835	0.898	0.865	0.723	0.871	0.824
	W Avg.	0.860	0.140	0.862	0.860	0.860	0.723	0.871	0.844

Cuadro 3.4: Medidas de rendimiento para el problema de clasificación de enlaces real/falso usando umbrales de localidad para la selección de candidatos.

Como muestra la tabla 3.4, los resultados obtenidos por la estrategia postulada supera el primer acercamiento. El uso de umbrales de localidad impacta en el rendimiento en varios puntos porcentuales, hecho que ilustra el balance entre *precision* y *recall*. Valores altos del umbral limitan la cantidad de semillas y en consecuencia, la cantidad de candidatos es menor. Por lo tanto, la precisión del método mejora pero a un costo de bajas tasas de *recall*. Por otro lado, el uso de valores bajos de umbral permite recuperar más semillas, reduciendo la precisión de los clasificadores.

Capítulo 4

Evolución de la red

En los capítulos anteriores se analizó la efectividad de la localidad en la predicción de enlaces, mostrando resultados acordes a los supuestos iniciales y a las propiedades inherentes de los grafos dispersos. Es por esto que en este capítulo se explorarán posibles estrategias a tomar para mejorar la predicción mediante un enfoque que aproxime la dinámica y evolución de la red.
Los análisis presentados en este capítulo se diseñan en base a los resultados del capítulo 3, donde se analiza la importancia de índices de localidad para el problema de predicción y de medidas asociadas con ciertas propiedades de los grafos. Dado que al delimitar el umbral de localidad se mejoraron los resultados de *precision*, ésta será la principal motivación para formular las configuraciones experimentales, aprovechando la localidad para mejorar la predicción.

4.1. Segmentación y ordenamiento

En los experimentos con datos sintéticos fueron generadas distintas redes en base a tres modelos generativos: modelos de grafos aleatorios, también conocidos como grafos Erdös-Rényi [12], Watts-Strogatz [51], y Barabási-Albert [3]. En la tabla 4.1 se pueden observar los parámetros utilizados para cada red.

Para cada uno de estos modelos se ejecutó un proceso de poda de acuerdo a un factor

Red	Parámetro	Valor
Barabási-Albert	n	N
	m	$\frac{Cluster_{coef} \cdot (n-1)}{2}$
	Dirigida	Falso
Erdös-Rényi	n	N
	p	$Cluster_{coef}$
	Dirigida	Falso
Watts-Strogatz	Dimensión	1
	n	N
	Alcance (vec.)	$\frac{Cluster_{coef} \cdot (n-1)}{2}$
	p	0.1

Cuadro 4.1: Parametros de configuración de redes

tomado como parámetro experimental que puede controlar la tasa de enlaces eliminados respecto de toda la red. Posteriormente, se procede a realizar predicción de enlaces a partir de los enlaces eliminados, donde los enlaces recomendados que fueron eliminados son considerados exitosos, y enlaces recomendados pero que no fueron removidos son considerados como casos de fracaso. De esta manera el factor de poda controla la tasa de soluciones reales respecto del total de candidatos. Un factor de poda bajo corresponde a una tasa menor de enlaces a ser predichos sobre el total de candidatos, haciendo el problema más difícil.
Cada una de las redes fue particionada con un algoritmo de *clustering* simple, en este caso en particular K-Means [32], usando cada una de las columnas de la matriz de adyacencia como vectores en el espacio n-dimensional. En la tabla 4.2 se puede ver un ejemplo de matriz de adyacencia, donde cada conexión entre par de nodos está representada por un 1, y un 0 en caso contrario. Al utilizar la matriz de adyacencia como representación vectorial se fuerza a que las particiones realizadas estén fuertemente enlazadas entre vecinos, a diferencia de usar distancia entre nodos, lo cual suaviza la restricción de localidad aplicada al método de *clustering*.

	1	2	3	4	5
1	0	0	1	0	1
2	0	0	0	1	0
3	1	0	0	1	0
4	0	1	1	0	0
5	1	0	0	0	0

Cuadro 4.2: Matriz de adyacencia de ejemplo

Posteriormente, usando una de las tres funciones de puntuación utilizadas en la red Skout (autoridad, grado, y transitividad) en el capítulo 3, se da paso a un ordenamiento intra-*cluster*. El algoritmo 4.1 muestra el proceso para llevar a cabo los experimentos sintéticos.

Algorithm 4.1 Algoritmo de *ranking* de enlaces con particionamiento previo

```
procedure LINK RANKING(prune_rate, k)
    g ← GenGraph()
    g_p ← PruneGraph(prune_rate)
    clusters ← Kmeans(g_p, k)
    cluster_measures ← ∅
    top_cand ← ∅
    for each cluster in clusters do
        cluster_measures ← calculates_measures(cluster)
    for each cluster in clusters do
        top_k ← RankCluster(cluster, cluster_measures[cluster])
```

4.2. Parámetros utilizados

Para los experimentos con datos sintéticos se utilizaron distintos parámetros para abordar distintos escenarios. En la tabla 4.3 se puede observar el rango de parámetros utilizados, resultando en una ejecución del experimento para cada posible combinación de parámetros. Es decir, se ejecutaron 1.875 instancias del experimento.

Parámetro	Valores
Tipo de red	Barabási; Erdös-Rényi; Watts Strogatz
Medida	Autoridad; Grado; Transitividad
Factor de poda	0.05-0.25
Cluster	3-7
Nodos	200-1000
Coef. *Clustering*	0.02-0.10
Candidatos	1-10

Cuadro 4.3: Parámetros utilizados para los experimentos sintéticos

4.3. Análisis de resultados

4.3.1. Análisis particular

Dentro de la variedad de ejecuciones de los experimentos realizados con datos sintéticos se generaron redes de 1000 nodos basadas en los 3 modelos generativos: Watts-Strogatz, Barabási-Albert y Erdös-Rényi. Para cada uno de estas redes se ejecutó el proceso descrito en el algoritmo 4.1. Las tablas 4.4 y 4.5 muestran resultados de *precision* y *recall* para 3 tipos de redes, donde E-R, W-S y B-A corresponden a Erdös-Rényi, Watts-Strogatz y a Barabási Albert respectivamente. Se muestra el valor promedio para las primeras 10 recomendaciones, las cuales son micro-promedios, es decir, los resultados corresponden a promedios para todos los nodos de la red. La tabla 4.4 muestra los resultados para un factor de poda de 0.1, y la tabla 4.5 muestra los resultados para un factor de poda de 0.25. En el sitio web de *www.pandasinc.cl/linkprediction* se pueden encontrar más resultados y combinaciones de parámetros.

	k	E-R P@10	B-A P@10	W-S P@10	E-R R@10	B-A R@10	W-S R@10
Autoridad	3	0.0126	0.0604	0.0516	0.0001	0.0022	0.0081
	4	0.0121	0.0462	0.0798	0.0005	0.0015	0.0169
	5	0.0121	0.0467	**0.1114**	0.0005	0.0010	0.0174
	6	0.0110	0.0328	**0.1310**	0.0005	0.0007	0.0232
	7	0.0104	0.0424	**0.1552**	0.0008	0.0010	0.0298
Grado	3	0.0110	0.0612	0.0312	0.0001	0.0026	0.0072
	4	0.0110	0.0464	0.0513	0.0004	0.0012	0.0143
	5	0.0110	0.0465	0.0400	0.0005	0.0010	0.0143
	6	0.0112	0.0326	0.0746	0.0005	0.0007	0.0212
	7	0.0112	0.0412	**0.1066**	0.0008	0.0010	0.0268
Transitividad	3	0.0050	0.0406	0.0236	0.0001	0.0026	0.0074
	4	0.0120	0.0412	0.0311	0.0005	0.0014	0.0143
	5	0.0102	0.0405	**0.0400**	0.0005	0.0010	0.0122
	6	0.0126	0.0333	**0.0441**	**0.0005**	0.0007	0.0204
	7	0.0114	0.0384	**0.0472**	**0.0008**	0.0010	0.0212

Cuadro 4.4: Resultados de *precision* y *recall* obtenidos con un factor de poda de 0.10

	k	E-R P@25	B-A P@25	W-S P@25	E-R R@25	B-A R@25	W-S R@25
Autoridad	3	0.0272	0.0712	0.0988	0.0018	0.0033	0.0144
	4	0.0241	0.0688	0.0155	0.0021	0.0025	0.0254
	5	0.0280	0.0612	**0.2028**	0.0022	0.0019	0.0255
	6	0.0282	0.0575	**0.2616**	0.0014	0.0018	**0.0339**
	7	0.0276	0.0510	**0.2901**	0.0017	0.0013	**0.0399**
Grado	3	0.0264	0.0791	0.0744	0.0018	0.0033	0.0122
	4	0.0224	0.0689	0.1135	0.0020	0.0024	0.0178
	5	0.0264	0.0616	0.1486	0.0022	0.0018	0.0231
	6	0.0270	0.0579	**0.2235**	0.0014	0.0018	**0.0336**
	7	0.0272	0.0520	**0.2735**	0.0017	0.0013	**0.0377**
Transitividad	3	0.0244	0.0512	0.0485	0.0016	0.0033	0.0102
	4	0.0212	0.0522	0.0603	0.0022	0.0026	0.0124
	5	0.0270	0.0614	0.0850	0.0022	0.0019	0.0146
	6	0.0292	0.0507	**0.1015**	0.0014	0.0018	**0.0147**
	7	0.0264	0.0472	**0.1177**	0.0017	0.0013	**0.0141**

Cuadro 4.5: Resultados de *precision* y *recall* obtenidos con un factor de poda de 0.25

4.3.2. Análisis general

En este trabajo experimental se aborda el problema de *Link Prediction* segmentando el grafo completo para luego analizar cada partición resultante en términos de grado, autoridad y transitividad. Algunos de los resultados de los experimentos con datos sintéticos muestran ciertas características relevantes. Una de ellas está relacionada con el parámetro de poda de enlaces, ya que este se relaciona con el tiempo de evolución de una red, es decir, un factor de poda bajo emula un período de observación corto para el problema de *Link Prediction*, por lo que se obtienen peores resultados si lo comparamos con factores de poda más altos, tal como se muestra en la figura 4.1. Esto se debe a que la predictibilidad mejora a medida que se amplía el horizonte de observación para problemas de predicción. En el caso particular de este experimento, lo que sucede es que al aumentar la razón de poda se está aumentando la cantidad de enlaces posibles que son candidatos a ser escogidos y además cada uno de estos enlaces candidatos que se agrega son ocurrencias de enlaces existentes en el grafo original, por lo que mejorarán los resultados al escogerlos.

Así como el factor de poda es un parámetro muy importante al momento de realizar mediciones sobre el grafo, también lo es el coeficiente de *clustering*, que regula que tan denso es el grafo. Mientras más denso un grafo, más posibilidades de éxito existen al predecir enlaces (ver figura 4.2), ya que existe un universo mayor de posibles candidatos. También es importante notar en la figura 4.2 que el valor promedio de *recall* va disminuyendo a medida que el coeficiente de *clustering* aumenta, esto debido a que se tiene una mayor cantidad de casos posibles, incrementándose el número de falsos negativos al momento de evaluar.

Por el lado de la segmentación los resultados muestran que este proceso es clave al momento de predecir enlaces, ya que se obtienen resultados muy distintos para diferentes segmentaciones. Es el caso del modelo de generación de grafos aleatorios Watts Strogatz

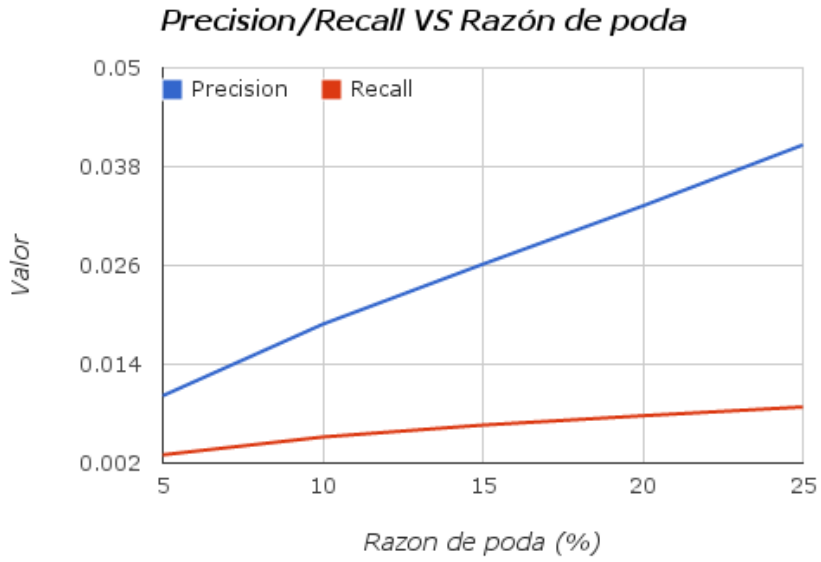


Figura 4.1: Variación de *precision* y *recall* respecto al factor de poda utilizado

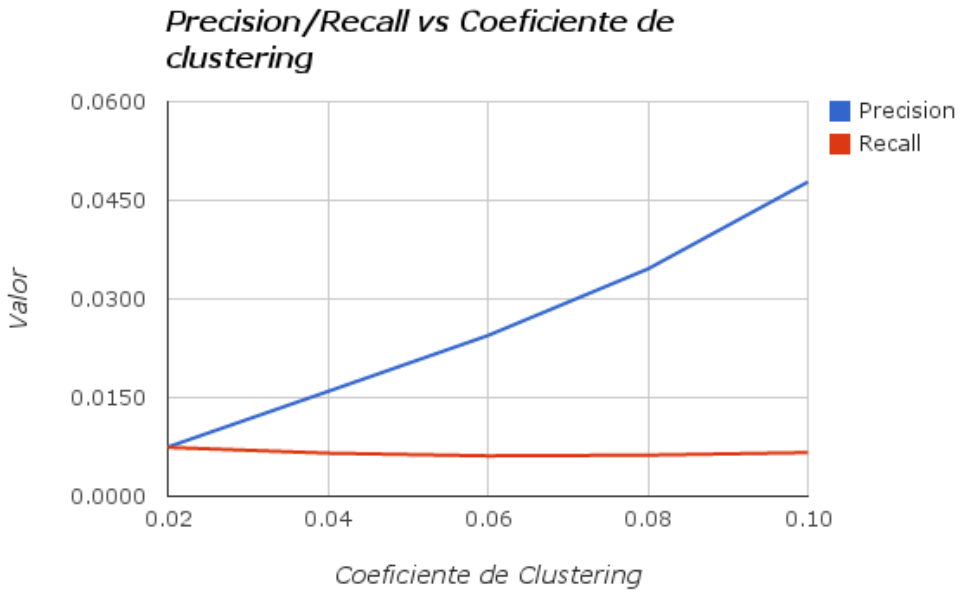


Figura 4.2: Variación de *precision* y *recall* promedio respecto al coeficiente de *clustering* utilizado.

que obtiene resultados de precisión cercanos a 0.01 con 3 *clusters* y resultados cercanos a 0.3 con 7 *clusters*. En la figura 4.3 se puede ver la variación de *precision* y *recall* promedio respecto del número de *clusters* utilizados. Finalmente, el número óptimo de *clusters* dependerá de la naturaleza de la red, la cantidad de nodos y la densidad de la red entre otros factores.

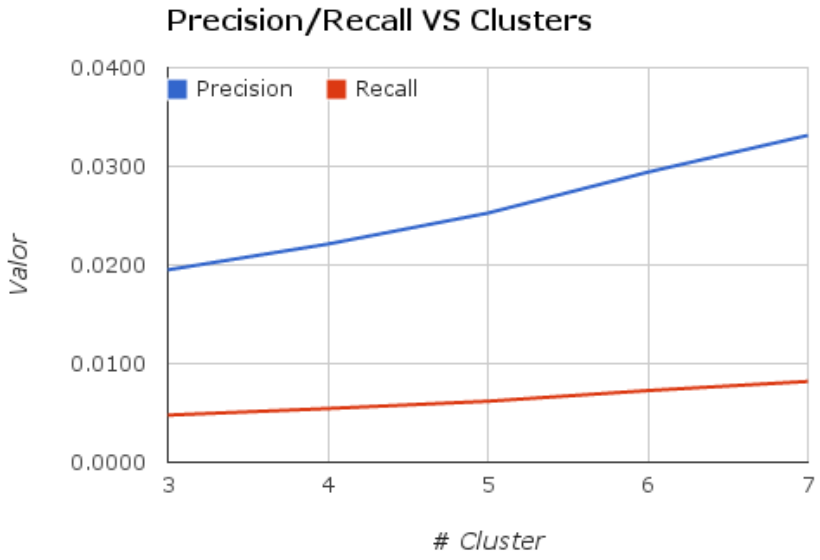


Figura 4.3: Variación de *precision* y *recall* respecto al número de clusters utilizado en el proceso de segmentación

El último de los parámetros que fue utilizado en un rango distinto de valores fue la elección de los mejores k candidatos, valor que influye en los valores obtenidos de *precision* y *recall*, generando un balance entre estas métricas. Tal como se muestra en la figura 4.4, a medida que el valor de k aumenta, el valor de *precision* va disminuyendo y en el caso de la medida de *recall* va aumentando (ver figura 4.5). Esto se debe principalmente a que a medida que se escogen más candidatos, se abarca una mayor proporción del conjunto total de soluciones posibles, empeorando la calidad de la recomendación al otorgar más resultados que no corresponden a enlaces creados efectivamente.

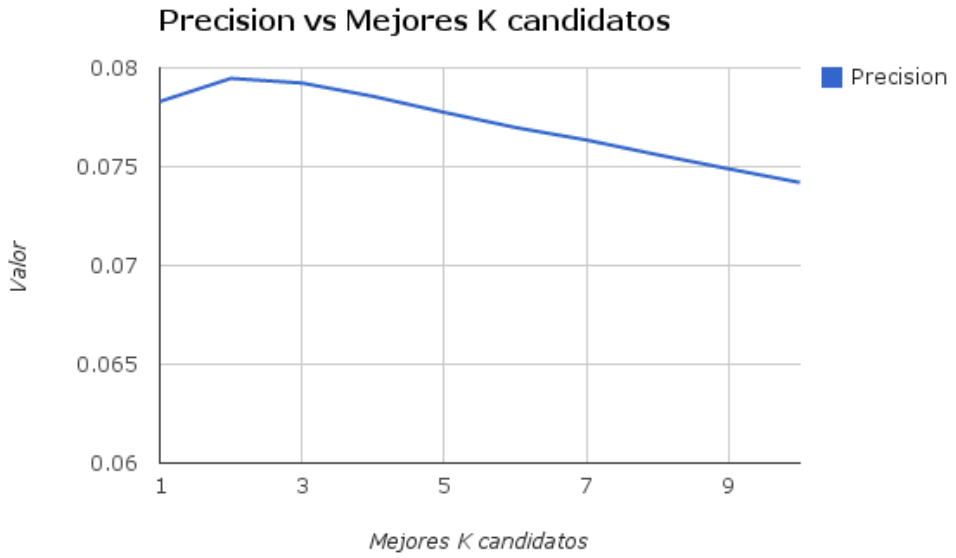


Figura 4.4: Variación de *precision* promedio respecto a los mejores top k candidatos

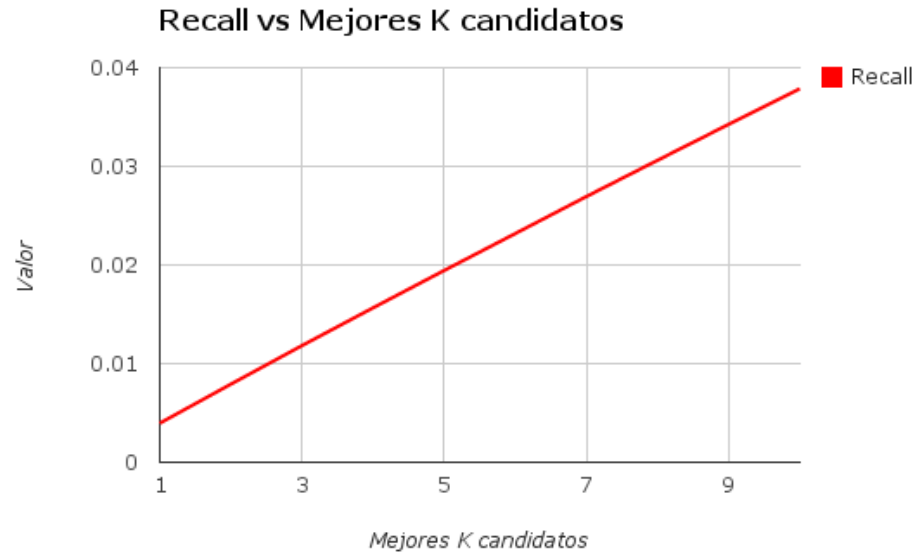


Figura 4.5: Variación de *recall* promedio respecto a los mejores top k candidatos

En *Link Prediction* explotar la localidad o explorar los amigos de los amigos es algo intuitivo para intentar obtener mejores resultados, pero los resultados obtenidos en este experimento sintético basados en transitividad muestran peores resultados que los basados en autoridad y grado (ver tabla 4.6). Una posible causa de esta situación es que la transitividad explota la localidad en la red mientras que el *clustering* explota las características estructurales globales de la red, obteniendo mejores resultados al segmentar la red. El caso contrario ocurre con las redes tipo Barabási, donde segmentar la red lleva a resultados pobres, empeorando los resultados de predicción a medida que se aumenta el número de *clusters* utilizados.

Tipo de Red	**Medida**	***Precision***	***Recall***
Barabási	Autoridad	0.0225	0.0026
Barabási	Transitividad	0.0150	0.0010
Barabási	**Grado**	**0.0242**	**0.0030**
Erdös-Rényi	Autoridad	0.0095	0.0015
Erdös-Rényi	Transitividad	0.0094	0.0014
Erdös-Rényi	Grado	0.0095	0.0015
Watts-Strogatz	**Autoridad**	**0.6512**	**0.2144**
Watts-Strogatz	Transitividad	0.3027	0.1005
Watts-Strogatz	Grado	0.4785	0.1580

Cuadro 4.6: Promedio de *precision* y *recall* según medida y modelos utilizados

Las medidas para caracterizar y escoger candidatos que mejor se comportan depende del tipo de red. En el caso de las redes tipo Barabási-Albert, la predicción basada en grado supera ligeramente a la basada en autoridad debido a que el modelo generativo de redes aleatorias Barabási-Albert se construye con probabilidad proporcional al grado de los nodos. En las redes tipo Watts Strogatz ocurre lo contrario, donde la predicción basada en autoridad supera levemente a la predicción basada en grado, resultados similares a los obtenidos con la red de Skout en el capítulo 3. Esto se debe a que la red es generada a partir de una red regular tipo anillo con cierta probabilidad p de ser modificada. Lo anterior sugiere que el modelo generativo Watts-Strogatz puede aproximar la evolución de la estructura de la red de Skout, otorgando un medio para realizar *Link Prediction* en una red segmenta-

da, complementando estrategias FOAF con estrategias de ranking basadas en la autoridad intra-cluster.

Capítulo 5

Conclusiones

5.1. Conclusiones

En la primera parte de este trabajo se exploró el uso de características basadas en la representación como grafo de la red social Skout. Los resultados muestran que el uso de umbrales de localidad es efectivo para el problema de *Link Prediction* y puede ser combinado exitosamente con medidas basadas en la topología del grafo, como lo son autoridad, grado y transitividad. En este trabajo en particular se obtuvieron grandes diferencias entre las características utilizadas, mostrando la importancia de elegir características apropiadas. En este trabajo se obtuvieron buenos resultados en términos de *precision* pero con el costo asociado de perder *recall* y viceversa. Esto obedece al balance entre *precision* y *recall*, es decir, a medida que se mejora la métrica de *precision*, el *recall* disminuye. Esto queda en evidencia en los dos acercamientos al problema de *Link Prediction* realizados en este trabajo.

En el caso de uso de Skout con datos reales, al disminuir la cantidad de semillas a utilizar con un umbral de localidad, se disminuye la cantidad posible de candidatos, por lo que la medida de *recall* disminuye a medida que el umbral de localidad es mayor, es decir, cuando este valor es más restrictivo. Sin embargo, al restringir el conjunto de candidatos, se está mejorando a su vez el valor de *precision*, ya que la recomendación se basa en un ordenamiento según las características utilizadas, por lo que la proporción entre soluciones válidas e inválidas es más favorable hacia las primeras. Cabe destacar que sólo se puede

notar una mejora en *precision* si las medidas mediante las cuáles se ordenan los candidatos son efectivas y adecuadas para realizar esta tarea de ordenamiento.
En el caso de los experimentos con datos sintéticos, se escoge explícitamente el valor k de candidatos a ser recomendados. Mientras más bajo es este valor de k, mejor es la métrica de *precision* que se obtiene en desmedro del *recall*, evidenciándose el mismo fenómeno observado con el caso de estudio de Skout.
Los resultados de *precision* obtenidos son bajos en comparación a otros métodos encontrados en la literatura, pero esto se debe a que en este trabajo sólo se utilizan características topológicas, sin combinar éstas con otro tipo de información, como la información de perfiles de usuarios, preferencias, transacciones, entre otras posibilidades.
Los resultados obtenidos en las pruebas reales en la sección 3.5.1 muestran que se obtiene un mejor rendimiento al utilizar medidas de autoridad y en las pruebas sintéticas se obtienen diferentes resultados dependiendo de la característica utilizada. En el caso de las pruebas sintéticas, las redes generadas con el modelo generativo Watts-Strogatz tuvieron mejor rendimiento utilizando medida de autoridad, lo que prueba la hipótesis de trabajo donde la naturaleza de la red determina el método que mejor se adecúa para el problema de *Link Prediction*.

5.2. Trabajo futuro

5.2.1. Aproximación del número óptimo de particiones de un grafo

El particionamiento espectral utilizado como método de predicción de enlaces en el trabajo de Symeonidis *et al* [48] muestra resultados prometedores. En este trabajo de tesis se exploró la utilización de este enfoque para mejorar la predicción de enlaces, pero el método es impracticable a gran escala, es decir, no puede ser aplicado en problemas reales. El método de particionamiento espectral consiste en calcular la matriz Laplaciana norma-

lizada, tal como se muestra en las ecuaciones 5.1 y 5.2, donde D es la matriz diagonal de grados y A es la matriz de adyacencia. Sobre la matriz L_{norm}, se calculan los k vectores propios con sus correspondientes k valores propios. Los k vectores son utilizados para particionar el grafo, caracterizando cada nodo del grafo según su ubicación y distancia en el espacio k-dimensional generado por los k vectores propios:

$$L = D - A \tag{5.1}$$

$$L_{norm} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} \tag{5.2}$$

El método funciona de manera óptima y entrega buenos resultados, pero depende del valor de k entregado. Determinar el valor de k *a priori* no es trivial, por lo que esta fase es primordial para el proceso posterior. Una técnica para determinar el número de particiones de un grafo es analizar la secuencia de valores propios en orden creciente, donde el número de valores propios cero o cercanos a cero en el caso de una aproximación corresponde al número de particiones adecuado. En la figura 5.1 se observa un grafo aleatorio tipo Erdös-Rényi de 40 nodos generado con probabilidad $p = 0.1$ sobre el cual se calculan los valores propios, conocidos como el espectro del grafo, el cual es mostrado en el gráfico de la figura 5.1. En éste gráfico se aprecia que existe un valor propio con valor cero y los valores siguientes tienen valores distintos de cero, lo que indica la existencia de solo una partición o componente.

Debido a lo anterior, sería oportuno explorar una manera de aproximar el número de valores propios con valor cero o cercanos a cero. Una manera tentativa de hacerlo es obtener los primeros k valores propios en orden creciente, hasta encontrar un cambio notable en los valores propios. Esto indicaría que nos encontramos frente al número de particiones adecuado a utilizar, dando paso al particionamiento espectral usando el número de valores

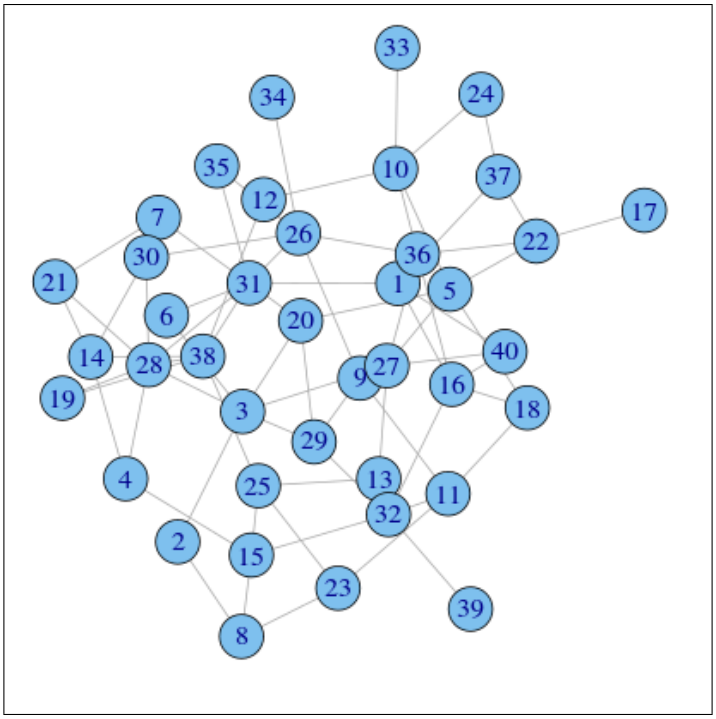


Figura 5.1: Grafo aleatorio Erdös-Rényi

propios existentes antes del umbral encontrado. Realizar esta tarea de encontrar un cambio notorio en los valores propios puede ser igual de complejo y costoso, por lo que una alternativa tentativa sería eliminar nodos hoja, es decir, aquellos que están conectados a un sólo nodo o los que son nodos extremos en un sub-grafo.

Realizar este trabajo puede abrir diversar posibilidades, siendo útil para distintas áreas, tal como se plantea en el trabajo de Jaouen *et al* [17].

5.2.2. Utilizar un índice de distancia diferente

Tal como se muestra en el trabajo de Symeonidis *et al* ([48]), se utiliza un índice de distancia denominado Bray-Curtis. En vez de utilizar este índice, sería interesante explorar un índice que escale a medida que se aumenta la dimensionalidad del problema, es decir,

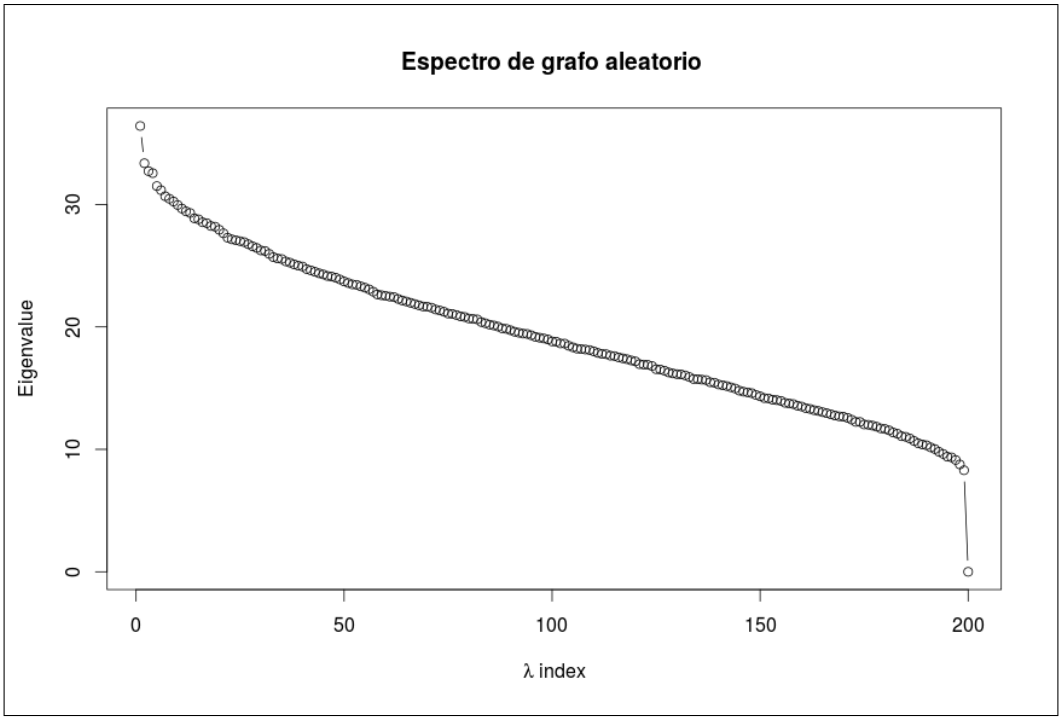


Figura 5.2: Espectro de grafo aleatorio Erdös-Rényi con $p = 0.1$

cuando los grafos son más grandes y están particionados por un número mayor de *clusters*. En la ecuación 5.3 se muestra un índice sugerido a explorar, el cual se comportaría mejor en situaciones en las que el grafo contiene 3 o más posibilidades de particionamiento, debido a que la multiplicación no superpone errores, evitando la anulación de errores opuestos.

$$D_{(x,y)} = \begin{cases} \frac{D_{(C_x,C_y)}}{D_{(x,C_x)} \cdot D_{(y,C_y)}}, & C_x \neq C_y \\ D_{(x,C_x)} \cdot D_{(y,C_y)}, & C_x = C_y \end{cases} \tag{5.3}$$

5.2.3. Estimar la naturaleza de una red

En el cápítulo 4 se evaluó el rendimiento de 3 medidas distintas para tres modelos generativos de redes distintos, obteniendo diferentes resultados según el modelo y la medida utilizada (ver Tabla 4.6). Como trabajo futuro sería interesante analizar representaciones de redes complejas reales y determinar a qué tipo de modelo generativo se asemejan con el objetivo de escoger la mejor combinación de técnicas y medidas para mejorar la predicción de nuevos enlaces.

Bibliografía

[1] Lada A. Adamic y Eytan Adar. "Friends and Neighbors on the Web". En: *Social Networks* 25.3 (2003), págs. 211-230.

[2] Lars Backstrom y J Leskovec. "Supervised random walks: predicting and recommending links in social networks". En: *Proceedings of the fourth ACM international ...* (2011), págs. 635-644.

[3] A. L. Barabasi y R. Albert. "Emergence of scaling in random networks". En: *Science* 286 (1999), págs. 509-512.

[4] S. Bastani, A.K. Jafarabad y M.H.F. Zarandi. "Fuzzy models for link prediction in social networks". En: *International Journal of Intelligent Systems* (), págs. 768-786. ISSN: 08848173 (ISSN).

[5] Catherine A Bliss y col. "An Evolutionary Algorithm Approach to Link Prediction in Dynamic Social Networks". En: (2013). arXiv: arXiv:1304.6257v2.

[6] John S. Breese, David Heckerman y Carl Kadie. "Empirical Analysis of Predictive Algorithms for Collaborative Filtering". En: *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. UAI'98. Morgan Kaufmann Publishers Inc., 1998, págs. 43-52. ISBN: 1-55860-555-X.

[7] Nitin Chiluka, Nazareno Andrade y Johan A. Pouwelse. "A Link Prediction Approach to Recommendations in Large-Scale User-Generated Content Systems." En: *ECIR*. Ed. por Paul Clough y col. Vol. 6611. Lecture Notes in Computer Science. Springer, 2011, págs. 189-200. ISBN: 978-3-642-20160-8.

[8] Alvin Chin y Mark Chignell. *Automatic detection of cohesive subgroups within social hypertext: A heuristic approach*. 2008. DOI: 10.1080/13614560802357180.

[9] Aaron Clauset, M. E. J. Newman y Cristopher Moore. "Finding community structure in very large networks". En: *Phys. Rev. E* 70 (6 dic. de 2004), pág. 066111. DOI: 10.1103/PhysRevE.70.066111.

[10] Charo I. Del Genio, Thilo Gross y Kevin E. Bassler. "All Scale-Free Networks Are Sparse". En: *Phys. Rev. Lett.* 107 (2011), pág. 178701. DOI: 10.1103/PhysRevLett.107.178701.

[11] Daniel M. Dunlavy, Tamara G. Kolda y Evrim Acar. "Temporal Link Prediction Using Matrix and Tensor Factorizations". En: *ACM Transactions on Knowledge Discovery from Data* 5.2 (feb. de 2011), págs. 1-27. ISSN: 15564681. DOI: 10.1145/1921632.1921636.

[12] P. Erdös y A. Rényi. "On random graphs, I". En: *Publicationes Mathematicae (Debrecen)* 6 (1959), págs. 290-297.

[13] Matias Estrada y Marcelo Mendoza. "Affinity Prediction in Online Social Networks". En: *IET* abs/1408.2871 (2014).

[14] Francois Fouss Francois Fouss y col. "Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation". En: *IEEE Transactions on Knowledge and Data Engineering* 19 (2007). ISSN: 1041-4347. DOI: 10.1109/TKDE.2007.46.

[15] P Jaccard. "La distribution de la flore dans la zone alpine". En: *Revue Générale des Sciences* (). ISSN: ¡null¿.

[16] SMAZ Jacobs, Winter Mason y Aaron Clauset. "Detecting Friendship Within Dynamic Online Interaction Networks". En: (mar. de 2013), pág. 11. arXiv: 1303.6372.

[17] Vincent Jaouen y col. "4DGVF segmentation of vector-valued images". En: *2014 IEEE International Conference on Image Processing, ICIP 2014, Paris, France, October 27-30, 2014*. 2014, págs. 11-15. DOI: 10.1109/ICIP.2014.7025001.

[18] Glen Jeh y Jennifer Widom. "SimRank: a measure of structural-context similarity". En: *Proceedings of the eighth ACM SIGKDD international ...* (2002), págs. 1-11. DOI: 10.1145/775047.775126.

[19] Indika Kahanda y Jennifer Neville. "Using Transactional Information to Predict Link Strength in Online Social Networks". En: *ICWSM*. 2009.

[20] Leo Katz. "A new status index derived from sociometric analysis". En: *Psychometrika* VOL. 18, NO. 1 (1953), págs. 39-43.

[21] J. Kleinberg. "Authoritative Sources in a Hyperlinked Environment". En: *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*. 1998.

[22] Gueorgi Kossinets. "Effects of missing data in social networks". En: *Social Networks* 28 (2006), págs. 247-268. ISSN: 03788733. DOI: 10.1016/j.socnet.2005.07.002. arXiv: 0306335 [cond-mat].

[23] Joonhee Kwon y Sungrim Kim. "Friend recommendation method using physical and social context". En: *International Journal of Computer Science and ...* 10.11 (2010), págs. 116-120.

[24] E A Leicht, Petter Holme y M E J Newman. "Vertex similarity in networks." En: *Physical review. E, Statistical, nonlinear, and soft matter physics* 73 (2006), pág. 026120. ISSN: 1539-3755. DOI: 10.1103/PhysRevE.73.026120. arXiv: 0510143 [physics].

[25] David Liben-Nowell y Jon Kleinberg. "The Link-prediction Problem for Social Networks". En: *J. Am. Soc. Inf. Sci. Technol.* 58.7 (mayo de 2007), págs. 1019-1031. ISSN: 1532-2882. DOI: 10.1002/asi.v58:7.

[26] Ryan N Lichtenwalter, Jake T Lussier y Nitesh V Chawla. "New perspectives and methods in link prediction". En: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2010, págs. 243-252.

[27] Dekang Lin. "An Information-Theoretic Definition of Similarity". En: *Quality*. Vol. 1. 1998, págs. 296-304. ISBN: 1558605568. DOI: 10.1.1.55.1832.

[28] L Linyuan. "Link Prediction in Complex Networks : A Survey". En: October 2010 (). arXiv: arXiv:1010.0725v1.

[29] Shuchuan Lo y Chingching Lin. "WMR–A Graph-Based Algorithm for Friend Recommendation". En: *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)* (dic. de 2006), págs. 121-128. DOI: 10.1109/WI.2006.202.

[30] Linyuan Lü, Ci-Hang Jin y Tao Zhou. "Similarity index based on local paths for link prediction of complex networks." En: *Physical review. E, Statistical, nonlinear, and soft matter physics* 80 (2009), pág. 046122. ISSN: 1539-3755. DOI: 10.1103/PhysRevE.80.046122.

[31] Zhengdong Lu y col. "Supervised Link Prediction Using Multiple Sources". En: *Data Mining (ICDM), 2010 ...* (2010), págs. 1-17.

[32] J. MacQueen. *Some methods for classification and analysis of multivariate observations*. Berkeley, Calif., 1967.

[33] Miller McPherson, Lynn Smith-Lovin y James M. Cook. "Birds of a Feather: Homophily in Social Networks". En: *Annu. Rev. Sociol.* 27.1 (2001). Ed. por reasons for social interactions how do people group together, págs. 415-444. DOI: 10.1146/annurev.soc.27.1.415.

[34] Tsuyoshi Murata y Sakiko Moriyasu. "Link Prediction based on Structural Properties of Online Social Networks". En: *New Generation Computing* 26.3 (jun. de 2008), págs. 245-257. ISSN: 0288-3635. DOI: 10.1007/s00354-008-0043-y.

[35] M. E. J. Newman. "Assortative Mixing in Networks". En: *Phys. Rev. Lett.* 89 (2002), pág. 208701.

[36] M. E. J. Newman. "Finding community structure in networks using the eigenvectors of matrices". En: *Phys. Rev. E* 74 (3 sep. de 2006), pág. 036104. DOI: 10.1103/PhysRevE.74.036104.

[37] M. E. J. Newman y M. Girvan. "Finding and evaluating community structure in networks". En: *Phys. Rev. E* 69.2 (feb. de 2004), pág. 026113. DOI: 10.1103/PhysRevE.69.026113.

[38] M.E.J. Newman. "Clustering and preferential attachment in growing networks". En: *Physical Review E* 2 (), pág. 025102.

[39] M.E.J. Newman. "The Structure and Function of Complex Networks". En: *SIAM review* 45.2 (2003), págs. 167-256. ISSN: 0036-1445. DOI: 10.1137/S003614450342480.

[40] Pascal Pons y Matthieu Latapy. "Computing Communities in Large Networks Using Random Walks". English. En: *Computer and Information Sciences - ISCIS 2005*. Ed. por pInar Yolum y col. Vol. 3733. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2005, págs. 284-293. ISBN: 978-3-540-29414-6. DOI: 10.1007/11569596_31. URL: http://dx.doi.org/10.1007/11569596_31.

[41] David M. W. Powers. "Applications and Explanations of Zipf's Law". En: *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*. NeMLaP3/CoNLL '98. Sydney, Australia: Association for Computational Linguistics, 1998, págs. 151-160. ISBN: 0-7258-0634-6.

[42] E Ravasz y col. "Hierarchical organization of modularity in metabolic networks." En: *Science (New York, N.Y.)* 297 (2002), págs. 1551-1555. ISSN: 00368075. DOI: 10.1126/science.1073374. arXiv: 0209244 [cond-mat].

[43] Maayan Roth y col. "Suggesting friends using the implicit social graph". En: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2010, págs. 233-242.

[44] G Salton y M J McGill. *Introduction to Modern Information Retrieval*. Vol. 22. 1983, xv, 448 p. ISBN: 0070544840. DOI: 10.1093/llc/fql026.

[45] Donghyuk Shin, Si Si e IS Dhillon. "Multi-scale link prediction". En: *Proceedings of the 21st ACM international ...* (), pág. 215. DOI: 10.1145/2396761.2396792. arXiv: arXiv:1206.1891v1.

[46] Nitai B. Silva y col. "A graph-based friend recommendation system using Genetic Algorithm". En: *IEEE Congress on Evolutionary Computation*. IEEE, jul. de 2010, págs. 1-7. ISBN: 978-1-4244-6909-3. DOI: 10.1109/CEC.2010.5586144.

[47] R. Sø rensen, U. Zinko y J. Seibert. *On the calculation of the topographic wetness index: evaluation of different methods based on field observations*. 2006. DOI: 10.5194/hess-10-101-2006.

[48] Panagiotis Symeonidis y col. "From biological to social networks: Link prediction based on multi-way spectral clustering". En: *Data & Knowledge Engineering* 87.0 (2013), págs. 226 -242. ISSN: 0169-023X. DOI: http://dx.doi.org/10.1016/j.datak.2013.05.008.

[49] Hanghang Tong Hanghang Tong, C. Faloutsos y J.-Y. Pan. "Fast Random Walk with Restart and Its Applications". En: *Sixth International Conference on Data Mining (ICDM'06)* (2006). ISSN: 1550-4786. DOI: 10.1109/ICDM.2006.70.

[50] Chao Wang, Venu Satuluri y Srinivasan Parthasarathy. "Local Probabilistic Models for Link Prediction". En: *Seventh IEEE International Conference on Data Mining (ICDM 2007)* (oct. de 2007), págs. 322-331. DOI: 10.1109/ICDM.2007.108.

[51] D J Watts y col. "Collective dynamics of 'small-world' networks." En: *Nature* 393 (1998), págs. 440-2. ISSN: 0028-0836. DOI: 10.1038/30918. arXiv: 0803.0939v1.

[52] Giovanni Zappella, Alexandros Karatzoglou y L Baltrunas. “Games of Friends: a game-theoretical approach for link prediction in online social networks”. En: *Workshops at the Twenty-Seventh ...* (2013).

[53] Yunpeng Zhao, Elizaveta Levina y Ji Zhu. “Link prediction for partially observed networks”. En: *arXiv preprint arXiv:1301.7047* (ene. de 2013), págs. 1-15. arXiv: 1301.7047.