Review for Bayesian Analysis

Title: The Hyperedge Event Model
Manuscript Number: BA1806-038RA0

The authors propose a continuous–time dynamic hyperedge event model which explicitly allows each sender to connect simultaneously with multiple receivers—and viceversa—with these relational events being monitored in continuous time. The probabilistic characterization of the candidates receivers process and the events' waiting times is developed with a particular attention to devise interpretable representations from a regression perspective. The authors provide also methodologies for posterior computation within a Bayesian inference framework, and strategies to evaluate the model plausibility. An application to temporal e-mail data highlights the benefits of the proposed modeling strategy.

I found the model interesting and potentially a relevant contribution to a somewhat less explored area in the framework of dynamic network inference. Indeed, most of the focus in this area is typically on modeling snapshots of the entire network collected at multiple times, instead of defining a stochastic process for relational events where each edge has its own time index. This is, therefore, a timely contribution, which further allows each sender to connect simultaneously with multiple receivers—a common feature of several networks. Despite this positive overview, I also think the contribution still requires major work, as seen in my comments below.

Comments

1. **PRESENTATION AND WRITING**: The paper requires major work in writing and presentation. I found many typos, sentences which require rephrasing or clarification, and poor bibliography. Below there are some examples, but many others can be found throughout the paper. The authors should pay more attention to these aspects.

   - TYPOS:
     receviers (page 2, line 6), cosponsorhip (page 2, line 9), t-test should be t-test (e.g. page 9, line 14), p-value should be p-value (e.g. page 9, line 14), fit should be fits (page 16, line 8), coavariates (page 17, line 3 from the end), help should be helps (page 20, line 13), etc ...
   - UNCLEAR PARTS:
     - Page 4, line 5 from the end: you write *features used to model the rate.* $V(\mu)$ ... The full–stop between *rate* and $V(\mu)$ is not clear to me. Please rephrase.
     - Page 6, line 2 from the end: $\mathrm{I}(\cdots)$ should be replaced with the simpler $\mathrm{I}(\cdot)$.
     - Page 9, line 8: *To keep the computational burden of re-running thousands of rounds of inference manageable.* The term *manageable* which apparently refers to *computational burden* is quite far away in the sentence and this makes it hard to read. Please rephrase.

  - Page 13, line 9 from the end: From what I have understood $N = 500$ is the number of times you repeat your predictive assessments (holding out each time a random subset of the data). Is this correct? If so, I would clarify it. I had to struggle a bit to understand the meaning of N.
  - Page 15, line 2: you write *missing timestamp bot be the median*. What do you mean by *bot* here? Is it a typo? If not, please clarify it.
  - Page 16, line 3: you write *outdegree distribution—the number of emails sent by each node, indegree distribution—the number of emails received by each node, receiver size distribution—the number of receivers on each emails, and a probability–probability (P–P) plot for time increments.* This sentence is quite hard to read. You use "—" many times and not always correctly. Please use it more parsimoniously and in a better way.
  - Page 18, line 2: you write *i sent n number of emails* and use similar words in other parts. I would simply say (here and elsewhere) *i sent n emails*.
  - Page 18, line 5: you write *a sender sends* and use similar words in other parts. I would try to avoid these repetitions (here and elsewhere).
  - Page 18, line 11: you write *those who have received a lot of emails a lot recently are likely to continue receiving a lot of emails.* This statement needs a careful rephrasing.
  - In some central equations (e.g. page 17 and 19) you write $+x \ldots + y$. I think you should write instead $+x + \cdots + y$.

- BIBLIOGRAPHY: It seems that the authors did not re–read the bibliography at all (but this is part of a paper as well). I could find 4 different issues in the first 5 references:
  - The book title *Numerical issues in statistical computing for the social scientist* should be capitalized.
  - ben Aaron should be ben–Aaron.
  - The title A RELATIONAL EVENT FRAMEWORK FOR SOCIAL ACTION should be lower-case.
  - bernoulli should be Bernoulli.

  There are MANY other issues in the rest of the bibliography. Please carefully correct them during the revision.

2. **LITERATURE REVIEW:** As already mentioned, the authors focus on the somewhat less explored area of continuous–time relational event models where each edge has its own different time index—instead of considering time–varying models for snapshots of networks collected on a pre–specified time grid. However, the contribution is still within the general class of dynamic network inference. In this respect, the literature review provides a poor picture for the state-of-the-art in this wider framework. I think the authors should provide a more comprehensive literature review including also temporal ERGMs and dynamic latent variable models (e.g. dynamic stochastic block models, dynamic mixed membership stochastic block models, dynamic latent space

models, . . .). Discussing your contribution in the light of these alternative (and quite different) models would further clarify the key novelties our the proposed methods.

3. **SECTION 2:** I fully understood the first paragraph in page 3 (summarizing HEMs) after reading the subsections 2.1, 2.2 and 2.3. This part should provide a much clear picture of your model instead of creating confusion. I suggest to improve it, leveraging also some intuitive illustrative figure. For example you could place Figure 1 much early and comment it while summarizing the HEM at the beginning of Section 2.

4. **PRIOR SPECIFICATION:** There are two important points to be clarified here.

   - You present the full conditionals in page 7 assuming uninformative $N(0, \infty)$ priors, but then you rely on weakly informative priors in the application (see page 17). I found this confusing. I'd present results in page 7 for generic Gaussian priors.
   - Given that the model is relatively complex, some sensitivity analyses should be carried out to check how much posterior inference is affected by the hyperparameters' settings, and, possibly, suggest some default values.

5. **POSTERIOR COMPUTATION:** The authors themselves claim that the model is relatively complex. In this respect, I appreciated section 3.2, but I was expecting also some more computational details. For instance:

   - There are many MH routines in the literature. Which type of MH do you consider? What is the proposal distribution? What about the acceptance rate? Is there any smart proposal in this case which helps in increasing the acceptance rate.
   - You rely on data augmentation MCMC, which has been shown to mix quite poorly (also in recent theoretical papers). Indeed, as expected, you end up thinning the chains every 40 samples in the application. However, there is no comment on this in the paper. I think it should be highlighted and not just hidden in the thinning.
   - Your quantitative assessments are based on 5 nodes in the simulation and 18 nodes in the application. These are quite small networks compared to those one would expect in real–world settings (such as those you list in the introduction). To what extent are your computational methods able to scale to much larger networks? An application to a bigger dataset would be useful. Moreover, more information on computational time should be provided.

6. **APPLICATION:** I have several remarks regarding the application.

   - You compare performance in predicting missing senders with random guess 1/18. This is a quite naive competitor and I am sure the authors can find much better ones—still relatively simple.
   - Related to the above point: since the proposed model seems to provide advances compared to SAOMs and their extensions, I think that the authors should include some comparisons against such formulations. It seems that this can be done at least in the predictive performance and posterior predictive checks part. This would provide a more convincing argument for the methods proposed here.

- In Figure 4c the choice of the logarithmic scale for MdAPE seems to hide a quite poor performance of your model in predicting the timestamps. It is true that the log–normal improves over the exponential, but the log–normal boxplot has still a third quartile providing an MdAPE of exp(7.36) which looks quite big. This is to me a negative result, which should be discussed and addressed in the paper.
- There seem to be two key issues in the interpretations in lines 11–12 of page 19.
    - Based on your comments in the paper you are interpreting the posterior expectations of $\exp(\eta_6)$ and $\exp(\eta_7)$—i.e. $\mathrm{E}[\exp(\eta_6) \mid \text{data}]$ and $\mathrm{E}[\exp(\eta_7) \mid \text{data}]$. However you seem to compute this as $\exp[\mathrm{E}(\exp(\eta_6) \mid \text{data})] = \exp(1.552)$ and $\exp[\mathrm{E}(\exp(\eta_7) \mid \text{data})] = \exp(0.980)$. By Jensen inequality this is wrong.
    - It is also not correct to claim that the e^{th} email is expected to take $\mathrm{E}[\exp(\eta) \mid \text{data}]$ *longer* compared to their counterpart. The term *longer* seems to refer to an additive effect on the event timing, which is not the case here since you assume a log–normal for the event timing and hence the effect is on the scale of the timing and not on the location—note that, if $X \sim \text{log–normal}(\mu, \sigma^2)$, then $\mathrm{E}(X) = \exp(\mu + \sigma^2/2)$. The authors should be more careful in the interpretations on the original time scale, or, more conservatively, they could simply comment on the effects on the log-time (as done immediately after).