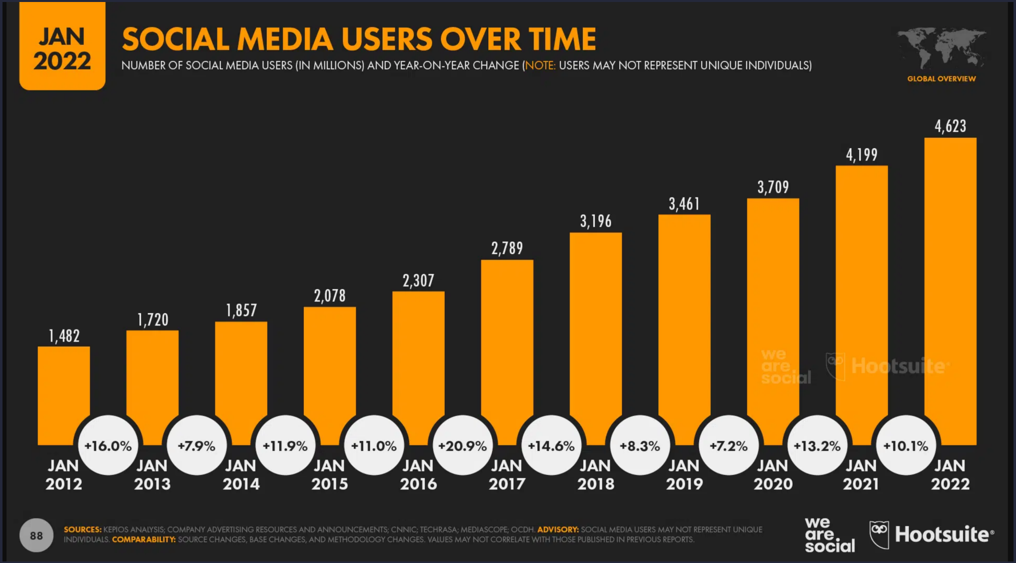
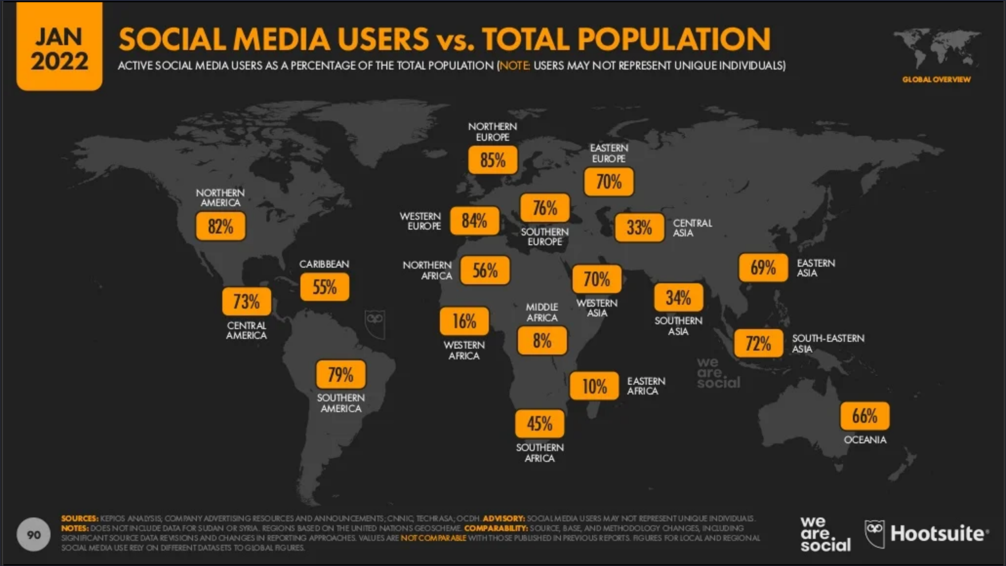
Social Media Data

DSIER [/dɪˈzaɪər/]

Julian HInz and Irene Iodice

Bielefeld University


JAN 2022
SOCIAL MEDIA USERS OVER TIME
NUMBER OF SOCIAL MEDIA USERS (IN MILLIONS) AND YEAR-ON-YEAR CHANGE (NOTE: USERS MAY NOT REPRESENT UNIQUE INDIVIDUALS)
GLOBAL OVERVIEW
1,482
1,720
1,857
2,078
2,307
2,789
3,196
3,461
3,709
4,199
4,623
+16.0%
+7.9%
+11.9%
+11.0%
+20.9%
+14.6%
+8.3%
+7.2%
+13.2%
+10.1%
JAN 2012
JAN 2013
JAN 2014
JAN 2015
JAN 2016
JAN 2017
JAN 2018
JAN 2019
JAN 2020
JAN 2021
JAN 2022
88
SOURCES: KEPIOS ANALYSIS; COMPANY ADVERTISING RESOURCES AND ANNOUNCEMENTS; CNNIC; TECHRASA; MEDIASCOPE; OCDH. ADVISORY: SOCIAL MEDIA USERS MAY NOT REPRESENT UNIQUE INDIVIDUALS. COMPARABILITY: SOURCE CHANGES, BASE CHANGES, AND METHODOLOGY CHANGES. VALUES MAY NOT CORRELATE WITH THOSE PUBLISHED IN PREVIOUS REPORTS.
we are social
Hootsuite

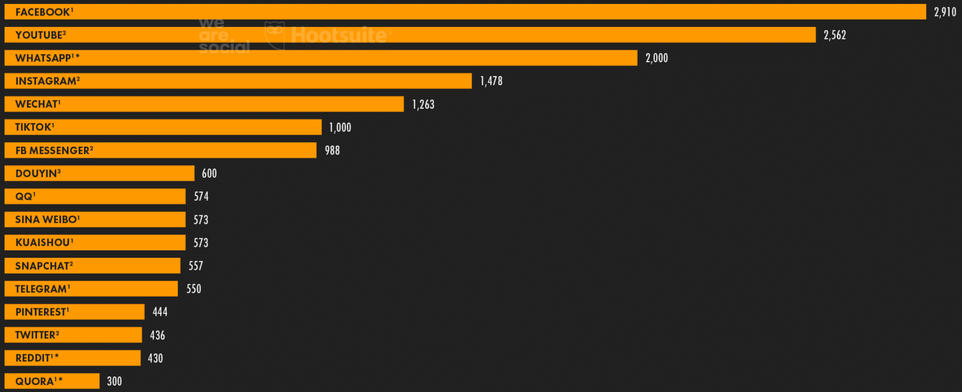

JAN 2022
SOCIAL MEDIA USERS vs. TOTAL POPULATION
ACTIVE SOCIAL MEDIA USERS AS A PERCENTAGE OF THE TOTAL POPULATION (NOTE: USERS MAY NOT REPRESENT UNIQUE INDIVIDUALS)
GLOBAL OVERVIEW
NORTHERN AMERICA 82%
CENTRAL AMERICA 73%
CARIBBEAN 55%
SOUTHERN AMERICA 79%
NORTHERN EUROPE 85%
WESTERN EUROPE 84%
SOUTHERN EUROPE 76%
EASTERN EUROPE 70%
NORTHERN AFRICA 56%
WESTERN AFRICA 16%
MIDDLE AFRICA 8%
EASTERN AFRICA 10%
SOUTHERN AFRICA 45%
WESTERN ASIA 70%
CENTRAL ASIA 33%
SOUTHERN ASIA 34%
EASTERN ASIA 69%
SOUTH-EASTERN ASIA 72%
OCEANIA 66%
90
SOURCES: KEPIOS ANALYSIS; COMPANY ADVERTISING RESOURCES AND ANNOUNCEMENTS; CNNIC; TECHRASA; OCDH. ADVISORY: SOCIAL MEDIA USERS MAY NOT REPRESENT UNIQUE INDIVIDUALS. NOTES: DOES NOT INCLUDE DATA FOR SUDAN OR SYRIA. REGIONS BASED ON THE UNITED NATIONS GEOSCHEME. COMPARABILITY: SOURCE, BASE, AND METHODOLOGY CHANGES, INCLUDING SIGNIFICANT SOURCE DATA REVISIONS AND CHANGES IN REPORTING APPROACHES. VALUES ARE NOT COMPARABLE WITH THOSE PUBLISHED IN PREVIOUS REPORTS. FIGURES FOR LOCAL AND REGIONAL SOCIAL MEDIA USE RELY ON DIFFERENT DATASETS TO GLOBAL FIGURES.
we are social
Hootsuite

JAN 2022

THE WORLD'S MOST-USED SOCIAL PLATFORMS

RANKING OF SOCIAL MEDIA PLATFORMS BY GLOBAL ACTIVE USER FIGURES (IN MILLIONS)

GLOBAL OVERVIEW



99

SOURCES: KEPIOS ANALYSIS OF (1) COMPANY ANNOUNCEMENTS OF MONTHLY ACTIVE USERS; (2) PLATFORMS' SELF-SERVICE ADVERTISING RESOURCES; (3) COMPANY ANNOUNCEMENTS OF DAILY ACTIVE USERS (NOTE THAT MONTHLY ACTIVE USER FIGURES MAY BE HIGHER). **ADVISORY:** USERS MAY NOT REPRESENT UNIQUE INDIVIDUALS. **COMPARABILITY:** PLATFORMS IDENTIFIED BY (*) HAVE NOT PUBLISHED UPDATED USER FIGURES IN THE PAST 12 MONTHS, SO FIGURES ARE LESS REPRESENTATIVE. BASE CHANGES AND METHODOLOGY CHANGES; DATA MAY NOT BE DIRECTLY COMPARABLE WITH PREVIOUS REPORTS.

we are social Hootsuite®

Online services

- Twitter, LinkedIn, Facebook, Instagram, TikTok, ...
- Content, but also metadata
- (Used to?) provide *some* data access

  $\rightarrow$ currently in flux

Online services

- Twitter, LinkedIn, Facebook, Instagram, TikTok, ...
- Content, but also metadata
- (Used to?) provide *some* data access

  $\rightarrow$ currently in flux

Online services

- Twitter, LinkedIn, Facebook, Instagram, TikTok, ...
- Content, but also metadata
- (Used to?) provide *some* data access

$\rightarrow$ currently in flux

Online services

- Twitter, LinkedIn, Facebook, Instagram, TikTok, ...
- Content, but also metadata
- (Used to?) provide *some* data access

  $\rightarrow$ currently in flux

Pros and Cons

- Facebook Data

  → large community, representative across income distribution

  → not accessible to users, not representative across age groups

- Twitter data

  → less large community, less representative across income distribution

Pros and Cons

- Facebook Data

  → large community, representative across income distribution

  → not accessible to users, not representative across age groups

- Twitter data

  → less large community, less representative across income distribution

Pros and Cons

- Facebook Data

  $\rightarrow$ large community, representative across income distribution

  $\rightarrow$ not accessible to users, not representative across age groups

- Twitter data

  $\rightarrow$ less large community, less representative across income distribution

Pros and Cons

- Facebook Data

  $\rightarrow$ large community, representative across income distribution

  $\rightarrow$ not accessible to users, not representative across age groups

- Twitter data

  $\rightarrow$ less large community, less representative across income distribution

Pros and Cons

- Facebook Data

  $\rightarrow$ large community, representative across income distribution

  $\rightarrow$ not accessible to users, not representative across age groups

- Twitter data

  $\rightarrow$ less large community, less representative across income distribution

  $\rightarrow$ freely accessible, rich data

Pros and Cons

- Facebook Data

  $\rightarrow$ large community, representative across income distribution

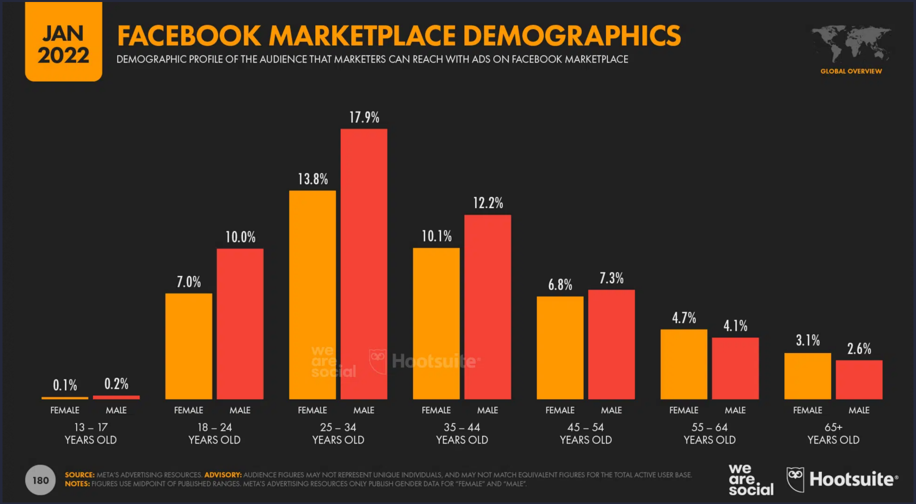
  $\rightarrow$ not accessible to users, not representative across age groups

- Twitter data

  $\rightarrow$ less large community, less representative across income distribution

  $\rightarrow$ ~~freely accessible~~, rich data

FACEBOOK DATA


JAN 2022
FACEBOOK MARKETPLACE DEMOGRAPHICS
DEMOGRAPHIC PROFILE OF THE AUDIENCE THAT MARKETERS CAN REACH WITH ADS ON FACEBOOK MARKETPLACE
GLOBAL OVERVIEW
0.1%
0.2%
7.0%
10.0%
13.8%
17.9%
10.1%
12.2%
6.8%
7.3%
4.7%
4.1%
3.1%
2.6%
FEMALE
MALE
13 – 17 YEARS OLD
18 – 24 YEARS OLD
25 – 34 YEARS OLD
35 – 44 YEARS OLD
45 – 54 YEARS OLD
55 – 64 YEARS OLD
65+ YEARS OLD
180
SOURCE: META'S ADVERTISING RESOURCES. ADVISORY: AUDIENCE FIGURES MAY NOT REPRESENT UNIQUE INDIVIDUALS, AND MAY NOT MATCH EQUIVALENT FIGURES FOR THE TOTAL ACTIVE USER BASE.
NOTES: FIGURES USE MIDPOINT OF PUBLISHED RANGES. META'S ADVERTISING RESOURCES ONLY PUBLISH GENDER DATA FOR "FEMALE" AND "MALE".
we are social
Hootsuite

Journal of Economic Perspectives—Volume 32, Number 3—Summer 2018—Pages 259–280

Social Connectedness: Measurement, Determinants, and Effects

Michael Bailey, Rachel Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong

Social networks can shape many aspects of social and economic activity: migration and trade, job-seeking, innovation, consumer preferences and sentiment, public health, social mobility, and more. In turn, social networks themselves are associated with geographic proximity, historical ties, political boundaries, and other factors. Traditionally, the unavailability of large-scale and representative data on social connectedness between individuals or geographic regions has posed a challenge for empirical research on social networks. More recently, a body of such research has begun to emerge using data on social connectedness from online social networking services such as Facebook, LinkedIn, and Twitter. To date, most

In a nutshell

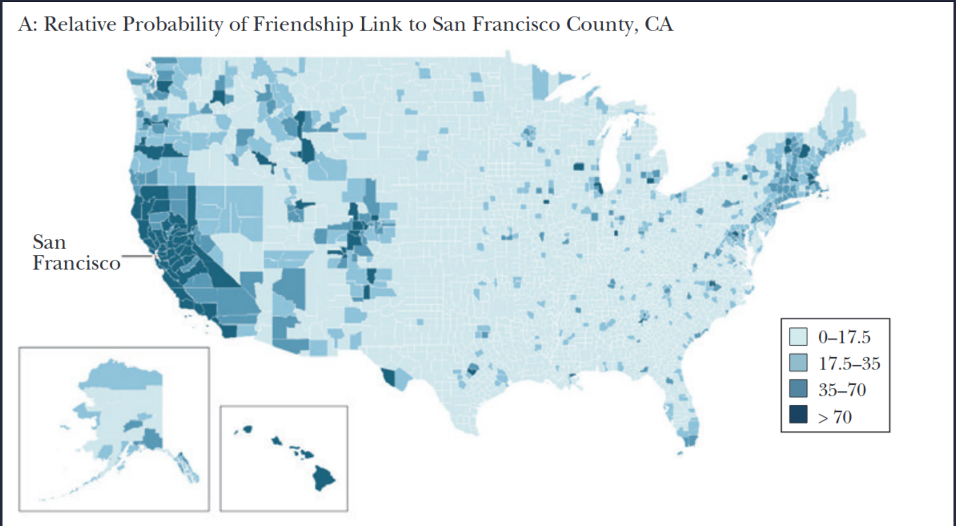
- Strength of connectedness between two geographic areas as represented by Facebook friendship ties
- Access data thanks to Micheal Bailey (Facebook)
- Validate their Social Connectedness Index (SCI):
  - SCI and geographic distance
  - concentration of social network and socio-economic charcteristics
  - social connectedness and bilateral economic ties (trade, innovation)
  - social connectedness and bilateral social activity (migration)
- SCI is openly available (upon request)

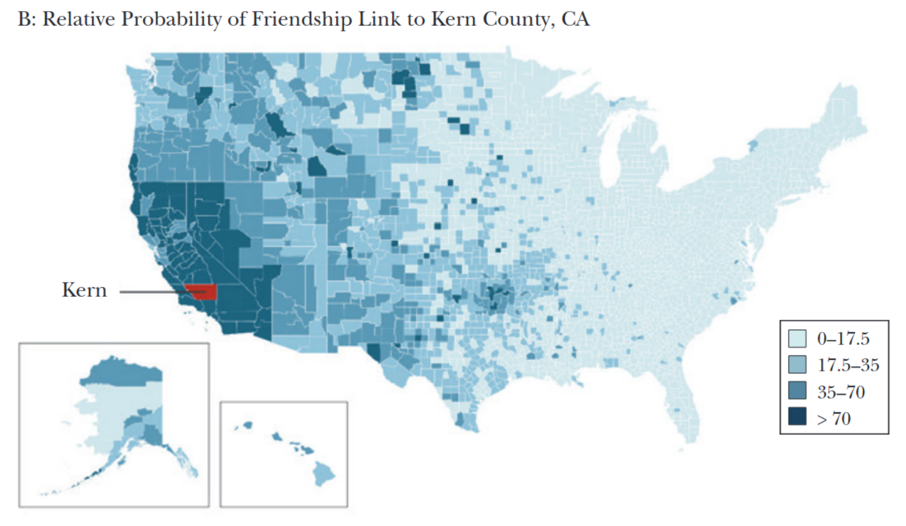
Social Conncectedness Index

1. Assign people to geographic areas
2. Calculate connectedness

$$SCI_{ij} = \frac{n_{ij}}{n_i \times n_j} \tag{1}$$

   where n_{ij} are the number of users in country i that are friends with j (friendship is symmetric in FB!), n_i FB users in i and n_j users in j
3. Drop small counts and add noise: remove all locations with a low number of observations and add random noise to the number of friendships between each set of locations to ensure no one can be re-identified.
4. Final sampling: The final SCI is the average scale of friendship ties across 10 random draws from 99% of active Facebook users to further protect privacy.


A: Relative Probability of Friendship Link to San Francisco County, CA
San Francisco
0–17.5
17.5–35
35–70
> 70


B: Relative Probability of Friendship Link to Kern County, CA
Kern
0–17.5
17.5–35
35–70
> 70

Determinants of Social Connectedness across County Pairs

	Dependent Variable: Log(SCI)				
	(1)	(2)	(3)	(4)	(5)
log(Distance in Miles)	–1.483***	–1.287***	–1.160***	–1.988***	–1.214***
	(0.065)	(0.061)	(0.059)	(0.043)	(0.055)
Same State		1.496***	1.271***	1.216***	1.496***
		(0.087)	(0.083)	(0.044)	(0.085)
Δ Income ($1,000)					–0.006***
					(0.001)
Δ Share Population White (%)					–0.012***
					(0.001)
Δ Share Population No High School (%)					–0.012***
					(0.002)
Δ 2008 Obama Vote Share (%)					–0.006***
					(0.001)
Δ Share Population Religious (%)					–0.002***
					(0.001)
County Fixed Effects	Y	Y	Y	Y	Y
Sample			>200 miles	<200 miles	
Number of observations	2,961,968	2,961,968	2,775,244	186,669	2,961,968
R^2	0.907	0.916	0.916	0.941	0.922

Note: Table shows results from a regression of the log of the Social Connectedness Index on a number of explanatory variables. The log of the geographic distance between the counties is the explanatory variable in column 1. In column 2, we include an additional control indicating whether both counties are within the same state. In columns 3 and 4, we restrict the sample to county-pairs that are more and less than 200 miles apart, respectively. The unit of observation is a county-pair. Standard errors are given in parentheses. The online Appendix (http://e-jep.org) provides more details on the data sources and exact specifications.
*, **, and *** indicate significance levels of $p < 0.1$, $p < 0.05$, and $p < 0.01$, respectively.

Network Concentration and County-Level Characteristics

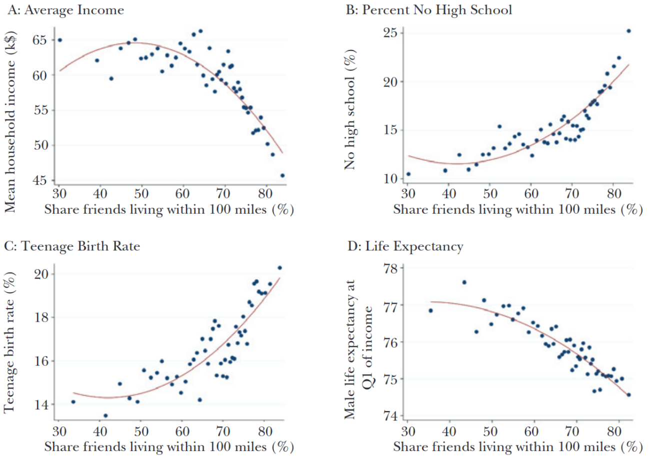


Table 3

Social Connectedness and Across-Region Economic Interactions

	(1)	(2)	(3)	(4)
Panel A: Dependent Variable: log(State-Level Trade Flows)				
log(Distance)	−1.057***		−0.531***	−0.533***
	(0.071)		(0.084)	(0.085)
log(SCI)		0.999***	0.643***	0.637***
		(0.051)	(0.071)	(0.060)
State Fixed Effects	Y	Y	Y	Y
Other State Differences	N	N	N	Y
Observations	2,219	2,220	2,219	2,219
R^2	0.912	0.918	0.926	0.930

Panel B: Dependent Variable: Indicator for Patent Citation

log(Distance)	–0.048***		–0.011**	–0.021**
	(0.002)		(0.005)	(0.009)
log(SCI)		0.063***	0.049***	0.066***
		(0.003)	(0.006)	(0.012)
Technological Category + County Fixed Effects	Y	Y	Y	Y
Cited + Issued Patent Fixed Effects, Other County Differences	N	N	N	Y
Observations	2,171,754	2,171,754	2,171,754	2,168,285
R^2	0.056	0.059	0.059	0.101

Panel C: Dependent Variable: log(County-Level Migration)

log(Distance)	−0.973***		0.023	0.031
	(0.048)		(0.021)	(0.021)
log(SCI)		1.134***	1.148***	1.159***
		(0.019)	(0.024)	(0.024)
County Fixed Effects	Y	Y	Y	Y
Other County Differences	N	N	N	Y
Observations	25,305	25,305	25,305	25,287
R^2	0.610	0.893	0.893	0.893

Food for thought

- What could one do with SCI data?
- You can access the data at the link https://data.humdata.org/dataset/social-connectedness-index

Food for thought

- What could one do with SCI data?
- You can access the data at the link https://data.humdata.org/dataset/social-connectedness-index

Contents lists available at ScienceDirect

Journal of International Economics

journal homepage: www.elsevier.com/locate/jie

International trade and social connectedness

Michael Bailey [a], Abhinav Gupta [b], Sebastian Hillenbrand [b], Theresa Kuchler [b], Robert Richmond [b,*], Johannes Stroebel [b]

[a] *Facebook, Inc, United States of America*
[b] *Stern School of Business, New York University, United States of America*

A R T I C L E I N F O

Article history:
Received 11 September 2020
Received in revised form 23 December 2020
Accepted 23 December 2020
Available online 29 December 2020

Dataset link: https://data.mendeley.com/datasets/7wddm84w9r/1

JEL codes:
F1
F5
F6

Keywords:

A B S T R A C T

We use de-identified data from Facebook to construct a new and publicly available measure of the pairwise social connectedness between 170 countries and 332 European regions. We find that two countries trade more when they are more socially connected, especially for goods where information frictions may be large. The social connections that predict trade in specific products are those between the regions where the product is produced in the exporting country and the regions where it is used in the importing country. Once we control for social connectedness, the estimated effects of geographic distance and country borders on trade decline substantially.

Table 2
Gravity Regressions - Goods Trade Heterogeneity in 2017.

	Dependent variable: Product-Specific Exports				
	(1)	(2)	(3)	(4)	(5)
log(SCI)	0.275***	0.299***	0.304***	0.281***	0.287***
	(0.027)	(0.028)	(0.024)	(0.031)	(0.025)
log(SCI) × Share Exchange-Traded		−0.179**	−0.148**		
		(0.080)	(0.070)		
log(SCI) × Rule of Law Destination				−0.014	−0.010
				(0.021)	(0.019)
log(SCI) × Rule of Law Origin				0.000	0.005
				(0.019)	(0.015)
Origin Country × Product FE	Y	Y	Y	Y	Y
Destination Country × Product FE	Y	Y	Y	Y	Y
Other Gravity Controls	Y	Y	Y	Y	Y
log(Distance) × Product FE	Y	Y		Y	
Distance Group × Product FE			Y		Y
R^2	0.932	0.933	0.946	0.932	0.946
N	2,597,760	2,597,760	2,597,760	2,597,760	2,597,760
N - Explained by FE	334,186	334,186	334,186	405,093	405,093

Note: Table shows results from regression 3. The dependent variable is exports of product category k from country i to country j in 2017. Product-level trade data are aggregated up to the first 2 digits of the HS96 product classification. Other gravity controls include a common border dummy, a common official language dummy, a dummy indicating whether the two countries had a common colonizer post-1945, and a dummy indicating whether the pair of countries was in a colonial relationship post-1945. We also separately control for the logarithm of distance interacted with product categories in columns 1, 2, 4 and for distance groups (dummies for percentiles of the distance distribution) interacted with product categories in columns 3 and 5. Share Exchange-Traded refers to the proportion of exchange-traded products—based on the conservative classification scheme in Rauch (1999)—within a product category. Rule of law is obtained from the World Governance Indicators published by the World Bank. All specifications include fixed effects for the importer and exporter country interacted with product categories. Standard errors are clustered by exporter and importer country. The data include 165 countries and 96 product categories, which amounts to 2,597,760 observations. Observations that are fully explained by the fixed effects are dropped before the PPML estimation. Significance levels: *(p<0.10), **(p<0.05), ***(p<0.01).

TWITTER DATA

- Twitter Streaming API: 1 % random sample of all tweets

  $\rightarrow$ filters: keyword, geolocation

  $\rightarrow$ between 40 and 60 per second

- 42 variables: text, username, user_lang, lang, followers, timezone, latitude, longitude, place, source,...

- Twitter Streaming API: 1 % random sample of all tweets

  $\rightarrow$ filters: keyword, geolocation

  $\rightarrow$ between 40 and 60 per second

- 42 variables: text, username, user_lang, lang, followers, timezone, latitude, longitude, place, source,...

- Twitter Streaming API: 1 % random sample of all tweets

  → filters: keyword, geolocation

  → between 40 and 60 per second

- 42 variables: text, username, user_lang, lang, followers, timezone, latitude, longitude, place, source,...

- Twitter Streaming API: 1 % random sample of all tweets

  $\rightarrow$ filters: keyword, geolocation

  $\rightarrow$ between 40 and 60 per second

- 42 variables: text, username, user_lang, lang, followers, timezone, latitude, longitude, place, source,...

Marvin || Runaways @ichmagdasnicht · 18 Apr 2017

Ich bin mit einer 2.0 in der Klausur einer der Schlechtesten. Ich studiere das Falsche, glaube ich langsam.

Translate from German

5 15

Timo Zander
@tinkengil

Follow

Replying to @ichmagdasnicht

offenbar nicht Mathematik 🙈

Translate from German

5:22 PM - 18 Apr 2017 from Kiel, Germany

1

Tweet your reply

Marvin || Runaways @ichmagdasnicht · 18 Apr 2017

Replying to @tinkengil

Wieso :D

1

Timo Zander @tinkengil · 18 Apr 2017

ich erinnere mich an eine Horrorklausur mit 80% Durchfallquote. Bestnote war 1,7. Völlig absurd was da gefordert wurde

Translate from German

1 1

Marvin || Runaways @ichmagdasnicht · 18 Apr 2017

Gab's bei uns auch anfangs. Richtiger Horror.

Translate from German

1

Timo Zander
@tinkengil

Follow

Replying to @ichmagdasnicht

offenbar nicht Mathematik 🙈

Translate from German

5:22 PM - 18 Apr 2017 from Kiel, Germany

1

example_tweet.json


```json
{
  "created_at": "Tue Apr 18 15:22:19 +0000 2017",
  "id": 854354410041991168,
  "id_str": "854354410041991168",
  "text": "@ichmagdasnicht offenbar nicht Mathematik 🙈",
  "display_text_range": [
    16,
    43
  ],
  "source": "<a href=\"http://tapbots.com/tweetbot\" rel=\"nofollow\">Tweetbot for iOS</a>",
  "truncated": false,
  "in_reply_to_status_id": 854247992186073088,
  "in_reply_to_status_id_str": "854247992186073088",
  "in_reply_to_user_id": 2535411248,
  "in_reply_to_user_id_str": "2535411248",
  "in_reply_to_screen_name": "ichmagdasnicht",
  "user": {
    "id": 19030252,
    "id_str": "19030252",
    "name": "Timo Zander",
    "screen_name": "tinkengil",
    "location": "Kiel",
    "url": "http://about.me/timozander",
    "description": "PhD-Student | Podcastet bei playtogether-podcast.de | bloggt gelegentlich bei insulinaspekte.de und http://tinkengil.com | http://instagram.com/tinkengil",
    "protected": false,
    "verified": false,
    "followers_count": 286,
    "friends_count": 344,
    "listed_count": 18,
    "favourites_count": 1830,
    "statuses_count": 12108,
    "created_at": "Thu Jan 15 17:40:27 +0000 2009",
    "utc_offset": 7200,
    "time_zone": "Bern",
    "geo_enabled": true,
    "lang": "en",
    "contributors_enabled": false,
    "is_translator": false,
    "profile_background_color": "EBEBEB",
    "profile_background_image_url": "http://pbs.twimg.com/profile_background_images/590786545/5vyvydxrk528xhz91w86.jpeg",
    "profile_background_image_url_https": "https://pbs.twimg.com/profile_background_images/590786545/5vyvydxrk528xhz91w86.jpeg",
    "profile_background_tile": true,
    "profile_link_color": "990000",
    "profile_sidebar_border_color": "FFFFFF",
    "profile_sidebar_fill_color": "F3F3F3",
    "profile_text_color": "333333",
    "profile_use_background_image": false,
    "profile_image_url": "http://pbs.twimg.com/profile_images/549318880876048384/zag6999H_normal.jpeg",
    "profile_image_url_https": "https://pbs.twimg.com/profile_images/549318880876048384/zag6999H_normal.jpeg",
    "default_profile": false,
```

```
"profile_background_image_url": "http://pbs.twimg.com/profile_background_images/590786545/5vyvydxrk528xhz91w86.jpeg",
"profile_background_image_url_https": "https://pbs.twimg.com/profile_background_images/590786545/5vyvydxrk528xhz91w86.jpeg",
"profile_background_tile": true,
"profile_link_color": "990000",
"profile_sidebar_border_color": "FFFFFF",
"profile_sidebar_fill_color": "F3F3F3",
"profile_text_color": "333333",
"profile_use_background_image": false,
"profile_image_url": "http://pbs.twimg.com/profile_images/549318880876048384/zag6999H_normal.jpeg",
"profile_image_url_https": "https://pbs.twimg.com/profile_images/549318880876048384/zag6999H_normal.jpeg",
"default_profile": false,
"default_profile_image": false,
"following": null,
"follow_request_sent": null,
"notifications": null
},
"geo": {
"type": "Point",
"coordinates": [
54.32436928,
10.12301066
]
},
"coordinates": {
"type": "Point",
"coordinates": [
10.12301066,
54.32436928
]
},
"place": {
"id": "1b9b5e83e647a7ed",
"url": "https://api.twitter.com/1.1/geo/id/1b9b5e83e647a7ed.json",
"place_type": "city",
"name": "Kiel",
"full_name": "Kiel, Germany",
"country_code": "DE",
"country": "Germany",
"bounding_box": {
"type": "Polygon",
"coordinates": [
[
[
10.032937,
54.250693
],
[
10.032937,
54.432916
],
[
```

```
      "coordinates": [
        [
          [
            10.032937,
            54.250693
          ],
          [
            10.032937,
            54.432916
          ],
          [
            10.218568,
            54.432916
          ],
          [
            10.218568,
            54.250693
          ]
        ]
      ]
    },
    "attributes": {}
  },
  "contributors": null,
  "is_quote_status": false,
  "retweet_count": 0,
  "favorite_count": 0,
  "entities": {
    "hashtags": [],
    "urls": [],
    "user_mentions": [
      {
        "screen_name": "ichmagdasnicht",
        "name": "Marvin || Runaways",
        "id": 2535411248,
        "id_str": "2535411248",
        "indices": [
          0,
          15
        ]
      }
    ],
    "symbols": []
  },
  "favorited": false,
  "retweeted": false,
  "filter_level": "low",
  "lang": "de",
  "timestamp_ms": "1492528939148"
}
```

Twitter data in research

- Obvious: Text-mining

  → Brexit, Trump election,.. Gorodnichenko et al. (2018), De Lyon et al. (2018), Halberstam and Knight (2016)

- Not so obvious: Metadata

  → Language distribution

  → Migration

Twitter data in research

- Obvious: Text-mining

  → Brexit, Trump election,.. Gorodnichenko et al. (2018), De Lyon et al. (2018), Halberstam and Knight (2016)

- Not so obvious: Metadata

  → Language distribution

  → Migration

Twitter data in research

- Obvious: Text-mining

  $\rightarrow$ Brexit, Trump election,.. Gorodnichenko et al. (2018), De Lyon et al. (2018), Halberstam and Knight (2016)

- Not so obvious: Metadata

  $\rightarrow$ Language distribution

  $\rightarrow$ Migration

Twitter data in research

- Obvious: Text-mining

  $\rightarrow$ Brexit, Trump election,.. Gorodnichenko et al. (2018), De Lyon et al. (2018), Halberstam and Knight (2016)

- Not so obvious: Metadata

  $\rightarrow$ Language distribution

  $\rightarrow$ Migration

Hinz and Leromain (2018): Languages and trade

- Spatial distribution of languages in Europe
- Geolocation from “coordinates”, and “user_lang” or “lang”

  → large heterogeneity across and within countries
- Coordinates provided either by the user’s device’s GPS coordinates, or a self-assigned location

  → Barratt, J. Cheshire, and E. Manley (2013) use similar data for NY boroughs

Hinz and Leromain (2018): Languages and trade

- Spatial distribution of languages in Europe
- Geolocation from “coordinates”, and “user_lang” or “lang”

  → large heterogeneity across and within countries

- Coordinates provided either by the user’s device’s GPS coordinates, or a self-assigned location

  → Barratt, J. Cheshire, and E. Manley (2013) use similar data for NY boroughs

Hinz and Leromain (2018): Languages and trade

- Spatial distribution of languages in Europe
- Geolocation from "coordinates", and "user_lang" or "lang"

$\rightarrow$ large heterogeneity across and within countries

- Coordinates provided either by the user's device's GPS coordinates, or a self-assigned location

$\rightarrow$ Barratt, J. Cheshire, and E. Manley (2013) use similar data for NY boroughs

Hinz and Leromain (2018): Languages and trade

- Spatial distribution of languages in Europe
- Geolocation from “coordinates”, and “user_lang” or “lang”

  → large heterogeneity across and within countries
- Coordinates provided either by the user’s device’s GPS coordinates, or a self-assigned location

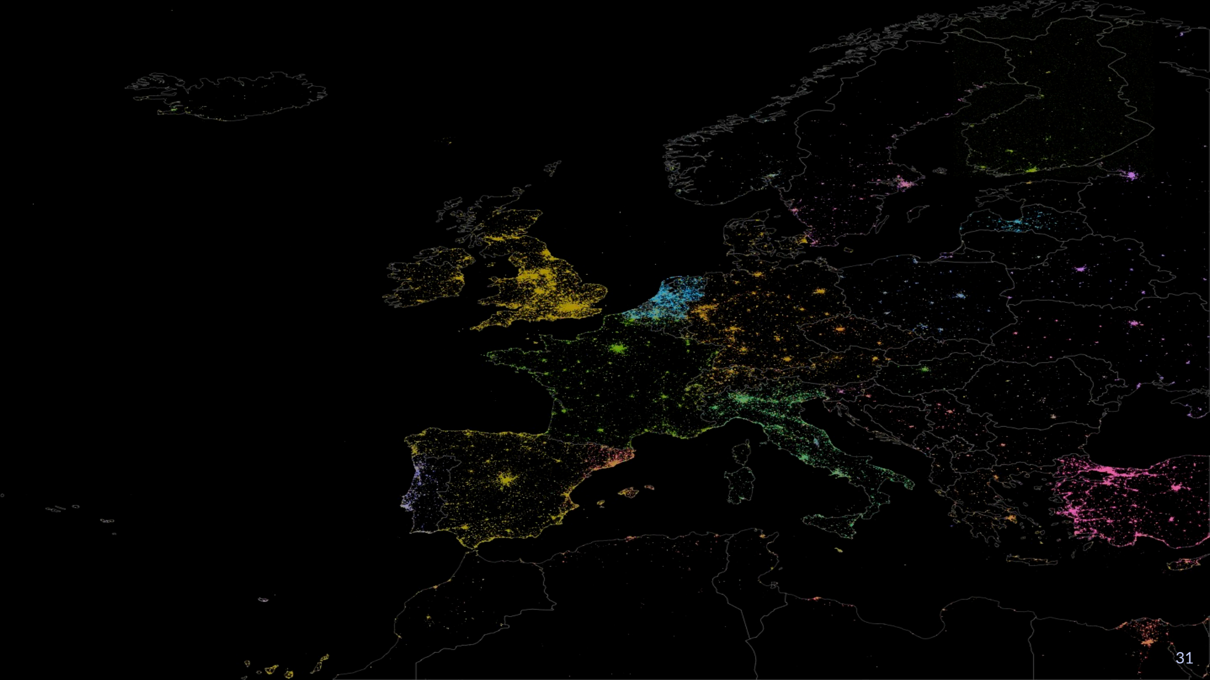
  → Barratt, J. Cheshire, and E. Manley (2013) use similar data for NY boroughs

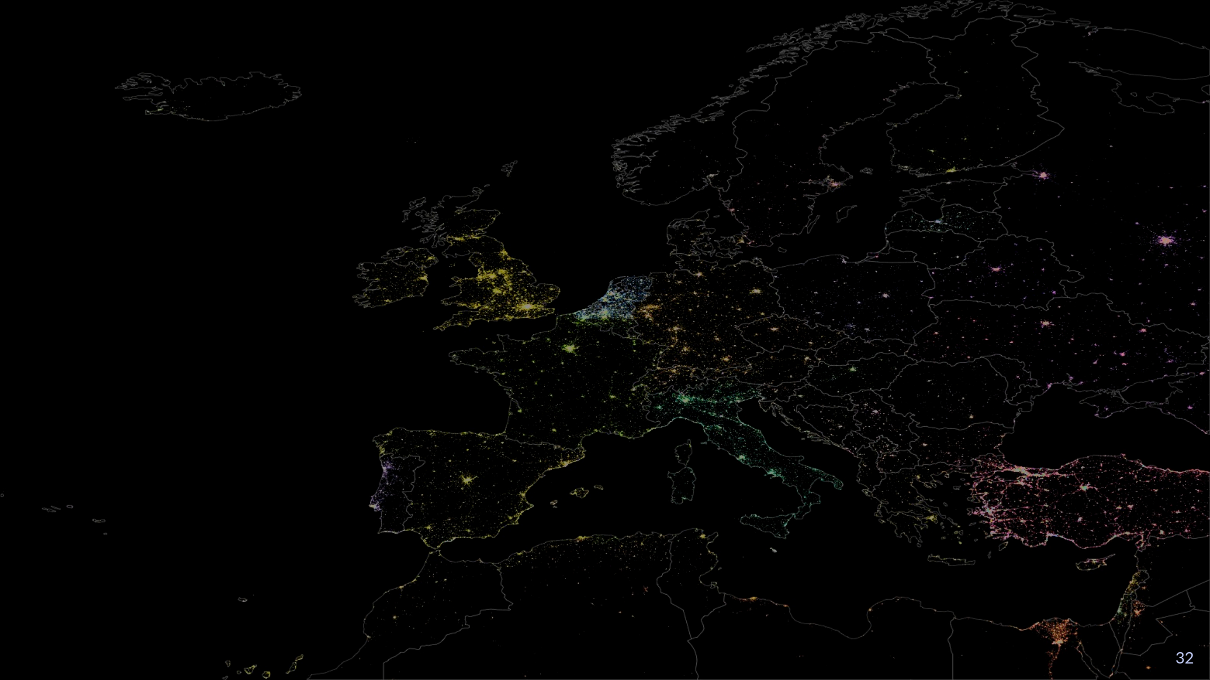
Hinz and Leromain (2018): Languages and trade

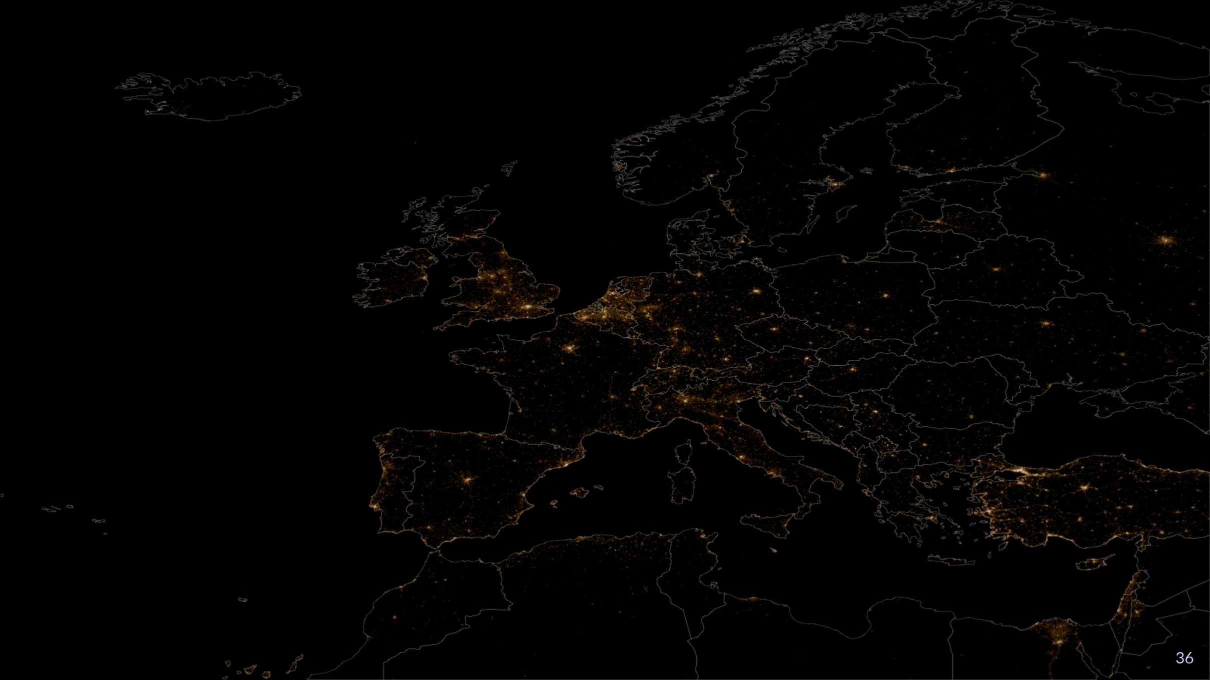
- Spatial distribution of languages in Europe
- Geolocation from “coordinates”, and “user_lang” or “lang”

  $\rightarrow$ large heterogeneity across and within countries
- Coordinates provided either by the user’s device’s GPS coordinates, or a self-assigned location

  $\rightarrow$ Barratt, J. Cheshire, and E. Manley (2013) use similar data for NY boroughs







Bots and human users

- Bots: an issue, Chu et al. (2012) suggest only taking those sent from smart phones and official app
- 6.6 million unique human Twitter users
- 481,720 unique human Twitter users in Europe
- 73 different languages
- 25 % tweet in more than 1 language, in Germany 31 %
- 958,071 unique language-user observations

Bots and human users

- Bots: an issue, Chu et al. (2012) suggest only taking those sent from smart phones and official app
- 6.6 million unique human Twitter users
- 481,720 unique human Twitter users in Europe
- 73 different languages
- 25 % tweet in more than 1 language, in Germany 31 %
- 958,071 unique language-user observations

Bots and human users

- Bots: an issue, Chu et al. (2012) suggest only taking those sent from smart phones and official app
- 6.6 million unique human Twitter users
- 481,720 unique human Twitter users in Europe
- 73 different languages
- 25 % tweet in more than 1 language, in Germany 31 %
- 958,071 unique language-user observations

Bots and human users

- Bots: an issue, Chu et al. (2012) suggest only taking those sent from smart phones and official app
- 6.6 million unique human Twitter users
- 481,720 unique human Twitter users in Europe
- 73 different languages
- 25 % tweet in more than 1 language, in Germany 31 %
- 958,071 unique language-user observations

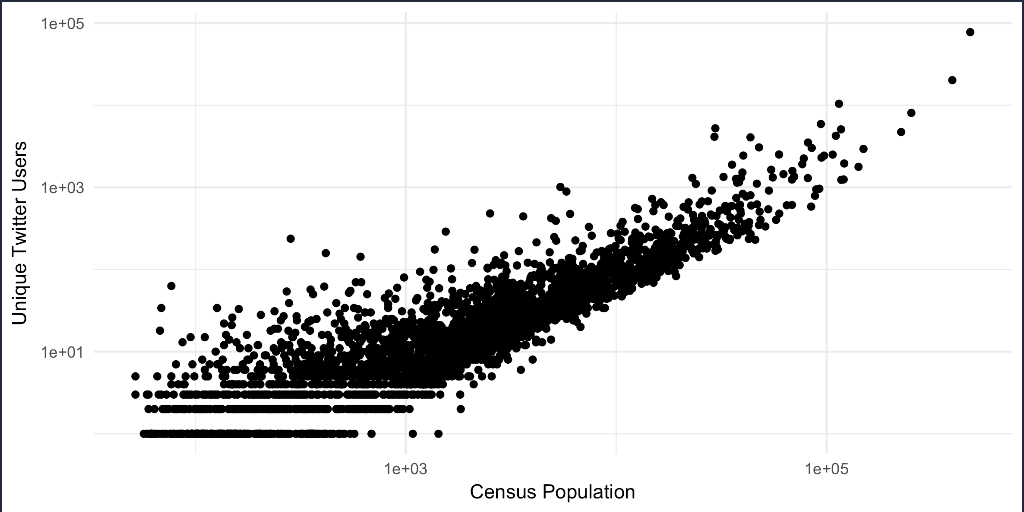
Bots and human users

- Bots: an issue, Chu et al. (2012) suggest only taking those sent from smart phones and official app
- 6.6 million unique human Twitter users
- 481,720 unique human Twitter users in Europe
- 73 different languages
- 25 % tweet in more than 1 language, in Germany 31 %
- 958,071 unique language-user observations

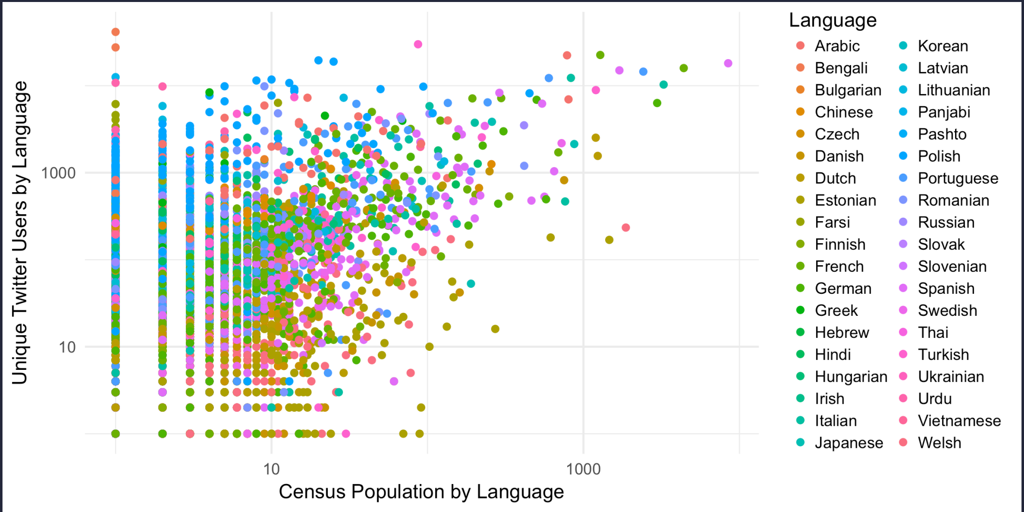
Bots and human users

- Bots: an issue, Chu et al. (2012) suggest only taking those sent from smart phones and official app
- 6.6 million unique human Twitter users
- 481,720 unique human Twitter users in Europe
- 73 different languages
- 25 % tweet in more than 1 language, in Germany 31 %
- 958,071 unique language-user observations

Twitter and UK Census Population

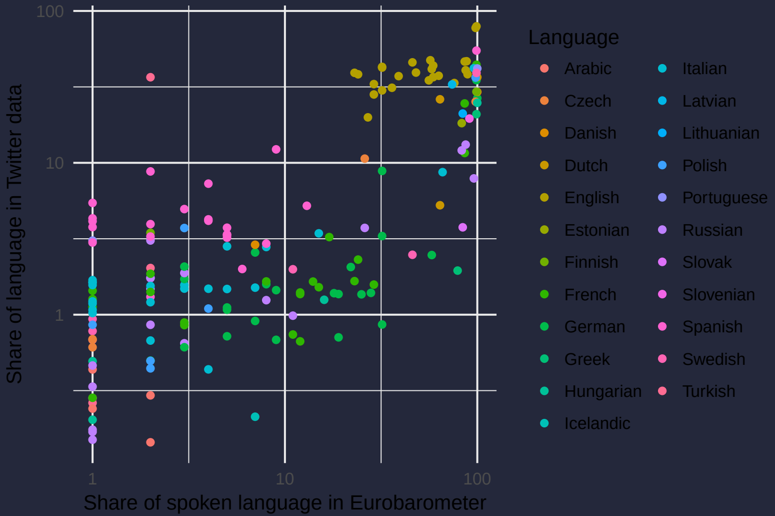


Twitter and UK Census Main Language

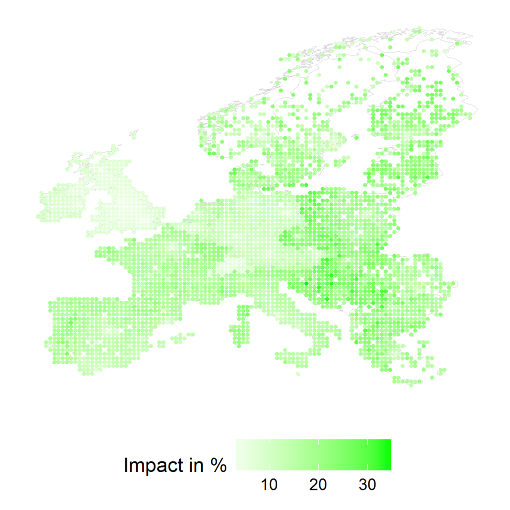


Language use on Twitter and UK census, correlation $= 0.49$.

Twitter and Eurobarometer



Language use on Twitter and Eurobarometer, correlation $= 0.74$.


Impact in %
10
20
30

Hausmann, Hinz and Yildirim (2018): Venezuelan emigration

- Economic crisis in Venezuela: Large (?) number of refugees
  $\rightarrow$ lack of official numbers

- Dataset of geolocalized Tweets of people that tweeted from Venezuela between February 2017 and May 2018
  $\rightarrow$ 5.4 million tweets
  $\rightarrow$ 490.000 tweets from 30.000 *human* Twitter users

- Idea: What location(s) do they tweet from over time?

Hausmann, Hinz and Yildirim (2018): Venezuelan emigration

- Economic crisis in Venezuela: Large (?) number of refugees
  $\rightarrow$ lack of official numbers
- Dataset of geolocalized Tweets of people that tweeted from Venezuela between February 2017 and May 2018
  $\rightarrow$ 5.4 million tweets
  $\rightarrow$ 490.000 tweets from 30.000 *human* Twitter users
- Idea: What location(s) do they tweet from over time?

Hausmann, Hinz and Yildirim (2018): Venezuelan emigration

- Economic crisis in Venezuela: Large (?) number of refugees
  $\rightarrow$ lack of official numbers
- Dataset of geolocalized Tweets of people that tweeted from Venezuela between February 2017 and May 2018
  $\rightarrow$ 5.4 million tweets
  $\rightarrow$ 490.000 tweets from 30.000 *human* Twitter users
- Idea: What location(s) do they tweet from over time?

Hausmann, Hinz and Yildirim (2018): Venezuelan emigration

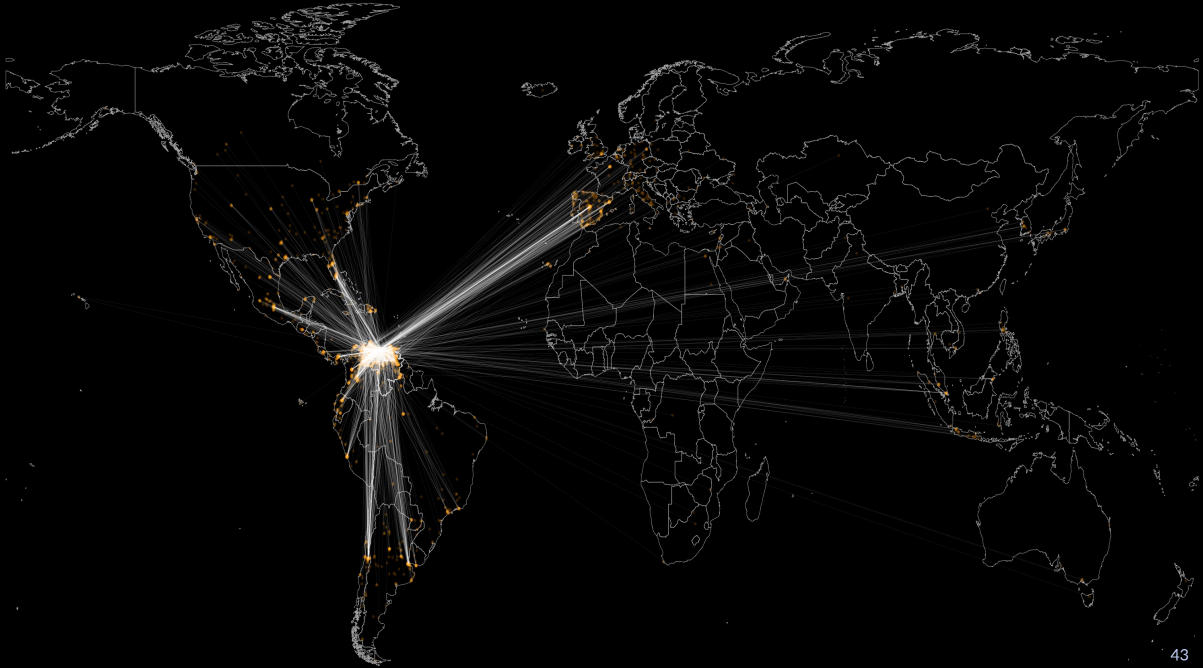
- Economic crisis in Venezuela: Large (?) number of refugees
  $\rightarrow$ lack of official numbers
- Dataset of geolocalized Tweets of people that tweeted from Venezuela between February 2017 and May 2018
  $\rightarrow$ 5.4 million tweets
  $\rightarrow$ 490.000 tweets from 30.000 *human* Twitter users
- Idea: What location(s) do they tweet from over time?

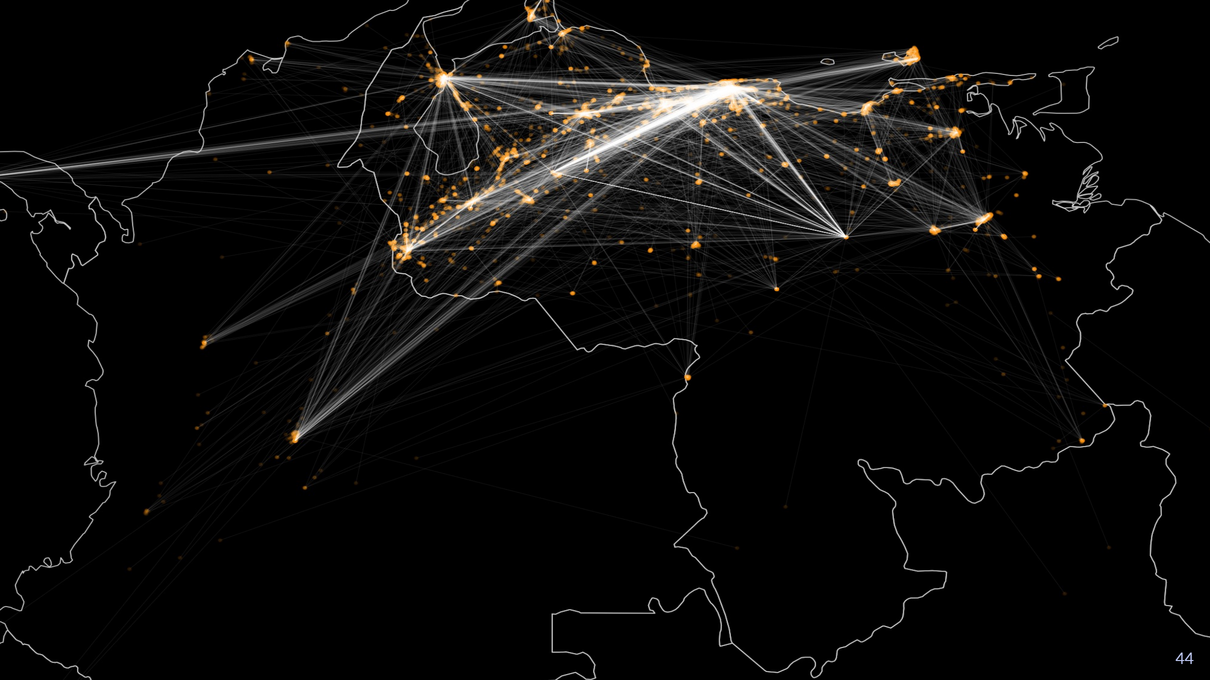
Hausmann, Hinz and Yildirim (2018): Venezuelan emigration

- Economic crisis in Venezuela: Large (?) number of refugees
  $\rightarrow$ lack of official numbers
- Dataset of geolocalized Tweets of people that tweeted from Venezuela between February 2017 and May 2018
  $\rightarrow$ 5.4 million tweets
  $\rightarrow$ 490.000 tweets from 30.000 *human* Twitter users
- Idea: What location(s) do they tweet from over time?

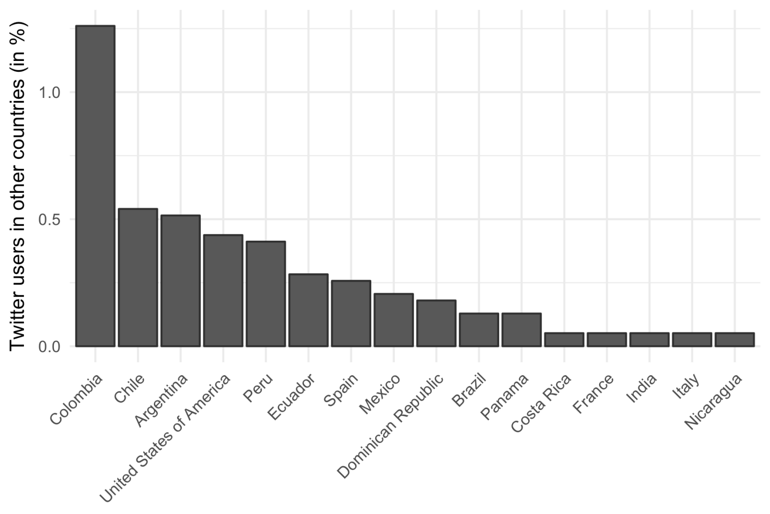
Hausmann, Hinz and Yildirim (2018): Venezuelan emigration

- Economic crisis in Venezuela: Large (?) number of refugees
  $\rightarrow$ lack of official numbers
- Dataset of geolocalized Tweets of people that tweeted from Venezuela between February 2017 and May 2018
  $\rightarrow$ 5.4 million tweets
  $\rightarrow$ 490.000 tweets from 30.000 *human* Twitter users
- Idea: What location(s) do they tweet from over time?





Distribution of countries



Distribution of countries of last recorded locations of users outside Venezuela

Migration and social media

- Hawelka (2014): global mobility patterns, tourism flows
- Jurdak (2015) city-to-city travel in Australia
- Morstatter (2013): random sample creates an accurate picture of the entire population of geolocated Tweets
- Question: How representative are geolocalized tweets?

Migration and social media

- Hawelka (2014): global mobility patterns, tourism flows
- Jurdak (2015) city-to-city travel in Australia
- Morstatter (2013): random sample creates an accurate picture of the entire population of geolocated Tweets
- Question: How representative are geolocalized tweets?

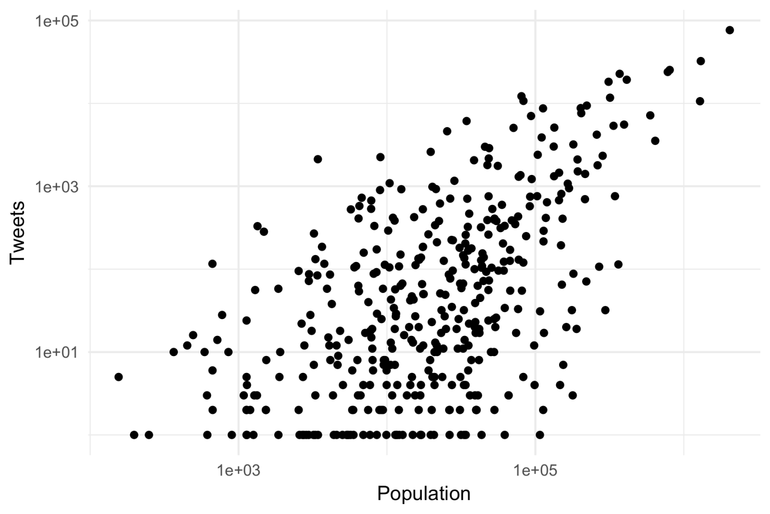
Migration and social media

- Hawelka (2014): global mobility patterns, tourism flows
- Jurdak (2015) city-to-city travel in Australia
- Morstatter (2013): random sample creates an accurate picture of the entire population of geolocated Tweets
- Question: How representative are geolocalized tweets?

Migration and social media

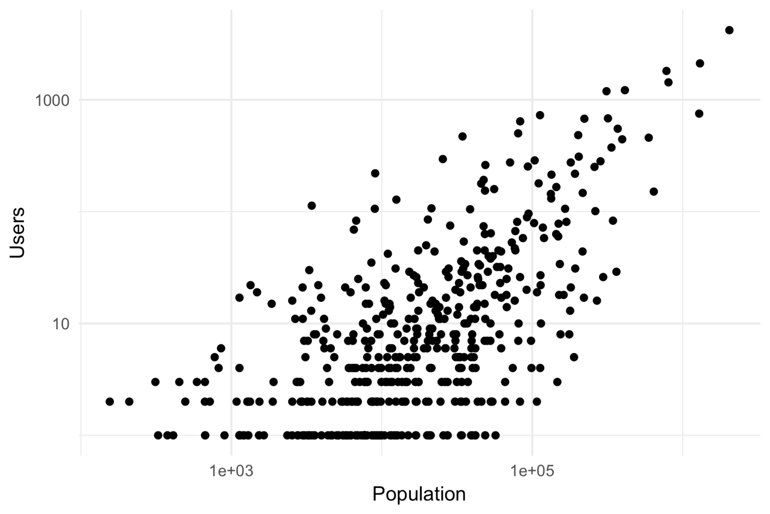
- Hawelka (2014): global mobility patterns, tourism flows
- Jurdak (2015) city-to-city travel in Australia
- Morstatter (2013): random sample creates an accurate picture of the entire population of geolocated Tweets
- Question: How representative are geolocalized tweets?

Population and Tweets



"Gridded Population of the World" and number of Tweets by location

Population and Users



“Gridded Population of the World” and number of Twitter users by location

Representativeness of Twitter users in Venezuela

- “Digital in 2017 Global Overview report”: 44% of Venezuelans social media, 35% from mobile device
- “ Tendencias Digitales”: 56% of internet users in Venezuela use Twitter or comparable social media services
- Twitter: penetration in Venezuela 26 %

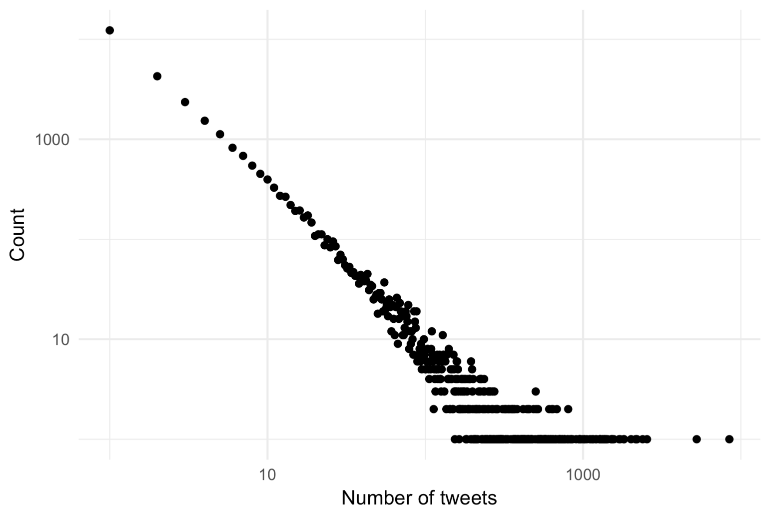
Representativeness of Twitter users in Venezuela

- “Digital in 2017 Global Overview report”: 44% of Venezuelans social media, 35% from mobile device
- “ Tendencias Digitales”: 56% of internet users in Venezuela use Twitter or comparable social media services
- Twitter: penetration in Venezuela 26 %

Representativeness of Twitter users in Venezuela

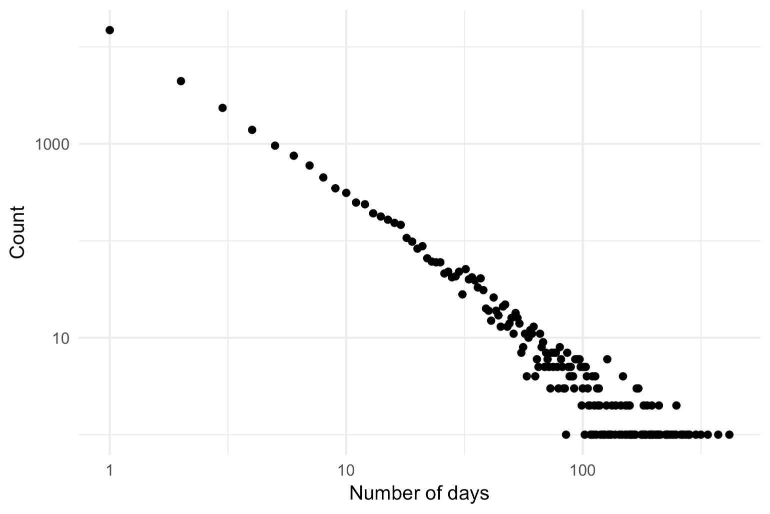
- “Digital in 2017 Global Overview report”: 44% of Venezuelans social media, 35% from mobile device
- “ Tendencias Digitales”: 56% of internet users in Venezuela use Twitter or comparable social media services
- Twitter: penetration in Venezuela 26 %

Tweets per users



Number of tweets per user in the dataset

Days per users



Number of days a user is observed in the dataset

How to make use not to capture tourists?

- narrow sample to users who
  → tweeted from Venezuela exclusively between Feb and May ’17 (Period 1)
  → tweeted from *a country* exclusively between Feb and May ’18 (Period 2)

- Everyone who is *not* in Venezuela in period 2: migrant

- reduces sample to 818 (!)
  → Problem: Large heterogeneity in tweet frequency

How to make use not to capture tourists?

- narrow sample to users who
  $\rightarrow$ tweeted from Venezuela exclusively between Feb and May '17 (Period 1)
  $\rightarrow$ tweeted from *a country* exclusively between Feb and May '18 (Period 2)
- Everyone who is *not* in Venezuela in period 2: migrant
- reduces sample to 818 (!)
  $\rightarrow$ Problem: Large heterogeneity in tweet frequency

How to make use not to capture tourists?

- narrow sample to users who
  $\rightarrow$ tweeted from Venezuela exclusively between Feb and May ’17 (Period 1)
  $\rightarrow$ tweeted from *a country* exclusively between Feb and May ’18 (Period 2)
- Everyone who is *not* in Venezuela in period 2: migrant
- reduces sample to 818 (!)
  $\rightarrow$ Problem: Large heterogeneity in tweet frequency

How to make use not to capture tourists?

- narrow sample to users who
  $\rightarrow$ tweeted from Venezuela exclusively between Feb and May ’17 (Period 1)
  $\rightarrow$ tweeted from *a country* exclusively between Feb and May ’18 (Period 2)

- Everyone who is *not* in Venezuela in period 2: migrant

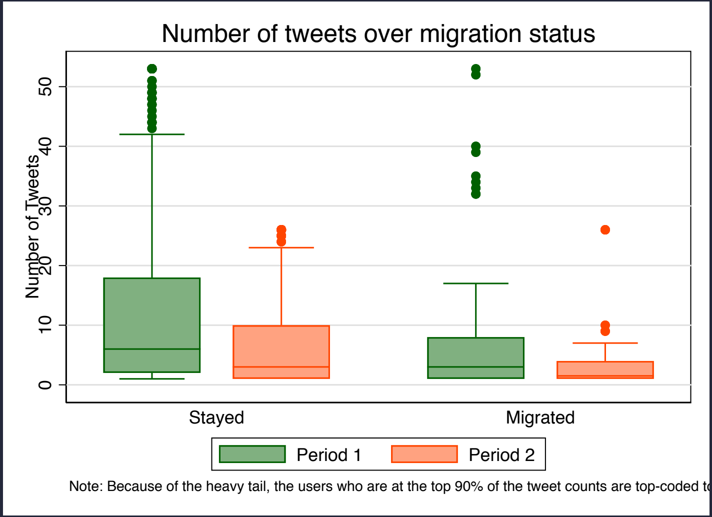
- reduces sample to 818 (!)
  $\rightarrow$ Problem: Large heterogeneity in tweet frequency

How to make use not to capture tourists?

- narrow sample to users who
  → tweeted from Venezuela exclusively between Feb and May '17 (Period 1)
  → tweeted from *a country* exclusively between Feb and May '18 (Period 2)
- Everyone who is *not* in Venezuela in period 2: migrant
- reduces sample to 818 (!)
  → Problem: Large heterogeneity in tweet frequency

How to make use not to capture tourists?

- narrow sample to users who
  $\rightarrow$ tweeted from Venezuela exclusively between Feb and May '17 (Period 1)
  $\rightarrow$ tweeted from *a country* exclusively between Feb and May '18 (Period 2)
- Everyone who is *not* in Venezuela in period 2: migrant
- reduces sample to 818 (!)
  $\rightarrow$ Problem: Large heterogeneity in tweet frequency



Tweets by migrants and non-migrants in two periods

Accounting for heterogeneity of Tweet frequency

- Need weight to correct for sampling bias
- Suppose probability of individual i tweeting exactly x tweets in three-month period given by

$$p_{i,x} = Pr\{tw_i = x\}$$

- tw_i random variable denoting tweets i
  $\rightarrow$ assume this probability distribution constant across periods

Accounting for heterogeneity of Tweet frequency

- Need weight to correct for sampling bias
- Suppose probability of individual i tweeting exactly x tweets in three-month period given by

$$p_{i,x} = Pr\{tw_i = x\}$$

- tw_i random variable denoting tweets i
  $\rightarrow$ assume this probability distribution constant across periods

Accounting for heterogeneity of Tweet frequency

- Need weight to correct for sampling bias
- Suppose probability of individual i tweeting exactly x tweets in three-month period given by

$$p_{i,x} = Pr\{tw_i = x\}$$

- tw_i random variable denoting tweets i
  $\rightarrow$ assume this probability distribution constant across periods

Accounting for heterogeneity of Tweet frequency

- Twitter provides $s = 0.01$ of all tweets, independent of user
  $\rightarrow q = (1 - s) = 99\%$ of Tweets not reported
- Denote U^1 (U^2) set all users observed at least once in period 1 (2)

Accounting for heterogeneity of Tweet frequency

- Twitter provides $s = 0.01$ of all tweets, independent of user
  $\rightarrow q = (1 - s) = 99\%$ of Tweets not reported
- Denote U^1 (U^2) set all users observed at least once in period 1 (2)

Accounting for heterogeneity of Tweet frequency

- Twitter provides $s = 0.01$ of all tweets, independent of user
  $\rightarrow q = (1 - s) = 99\%$ of Tweets not reported
- Denote U^1 (U^2) set all users observed at least once in period 1 (2)

Accounting for heterogeneity of Tweet frequency

- Probability of observing an individual who tweeted x_i times in period 1

$$Pr\{i \in U^1 | tw_i^1 = x\} = 1 - q^x.$$

- Probability of observing same individual who tweeted y_i times in period 2

$$Pr\{i \in U^2 | tw_i^2 = y\} = 1 - q^y.$$

Accounting for heterogeneity of Tweet frequency

- Probability of observing an individual who tweeted x_i times in period 1

$$Pr\{i \in U^1 | tw_i^1 = x\} = 1 - q^x.$$

- Probability of observing same individual who tweeted y_i times in period 2

$$Pr\{i \in U^2 | tw_i^2 = y\} = 1 - q^y.$$

Accounting for heterogeneity of Tweet frequency

- Assuming independence between the two sample, probability to be observed in both periods

$$
\begin{aligned}
Pr\{i \in U^1 \text{ and } i \in U^2\} &= \sum_{x=0}^{\infty}\sum_{y=0}^{\infty} Pr\{i \in U^1 | tw_i^1 = x\} Pr\{tw_i^1 = x\} \times \\
&\qquad Pr\{i \in U^2 | tw_i^2 = y\} Pr\{tw_i^2 = y\} \\
&= \sum_{x=0}^{\infty} p_{i,x}(1-q^x) \sum_{y=0}^{\infty} p_{i,y}(1-q^y) \\
&= (1 - E_i[q^x])^2 = (1 - G_i(q))^2
\end{aligned}
$$

- $G_i(q)$ probability generating function

Accounting for heterogeneity of Tweet frequency

- Model the individuals' tweeting behavior as a Poisson process
- Assume each individual has Poisson tweet rate in a three month period λ_i
- With Poisson distribution, rewrite the probability generating function as

$$G_i(q) = e^{-\lambda_i(1-q)} = e^{-\lambda_i s}.$$

Accounting for heterogeneity of Tweet frequency

- Model the individuals' tweeting behavior as a Poisson process
- Assume each individual has Poisson tweet rate in a three month period λ_i
- With Poisson distribution, rewrite the probability generating function as

$$G_i(q) = e^{-\lambda_i(1-q)} = e^{-\lambda_i s}.$$

Accounting for heterogeneity of Tweet frequency

- Model the individuals' tweeting behavior as a Poisson process
- Assume each individual has Poisson tweet rate in a three month period λ_i
- With Poisson distribution, rewrite the probability generating function as

$$G_i(q) = e^{-\lambda_i(1-q)} = e^{-\lambda_i s}.$$

Accounting for heterogeneity of Tweet frequency

- Hence probability of being observed in both periods

$$Pr\{i \in U^0 \text{ and } i \in U^1\} = (1 - e^{-\lambda_i s})^2 \tag{2}$$

with $s = 0.01$ in our case.

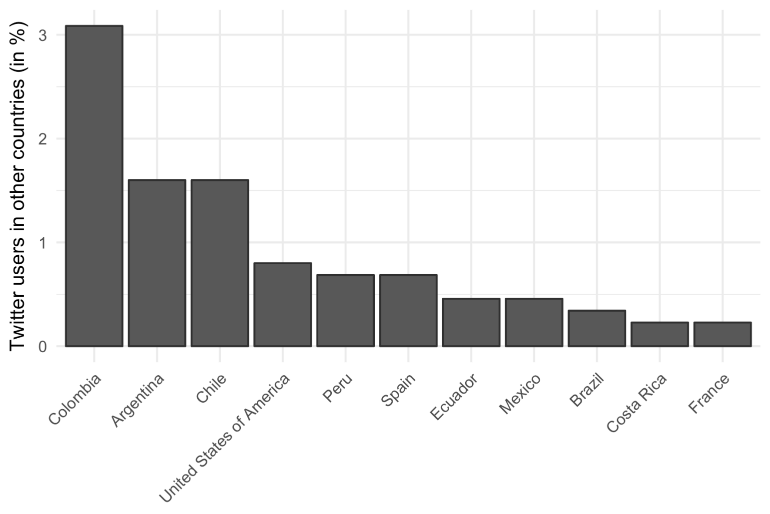
Net outflow over time

		(1) Venezuela	(2) Colombia	(3) Argentina	(4) Brazil	(5) Germany	(6) Venezuela	(7) Colombia
Emigration	*unweighted*	6,76%	7,78%	7,62%	3,88%	11,59%	6,99%	6,06%
	weighted	9,59%	7,84%	7,92%	3,97%	13,18%	7,98%	6,10%
Immigration	*unweighted*	2,01%	5,21%	10,48%	3,59%	11,27%	1,77%	5,21%
	weighted	2,22%	5,48%	10,70%	3,67%	12,41%	1,70%	5,37%
Difference	*unweighted*	-4,75%	-2,57%	2,86%	-0,29%	-0,32%	-5,22%	-0,85%
	weighted	-7,37%	-2,36%	2,78%	-0,30%	-0,77%	-6,28%	-0,73%
Annualized weighted perc.		-9,7%	-3,1%	3,7%	-0,4%	-1%	-12,1%	-1,4%
Period 1		02–04/17	02–04/17	02–04/17	02–04/17	02–04/17	12/16–04/17	12/16–04/17
Period 2		02–04/18	02–04/18	02–04/18	02–04/18	02–04/18	12/17–04/18	12/17–04/18

Source: Authors' calculations.

Computed emigration and immigration numbers

Distribution of countries



Distribution of countries of users between February and April ’18

Conclusion

- Social media data allows researchers to observe people, revealed preferences
- Design of exercise important: Endogeneity, sampling, ...

Social Media Data

DSIER [/dɪˈzaɪər/]

Julian HInz and Irene Iodice

Bielefeld University