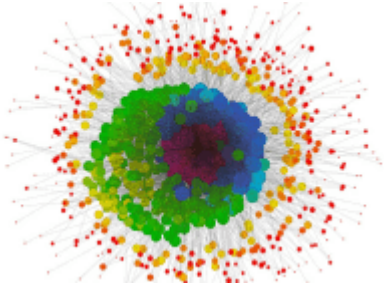
Datanami article – Multi-Dimensional Graph Data Opens the Door to New Applications

David S. Frankel



As the use of graph databases has grown in recent years, ever more applications of this technology involve storing, searching, and reasoning about events. In fact, many companies use this technology for this purpose, and the size of these databases is rising in many cases to billions of events. Now, there is advanced technology which overcomes performance problems that emerge when searching and reasoning over event databases of such size.

The kinds of events that graph databases manage typically have at least the following elements:

- A type, such as a phone call, a text message, a bank transaction, an observation of a moving vehicle, and so on
- A start and end time, or a single instant of time (the temporal dimensions)
- Location coordinates (the geospatial dimensions)
- A set of actors, such as sender and receiver, payer and payee, vehicle and operator, and so on (the social network dimensions).

Some important advances in semantic graph technology have improved the ability to store, search, and reason about geospatial, temporal, and social network data. However, the

advances in these three areas remained quite isolated from each other for some time.

The next logical step was to bring these advanced capabilities together to support searching and reasoning over records that combine all of these dimensions. Recent technical innovations continue the progress along this path by enabling highly efficient applications dealing with vast amounts of such multi-dimensional data. Diverse kinds of applications can benefit by harnessing this newly available power, such as tracking moving objects in time and space, managing weather data, detecting fraud or other criminal activity, and more.

Key Kinds of Data and Reasoning

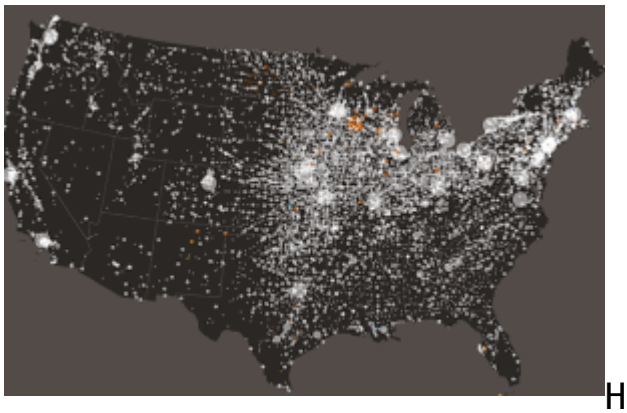
Before we delve into unified multi-dimensional facilities, it is useful to summarize the key characteristics of geospatial, temporal, and social network data and the kinds of reasoning we do about such data.

Geospatial Data

Geospatial data is about location in space. Two-dimensional geospatial databases describe location in terms of latitude and longitude or in terms of x- and y-axes on a grid. Three-dimensional geospatial databases might add a dimension for altitude, height, or simply a z-axis in a 3D grid.

We can ask the following questions about locations and shapes that we have stored in a database:

- What are all the events that occurred within a specified radius of a given location?

-  How far are two given locations from each other?

With prior technology, support for searching and reasoning about geospatial data was limited to two-dimensional coordinates. More advanced geospatial technology supports three-dimensional coordinates. As an example of the practical ramifications of this enhancement, tools based on the newer technology can search for and reason about objects moving through three-dimensional space, such as airplanes; whereas previous versions could only deal with objects moving through two-dimensional space, such as automobiles.

Temporal Data

Temporal data is about time. Key questions we ask about time often have to do with time intervals. Given two time intervals, we can, for example, ask the following questions:

- Does one interval occur entirely before the other?
- Do the intervals meet (meaning one interval starts where the other interval leaves off)?
- Do the intervals overlap?

Temporal data technology generally supports this kind of reasoning about time intervals. However, typically when searching event databases we are simply looking for events that occurred within a specific time interval.

Social Network Data

Social network data captures connections between actors, such as the fact that one person is a friend of another. But social networks do not have to be about people; social network technology is proving useful in other fields such as life sciences, where, for example, researchers study protein interaction patterns as social networks in which the actors are proteins.

We can ask the following kinds of questions about a group of actors and the connections among them that we have stored in a database:

- How far apart in the network are two given actors, and how strong is the relation?
- What are the cliques and ego groups?
- How important is a given actor in the group?
- How cohesive is the group?

Multi-Dimensional Graph Data – Bringing it All Together

Reasoning about each of these kinds of data as described above is useful, but, as we have seen, event databases require combining these facilities.

Geospatial, Temporal, Social Network, and More

For example, a log of cellular phone calls may well include the following data for a given call:

- The latitude and longitude of the call originator's location and receiver's location
- Start time and finish time
- The calling and receiving phone numbers.

In this case, the call log entries in the database – which are event records, where a phone call is an event – clearly have both geospatial and temporal data. We can use such data to answer questions such as:

- What calls have an originating location within a given radius of a given location within a given time interval?
- Did a given phone number place a call within a given radius of a given location within a given time interval?
- What calls to a given area code were made by a given phone number within a given time interval?

Figure 1 is a snapshot of a screen from a phone call app

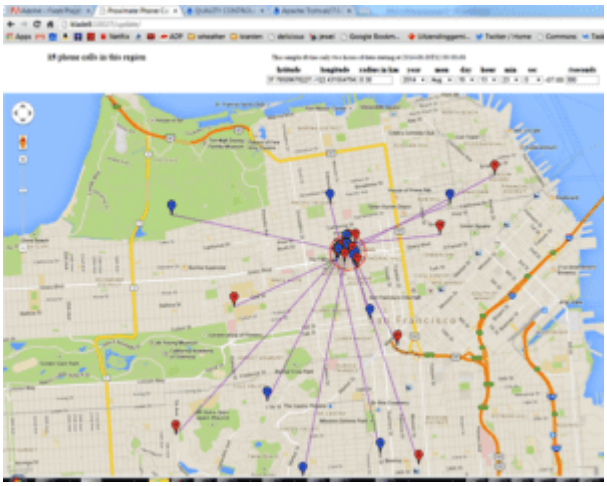


Figure 1: Phone Call Tracking Application

ic
at
io
n
of
th
is
ge
nr
e.
Th
e
ap
pl
ic
at
io
n
di
sp
la
ys
a
Go
og
le
ma
p.
Th
e
us
er
se
ts
th
e
ra

di
us
an
d
da
te
/t
im
e
in
te
rv
al
an
d
cl
ic
ks
on
a
lo
ca
ti
on
on
th
e
ma
p.
Th
e
ap
pl
ic
at
io
n

th
en
di
sp
la
ys
th
e
lo
ca
ti
on
s
of
ca
ll
er
s
an
d
re
ce
iv
er
s
of
ca
ll
s
th
at
or
ig
in
at
ed
wi

th
in
th
at
ra
di
us
wi
th
in
th
e
da
te
/t
im
e
in
te
rv
al

For an additional enhancement, we could construct a social network of connections between phone numbers, making it possible to pose questions such as finding the phone numbers (and, implicitly, their owners) that are the most central for phone call traffic in a given radius of a given location for a given subgroup of phone numbers.

Consider another example, where we store observations of airplanes moving through space. At regular periods we record the latitude, longitude, altitude, and heading of flying airplanes and we time stamp each observation. This data enables us to ask questions such as how many airplanes were within a given altitude range with a given heading during a given time interval.

N-dimensional Data

Clearly, databases that support these scenarios are multi-dimensional. They have geospatial dimensions, temporal dimensions, and social networking dimensions. Moreover, there are strong use cases for adding additional dimensions to such databases. For example, in the case of airplane tracking, each time-stamped observation may also include weather readings such as outside air temperature, wind speed, and barometric pressure.

As mentioned earlier, previous technology supported searching and reasoning over two-dimensional geospatial data, whereas more recent technology supports three-dimensional geospatial data. But new technology, such as AllegroGraph version 5, can search and reason over an open-ended number of additional dimensions. Thus these new facilities are not merely three-dimensional, because there is no restriction to three dimensions. It is more accurate to use the term N-dimensional to describe the nature of graph databases and related applications that use these new facilities.

Efficiency Breakthrough

The idea of combining geospatial, temporal, social networking, and other dimensions in a database record is not new, but up to now implementation of this idea has been limited. The roadblock has been serious performance degradation as multi-dimensional databases grow to enormous sizes. Despite the fact that graph databases have known efficiency advantages over relational databases for dealing with geospatial, temporal, and social network data, simply using a graph database is not enough to get over the performance hurdle with gigantic multi-dimensional databases. The performance hit is most severe when search parameters are about proximity, such as searching for events that occurred within a specified radius of a given location or within a given time interval, or within a temperature range.

But performance is another area that has recently seen the addition of innovative technology that can answer complex proximity questions across multiple dimensions over billions of records in sub-second time. A key characteristic of this new technology is that, with proper database design, the time required to execute a search does not increase substantially as the size of the database increases.

Highly Scalable Applications

We are just beginning to tap the potential of this powerful technology. Here are a few examples of what we can do with the new N-dimensional search and reasoning capabilities:

- Insider Threat Detection: Quickly identify risks and the potential impact that an individual's actions pose to the public or an organization. New semantic-based behavior models can empower companies to gain the critical knowledge necessary to predict high-risk events to prevent or aid in crisis situations.
- Precision Medicine: Integrate information from structured and unstructured data (and integrate different types of data – patient information with socio-economic and genetic information, etc.) to improve efficiencies and personalize care. Provide graphical analysis of genetic info, images, clinical trials, and public health data to help fuel discoveries, improve patient care and cut the overall cost of healthcare.
- Law Enforcement/Homeland Security: An application tied to a constantly updated database of telephone calls and text messages could use the location data, the time stamps, and the social network represented by the phone numbers to determine the focal points of a criminal enterprise and monitor the movements of the key actors in near real time as their centrality emerges from the data.

As the need to manage vast numbers of event records increases and organizations begin to understand the implications of high capacity multi-dimensional graph database technology, people will think of all sorts of applications that even the creators of the technology have not contemplated. We may one day look back on the emergence of this technology as an inflection point that took us to a new level of powerful data management.

About the Author: David S. Frankel has over 30 years of experience in the software industry as a technical strategist, architect, and programmer. He is recognized as a pioneer and international authority on the subject of model-driven systems and semantic data modeling. David has made major contributions to a number of industry standards, including XBRL, ISO 20022, BIAN, and UML. You can read more at his website: http://www.dfrankelconsulting.com