We thank the reviewers for their useful feedback. Please find below responses to the comments by each reviewer. Also kindly note that the new results referred to in the following will be added to the final version of the paper.

R1 and R3: We empirically varied both α and γ in the range $\{10^{-3}, 10^{-2}, 0.1, 1\}$ for LDA for all data sets. We found that $\gamma \leq 0.01$ (inducing sparsity) is preferred for topics (larger values produced useless results despite different α). Although too small γ increases computational complexity and the risk of getting stuck in a locally optimal mode. For a sparse topic prior, increasing α (i) decreases topic entropies (inferring sparser topics), (ii) coherences improve marginally or remain the same and (iii) **both recall and precision decrease**. Recall and precision curves for different γ as a function of α have similar form peaking at $\alpha = \{0.01, 0.1\}$ and $\gamma = \{10^{-3}, 10^{-2}\}$, verifying that the adopted setting for LDA is competitive. Despite the tuning, the precisions and coherences are worse than for our model. As suggested by the reviewer 3, we repeated the experiment for an asymmetric topic prior, matching the prior strength to equal the strength of the symmetric variant for varying γ. The results for the asymmetric prior are very similar to the symmetric prior showing that the asymmetric prior is ineffective to boost precision. To conclude, tuning of the LDA hyperparameters is ineffective to trade-off recall and precision. Intuitively, prior tuning is unable to overcome problems of the likelihood function; here the sensitivity of the $\mathrm{KL}(\mathbf{p}, \mathbf{q})$ to misses. We fix this issue by modifying directly the likelihood function.

R1: We emphasise that we are essentially clustering the documents based on the inferred topics and using ARI to compute similarity between inferred and true clusters. As opposed to classification we do not assume the number of clusters to be known; the number of potential clusters is constrained by the number of topics. We argue the clustering scenario is more interesting than the classification set-up, which additionally induces classification algorithm bias.

Predictive accuracy index (PAI) penalises the values by the predicted area size, giving large values for crime hotspots using the smallest area. Predictive efficiency index (PEI) computes a ratio between the number of crimes occurred in the predicted hotspots and the maximum number of crimes that could have occurred in same area size. In general, PAI and PEI may be interpreted as generalisations of precision and recall, correspondingly, for spatial crime hotspot prediction.

Given the ease of implementation and insensitivity of setting λ for obtaining better performance than LDA, which is found in nearly every data scientists' toolbox, we argue our new topic model would serve as a new standard for topic modelling.

R2: We found that the results for SWB are marginally worse or similar to SW, suggesting that including a common background term is not effective for improving performance, as suggested by the reviewer. The key difference between SW(B) and our model is specification of the background distributions and λ. We point out that our model performs better than SW(B) model that has no connection to K-divergence and is not able to trade-off recall and precision.

We agree that it is in theory straightforward to place an informative prior for λ instead of fixing a value for it but tuning of the corresponding hyperparameters is less straightforward in practice; in addition to mean also the strength of the prior needs to be specified. This tuning can be expensive and is further data-set dependent.

The conclusions drawn from Table 2 hold exactly for $\lambda \in (0.07, 0.11)$, further showing that obtaining good results is not sensitive for particular λ. For larger λ the coherences for larger thresholds (T) may become less meaningful because of increased topic sparsity, noting that coherence computation requires ordering the top-T words. Focusing for smaller thresholds (here for $T = 5$), the conclusions hold for $\lambda \in (0.07, 0.14)$.

R3: We adopted fixed symmetric priors for computational reasons, for permitting a fair comparison to SW(B) models and for experimenting the effect of different hyperparameters for LDA (see above).

We use the uniformity assumption to demonstrate the connection between KL-divergences to standard recall and precision, that only work for binary relevances; a term is either relevant or not. KL-divergences are suitable for graded relevances and do not assume uniformity; here the $\log$-function acts as a barrier function giving a large penalty for $p \log(q)$ for non-zero p and small q (corresponding to misses).

Often hand-selected topics, that make sense, are shown in research papers avoiding the issue that top topics according to cardinality may not be meaningful.