A Network Model for Dynamic Textual Communications with Application to Government Email Corpora

Bomin Kim[1], Aaron Schein[3], Bruce Desmarais[1], and Hanna Wallach[2,3]

[1]Pennsylvania State University
[2]Microsoft Research NYC
[3]University of Massachusetts Amherst

May 28, 2017

1 Tie Generating Process

We assume the following generative process for each document d in a corpus D:

1. Choose the number of recipients
$$R^{(d)} \sim \text{Poission}(\delta), \tag{1}$$
which is analogous to the number of words in LDA (use separate parameter not involving λ's). For simplicity, we can assume $R^{(d)}$ is known such that we do not have to infer δ, as we do for the number of words.

2. (Data augmentation) For each sender $i \in \{1, ..., A\}$, create a list of receivers J_i by applying multivariate Wallenius' noncentral hypergeometric distribution (MWNCHypergeo) to every $j \in \mathcal{A}_{\setminus i}$
$$J_i^{(d)} \sim \text{MWNCHypergeo}\Big(\mathbf{m} = \mathbf{1}_{A-1}, N = R^{(d)}, \boldsymbol{\omega} = \{\lambda_{ij}^{(d)}\}_{j \in \mathcal{A}_{\setminus i}}\Big), \tag{2}$$
where $\mathbf{m}$ is the vector of availability (we have maximum 1 available for each actor except the sender), N is the total number of receivers to be sampled, and $\boldsymbol{\omega}$ is the weight for each actor to be sampled. Same as before, $\lambda_{ij}^{(d)}$ is evaluated at time $t_+^{(d-1)}$. Note that $_+$ denotes including the timepoint itself, meaning that λ_{ij} is obtained using the history of interactions until and including the timestamp $t^{(d-1)}$. For example, we can use R function

```
library(BiasedUrn)
J_i = rMFNCHypergeo(nran = 1, m = c(1,1,1,1), n = 2, odds = c(0.1, 0.2, 0.3, 0.4))
```

3. For every sender $i \in \mathcal{A}$, generate the time increments
$$\Delta T_{iJ_i}^{(d)} \sim \text{Exp}(\lambda_{iJ_i}^{(d)}), \tag{3}$$
where $\lambda_{iJ_i}^{(d)}(t) = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{\lambda_0^{(c)} + \frac{1}{|J_i|} \sum_{j \in J_i} \boldsymbol{b}^{(c)T} \boldsymbol{x}_t^{(c)}(i,j)\Big\} \cdot \prod_{j \in J_i} 1\{j \in \mathcal{A}_{\setminus i}\}$.

4. Set timestamp, sender, and receivers simultaneously (NOTE: $t^{(0)} = 0$):
$$\begin{aligned} t^{(d)} &= t^{(d-1)} + \min(\Delta T_{iJ_i}), \\ i^{(d)} &= i_{\min(\Delta T_{iJ_i})}, \\ J^{(d)} &= J_{i^{(d)}}. \end{aligned} \tag{4}$$

NOTE: We have to come up with the way to treat $R^{(d)} = 0$ case if we want to assum $R^{(d)}$ unknown. One possible option is $R^{(d)} = 1 + \text{Poisson}(\delta)$, however, this will give biased estimate of δ from the inference. Or, if we allow empty receiver documents (such as empty token documents), we actually generate documents that are never observed in the real dataset.

2 Tie Generating Process - No data augmentation

We assume the following generative process for each document d in a corpus D:

1. Choose the number of recipients
$$R^{(d)} \sim \text{Poission}(\delta), \tag{5}$$
which is analogous to the number of words in LDA (use separate parameter not involving λ's). For simplicity, we can assume $R^{(d)}$ is known such that we do not have to infer δ, as we do for the number of words.

2. Choose the sender:
$$i^{(d)} \sim \text{Multinomial}(\{\frac{\lambda_i^{(d)}}{\sum_{i=1}^{A}\lambda_i^{(d)}}\}_{i=1}^{A}), \tag{6}$$
where $\lambda_i^{(d)}(t) = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{\lambda_0^{(c)} + \frac{1}{A-1}\sum_{j\in\mathcal{A}_{\setminus i}} \boldsymbol{b}^{(c)T}\boldsymbol{x}_t^{(c)}(i,j)\Big\} \cdot \prod_{j\in\mathcal{A}_{\setminus i}} 1\{j \in \mathcal{A}_{\setminus i}\}$ is the stochastic intensity of the actor i, calculated from the average of all possible recipients.

3. For the chosen sender $i^{(d)}$, choose the recipient $J^{(d)}$ by applying multivariate Wallenius' non-central hypergeometric distribution (MWNCHypergeo) to every $j \in \mathcal{A}_{\setminus i}$
$$J^{(d)} \sim \text{MWNCHypergeo}\Big(\mathbf{m} = \mathbf{1}_{A-1}, N = R^{(d)}, \boldsymbol{\omega} = \{\lambda_{i^{(d)}j}^{(d)}\}_{j\in\mathcal{A}_{\setminus i^{(d)}}}\Big), \tag{7}$$
where $\mathbf{m}$ is the vector of availability (we have maximum 1 available for each actor except the sender), N is the total number of receivers to be sampled, and $\boldsymbol{\omega}$ is the weight for each actor to be sampled. Same as before, $\lambda_{ij}^{(d)}$ is evaluated at time $t_+^{(d-1)}$. Note that $_+$ denotes including the timepoint itself, meaning that λ_{ij} is obtained using the history of interactions until and including the timestamp $t^{(d-1)}$. For example, we can use R function

```
library(BiasedUrn)
J_i = rMFNCHypergeo(nran = 1, m = c(1,1,1,1), n = 2, odds = c(0.1, 0.2, 0.3, 0.4))
```

4. For the chosen sender and recipients $(i^{(d)}, J^{(d)})$, generate one time increment
$$\Delta T_{i^{(d)}J^{(d)}}^{(d)} \sim \text{Exp}(\lambda_{i^{(d)}J^{(d)}}^{(d)}), \tag{8}$$
where $\lambda_{i^{(d)}J^{(d)}}^{(d)} = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{\lambda_0^{(c)} + \frac{1}{|J^{(d)}|}\sum_{j\in J^{(d)}} \boldsymbol{b}^{(c)T}\boldsymbol{x}_t^{(c)}(i^{(d)},j)\Big\} \cdot \prod_{j\in J^{(d)}} 1\{j \in \mathcal{A}_{\setminus i^{(d)}}\}$.

5. Set timestamp, sender, and receivers from 2, 3, and 4, (NOTE: $t^{(0)} = 0$):
$$\begin{aligned} t^{(d)} &= t^{(d-1)} + \Delta T_{i^{(d)}J^{(d)}}^{(d)} \text{ from 4} \\ i^{(d)} &= i^{(d)} \text{ from 2} \\ J^{(d)} &= J^{(d)} \text{ from 3} \end{aligned} \tag{9}$$

The second version is very similar to the collapsed-time generative process we derived, but this one does not generate the latent edges at all. This one might be equation-wise simpler and computaiton-ally faster, since it does not involve latent edge generating steps. However, I do not like the idea of using averages (over all possible receivers) in 2, because the urgency of document should depend on 'who are the recipients', instead of 'general desire of sending the document'.