

Bridging Nations: Quantifying the Role of Multilinguals in Communication on Social Media

Julia Mendelsohn

🌐 juliamentelsohn.github.io
🐦 [@jmendezsohn2](https://twitter.com/jmendezsohn2)
✉️ juliame@umich.edu



Sayan Ghosh

🐦 [@sayan__ghosh](https://twitter.com/sayan__ghosh)
✉️ ghoshsay@usc.edu



David Jurgens

🌐 jurgens.people.si.umich.edu
🐦 [@david__jurgens](https://twitter.com/david__jurgens)
✉️ jurgens@umich.edu



Ceren Budak

🌐 cbudak.com
🐦 [@cerenbudak](https://twitter.com/cerenbudak)
✉️ cbudak@umich.edu



Information shared on social media can spread quickly across the globe, crossing linguistic and national borders.



Information shared on social media can spread quickly across the globe, crossing linguistic and national borders.

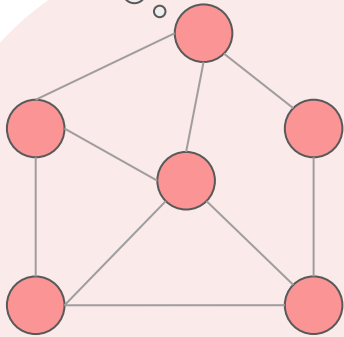


But we don't really know how...

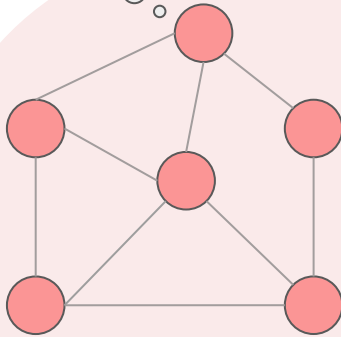


We quantify the **role of multilinguals** in cross-lingual information exchange on European Twitter

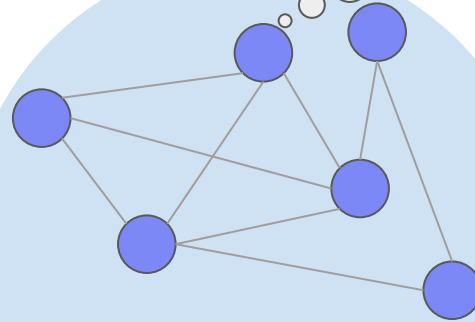
Hallo!
#cdu



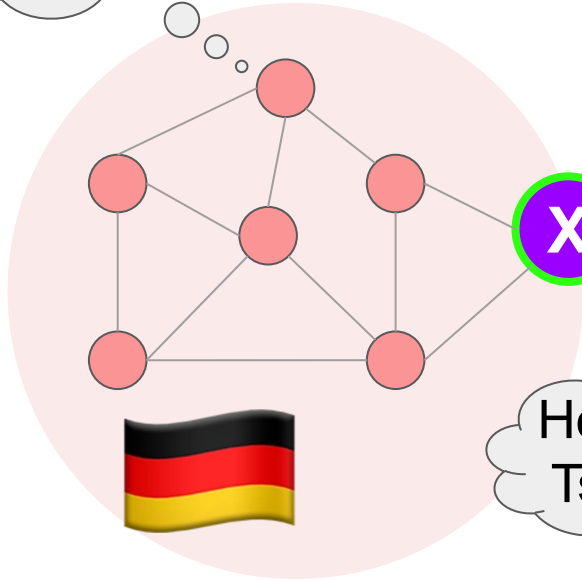
Hallo!
#cdu



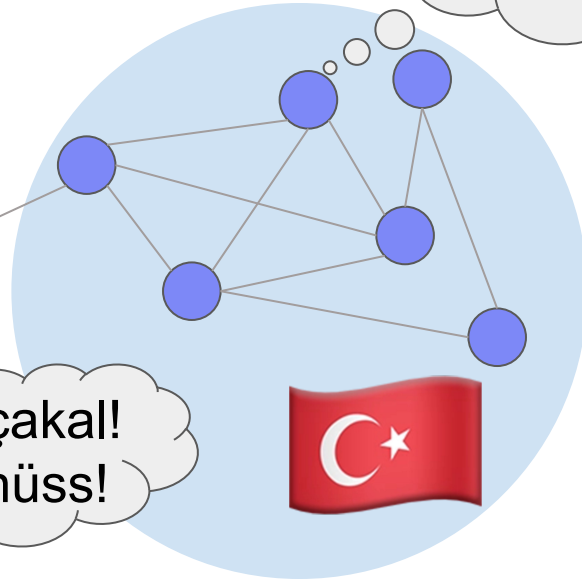
Merhaba!
#pazartesi



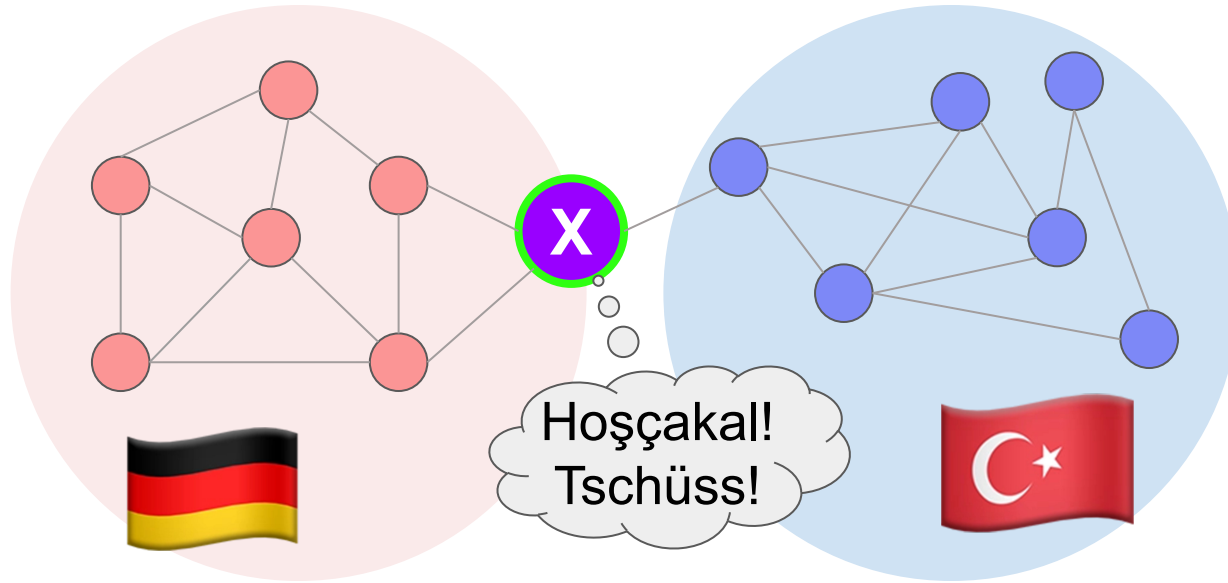
Hallo!
#cdu



Merhaba!
#pazartesi

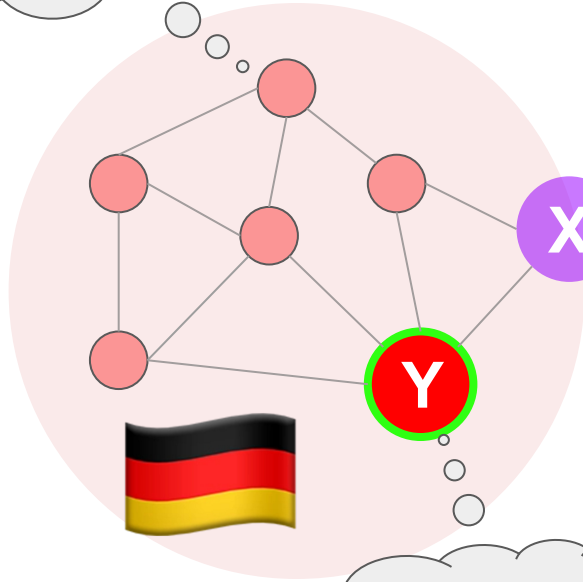


Hoşçakal!
Tschüss!

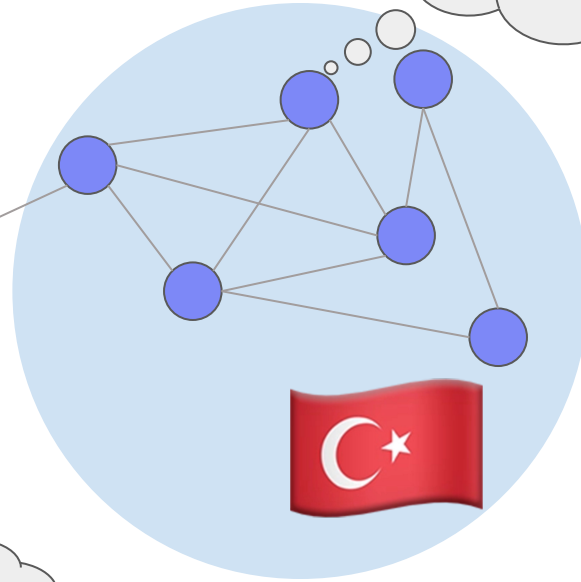


Structural Role: To what extent are multilingual users bridges (positioned to share information)?

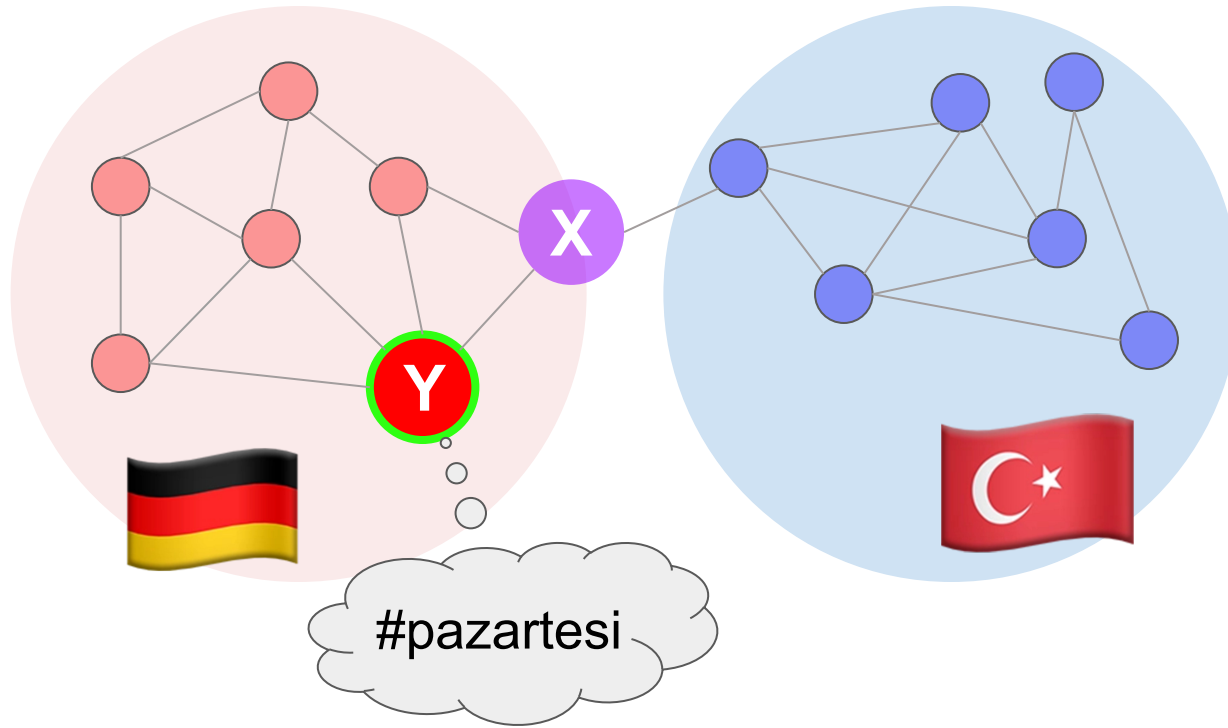
Hallo!
#cdu



Merhaba!
#pazartesi

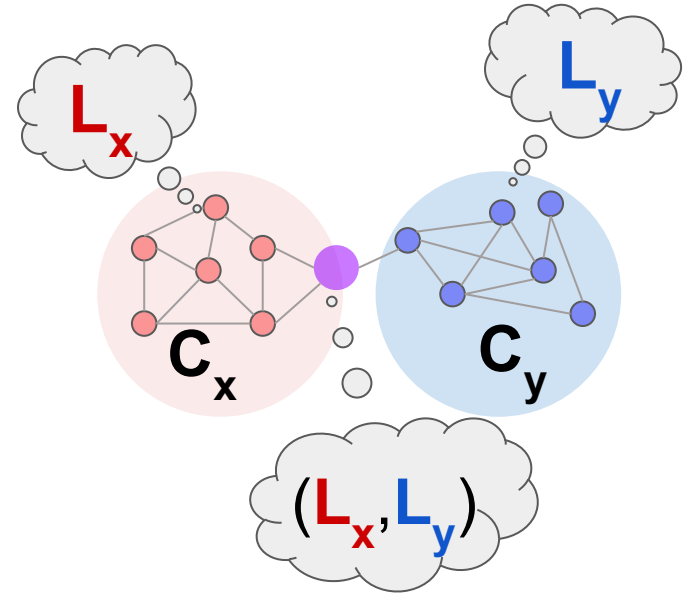


#pazartesi



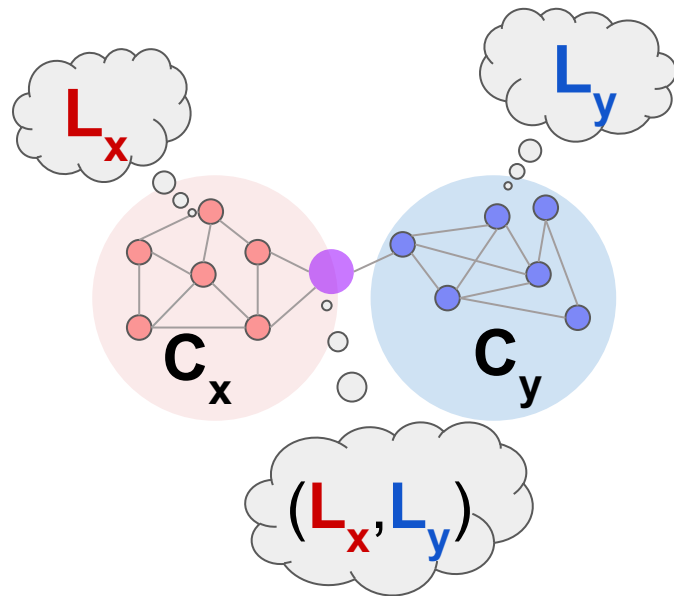
Communication Influence: How does having a multilingual friend impact one's sharing behavior?

Multilingual country pair (MCP) networks



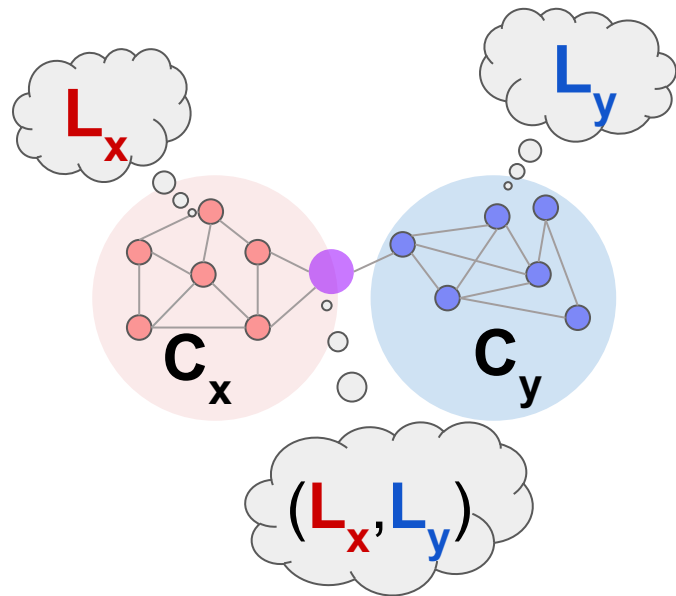
Multilingual country pair (MCP) networks

- Undirected network of mutual mentions from Decahose



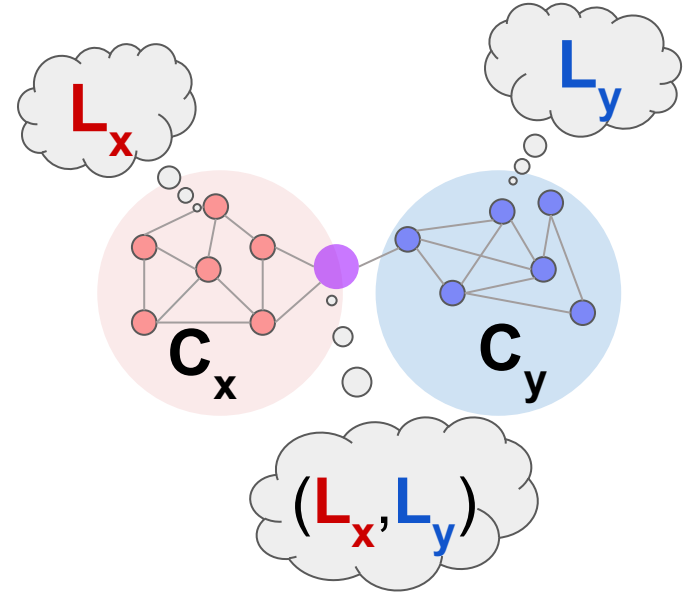
Multilingual country pair (MCP) networks

- Undirected network of mutual mentions from Decahose
- Location inference to get nodes and edges from $C_x \cup C_y$



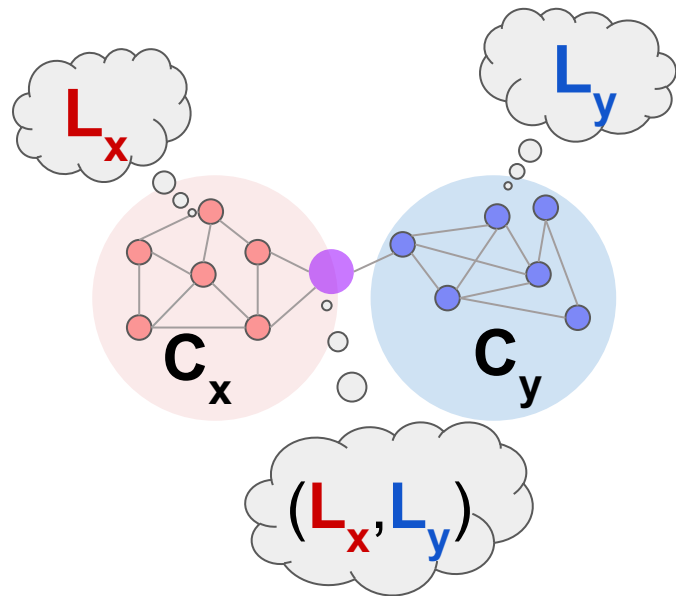
Multilingual country pair (MCP) networks

- Undirected network of mutual mentions from Decahose
- Location inference to get nodes and edges from $C_x \cup C_y$
- C_x & C_y must have single, distinct dominant languages



Multilingual country pair (MCP) networks

- Undirected network of mutual mentions from Decahose
- Location inference to get nodes and edges from $C_x \cup C_y$
- C_x & C_y must have single, distinct dominant languages
- ~250 MCPs from Europe



Identifying multilingual Twitter users

- We determine a user's language use based on tweet text using Twitter's LangID

Identifying multilingual Twitter users

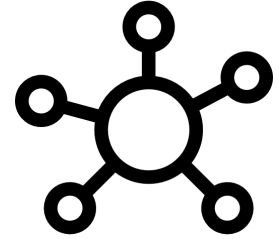
- We determine a user's language use based on tweet text using Twitter's LangID
- A multilingual user of L_x and L_y has at least 10% of their tweets in each language

Identifying multilingual Twitter users

- We determine a user's language use based on tweet text using Twitter's LangID
- A multilingual user of L_x and L_y has at least 10% of their tweets in each language
- This captures language *performances* on Twitter. We know nothing about offline multilingualism

Measuring **structural role**

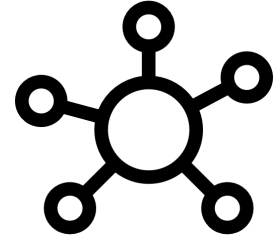
To what extent are multilingual users bridges (positioned to share information)?



Unit	
Treatment	
Outcome	

Measuring **structural role**

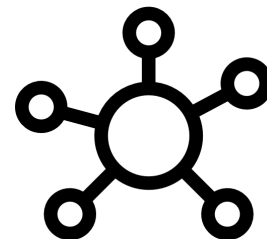
To what extent are multilingual users bridges (positioned to share information)?



Unit	Users from C_x who post in L_x
Treatment	
Outcome	

Measuring **structural role**

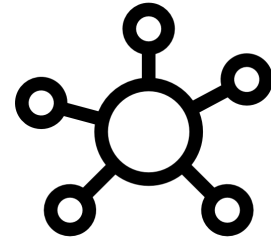
To what extent are multilingual users bridges (positioned to share information)?



Unit	Users from C_x who post in L_x
Treatment	Posting in L_x and L_y
Outcome	

Measuring **structural role**

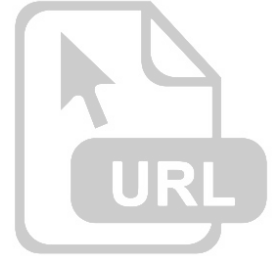
To what extent are multilingual users bridges (positioned to share information)?



Unit	Users from C_x who post in L_x
Treatment	Posting in L_x and L_y
Outcome	Betweenness centrality (log)

Measuring **communication influence**

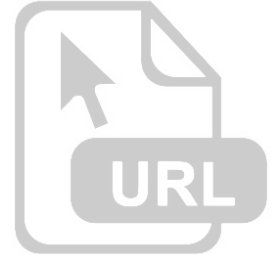
How does having a multilingual friend impact one's sharing behavior?



Unit	
Treatment	
Outcome	

Measuring **communication influence**

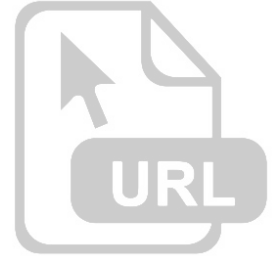
How does having a multilingual friend impact one's sharing behavior?



Unit	Monolinguals from C_x who use L_x
Treatment	
Outcome	

Measuring **communication influence**

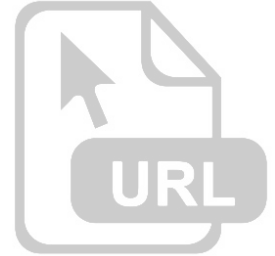
How does having a multilingual friend impact one's sharing behavior?



Unit	Monolinguals from C_x who use L_x
Treatment	Having a multilingual friend who posts in L_x and L_y
Outcome	

Measuring **communication influence**



How does having a multilingual friend impact one's sharing behavior?



Unit	Monolinguals from C_x who use L_x
Treatment	Having a multilingual friend who posts in L_x and L_y
Outcome	Sharing a hashtag associated with L_y

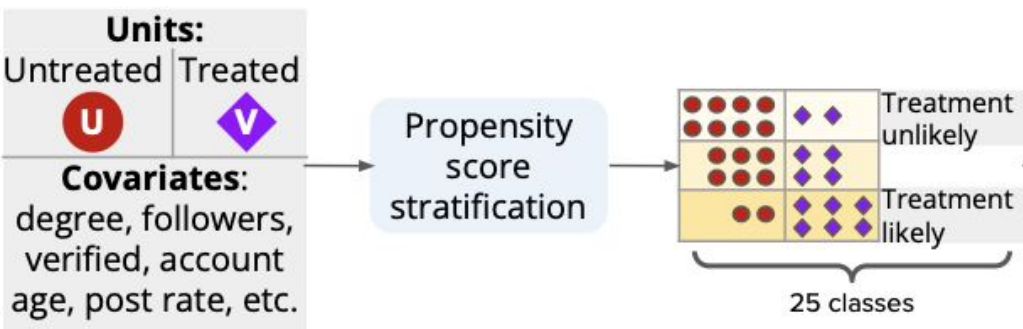
Causal Inference Design

How would **betweenness centrality** be different in a counterfactual world where a **multilingual** were **monolingual** instead?

Units:	
Untreated	Treated
	
Covariates:	
degree, followers, verified, account age, post rate, etc.	

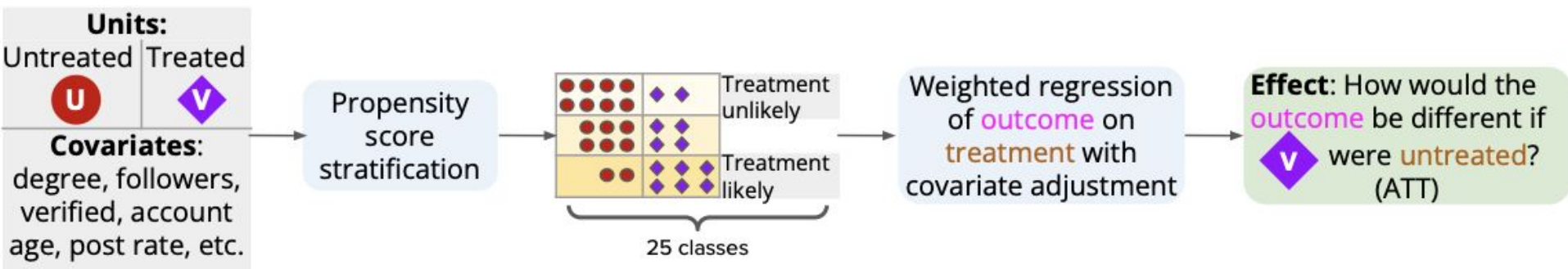
Causal Inference Design

How would **betweenness centrality** be different in a counterfactual world where a **multilingual** were **monolingual** instead?



Causal Inference Design

How would **betweenness centrality** be different in a counterfactual world where a **multilingual** were **monolingual** instead?



Multilinguals are important!

For a user in C_x , posting in both L_x & L_y increases:

Multilinguals are important!

For a user in C_x , posting in both L_x & L_y increases:

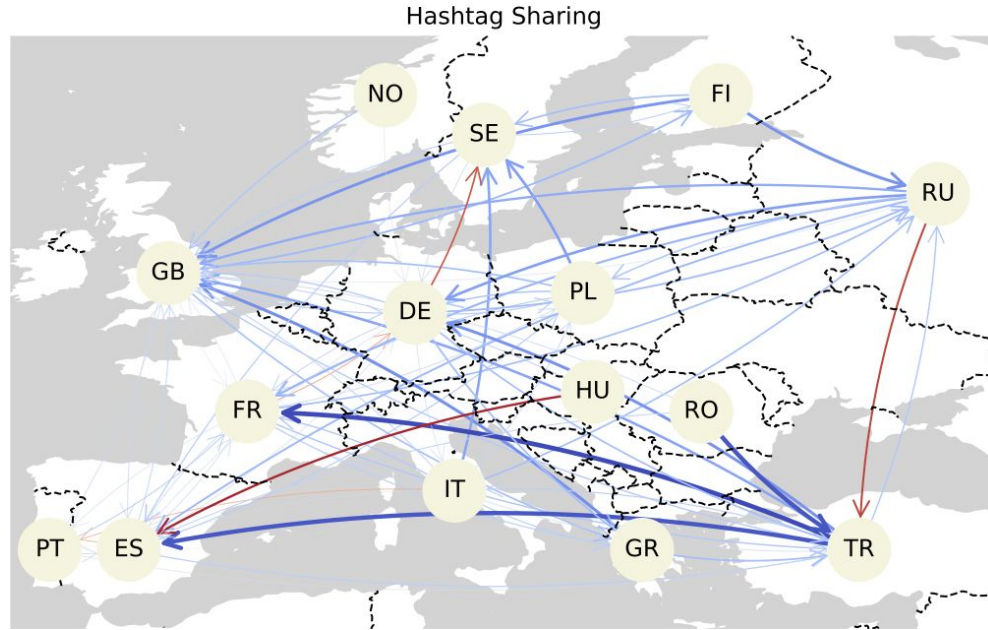
- Betweenness centrality by **13.5%**

Multilinguals are important!

For a user in C_x , posting in both L_x & L_y increases:

- Betweenness centrality by **13.5%**
- Odds of a L_x friend sharing L_y hashtags **4-fold**

But there's a lot of variation across pairs

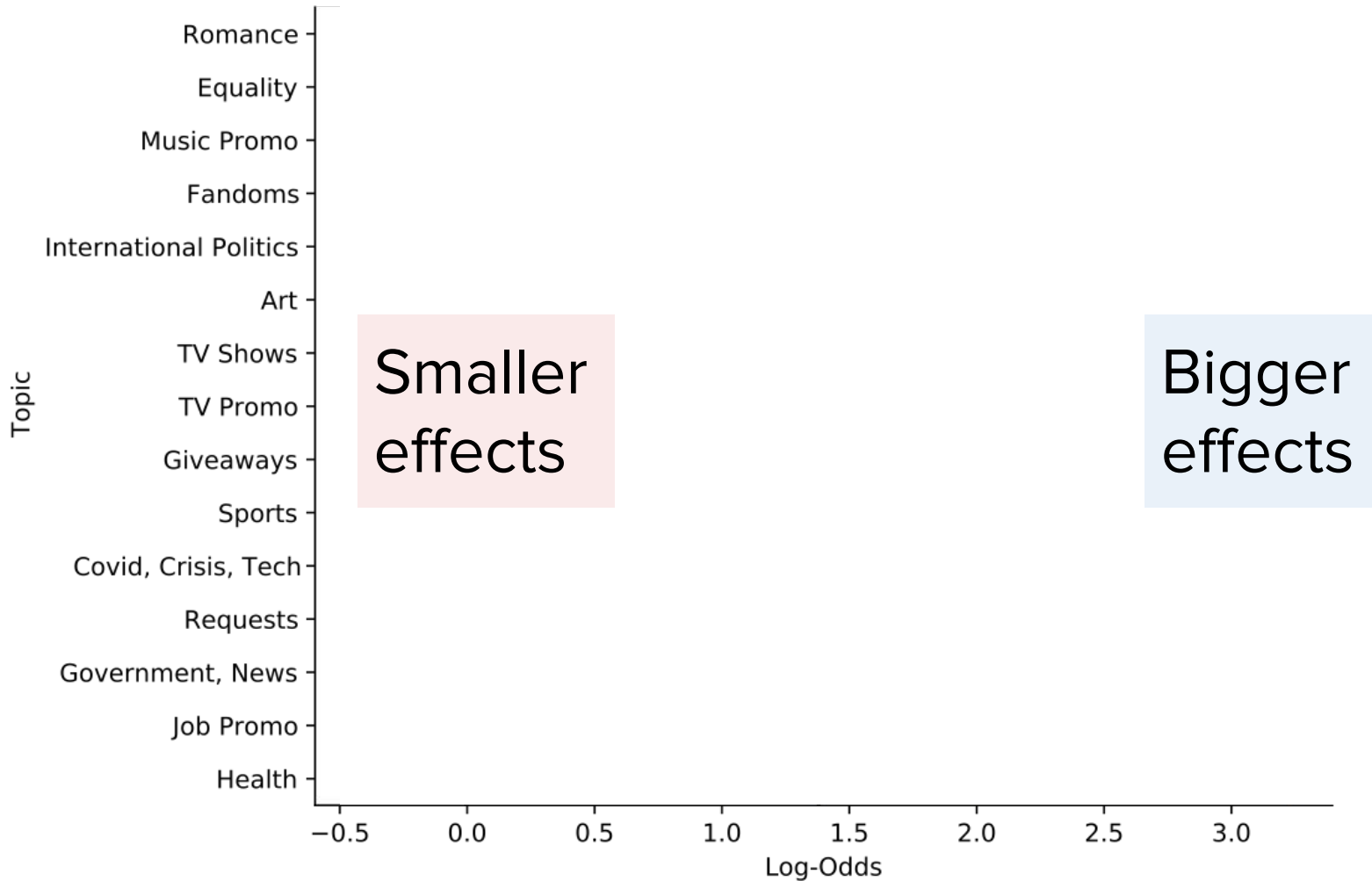


See paper for analyses of variation across geographic, demographic, political, economic, and linguistic relationships between countries

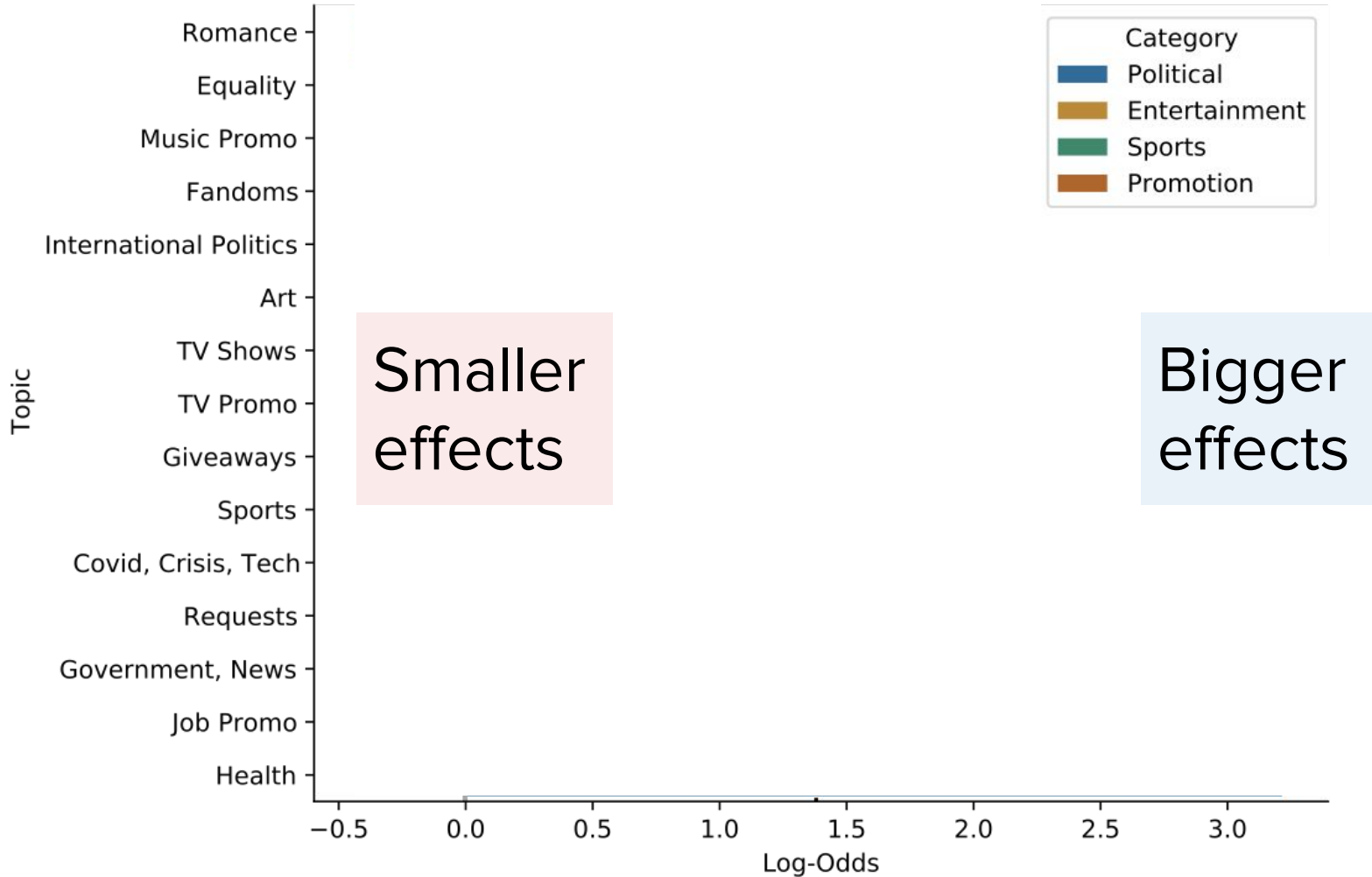
How does the effect of multilinguals varies across **content topic**?

- Multilingual contextualized topic model (CTM) to identify 50 topics [Bianchi et al., 2021]
- Assign hashtags to topic
- Topic-intrusion in 5 languages for evaluation

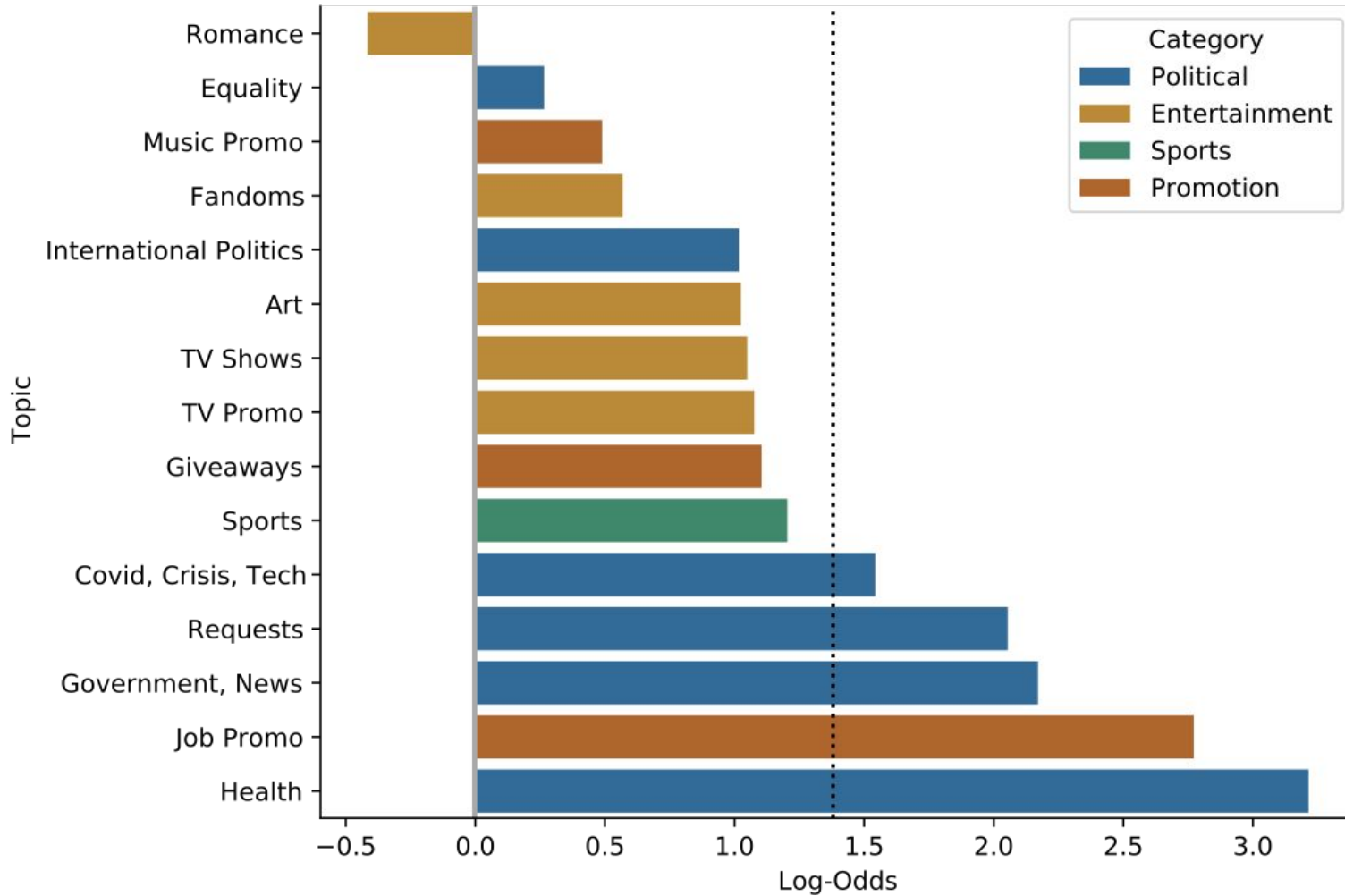
Causal effect of having a multilingual friend by topic



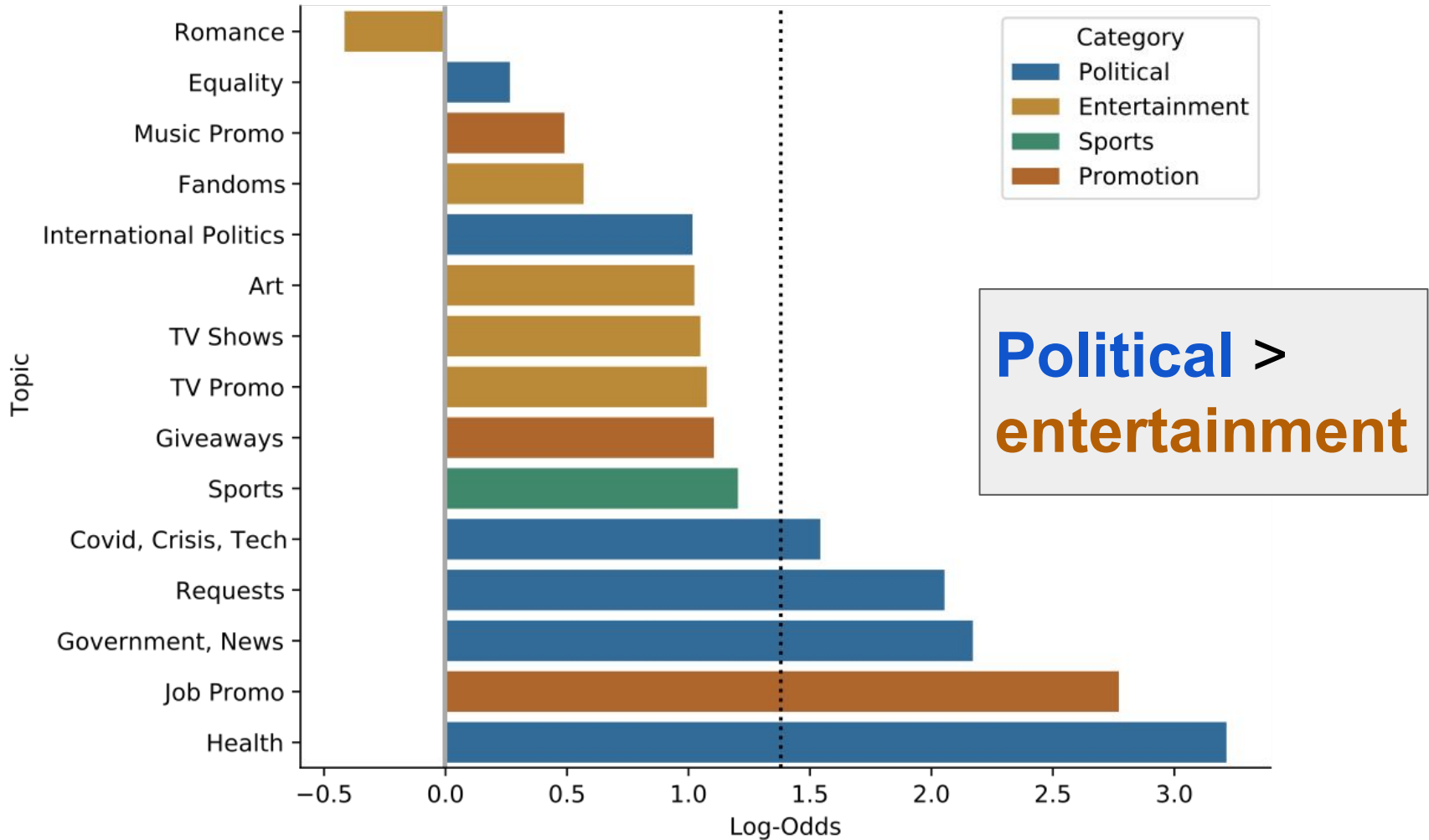
Causal effect of having a multilingual friend by topic



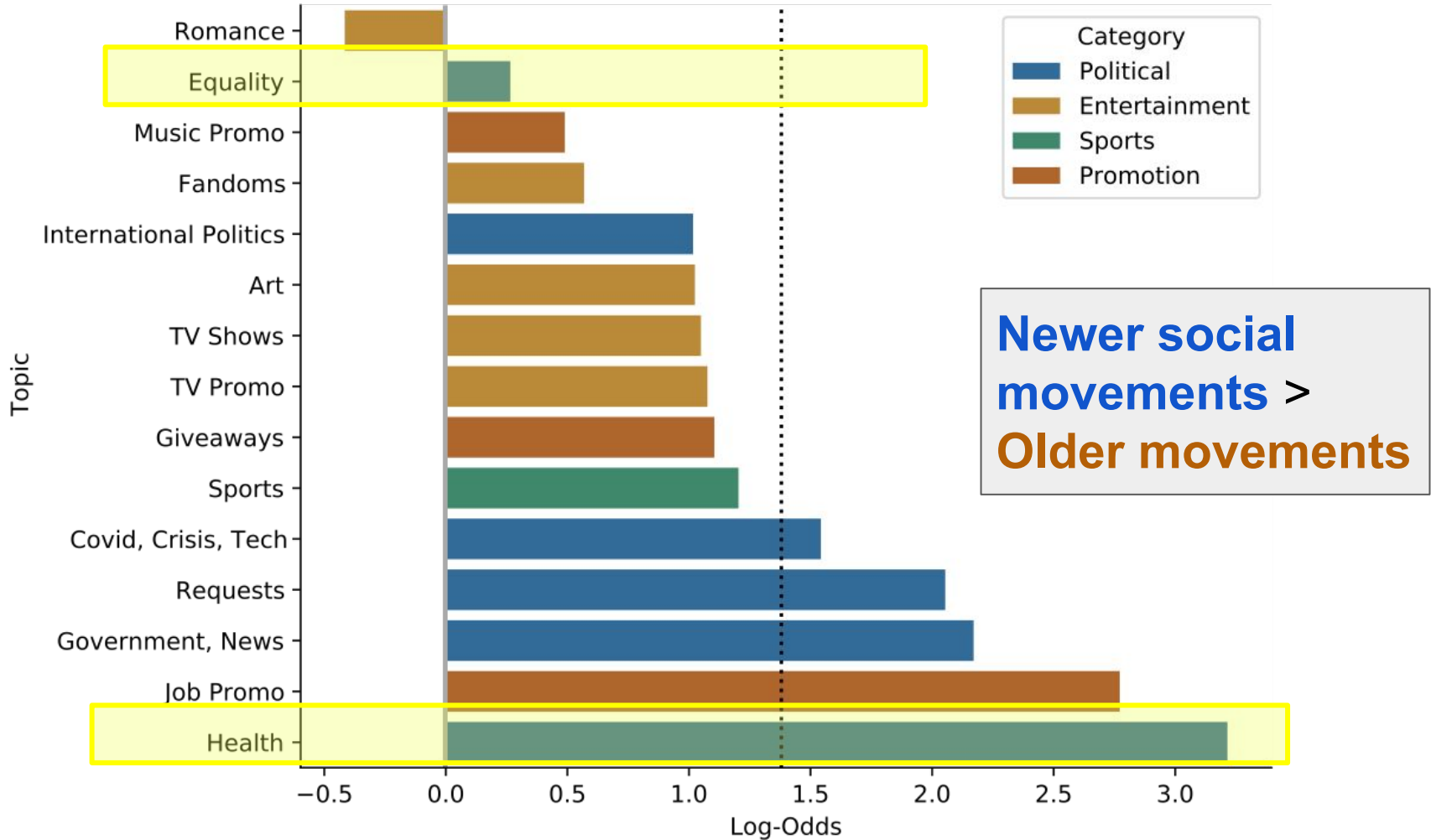
Causal effect of having a multilingual friend by topic



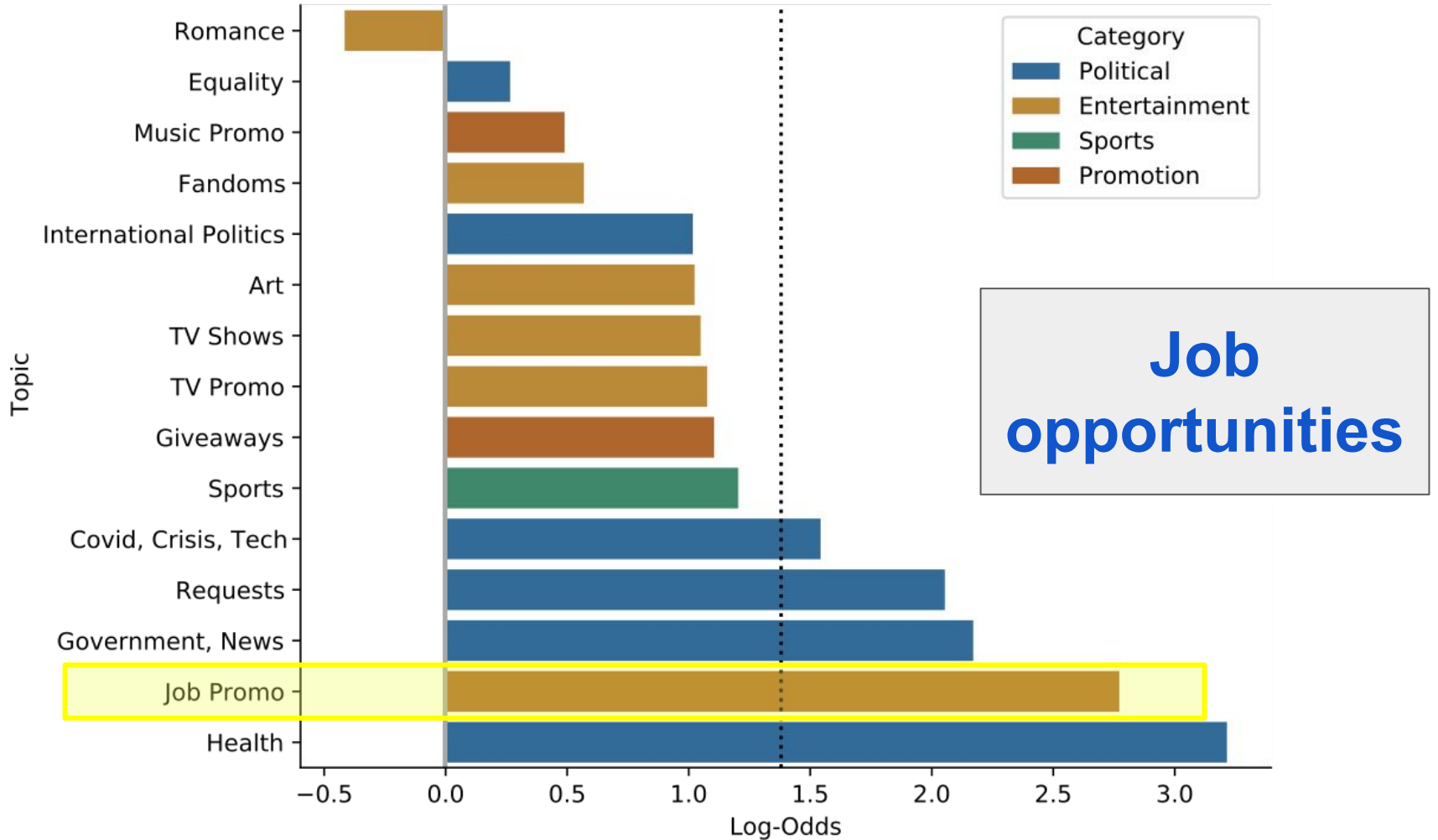
Causal effect of having a multilingual friend by topic



Causal effect of having a multilingual friend by topic



Causal effect of having a multilingual friend by topic



Conclusion

- **Structural role** and **communication influence** of multilinguals in cross-lingual information exchange

Conclusion

- **Structural role** and **communication influence** of multilinguals in cross-lingual information exchange
- Multilinguals play important role overall, with the biggest effects in spreading otherwise less-accessible information

Conclusion

- **Structural role** and **communication influence** of multilinguals in cross-lingual information exchange
- Multilinguals play important role overall, with the biggest effects in spreading otherwise less-accessible information
- Future: nuancing multilingualism, platform implications

Conclusion

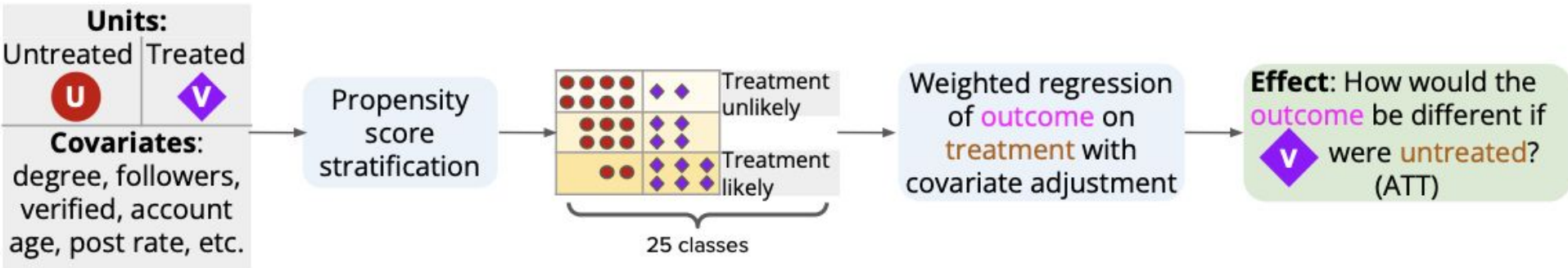
- **Structural role** and **communication influence** of multilinguals in cross-lingual information exchange
- Multilinguals play important role overall, with the biggest effects in spreading otherwise less-accessible information
- Future: nuancing multilingualism, platform implications

Thank you! Ευχαριστώ!

Julia Mendelsohn, Sayan Ghosh, David Jurgens, Ceren Budak

🌐 juliamentelsohn.github.io 🐦 [jmendelsohn2](https://twitter.com/jmendelsohn2) ✉️ juliame@umich.edu

Additional Slides

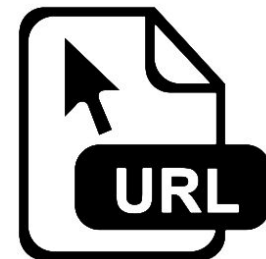


	Study 1	Study 2	Study 3
Units	C_x users posting in L_x	Monolinguals in L_x from C_x	Monolinguals in L_x from C_x
Treatment	Posting multilingually in (L_x, L_y)	Having ≥ 1 (L_x, L_y) multilingual neighbor	Having ≥ 1 (L_x, L_y) multilingual neighbor
Outcome	Betweenness centrality	Sharing domain from L_y	Sharing hashtag from L_y

Table S3: Intercoder agreement (Krippendorff’s α) between each pair of LangID models. Bolded values represent the highest agreement score for each column. Twitter’s LangID model has the highest average agreement with other models.

Model	Twitter	fastText	langid.py	langdetect	CLD2	CLD3
Twitter	-	0.87	0.84	0.82	0.82	0.77
fastText	0.87	-	0.86	0.80	0.79	0.75
langid.py	0.84	0.86	-	0.79	0.76	0.74
langdetect	0.82	0.80	0.79	-	0.70	0.70
CLD2	0.82	0.79	0.76	0.70	-	0.71
CLD3	0.77	0.75	0.74	0.70	0.71	-
Mean	0.824	0.814	0.798	0.763	0.760	0.734

Identifying domains by language



- We include retweets to determine top domains/hashtags by language and users' sharing of domains/hashtags

Identifying hashtags by language



- Trending hashtags change fast, so top hashtags calculated for 14 day intervals.
- A user shares a hashtag from language L if they share a top L hashtag from interval i during interval i or $i+1$

Table S4: Examples of hashtags (from one selected time interval) and domains associated with five different languages.

German	Portuguese	Turkish	Polish	English
cdu	fcporto	çağlarertuğrul	pis	oddoneout
spd	todosportugal	sustunuz	konwencjapis	remain
merkel	capricórnio	pazartesi	topmodel	eastenders
klimaschutz	aquário	cumartesi	thevoiceofpoland	liarjohnson
noafd	sportingcp	burcuözberk	kaczyński	ncfc
tagesschau.de	publico.pt	tele1.com.tr	wpolityce.pl	manchestereveningnews.co.uk
faz.net	record.pt	haber.sol.org.tr	niezalezna.pl	whounfollowedme.org
spiegel.de	maisfutebol.iol.pt	diken.com.tr	dorzeczy.pl	theneweuropen.co.uk

But there's a lot of variation across pairs

	Betweenness Centrality (Study 1)	Domain Sharing (Study 2)	Hashtag Sharing (Study 3)
# Eligible MCPs	214	158	199
# Eligible Loci	317	205	284
% Loci w/ sig. pos ATT	46.37%	56.10%	50.00%
% Loci w/ no sig. ATT	51.42%	40.49%	46.48%
% Loci w/ sig. neg ATT	2.21%	3.41%	3.52%

But there's a lot of variation across pairs

	Betweenness Centrality (Study 1)	Domain Sharing (Study 2)	Hashtag Sharing (Study 3)
# Eligible MCPs	214	158	199
# Eligible Loci	317	205	284
% Loci w/ sig. pos ATT	46.37%	56.10%	50.00%
% Loci w/ no sig. ATT	51.42%	40.49%	46.48%
% Loci w/ sig. neg ATT	2.21%	3.41%	3.52%

Larger effects when C_x and C_y are far apart

	Betweenness Centrality	Domain Sharing	Hashtag Sharing
Geographic distance	0.020***	0.434	0.412***
Time difference	0.015**	0.815***	0.001
Pop. C_x / C_y	-0.030***	-0.031	0.017
% C_y foreign-born	0.021**	-1.100**	-0.058
% C_y pop. born in C_x	0.017**	-0.044	-0.043
% C_x foreign-born	-0.002	-0.031	0.064
% C_x pop. born in C_y	0.007	-0.197	-0.040
GDP per capita C_x / C_y	0.038***	0.318	1.171***
RTA	0.010	-0.411	0.216*
Tradeflow per capita	-0.013*	0.425	0.109
% C_x 's conflicts vs. C_y	0.019**	-0.159	-0.032
% C_y 's conflicts vs. C_x	0.002	-0.257	0.088
Linguistic distance	-0.027***	0.449	0.146*
Observations	317	205	284
R ²	0.266	0.193	0.448

Western European multilinguals who post in Eastern European languages have an especially big influence

Measuring linguistic closeness

Level 0: no established relationship

- (German, Turkish), (Spanish, Hungarian)

Level 1: same family

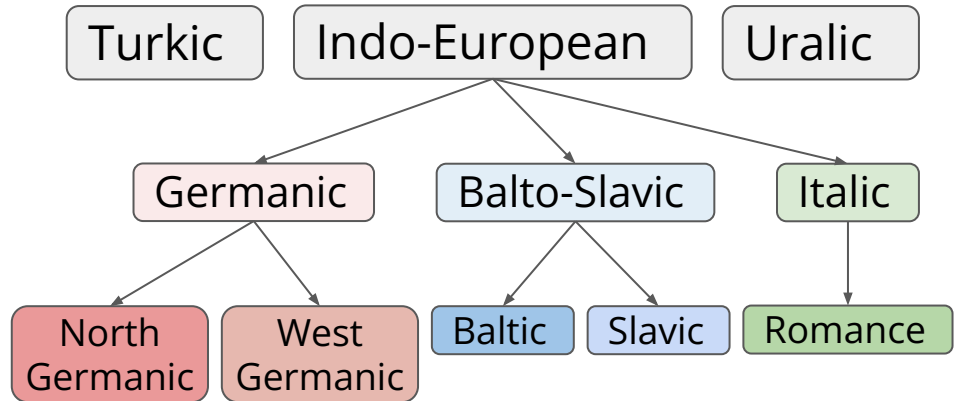
- (German, Polish), (Spanish, Russian)

Level 2: same branch

- (German, Swedish), (Russian, Latvian)

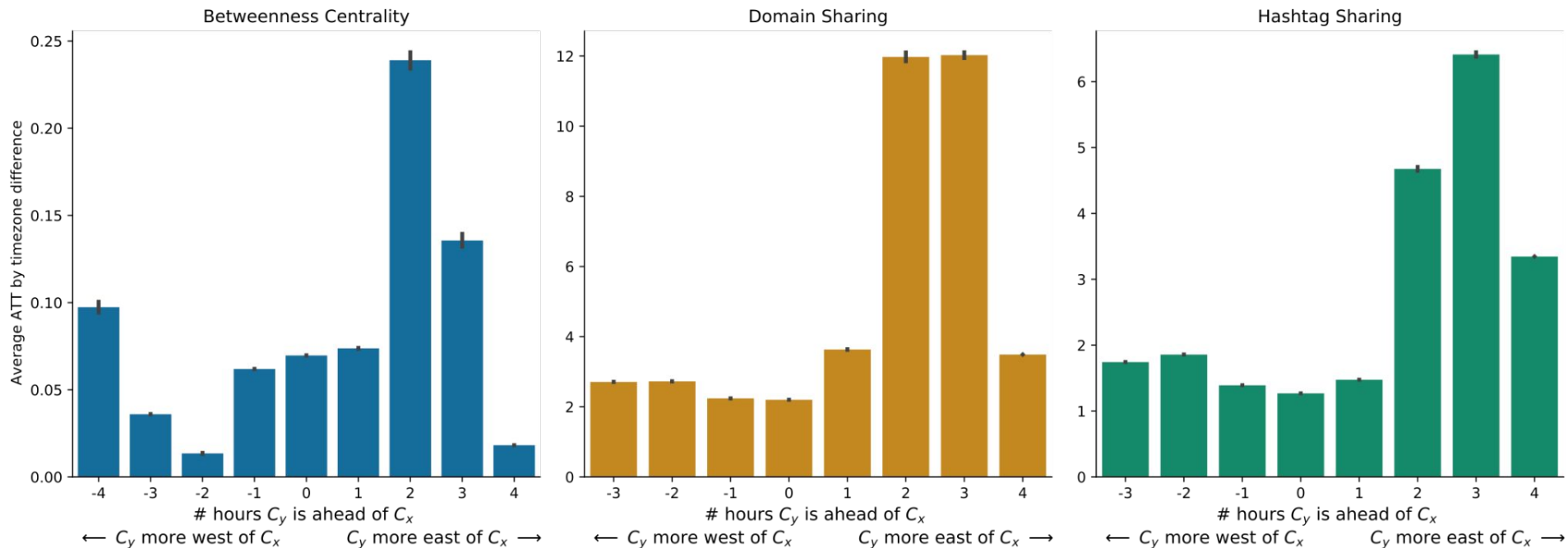
Level 3: same sub-branch

- (Spanish, Portuguese), (Russian, Polish)



Does not consider similarities due to language contact (e.g. lexical borrowing)

Time differences and ATT



Identifying topics in cross-lingual diffusion

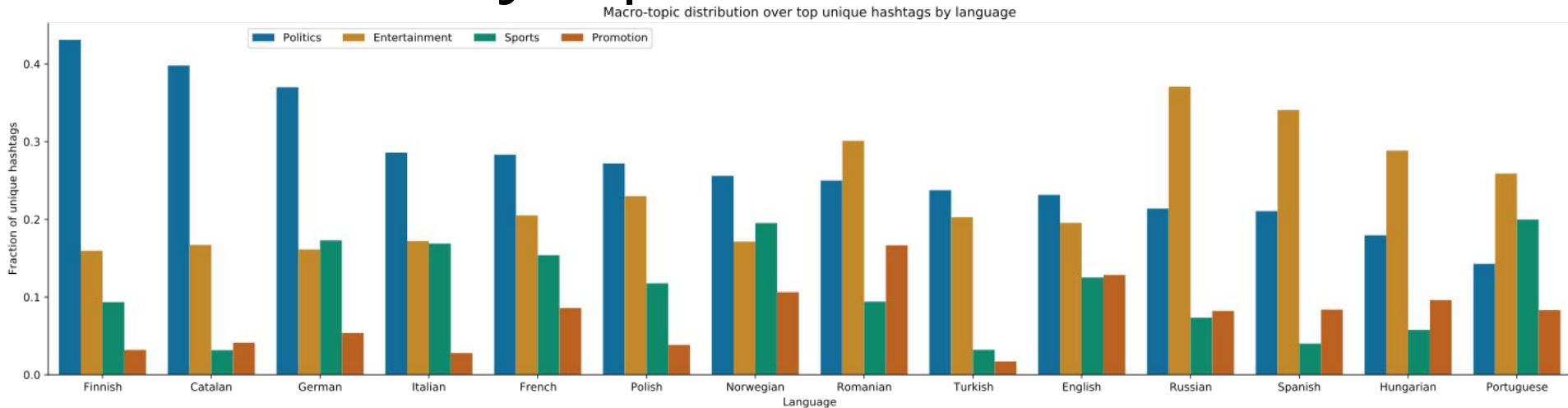
- Train multilingual contextualized topic model (CTM) to identify 50 topics (Bianchi et al., 2021)
- Assign tweets to the most highly-weighted topic
- Assign hashtags to the most common topic of tweets in which they appear.
- Evaluated with topic-intrusion test in 5 languages
- Re-run hashtag diffusion study for each topic

Hashtag intrusion test evaluation

Language	English	Spanish	German	Turkish	Italian
Accuracy	0.731	0.747	0.813	0.587	0.627
Krippendorff α	0.670	0.682	0.767	0.484	0.536

ID	Description	Example Hashtags
1	TV Shows	<i>skamitalia, masterchefgr, thearchers, ibes</i>
10	Fandoms	<i>jungkook, choicefandom, saveshadowhunters</i>
19	Art	<i>painting, etsy, vintage, fasion, architecture, arte</i>
24	Romance TV	<i>loveisland, poweroflovegr, liebesgeschichten</i>
44	TV Promo	<i>comingsoon, luciferonnetflix, skyupnext</i>
3	Job Promo	<i>career, hiring, jobs, startup, sales, jobsearch</i>
16	Giveaways	<i>giveaway, freebiefriday, sorteo, winwin, free</i>
31	Music Promo	<i>radio, youtube, hits, newmusic, magicfm, live</i>
11	Government, News	<i>bbcnews, afd, noafd, labour, parlament, orban</i>
23	Covid, Crisis, Tech	<i>covid19france, koronawirus, polizei, tech, gdpr</i>
30	International Politics	<i>france, syria, venezuela, eeuu, isis, migranti</i>
41	Health	<i>mentalhealth, autism, clapfornhs, discapacidad</i>
46	Requests	<i>stop, shoplocal, stopgiletsjaunes, helpme</i>
47	Equality	<i>metoo, 8demarzo, weltfrauentag, racisme, lgbt</i>
48	Sports	<i>arsenal, halamadrid, futbol, fcporto, rusia2018</i>

Cross-country topical variation



The distribution of hashtags in these macro-topics varies substantially across countries, thus presenting a possible confound when considering country pair-level effects.