Fórmula 1: un análisis a la historia con Neo4j

Dara Silvera
Facultad de Ingeniería, Universidad de la República
Montevideo, Uruguay
dara.silvera@fing.edu.uy

Santiago Rodriguez
Facultad de Ingeniería, Universidad de la República
Montevideo, Uruguay
santiago.rodriguez.barthe@fing.edu.uy

***Resumen*—La Fórmula 1 es uno de los deportes más populares del mundo. Su rica historia consta de años de pilotos de todas partes del mundo compitiendo en multitud de carreras en los distintos continentes. En este trabajo se buscó el modelado y armado de una base de datos de grafos usando *Neo4j* que contenga los datos históricos de la Fórmula 1. Además, se realizaron experimentos como distintas consultas para acceder a esta información, búsqueda de caminos entre pilotos o distintas visualizaciones de redes de grafos armados en base a la base de datos.**

I. Introducción

La Fórmula 1 es uno de los deportes automovilísticos de más alto nivel y prestigio en el mundo, atrayendo a millones de fanáticos cada carrera tanto presencialmente como audiencia de televisión y digital [Fórmula 1 2022]. Los autos en este deporte son de los más rápidos del mundo, alcanzando velocidades de 350 km/h. El campeonato mundial, celebrado en cada temporada, consiste de una serie de carreras denominadas *Grand Prix*, disputadas en circuitos específicos a lo largo del mundo. Cada carrera otorga una determinada cantidad de puntos al piloto y a su respectiva escudería, según la posición en la que éste finalice, los cuales sumarán al final para el campeonato mundial de pilotos y constructores, respectivamente. Al finalizar la temporada, el piloto y el constructor serán consagrados campeones de estos campeonatos. Los comienzos de este deporte se remontan al 1950, cuando se organizó por primera vez el campeonato mundial de pilotos, que constaba de siete carreras oficiales. Desde entonces, 72 campeonatos se han disputado cada año. En la actualidad, un total de diez escuderías y veinte pilotos (dos por cada escudería) pelean por los títulos año a año.

En este trabajo, se creó una base de datos de grafos que contenga información acerca de todos los pilotos, las escuderías, las carreras, los circuitos y las temporadas desde los inicios de este deporte hasta la actualidad. Luego, se realizaron consultas a la base de datos con la intención de estudiar el poder de la herramienta *Neo4j* para la obtención de datos.

El artículo se organiza de la siguiente manera. En la sección II se presentan trabajos relacionados al análisis de datos sobre el mundo de la Fórmula 1, así como también otros estudios usando bases de datos de grafos en otros deportes. La sección III presenta el proceso de obtención de los datos históricos de la Fórmula 1 y el posterior modelado y carga de la base de datos. La sección IV detalla los diferentes experimentos realizados sobre la base de datos, que van desde diversas consultas hasta distintas proyecciones sobre el grafo y estudios de modularidad y centralidad sobre los mismos. Por último, en la sección V se destacan algunas conclusiones y se señalan posibles trabajos futuros que se podrían hacer sobre la base de datos.

II. Trabajos relacionados

Se hizo una búsqueda de trabajos relacionados acerca de la manipulación y análisis de datos de la Fórmula 1, ya sea usando una base de datos de grafos o otras alternativas. Además, no se limitó la búsqueda solamente a la Fórmula 1 si no que también se consideraron trabajos que usaron bases de datos de grafos para analizar datos relacionados a otros deportes. A continuación de se listará una serie de artículos que tuvieron una gran influencia en este trabajo.
Por el lado de la Fórmula 1, no se encontraron prácticamente artículos que usaran bases de datos de grafos para analizar los datos. En [Abdon] se describe el modelado de una base de datos orientada a grafos de la temporada 2013 de la Fórmula 1 usando Neo4j, así como también detalla consultas de ejemplo que se pueden realizar sobre la misma. Por otro lado, en [Serapiglia2018] se detalla el procedimiento de diseño y carga de una base de datos relacional para los datos de una temporada de la Fórmula 1, aunque sin entrar en detalles.
Sin embargo, hay bastantes trabajos relacionados a otros deportes que se centran en el análisis de datos usando grafos. Por ejemplo, en [Schnaiderman and Mignaco2021] los autores lograron, sobre una base de datos de grafos de equipos y jugadores de fútbol, hacer un análisis de caminos entre nodos y también estudiaron la detección de comunidades usando múltiples algoritmos como el método de Louvain [Neo4jb], el algoritmo de detección de componentes débilmente conexas [Neo4jd] y el algoritmo de conteo de triángulos [Neo4jc]. Además, los trabajos realizados en [Park and Yilmaz2010] y [Clemente et al.2015] describen el uso de algoritmos sobre grafos que describen las interacciones de los jugadores de un equipo de fútbol a lo largo de un partido, para analizar la correspondencia entre los valores de modularidad y la centralidad del grafo y el rendimiento del equipo. De la misma forma, en [Bourbousson et al.2010] se utilizan grafos para analizar cómo se conectan e interactúan jugadores de basketbol de un equipo durante un partido.

III. Desarrollo

III-A. Obtención de datos

Para obtener los datos de toda la historia de la Formula 1, se utilizó un conjunto de datos publicado en *Kaggle* [Vopani2022]. Este conjunto de datos está organizado en varios archivos de formato *cvs* y contiene información de las carreras, pilotos, constructores, campeonatos y más, que abarca desde los inicios del deporte hasta la actualidad.

III-B. Diseño y cargado de datos

La Figura 1 ilustra el modelo de realidad diseñado. Este está constituido por siete nodos y nueve relaciones.

- *Driver*: Todos los pilotos de la historia de la Fórmula 1.
- *Constructor*: Todos los constructores de la historia de la Fórmula 1.
- *Season*: Todas las temporadas de la historia de la Fórmula 1.
- *DriversLineUp*: Alineación de pilotos para cada constructor en cada temporada.
- *Race*: Todas las carreras de la historia de la Fórmula 1.
- *Circuit*: Todos los circuitos en los que se corrió una carrera de Fórmula 1.
- *Country*: Países del mundo relevantes a los datos.

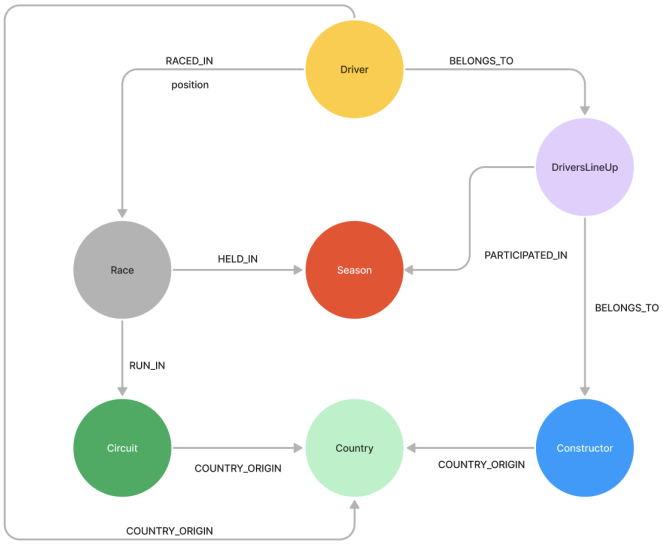

Figura 1: Modelo de la base de datos.

Como se mencionó anteriormente, en cada temporada dos pilotos corren para un equipo constructor. Debido a esto, decidimos crear el nodo *DriversLineUp* que representa la alineación de pilotos para un equipo constructor en una temporada, ya que era de interés saber, dado un piloto, su compañero de equipo en una temporada o a lo largo de su carrera.

Los pilotos corren carreras, las cuales se disputan en distintos circuitos ubicados en distintos países. Por lo general, en un circuito se realiza una carrera por temporada, aunque en ciertas ocasiones ha sucedido que en un mismo circuito se corren varias carreras en una temporada. Debido a esto, se agregó la propiedad *season_round* en la relación *HELD_IN* entre los nodos *Race* y *Season*, utilizada para identificar carreras en un mismo circuito dentro de una temporada.

Además, en la relación entre piloto y carrera, se agregó el atributo *position* para modelar la posición que un piloto termina una carrera.

Por último, los pilotos tienen una nacionalidad, mientras que los circuitos se geolocalizan en un país y los constructores tienen de origen un país. Cabe destacar que en la base de datos obtenida, los pilotos tienen su país representado como nacionalidad. Por lo tanto, se tuvo que hacer un mapeo de las nacionalidades con sus países correspondientes.

La tecnología que se utilizó para el cargado de datos es *Python* y *Neo4j Python Driver*[1]. El código utilizado para cargar la base puede encontrarse en el repositorio de gitlab[2].

IV. Experimentación

En esta sección se presentarán las distintas consultas a modo de experimentación que se hicieron sobre la base de datos, detallando las implementaciones de las mismas y los resultados obtenidos. Las consultas fueron realizadas utilizando la herramienta *Neo4j Desktop* en su versión *1.4.15*.

IV-A. Grafo de pilotos conectados a los constructores con los que ganaron una carrera

Una de las primeras consultas que se tuvo en cuenta fue, para cada piloto, obtener todos los constructores con los que ganó una carrera, así como también cuántas ganó con cada uno. Se decidió proyectar un grafo con el resultado de esta consulta, con el objetivo de lograr una mejor visualización de los datos y de realizar un análisis más profundo utilizando algoritmos de grafo.

Listado 1 Proyección del grafo de pilotos y los constructores con los que ganó una carrera

```
MATCH (constructor:Constructor)<-[:BELONGS_TO]-
  (dlu:DriversLineUp)<-[:BELONGS_TO]-(driver:Driver)
CALL {
   WITH driver, dlu
   MATCH (driver)-[f:RACED_IN { position: '1' }]->(:Race)-
     [:HELD_IN]->(season)<-[:PARTICIPATED_IN]-(dlu)
   RETURN season.year AS season, count(*) AS season_wins
}
WITH constructor, driver, sum(season_wins) as total_wins
WITH gds.alpha.graph.project(
  'driverNetwork',
  driver,
  constructor,
  {
    sourceNodeLabels: labels(driver),
    targetNodeLabels: labels(constructor)
  },
  {
    relationshipType: 'WON_WITH',
    properties: { total_wins: total_wins }
  }
) AS g
RETURN
  g.graphName AS graph,
  g.nodeCount AS nodes,
  g.relationshipCount AS rels
```

[1] https://neo4j.com/docs/python-manual/current/
[2] https://gitlab.fing.edu.uy/santiago.rodriguez.barthe/bdnr-proyecto-final

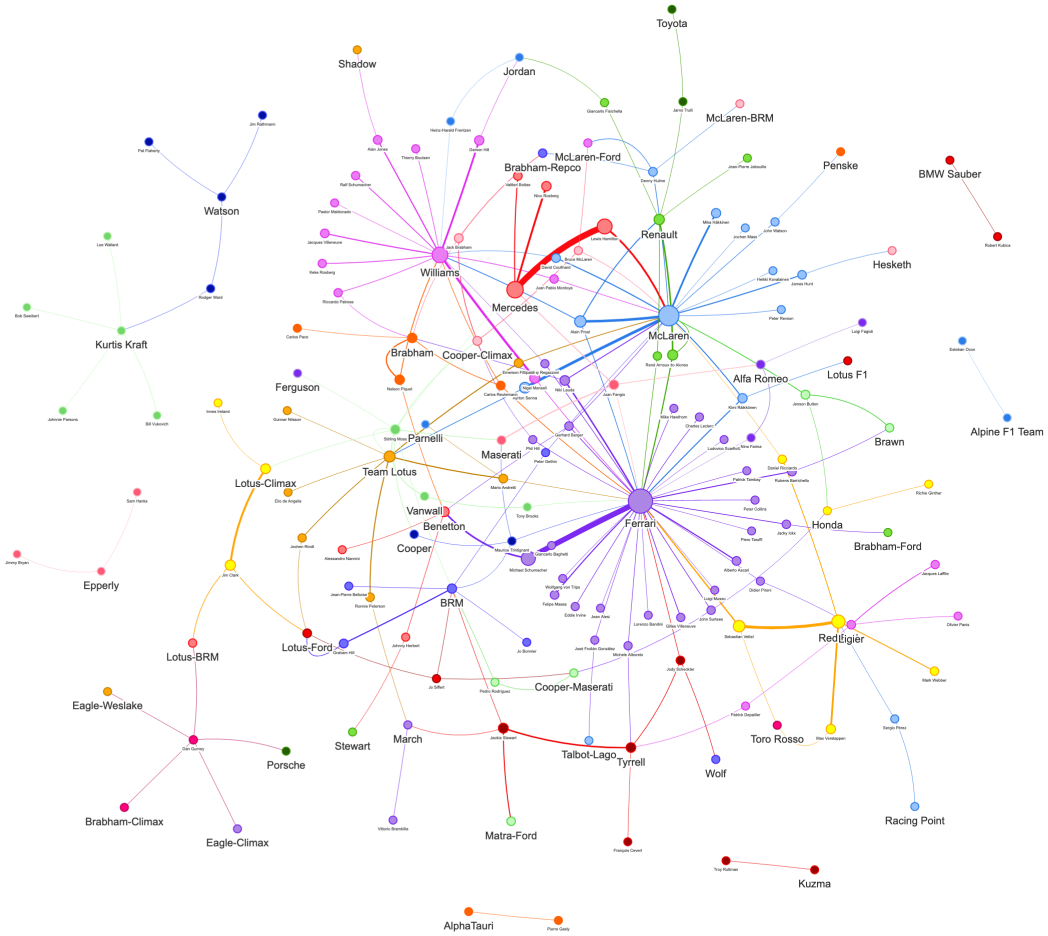

Figura 2: Red de pilotos conectados a los constructores con los que ganaron una carrera.

El grafo generado puede verse en la Figura 2. Se puede ver que los nodos presentan diferentes colores, los cuales se usan para identificar las comunidades en el grafo. También se puede apreciar que cada nodo presenta un tamaño distinto, el cual representa la importancia del nodo en el grafo, calculada con un algoritmo de centralidad de grado. De igual manera, las aristas presentan distinto grosor, de manera que mientras más gruesa es la arista, más victorias con ese constructor tuvo el piloto.

Para identificar las comunidades en el grafo usamos el método de Louvain [Neo4jb], ya que intenta maximizar la modularidad del grafo y soporta el uso de la cantidad de victorias de cada piloto con un constructor como peso. El algoritmo detectó 53 comunidades, siendo la más grande de ellas la perteneciente al equipo Ferrari, que consta del 16 % de los nodos. Como se puede observar, dada su estructura, el grafo presenta un nivel alto de modularidad al rededor de los constructores, con una topología de estrella [Weisstein]. Esto se corresponde con la alta modularidad (0.6232) reportada por *Neo4j* al correr el algoritmo.

Por otro lado, el tamaño de los nodos fue computado en base a un algoritmo de centralidad de grado [Neo4ja], el cual mide la importancia de un nodo según la cantidad de relaciones entrantes y salientes. Nótese que se tiene en cuenta la cantidad de victorias del piloto con el constructor para calcular este valor. Nuevamente, el nodo con un valor mayor de centralidad es Ferrari, mientras que el piloto con un valor más grande es Lewis Hamilton.

Para obtener la visualización del grafo se actualizaron los respectivos nodos en la base de datos con su identificador de comunidad y con su valor de centralidad. Luego, se realiza una consulta sobre la base de datos misma (en lugar de hacerla sobre la proyección) y se ilustra el resultado de la misma usando la librería de Javascript *NeoVis* [3].

IV-B. Grafo de compañeros de equipo

Se quiere crear un grafo para visualizar los compañeros de equipos de todos los pilotos. En la actualidad, por temporada los pilotos tienen un compañero de equipo, aunque en algunos casos este número puede incrementar dependiendo de las circunstancias (por ejemplo, por problemas de salud, un piloto de la alineación titular podría quedar imposibilitado

[3] https://github.com/neo4j-contrib/neovis.js

para competir por lo que un piloto de reserva se presenta en la competencia). Sin embargo, esto no era así en los primeros años de la competición, ya que las escuderías solían presentar varios pilotos para que compitieran bajo su nombre.

Para su visualización se realiza una proyección, colapsando los nodos intermedios *DriversLineUp* para crear la relación *TEAMMATE_WITH*, como se puede ver en la Figura 3 con un grafo reducido.

Se ejecutó el algoritmo de recuento de triángulos sobre la proyección de grafos de compañeros de equipo, código que podemos ver en el Listado 2. Este algoritmo es usado para detectar comunidades y medir la cohesión de las mismas [Neo4jd]. Los resultados del conteo fueron inesperados, dado que varios nodos presentaron una alta cantidad de triángulos. Por ejemplo, se encontró el caso de Maurice Trintignant, al cual se le contaron al rededor de 1700 triángulos. Si se observa la carrera del mismo, se puede ver que corrió en la fórmula 1 desde 1950 a 1964 y tuvo 126 compañeros distintos a lo largo de estos años. A modo de ejemplo, en el año 63 Maurice tuvo 10 compañeros distintos.

Listado 2 Proyección del grafo de la relación *TEAMMATE_WITH*

```
CALL gds.graph.project.cypher(
      'undirected_interactions',
      'MATCH (d1:Driver) RETURN id(d1) AS id',
      'MATCH (d1:Driver)-[r:BELONGS_TO*2]-(d2:Driver)
      WITH d1, d2, count(*) as times
      RETURN ID(d1) as source,
      ID(d2) as target, 'TEAMMATE_WITH' as type'
)
```

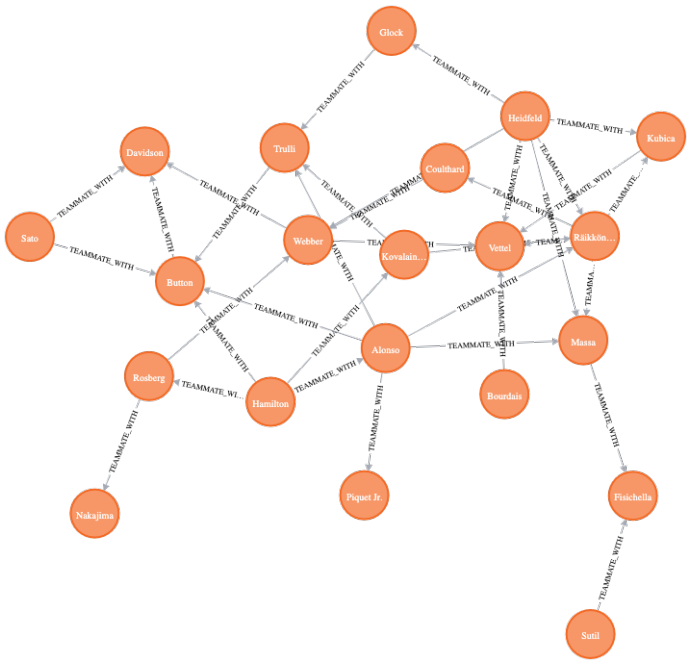

Figura 3: Red de pilotos conectados a sus compañeros.

IV-C. Consultas de interés

En esta sección se introducirán algunas consultas adicionales que resultaron de interés para demostrar la capacidad de la herramienta *Neo4j*.

Una de las primeras consultas con las que se decidió experimentar fue la de armar una lista con los países que tengan más pilotos que hayan ganado una carrera.

Listado 3 Países que tiene más victorias de pilotos

```
MATCH (c:Country)<-[:COUNTRY_ORIGIN]-(:Driver)-
   [:RACED_IN {position:'1'}]->(:Race)
RETURN c.name as country, count(c) as cantidad
ORDER BY cantidad DESC
```

El resultado no es inesperado, Reino Unido lidera por lejos con 308 victorias, de las cuales 103 de ellas fueron hechas Lewis Hamilton, el piloto que tiene el récord de carreras ganadas. Luego se encuentra Alemania en segundo lugar, país en el cual nació el famoso piloto Michael Schumacher, quién es el segundo piloto con más victorias. También alemán es el piloto Sebastian Vettel, tercero con mas carreras ganadas.

Por otro lado, otra consulta de interés fue la de determinar quiénes fueron los pilotos que más carreras ganaron en los circuitos ubicados en su país de origen.

Listado 4 Los pilotos que han ganado más grandes premios en su país de origen

```
MATCH (c1:Country)<-[:COUNTRY_ORIGIN]-(driver:Driver)-
   [:RACED_IN {position:'1'}]->(:Race)-[:RUN_IN]->
   (:Circuit)-[:COUNTRY_ORIGIN]->(c2:Country)
WHERE c1.name = c2.name
RETURN c1.name as country, driver.last_name,
   driver.first_name, count(c1) as cantidad
ORDER BY cantidad DESC
```

Los resultados son similares a la consulta anterior: Michael Schumacher es el piloto con más victorias en su país, habiendo ganado nueve carreras en circuitos ubicados en Alemania. En segundo lugar le sigue Lewis Hamilton con ocho carreras ganadas en Reino Unido.

Finalmente, la última consulta realizada fue la de, dada una temporada, determinar el himno que más veces se reprodució en los podios en toda la temporada. Vale la pena mencionar que, al finalizar una carrera, se realiza una ceremonia donde se le entrega el premio al piloto ganador y los pilotos se suben al podio para posteriormente escuchar los himnos de los pilotos y escudería ganadores. En el caso de que piloto y escudería compartan el mismo país de origen, el himno de ambos se reproduce una única vez.

Listado 5 El himno que más ha sonado en un podio dada una temporada

```
MATCH (cont:Country), (s:Season {year: '2008'})
CALL {
   WITH cont, s
   MATCH (race:Race)
   WHERE (race)-[:HELD_IN]->(s)
   AND (EXISTS {
      MATCH (race)<-[:RACED_IN {position: '1'}]-
        (:Driver)-[:COUNTRY_ORIGIN]->(cont)
   } OR EXISTS {
      MATCH (race)<-[:RACED_IN {position: '1'}]-
        (:Driver)-[:BELONGS_TO]->(dlu:DriversLineUp)-
        [:BELONGS_TO]->(:Constructor)-
        [:COUNTRY_ORIGIN]->(cont)
      WHERE (dlu)-[:PARTICIPATED_IN]->(s)
   })
   RETURN count(*) as cant_wins_per_country
}
WITH cant_wins_per_country, cont
WHERE cant_wins_per_country > 1
RETURN cont.name, cant_wins_per_country
ORDER BY cant_wins_per_country DESC
```

A modo de ejemplo, en el caso de la temporada 2008, el himno más escuchado fue el de Italia dado que Felipe Massa y Kimi Räikkönen ganaron ocho carreras para Ferrari (seis y dos respectivamente) y Sebastian Vettel ganó una para la escudería Toro Rosso-Ferrari.

V. Conclusiones y trabajo futuro

Como parte de este trabajo, se diseño y armó una base de datos de grafos con datos históricos de la Fórmula 1 usando *Neo4j*. El diseño utilizado demostró ser adecuado a las necesidades de los experimentos que posteriormente se realizaron, facilitando la implementación de las consultas. También se realizaron varias proyecciones de grafo, utilizando algoritmos de detección de comunidad y de centralidad para analizarlos y ayudar su visualización. Además, se realizaron distintas consultas a la base de datos, demostrando la capacidad de la herramienta y de las bases de datos de grafos para obtener información.

Los resultados obtenidos fueron muy satisfactorios e interesantes, tanto en cuanto a la información obtenida como el potencial extraído al usar la herramienta, desarrollando formas claras e intuitivas de observar la información.

A pesar de que la experiencia usando bases de datos de grafo fue muy buena, se podría sacarle más provecho al conjunto de datos obtenido ya que no todos sus datos fueron utilizados. Información como las paradas en boxes, las vueltas más rápidas o las posiciones al finalizar la temporada podrían abrir la puerta a consultas más específicas o a otras proyecciones interesantes. Se podría estudiar incluso la posibilidad de usar bases de datos de grafos para analizar los rendimientos de los pilotos o los equipos, tal y como se vió para otros deportes en los trabajos relacionados.

Referencias

[Abdon] Abdon. Formula 1 2013 Season - graphgists. URL: https://neo4j.com/graphgists/formula-1-2013-season/.

[Bourbousson et al.2010] Bourbousson, J., Poizat, G., Saury, J., and Carole, S. (2010). Team coordination in basketball: Description of the cognitive connections among teammates. *Journal of Applied Sport Psychology*, 22:150–166.

[Clemente et al.2015] Clemente, F. M., Couceiro, M. S., Martins, F. M. L., and Mendes, R. S. (2015). Using network metrics in soccer: a macro-analysis. *Journal of human kinetics*, 45:123.

[Fórmula 1 2022] Fórmula 1 (2022). Formula 1 announces TV, race attendance and digital audience figures for 2021 — Formula 1®. URL: https://www.formula1.com/en/latest/article.formula-1-announces-tv-race-attendance-and-digital-audience-figures-for-2021.1YDpVJIOHGNuok907sWcKW.html.

[Neo4ja] Neo4j. Degree Centrality - Neo4j Graph Data Science. URL: https://neo4j.com/docs/graph-data-science/current/algorithms/degree-centrality/.

[Neo4jb] Neo4j. Louvain - Neo4j Graph Data Science. URL: https://neo4j.com/docs/graph-data-science/current/algorithms/louvain/.

[Neo4jc] Neo4j. Triangle Count - Neo4j Graph Data Science. URL: https://neo4j.com/docs/graph-data-science/current/algorithms/triangle-count/.

[Neo4jd] Neo4j. Weakly Connected Components - Neo4j Graph Data Science. URL: https://neo4j.com/docs/graph-data-science/current/algorithms/wcc/.

[Park and Yilmaz2010] Park, K. and Yilmaz, A. (2010). Social network approach to analysis of soccer game. pages 3935–3938.

[Schnaiderman and Mignaco2021] Schnaiderman, M. and Mignaco, J. (2021). El Número de Abreu.

[Serapiglia2018] Serapiglia, A. (2018). Formula one–a database project from start to finish. *Information Systems Education Journal*, 16(2):34.

[Vopani2022] Vopani (2022). Formula 1 World Championship (1950 - 2022). URL: https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020.

[Weisstein] Weisstein, E. Star Graph. URL: https://mathworld.wolfram.com/StarGraph.html.