Mixed-Membership Stochastic Block-Models for Transactional Networks

BIHAN ZHUANG

LIN ZUO

Transactional Data

- A list of one-to-many communications (e.g. email) among nodes in a social network
- The assumptions that relations are binary-valued and occur between pairs of nodes no longer holds
- Depending on the type of transactional data, additional information on each transaction include: timestamps, message content, recipient classes(To/Cc/Bcc)

Transactional Data

- Structure of the network data we seek to model
  - M nodes (people)
  - N transactions, each of which involves 1 to (M-1) recipients and one sender
  - Additional transactional information will not be used

- Assumptions:
  - Each node can play different roles while interacting with different nodes
  - Likelihood of interaction between two nodes depend on the roles they play at the time of their interaction (e.g. phd/RA/TA)

(a) Transactions

Sender	A	B	C	D
A	.	1	0	0
A	.	1	0	1
C	0	1	.	0
B	1	.	0	1
C	0	1	.	0

(b) Transaction counts

	(recipients)			
(sender)	A	B	C	D
A	.	2	0	1
B	1	.	0	1
C	0	2	.	0
D	0	0	0	.

(c) Binary relations

	(recipients)			
(sender)	A	B	C	D
A	.	1	0	1
B	1	.	0	1
C	0	1	.	0
D	0	0	0	.

Related Research

Problems with previous work about transactional data

- Lost information about co-recipient of the same message
- Lost information about frequency of interactions between nodes
  - Counts thresholded in socio-matrix

Work that inspired this paper to build network model for transactional data

- Mixed membership stochastic block-model

  -- E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. Journal of Machine Learning Research, 9:1981–2014, 2008.
- Network transactional feature in the network is important for predicting links

  -- I. Kahanda and J. Neville. Using transactional information to predict link strength in online social networks. In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media
- Frequency of interactions could improve the accuracy of modeling

  -- K. Kurihara, Y. Kameya, and T. Sato. A frequency-based stochastic blockmodel. In Workshop on Information Based Induction Sciences, 2006.

Transactional Mixed Membership Stochastic Block-Model Set-up

- **N** messages are sent within a network of **M** nodes
- Each message **n** has a sender **Sn** and **Sn** itself can't be a recipient
- Each message **n** has a recipient list represented by **M** binary variables $\mathbf{Y_{n1}, \ldots, Y_{nM}}$
  - $\mathbf{Y_{nm}}$ = 1 when node **m** received message **n** from **Sn**
  - $\mathbf{Y_{nm}}$ = 0 when node **m** didn't receive message **n** from **Sn**
- **K** groups in the network
- Each node **i** has a K-dimensional membership probability $\boldsymbol{\pi_i}$, with $\sum_{k=1}^{K} \pi_{ik} = 1$
- Each element $\mathbf{B_{kl}}$ in the interaction matrix **B** represents the probability of a node **i** in group **k** sending a message to a node **j** in group **l**.

TMMSB Generating Process

1. For each node i, draw mixed-membership vector $\pi_i \sim \text{Dirichlet}(\alpha)$
2. For each node i, draw its friendship value $\lambda_i \sim N(\mu, \delta)$
3. Choose $N \sim Poisson(\varepsilon)$: number of emails
4. For each email n
   (a) For each node i, draw $z_{ni} \sim \text{Multinomial}(\pi_i)$
   (b) Pick node u as sender (i.e., $S_n = u$) among all the nodes with probability $\frac{exp(\lambda_u)}{\sum_j exp(\lambda_j)}$
   (c) For each node $j \neq u$, draw $Y_{n,j} \sim \text{Bernoulli}(z_{nu} B z_{nj}^T)$

Inference

- $\{\pi_{M*K}, Z_{N*M*K}\} \equiv \theta$ as random latent variables
- $\{\alpha, B\} \equiv \beta$ as fixed parameters that we need to estimate
- Estimate the posterior distribution

$$\mathrm{p}(\theta \mid Y, \beta) = \frac{p(Y|\,\theta,\beta)p(\theta|\beta)}{p(Y|\beta)}$$

  by using Mean-field Variational Bayesian approximation
- Variational distribution

$$q(\pi_{1:M}, Z_{1:N,1:M}) = \prod_{m=1}^{M} q_1(\pi_m|\gamma_m) \prod_{n=1}^{N} \prod_{m=1}^{M} q_2(z_{n,m}|\phi_{n,m})$$

  where q_1 is a Dirichlet and q_2 is a Multinomial, approximates the posterior distribution in terms of Kullback-Leibler divergence

Inference

Brief VB Algorithm

1. Initialize $B^{(0)}, \alpha^{(0)}, \gamma_{1:M}^{(0)}, \phi_{1:N,1:M}^{(0)}$
2. E-step:
   i. Update $\gamma_i^{(j)}$ for $i = 1, \ldots, N$
   ii. Update $\phi_{n,m}^{(j)}$ for all n, m
   iii. Until convergence
3. M-step: Update $B^{(j)}$
4. Until convergence

$$\phi_{nm,k} \propto \mathbb{E}_q(\log(\pi_{m,k}))\times$$
$$\mathbb{1}_{[m \neq S_n]} \cdot \prod_{l=1}^{K} \left(B_{lk}^{Y_{nm}} . (1 - B_{lk})^{1-Y_{nm}}\right)^{\phi_{nS_n,l}} \times$$
$$\mathbb{1}_{[m = S_n]} \cdot \prod_{m' \neq m} \prod_{l=1}^{K} \left(B_{kl}^{Y_{nm'}} . (1 - B_{kl})^{1-Y_{nm'}}\right)^{\phi_{nm',l}}$$

$$\gamma_{m,k} = \alpha_k + \sum_{n=1}^{N} \phi_{nm,k}$$

$$B_{k,l} = \frac{\sum_{n=1}^{N} \sum_{m=1, m \neq S_n}^{M} \phi_{nS_n,k} \phi_{nm,l} Y_{nm}}{\sum_{n=1}^{N} \sum_{m=1, m \neq S_n}^{M} \phi_{nS_n,k} \phi_{nm,l}}$$

Model Choice

- A BIC criterion was developed in order to choose the number of clusters

$$BIC = 2.\log \mathcal{L} - (K^2 + K).\log(|Y|)$$

where $\mathcal{L} = \prod_{n=1}^{N} \prod_{j \in 1..M, j \neq S_n} p_{ij}^{y_{nj}} (1 - p_{ij})^{1-y_{nj}}$ and p_{ij} = Pr(j receives | i sends) = $\pi_i B \pi_j^T$

Simulation Results from paper

0.01	**0.2**	0.01	0.01
0.01	**0.3**	**0.2**	**0.1**
0.1	0.01	0.01	**0.3**
0.1	0.01	0.01	**0.3**

True B matrix

0.0127	**0.2012**	0.0149	0.0115
0.0064	**0.3055**	**0.2064**	**0.0802**
0.0964	0.0207	0.0146	**0.2959**
0.0979	0.0243	0.0164	**0.2733**

Estimated B matrix

Simulation Input

M <- 65

N <- 650

α <- 0.25

K <- 4

True Adjacency Matrix

Recovered Adjacency Matrix

K^*	2	3	4	5	6	7
BIC ($\times 10^4$)	-3.0357	-2.9660	-2.9229	-2.9229	-2.9339	-2.9501

Reproduce results

Simulation Input
M <- 65
N <- 650
α <- 0.25
K <- 4

0.01	**0.2**	0.01	0.01
0.01	**0.3**	**0.2**	**0.1**
0.1	0.01	0.01	**0.3**
0.1	0.01	0.01	**0.3**

True B matrix

0.0127	**0.2012**	0.0149	0.0115
0.0064	**0.3055**	**0.2064**	**0.0802**
0.0964	0.0207	0.0146	**0.2959**
0.0979	0.0243	0.0164	**0.2733**

Estimated B matrix TMMSB from paper

0.1332	**0.1345**	0.0619	**0.0972**
0.0696	0.0703	0.0311	0.0498
0.1779	**0.1796**	**0.0849**	**0.1315**
0.1104	**0.1115**	0.0506	0.0799

Estimated B matrix TMMSB (randomly initialize B)

0.0469	0.0276	0.0575	**0.0866**
0.0781	0.0465	**0.0951**	**0.1403**
0.1251	0.0758	**0.1506**	**0.2162**
0.0937	0.0562	**0.1137**	**0.1661**

Estimated B matrix TMMSB (randomly initialize B, phi)

0.7	**0.3175**	**0.1778**	**0.2837**
0.9925	**0.9809**	**0.8319**	**0.9386**
0.6120	**0.5932**	0.0274	**0.4708**
0.0206	0.009	0	0.0029

Estimated B matrix MMSB with 100,000 iterations