A Supplemental Material

A.1 Data Collection

We collected a significant amount of data from Twitter. We used Snowball sampling to identify Twitter accounts of local citizens. Crawling publicly available tweets from user profiles enables us to collect significantly more amount of data in comparison to collecting streaming real-time tweets of the Columbus area. For the purpose of our study, 63 Twitter accounts that mostly posted news and information about Columbus city were identified and used as the seed users. Many local residents follow such accounts to stay informed about the local events (Kwak et al. 2010). The seed set included the twitter account of several organizations including major universities, recreational centers, medical centers, newspapers, local bloggers, local reporters, police, libraries, restaurants as well as the local sports teams. Using Twitter's streaming API, the followers of the seed accounts were collected. Following this step, we explored user's profiles and identified 54K public profiles that marked their locations as Columbus or one of the suburban areas included in the AHDC study. The AHDC study area included several populous suburbs. Collectively, 50 million publicly available tweets were collected from these accounts. In another wave of data collection, we collected publicly available geotagged tweets for a period of May-August of 2018. This resulted to additional 2.8 million tweets. Next, LNEx was used for location name extraction from tweets and associating tweets to neighborhoods. There are cases in which ambiguous locations were reported by LNEx. In our study, we exclude tweets containing ambiguous location entities. The location ambiguities were observed in a following cases:

- A location entity may have several matches in the gazetteer. For example, *Holiday Inn* and *Gamestop*.
- A location entity having a single gazetteer entry can potentially refer to a huge area. For example, gazetteer entry of Trans-Siberian Highway in Russia spans from St. Petersberg to Vladivastok. Such entities cannot be mapped to a single neighborhood.
- Location entities extracted by LNEx having a gazetteer entry but not referring to a location in the context. For example, *American Girl*, *Modern Male* etc.

Such mentions were identified manually and excluded from the study. This pruning step resulted in 4846 unique locations in the area that were spotted in 545k tweets and were mapped to 424 neighborhoods.

A.2 TF-IDF of crime related words

We tokenized each tweet in our train set preserving the hashtags, handles, and emojis as separate words. We then removed the stopwords and lemmatized the tokens. Bigrams of the tweets were added to the token set. The top 100 crime-related terms that had the most frequency across the tweets were chosen as our vocabulary set. For the test set, we concatenated all the tweets in each neighborhood to get a single corpus per each neighborhood. We then, transformed each corpus to get the corresponding term-document vector.

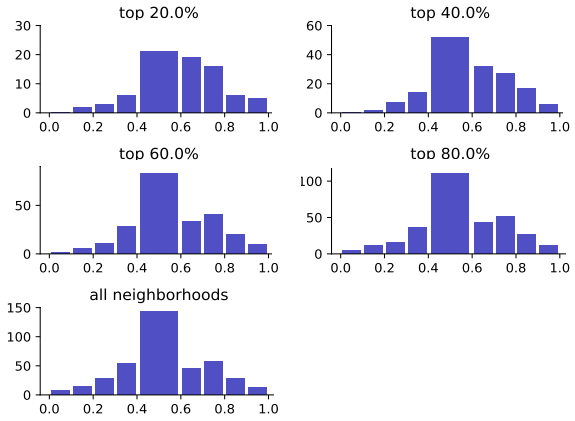

Figure 4: Distribution of the collective efficacy values in each set of neighborhoods. The collective efficacy values are computed from the survey study and are normalized in the 0 to 1 range. Refer to Section 3 for more details.

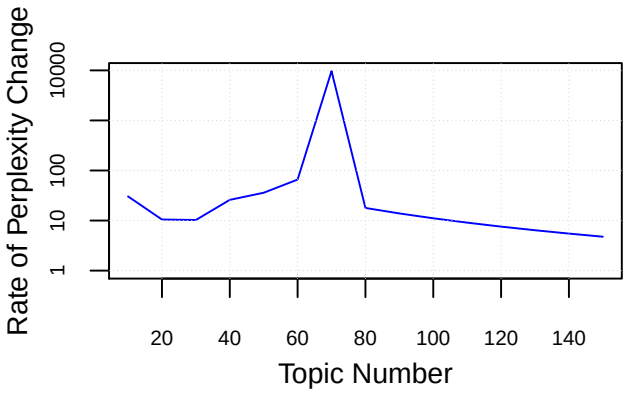

Figure 5: RPC was used to determine the appropriate number of topics. RPC is maximized for 70 topics.

A.3 Distribution of spatio-temporal urban activities

Prior to feeding the corpus to the LDA module we tokenized the tweets using a tokenizer adapted for tweets[6], removed stop words, lemmatized the tokens, and added the bi-grams that appeared in more that 20 tweets to our set of tokens. Next, we removed the words that appeared in less than 20 tweets (rare words) or more than 50% of the tweets. Employing RPC, we used an increment of 10 and varied the number of topics from 10 to 150 and trained LDA model on a corpus of 5M tweets collected from users' profiles. As depicted in Figure 5 RPC in maximized at 70 topics. Thus, we used 70 as the optimal number of topics for our model.

A.4 Document Embedding

We tokenized, lemmatized, and removed the stop words of 5M tweets collected from user profiles. Subsequently, we fit a Doc2vec model on this corpus. We set the vector size to 50.

[6]We used the open source tokenizer presented in https://github.com/erikavaris/tokenizer

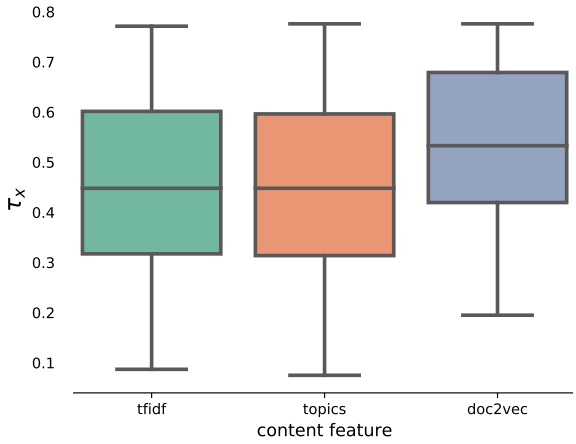

Figure 6: The box plot of τ_x values for TF-IDF, distribution of topics and doc2vec baselines on Top 40% of tweeted neighborhoods with tie threshold from 0 to 1 with an interval of 0.2. Three classifiers including random forest, multi layer perceptron, and logistic regression were used to conduct overall of 54 experiments per content feature.

For each neighborhood we concatenate all of the associated tweets and generate the embedding using the trained model.

A.5 Sentiment Distribution

We applied the 5 sentiment analysis tools to each tweet and normalized the values in a range of -1 to 1. Most of these tools predict the sentiment value using a predefined lexicon. Thus, they cannot perform accurately in the absence of sentiment lexicons in the tweets. To account for this, for each tweet, we only consider the non-zero outputs and compute the average value of them. Subsequently, we use a binning step to put the tweets associated with a neighborhood in four bins - highly negative, negative, positive, and highly positive. We normalized the value of bins by dividing the counts by the total number of tweets of the neighborhood. At the end of this step, for each neighborhood, we report the distribution of sentiment of all the tweets mentioning a venue located inside the boundaries of the neighborhood.

A.6 Distribution of Collective Efficacy

The distribution of collective efficacy has been presented in Figure 4. As it can be seen in the plot, in all of neighborhood sets, the distribution of collective efficacy ground truth values in approximately similar to set of all neighborhoods. Also, it can be seen that most of the block groups have a collective efficacy value between 0.4 and 0.6.

A.7 Tied Neighborhoods

As discussed in section 4.1 we define tied neighborhoods as the ones having a significantly small difference in collective efficacy value. Tied neighborhoods are considered interchangeable in the ranking. We define the ties based on a threshold on collective efficacy difference. We refer to this value as "Tie Threshold". We compute the standard deviation of the collective efficacy value of the neighborhoods in

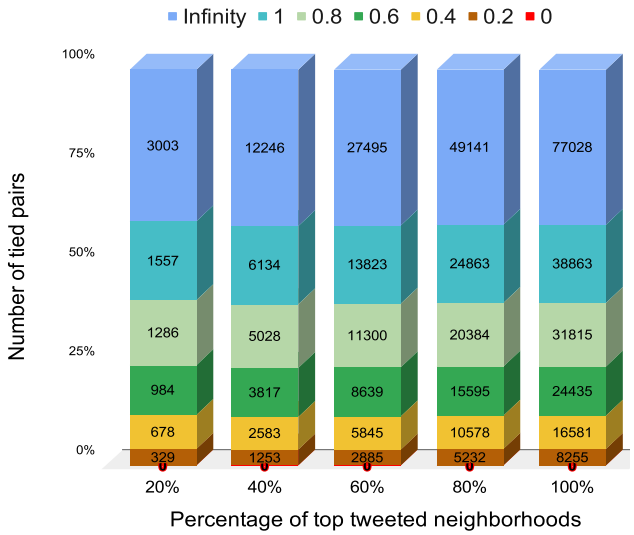

Figure 7: Number of tied neighborhood pairs at each tie coefficient. Tied neighborhoods are not ranked against each other. For N number of neighborhoods at each set, the number of paired is $N \times N - 1$. Total number of pairs at each tie threshold is shown with label "infinity". By increasing the tie threshold, the number of tied neighborhoods increases.

our study and define our threshold based on different coefficients of the standard deviation of the collective efficacy. We refer to these coefficient as "Tie Coefficient". We vary the coefficient from 0 to 1 with 0.2 increments and evaluate the ranking consensus using a ranking correlation metric discussed in section 4.5. The number of neighborhoods that are considered as "tied" at each tie threshold is shown in Figure 7. As indicated in the plot, by increasing the tie threshold, number of tied pairs increases.

A.8 Results

In order to find the best context feature, we experimented with features in this group namely TF-IDF of crime lexicon, topic distribution of urban activities, and doc2vec on the top 40% highly tweeted neighborhoods. We performed experiments with all different combinations of our 3 content factors. Each content feature is enabled in 3 combinations and disabled in 3 other corresponding paired combinations. We conducted each factorial experiment with 3 classifiers for the local ranking module including random forest, multi layer perceptron, and logistic regression. We repeated this process for 6 tie coefficients. Tie coefficients varied from 0 to 1 with an interval of 0.2. The number of tied neighborhoods at each tie coefficient is shown in Figure 7. The cross product of these parameters resulted to $3 \times 3 \times 6 = 54$ experiments in which a content feature is enabled. The box plot of the observed ranking consensus for these 54 experiment for each content feature is presented in Figure 6. As it can be seen in the figure, by characterizing neighborhood's tweets using doc2vec we consistently generate better rankings in comparison to TF-IDF of crime lexicon and topic distribution of urban activities.