August 21, 2009

Social network analysis, aka relationship analytics

- Home
- About
- Contact
- Feeds

A number of applications lend themselves to graph-oriented analytics, including:

- Finding bad guys (national intelligence)
- Finding bad guys (anti-fraud)
- Data mining the social graph (e.g., for advertising optimization on social networks, or to identify influencers)

There are plenty more graph-oriented applications, of course, such as the identification of biochemical pathways. But I want to focus for now on ones like those on my list. My key points are:

- **There are Big Data problems that lend themselves to graphical data models.**
- So far as I can tell, **the database management community isn't doing enough to address them.** (If I'm wrong about that, please tell me. I plan to arrive in Lyon for VLDB/XLDB Wednesday of next week, and of course I can always be reached by email.)

Here's what I mean.

Applications that analyze relationship graphs are commonly grouped under the name *social network analysis*. As I frequently point out, category names and definitions tend to be imperfect, and that one is no exception. In particular — and the Wikipedia article on social networks and social network analysis is an excellent example of this – the term tends to be construed to cover the linkages between people or organizations, but not between, say, physical addresses, email addresses, and all the other stuff those intelligence applications actually track. I tried to introduce the term relationship analytics a while back, but it unfortunately didn't stick.

I only ever got familiar with one company that tried to do a true graph-oriented database management system, suitable for social network analysis/relationship analytics. It was called Cogito, and had some interesting ideas about graphical data structures. Unfortunately, Cogito didn't stick either.

As per the "Metrics" section of the Wikipedia article linked above, there are a number of well-established metrics about the relationships of pairs or groups of node to each other. The usual way to calculate these metrics is to load the graph into memory and get to work. (Indeed, such uses seem to be driving a lot of the national intelligence adoption of Hadoop.) And while I'm perfectly willing to believe that relational database management systems can do a fine job of managing generic RDF, it's less obvious that they're well-suited to support standard graph-analysis metric computations.

The reason, in a nutshell, is that the relational approaches usually boil down to maintaining a table with a row for every node-edge-node triple, and then doing a lot of fast self-joins to identify paths. That can work if connectivity is low and paths are sparse. But for higher degrees of connectivity, such strategies can lead –

BOOM! — to serious combinatorial explosion. And that's not good, because a lot of this analysis focuses on finding exactly the parts of the graph where the connections run thickest.

Cogito's idea was to say "What if, for every node, you could retrieve in only a few blocks all the paths leading from it, at least up to pathlength N?" Unfortunately, Cogito's approach to creating this effect had too little to do with optimizer development or selectively redundant data storage, and too much to do with wishful thinking; not coincidentally, Cogito is no longer around. (I haven't kept in touch with Cogito's successor 7 Degrees, and the reason hasn't been lack of effort or interest on my part.)

But suppose the idea had worked. Then – unlike today – it might be realistic to do on-the-fly analytics on Very Large Graphs, just as we do operational business intelligence of a more relational or MOLAP nature. That would be cool.

How cool would it be? Well, that's a bit hard to say. Look again at the list of applications I put up top. Those are NOT ones people generally talk a lot about. Spooks and fraud-fighters are two very secretive kinds of folks. And, for a variety of reasons, the owners of the largest websites also are reluctant to publicize details of how they do or don't profile individual users in vivid detail. And then there's also the question of whether we even want to help improve technology whose main use is to improve the precision with which computers track individuals – but I don't think that's the front on which the privacy wars are best fought.

But if I were a computer science researcher right now, graph databases – optimized to support graph-analytic metrics — are one of the areas I'd look at to see if I could make an impact.

Categories: Analytic technologies, Cogito and 7 Degrees, Data models and architecture, Data types, RDF and graphs, Theory and architecture
Subscribe to our complete feed!

Comments

16 Responses to "Social network analysis, aka relationship analytics"

1. dave on August 21st, 2009 7:31 am

   there was a time when both spoke and visible path were doing well (with funding) selling social network analysis software to the enterprise, and there was interest from areas such as hr, sales, etc – and even the privacy issues had been resolved/addressed – but in the end, i more than agree, there was no serious database thinking around these orgs, but oddly, this all lends itself more to hellerstein's paper (with greenplum/FAN) that talks about the changing role of the analyst…sounds like an opportunity for a BI vendor perhaps, or on a different pass, an opportunity for a CRM vendor (think salesforce) to bring in such thinking to expand the offerings – far beyond the traditional linkedin approach (which in itself is a decent first step, though the FB friend analysis is pretty decent as well)…7 degrees (aka "people maps") strikes me as just another iteration of wasserman's work behind visible path…

   oh, and beyond your list (catching bad guys, etc), i'd add: modeling the spread of infectious disease (a huge usage case, think WHO and NIH and CDC – they do use it)

2. Jesse on August 21st, 2009 9:58 am

   Great article, I think you might be interested to see the kind of work my research group does at Carnegie Mellon. While we don't have anyone taking a database perspective on this we do a ton of work with the computational aspects of SNA and Network Science. We also pay particular attention to multi-modal data. You can check us out at http://www.casos.cmu.edu

Cheers.

3. Curt Monash on August 21st, 2009 10:12 am

   Hi. That link isn't working well. Googling gets one to your group pretty quickly, but it's not immediately obvious which projects you're referring to in which parts of your comment. 😉

4. Jesse on August 21st, 2009 10:46 am

   Yeah, I tried to post a correction but I think the spam filter got me =/. The actual address is: http://www.casos.cs.cmu.edu

   Sorry about that.

   ORA is our most popular network analysis tool; that is probably a good place to start: http://www.casos.cs.cmu.edu/projects/ora/

   Aside from this we develop a good deal of new and interesting computational analysis and simulation techniques. The different research arcs are better represented in the publication list rather than the project list, as many of the research arcs get integrated into our tools as they mature.

   One of the reasons we don't currently look at graph databases is that for our current purposes sparse/dense matrix representations are sufficient and, more importantly, fast. Though as the size of our data increases we will have to pursue efficient graph database representations. If we get someone interested in the subject I wouldn't be surprised to see that dissertation come out of our group in the next few years.

5. Jeff Hammerbacher on August 21st, 2009 11:43 am

   Hey Curt,

   At Facebook, our group used Hadoop for some fairly sophisticated social network analysis. The team published two academic papers detailing their work, one of which won the best paper award at this year's ICWSM. See http://cameronmarlow.com/papers for the details.

   It's not an uncommon use of Hadoop: Jake Hofman of Yahoo will be speaking about his work on SNA with Hadoop at the upcoming Hadoop World NYC conference, and there's even an open source package for SNA with Hadoop up on Sourceforge: http://sourceforge.net/projects/xrime/.

   Some of the more interesting work is being done at Carnegie Mellon in the "Peta Graph Mining Project (PEGASUS)": http://www.cs.cmu.edu/~ctsourak/projects.html.

   While Hadoop is great for some kinds of large graph analysis, Google has developed Pregel as a more efficient system for iterative computations on graphs. See http://googleresearch.blogspot.com/2009/06/large-scale-graph-computing-at-google.html for more.

   Glad to see you casting light on this subject!

   Regards,
   Jeff

6. Zman on August 21st, 2009 12:24 pm

For a non-RDBMS non-SQL alternative you can look at neo4j

7. Michael E Driscoll on August 21st, 2009 1:12 pm

   Toby Segaran and Jamie Taylor of Freebase — which is developing what may become the world's largest graph database — just published a book on working with graph data with O'Reilly.

   Programming the Semantic Web

8. Kim Rees on August 21st, 2009 2:02 pm

   Hi,

   Great post. I think this topic deserves a lot of discussion. I would love to see more work being done in this area.

   The few bits I know of are:
   1) A guy from OLAF gave a presentation at the Tableau conference last month that described their anti-fraud efforts:
   http://conference.tableausoftware.com/speakers.html#jurgen

   2) i2. I read an article about these folks once that was a case study of thwarting crime by data analysis.
   http://www.i2inc.com

   3) I know that Interpol does a lot. Of course, it's top secret, I'm sure.

   I would love to see a wider discussion of this. Thanks for bringing it up.

9. Valdis Krebs on August 21st, 2009 2:42 pm

   Here is a page full of various cases of social network analysis applied both good guys and bad guys…

   http://orgnet.com/cases.html

   And this blog about network analysis applied to various business issues

   http://thenetworkthinker.com

10. Hans on August 22nd, 2009 2:21 pm

    I would also prefer a term like relationship analytics. The graph structure is but one part of analyzing relationships.

    Constructing and updating the graph is usually not as easy as Facebook, where the edges are conveniently input by the users. Many methods for identifying relationships among disparate data do not rely on graphs structures per-se.

    A strategy for analyzing something like crime or national security would most certainly rely on mixture methods, some of which are explicitly graph oriented and some are not.

    In applications like this, you will see long job chains, for example: reducing text data into computable summaries, running analytics over the computable data to produce graph structures, then analysis of the graph structures.

It will be interesting to see if and how this kind of thing is supported under a single product.

11. Teradata's Active Enterprise Data Warehouse story | DBMS2 -- DataBase Management System Services on August 24th, 2009 4:35 am

    [...] Teradata was able to quickly cite examples in both social network analysis and anti-fraud — but not in the use of social network analysis for law enforcement or fraud [...]

12. Getting Closer to Real Time With Hadoop | Digital Asset Management on September 21st, 2009 2:49 am

    [...] Social network analysis, aka relationship analytics (dbms2.com) [...]

13. Warren Davidson on December 11th, 2009 11:30 am

    Very interesting article. As our company has had experience in dealing with graph data, and addressing the issue of multiple degrees of separation I agree this is needed in the market but not well addressed. In fact looking at how some people are going about addressing this for social networking looks far more complicated than it needs to be.
    Here is a link to one small demonstration of what we are doing for an agency currently, with very insetting results.

    http://www.objectivity.com/media/link-hunter/default.asp

14. Open issues in database and analytic technology | DBMS2 -- DataBase Management System Services on February 1st, 2010 6:05 pm

    [...] increasingly persuaded that graph analytics can be handled without a graph-centric data model. But right now, it isn't being handled well [...]

15. Social Network Analysis Software - Topic Research, Trends and Surveys on February 3rd, 2010 5:26 pm

    [...] systems. SNA helps … Read More RECOMMENDED BOOKS REVIEWS AND OPINIONS Social network analysis, aka relationship analytics | DBMS2 … There are plenty more graph-oriented applications, of course, such as the identification of [...]

16. David Ingersoll on April 8th, 2010 5:16 pm

    We have the original graph oriented database at Versant, which is used for all the above use cases. We own the network and element management space at customers, Alcatel-Lucent, Avaya, Ciena, Ericsson, NEC, Nokia-Siemens, Samsung. Verizon uses Versant for their Fraud Detection system in conjuction with the NASA Open Source Rules Engine, 400 Million CDR's processed in 10 hours.

Leave a Reply

Name (required)

Email Address(required)

Website

Submit Comment

Subscribe to the Monash Research feed via RSS or email:

Enter address here Subscribe! Login

Search our blogs and white papers

Search

Monash Research blogs

- DBMS2 covers database management, analytics, and related technologies.
- Text Technologies covers text mining, search, and social software.
- Strategic Messaging analyzes marketing and messaging strategy.
- The Monash Report examines technology and public policy issues.
- Software Memories recounts the history of the software industry.

User consulting

Building a short list? Refining your strategic plan? We can help.

Vendor advisory

We tell vendors what's happening -- and, more important, what they should do about it.

Monash Research highlights

Learn about white papers, webcasts, and blog highlights, by RSS or email.

- **Recent posts**

- Fun with quotes in the VectorWise press release
- The most important part of the “social graph” is neither social nor a graph
- Algebraix
- Extended set theory, aka “What is a tuple anyway?”
- VoltDB finally launches

Categories

- About this blog
- Analytic technologies
    - Business intelligence
    - Data mart outsourcing
    - Data warehousing
    - MOLAP
- Application areas
    - Games and virtual worlds
    - Investment research and trading
    - Log analysis
    - Scientific research
    - Telecommunications
    - Web analytics
- Buying processes
    - Benchmarks and POCs
- Companies and products
    - 1010data
    - Ab Initio Software
    - Akiban
    - Aleri and Coral8
    - Algebraix
    - Alpha Five
    - Amazon and its cloud
    - ANTs Software
    - Aster Data
    - Business Objects
    - Calpont
    - Cassandra
    - Cast Iron Systems
    - Clearpace
    - Cloudera
    - Clustrix
    - Cogito and 7 Degrees
    - Cognos
    - Continuent
    - CouchDB
    - DATAllegro
    - Datameer
    - Dataupia
    - Elastra
    - EMC
    - EnterpriseDB and Postgres Plus

- Exasol
- Expressor
- FileMaker
- Gooddata
- Google
- Greenplum
- Groovy Corporation
- Hadoop
- HP and Neoview
- IBM and DB2
  - pureXML
- illuminate Solutions
- Infobright
- Informatica
- Information Builders
- Inforsense
- Ingres
- Intel
- Intersystems and Cache'
- Jaspersoft
- Kalido
- Kickfire
- Kognitio
- Mark Logic
- McObject
- memcached
- Microsoft and SQL*Server
- MonetDB
- MySQL
- Netezza
- Northscale
- Oracle
  - Exadata
- Oracle TimesTen
- ParAccel
- Pentaho
- Pervasive Software
- PostgreSQL
- Progress, Apama, and DataDirect
- QlikTech and QlikView
- SAND Technology
- SAP AG
- SAS Institute
- ScaleDB
- SciDB
- SenSage
- Software AG
- solidDB
- Splunk
- StreamBase
- Sybase

  - Tableau Software
  - Talend
  - Teradata
  - Tokutek
  - Truviso
  - VectorWise
  - Vertica Systems
  - VoltDB and H-Store
  - Xkoto
  - XtremeData
- Data integration and middleware
  - Application servers
  - EAI, EII, ETL, ELT, ETLT
- Data types
  - GIS and geospatial
  - Object
  - RDF and graphs
  - Structured documents
  - Text
- DBMS product categories
  - Archiving and information preservation
  - Data warehouse appliances
  - Mid-range
  - OLTP
  - Open source
- Emulation, transparency, portability
- Fun stuff
  - Humor
- Liberty and privacy
- Market share
- Memory-centric data management
  - Cache
  - Complex event processing (CEP)
  - In-memory DBMS
- Michael Stonebraker
- Parallelization
  - Clustering
  - MapReduce
- Presentations
- Pricing
- Software as a Service (SaaS)
  - Cloud computing
- Specific users
  - eBay
  - Facebook
  - Fox and MySpace
  - TEOCO
  - Yahoo
- Storage
  - Solid-state memory
- Theory and architecture

    - Columnar database management
    - Data models and architecture
    - Database compression
    - Database diversity
    - NoSQL
    - Petabyte-scale data management
  - TransRelational

- **Date archives**

  Select Month

- Links
  - Monash Research
  - Webcasts
  - White Papers

- **Admin**
  - Log in

- Home
- About
- Contact
- Feeds