Mining Massive Graphs for Telecommunication Applications

Chris Volinsky
AT&T Labs-Research

Workshop on Mining and Learning with Graphs
July 25, 2010

Outline

- Telecommunications Network Traffic
- Current Research and Directions
- Our approach to Telecommunications Graphs
- Calculating Communities of Interest Signatures
- Applications
  - Fraud: Guilt By Association, Repetitive Debtors
  - Marketing: Network-based Marketing, Loyalty Studies
  - Other: Proximity Models, Connection Subgraphs
- Influence in Networks
- Conclusions

Telecommunications Network Traffic

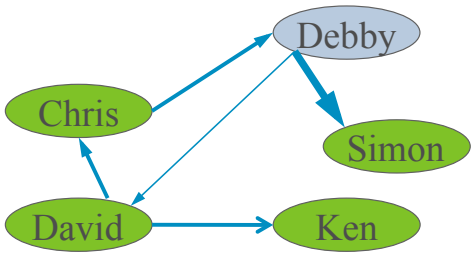
Any transactional data can be represented as a graph

Transactional Data = (originator, terminator, timestamp, duration)

Can be bipartite, multimodal, etc.

- Phone call records (call detail)
- SMS
- Weblog data
- Network Packet Traffic

Orig	Term	Time Stamp	Dur
Chris	Debby	6/21/2010 20:21	4.5
Debby	Simon	7/21/2010 10:11	134.2
Debby	David	7/23/2010 12:11	0.1
David	Chris	7/26/2010 03:30	10.5
David	Ken	7/26/2010 10:01	14.5



Network Traffic at AT&T

- Telephony data is Large, Dynamic, and Sparse
- Call Detail Records
  - 4 Billion per day
  - Incl. 2 Billion SMS
  - > 400 Million unique numbers
- Dynamic
  - Cellular numbers can recycle in two weeks
  - 0.2% numbers disappear and 0.3% appear every day!
- Heterogeneous, but typically Sparse

Our Goals

By looking at our call detail records as a large communication graph:

We get insight into product diffusion

We learn about how fraud clusters

We can see influence of social networks

Current Research in Communications Graphs

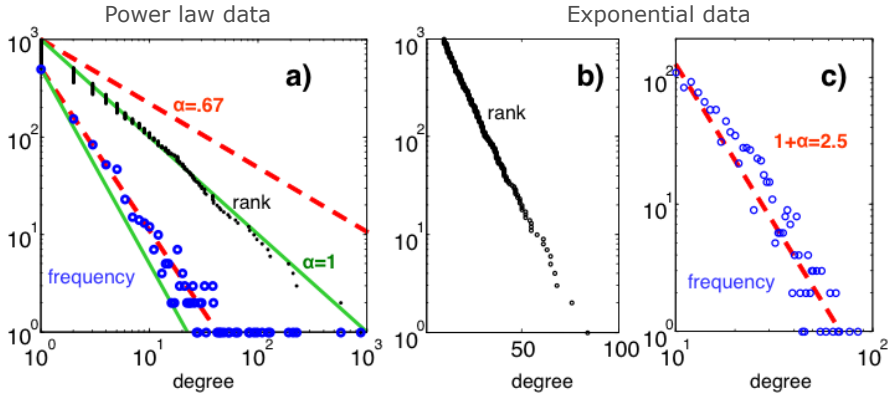
- Global graph properties
  - Clustering coefficients, diameters, power laws
  - Generative models for graphs (pref. attachment, forest fire, etc)
  - Graph properties are interesting for novel or massive graphs
    - (Leskovic and Horvitz 2008)
  - Our experience: not so useful for our questions of interest

  - Please be careful with power laws!!

Be Careful with Power Laws…

•From *Towards a Theory of Scale-Free Graphs*: Definition, Properties, and Implications. (Li, Alderson, Doyle, Willinger 2005)

> "...even when the underlying random variable X is [scale-free], size-frequency plots systematically underestimate α, and worse, have a tendency to suggest that scaling exists where it does not."



Current Research and Directions

- Community Detection
  - Large body of work on finding cliques, pseudo-cliques, dense communities
  - Our experience: similar to Leskovec, Lang, DasGupta, Mahoney (2008)
    - Communities distinct from the rest of the graph only happen in small groups
    - 'Larger size scales gradually "blend into" the expander-like core of the network and thus become less "community-like."'
- Social Effects and Influence
  - Effects of social networks on product adoption, churn
  - Identifying influential members of a network for targeting or retention or information dissemination
  - Our experience: very application dependent
- Estimation of nodes and edges
  - Classification, collective inference
  - Sampling of graphs to enable inference
  - Our experience: problems with scalability

Our Approach to Telecommunications Graphs

We address the three main characteristics of our communication graphs

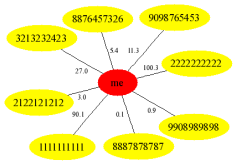
- Large Size
- Dynamic Nodes and Edges
- Sparsity of Connections

Our Approach: Size

Storing hundreds of millions of small graphs is much more *efficient* than storing one massive graph.

Define a social network signature for each node

These are our atomic units of analysis, the local behaviro of each node.



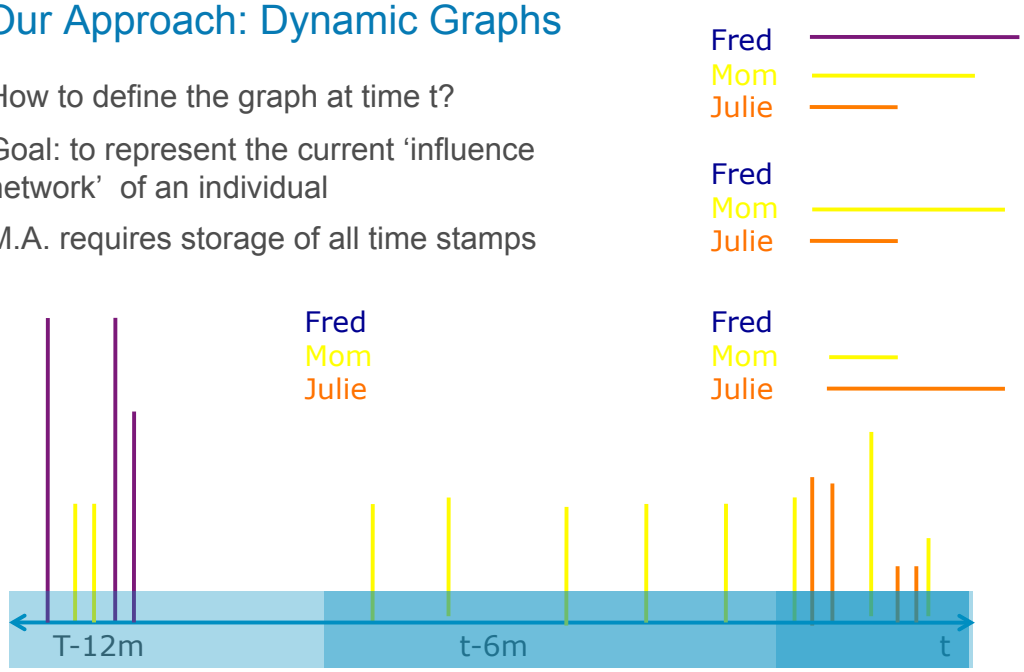
2222222222	100.3
1111111111	90.1
3213232423	27.0
9098765453	11.3
8876457322	5.4
2122121212	3.0
9908989898	0.9
8887878787	0.1

Our Approach: Dynamic Graphs

How to define the graph at time t?

Goal: to represent the current 'influence network' of an individual

M.A. requires storage of all time stamps



Our Approach: Dynamic Graphs

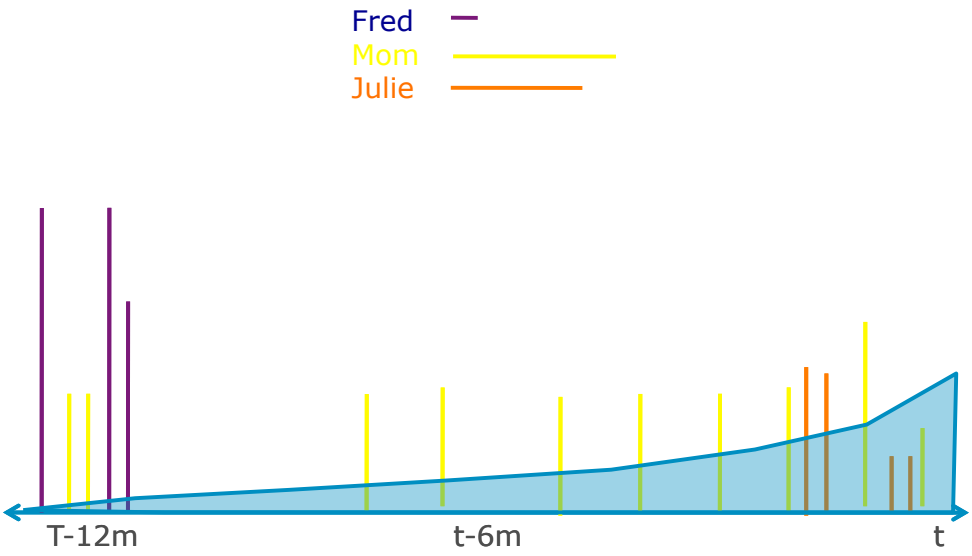
We want our graph to:

- emphasize recent behavior
- change smoothly
- be efficient to store

We adopt an Exponentially Weighted Moving Average (EWMA):

$$G_t = \theta G_{t-1} + (1-\theta) g_t$$

Where:
G_t = Today's Graph
G_{t-1} = Yesterdays's Data
g_t = Today's data
θ in (0,1) is a scalar decay parameter

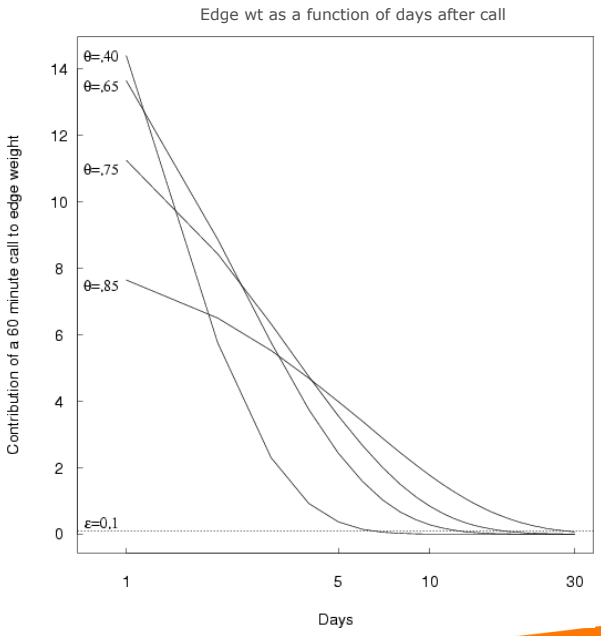


Our Approach: Defining dynamic graphs

θ closer to 1
- calls decay slower
- more historical data included
- smoother

θ closer to 0
- faster decay
- recent calls count more
- more power to detect changes
- less smooth

For a 60 minute call (0.1 thresh)
θ =0.8 => 22 days
θ =0.9 => 120 days
θ =0.99 => 180 days



Our Approach: Sparsity

Sparsity works to our advantage:

- humans' communication patterns are (roughly) Zipfian.
- 95% of consumers have 95% of their calls among 20 contacts
  - (even though median = 34, 95th percentile = 171!)

Approximations to account for sparsity:

- Global pruning of edges – overall threshold (ε) below which edges are removed from the graph
- Local pruning of edges – designate a maximal in and out degree (k) for each signature, and assign an overflow bin

Reduces effect of supernodes

Increases efficiency

Preserves total weight

Changes global properties

1111111111	92.1
2222222222	90.3
3213232423	24.3
9098765453	10.1
8876457326	4.9
2122121212	3.7
9991119999	0.5
3990898989	0.8
8887878787	0.09

=

1111111111	92.1
2222222222	90.3
3213232423	24.3
9098765453	10.1
8876457326	4.9
2122121212	3.7
Other	1.4

Our Approach: Implementation

We have now defined a representation of a dynamic graph by three parameters:

- θ - controls the decay of edges and edge weights
- ε - global pruning parameter
- k – local pruning parameter

Application-specific; parameter values are set using grid search over a training set

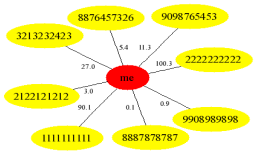
Typical settings for telephony data:

- $\theta = 0.97-0.99$
- $k = 20-50$
- $\varepsilon = 0.1$

DB ~ 30 GB

Our Approach: Communities of Interest

- Often the signature contains things we don't want:
  - Businesses, High weight nodes, Wrong numbers
- Often the signature doesn't contain things we do want:
  - Other carrier calls, other modes of communications
- Starting with the signature, we create a COI by:
  1. Recursively expanding the COI signature
  2. (maybe) adding edges (Agarwal and Pregibon 2004, Latent Space Models)
     - But probably not – too complex
  3. Pruning edges



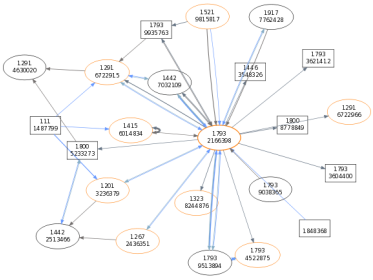
Summary of Methodology

Our 3 parameter implementation is a flexible way of accounting for scale, dynamics, and sparsity

For every node on our network these COI are updated daily

Applications:

- Fraud – Guilt By Association
- Fraud – Repetitive Debtors
- Marketing – Viral Marketing
- Marketing - Loyalty



w/ D.Pregibon, C.Cortes

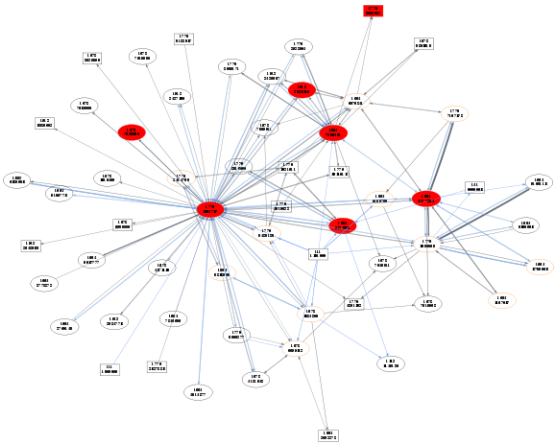
Applications: Fraud

Guilt by Association

Look at the COI for all new customers after 2 days/2 weeks.

If enough fraudulent or risky numbers in their COI, risk score is increased.

Note: even if fraudulent numbers are defunct, the nodes may still exist in the COI



Applications: Fraud – Repetitive Debtors

Repetitive Debtors:

Looking for customers who are trying to avoid payment:

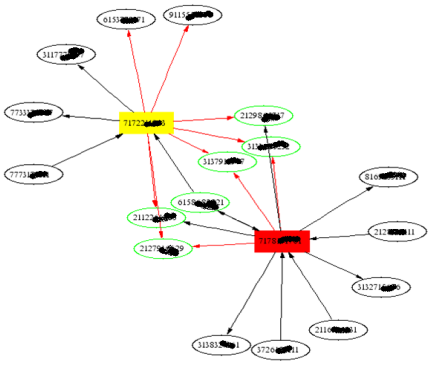
Name	Ted Hanley
Address	14 Pearl Dr St Peters, MN
Balance	$208.00
Disconnected	2/19/08 (nonpayment)

Name	Debra Handley
Address	14 Pearl Dr St Peters, MN
Balance	$142.00
Connected	2/22/08

w/ D. Pregibon, C. Cortes, R. Bell, D. Agarwal, S. Hill

Applications: Fraud – Repetitive Debtors

COI allows us to determine if two accounts have the same underlying person



Need a metric for similarity between graphs

can use Weighted Dice:

$$WD(A,B) = \frac{I_{j \in A \cap B}(w_A(j) + w_B(j))}{1 + \sum_j w_A(j)}$$

We generate cases for the fraud team to further investigate

Implemented in production fraud system:

Determined to be 95% accurate

at&t

Applications: Marketing

Define a “viral” prospect as one who has an early adopter of new product in their COI

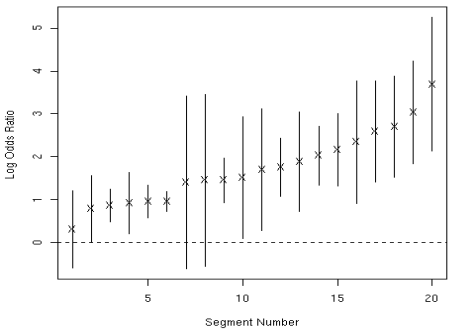
Two Setups:

Effect of COI under direct marketing event

Factor of 4.8 more sales for virals

‘Organic’ effects of COI

Factor of 3.5 more sales for virals



Other Findings:

- Virals less likely to call customer care
- More likely to order on the web
- Higher revenue
- More tech-y products have more viral behavior

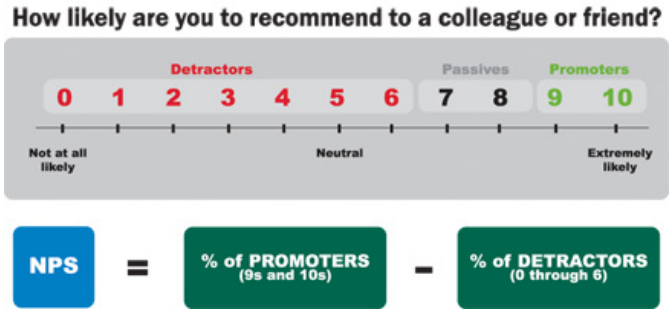
w/ S.Hill, F. Provost, D. Paul

at&t

Applications: Marketing - Loyalty

w/ D. Paul, T. Keiningham, L. Aksoy, B. Cooil

- Loyalty commonly measured by Net Promoter Score
  - 2003 Harvard Business Review Article by Fred Reichheld
  - "The One Number you Need to Grow"
    - Claimed that is the BEST predictor of firm growth
    - Is the ONLY metric needed to predict growth



Applications: Marketing - Loyalty

Do we see evidence of recommendations from those who say they "Definitely would recommend"?

Variable	Model 0	Model 1	Model 2	Model 3	Model 4
Variable					
Intercept	**-4.601 ***	**-4.571 ***	**-4.578 ***	**-4.545 ***	**-4.551 ***
Survey Household					
Call Variable 1	**3.1E-5 ***	**3.1E-5 ***	**3.1E-5 ***	**3.2E-5 ***	**3.2E-5 ***
Definitely Recommend		-0.063	-0.096	-0.118	-0.1
Tenure24			0.01		0.008
DefinitelyRec X Tenure24			0.045		-0.033
Link Level					
Call Variable 2	**0.002 ***	**0.002 ***	**0.002 ***	**0.002 ***	**0.002 ***
DSL Presence	**1.731 ***	**1.724 ***	**1.725 ***	**1.722 ***	**1.725 ***
LinkTenure	**-0.002 ***	**-0.002 ***	**-0.002 ***	**-0.002 ***	**-0.002 ***
StrongLink				-0.285	-0.314
Cross-Level Interactions					
DefinitelyRec X StrongLink				0.606	
DefinitelyRec X Tenure24 X StrongLink					**0.832 ***
					p < .05

- No 'overall' impact
- Strong ties to tenured links showed impact
- No evidence that negative response inhibited sales in COI
- No evidence of "special connectors" or influencers

Influence in Networks

Some thoughts on influence:

Lots of evidence in literature on influence of social networks on behavior (adoption, churn, etc)

Strength of ties matter, size of network matters

consistent with constant influence per minute

Finding the ‘influencers’ is the holy grail!

- that’s what the marketers (at least at AT&T) want!
- very little evidence of Ferris
- Is ‘influence’ just a side effect of ‘connectivity’?
- Do second order effects matter?
  - Collective inference models useful here

at&t

Thank you!

at&t