Simultaneous text & Network modeling in MMSB for Learning Structured Interactions on Online Forums

Anonymous Author(s)
Affiliation
Address
email

1 Tentative Ideas

We discuss here the ideas and questions that are important for modeling structure in online forums. Following are the questions that we want answered:

1. When two persons interact in a thread or a post which topic/community they each belong to
2. When a user U is in communnity C what type of text does he use to communicate
3. Multi-user-Interaction: In a thread a user can post by addressing to a specific user but he is also talking to other users in the thread simultaneously. Can we model this phenomenon
4. There is an inherent bias towards the thread starter or in turn topic of the theread; can such an information be utilised in some form of a prior value/input
5. Multi-layer-Interaction: On the network side of things there are multiple signals which cannt be simply added to make a single signal e.g. different types of edges in the graphs (user calling by username and nick-name). Can the model take this into account. We are not doing this at present.
6. User posts aggregation; there are multiple ways to aggreagte
   (a) Network Layer aggregation: We call all types of edges as a single edge type and use this combined signal.
   (b) aggregating user posts accross multiple threads in the forum.
   (c) Aggregating user post only in the same thread
   (d) Aggregating user post only for same user-user pair interaction; i.e. a user might have posted multiple replies to another user and we aggregate all such replies into one for this user pair interaction.
   (e) No aggregation at all.

2 Graphical Model & Generative Story

Based on the discussions above we came up with the following final model shown in figure 1. In this model, figure 1 below, we aggregate the posts of a given user in a given thread into one document called R_p.

The generative process for the figure is as follows:

Assuming that there are total N_t users in the thread t.

- For each Thread t
  - For each user $p \in \mathcal{N}_t$

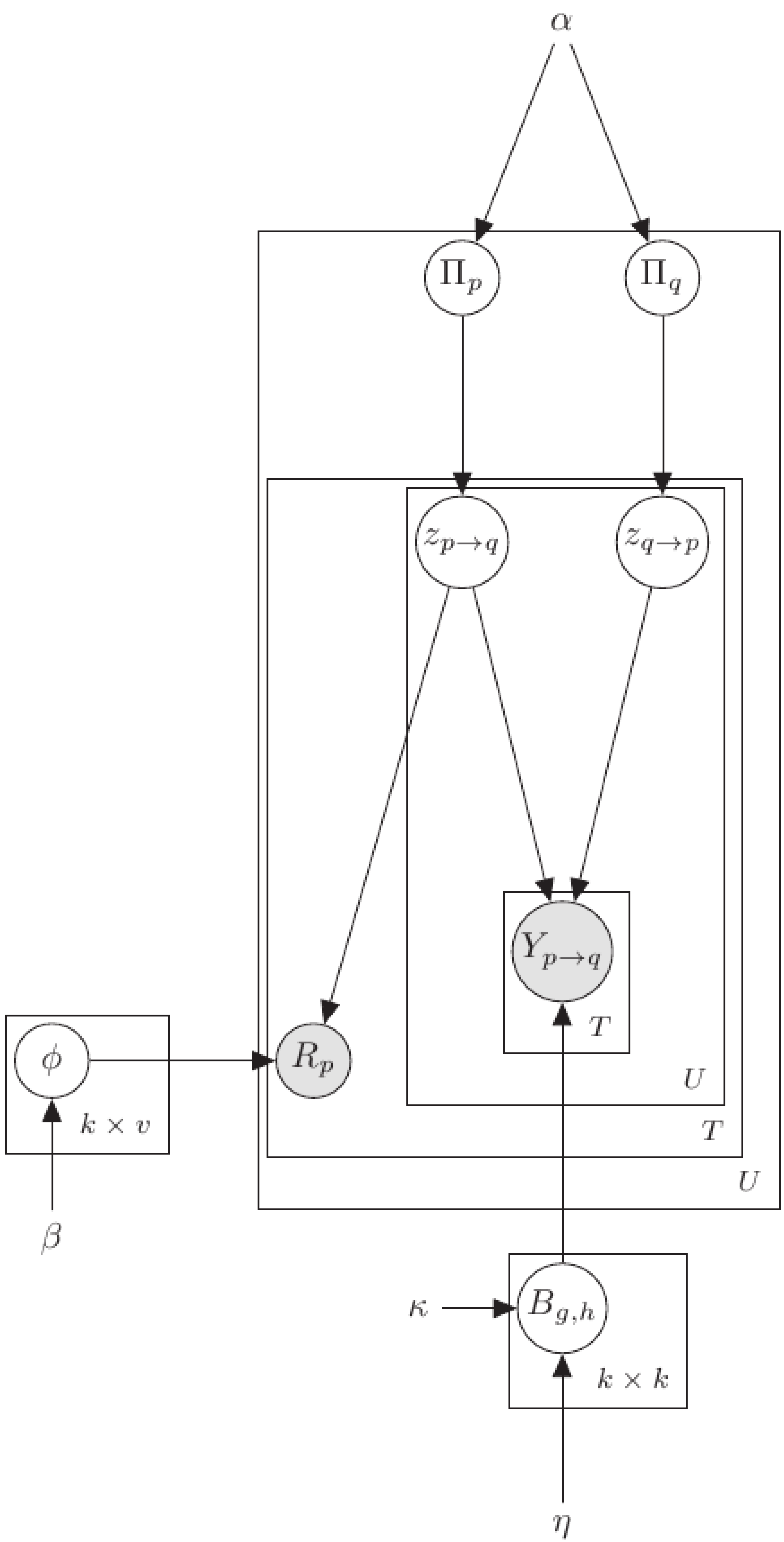

Figure 1: This graphical model takes into account multi-way interaction among users in a thread simultaneously

* Draw a K dimensional mixed membership vector $\vec{\pi}_p \sim$ Dirichlet(α)
* Draw $B(g,h) \sim Gamma(\kappa, \eta)$; where κ, η are parameters of the gamma distribution.

– For each pair of users $(p,q) \in \mathcal{N}_t \times \mathcal{N}_t$:
  * Draw membership indicator for the indicator, $\vec{z}_{(p\to q,t)} \sim$ Multinomial(π_p).
  * Draw membership indicator for the receiver, $\vec{z}_{(q\to p,t)} \sim$ Multinomial(π_q).
  * Sample the value of their interaction, $Y(p,q,t) \sim$ Poisson($\vec{z}^{\top}_{(p\to q,t)} B \vec{z}_{(p\leftarrow q,t)}$).

– For each user $p \in \mathcal{N}_t$
  * Draw ϕ_k from $Dirichlet(\beta)$.
  * Form the set $Q_{p,t}$ that contains all the users that p interacts to on thread t
    · For each word $w \in R_{p,t}$
    · Draw $w \sim \phi(w|z_{(p\to q,t)}, \forall q \in Q_{p,t})$

The data likelihood for the model in figure 1

$$P(Y, R_p|\alpha,\beta,\kappa,\eta) = \int_\Phi\int_\Pi \sum_z P(Y, R_p, z_{p\to q}, z_{p\leftarrow q}, \Phi, \Pi|\alpha,\beta,\kappa,\eta)$$
$$= \int_\Phi\int_\Pi \sum_z \Big[\prod_{p,q}\prod_t P(Y^t_{pq}|z^t_{p\to q}, z^t_{p\leftarrow q}, B)\cdot P(z^t_{p\to q}|\Pi_p)\cdot P(z^t_{p\leftarrow q}|\Pi_q)$$
$$\cdot \Bigg(\prod_p P(\Pi_p|\alpha) \prod_t\prod_p P(R^t_p|z^t_{p\to q}, \Phi)\cdot \prod_k P(\Phi_k|\beta)\Bigg)\cdot \prod_{g,h} P(B_{gh}|\eta,\kappa)\Big] \tag{1}$$

The complete log likeliood of the model is:

$$\log P(Y, W, z_\to, z_\leftarrow, \Phi, \Pi, B|\kappa,\eta,\beta,\alpha) = \sum_t\sum_{p,q} \log P(Y^t_{pq}|z^t_{p\to q}, z^t_{p\leftarrow q}, B) +$$
$$\sum_t\sum_{p,q} (\log P(z^t_{p\to q}|\Pi_p) + \log P(z^t_{p\leftarrow q}|\Pi_p)) + \sum_p \log P(\Pi_p|\alpha) +$$
$$\sum_t\sum_p\sum_{w\in R^t_p} \log P(w|z_{p\to}, \Phi) + \sum_k \log P(\Phi_k|\beta) + \sum_{gh} \log P(B_{gh}|\eta,\kappa) \tag{2}$$

The mean field variational approximation for the posterior is

$$q(z, \Phi, \Pi, B|\Delta_{z_\to}, \Delta_\Phi, \Delta_B, \Delta_{z_\leftarrow}, \Delta_{B_\kappa}) = \prod_t\prod_{p,q}\Big(q_1(z^t_{p\to q}|\Delta_{z_{p\to q}}) + q_1(z^t_{p\leftarrow q}|\Delta_{z_{p\leftarrow q}})\Big)$$
$$\cdot \prod_p q_4(\Pi_p|\Delta_{\Pi_p}) \prod_k q_3(\Phi_k|\Delta_{\Phi_k}) \prod_{g,h} q(B_{g,h}|\Delta_{B_\eta}, \Delta_{B_\kappa}) \tag{3}$$

The lower bound for the data log-likelihood from jensen's inequality is:

$$L_\Delta = E_q\Big[\log P(Y, W, z_\to, z_\leftarrow, \Phi, \Pi, B|\kappa,\eta,\beta,\alpha) - \log q \Big] \tag{4}$$

$$L_\Delta = E_q\Bigg[\sum_t\sum_{p,q}\log\left(B_{g,h}^{Y_{p,q}^t}\frac{e^{-B_{gh}}}{Y_{pq}^t!}\right) + \sum_t\sum_{pq}\log\left(\prod_k(\pi_{p,k}^{z_{p\to q}=k})\right) + \sum_t\sum_{p,q}\log\Bigg(\prod_k(\pi_{q,k})^{z_{p\leftarrow q}=k}\Bigg) + \sum_p\log\left[\prod_k(\Pi_{p,k})^{\alpha_k-1}\cdot\frac{\Gamma(\sum\alpha_k)}{\prod_k\Gamma(\alpha_k)}\right] + \sum_t\sum_p\sum_{w\in R_p^t}\log\Bigg(\prod_{u\in V}(\bar{z}^T\phi_u)^{w=u}\Bigg) + \sum_k\log\left[\prod_{u\in V}(\phi_{k,u})^{\beta_k-1}\cdot\frac{\Gamma(\sum\beta_k)}{\prod_k\Gamma(\beta_k)}\right] + \sum_{g,h}\log\Big(B_{g,h}^{\kappa-1}/\eta^\kappa\Gamma(\kappa)\cdot\exp(-B_{g,h}/\eta)\Big)$$

$$- E_q\Bigg[\sum_t\sum_{p,q}\log\big(\prod_k(\Delta_{z_{p\to q},k})^{z_{p\to q}=k}\big) + \sum_t\sum_{p,q}\log\big(\prod_k(\Delta_{z_{p\leftarrow q},k})^{z_{p\leftarrow q}=k}\big) + \sum\log\big[\prod_k(\Pi_{p,k})^{\Delta_{\pi_{pk}}-1}\frac{\Gamma(\Delta_{\Pi_p})}{\prod_{k=1}\Gamma(\Delta_{\Pi_{p,k}})}\big] + \sum_k\log\big[\prod_{u\in v}(\Phi_{k,u})^{\Delta_{\Phi_{ku}}-1)}\frac{\Gamma(\Delta_{\Phi_k})}{\prod_{u\in v}\Gamma(\Delta_{\Phi_{k,u}})}\big] + \sum_{g,h}\log\big[\frac{B_{g,h}^{\Delta_{\kappa=1}}}{\Delta_\eta^{\Delta_\kappa}\Gamma(\Delta_\kappa)}\exp(-B_{g,h}/\Delta_\eta)\big] \tag{5}$$

Equation **??** is the variational lower bound of the log likelihood function which is to be maximized. There are terms like $E_q\left[\sum_{g,h}\log\Big(B_{g,h}^{\kappa-1}/\eta^\kappa\Gamma(\kappa)\cdot\exp(-B_{g,h}/\eta)\Big)\right]$ which can be obtained by taking derivation of the partition function of the exponential family form of gamma distribution. I still have to figure out an effective way to evaluate $E_q\left[\sum_t\sum_p\sum_{w\in R_p^t}\log\Bigg(\prod_{u\in V}(\bar{z}^T\phi_u)^{w=u}\Bigg)\right]$ which Chong suggested (and Eric too in the last meeting) to evaluate by intriducing an additional latent variable $\bar{z}_p$ which is a realization of the average $\frac{\sum_{q\in Q}z_{p\to q}}{|Q|}$.