Fantastic Beasts and Whether They Matter:

How Pro-Government "Trolls" Influence Political Conversations in Russia

Anton Sobolev[1]

[1]UCLA

COMPASS, UCLA 2018

Fake agents

- Kuran (1989), Lohmann (1994): the role of public signals in political behavior
  - The key factor of the decision to act: the ability to learn the distribution of preferences
- "Fake agents" generate informational noise and can complicate political learning

"DISSIDENTS DECIDED TO ORGANIZE A MEETING AT PUSHKIN SQUARE. WHAT SHOULD HAVE WE DONE? WE MOBILIZED "VOLUNTEERS AMONG COMMUNISTS" AND ORGANIZED OUR OWN MEETING TO OCCUPY THE SAME PLACE."

"DISSIDENTS DECIDED TO ORGANIZE A MEETING AT PUSHKIN SQUARE. WHAT SHOULD HAVE WE DONE? WE MOBILIZED "VOLUNTEERS AMONG COMMUNISTS" AND ORGANIZED OUR OWN MEETING TO OCCUPY THE SAME PLACE."

VLADIMIR PUTIN, 2000

Fake agents on social media

What we DO know

- **Extremely widespread phenomenon.** Venezuela (Munger et al, 2015), South Korea (Keller et al, 2017), China (King et al. 2016, Miller, 2017), Russia (Sanovich et al 2017)
- **Different political contexts:** from elections to protests

Political astroturfing on social media

What we DO NOT know

- Governments spend huge amount of money to employ "fake commentators", but
  - The impact of political astroturfing is unclear
  - Do regular Internet users pay attention to what pro-government agents write? Can pro-government agents engage others with the loyalist agenda? Can they divert ordinary users from criticizing the government?

Political astroturfing on social media

What we DO NOT know

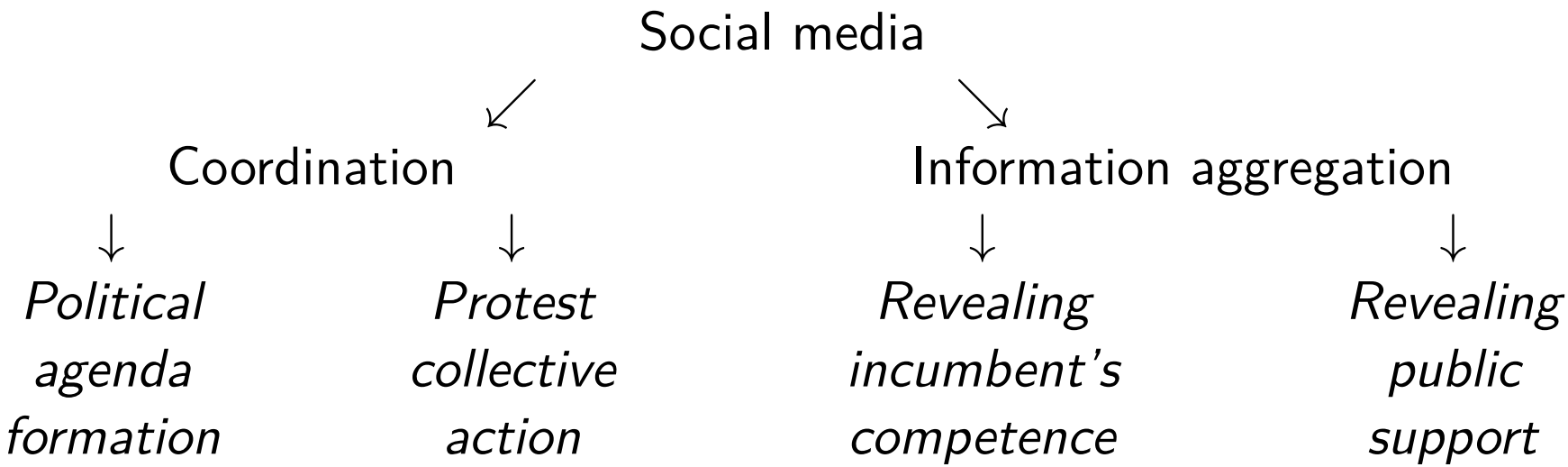
- Governments spend huge amount of money to employ "fake commentators", but
  - The impact of political astroturfing is unclear
  - Do regular Internet users pay attention to what pro-government agents write? Can pro-government agents engage others with the loyalist agenda? Can they divert ordinary users from criticizing the government?
- **This paper:** the case of Russian trolls' online activity in Russia, leaked documents of social media accounts of paid commentators
  - Present a novel dataset of trolls' activities on *LiveJournal*
  - A framework for analyzing politically charged conversations on social media
  - Classification of goals of online political astroturfing
  - A method of estimating the effect of interventions on critically charged online discussions
  - Do trolls change online conversations?

Political astroturfing on social media

What we DO NOT know

- Governments spend huge amount of money to employ "fake commentators", but
  - The impact of political astroturfing is unclear
  - Do regular Internet users pay attention to what pro-government agents write? Can pro-government agents engage others with the loyalist agenda? Can they divert ordinary users from criticizing the government?
- **This paper:** the case of Russian trolls' online activity in Russia, leaked documents of social media accounts of paid commentators
  - Present a novel dataset of trolls' activities on *LiveJournal*
  - A framework for analyzing politically charged conversations on social media
  - Classification of goals of online political astroturfing
  - A method of estimating the effect of interventions on critically charged online discussions
  - Do trolls change online conversations? **Yes. sometimes...**

Social media and contentious politics



Government responses

	$\longrightarrow$	*Legal restrictions*
Censorship	$\longrightarrow$	*Intimidation*
	$\longrightarrow$	*Black lists and content filtering*
Propaganda	$\longrightarrow$	*Exposure to biased news reports*
	$\longrightarrow$	*Exposure to fake regime supporters*
Astroturfing	$\longrightarrow$	*Exposure to fake dissidents*
	$\longrightarrow$	*Exposure to fake median citizens*

Government responses

	$\longrightarrow$	*Legal restrictions*
Censorship	$\longrightarrow$	*Intimidation*
	$\longrightarrow$	*Black lists and content filtering*
Propaganda	$\longrightarrow$	*Exposure to biased news reports*
	$\longrightarrow$	*Exposure to fake regime supporters*
Astroturfing	$\longrightarrow$	*Exposure to fake dissidents*
	$\longrightarrow$	*Exposure to fake median citizens*

Targets and goals of paid commentators

- *Targets:*
  - authors of blog posts
  - participants of online conversations
  - readers of online conversations and posts
  - online public narrative

Targets and goals of paid commentators

- *Targets:*
  - authors of blog posts
  - participants of online conversations
  - readers of online conversations and posts
  - online public narrative

Targets and goals of paid commentators

- *Targets:*
  - authors of blog posts
  - participants of online conversations
  - readers of online conversations and posts
  - online public narrative
- *Different goals:*
  - Project strength (*"We see you and know where you are"*)
  - Project popularity (*"86% of Russians support Putin"*)
  - Imitate Anti-government Extremism
  - Promote pro-government agenda
  - Distract users from discussing critically charged topics

Targets and goals of paid commentators

- *Targets:*
  - authors of blog posts
  - participants of online conversations
  - readers of online conversations and posts
  - online public narrative
- *Different goals:*
  - Project strength (*"We see you and know where you are"*)
  - Project popularity (*"86% of Russians support Putin"*)
  - Imitate Anti-government Extremism
  - Promote pro-government agenda
  - Distract users from discussing critically charged topics

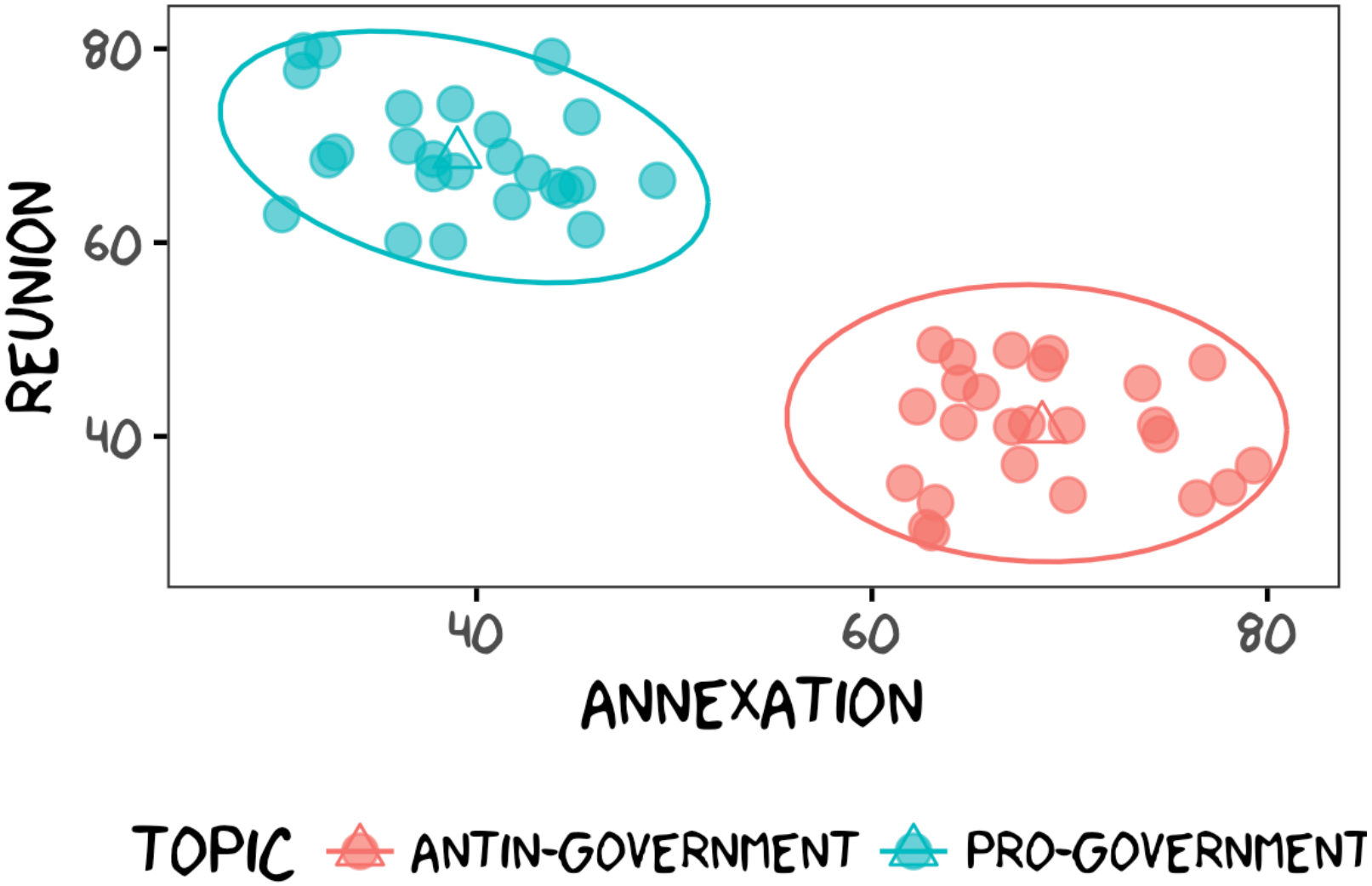
Example

- a post "*Fasting Will Be Less Pleasant*" by Deacon Andrei Kuraev
  - food ban negatively affects Orthodox Russians Great Lent before Easter
  - a follow-up discussion of negative effects of sanctions
  - Глусмапгода "problem would be solved in a couple of years, because Crimea is a perfect place to grow olives"
  - Глусмапгода is a troll accont
- Two scenarios of interest:
  - a shift from discussion of negative effects of sanctions toward less sensitive topics
  - a shift toward discussion of positive aspects of Crimea annexation

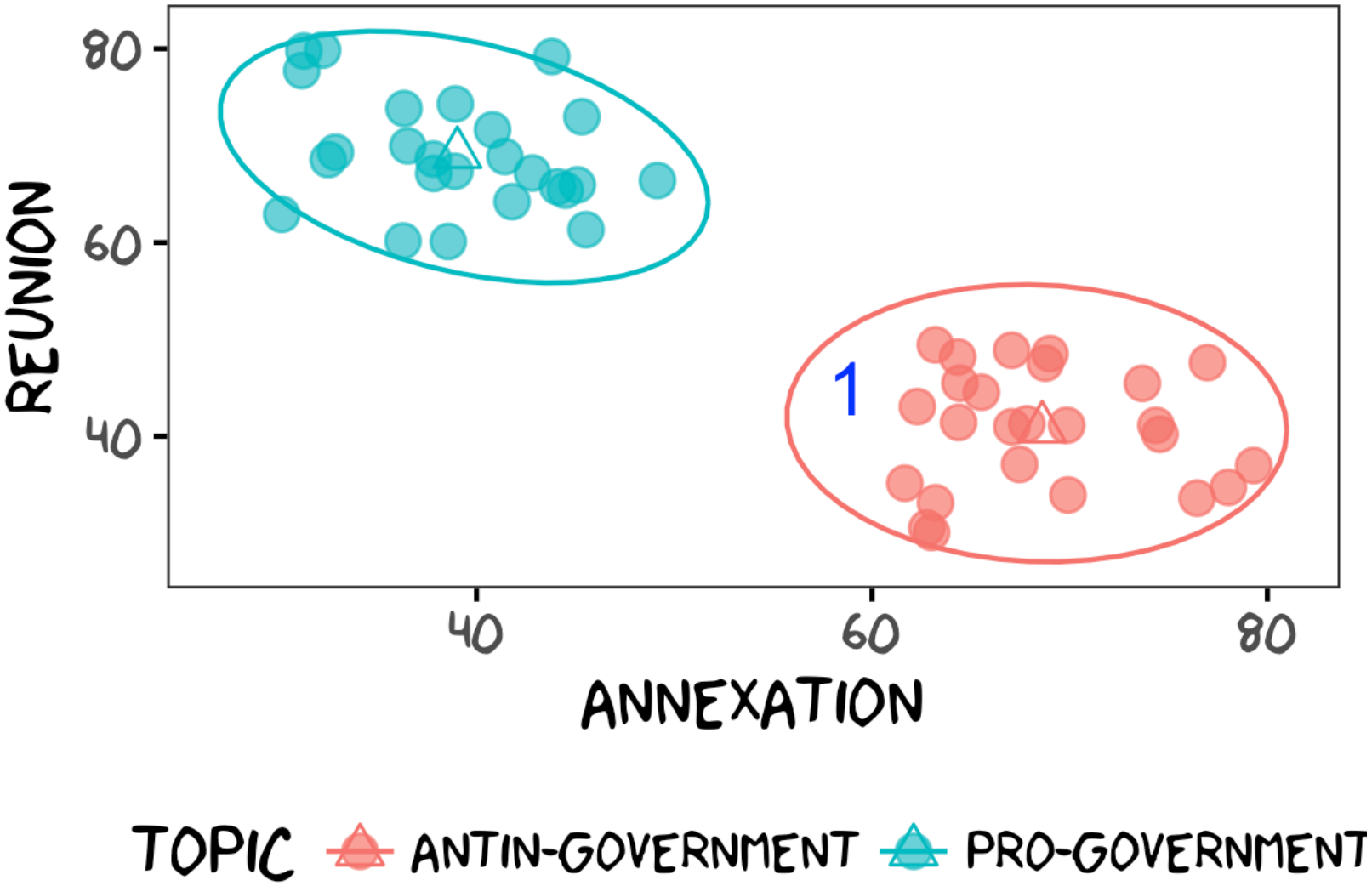
Example

- Post "*Fasting Will Be Less Pleasant*" by Deacon Andrei Kuraev
  - food ban negatively affects Orthodox Russians Great Lent before Easter
  - a follow-up discussion
  - Глеб Лягушонок "problem would be solved in a couple of years, because Crimea is a perfect place to grow olives".
- Two scenarios of interest:
  - shift from discussion of negative effects of sanctions toward less sensitive topics **diversion**
  - shift toward discussion of positive aspects of Crimea annexation **promotion**

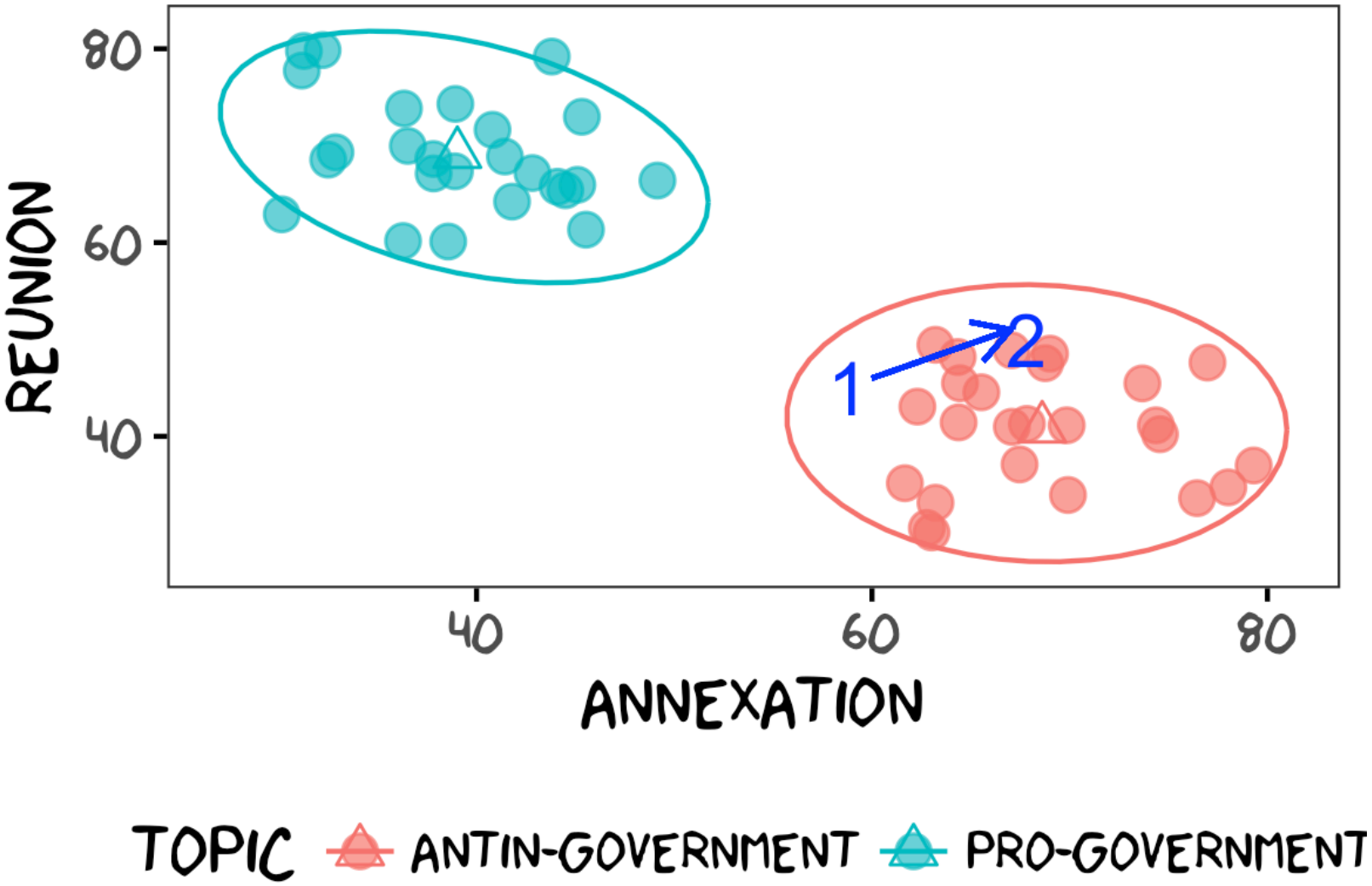
How to track the shift in online conversations



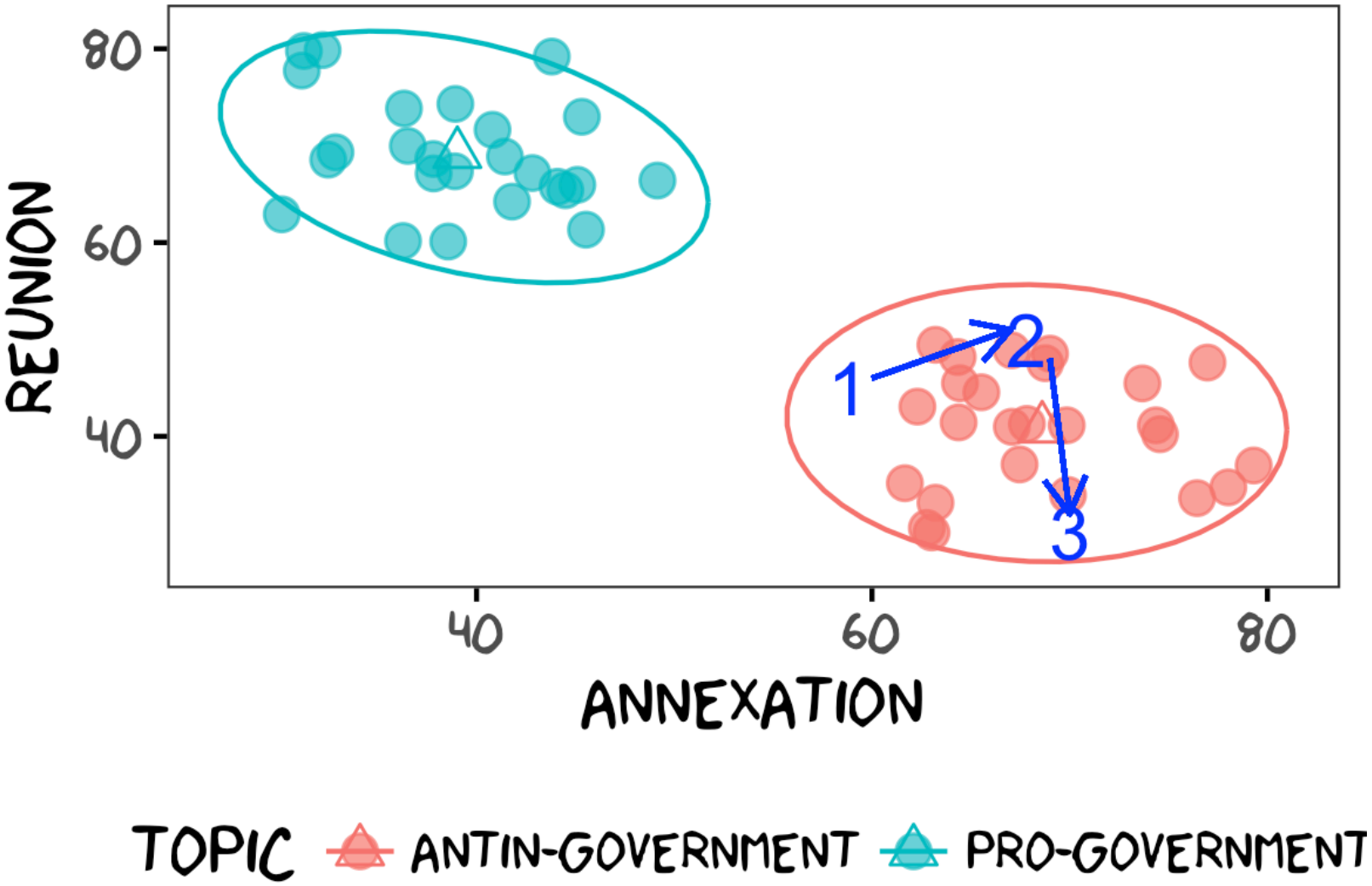
How to track evolution of online conversations



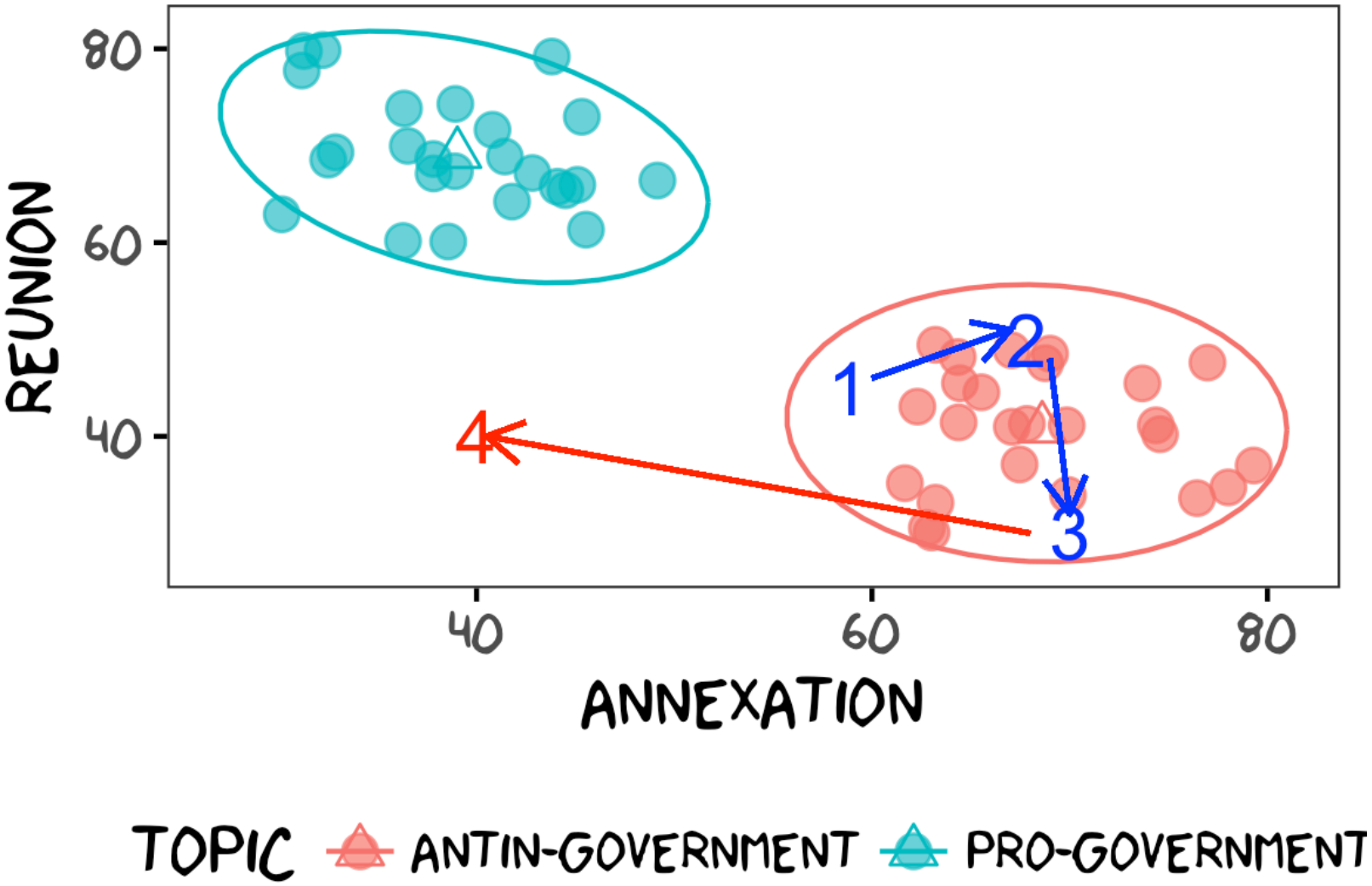
How to track evolution of online conversations



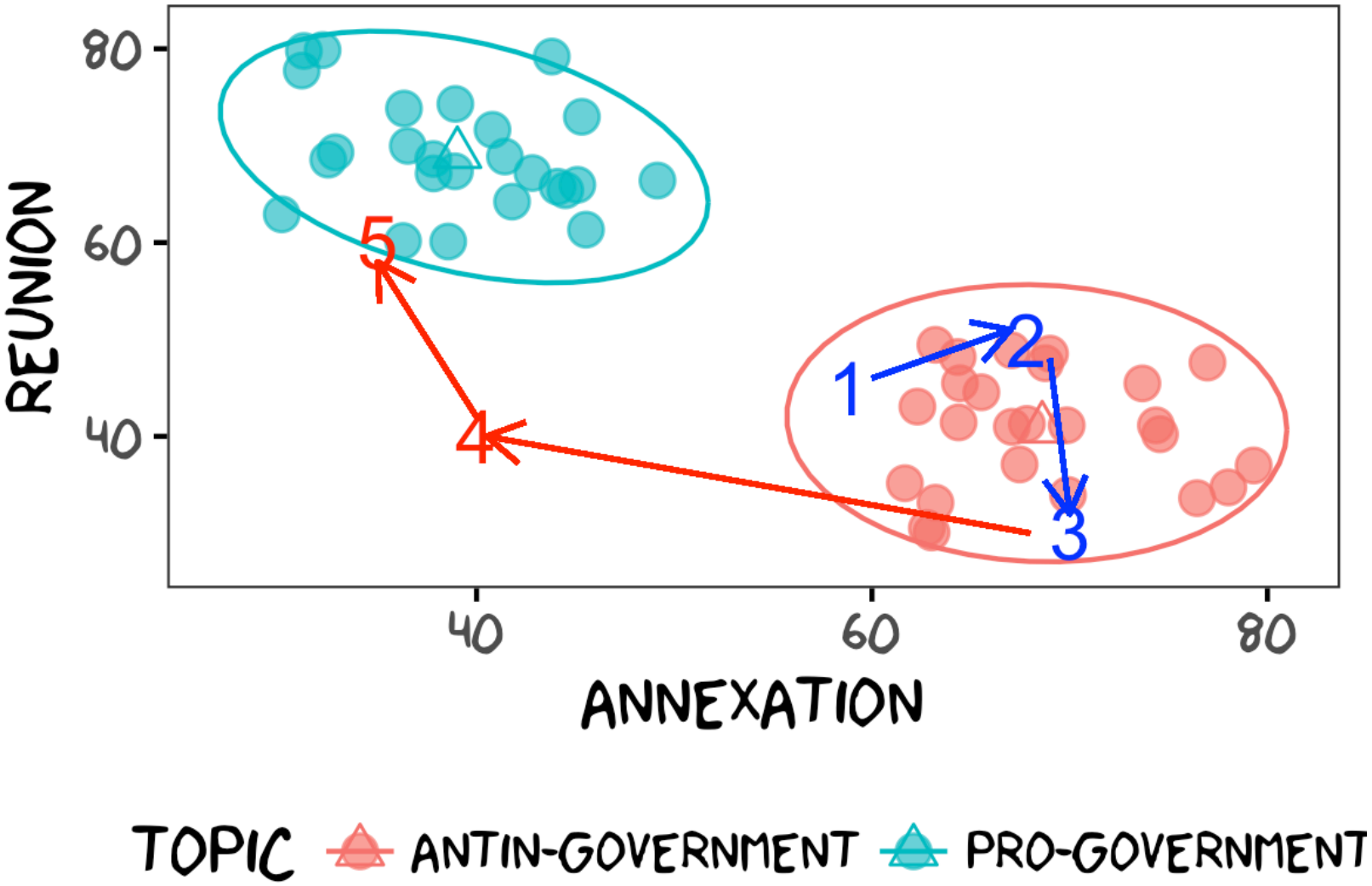
How to track evolution of online conversations



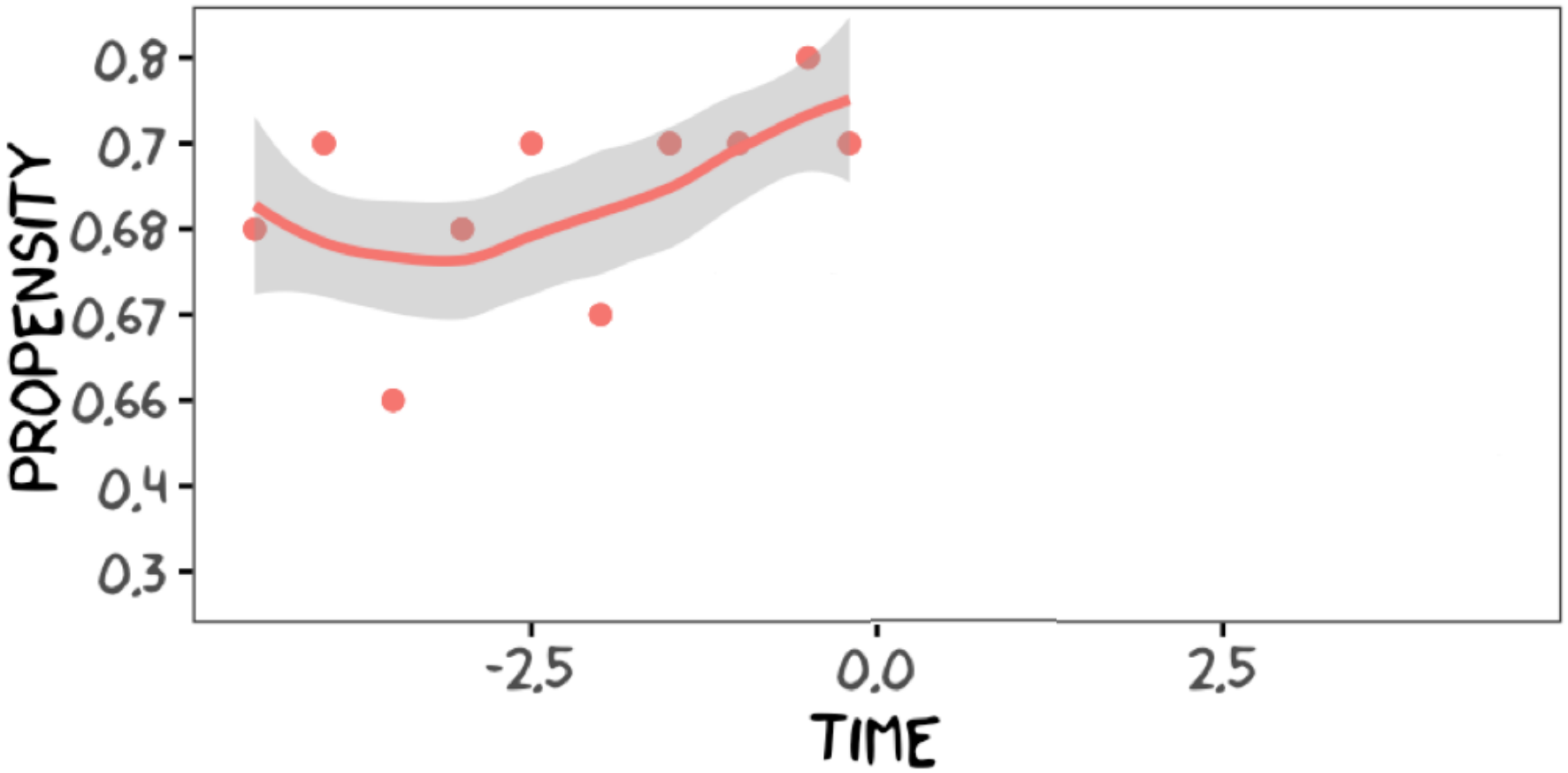
How to track evolution of online conversations



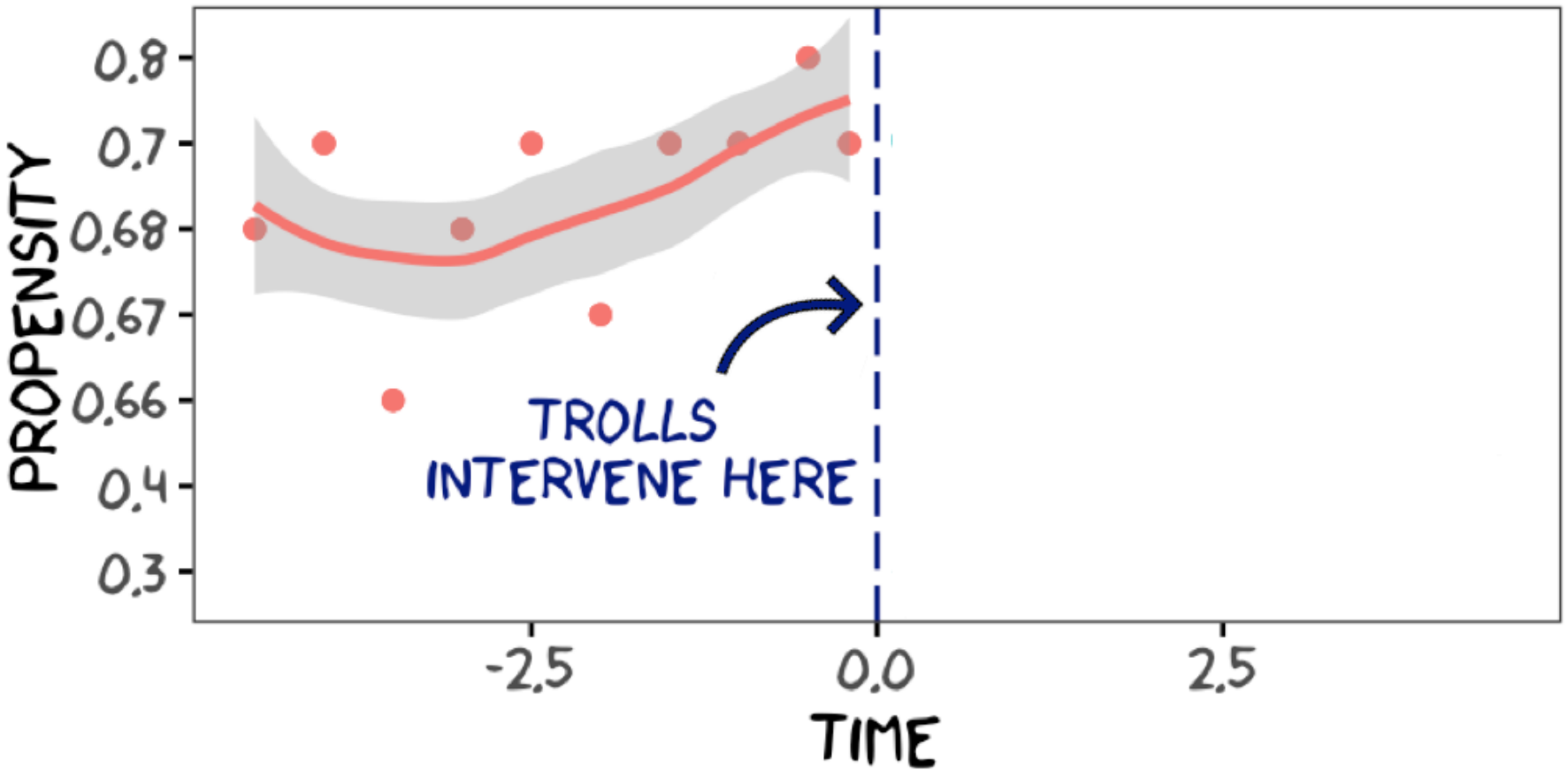
How to track evolution of online conversations



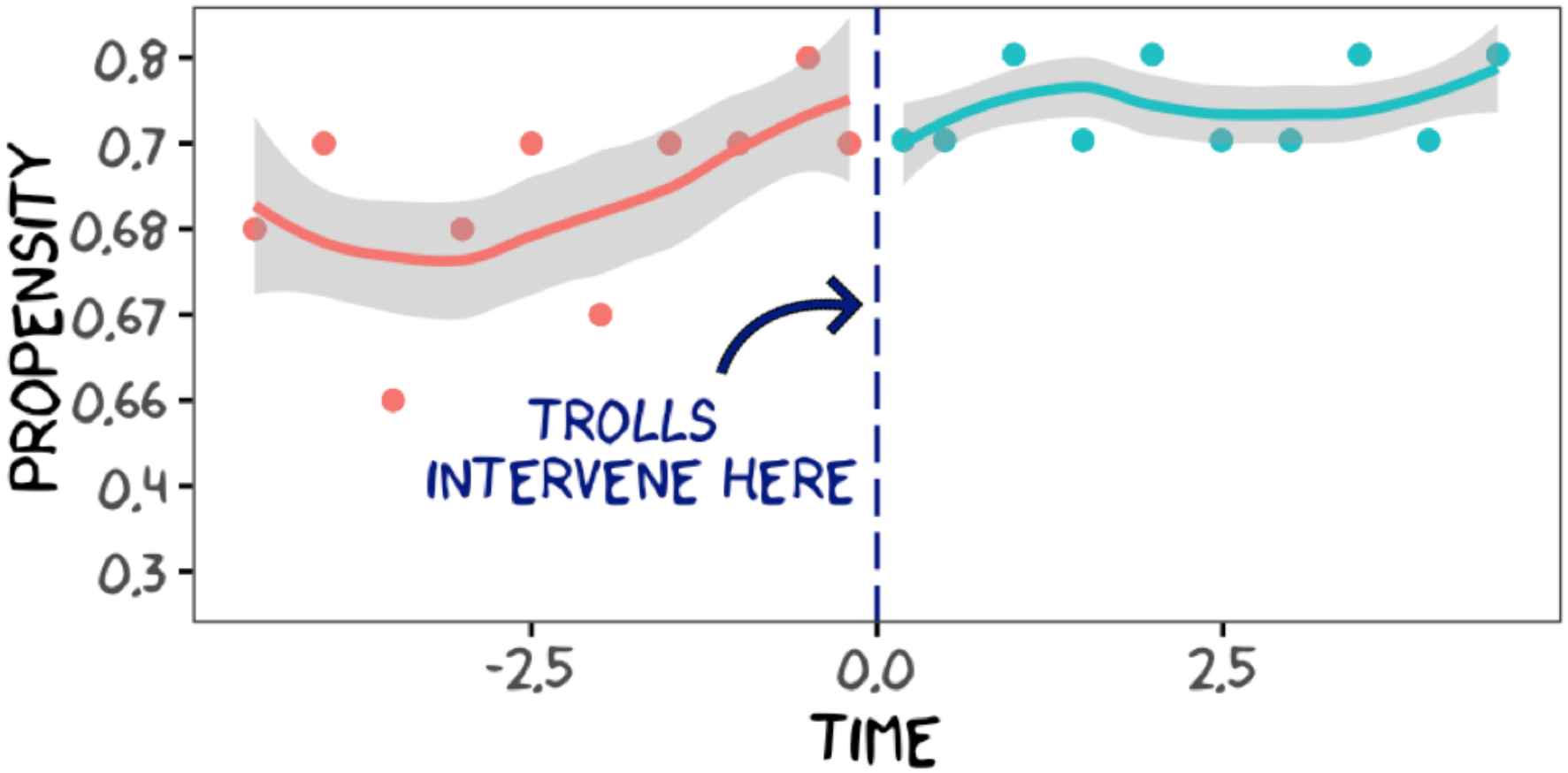
How to track evolution of online conversations



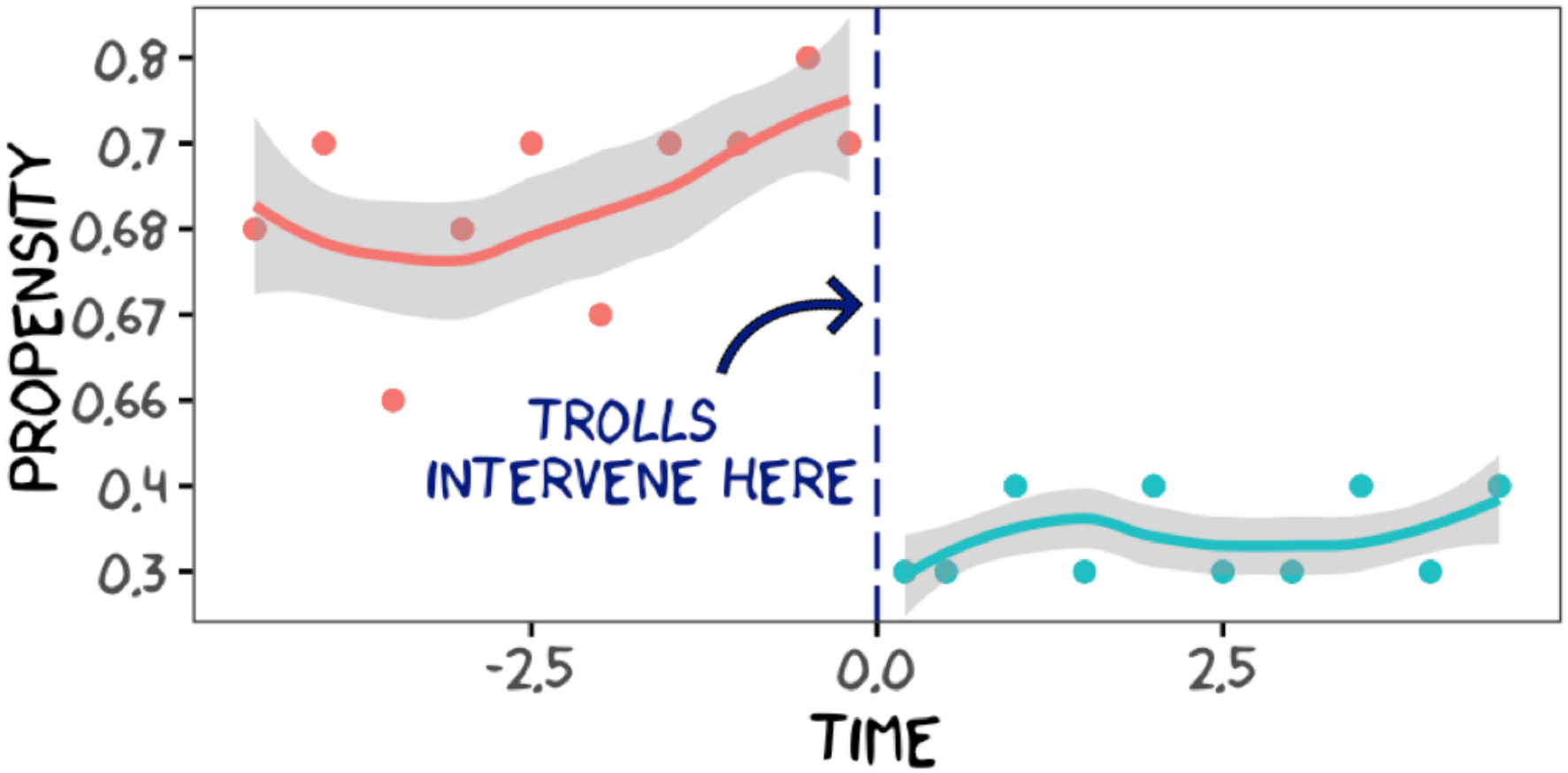
How to track evolution of online conversations



How to track evolution of online conversations



How to track evolution of online conversations



Data - 1

- ***LiveJournal, 2014-2015***
  - leads in the amount of discussions concerning current affairs (Koltsova and Shcherbak, 2014)
  - has happened to be the most common social media platform for dissidents of the regime

Data - 1

- ***LiveJournal, 2014-2015***
  - leads in the amount of discussions concerning current affairs (Koltsova and Shcherbak, 2014)
  - has happened to be the most common social media platform for dissidents of the regime
  - in March 2015, *Novaya Gazeta* published an investigation about the activities of "Trolls Factory" on *LJ*.
    - List of 700 trolls' accounts

Data - 2

- a dataset of almost 438k trolls' posts
- a dataset of 80k posts and follow-up discussions infiltrated by those trolls
  - 60k did not refer to any current affairs
  - 12k discussed politics
  - 8k discussed problems with Russian national economy

Data - 2

- a dataset of almost 438k trolls' posts
- a dataset of 80k posts and follow-up discussions infiltrated by those trolls
  - 60k did not refer to any current affairs
  - 12k discussed politics
  - 8k discussed problems with Russian national economy

Data processing

	Author	Time		Post Title
		Post		
	User 1	Time	Comment	Pooled Comments $t = -2$
	User 2	Time	Comment	
	User 1	Time	Comment	Pooled Comments $t = -1$
	User 2	Time	Comment	
drop ←	Troll 1	Time	Comment	$t = 0$
	User 2	Time	Comment	Pooled Comments $t = +1$
drop ←	Troll 2	Time	Comment	
	User 1	Time	Comment	
	User 3	Time	Comment	

Data processing

- For each group of posts:
  - Split the follow-up conversations into 30-minute slices (+- 5 hours around the time of the first troll comment)
  - Estimate the distribution of topics in these slices with Latent Dirichlet Allocation model (Blei, 2003)
  - Choose topics of interest

Topics of interest

	Economic discussions	**Politics discussions**
Anti-government topic	*"ruble" + "price" +"oil"+ "USD" + "exchange rate" + "Economy" + "crisis" + "Putin"*	*"war" + "Ukraine" + "military" + "Donbas" + "Donetzk" + "Boeing"*
Pro-government topic	*"good" + "salary" + "employed" + "better" + "jobs" + "can afford"*	*"Ukraine" + "USA" + "plot" + "Crimea" + "great" + "peace"*

Empirical strategy

- Hypotheses:
  - **Diversion hypothesis:** *the propensity of the online conversation to cover an anti-government topic goes down after trolls intervene*
  - **Promotion hypothesis:** *the propensity of the online conversation to cover a pro-government topic goes up after trolls intervene*
- Treat the time of the first troll comment as a cut-point
- Fit RD model
  - conversation-level clustered errors
  - Imbens-Kalyanaraman optimal bandwidth calculation (Imbens and Kalyanaraman, 2012)
- Result: LATE on the change in the prominence of the topic while taking into account the natural evolution of this topic in the discussion

Identification

- **Assumption:** in the narrow time frame, the timing the intervention is as good as if random

Figure: A plausible contrafactual for treated under the *narrow window* assumption

Identification

- **Assumption:** in the narrow time frame, the timing the intervention is as good as if random

Figure: A plausible contrafactual for treated under the *narrow window* assumption

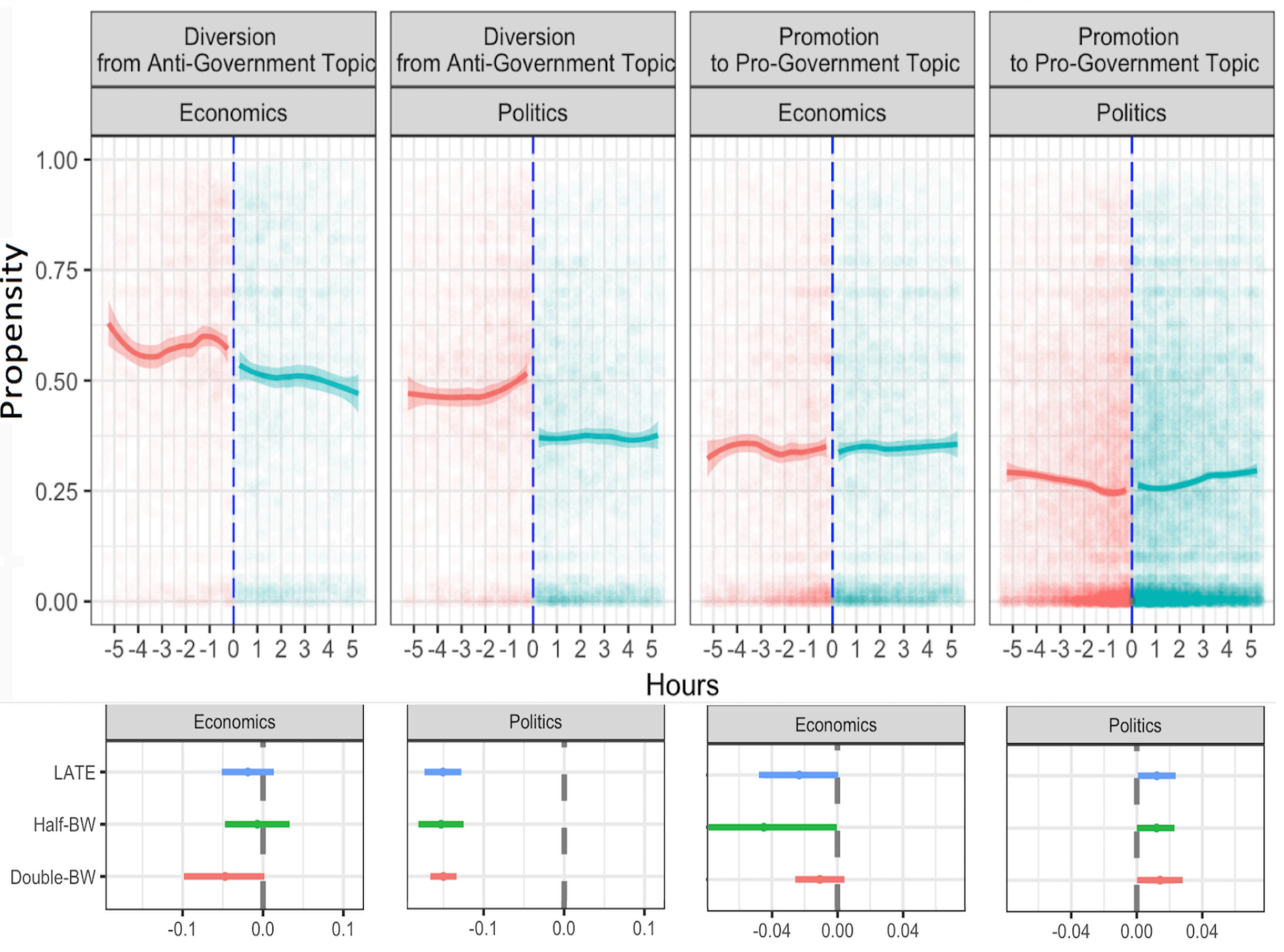
Identification

- **Assumption:** in the narrow time frame, the timing the intervention is as good as if random

Figure: A plausible contrafactual for treated under the *narrow window* assumption

Identification

- Is this assumption plausible?
  - Evidence from data: *trolls are not different from regular users in terms of the time of their first comment*
  - Evidence from the leaked documents:
    - Trolls have fixed working hours
    - Trolls are required to post numerous comments on numerous posts per day
    - They required to attack posts of specific content
    - *Trolls need to manually read a large number of post abstracts via LJ search engine to identify the valid posts for targeting*


Diversion from Anti-Government Topic
Economics
Diversion from Anti-Government Topic
Politics
Promotion to Pro-Government Topic
Economics
Promotion to Pro-Government Topic
Politics
Propensity
1.00
0.75
0.50
0.25
0.00
-5 -4 -3 -2 -1 0 1 2 3 4 5
Hours
Economics
Politics
Economics
Politics
LATE
Half-BW
Double-BW
-0.1
0.0
0.1
-0.04
0.00
0.04

Threats to the validity of the results - 1: The effect of a new user

Threats to the validity of the results - 1: The effect of a new user

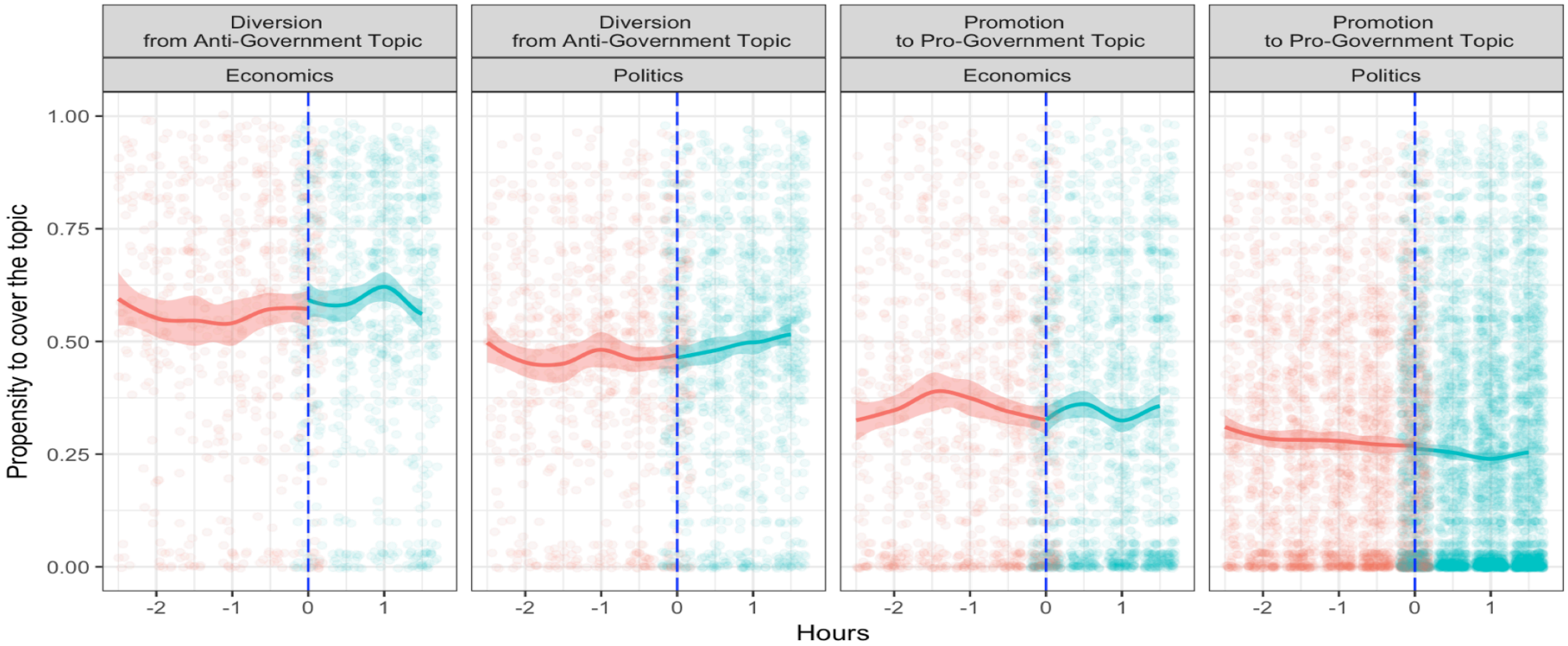


Figure: Effect of a random user

Threats to the validity of the results - 2: The incomplete list of trolls

- The actual trolls who did not appear on the *Novaya Gazeta* list were a treated as regular participants of an online conversation
  - False positive results

Threats to the validity of the results - 2: The incomplete list of trolls

- The actual trolls who did not appear on the *Novaya Gazeta* list were a treated as regular participants of an online conversation
  - False positive results
  - Evidence of the opposite:
    - *From the leaked documents:* four groups of trolls; they worked in twelve hours shifts (a standard schedule); fixed number of posts

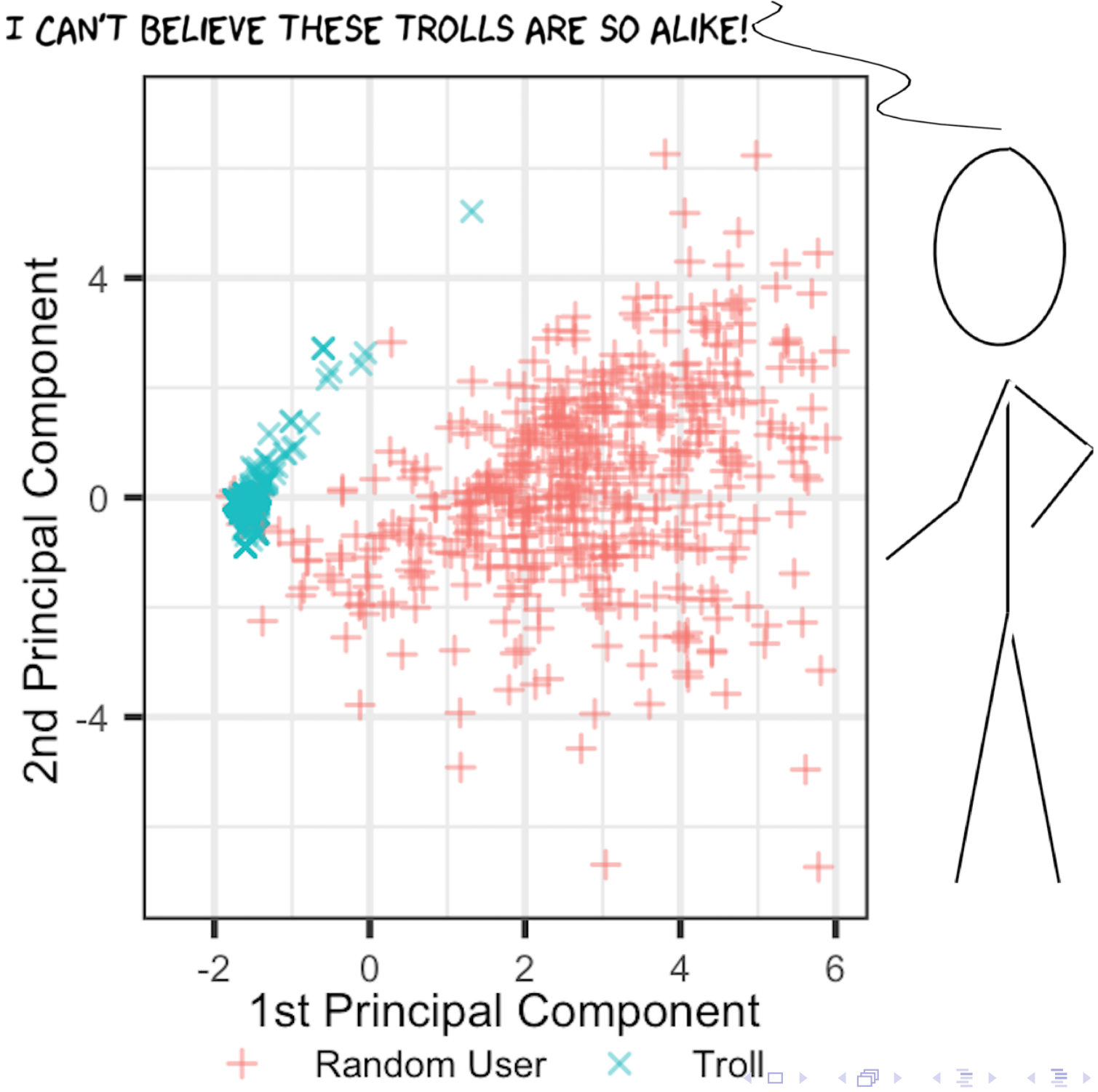
Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Group 1	Group 3	Group 1	Group 3	Group 1	Group 3	Group 1
Group 2	Group 4	Group 2	Group 4	Group 2	Group 4	Group 2

Table: Potential working schedule of trolls' shifts

- *From data:* the trolls published approximately the same number of posts on each day and night of the week

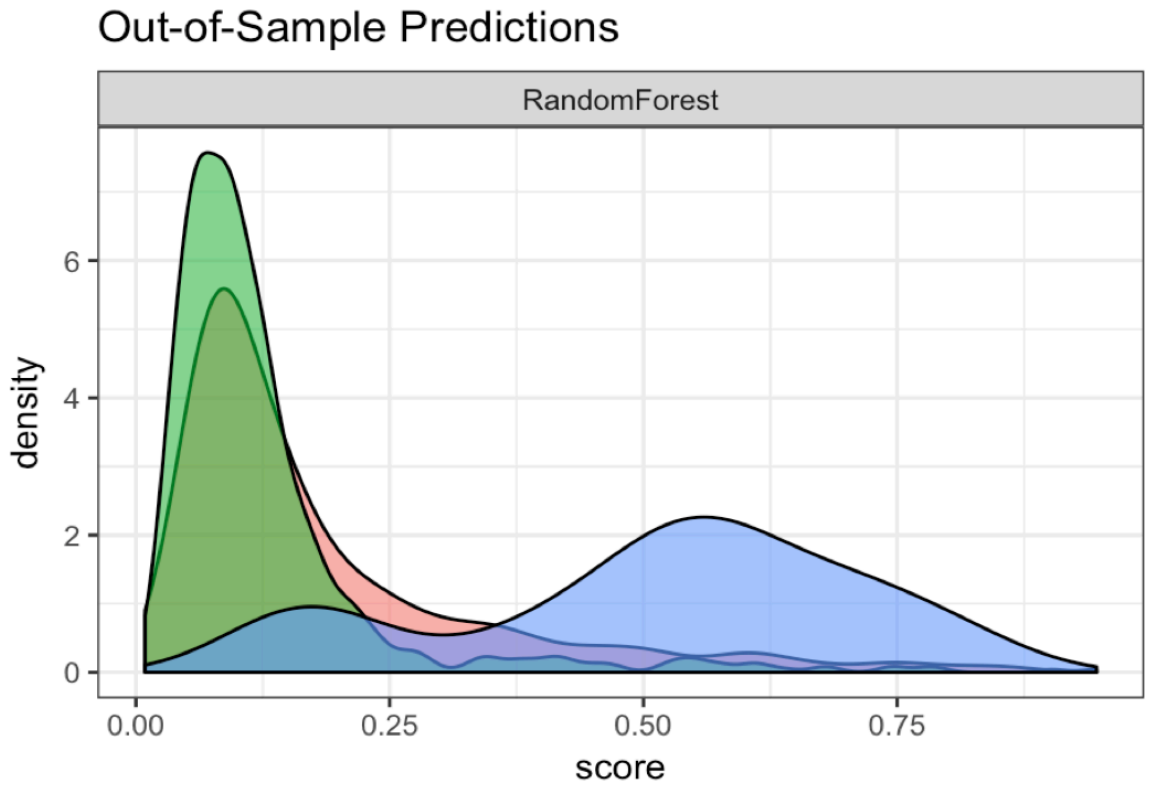
Threats to the validity of the results - 3: The incomplete list of trolls

- The overall number of troll accounts should be negligible in comparison to the overall *LJ* community
  - Random sample 900 Cyrillic *LJ* accounts
- Assume a randomly drawn account does not belong to a troll
- Collect data from these accounts (300k posts) from 2014-15 and bundle with the data on posts made by trolls (500k posts)
- Feature extraction: TF-IDF, LDA, user-level features
- Train classification models
- Sample 650 non-troll participants of online conversations
- Apply the trained model to calculate their propensity scores to be *de-facto* trolls


I CAN'T BELIEVE THESE TROLLS ARE SO ALIKE!
2nd Principal Component
4
0
-4
-2
0
2
4
6
1st Principal Component
Random User
Troll

Threats to the validity of the results - 3: The incomplete list of trolls

- Less than 3% percent of the participants of the targeted conversations have a propensity to be a troll higher that 50%,
- Less than .5% have a propensity that exceeds 60%.



Conclusions

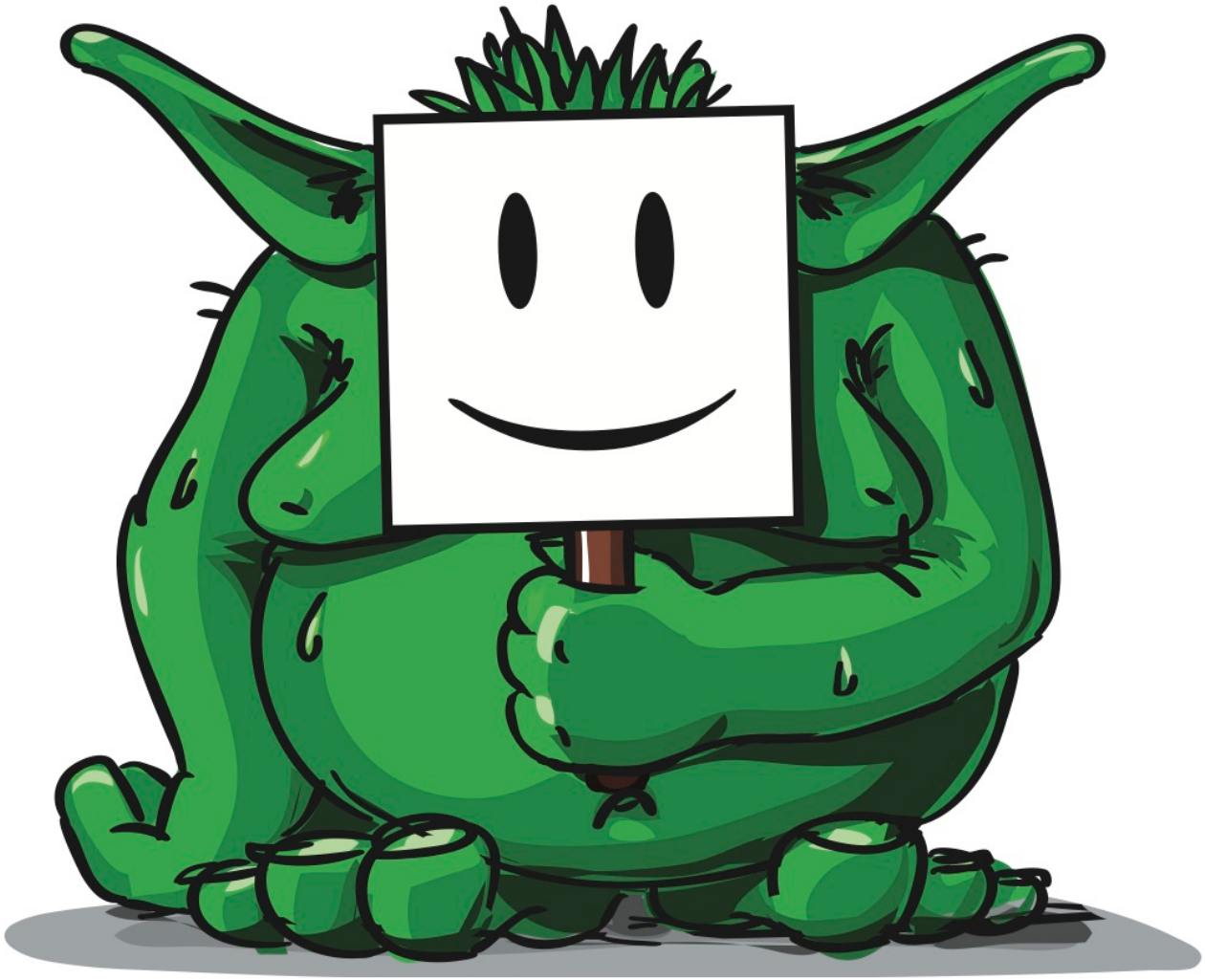
- "Fake agents" are believed to be important limit the ability of citizens to learn public preferences and to coordinate
- "Fake agents"appear to be successful in diverting the discussions from critically charged topics if they discuss political affairs, but not the problems of national economy
- Evidence of no promotion effect of pro-government agenda

Conclusions

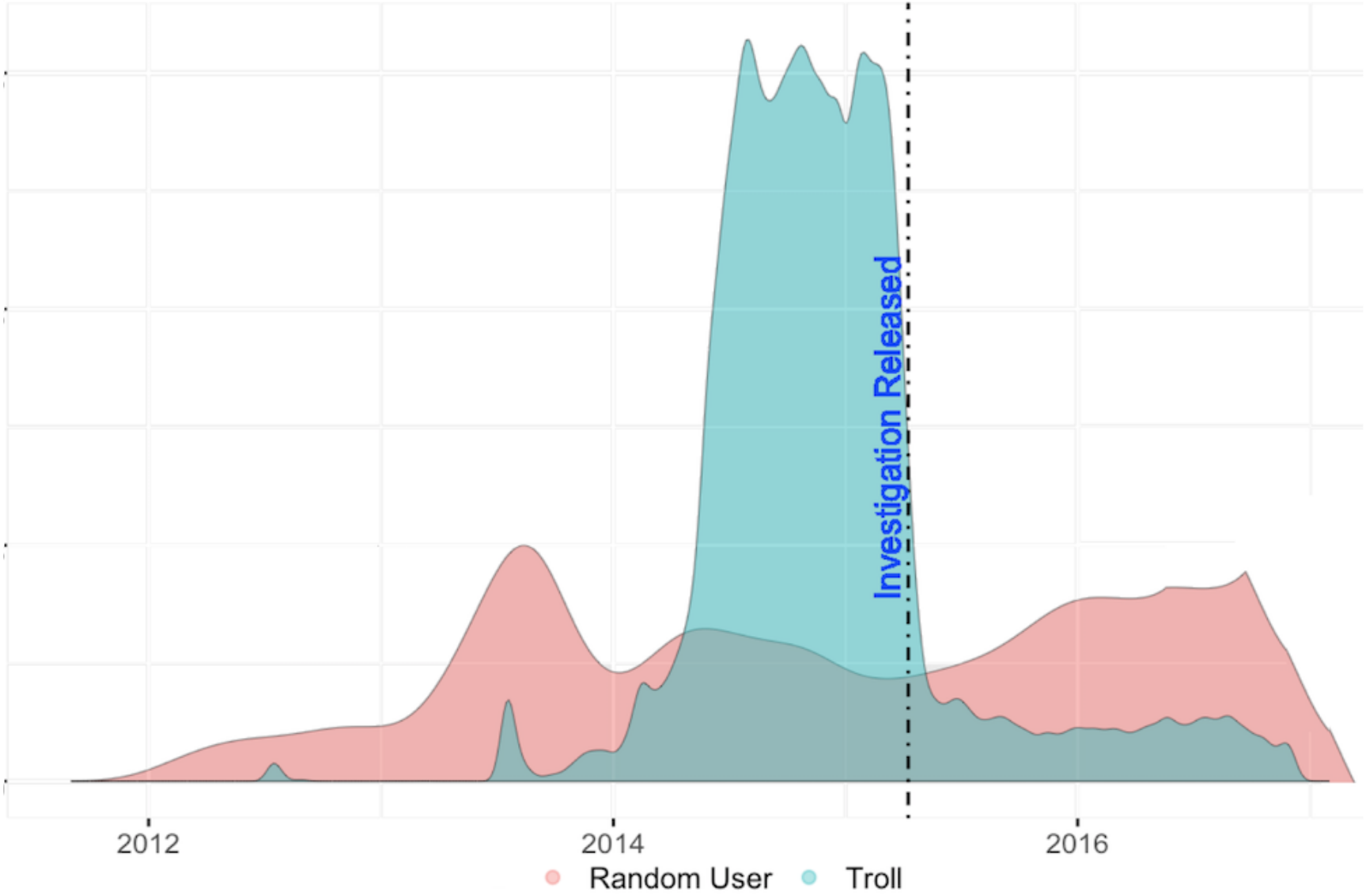
- "Fake agents" are believed to be important limit the ability of citizens to learn public preferences and to coordinate
- "Fake agents"appear to be successful in diverting the discussions from critically charged topics if they discuss political affairs, but not the problems of national economy
- Evidence of no promotion effect of pro-government agenda
- *Why trolls succeed in political, not economic conversations?*
  - Private information in discussions of economic issues
- The analysis is limited in scope:
  - it considers only two potential effects of trolls' interventions in online conversations
  - it does not consider effects of such interventions on a larger audience of social media users
  - does not provide the evidence that trolls can change the preferences or the offline political behavior of users

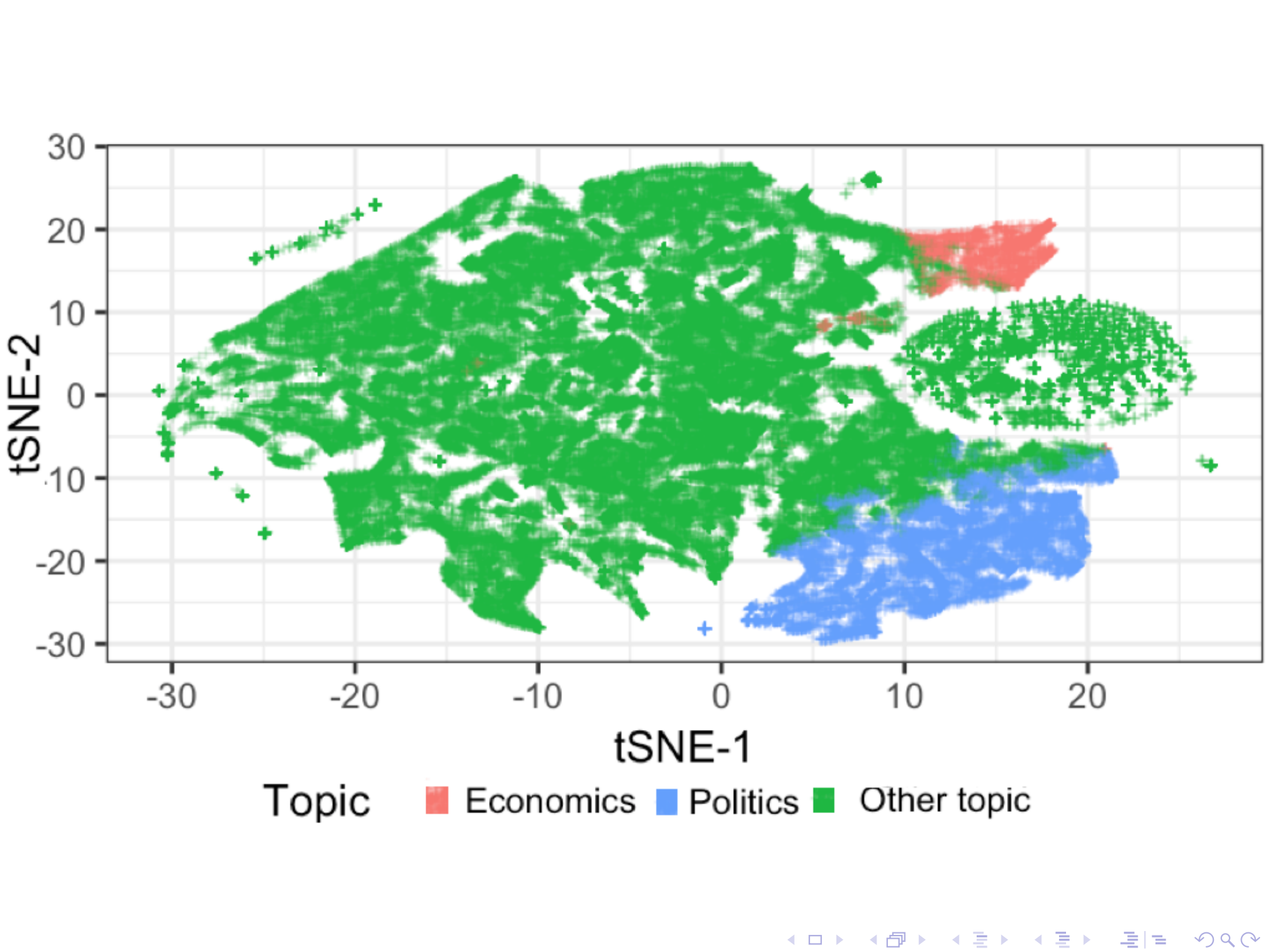
THANK YOU FOR YOUR PATIENCE!

WWW.ASOBOLEV.COM



Right on time!




30
20
10
0
-10
-20
-30
tSNE-2
-30
-20
-10
0
10
20
tSNE-1
Topic
Economics
Politics
Other topic