Review of “The hyperedge event model"

August 9, 2018

This paper proposes a novel Bayesian approach for dynamic multicast interaction data, or more generally dynamic hyperedges interaction data. The authors derive a Markov chain Monte Carlo algorithm for posterior inference, and demonstrate the usefulness of the approach on email interaction data.

There has been relatively few papers on the modeling of multicast interaction data, and this paper offers a useful and timely contribution. While the overall approach looks sensible, I have several concerns about the article that would require work going beyond a major revision.

- Presentation: I found the presentation rather lacking; in particular, the description of the model in Section 2 is very difficult to follow, due to the numerous variables introduced, the order in which the different elements of the model are introduced, and the lack of definition for some of them. I think a major rewriting is needed here. More precisely:
  - Some notations are rather unusual, for example A for nodes. Although this is not indicated in the text, I assume this comes from the term "actor" in the stochastic actor-oriented model (SAOM) of Snijders, cited by the author? If this is the case, it would be worth mentioning it so that the reader recalls this later on. y for the covariates is also rather unsual. Some notations are inconsistent. For example u_{ie} denotes the first line of the matrix u_e, but $\tau_e = \min_i(\tau_{ie})$.
  - In section 2.1, I think it makes more sense to first introduce Equation (2.2) then (2.1). The term ‘intensity' used for λ_{iej} is rather confusing. As the paper is concerned with a continuous-time model, one would expect that the term intensity refers to some point process, which is not the case here.
  - It is difficult to understand Section 2.2 without reading section 2.3. The authors introduce the notation τ_{ie} above Equation (2.5), but do not explain what this represents (at this stage I thought it was a component of the vector τ_e). I do not see what the variable μ represents in Equation (2.5), and if it is related to μ_{ie} in Equation (2.4)

  - Some sentences are difficult to understand:
    Page 3: "we define a probability measure “MBG” motivated by the Gibbs measure"

- Discussion of relevant literature: The authors should provide a better discussion of how their model differs from other approaches, in particular the approach of Perry and Wolfe (2013), and Snijders (1996), based on point processes.

  - Perry and Wolfe propose a specific model for multicast interaction (Equation (6) in the arxiv version of their paper). What are the differences/advantages between both constructions?
  - In Section 2.3, the author cite Snijders (1996) when they introduce the model for the interaction arrival times. I am not sure what is meant here: is this part of the model already introduced in Snijders (1996)? The authors should be more specific here.

- Model, covariates and missing data: This is not explicitly mentioned in the definition of the model in Section 2, but the covariates x_e and y_e depend on the observed interaction data $(s_{e'}, r_{e'}, t_{e'})$ in the last 7 days, as written in Section 4.1. Hence the observations $(s_e, r_e, t_e)_{e=1,\dots,E}$ are not conditionally independent given the model parameters. This should be clearly stated in Section 2. For this reason, I do not think that the out-of-sample algorithm described in Algorithm 3 is correct. The conditional distribution for the missing observations s_e,r_e and t_e should depend on $(s_{e'}, r_{e'}, t_{e'})_{e':t_e<t_{e'}<t_e+\ell_e}$, which is not what is done.

- MCMC sampler.

  - The authors should provide more details on the Metropolis-Hastings proposals for the parameters b and η in Section 3.2.
  - The sampler uses a blocked Gibbs sampler for the u_{iej}. Given the constraint that $\sum_j u_{iej} > 0$, this sampler may be rather inefficient in the case where there are many one-to-one interactions (as is the case in the application, where 83% of the interactions are dyadic). In order to go from the state $u_{ie} = (1,0,0,0,0,0)$ to $u_{ie} = (0,1,0,0,0,0)$ one needs to go through a state where two receivers are activated, which has low probability in this case. It would be good to comment on this, and in general on the mixing of the MCMC sampler.
  - The authors provide in Section 3.2. some sanity checks on the sampler. While this is good practice to perform such checks, I am not sure it is very useful to include this in the main body (could be moved to the appendix), as this is done on a very small scale example with 5 nodes and 100 events, and does not really give an indication on the convergence properties of the algorithm in a more realistic scenario.

    I suggest the authors perform a simulation study with a larger number of nodes and events to demonstrate that the algorithm is able to approximate the posterior distribution well in that case.

  - The authors should provide some indication of the computational complexity per iterations of their sampler. The application to email interaction data is rather small (18 nodes and 680 emails), so it would be good to know how many nodes/events the proposed approach can handle.

- Typos: The article contains numerous typos. Here are some of them:

  - Page 4, section 2.1: "$u_{ie} \in [0,1]^{A}$” should be $\{0,1\}^{A}$
  - Page 4: "sender-specfic mean"
  - Page 4: "statistify"
  - Page 4: "the features used to model the rate. $V(\mu)$ - e.g. the rate..." should be rephrased
  - Page 6: "hyperparamters"
  - Page 15: "missing timestamp bot be the median", rephrase