We thank the reviewers for their thoughtful feedback that shows they understood the key points in our paper. We are glad that they found our contributions to be timely (R1, R2), relevant (R1, R4), novel (R4), and our paper to be well-written (R1, R2, R4). We are particularly encouraged that they found both the theoretical rigor (R1, R3) and empirical results (R2, R3) to be strong. An area of concern relates to the number of communities k (R1, R2, R4)—we first address this concern then respond to some other reviewer comments below.

@R1 - The estimation of k is not handled; @R4 - Are there better approaches to find the optimal k? We thank the reviewers for pointing this out. Our estimator uses spectral clustering on the weighted adjacency matrix N so model selection approaches for static block models can be used. We used the eigengap heuristic for the exploratory analysis in Section 5.4 and in C.2.3 and C.2.4 of the supplementary, but more sophisticated methods including using eigenvalues of the non-backtracking matrix and Bethe hessian matrix (Le & Levina, 2015), and network cross validation (Chen & Lei, 2018; Li, Levina, & Zhu, 2020) could be used. Another approach mentioned by R4 specific to timestamped networks, is to *hold out a portion of the events and select the k that maximizes test log-likelihood*, which we used in Table 1. As shown in Figure 1, for $k < 100$, there is hardly any increase in the runtime, and it is manageable even for $k = 1,000$. We would add this discussion to the paper.

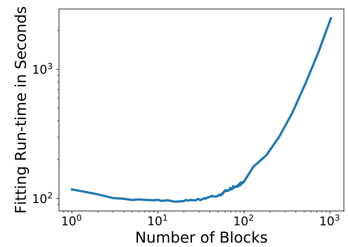

Figure 1: CHIP's fitting runtime on the Facebook data on a log-log scale with increasing k.

@R2, R4 - The number of communities k and community assignments are fixed over time which prevents the model to be used for dynamic network analysis: Both k and community assignments are indeed fixed in the CHIP model and in most other continuous-time block models [1, 3-5, 7, 8]. This is a current limitation of continuous-time block models compared to discrete-time models that often allow changes in communities over time [9-13]. However, we disagree that this prevents the model from being used for dynamic network analysis because the temporal dynamics are being captured by the Hawkes processes. *Thus, the CHIP model still captures time-varying behavior due to their self-exciting nature despite the fixed communities.* Since the paper submission, we became aware of the continuous-time block model of Corneli, Latouche, & Rossi (2018) that divides time into D equally-spaced change points where community structure can change. Such an approach could be used also with the CHIP model.

@R1 - **No detail on the likelihood estimation scheme proposed for α and β and their theoretical properties:** The estimation procedure is discussed in detail in Section A.3 in the supplementary. We have no guarantees for α and β but demonstrate in Section 5.2 through simulation that the MSEs of their estimators with decrease quadratically with n.

@R1 - **The paper concentrates on dense graphs. The dependence of parameter μ of Hawkes process on the node size n is not discussed in detail:** We provide results for the sparse regime in Section B.1.1 in the supplementary. We let $\mu \asymp \frac{1}{f(n)g(T)}$, a function of n and T and explore various sparsity settings by varying f and g. Our proofs allow μ to vary with n and T and can be as small as $\log(n)/(nT)$, as **R1** suggested. In particular, in the last paragraph we wrote, "if we set $g(T) \asymp T$ and $f(n) = \frac{n}{\log n}$, such that $\mu_1 \asymp \mu_2 \asymp \frac{\log n}{nT}$, then the expected number of events between a node pair is $O(\frac{\log n}{n})$. In that case, $r(T) \lesssim \frac{k^2}{\log n (c_1 - c_2)^2}$, and consistent community detection is possible as long as $k = o(\sqrt{\log n}|c_1 - c_2|)$." We will add a discussion on the sparse graph setting and a reference to the supplementary.

@R2 - Why does finding only 1 or 2 clusters suggests independence? In CHIP, a small number of communities (e.g. 1 in the case of Reality Mining data) suggests a weak community structure, but not necessarily independence. That conclusion was mostly derived from the fact that BHM (which models dependence of node pairs within block pairs) achieves its best test log-likelihood on the same dataset for extremely large $k = 50$ on a network with only 70 nodes!

@R3 - Is there a stronger case made for the utility of a good predictive model (in CHIP)? We thank R3 for this suggestion. Two potential use cases are for time-to-event prediction, i.e. the time until the next event between a pair of nodes, and predicting the number of events between a pair of nodes in a future time period.

@R3 - Why can't BHM turn into CHIP by a simple modification? Why such a high difference in log-likelihood even when $k = 1$? The BHM uses a single Hawkes process for each block pair then randomly assigns events to node pairs so that the dependence between node pairs cannot be relaxed. On the other hand, CHIP assumes independent node pairs in a block pair that share the same parameters. The closest the BHM can get to CHIP is for $k = 1$, where the BHM shares parameters but has dependence, and for $k = n$, the BHM has independence but no parameter sharing.

@R4 - What is the motivation for the simplified estimation procedure that ignores timestamps? The main advantage of ignoring timestamps is scalability—our estimators for the μ and m parameters scale independent of the number of events (beyond the trivial computation of the count matrix N), while the standard MLE using the timestamps (e.g. in the BHM) requires solving a non-convex optimization problem that depends on the number of events.

We especially thank R4 for the very detailed comments and will incorporate them despite lack of space to respond here.