In this appendix, we consider the application of LDA to the dataset used in the case study of the 2016 Charlotte protests for our work. Our primary analysis focuses on the probabilistic requirements for LDA, and how they force complexity on the relationship between users, hashtags, and communities that problematizes in-depth qualitative interpretations.

Recall, our goal is to co-cluster users and hashtags into substantively meaningful clusters that can facilitate further, in-depth qualitative research. Within the context of mixture models, like LDA, researchers typically deal with the complexities of the combination of user mixtures and hashtag mixtures by setting thresholds. More specifically, researchers define topics as the top K hashtags most likely to be assigned to each topic, and then define which topic(s) a user belongs to by identifying the most likely topic(s) for that user. In this way, we can identify communities by determining the set of users aligned on the same topic, and determine the ideological leaning of that community by the hashtags representing that topic.

A long line of work shows how such ad hoc decisions can be biased, even when formal attempts are made to determine the quality of various thresholding decisions (Morstatter & Liu, 2017), and, as we demonstrate later, thresholding introduces conceptual problems with soundly interpreting the resulting clusters. If we want to study communities of people and hashtags, we would ideally leverage an algorithm that makes these decisions for us. By modeling the duality of users and hashtags, this is exactly what bi-spectral clustering does. However, bi-spectral clustering additionally defines a mathematical model in which communities of users and hashtags are explicitly identified and relates it formally to the problem of partitioning a bipartite graph. Consequently, the communities resulting from bi-spectral clustering are well-defined from a mathematical perspective and, more importantly, consistent from a socio-theoretic perspective.

Comparison with LDA Clustering of #Charlotte Data

For qualitative and quantitative comparison, we also trained an LDA model over the #Charlotte user-hashtag data. Unlike bi-spectral clustering, LDA is a probabilistic model, meaning that each run of an LDA model will produce different results. Using the Python package Gensim (https://radimrehurek.com/gensim/), we ran 100 LDA topic models of 100 topics each. We ran each model for 1000 iterations, and specified 100 topics in order to hold all things equal other than the different algorithms themselves, and to give LDA the opportunity to be as granular as the bi-spectral clustering. From the 100 LDA topic models, we chose the model with the lowest perplexity as our final model. Although it is well known that perplexity does not correlate with human assessment of topic quality (Chang et al., 2009), it is the most commonly used metric for LDA model selection because it measures the "unpredictability" of text as described by a model (Wallach, Mimno, & McCallum, 2009). The lower the perplexity of a topic model over a corpus, the less unpredictable that corpus is to the topic model.

Within the ad-mixture framework of LDA, communities are characterized by distributions of hashtags and users. From these distributions, we would like to construct coherent, stable clusters of users and hashtags. However, since, technically, every hashtag is in every community and every community is expressed by every user with some probability, clusters of users and hashtags are often determined by setting an ad hoc threshold on the topic-word (community-hashtag) and document-topic (user-community) probabilities. Figure 2 shows how different plausible thresholds (ranging from 10^{-1} to 10^{-5}) affect the number of hashtags and users in every topic cluster.

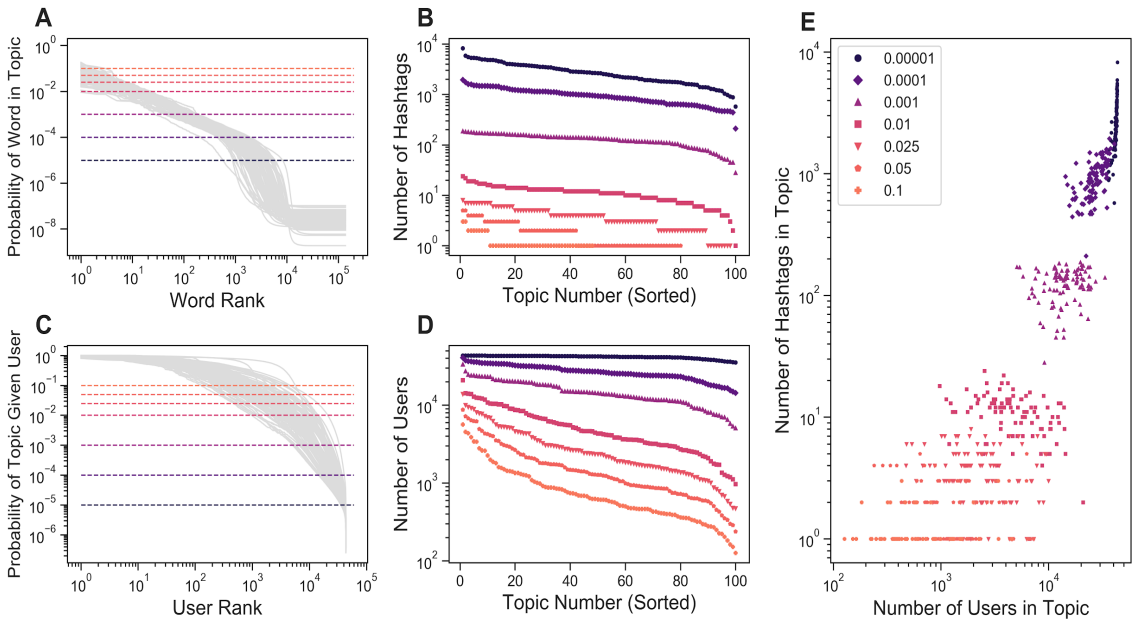

Figure 2. Effects of thresholding on LDA clusters with 100 clusters. Note logarithmic axes. **A)** Gray lines indicate the probability distribution of words within topics ranked by probability. Dashed lines indicate thresholds of 0.1, 0.05, 0.025, 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. **B)** The number of hashtags per topic under each threshold. **C-D)** Distribution of most probable topics for users by rank and the number of users per topic under each threshold. **E)** The number of users and hashtags per topic under each threshold.

From Figure 2A--D, we see that the sizes of user and hashtag topic clusters produced by LDA are sensitive to the threshold that is chosen. For higher thresholds, the user clusters span several orders of granularity with some clusters containing only 100 users and others containing over 10,000 users, similar to bi-spectral clustering. However, for lower thresholds, all clusters begin to concentrate into large sizes of 1,000 to 10,000 users each. The number of hashtags per topic cluster are more constricted in their range, often spanning less than one order of magnitude for any given threshold. For all but the lowest thresholds, the number of hashtags per cluster is

largely independent of the number of users per cluster; the same number of hashtags are used to describe user clusters of vastly different sizes.

This may initially appear to be a benefit of LDA: at higher thresholds the user community clusters span several resolutions but are all succinctly described by a small number of hashtags. However, we cannot be confident that these high probability hashtags accurately summarize the users that express these communities with high probability, because the probability a user expresses a community is a function of *all* of the hashtags. As a consequence, the many hashtags with low probability within each topic can contribute to a user having a high community probability (Schmidt, 2012). This is demonstrated in Figure 3, which shows the pairwise overlaps between hashtag clusters and user clusters as the threshold is varied. Even at low thresholds, there is limited overlap between different hashtag clusters, but the user clusters almost all completely overlap. Since there is minimal overlap between the hashtag clusters, the fact that nearly all users belong to nearly all community clusters can only be explained by the remaining lower probability hashtags. Overall, we find that at high thresholds a few high probability hashtags are not sufficient to describe large corresponding user community clusters, and at lower thresholds the clusters grow and overlap so much that they cannot be properly disentangled. Thus, while LDA is a de facto method in text analysis, it defines clusters of hashtags that are unrepresentative or clusters of users that are unwieldy, limiting its ability to support additional qualitative analyses describing the diversity of communities that coalesce around online social protests.

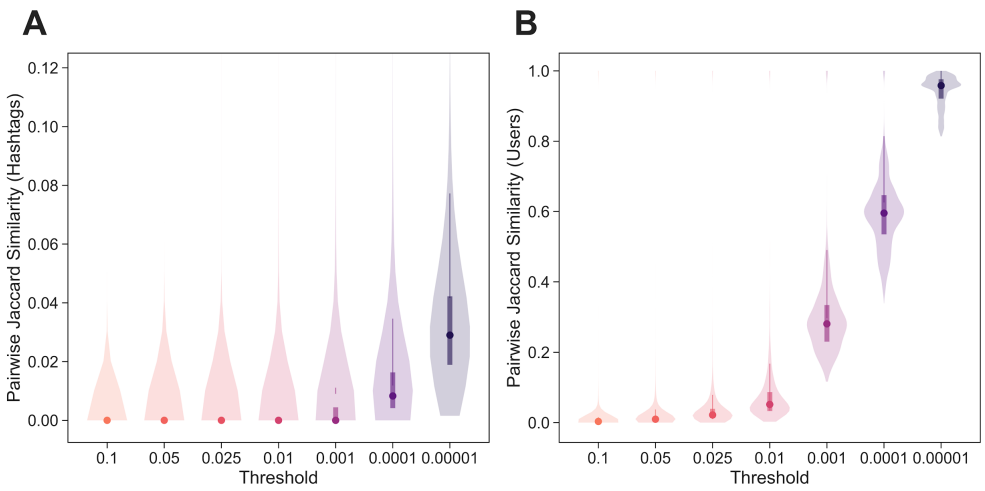

Figure 3. Violin plots with inset box plots of the pairwise Jaccard similarities between each topic cluster of hashtags (**A**) and users (**B**) as the threshold varies. Note the different vertical axis scales. The Jaccard similarity of two hashtag clusters is calculated as the ratio between the number of shared hashtags between the clusters and the number of unique hashtags in total between the clusters, and similarly for user clusters. The dot embedded in each box plot indicates the median similarity.

Alternatively, for the user clusters, one could place each user in the single cluster that is most probable given their hashtag usage. Doing so results in clusters that are of modest sizes of 100-500 users each (median 258, mean 435) with only 25.9% of users falling into the 5 largest clusters, which is half the percentage of users in the largest clusters when using bi-spectral clustering. While this is a less arbitrary approach to constructing user clusters, it still has two disadvantages. First, it walks back the intent of LDA to model each user as a mixture of communities and discards the information that was used to formulate those mixtures. Second, it does not solve the issue of hashtag thresholding and how representative they are of the clusters,

or the related concern that some users may have a high community probability because of the long tail of hashtags outside of the top hashtags.

In sum, while we are not arguing that there is no way to properly interpret the LDA user-hashtag communities, we have demonstrated that, unlike bi-spectral clustering, researchers need to make a number of ad hoc decisions to properly construct hashtag and user clusters. The abundance of these decisions can significantly hamper the process of qualitative content analysis because they introduce a variety of concerns about the representativeness and stability of the underlying clusters. Researchers still must make one such ad hoc decision when using bi-spectral clustering, which is how to determine the most important or representative users and hashtags per community. In the case of LDA, this decision is based on probabilities which, as we have demonstrated, can be misleading. In contrast, for bi-spectral clustering, we can use a considerably easier metric--namely, popularity--because community boundaries are already pre-specified. Thus, bi-spectral clustering changes and, we argue, makes easier the decision of how to identify important hashtags and users relevant to communities identified by the algorithm.