Joint Modeling of Content-Partitioned Multinetwork Embeddings (CPME) and Point Process Approach

Bomin Kim

May 19, 2016

Abstract

Your abstract.

1 Ideas

Current CPME model does not involve any of temporal component, which plays a key role in email interactions. Intuitively, past interaction behaviors significantly influence future ones; for example, if an actor i sent an email to actor j, then j is highly likely to send an email back to i as a response (i.e. reciprocity). Moreover, the recency and frequency of past interactions can also be considered to effectively predict future interactions. Thus, as an exploratory data analysis, point process model for directional interaction is applied to the North Carolina email data. Starting from the existing framework focused on the analysis of content-partitioned subnetworks, I would suggest an extended approach to analyze the data using the timestamps in the email, aiming to develop a joint dynamic or longitudinal model of text-valued ties.

CPME model is a Bayesian framework using two well-known methods: Latent Dirichlet Allocation (LDA) and Latent Space Model (LSM). Basically, existence of edge depends on topic assignment t (LDA) and its corresponding interaction pattern c. Each topic t=1,. . . ,T has one interaction pattern c=1,. . . ,C, and each interaction pattern posits unique latent space (LSM), thus generating $A \times A$ matrix of probabilities $P^{(c)}$ that a message author a will include recipient r on the message, given that it is about a topic in cluster c. Incorporating point process approach, now assume that under each interaction pattern, we have $A \times A$ matrix of stochastic intensities $\lambda^{(c)}(t)$ which depend on the history of interaction between the sender and receiver.

2 CPME + Point Process Model

Before we build up the ultimate joint model of LDA, LSM, and point process approach, we first start with simpler model which combines LDA and point process approach.

2.1 General framework

In this section, we introduce multiplicative Cox regression model for the edge formation process in a longitudinal communication network. For concreteness, we frame our discussion of this model in terms of email data, although it is generally applicable to any similarly-structured communication data.

A single email, indexed by d, is represented by a set of tokens $w^{(d)} = \{w_n^{(d)}\}_{n=1}^{N^{(d)}}$ that comprise the text of that email, an integer $a^{(d)} \in \{1, ..., A\}$ indicating the identity of that email's author, and an integer $t^{(d)} \in [0, T]$ indicating the (unix time-based) timestamp of that email.

As in LDA, each email, indexed by d, has a discrete distribution over topics $\theta^{(d)}$. A Dirichlet prior with concentration parameter α is placed (i.e. $\theta^{(d)} \sim \text{Dir}(\alpha, m)$). Then, a "topic" t is characterized by a discrete distribution over V word types with probability vector $\phi^{(t)}$. A symmetric Dirichlet prior with concentration parameter β is placed (i.e. $\phi^{(t)} \sim \text{Dir}(\beta, n)$). Now, each topic t is associated with one "interaction pattern" among C different types, which is assigned by $l_t \sim \text{Unif}(1, C)$.

To capture the relationship between the interaction patterns expressed in an email and that email's recipients, topics that share C are associated with an $A \times A$ matrix of $N_{ar}^{(c)}(t)$, a counting process denoting the number of edges (emails) from actor a to actor r up to time t. NOTE: We use the partition C since we expect that some topics/interaction patterns have little variation among the pairs of actors (e.g. broadcasting), while some have large variation (e.g. meeting scheduling, personal affairs). Combining the individual counting processes of all potential edges, $\mathbf{N}^{(c)}(t)$ is the multivariate counting process with $\mathbf{N}^{(c)}(t) = (N_{ar}^{(c)}(t) : a, r \in 1, ..., A, a \neq r)$. Here we make no assumption about the independence of individual edge counting process. As in Vu et al. (2011b), we model the multivariate counting process via Doob-Meyer decomposition:

$$\mathbf{N}^{(c)}(t) = \int_0^t \boldsymbol{\lambda}^{(c)}(s)ds + \mathbf{M}(t) \tag{1}$$

where essentially $\boldsymbol{\lambda}^{(c)}(t)$ and $\mathbf{M}(t)$ may be viewed as the (deterministic) signal and (martingale) noise, respectively.

Following the multiplicative Cox model of the intensity process $\boldsymbol{\lambda}^{(c)}(t)$ given $\boldsymbol{H}_{t-}^{(c)}$, the entire past of the network related to C up to but not including time t, we consider for each potential directed edge (a, r) the intensity forms:

$$\lambda_{ar}^{(c)}(t|\boldsymbol{H}_{t-}^{(c)}) = \lambda_0^{(c)}(t) \cdot \exp\{\beta^{(c)T} x^{(c)}(a, r, t)\}\} \cdot 1\{r \in \mathcal{A}_{(a,t)}^{(c)}\} \tag{2}$$

where $\lambda_0^{(c)}(t)$ is the baseline hazards for the interactin pattern C, $\beta^{(c)}$ is an unknown vector of coefficients in R^p, $x^{(c)}(a,r,t)$ is a vector of p statistics for directed edge (a,r) constructed based on $\boldsymbol{H}_{t-}^{(c)}$ (examples of these statistics are given in the next section), and $\mathcal{A}_{(a,t)}^{(c)}$ is the predictable receiver set of sender a at time t within all actors $A^{(c)}$. NOTE: We assume that all possible actor set $A^{(c)}$ varies depending on the interaction pattern (e.g. confidential communication do not move outside of certain actors)

2.2 Dynamic covariates to measure network effects

The network statistics $x^{(c)}(a,r,t)$ of equations (2), corresponding to the ordered pair (a,r), can be time-invariant (such as gender) or time-dependent (such as the number of two-paths from a to r just before time t). Since time-invariant covariates can be easily specified in various manners (e. g. homophily or group-level effects), here we only consider specification of dynamic covariates.

Following Perry and Wolfe (2013), we use 6 effects as components of $x^{(c)}(a,r,t)$. The first two behaviors (send and receive) are dyadic, involving exactly two actors, while the last four (2-send, 2-receive, sibling, and cosibling) are triadic, involving exactly three actors. However, different from Perry and Wolfe (2013), we define the effects not based on finite sub-interval, which require large number of dimention. Instead, we create single statistic for each effect by incorporating the recency of event into the statistic itself. Moreover, we take the interaction pattern C into account as well, based on the topic-token-interaction pattern assignment from LDA.

1. $\textbf{send}^{(c)}(a,r,t) = \sum_{d:t_d<t} \frac{\bar{N}^{(c|d)}}{\bar{N}^{(d)}} \cdot I\{a \to r\} \cdot g(t-t_d)$
2. $\textbf{receive}^{(c)}(a,r,t) = \sum_{d:t_d<t} \frac{\bar{N}^{(c|d)}}{\bar{N}^{(d)}} \cdot I\{r \to a\} \cdot g(t-t_d)$
3. $\textbf{2-send}^{(c)}(a,r,t) = \sum_{h\neq a,r} (\sum_{d:t_d<t} \frac{\bar{N}^{(c|d)}}{\bar{N}^{(d)}} \cdot I\{a \to h\} \cdot g(t-t_d))(\sum_{d:t_d<t} \frac{\bar{N}^{(c|d)}}{\bar{N}^{(d)}} \cdot I\{h \to r\} \cdot g(t-t_d))$
4. $\textbf{2-receive}^{(c)}(a,r,t) = \sum_{h\neq a,r} (\sum_{d:t_d<t} \frac{\bar{N}^{(c|d)}}{\bar{N}^{(d)}} \cdot I\{h \to a\} \cdot g(t-t_d))(\sum_{d:t_d<t} \frac{\bar{N}^{(c|d)}}{\bar{N}^{(d)}} \cdot I\{r \to h\} \cdot g(t-t_d))$
5. $\textbf{sibling}^{(c)}(a,r,t) = \sum_{h\neq a,r} (\sum_{d:t_d<t} \frac{\bar{N}^{(c|d)}}{\bar{N}^{(d)}} \cdot I\{h \to a\} \cdot g(t-t_d))(\sum_{d:t_d<t} \frac{\bar{N}^{(c|d)}}{\bar{N}^{(d)}} \cdot I\{h \to r\} \cdot g(t-t_d))$
6. $\textbf{cosibling}^{(c)}(a,r,t) = \sum_{h\neq a,r} (\sum_{d:t_d<t} \frac{\bar{N}^{(c|d)}}{\bar{N}^{(d)}} \cdot I\{a \to h\} \cdot g(t-t_d))(\sum_{d:t_d<t} \frac{\bar{N}^{(c|d)}}{\bar{N}^{(d)}} \cdot I\{r \to h\} \cdot g(t-t_d))$

Here, $g(t-t_d)$ reflects the difference between current time t and the timestamp of previous email t_d, thus measuring the recency. Inspired by the self-exciting Hawkes process, which is often used to model the temporal effect of email data, we can take the exponential kernel $g(t-t_d) = we^{-w(t-t_d)}$ where w is the parameter of speed at which sender replies to emails, with larger values indicating faster response times. Indeed, w^{-1} is the expected number of hours it takes to reply to a typical email. NOTE: First, we can let w to be hyperparameter to

be specified (since β will be adjusted to the scale of w). However, later we can try to estimate it via Bayesian approach, or even consider sender-specific w_a or pair-specific w_{ar} and estimate them.

2.3 Inference

Following Perry and Wolfe (2013), after treating the baseline rate $\lambda_0(t)$ as a nuisance parameter, we can estimate the coefficient vector β using the log-partial likelihood at time t:

$$\log PL_t(\beta) = \sum_{t_m \leq t} \Big\{ \beta^T x(a_m, r_m, t_m) - \log\big[\sum_{j \in \mathcal{J}_{t_m}(i_m)} \exp\{\beta_0^T x_{t_m}(i_m, j)\}\big] \Big\}. \quad (3)$$

3 Preliminary Analysis

Hurricane Sandy was the most destructive hurricane in 2012, which hit North Carolina on late October (October 28, Governor Bev Perdue declared a state of emergency in 24 western counties due to snow and strong winds). In our dataset, there are three counties which cover the date of Hurricane Sandy (October 22, 2012 – November 2, 2012), so we focus on the three counties, since the timestamp of email in this case is much more important than usual case without any disastrous event.

3.1 Dare County

Period	**Before Sandy**	**During Sandy**	**After Sandy**	**Overall**
# emails	1933	1563	1467	4963

Table 1: Summary of Dare county email data based on time period

Before Sandy ranges from 2012-09-01 to 2012-10-21 (7 weeks), During Sandy ranges from 2012-10-22 to 2012-11-02 (2 weeks), and After Sandy ranges from 2012-11-03 to 2012-11-30 (4 weeks).

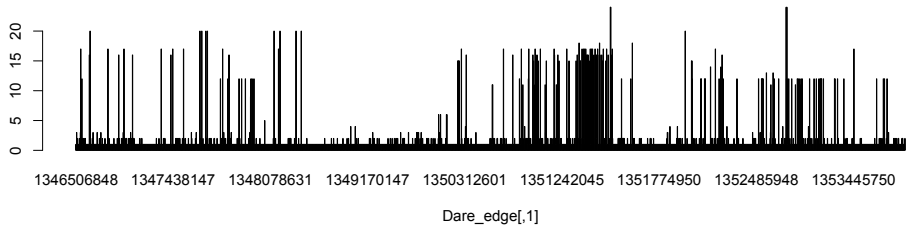

Figure 1: Frequency of Dare county emails from 2012-09-01 to 2012-11-30

Time Interval	send	receive
$[-\infty, t)$	2.128, 2.659, 2.355, 2.919	0.292, 0.257, 0.047, 0.110
$[t-30m, t)$	0.262, -0.064, 0.782, 0.317	2.087, 1.287 , 2.346, 1.870
$[t-2h, t-30m)$	0.383, 0.157 , 0.024, -0.045	0.553, 0.082, 0.794, 0.269
$[t-8h, t-2h)$	0.816, 0.054 , 0.077, 0.381	-0.221, 0.048, 0.298, -0.012
$[t-32h, t-8h)$	0.085, 0.014, 0.228, 0.070	0.101, 0.017, -0.033, 0.019
$[t-5.33d, t-32h)$	0.103, 0.025, 0.092, 0.008	-0.027, -0.016, -0.033, -0.009
$[t-21.33d, t-5.33d)$	0.052, 0.000, 0.059, 0.010	0.013, 0.030 , -0.016, 0.013
$[-\infty, t-21.33d)$	0.052, 0.103, 0.027, 0.021	0.008, 0.000, 0.020, -0.005

Table 2: Estimated coefficients and approximate standard errors for dyadic effects of Dare county data (before Sandy, during Sandy, after Sandy, overall)

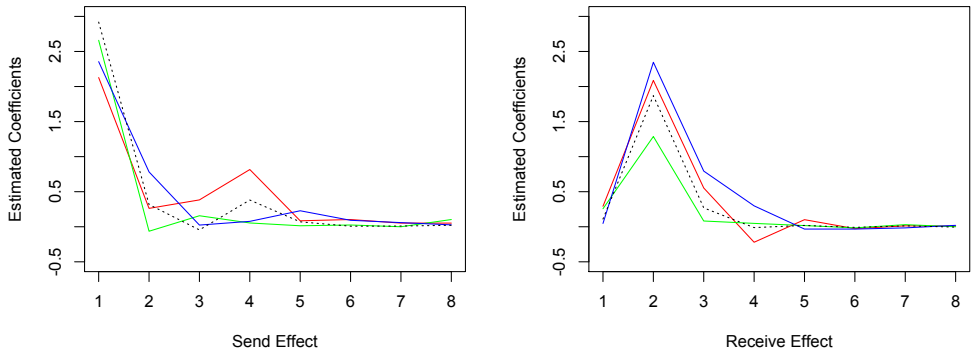

Figure 2: Comparison of Send (left) and Receive (right) effect based on periods in Table 1. (Red=Before, Green=During, Blue=After, and dot=Overall)

3.2 Lenoir County

Before Sandy ranges from 2012-10-01 to 2012-10-21 (3 weeks), During Sandy ranges from 2012-10-22 to 2012-11-02 (2 weeks), and After Sandy ranges from 2012-11-03 to 2012-12-31 (8 weeks).

Period	Before Sandy	During Sandy	After Sandy	Overall
# emails	216	83	302	601

Table 3: Summary of Lenoir county email data based on time period

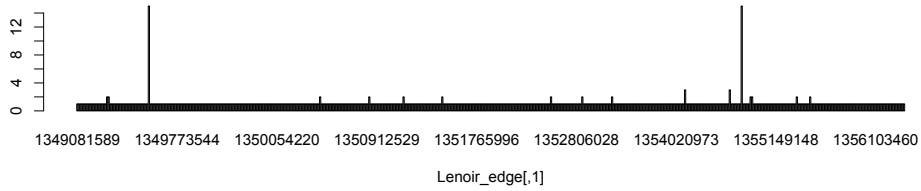

Figure 3: Frequency of Lenoir county emails from 2012-10-01 to 2012-12-31

3.3 Vance County

Period	Before Sandy	During Sandy	After Sandy	Overall
# emails	198	18	55	271

Table 4: Summary of Vance county email data based on time period

Before Sandy ranges from 2012-09-04 to 2012-10-21 (7 weeks), During Sandy ranges from 2012-10-22 to 2012-11-02 (2 weeks), and After Sandy ranges from 2012-11-03 to 2012-11-30 (4 weeks).

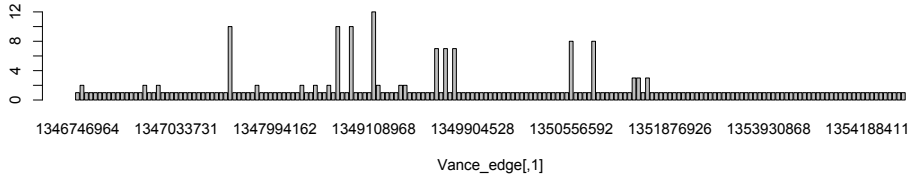

Figure 4: Frequency of Vance county emails from 2012-09-04 to 2012-11-30