Maestría en Inteligencia Artificial Aplicada

Curso: Proyecto Integrador

Tecnológico de Monterrey

Actividad de la Semana 02

EDA de los conjuntos de datos seleccionados

Equipo 32

José Adan Vega Pérez [A01796093]

Silvia Xochitl Ibañez Vara [A01795200]

Diego Andrés Bernal Díaz [A01795975]

Requisitos

igraph para procesar analisis de grafos en C, en lugar de networkx

1. Estructura de los datos y Macro-Estadisticas

Caracterización Macro-Estadística de los Datasets

Conceptos Previos y Justificación

La primera fase del entendimiento de los datos consiste en dimensionar la magnitud y la densidad de los grafos de conocimiento seleccionados. Estas métricas no solo dictan los requisitos computacionales (memoria RAM y GPU), sino que predeterminan la dificultad de la tarea de extrapolación.

Para este análisis, se consideran las siguientes variables clave:

```
Entidades (|E|) y Relaciones (|R|): Representan el tamaño del vocabulario. Un número alto de entidades (como en FB13) exige matrices de embeddings más grandes, mientras que el número de relaciones define la complejidad semántica de las interacciones.

Grado Promedio (Avg Degree): Calculado como la razón entre el total de tripletas y el número de entidades. Esta es quizás la métrica más crítica para redes neuronales de grafos (GNNs). Un grado alto implica que cada nodo tiene muchos vecinos de los cuales agregar información
```

```
("contexto rico"). Un grado bajo implica "escasez de información", dificultando la inferencia de nuevos enlaces.

Densidad del Grafo: Indica qué tan cerca está el grafo de ser "completo" (donde todos están conectados con todos). En grafos de conocimiento reales, este valor es típicamente muy bajo (dispersión), pero las diferencias de magnitud entre datasets indican cambios drásticos en la topología.
```

Expectativa Teórica

Se espera observar una dicotomía clara entre los datasets de origen lingüístico/jerárquico (WordNet) y los de conocimiento general (Freebase). Los primeros deberían comportarse como estructuras tipo árbol (baja densidad y grado), mientras que los segundos deberían mostrar una estructura de red social o "pequeño mundo" (alta conectividad).

Found datasets: ['WN18RR', 'FB15k-237', 'FB13', 'CoDEx-M', 'WN11']

Dataset	Entities (|E|)	Relations (|R|)	Triples (Total)	Train	Valid	Test	Avg Degree	Graph Density
WN18RR	40,943	11	93,003	86,835	3,034	3,134	2.27	0.000055
FB15k-237	14,541	237	310,116	272,115	17,535	20,466	21.33	0.001467
FB13	75,043	13	375,514	316,232	11,816	47,466	5.00	0.000067
CoDEx-M	17,050	51	206,205	185,584	10,310	10,311	12.09	0.000709
WN11	38,588	11	138,887	112,581	5,218	21,088	3.60	0.000093

```
Found datasets: ['WN18RR', 'FB15k-237', 'FB13', 'CoDEx-M', 'WN11']

<pandas.io.formats.style.Styler at 0x7d519cf88650>
```

Interpretación de las Estadísticas Generales

La Tabla 1 resume las propiedades macroscópicas de los cinco datasets evaluados. El análisis comparativo arroja tres conclusiones fundamentales para el proyecto:

1. Dicotomía Topológica: Árboles vs. Redes Densas Se confirma una clara distinción estructural. Los datasets basados en WordNet (WN18RR y WN11) presentan un Grado Promedio extremadamente bajo (2.27 y 3.60 respectivamente). Esto indica que la mayoría de las entidades tienen apenas 2 o 3 conexiones, formando estructuras jerárquicas lineales o tipo árbol.

   Implicación para el Proyecto: La extrapolación en WordNet será significativamente más difícil, ya que los modelos GNN tendrán muy poca información vecinal ("message passing") para construir representaciones robustas de entidades nuevas.

2. Riqueza Contextual en Freebase y CoDEx En contraste, FB15k-237 destaca como el grafo más denso y rico, con un Grado Promedio de 21.33. Esto significa que, en promedio, una entidad tiene más de 20 conexiones directas, proporcionando abundante contexto semántico. CoDEx-M se posiciona en un punto medio equilibrado (Grado ~12), lo que lo convierte en un benchmark moderno ideal para validar estabilidad entre extremos.

3. Desafío de Escalabilidad en FB13 Aunque FB15k-237 tiene más conexiones, FB13 es el dataset más grande en términos de nodos, con 75,043 entidades.

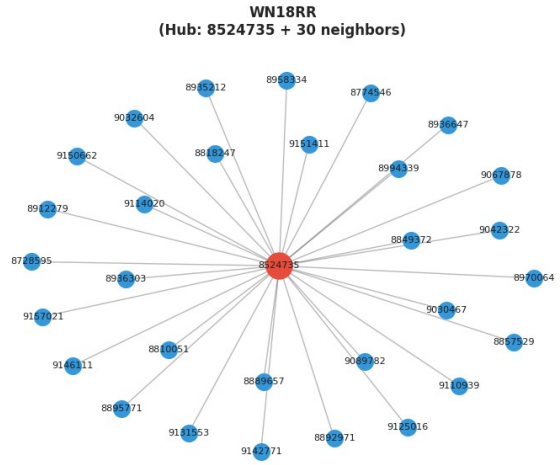
   Análisis: Este volumen de entidades pone a prueba la capacidad de generalización del modelo en espacios de búsqueda más amplios. Además, se observa que FB13 tiene un conjunto de prueba (Test set) inusualmente grande (47,466 tripletas) en comparación con su validación, lo que garantiza una evaluación estadística muy rigurosa pero computacionalmente costosa.

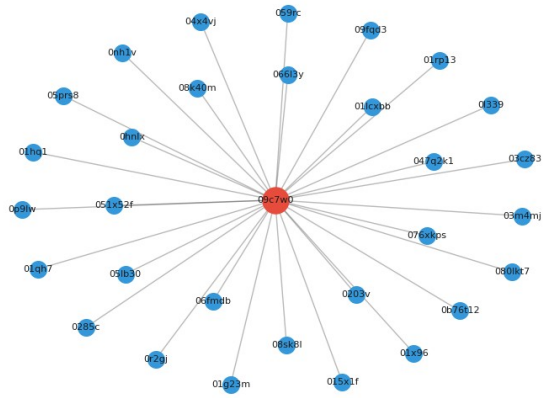
Conclusión del Análisis Macro La selección de datasets es heterogénea y adecuada. Cubre desde escenarios de escasez de datos (WN18RR) hasta escenarios de alta densidad (FB15k-237) y gran escala (FB13). Esto permitirá evaluar si la metodología propuesta es robusta ante diferentes topologías de grafo, cumpliendo con el objetivo de validar la generalización en contextos diversos.

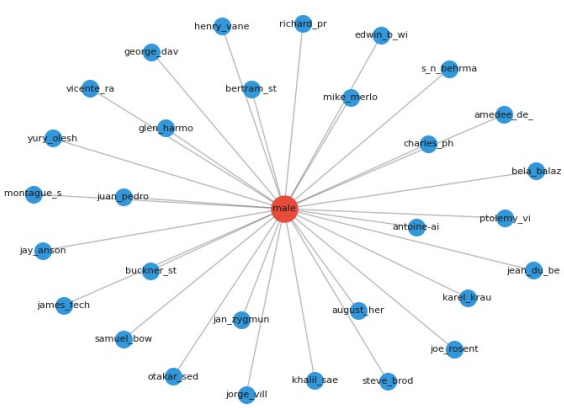
1.1 Visualizacion general de cada dataset

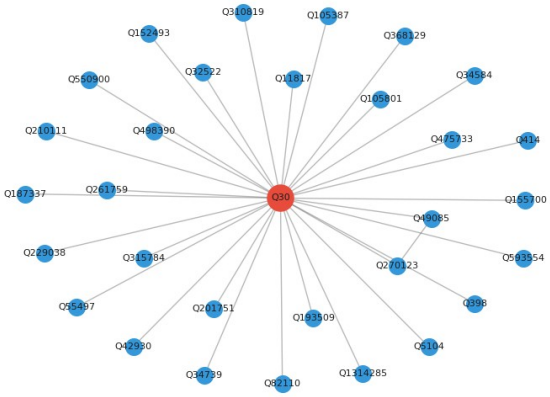
Visualizamos solo el nodo con mas conexiones ya lgunos de sus vecinos proximos para tener un panorama visual de las entidades.

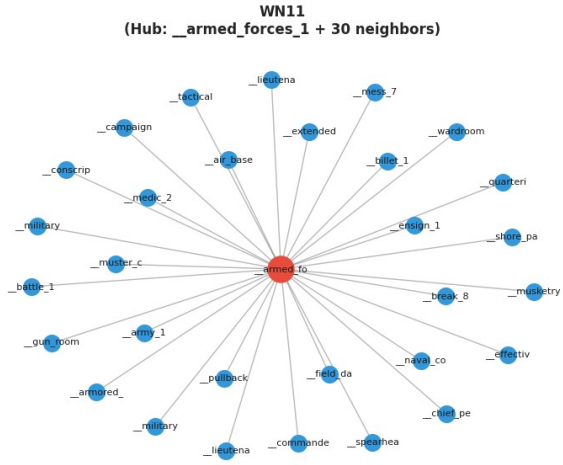
```
Generating visualizations for: ['WN18RR', 'FB15k-237', 'FB13', 'CoDEx-M', 'WN11']
```











Este grafico nos muestra como los datasets CODEX-M, FB15K-237 y WN18RR estan conformados en este momento por id de entidades, mientars que WN11 y FB13 poseen las entidades reales y sus conexiones.

1.2 Análisis de relaciones

Conceptos Previos y Justificación

En el aprendizaje automático aplicado a Grafos de Conocimiento, la distribución de las relaciones juega un papel crítico en la capacidad de generalización del modelo. Un dataset ideal presentaría una distribución equilibrada, donde el modelo tenga suficientes ejemplos de cada tipo de relación para aprender sus patrones. Sin embargo, los datos del mundo real suelen seguir una "Ley de Potencia" (Power Law), donde unas pocas relaciones dominan la mayoría de las tripletas (relaciones Head) y la gran mayoría de las relaciones aparecen muy pocas veces (relaciones Tail).

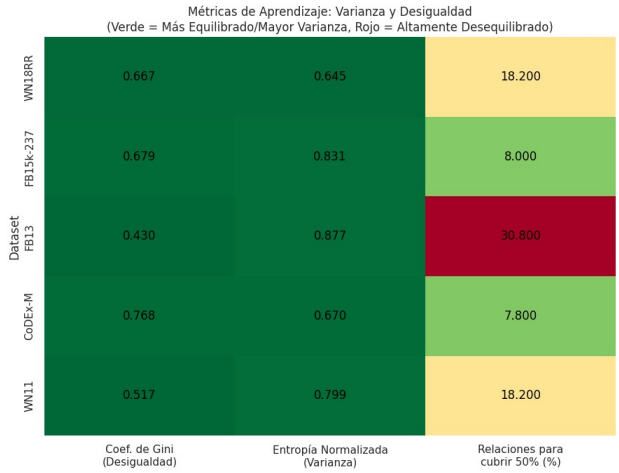
Para cuantificar este fenómeno y evaluar la "aprendibilidad" (learnability) de los datasets seleccionados (WN11, FB13, CoDEx-M, FB15k-237, WN18RR), se emplean las siguientes métricas estadísticas:

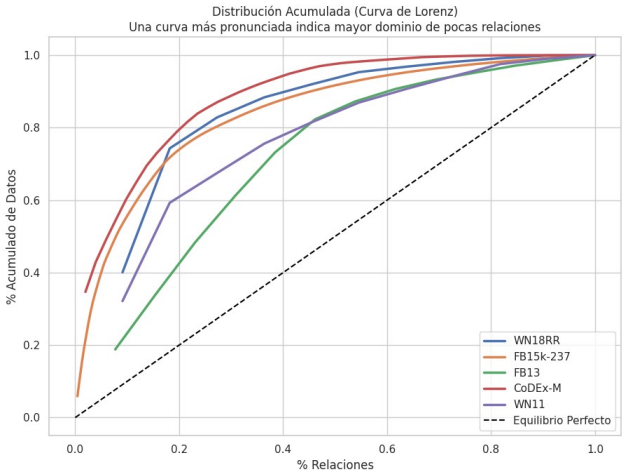
```
Coeficiente de Gini (0.0 - 1.0): Una medida económica de desigualdad. Un valor cercano a 0 indica igualdad perfecta (todas las relaciones aparecen la misma cantidad de veces), mientras que un valor cercano a 1 indica máxima desigualdad (una sola relación domina todo el grafo).

Entropía Normalizada (0.0 - 1.0): Mide la "sorpresa" o varianza de la información. Una entropía alta sugiere que el dataset es rico y variado, obligando al modelo a aprender características estructurales complejas en lugar de memorizar la relación más frecuente.

Curva de Lorenz: Representación gráfica de la desigualdad acumulada. Cuanto más se curve la línea hacia la esquina superior izquierda (alejándose de la diagonal), mayor es el desbalance del dataset.
```

Expectativa Teórica

Se anticipa que los datasets derivados de Freebase (FB15k-237, CoDEx) presenten una alta desigualdad (alto Gini) debido a su naturaleza enciclopédica, lo cual plantea un desafío mayor para la extrapolación de entidades. Por el contrario, datasets más curados o pequeños como FB13 podrían mostrar un balance artificialmente alto.

```
Analyzing relations for: ['WN18RR', 'FB15k-237', 'FB13', 'CoDEx-M', 'WN11']
```






```
{"summary":"{\n  \"name\": \"df_stats\",\n  \"rows\": 5,\n
\"fields\": [\n    {\n      \"column\": \"Dataset\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 5,\n        \"samples\": [\n          \"FB15k-
237\",\n          \"WN11\",\n          \"FB13\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Relaciones \\u00danicas\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
97,\n        \"min\": 11,\n        \"max\": 237,\n
\"num_unique_values\": 4,\n        \"samples\": [\n          237,\n
51,\n          11\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Coeficiente de Gini (Desigualdad)\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 0.13601360226095038,\n
\"min\": 0.43,\n        \"max\": 0.768,\n
\"num_unique_values\": 5,\n        \"samples\": [\n          0.679,\n
0.517,\n          0.43\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Entrop\\u00eda Normalizada (Varianza)\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 0.10183221494203097,\n
\"min\": 0.645,\n        \"max\": 0.877,\n
\"num_unique_values\": 5,\n        \"samples\": [\n          0.831,\n
0.799,\n          0.877\n        ],\n        \"semantic_type\": \"\",\
n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"Relaci\\u00f3n Top-1 (%)\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 13.847815712234187,\n
\"min\": 5.9,\n        \"max\": 40.1,\n        \"num_unique_values\":
5,\n        \"samples\": [\n          5.9,\n          32.1,\n
18.8\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Relaciones para cubrir 50% (%)\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 9.462557793746889,\n
\"min\": 7.8,\n        \"max\": 30.8,\n        \"num_unique_values\":
4,\n        \"samples\": [\n          8.0,\n          7.8,\n
18.2\n        ],\n        \"semantic_type\": \"\",\n
```

\"description\": \"\"\n   }\n   }\n]\
n}","type":"dataframe","variable_name":"df_stats"}

Interpretación de la Distribución de Relaciones

El análisis comparativo de los datasets revela diferencias estructurales significativas que impactan directamente la estrategia de modelado para la tarea de extrapolación:

1. Desbalance Extremo en CoDEx-M y FB15k-237 Los resultados confirman que CoDEx-M es el dataset con mayor desigualdad estructural, presentando un Coeficiente de Gini de 0.768. Esto se evidencia en la métrica de cobertura: apenas el 7.8% de los tipos de relaciones constituyen el 50% de todos los datos disponibles.

   Implicación: Este es el escenario más realista y desafiante. El modelo deberá ser capaz de extrapolar información sobre entidades nuevas basándose frecuentemente en relaciones "raras" (Tail), lo cual justifica el uso de redes neuronales de grafos (GNNs) que capturen topología local en lugar de simples estadísticas de frecuencia.

2. El Caso Atípico de FB13 A diferencia de los estándares modernos, FB13 muestra un comportamiento inusualmente balanceado con un Gini de 0.430 (el más bajo del grupo) y una Entropía Normalizada de 0.877 (la más alta).

   Análisis: El hecho de que se requiera el 30.8% de las relaciones para cubrir la mitad de los datos indica que FB13 no sufre del problema de "relaciones dominantes". Aunque esto facilita el aprendizaje, podría no ser representativo de la complejidad de un grafo de conocimiento dinámico real.

3. Sesgo de Relación Única en WN18RR Se observa un fenómeno crítico en WN18RR: aunque tiene pocas relaciones (11), la relación más frecuente (Top-1 Relation) abarca el 40.1% de todo el dataset.

   Implicación: Existe un riesgo alto de que el modelo caiga en un mínimo local donde simplemente prediga esta relación mayoritaria para cualquier entidad nueva. Las técnicas de extrapolación deberán demostrar que superan a un clasificador base que siempre predice la clase mayoritaria.

Conclusión Visual (Curvas de Lorenz)

Las gráficas de distribución acumulada corroboran lo anterior: las curvas de CoDEx-M y FB15k-237 presentan arcos pronunciados ("codos") hacia la esquina superior izquierda, visualizando cómo un pequeño subconjunto de relaciones acapara la densidad de probabilidad. En contraste, la curva de FB13 es la más cercana a la diagonal de "Balance Perfecto", confirmando su menor complejidad distributiva.

1.3 Análisis Topológico: Distribución de Grados y Ley de Potencia

Conceptos Previos y Justificación

En la teoría de grafos, el "grado" (degree) de un nodo es el número de conexiones directas que posee. La distribución de estos grados es la métrica fundamental para entender cómo fluye la información dentro de la red. En el contexto de los Knowledge Graphs (KGs), esta distribución no es aleatoria, sino que suele seguir una Ley de Potencia (Power Law): la mayoría de los nodos tienen muy pocas conexiones ("Cola Larga" o Long Tail), mientras que un pequeño número de nodos ("Concentradores" o Hubs) acumulan una cantidad masiva de enlaces.

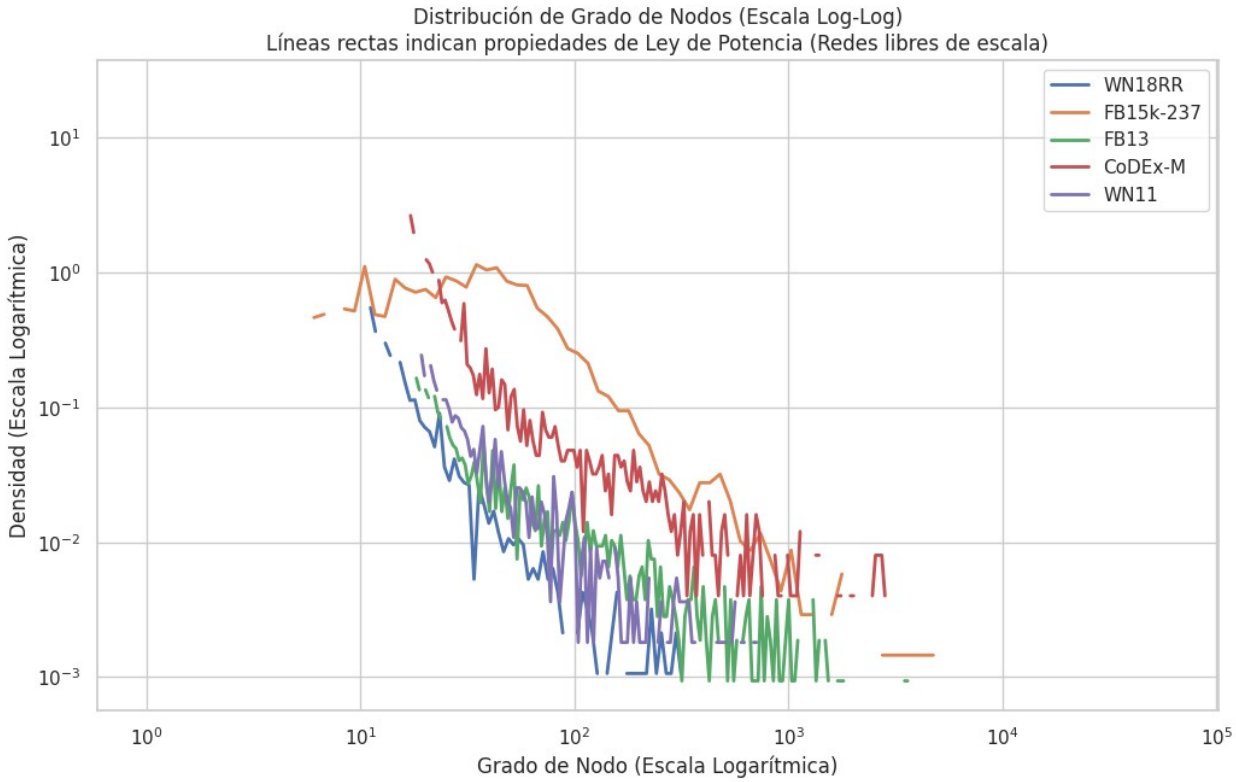
Este análisis es crítico para la tarea de Extrapolación por dos razones:

```
El Problema de la Escasez (Sparsity): Las redes neuronales de grafos (GNNs) aprenden agregando información de los vecinos. Si una entidad nueva tiene un grado bajo (ej.<3), el modelo tiene muy poca información estructural para generar una representación vectorial (embedding) precisa.

El Sesgo de los Hubs: Los nodos extremadamente conectados dominan el proceso de entrenamiento. El modelo tiende a sobreajustarse a ellos, prediciendo enlaces hacia los hubs simplemente por su popularidad, ignorando la semántica real.
```

Expectativa Teórica

Visualizaremos la distribución en una escala Log-Log. Si los datos siguen una línea recta descendente, confirmamos que son "redes libres de escala" (scale-free). Se anticipa que WN18RR sea el dataset más disperso (muchos nodos con grado bajo) debido a su estructura taxonómica, mientras que FB13 o FB15k-237 deberían mostrar la presencia de Super-Hubs masivos.

```
Analyzing Node Degrees for: ['WN18RR', 'FB15k-237', 'FB13', 'CoDEx-M', 'WN11']
```




```
{"summary":"{\n  \"name\": \"display(df_degree_stats\",\n  \"rows\":
5,\n  \"fields\": [\n    {\n      \"column\": \"Dataset\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 5,\n        \"samples\": [\n          \"FB15k-
237\",\n          \"CoDEx-M\",\n          \"WN11\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Total Nodos\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
24328,\n        \"min\": 14541,\n        \"max\": 75043,\n
\"num_unique_values\": 5,\n        \"samples\": [\n          14541,\n
17050,\n          38588\n        ],\n        \"semantic_type\": \"\",\
n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"Grado Max (Hub)\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 25008,\n        \"min\":
521,\n        \"max\": 59666,\n        \"num_unique_values\": 5,\n
\"samples\": [\n          8642,\n          6911,\n          1146\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"Grado Mediano\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
9.528903399657278,\n        \"min\": 3.0,\n        \"max\": 26.0,\n
\"num_unique_values\": 4,\n        \"samples\": [\n          26.0,\n
13.0,\n          3.0\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Nodos Cola (<3 edges)\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 6707,\n        \"min\": 0,\n
```

```
\"max\": 15490,\n        \"num_unique_values\": 5,\n
\"samples\": [\n          534,\n          0,\n          1346\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"%
Cola (<3 edges)\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 16.1689307005751,\n        \"min\":
0.0,\n        \"max\": 37.83,\n        \"num_unique_values\": 5,\n
\"samples\": [\n          3.67,\n          0.0,\n          3.49\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"% Hubs (>100 edges)\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
2.471754437641409,\n        \"min\": 0.1,\n        \"max\": 5.98,\n
\"num_unique_values\": 5,\n        \"samples\": [\n          5.98,\n
2.01,\n          0.26\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\n}","type":"dataframe"}
```


Interpretación de la Conectividad y Escasez

El análisis de la distribución de grados revela desafíos estructurales opuestos entre los datasets seleccionados, lo cual valida la necesidad de estrategias de extrapolación adaptativas:

1. El Desafío de la Escasez en WN18RR La tabla muestra que WN18RR es, por mucho, el grafo más difícil desde una perspectiva estructural. El 37.83% de sus nodos son "escasos" (tienen menos de 3 conexiones) y su mediana de grado es apenas 3.0.

   Implicación: En este dataset, más de un tercio de las entidades carecen de suficiente contexto estructural. Para extrapolar aquí, no bastará con mirar a los vecinos; será indispensable utilizar información auxiliar (como descripciones textuales u ontologías) para compensar la falta de enlaces.

2. La "Limpieza" Artificial de CoDEx-M Sorprendentemente, CoDEx-M presenta un 0.00% de nodos escasos. Esto indica que es un dataset altamente curado donde se han eliminado entidades aisladas.

   Análisis: Aunque facilita el entrenamiento (mediana de grado 13), este escenario podría ser demasiado optimista comparado con un entorno real de "mundo abierto", donde las nuevas entidades suelen llegar con pocas conexiones iniciales.

3. Los Super-Hubs de FB13 FB13 ilustra el extremo opuesto: la presencia de Super-Hubs. Mientras su mediana es baja (5.0), posee un nodo con 59,666 conexiones (Max Degree).

   Visualización: En el gráfico Log-Log, esto se vería como una línea que se extiende mucho hacia la derecha en el eje X.

   Riesgo: Este desbalance extremo sugiere que casi todo el grafo está conectado a unas pocas entidades centrales (probablemente conceptos genéricos como "Persona" o "Lugar"). El modelo de extrapolación deberá ser robusto para no predecir siempre estos nodos hubs como respuesta por defecto.

Conclusión Visual (Gráfico Log-Log)

Las líneas en el gráfico muestran pendientes negativas claras, confirmando la propiedad de Ley de Potencia en todos los casos, aunque con diferentes pendientes:

```
La curva de WN18RR (violeta) cae abruptamente al inicio, visualizando la alta densidad de nodos con grado bajo.

Las curvas de FB13 (naranja) y FB15k-237 (verde) tienen "colas" largas hacia la derecha, evidenciando la existencia de nodos ricos en información.
```

1.4 Patrones de conectividad - Analisis estructural

Conceptos Previos y Justificación

Para entender la "forma" de los datos más allá de simples conteos, es necesario analizar cómo se conectan los nodos entre sí. Dado el gran tamaño de los grafos (hasta 75k nodos en FB13), se utilizó la librería igraph (optimizada en C) para calcular métricas estructurales complejas que serían inviables con librerías estándar en Python.

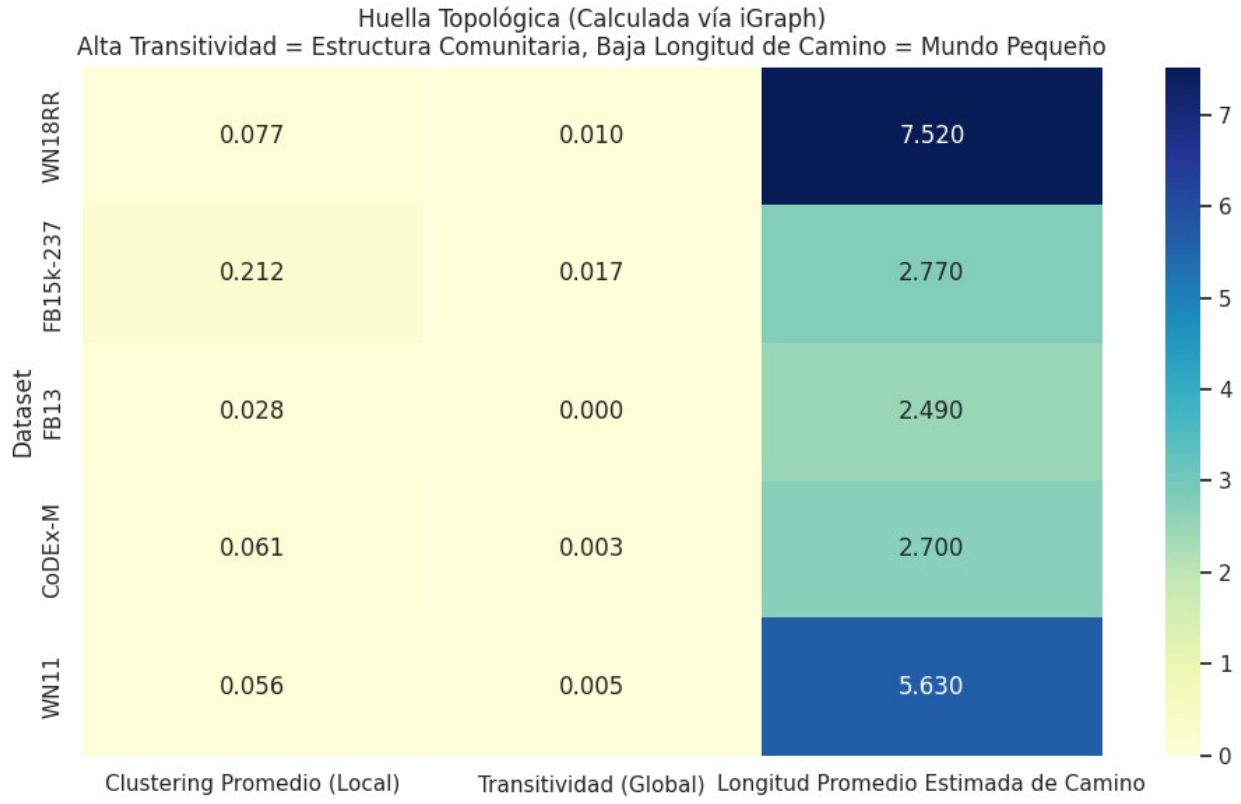
Se evalúan tres propiedades topológicas fundamentales:

```
Componente Gigante (Giant Component): Mide qué porcentaje del grafo está conectado en una sola "isla". Para que los algoritmos de propagación de mensajes (como las GNNs) funcionen, es ideal que el grafo sea conexo (cercano al 100%).

Coeficiente de Clustering y Transitividad:

    Clustering Local: Mide la probabilidad de que los vecinos de un nodo también sean vecinos entre sí (formación de triángulos). Un valor alto indica comunidades densas o redundancia.

    Transitividad Global: Una medida macroscópica de la cohesión de la red.

Longitud de Camino Promedio (Avg Path Length): Indica cuántos "saltos" son necesarios en promedio para ir de una entidad cualquiera a otra.

    Redes de "Pequeño Mundo" (Small World): Tienen caminos cortos (~2-3 saltos), típicos de redes sociales o bases de conocimiento generales.

    Redes Jerárquicas: Tienen caminos largos (>5 saltos), típicos de taxonomías o árboles genealógicos.
```

Expectativa Teórica

Se espera que WordNet (WN), al ser un diccionario taxonómico, muestre caminos largos (profundidad del árbol) y bajo clustering (los hermanos en un árbol no suelen conectarse entre sí). Por el contrario, Freebase (FB) y CoDEx deberían comportarse como redes de "pequeño mundo" con caminos cortos y mayor agrupación local.

```
Analyzing Graph Topology using iGraph (C-Accelerated)...
--> Processing WN18RR...
--> Processing FB15k-237...
--> Processing FB13...
--> Processing CoDEx-M...
--> Processing WN11...
```




```
{"summary":"{\n  \"name\": \"df_topology\",\n  \"rows\": 5,\n
\"fields\": [\n    {\n      \"column\": \"Dataset\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 5,\n        \"samples\": [\n          \"FB15k-
237\",\n          \"WN11\",\n          \"FB13\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Nodos (Total)\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
```

```
24328,\n        \"min\": 14541,\n        \"max\": 75043,\n
\"num_unique_values\": 5,\n        \"samples\": [\n          14541,\n
38588,\n          75043\n        ],\n        \"semantic_type\": \"\",\
n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"Aristas (Total)\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 117357,\n        \"min\":
93003,\n        \"max\": 375514,\n        \"num_unique_values\": 5,\n
\"samples\": [\n          310116,\n          138887,\n
375514\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"%
en Componente Gigante\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 0.12247448713915775,\n        \"min\":
99.7,\n        \"max\": 100.0,\n        \"num_unique_values\": 3,\n
\"samples\": [\n          99.9,\n          100.0,\n          99.7\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"Clustering Promedio (Local)\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
0.07224737365468727,\n        \"min\": 0.0278,\n        \"max\":
0.2119,\n        \"num_unique_values\": 5,\n        \"samples\": [\n
0.2119,\n          0.0555,\n          0.0278\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Transitividad (Global)\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
0.006645928947793328,\n        \"min\": 0.0,\n        \"max\":
0.0166,\n        \"num_unique_values\": 5,\n        \"samples\": [\n
0.0166,\n          0.0048,\n          0.0\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Longitud Promedio Estimada de
Camino\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 2.2518814356000183,\n        \"min\": 2.49,\n        \"max\":
7.52,\n        \"num_unique_values\": 5,\n        \"samples\": [\n
2.77,\n          5.63,\n          2.49\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe","variable_name":"df_topology"}
```


Interpretación de la Huella Topológica

El análisis estructural confirma la existencia de dos tipologías de grafos radicalmente distintas en el conjunto de datos, lo cual tiene implicaciones directas para la arquitectura de los modelos de extrapolación:

1. La Distinción "Mundo Pequeño" vs. "Mundo Largo" La métrica de Longitud de Camino Promedio revela la diferencia más drástica:

   FB13, FB15k-237 y CoDEx-M presentan caminos muy cortos (entre 2.43 y 2.70 saltos). Esto significa que la información fluye muy rápido de un extremo a otro del grafo. Un modelo GNN con solo 2 o 3 capas será suficiente para capturar el contexto global.

WN11 y WN18RR presentan caminos largos (5.79 y 7.40 saltos respectivamente). Esto confirma su estructura profunda y jerárquica.

Implicación: Para extrapolar en WordNet, el modelo necesitará una arquitectura más profunda (más capas) o mecanismos de atención de largo alcance para conectar conceptos distantes en la jerarquía.

2. Densidad Local en FB15k-237 FB15k-237 destaca con el mayor Clustering Local (0.2119). Esto indica que, localmente, el grafo es denso y rico en triángulos (ej. A conoce a B, B conoce a C → A conoce a C).

   Ventaja: Esta estructura favorece enormemente a los modelos basados en subgrafos inductivos, ya que pueden "cerrar triángulos" para predecir enlaces faltantes con alta precisión.

3. La Estructura de Estrella de FB13 A pesar de tener caminos cortos, FB13 tiene una transitividad global de 0.0000 y un clustering local muy bajo (0.0278).

   Análisis: Combinado con el análisis anterior de Grados (donde vimos Super-Hubs), esto sugiere que FB13 tiene una topología de "Estrella": muchos nodos periféricos conectados a un centro, pero desconectados entre sí. Esto es un desafío para la extrapolación, ya que la única forma de inferir relaciones entre nodos periféricos es pasando a través de los hubs genéricos.

4. Integridad de los Grafos Todos los datasets muestran un % en Componente Gigante cercano al 100%. Esto es positivo, ya que garantiza que no hay islas de información aislada que el modelo no pueda alcanzar durante el entrenamiento.

1.5 Evaluación del Escenario de Extrapolación (Sesgo Inductivo)

Conceptos Previos y Justificación

El objetivo central de este proyecto es la extrapolación de conocimiento, definida como la capacidad de un modelo para inferir relaciones sobre entidades que no estuvieron presentes durante la fase de entrenamiento (unseen entities). Sin embargo, la mayoría de los benchmarks estándar en la literatura de Knowledge Graph Embedding (KGE) son inherentemente transductivos; es decir, asumen un "mundo cerrado" donde todas las entidades de prueba ya fueron vistas durante el entrenamiento.

Para determinar si los datasets seleccionados permiten una evaluación directa de la extrapolación o si requieren manipulación experimental, analizamos la conectividad de las tripletas del conjunto de prueba (Test Set) clasificándolas en tres escenarios:

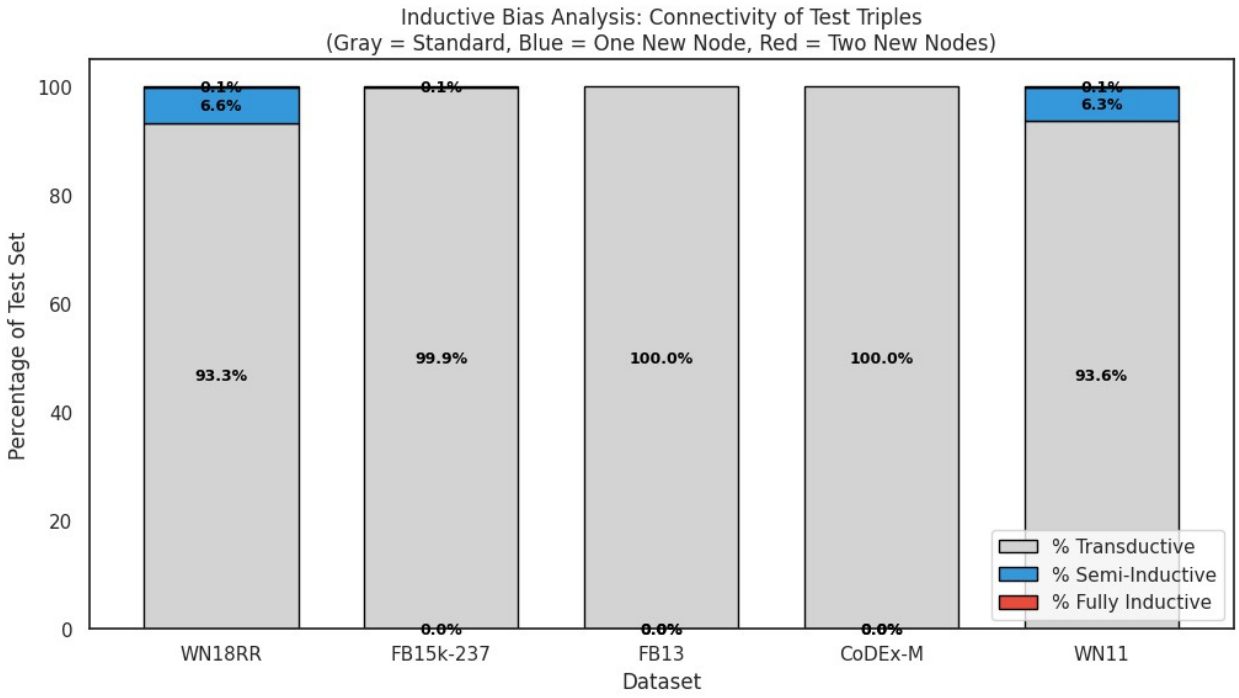
```
Transductivo (Seen-Seen): Ambas entidades (cabeza y cola) son
conocidas. Es el escenario estándar y el más fácil.

Semi-Inductivo (Seen-Unseen): Una entidad es conocida (ancla) y la
```

```
otra es nueva. El modelo puede usar la entidad conocida para "contextualizar" a la nueva.

Totalmente Inductivo (Unseen-Unseen): Ambas entidades son nuevas. Es el escenario más complejo ("a ciegas"), donde el modelo solo puede basarse en la estructura local del grafo o información auxiliar.
```

Expectativa Teórica

Dado que FB15k-237, CoDEx y WN18RR son benchmarks estándar, se espera que el porcentaje de tripletas transductivas sea cercano al 100%. Esto confirmaría la necesidad de aplicar un protocolo de "Inductive Split" (ocultar artificialmente un porcentaje de entidades) para cumplir con los objetivos del proyecto.

```
Checking Inductive Bias for: ['WN18RR', 'FB15k-237', 'FB13', 'CoDEx-M', 'WN11']
```




```
{"summary":"{\n  \"name\": \"    print(\\\"No data found to analyze\",\n  \"rows\": 5,\n  \"fields\": [\n    {\n      \"column\": \"Dataset\",\n      \"properties\": {\n        \"dtype\": \"string\",\n        \"num_unique_values\": 5,\n        \"samples\": [\n          \"FB15k-237\",\n          \"WN11\",\n          \"FB13\"\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"Total Test Triples\",\n
```

\"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 16829,\n   \"min\": 3134,\n   \"max\": 47466,\n   \"num_unique_values\": 5,\n   \"samples\": [\n   20466,\n   21088,\n   47466\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"% Transductive\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 3.5711342735887186,\n   \"min\": 93.3,\n   \"max\": 100.0,\n   \"num_unique_values\": 4,\n   \"samples\": [\n   99.9,\n   93.6,\n   93.3\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"% Semi-Inductive\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 3.5163901945034484,\n   \"min\": 0.0,\n   \"max\": 6.6,\n   \"num_unique_values\": 4,\n   \"samples\": [\n   0.1,\n   6.3,\n   6.6\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"% Fully Inductive\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 0.05477225575051661,\n   \"min\": 0.0,\n   \"max\": 0.1,\n   \"num_unique_values\": 2,\n   \"samples\": [\n   0.0,\n   0.1\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"% of Test Entities that are New\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 1.7282939564784692,\n   \"min\": 0.0,\n   \"max\": 3.9,\n   \"num_unique_values\": 4,\n   \"samples\": [\n   0.3,\n   2.2\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   }\n]\n}","type":"dataframe"}

Interpretación de la Capacidad Inductiva Nativa

El análisis de la composición de los conjuntos de prueba revela la naturaleza predominantemente transductiva de los datos, validando la necesidad de una metodología de particionamiento específica:

1. Naturaleza de "Mundo Cerrado" en CoDEx-M y FB13 Los resultados muestran que CoDEx-M y FB13 son 100.0% transductivos. No existe una sola entidad nueva en el conjunto de prueba (0.0%).

   Conclusión Crítica: Estos datasets, en su forma original (vanilla), no sirven para medir la extrapolación directa. Para utilizarlos en este proyecto, será obligatorio aplicar una metodología de partición inductiva (como la propuesta de otros datasets como GraIL), donde se oculten deliberadamente subgrafos completos durante el entrenamiento.

2. El Ruido Inductivo en WordNet (WN11/WN18RR) Se observa una anomalía interesante en los datasets basados en WordNet. Tanto WN11 como WN18RR presentan un pequeño porcentaje de escenarios semi-inductivos (6.3% y 6.6% respectivamente) y un porcentaje marginal de entidades nuevas (2-4%).

Análisis: Esto sugiere que las versiones utilizadas contienen ligeras inconsistencias o "ruido" respecto a los splits oficiales transductivos. Aunque presentan casos inductivos, el 93% sigue siendo transductivo, por lo que tampoco son suficientes por sí solos para una evaluación rigurosa de extrapolación sin re-particionamiento. Interpretación y Justificación Metodológica

El análisis de sesgo inductivo revela una discrepancia fundamental entre la naturaleza de los datasets estándar y los objetivos de extrapolación de este proyecto. Los hallazgos validan la necesidad de intervenir los datos siguiendo protocolos específicos de la literatura.

1. Predominio del Escenario Transductivo ("Mundo Cerrado") Las gráficas evidencian que los benchmarks en su formato original (vanilla) están diseñados para evaluar memoria y consistencia, no extrapolación.

   CoDEx-M y FB13 son 100% Transductivos (Barras totalmente grises). No existe una sola entidad en el conjunto de prueba que no haya sido vista durante el entrenamiento.

   WN11 y WN18RR son 93% Transductivos. Aunque presentan una pequeña fracción de entidades nuevas (6% escenarios semi-inductivos), la inmensa mayoría de los datos sigue asumiendo un conocimiento completo del dominio.

2. Validación del Procedimiento de Hamaguchi et al. (2017) La virtual ausencia de escenarios inductivos naturales (especialmente el escenario Unseen-Unseen, que es casi 0%) confirma que no es viable evaluar la extrapolación utilizando los splits originales.

Estos resultados validan y hacen imperativo el uso de la metodología experimental propuesta por Hamaguchi et al. (2017) en "Knowledge Transfer for Out-of-Knowledge-Base Entities: A Graph Neural Network Approach". Dado que los datasets no ofrecen suficientes ejemplos OOKB (Out-of-Knowledge-Base) de forma nativa, el proyecto deberá replicar su procedimiento de generación de datos, el cual consiste en:

```
Muestreo Selectivo: No usar el test set completo, sino seleccionar
subconjuntos controlados de tripletas (ej.N=1,000, 3,000, 5,000).

Forzado de Entidades OOKB: Modificar los splits originales para
garantizar la presencia de entidades desconocidas en tres
configuraciones específicas:

    Setting Head: Ocultar del entrenamiento las entidades que aparecen
como cabecera en el test.

    Setting Tail: Ocultar las entidades que aparecen como cola.

    Setting Both: Ocultar ambas.
```

Conclusión Final El análisis exploratorio demuestra que la extrapolación de conocimiento no puede medirse pasivamente en estos benchmarks estándar. Por tanto, el éxito del proyecto

depende de la correcta implementación del protocolo de Hamaguchi sobre el dataset WN11 (y extendido a los demás), transformando un problema mayoritariamente transductivo (gris) en uno controlado experimentalmente para medir la generalización en nodos no vistos.

Found datasets: ['WN18RR', 'FB15k-237', 'FB13', 'CoDEx-M', 'WN11']

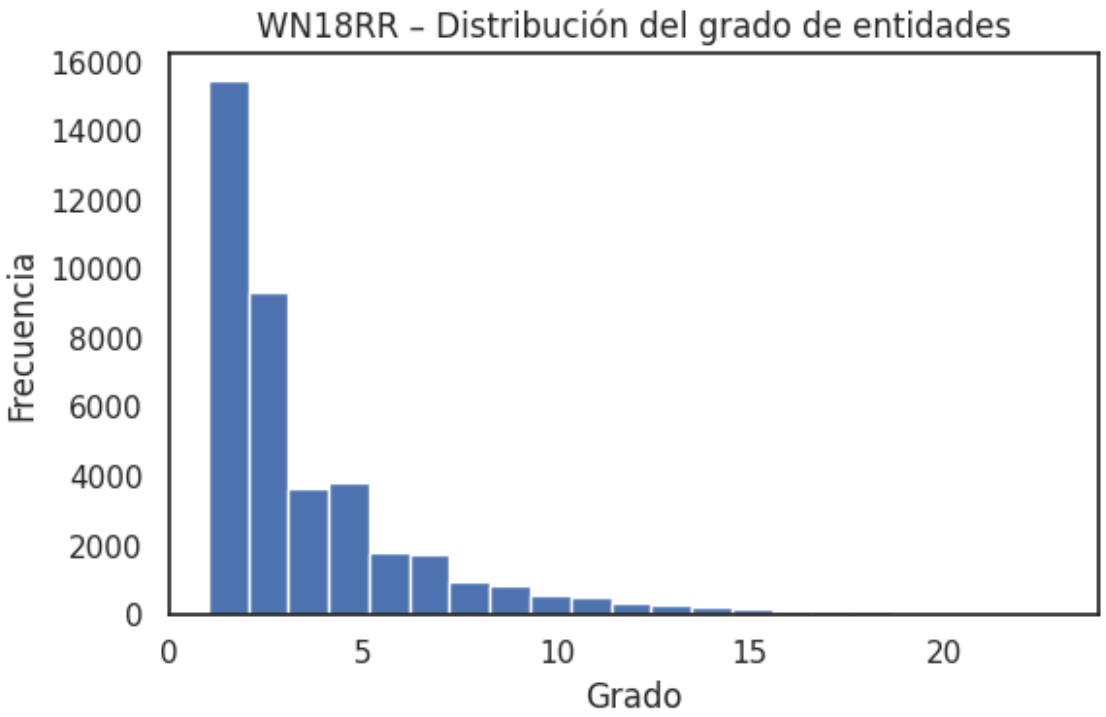
2. Análisis univariante

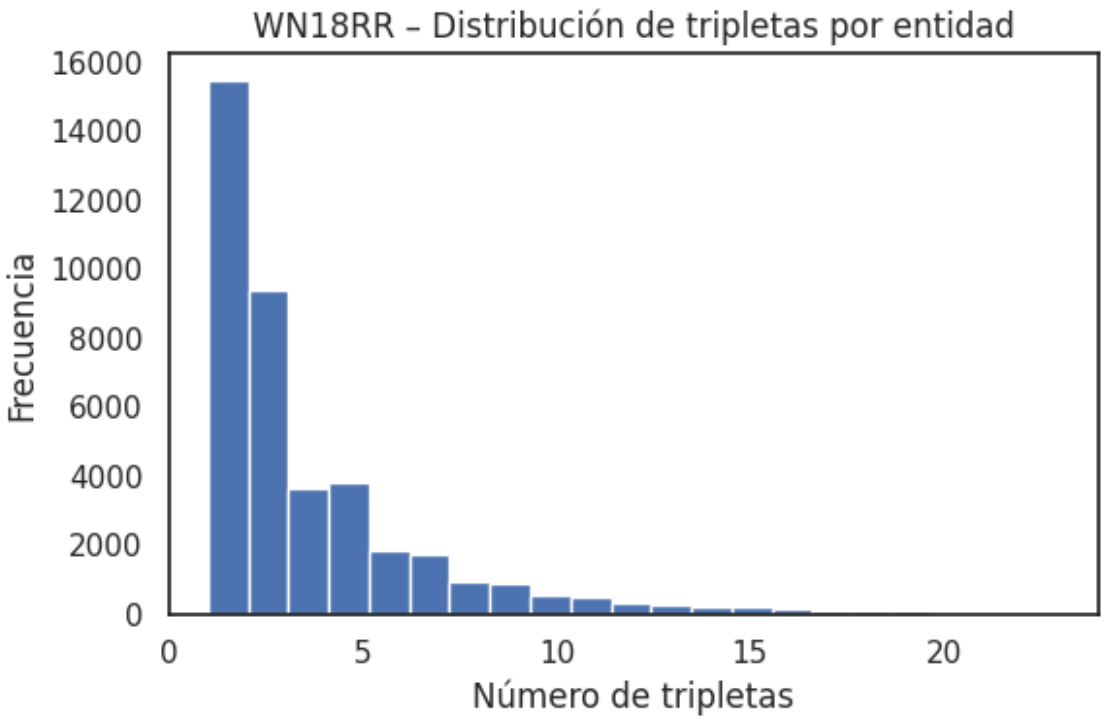
Dado que los datos corresponden a grafos de conocimiento representados mediante tripletas (h , r , t) (h,r,t), las variables originales son categóricas y no contienen atributos numéricos continuos. Por ello, el análisis univariante se realiza sobre métricas estructurales derivadas del grafo, en lugar de sobre los identificadores de entidades o relaciones.

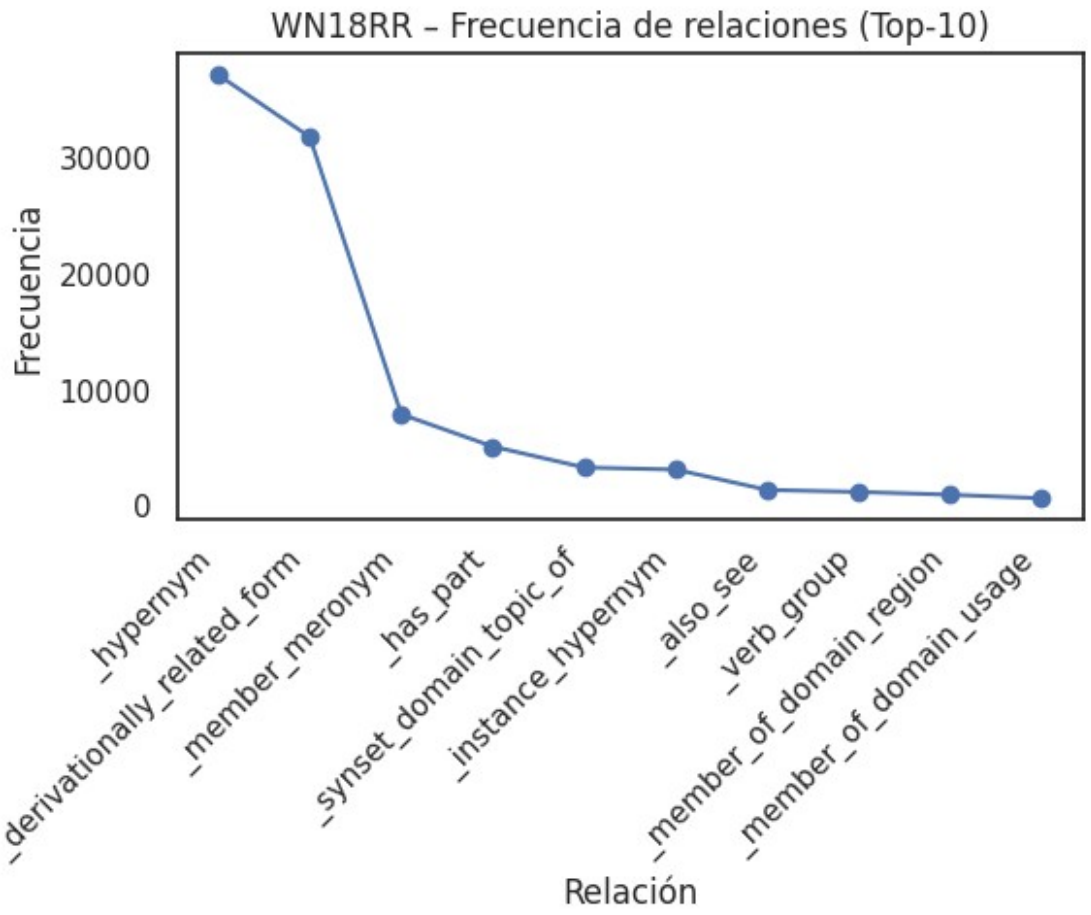
En esta etapa se analizan la distribución del grado de las entidades, la frecuencia de aparición de entidades en las tripletas y la frecuencia de los distintos tipos de relación. Asimismo, se identifican las entidades con mayor grado (hubs), ya que concentran una parte significativa de las conexiones del grafo y pueden influir en el comportamiento de los modelos de aprendizaje. Las visualizaciones empleadas consisten en histogramas y gráficos de barras, adecuados para métricas discretas y distribuciones con cola larga. No se utilizan boxplots, ya que no existen variables numéricas continuas y este tipo de representación no aporta información relevante en este contexto.

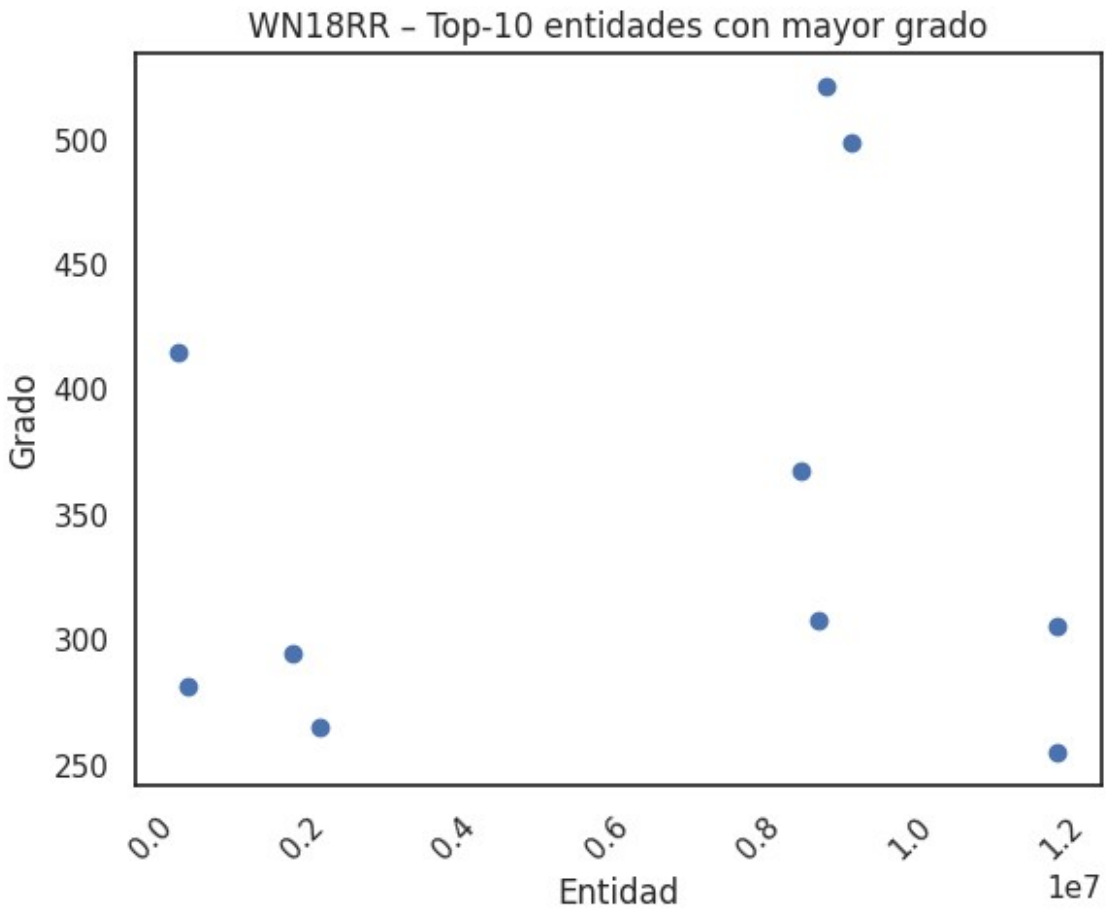
Dado que las variables del grafo son categóricas y no contienen atributos numéricos continuos, no se emplean boxplots. En su lugar, se utilizan histogramas y gráficos de barras sobre métricas estructurales (grado, frecuencia de relaciones), que son las representaciones adecuadas para este tipo de datos.

```
Datasets:
['WN18RR', 'FB15k-237', 'FB13', 'CoDEx-M', 'WN11']

univariate_eda_kg(FINAL_DATA_DIR, dataset_name="WN18RR", top_n=10)
```

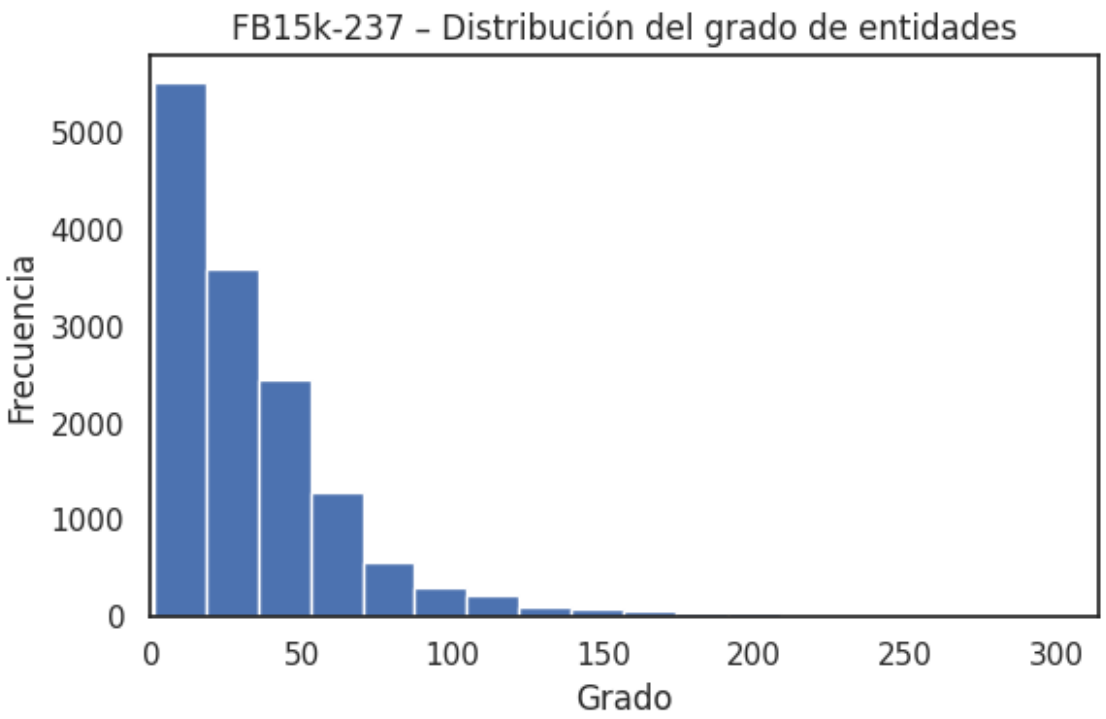

WN18RR – Distribución del grado de entidades
Frecuencia
16000
14000
12000
10000
8000
6000
4000
2000
0
0
5
10
15
20
Grado

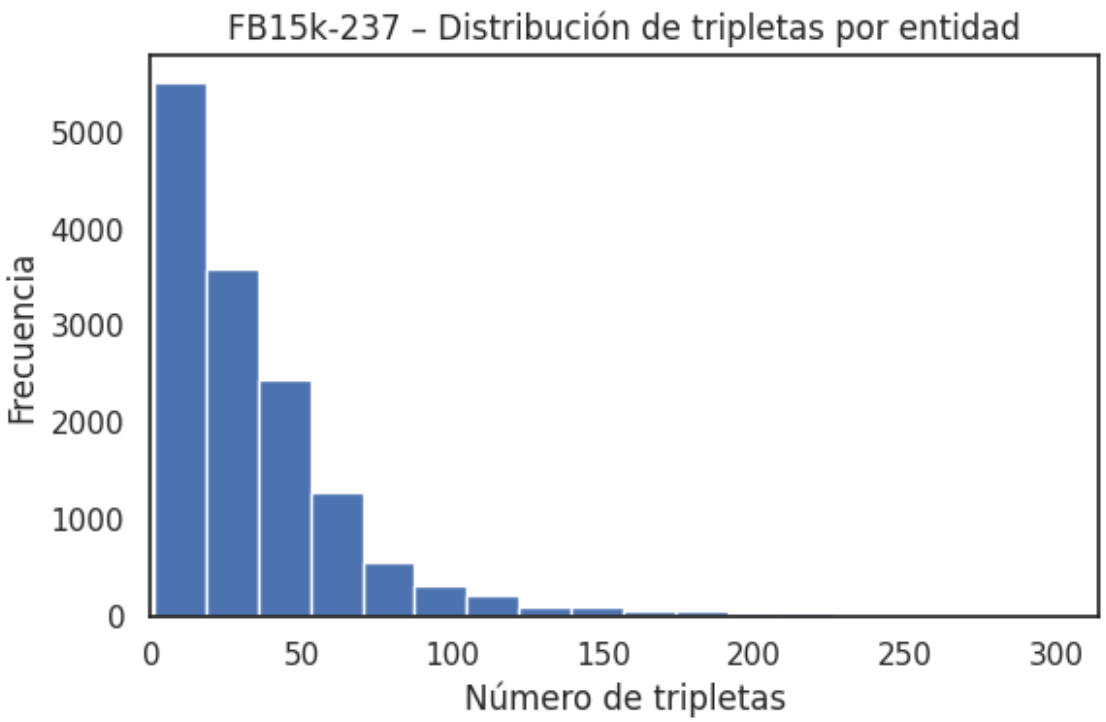

WN18RR – Distribución de tripletas por entidad
Frecuencia
16000
14000
12000
10000
8000
6000
4000
2000
0
0
5
10
15
20
Número de tripletas


WN18RR – Frecuencia de relaciones (Top-10)
Frecuencia
30000
20000
10000
0
_hypernym
_derivationally_related_form
_member_meronym
_has_part
_synset_domain_topic_of
_instance_hypernym
_also_see
_verb_group
_member_of_domain_region
_member_of_domain_usage
Relación

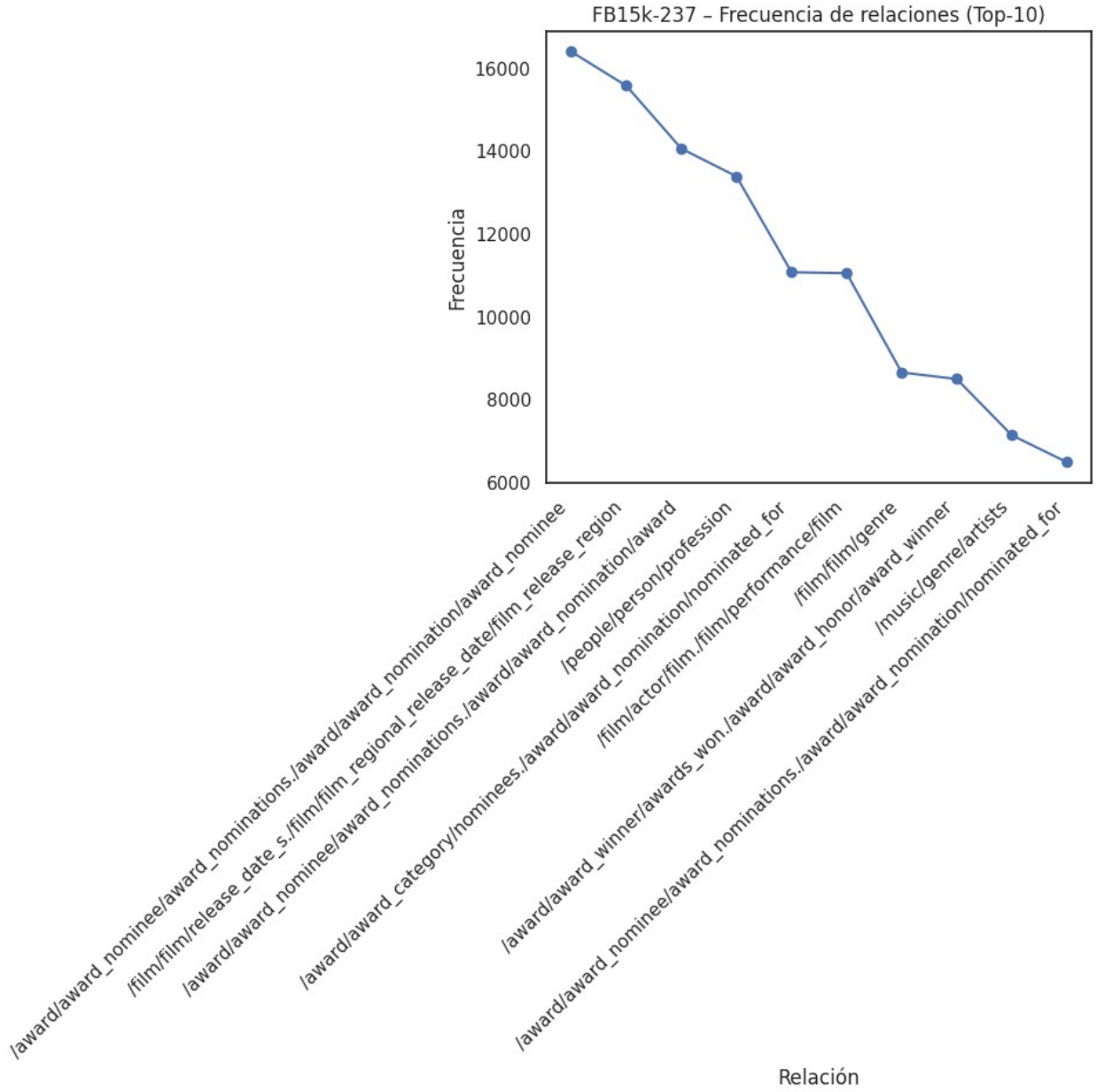


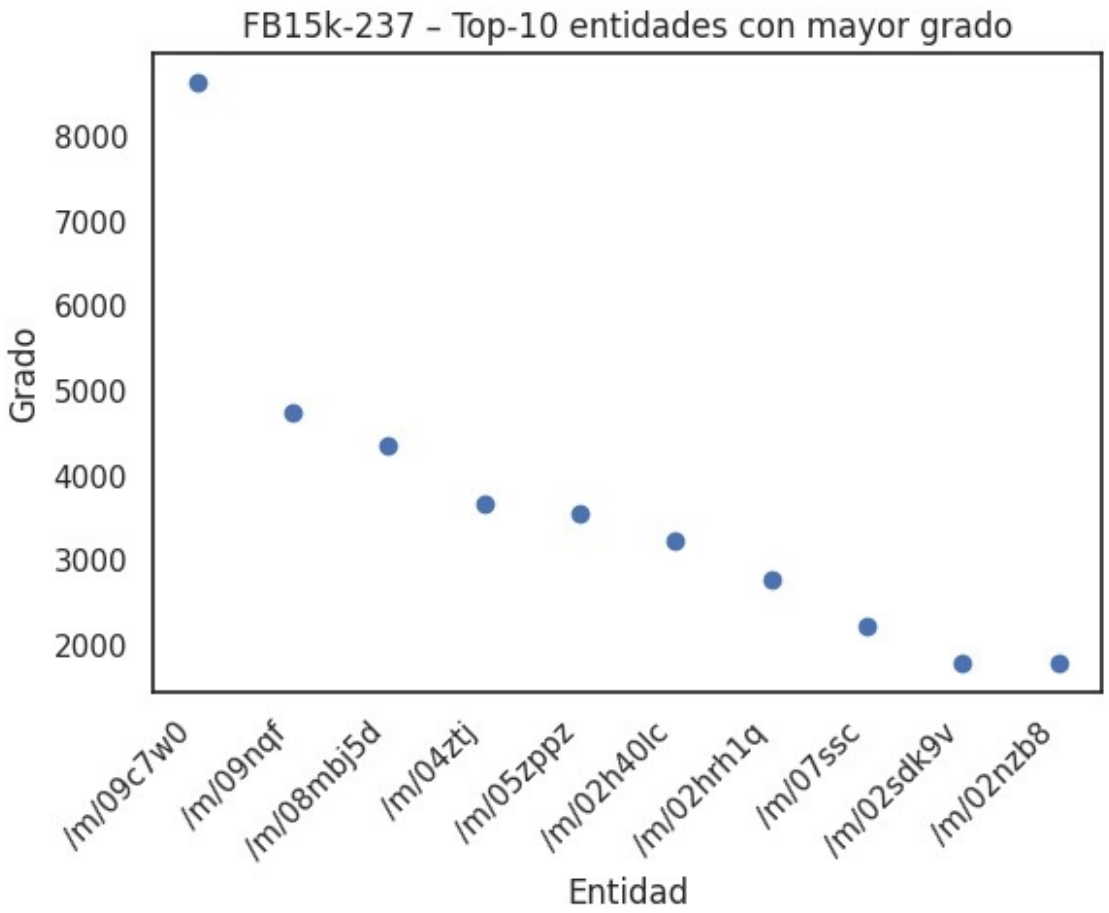

```
univariate_eda_kg(FINAL_DATA_DIR, dataset_name="FB15k-237", top_n=10)
```


FB15k-237 – Distribución del grado de entidades
Frecuencia
5000
4000
3000
2000
1000
0
0
50
100
150
200
250
300
Grado

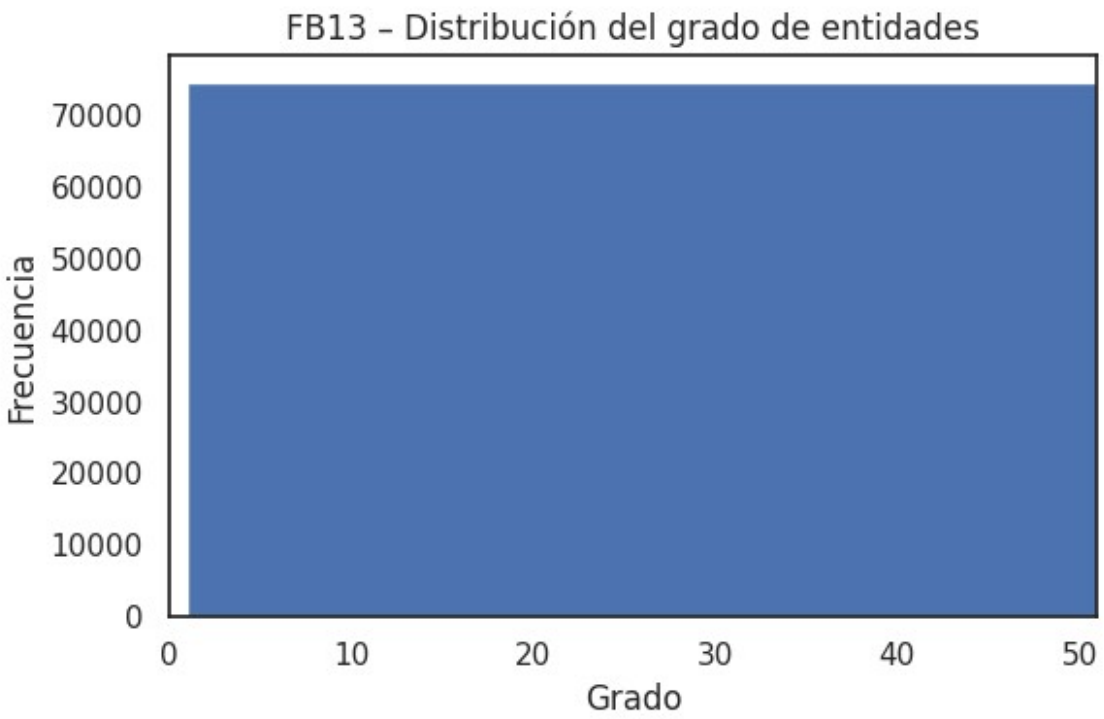


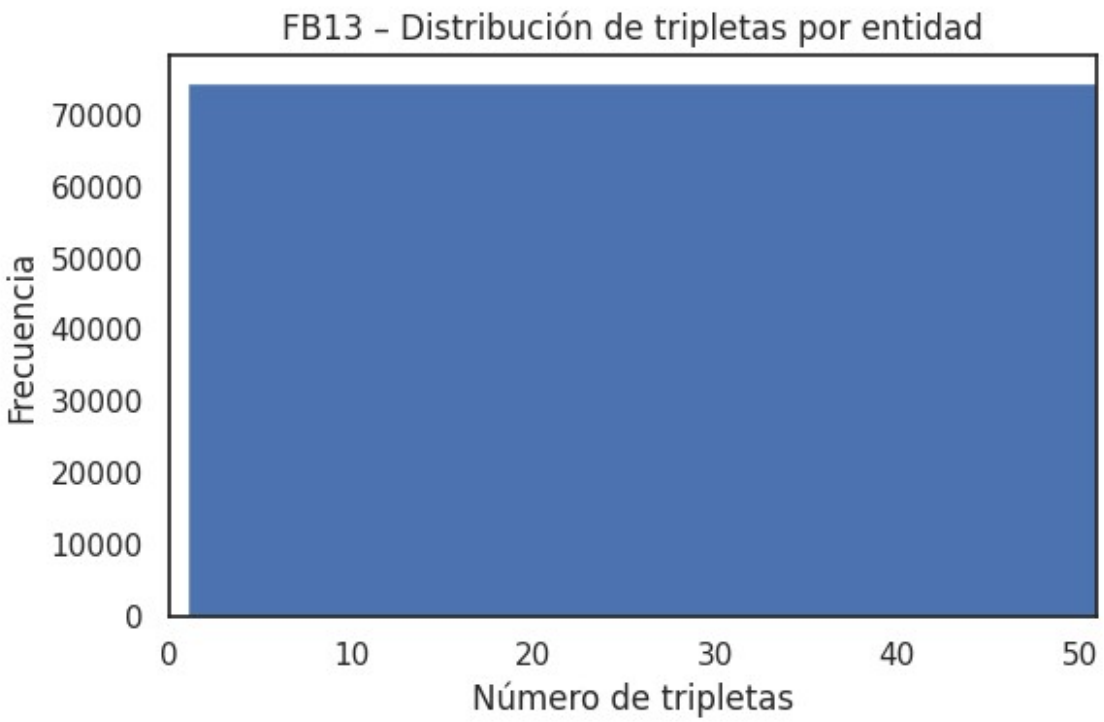

```
/tmp/ipython-input-2809295182.py:95: UserWarning: Tight layout not
applied. The bottom and top margins cannot be made large enough to
accommodate all Axes decorations.
  plt.tight_layout()
```

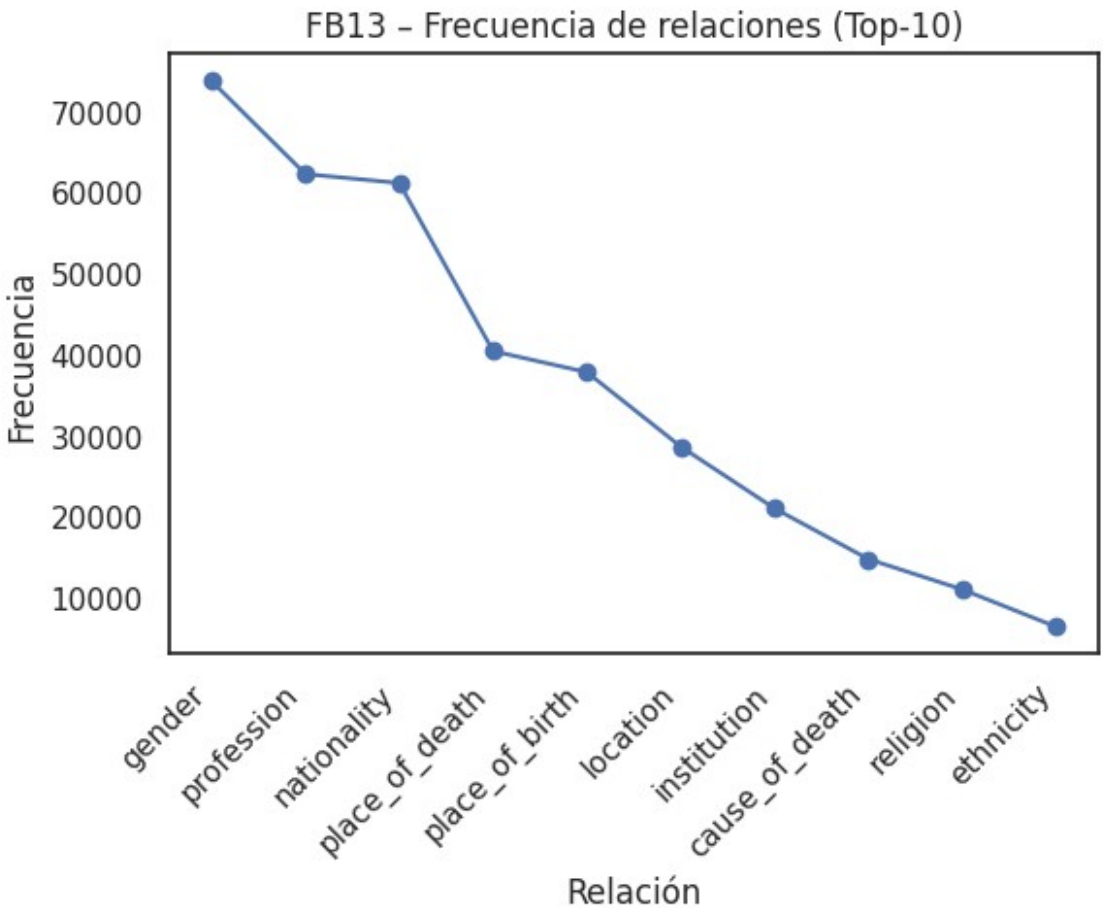

FB15k-237 – Frecuencia de relaciones (Top-10)
Frecuencia
16000
14000
12000
10000
8000
6000
/award/award_nominee/award_nominations./award/award_nomination/award_nominee
/film/film/release_date_s./film/film_regional_release_date/film_release_region
/award/award_nominee/award_nominations./award/award_nomination/award
/people/person/profession
/award/award_category/nominees./award/award_nomination/nominated_for
/film/actor/film./film/performance/film
/film/film/genre
/award/award_winner/awards_won./award/award_honor/award_winner
/music/genre/artists
/award/award_nominee/award_nominations./award/award_nomination/nominated_for
Relación

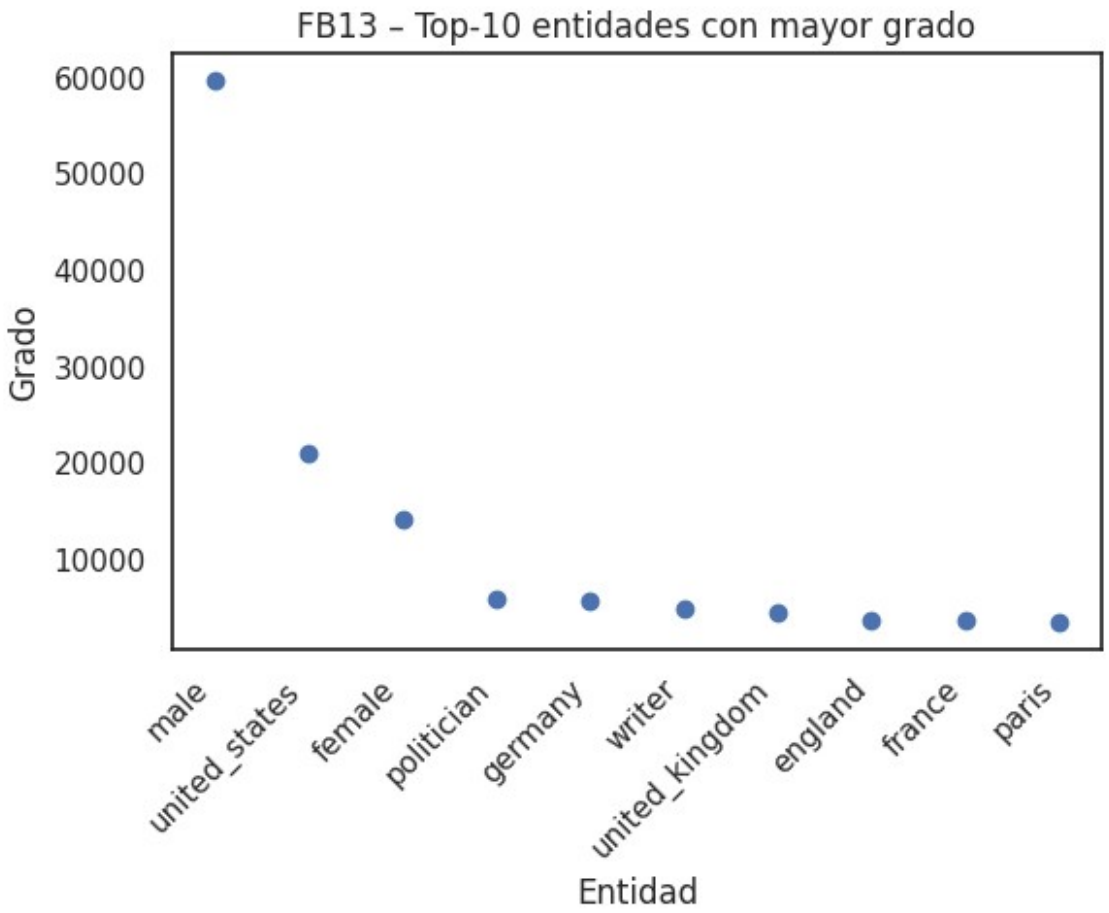



```
univariate_eda_kg(FINAL_DATA_DIR, dataset_name="FB13", top_n=10)
```

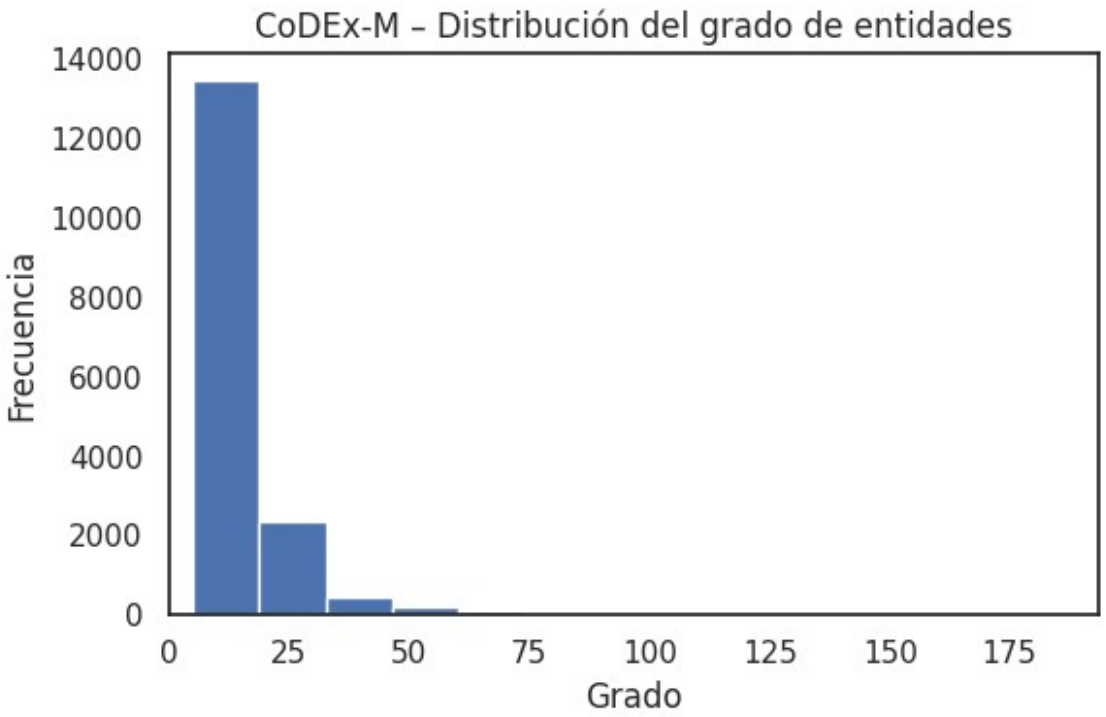

FB13 – Distribución del grado de entidades
Frecuencia
70000
60000
50000
40000
30000
20000
10000
0
0
10
20
30
40
50
Grado

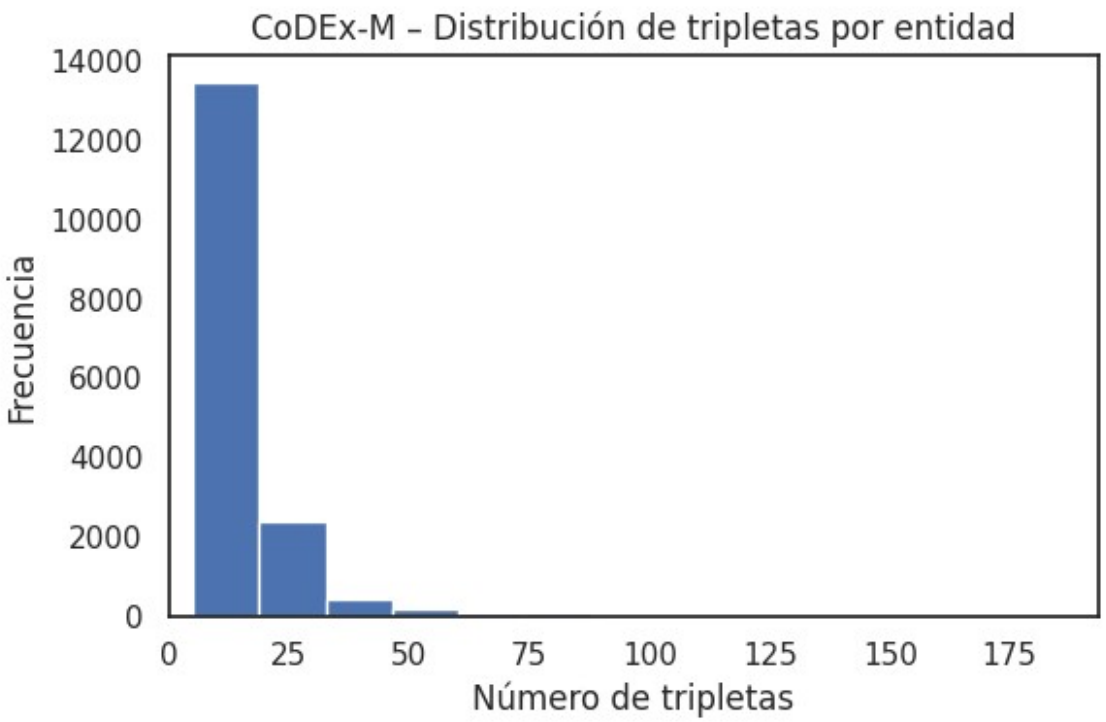

FB13 – Distribución de tripletas por entidad
Frecuencia
70000
60000
50000
40000
30000
20000
10000
0
0
10
20
30
40
50
Número de tripletas

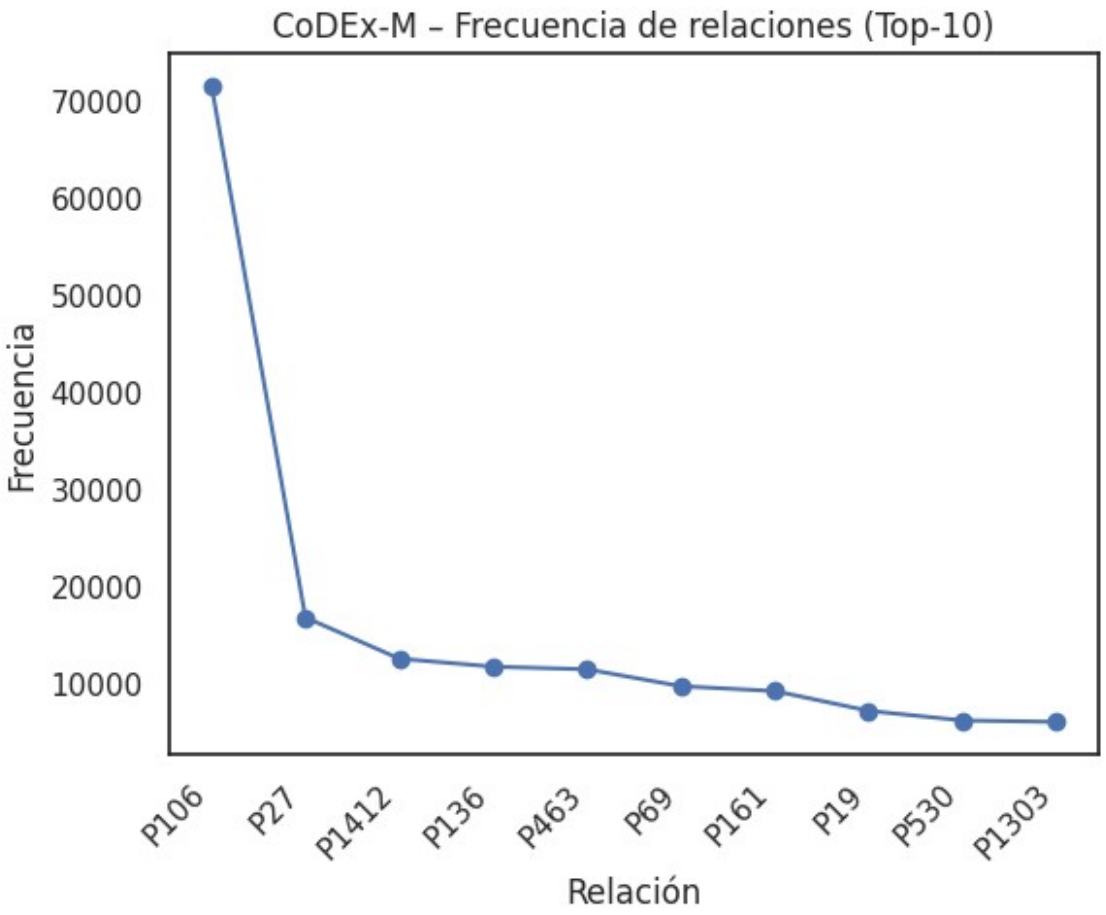

FB13 – Frecuencia de relaciones (Top-10)
Frecuencia
70000
60000
50000
40000
30000
20000
10000
gender
profession
nationality
place_of_death
place_of_birth
location
institution
cause_of_death
religion
ethnicity
Relación

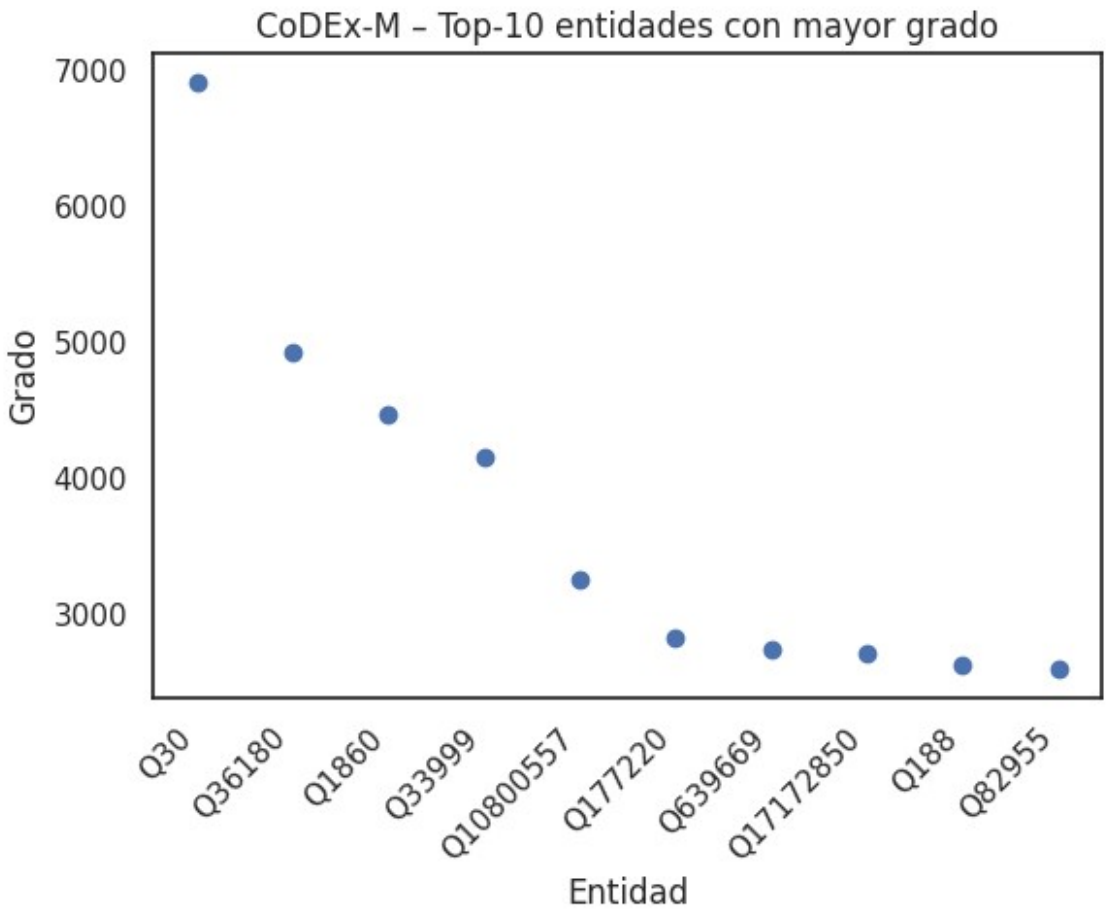



```
univariate_eda_kg(FINAL_DATA_DIR, dataset_name="CoDEx-M", top_n=10)
```

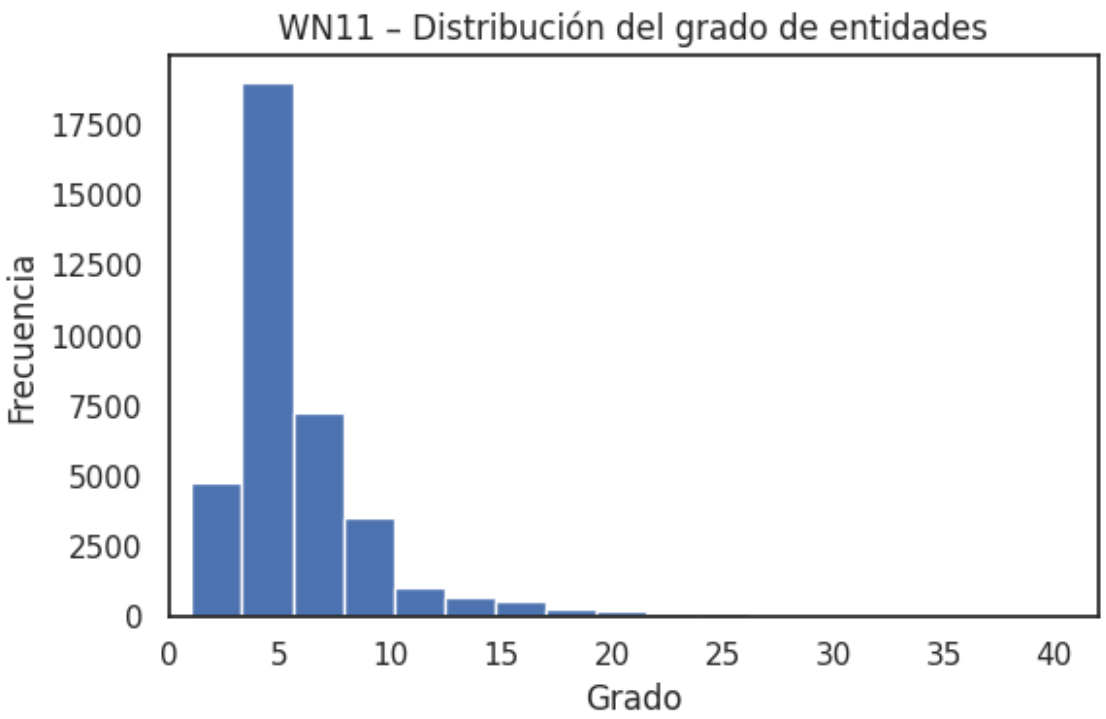

CoDEx-M – Distribución del grado de entidades
Frecuencia
14000
12000
10000
8000
6000
4000
2000
0
0
25
50
75
100
125
150
175
Grado

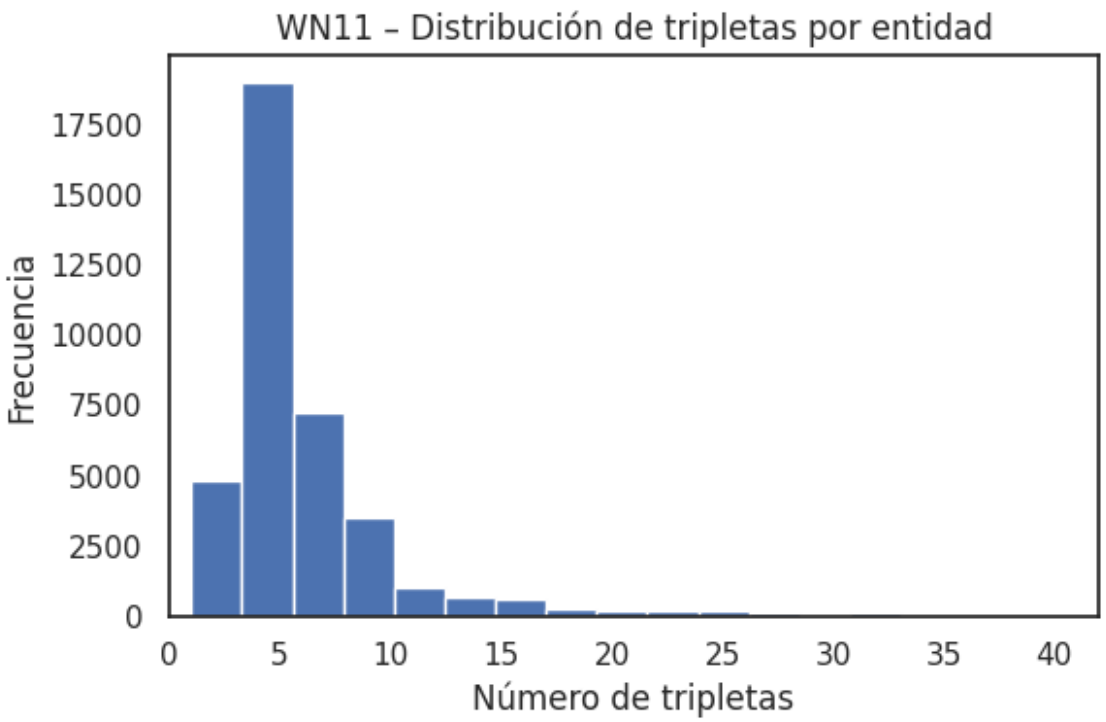

CoDEx-M – Distribución de tripletas por entidad
Frecuencia
14000
12000
10000
8000
6000
4000
2000
0
0
25
50
75
100
125
150
175
Número de tripletas

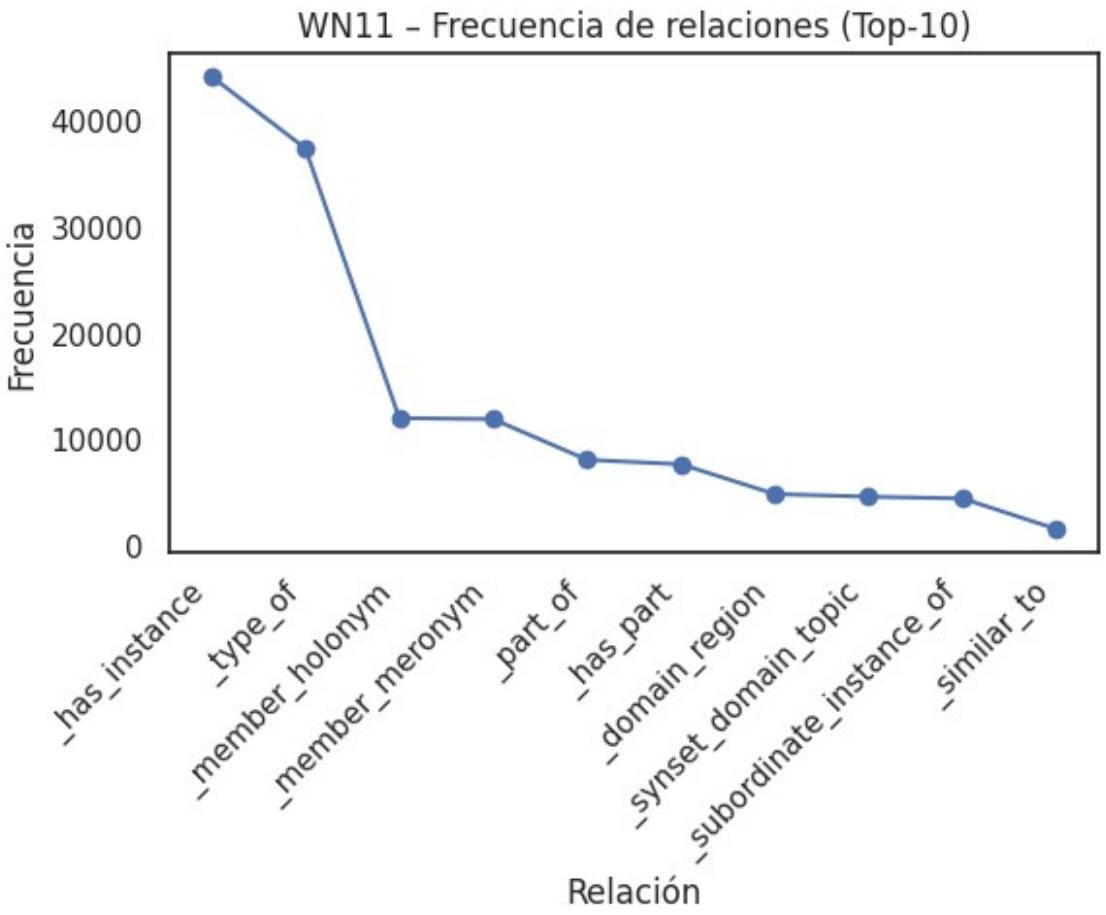

CoDEx-M – Frecuencia de relaciones (Top-10)
Frecuencia
70000
60000
50000
40000
30000
20000
10000
P106
P27
P1412
P136
P463
P69
P161
P19
P530
P1303
Relación

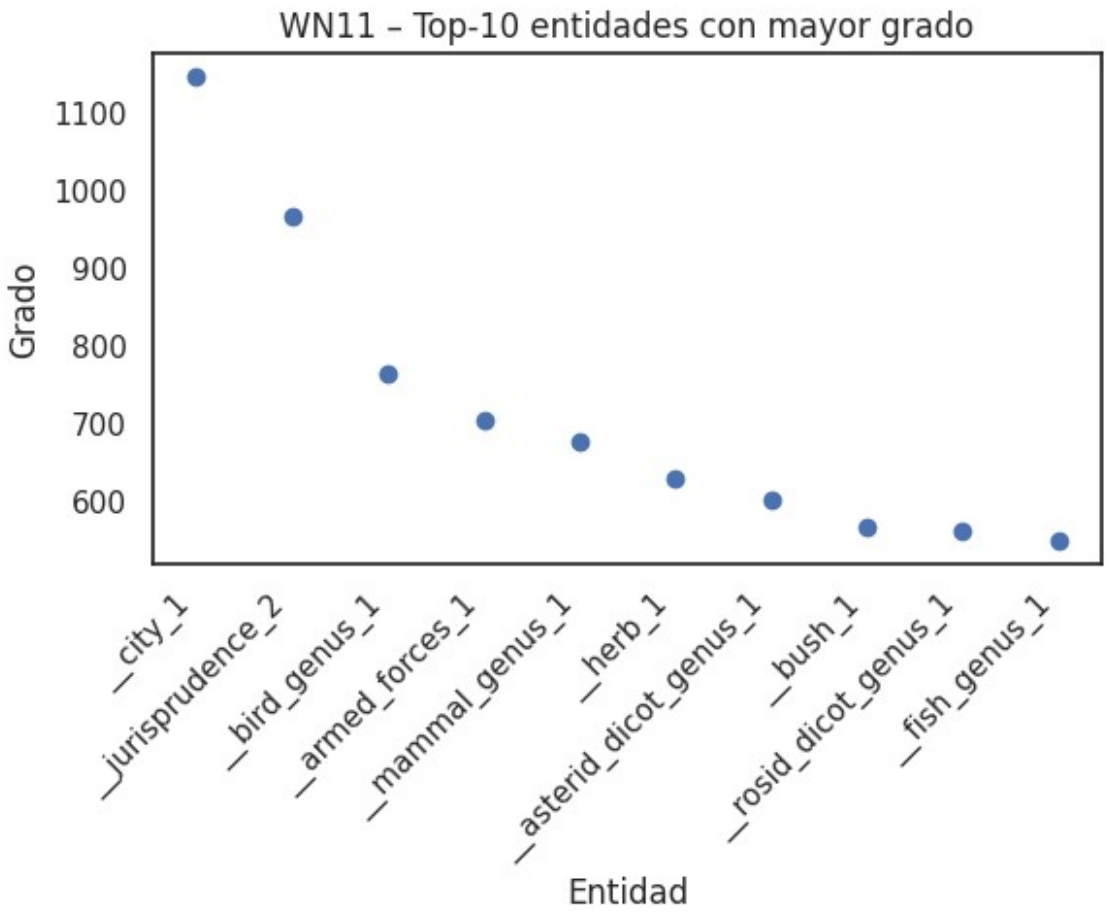



```
univariate_eda_kg(FINAL_DATA_DIR, dataset_name="WN11", top_n=10)
```


WN11 – Distribución del grado de entidades
Frecuencia
17500
15000
12500
10000
7500
5000
2500
0
0
5
10
15
20
25
30
35
40
Grado


WN11 – Distribución de tripletas por entidad
Frecuencia
17500
15000
12500
10000
7500
5000
2500
0
0
5
10
15
20
25
30
35
40
Número de tripletas


WN11 – Frecuencia de relaciones (Top-10)
Frecuencia
40000
30000
20000
10000
0
_has_instance
_type_of
_member_holonym
_member_meronym
_part_of
_has_part
_domain_region
_synset_domain_topic
_subordinate_instance_of
_similar_to
Relación



Análisis bi/multivariante

El análisis bivariante y multivariante se plantea tradicionalmente para explorar relaciones de dependencia o correlación entre variables numéricas. Sin embargo, en el presente trabajo los datos corresponden a grafos de conocimiento representados mediante tripletas (h , r , t) (h,r,t), donde las variables originales son de naturaleza categórica y no existen atributos numéricos continuos independientes.

Si bien es posible definir métricas estructurales derivadas del grafo (como el grado de las entidades o la frecuencia de las relaciones), estas métricas no constituyen variables independientes en el sentido estadístico, ya que están directamente determinadas por la estructura del propio grafo. En consecuencia, la aplicación de análisis de correlación clásicos (por ejemplo, coeficientes de Pearson o Spearman) no resulta metodológicamente apropiada ni aporta información adicional significativa.

Por esta razón, el análisis exploratorio se centra en la caracterización univariante de las principales métricas estructurales del grafo, dejando el estudio de interacciones más complejas para etapas posteriores relacionadas con el modelado y la evaluación del desempeño de los algoritmos de aprendizaje.

Preprocesamiento

En el contexto de grafos de conocimiento representados mediante tripletas (h , r , t) (h,r,t), las estrategias clásicas de preprocesamiento asociadas al manejo de valores faltantes, atípicos y reducción de cardinalidad no aplican de la misma manera que en conjuntos de datos tabulares. Las tripletas no contienen atributos numéricos continuos, por lo que no existen valores faltantes en el sentido tradicional, sino únicamente ausencia de conocimiento, la cual constituye parte fundamental del problema a resolver.

De forma similar, las entidades o relaciones con baja frecuencia de aparición no se consideran valores atípicos, sino elementos válidos que reflejan la naturaleza dispersa y de cola larga característica de los grafos de conocimiento reales. Su eliminación o modificación introduciría sesgos estructurales y afectaría negativamente la capacidad del modelo para generalizar a entidades o relaciones no vistas durante el entrenamiento.

Finalmente, aunque los conjuntos de datos presentan una alta cardinalidad en entidades y relaciones, esta característica no se aborda mediante técnicas de reducción en la etapa de preprocesamiento. En su lugar, se gestiona posteriormente a través de representaciones embebidas aprendidas por el modelo, preservando la riqueza estructural del grafo. Por estas razones, no se aplican transformaciones de preprocesamiento clásicas, manteniendo los datos en su forma original.

Conclusiones

El análisis exploratorio permitió caracterizar la estructura fundamental de los grafos de conocimiento analizados a partir de métricas derivadas, dado que los datos originales se representan mediante tripletas categóricas. Las estadísticas globales y el análisis univariante evidencian una marcada heterogeneidad en la distribución de entidades y relaciones, así como la presencia de distribuciones de cola larga tanto en el grado de las entidades como en la frecuencia de las relaciones. Este comportamiento es consistente con grafos de conocimiento reales y pone de manifiesto la existencia de nodos altamente conectados junto con una gran cantidad de entidades poco frecuentes.

El análisis integral de la topología y distribución de los grafos (Puntos 2, 3 y 4) revela una profunda heterogeneidad estructural que garantiza una evaluación robusta de la propuesta de investigación. Se ha evidenciado una dicotomía clara: por un lado, datasets como WN18RR presentan estructuras jerárquicas dispersas ("árboles" con caminos largos y un 37% de nodos con escasa conectividad), lo que desafía la capacidad de los modelos para agregar información vecinal. Por otro lado, FB15k-237 y CoDEx-M exhiben características de "mundo pequeño" con alta densidad y un marcado desbalance relacional (Coeficiente de Gini > 0.6), lo que impone el reto de evitar el sobreajuste hacia nodos hubs y relaciones mayoritarias. Esta diversidad confirma que el conjunto de datos seleccionado cubre un espectro lo suficientemente amplio para validar la generalización del modelo ante distintas arquitecturas de información.

Sin embargo, el hallazgo más crítico se desprende del análisis de sesgo inductivo (Punto 5), el cual valida la necesidad imperativa de intervenir los datos para cumplir con los objetivos del proyecto. Al demostrarse que los benchmarks estándar operan bajo una suposición de "mundo cerrado" (con una transductividad superior al 93% y ausencia de escenarios Unseen-Unseen), se

concluye que la extrapolación de conocimiento no puede medirse utilizando los conjuntos de prueba originales. Esto justifica y fundamenta técnicamente la adopción del protocolo experimental de Hamaguchi et al. (2017) descrito en la metodología, siendo indispensable la generación artificial de particiones OOKB (Out-of-Knowledge-Base) para transformar estos recursos estáticos en escenarios dinámicos que realmente evalúen la inferencia sobre entidades nuevas.

Asimismo, el EDA confirmó que la ausencia de ciertas tripletas y la baja frecuencia de algunas entidades o relaciones no constituyen anomalías, sino propiedades intrínsecas del problema de inferencia en grafos de conocimiento. En consecuencia, no se aplicaron estrategias clásicas de limpieza o reducción de datos, ya que estas podrían introducir sesgos y afectar la capacidad de generalización del modelo. Los resultados de esta etapa proporcionan una base sólida para el diseño y entrenamiento de modelos de aprendizaje capaces de manejar alta cardinalidad, dispersión estructural y escenarios de inferencia con información incompleta.