Generating Synthetic Event Data

Chaitan Baru, SDSC
Chen Xie, Yilin Jin, CSE Dept

University of California San Diego
Acknowledgement

- Work performed as part of CSE Master’s projects
  - Chen Xie (now at Zynga)
  - Yi Lin (now at Google)
Background

- Deep Analytics Pipeline is based on event data
  - Clickstreams at a social site are a good example of such data
  - The application being modeled: Building predictive models to detect events of interest
    - Event of interest: An ad click
    - Application might look for click “prefix” sequences and other conditions that predict click on the ad (details TBD)
- The project: Generate synthetic clickstream data
Data Model

• User home pages
  ▫ Links to other Users (the social network)
  ▫ Comments / posts
  ▫ Links to Photos
  ▫ Links to Ads

• Photos
  ▫ Comments
Data Schema

- **Users**
  - User Profile
  - Comments
  - Friends
  - Friend Requests
  - Photos
- **Photos**
  - Photo tags (comment)
- **Ads**
- **Links among entities**
The Event Model

• Users click on web pages, generating the clickstream
  ▫ User traverses multiple links within a session
  ▫ Captured in logs
• Some of the clicks are on ads
  ▫ Want to model these click sequences
Data Generation

• Generate data for entities in the social network
  ▫ E.g. Users, Photos, Comments, Ads
• Generate a graph linking entities (Users)
  ▫ Modeling the social network graph
• Generate the Clickstream dataset
  ▫ A set of (sub)paths in the social network graph, some of which result in clicks on ads
Data Generation

• Populate tables
  ▫ Use config files to specify constraints on columns values, domain values, etc.
  ▫ Similar to TPC DataGen

• Create links between User and Photo, Ad tables
• Create User to User links, i.e. the social network graph
Graph Generation

• Construct a synthetic social network graph that can model real-world social networks
• Solution: *Multiplicative Attribute Graph (MAG) Model*, Kim and Leskovic

• Key Components:
  ▫ Node-attribute Vector
  ▫ Attribute Link-affinity Matrices
• Generating large graphs
  ▫ Modify their data generator to create a parallel data generator for large graph sizes
Multiplicative Attribute Graph Model

- Each node in graph associated with a Node-Attribute Vector
  - Yes/No answers to series of questions, e.g. Are you older than 45? Are you Female? ...
  - Binary Attributes

\[ \begin{array}{c|c|c|c|c} 
\text{Age (0: <45/ 1: >45)} & \text{Gender (0: Male / 1: Female)} & \text{Daily Web Clicks:} \\
1 & 0 & 1 & … & 1 & 0 & 1 & 1 \\
\end{array} \]

\(k\) binary attributes
“Attribute link affinity”

- If two nodes in the graph have the same values for the $i^{th}$ attribute, what is the probability of a link between the two
  - Model homophily as well as heterophily
- Attribute link-affinity matrix $\Theta_i$
  - Affinity of the $i^{th}$ attribute to form a link between a pair of nodes given the value of attribute $i$ for both nodes
Attribute Link-affinity Matrices

- For $k = 4$ attributes: $\Theta_0, \Theta_1, \Theta_2, \Theta_3$

- Binary Attribute vectors for nodes $u$ and $v$

$$ \begin{array}{ccc}
  a_0(u) = 0 & a_0(v) = 0 \\
  a_0(u) = 1 & a_0(v) = 1 \\
\end{array} $$

<table>
<thead>
<tr>
<th></th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u$</td>
<td>0.9</td>
<td>0.2</td>
<td>0.5</td>
</tr>
<tr>
<td>$v$</td>
<td>0.1</td>
<td>0.9</td>
<td>0.5</td>
</tr>
</tbody>
</table>

$\theta_0$ $\theta_1$ $\theta_2$ $\theta_3$

<table>
<thead>
<tr>
<th>0</th>
<th>0</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Probability of an Edge

- Computing the probability of an edge between $u$ and $v$

$$a(u) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad a(v) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad P(u,v) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Theta_i = \begin{array}{cccc}
\alpha_1 & \beta_1 \\
\beta_1 & \gamma_1 \\
\alpha_2 & \beta_2 \\
\beta_2 & \gamma_2 \\
\alpha_3 & \beta_3 \\
\beta_3 & \gamma_3 \\
\alpha_4 & \beta_4 \\
\beta_4 & \gamma_4 \\
\end{array}$$

- Randomly generate a number, $p$, between (0, 1)
  - If $p < P(u,v)$, form the edge; else not.
Multiplicative Attribute Graph Model

- Different matrix values can capture different structure patterns

(a) Homophily

(b) Heterophily

(c) Core-Periphery

(d) Random

- Value of Vector and Matrices can be trained from real-world graph data.
Simplified MAG

- $M(n, k, \mu, \theta)$
  - $n$ – number of nodes
  - $k$ – number of attributes of each node
  - $\mu$ – probability that an attribute takes the value zero
  - $\theta$ – link affinity matrix $[ \alpha, \beta; \beta \gamma ]$
    - Models a core-periphery graph
    - Most large real-world networks have a common “onion”-like core–periphery structure

Single-Machine Generator: MAGGEN

- Developed by SNAP Group at Stanford
- In C++

```
O(n)

O(n^2)
```

```
Generate nodes

Generate Edges

Return Results

Finish
```
Design of Distributed Synthetic Network Generator (DSNGEN)

- **Compute Server**
  - Generate Nodes
  - Generate Edges
- **Barrier Server**
  - Workflow Control
  - Collect Results
Experiments

- Correctness (Yahoo! Flickr Data)
  - 10k Simple Network
    - Same values for $\vec{\mu}$ and $\vec{\theta}$ vector for all nodes
      \[ n = 10240, \: k = 8, \: \mu = 0.45, \: \theta = [0.85, 0.30; 0.30, 0.25] \]
  - 10k Bernoulli Network
    \[ n = 10240, \: k = 9 \]

<table>
<thead>
<tr>
<th>Attribute (i)</th>
<th>$\mu_i$</th>
<th>$\theta_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.60</td>
<td>[0.9999 0.0432; 0.0505 0.9999]</td>
</tr>
<tr>
<td>1</td>
<td>0.04</td>
<td>[0.9999 0.9999; 0.9999 0.1506]</td>
</tr>
<tr>
<td>2</td>
<td>0.24</td>
<td>[0.9999 0.9999; 0.9999 0.2803]</td>
</tr>
<tr>
<td>3</td>
<td>0.17</td>
<td>[0.9999 0.9999; 0.9999 0.2833]</td>
</tr>
<tr>
<td>4</td>
<td>0.62</td>
<td>[0.9999 0.0476; 0.0563 0.9999]</td>
</tr>
<tr>
<td>5</td>
<td>0.08</td>
<td>[0.9999 0.9999; 0.9999 0.1319]</td>
</tr>
<tr>
<td>6</td>
<td>0.57</td>
<td>[0.9999 0.1246; 0.1402 0.9999]</td>
</tr>
<tr>
<td>7</td>
<td>0.57</td>
<td>[0.9999 0.1186; 0.1364 0.9999]</td>
</tr>
<tr>
<td>8</td>
<td>0.40</td>
<td>[0.9999 0.1757; 0.1535 0.9999]</td>
</tr>
</tbody>
</table>
Current Status

• Early version of code has been developed by students

• Awaiting new students to take over this work

• Open Questions:
  ▫ How about perturbing the data, to create anamolies?
    • Creating noisy, dirty data as part of data gen
  ▫ ...