Elites Tweet to Get Feet Off the Streets

Kevin Munger, Rich Bonneau,
Jonathan Nagler, Joshua Tucker[1]

February 7, 2017

[1] This paper was written in conjunction with NYUs Social Media and Political Participation (SMaPP) lab, of which Munger is a PhD student member and the remaining authors are Principal Invesetigators. km2713@nyu.edu

Motivating Question

How do elites in non-democratic contexts use information to promote or repress protest/revolution?

Background: Chavismo and Maduro

Background: Chavismo and Maduro

2014 Protests

- Soaring inflation and one of the highest violent crime rates
- Lack of freedom of speech and economic freedom
- Sparked by repression in of student protests in a regional capital
- Paralyzed Caracas for months in early 2014
- Ultimately unsuccessful

Overview

- We use a difference-in-difference approach

Overview

- We use a difference-in-difference approach
- Show that the regime and opposition elites are quantitatively similar prior to the protests

Overview

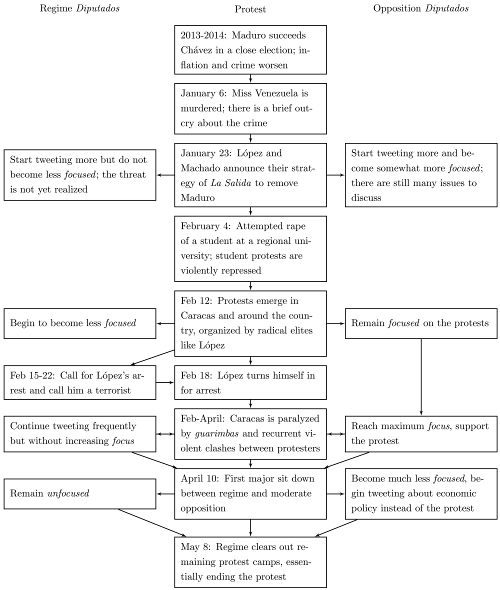
- We use a difference-in-difference approach
- Show that the regime and opposition elites are quantitatively similar prior to the protests
- But the (largely exogenous) onset of the protests causes their behavior to change

Overview

- We use a difference-in-difference approach
- Show that the regime and opposition elites are quantitatively similar prior to the protests
- But the (largely exogenous) onset of the protests causes their behavior to change
- The opposition elites keep talking about the protests, while the regime tries to advance various other narratives, to distract from the protests

Overview

- We use a difference-in-difference approach
- Show that the regime and opposition elites are quantitatively similar prior to the protests
- But the (largely exogenous) onset of the protests causes their behavior to change
- The opposition elites keep talking about the protests, while the regime tries to advance various other narratives, to distract from the protests
- Relative to the "null hypothesis" that both regimes will respond the same way to the protest


Regime *Diputados*
Protest
Opposition *Diputados*
2013-2014: Maduro succeeds Chávez in a close election; inflation and crime worsen
January 6: Miss Venezuela is murdered; there is a brief outcry about the crime
Start tweeting more but do not become less *focused*; the threat is not yet realized
January 23: López and Machado announce their strategy of *La Salida* to remove Maduro
Start tweeting more and become somewhat more *focused*; there are still many issues to discuss
February 4: Attempted rape of a student at a regional university; student protests are violently repressed
Begin to become less *focused*
Feb 12: Protests emerge in Caracas and around the country, organized by radical elites like López
Remain *focused* on the protests
Feb 15-22: Call for López's arrest and call him a terrorist
Feb 18: López turns himself in for arrest
Continue tweeting frequently but without increasing *focus*
Feb-April: Caracas is paralyzed by *guarimbas* and recurrent violent clashes between protesters
Reach maximum *focus*, support the protest
Remain *unfocused*
April 10: First major sit down between regime and moderate opposition
Become much less *focused*, begin tweeting about economic policy instead of the protest
May 8: Regime clears out remaining protest camps, essentially ending the protest

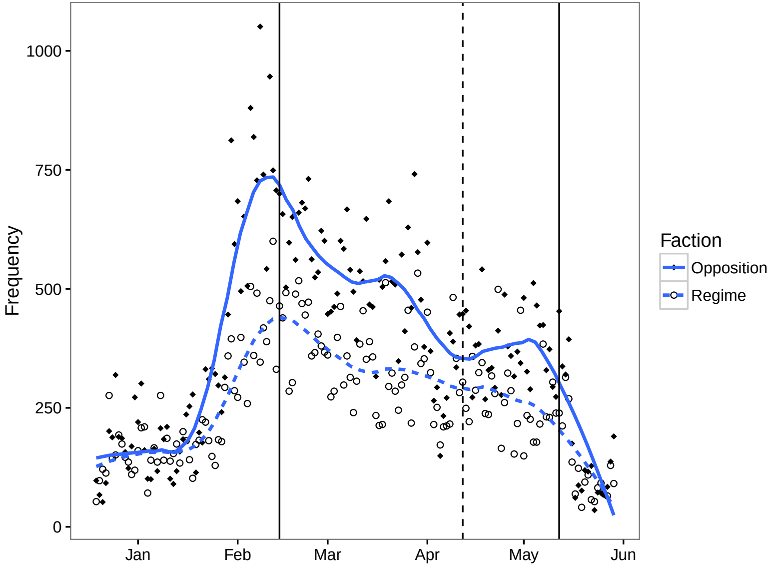
Data Collection

Table: **Number of Tweets by Venezuelan Diputados**

Diputados	*N*	25th	Median	75th	Mean	Total Tweets
Regime	65	109	299	794	663	43,093
Opposition	56	211	584	1,231	1,115	62,423

Period of Analysis: December 19, 2013 - May 29, 2014

Figure: Tweets per Day by Each Coalition



Hypotheses

1. Compared to their pre-protest levels of focus, the opposition diputados' tweets will become more focused and their topic diversity scores will decrease, while the regime diputados' tweets will become less focused and their topic diversity scores will increase.

Hypotheses

1. Compared to their pre-protest levels of focus, the opposition diputados' tweets will become more focused and their topic diversity scores will decrease, while the regime diputados' tweets will become less focused and their topic diversity scores will increase.
2. During the protests, the regime diputados will use long hashtags more frequently than the opposition

Hypotheses

1. Compared to their pre-protest levels of focus, the opposition diputados' tweets will become more focused and their topic diversity scores will decrease, while the regime diputados' tweets will become less focused and their topic diversity scores will increase.
2. During the protests, the regime diputados will use long hashtags more frequently than the opposition
3. During the protests, the regime diputados will send tweets containing multiple hashtags less frequently than the opposition

Topic Models

- There is so much text in the world, usually divided into "documents"–papers, speeches, tweets

Topic Models

- There is so much text in the world, usually divided into "documents"–papers, speeches, tweets
- We'd like to summarize the information in these documents, so we use topic models to create topics and assign them to documents

Topic Models

- There is so much text in the world, usually divided into "documents"–papers, speeches, tweets
- We'd like to summarize the information in these documents, so we use topic models to create topics and assign them to documents
- This method is "unsupervised machine learning," and this reduces the role of researcher biases (more on this later)

Topic Models

- There is so much text in the world, usually divided into "documents"–papers, speeches, tweets
- We'd like to summarize the information in these documents, so we use topic models to create topics and assign them to documents
- This method is "unsupervised machine learning," and this reduces the role of researcher biases (more on this later)
  - This means you don't choose "topics"

Topic Models

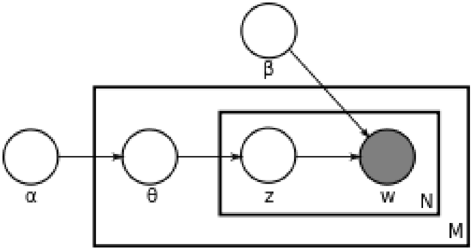
- There is so much text in the world, usually divided into "documents"–papers, speeches, tweets
- We'd like to summarize the information in these documents, so we use topic models to create topics and assign them to documents
- This method is "unsupervised machine learning," and this reduces the role of researcher biases (more on this later)
  - This means you don't choose "topics"
  - In fact, "topics" is sometimes begging the question

Topic Models

- There is so much text in the world, usually divided into "documents"–papers, speeches, tweets
- We'd like to summarize the information in these documents, so we use topic models to create topics and assign them to documents
- This method is "unsupervised machine learning," and this reduces the role of researcher biases (more on this later)
  - This means you don't choose "topics"
  - In fact, "topics" is sometimes begging the question
  - Also means you want a LOT of data

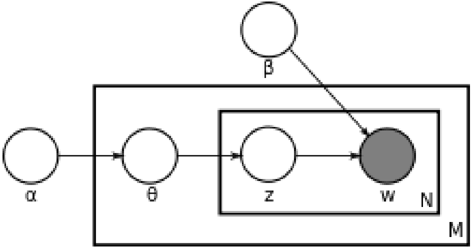
Latent Dirchlet Allocation

- Given only a number of topics, concentration parameter α and a collection of documents, produces a distribution of "topics" over those documents
- Does *not* incorporate covariates about the documents



Latent Dirchlet Allocation

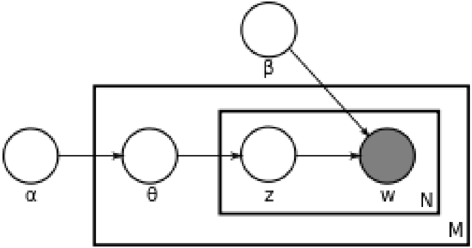
- Given only a number of topics, concentration parameter α and a collection of documents, produces a distribution of "topics" over those documents
- Does *not* incorporate covariates about the documents



- LDA is a *generative topic model*–the fundamental assumption is that each document is created via draws from some distribution

Latent Dirchlet Allocation

- Given only a number of topics, concentration parameter α and a collection of documents, produces a distribution of "topics" over those documents
- Does *not* incorporate covariates about the documents



- LDA is a *generative topic model*–the fundamental assumption is that each document is created via draws from some distribution
- With the caveat that word order doesn't matter–"bag of words"

From text to topics

- What are the concrete steps of taking documents and generating topics using LDA?

From text to topics

- What are the concrete steps of taking documents and generating topics using LDA?
- Since we're using tweets, first step is to combine each day's worth of tweets from each coalition into a document

From text to topics

- What are the concrete steps of taking documents and generating topics using LDA?
- Since we're using tweets, first step is to combine each day's worth of tweets from each coalition into a document
    - Turn each document into a Document-Term Matrix

From text to topics

- What are the concrete steps of taking documents and generating topics using LDA?
- Since we're using tweets, first step is to combine each day's worth of tweets from each coalition into a document
  - Turn each document into a Document-Term Matrix
  - Remove stop-words

From text to topics

- What are the concrete steps of taking documents and generating topics using LDA?
- Since we're using tweets, first step is to combine each day's worth of tweets from each coalition into a document
    - Turn each document into a Document-Term Matrix
    - Remove stop-words
    - Remove rare words

From text to topics

- What are the concrete steps of taking documents and generating topics using LDA?
- Since we're using tweets, first step is to combine each day's worth of tweets from each coalition into a document
  - Turn each document into a Document-Term Matrix
  - Remove stop-words
  - Remove rare words
  - Stemming

From text to topics

- What are the concrete steps of taking documents and generating topics using LDA?
- Since we're using tweets, first step is to combine each day's worth of tweets from each coalition into a document
  - Turn each document into a Document-Term Matrix
  - Remove stop-words
  - Remove rare words
  - Stemming
  - Choose the number of topics, K

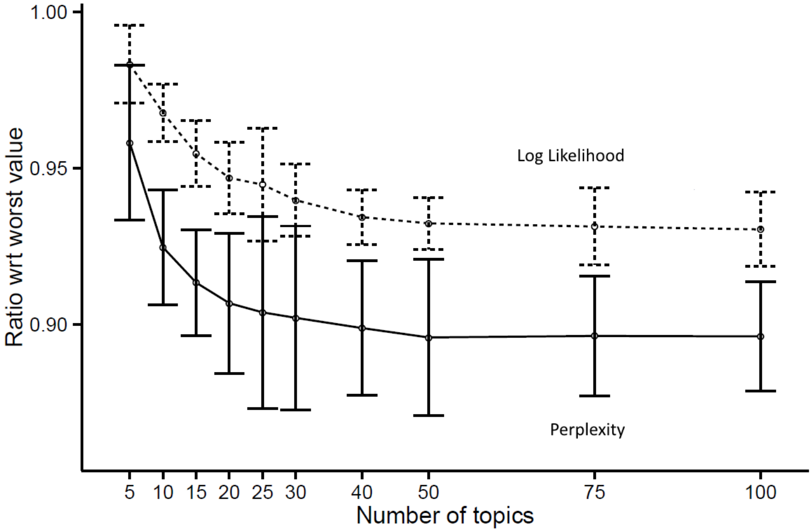
From text to topics

- What are the concrete steps of taking documents and generating topics using LDA?
- Since we're using tweets, first step is to combine each day's worth of tweets from each coalition into a document
  - Turn each document into a Document-Term Matrix
  - Remove stop-words
  - Remove rare words
  - Stemming
  - Choose the number of topics, K
    - This is the hard part

From text to topics

- What are the concrete steps of taking documents and generating topics using LDA?
- Since we're using tweets, first step is to combine each day's worth of tweets from each coalition into a document
  - Turn each document into a Document-Term Matrix
  - Remove stop-words
  - Remove rare words
  - Stemming
  - Choose the number of topics, K
    - This is the hard part

Figure: Testing Different Numbers of Topics



From text to topics

- LDA creates a probability distribution that each document pertains to each topic

From text to topics

- LDA creates a probability distribution that each document pertains to each topic
  - These are the θ parameters

From text to topics

- LDA creates a probability distribution that each document pertains to each topic
  - These are the θ parameters
  - For each document, there is a θ score given to each topic; these sum to 1

From text to topics

- LDA creates a probability distribution that each document pertains to each topic
  - These are the θ parameters
  - For each document, there is a θ score given to each topic; these sum to 1
  - Calculate the Shannon Entropy of this θ distribution:

From text to topics

- LDA creates a probability distribution that each document pertains to each topic
  - These are the θ parameters
  - For each document, there is a θ score given to each topic; these sum to 1
  - Calculate the Shannon Entropy of this θ distribution:

$$\text{Topic Diversity} = -\sum_{1}^{k} \theta_{w,k} \log_2(\theta_{wk})$$

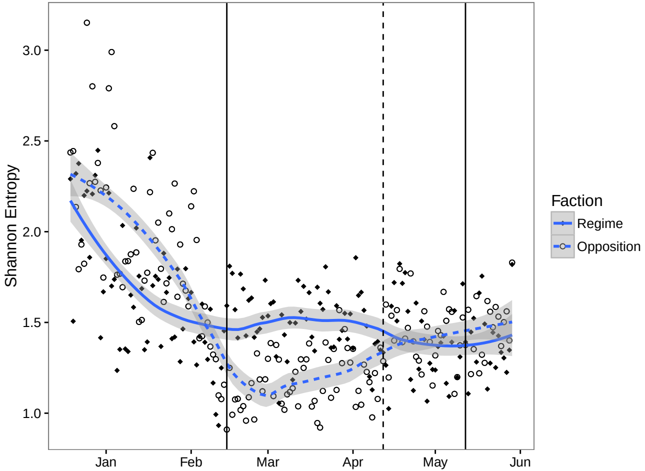
From text to topics

- LDA creates a probability distribution that each document pertains to each topic
  - These are the θ parameters
  - For each document, there is a θ score given to each topic; these sum to 1
  - Calculate the Shannon Entropy of this θ distribution:

$$\text{Topic Diversity} = -\sum_{1}^{k} \theta_{w,k} \log_2(\theta_{wk})$$

  - Higher values = more diversity/less focus

Analysis–Topic Diversity as a Measure of Focus



Each point represents the Topic Diversity score for the opposition and regime tweets respectively, compuated based on their tweets. The vertical lines correspond to the murder of Miss Venezuela, the arrest of López, and Beginning of the Independence Movement Day, respectively.

Restricting to Coherent Topics

- Excellent reviewer comment (the system works!)

Restricting to Coherent Topics

- Excellent reviewer comment (the system works!)
- Selecting k with cross-validation can lead to semantically meaningless topics

Restricting to Coherent Topics

- Excellent reviewer comment (the system works!)
- Selecting k with cross-validation can lead to semantically meaningless topics
- Calculate *semantic coherence* (Roberts, Stewart and Tingley, 2014):

Restricting to Coherent Topics

- Excellent reviewer comment (the system works!)
- Selecting k with cross-validation can lead to semantically meaningless topics
- Calculate *semantic coherence* (Roberts, Stewart and Tingley, 2014):
  - Look at top N words in a topic

Restricting to Coherent Topics

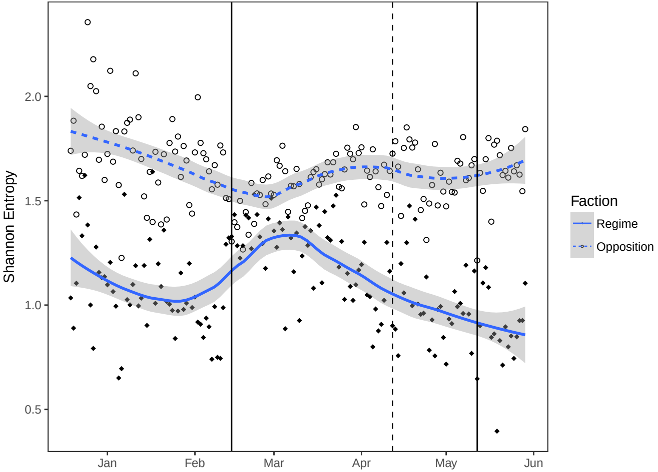
- Excellent reviewer comment (the system works!)
- Selecting k with cross-validation can lead to semantically meaningless topics
- Calculate *semantic coherence* (Roberts, Stewart and Tingley, 2014):
  - Look at top N words in a topic $\rightsquigarrow$ how often they co-occur in documents

Restricting to Coherent Topics

Table: **Top Terms for Relevant Topics**

Top Government Topics	
#	Terms
44	nicolasmadur, venezuel, puebl, psuv, paz, president, nuev, chávez, dcabellor, maduro
76	suramerican, eduardopinata, zerp, campament, cati, háro, may, rendón, oro, amesesdetusiembracomandant
77	celac, ener, pacificación, cumbr, contrab, caribe, natalici, haban, hagamoslapaz, magallan
Top Opposition Topics	
#	Terms
74	**papel, periód, violenci, lasal, segur, matern, sinpapelnohayperiód, medi, pilieri, biagi**
94	estudiant, venezuel, protest, puebl, paz, march, call, maduro, hcapril, hoy
72	prmerojustici, gobiern, diput, hoy, más, juliocmontoy, país, dip, vía

Coherent Topic Diversity as a Measure of Focus



Each point represents the Topic Diversity score for the opposition and regime tweets respectively, compuated based on their tweets. The vertical lines correspond to the murder of Miss Venezuela, the arrest of López, and Beginning of the Independence Movement Day, respectively.

Alternative Model: Correlated Topic Model (CTM)

- An extension of the previous model (Blei and Lafferty, 2007)

Alternative Model: Correlated Topic Model (CTM)

- An extension of the previous model (Blei and Lafferty, 2007)
- Unlike LDA, explicitly models the co-occurance of different topics

Alternative Model: Correlated Topic Model (CTM)

- An extension of the previous model (Blei and Lafferty, 2007)
- Unlike LDA, explicitly models the co-occurance of different topics
- Trade-off between semantic coherence and *exclusivity*:

Alternative Model: Correlated Topic Model (CTM)

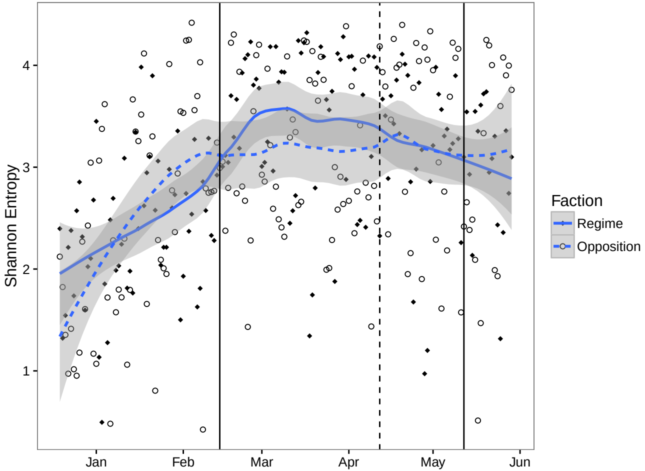
- An extension of the previous model (Blei and Lafferty, 2007)
- Unlike LDA, explicitly models the co-occurance of different topics
- Trade-off between semantic coherence and *exclusivity*:
  - The % of the top N words in a given topic that are also in the top *N* word

Topic Exclusivity

Table: **Top Terms for Relevant Topics**

	Top Government Topics
#	Terms
44	nicolasmadur, **venezuel**, **puebl**, psuv, **paz**, president, nuev, chávez, dcabellor, **maduro**
76	suramerican, eduardopinata, zerp, campament, cati, háro, may, rendón, oro, amesesdetusiembracomandant
77	celac, ener, pacificación, cumbr, contrab, caribe, natalici, haban, hagamoslapaz, magallan
	Top Opposition Topics
#	Terms
74	papel, periód, violenci, lasal, segur, matern, sinpapelnohayperiód, medi, pilieri, biagi
94	estudiant, **venezuel**, protest, **puebl**, **paz**, march, call, **maduro**, hcapril, **hoy**
72	prmerojustici, gobiern, diput, **hoy**, más, juliocmontoy, país, dip, vía

Correlated Topic Diversity



Each point represents the Topic Diversity score for the opposition and regime tweets respectively, compuated based on their tweets. The vertical lines correspond to the murder of Miss Venezuela, the arrest of López, and Beginning of the Independence Movement Day, respectively.

Hashtags

- Hashtags are specific to Twitter but are also essential to it

Hashtags

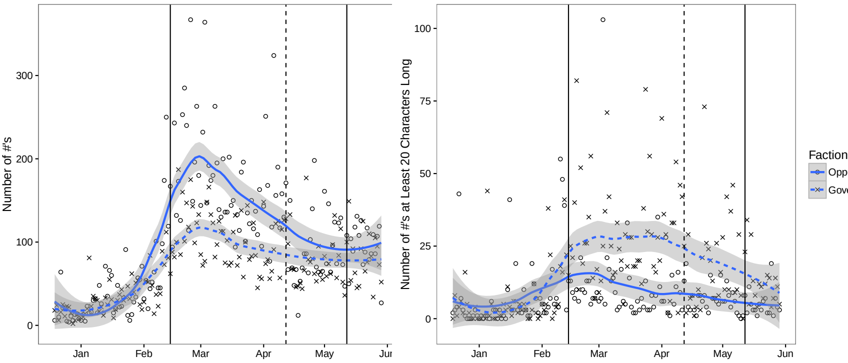
- Hashtags are specific to Twitter but are also essential to it
- Longer hashtags $\rightarrow$ attempt to create a topic of discussion

Hashtags

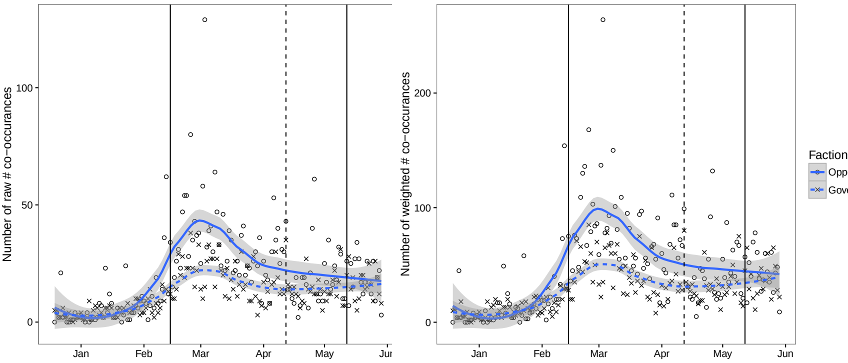
- Hashtags are specific to Twitter but are also essential to it
- Longer hashtags → attempt to create a topic of discussion
    - #DesconectateDeLaGuarimba
    - #VzlaBajoAtaqueMediatico
    - #GringosyFascistasRespeten
- Multiple hashtags in one tweet → attempt to unite ongoing discussions

Hashtags

- Hashtags are specific to Twitter but are also essential to it
- Longer hashtags → attempt to create a topic of discussion
  - #DesconectateDeLaGuarimba
  - #VzlaBajoAtaqueMediatico
  - #GringosyFascistasRespeten
- Multiple hashtags in one tweet → attempt to unite ongoing discussions
  - #caracas
  - #sucre
  - #guayana
  - #petare
  - #juntosgobernamos
  - #12f
  - #hayuncamino
  - #8d



Left graph shows the total number of hashtags tweeted per day by the two coalitions. Right graph restricts this analysis to the number of hashtags that were at least twenty characters long.



Left graph shows the raw number of tweets containing multiple different hashtags tweeted per day by the two coalitions. Right graph weights this number by the number of hashtags each of those tweets contained.

Conclusion

- We have shown how to analyze elite communication to infer regime strategies

Conclusion

- We have shown how to analyze elite communication to infer regime strategies
- Unsupervised approach is generalizable and replicable

Conclusion

- We have shown how to analyze elite communication to infer regime strategies
- Unsupervised approach is generalizable and replicable
  - Caveat: might look very different in a democratic context
- Topic models are good at summarizing the information in a text, less good at identifying "topics" per se

Conclusion

- We have shown how to analyze elite communication to infer regime strategies
- Unsupervised approach is generalizable and replicable
    - Caveat: might look very different in a democratic context
- Topic models are good at summarizing the information in a text, less good at identifying "topics" per se
- More research on Twitter-specific features like hashtags