June 8, 2010

The most important part of the “social graph” is neither social nor a graph

- Home
- About
- Contact
- Feeds

“Social graph” is a highly misleading term, and so is “social network analysis.” By this I mean:

There’s something akin to “social graphs” and “social network analysis” that is more or less worthy of all the current hype – but graphs and network analysis are only a minor part of the whole story.

In particular, **the most important parts of the Facebook “social graph” are neither social nor a graph.** Rather, what’s really important is an aggregate **Profile of Revealed Preferences**, of which person-to-person connections or other things best modeled by a graph play only a small part.

Let me hasten to note that – even when viewed narrowly — the ideas of “social graph”and “social network analysis” do have significance. Nontrivial use cases to date for big data social network analysis include:

- Intelligence agencies identify and analyze terrorist networks. Corporations and civilian law enforcement do the same for fraud networks.
- Telephone companies use calling data to figure out which of their customers are most likely to influence which other customers in the decision to keep or change service providers. (Frankly, I find that rather creepy.)
- Social networks figure out which other members you’re likely to know, and encourage you to connect with them.

Epidemiologists aspire to add to that list, based on their success to date using much more micro forms of social network analysis. But after that, I’m running out of examples. Sure, graph analytics is good for a bunch of other things (e.g., biology at the genetic or molecular level), but those have little or nothing to do with “social graphs” or social network analysis as they are commonly understood.

Note: Of course, it is also the case that everything can be modeled by entity-attribute-value triples, and those can always be modeled by graphs. But so what?

Let’s consider what, in a marketer’s ideal world, would go into your Profile of Revealed Preferences. Raw data might include:

- **Personally identifyING information.** Duh. This is what makes everything else possible.
- **Purchase transaction data.** Lots of it. Like, everything on your credit card statements.
- **Demographic and lifestyle information.** Address, date of birth, educational history, race, household composition, and so on.
- **Affiliations.** Politics, religion, group membership of any kind. (OK, that’s partly social.)
- **Explicitly stated opinions, preferences and desires,** including:
    - Any surveys you have filled out.

  - **Any recommendations you have made** (e.g., through the Facebook Like feature).
  - The text of anything you've written and posted – and, very ideally, of your private emails as well.
  - Any **wish lists** you've filled in.
- **Attention information.** What you clicked on, what you looked at, and all that stuff website owners track.
- **Your movements,** to the extent they are tracked. (E.g., via Foursquare and the like.)
- **Your gaming activities** and the like. (This is social mainly to the extent it overlaps with other categories I've already mentioned.)
- **Your medical information.**
- **Who you communicate with, and what you communicate with them about.** (Hey! There's something else social!)
- Similar **information about the people you communicate with.**

My core **privacy** thoughts about that data include:

- **Individuals deserve the right to control all that information.** At a minimum, they deserve total control over how the data (raw or derived) is passed from the service – e.g., website – where it naturally resides (e.g., where it is originated) to any other place.
- Given a chance, **individuals would make fine-grained choices about what parts of their Profile of Revealed Preferences are available to which organizations.** Reasons include:
- Individuals have rather complex trust relationships with different kinds of merchants and marketers.
- Consumers get different benefits from sharing information with different kinds of merchants and marketers. (Sometimes personalization is a large benefit. Sometimes it's just creepy. And some companies actively bribe you to give them information they can use to sell to you.)

When one frame things this way, two rather difficult technological questions naturally arise.

1. Suppose, implausibly, that a single entity were allowed to control and use (for marketing) all of your Profile of Revealed Preferences information. How would they store and analyze it?
2. How does the answer to #1 change because control over the information will, in fact, be fragmented?

It's tough enough to answer these questions for data about one person. Trying to include all but the simplest information about other people is and will for years remain quite infeasible. So, for the most part, **this is not "social" information.**

It's also **not naturally a "graph."** Similarly, it is **not a good candidate for network analysis.** To see why, let me outline **why I used the name "Profile of Revealed Preferences":**

- The reason marketers want this data is, mainly, because they want to know what appeals to you, and how strongly you feel about it.
- The analytic process often entails taking explicit choices you have made, and inferring other preferences from them.
- The output of the analytic process is often one or more "scores" that then get plugged into various selection algorithms to determine what you should be shown or offered. At least implicitly, these algorithms are predicting what you will or won't respond well to.

Not much graph-like there.

This post has gotten pretty long, so I'll stop here without spelling anything else out. But questions I still hope to address down the road include:

- How should Profile of Revealed Preferences data be stored?

- Suppose we want to pass around derived results and not the raw data. How could we ever get to standards that would make such interchange realistic?
- If we only have raw data to pass around, what are the implications for privacy, liberty, and the structure of the online industries?

Categories: Analytic technologies, Facebook, Games and virtual worlds, Liberty and privacy, RDF and graphs, Web analytics
Subscribe to our complete feed!

Comments

3 Responses to “The most important part of the “social graph” is neither social nor a graph”

1. J. Andrew Rogers on June 8th, 2010 3:24 am

   Great topic with several interesting tangents that could be written on at length. A few semi-random thoughts:

   - The real value of graphs is that they can be a universal representation and access method in theory. You can seamlessly mix-and-match your data with anyone else’s. In practice, typical graph implementations scale so badly that rigid, non-universal representations and access methods are more attractive.

   - Almost no one does true graph analysis due to the difficulty of scaling operations like transitive closures. I could easily add another half-dozen industries to your list that badly want true graph analytics but can’t get the scale to make it economical. Graphs tend to devolve to a key-value store as scale requirements grow.

   - The idea of “graphs” in mathematics has a broader scope and is more exotic (and powerful) than the simple point-link-point model ubiquitously implemented. It is easy to see why this simple model is used since the more exotic models are not visualizable; you have to retreat to “data structure design by obscure mathematics” that is difficult to reason about.

   - An aspect of graphs that is not immediately obvious is that they have a direct relationship to algorithmic information theory that allows very powerful types of computational induction that can pull potent patterns out of surprisingly diffuse bits. However, this type of use case is provably intractable using trivial graph constructs (but not for exotic ones). Most people working with social graphs intuit the possibility of something like this even if unfamiliar with the mathematics.

   - If someone really figures out the “induction over exotic graphs” angle, the privacy arguments almost become moot because sufficiently competent induction can reconstruct most of it from latent bits of we unavoidably leak everywhere in our lives. Even at the level of mathematics published now, the ability to reconstruct entities from diffuse environmental bits has been getting very good, very quickly. Most people have not thought about the implications of this, they are assuming that personal information is aggregated via conventional record sharing channels rather than exotic information theoretic reconstruction. Still a mostly correct intuition but not for long.

   Social graphs are boring in large part because doing something not boring and scalable is very non-trivial. A lot of the social graph companies are hoping to be sitting on the killer data set as it gets solved but they aren’t considering the implication of my last point and the fact that all services are inherently leaky.

This comment became much too long…

2. J. Andrew Rogers on June 8th, 2010 3:46 am

   Let me add that while I'm using graphs generically here, I've seen incredibly detailed social graph models that go far beyond the friend-of-a-friend case. When trying to develop deep contextual models, the social graphs become complicated. The interactions between two individuals is naturally very dynamic and contextual, so for the kinds of behavioral prediction organizations are interested in this is important.

   A really good social behavior model is so graph-like that it is intractable for anything useful even at small scales. Yeah, we get to see the "FoaF" model as consumers, but the internal models are often accumulating more detail than that about the influence of network dynamics on behavior.

3. Alan on June 8th, 2010 4:07 am

   'But after that, I'm running out of examples.' Two examples I've come across.

   Preferential customer service – similar to the 'Telephone companies use calling data'. Based on estimated net worth of the social network, a fin services firm goes to extra-ordinary lengths to deliver customer service. You never know who will marry a Kennedy these days – and they *all* talk.

   Targeted promotions. Think travel for 'Social networks figure out which other members you're likely to know'. Based on past history and geography of the social network there's a limited time offer for discounted travel rates. This isn't new, it's just one more source of data for targeting.

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