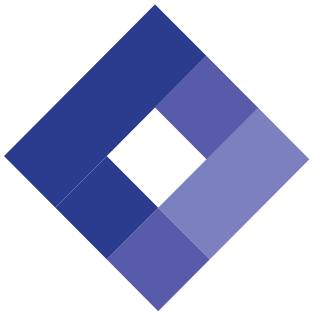

Instituto
International de
Ciencia de datos

Análisis de Redes en la era del Big Data

Rodrigo Aguilar

Introducción

> *Our Sun is a second- or third-generation star. All of the rocky and metallic material we stand on, the iron in our blood, the calcium in our teeth, the carbon in our genes were produced billions of years ago in the interior of a red giant star. We are made of star-stuff.*
>
> Carl Sagan

El mundo en el que vivimos está formado de materia interconectada de diversas formas: la masa lo hace a través de ligas de fuerza de atracción atómica fuerte, débil, a través de fuerza magnética, eléctrica y gravitacional. Todos y cada uno de nuestros átomos están vinculados entre sí y también al resto de los que existen en el planeta.

Así como estamos conectados a nivel atómico, lo estamos también en muchas otras capas, por ejemplo, con la información genética que compartimos con nuestros parientes, con los profesores que nos impartieron clase, con los comercios donde hacemos transacciones o conectamos con los gustos de las demás personas que miran los mismos videos o leen los mismos artículos que nosotros. También hay grupos de personas interconectados a través de carreteras y autopistas o que comparten la red eléctrica o hidráulica que les provee servicio.

Todos los ejemplos mencionados, incluso la lista de ejemplos de red, son entonces, ejemplos de una red. Las redes por su estructura se dividen en aleatorias y determinísticas, y por su dinámica se dividen en simples y complejas, la complejidad crece cuando las características topológicas (se refiere a la forma de la red) se distinguen de las que tiene una red aleatoria.

Tener la capacidad de comprender la estructura y la dinámica de las redes, así como su correcta similitud con las redes sociales y comerciales, nos permitirá intervenir de una forma más oportuna en los retos que se develan en la era del Big Data.

A continuación, realizaremos un breve repaso a las ventajas que tiene conocer el panorama que ofrece la teoría de redes, incluyendo varias aplicaciones de interés actual en la industria y el mercado.

Antecedentes

Esta forma de entender las relaciones entre distintos entes fue entendida y estudiada desde hace ya varios siglos. Una de las historias más famosas al respecto es la de los siete puentes de Königsberg: En dicha ciudad existían 7 puentes que la conectaban. ¿Es posible encontrar una ruta en la cual solo se transite por cada puente una sola vez? El matemático suizo Leonard Euler dio con la respuesta a este problema y sentó las bases para el estudio de las redes más complejas, las que ahora residen en los servidores de bases de datos que hay en todo el mundo.

Para que Euler pudiera dar con la respuesta, primero se abstrajo de todos los elementos que eran irrelevantes en cuanto al camino se refiere, así pues, la estructura de las calles, la geografía y la orografía de la ciudad fueron eliminadas, para contar el paso por puentes, también es irrelevante. Al final, la representación entre las regiones de la ciudad interconectadas por los puentes, queda dada por dos elementos: los puntos que se interconectan y las conexiones, es decir, lo que actualmente llamamos nodos y vértices.

El resultado de la existencia de un camino que pase solo por un puente es negativo, resultado que tiene que ver con la estructura intrínseca de este conjunto de nodos y vértices, objetos a los cuales a partir de ahora, les llamaremos grafos (graphs). El razonamiento que dio para probarlo tiene que ver con que, excepto por los puntos de inicio y fin, por todos los demás nodos tienen que haber exactamente el mismo número de vértices de entrada que de salida (un número par de ellos). El número de vértices de entrada y salida se llaman hoy el grado que tiene cada nodo y esta característica es una de las más importantes que nos permiten catalogar los distintos tipos de redes o grafos.

Desde ese problema en 1746 y hasta la fecha, el entendimiento de las estructuras de las redes ha crecido y ha permitido aplicar este conocimiento en la optimización del funcionamiento de todo lo que podemos entender como grafos (nodos y vértices). Algunos ejemplos de esto son la red de distribución eléctrica en una ciudad, la utilización de la interacción social en una red virtual o real para promover la compra de un artículo, la optimización de recursos en el tránsito aéreo, la optimización de venta de vuelos comerciales, la simulación del cerebro de las lombrices, el diseño de material genético (ver figura 1).

Figura 1. Abstracción del problema de los puentes de Königsberg

¿Cómo se conecta esta área de conocimiento al mundo de los negocios? Lo hace en distintos ámbitos, desde el alto nivel, siendo un método estratégico de entender relaciones entre grandes empresas, o entre empresas y empleados, o entre empresas y Estado, y lo hace también con detalle, en el nivel de las transacciones de sus clientes, como la identificación de comunidades que usan con más frecuencia las líneas telefónicas móviles o fijas, o distinguir las operaciones bancarias cotidianas de las que son susceptibles de formar parte de operaciones fraudulentas.

En las siguientes secciones, se utilizará el término gráfico, grafo o red de manera indistinta, puesto que para el tema que nos compete, las tres son sinónimos.

Aplicaciones estratégicas

Para cualquier tomador de decisiones es fundamental entender el panorama que lo rodea, teniendo claro las posiciones de los jugadores y la relación que existe entre todos ellos.

Adam Brandenburger y Barry Nalebuff, notables economistas del siglo pasado, desarrollaron la teoría de la coopetición, en la cual el valor total de la red de una empresa se calcula a partir de las relaciones que existen entre sus clientes y sus proveedores, así como entre sus competidores y sus complementos. Estos entes en su conjunto conforman un grafo de 5 nodos y grado 3. Los vértices, identificados con la relación que se tiene con cada uno de sus vértices, pueden representar distintos conceptos, como ventas, compras, descuentos y en cada caso, el valor neto de la red se calculará con la suma del valor de los vértices.

Existen también distintos indicadores que dependen únicamente de la estructura de la gráfica, como saber si el grafo está conectado, si existen trayectos entre todos sus puntos y qué tan difícil es “llegar” (ojo, que el término “llegar” va a depender de la relación definida en los vértices).

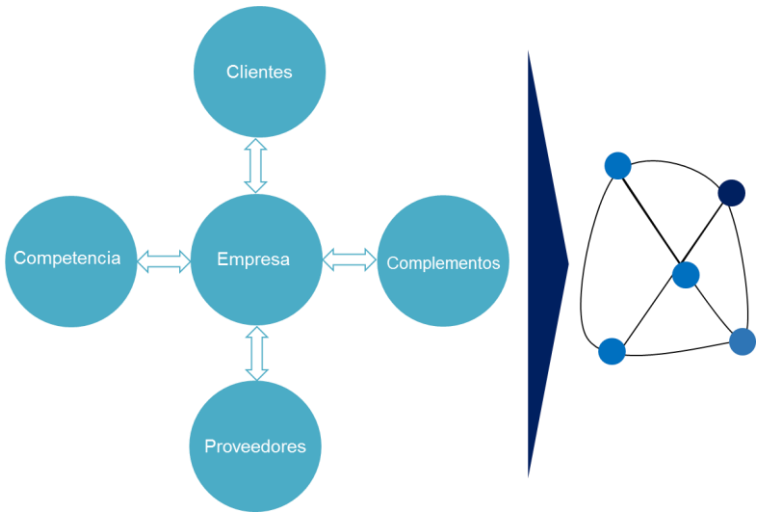

Figura 2. Traducción de la teoría de coopetición a un grafo de 5 nodos y orden 3.

Dos aplicaciones más del análisis de redes más a nivel estratégico son un marco de referencia para los negocios electrónicos (e-business) y el que pueden utilizar las áreas de recursos humanos para mapear a los individuos y las interrelaciones formales e informales que existen entre los puestos que ejercen.

El primer marco de referencia se enfoca en tres elementos: participantes (empresas, clientes, proveedores y colaboradores), relaciones (todo tipo de relaciones electrónicas y no electrónicas) y flujos (incluyendo flujos de capital, flujos de información y flujos de productos y servicios). Gordjin y Akkermans, de la Universidad Libre de Amsterdam modelan a los actores que forman parte del núcleo del negocio electrónico: actor, interfaz de valor, segmento de mercado y por otro lado, el empalme, objeto de valor, segmento escénico, intercambio de

valor, puerto de valor y estímulo final. Este marco de referencia facilita cálculos de escenarios hipotéticos de indicadores clave, la cual depende en una manera compleja de la relación que conservan con distintos entes (nodos y vértices).

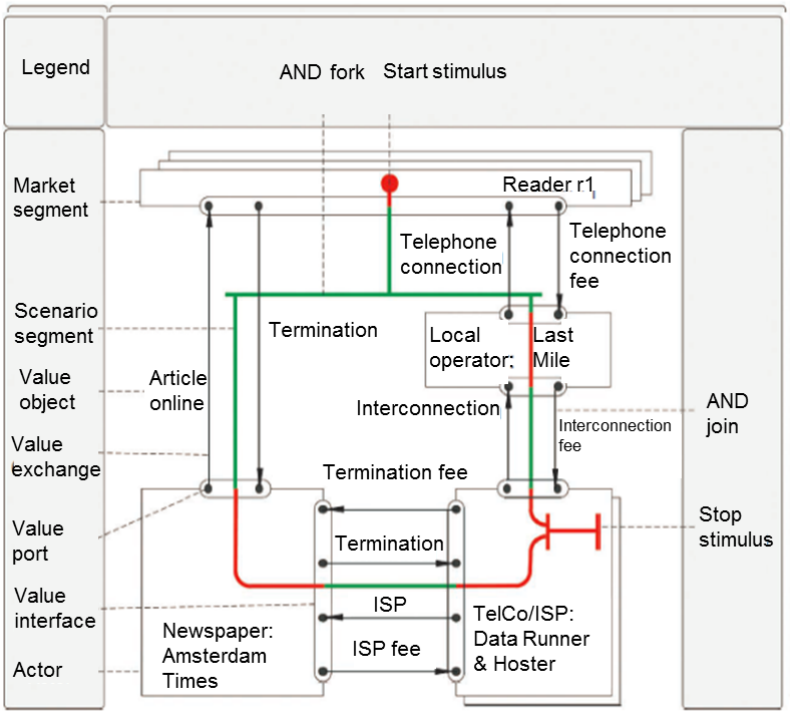

Figura 3. Modelo de Negocios Electrónicos (Gordjin, Designing and Evaluating E-Business)

Por otro lado, el Análisis de la Red Organizacional (ONA por sus siglas en inglés), permite comprender las relaciones formales (estructura empresarial) y las relaciones informales (construcción de la marca propia) a través de una metodología basada la medición de las relaciones entre las personas, los grupos y las organizaciones con los recursos, conocimiento y tareas con las que se desempeñan. La técnica que se utiliza para mapear estas entidades (nodos y vértices) toman en cuenta la complejidad de las organizaciones humanas y sus interacciones, las cuales no podrían ser representadas sin conceptos relacionales (o de teoría de redes). Los resultados de estos análisis permiten a los tomadores de decisiones en dichas organizaciones el poder comprender los factores de desempeño críticos, así como la ruta de difusión, la velocidad, la calidad y la precisión que requiere y sigue la información dentro de la organización.

Aplicaciones técnicas: Análisis de Redes y Big Data

Las aplicaciones técnicas del análisis de redes son tan grandes como la industria misma: desde calcular la eficiencia de la distribución de las cargas eléctricas en la red de una ciudad, hasta la planeación de la demanda de la cantidad de cada producto (vértices) a nivel sku (nodo) a través de centros de distribución (nodo) de los más grandes retailers, garantizando que los tiempos (vértices) en los que se distribuyen dichos materiales sean adecuados. Y ni qué decir del potencial que encierra el porvenir del Internet de las Cosas, o el Internet de Todo, o el Internet Industrial, o la evolución de hardware, como guste llamarlo el amable lector.

El área de Tecnologías de la Información ha estado particularmente muy activa al respecto en estos últimos 10 años. La explosión del uso del internet y las formas en las que se usa hoy día, incluyendo las redes sociales virtuales, la navegación para la investigación de conceptos y la investigación de precios de artículos que se venden en línea son manifestaciones de relaciones entre entes (vértices y nodos). En una gran diversidad de industrias, el análisis de redes habilita las actividades encaminadas a combatir el fraude y el terrorismo, permite entender la influencia de los líderes de pensamiento, ayuda a monitorear el sentimiento (sentiment analysis), caracterizar los segmentos de mercado,

ptimizar el compromiso y la experiencia con los usuarios de las redes sociales y en general, cualquier aplicación en la que sea necesario identificar rápidamente patrones complejos de comportamiento.

En un grafo, cada nodo tiene un valor y un peso específico dentro del mismo. Las medidas de centralidad de un nodo son aquellas que definen con precisión dichos valores y pesos. Las aplicaciones de estas medidas son en muchos casos inmediatas, así pues, si hablamos del estudio de una red social, algunas de las medidas que dan información valiosa y no trivial de un nodo (que en este caso, es una persona), son las siguientes:

Centralidad	**Aplicación**
Grado	Personas a con las que directamente está conectado
Cercanía	¿Qué tan rápido puede alcanzar una persona a todas las demás de la red?
Intermediación	¿Qué tan probable es que esta persona sea un punto intermedio en la ruta entre otras dos personas?
Valor Propio	¿Qué tan *bien* está conectada esta persona con otras personas que están *bien conectadas*?

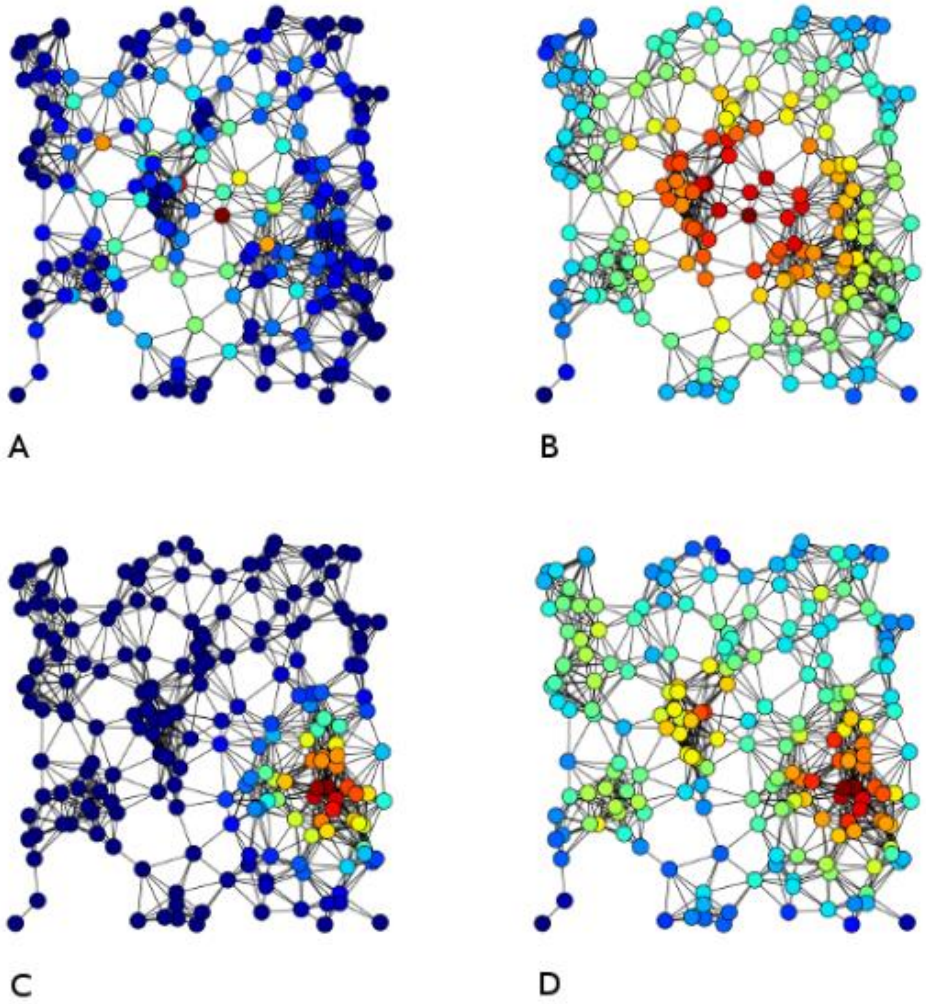

Figura 4. La expresión gráfica de las cuatro medidas de centralidad. Los colores fríos indican un índice bajo y los colores cálidos un índice alto. A) Intermediación B) Cercanía C) Valor Propio D) Grado

Muchas de las aplicaciones del análisis de redes en la industria se basan en estas medidas de centralidad y se complementan con las que se definen de acuerdo a las necesidades de cada negocio. El tipo de preguntas que se responden con estos análisis son:

¿Dónde sucede el trabajo? ¿Dónde se encuentran concentradas las relaciones? ¿Dónde reside el conocimiento? ¿Quiénes son los elementos claves en una relación? Cuando hablamos de relación, se entiende que esta es de transaccionalidad bancaria, compra en tienda, compra en línea, las labores diarias en la compañía.

Bases de Datos Gráficas

El análisis de redes por su naturaleza, consumirá una gran cantidad de datos. En estos momentos ya se oye acerca de infraestructuras para análisis paralelo masivo de redes en el sector público de Estados Unidos e Inglaterra, donde se ejecutan 4.4 billones de nodos (registros) y 70 billones de vértices (relaciones entre dichos nodos). Facebook ha desarrollado la capacidad de poder buscar en los posts individuales, gracias a una serie de capas de procesamiento, en las cuales se incluye una de procesamiento de lenguaje, que prepara la consulta y, posteriormente, a través “Unicornio”, el motor de búsqueda, se rastrean las entidades y relaciones correspondientes a la búsqueda. Por ejemplo, “friend:7” es el índice que caracteriza a todos los amigos del usuario con id 7, de forma similiar, Facebook maneja alrededor de 100 etiquetas con características (relaciones) del estilo “fotos tomadas por el usuario 12”, “usuarios que le dieron like a la publicación con id 23”. Una forma muy eficaz de cómo solucionar estas consultas es a través del uso de bases de datos gráficas.

Una de las tecnologías en bases de datos que habilita estas búsquedas asociadas a relaciones en una red, son las bases de datos gráficas. Estas bases de datos forman parte de la familia de tecnología conocida como NOSQL (Not Only SQL), que son capaces de administrar y trabajar eficientemente con datos semiestructurados y no estructurados. Otras tres tecnologías de esta familia NOSQL son los Document Stores (MongoDB o CouchDB), Key Value Stores (redis o amazon DynamoDB) y Bases de Familias de Columnas (HBase, riak, Cassandra). El empleo de estas tres tecnologías será abordado en posteriores artículos.

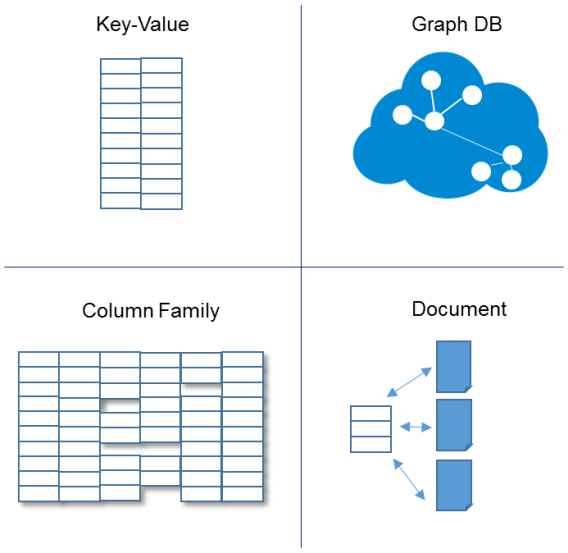

Figura 5. El cuadrante de la familia NOSQL

En el caso de las bases de datos gráficas, existen dos propiedades que les permiten trabajar mejor con modelos de análisis de redes:

El procesamiento de grafos. Existen distintas formas de codificar y calcular las relaciones de los grafos, entre ellas, se dice que un sistema administrador de bases de datos tiene capacidades nativas cuando su diseño se basa en la adyacencia sin índice. En estas condiciones, cada nodo funciona como un micro-índice de sus nodos adyacentes, lo que resulta mucho más económico en términos computacionales que mantener un índice global. Esta forma de procesar datos es mucho más eficiente, ya que dependiendo de la implementación, las búsquedas de índices puede ser de orden O(log n) en complejidad algorítmica, mientras que la búsqueda de relaciones inmediatas es del orden O(1). Navegar por m pasos de una red, el costo de la búsqueda indexada es O(m log n), por mucho, mucho mayor que el costo computacional de una búsqueda de adyacencia sin índice de O(m). En resumen, las intersecciones bidireccionales se precalculan y guardan en la base de datos como una relación, a diferencia de lo que sucede en un sistema indexado, donde adicionalmente, existe la dificultad adicional de caracterizar las “relaciones inversas” entre entidades.

El almacenamiento de grafos. La forma en cómo se guardan los datos es un elemento clave para poder explotar los beneficios de utilizar bases de datos gráficas. Una forma de almacenar los datos es a través de archivos de almacenamiento para cada parte específica del grafo. Es decir, hay archivos separados para los nodos, las relaciones, las etiquetas y las propiedades.

Además de la optimización del almacenamiento, en cada implementación es necesario tomar en cuenta el hardware, puesto que no obstante el costo económico de la memoria RAM ha disminuido consistentemente, el tamaño de una red de regular tamaño sigue superando la capacidad de dicha memoria. Si bien los discos de estado sólido son una buena opción, la latencia entre el disco y el CPU sigue siendo mucho mayor que la del chaché L2 o el RAM. Actualmente la solución que utilizan las bases de datos gráficas es la de trabajar con archivos caché, con lo cual, se puede trabajar con redes de billones de nodos y decenas de billones de vértices.

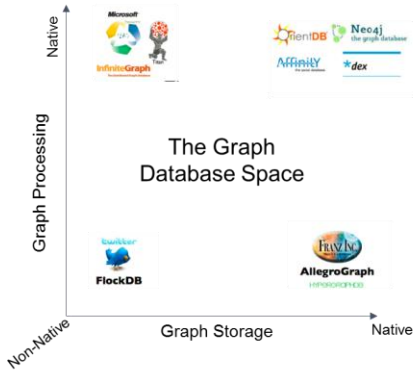

Figura 6. Ecosistema de Bases de Datos Gráficas

Existe una lista bastante nutrida de bases de datos gráficas que, satisfacen necesidades específicas de distintos usos. Las hay comerciales y de código abierto, nativas y no nativas por su almacenamiento y por su procesamiento. Con una buena implementación, todas ellas son capaces de ayudar a resolver preguntas importantes de negocio.

Casos de uso de las Bases de Datos Gráficas

Detección de Fraude. Los bancos y compañías aseguradoras pierden fuertes cantidades de dinero debido a los fraudes. En México se estima que las aseguradoras pierden 12 mil millones de pesos anuales debido a este rubro y una estimación de Forbes afirma que el 40% de las tarjetas de crédito son víctimas de fraude. Muchas técnicas tradicionales que utilizan análisis discreto y ayudan a detectar fraudes, adolecen de un alto registro de errores de tipo I y II (falsos positivos y falsos negativos). Esta debilidad del modelo es utilizada por quien realiza fraudes de forma cada vez más sofisticada. El análisis de redes, habilitado en grandes conjuntos de nodos y vértices a través de bases de datos gráficas, permite desarrollar métodos, como el análisis de enlaces contextuales avanzado, para descubrir comunidades completas de fraude y algunas otras formas más complejas de estafa a los tarjetahabientes y a las empresas.

En el comercio electrónico, que se ha vuelto un estándar para muchos consumidores, los defraudadores se han adaptado y se ha adaptado para aprovechar distintas artimañas fraudulentas en sistemas de pago en línea. Mientras que la mayoría de los fraudes en los pagos electrónicos son ejecutados por amplias redes criminales, incluso un defraudador solitario y bien informado, puede crear una cantidad grande de identidades sintéticas con las que logre defraudar una cantidad considerable. Las transacciones en línea contienen mínimamente los siguientes campos: identificación de usuario, dirección IP, geolocalización, un número de tarjeta y una cookie. Salvo algunas excepciones, como cuando una familia utiliza una misma tarjeta o cuando un individuo utiliza distintas locaciones e IPs, las relaciones entre estos campos debería ser uno a uno. No obstante, cuando la relación entre variables excede considerablemente un cierto número de ocasiones, es sano indagar si nos encontramos ante un caso de fraude. Mientras más conexiones tenga un nodo, más posible es que se trate de un fraude. Con un esquema correcto de alarmas y validaciones, estas redes pueden descubrirse antes de que inflija un daño sensible. Algunas de las validaciones pueden incluir eventos como intentos de acceso, hacer un pedido en línea o registrar una tarjeta de crédito, todos estos eventos se pueden comparar con la estructura de un esquema de fraude conocido y crecer los casos conocidos de dichos esquemas.

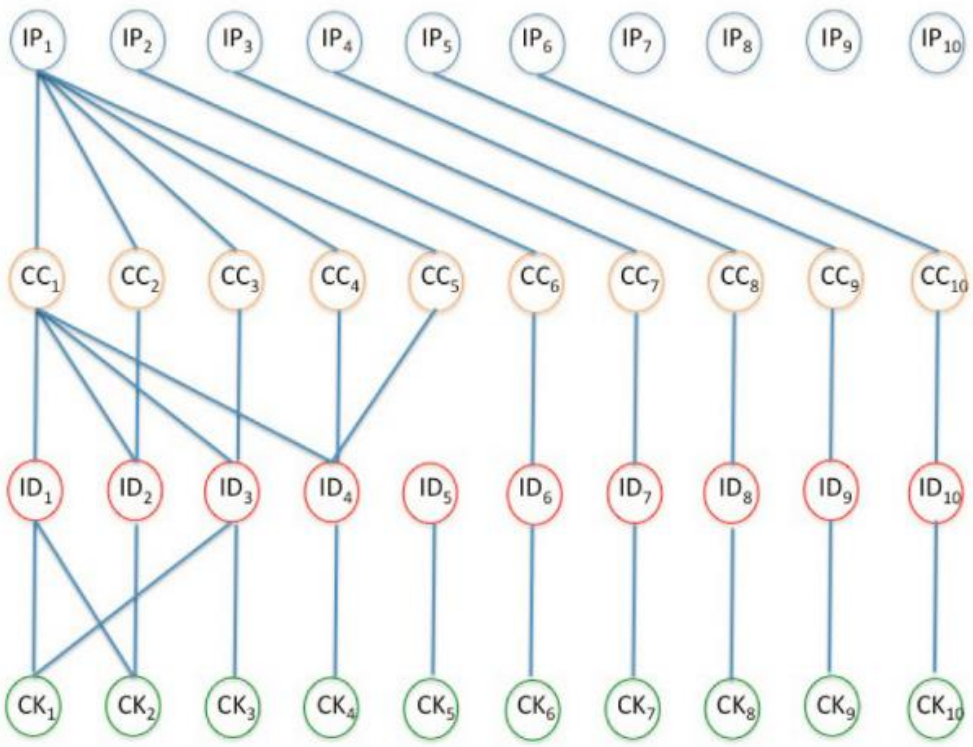

Figura 7. Representación gráfica de redes, donde existe una alta probabilidad de que se esté cometiendo un fraude desde la IP1 , que está haciendo transacciones con 5 tarjetas de crédito distintas (CC).

El elemento que enriquece la detección de fraude con análisis de redes, es el análisis de enlaces. Por tanto, mientras los negocios electrónicos se vuelven más rápidos y más automatizados, el intervalo de tiempo necesario para detectar un fraude se acorta, acercándose a la necesidad de dar solución en tiempo real. Aquí el análisis de redes a través de bases de datos gráficos agregan valor para hacer detecciones más eficientes.

Recomendaciones en Tiempo Real para Retail. La industria de Retail en México es un negocio del orden de los 40 mil millones de dólares y da servicio a millones de clientes. Una de las aplicaciones que utilizan a nivel global en este sector es la de recomendaciones en tiempo real, basada en las preferencias de los clientes que visitan las tiendas virtuales. Gigantes como eBay han implementado de manera muy efectiva modelos de recomendación basados en análisis de redes con bases de datos gráficas.

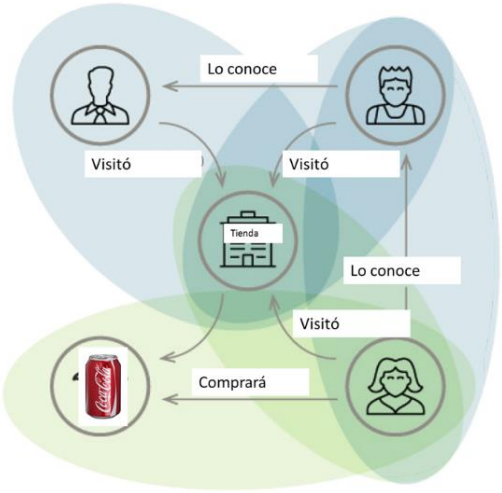

Figura 8. Relaciones que permiten hacer recomendaciones en tiempo real.

Guardando y utilizando datos de preferencias y relaciones en una base de datos gráfica, es posible proveer un servicio en tiempo real, satisfaciendo las exigencias y expectativas de clientes cada vez más estrictos e impacientes.

Tendencias en la industria, como el Big Data que crece cada vez a mayor velocidad, implican cambios cuantitativos y cualitativos en las empresas, para los cuales deben estar preparados. La generación, administración y acumulación de datos es necesaria, pero no suficiente para poder tomar las decisiones correctas que permitan a las empresas crecer y prevalecer sobre la competencia y los acontecimientos externos. No basta recolectar los puntos, sino que hay que conectarlos; la relación entre los datos registrados es tan valiosa en términos de información, como los registros mismos. Para entender y utilizar esas relaciones entre los distintos datos de una organización, requiere la tecnología correcta. Para utilizar el poder de los modelos de análisis de redes en gran escala, dicha tecnología es la de las bases de datos gráficas. En general, en un ambiente en el que la aparición de nuevas fuentes de datos que sirven a las empresas es constante, toda la familia de bases de datos NOSQL permiten desarrollar un ambiente robusto (Data Lakes) en el cual es más fácil la agregación y administración de dichas fuentes, habilitando así una familia también nueva de soluciones rápidas y eficaces, que responden a las necesidades actuales de los consumidores.

El análisis de redes en la era del Big Data potencia el entendimiento y la capacidad de rápida transformación que caracteriza las necesidades de nuestro mundo actual. Para las industrias, dominar el análisis de las redes que se crean y se almacenan en sus bases de datos, es un requerimiento necesario para mantener su competitividad. Poder saber cómo se concatenan individuos y transacciones a lo largo y ancho del mundo, favorecerá sin duda a las condiciones que permitan a las empresas y a los individuos tomar decisiones más informadas.

Acerca del autor

Rodrigo Aguilar es matemático y un temprano promotor de la aplicación del método científico con grandes datos. Utiliza su experencia para encontrar la mejor manera de adoptar nuevas tecnologías y paradigmas, como blockchain o conocimiento abierto, y ayudar así a convertir datos en información valiosa en la toma de decisiones.

Como Gerente Senior de la práctica de Data Analytics en PwC México, ha construido y consolidado al equipo multidisciplinar de especialistas para el desarrollo de productos y servicios analíticos que satisfagan necesidades y requerimientos del mercado.

Uno de sus principales objetivos es utilizar el poder de la ciencia aplicada en grandes datos para optimizar y aplicar los conocimientos colectivos y construir así un mundo mejor para todos.

Acerca del Instituto

El Instituto Internacional de la Ciencia de Datos busca resolver preguntas importantes de la sociedad a partir de los datos que genera la misma. Construye sinergias entre la creciente comunidad de entusiastas y científicos de datos y fomenta la correcta ejecución de la apertura de datos y conocimiento.

https://www.linkedin.com/company/instituto-de-la-ciencia-de-datos

@the_i2ds