Referee report for manuscript EJS1807 titled "A Preferential Attachment Model for the Stellar Initial Mass Function"

The manuscript "A Preferential Attachment Model for the Stellar Initial Mass Function" introduces a generative simulation model (depending on three parameters $(\alpha, \gamma, \lambda)$) for the stellar initial mass function science. ABC is used to obtain approximate posterior samples of these three parameters on simulations that incorporate complicated observational effects which are difficult to include without the ABC technology.

In general I like this manuscript and think EJS is a good fit for publication. There appears to be two main contributions. The first is the introduction of the preferential attachment model (apparently a Yule-Simon stochastic model) as a more direct approach—compared with parametric stellar initial mass functions—for modeling the distribution of star masses in a cluster. It seems as though this could be a useful tool for astronomers to have in their modeling tool box. I would also predict the generative procedure is general enough for astronomers to be able to add tweaks to the stochastic procedure which could incorporate other physical effects. The second contribution is the exposition of how ABC can be used with this generative model, along with complicated observational effects.

As for revisions, I think the paper needs to be tighten up a bit. The writing can be somewhat wordy at times and there seems to be a couple superfluous sections that could probably be dropped. Also, there are 17 figures, some with multiple panels and a lot of white space. I would encourage the authors to work on filtering out the diagrams which do not significantly contribute to the main points of the paper or help the reader understand the exposition. For example, Figure 1 is mentioned in one sentence on page 5 and doesn't seem to give the reader much more than isn't already written in the proceeding text. Another example comes in Figure 3 where the x-axis scale is chosen so wide that it's hard to discern multiple densities. Why have all that white space? Some of this boils down to reader preference, and of course the authors can't please everyone, but I think some editing iteration could really improve the paper. I've included other comments below (in no particular order). Again, many correspond to personal preferences which I think would improve the paper. I'll leave it up to the editors to evaluate what of these comments are important for publication.

Further comments

1. Only after I read the Yule-Simon simulation procedure did I actually get a clear picture what the stellar initial mass function is: a continuous density describing the histogram of the list of star masses in a cluster. The exposition preceding somehow never quite gets to the point that an IMF

isn't necessarily a physical model describing how the star masses are formed in the dynamic evolution of the cluster, but rather a non-physical summary of such a list of star masses. Also that the PA model avoids specification of a parametric IMF using a simple non-physical simulation procedure to create such a list. Perhaps this should be emphasized someplace in the first couple paragraphs of the paper.

2. In (3.1) why write $\pi_t = \min(1, \alpha)$ then restrict $\alpha \in [0, 1)$ in the next sentence? Why write π_t depending on t? Why even have a new symbol for what amounts to α anyway?

3. Many of the estimated posterior densities which summarize the ABC output look to have small scale local fluctuations. I'm guessing these bumps and wiggles are just artifacts from the finite number of posterior samples, but I think it runs the risk of suggesting to the reader that ABC is less accurate than it actually is. It would really strengthen the authors main points if the plots didn't have as much visual finite sample fluctuation. Perhaps step-line histograms with wide enough bins would give a sufficient visual description of the posterior samples and also suggest to the reader that the actual samples may be very accurate but not necessarily the histogram derived from it.

   On a related note, I wonder if Rao-Blackwellisation can be used here for reducing the sample variability in the posterior density estimates from ABC. The final posterior samples $(\theta_1, \gamma_1, \alpha_1), \ldots, (\theta_n, \gamma_n, \alpha_n)$ carry along with them simulated auxiliary variables used in the process of generating a data sample that can be used to, effectively, make a better kernel smoother estimate of the marginal density. In particular, let N_i denote the number of stars generated by the particular data generating process associated with the ABC sample $(\theta_i, \gamma_i, \alpha_i)$. Notice that $N_i \sim \text{Poi}(M_{tot}/\lambda_i)$, or at least if your using M_{tot} to denote the upper limit for the mass of the system. Now instead of plotting a histogram or kernel density estimate based on the approximate posterior samples $\lambda_1^{-1}, \ldots, \lambda_n^{-1}$, the rao-blackwell density estimate would be

$$\lambda^{-1} \mapsto P(\lambda^{-1}|\text{data}) \approx \frac{1}{n} \sum_{i=1}^{n} P(\lambda^{-1}|N_i)$$

   where $P(\lambda^{-1}|N_i)$ is easy to compute using the Poisson likelihood and the prior π. It is not entirely clear to me that this will be easy for the marginal posterior density of the other parameters. If it is easy and it makes the plots in this manuscript less variable, it might be worth while adding it to the paper. However, I think it would be certainty sufficient to instead simply clean up the plots by replacing the kernel density estimates with an appropriately binned histogram.

4. Section 3.1.1 seems like a bit of an afterthought. Not really sure I follow

what I should get out of it. Is this probing the flexibility of the PA model or the ability of the data to constrain γ. If the main point of this section is something of direct relevance to astronomers, then I would suggest being a bit more clear and to the point what the authors are trying to get accross. Otherwise, I would suggest dropping it altogether.

5. In the beginning of Section 3.2 the authors write: "*The PA model describes the formation of a star cluster at initial formation. However, we are not generally able to observe the star cluster after initial formation due to observational uncertainties, measurement uncertainties, and aging and dynamical evolution of the cluster.* "

   This statement confuses me a bit. It seems to suggest that we only observe the *initial* formation of star clusters. I would expect the opposite, i.e. that the star clusters we observe are a mix of old and new clusters that have developed over different time ranges. Perhaps the authors are getting at a selection effect where the old star clusters are more dim and are effectively censored due to the sensitivity of our instruments. Either way, this paragraph could use some cleaning up.

6. In display eqn (3.4) the letter 'm' appearing on the left hand side should be italic.

7. I think Section 4.1, which describes the ABC sampling algorithm, can go into an appendix. In fact, I would remove Section 4 (titled "Methods") altogether, moving Section 4.1 to an appendix and Section 4.2 to the next section on simulations. This will allow the authors to have all the simulations in one place and can sharpen their main points and conclusions.

   Regarding the three simulations presented in the paper (currently given in Sections 4.2, 4.2.1 and 5), do the authors really need a preliminary simulation study, given before 4.2.1, which doesn't include observational effects? Why not just include two sets of simulations: (1) using the PA model to generate the data which includes observation effects; (2) using the astrophysical simulation to generate the data (also including observational effects). Reducing the length of the exposition, number of plots to examine and focusing on the main conclusions seems like it would greatly improve the manuscript.