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

Anonymous Authors[1]

1. Normalizing constant of Gibbs measure

The non-empty Gibbs measure defines the probability of author i selecting the binary recipient vector $\boldsymbol{u}_{id}$ as

$$\begin{aligned} &P(\boldsymbol{u}_{id}|\delta, \boldsymbol{\lambda}_{id}) \\ &= \frac{\exp\Big\{\log\big(\mathrm{I}(\|\boldsymbol{u}_{id}\|_1 > 0)\big) + \sum_{j\neq i}(\delta+\lambda_{idj})u_{idj}\Big\}}{Z(\delta, \boldsymbol{\lambda}_{id})}. \end{aligned}$$

To use this distribution efficiently, we derive a closed-form expression for $Z(\delta, \boldsymbol{\lambda}_{id})$ that does not require brute-force summation over the support of $\boldsymbol{u}_{id}$ (*i.e.* $\forall \boldsymbol{u}_{id} \in [0,1]^A$). We recognize that if $\boldsymbol{u}_{id}$ were drawn via independent Bernoulli distributions in which $P(u_{idj} = 1|\delta, \boldsymbol{\lambda}_{id})$ was given by $\mathrm{logit}(\delta + \lambda_{idj})$, then

$$P(\boldsymbol{u}_{id}|\delta, \boldsymbol{\lambda}_{id}) \propto \exp\Big\{\sum_{j\neq i}(\delta+\lambda_{idj})u_{idj}\Big\}.$$

This is straightforward to verify by looking at

$$P(u_{idj} = 1|\boldsymbol{u}_{id[-j]}, \delta, \boldsymbol{\lambda}_{id}) = \frac{\exp{(\delta+\lambda_{idj})}}{\exp{(\delta+\lambda_{idj})}+1}.$$

We denote the logistic-Bernoulli normalizing constant as $Z^l(\delta, \boldsymbol{\lambda}_{id})$, which is defined as

$$Z^l(\delta, \boldsymbol{\lambda}_{id}) = \sum_{\boldsymbol{u}_{id}\in[0,1]^A} \exp\Big\{\sum_{j\neq i}(\delta+\lambda_{idj})u_{idj}\Big\}.$$

Now, since

$$\begin{aligned} &\exp\Big\{\log\Big(\mathrm{I}(\|\boldsymbol{u}_{id}\|_1 > 0)\Big) + \sum_{j\neq i}(\delta+\lambda_{idj})u_{idj}\Big\} \\ &= \exp\Big\{\sum_{j\neq i}(\delta+\lambda_{idj})u_{idj}\Big\}, \end{aligned}$$

except when $\|\boldsymbol{u}_{id}\|_1 = 0$, we note that

$$\begin{aligned} Z(\delta, \boldsymbol{\lambda}_{id}) &= Z^l(\delta, \boldsymbol{\lambda}_{id}) - \exp\Big\{\sum_{\forall u_{idj}=0}(\delta+\lambda_{idj})u_{idj}\Big\} \\ &= Z^l(\delta, \lambda_i^{(d)}) - 1. \end{aligned}$$

We can therefore derive a closed form expression for $Z(\delta, \boldsymbol{\lambda}_{id})$ via a closed form expression for $Z^l(\delta, \boldsymbol{\lambda}_{id})$. This can be done by looking at the probability of the zero vector under the logistic-Bernoulli model:

$$\frac{\exp\Big\{\sum\limits_{\forall u_{idj}=0}(\delta+\lambda_{idj})u_{idj}\Big\}}{Z^l(\delta, \boldsymbol{\lambda}_{id})} = \prod_{j\neq i}\Big(1 - \frac{\exp{(\delta+\lambda_{idj})}}{\exp{(\delta+\lambda_{idj})}+1}\Big).$$

Then, we have

$$\frac{1}{Z^l(\delta, \boldsymbol{\lambda}_{id})} = \prod_{j\neq i}\frac{1}{\exp(\delta+\lambda_{idj})+1}.$$

Finally, the closed form expression for the normalizing constant under the non-empty Gibbs measure is

$$Z(\delta, \boldsymbol{\lambda}_{id}) = \prod_{j\neq i}\big(\exp\{\delta+\lambda_{idj}\}+1\big) - 1.$$

2. Specification of Network Features

We provide the details on the specification of $\boldsymbol{x}_{idjc} = (x^1_{idjc}, x^2_{idjc}, x^3_{idjc})$ in Section 4.2. We first partitioned the interval $[t_d - 16d, t_d)$ into $L = 3$ sub-intervals with equal length in the log-scale, by setting the difference Δ_l = (6 hours) $\times 4^l$ for $l = 1, 2, 3$. In other words, we define the intervals I^l_d by

$$\begin{aligned} &[t_d - 384h, t_d) \\ &= [t_d - 384h, t_d - 96h) \cup [t_d - 96h, t_d - 24h) \cup [t_d - 24h, t_d) \\ &= I^3_d \cup I^2_d \cup I^1_d, \end{aligned}$$

where I^l_d is the half-open interval $[t_d - \Delta_l, t_d - \Delta_{l-1})$.

Then, for each time interval $l = 1, 2, 3$, the degree and dyadic statistics are defined as:

1. $\textbf{outdegree}^l_{id\cdot c} = \sum\limits_{d\in I^l_d}\pi_{dc}I\{i\rightarrow\forall j\}$
2. $\textbf{indegree}^l_{\cdot djc} = \sum\limits_{d\in I^l_d}\pi_{dc}I\{\forall i\rightarrow j\}$
3. $\textbf{send}^l_{idjc} = \sum\limits_{d\in I^l_d}\pi_{dc}I\{i\rightarrow j\}$

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

4. **receive**$^l_{idjc} = \sum_{d \in I^l_d} \pi_{dc} I\{j \rightarrow i\}$

Next, we define four triadic statistics involving pairs of messages, which are analogous to 2-path statistics commonly used in the network science literature. While earlier works (Perry & Wolfe, 2013) adapted full sets of triadic statistics for each combination of time intervals (e.g. $3 \times 3 = 9$), we maintain 3 intervals per each statistic, by defining 3×3 time windows and sum the combination-specific statistics based on the interval where the triads are closed (Refer to Figure 1). As a result, our interval-adjusted definitions of triadic effects become

5. **2-send**$^l_{idjc} =$

$$\sum_{\max(l_1,l_2)=l} \sum_{\substack{h \neq i \\ h \neq j}} \Big(\sum_{d \in I^{l_1}_d} \pi_{dc} I\{i \rightarrow h\}\Big)\Big(\sum_{d \in I^{l_2}_d} \pi_{dc} I\{h \rightarrow j\}\Big)$$

6. **2-receive**$^l_{idjc} =$

$$\sum_{\max(l_1,l_2)=l} \sum_{\substack{h \neq i \\ h \neq j}} \Big(\sum_{d \in I^{l_1}_d} \pi_{dc} I\{h \rightarrow i\}\Big)\Big(\sum_{d \in I^{l_2}_d} \pi_{dc} I\{j \rightarrow h\}\Big)$$

7. **sibling**$^l_{idjc} =$

$$\sum_{\max(l_1,l_2)=l} \sum_{\substack{h \neq i \\ h \neq j}} \Big(\sum_{d \in I^{l_1}_d} \pi_{dc} I\{h \rightarrow i\}\Big)\Big(\sum_{d \in I^{l_2}_d} \pi_{dc} I\{h \rightarrow j\}\Big)$$

8. **cosibling**$^l_{idjc} =$

$$\sum_{\max(l_1,l_2)=l} \sum_{\substack{h \neq i \\ h \neq j}} \Big(\sum_{d \in I^{l_1}_d} \pi_{dc} I\{i \rightarrow h\}\Big)\Big(\sum_{d \in I^{l_2}_d} \pi_{dc} I\{j \rightarrow h\}\Big)$$

where $l_1 = 1, 2, 3$ and $l_2 = 1, 2, 3$.

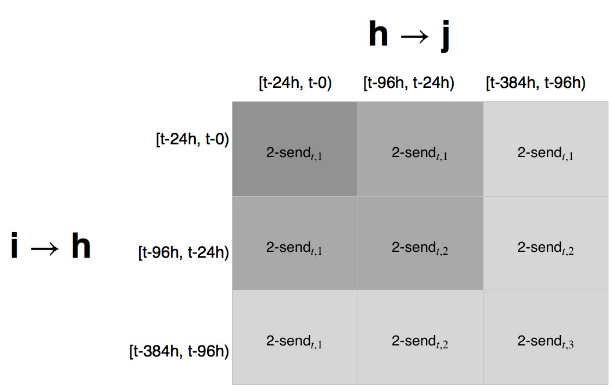

Figure 1. 2-send defined for each interval $l = 1, 2, 3$. Cells with same shades sum up to one statistic.