A Bayesian Framework for Estimating Properties of Network Diffusions

Varun Embar\textsuperscript{1} \quad Rama Kumar Pasumarthi\textsuperscript{2} \\
Indrajit Bhattacharya\textsuperscript{1}

\textsuperscript{1}IBM Research India \\
\textsuperscript{2}American Express (work done when at IBM)

ACM SIGKDD 2014
Network Diffusion

Example 1: Spread of Ebola in West Africa
Example 2: Spread of hashtags among Twitter users

Entities of interest

- **Network**: Captures connections between people / Twitter users
- **Diffusion Process**: Stochastic mechanism of ‘infection’ spread
- **Diffusion Cascades**: Time-stamped infection paths over network

Long studied in epidemiology, sociology, econometrics, marketing
Recent interest in computer science
Problems in Network Diffusion Analysis

Evaluating Properties of Observed Network and Cascades

- Centrality and reach of individual nodes
- Viral marketing seeds (Kempe KDD03, Goyal CIKM08, PVLDB11)
- Community structures (Mehmood ECML13, Barbieri ICDM13)
- Likelier diffusion mechanism (Milling SIGMetrix12)

Inferring Network from Partially-observed Cascades

- Estimate network connections and strengths given cascades
- Primarily assuming independent cascade diffusion model
- Maximum likelihood estimation
- Saito (AML09), Gomez-Rodriguez (KDD10, ICML11,13, WSDM13), Du (NIPS12), Netrapalli (SIGMetrix12), Wang (ECML12), Kutzkov (KDD13), Daneshmand (ICML14)
Example Property: Leaders of tribes (LoT)

**Leader of Tribes** *(Goyal CIKM08)*

- Network property: High-weight paths to large ‘tribe’ of nodes
- Cascade property: Frequent transmissions over these paths

Not tractable even given complete observations

**Weak LoT**

- Network property: High-weight edges to tribe nodes
- Cascade property: Frequent transmissions over these edges

Easy to compute given complete observations
Still interesting for marketing, epidemiology
Evaluating Joint Properties: Challenges

Network and Diffusion Cascade

- $\alpha_{uv} \in \mathbb{R}_+$: connection strength between nodes $u, v$
- Cascade of infections: infected node $u_i$, parent node $z_i$, time $t_i$

Diffusion Process: Independent Cascade Model (ICM)

- Infected node proposes infection time for uninfected neighbors
- Uninfected node catches infection with earliest proposed time
- Multiple infections of same node (Splitting model) (Wang ECML12)

Hidden variables

- Network: Connection strengths and sometimes edges unobserved
- Cascades: Cascades do not record infection sources $z_i$
Expectations of Network Diffusion Properties

Network Diffusion Property

Function \( f(\alpha, z) \) defined on network \( \alpha \) and cascade \( z \)

Frequentist Plug-in

Evaluate property using point estimates (MLE) of \( \alpha \) and \( z \)

- MLE overfits for infrequent edges
- Most likely property may not correspond to mostly likely \( \alpha \) and \( z \)

Expectation of Network Diffusion Properties

\[
\bar{f}(z, \alpha) = E_{p(z,\alpha|\{c^o\})}[f(z, \alpha)]
\]
**Conjugate prior for Network Strengths**

- IID Gamma prior over network strengths $\alpha$
  
  $$p(\alpha) = \prod_{uv} p(\alpha_{uv}) ; \quad p(\alpha_{uv}) = Gamma(\alpha_{uv}; a, b)$$

- Gamma conjugate for Rayleigh and Exponential delays in ICM

- Posterior also Gamma distributed
  
  $$p(\alpha|\{c^0\}, z) = \prod_{uv} Gamma(a + \rho_{uv}(z), b + \Delta_{uv})$$

  $\rho_{uv}(z)$: #u-v infections; $\Delta_{uv}$: Cumulative u-v infection delay

Parameter settings for sparse network or non-informative prior
Expectation Computation

- Integration over network $\alpha$; $p(\alpha | \{c^o\}, z)$ decomposes over $\alpha_{uv}$
- Summation over parents $z$; $p(z | \{c^o\}, \alpha)$ decomposes over $z_i$

Characterization of properties (for Independent Cascade Model)

- **Network-nice**: When property decomposes over $\alpha_{uv}$ and $z$, and $\alpha_{uv}$ terms are Gamma integrable, network integration is $O(|E|)$
- **Cascade-nice**: When the property decomposes over $z_i$, cascade path summation is $O(|C|)$
- **Totally-nice**: When the property is such that $\alpha_{uv}$ and $z_i$ terms are decoupled in the expectation, and $\alpha_{uv}$ terms are Gamma integrable, joint marginalization is $O(|E| + |C|)$
Interesting Network Diffusion Properties

Network-centric Properties: Scores for nodes, edges, etc
- Nodes with large path-reach (approx. LoT): not nice at all
- Nodes with large edge-reach (weak LoT): network-nice only
- Strong frequent edges: network-nice only
- Network Inference: network-nice, cascade-nice, not totally nice
- Edges that are strong or frequent but not both: totally nice

Cascade-centric Properties: Scores for individual infections, infection paths, etc
- Infections by strongest neighbor: cascade-nice only
- Infection parent inference: cascade-nice only
- Complete likelihood, Likelihood: cascade-nice only
Approximate Evaluation of Not-nice Properties

**MCMC Marginalization**
- Only network-nice: Efficiently marginalize $\alpha$, MCMC over $z$
- Only cascade-nice: Efficiently marginalize $z$, MCMC over $\alpha$
- Not nice: MCMC over both $\alpha$ and $z$

**Gibbs Sampling for Network Diffusion**

Full / Uncollapsed Gibbs Sampling

\[
p(z_i = j|z_{-i}, \alpha, \{c^o\}) \propto \alpha_{ji}
\]
\[
p(\alpha_{uv}|z, \alpha_{-uv}, \{c^o\}) \sim Gamma(\rho_{uv} + a, \Delta_{uv} + b)
\]

Collapsed Gibbs sampling for network-nice properties

\[
p(z_i = j|z_{-i}, \{c^o\}) \propto \frac{(\rho_{uv}^i(z)+a)}{\Delta_{uv}+b}
\]
**Experiments: Algorithms**

**Bayesian Expectation (BE)**

Gamma prior parameters: $a = 0.00001$, $b = 0.1$

**Frequentist Plug-in (FP)**

- Take point estimate $\hat{\alpha}$ of network
- Most likely infection parents given $\hat{\alpha}$: $\hat{z} = \arg \max_z p(z|\hat{\alpha}, \{c^0\})$
- Evaluate $f(\hat{\alpha}, \hat{z})$
- For $\hat{\alpha}$, use MONET

**Exponential distribution for all experiments**
## Synthetic Data Experiments

### Data Generation
- Forest Fire, Random, Hierarchical, Core-Periphery
- 1000 nodes, $\sim$ 2000 edges
- $\alpha_{uv} \sim U(0.01, 10)$
- 20 splitting cascades with 2 random seeds, $\sim$ 50,000 infections

### Evaluation
- Parent inference: accuracy against true parent $z^*$
- Network inference: Error wrt best possible given true parents ($\alpha^*$)
- Property: Error wrt $f(\alpha^*, z^*)$
Bayesian approach recovers the signature shape of the distribution, by considering less likely networks.
Synthetic Data Experiments: One-to-one Properties

Loglikelihood

<table>
<thead>
<tr>
<th></th>
<th>CorePeriphery</th>
<th>Hierarchical</th>
<th>Random</th>
<th>ForestFire</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BE, FP</td>
<td>BE, FP</td>
<td>BE, FP</td>
<td>BE, FP</td>
</tr>
<tr>
<td>Test</td>
<td>1.0e4, 0.6e4</td>
<td>6.5e3, 2.4e3</td>
<td>1.1e4, -1.5e4</td>
<td>1.2e4, 926</td>
</tr>
<tr>
<td>Train</td>
<td>2.8e4, 3.6e4</td>
<td>2.0e4, 2.2e4</td>
<td>2.3e4, 2.9e4</td>
<td>2.8e4, 3.3e4</td>
</tr>
</tbody>
</table>

Network and Parent Inference

<table>
<thead>
<tr>
<th></th>
<th>CorePeriphery</th>
<th>Hierarchical</th>
<th>Random</th>
<th>ForestFire</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BE</td>
<td>BE</td>
<td>BE</td>
<td>BE</td>
</tr>
<tr>
<td>Network Inf</td>
<td>0.116</td>
<td>0.884</td>
<td>0.147</td>
<td>0.329</td>
</tr>
<tr>
<td>Parent Inf</td>
<td>0.533</td>
<td>0.861</td>
<td>0.757</td>
<td>0.770</td>
</tr>
</tbody>
</table>

Bayesian approach avoids overfitting for one-to-one properties
Real Data Experiments

Meme-Tracker

- Meme diffusion between 5000 blogs, news sites (Mar 11 - Feb 12)
- 5 topics: Basketball, Alcohol, Technology, NBA, Occupy
- Long cascades: length > 30
- 80-20 train-test split; infections of new users pruned in test

Test Loglikelihood

<table>
<thead>
<tr>
<th></th>
<th>Bball</th>
<th>Alcohol</th>
<th>Tech</th>
<th>NBA</th>
<th>Occupy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BE</td>
<td>-1.5e6</td>
<td>-5.8e5</td>
<td>-6.6e5</td>
<td>-8.9e5</td>
<td>-5.1e5</td>
</tr>
<tr>
<td>FP</td>
<td>-3.5e6</td>
<td>-8.9e5</td>
<td>-2.6e6</td>
<td>-1.1e7</td>
<td>-1.2e6</td>
</tr>
</tbody>
</table>

Bayesian approach generalizes much better
Scaling Experiments

- Map reduce implementation; 12 core server
- Randomly sampled cascades from Meme-Tracker

Scaling (roughly) linear in no. of cores
Closing Thoughts

Summary of this paper

- Evaluating network diffusion properties from partial observations
- Bayesian framework: Expected property under posterior distribution of hidden network and infection parents
- Characterization of properties in terms of computation cost
- Significantly better reconstruction than frequentist baseline

Future directions

- Optimize for specific properties
- Consider more sophisticated random graph priors
- Analyze alternative diffusion models, e.g. linear threshold