We thank all reviewers for carefully reading our manuscript and supplement. We also want to thank all referees for pointing out some minor issues, which we will not address individually in this rebuttal, but which will be implemented in the next version of the manuscript.

To reviewer 1 We thank the reviewer for the positive assessment. The purpose of the BHPS experiment is to demonstrate better scalability of our approach compared to previous work as, e.g., in [4], [11] or [16]. With the previous methods, it would not be possible to infer a network of 15 subnodes (as we did here). However, inferring CTBNs from incomplete data inherently requires substantial computational efforts, as each trajectory has to be processed individually. This is the reason why we down-sampled the BHPS data (as was also done in [16], but for far fewer subnodes). We will add a more detailed experiment in the next version of the manuscript. Here, we will vary the number of trajectories to investigate the stability of our prediction. For a discussion about the similarity of our approach with [4] we refer to [11], where the weak-coupling approximation (for non-mixture CTBNs) was derived. The reviewer is correct that a more scalable structure learning method than in [16] can be constructed in combination with [4] (or [11]). Structure estimation in CTBNs from incomplete data has two distinct bottlenecks: 1.) Latent state estimation (scales exponential in the number of nodes). 2.) Iteration over candidate structures (scales super-exponential in the number of nodes). While 1.) has been overcome in [4] and [11] to different orders of accuracy, here we are the first to touch on bottleneck 2.) by introducing mixtures of generators. Our method is inherently more scalable than SEM ([4] or [11]) due to the additional approximation step (Jensen's inequality in line 103-106). We show improved scalability in figure c), where we compare the run-time of the greedy version of our method with greedy hill-climbing (maximum number of parents is two) for SEM+variational inference [11]. We did account for the $\sqrt{2\pi/z}$ term. However, an approximation, which holds for $\bar{\alpha} \gg 1$, is omitted in the current version.

To reviewer 2 We thank the reviewer for the positive assessment. For a discussion about the likelihood function, we would like to refer to [15] where the likelihood for a CTBN is derived. We would like to point out here that $\bar{\alpha} \gg 1$ is a quite natural assumption, even for situations where data is sparse. Here, a Bayesian perspective implies that sufficiently strong prior information has to be leveraged to regularize the problem and allow for inference. We exactly do this when assuming $\bar{\alpha} = M + \alpha \gg 1$ by incorporating a sufficiently strong prior α on the number of recorded transitions M. While inference in a learned CTBN is not our main focus in this manuscript, we employ smoothing under a mixture CTBN by solving eq. (10) and (11), to calculate the expected sufficient statistics. This algorithm can also be applied to the learned model, to perform inference. Based on the suggestions of the reviewer we ran an experiment on a minimal CTBN example (2 nodes with a bidirectional coupling) in figure $a)$ given below. We plotted the normalized lower bound of the marginal posterior $\mathcal{F}_i$ (eq. 5) for the node $i = 1$ (color-coded) vs the mixture probability $\pi_1[1]$ (x-axis) for the node of having one parent (left, $\pi_1[1] = 0$) and having no parent (right, $\pi_1[1] = 1$) and the dashed sample mean to indicate the trend. The y-axis denotes the number of trajectories. While for a large amount of trajectories the mass is allocated at the ground-truth $\pi_1^*[1] = 0$ for all concentration parameters c in the Dirichlet prior, for a small number of trajectories c can either force selection ($c = 2$) or force a mixture through the convexity of the profile ($c = 0$).

To reviewer 3 We thank the reviewer for the detailed review of our work. Based on his suggestions, we will improve the readability in the next version of our manuscript. The expression for τ_i can be found in the main text in line 164. The derivation was shifted to the supplement due to length constraints. To avoid confusion, we will feature this expression as an equation in the next version. The solutions of the ODEs (10,11) do depend on the data Y, as it is explained in lines 165-168 in the main text. Here, the observations are explained to be implemented via jump conditions on the Lagrange multipliers ρ_i (line 168). This is in line with existing works, e.g., [4], [11] and [17] (in [17] it is derived in the main text). As mentioned in the main text, we solve the set of ODEs in the same way as in these works. The equations to be referenced in Alg. 1 are indeed (10), (11) and the jump condition on ρ_i (line 168), thank you for pointing this out. We refer to [5] in line 127 as a variational approach. The authors of [5] themselves refer to their method as a variational approach, e.g., in their abstract. We did not report ROC scores for the BHPS, as no ground-truth is available (because the true dependencies are unknown). We ran this experiment only to demonstrate scalability. In case the reviewer is referring to Fig. 2 a) and b), these subfigures refer to a separate experiment on synthetic data in a larger system using greedy search, as explained in the main text. Based on the suggestion of the reviewer, we investigated the accuracy of the lower bound of the marginal posterior (see eq. 5) against the true marginal posterior (calculated via numerical integration) for mixtures of different entropies, different amounts of trajectories for $c = 1$, see figure $b)$ below. This experiment is performed on the graph ensemble as in the first synthetic experiment of our main paper with rates drawn from a Gamma distribution ($\alpha = 5$ and $\beta = 10$). For scalability, we refer to our response to reviewer 1.