

GRAPH MACHINE LEARNING: Amazon by PyTorch Geometric Datasets



Camila Andrea Arias Vargas
Felipe Clavijo Acosta
Joel Alfredo Márquez Álvarez
Juan Camilo Ramírez Restrepo
Juan Pablo Cuellar Solano

ENTENDIMIENTO DE NEGOCIO



FUNDADA EN 1994

Como una librería en linea, ofreciendo una selección extensa de libros que se organizaban en categorías y subcategorías para facilitar la búsqueda y la navegación.

EXPANSIÓN A OTROS PRODUCTOS

Venta de productos de tecnología, moda, hogar, etc...

NUEVOS SERVICIOS

Migración a servicios de Cloud, almacenamiento, dispositivos inteligentes, streaming

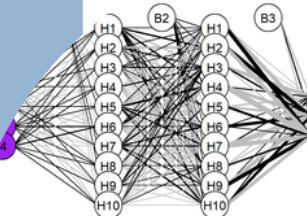
MACHINE LEARNING

Actualmente emplea una combinación de enfoques tradicionales de clasificación y algoritmos avanzados de aprendizaje automático para categorizar sus productos y entregar un mejor servicio.

JUSTIFICACIÓN USO DE GRAFOS

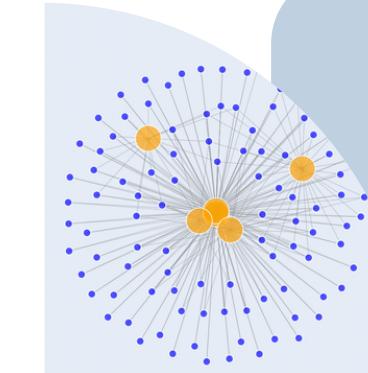
CAPTURA DE RELACIONES COMPLEJAS

Permite representar las relaciones entre productos (como compras conjuntas o características compartidas) de forma natural, mejorando la precisión en recomendaciones y el descubrimiento de productos.



REPRESENTACIÓN INTEGRAL DE PRODUCTOS

Cada producto es un nodo con atributos específicos, y sus conexiones representan relaciones con otros productos, ofreciendo una visión completa que potencia la clasificación y relevancia en las recomendaciones.



ESCALABILIDAD PARA NUEVOS PRODUCTOS

Los grafos permiten una alta escalabilidad, integrando nuevos productos y conexiones sin necesidad de reconstruir el modelo, asegurando recomendaciones actualizadas en tiempo real.



OPTIMIZACIÓN DE LA EXPERIENCIA DEL CLIENTE

Con relaciones de productos mejor representadas, Amazon puede proporcionar recomendaciones que refuerzen la satisfacción y lealtad del cliente en su plataforma.



OBJETIVOS DE NEGOCIO

- Mejorar la experiencia del usuario con recomendaciones personalizadas.
- Aumentar la fidelidad de los clientes y ventas.

1

CUSTOMER LIFETIME VALUE

Estimador del valor total que un cliente generará para la empresa durante toda su relación.

Incrementar del 15-20% en el CLV en un período de 1 año.

2

RECOMMENDATION DIVERSITY

Mide la variedad de productos recomendados a los usuarios, evaluando si las recomendaciones son variadas o si el sistema tiende a recomendar productos similares repetidamente.

Mantener la diversidad en un rango alto, es decir, por encima del 0,6.

3

TIEMPO DE DURACIÓN EN LA APP

Duración en minutos y segundos de las personas dentro de la APP

Aumentar un 10% el tiempo en uso de la APP, superando la barrera de los 4.35 minutos en promedio de uso.



OBJETIVO MINERIA

DESARROLLAR UN SISTEMA DE CLASIFICACIÓN DE PRODUCTOS (COMPUTADORES) BASADO EN GRAFOS QUE PERMITA CATEGORIZARLOS CORRECTAMENTE. ESTE SISTEMA SE DISEÑARÁ PARA SER ESCALABLE Y ADAPTABLE PARA SU USO EN TIEMPO REAL, CONTRIBUYENDO A MEJORAR LA EXPERIENCIA DE USUARIO MEDIANTE CATEGORÍAS PRECISAS Y RELEVANTES, AL FACILITAR LA NAVEGACIÓN Y EXPLORACIÓN DE PRODUCTOS EN LA PLATAFORMA.

ACCURACY

Mantener la métrica en un rango alto, por encima del 0,85.

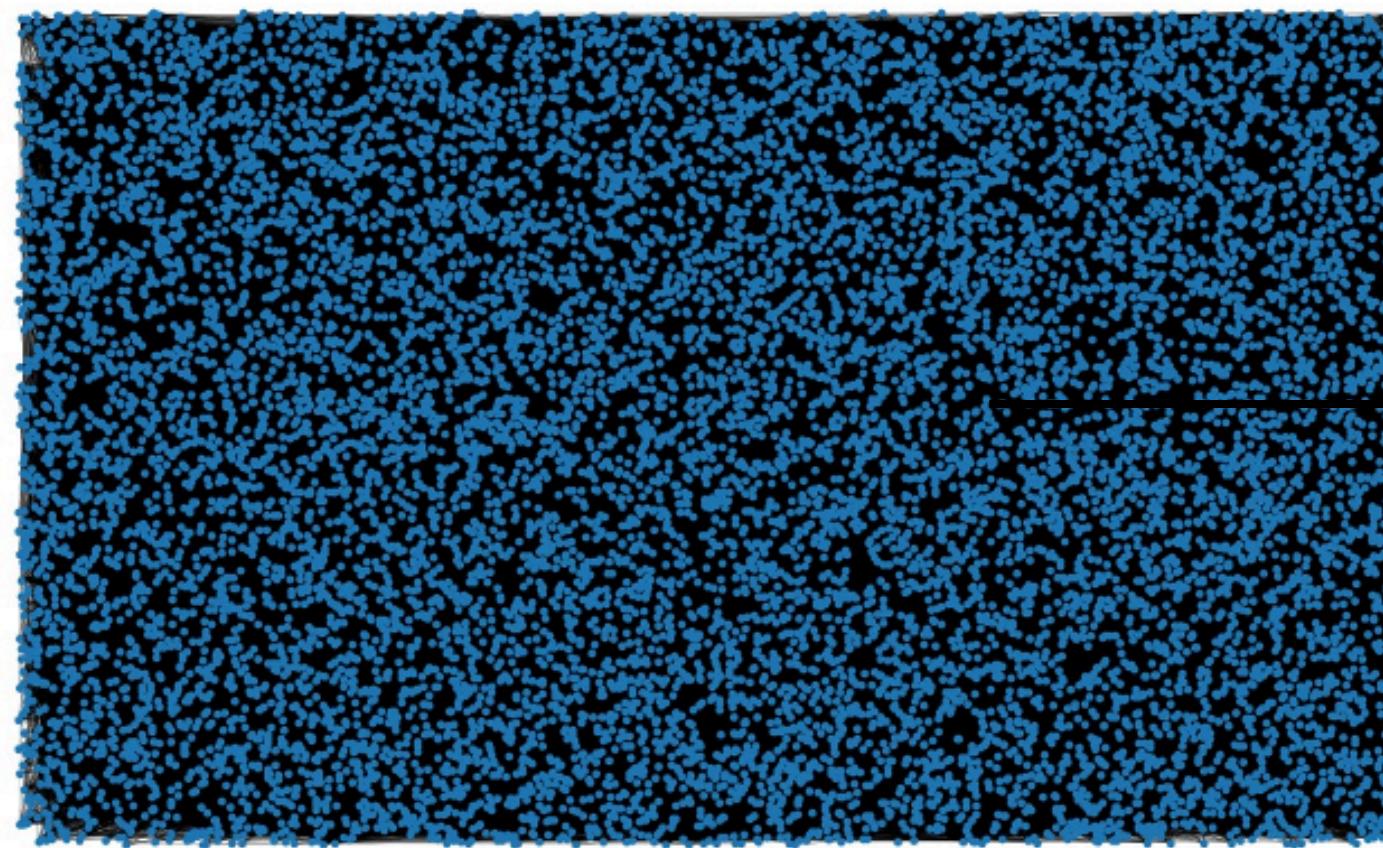
ENTENDIMIENTO DE LOS DATOS

FUENTE: DATASET DE AMAZON EN PYTORCH GEOMETRIC, INCLUYENDO 13,752 NODOS DE PRODUCTOS Y 491,722 CONEXIONES ENTRE ELLOS

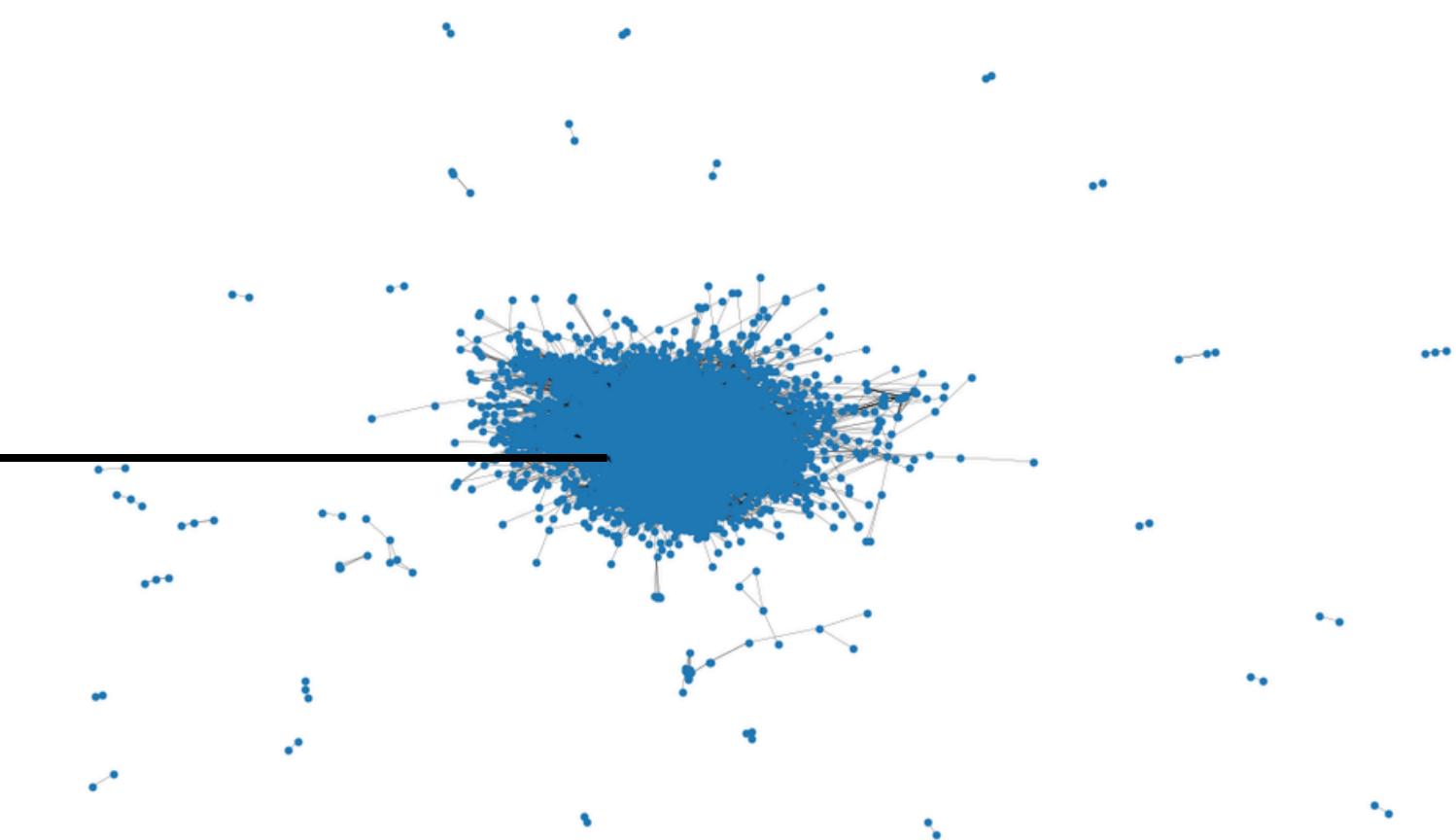
CADA NODO CUENTA CON 767 CARACTERÍSTICAS QUE CORRESPONDEN A LAS RESEÑAS DE PRODUCTOS TRANSFORMADAS EN UNA REPRESENTACIÓN DE BOLSA DE PALABRAS (BAG-OF-WORDS), Y EL CONJUNTO DE DATOS INCLUYE 10 CATEGORÍAS DE PRODUCTOS DIFERENTES.

Name	#nodes	#edges	#features	#classes
Computers	13,752	491,722	767	10

ENTENDIMIENTO DE LOS DATOS

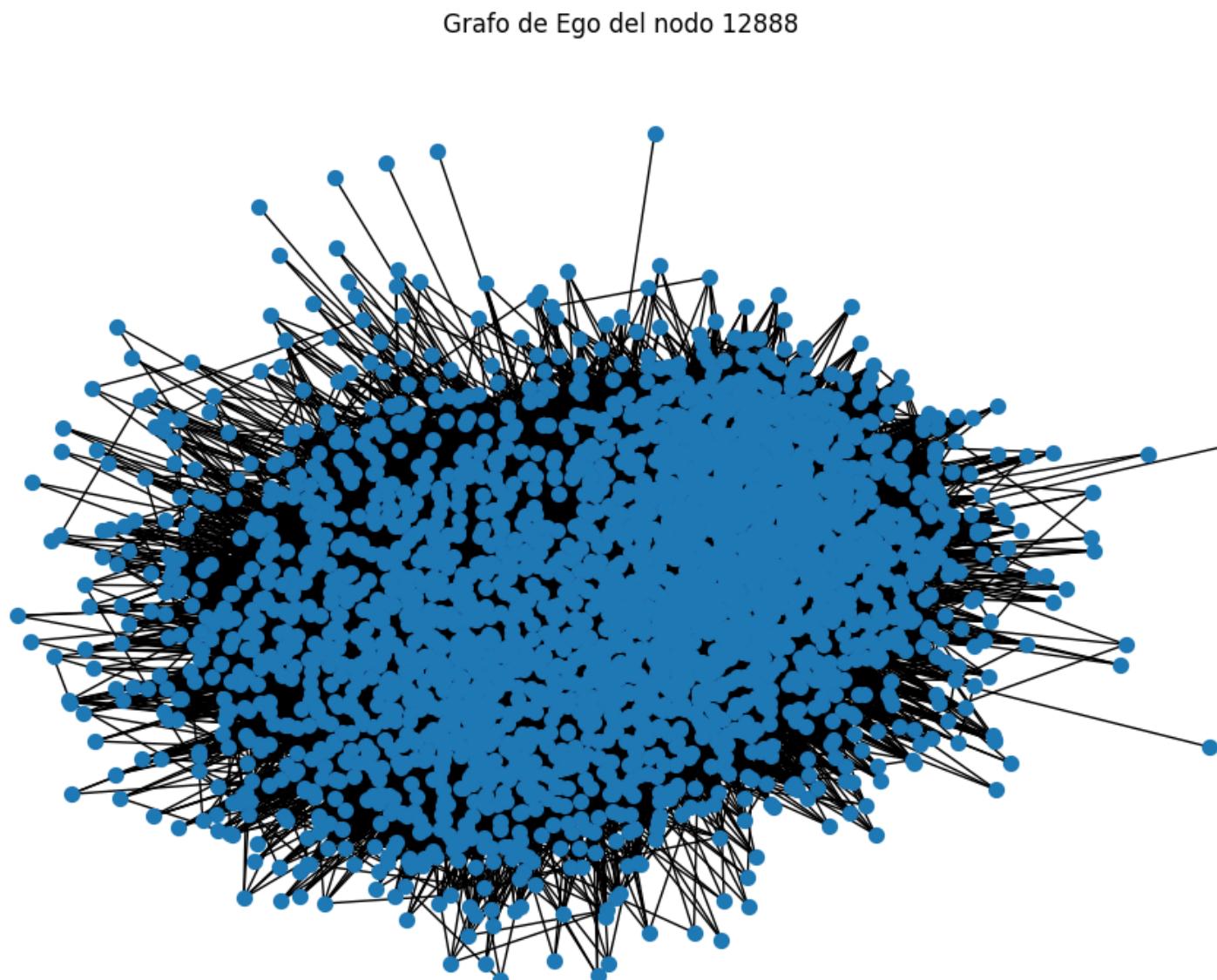


Grafo productos de computación Amazon



Grafo de ego con Spring Layout

ENTENDIMIENTO DE LOS DATOS

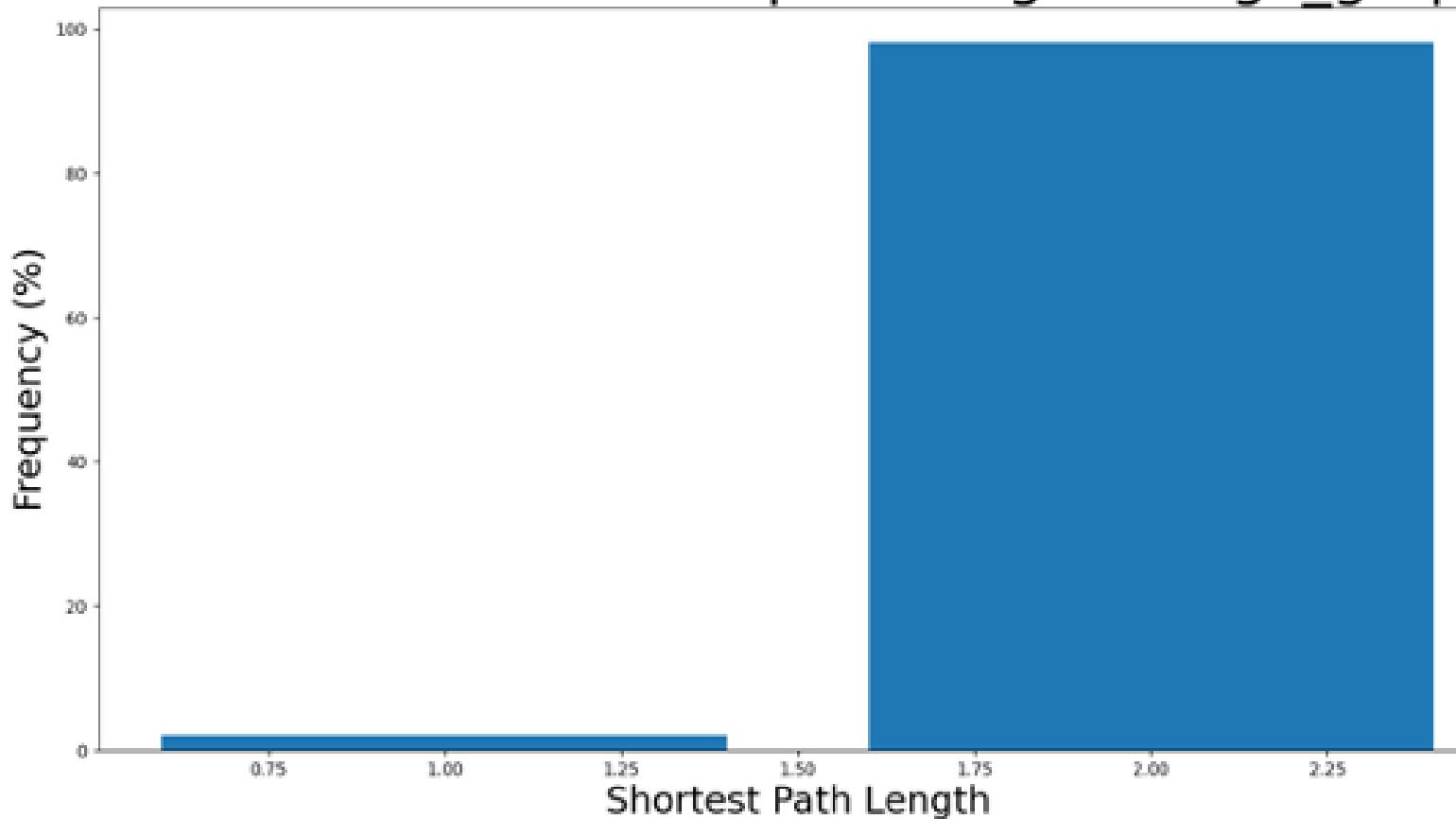


Nodo de mayor grado

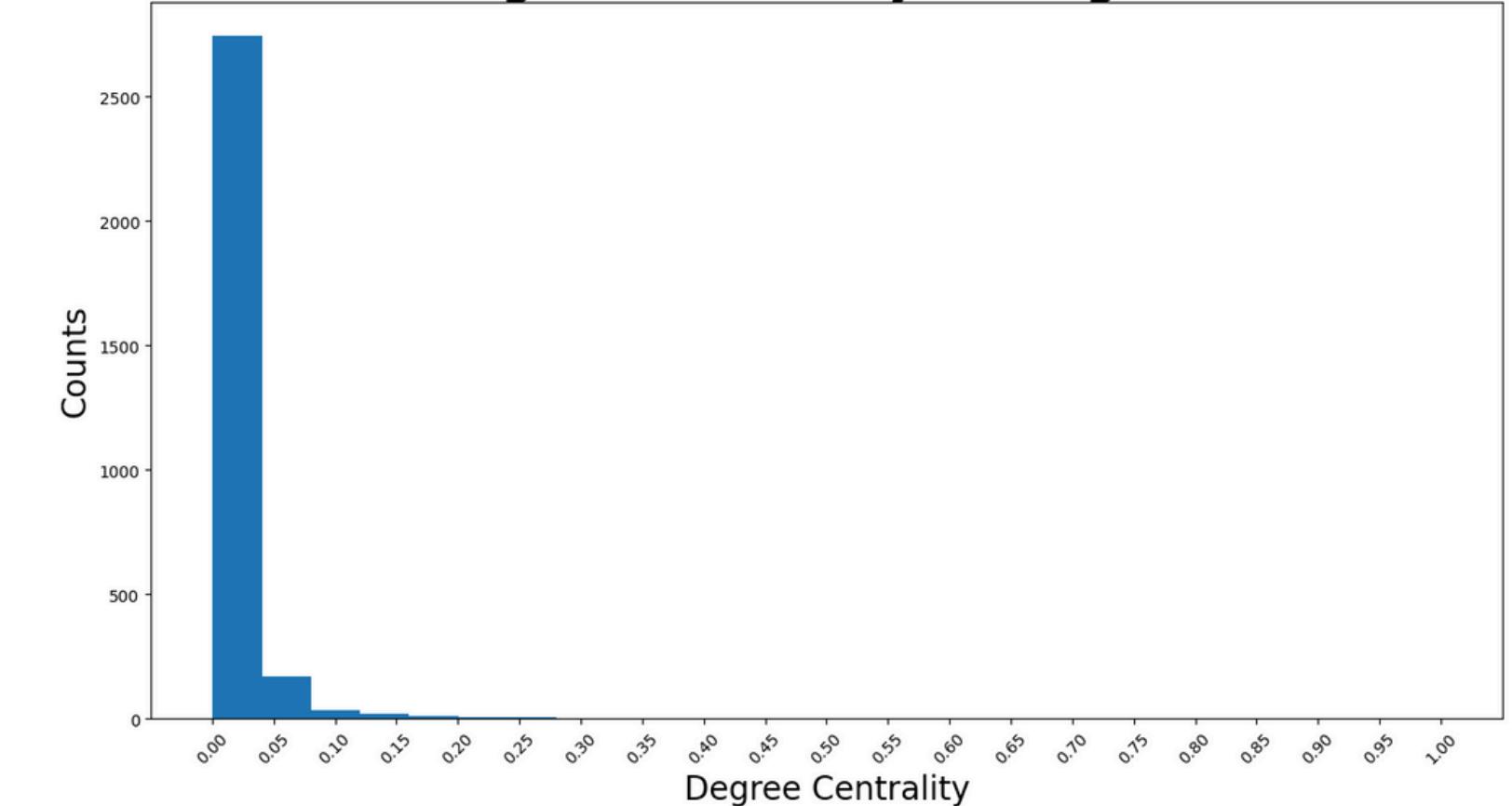
- En promedio, cada nodo está conectado a 57.69 otros nodos
- La longitud promedio de las rutas más cortas fue de aproximadamente 1.98

ENTENDIMIENTO DE LOS DATOS

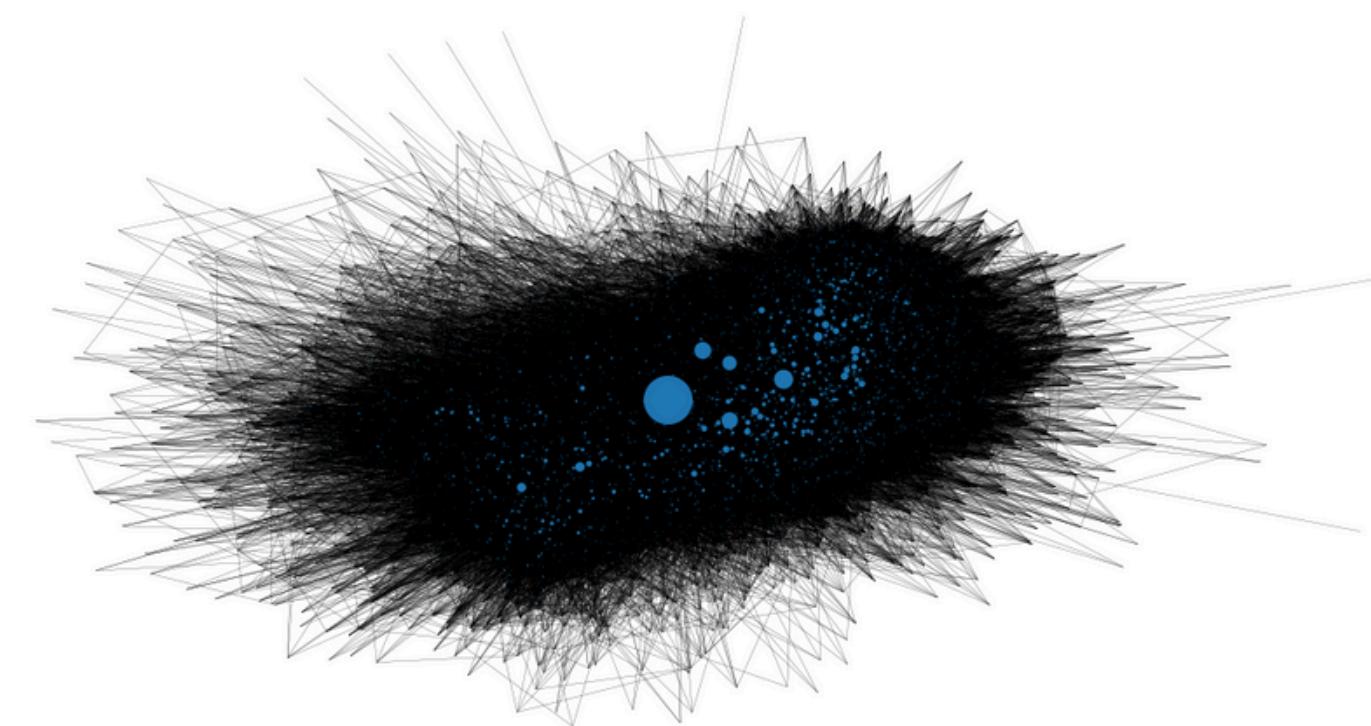
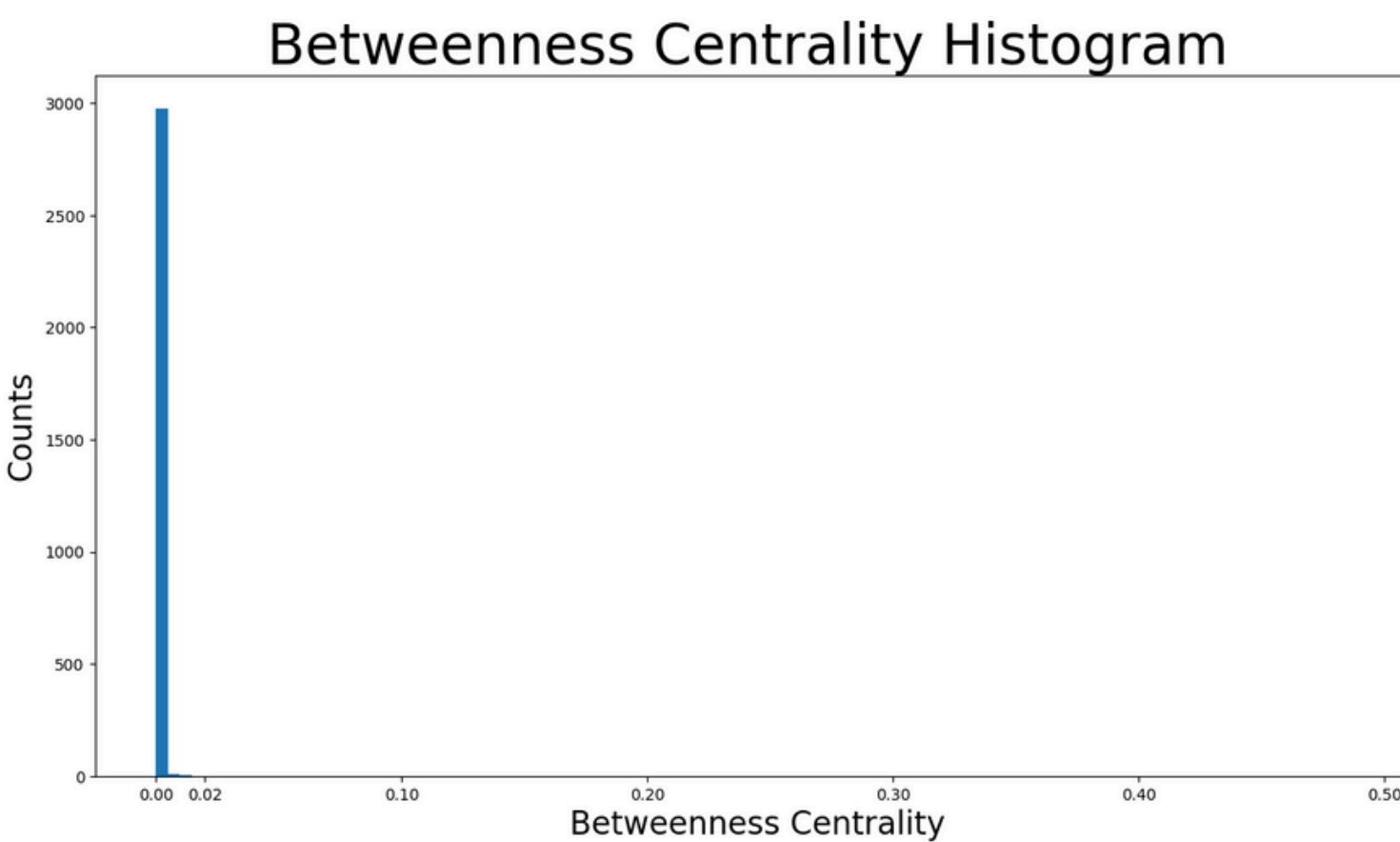
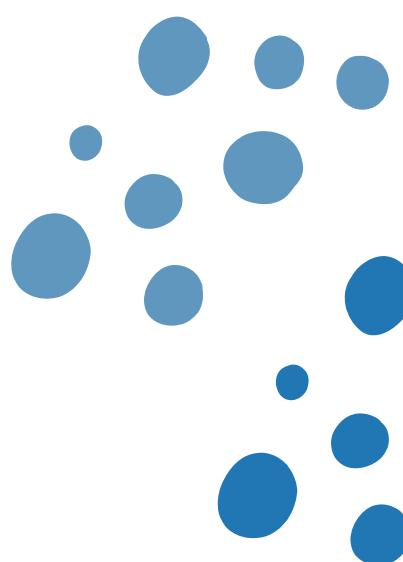
Distribution of shortest path length in ego_graph



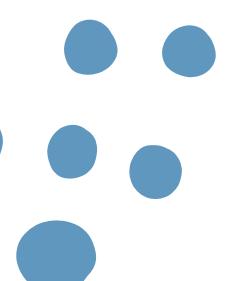
Degree Centrality Histogram



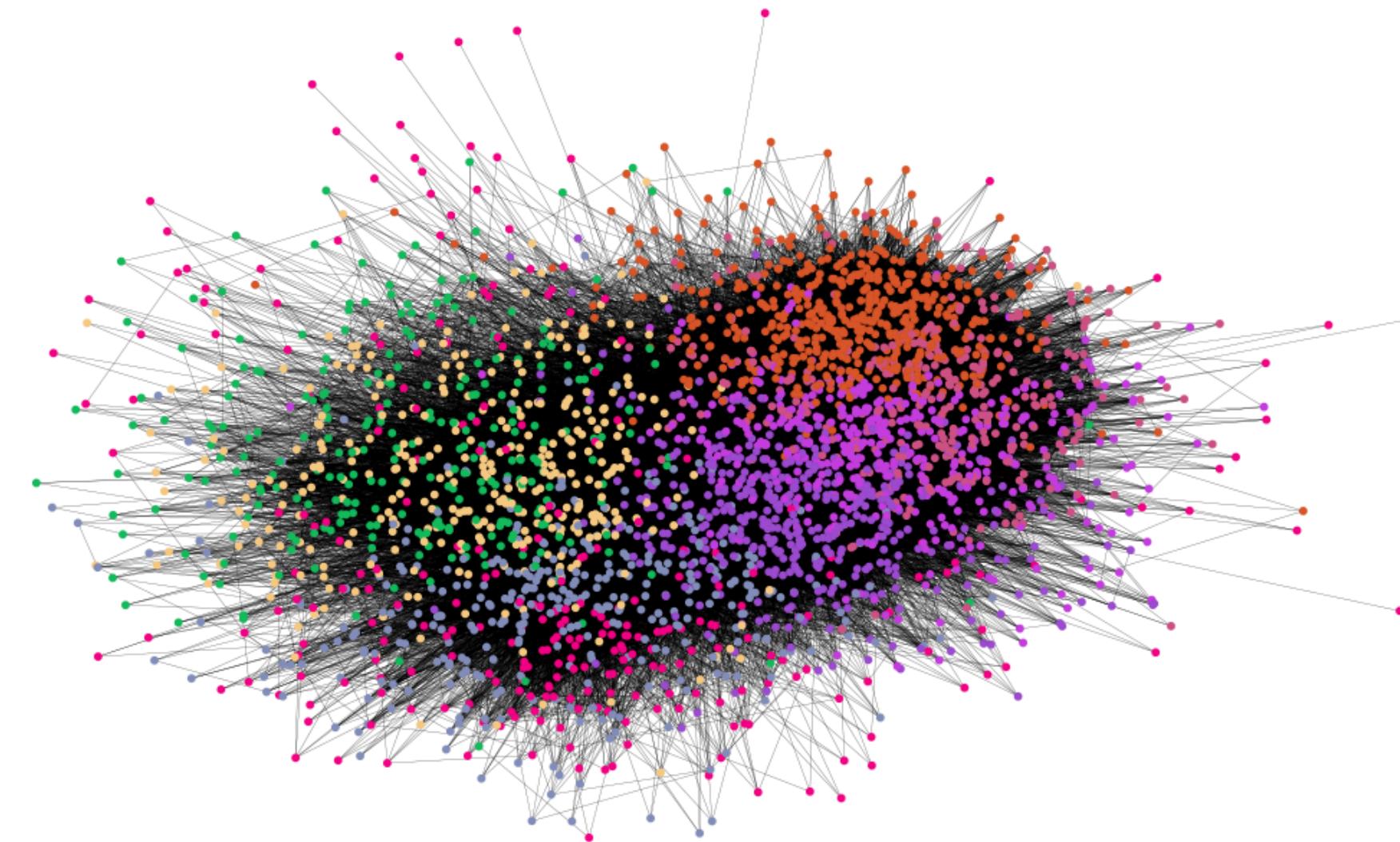
ENTENDIMIENTO DE LOS DATOS



Grafo con nodos con mayor centralidad de intermedición



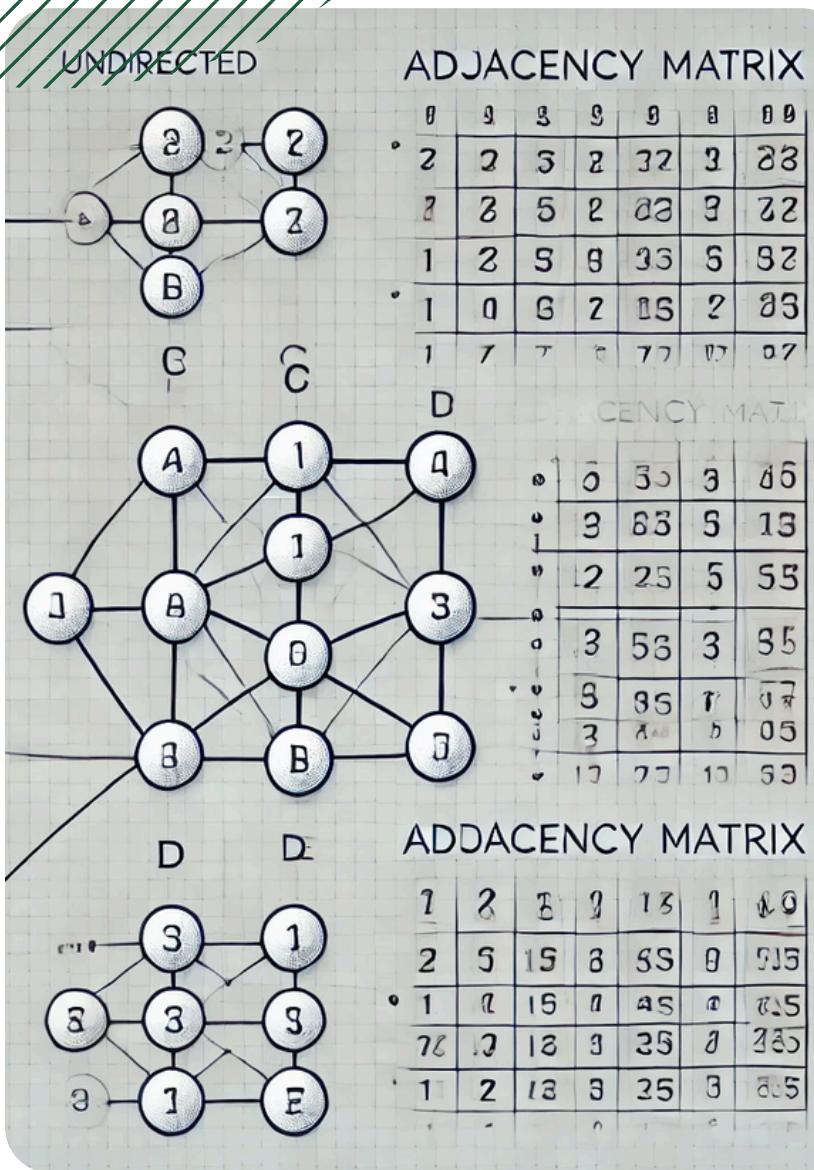
ENTENDIMIENTO DE LOS DATOS



Grafo de comunidades

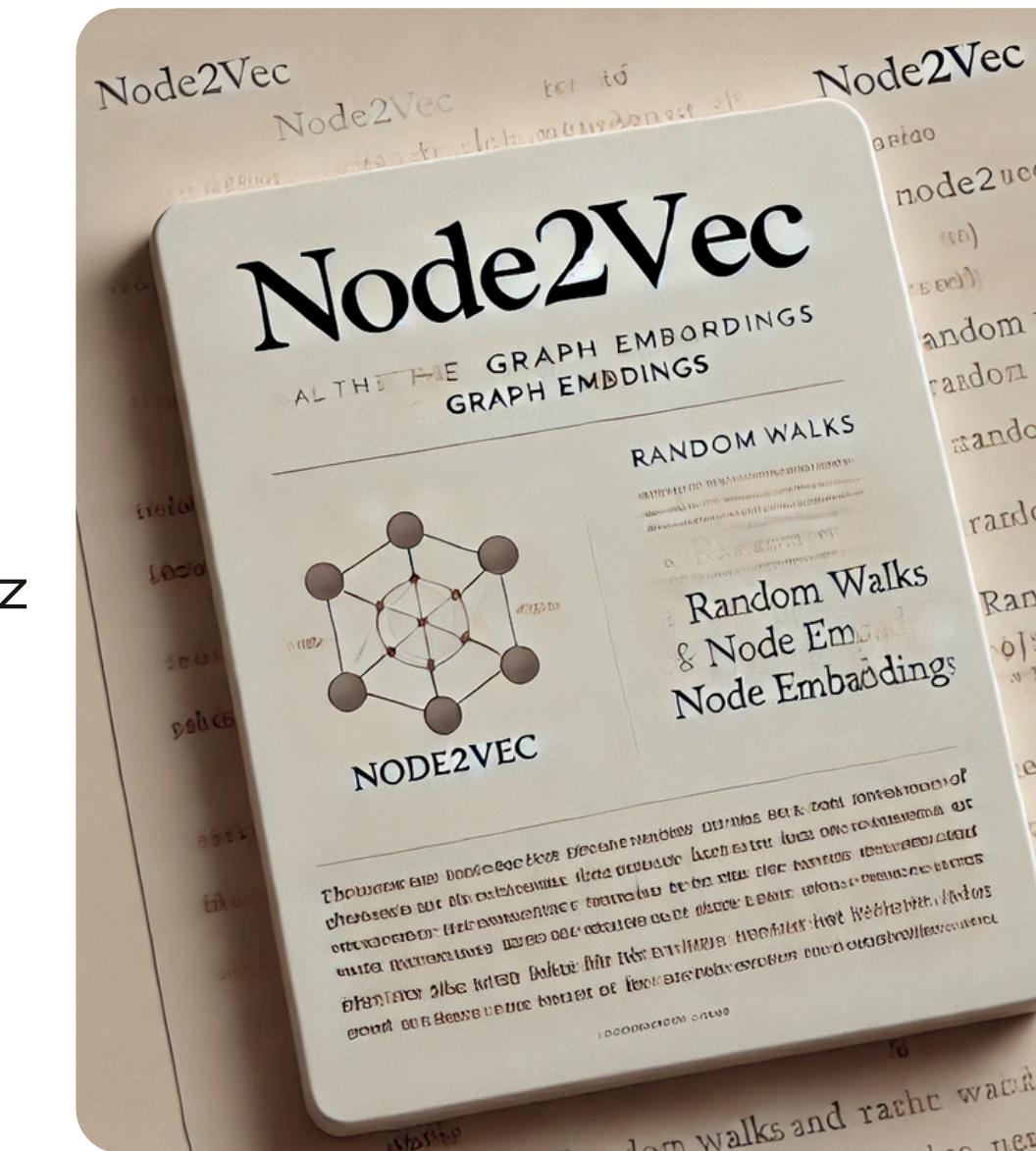
PREPARACIÓN DE LOS DATOS

Generación de la matriz de adyacencia y la obtención de embeddings de nodos mediante Node2Vec.



Matriz de Adyacencia

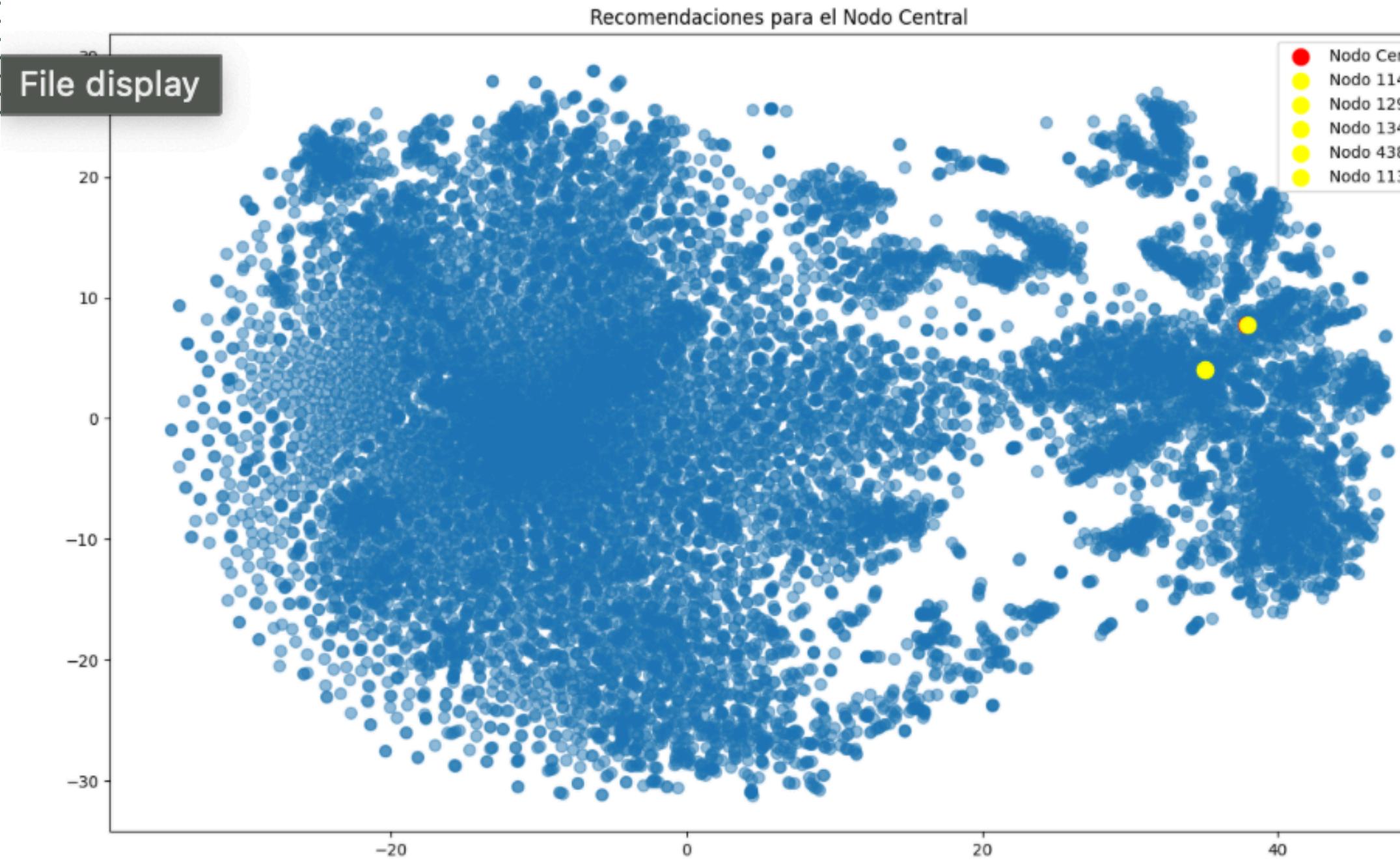
Se generó la matriz de adyacencia del grafo, la cual representa las conexiones entre los nodos en forma de matriz binaria.



Node2Vec

Se generaron embeddings de nodos con Node2Vec mediante caminatas aleatorias parametrizadas, creando secuencias de nodos utilizadas para entrenar el modelo.

SISTEMA DE RECOMENDACIÓN



MODELOS

Para el desarrollo del sistema de clasificación de productos basados en gráficos, se implementarán y evaluarán distintos modelos de aprendizaje automático y de aprendizaje profundo, tanto tradicionales como basados en gráficos.

0.8612

Regresión Logistica
Modelo base

0.8169

Random Forest
Conjunto de árboles de decisión para datos no lineales.

0.8844

GCN*

Extrae características de nodos en grafos para clasificar productos.

0.9106

GAT**

Mejora GCN con atención para priorizar relaciones clave.

0.9193

GATv2Conv

Atención avanzada para captar relaciones complejas en grafos.

0.9019
Con matriz

0.8292
Con matriz

CONCLUSIONES

BENEFICIOS

- Facilitan una gestión eficiente del inventario al modelar las interacciones complejas entre productos.
- Mejoran la precisión en clasificación y recomendaciones, ofreciendo una experiencia personalizada.
- Refuerzan la fidelidad del cliente y optimizan ventas al presentar recomendaciones más relevantes.

DESAFIOS

- Implican un alto costo computacional para ejecutar y mantener grafos densamente conectados.
- Requieren balancear precisión y eficiencia, minimizando costos sin comprometer el rendimiento.
- Demandan un procesamiento y análisis de datos robusto para satisfacer las necesidades del negocio.

