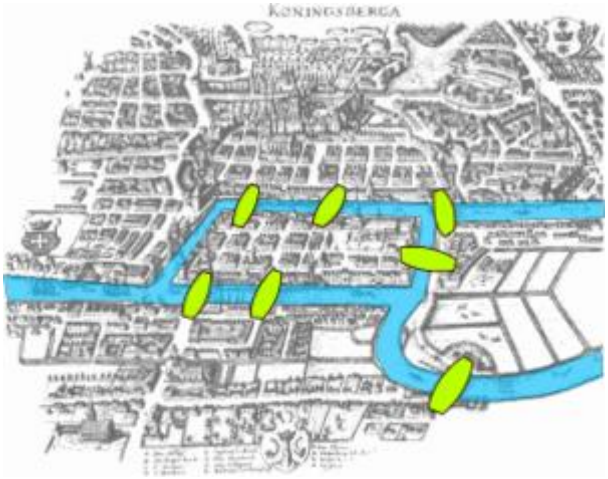
Creado por:

Isabel Maniega

Teoría de Grafos

Según la historia, la teoría de grafos surgió en el siglo XVIII a partir de un problema con unos puentes conocidos como los puentes de Königsberg. Resulta que en aquella época Prusia Oriental estaba dividida en cuatro zonas por el río Pregel. Había siete puentes que comunicaban estas regiones, tal y como se muestra en el dibujo.



El interrogante era como se podía caminar por la ciudad, cruzando cada puente una sola vez, y regresando al lugar de partida.1 Para ello el matemático Leonard Euler se dedicó a buscar una respuesta, lo que posteriormente daría origen a la teoría de Grafos.

Para resolver el problema Euler representó las cuatro zonas de la ciudad como cuatro puntos, y los puentes como aristas que unían las zonas. A partir del razonamiento de Euler se entendió que un grafo se compone de dos elementos principales: Vértices y Aristas. Y se comenzó a utilizar el mismo razonamiento para entender las relaciones entre cosas representadas por aristas.

V ------- > Vértice

E --------> Edge (Aristas)

Todo resultó en una fórmula matemática que se representa así:

G= (V, E)

Entonces, según el razonamiento de Euler se tenían cuatro vértices y las aristas eran todas las conexiones que se podían usar entre ellas, es decir la formas de ir de una zona de la ciudad a otra. Ahora bien, dejando un poco la historia de lado y regresando a nuestro tiempo, los grafos se pueden utilizar en casi cualquier ejemplo del día a día. Básicamente se proponen vértices y se buscan las aristas para ver cómo se conectan.

- Tipologías de grafos:

• **Grafo simple** o simplemente **grafo** es aquel que acepta una sola una arista uniendo dos vértices cualesquiera. Esto es equivalente a decir que una arista cualquiera es la única que une dos vértices específicos. Es la definición estándar de un grafo. Ejemplo: Tengo cuatro zonas en la cuidad y siete puentes.

• **Multigrafo o pseudografo** son grafos que aceptan más de una arista entre dos vértices. Estas aristas se llaman múltiples o lazos (loops). Los grafos simples son una subclase de esta categoría de grafos. También se les llama grafos no-dirigido. Por ejemplo, para llegar a Valencia de Madrid nos podemos ir por la autovía, o podemos tomar carreteras secundarias. Existen varias formas de llegar y cada una tendrá una distancia y un kilometraje distinto.

• **Grafo dirigido**: Son grafos en los cuales se ha añadido una orientación a las aristas, representada gráficamente por una flecha. Ejemplo, es como comprar un billete de Ave solo de ida de Madrid a Barcelona, porque nos vamos a vivir allá. Realmente nos interesa la distancia y el tiempo que se dura hasta nuestro destino y no de regreso. Esto resulta muy importante a la hora de contar el número de aristas en cada vértice.

• **Grafo etiquetado**: Grafos en los cuales se ha añadido un peso a las aristas (número entero generalmente) o un etiquetado a los vértices. Por ejemplo, en una red social como FaceBook. Cada individuo tiene su cuenta y agrega a sus amigos o conocidos. Existe la posibilidad de que cada persona tenga amigos en común. En algunos casos hay más amigos en común que en otros. Aquí las aristas tienen un valor numérico que representa la cantidad de amigos en común de cada vértice.

• **Grafo aleatorio**: Grafo cuyas aristas están asociadas a una probabilidad.

• **Hipergrafo**: Grafos en los cuales las aristas tienen más de dos extremos, es decir, las aristas son incidentes a 3 o más vértices.

• **Grafo infinito**: Grafos con conjunto de vértices y aristas de cardinal infinito.

Big Data y la teoría de grafos ¿Cómo se relacionan?

La relación entre Big Data y la teoría de grafos se fundamenta en la utilidad que aportan los grafos a la hora de dar explicación y almacenar grandes cantidades de datos. Para comprender un poco mejor la relación que guardan ambos conceptos es preciso saber con claridad de qué se trata cada uno por separado. Así pues, se puede comenzar diciendo que el Big Data, además de ser un conjunto de técnicas, es también un término utilizado para hacer referencia a los grandes volúmenes de datos. Estos grandes volúmenes de datos pueden componerse tanto de datos estructurados como de datos no estructurados. Haciendo más compleja la tarea de análisis e interpretación.

Pero, entre los diversos mecanismos utilizados para almacenar y pulir información valiosa destacan los grafos. Lógicamente, para seguir descubriendo la relación e importancia que tienen el Big Data y la teoría de grafos, es esencial definir de qué se trata esta última.

Los grafos y su utilidad

Como se ha mencionado, los grafos son una invención que data de hace muchos siglos atrás. Las matemáticas, en su afán de hacer una representación lógica de la realidad de manera gráfica, han dado origen a la teoría de grafos. Los grafos son estructuras con la capacidad suficiente para albergar una gran cantidad de datos. Lo cual, en principio, cae como anillo al dedo para efectos del Big Data.

Los grafos a su vez permiten conocer con exactitud las conexiones entre los datos que son capaces de almacenar y que se encuentran interconectados. Además, los grafos se componen de nodos, los cuales pueden localizar los diferentes tipos de datos. Dichos datos son interconectados a través de aristas gracias a los grafos. La relación entre los grafos y la analítica de datos Big Data y la teoría de grafos cobran especial importancia en este particular, puesto que los grafos pueden dar una explicación detallada acerca de toda clase de datos en grandes cantidades. Esto permite que la información pueda ser comprensible, independientemente de la naturaleza de los datos almacenados.

Los grafos resuelven muchos de los grandes retos del Big Data y la Data Analytics. Pero a su vez, para muchos dentro de la comunidad científica son considerados insuficientes. Esto último se debe en gran parte a que las relaciones que describen los grafos son representaciones binarias. Pero ¿qué sucede con las llamadas “relaciones de orden superior”?

Aquellas relaciones influenciadas por la dinámica grupal y que van más allá de las relaciones binarias.

Big data y la teoría de grafos: los hipergráficos

Hipergráficos es el término que los matemáticos han empleado para describir las relaciones que van más allá y que la concepción tradicional acerca de los grafos no puede explicar. Existen relaciones de orden superior, las cuales se ven afectadas por la dinámica que proponen las interacciones grupales. Algo que es mucho más complejo que el esquema binario que proponen los grafos.

Los grafos pueden explicar mediante aristas las relaciones que existen entre dos elementos de un conjunto. Dejando a un lado las relaciones subyacentes más complejas.

Estos fenómenos matemáticos están presentes prácticamente en todo lo que nos rodea. Por ejemplo, los grafos pueden dar explicación a la relación entre dos elementos individuales en un sistema. Pero no a la combinación entre estos y otro número ilimitado de elementos del mismo sistema.

La ciencia está evolucionando a una velocidad nunca vista. Esto lleva a la necesidad de profundizar cada vez más en nuevos conceptos y estar siempre actualizado para poder tener un seguimiento correcto de los detalles asociados a los nuevos paradigmas coligados a las nuevas tecnologías. No obstante, no todos estos conceptos son totalmente nuevos, sino que se encuentran asociados a ramas como las matemáticas o la computación. Así mismo, cada vez es más importante la visualización de los resultados para una fácil interpretación y es aquí donde gana importancia la teoría de grafos.

Históricamente el grafo se ha definido como un dibujo o bien una imagen. Pero, en las matemáticas o en la ciencia de la computación se entiende como grafo al conjunto de objetos llamados nodos o vértices. Más concretamente un grafo es un conjunto no vacío que se une en un diagrama de pares de vértices llamados aristas, que pueden encontrarse orientadas o no.

Nos surge la siguiente cuestión ¿Cuáles son las aplicaciones más comunes de la teoría de grafos en el entorno empresarial? Para dar respuesta dicha pregunta debemos saber que éstos permiten modelar problemas de la vida cotidiana. Esta afirmación suena a imposible, pero realmente ¿es inverosímil? La respuesta es “no”, ya que los grafos permiten de manera gráfica unificar nodos que nos permite observar la relación existente entre los vértices. Es decir, y a modo de ejemplo, podemos tomar las redes sociales donde se podría indicar que la unión de los vértices permite observar las interrelaciones y conexiones entre los distintos usuarios (Almagro y Ordóñez, 2014). Si continuamos con dicho ejemplo podemos indicar como M. Zuckerberg, fundador y CEO de Facebook utilizó un grafo o la teoría de grafos para la creación de su red social como facilitador del análisis y representación de la información (Almagro y Ordóñez, 2014).

Aplicaciones de la teoría de grafos

Pero ¿Cuáles son las principales aplicaciones de la teoría de grafos? Para obtener una respuesta a dicha mostramos la siguiente lista sobre la selección de unos cuantos hitos importantes que permite el análisis de la teoría de grafos (Puchades Cortés et al., 2008; Robledo et al., 2014; Jiménez Motte, 2017):

• La adaptación de la estrategia comercial por parte de las empresas. • La detección de comunidades en las redes sociales. • El reconocimiento de patrones que permita la extracción de información para la toma de decisiones. • La organización con respecto a la producción, y sus factores. • El análisis de la internacionalización de la empresa. • Todo aquello que se asocia a la demostración de teoremas. • Análisis de los problemas asociados al transporte de mercancías. • La creación, desarrollo y producción de software.

Tal y como podemos comprobar la teoría de grafos tiene una múltiple utilización en el conjunto matemático y computacional, lo que nos permite abrir un extenso abanico de usos, dado que tenemos una herramienta completa y relacional.

Es por ello que la utilización y desarrollo de base de datos de grafos ha demostrado una gran eficacia como herramienta de manipulación de grandes cantidades de datos, haciendo fácil lo difícil a la hora de implementar estrategias de manejo de dicha información (Pinilla et al., 2017).

Llegados a este punto deberíamos obtener una clara respuesta a la siguiente cuestión ¿Qué ventajas aporta el uso de bases de datos orientadas a grafos frente al uso de bases de datos relacionales conocidas como las tradicionales? La importancia de dicha pregunta viene asociada a la diferencia existente entre ambas bases de datos, así se encuentran dos puntos esenciales que nos diferencian la teoría de grafos y la tradicional.

En primer lugar, cabe destacar como la Teoría de grafos permite representar las interrelaciones de forma explícita en la base de datos. A diferencia del uso de base de datos relacionales los grafos muestran la relación entre los nodos que quedan almacenadas en el disco como una serie de punteros entre los nodos relacionados. Esto conlleva que la recuperación de los diferentes elementos interrelacionados es más simple. Esto implica una mayor eficiencia y eficacia para realizar una consultar de la información entre los nodos relacionados por parte de la teoría de grafos en relación con la base de datos tradicional. En segundo lugar, las bases de datos en grafo son también schemaless, lo que implica mayor flexibilidad a los potenciales cambios en los esquemas de datos. Con la facilidad asociada a la incorporación de nuevos nodos de manera mucho más sencilla en comparación a la tradicional.

En líneas generales, se puede afirmar que la utilización de Teoría de grafos es mucho más eficiente y eficaz que la utilización de base de datos tradicionales.

El Rol de los grafos

Los grafos son estructuras de especial importancia ya que podemos almacenar cantidades grandes de información, permitiendo entender las conexiones entre los datos que albergan. Los grafos están compuestos por nodos en los que podemos encontrar diferentes tipos de datos y estos pueden estar relacionados o conectados a otros nodos a través de aristas.

Al poder hacer un seguimiento detallado y con el apoyo de poderosas herramientas, darle rango visual para el ojo humano, podemos entender de mejor manera los datos que tenemos frente a nuestros ojos para tomar decisiones correctas y ajustadas a la realidad.

Los grafos permiten estudiar y determinar las interrelaciones entre unidades de datos, siendo de especial importancia en los análisis de big data. Como ya hemos dicho el Big Data concentra, analiza o gestiona cantidades gigantescas de datos y los grafos nos permiten entenderlos para tener un diagnóstico acertado y de calidad de ellos. Los grafos pueden ayudar, si son desarrollados de forma acertada a atender las principales dificultades del big data.

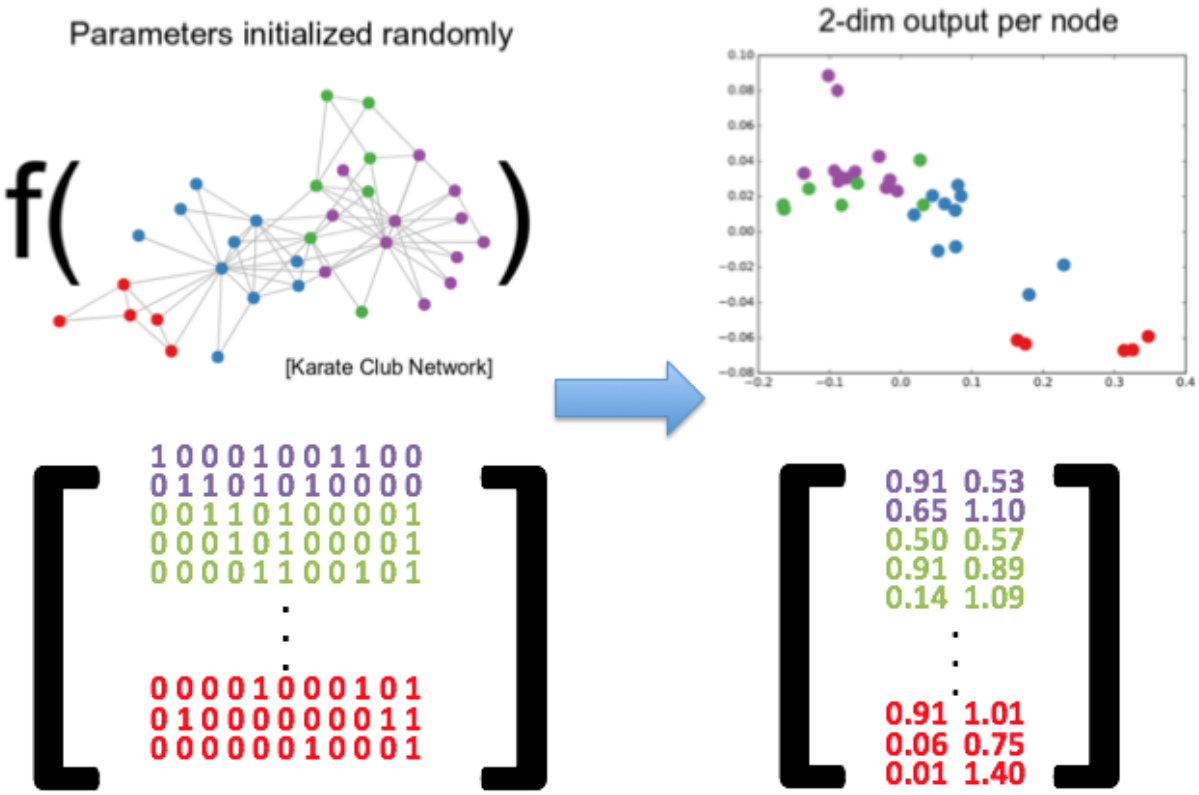
En instancias de gran volumen de datos, los diferentes métodos de análisis pueden ayudar a recolectar, limpiar, integrar y obtener datos de calidad en tiempos cortos de análisis.

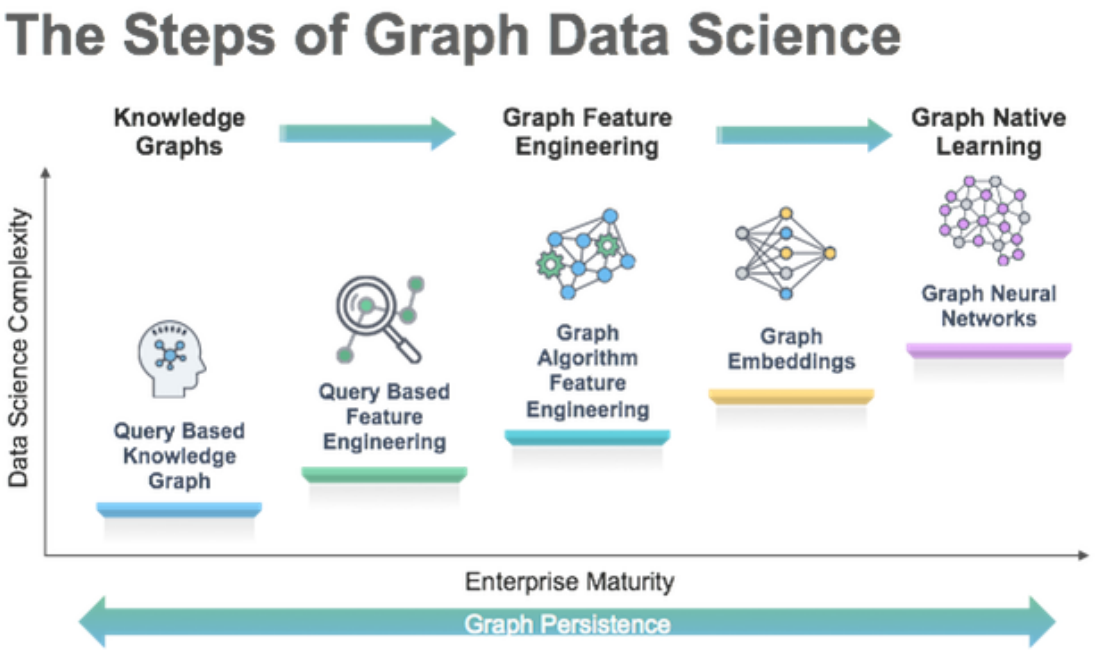
Además, esta reducción de tiempo en procesos de estudio permite que la toma de decisiones pueda ser desarrollada a tiempo, disminuyendo el porcentaje de probabilidad de que ejecutemos tareas con datos caducados.

Lo más importante en la vinculación de estas dos áreas tan potentes del desarrollo tecnológico es que pueden ser extrapoladas a diferentes áreas y ser aprovechadas por nuestras empresas de forma rápida para solucionar problemas. Esta área está en pleno desarrollo y crecimiento; debemos tomar la iniciativa desde ya para sacarle el máximo provecho.

Grafos en ML

El aprendizaje automatizado de representaciones (o Embeddings). Las Embeddings transforman los nodos de un gráfico en un vector, o un conjunto de vectores, preservando así la topología, la conectividad y los atributos de los nodos y bordes del gráfico. Estos vectores se pueden usar como características para que un clasificador prediga sus etiquetas o para un agrupamiento no supervisado para identificar comunidades entre los nodos.





En este caso, el primer nivel de evolución del uso de los grafos es realizarle consultas al mismo. Una vez tenemos los datos conectados dentro de un sistema gestor nativo de grafo obtener conocimiento del grafo es

sencillo. Usando consultas podemos interrogar a la base de datos y obtener cualquier información. Pero a veces esto no es suficiente, y lo que queremos es alimentar nuestro Pipeline de Machine Learning con Features que provienen precisamente del grafo. Estás features o variables, incorporan conocimiento de la estructura del grafo, como por ejemplo el número de vecinos. Esta información sería muy compleja de obtener de otro modo.

El siguiente nivel de evolución es el “Graph Algorithm Feature Engineering”, y aquí la alimentación del modelo de ML viene dada no sólo por consultas más o menos sencillas sobre el grafo, sino con datos provenientes de la ejecución de complejos algoritmos, como PageRank, Label Propagation, etc, dando como resultado Features relevantes para el cálculo del modelo en producción.

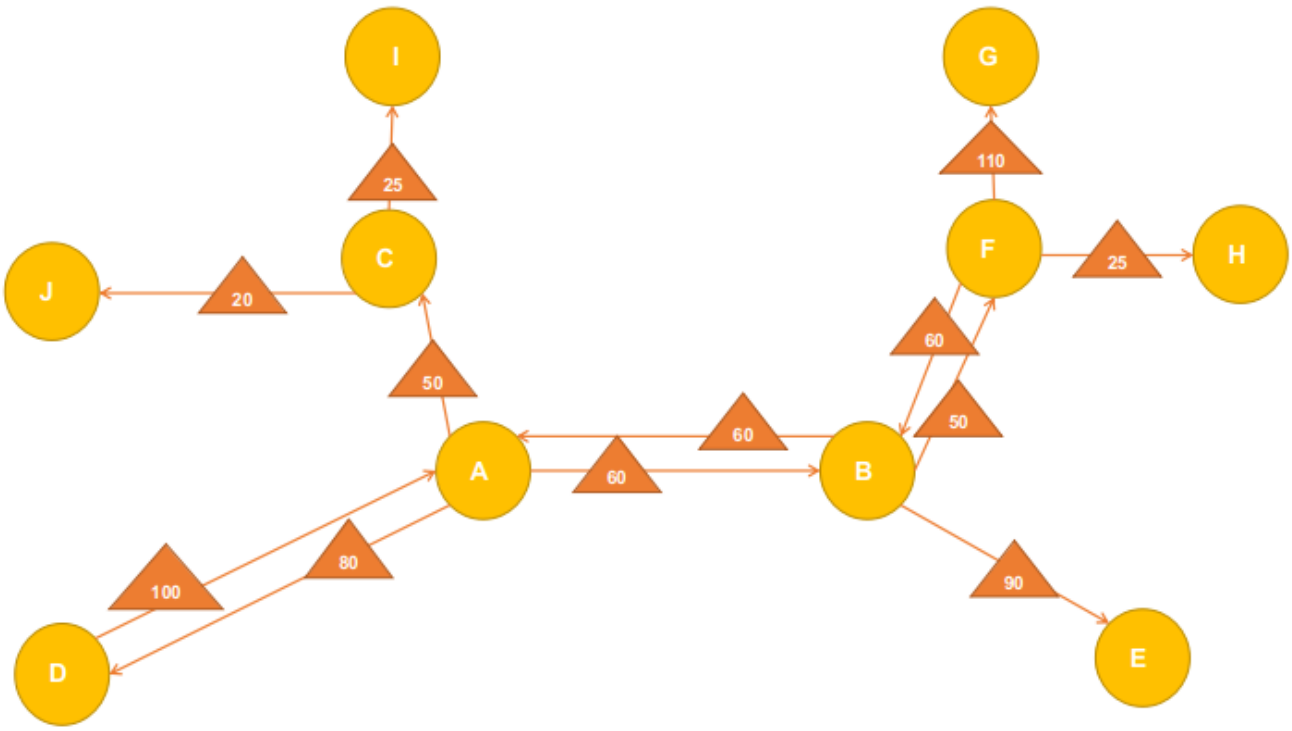
Los siguientes estadios son los más evolucionados, y requieren de mucha experiencia dentro de la organización. Graph Embedings, así como las Redes Neuronales de grafos, ayudan a sacar el máximo partido del hecho de tener un grafo como parte de la arquitectura de datos de la organización.

El grafo es ese repositorio de relaciones que permitirá a la organización dar un salto cualitativo en la capacidad de analizar nuestros datos.

Spark GraphX

GraphFrame admite el procesamiento general de gráficos, que es similar a la biblioteca GraphX de Apache Spark. Además, GraphFrames se basa en Spark DataFrames, que tiene las siguientes ventajas:

- API de Python, Java y Scala: GraphFrames proporciona interfaces API comunes para los tres lenguajes. Es la primera vez que todos los algoritmos implementados en GraphX se pueden usar en Python y Java.
- Consultas potentes: GraphFrames permite consultas breves, al igual que las consultas potentes en Spark SQL y DataFrame.
- Guardar y cargar modelos de gráficos: GraphFrames es totalmente compatible con las fuentes de datos de estructura DataFrame, lo que permite el uso de Parquet, JSON y CSV familiares para leer y escribir gráficos.



Estructura del graph

Existen distintos algoritmos para calcular los grafos, realizaremos una breve mención de todos ellos, ya que los calcularemos sobre el ejemplo anterior a través de caso práctico.

PageRank

Permite calcular los pesos para cada nodo o lo que es lo mismo asignar de forma numérica la relevancia de los nodos.

Label Propagation Algorithm (LPA)

Poder obtener la relación entre nodos obteniendo los grupos que lo forman, asignando una etiqueta a cada grupo.

Connected Components

Observar cuales la conexión entre los nodos, si hay nodos aislados o no.

Strongly Connected Components

Ver que nodos están fuertemente conectados.

Triangle count

Cuenta el número de triángulos para cada nodo en el gráfico y calcula el coeficiente de agrupamiento promedio para la red de nodos resultante. Un triángulo se define como tres nodos que están conectados por tres bordes (a-b, b-c, c-a).

Shortest paths

Calcula la ruta más corta (ponderada) entre un par de nodos.

Breadth-first search (BFS)

Recorrer el gráfico desde el nodo raíz y explora todos los nodos vecinos. Luego, selecciona el nodo más cercano y explora todos los nodos inexplorados. El algoritmo sigue el mismo proceso para cada uno de los nodos más cercanos hasta que encuentra el objetivo.

Subgraphs

Selecciona los nodos que cumplen una serie de normas.

Motif finding

Se conoce como coincidencia de patrones de gráficos. La coincidencia de patrones encuentra algún patrón dentro del gráfico. El patrón es una expresión que se usa para definir algunos vértices conectados.

Creado por:

Isabel Maniega