The Hyperedge Event Model

Bomin Kim

Department of Statistics
Pensylvania State University

July 29, 2018

Joint Statistical Meetings 2018

Collaborators

- Aaron Schein

  College of Information and Computer Sciences, UMass Amherst
- Bruce Desmarais

  Department of Political Science, Pennylvania State University
- Hanna Wallach

  Microsoft Research NYC

Motivations: Hyperedge Event Model

- Hyperedge: directed edges from one sender to multiple receivers or from multiple senders to one receiver
- Event: timestamped events in the continous-time scale
- Model: statistical framework to jointly understand

“who interacts with whom, and when?”

Generative Process: "Who Interacts with Whom"

For event $e = 1, \ldots, E$, between $i \in \{1, \ldots, A\}$ & $j \in \{1, \ldots, A\}$,

- Receiver intensity for every sender-receiver pair $(i, j)_{i \neq j}$

$$\lambda_{iej} = \boldsymbol{b}^T \boldsymbol{x}_{iej},$$

  where $\boldsymbol{x}_{iej}$ is a set of receiver selection features or covariates and $\boldsymbol{b}$ is the corresponding P-dimensional coefficient.

- Every sender i selects candidate receivers from non-empty multivariate Bernoulli distribution[1] $\boldsymbol{u}_{ie} \sim \text{MB}_G(\lambda_{ie1}, \ldots, \lambda_{ieA})$

$$P(\boldsymbol{u}_{ie} | \boldsymbol{b}, \boldsymbol{x}_{iej}) \propto \exp\Big(\log(I(||\boldsymbol{u}_{ie}||_1 > 0)) + \sum_{j \neq i} \lambda_{iej} u_{iej}\Big)$$

[1] Fellows and Handcock (2017); Dai et al. (2013)

Generative Process: "and When"

- Timing rate for each sender i

$$\mu_{ie} = g^{-1}(\boldsymbol{\eta}^T \boldsymbol{y}_{ie}),$$

where $\boldsymbol{y}_{ie}$ is a set of timing features or covariates and $\boldsymbol{\eta}$ is the corresponding Q-dimensional coefficient.

- Generalized linear model (GLM) for time increment τ_{ie} so that

$$E(\tau_{ie}) = \mu_{ie} \text{ and } V(\tau_{ie}) = V(\mu_{ie}),$$

with a choice of distribution from exponential family.

- Select the sender-receiver-set with the smallest time increment[2]

$$s_e = \text{argmin}_i(\tau_{ie}),$$
$$\boldsymbol{r}_e = \boldsymbol{u}_{s_e e},$$
$$t_e = t_{e-1} + \tau_{s_e e}.$$

[2] Snijders (1996)

Generative Process: Sender, Receivers, and Timestamps[3]

Candidate Receivers

i \ j	1	2	3	4	5
1	0	1	0	0	0
2	1	0	1	0	0
3	0	0	0	0	1
4	1	1	1	0	1
5	0	1	1	1	0

Candidate Timestamps

i	t
1	0.1
2	1
3	10
4	5
5	0.2

Generate Data

sender: 1
receiver: 2
timestamp:0.1

- Bayesian inference: invert the generative process to obtain the posterior distribution over the latent variables—i.e., $(\boldsymbol{u}, \boldsymbol{b}, \boldsymbol{\eta})$.

[3] Assuming $t_{e-1} = 0$ for simplicity.

Application: Montgomery County Government Email Data[4]

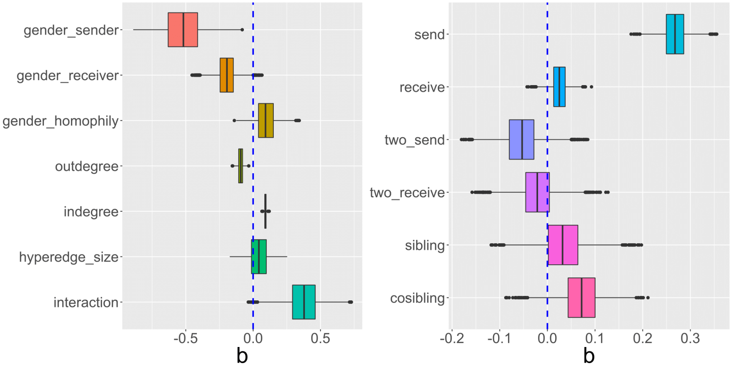
- Email corpora covering inboxes and outboxes of Montgomery county government managers in North Carolina
- Contains $E = 680$ emails, sent and received by $A = 18$ department managers over 3 months (March–May) in 2012.

"To what extent are nodal, dyadic or triadic network effects relevant to predicting future emails?"

[4] ben Aaron et al. (2017)

Results: Exploratory Analysis on *b*

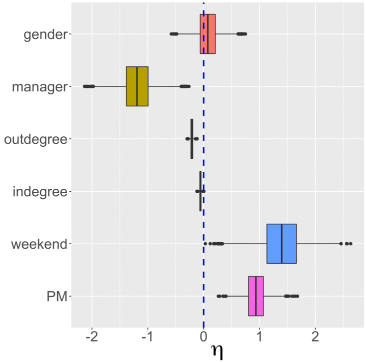
$$\text{logit}(\lambda_{iej}) = \log\left(\frac{\lambda_{iej}}{1-\lambda_{iej}}\right) = b_1 + b_2 x_{iej2} \ldots + b_{14} x_{iej14},$$



- Log odds is two times less if the sender is a woman.
- If i sent n number of emails to j last week, then i is $e^{0.27n} \approx (1.32)^n$ times more likely to send an email to j.

Results: Exploratory Analysis on $\boldsymbol{\eta}$

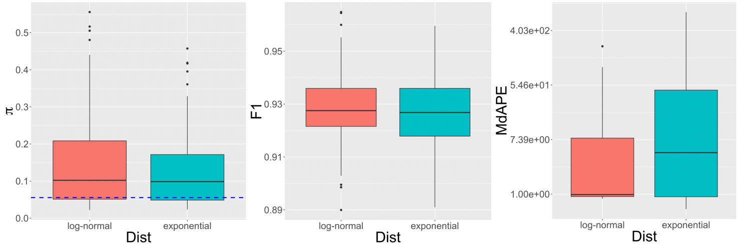
$$\log(\tau_{ie}) \sim N(\mu_{ie}, \sigma_\tau^2), \text{ with } \mu_{ie} = \eta_1 + \eta_2 y_{ie2} \ldots + \eta_7 y_{ie7}.$$



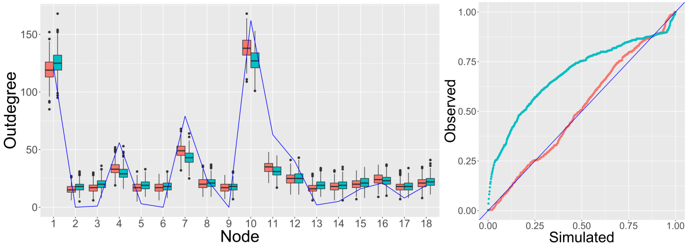
- If an email was sent during weekend or PM, then time to next email takes $e^{1.55} \approx 4.72$ and $e^{0.98} \approx 2.67$ hours longer.
- manager, outdegree, and indegree shorten time to next email.

Comparison: Lognormal vs. Exponential

- Out-of-sample predictions: sender, receiver, and timestamp



- Posterior predictive checks (PPC)



Conclusion and Discussion

- Account for hyperedges without duplications
- Flexible choice of continuous-time distribution via GLM
- Reverse the process for multiple senders to one receiver (e.g., international sanctions and co-sponsorship of bills)
- Sources: http://arxiv.org/abs/1807.08225 https://github.com/desmarais-lab/MulticastNetwork