

Pontificia Universidad Javeriana Cali  
Facultad de Ingeniería y Ciencias  
Ingeniería de Sistemas y Computación  
Matemáticas Aplicadas  
Trabajo de Grado

# Sistemas de Recomendación Basados en la Inferencia de Redes Sociales

Santiago Uribe Pastás

Director: Jorge Finke Ortiz  
Co-director: Carlos Ernesto Ramírez Ovalle

Enero 2022





Santiago de Cali, Enero 2022.

Señores

**Pontificia Universidad Javeriana Cali.**

Gerardo Mauricio Sarria Montemiranda

Director Carrera de Ingeniería de Sistemas y Computación.

Diana Haidive Bueno Carreño

Directora Carrera de Matemáticas Aplicadas.

Cali.

Cordial Saludo.


Por medio de la presente me permito informarle que el estudiante de Ingeniería de Sistemas y Matemáticas Aplicadas, Santiago Uribe Pastás (cod: 8925546), trabaja bajo nuestra dirección, en el proyecto de grado titulado “Sistemas de Recomendación Basados en la Inferencia de Redes Sociales”.

Atentamente,



---

Jorge Finke Ortiz



---

Carlos Ernesto Ramírez Ovalle

Santiago de Cali, Enero 2022.

Señores

**Pontificia Universidad Javeriana Cali.**

Gerardo Mauricio Sarria Montemiranda

Director Carrera de Ingeniería de Sistemas y Computación.

Diana Haidive Bueno Carreño

Directora Carrera de Matemáticas Aplicadas.

Cali.

Cordial Saludo.

Me permito presentar a su consideración el trabajo de grado titulado “Sistemas de Recomendación Basados en la Inferencia de Redes Sociales” con el fin de cumplir con los requisitos exigidos por la Universidad para optar al título de Ingeniero de Sistemas y Matemático Aplicado.

Al firmar aquí, doy fe que entiendo y conozco las directrices para la presentación de trabajos de grado de la Facultad de Ingeniería aprobadas el 26 de Noviembre de 2009, donde se establecen los plazos y normas para el desarrollo del anteproyecto y del trabajo de grado.

Atentamente,

Santiago Uribe P.

Santiago Uribe Pastás

Código: 8925546

# Agradecimientos

En primer lugar, le agradezco a Dios por permitirme realizar este Trabajo de Grado y capacitarme en cada uno de los retos que se me presentaron durante todo el proceso. Agradezco a mis directores de tesis, Jorge Finke y Carlos Ramírez, por su guía durante la elaboración de este proyecto. Su amplio conocimiento y observaciones me ayudaron a encaminar las ideas que tuve. Me alegra haber coincidido con ellos en el interés por este tema apasionante y retador.

Doy gracias a mi familia, a mis padres Jaime y Maribel, y a mi hermana Catalina, por ser mi apoyo emocional a lo largo de mi vida y por enseñarme a amar las cosas que hago. También, les agradezco a mis amigos por sus palabras de ánimo y su compañía en el transcurso de este camino.



# Resumen

Los sistemas de recomendación son una parte vital para empresas con una activa participación en la web. Dichas empresas requieren de estrategias que les permitan apalancarse en las calificaciones de productos por parte de usuarios, para poder brindar futuras recomendaciones a otros usuarios. En la última década se han desarrollado diversos algoritmos para la recomendación de películas, siendo el algoritmo llamado *Matrix Factorization* uno de los más populares. El enfoque de esta tesis es evaluar el desempeño de este algoritmo de recomendación en escenarios en los que se pueden inferir redes sociales subyacentes, que caracterizan cierto tipo de interacción entre usuarios. En particular se utiliza el conjunto de datos *MovieLens*, el cual consta de aproximadamente cien mil valoraciones por parte de 671 usuarios sobre 9066 películas, durante el periodo del 29 de marzo de 1996 al 24 de septiembre de 2018. En esta tesis se investigará diferentes métodos para la inferencia de redes sociales, así como diferentes medidas topológicas sobre estas redes inferidas. Y con base en esta investigación, escoger y analizar un algoritmo que permita entender cómo el uso de la información de las redes sociales afecta el desempeño del algoritmo *Matrix Factorization*.

**Palabras Clave:** Sistemas de Recomendación, Redes Sociales Inferidas, *Matrix Factorization*, Grafos.





# Índice general

<b>1. Descripción del Problema</b>	<b>13</b>
1.1. Planteamiento del Problema . . . . .	13
1.1.1. Formulación . . . . .	14
1.1.2. Sistematización . . . . .	14
1.2. Objetivos . . . . .	14
1.2.1. Objetivo General . . . . .	14
1.2.2. Objetivos Específicos . . . . .	14
1.3. Justificación . . . . .	14
1.4. Delimitaciones y Alcances . . . . .	15
1.4.1. Entregables . . . . .	15
<b>2. Investigación de la Literatura</b>	<b>17</b>
2.1. Áreas Temáticas . . . . .	17
2.2. Trabajos Relacionados . . . . .	17
2.3. Métodos para Inferir Redes Sociales . . . . .	18
<b>3. Marco Teórico</b>	<b>21</b>
3.1. Factorización de Matrices . . . . .	21
3.2. Grafo . . . . .	23
3.3. Subgrafo Inducido . . . . .	23
3.4. Grafo Bipartito . . . . .	23
3.4.1. Proyección de un Grafo Bipartito . . . . .	24
3.5. Redes Sociales . . . . .	24
3.6. Medidas Topológicas de Redes Sociales . . . . .	24
3.7. Inferencia de Redes Sociales ( <i>NETINF</i> ) . . . . .	26
<b>4. Implementación</b>	<b>29</b>
4.1. Recomendación de Películas . . . . .	29
4.2. Datos en Representación de Grafos . . . . .	30
4.2.1. Grafo Bipartito . . . . .	30
4.2.2. Proyección del Grafo con Respecto a Usuarios . . . . .	31
4.3. Inferencia de Redes . . . . .	35
4.4. Recomendación de Películas Usando Redes Sociales Inferidas y sus Propiedades . . . . .	36
4.5. Observaciones Teóricas sobre los Algoritmos Usados . . . . .	37
<b>5. Conclusiones</b>	<b>39</b>
<b>Bibliografía</b>	<b>41</b>



# Introducción

Con el constante aumento del volumen de la información en línea, los sistemas de recomendación (SR) han sido una herramienta utilizada por un amplio conjunto de empresas, para facilitar el encontrar información relevante que satisfaga las necesidades e intereses de usuarios, y así incrementar sus niveles de satisfacción en la interacción con diferentes marcas y servicios.

Diferentes implementaciones de SR intentan sugerir todo tipo de productos para los usuarios. Hoy en día se encuentran en la mayoría de las plataformas de empresas con alta participación en la web. Se utilizan para recomendar música en Spotify, productos varios por Amazon, vídeos en YouTube, y hasta artículos de investigación a la comunidad científica [1]. En el caso del material audiovisual, plataformas de *streaming* (como Netflix, Amazon Prime Video y HBO) aprovechan una variedad de conjuntos de datos y técnicas de *Machine Learning* para generar las recomendaciones [2].

En los últimos años, se han generado diferentes algoritmos para la recomendación de películas, los cuales han demostrado una alta efectividad [3]. El filtrado colaborativo es uno de los enfoques más usados en SR. Este tipo de algoritmo parte de la base que si los usuarios han valorado artículos de forma similar en el pasado, es probable que valoren artículos semejantes de forma similar en el futuro. Aunque los modelos de filtrado colaborativo generan resultados prometedores en diferentes escenarios, no explotan de forma explícita patrones ocultos en los datos de redes sociales que podrían agregar valor a la predicción de las valoraciones de los usuarios.

Inferir las redes sociales subyacentes a un conjunto de usuarios permite caracterizar formalmente cómo se encuentran conectados los usuarios. Esto es útil para determinar, por ejemplo, cómo se difunde la información de cierto producto entre usuarios, lo que permitiría recomendar nuevos productos a los usuarios o maximizar las ventas de un producto en específico.

Sinha *et al.* [4] mostraron que, si se da a elegir entre las recomendaciones de la red de amigos y las de los usuarios en general, en términos de calidad y utilidad, se prefieren las recomendaciones de los amigos. En consecuencia, los SR que ignoran la estructura de la red social de los usuarios muestran un poder de predicción más limitado.

Por lo anterior, el propósito de esta investigación es evaluar el desempeño del algoritmo de *Matrix Factorization* para la recomendación de películas, teniendo en cuenta modelos de redes sociales inferidas y las medidas topológicas de estas redes.



# Descripción del Problema

---

## 1.1. Planteamiento del Problema

Actualmente la información en internet crece exponencialmente. Diversas páginas de ventas o servicios contienen diferente material para mostrar a una persona que ingrese a sus sitios web. Los sistemas de recomendación le facilitan a estas empresas a limitar el foco de sugerencias de productos a solo la parte más relevante y de mayor impacto. Estos sistemas son algoritmos que generan recomendaciones a partir de los meta-datos del producto (o del usuario), para ayudar a las personas en procesos de toma de decisiones.

En 2012 Gomez-Rodriguez *et al.* [5] propusieron que detrás de los conjuntos de datos existen enlaces subyacentes desconocidos, sobre los cuales se propaga la información. Al propagarse sobre los enlaces se genera una *cascada* y, gracias a estas, junto con otros datos, se puede inferir redes sociales. El algoritmo desarrollado por Gomez-Rodriguez *et al.* se llama *NETINF*.

Fan *et al.* extendieron *NETINF* al fusionar el trabajo de Gomez-Rodriguez *et al.* junto con datos de calificación de películas, para evaluar el algoritmo de inferencia en función de las recomendaciones [6]. Los autores sugirieron evaluar el algoritmo de Gomez-Rodriguez *et al.* en un conjunto de datos sin red social, como *MovieLens*, para comparar la precisión de las recomendaciones.

En 2017 Kose *et al.* compararon los algoritmos de *User-Based Collaborative Filtering*, *Matrix Factorization* y *Yehuda Koren's Integrated Model*, para predecir las calificaciones que usuarios darían a películas [3]. Esta investigación dio como resultado un desempeño similar para los tres algoritmos, con respecto a su raíz del error cuadrático medio, o *RMSE* (por sus siglas en inglés). Cabe resaltar que se destaca el algoritmo de *Matrix Factorization* gracias a sus buenos resultados y a su sencilla implementación.

En general, los sistemas de recomendación actuales no toman en cuenta las estructuras internas que existen detrás de los conjuntos de datos, las cuales permiten saber cómo se encuentran conectados los usuarios. En consecuencia, surge la pregunta ¿De qué manera se ve afectado el desempeño de los algoritmos de recomendación, haciendo uso de redes sociales inferidas y de las medidas topológicas de estas redes?

### 1.1.1. Formulación

¿Cómo se ve afectado el desempeño del algoritmo *Matrix Factorization* para la recomendación de películas, al añadir medidas topológicas de redes sociales inferidas?

### 1.1.2. Sistematización

Se busca responder a las siguientes preguntas:

- ¿Qué algoritmos o métodos existen para inferir redes sociales?
- ¿Qué modelos de redes sociales se pueden inferir a partir de un conjunto de datos en particular?
- ¿Cómo se ve afectada la precisión en las predicciones del algoritmo *Matrix Factorization*, agregando medidas topológicas de redes sociales inferidas?

## 1.2. Objetivos

### 1.2.1. Objetivo General

El objetivo de esta tesis es investigar diferentes métodos para la inferencia de redes sociales (con base en el conjunto de datos *MovieLens* [7]), para escoger y analizar un algoritmo, con el fin de evaluar el beneficio de inferir una red social y utilizar las medidas topológicas de esta red, al realizar recomendaciones utilizando el algoritmo *Matrix Factorization*.

### 1.2.2. Objetivos Específicos

- **O1:** Investigar diferentes métodos o algoritmos para la inferencia de redes sociales.
- **O2:** Explorar los diferentes modelos de redes sociales que se pueden inferir a partir del conjunto de datos de *MovieLens*.
- **O3:** Analizar las métricas de evaluación para el algoritmo *Matrix Factorization*, con el fin de determinar los beneficios de añadir propiedades de redes sociales inferidas.

## 1.3. Justificación

El centro de esta investigación se fundamenta en encontrar diferentes modelos de redes sociales en el conjunto de datos de *MovieLens*, para evaluar el beneficio de inferir una red social a partir de valoraciones, y utilizar las medidas topológicas de esta red en el algoritmo de recomendación *Matrix Factorization*.

La finalidad de este trabajo es permitir a cualquier organización de servicio usar los algoritmos a desarrollar para mejorar sus sistemas de recomendación. Esto se podría reflejar en recomendaciones para estructuras de comunidad de la red social y no necesariamente a un único usuario.

Mejorar los sistemas de recomendación implica una optimización de recursos, una ampliación de la cobertura y la accesibilidad a sus clientes; lo cual se refleja en mayores beneficios financieros para las organizaciones que adopten estos algoritmos.

## 1.4. Delimitaciones y Alcances

Teniendo en cuenta los conocimientos en teoría de redes, algoritmia y aprendizaje automático adquiridos en las carreras de Matemáticas Aplicadas e Ingeniería de Sistemas y Computación se pretende extraer diferentes modelos de redes sociales, del conjunto de datos *MovieLens*, para agregar las propiedades de estas redes al algoritmo de recomendación *Matrix Factorization* y comparar las métricas de evaluación de este algoritmo, para determinar los cambios en su desempeño.

De esta forma se alcanzaría el objetivo de evaluar qué tan útil, es inferir una red social a partir de las valoraciones y utilizar las métricas de esta red para realizar recomendaciones de películas. Se prevé que el algoritmo puede ser usado para recomendar no solo películas, sino servicios y productos en general.

### 1.4.1. Entregables

Archivos ejecutables<sup>1</sup> que contienen el análisis del algoritmo escogido para la inferencia de redes sociales, junto con el cálculo de las medidas topológicas de estas redes, las cuales sirven como insumo para realizar recomendaciones usando el algoritmo *Matrix Factorization*.

---

<sup>1</sup><https://github.com/suribe06/BSc-Thesis>





# Investigación de la Literatura

---

## 2.1. Áreas Temáticas

- Information systems → Information systems applications → Data mining → Collaborative filtering.
- Mathematics of computing → Discrete mathematics → Graph theory.
- Networks → Social Networks → Network Properties.
- Computing methodologies → Machine learning → Machine learning approaches → Factorization methods → Matrix factorization.

## 2.2. Trabajos Relacionados

En la exploración y búsqueda de información para el desarrollo de este proyecto, se revisó la literatura existente a través de bases de datos, antecedentes de estudios, investigaciones o artículos similares, que permitan construir una base sólida para el desarrollo del presente proyecto de investigación.

- Koren *et al.* [8] presentan diferentes técnicas de factorización matricial para sistemas de recomendación, como la descomposición en valores singulares (SVD por sus siglas en inglés), el descenso de gradiente estocástico y mínimos cuadrados alternos. Debido a su primer puesto en *Netflix Prize Competition* (competencia que buscaba el mejor algoritmo de filtrado colaborativo, para predecir las valoraciones de los usuarios sobre películas) en 2007 y 2008, Koren *et al.* indican que estos son los métodos más exitosos y dominantes dentro de los recomendadores de filtrado colaborativo.
- Ma *et al.* [9] basándose en que la red social de un usuario afectará los comportamientos de dicho usuario, presentan un marco de recomendación social, que fusiona una matriz de calificación de elementos de usuario, con la red social del usuario, utilizando factorización de matriz probabilística.
- He *et al.* [10] proponen un sistema de recomendación basado en redes sociales. Este sistema utiliza la información de las redes sociales, incluidas las preferencias de los usuarios, la aceptación general de los artículos y la influencia de los amigos sociales. A partir de esa información desarrollan un modelo probabilístico, que les permite hacer recomendaciones personalizadas.

- Liu *et al.* [11] plantean que las redes sociales de los usuarios y las opiniones de estos se pueden incorporar para mejorar la precisión de la predicción en el ámbito de la recomendación de películas. Proponen entonces un método de recomendación basado en la red local de confianza (*LTN*). Para esto extraen patrones de redes sociales latentes y las múltiples fuentes de opiniones de los usuarios, para generar la red local de confianza de usuarios y poder realizar la predicción basándose en esta red. El algoritmo *LTN* supera significativamente el rendimiento de la recomendación estándar del Filtrado Colaborativo con respecto al *RMSE* y al *MAE*.
- Wang *et al.* [12] proponen ver el problema de recomendaciones como un problema de predicción de enlaces en redes bipartitas. Para dar solución a esto, propusieron el algoritmo *SRNMF*, que toma explícitamente las características latentes de los nodos, junto con su estructura topológica intrínseca y codifica la información geométrica de la red, mediante la construcción de una matriz basada en similitudes. En comparación con otros 17 métodos de predicción de enlaces, el método *SRNMF* es significativamente superior en términos de precisión y estabilidad.
- Li *et al.* [13] plantean que la factorización matricial (FM) estándar no captura las correlaciones estructurales jerárquicas, por ende, proponen una técnica de Factorización Matricial Jerárquica Oculta (*HHMF*), que aprende la estructura jerárquica oculta a partir de los registros de usuario-artículo. Esta técnica no requiere el conocimiento previo de la estructura jerárquica; por lo tanto, puede aplicarse cuando esta información sea explícita o implícita. *HHMF* supera a los métodos tradicionales de FM, a los métodos de FM jerárquica y a los métodos basados en redes neuronales.
- Gasparetti *et al.* [14] revisan la técnica de recomendación social basada en la detección de comunidades. Hacen énfasis en que grupos de usuarios que tienen características sociales más similares, pueden proporcionar evidencia valiosa y adicional para el proceso de recomendación. Sin embargo, debido a que se requieren varios pasos en las técnicas de detección de comunidades en sistemas de recomendación, esto hace que sea un objetivo principalmente teórico y prácticamente desafiante.

## 2.3. Métodos para Inferir Redes Sociales

- Gomez-Rodriguez *et al.* [5] propusieron que detrás de ciertos conjuntos de datos existe alguna red estática desconocida subyacente, sobre la cual se propaga la información. Al propagarse sobre la red se genera una *cascada* y gracias a estas, junto con otros datos, se puede inferir redes sociales. Como resultado, basándose en un modelo probabilístico generador, desarrollaron un algoritmo escalable para la inferencia de redes sociales llamado *NETINF*.
- Fan *et al.* [6] plantearon fusionar el trabajo de Gomez-Rodriguez *et al.* junto con datos de calificación de películas. Propusieron el algoritmo *MOVINF* basándose en [5], para predecir una red social mediante cascadas de calificación de películas. El algoritmo propuesto por Fan

*et al.* utiliza el concepto de  $k$  vecinos mas cercanos en el modelo probabilístico generador usado en *NETINF* para calcular la probabilidad de que una cascada se propague entre  $k$  vecinos mas cercanos. Este algoritmo mejora la complejidad temporal y la precisión del algoritmo *NETINF* bajo ciertas condiciones.

- Alpay *et al.* [15] desarrollan el algoritmo *FASTINF* basado en *NETINF*, para usarlo en un entorno de big data en la industria de grandes redes sociales en linea como Facebook o Twitter. El modelo probabilístico generador usado por Alpay *et al.* es similar al trabajo de [5], pero en lugar de utilizar la estrategia de árbol de propagación mas probable, consideran la cascada como un conjunto de bordes dispersos en un grafo y, por lo tanto, el algoritmo infiere arcos en lugar del árbol. El algoritmo desarrollado obtiene ordenes de complejidad temporal mas rápidos sin sacrificar la precisión.
- Gomez-Rodriguez *et al.* [16] proponen un algoritmo paralelo denominado *NETRATE*, el cual se basa en su algoritmo *NETINF*. La diferencia radica que en [5] la velocidad de transmisión es fija, y no inferida. Este nuevo algoritmo se basa entonces en encontrar la red óptima y las velocidades de transmisión que maximizan la probabilidad de un conjunto observado de cascadas. La principal novedad de este método es modelar la difusión como una red de sucesos temporales continuos e independientes que ocurren a diferentes velocidades.
- Gomez-Rodriguez *et al.* [17] proponen un nuevo algoritmo el cual extiende su algoritmo previo *NETINF*, la principal innovación de este algoritmo es abordar el problema de la inferencia de red considerando todos los árboles posibles admitidos por un grupo de cascadas. Los autores presentan un algoritmo de inferencia de red que puede ser capaz de inferir redes del orden de cientos de miles de nodos con una pequeña cantidad de cascadas observadas. Este algoritmo no infiere probabilidades previas de infección ni tasas de transmisión, sino solo la conectividad de la red.
- Gray *et al.* [18] utilizan el modelo de cascada independiente (aunque se puede extender a cualquier modelo en cascada) para desarrollar un método de inferencia bayesiano basado en la Cadena de Markov Monte Carlo, teniendo en cuenta las incertidumbres inherentes que surgen en la inferencia y las observaciones de datos. Los autores utilizan el modelo de inferencia bayesiano junto al “Teorema del Árbol de Matrices” de Kirchoff, extendido a árboles dirigidos por Tutte [19], en el proceso de la inferencia de la red. Esta técnica es probada con un pequeño número de cascadas simuladas para las cuales *NETINF* no produce un resultado.



# Marco Teórico

---

Para el desarrollo de la presente propuesta, se deben tener en cuenta conocimientos tanto de ciencias de la computación como de matemáticas. A continuación se presentan las bases teóricas sobre las que se trabajará en el desarrollo del proyecto.

## 3.1. Factorización de Matrices

Sea  $U = \{u_1, u_2, \dots, u_m\}$  un conjunto de usuarios,  $I = \{i_1, i_2, \dots, i_n\}$  un conjunto de ítems (en este caso películas) y sea  $T$  un conjunto finito dado:

$$T \subset U \times I \times \mathbb{R}_{\geq 0}$$

Cada tripleta  $(u, i, r) \in T$ , con  $u \in U$ ,  $i \in I$  y  $0 \leq r \leq 5$ ; significa que el usuario  $u$  calificó el ítem  $i$  con el valor  $r$ . Note que  $|T| < m \cdot n$  debido a que la mayoría de usuarios sólo valoran algunos ítems. Ahora, sea  $R$  una matriz  $m \times n$  que representa a  $T$ , donde cada entrada  $r_{u,i}$  ( $1 \leq u \leq m$  y  $1 \leq i \leq n$ ) tiene el valor de la calificación  $r$  del usuario  $u$  sobre el ítem  $i$ .

Dado un usuario  $u \in U$  y un ítem  $i \in I$  de quien se desconoce  $r_{u,i}$ , el objetivo es predecir la valoración de  $u$  sobre el ítem  $i$ . Para esto, se emplean técnicas de factorización matricial, para aprender las características latentes de los usuarios y los ítems, a fin de predecir las valoraciones desconocidas, utilizando estas características latentes [20]. Como presenta Cortes [21], se pueden realizar recomendaciones utilizando modelos de factorización matricial colectiva (CMF por sus siglas en inglés), los cuales toman en cuenta información adicional sobre los usuarios y/o ítems.

Para poder entender la factorización matricial colectiva es importante entender cómo funcionan los modelos de factorización matricial de bajo rango. Estos intentan factorizar la matriz  $R$ , en el producto de dos matrices de menor dimensión  $\mathbf{P} \in \mathbb{R}^{m \times k}$  y  $\mathbf{Q} \in \mathbb{R}^{k \times n}$ , donde  $k$  corresponde a ciertas características latentes. Es decir:

$$\mathbf{R} \approx \mathbf{P}\mathbf{Q} = \hat{\mathbf{R}}$$

La matriz  $\mathbf{P}$  representa la asociación entre un usuario y las características, mientras que la matriz  $\mathbf{Q}$  representa la asociación entre un ítem y las características.

Se puede obtener la predicción  $\hat{r}_{u,i}$  de la valoración de un ítem  $i$  por parte de un usuario  $u$  mediante el cálculo del producto punto de los dos vectores correspondientes a  $u$  e  $i$ .

$$\hat{r}_{u,i} = p_u q_i = \sum_{d=1}^k p_{ud} \cdot q_{di}$$

Si esto se hace para todos los usuarios e ítems, la matriz de valoración completa usuario-ítem se convierte en:

$$\begin{bmatrix} p_1 \\ \vdots \\ p_m \end{bmatrix} \begin{bmatrix} q_1 & \cdots & q_n \end{bmatrix} = \begin{bmatrix} \hat{r}_{1,1} & \cdots & \hat{r}_{1,n} \\ \vdots & \ddots & \vdots \\ \hat{r}_{m,1} & \cdots & \hat{r}_{m,n} \end{bmatrix} = \hat{\mathbf{R}}$$

Para aprender a reconocer los vectores de características ( $p_u$  y  $q_i$ ), el sistema minimiza la suma del error cuadrático medio ( $RMSE$ , por sus siglas en inglés), en el conjunto de valoraciones conocidas:

$$SSE = \sum_{(u,i,r) \in T} (r_{u,i} - \hat{r}_{u,i})^2 = \sum_{(u,i,r) \in T} \left( r_{u,i} - \sum_{d=1}^k p_{ud} \cdot q_{di} \right)^2$$

$$RMSE = \sqrt{SSE/|T|}$$

$$(\mathbf{P}^*, \mathbf{Q}^*) = \arg \min_{(\mathbf{P}, \mathbf{Q})} RMSE^1 \quad (1)$$

Koren *et al.* [8] mostraron que estos modelos son muy propensos al sobreajuste si se estiman tal cual, por lo que a menudo se estiman con regularización. Dos enfoques para resolver la ecuación (1) son el descenso de gradiente estocástico y los mínimos cuadrados alternos.

Ahora bien, la CMF factoriza conjuntamente la matriz  $R$  junto con la matriz de atributos de los usuarios y la matriz de atributos de los ítems (en este caso solo se hará uso de la matriz de atributos de usuarios). Sea entonces  $\mathbf{A} \in \mathbb{R}^{m \times s}$  la matriz de  $s$  atributos de los usuarios y la matriz  $\mathbf{C} \in \mathbb{R}^{k \times s}$ , entonces la ecuación (1) se convierte en:

$$SSE = \sum_{(u,i,r) \in T} \left( r_{u,i} - \sum_{d=1}^k p_{ud} \cdot q_{di} \right)^2 + \sum_{0 < u \leq m, 0 < l \leq s} \left( a_{u,l} - \sum_{d=1}^k p_{ud} \cdot c_{dl} \right)^2$$

$$RMSE = \sqrt{SSE/|T|}$$

$$(\mathbf{P}^*, \mathbf{Q}^*, \mathbf{C}^*) = \arg \min_{(\mathbf{P}, \mathbf{Q}, \mathbf{C})} RMSE \quad (2)$$

---

<sup>1</sup>Las matrices  $\mathbf{P}$ ,  $\mathbf{Q}$  para las que el  $RMSE$  alcanza su valor mínimo.

### 3.2. Grafo

Un grafo  $G$  es una pareja ordenada  $G = (V, E)$ , donde  $V$  es un conjunto finito de nodos o vértices y  $E$  es un conjunto finito de aristas, que relacionan los nodos. Un grafo  $G$  con  $|V| = n$  puede ser representado mediante su matriz de adyacencia  $A$ , que es una matriz cuadrada  $n \times n$ , donde  $a_{ij} = 1$  si hay una arista del vértice  $u_i$  al vértice  $u_j$ , y cero de lo contrario.

### 3.3. Subgrafo Inducido

Sea  $G = (V, E)$  un grafo y  $S \subset V$  un subconjunto de nodos de  $G$ . El subgrafo inducido en  $G$  por  $S$  (denotado  $G[S]$ ), es aquel que se obtiene tomando los vértices de  $S$  y las aristas de  $G$  que son incidentes a todos los vértices en  $S$ .

### 3.4. Grafo Bipartito

Un grafo bipartito se define como  $B = (U, V, E)$  donde  $U, V$  son conjuntos de nodos y  $U \cap V = \emptyset$ , es decir, son conjuntos disjuntos. De manera que cada arista  $e \in E$  conecta vértices del conjunto  $U$  con vértices del conjunto  $V$ .

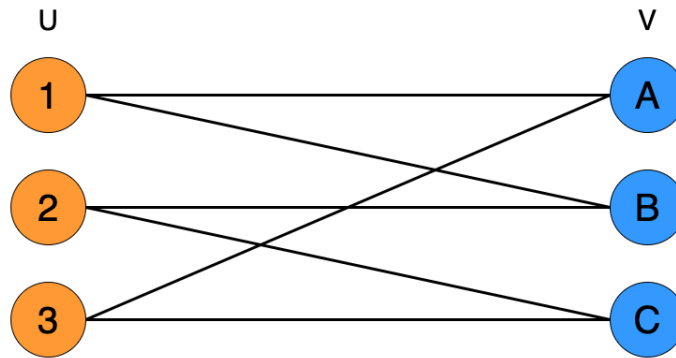


Figura 3.1: Ejemplo de un Grafo Bipartito

La matriz de adyacencia  $A$  de un grafo bipartito  $B$ , dándole un orden adecuado a los nodos (en este caso primero  $U$ , luego  $V$ ), tiene la forma de una matriz por bloques  $2 \times 2$ ,

$$A = \begin{pmatrix} 0 & C \\ C^T & 0 \end{pmatrix}$$

en la que los bloques de la diagonal esta formada por matrices cero (ya que no hay aristas entre nodos del mismo conjunto). El bloque inferior izquierdo es la transpuesta  $C^T$  de la matriz  $C$  de entradas del bloque superior derecho.  $C$  se conoce como la matriz de biadyacencia y  $c_{i,j} = 1$  si y solo si  $(u_i, v_j) \in E$ .

### 3.4.1. Proyección de un Grafo Bipartito

La proyección de un grafo es un método utilizado para comprimir información sobre grafos bipartitos. Esto significa que el grafo resultante contiene nodos de solo cualquiera de los dos conjuntos  $U$  o  $V$ , en donde dos nodos están conectados si tienen al menos un vecino en común del otro conjunto de nodos. Dado que  $A = A^T$ , se tiene que:

$$A^T \cdot A = A \cdot A = \begin{pmatrix} C \cdot C^T & 0 \\ 0 & C^T \cdot C \end{pmatrix}$$

Donde  $C \cdot C^T$  es la matriz de adyacencia de la proyección sobre el conjunto de nodos  $U$ , y  $C^T \cdot C$  es la matriz de adyacencia de la proyección sobre el conjunto de nodos  $V$ . Uno de los muchos métodos para proyectar un grafo bipartito se conoce como **ponderación simple**, el cual pondera las aristas directamente con el número de veces que se repite la asociación entre los nodos en cuestión. En la Figura 3.2 se puede apreciar las diferentes proyecciones de un grafo bipartito con el método de ponderación simple.

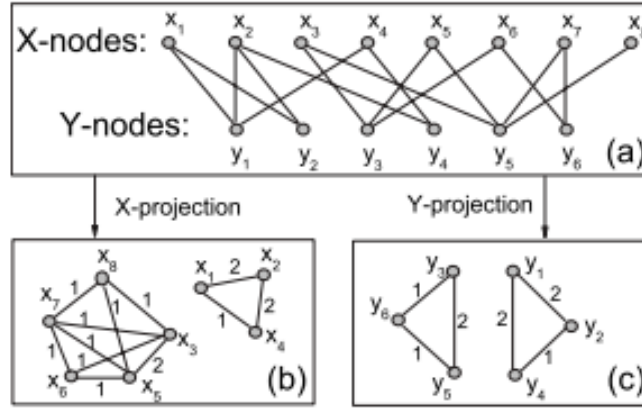


Figura 3.2: Ilustración de un grafo bipartito (a), su proyección con respecto a  $X$  (b) y su proyección con respecto a  $Y$  (c) [22]

## 3.5. Redes Sociales

Una red  $N$  puede definirse como  $N = (V, E, f_V, f_E)$ , que contiene un grafo  $G = (V, E)$ , extendido con una función  $f_V$  que especifica propiedades de los vértices ( $f_V : V \rightarrow X$ ) y una función  $f_E$  que especifica propiedades de las aristas ( $f_E : E \rightarrow Y$ ) [23].

## 3.6. Medidas Topológicas de Redes Sociales

Nacher [24] señala que las medidas topológicas de una red se refieren a una variedad de medidas matemáticas, que utilizan la matriz de adyacencia de un grafo  $G$  para capturar propiedades



específicas de la topología de la red. Algunas de estas medidas son:

- **Grado de un nodo:** Sea un nodo  $u \in V$ , el grado de  $u$  se denota por  $\deg(u) = |\{v \in V : \{u, v\} \in E\}|$ , e indica el número de aristas conectadas al nodo  $u$ . Nodos muy conectados se denominan *hubs*.
- **Distancia:** La distancia  $d(u, v)$  entre dos nodos  $u, v \in V$  se define como el número más corto de aristas a recorrer, desde el nodo  $u$  al nodo  $v$ . Si los nodos no son accesibles entonces  $d(u, v) = \infty$ .
- **Excentricidad:** La excentricidad de un nodo es la distancia de su camino más corto desde el otro nodo más lejano en el grafo. Se denota por  $e(u) = \max\{d(u, v) : v \in V\}$
- **Coeficiente de agrupamiento:** Mide la proporción del número de aristas entre los vecinos del nodo  $u$  y el número máximo de aristas que podría existir entre los vecinos del nodo  $u$ . Sirve para cuantificar qué tanto está agrupado un nodo  $u$  con sus vecinos.
- **Medidas de centralidad:** Sirven para calcular la importancia de un nodo  $u$  en la red.
  - Centralidad de Cercanía (o *Closeness Centrality* por su nombre en inglés): Puntúa un nodo  $u$  en función de su cercanía a todos los demás nodos de la red. Está dada por:

$$CC(u) = \frac{n}{\sum_{v \in V} d(u, v)}$$

Donde  $n$  es el número de nodos del grafo.

- Centralidad de Intermediación (o *Betweenness Centrality* por su nombre en inglés): Cuantifica el número de veces que un nodo se encuentra en el camino más corto entre otros dos nodos. Está dada por:

$$BC(u) = \sum_{u \neq v \neq s \in V} \frac{\sigma_{sv}(u)}{\sigma_{sv}}$$

Donde  $\sigma_{sv}$  es el número total de caminos más cortos, desde el nodo  $s$  al nodo  $v$  y  $\sigma_{sv}(u)$  es el número de esos caminos que pasan por  $u$ .

- Centralidad del Vector Propio (o *Eigenvector Centrality* por su nombre en inglés): Esta propiedad mide la influencia de un nodo en una red. Puntúa un nodo  $u$  en función de la centralidad de sus vecinos. Está dada por:

$$EC(u) = \frac{1}{\lambda} \sum_{v \in N(u)} EC(v) = \frac{1}{\lambda} \sum_{v \in V} a_{u,v} \cdot EC(v)$$

Donde  $\lambda$  es una constante,  $N(u)$  es el conjunto de los vecinos de  $u$  y  $a_{u,v}$  es una entrada de la matriz de adyacencia  $A$  del grafo  $G$ .

- Centralidad de Lejanía (o *Farness Centrality* por su nombre en inglés): Esta medida es recíproca a la centralidad de cercanía (de modo que si la cercanía es pequeña, la lejanía es grande y viceversa), es decir, es la suma de la distancia de un nodo a todos los demás nodos del grafo.

Muchas otras métricas que se podrían usar en este proyecto se encuentran definidas en [24].

De igual forma se piensa hacer uso de estructuras de comunidad, que nos permiten revelar relaciones ocultas entre los nodos de la red. Fortunato [25] define las comunidades como grupos de nodos, que probablemente comparten propiedades comunes y/o desempeñan funciones similares dentro del grafo.

### 3.7. Inferencia de Redes Sociales (*NETINF*)

Gomez-Rodriguez *et al.* presentan cómo inferir una red basándose en un modelo probabilístico generador y heurísticas. Definen una cascada  $c$  como una pieza diferente de información que se propaga sobre un grafo  $G$ . A continuación, se presenta el modelo de transmisión de cascadas descrito en [5].

**Definición 3.7.1** (Modelo de Transmisión de Cascadas). Para un conjunto de cascadas  $C$  que se propagan sobre un grafo  $G$ , se observa una pareja  $(v, t_v)_c$ , que describe el momento en que un nodo  $v$  se infecto con una cascada  $c$  en el tiempo  $t_v$  (si  $t_v = \infty$  significa que el nodo  $v$  no es alcanzado por la cascada  $c$ ), pero no se sabe quién los infecto, las infecciones son binarias, es decir, un nodo está infectado o no lo está. Gomez-Rodriguez *et al.* utilizan el modelo de cascada independiente propuesto por Tardos *et al.* [26], que indica que un nodo infectado infecta a cada uno de sus vecinos en el grafo  $G$  de forma independiente y aleatoria con alguna pequeña probabilidad. El modelo asume implícitamente que cada nodo  $v$  en la cascada  $c$  es infectado como máximo por un nodo  $u$ , es decir, no toma en cuenta las infecciones que vienen después. Así, la estructura de una cascada está descrita completamente por un árbol  $T$ , contenido en el grafo  $G$ . El árbol de propagación  $T$  especifica qué nodos infectaron a otros nodos.

Sea  $\hat{G}$  un grafo oculto sobre el que se extienden múltiples cascadas. Para poder inferir este grafo se utiliza el Modelo de Transmisión de Cascadas. Primero se define el modelo de transmisión de cascada por pares  $P_c(u, v)$ , que describe la probabilidad de que el nodo  $u$  propague la cascada  $c$ , al nodo  $v$ . El algoritmo *NETINF* consta de 3 modelos distintos para la probabilidad de transmisión por pares entre nodos; estos modelos se presentan en la Tabla (3.1), donde  $\alpha$  corresponde a la tasa de transmisión. Luego se describe la probabilidad  $P(c|T)$  que la cascada  $c$  se propague en un determinado árbol  $T = (V_T, E_T)$ . Asumiendo independencia entre aristas, la cascada  $c$  se propaga con probabilidad  $\beta$  y se detiene con probabilidad  $1 - \beta$  sobre las aristas del árbol  $T$ , por lo tanto:

$$P(c|T) = \prod_{(u,v) \in E_T} \beta P_c(u, v) \prod_{u \in V_T, (u,x) \in E \setminus E_T} (1 - \beta) \quad (3)$$

La ecuación (3) puede reescribirse como:

$$P(c|T) = \beta^q (1 - \beta)^r \prod_{(u,v) \in E_T} P_c(u, v) \quad (4)$$

Donde  $q = |E_T| = |V_T| - 1$  es el número de aristas en  $T$  y cuenta las aristas sobre las cuales la cascada  $c$  se ha propagado con éxito. Del mismo modo,  $r$  es el número de aristas que no transmitieron la cascada:  $r = \sum_{u \in V_T} d_{out}(u) - q$ , y  $d_{out}(u)$  es el grado de salida del nodo  $u$  en el grafo  $G$ .

Ahora, se define  $P(c|G)$ , como la probabilidad que la cascada  $c$  ocurra en un grafo  $G$ . Esta se puede calcular basándose en  $P(c|T)$ . Se buscan todas las posibles estructuras de arboles  $T$  en  $\mathcal{T}_c(G)$  en donde la cascada  $c$  puede propagarse. Aquí,  $\mathcal{T}_c(G)$  es el conjunto de todos los posibles árboles de propagación en un subgrafo de  $G$  inducido por los nodos afectados por la cascada  $c$ . La ecuación (5) calcula la probabilidad que una sola cascada  $c$  ocurra en un grafo  $G$ .

$$P(c|G) = \sum_{T \in \mathcal{T}_c(G)} P(c|T) P(T|G) \quad (5)$$

Note que aunque la suma se extiende sobre todos los posibles árboles de propagación  $T \in \mathcal{T}_c(G)$ , en caso de que  $T$  sea inconsistente con los datos observados, entonces  $P(c|T) = 0$ . Usando la ecuación (4), la ecuación (5) puede reescribirse como:

$$P(c|G) = \sum_{T \in \mathcal{T}_c(G)} \prod_{(u,v) \in E_T} P_c(u, v) \quad (6)$$

Por último, la probabilidad que un conjunto de cascadas  $C$  ocurra en un grafo  $G$  se define como:

$$P(C|G) = \prod_{c \in C} P(c|G) \quad (7)$$

La ecuación (7) describe el Modelo de Transmisión de Cascadas. Bajo este modelo, se busca estimar un grafo  $\hat{G}$  que (aproximadamente) maximice la probabilidad que  $C$  ocurra en él, es decir, encontrar  $\hat{G}$  que resuelva el siguiente problema de optimización.

$$\hat{G} = \arg \max_{|G| \leq e} P(C|G) \quad (8)$$

Note que la maximización es sobre todos los grafos dirigidos  $G$  de a lo sumo  $e$  aristas, y  $P(C|G)$  se define mediante la ecuación (7).

Este problema de optimización parece intratable. Para evaluar la ecuación (7), hay que calcular la ecuación (5) para cada cascada  $c \in C$  en todos los posibles arboles  $T \in \mathcal{T}_c(G)$  y el número de árboles puede ser exponencial, con respecto al tamaño de  $G$ . Por esta razón, los autores introducen dos conceptos para que la ecuación (8) sea tratable. El primer concepto es haciendo uso únicamente del árbol de propagación más probable por cada cascada. El segundo concepto lo definen como

“ $\varepsilon$ -bordes”, el cual hace referencia a que los nodos pueden infectarse por razones diferentes a la influencia de la red. Los detalles técnicos y las demostraciones se encuentran en [5].

Modelo	Probabilidad de Transmisión
Exponencial	$\begin{cases} \alpha e^{-\alpha(t_i - t_j)} & \text{Si } t_j > t_i \\ 0 & \text{en otro caso} \end{cases}$
Ley de Potencias	$\begin{cases} (\alpha - 1)(t_i - t_j)^{-\alpha} & \text{Si } t_j + 1 < t_i \\ 0 & \text{en otro caso} \end{cases}$
Rayleigh	$\begin{cases} \alpha(t_i - t_j)e^{-\frac{1}{2}\alpha(t_i - t_j)^2} & \text{Si } t_j > t_i \\ 0 & \text{en otro caso} \end{cases}$

Tabla 3.1: Modelos de Probabilidad de Transmisión por Pares



# Implementación

## 4.1. Recomendación de Películas

Para entender mejor como se afectan las recomendaciones al hacer uso de redes sociales inferidas y de las propiedades topológicas de estas redes, es necesario hacer un análisis previo de la variable objetivo (el rating de una película) y así obtener las métricas del algoritmo *Matrix Factorization*. Por lo anterior, se presenta en la Figura 4.1 y la Tabla 4.1 el análisis realizado para la variable objetivo.

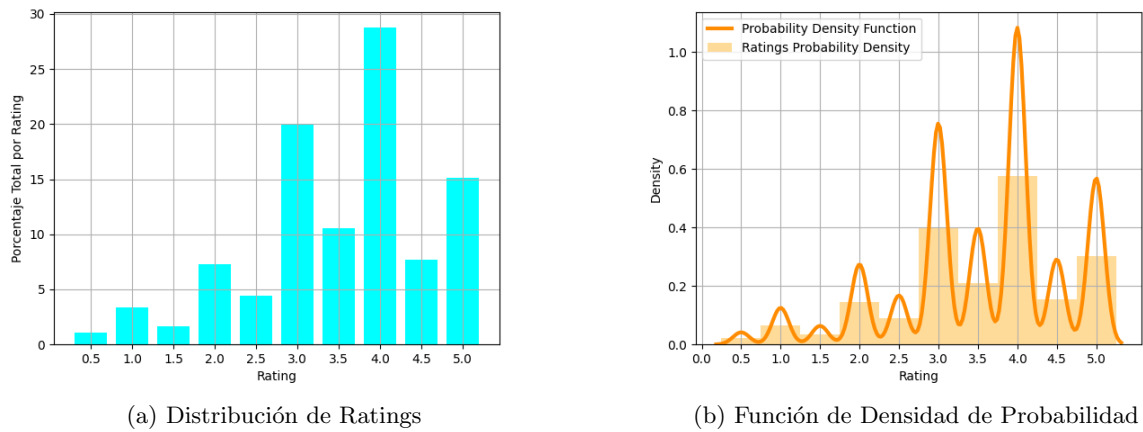


Figura 4.1: Análisis de la Variable Objetivo

Media:	3.543608
Desviación Estándar:	1.058064
Mínimo:	0,5
25 %:	3,0
50 %:	4,0
75 %:	4,0
Máximo:	5,0

Tabla 4.1: Medidas de Tendencia Central Variable Objetivo

Una vez realizado el estudio de la variable objetivo, se utiliza la librería `cmfrec` [21] para realizar las recomendaciones, haciendo uso del algoritmo *Matrix Factorization*. En primer lugar, se utiliza la técnica *Grid Search* para calcular los valores óptimos de los hiperparámetros del modelo, con lo que se obtiene  $k = 20$  y  $\lambda = 10$ , los cuales hacen referencia a los factores latentes para la factorización matricial y el parámetro de regularización respectivamente. Posteriormente se realiza la recomendación de películas sin tener en cuenta propiedades de grafo, después de 20 iteraciones los resultados obtenidos se muestran en la Tabla 4.2.

Métrica	Promedio	Desv. Estándar
$RMSE$	0,8983	0,1867
$MAE$	0,7126	0,0427
$R^2$	0,1964	0,1715

Tabla 4.2: Métricas Obtenidas para la Recomendación

## 4.2. Datos en Representación de Grafos

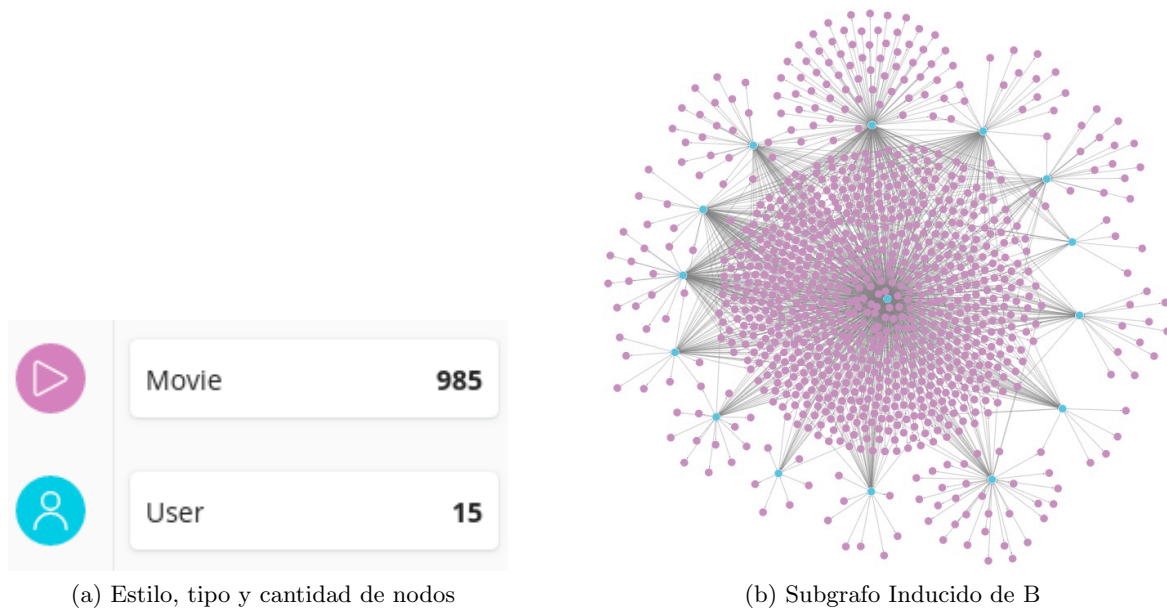
### 4.2.1. Grafo Bipartito

El conjunto de datos de *MovieLens* se puede ver como un grafo bipartito  $B = (U, M, R)$  donde  $U$  es un conjunto de usuarios,  $M$  un conjunto de películas y cada arista  $e \in R$  representa la valoración de un usuario a una película.

Conjunto	Cardinalidad
$U$	671
$M$	9.066
$R$	100.004

Tabla 4.3: Datos del Grafo Bipartito  $B$

En la figura 4.2 se puede observar un subgrafo inducido de  $B$  visto desde la interfaz Bloom de Neo4j.

Figura 4.2: Subgrafo inducido de  $B$  visto desde Neo4j

#### 4.2.2. Proyección del Grafo con Respecto a Usuarios

Previo a la inferencia de la red con el algoritmo de *NETINF* es importante realizar un primer acercamiento a un posible modelo de red social entre usuarios. Para esto, se hará uso de la proyección del grafo  $B$  con respecto al conjunto  $U$ . Con ayuda de los algoritmos de la librería *networkx* de *Python* se realizó la proyección de  $B$  obteniendo lo siguiente:

Numero de Nodos:	671
Numero de Aristas:	197.780

Tabla 4.4: Datos del Grafo Bipartito  $B$  Proyectado en  $U$ 

Con la ayuda del software Gephi se pudo graficar un subgrafo inducido del grafo  $B$  proyectado con respecto a los usuarios, el cual consta de 307 nodos y 887 aristas. Esto se puede apreciar en la Figura 4.3.



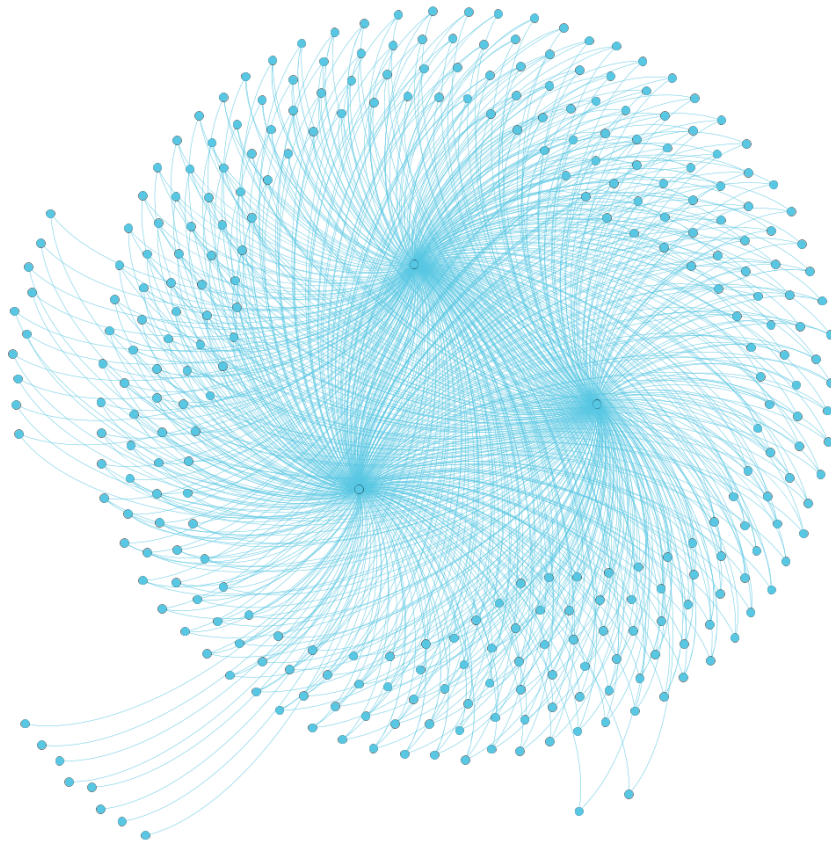


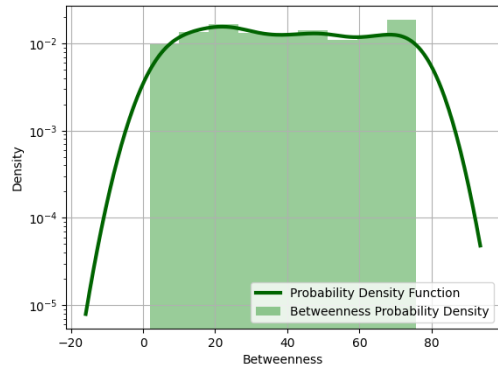
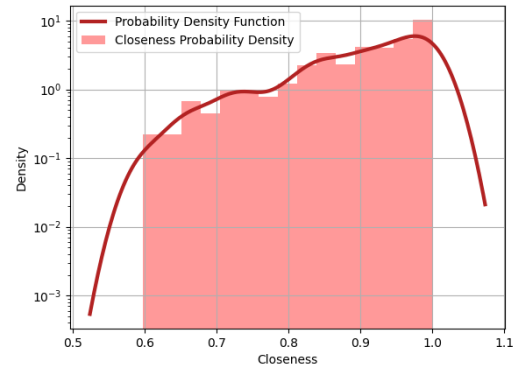
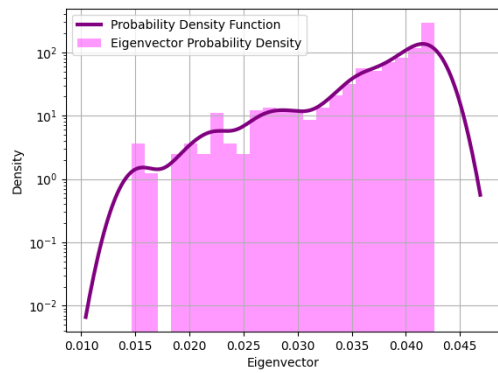
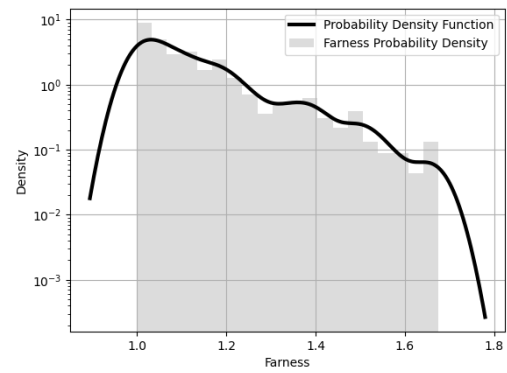
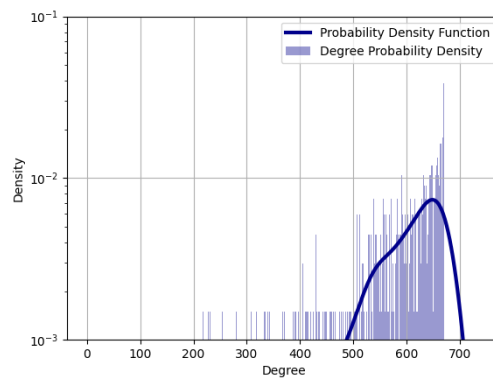
Figura 4.3: Subgrafo inducido de  $B$  proyectado en  $U$

Para poder comprender la proyección realizada se obtuvieron las siguientes medidas topológicas:

Densidad:	87,9862 %
Diámetro:	2
Coef. de Agrupamiento Promedio:	0,919756

Tabla 4.5: Datos del Grafo Bipartito  $B$  Proyectado en  $U$

Y las siguientes medidas de centralidad del grafo:

(a) *Betweenness Centrality*(b) *Closeness Centrality*(c) *Eigenvector Centrality*(d) *Farness Centrality*(e) *Degree*Figura 4.4: Medidas de Centralidad para el grafo  $B$  proyectado en  $U$

Con la Centralidad de Intermediación (Figura 4.4a) se puede calcular la frecuencia en que un nodo actúa como “puente” dentro de una ruta corta entre dos nodos determinados. Los nodos que poseen alta intermediación pueden interpretarse como nodos reguladores de flujos de información dentro de la red. El *Betweenness* promedio del grafo proyectado en usuarios corresponde a 40,2459, dando a entender que los nodos tienen más control sobre la red, ya que más valoraciones de películas, pasarán a través de ellos a los demás nodos.

Gracias a la Centralidad de Cercanía se pueden clasificar los nodos con respecto a las distancias más cortas a los demás nodos. Lo anterior puede interpretarse como la rapidez en que se propaga la información, en este caso la calificación de una película, desde un nodo (o usuario) al resto. En la Figura 4.4b se puede observar que para los usuarios su centralidad de cercanía varía entre 0.6 y 1.0 con un promedio de 0.9029, lo que indica que las valoraciones de las películas se difunden rápidamente a través de los nodos de la red.

En lo que respecta la Centralidad del Vector Propio, con esta se puede medir la influencia de un nodo dentro de la red. En la Figura 4.4c una puntuación alta de vector propio significa que un nodo está conectado a muchos nodos que, a su vez, tienen puntuaciones altas; siendo entonces capaz de difundir la información (valoraciones de películas) a los demás nodos de forma fluida. Para esta medida se tiene un promedio de 0.0382.

Por la Centralidad de Lejanía (Figura 4.4d) se puede determinar si un nodo tiene relaciones estrechas con los demás nodos, lo que se puede interpretar en vértices conectados al grafo, pero que están alejados de la mayoría de los demás vértices.

En la Distribución de Grados expuesta en la Figura 4.4e se puede distinguir que la mayoría de grados varían entre 500 y 671, generando así un grado promedio de 589,5 lo que indica que una gran parte de los nodos del grafo proyectado son *hubs*, esto también se puede percibir con respecto a la densidad del grafo que equivale a un 87,98 % y en el Coeficiente de Agrupamiento Promedio que es de 0,9197.

Debido a que el grafo proyectado obtenido es muy denso y con un grado promedio elevado (en comparación al número total de nodos), esto da a entender que un usuario al azar, tendría en promedio en su red social cerca al 88 % de los demás usuarios, lo que en la vida real es muy poco probable. Por esta razón lo que se busca obtener es un grafo disperso, que modele de manera más acertada la realidad, pues una persona no es amiga de todo el mundo, sino sólo de algunas personas.

### 4.3. Inferencia de Redes

Gracias al conjunto de datos de *MovieLens* se puede saber cuando un usuario califica una película, pero no se sabe quien o que influye sobre el usuario para calificar dicha película. Es así como el algoritmo de *NETINF* toma gran importancia en este ámbito. El problema consiste entonces en inferir la red social subyacente de usuarios sobre la que se propagan las cascadas, en este sentido, se utiliza el análisis de cascadas de películas propuesto por Fan *et al.* [6] donde todas las valoraciones de una película se representarán como una cascada, para un total de 9066 cascadas. Cada cascada  $c$  contiene todas las valoraciones  $r_i$  de la película asociada  $m_c$ , dicha valoración puede representarse como  $(u_i, t_i)_c$ , lo que significa que el usuario  $u_i$  valora la película  $m_c$  en el momento  $t_i$ .

Debido a los diferentes modelos de probabilidad con los que *NETINF* infiere la transmisión por pares (Tabla 3.1), en la Tabla 4.6 se muestra los modelos usados y los resultados obtenidos.

Modelo	$\alpha$	Arcos Inferidos	Tiempo de Ejecución
Exponencial	$1 \cdot 10^{-4}$	3866	4.27 seg
	$3 \cdot 10^{-5}$	7369	7.68 seg
	$1 \cdot 10^{-5}$	11368	12.77 seg
	$3 \cdot 10^{-6}$	15496	20.76 seg
	$1 \cdot 10^{-7}$	19762	41.2 seg
Ley de Potencias	7.5	19156	43.59 seg
	8.5	15215	21.58 seg
	9	12042	14.59 seg
	9.5	8859	9.65 seg
	11	2726	3.48 seg
Rayleigh	$1 \cdot 10^{-10}$	4349	4.78 seg
	$2 \cdot 10^{-11}$	6776	6.97 seg
	$3 \cdot 10^{-12}$	10144	10.95 seg
	$1 \cdot 10^{-12}$	12323	14.76 seg
	$1 \cdot 10^{-14}$	19460	34.31 seg

Tabla 4.6: Resultados de la Inferencia de Redes con *NETINF*

#### 4.4. Recomendación de Películas Usando Redes Sociales Inferidas y sus Propiedades

Una vez obtenidos los diferentes grafos para los respectivos modelos (Tabla 4.6), se utilizó la librería SNAP [27] para obtener las propiedades del grafo en cuestión. Seguidamente, se generaron conjuntos de datos con las propiedades obtenidas de dichos grafos, para así, agregarlas al modelo de recomendación. Hecho esto, se realizó la recomendación, esta vez tomando en cuenta las propiedades del grafo. Los resultados obtenidos después de 12 iteraciones en cada modelo se muestran en la Figura 4.5.

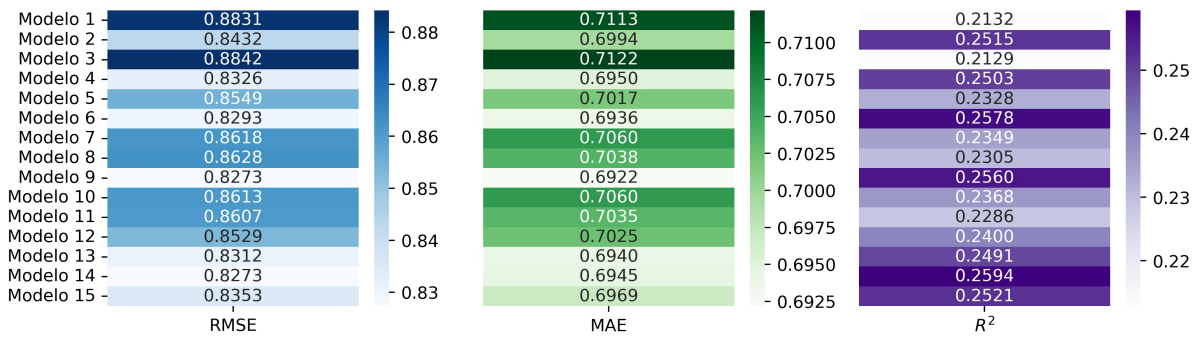
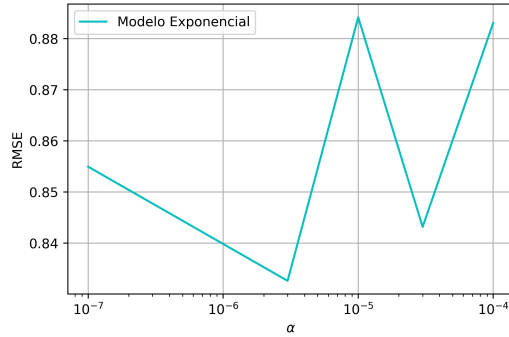
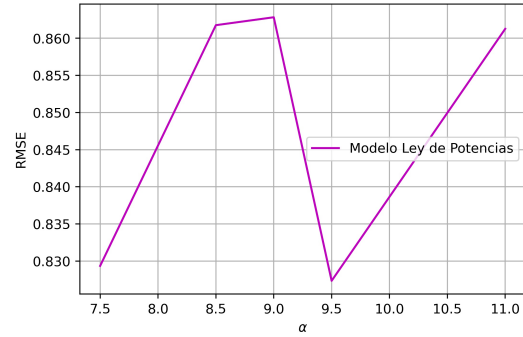
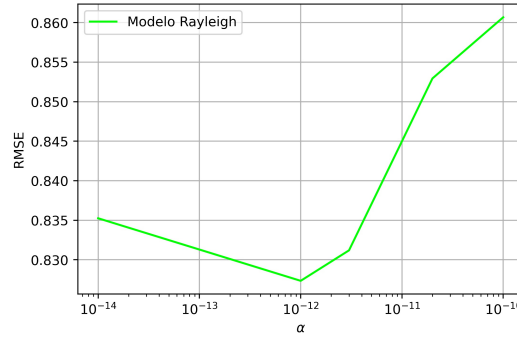


Figura 4.5: Resultados de la Recomendación con Propiedades de Grafo

Para analizar de mejor manera como se ve afectada la recomendación de películas, teniendo en cuenta propiedades de grafo, se toma como referencia la métrica del error cuadrático medio ( $RMSE$  por sus siglas en inglés). Note que para los 15 modelos se presenta una mejora en el  $RMSE$  con respecto a la recomendación sin propiedades de grafo, la cual corresponde a 0,8983. Así mismo, es importante destacar la mejora más notoria, obtenida en el Modelo 9 (Ley de Potencias y  $\alpha = 9,5$ ) con un  $RMSE$  de 0.8273.

Debido a que el algoritmo *NETINF* posee un único parámetro variable  $\alpha$  (tasa de transmisión) en común para los 3 modelos (Tabla 3.1), el cual afecta la probabilidad de transmisión por pares entre nodos; en la Figura 4.6 se presenta como influye este parámetro en el  $RMSE$  de los diferentes modelos.

(a)  $RMSE$  para el Modelo Exponencial(b)  $RMSE$  para el Modelo Ley de Potencias(c)  $RMSE$  para el Modelo RayleighFigura 4.6:  $RMSE$  vs.  $\alpha$  para los Diferentes Modelos de Redes Inferidas

## 4.5. Observaciones Teóricas sobre los Algoritmos Usados

El algoritmo *NETINF* recoge los modelos de difusión de información introducidos por Tardos *et al.* [26], donde estudiaron el problema de encontrar un conjunto de nodos influyentes iniciales, que puedan maximizar el número de nodos que serán influenciados en una red. Gomez-Rodriguez *et al.* desarrollaron un algoritmo voraz para aproximar el problema NP-Hard resultante. Basándose en [26], primero formalizan el problema en un marco de optimización, como se aprecia en la ecuación (8), y demuestran que es NP-Hard por reducción al problema de cobertura máxima (*MAX-k-COVER*). Luego muestran que el problema satisface submodularidad, para que se pueda encontrar una solución casi óptima (de al menos una fracción constante de  $(1 - 1/e) \approx 63\%$  del valor óptimo alcanzable empleando  $k$  aristas), utilizando un algoritmo voraz (los detalles técnicos se encuentran en [5]). Debido a que un análisis de complejidad espacial y temporal depende en gran medida de la estructura de la red y de los supuestos que se hacen de esta, los autores no realizan los análisis

correspondientes.

Mínimos cuadrados alternos (ALS por sus siglas en inglés) es el método utilizado por el algoritmo de factorización matricial de la librería `cmfrec` para la recomendación. Este método calcula una de las matrices a encontrar en la ecuación (2) mientras fija las demás, una vez hecho esto, actualiza la matriz y va alternando con las matrices restantes para calcularlas todas. De esta forma se asegura que en cada paso, disminuye la ecuación hasta su convergencia (los detalles técnicos se encuentran disponibles en [21]). Si la convergencia no ocurre en el número de iteraciones estándar de la librería `cmfrec`, esta provee un parámetro para el número de iteraciones máximas que el método ejecuta. La implementación de este método es de código abierto y disponible gratuitamente<sup>1</sup>.

---

<sup>1</sup><https://github.com/david-cortes/cmfrec>

# Conclusiones

---

En la etapa de la revisión de la literatura para métodos de inferencia de redes sociales [5], [6], [15]-[18] se tuvieron en cuenta varios trabajos de diversos autores, encontrando que todos ellos se basan en el modelo de cascada independiente y los modelos de difusión de información introducidos por Tardos *et al.* [26].

Cada uno de los autores revisados realizan una extensión y/o mejora del algoritmo *NETINF* propuesto por Gomez-Rodriguez *et al.* [5]. Fan *et al.* [6] plantearon usar *NETINF* bajo ciertas condiciones y un específico análisis de cascadas de películas en el que se obtiene mejoras en la complejidad temporal y en la precisión. Por otro lado, Alpay *et al.* [15] desarrollan el algoritmo *FASTINF*, el cual busca ser más rápido y escalable para grandes redes sociales en línea en comparación a *NETINF*. Gray *et al.* [18] desarrollan un método bayesiano para la inferencia de la red, con el fin de probarlo en un pequeño número de cascadas simuladas para las que *NETINF* no produce un resultado. Con base en los resultados obtenidos por estos autores, Gomez-Rodriguez *et al.* decidieron mejorar su algoritmo como se aprecia en [16] y [17]. Por esta razón se concluye que la propuesta de Gomez-Rodriguez *et al.* ha sido la más desarrollada y la más completa, para llevar a cabo esta tesis de grado.

Con respecto a los modelos de inferencia de redes sociales (Tabla 3.1) en la fase de implementación, se usaron distintos valores para la tasa de transmisión ( $\alpha$ ), para obtener diversos modelos de redes sociales, con diferente cantidad de arcos inferidos; con el fin de explorar varios grafos que modelen de la mejor manera la estructura subyacente, en donde se propagan las recomendaciones de películas entre los usuarios.

El presente proyecto de grado desarrolla una técnica para realizar recomendaciones (en este caso de películas) usando redes sociales inferidas y las propiedades topológicas que estas presentan. Para el desarrollo se usó dos algoritmos, *NETINF* y *Matrix Factorization*, para la inferencia de redes sociales y para la recomendación respectivamente. Como se puede observar en la Figura 4.5 el método aquí propuesto genera mejores resultados, que una recomendación común de filtrado colaborativo (únicamente usando algoritmos de factorización matricial). Sin embargo, las mejoras en el *RMSE* son pequeñas, ya que el rating de una película varía en un intervalo de  $[0, 5]$  (para el conjunto de datos de *MovieLens*), por el contrario, en recomendaciones donde el rango de predicción sea más amplio, se notaran resultados más significativos.





# Bibliografía

- [1] Y. Fang and L. Si, “Matrix Co-factorization for Recommendation with Rich Side Information and Implicit Feedback,” in *Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems - HetRec '11*, (New York, New York, USA), pp. 65–69, ACM Press, oct 2011.
- [2] Netflix, “How Netflix’s Recommendations System Works.” <https://help.netflix.com/en/node/100639/us>. Online. [Accessed: 27- Mar- 2021].
- [3] A. Kose, C. Kanbak, and N. Evirgen, “Performance comparison of algorithms for movie rating estimation,” in *Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017*, pp. 955–959, Institute of Electrical and Electronics Engineers Inc., 2017.
- [4] R. R. Sinha and K. Swearingen, “Comparing Recommendations Made by Online Systems and Friends,” in *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries*, (Dublin, Ireland), jun 2001.
- [5] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, “Inferring networks of diffusion and influence,” in *ACM Transactions on Knowledge Discovery from Data*, vol. 5, (New York, New York, USA), pp. 1–37, ACM PUB27, feb 2012.
- [6] C. Fan and L. Yu, “Inferring Social Networks Based on Movie Rating Data,” tech. rep., Stanford University, CA, 2011.
- [7] F. M. Harper and J. A. Konstan, “The movielens datasets: History and context,” *ACM Transactions on Interactive Intelligent Systems*, vol. 5, pp. 1–19, dec 2015.
- [8] Y. Koren, R. Bell, and C. Volinsky, “Matrix factorization techniques for recommender systems,” *Computer*, vol. 42, no. 8, pp. 30–37, 2009.
- [9] H. Ma, H. Yang, M. R. Lyu, and I. King, “SoRec: Social recommendation using probabilistic matrix factorization,” in *International Conference on Information and Knowledge Management, Proceedings*, (New York, New York, USA), pp. 931–940, ACM Press, 2008.
- [10] J. He and W. Chu, “A social network-based recommender system (snrs),” in *Data Mining for Social Network Data*, vol. 12, pp. 47–74, jun 2010.
- [11] B. Liu and Z. Yuan, “Incorporating Social Networks and User Opinions for Collaborative Recommendation: Local Trust Network based Method,” in *Proceedings of the Workshop on Context-Aware Movie Recommendation - CAMRa '10*, (New York, New York, USA), ACM Press, 2010.

- [12] W. Wang, X. Chen, P. Jiao, and D. Jin, "Similarity-based Regularized Latent Feature Model for Link Prediction in Bipartite Networks," *Scientific Reports*, vol. 7, dec 2017.
- [13] H. Li, Y. Liu, Y. Qian, N. Mamoulis, W. Tu, and D. W. Cheung, "HHMF: hidden hierarchical matrix factorization for recommender systems," *Data Mining and Knowledge Discovery*, vol. 33, pp. 1548–1582, nov 2019.
- [14] F. Gasparetti, G. Sansonetti, and A. Micarelli, "Community detection in social recommender systems: a survey," *Applied Intelligence*, pp. 1–21, nov 2020.
- [15] A. Alpay, D. Demir, and J. Yang, "FastInf: A Fast Algorithm to Infer Social Networks from Cascades," tech. rep., Stanford University, CA, 2011.
- [16] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf, "Uncovering the temporal dynamics of diffusion networks," in *Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11*, (Madison, WI, USA), p. 561–568, Omnipress, 2011.
- [17] M. Gomez-Rodriguez and B. Schölkopf, "Submodular inference of diffusion networks from multiple trees," *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, vol. 1, pp. 489–496, may 2012.
- [18] C. Gray, L. Mitchell, and M. Roughan, "Bayesian inference of network structure from information cascades," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 6, pp. 371–381, apr 2020.
- [19] W. T. Tutte, "The dissection of equilateral triangles into equilateral triangles," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 44, no. 4, p. 463–482, 1948.
- [20] G. Takács, I. Pilászy, B. Németh, and D. Tikk, "Matrix factorization and neighbor based algorithms for the netflix prize problem," in *RecSys'08: Proceedings of the 2008 ACM Conference on Recommender Systems*, (New York, New York, USA), pp. 267–274, ACM Press, 2008.
- [21] D. Cortes, "Cold-start recommendations in Collective Matrix Factorization," *ArXiv*, vol. abs/1809.00366, 2018.
- [22] T. Zhou, J. Ren, M. Medo, and Y. C. Zhang, "Bipartite network projection and personal recommendation," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 76, 10 2007.
- [23] W. de Nooy, "Social Network Analysis, Graph Theoretical Approaches to," in *Encyclopedia of Complexity and Systems Science*, pp. 8232–8233, Springer New York, jul 2009.
- [24] J. C. Nacher, "Network Metrics," in *Encyclopedia of Systems Biology*, pp. 1516–1517, Springer New York, jun 2013.

- 
- [25] S. Fortunato, “Community detection in graphs,” *Physics Reports*, vol. 486, pp. 75–174, jun 2009.
  - [26] E. Tardos, D. Kempe, and J. Kleinberg, “Maximizing the spread of influence through a social network,” *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 137–146, 2003.
  - [27] J. Leskovec and R. Sosič, “SNAP: A general-purpose network analysis and graph-mining library,” *ACM Transactions on Intelligent Systems and Technology*, vol. 8, pp. 1–20, jul 2016.