

GRAPH-BASED RELATIONSHIP ANALYTICS: ELIMINATING DATA ANALYSIS BARRIERS

Cogito Knowledge Center provides powerful means to structure, navigate, and discover hidden information.

Today's media and information technologies produce an avalanche of data. Yet finding

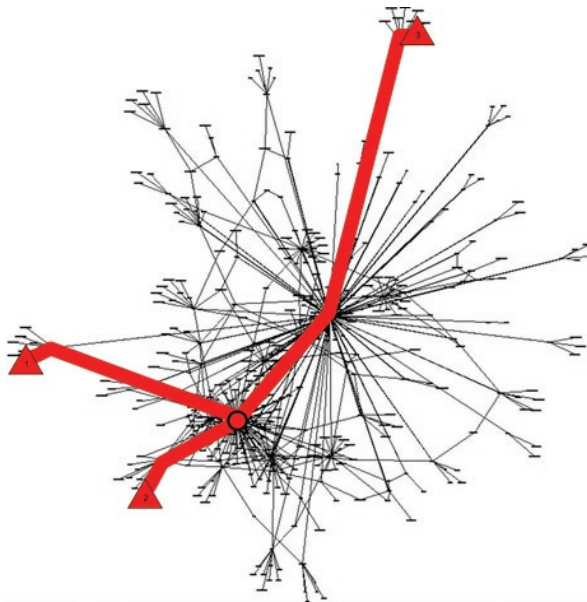


Figure 1: Relationship Analytics shows gatekeepers in a money transfer network

meaning in this data is increasingly difficult using traditional database technologies and associated data mining & analysis tools. New requirements for compliance monitoring and data analysis in virtually every industry have been met with “brute force” techniques, straining resources and showing the need for a different approach to be all the more urgent.

To meet this challenge, Cogito developed an innovative graph-based relationship analytics (GBRA) solution that provides tremendous power in modeling, querying and viewing data relationships. Using the Cogito Knowledge Center, information analysts—security experts, financial analysts, database administrators and researchers—can rapidly search, match, compare, and discover patterns in structured and unstructured data. Data is put in context, so relationships are more easily discovered, particularly in data sets that are large, complex and dynamic.

This white paper summarizes graph-based relationship analytics capabilities and how they complement and extend the data management solutions currently in wide use.

REAL-WORLD APPLICATIONS

Graph-based Relationship Analytics can be valuable for many applications in environments such as intelligence/ security, life sciences, financial management, and manufacturing (Figure 4). It is particularly useful in determining N-degree relationships, data mining, optimizing data management, advanced search, and pattern matching. Here are several common scenarios.

Sense and Respond

Significant effort is being applied to efforts in the military and intelligence communities to create “sense and respond” systems. This requires the ability to cull through massive amounts of data, looking for activity patterns in what often appears at the surface to be unrelated elements. Node/ arc relationships between such elements enable researchers to establish connections that may only be related through several degrees of separation.

For example, a late return on a rental truck may appear to have no relevance until it is connected to a driver who works at a plant where another worker who has made foreign cell phone calls also has a roommate who recently purchased fertilizer. This kind of timely, context-related information provides organizations with the ability to sense potential change or disruptions and quickly respond to threats, whether they are market, competitive, or terrorist in nature.

Static Data Analysis

Organizations often have stockpiles of aged data. In many cases, this information is used to develop predictive models in the wake of actual events. Applying relationship analytics to this stockpile of existing data enables a new dimension of analysis and visualization to identify converging factors or discover trends. At an abstract level, graph technology provides the ability to fully re-index a stockpile of data for easier access to previously obscured patterns.

For example, data can be ingested based on a series of different models that show relationships and context. The ability to look at production timing statistics with people as process checkpoints in one model and machines as process checkpoints in another can help easily identify process flow bottlenecks. Given a known defect, graphs help quickly determine the set of connections that may have been the source of the defect.

Enterprise Data Optimization

The work of normalizing data in relational tables and creating queries for reporting is non-trivial. With graph-based relationship analytics, this effort can be significantly reduced. The process of modeling (determining classes, attributes, instances, and relationships) is simpler than relational table normalization. In addition, if new attributes or relationships are added after the fact, there is no requirement to restructure tables or queries.

A human resources database could be modeled with individuals as nodes, teams as nodes, and arcs defining an employee’s superiors, subordinates, and team members; arc types would indicate the nature of the relationship. Adding a new variable such as workplace location does not require restructuring tables or reformatting data. A new location node is created with a “works-at” arc type linking the appropriate individual nodes to the new location.

Navigation and search

Instead of retrieving information based on a specified query or range of values, graph-based relationship analytics allows search and navigation according to patterns. A simple query could be, “locate all of the nodes of class ‘student’ that are connected to nodes of class ‘dorm 20’ with an arc type of ‘lives-at’.” A more complex pattern match could extend this simple query to include a chain of connections and a pattern of chains with constraints based on attribute values and arc types.

Searches of this type enable an analyst to search for a pattern such as, “locate individuals connected through bloodlines that have had exposure to religious radicalism at a certain site, are attending flight school, and have had recent foreign account deposits.” The actual structure of the relationships does not need to be known in advance in order to determine connection—graph query facilitates linking through an undetermined number of connections constrained by a given set of matching attributes.

Actionable intelligence

A complete analysis requires the fusion of large amounts of disparate data. This provides a complete picture for the analyst, but it also tends to obscure the information sought. With graph-based relationship analytics, the data is brought into an easy query format that enables the segregation of relevant information from the much larger collection of irrelevant. This enables the analyst to get to actionable intelligence quicker and with greater accuracy.

WHAT IS RELATIONSHIP ANALYTICS?

Relationship Analytics greatly simplifies discovery while unleashing extraordinary analytical power. Relationship Analytics is fundamentally different from classic search—it requires no hypothesis and enables the analyst to follow connections wherever they lead and evaluate their relative strength and affiliation.

Relational database management systems (RDBMS) and their related data mining tools are basically about data storage and extraction. This is very useful in applications such as payroll, where the data is used in a predictable way. But what if that payroll information could provide a critical clue in an internal investigation? How could the data be queried in new and unpredictable ways to find patterns and links with other data in other systems?

Relationship Analytics employs many techniques not typically associated with RDBMS and data mining solutions.

Link Discovery

The first step in an investigation is typically to determine if two entities are related in some as yet unknown way. This relationship could be through one or more connections such as family lines, business associations, phone calls, or locations.

Link Analysis

Once a connection has been identified, Relationship Analytics lets the analyst evaluate how strong or weak it is, and whether it is unique.

Path Analysis

Sometimes it is important to determine the probability of events occurring within a given timeframe. Path analysis helps define that probability.

Group Finding

Once relationships are identified it is possible to find groups based on inferred relationship by association. This is a technique for finding hidden members of known groups, and detecting individuals tied to events of interest.

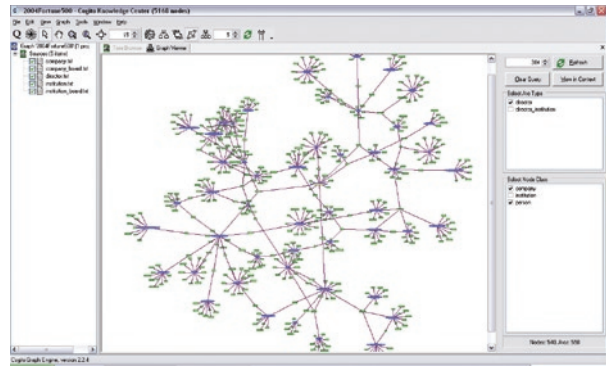


Figure 1: An analysis of companies (blue rectangles) and board members (green rectangles) shows that some people serve on multiple corporate boards, creating a network of relationships.

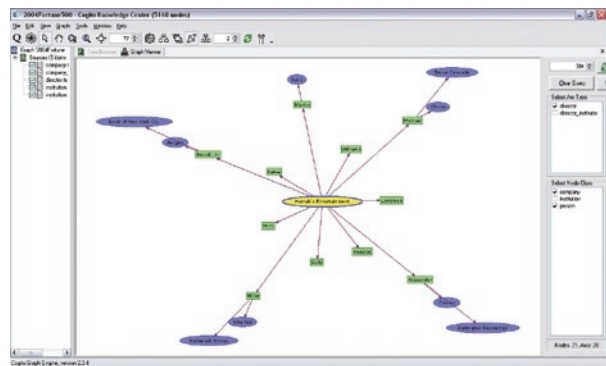


Figure 2: Some board members of Harrah's Entertainment (green rectangles) also serve on the boards of other companies (blue rectangles).

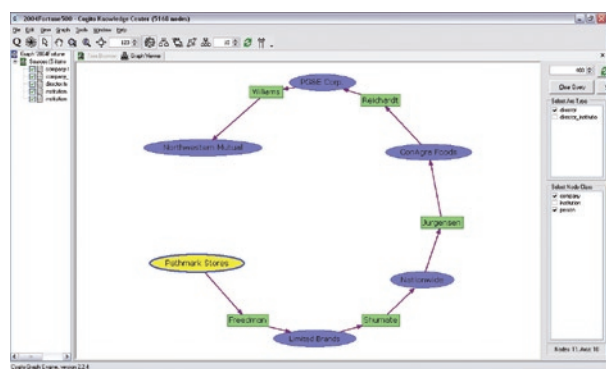


Figure 3: A chain of corporate directors connects Pathmark Stores and Northwestern Mutual.

Pattern Matching

Humans are creatures of habit, leaving patterns as clues to behavior. By identifying a pattern, an analyst can then use Relationship Analytics to search for that pattern across massive data sets. Information can be viewed and filtered visually (specific relationships turned on/off at varying degrees of connection) for immediate observation of patterns. This visual query capability allows analysts to see at a glance relationships that exist within a data set.

Data can also be queried and filtered using any of several query options including shortest path between nodes, common nodes, and sub-graph pattern matching. Full and partial sub-graph pattern matching is particularly powerful for discovering relationships between what may appear as unrelated data (a common social network analysis challenge). Information is then available for further analysis, pattern discovery, reporting, and distribution as needed.

Pattern Detection

Just as an analyst may not always have a hypothesis, that analyst may not have a pattern. Relationship Analytics can look for patterns of activity, networks and transactions. Modeled relationships are retained as part of the stored data structure and as a result, patterns may appear that were not anticipated. Simple visualization of data can produce vivid illustrations of patterns showing concentrations of activity, visually traceable connection paths, and multiple degrees of connectivity.

Data stored in relational tables tends to lose a portion of its structural integrity. The only way to determine trends or to extract information is to author a specific query that takes into account the known data structure. In effect, one only gets out of a relational database what one knows to ask.

Pattern Analysis

This higher level of analysis evaluates the pattern itself and its consistency, replication and condensability. This can help determine how prevalent a pattern is, what other patterns it links with and whether similar patterns exist on a different level.

Cluster Analysis

Relationship Analytics often detects clusters of people or activity. Analysis of these clusters can find information brokers, transmitters, ringleaders, and boundary spanners in gangs, drug rings, terror networks and organized crime.

Cycle and Flow Analysis

Using Relationship Analytics, an analyst can test for cycles and breaks in cycles. Maximum, minimum and minimum cut flows through a cyclic network can also be calculated.

Alias Detection

This solves the problem of identifying labels or names that may be intentionally or unintentionally associated with a specific entity. This is useful for finding terrorists, criminals and spies who are using other than their real names. This uses social network information to establish context identity.

Inference Establishment

This capability enables the analyst to make assumptions in the data and analyze from there. For example, all inmates in the same cell block over the same time period can be assumed to 'know each other'.

ADVANTAGES OF GRAPH-BASED RELATIONSHIP ANALYTICS

Graph technology has several advantages when addressing today's expanding data challenges, including relationships in context, faster query performance, flexible data organization, and hidden pattern discovery.

Relationships in Context

Data must (to a certain extent) be 'destructured' in order to place it in traditional relational database table form. This process includes breaking apart inherent data relationships to form separate tables, rows, and columns in an effort to minimize the amount of replicated data. Recovering a specific desired data set requires the creation of a structured query language (SQL) statement.

In sophisticated data sets, the creation of multiple tables plus the joins and queries required to extract desired information can be very complex. Analysts or database administrators must know, to a fairly precise degree, the structure of the information they are filtering to get the specific results they seek. Even then the queries will take some time to complete given the recursive nature of a deep query across multiple joins.

Graph databases eliminate the need for related tables, joins, and keys. Depending on the data mapping schema, data can be fused with all relevant relationships already established and linked. The destructure/restructure exercise is eliminated with no need to create normalized tables or cubes, or to construct queries based on these different tables. In effect all joins are pre-computed.

Faster Query Performance

Extracting precise information can be time consuming and expensive when working with complex data sets. Administrators using relational database technology strive to optimize queries across multiple tables, but even this often involves iterative cycles for filtering out irrelevant information and structuring statements that reduce the answer set based on ordered sequences. Because of this, relational queries through chained data are often limited to four or five connection levels. In many cases, a four or

five degree search becomes unmanageable, overly time consuming, and requires additional hardware and software.

Queries when using graph-based relationship analytics are significantly simpler, with the ability to traverse through data that was never destructured to fit in tables. To a large degree, data in a graph follows its natural pattern of existence and/or the model with relevant information related through close association. This pattern follows even as the data is committed to disk.

To illustrate, assume a large data set with records indicating parent-child relationships but no extended family relationships. The objective is to find a common ancestor among two individuals who are not known to be related. To search parents on both sides going back seventy-two generations requires a search with 272 iterations. Given standard server class hardware and relational database technology, the problem can take hours. The same exercise with graph technology can be performed in seconds—the operation is a simple node walk to find a common ancestor. Graph-based relationship analytics makes parent/child relationships inherent in the data structure and closely located through arcs. This same query performance is possible with any type of related information.

Flexible Data Organization

With graph-based relationship analytics, if a new variable is added, the model is adjusted to accommodate the new variable. New node and arc definitions are added and all succeeding data imported to the graph will include the new association.

With relational tables, the schema is static. Making a change to a relational database schema involves restructuring the data with (at minimum) the addition of another field and the rewriting of all related queries. Schema changes are often accommodated using brute force by creating a new table with more replicated data and new queries. The net advantage is that database maintenance and administration costs can be reduced by a factor of up to 60 percent for the average database.

SUMMARY

The benefits of graph-based relationship analytics can be applied to any organization in any field. For all applications, Cogito solutions provide faster answers, more comprehensive analytics, increased responsiveness to customers, reduced total cost of ownership, optimized database resources, and overall operational excellence.

Cogito's Graph-Based Relationship Analytics products allow organizations to leverage their existing data structures and applications and identify previously unknown links and sub groups buried deep in generations of structured and unstructured data. The Cogito Knowledge Center provides a robust framework of application services that allow businesses to improve their data analytics and decision-making capabilities by overcoming the limitations of SQL and relational database technology. The Cogito Knowledge Center allows a dynamic schema that can be updated, changed or reorganized at a moment's notice to reflect an ever-changing business environment.

More information on Graph-based Relationship Analytics and the Cogito Knowledge Center is available at:

Cogito Incorporated

170 West Election Road Suite 205
Draper, Utah 84020

801-858-1000
www.cogitoinc.com
sales@cogitoinc.com