Scaling choice models of relational social data

Jan Overgoor · Stanford University
SIAM-NS
July 09, 2020

Slides: bit.ly/c2g-venmo
Events on networks

Rosie Cima paid Jan Overgoor
December 16, 2019, 10:13 PM

Be the first to like this.

Dave Holtz @daveholtz · Jul 29, 2014
Is there a German word for the fear that you may tweet all of your life’s best tweets while you still don't have many followers?
Observed data
"Choosing to Grow a Graph"  [Overgoor, Benson & Ugander, WWW’19]

• Model edges as choices

• **Conditional** on $i$ initiating an edge, which $j$ to pick from choice set $C$?

• Conditional Logit model:  
  $$P_i(j, C) = \frac{\exp \theta^T x_j}{\sum_{\ell \in C} \exp \theta^T x_\ell}$$
Conditional Logit choice process

$t = 0$
"Choosing to Grow a Graph"  [Overgoor, Benson & Ugander, WWW’19]

- Generalizes multiple known formation models and dynamics
  preferential attachment, local search, fitness, homophily, ...

- Efficient maximum likelihood estimation of model parameters,
  existing tools

<table>
<thead>
<tr>
<th>Process</th>
<th>$u_{i,j}$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform attachment [10]</td>
<td>1</td>
<td>$V$</td>
</tr>
<tr>
<td>Preferential attachment [2, 32]</td>
<td>$\alpha \log d_j$</td>
<td>$V$</td>
</tr>
<tr>
<td>Non-parametric PA [50, 54, 58]</td>
<td>$\theta_{d_j}$</td>
<td>$V$</td>
</tr>
<tr>
<td>Triadic closure [57]</td>
<td>1</td>
<td>${j : FoF_{i,j}}$</td>
</tr>
<tr>
<td>FoF attachment [28, 61, 73]</td>
<td>$\alpha \log \eta_{i,j}$</td>
<td>$V$</td>
</tr>
<tr>
<td>PA, FoFs only</td>
<td>$\alpha \log d_j$</td>
<td>${j : FoF_{i,j}}$</td>
</tr>
<tr>
<td>Individual node fitness [9]</td>
<td>$\theta_j$</td>
<td>$V$</td>
</tr>
<tr>
<td>Latent space [20, 38, 51]</td>
<td>$\beta \cdot d(i,j)$</td>
<td>$V$</td>
</tr>
<tr>
<td>Stochastic block model [30]</td>
<td>$\omega_{g_i, g_j}$</td>
<td>$V$</td>
</tr>
<tr>
<td>Homophily [45]</td>
<td>$h \cdot 1{g_i = g_j}$</td>
<td>$V$</td>
</tr>
</tbody>
</table>
"Choosing to Grow a Graph" [Overgoor, Benson & Ugander, WWW’19]

- Generalizes multiple known formation models and dynamics
  preferential attachment, local search, fitness, homophily, ...

- Efficient maximum likelihood estimation of model parameters, existing tools

- Straightforward extension to events
Two problems at scale

1. Estimation on large networks infeasible as $n$ options for all $m$ choices - features change at each event
Two problems at scale

1. Estimation on large networks infeasible as $n$ options for all $m$ choices
2. Conditional logit model class less realistic
   - availability assumption of complete information
Solution to Problem #1 – Negative sampling

• Sample non-chosen alternatives and do estimation on the reduced choice set $\tilde{C} \subset C$, $|\tilde{C}| = s$

  also called case-control sampling (see Vu 2015, Lerner 2019)

• Update likelihood with sampling probabilities $q_j$ of data points:

$$P_i(j, \tilde{C}) = \frac{\exp \left( \theta^T x_j - \log q_j \right)}{\sum_{\ell \in \tilde{C}} \exp \left( \theta^T x_\ell - \log q_\ell \right)}$$

• Estimates on data with reduced choice sets generated with importance sampling are consistent for the estimates using complete choice sets. [McFadden 1977]
Negative sampling strategies

**Uniform** sampling

+ no adjustment necessary, weights cancel out
- inefficient for rare (but important) features
Negative sampling strategies

**Uniform** sampling
- no adjustment necessary, weights cancel out
- inefficient for rare (but important) features

**Stratified** sampling
sample according to strata, adjust with \( q_x = \frac{s}{n_G} \)
Negative sampling strategies

**Uniform** sampling
- no adjustment necessary, weights cancel out
- inefficient for rare (but important) features

**Stratified** sampling
- sample according to strata, adjust with \( q_x = \frac{s}{n_G} \)

**Importance** sampling
- sample according to likelihood of being chosen
- optimal weights are what we’re trying to estimate
Sampling with synthetic data

- Simulate 160k events with 5k nodes
- Utility function with popularity, repetition, reciprocity, and FoFs
- Estimate known parameter values
- **Samples** \( n \) constant at 10k, vary \( s \)
- Stratification requires factors less negative samples for comparable MSE

![Graph showing MSE vs. number of samples (s)]
Run time is linear in $n$ and $s$
Sampling with synthetic data

- Simulate 160k events with 5k nodes
- Utility function with popularity, repetition, reciprocity, and FoFs
- Estimate known parameter values

- Value of $n$ and $s$ at constant $n*s$ budget
- More choice samples ($n$) is better, but diminishing returns below $s = 24$
Back to problem #2

2. Conditional logit model class less realistic
Mixed Logit

• Combines multiple \textbf{latent} logits

• Each "mode" has its own utility function and choice set
  \textit{for example: social neighborhood}

\[
P_i(j, C) = \sum_{m=1}^{M} \pi_m \frac{\exp \theta_m^T x_j}{\sum_{\ell \in C_m} \exp \theta_m^T x_\ell} 1[j \in C_m]
\]

Problems:
• Log-likelihood not convex in general, need much slower EM
• No sampling guarantees
Solution to Problem #2 – De-mixed logit

- Simplify: assume that each mode has a disjoint choice set
- Reduces to $m$ individual conditional logits, simple to estimate
- The chosen item indicates the mode
De-mixed logit choice process

t = 0

chooser
neighborhood
De-mixing with synthetic data

- Simulate 80k events with 5k nodes
- "local" and "rest" mode with different utility functions $\pi_{\text{local}} = 0.75$
De-mixing with synthetic data

- Simulate 80k events with 5k nodes
- "local" and "rest" mode with different utility functions $\pi_{\text{local}} = 0.75$

- **Conditional logit**
  - Estimates in between the two modes (true values are 0.5 and 1.0)
  - Importance sampling doesn’t help accuracy
De-mixing with synthetic data

- Simulate 80k events with 5k nodes
- "local" and "rest" mode with different utility functions $\pi_{\text{local}} = 0.75$

- **Conditional logit**
- Estimates not stable for different values of $s$ outside the model class

![Graph](image)
De-mixing with synthetic data

• Simulate 80k events with 5k nodes
• "local" and "rest" mode with different utility functions $\pi_{\text{local}} = 0.75$

• De-mixed logit
• Estimates accurate and stable
Venmo Data

- Scraped public transactions
- 25M users and 501M transactions
- 80% transactions are “local”
- Analyze stratified CL and de-mixed CL
Venmo Non-parametric estimates

- Easy to test hypotheses over different modes.
- Degree is number of incoming transactions.
- Degree is less important within social neighborhood, super-linear outside.

![Graph showing relative probability versus in-degree for local and non-local transactions.](image-url)
Discussion

- Leverage existing results from sampling and econometrics literatures
- Make feasible to estimate complex models on very large graphs
- Think carefully about limitations of model class

Future work

- Theory on “to sample or to negatively sample?”
- Sampling guarantees for mixed logit
- Empirical comparison with similar modeling frameworks (SAOM, REM)
- More applications

THANKS!

🔗 bit.ly/c2g-code
✉️ overgoor@stanford.edu