Media Texts and Processing

News stories originated in the ICEWS database and were selected by first querying the database for stories located in Yemen. This resulted in 47,385 stories ranging from January 15, 1991 through January 4, 2015. I then selected only "violent" events, defined as events that fall into one of the following ICEWS event types: "Threaten with military force," "Use unconventional violence," "Violate ceasefire," "Use as human shield," "Threaten," "Occupy Territory," "Physically assault," "Mobilize or increase armed forces," "Engage in violent protest for leadership change," "Engage in mass killing," "Conduct suicide, car, or other non-military bombing," "Carry out suicide bombing," "Attempt to assassinate," "Assassinate," "Abduct, hijack, or take hostage," "fight with small arms and light weapons," "fight with artillery and tanks."

This resulted in 10,818 stories, of which I took a random sample of 1772 stories. ICEWS codes for event date, event type, source actor, and target actor. However, the source and target actor codes typically characterize the actor by their role, such as "Armed Rebel" or "Militant," rather than by group affiliation.

An example of stories featured in the corpus can be seen below. The two stories were chosen at random from the corpus.

A June 28, 2011 story from *Xinhua* news describing an al-Qaeda operation highlights the target of an al-Qaeda offensive (a military hospital), remarks on a distinctive weapon choice (heavy machine guns).

> "Al-Qaida militants carried out offensive attacks targeting the 25th Mechanized Brigade in the east of Zinjibar city, which aroused heavy clashes, leaving three soldiers and eight al-Qaida militants killed in addition to injuring dozens of others from both sides, the official told Xinhua, who asked to remain anonymous. Fierce battles are still ongoing around the military brigade, which was surrounded by

> the militants, the official said. Al-Qaida group was trying to bring down the military brigade by using heavy machine guns, he ad ded.
>
> Meanwhile, a local medic at the Basuhib military hospital in Aden said that dozens of injured soldiers were receiving treatment from the clashes..."

Similarly, a September 22, 2013 article from Agence France Presse about activities attributed to Ansar al-Shariah likewise identified the target and tactics of the attack, and indicated that the method of assassination has been a frequent strategy of Ansar al-Shariah.

> "ADEN: Gunmen have shot dead a Yemeni man in the south of the country because they suspected he was homosexual, police said yesterday, in the sixth such killing this year. One of two men on a motorbike opened fire at the man in his 20s outside his house in Huta, the capital of Lahij province, killing him on the spot. Police said the attackers, presumed Islamist militants escaped. Similar murders in the nation's provinces of Abyan and Aden have been blamed on an al-Qa'ida-affiliated group, Ansar al-Sharia. For the past year,Ansar al-Shariah has imposed Islamic law on areas of Abyan where it still holds sway. Its so-called courts have condemned to death several people. Others have had hands amputated after being 'convicted' of theft. After an army offensive in May last year ousted the militants from areas they controlled, they holed up in mountainous regions of south and southeast Yemen."

Notably, in the two articles the two Sunni groups are presented as operating in different spheres. AQAP is described as a significant military threat, using heavy weaponry against a hard military target, with Ansar al-Shariah trying to impose local administration and social behaviors via a campaign of individual-level violence and assassination.

In order to generate data on how groups operate, I re-coded the reports to include a variable for group or movement affiliation. I first sent the sample to Amazon's Mechanical Turk

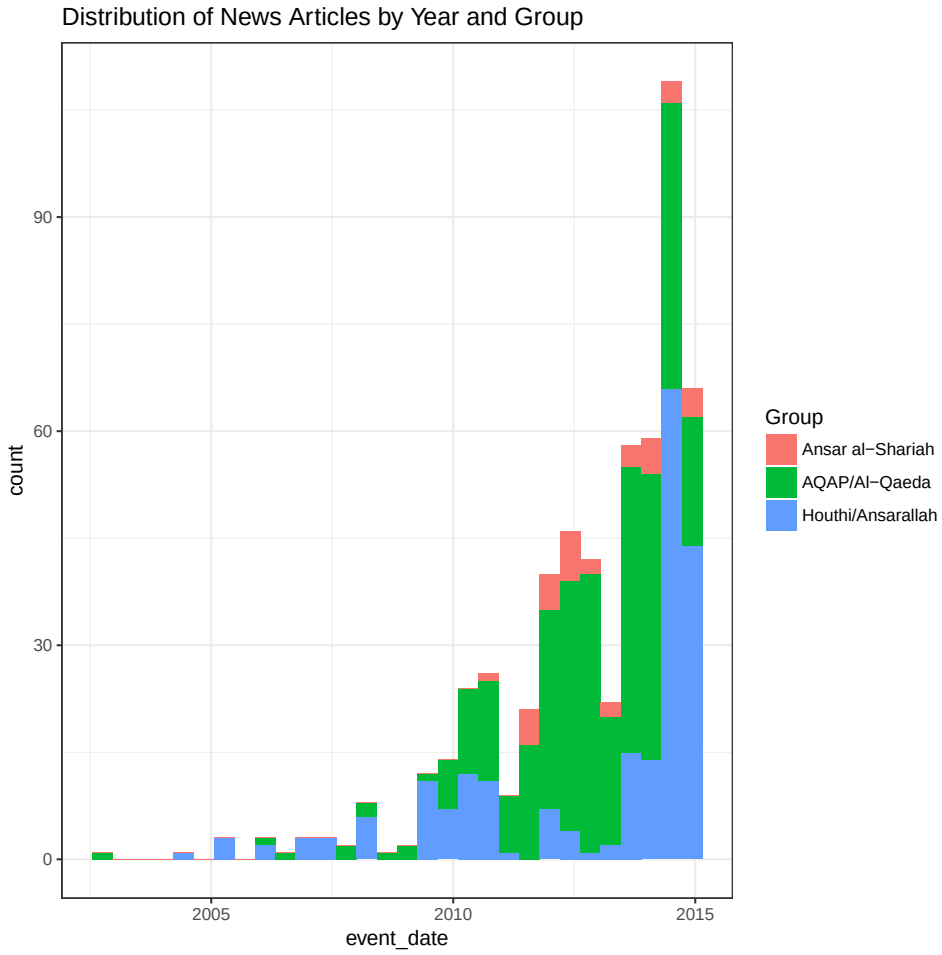

Figure 1: Distribution of News Articles

platform, asking the workers to categorize the stories as relating to an action carried out by Ansar al-Shariah, AQAP, Houthi/Ansarallah, Yemeni Government, Tribal Uprising, Other, Multiple Actors, or Unknown. I kept the tags for the 283 stories that both coders agreed on, and hand-coded the remaining 1489 stories. The temporal distribution of the news articles and actor labels can be seen in Figure 1. I then further subset the data to keep only the stories tagged as describing a violent event carried out by one of the three militias of interest. This produced the final 720-story corpus of news events. I randomly divided these into a development-validation-test set at a 60%, 20%, 20% ratio. The analysis presented in the manuscript uses the development and validation sets as effective training and test sets, and reserves the remaining set of 144 held-out stories as a clean test set for future refinement.

For each of the development, validation, and test sets, I used the tm() package to tokenize the words in each story and remove numbers, standard English stopwords, whitespace, and stray HTML markup. I additionally removed a custom list of stopwords that strongly signal the group, such as variations on the group name and signifiers of sectarian identity. These custom stopwords are comprised of: of word that signaled AQAP: "qaeda," "alqaida," "alqaeda," "qaida"; words that signaled Houthis: "houthi," "huthi," "houthis," "zaidi," "alhouthi"; terms that signaled Ansar al-Shariah: "ansar," "sharia," "alsharia";[1] terms that suggest an al-Qaeda affiliation: "laden," "osama"; words often used to summarize location of action for one of the groups: "peninsula," "northern," "southern," "arabian," "yemenbased"; and finally terms that denote sectarian identity: "sunni," "shia," "shiite." Word frequency was normalized via term frequency-inverse document frequency (tf-idf), producing a pair of tf-idf matrices, from which I took the intersection of features (i.e. words).[2] This generated a set of 2,222 "features" for classification in the texts. This step reduced the available terms significantly, but is necessarily to test models across the training, validation, and test sets.

[1]The robustness models also remove "alshariah" with little change in results.

[2]Following the insights from (Denny and Spirling, 2018) I minimally process the texts as I am looking for potentially subtle differences in the presentation of the groups, rather than thematic similarities among the texts.

	Absolute Frequency	Proportion of Documents
Ansar al-Shariah	27	5.8%
AQAP/Al-Qaeda	260	56.3%
Houthi/Ansarallah	174	37.7%
Total	461	99.8%

Table 1: Distribution of group labels in "development" set

I then reattached metadata to each of the term document matrices. Metadata included group label, date, and whether the story was coded by Mechanical Turk workers.

The distribution of group labels in the development set can be seen in Table 1 with the corresponding distribution from the validation set in Table 2.[3]

	Absolute Frequency	Proportion of Documents
Ansar al-Shariah	9	7.8%
AQAP/Al-Qaeda	67	58.2%
Houthi/Ansarallah	39	33.9%
Total	115	99.9%

Table 2: Distribution of group labels in "validation" set

Machine Learning Classifiers

The supervised machine learning techniques used in the paper provide a strategy to adjudicate between the counterfactuals introduced in the theory and qualitative sections. In particular, the clustering analysis indicates that international and local journalists writing about events in Yemen use similar terms when describing the activities of AQAP and Ansar al-Shariah. This suggests that AQAP has been unable to maintain a local spin-off with a distinctive operational profile.

However, one significant caveat is that these techniques are unable to distinguish between AQAP acting like Ansar al-Shariah, Ansar al-Shariah acting like AQAP,[4] or journalists

[3]Deviations from 100% in the relative frequency sums is due to rounding.

[4]One natural counterfactual in which an influx of local fighters is followed by behavioral convergence of AQAP and Ansar al-Shariah but which does not follow the mechanism hy- pothesized by the bottom-up transformation theory could be that AQAP's socialization has been so successful that the group has changed

conflating Sunni insurgent groups. Distinguishing between the three possibilities is important to assess the theoretical expectation that an inflow of recruits should pressure AQAP's leadership to adopt a local emphasis. The topic modeling section addresses concerns about the direction of convergence.

The following section provides technical details about the implementation of the tSNE visualization and the SVM and Random Forest classifiers.

tSNE Hyperparameter Selection

Figure 2 presents tSNE clustering for all stories published between 2011 and 2014, presented according to year of publication. These dates provide a snapshot of writing about each of AQAP (green), Ansar al-Shariah (red), and the Houthis (blue), and provide a high-level visualization of the separation or convergence among the words used to describe each of the three groups. The yearly clustering displayed in 2 features one point per story, and indicates that across the time period, stories about the Houthi insurgency appear to be systematically different from stories about the two Sunni groups, and is suggestive of a pattern in which from 2012 through 2014, the Ansar al-Shariah stories become progressively more similar to AQAP stories.

The visualization presented here was generated by running the tSNE algorithm on for 5,000 iterations on the pooled data. The perplexity hyperparameter presented below was selected after grid sweeping from 5-50, at intervals of 5. Sweeping the perplexity hyperparameter changes the exact outcome, as expected from a probabilistic approach to summarizing structure in complex high dimensional data, the conclusions are broadly consistent across the specifications. To address concerns that the observed clustering is random noise or driven by a specific initialization, clustering was carried out in parallel on two different machines using the same specifications but different starting points. The results were broadly similar,

the preferences of the communities in which they operate. In this scenario, the local Ansar al-Shariah should gain a greater international focus as local actors are socialized into the transnational jihadi ideology.

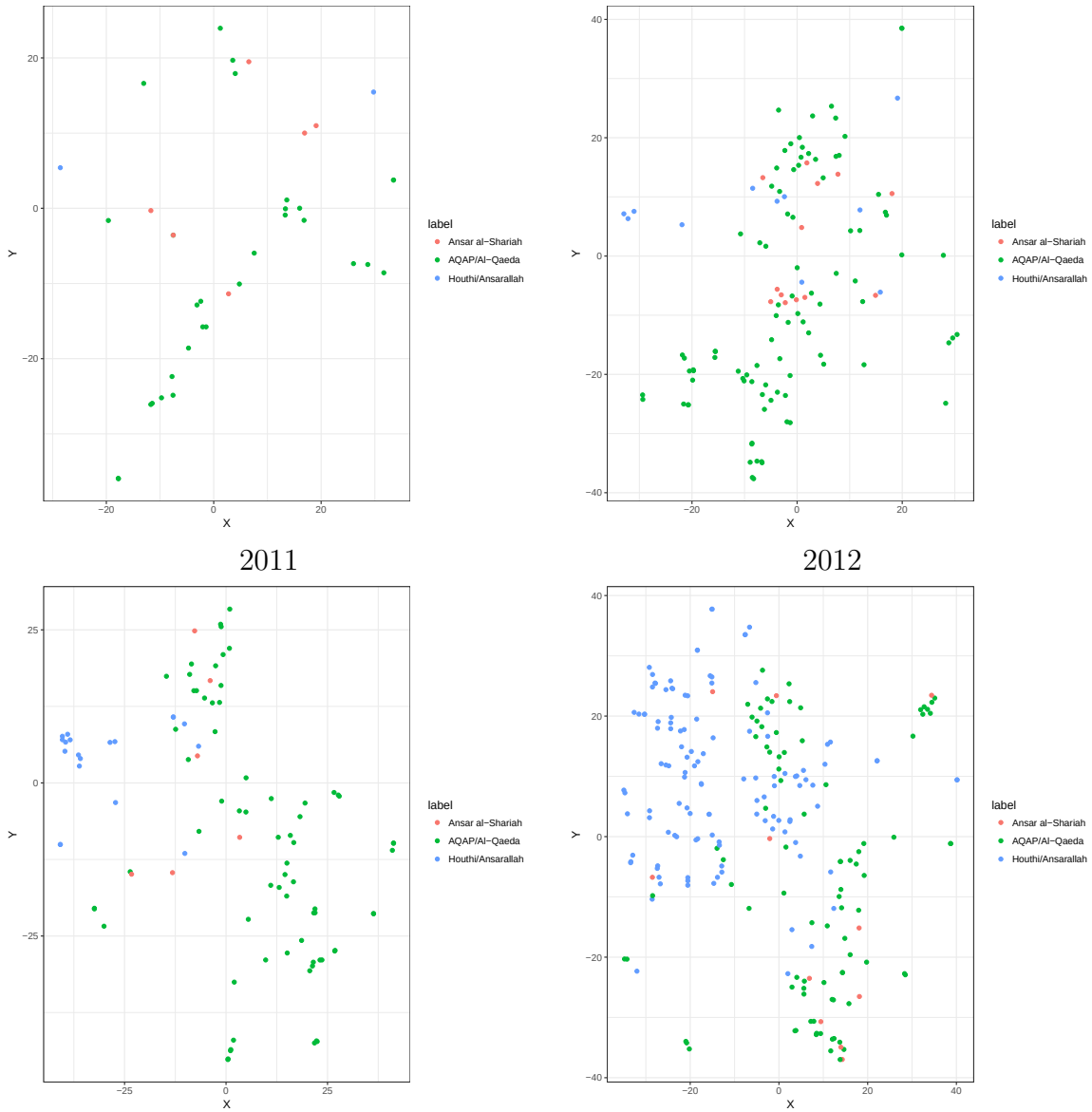

Figure 2: tSNE Clustering, All Stories 2011-2014

namely clear separation between the Sunni and Houthi stories but lack of clear separation among stories about AQAP and Ansar al-Shariah. As the diffusion and relative positioning of clusters generated using tSNE are not inherently meaningful, comparison across the runs is only impressionistic and presenting averaged results would not be meaningful.

Figure 2 presents tSNE clustering for AQAP and Ansar al-Shariah stories published between 2011 and 2014, presented according to year of publication. These dates provide a snapshot of writing about each of AQAP (green), Ansar al-Shariah (red), and the Houthis (blue), and provide a high-level visualization of the separation or convergence among the words used to describe each of the three groups. The yearly clustering displayed in Figure 2 features one point per story, and indicates that across the time period, stories about the Houthi insurgency appear to be systematically different from stories about the two Sunni groups, and is suggestive of a pattern in which from 2012 through 2014, the Ansar al-Shariah stories become progressively more similar to AQAP stories. Figure 3 focuses only on whether or not the tSNE visualization differentiates between Ansar al-Shariah and AQAP stories. As compared to the separation of the Houthi stories in the corpus, the two Sunni groups demonstrate no apparent separation. This implies that stories about the two groups are much more similar than are stories about AQAP and the Houthi insurgency.

Random Forest Parameter Selection

The random forest classifier used the randomforest() method from the randomForest R package. The specification used the development data as training data, and the validation data as a test set, with story label as the classifier to predict. The model grew 500 trees, from which is generated a proximity measure for each document. The variable importance plot was generated after extracting the variable importance measure and plotted using the varImpPlot() method native to randomForest. As the only predicted categories for the

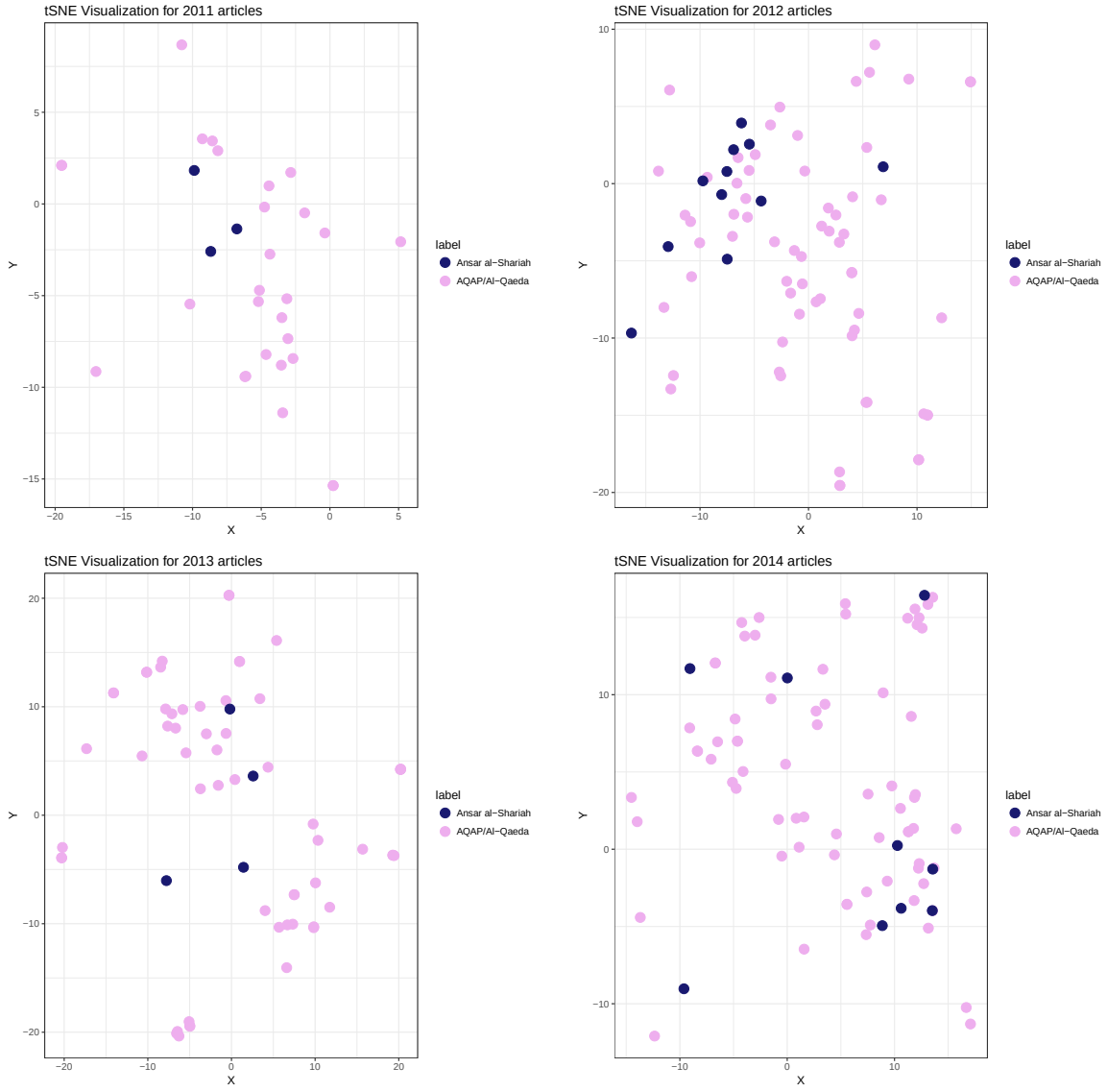

Figure 3: tSNE Visualization Sunni Groups, 2011-2014

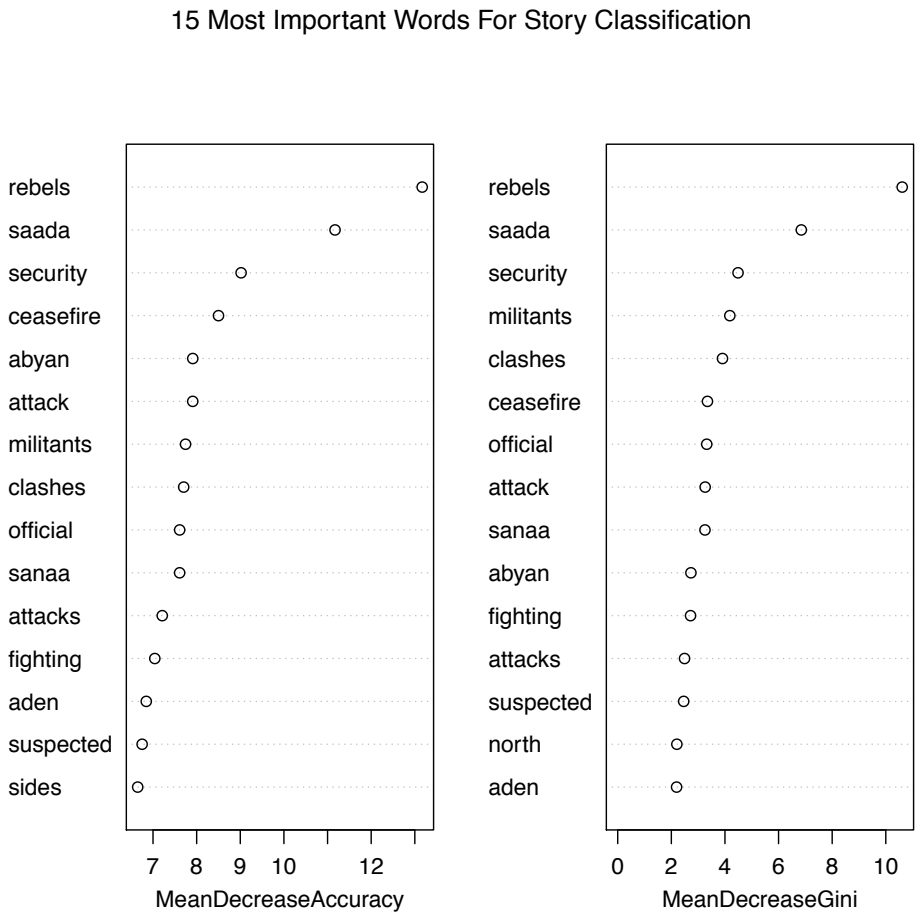

Figure 4: Important Words for Classification

stories were AQAP and HouthiAnsarallah, the interpretation that the random forest was identifying Houthi stories based on widespread description of the Shia group as a rebellion while the AQAP activities are more typically described as militant activity.

A second version of the random forest model estimated the model only on Ansar al-Shariah and AQAP stories. As with the full data, this model failed to predict any stories for the Ansar al-Shariah label. However, it did predict two AQAP-labeled stories as being likely Ansar al-Shariah stories. These stories were an August 3, 2014 story about a clash between al-Qaeda militants and police forces and a December 17, 2014 story about a car bomb carried out against Yemeni police officers. Notably, both stories were one of the 11 stories

from Xinhua News that used the names AQAP and Ansar al-Shariah interchangeably.

One advantage of the random forest approach to classifying text is that the features used for classification are also terms in the document, which provide interpretable insights into what words drive separation among stories in the data. The fifteen most important words for the random forest classification—after removing stopwords that describe or name the active group— suggest that the reason for the clean split across the Sunni and Shia movements lies in framing. The list of these terms, ranked in order of their importance for decreasing the classifier accuracy (left) and importance for decreasing homogeneity in the final nodes of the classification (right), can be seen in Figure 4. The importance of "rebels" as the top term for both accuracy and Gini coefficient is revealing: indeed in the texts, Houthis are consistently described as "rebels" while the Sunni fighters are frequently presented as "militants." After the rebel/militant split, Words that describe the location of operations and military occupations are, unsurprisingly, important classifiers: Aden, Saada and Sanaa are regions associated with Houthi territorial gains, while Abyan is more closely linked to AQAP and Ansar al-Shariah activities.[5]

Principle Component Analysis

Another view of the separation among the stories is shown in Figure 5, which uses a principle component analysis (PCA) to plot proximity of stories, as measured by the proportion of times that individual stories are in the same terminal node (Jones and Linder, 2016). This reaffirms the takeaway from the confusion matrix: Houthi stories are distinct from AQAP stories, but Ansar al-Shariah stories contain enough words in common with AQAP stories that the two are difficult to distinguish via the random forest's iterated decision trees.

[5]Although the custom stopword list described in previous sections removed many terms relating to location, I did not remove areas of operation from the texts as the goal of the classifiers was to seek discussion of operational differences. Retaining location identifiers such as "Arabian Peninsula" would provide a too-easy test for the classifier, which could simply look for that element of AQAP's group name, without necessarily illuminating underlying differences. However, locations of operation are substantively meaningful.

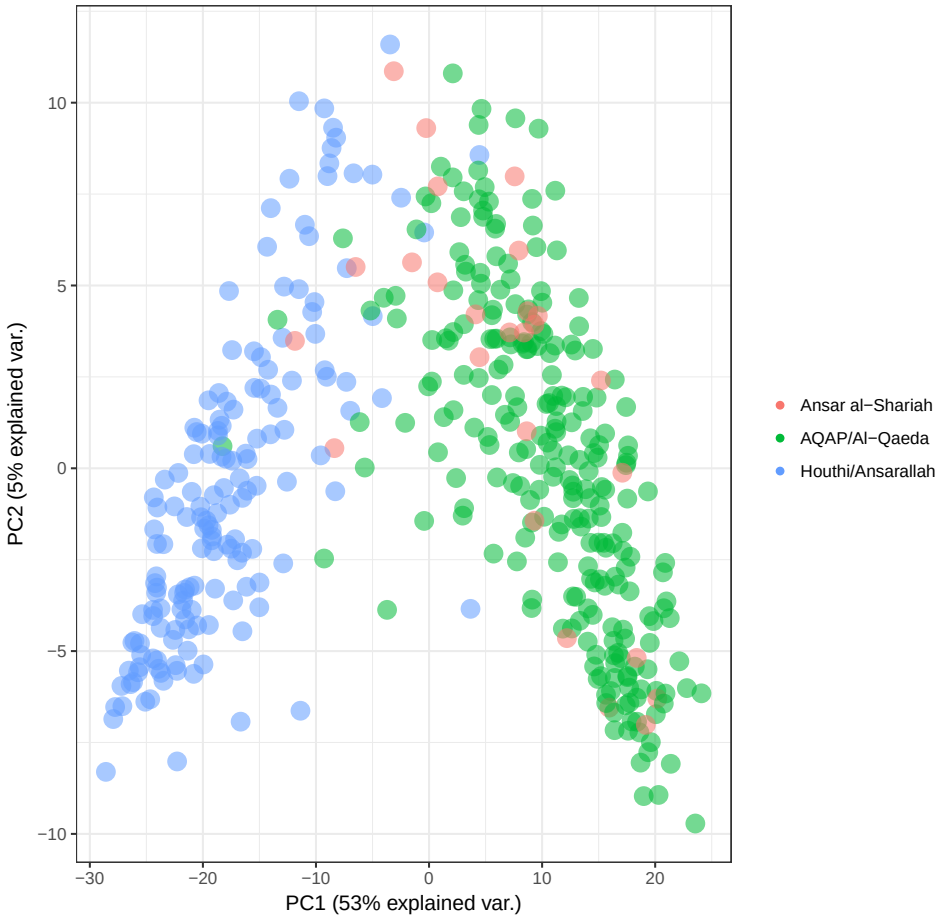

Figure 5: PCA Visualization of Group Classification

The principle component analysis presentation was done using the extract_proximity() and prcomp() methods from the edarf R package. The proximity extraction categorizes document proximity as the proportion of times that two observations are in the same terminal node across each estimated tree. This generates an NxN matrix of location similarities. This proximity is visualized using a principle components analysis of the proximity matrix, implemented in the edarf package.

SVM Specification

The support vector machine classification was developed using the ksvm() method from the kernlab R package. I used the "development" data as a training set and the "validation" data as the test set, with "label" as the classification attribute to predict and the words from each term document matrix as the features used in the classification. The kernel specified was the radial basis function (RBF) Gaussian kernel via kernlab's "rbfdot" implementation. The RBF kernel is defined as:$K(x, x\prime) = \exp(-\frac{||x-x\prime||^2}{2\sigma^2})$. This kernel is based on squared Euclidean distance between feature vectors for each document, and thus is amenable to interpretation as distance between the documents. The svm carried out a 3-fold cross-validation on the test set, and used the resulting model to predict category assignment in the test (validation) set. In the analysis presented in the paper, each class was given a weight of 1. The paper reports predicted probability of label assignment.

Given the imbalance in the data, I developed two additional versions of the svm model, subsetting the development-validation data subset to include only AQAP and Ansar al-Shariah stories. One of these replicated the original support vector machine with an equal weighting on each possible class and one of which assigned class weights to the labels in the 90-10% distribution of AQAP and Ansar al-Shariah stories. Results of these two models were similar to the model presented in the paper: given the words in the news stories as features, the SVM does not distinguish Ansar al-Shariah stories from AQAP stories.

Figure 6 provides a closer look at whether the support vector machine's confidence in whether to assign stories to Ansar al-Shariah or AQAP change over time. Ideally, support for the transformation theory would indicate the SVM assigning increasing weight to the Ansar al-Shariah label for news stories about AQAP activities as AQAP members push the group to engage with local concerns. The plot focuses on stories with a true label of AQAP and depicts the predicted probability that an article would be assigned to the label of Ansar al-Shariah given that it is a story about AQAP actions (red) as well as the SVM's confidence in the classification for AQAP (blue) for the 144 stories in the test set. As the SVM consistently predicts all AQAP and Ansar al-Shariah stories for AQAP, the expected probability of assignment to Ansar al-Shariah remains constant at approximately $p = .15$. The classification predictions provide mixed support for the expectation that an influx of local members should increase the difficulty of assigning group labels: although the SVM's confidence in predicting the AQAP label becomes more variable over time, the predicted probabilty of assignment to Ansar al-Shariah remains constant over the time period

Appendix: STM Model Selection and Diagnostics

The three STM models are based on a corpus of 1353 documents, spanning October 25, 2005 through September 21, 2016. Approximately 500 documents are associated with as-Sahab. A histogram of the distribution of can be seen in Figure 7.

Data

I focus on media released online to jihadi media platforms and outlets. Preprocessing removed words that occurred in fewer than two or more 70% of the documents in the corpus.[6]

[6]In the AQAP corpus, there was no change to the number of tokens in the corpus for an upper bound threshold between 70-95%. I evaluated coherence and exclusivity at an upper threshold of 50%, but did not find results that would suggest either a coherence or exclusivity benefit from the additional reduction in corpus size.

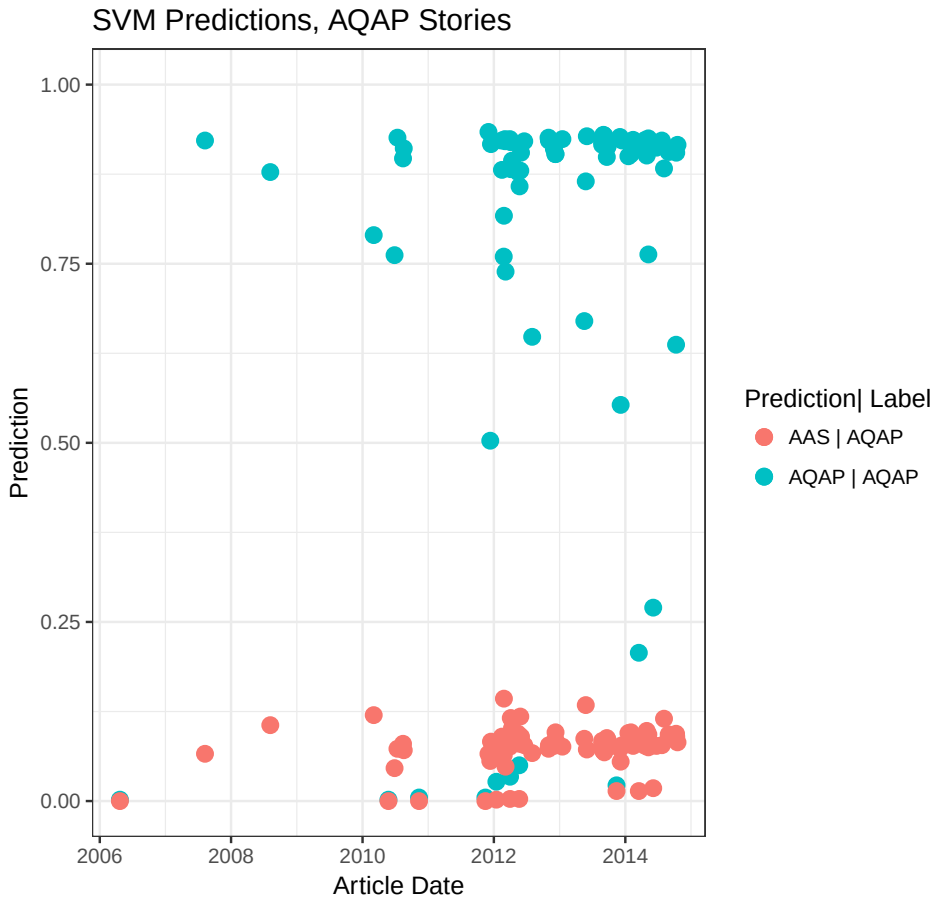

Figure 6: SVM Predictions Over Time For AQAP Stories

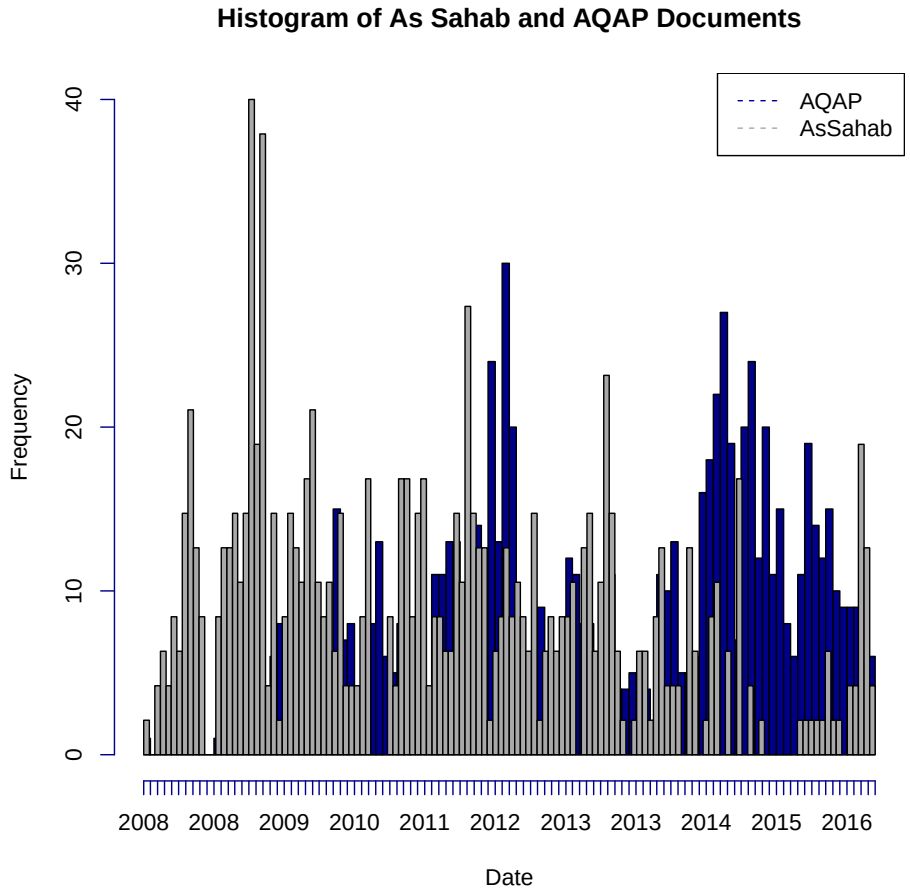

Figure 7: Distribution of AQAP and AQC Materials

Models One, Two, and Three each use time as a covariate. The time variable is expressed in the data as a running counter of days from the oldest document in each corpus. Thus, the date of the oldest document is given as 1 and the "day" of each subsequent document is modeled as the numer of days between the first date and the date of the individual document. These dates are linked to the translation date rather than release date as the former can be accurately pinpointed for each document in the corpus. For the vast majority of the corpus, the translation date closely coincides with the date that the document was released to online jihadi media outlets. The precision of release dates contained in the original Arabic text can vary according to type of document: communiques are typically dated to a specific day, while strategy documents or promotional magazines can be dated with a day, a month, or even a season. Thus, for consistency, the date covariate is linked to translation date.

Model Selection

Topic models rely on the user to prespecify a number topics for the algorithm to search for. However, this parameter fundamentally influences the themes that will be identified in the documents. For models one and two, I selected the number of topics by doing a sweep of model specifications with 10 to 30 topics. I then selected a topic number that performed best on both semantic coherence and exclusivity.[7] After this process, a model with 18 topics appeared to present the greatest gains to semantic coherence without trading off exclusivity. Moreover, the 18-topic model identified topics that were particularly substantively coherent.

For the joint model, after comparing the semantic coherence-exclusivity trade-off for models across a sweep from 10 to 40 topics, I set the number of topics to discover at 34. The increase of topics reflects the expectation that the two organizations are already rhetorically distinct, and so the joint corpus should require more topics. Specifically choosing an output that

[7]Ideally, the selected number of topics would have relatively high exclusivity and semantic coherence. I often faced a trade-off between the two. When determining the trade-off, I prioritized semantic coherence over exclusivity. The exclusivity bands were, overall, narrow while coherence varied substantially.

doubled the number of topics slightly penalized semantic coherence over a model with fewer topics, but allowed for a more precise comparison of topics between each group.

As topic models are, by nature, non-deterministic, each implementation of a given model will produce slightly different results. Thus, after selecting the number of topics for the STM to identify, I ran each model specification ten times to create a range of possible output models for analysis. I compared the average semantic coherence and exclusivity for each of the models. For each of the three models below, I found that the averages within each ten-model set were nearly identical. To avoid biasing my results by selectively choosing the output that best confirms my theoretical expectations, I chose which specific models to analyze by maximizing average coherence and exclusivity metrics. As no model clearly dominated the coherence-exclusivity trade-off, I assigned a relatively stronger weighting to semantic coherence when selecting a specific iteration to present. I then selected a model to present before qualitatively evaluating any of the topics. This decision was intended to avoid bias in choosing how to prioritize coherence gains against exclusivity losses.[8]

Finally, after selecting which model to present, I evaluated the remaining models to ensure that the output was consistent across the set of ten results for each model. In particular, I verified general agreement on the thematic content identified across the runs.

Full Topics In STM Model Two

As with Model One, Model Two is an 18-topic Structural Topic Model estimated on the 875-document AQAP corpus. Unlike the first model, Model Two includes a binary indicator of whether or not a document was issued under the name "Ansar al-Shariah." 145 documents in the corpus had this tag, distributed in three temporal waves across the corpus. Figure 8 shows the distribution of Ansar al-Shariah documents in the corpus. As Figure 8 indicates, Ansar al-Shariah documents occur in three roughly equal waves, starting on August 31,

[8] A plot of average semantic coherence and exclusivity scores is available upon request.

2011 and ending on April 9, 2016. This balanced pattern of date distributions can help to assuage concerns that the localizing trend from Model One is an artifact of the concentration of local conflict themes in Ansar al-Shariah texts. If Ansar al-Shariah branded documents were concentrated towards the end of the corpus time frame and were strongly associated with localizing trends, then the localizing trend from Model One might simply reflect the influence of the locally-focused Ansar al-Shariah brand. However, as the Ansar al-Shariah documents occur in three separate waves in the data and end just as the local influence is most pronounced, the findings from Model One are more likely to reflect actual trends in AQAP's messaging.

Clustering and FREX words for the substantively interesting topics of the second STM model are summarized in Figure 9.

As-Sahab vs AQAP Corpus

Due to the complexity of a 34-topic model, the section briefly summarizes the results of the STM model from the joint AQAP and As-Sahab corpus. Twenty-nine of the substantively interesting topics from the model are featured in Figure 11.[9]

For ease of interpretation, the topics are clustered into four general categories: those primarily relating to Yemen, topics with strong religious overtones, topics that suggest engagement with global jihadi issues, and topics that address specific countries and battlefields other than Yemen. This clustering was done on the unsupervised STM output using the author's substantive expertise. Within each cluster, topics are summarized according to the top FREX words.

Rather than present expected topic proportions, which are difficult to process visually for such a large number of topics, Figure 11 presents the topics according to their likelihood

[9]Of the five topics excluded from the summary, four were associated with editing and document preparation and the fifth consists of declarations of defiance and intention. FREX words for this last topic include: "know," "think," "can," "see," "thing," "happen," and "now".

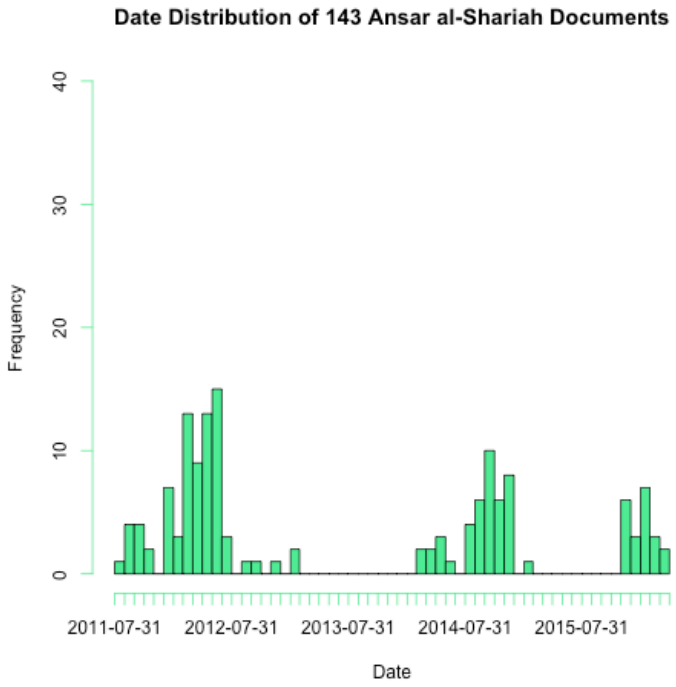

Figure 8: Histogram of Ansar al-Shariah Documents in AQAP Corpus

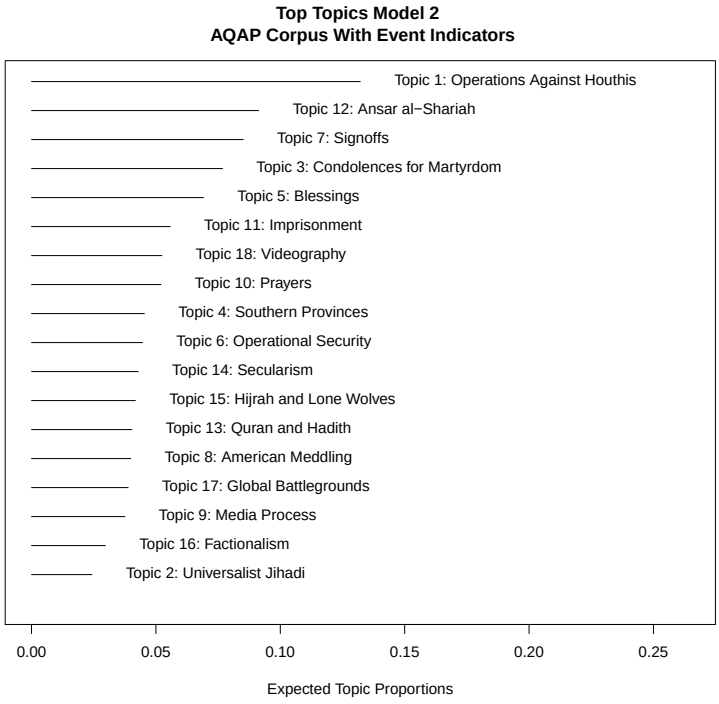

Figure 9: Expected Proportion of Topics in Model Two

of being associated with either AQAP or As-Sahab. Positive values indicate a stronger association with AQAP's corpus, while negative values indicate topics more closely associated with As-Sahab. The point estimates of the difference in topic prevalence are presented along with 95% confidence intervals. Unsurprisingly, topics associated with battlegrounds other than Yemen, including Afghanistan, the Indian subcontinent, Libya, Somalia, and Syria, are all statistically more likely to have occurred in the As-Sahab corpus. Conversely, AQAP is significantly more likely to use terms that refer to specific events and locations in Yemen.

The two topics highlighted in the main text, named as "Crusader and Zionists" and "Defending the Weak" both occur in the "Global Jihadi Topics" cluster. Topics in this cluster are, on the whole, not statistically associated with either media group or the other, and thus group-level differences in attention to them over time are unlikely to be an artifact of frequency of messages from one of the two organizations. Indeed, with a point estimate near zero and confidence intervals crossing the null line, the across-time group association of the

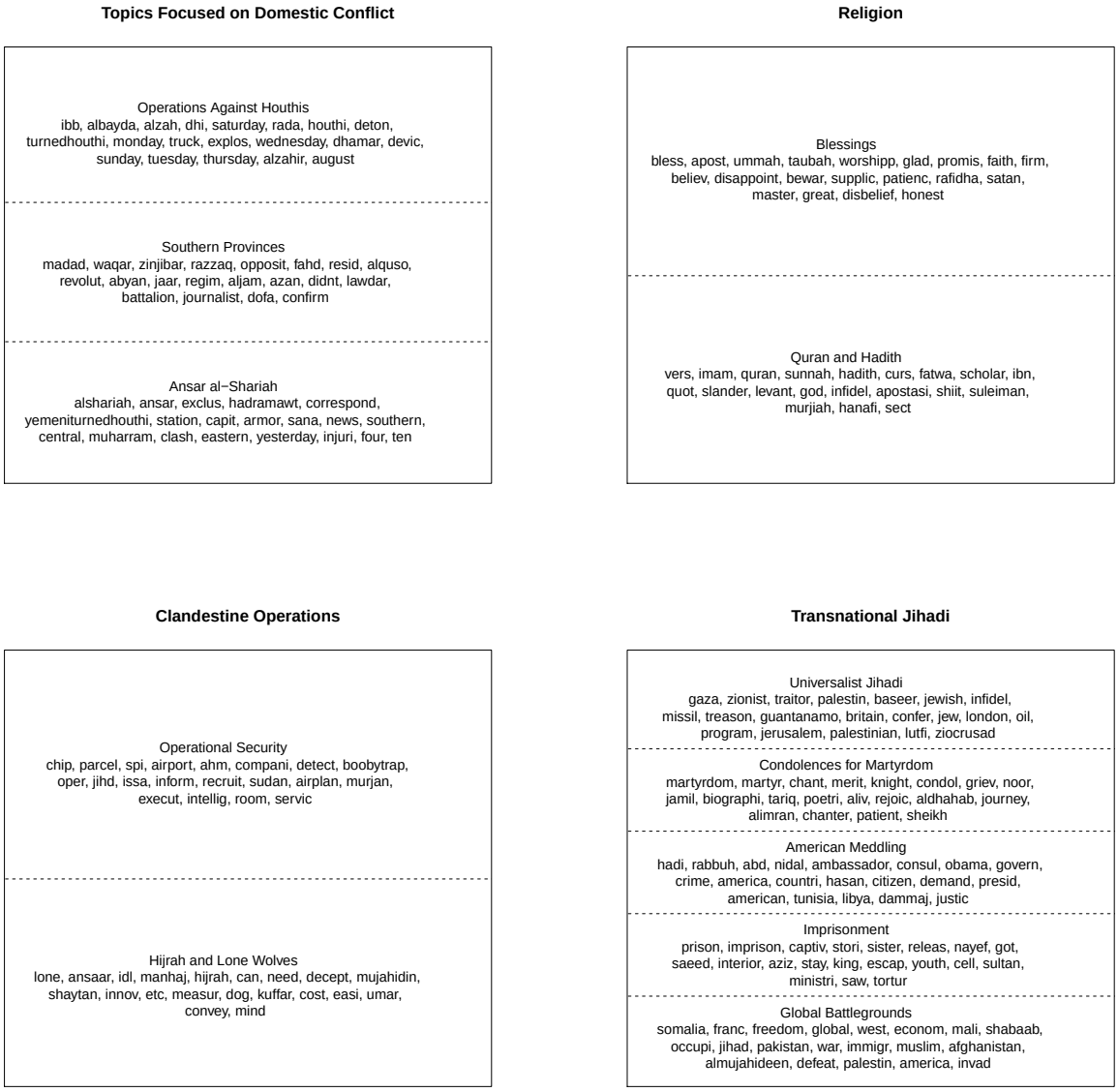

Figure 10: Groupings of Substantive Topics in Event Covariate Model

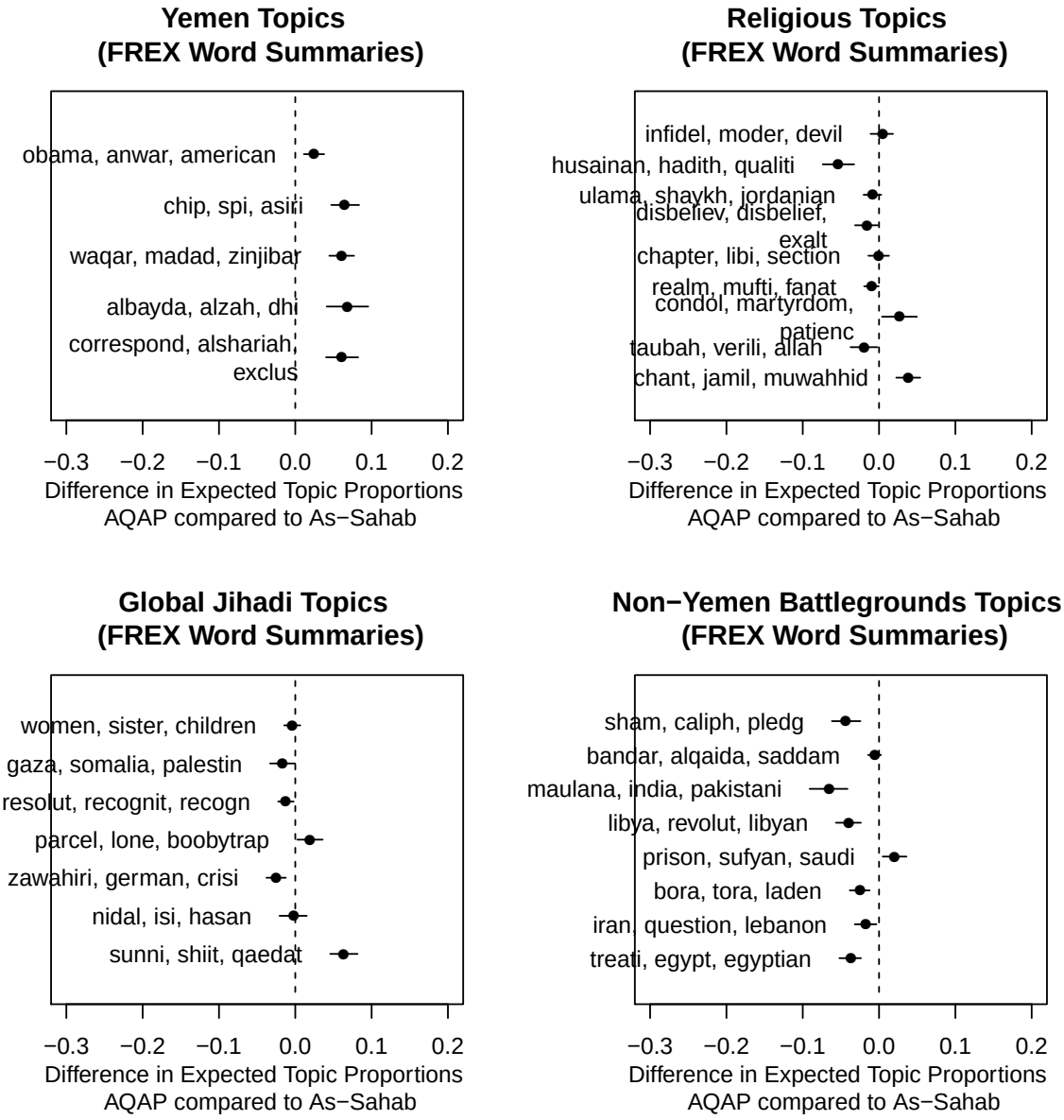

Figure 11: Summary of 34-Topic AQAP and As-Sahab Model

"Defending the Weak" topic, summarized with the FREX terms "women, sister, children" is statistically indistinguishable for AQAP and As-Sahab. Thus, it is notable that when disaggregating the topic prevalence across the dates of the corpus, the topic is statistically more likely to feature in As-Sahab propaganda. Similarly, the group-association point estimates for the "Crusader and Zionists" topic is featured directly under that of the "Defending the Weak" topic. It is marginally associated with As-Sahab in the plot in Figure 11, but as the "AQAP and As-Sahab Divergence" figure in the main document indicates, clearly statistically significantly more likely to feature in media from As-Sahab. Taken together, these results suggest that the groups started out the time period with relatively similar attention to these topics, but that over time AQAP has directed their propaganda attention completely away from this topic as they became more preoccupied with the local Yemeni conflict.

References

Denny, Matthew J and Arthur Spirling. 2018. "Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it." *Political Analysis* pp. 1–22.

Jones, Zachary and Fridolin Linder. 2016. edarf:Exploratory data analysis using random forests. Vol. 1.