

A Torres, Michelle. SI for “A framework for the unsupervised and semi-supervised analysis of images”

This Appendix provides additional analyses, diagnoses, and details regarding the application and plots presented in the main text. It is organized into 9 sections.

- **BoVW and CNNs (A1):** pp. 2-7
 - Discussion and comparison of CNN approaches for unsupervised learning and BoVW+STM framework.
- **Technical details: detection and description of key points (A2):** pp. 8-9
 - This section provides technical details and extended information regarding the location and description of key points in images.
- **Building a vocabulary: sample of features and clustering (A3):** p. 10
 - Extended information regarding the sampling of features and clustering for the construction of visual vocabularies.
- **Emulating the Document-Term matrix: the Image-Visual Word Matrix (A4):** p. 11
 - Illustration of similarities between Document-Term Matrix and Image-Visual Word Matrix.
- **Extension and Supplementary Information of Application (Framing a Movement) (A5):** pp. 12-16
 - Information and extended analysis of the running example in the main text regarding the migrant caravans from Central America.
- **Structural Topic Models and Latent Dirichlet Allocation (A6):** pp. 17-18.
 - Discussion of the use of STMs for studying framing, theoretical and practical comparison of both methods using the example of the Caravan of migrants.
- **Topic modeling – label validation (A7):** pp. 19-20.
 - Table with responses to coding tasks and word clouds of words associated with the content of the most representative images of each topic.
- **Diagnosis and practical guidance – parameter selection (A8):** pp. 21-28.
 - Details and results from a set of diagnoses analyzing the consequences of parameter choice in the BoVW process.
- **Credits for images/photos used in the main text and appendix (A9):** p. 29-30.
- **References used in the Appendix (A10):** p. 31-32.

A.1 BoVW and CNNs

In this section, I present a brief comparison of the performance and results from analyzing images using a BoVW and a CNN. First, I discuss the differences in the results that arise from using a binary indicator to identify a given object (in this case a “dense crowd”) in contrast to the proportion that the object as a theme occupies in the picture. Then, I use the `Google Vision API` to extract content from the images (in this case the objects and their “topicality”) and compare the output to the topic proportions from a visual STM. Finally, I use a pre-trained CNN to learn and extract features from the images, and cluster the images according to them. I present a brief comparison and discussion between this output and the one from the BoVW approach.

A.1.1 Continuous vs. Binary classifications

Binary quantifiers of a picture such as “Presence of a crowd” might be fitting for certain projects. However, as in the case of visual framing, if researchers do not only care about the mere presence but also the way in which a given object or theme is included in a picture, then other measures are more adequate. For example, a proportion of a given object in a picture is more relevant for a researcher analyzing the salient content and focus of visual coverage rather than just the knowledge of whether a given object exists. This is a common issue in fields like political communication, journalism and advertising and a prevalent case in the pictures under analysis (as Figure 7 in the main text shows).

However, below I provide example of how the two different measures can yield different results. In the following table, I regress the proportions of the “dense crowd” topic per image (Column 1 of Table [A.1](#)) and the indicator of whether there is a crowd (manually coded, Column 2 of the same table) on the ideology of the news outlet. The use of a continuous measure, the proportion of a frame, might not only be useful for research objectives but may also uncover interesting results.

In this case, the proportion of crowds that center left and center right outlets use is significantly lower than the proportion used by right leaning outlets. Although the model with a binary indicator also shows a negative coefficient for these outlets, the coefficients are not distinguishable from zero at conventional levels. However, it is also important to note that both models show a reliable and negative effect of being a left leaning outlet on the prevalence of dense crowds in pictures about the Caravan. If the theoretical expectation of a researcher was merely to test this difference, using either of the two approaches would have led to the same conclusion.

However, if the exercise is exploratory or the research question deals with the comparison of the right to other positions in the ideological spectrum, then the binary indicator approach hides some valuable insights (that are in line with theoretical expectations) due to a loss of nuisance in the main outcome variable. Further, either manually coding whether there is a crowd in each image or finding an appropriate CNN to locate a dense crowd accurately in a picture would have demanded higher costs than using the BoVW framework (e.g., finding coders, time to label, discrepancies in coding, available CNNs for detection, monetary costs if using an API, etc.).

Table A.1: Binary vs. continuous measures of “crowds”

	Dense Crowd proportion	Presence of a crowd
	OLS	Logistic
	(1)	(2)
Center right	-0.114 (0.047)	-0.096 (0.528)
Center	-0.090 (0.035)	-0.784 (0.397)
Center left	-0.078 (0.035)	-0.734 (0.413)
Left	-0.106 (0.039)	-1.469 (0.482)
Not rated	-0.211 (0.081)	0.656 (0.927)
Constant	0.166 (0.209)	-16.470 (2,399.545)
Date	✓	✓
N	688	688
R ²	0.042	
Adjusted R ²	-0.007	
Log Likelihood		-301.910
AIC		671.820

Standard errors in parentheses. **Bolded** coefficients indicate statistical significance at conventional levels.

A.1.2 Object detection and BoVW: topicality vs. topics

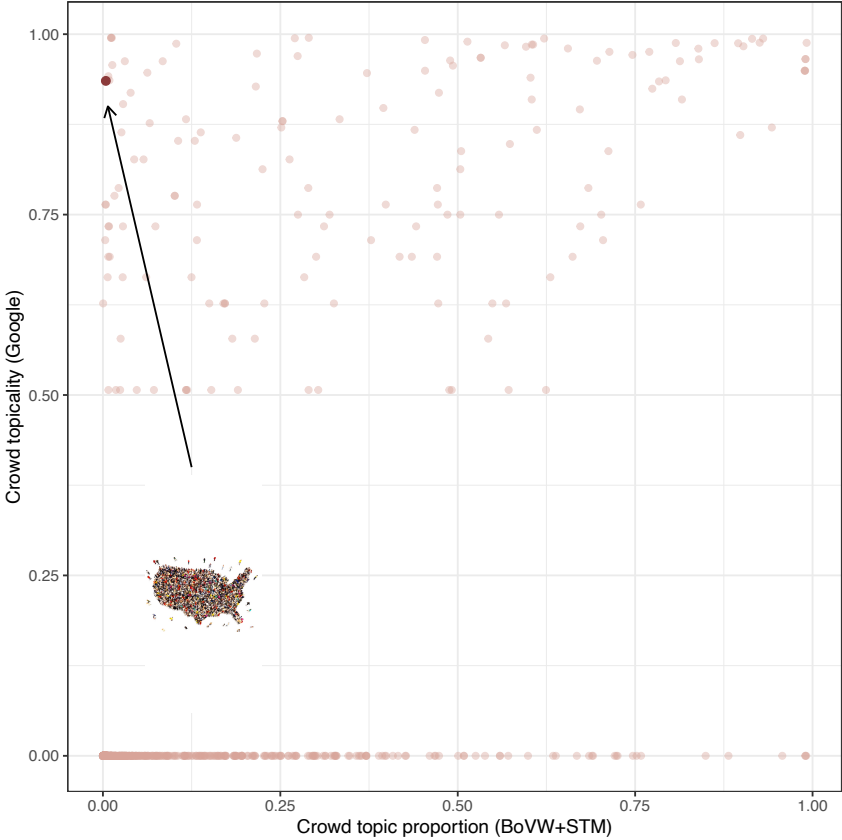
An alternative approach to the BoVW for measuring “proportions” of certain elements in a picture is to use CNNs and deep learning methods to detect all the objects in an image. In fact, [Karpathy and Fei-Fei \(2015\)](#) conduct this approach to describe content in an image using a CNN that detects objects in pictures. However, this approach requires the construction and coding of a very complex and customized network able to detect all potential objects in a picture. This might be necessary despite the existence of commonly used pre-trained networks. This is due to the type of pictures used to train those models (which generally include basic objects and colloquial scenes) that differ from those with political content.

Alternatively, researchers might want to use tools like the Google or Microsoft APIs with CNNs embedded for object detection. These platforms offer a wide variety of services for image analysis such as face and emotion detection, landmark identification, labeling, and even web search. Among these, the object detection feature from the [Google Vision API](#) provides a summary of the objects in a picture together with an indicator of “topicality”. However, there are two issues with this measure. First, although the name suggests that it captures the same thing as a topic proportion, the topicality of an object x is the probability that a CNN finds that object in the scene. Thus, the topicalities available do not sum to one or provide information about the proportion that the object occupies in the picture. Second, the topicalities are available for objects that are likely to appear in the picture; that is, objects with probability over 0.6. Thus, less salient objects are not included in the list

provided by the API, which could also be informative for the full analysis of the picture. A final minor issue is that using the API is not free, and although the cost of analyzing batches of images is not excessive, when dealing with high volumes the final cost is not negligible.

To compare the topics and the “topicality” of a crowd, I ran the images through the Google API and obtained this topicality measure. Then I compare that quantity (y -axis in Figure A.1) to the topic proportion of “dense crowd” from the STM (x -axis in Figure A.1). First, it is possible to see that although some pictures have a proportion of crowds as indicated by the STM, if this element is not too salient the topicality is not going to be high enough to be included in the list of objects describing the picture (i.e., there are many observations with $y = 0$ but $x > 0$). Further, in other cases, although the topicality is high (indicating a high likelihood of finding a crowd in the picture), the topic proportion might be appropriately low when we do not actually observe a crowd in the images.

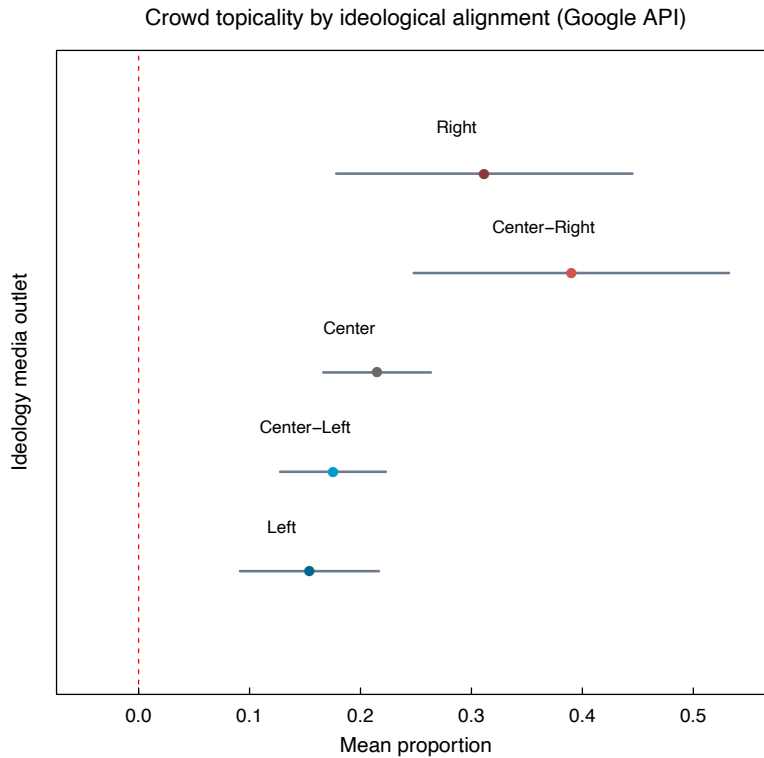
Figure A.1: Topicality (Google API) vs. Topics



What is also interesting is to observe what happens when instead of analyzing the mean proportion of topic crowd by ideological alignment, I use “crowd topicality” from the API. In Figure A.2 I present the results which differ substantially from those presented in Figure 12.

Finally, I would like to briefly discuss another method used by Kriegsmann et al. (2020). In their paper, these authors use a CNN to classify the type of tumor present in pictures of brains. As a validation exercise, they broke the images into pieces and used the

Figure A.2: Ideology and topicality



CNN to classify each of the pieces. Then they estimated the proportion of cancer types by picture, as a robustness check of the initial classification (i.e. higher proportions of type A pieces in an image are reassuring of the overall classification of it as type A).

This approach is extremely creative and can deliver quantities similar to the those suggested by this article in certain instances. For example, I can break the images of the caravan into segments, and use a CNN to detect if a crowd appears in each of them. Then, it is possible to aggregate the information at the picture level and estimate the proportion of pictures featuring a crowd. Just as with pieces belonging to brain tumors where a single patch showing a piece of the concept of interest resembles some of the full images used to train the CNN, the approach can perform optimally in similar instances. But in others this might not be the case. For example, if the topic of interest is “portraits” or “two-people meetings”, the pieces conforming a picture of two politicians shaking hands are not meaningful for the estimation of proportions: I could have a hand, an eye or a piece of a head that the CNN will not only have to correctly identify but also to aggregate to create meaningful concepts.

When the visual content is more heterogeneous and complex, this approach might not be adequate, or might require models that are able to recognize a very large number of objects and even characteristics or parts of them.

A.1.3 Using features from CNNs for clustering

Most of the classification of images using deep learning methods have focused on supervised tasks. However, there are tools that take advantage of the models that have already been extensively trained to learn about the features in the images to subsequently cluster them and identify groups. In particular, it is possible to use the weights (or “coefficients”) and feature maps from pre-trained models to extract features from a sample of interest as input for other clustering or grouping algorithms.

As I discussed in the main text, the features derived are not easily interpretable and thus cannot be coerced into a format that allows the use of mixed-member classification methods like STMs. However, these features can be used to feed regular clustering routines like k-means.

To compare the results from this method with the creation of topics through the visual STM presented in the current manuscript, I used a pre-trained CNN, a ResNet50, to extract features from the images of the Caravan. A ResNet50 has 48 convolutional layers and has been trained on a million images from the ImageNet database. In general, this CNN yields a single label (with the highest probability) out of 1,000 categories for each image under analysis. The model determines these labels after considering the weights and features of the images. For the unsupervised setting, the idea is to use those features without reaching the final classification step. Thus, I will extract the final vector of features to then use them as input for the clustering method of choice (in this case a regular k-means with 15 classes).

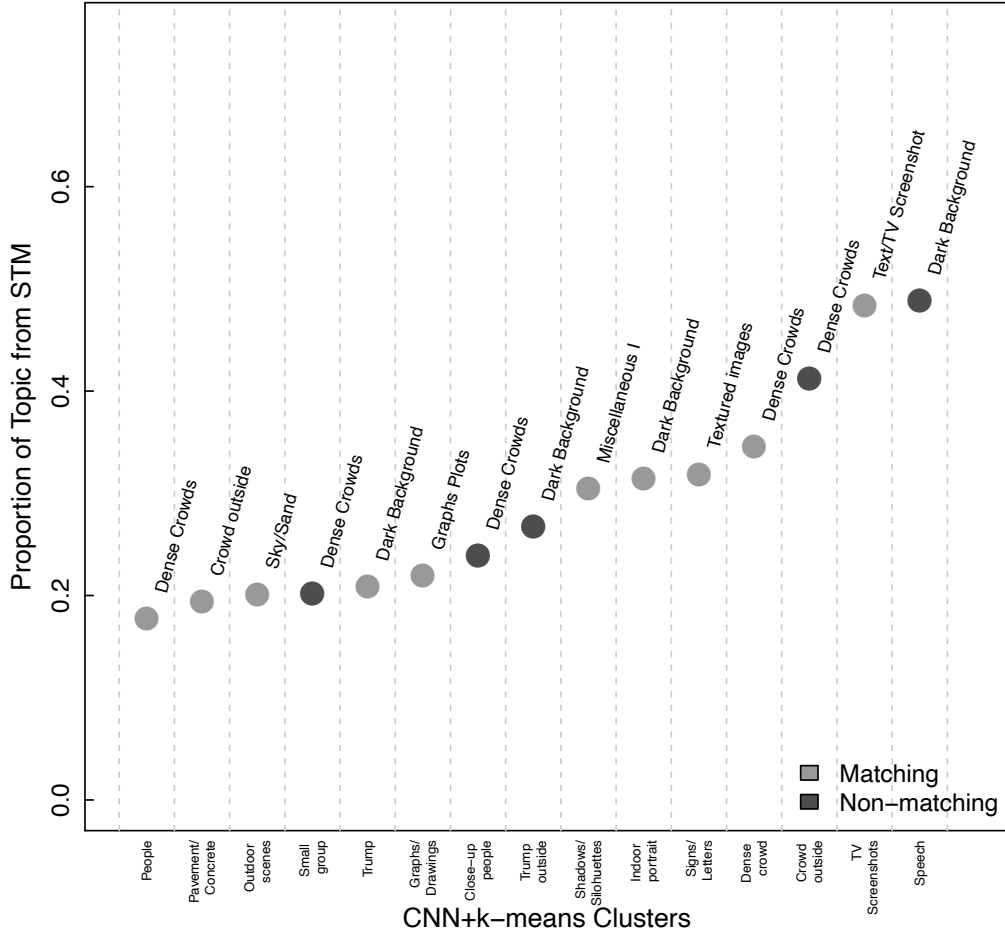
Although the output from this routine is the groups or clusters to which each image belongs, I can still inspect whether there is overlap with the topics from the STM, as well as the connection between the two. Figure [A.3](#) illustrates this relationship. For each cluster from the CNN+k-means procedure, cl , I took the means of all K topic proportions in the images belonging to that cluster. For example, for cluster 1, “Close-up of people”, I computed the mean proportion of each of the 15 topics of the STM. Then, I identified the topic with the highest proportion. Thus, for each cl in the x -axis, the gray point represents the proportion of the most common topic among the images from that cluster. For example, the most prevalent topic among the images belonging to the “People cluster” was the “Dense Crowds” (with a mean proportion of this topic of about 0.19).

Although, it is difficult to properly compare cluster membership to an actual proportion, the main takeaway from this plot is that there is a very decent correspondence between the topics identified by the visual STM and the content of the CNN clusters. Moreover, there are relevant dimensions, like “Border/Fence” that the BoVW identified but the CNN clustering did not. This could be as a result of the single membership process where other more salient features define the classification.

A.1.4 Other advantages: Fully unsupervised approach

The BoVW has a feature extraction step that is separate from the final classification. This means labeled data is not necessary. Thus, a natural application of this approach is for fully unsupervised analysis. Although CNNs are designed to work as supervised methods, researchers are exploiting their ability to learn complex features to conduct unsupervised analyses. Some approaches use popular pre-trained CNNs to learn features from unlabeled

Figure A.3: Clustering based on CNN features: comparison with STM output



data that are subsequently fed to unsupervised algorithms (Kielbaso and Bottou 2014; Zhang et al. 2018). Others use metadata, like text, to leverage the identification of patterns from CNNs (Gomez et al. 2017).

While these approaches might be useful for certain tasks, there is still a reliance on pre-trained models or sources that are based on labeled data or expected categories. This makes them unfit for other fully unsupervised and exploratory endeavors. Moreover, in cases where the pool of images and the object of study differ considerably from those used to train a given model, the classification results might be biased or inaccurate (Caron et al. 2019). In turn, the BoVW can be used in a fully unsupervised setting with only the unlabeled sample under analysis as input.

A.2 Technical details: detection and description of key points

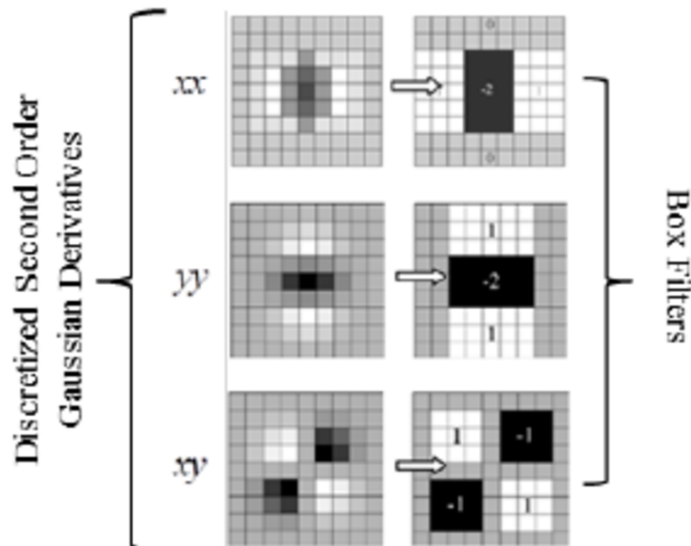
A.2.1 Key-point detection

The FAST Hessian is suitable detector for the purposes of this article given its two key properties: scale invariance (i.e. key points should be both repeatable and recognizable at different scales of the image), and high computational speed. To identify key points while preserving the scale invariance property, the FAST Hessian relies on the approximation of the Hessian matrix of a scale-space function, where space is measured by $\mathbf{x} = (x, y)$, and scale by σ . Let $I(x, y)$ be the intensity of the pixel located at coordinates (x, y) . Ideally, the process starts by calculating the second order partial derivatives of the image, by convoluting it with a second order scale normalized Gaussian kernel. Thus, the “ideal” Hessian matrix has the form:

$$\mathcal{H}(\mathbf{x}, \sigma) = \begin{bmatrix} L_{xx}(\mathbf{x}; \sigma) & L_{xy}(\mathbf{x}; \sigma) \\ L_{xy}(\mathbf{x}; \sigma) & L_{yy}(\mathbf{x}; \sigma) \end{bmatrix},$$

where, for example, $L_{xy}(\mathbf{x}; \sigma)$ is the convolution of the Gaussian second order derivative, $\frac{\partial^2 g(\sigma)}{\partial x^2}$, with the image I in point \mathbf{x} .¹ The determinant of the Hessian of each pixel will then be used to determine salient points. However, the estimation of this Hessian is computationally expensive, especially as the size of the kernel grows. Thus, Bay et al. (2006), proposed an approximation of the second derivative kernels by using “box filter” representations of those matrices. Figure A.4 illustrates the original and approximated filters.

Figure A.4: Original second order derivative Gaussian filters and approximations



These box filter approximations of L_{xx} , L_{xy} and L_{yy} , denoted as D_{xx} , D_{xy} and D_{yy} increase efficiency and speed considerably, and allow us to estimate the determinant of the

¹Where $g(\sigma)$ is the pdf of a normal distribution with $\mu = 0$ and standard deviation σ .

approximated Hessian as follows:

$$\det(\mathcal{H}_{approx}) = D_{xx}D_{yy} - (0.9D_{xy})^2$$

To detect key points, I will build layers of the image by using increasing sizes of kernels as a way of varying the scale of the original picture (for example, the smallest kernels possible of size 9×9 , will correspond to a real valued Gaussian with $\sigma = 1.2$). Once this scale-space 3D structure is built, a maximal suppression is performed to find the salient points. In other words, a pixel is considered a key point if its intensity is higher than the one of its 26 neighbors, comprised in the $3 \times 3 \times 3$ cube that surrounds it: 8 along the x and y axis plane, and 9 across scale layers. The final step involves interpolation of the data surrounding the key points in order to reach sub-pixel accuracy. Figure 2 in the main text shows an example of the key points that are found in one of the images in my sample. The green circles represent the coordinates of the key points. The figure illustrates how most of the key points are representing edges, corners or regions where color changes significantly. Once the key points are identified, as in the case of this image, I proceed to extract its features.

In this article I use the RootSIFT descriptor. This is an extension of one of the most popular descriptors in computer science: the Scale Invariant Feature Transform (Lowe 1999, SIFT) which has the advantage of being invariant to image translation, scaling, rotation, and even partially invariant to illumination changes. The RootSIFT was developed by Arandjelović and Zisserman (2012) who added two extra steps to the regular SIFT implementation to drastically improve accuracy: a L1-normalization of the SIFT vectors, and the calculation of the square root of the elements in each of those normalized vectors.

A.3 Building a vocabulary: sample of features and clustering

The BoVW requires the researcher to select and cluster a sample of the total features identified in the images under analysis. Why does the process involve lumping together a subset of the features instead of using the full set? Suppose that a sample of interest contains images of dogs, flowers, and humans and that researchers are interested in classifying this pool according to the actor that each observation depicts. For simplification purposes, imagine that after completing the steps above a finding is that one common neighborhood across human photos is (unsurprisingly) a human nose. However, although similar, it is extremely hard to find two identical noses; even two pictures of the same person would look different due to lighting, position, angles, etc. Therefore, the *average* of those noses is necessary to accurately represent a general concept of a nose. Thus, we can cluster the features associated with the nose and take the feature vector of the centroid as the representation of our “visual word”.

To achieve higher levels of speed and efficiency I form this vocabulary based on a random sample of the feature vectors. In general, taking 10-30% of the feature vectors is accepted as a common practice and it avoids having clusters with extremely similar features from the same image. However, this number will depend on speed necessities and size of the data. Given the complexity of the images under analysis, especially in comparison to more standard canonical datasets, for all the models in this article, I sampled 30% of the features.

A.4 Emulating the Document-Term matrix: the Image-Visual word Matrix

The Image-Visual Word matrix (IVWM) emulates the Document-Term matrix (DTM) in text analysis. Their underlying logic and structure is similar: the units of analysis are in rows, while each column has an element contained in the full sample. A cell in row i and column j indicates the number of times that the element in column j appears in observation i . This can be a count or proportion, either weighted or unweighted.







In the case of a DTM, each row represents a text under analysis, while the columns are generally words, word stems, sentences, n -grams, etc. that appear in the full pool of texts. In the IVWM, the rows are images, and the columns are visual words. Figure [A.5](#) illustrates both.

Figure A.5: DTM and IVWM

(a) Document-Term Matrix

Document/Term	President	elections	...	migrants	troops	Central
President Donald Trump has focused heavily on issues related to immigration in the run-up to the midterm elections, warning of an "invasion" of Central American migrants, and sending thousands of troops to the border.	1	1	...	1	1	1
President Donald Trump is trying to frame the upcoming midterm elections as a national referendum on immigration issues. The President complains that Mexico is not doing enough to stop the caravan of migrants.	2	1	...	1	0	0
Thousands of Central American migrants have again resumed their trek through southern Mexico after failing find buses to carry them. President Donald Trump said Wednesday that the deployment of active troops to the southern U.S. border could increase dramatically.	1	1	...	1	1	1

(b) Image-Visual Word matrix

Image/Visual Word			...	
	20	0	...	3
	0	7	...	5
	12	9	...	0

A.5 Extension and Supplementary Information of Application (Framing a Movement)

The data used in this article was collected using the `News API` and manually curated by the author. The `News API` is a tool that allows users to retrieve information of events and news from more than 30,000 sources worldwide. I limited the search to sources in the U.S. The reports and news are extracted from websites of several prominent outlets such as ABC, Politico, The New York Times, Fox News, Huffington Post, etc. The metadata includes date, author, image, headline, the truncated text of the article, original length of the article, and its URL.

The data was used to feed a structural topic model with 15 topics and three prevalence covariates: news outlet, date, and ideology of the news outlet. The scores for ideology are provided by *All Sides* and are based on surveys asking respondents about their own bias and how they rate the bias of news sites. Then, this information and the aggregation of the rankings by ideological group and news outlet are used to determine the average bias rating of a source. Robertson et al. (2018) show that *All Sides* scores have a strong correlation with other validated measures of media bias. The data for this article include 424 articles published by 33 sites with different ideological groups: left leaning (center-left and left, n=16), center (n=10), and right leaning (center-right and right, n=5). The other two have not been rated. Examples of outlets in the “Right” category include *Breitbart*, *Fox News* and the *Washington Times* whereas the “Left” includes outlets such as the *Huffington Post*, *Politico*, and *MSNBC*. The “Center” covers outlets like *Bloomberg*, *CNBC*, and *USA Today*. For more information regarding the distribution of number of articles per ideological group, see the Appendix.

A.5.1 Descriptive Statistics: BoVW

Table A.2: Summary of number of words in the IVWM

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Number of visual words (per image)	688	233.960	130.270	16	132	310.500	816
Mentions of visual words	2,000	80.480	38.700	0	54	101	452

A.5.2 STM Results (cont.)

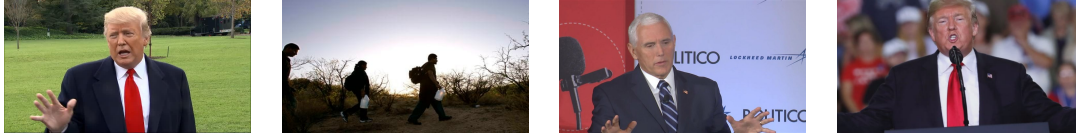
The most representative images per topic, and most frequent and exclusive words for topics 7-15 are presented below in Figure [A.6](#) and Figure [A.7](#).

Although most of the topics are sensible and meaningful for the study of visual framing, I also find a few that are less relevant than the others. For example, there is a topic whose most representative images are pictures with banners and ribbons that typically appear at the bottom of news shows. Although this “TV screenshots” topic is not a politically

relevant dimension for the study of portrayals of immigration, it makes sense from the computational viewpoint: there are elements of pictures with high proportions of this frame such as figures with text at the bottom that make them look similar.

Figure A.6: Most representative images per topic

Miscellaneous I: Indoor Portraits



Miscellaneous II: People with fuzzy backgrounds



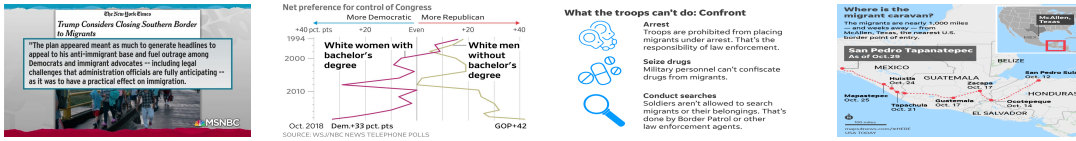
Text/TV Screenshots



Crowds Outside



Infographics



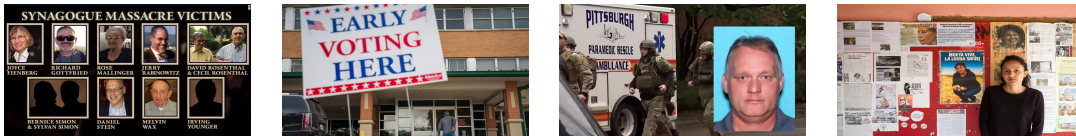
Small group



Images with texture



Rectangular shapes



Graphs and plots



Figure A.7: FREX Visual Words per Topic

Miscellaneous I: Indoor Portraits



Miscellaneous II: People with fuzzy backgrounds



Text/TV Screenshots



Crowds Outside



Infographics



Small group



Images with texture



Rectangular shapes



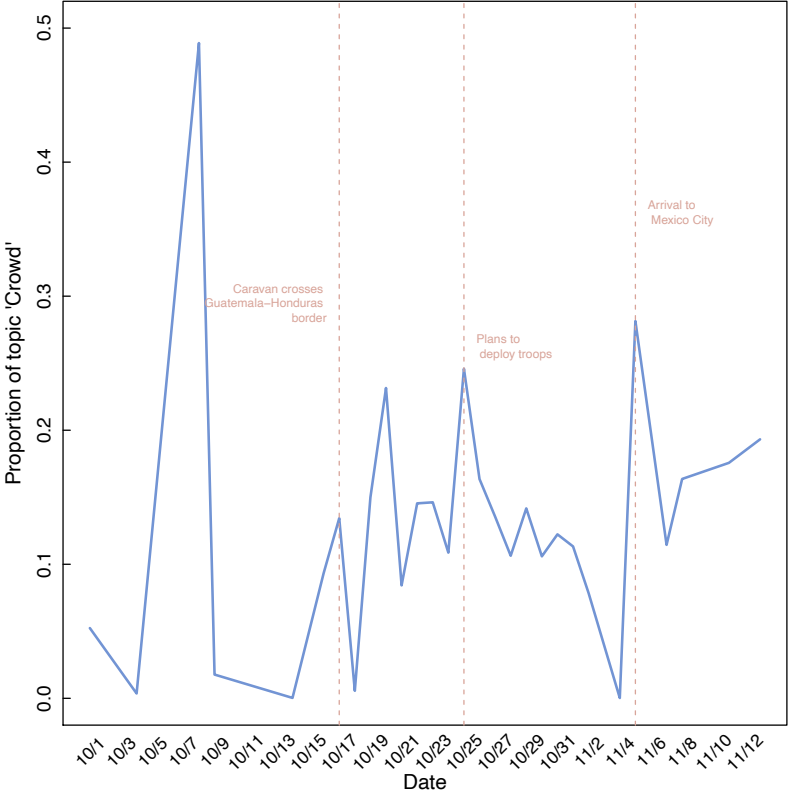
Graphs and plots



Note: The numbers of the topics in the replication file are 2, 3, 5, 7, 8, 10, 12, 14 & 15.

I analyze the use of the topic “crowd” over time to identify whether there is any variation in the coverage of the magnitude dimension of the caravan. Figure A.8 shows on the y -axis the proportion of topic “dense crowd” in the full corpus of images at different points in time (x -axis). The red dashed lines show relevant events that received wide coverage in the U.S. media, such as the arrival of the caravan to Mexico City. It is interesting to notice that these events correspond to peaks in the dataset, suggesting a stronger focus on the size and magnitude of the caravan when its salience in the media market is higher.

Figure A.8: Use of topic “crowd” over time (2018)



Note: The line shows the trend of the topic “crowd” from October to November of 2018. The gaps between points indicate that there was no coverage in that period. The dashed lines indicate important dates in the time line of the migrant caravan coverage and development.

A.6 Structural Topic Models and Latent Dirichlet Allocation

In this article, I present an application of the framework using a Structural Topic Model for the identification of “frames” in images. Although the method of constructing the Bag of Visual Words as an input of several analyses is the main focus of this article, in this section I offer a discussion of the suitability of STMs for the particular substantive application that I present in the main text.

First, to study the way in which media uses visuals to frame stories, we must understand the type and prevalence of the themes that these visuals contain. Just as in text, topics are a natural way to capture these themes. Further, given that I am interested in how the ideological leaning of a media outlet affects the particular distribution of topics in each image, a topic modeling tool seems appropriate for the endeavor.

Notice that the topic modeling approach outputs *proportions* of different topics in each image (a continuous measure bounded between 0 and 1 for each of the k topics), rather than a global set of descriptors of whether a given object or theme k *appears* in an image. Thus the proportion of images featuring a crowd that an outlet j publishes is not going to be equal or comparable to the mean proportion that a crowd occupies in the pictures that j publishes.

Although LDA provides those proportions, STM is more suitable for testing whether ideology affects the proportion of topic crowd in an image for the following reasons. First, LDA assumes that topics within a document are independent of one another, an assumption that is hard to fulfill in most of the applications in social sciences. STM weakens this assumption and allows for independence conditional on covariates. Second, LDA implies that a topic k for document i is defined by the same words as this topic in other documents $-i$. Once again, this is a stringent assumption that STM does not require and that fits our purposes given that the theoretical motivation is that images might portray the same visual topic (e.g., water) but by using different words (e.g., river, lake). Finally, LDA states that topics can be modeled entirely based on the text of the document. This will imply that the frames that we see in the images are only a function of the visual words that we created, and are not affected by other factors (including the characteristics of whoever produced the picture, time of the events, date, location, etc.). The theoretical motivation and expectations of the research exercise is to indicate that metadata, such as date and, in particular, ideological stands of the news outlets producing and selecting the images, influence the words and the topics that we identify in them. Thus, a STM provides the set up for including this information and testing their effect in the generation of topics. This is a feature that also captures the hypothesis *Ideology* \rightarrow *VisualFrames* more appropriately.

Beyond these arguments, the STM set up allows me to illustrate the benefits of the BoVW as a dimension reduction technique for images given that social scientists use the tool often and are familiar with their structure, implementation, and results. As of August 10, 2022, Google Scholar reports that the three most prominent pieces related to STMs (Roberts, Stewart, and Tingley 2019, Roberts et al. 2014, and Roberts, Stewart and Airolidi 2016) have collected 2,817 citations. Further, the use of STM extends and reinforces the main contribution of the article regarding the “translation” of tools for text to images.

However, there are concerns regarding the potential sensitivity of the topic distribution to the prevalence covariates in the structural model setting that the researcher selects.

To ameliorate this, I conduct a regular Latent Dirichlet Allocation (LDA) analysis and a subsequent regression of the identified “crowd” proportions on the ideological leaning of the media outlets. The topics, most frequent words, and most representative images are very similar to those from the STM.

The table below shows two columns with the results from regressing two outcome variables on ideological slant: the first column, the proportions of topic crowd per image from the STM model with prevalence covariates, and the second column, the proportions of topic crowd from a vanilla LDA model. The results are remarkably similar and lead to the same substantive conclusions. This is one case where the decision between LDA or STM does not affect the implications of a study, but the decision to use one over another was impacted by underlying assumptions and a specific theoretical framework.

Table A.3: Comparison of results STM vs LDA: Relationship between outlet’s ideology and proportion of topic crowd in images

	STM	LDA
	(1)	(2)
Center right	-0.114*	-0.114*
	(0.047)	(0.047)
Center	-0.090*	-0.091*
	(0.035)	(0.035)
Center left	-0.078*	-0.079*
	(0.035)	(0.036)
Left	-0.106*	-0.106*
	(0.039)	(0.039)
Not rated	-0.211*	-0.211*
	(0.081)	(0.081)
Constant	0.166	0.169
	(0.209)	(0.210)
Date	Yes	Yes
N	688	688
R ²	0.042	0.042
Adjusted R ²	-0.007	-0.007
F Statistic (df = 33; 654)	0.858	0.862

Bolded coefficients indicate $p < 0.05$

A.7 Topic modeling – label validation

Looking at both the most representative images and the most frequent visual words help with the labeling of the topics. This, however, is not a trivial task and is not immune to the user’s biases and subjective considerations. Thus, a labeling strategy is to expose human coders to several representative images of each topic and ask them to provide words describing the content of each image. After that, they can assess a cluster of images belonging to a particular topic and give a label to it. The results from this process for labeling the topics of the application are in Table A.4 and Figure A.9 below.

Four coders (three graduate students and one professor) completed those two labeling tasks. I determined the final label of each topic based on the responses given by the coders, as well as my own qualitative inspection of both the most representative images and visual words per topic. I noticed that offering the visual words had an impact on the coders’ answers and gave some qualifiers to their original responses. Instead of saying “people talking”, the visual words helped to complement the answers with other information “people talking in an indoor/dark setting”.

Table A.4: Top words and labels per topic

Topic	Most frequent words	Labels by coders	Final label
1	wall, people, fence, vertical, ground, bars	border, fence, graph, speech	Border
2	people, walking, plants, pavement, road, shoes	caravan, walking, America, kids	Groups walking
3	sky, mosaic, people, map, cloud, pavement	control, sky, cloud, Graphs/Maps and Landscapes	Sky/Sand
4	people, plants, walking, water, night, person	domestic, walking, Contrast images (Dark/Light), interview	Small groups/Individuals
5	dark, person, trump, background, tie, face	politics, Trump, press conference, Clinton	Portrait with dark background
6	people, crowd, plants, rally, trump, walking	sides, crowd, packed, rally	Dense crowd

Figure A.9 show the word clouds of all the responses given by coders (weighted by frequency).

This is only one way of interpreting the labels and results from the visual STM. However, there are other methods like the one proposed by Ying, Montgomery, and Stewart (Forthcoming) that take advantage of crowd sourcing for the labeling and validation of topics.

A.8 Diagnosis and practical guidance – the impact of parameters on topic discovery and estimation

In this section, I present an analysis of the impact that several decisions throughout the process of the BoVW have on 1) topic quality, and 2) the estimates of the effect that prevalence covariates, like news outlet’s ideology, have on the generation of visual frames. For the former, I use statistics like semantic coherence, exclusivity and residuals as Roberts, Steward, and Tingley (2014) suggest². For the latter, I estimate the differences in mean “dense crowd” proportions between right leaning outlets and left leaning outlets. Recall that Figure 12 in the main text shows a positive and reliable difference between these two types of outlets indicating that right leaning outlets tend to publish higher proportions of the “dense crowds” frame in the articles they publish about the migrant caravan.

The areas I explore and compare are a) number of key points detected in an image, b) number of visual words in the vocabulary, c) number of topics in the visual STM, and d) labeling the final topics.

A.8.1 Detecting key points

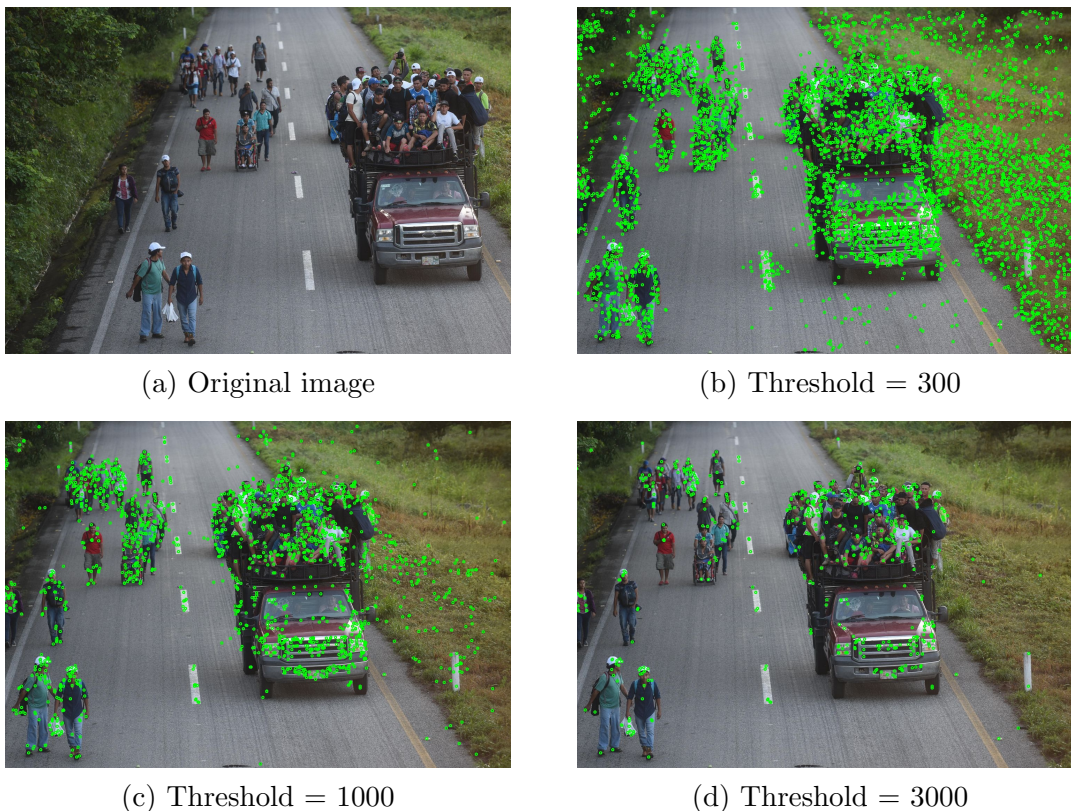
There are two important things to consider during the key point detection: 1) the type of salient regions to identify, and 2) the precision of this identification. The first one relates to the definition of what constitutes a salient region. For example, certain applications require an accurate identification of “corners” in an image (e.g., identification of buildings and houses, captured with rectangular shapes). Others, however, rely more heavily on “blobs” (e.g., classification of texture). For more complex content, as in the case of most social sciences applications, the detection of all edges, blobs, and corners is in general more fitting (Olaode, Naghdy, and Todd 2014). The final definition of a salient region determines the type of detector to use. For example, the FAST and GTTT detectors are used to detect corners in images, while DoG and FAST Hessian focus on the detection of corners, edges, and the combination of both.

The second feature, the precision, will have an impact on the number of key points that the detector identifies in each image. The decision should consider what features are substantively relevant to the objective of the study. For example, the FAST Hessian with a low threshold captures even small changes in pixel intensity and therefore yields a large number of key points. A visual inspection of a small sample of images and the key points detected in each of them using different thresholds is helpful to address this issue. Consider Figure A.10 where a low threshold of 300 yields a very large number of key points detecting finer features such as lines on the pavement. In contrast, a higher threshold of 3000 leads to the identification of more prominent features like the people and cars in the picture.

The Hessian threshold is a parameter that regulates the accuracy of the key point detection. In other words, only features whose Hessian threshold is larger than the defined value are retained by the SURF detector. This means that the larger the threshold, the lower the number of key points identified. Figure A.10 in the main text illustrates the points that a SURF detector identifies when this threshold changes. It is clear from these pictures that a low threshold of 300 yields a high number of points detected. Substantively this means that

²Held-out likelihood should be used for model selection but I will not include it for comparison purposes

Figure A.10: Comparison of key point detection outputs with different thresholds



Credit, original image: Johan Ordonez (AFP/Getty Images).

the detector considers even minor changes in pixel intensity as salient. In contrast, a detector with a large threshold focuses on the most prominent pixel intensity changes and therefore retains only the most salient key points in an image. The decision to have a higher or lower threshold depends on the research objectives of the user as well as on the nature of the images under study. For example, images in canonical computer vision datasets like CALTECH 101 tend to depict one or a few close-up objects with low complexity. In this case, a threshold between 300 and 500 would retain an adequate number of features to make subsequent classifications. However, as Figure [A.10](#) shows, the complexity in the composition of the images under analysis is high, which leads to a very high number of key points identified when using a low threshold. Several of these key points do not contribute with meaningful information about the image (e.g., points along the lines on the pavement). However, a high threshold misses a few key points that provide information about the environment and set up of the event depicted in the picture. Thus, a number in the middle is an adequate option.

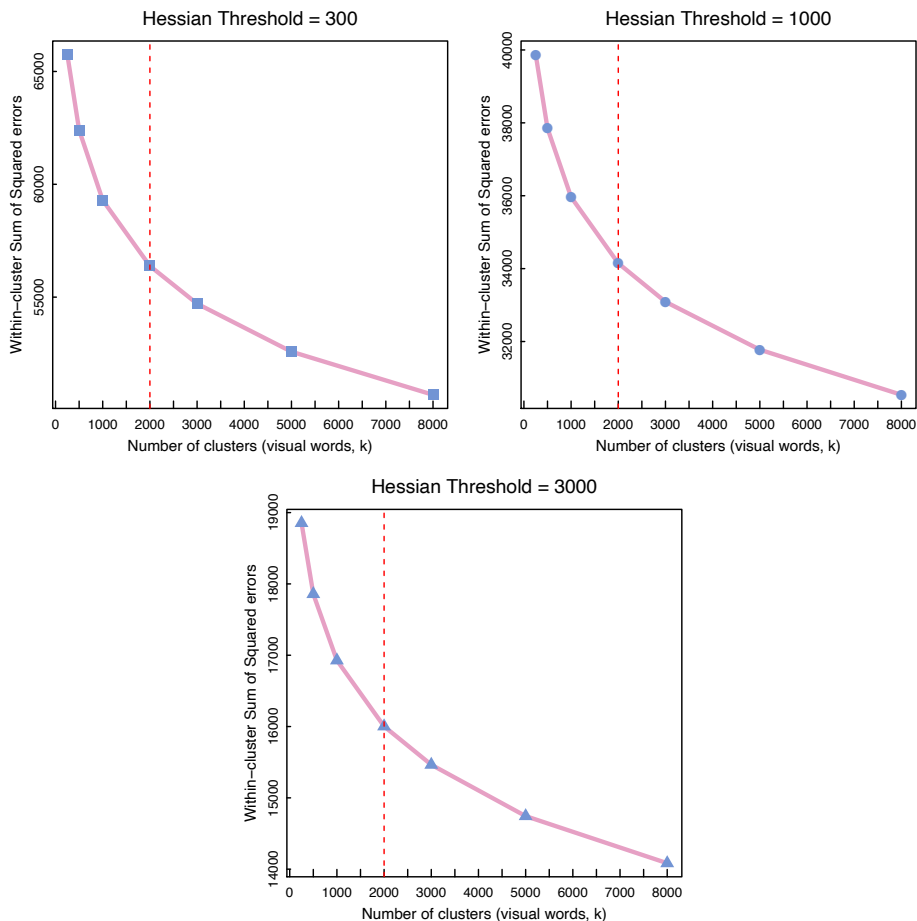
How would the number of key points impact the construction and application of the BoVW? In a nutshell, more key points represent more information about the picture. This extra information is represented by a larger number of features that especially impact the clustering process. First, the process is more computationally expensive due to the higher volume of features. Second, for a fixed number of clusters (visual words) v , the within-cluster sum of squared errors (WSSE) will be larger as the number of key points increases. Thus, if the objective is to achieve low WSSE while also keeping a low number of visual words, a

lower number of key points is preferred.

Consider Figure A.11. For the same number of clusters v (on the x -axis), the WSSE is the lowest when using a threshold of 3,000 (around 19,000), and the highest with one of 300 (around 66,000). However, it is interesting to notice that while in absolute terms, as expected, the model in the bottom panel minimizes the WSSE, all models show a similar trend in the reduction of WSSE as the number of clusters increases. From the plots and using the “elbow” method, the chosen number of clusters would be 2,000 in all cases.

The effects on topic quality are illustrated in Figure A.13, where the different colors correspond to the different thresholds. Although the differences in exclusivity between the models are not stark, it is worth highlighting that this indicator is higher when there are more key points. For semantic coherence, the differences are larger even when comparing models with the same number of visual words. In this case, those with lower thresholds show more semantic coherence. Given that more key points retain more information about the picture, it is not surprising to find that the topics are better described when using additional features. However, this is also conditional on increasing the number of visual words in the vocabulary to preserve the cohesiveness of the clusters identified.

Figure A.11: Comparison of WSSE by Hessian threshold



A.8.2 Building a vocabulary

For the visual vocabulary, the first consideration based on the particular research question should be from what images it will be constructed: from a *corpus of reference* or the *corpus of interest*. If the former, the next choice is to determine the type of images that will form it. As explained above, each of the feature vectors in an image is associated with a visual word. Thus, it could be the case that a given feature is linked to a visual word that does not properly represent it if there are no better candidates.³ Therefore, it is crucial to build a vocabulary with images relevant to the target pool under study.

Another consideration is the number of clusters or “visual words” to extract. A richer vocabulary has more power to discriminate and distinguish features, but a more parsimonious one focuses less on the details that each visual word is capturing. The decision to have a more fine-grained vocabulary depends on the substantive motivation of a project: is it relevant to keep features of a given object with light and dark backgrounds as separate visual words? If only the object itself is a relevant component of the visual theme then it might be adequate to keep them together. If instead the background is relevant as a time-space indicator, then the separation is more useful.

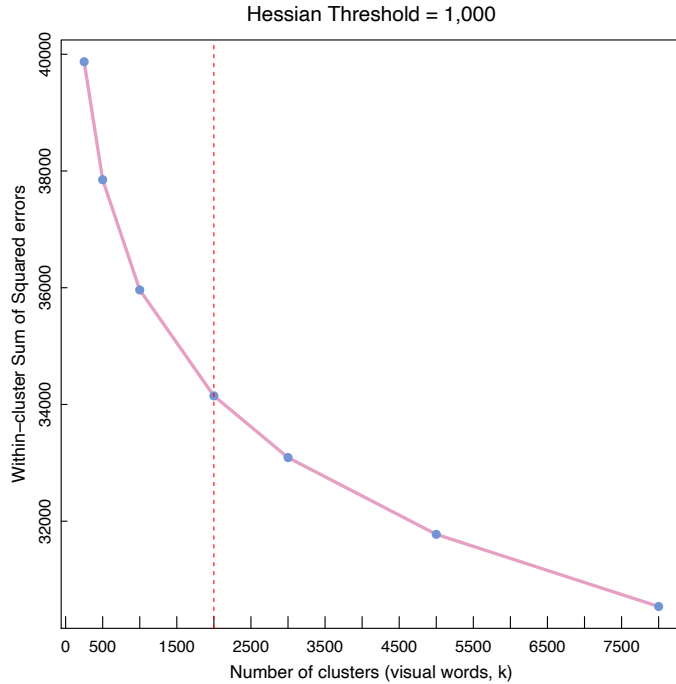
Researchers can also rely on commonly used tests providing an “optimal” number of clusters. Figure A.12 illustrates the “elbow” method. The x -axis of the figure shows the number of potential clusters/visual words, and the y -axis the within-cluster sum of squared errors (WSSE). A lower WSSE is desirable suggesting more cohesiveness of a cluster (i.e. smaller distances between the features and the cluster centroid). The elbow method consists of determining a “break point” or “elbow” in the plot where the WSSE decreases sharply and starts flattening afterwards. In Figure A.12 we observe a strong jump between 1000 and 2000 visual words but a less intense change from 2000 to 3000. Thus, this test informs our decision to select an appropriate number of visual words in our vocabulary (in this case, 2000 visual words).

Using the elbow method, I find that the number of clusters have an effect on the discovery of topics. While there are topics and frames that can be clearly identified regardless of the parameter definition, even in small models like a STM with 5 topics, there are still differences in the composition of the topics across the different vocabularies. In this particular case, while topics like “dense crowd”, “sky” or “dark background” were constant across the three visual vocabulary categories, the model with the smallest vocabulary yielded a topic whose distinctive features corresponded to light backgrounds and “sand” like texture. In contrast, the longer vocabularies containing pieces with text and drawings on similar backgrounds contributed to the formation of a topic with “maps and infographics” rather than just “light backgrounds”. When the number of key points and visual words increase, the information retained from the pictures also increase and lead to a finer distinction between pictures. This might be something desirable depending on the characteristics of the study and the research objectives of the user.

Here I analyze the impact that this number might have on topic quality as well as on the results from effect estimation of prevalence covariates in topic models. Figure A.13 shows the result for the former. With respect to exclusivity, the lowest levels come from the

³Although this could be improved by using “acceptance thresholds,” it is still advisable to build sensible and conceptually coherent vocabularies.

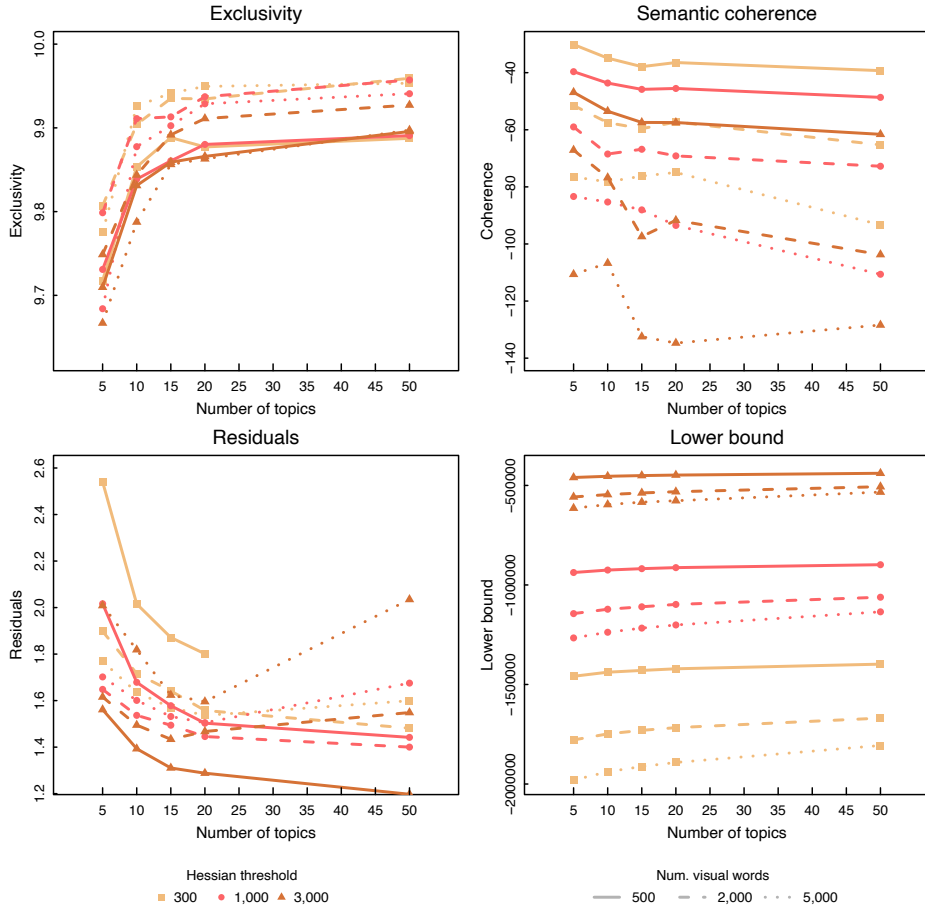
Figure A.12: Within-cluster sum of squared errors by number of clusters



models with the smallest number of visual words. The relationship between this parameter and exclusivity does not seem to be linear. While the simpler vocabulary has the lowest levels of exclusivity, the models with 2,000 words (middle category) evidence the highest levels of exclusivity. In contrast, the semantic coherence seems to be the highest among models with the shortest vocabularies, and it decreases as the number of visual words grows. This is in line with the intuition that more words, and the added complexity that these represent, make topics less cohesive given the increased number of components to consider. However, it is important to highlight that these statistics also interact with other parameters such as key point detection. For example, the differences in semantic coherence are not large between models with 500 visual words and Hessian thresholds of 300 or 1,000. However, this indicator decreases sharply when the Hessian threshold is 3,000. The lowest levels of semantic coherence are found in models with low number of key points but high levels of clustering. Overall, these findings highlight the importance of running multiple diagnosis exercises and comparisons between parameters to understand the particular effects of certain choices on indicators like the ones discussed in this section.

With respect to the impact that the number of clusters has on the estimation of an effect of a prevalence covariate on a topic of interest, the results are remarkably similar within number of topic and Hessian threshold groups. Figure [A.14](#) shows three plots corresponding to different Hessian thresholds. The lowest with 300 contains the highest number of key points while the one with 3,000 has the lowest. The rows in each plot correspond to a STM initialized with k topics. The estimates with their respective confidence intervals correspond to the difference in “dense crowd” proportions used in images published by right leaning and left leaning outlets. While there are strong differences between the results of models

Figure A.13: Quality indicators across BoVW designs

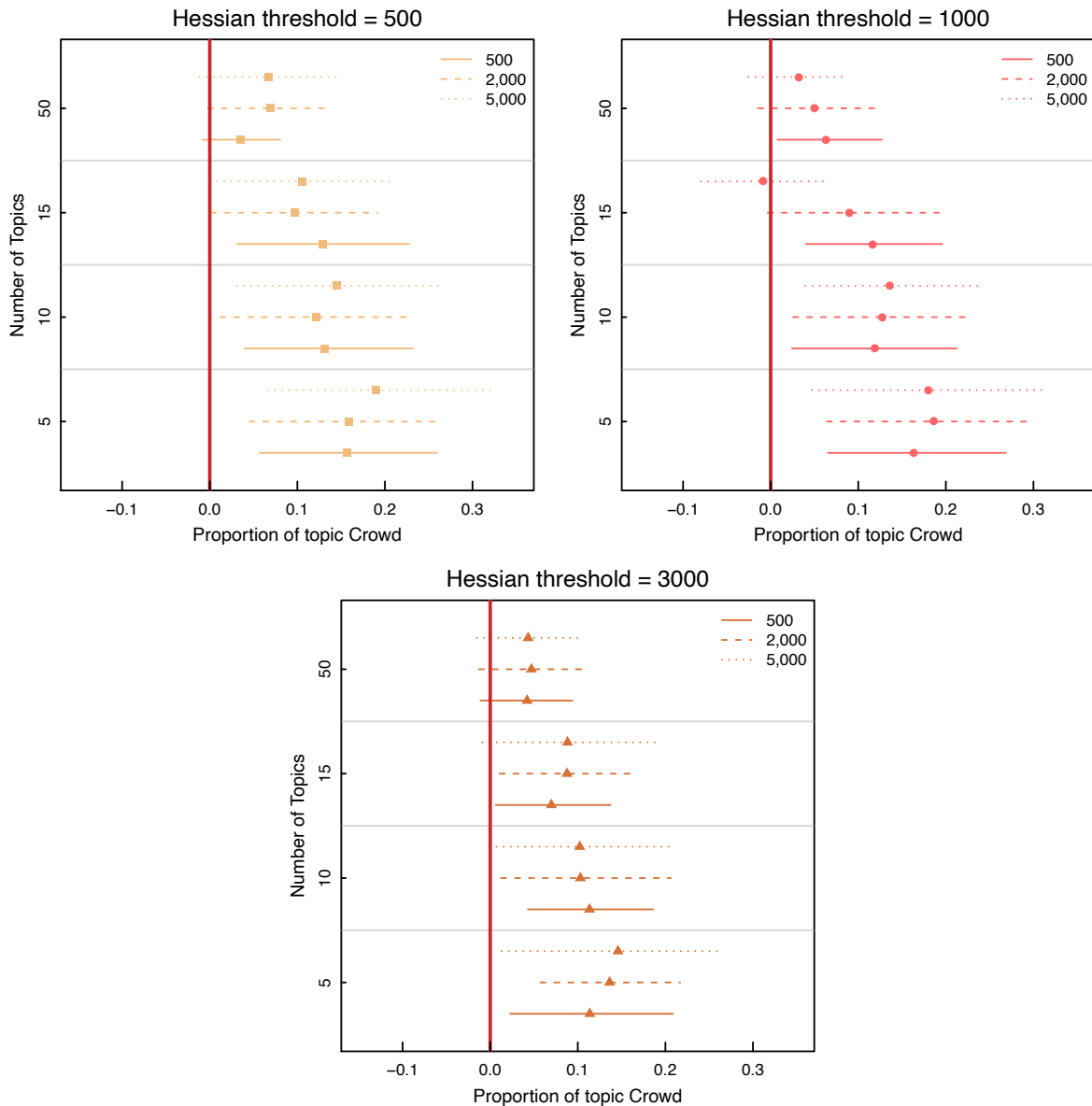


with high and low numbers of topics, within each of these categories the differences between number of visual words are very small. In a few cases, the results and their implications are different depending on the length of the vocabulary (based on statistical significance), but for the most part the results and findings are the same across categories. This is also due to the distinctiveness of the topic under analysis. “Dense crowd” was a topic that was clearly identified in all the topic models regardless of the specification of the parameters. However, these results might differ in the cases where the parameters substantially affect the topic structure or composition. Further investigation regarding these effects is required.

A.8.3 Determining the number of visual frames

Another parameter to consider is the number of topics to extract from the visual STM. Here, I follow the practices recommended by Roberts, Stewart, and Tingley (2014) involving the evaluation of aspects like semantic coherence and exclusivity. Using the functions embedded in the STM package paired with the functions provided in the replication code of this article, researchers can obtain average measures of mean held-out likelihood, semantic coherence, exclusivity, and residuals from models with a varying number of topics. Visual inspection

Figure A.14: Comparison of “ideology” effects by number of topics, vocabulary, and Hessian threshold



of plots with this information such as those in Figure [A.13](#) allows researchers to select an “optimal” number of topics that preserve parsimony, and maximizes semantic coherence, held-out likelihood and exclusivity. Researchers can take these statistics as guidance for their decision but should also complement them with considerations of their research needs and qualitative inspection of the topics using the FREX visual words and most representative topics.

To inspect the effect of this parameter, I implemented a series of topic models varying the number of topics, and across different BoVW specifications with respect to number of key points detected (Hessian threshold) and number of visual words in the vocabulary. From

visual inspection of the topic composition, there are substantive differences between the models. While smaller models with 5 topics cluster many different pictures into a “crowd” topic containing visual words with small human bodies and granular textures, increasing the number of topics allows for a more careful differentiation of topics like “crowd”. For example, setting the STM to 15 topics allows for the identification of different types of crowds: “groups walking” that not only contain visual words of people but also of pavement, or “outside crowds” where the focus is more on environmental factors like the sky than the people. The differences between these frames and the role they play in the research projects should help with the user’s decision regarding the selection of topics. Other methods aimed to assess topic quality, such as semantic coherence, held-out likelihood, and exclusivity are also suggested as part of the STM routine to evaluate the optimal number of topics to keep as illustrated in Figure [A.13](#).

Finally, beyond the differences that exist in topic quality and composition, it is also important to assess how the number of topics affects some of the potential inferences that researchers can make about the effect of prevalence covariates on the generation of visual frames. In this case, if I focus on whether the ideology of the news outlet (right vs. left) has an effect on the proportion of topic “dense crowds” in the pictures that such outlets use, and compare these differences between topic models, I get some interesting results. First, although the results are different between model set ups, in most of them the substantive finding that right leaning outlets publish pictures with higher proportions of dense crowds than left leaning outlets does not change. Second, the effect size decreases as the number of topics increases. This is a natural consequence of increasing the number of topics: the proportions of a single topic go down as the possibility of containing other similar but new themes arises. Thus, finding a difference distinguishable from zero as the proportions of a given topic become smaller gets more complicated as the number of topics increase. In Figure [A.14](#), it is possible to observe that although there is still a positive and reliable difference between right and left outlets, this differences becomes indistinguishable from zero when there are 50 or more topics. This is solely based on the estimation of the proportion “dense crowd” topic and does not consider summing up the proportions of topics associated with “crowds”.

A.9 Credits for images used in main article and appendix

A.9.1 Figure 9

Border/Fence Gregory Bull (AP)	Getty Images	Mark (AFP/Getty ages)	Ralston Im-	Gregory Bull (AP)
--	--------------	-----------------------------	----------------	-------------------

People walking AFP	Gary (Getty Images)	Williams	Rebecca (AP)	Blackwell	Sandra Cuffe (Al Jazeera)
------------------------------	------------------------	----------	-----------------	-----------	------------------------------

Sky/Sand Moises Castillo (AP)	Ross (AP)	D. Franklin	Mauro (AP)	Bucarello	Department of De- fense
---	--------------	-------------	---------------	-----------	----------------------------

Small groups/Individuals Michelle Frankfurter	Pedro Pardo (AP)	Evan Vucci (AP)	USA Today
---	------------------	-----------------	-----------

Portrait with dark background CNN	Department of De- fense	AFP	Getty Images
---	----------------------------	-----	--------------

Dense crowd Ueslei Marcelino (Reuters)	Pedro Pardo (AP)	Pedro Pardo (AP)	Mandel (AFP/Getty ages)	Ngan Im-
---	------------------	------------------	-------------------------------	-------------

A.9.2 Figure A.6

Miscellaneous I: Indoor Portraits			
CNN	Yair Oded	NBC	CNBC

Miscellaneous II: People with fuzzy backgrounds			
AP	Bram Janssen (AP)	CNN	AP

Text/TV Screenshots			
HBO	Fox News	MSNBC	MSNBC

Crowds Outside			
M. Castillo	Nick Oza	CNBC	David McNew (Getty Images)

Infographics			
MSNBC	Wall Street Journal/NBC News	USA Today	USA Today

Small group			
Sandra Cuffe (Al Jazeera)	Rebecca Blackwell (AP)	CNN	Rebecca Blackwell (AP)

Images with texture			
AFP/Newsweek	Getty Images	CBS/AP	Ryan F. Smith (AFP/Getty Images)

Rectangular shapes			
CNN	Drew Angerer (Getty Images)	NBC	Global Initiative

Graphs and plots			
AP	The Wall Street Journal	USA Today	Vice News

A.10 References used in the Appendix

- Arandjelović, Relja, and Andrew Zisserman. 2012. Three things everyone should know to improve object retrieval. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE: pp. 2911-2918.
- Bay, Herbert, Tinne Tuytelaars, and Luc Van Gool. 2006. Surf: Speeded up robust features. In *European conference on computer vision*. Springer pp. 404-417.
- Caron, Mathilde, Piotr Bojanowski, Julien Mairal, and Armand Joulin. 2019. “Unsupervised pre-training of image features on non-curated data.” In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE: pp. 2959-2968.
- Gomez, Lluís, Yash Patel, Marçal Rusiñol, Dimosthenis Karatzas, and CV Jawahar. 2017. “Self-supervised learning of visual features through embedding images into text topic spaces.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE: pp. 4230–4239.
- Karpathy, Andrej, and Li Fei-Fei. 2015. “Deep visual-semantic alignments for generating image descriptions.” In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 3128-3137.
- Kriegsmann, Mark, Christian Haag, Cleo-Aron Weis, Georg Steinbuss, Arne Warth, Christiane Zgorzelski, Thomas Muley, Hauke Winter, Martin E Eichhorn, and Florian Eichhorn. 2020. “Deep Learning for the Classification of Small-Cell and Non-Small-Cell Lung Cancer.” *Cancers* 12(6): 1604.
- Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. “Structural topic models for open-ended survey responses.” *American Journal of Political Science* 58(4): 1064-1082.
- Roberts, Margaret E, Brandon M Stewart, and Dustin Tingley. 2014. “stm: R package for structural topic models.” *Journal of Statistical Software* 10(2): 1-40.
- Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airolidi. 2016. “A model of

text for experimentation in the social sciences.” *Journal of the American Statistical Association* 111(515): 988-1003.

Robertson, Ronald E, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. “Auditing partisan audience bias within google search.” *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW): 148.

Ying, Luwei, Jacob M Montgomery, and Brandon M Stewart. Forthcoming. “Inferring concepts from topics: Towards procedures for validating topics as measures.” *Political Analysis*.