MAX-PLANCK-INSTITUT
FÜR DEMOGRAFISCHE
FORSCHUNG

MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC
RESEARCH

MAX-PLANCK-INSTITUT
FÜR DEMOGRAFISCHE
FORSCHUNG

MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC
RESEARCH

Combining classic and emerging data sources for migration

Emilio Zagheni

UNECE-Eurostat Work Session on Migration Statistics
Geneva, Oct. 25, 2018

Acknowledgement: Main collaborators

Guy Abel

Monica Alexander

Francesco Billari

Jixuan Cai

Antoine Dubois

Lee Fiorio

René Flores

Kiran Garimella

Krishna Gummadi

Kivan Polimis

Tim Riffe

Ian Stewart

Ingmar Weber

How many people moved from Italy to Spain in 2017?

- We don't know with certainty, but...

⇒ Flows from Italy to Spain in 2015 (according to Spain): 17,350

⇒ Flows from Italy to Spain in 2015 (according to Italy): 5,003

- Flows are difficult to quantify, data are not timely and multiple sources are needed.

How many people moved from Italy to Spain in 2017?

- We don’t know with certainty, but...

⇒ Flows from Italy to Spain in 2015 (according to Spain): 17,350

⇒ Flows from Italy to Spain in 2015 (according to Italy): 5,003

- Flows are difficult to quantify, data are not timely and multiple sources are needed.

How many people moved from Italy to Spain in 2017?

- We don't know with certainty, but...

$\Rightarrow$ Flows from Italy to Spain in 2015 (according to Spain): 17,350

$\Rightarrow$ Flows from Italy to Spain in 2015 (according to Italy): 5,003

- Flows are difficult to quantify, data are not timely and multiple sources are needed.

How many people moved from Italy to Spain in 2017?

- We don't know with certainty, but...

$\Rightarrow$ Flows from Italy to Spain in 2015 (according to Spain): 17,350

$\Rightarrow$ Flows from Italy to Spain in 2015 (according to Italy): 5,003

- Flows are difficult to quantify, data are not timely and multiple sources are needed.

How many people moved from Italy to Spain in 2017?

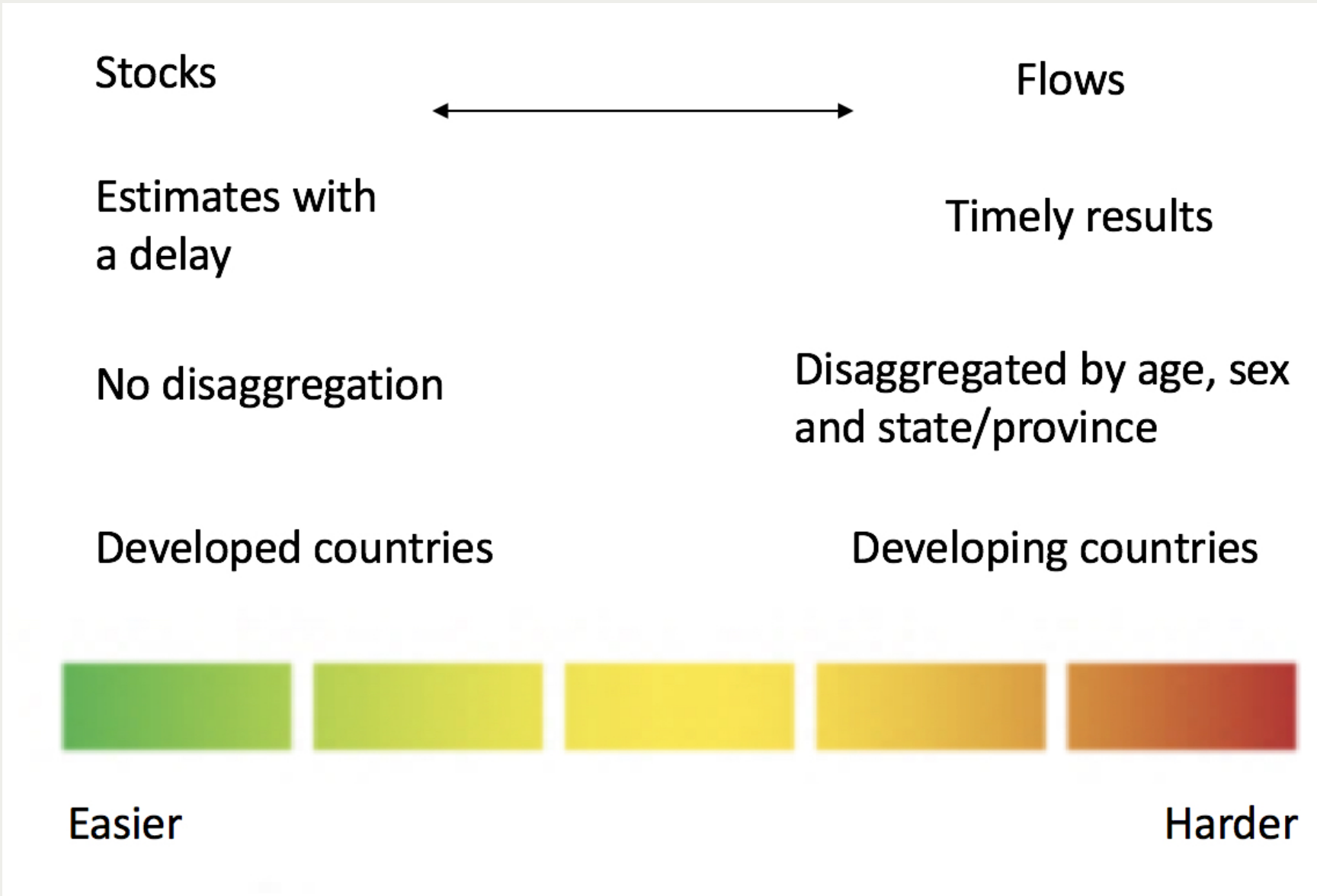
- We don't know with certainty, but...

$\Rightarrow$ Flows from Italy to Spain in 2015 (according to Spain): 17,350

$\Rightarrow$ Flows from Italy to Spain in 2015 (according to Italy): 5,003

- Flows are difficult to quantify, data are not timely and multiple sources are needed.

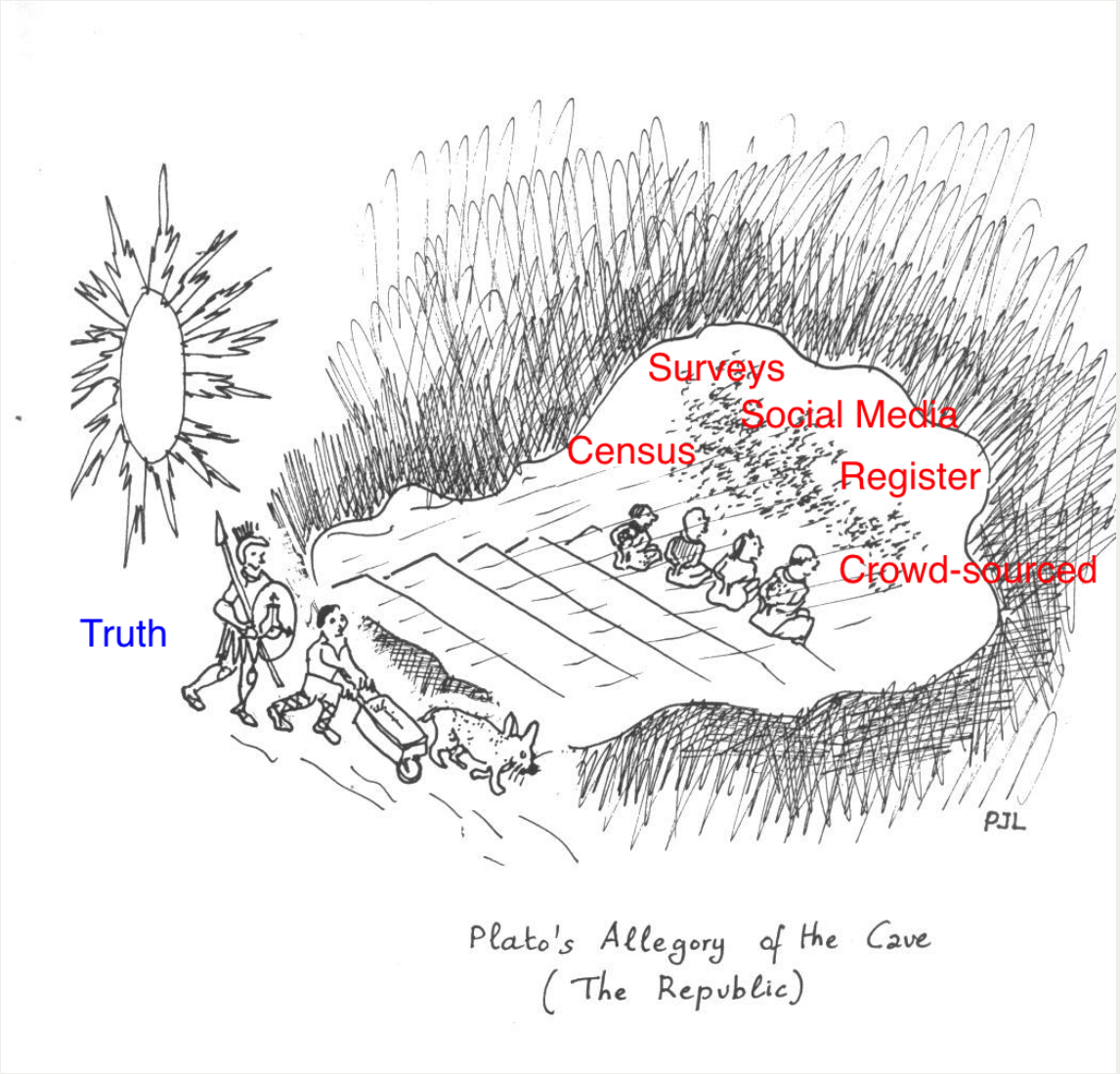
A gradient of difficulty for different types of questions



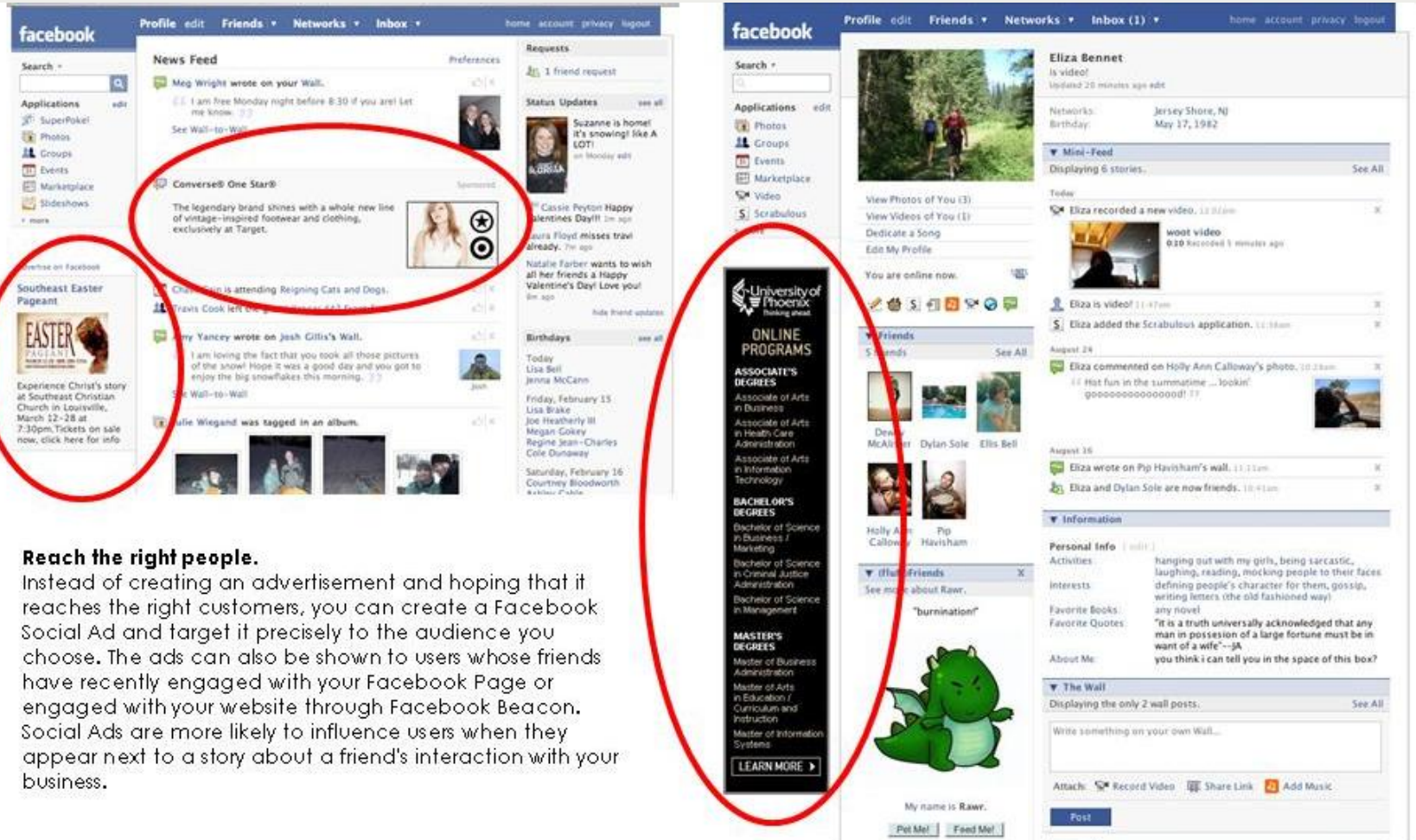
This presentation

1. Measuring stocks
2. Harmonizing flows
3. Evaluating cultural assimilation

Overarching framework: Combining Data Sources to Improve Estimates and Predictions

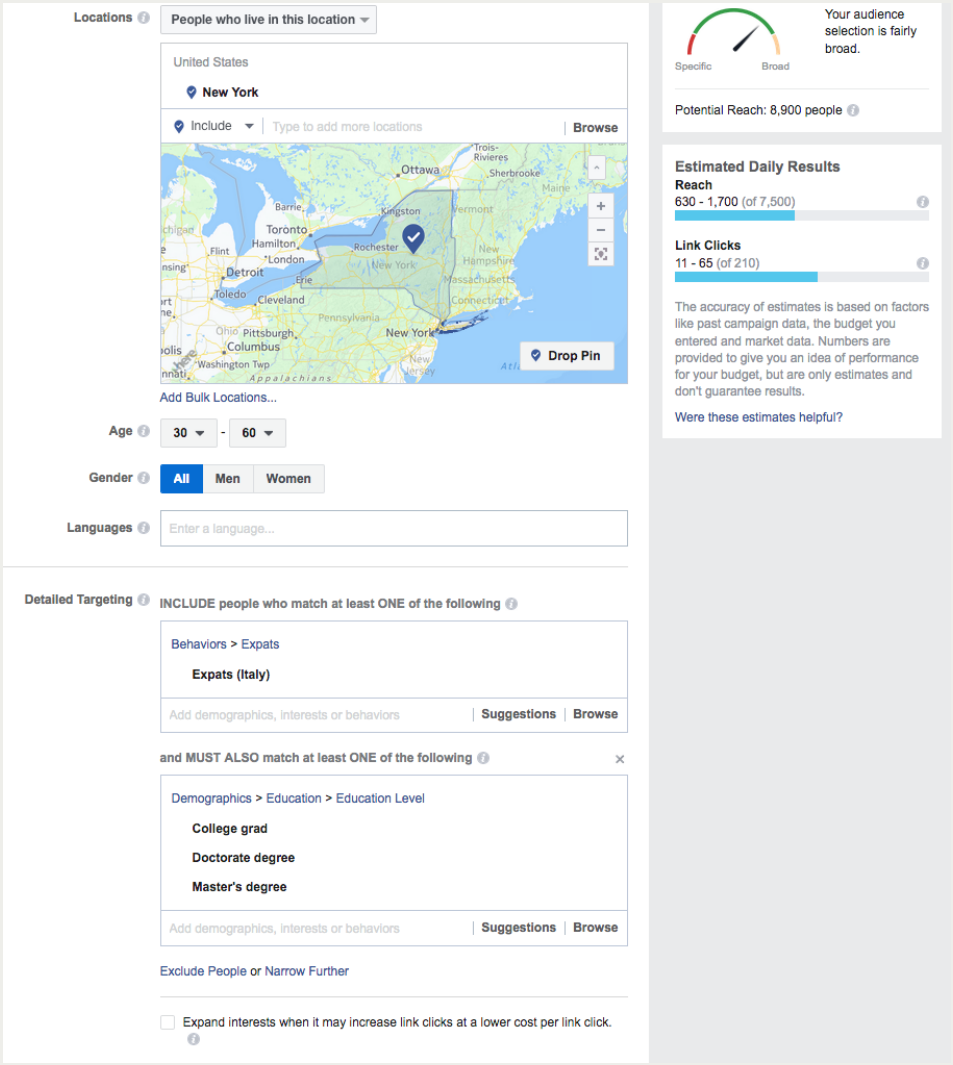


1. Measuring stocks of migrants with online (targeted) advertisement data



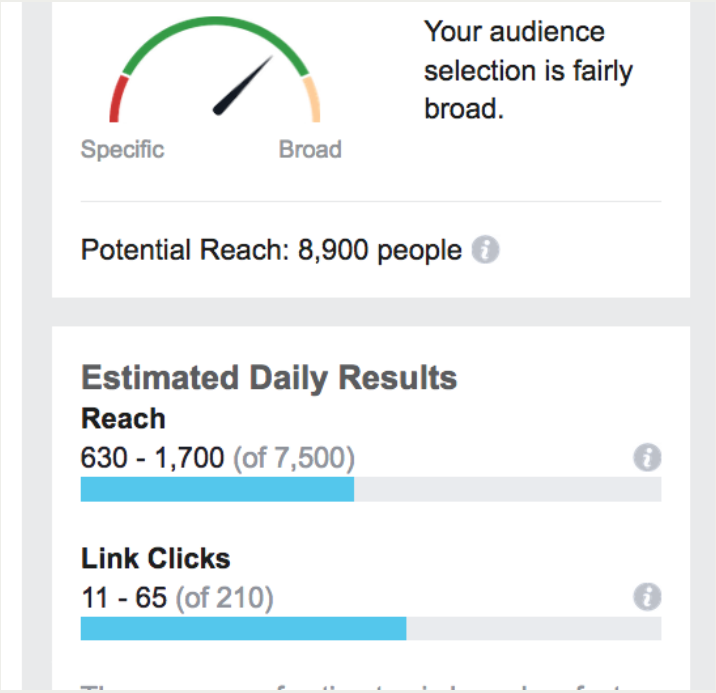
Reach the right people.
Instead of creating an advertisement and hoping that it reaches the right customers, you can create a Facebook Social Ad and target it precisely to the audience you choose. The ads can also be shown to users whose friends have recently engaged with your Facebook Page or engaged with your website through Facebook Beacon. Social Ads are more likely to influence users when they appear next to a story about a friend's interaction with your business.

Targeting a demographic group on Facebook



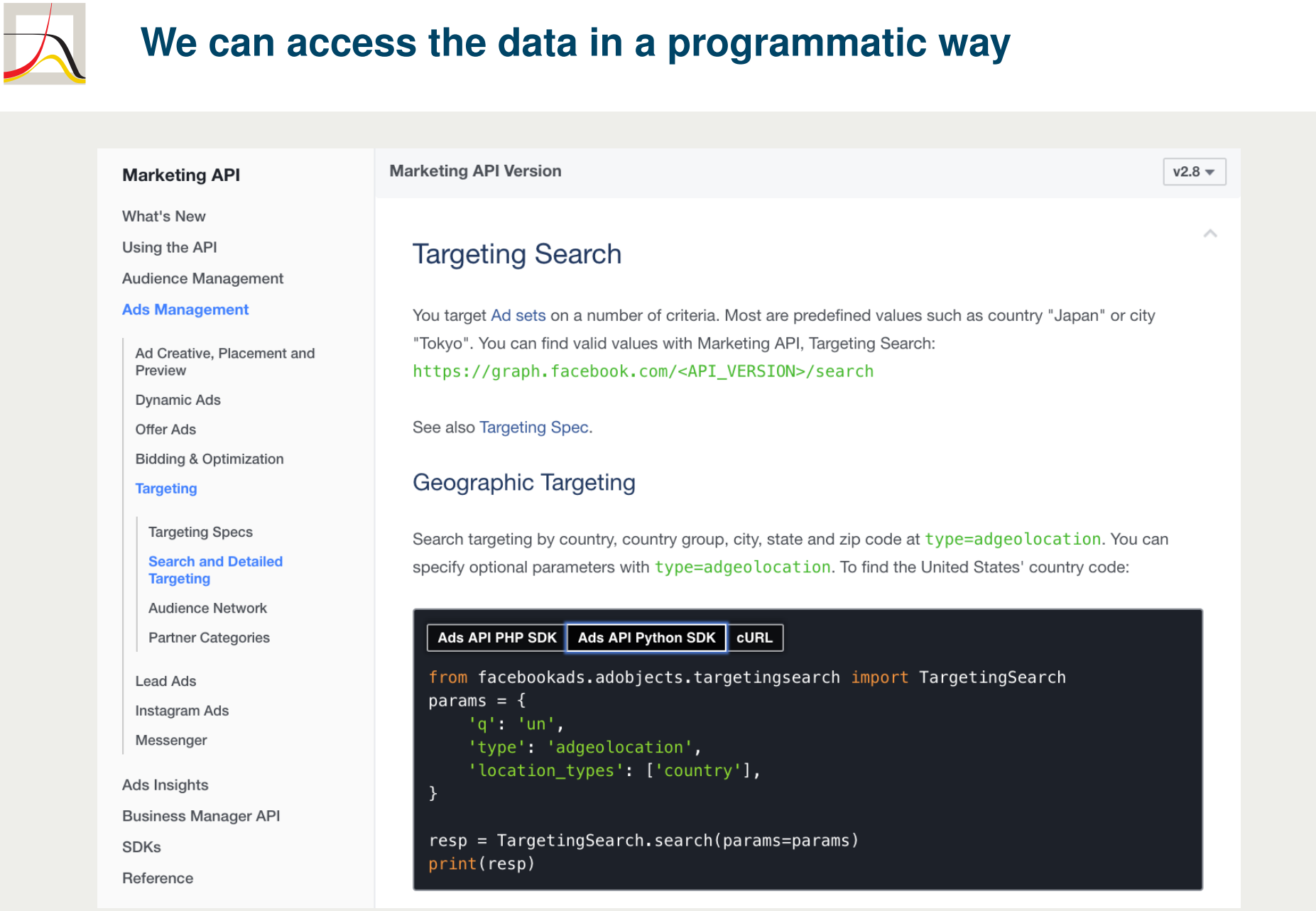
http://www.facebook.com/business

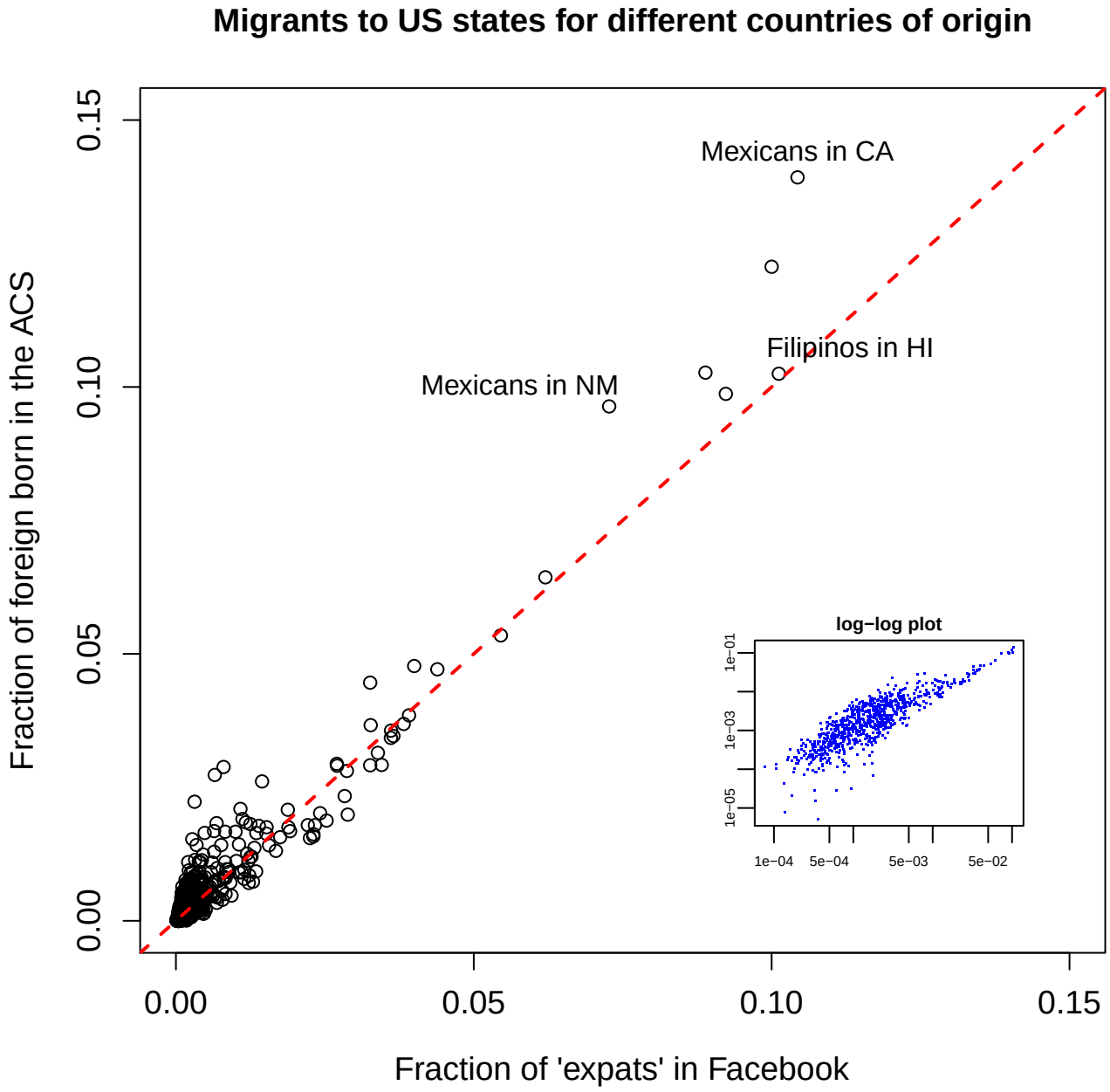
Targeting a demographic group on Facebook

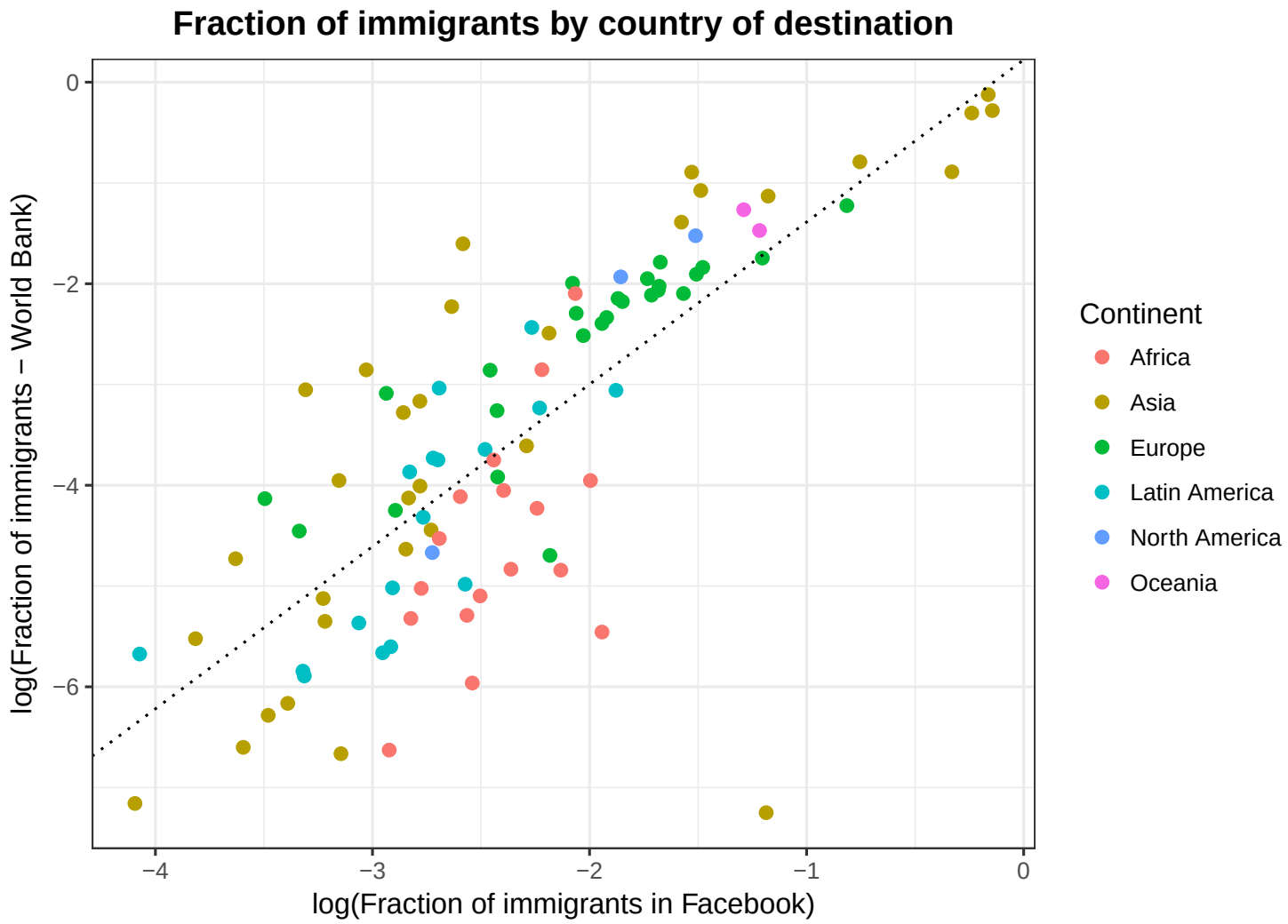


http://www.facebook.com/business

We can access the data in a programmatic way

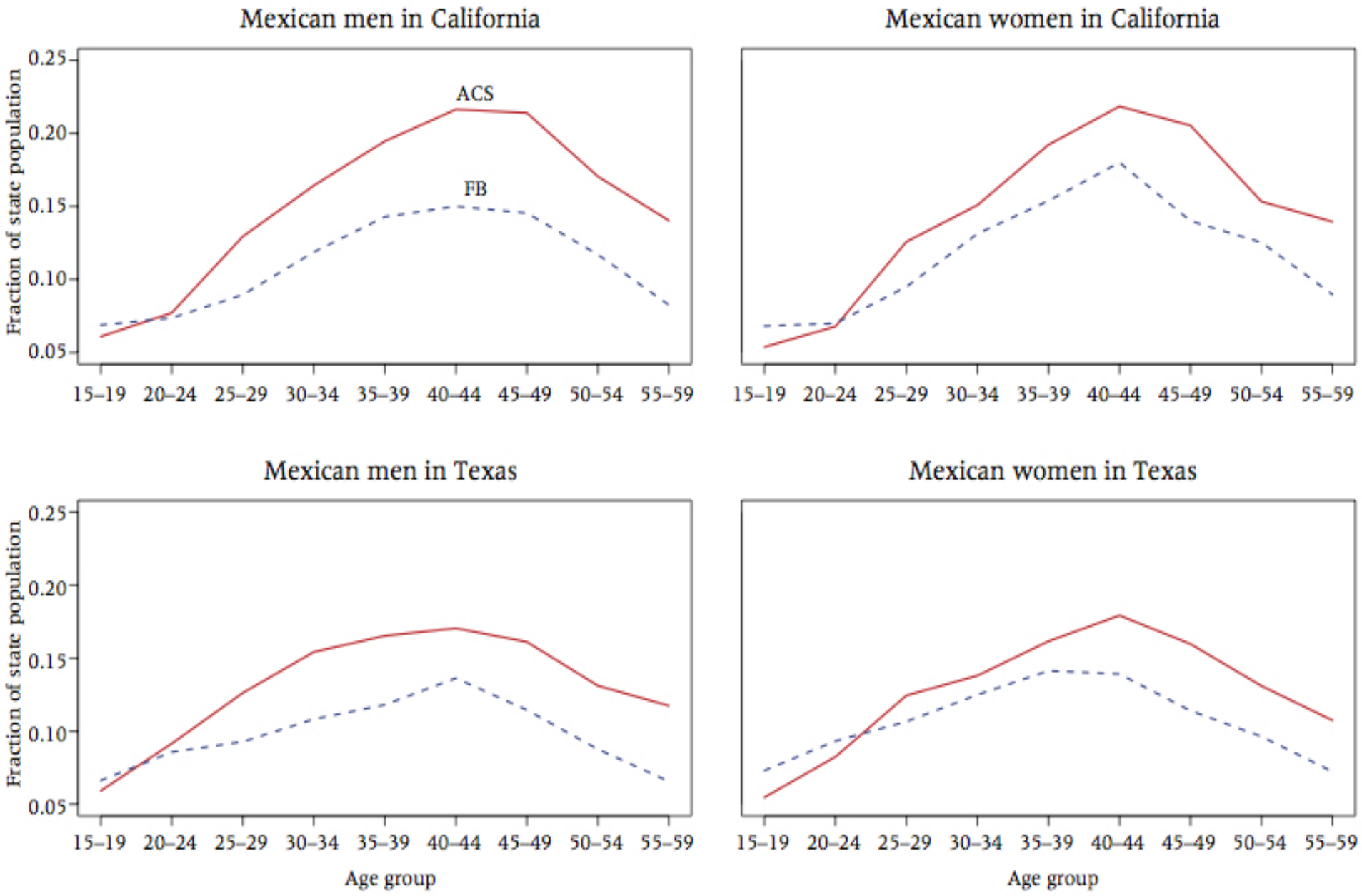



Migrants to US states for different countries of origin
Mexicans in CA
Filipinos in HI
Mexicans in NM
log−log plot
Fraction of foreign born in the ACS
Fraction of 'expats' in Facebook
0.00
0.05
0.10
0.15
1e−01
1e−03
1e−05
1e−04
5e−04
5e−03
5e−02


Fraction of immigrants by country of destination
log(Fraction of immigrants – World Bank)
0
−2
−4
−6
−4
−3
−2
−1
0
log(Fraction of immigrants in Facebook)
Continent
Africa
Asia
Europe
Latin America
North America
Oceania

Profiles by age and sex

FIGURE 3 Facebook and ACS profiles of stocks of migrants by age and sex for Mexicans in California and in Texas



Evaluating patterns in the bias

$$
\begin{aligned}
log(\text{ACS foreign born pop}_{ij}^{z}) = \beta_0 &+ \beta_1 log(\text{Facebook expats}_{ij}^{z}) + \\
&+ \beta_2 \mathbb{1}(\text{Origin 1}) + \cdots + \beta_{30} \mathbb{1}(\text{Origin 29}) + \\
&+ \beta_{31} \mathbb{1}(\text{Age group 1}) + \cdots + \beta_{38} \mathbb{1}(\text{Age group 8}) + \\
&+ \epsilon_{ij}^{z}
\end{aligned}
$$

Age pattern

Age group (20-24)	-0.483^{***} (0.032)
Age group (25-29)	-0.291^{***} (0.032)
Age group (30-34)	-0.010 (0.031)
Age group (35-39)	0.094^{**} (0.031)
Age group (40-44)	0.301^{***} (0.031)
Age group (45-49)	0.309^{***} (0.031)
Age group (50-54)	0.460^{***} (0.031)
Age group (55-59)	0.519^{***} (0.031)

Predictive capacity

- Goal: Predict the total number of foreign born from country *i* living in US state *j* (e.g, what is the stock of Mexicans in CA, Italians in NY?)
- Split the data into training set (80% of US states) and test set (remaining 20% of US states)
- Estimate a model with no age & country of origin disaggregation and a model with disaggregation

Predictive capacity

- Goal: Predict the total number of foreign born from country *i* living in US state *j* (e.g, what is the stock of Mexicans in CA, Italians in NY?)
- Split the data into training set (80% of US states) and test set (remaining 20% of US states)
- Estimate a model with no age & country of origin disaggregation and a model with disaggregation

Predictive capacity (continued)

The average out-of-sample Mean Absolute Percentage Error (MAPE) for total number of foreign born from country *i* living in state *j*:

- MAPE with no disaggregation by age-origin = 56%
- MAPE with disaggregation by age-origin = 37%

$\Rightarrow$ Indication that accounting for biases for subgroups of the population, in a way analogous to post-stratification, helps with predictions

Predictive capacity (continued)

The average out-of-sample Mean Absolute Percentage Error (MAPE) for total number of foreign born from country *i* living in state *j*:

- MAPE with no disaggregation by age-origin = 56%
- MAPE with disaggregation by age-origin = 37%

⇒ Indication that accounting for biases for subgroups of the population, in a way analogous to post-stratification, helps with predictions

Moving further: Combining Survey and FB data within a Bayesian framework

1. Calibrate Facebook data against the ACS $\Rightarrow$ use the latest Facebook data to generate predictions of the present
2. Model trends in age schedules by using ACS data and a principal component approach $\Rightarrow$ generate predictions in a way similar to the Lee-Carter model, but for migration schedules
3. Combine 1 and 2 within a Bayesian hierarchical model

Moving further: Combining Survey and FB data within a Bayesian framework

1. Calibrate Facebook data against the ACS $\Rightarrow$ use the latest Facebook data to generate predictions of the present
2. Model trends in age schedules by using ACS data and a principal component approach $\Rightarrow$ generate predictions in a way similar to the Lee-Carter model, but for migration schedules
3. Combine 1 and 2 within a Bayesian hierarchical model

Moving further: Combining Survey and FB data within a Bayesian framework

1. Calibrate Facebook data against the ACS $\Rightarrow$ use the latest Facebook data to generate predictions of the present
2. Model trends in age schedules by using ACS data and a principal component approach $\Rightarrow$ generate predictions in a way similar to the Lee-Carter model, but for migration schedules
3. Combine 1 and 2 within a Bayesian hierarchical model

Moving further: Combining Survey and FB data within a Bayesian framework

1. Calibrate Facebook data against the ACS $\Rightarrow$ use the latest Facebook data to generate predictions of the present
2. Model trends in age schedules by using ACS data and a principal component approach $\Rightarrow$ generate predictions in a way similar to the Lee-Carter model, but for migration schedules
3. Combine 1 and 2 within a Bayesian hierarchical model

1. Modeling Facebook Data

Consider e.g. Mexicans in California:

Fraction of migrants for age x at time t, from ACS data

$$logit(p_{xt}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot logit(p_{xt}^{\text{facebook}}) + X\hat{\beta}$$

Fraction of migrants for age x at time t from FB data

Indicator variables for age groups and country of origin

2. Modeling trends in age schedules from the ACS

Consider e.g. Mexicans in California:

Column vector of fraction of migrants by age x at time t

Average age schedule over time

Error

$$logit(p_{xt}) = a_x + \kappa_t b_x + \varepsilon_{xt}$$

Estimate via SVD

Time-varying scalar

Age deviations from the average schedule

Assumption that second differences for the time parameter are small ~ linear trends for short periods of time

$$\Delta^2 \kappa_t \sim N(0, \sigma_\kappa^2)$$

Assumption of autoregressive errors

$$\varepsilon_{xt} \sim N(\rho_x \varepsilon_{x,t-1}, \sigma_\varepsilon^2)$$

3. Combined Bayesian Hierarchical Model

$$
\begin{aligned}
\text{logit}\, p_{xt} &\sim N(\mu_{xt}, \sigma^2) \\
\mu_{xt} &= a_x + \kappa_t b_x + \varepsilon_{xt} \\
\Delta^2 \kappa_t &\sim N(0, \sigma_\kappa^2) \\
\varepsilon_{xt} &\sim N(\rho_x \varepsilon_{x,t-1}, \sigma_\varepsilon^2) \\
\rho_x &\sim U(-1, 1)
\end{aligned}
$$

2. Model for time series from ACS

Variance for nowcasts includes:

a. Sampling error from ACS
b. Uncertainty related to bias adjustment of FB data
c. Variability across FB data collections

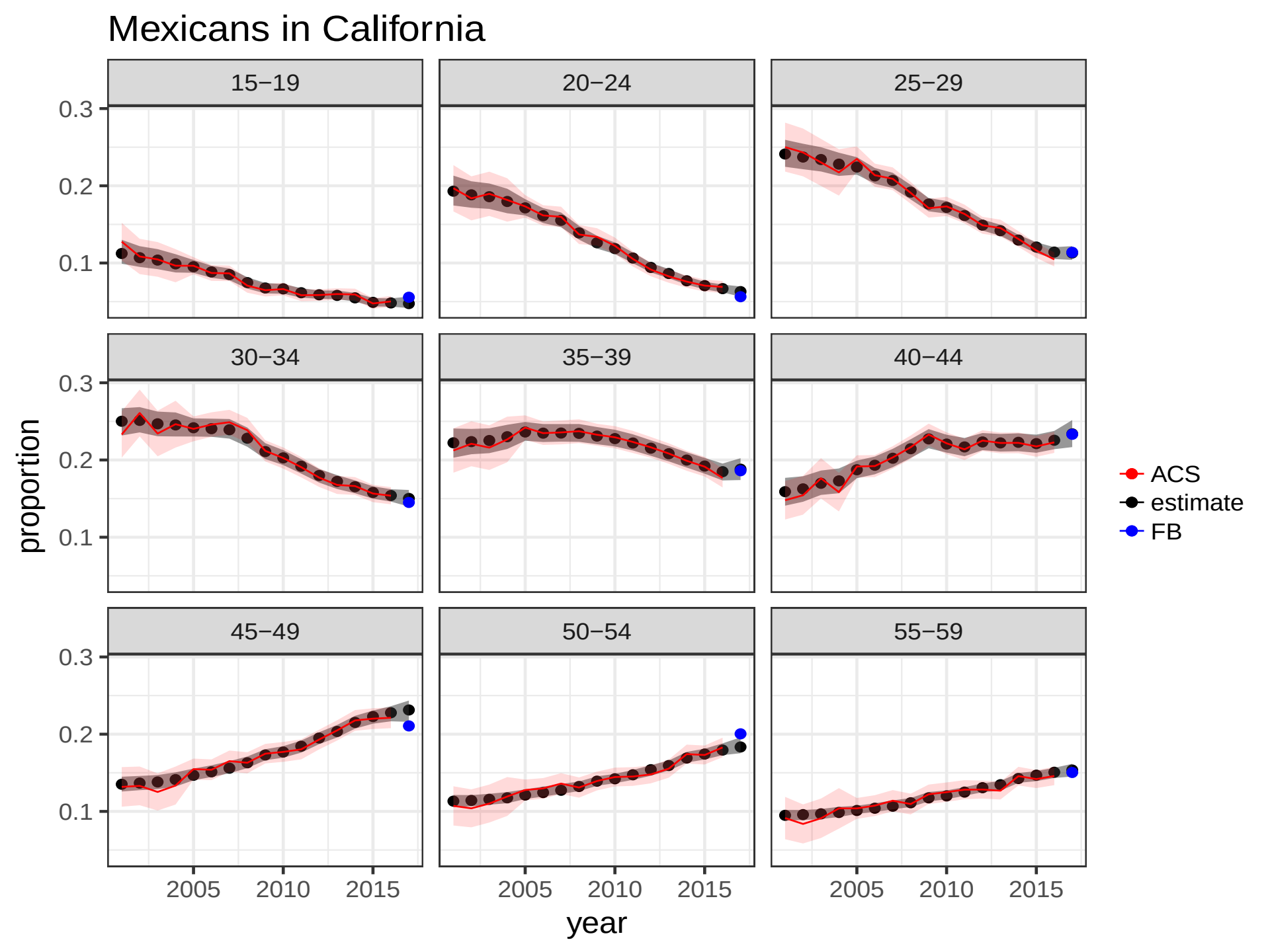
$$
\sigma^2 = \begin{cases} \sigma_p^2, & \text{if } 2001 \le t \le 2016 \\ \sigma_p^2 + \sigma_{bias}^2 + \sigma_{ns}^2, & \text{if } t = 2017 \end{cases}
$$

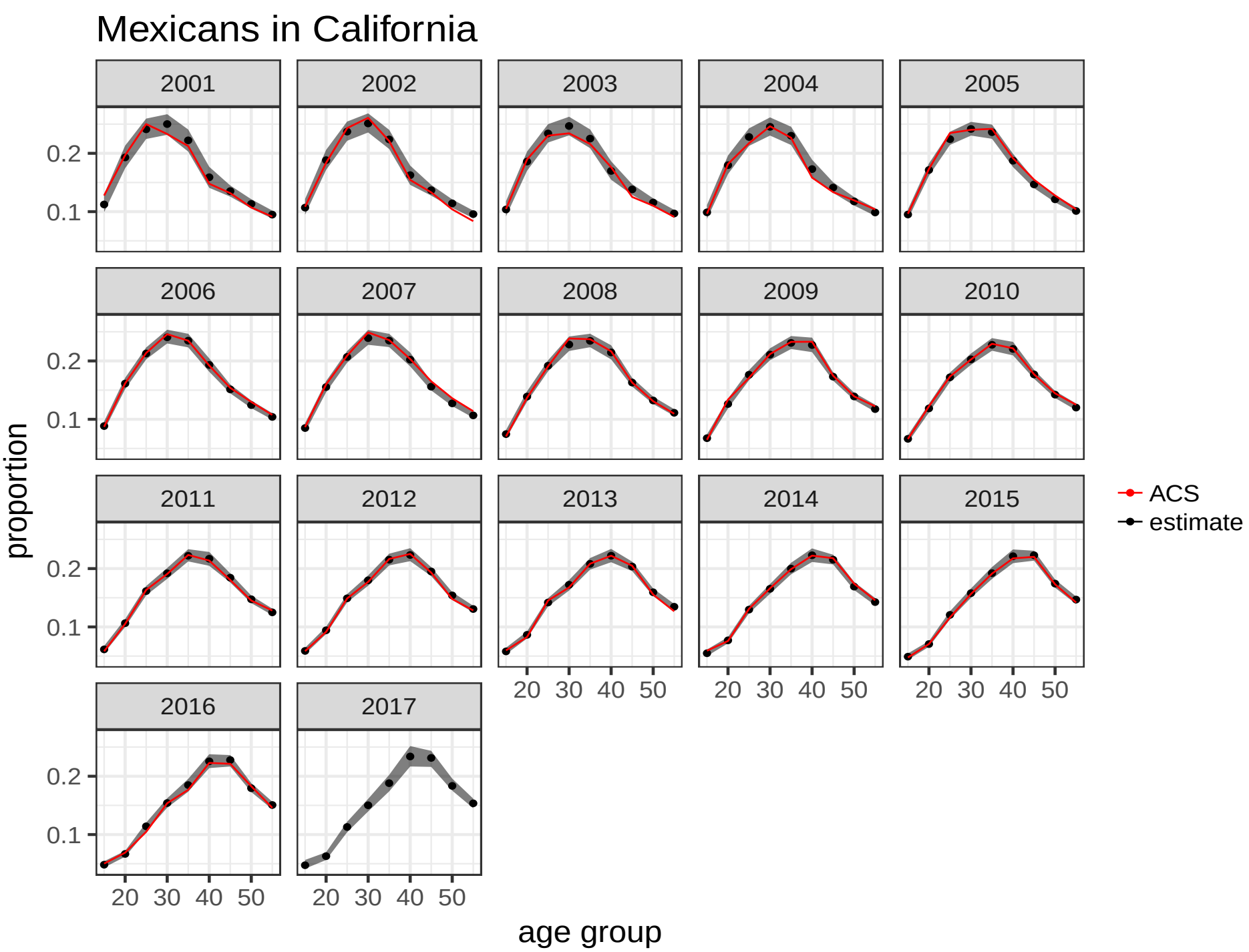
$$
p_{xt} = \begin{cases} \text{from ACS}, & \text{if } 2001 \le t \le 2016 \\ p_{xt}^* \text{ (estimated proportion)}, & \text{if } t = 2017 \end{cases}
$$

$$
\begin{aligned}
p_{xt}^* &= \overline{p_{xtj}^*} \\
p_{xtj}^* &= \hat{\beta}_0 + \hat{\beta}_1 \cdot p_{xtj}^{\text{facebook}} + X\hat{\beta}
\end{aligned}
$$

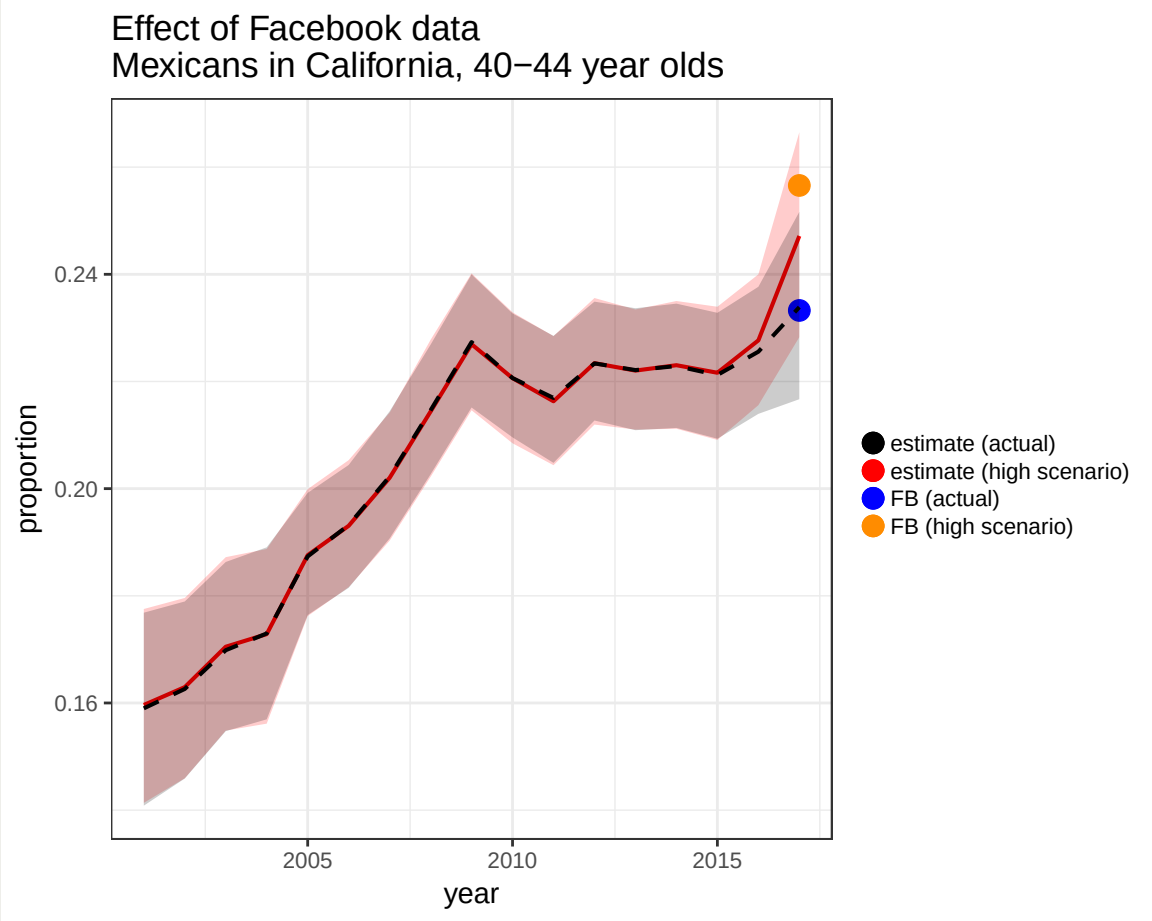
1. Model for Facebook data

Some Results


Mexicans in California
15–19
20–24
25–29
30–34
35–39
40–44
45–49
50–54
55–59
0.3
0.2
0.1
2005
2010
2015
proportion
year
ACS
estimate
FB


Mexicans in California
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
0.2
0.1
20
30
40
50
proportion
age group
ACS
estimate

Illustrative example: What if Facebook values suddenly increased by 10%?



The estimates from the Bayesian model would be 2.6% higher, on average, across all age groups

Key aspect: many different types of data can be combined within the same framework

2. Flows

Data

- Approximately 62,000 Twitter users who posted at least 8 geo-tagged tweets within the US between 2010 and 2013
- Their series of geo-located tweets were extended until September 2016
- Geographic coordinates were then coded at the level of Census Division

We considered flows between US Census divisions as a testbed for harmonization

Goal

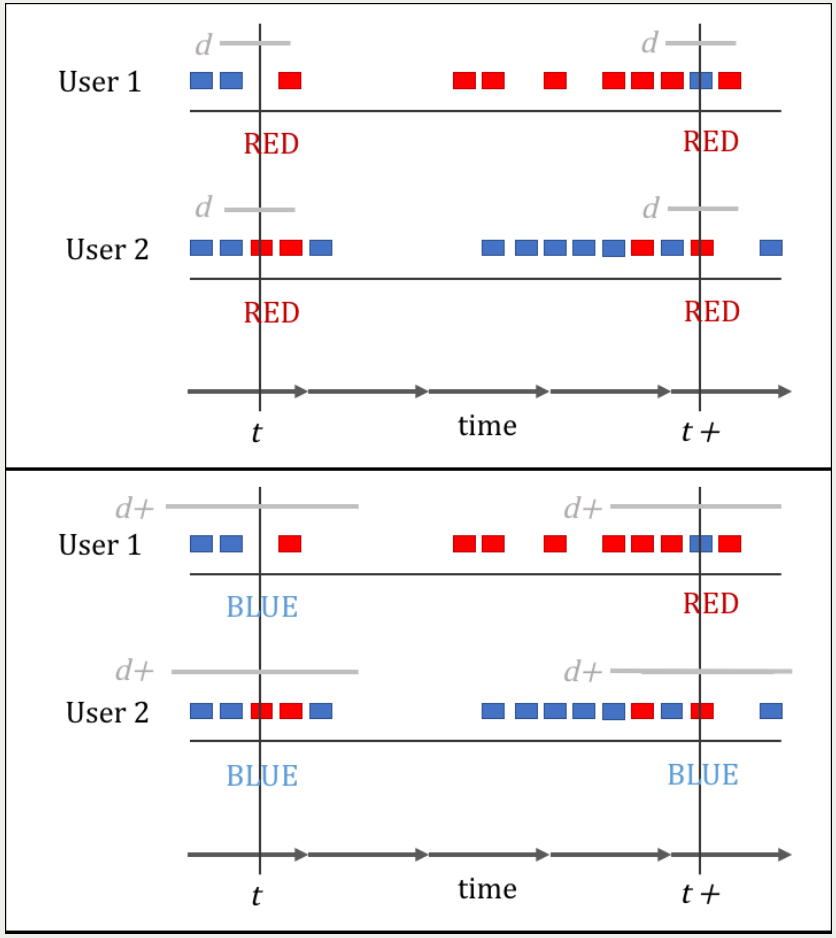
- To evaluate how mobility/migration rates vary as we change the definition of mobility/migration

$\Rightarrow$ Relevant for theory of mobility/migration and for harmonization of statistics across countries that use different definitions

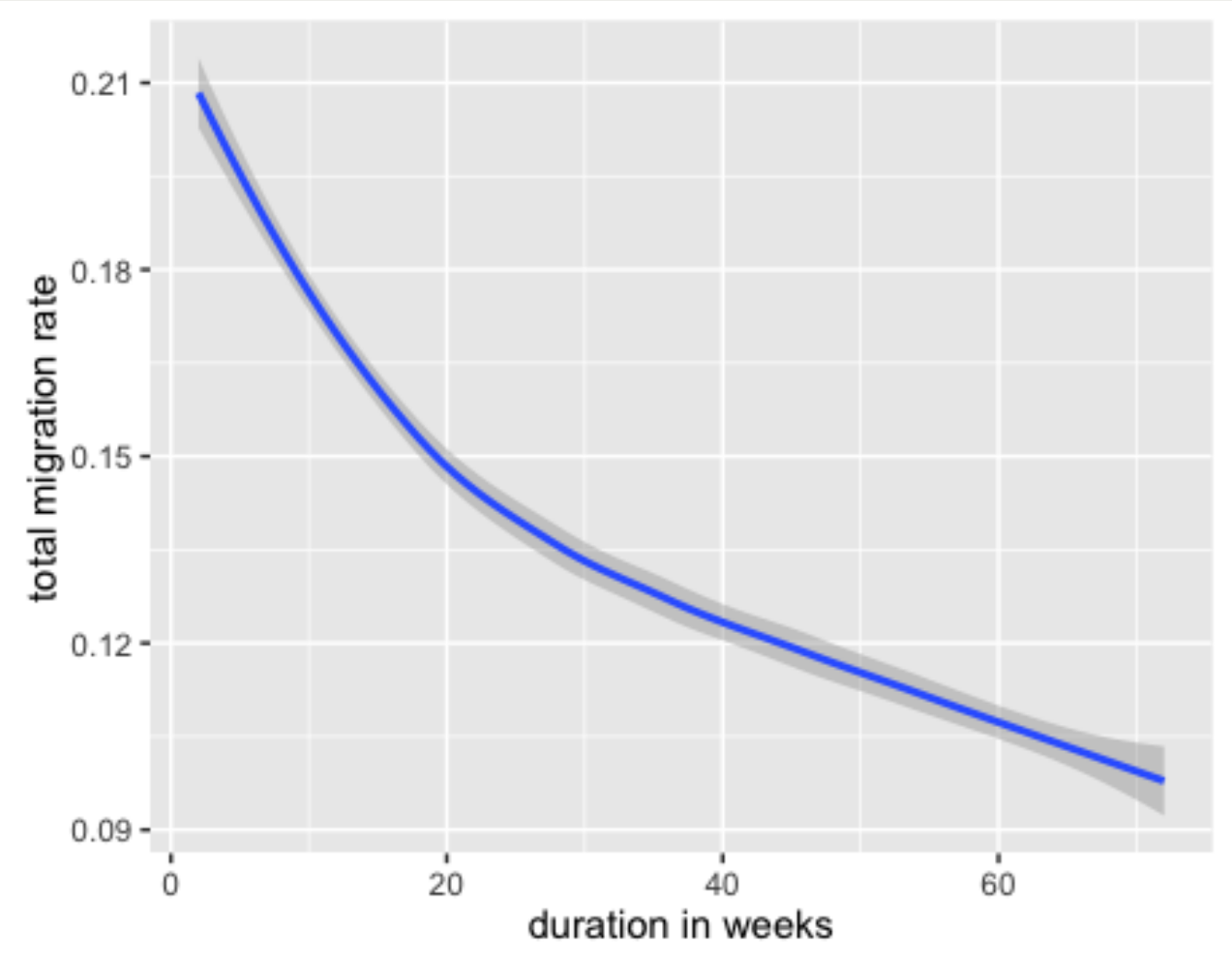
Two key concepts

- Buffer: the duration or window around a particular date used to impute residency
- Interval: the period between two dates used to evaluate migration

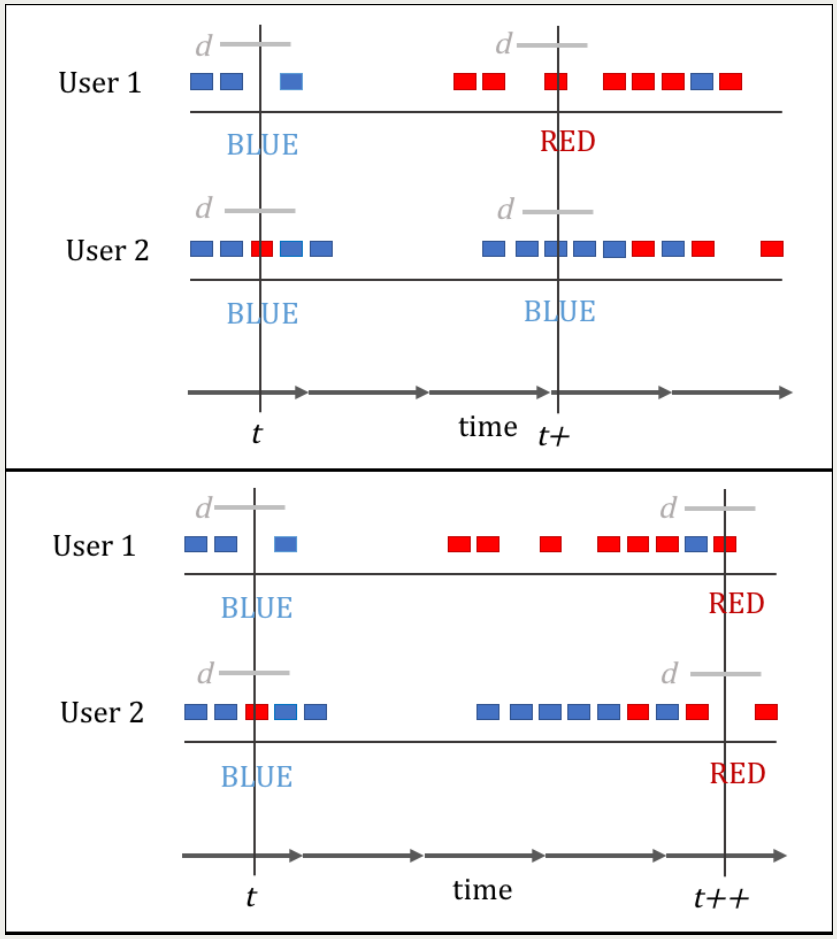
How do estimates of migration change when the 'buffer' changes?



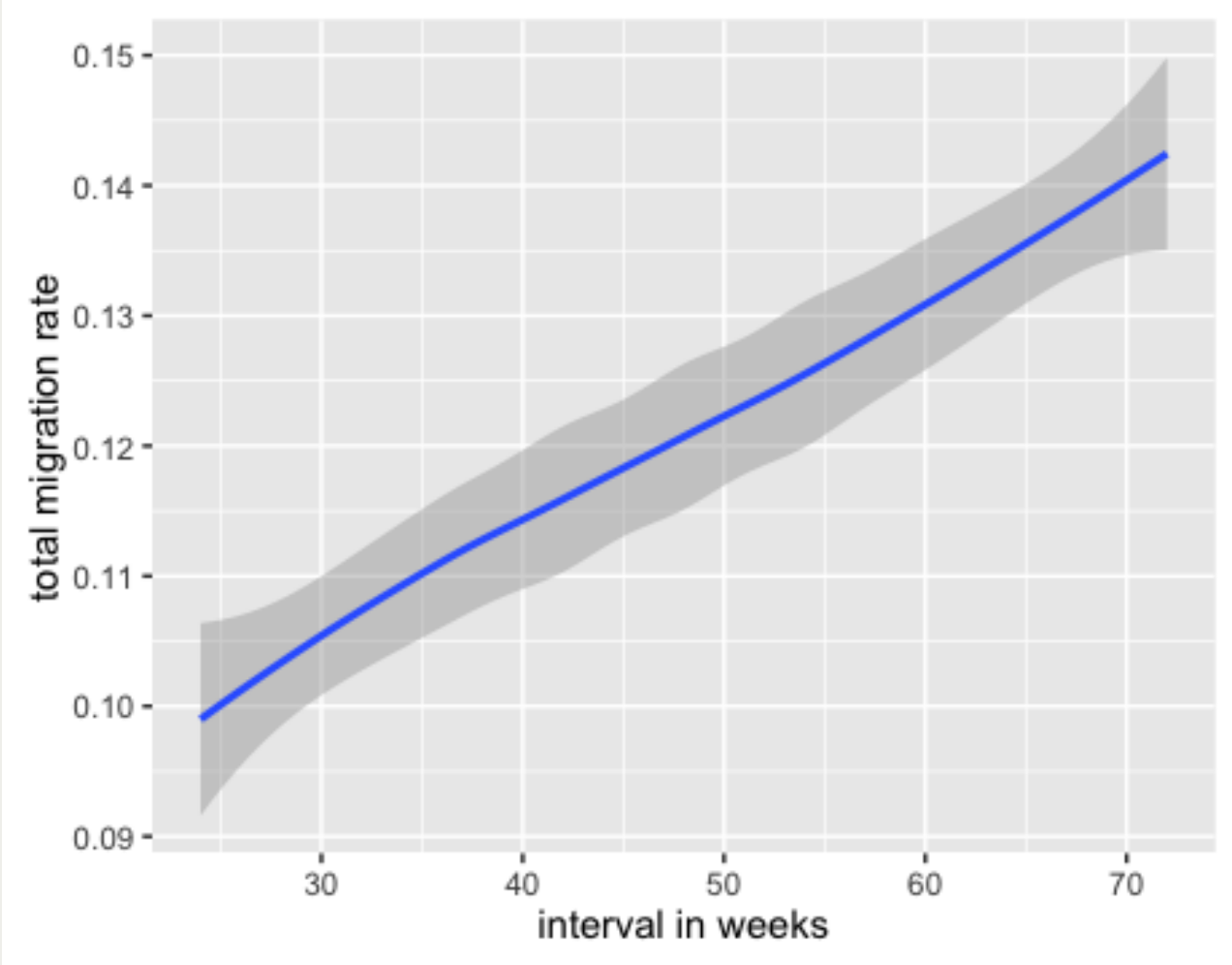
Migration rates as a function of the ‘buffer’ in our sample



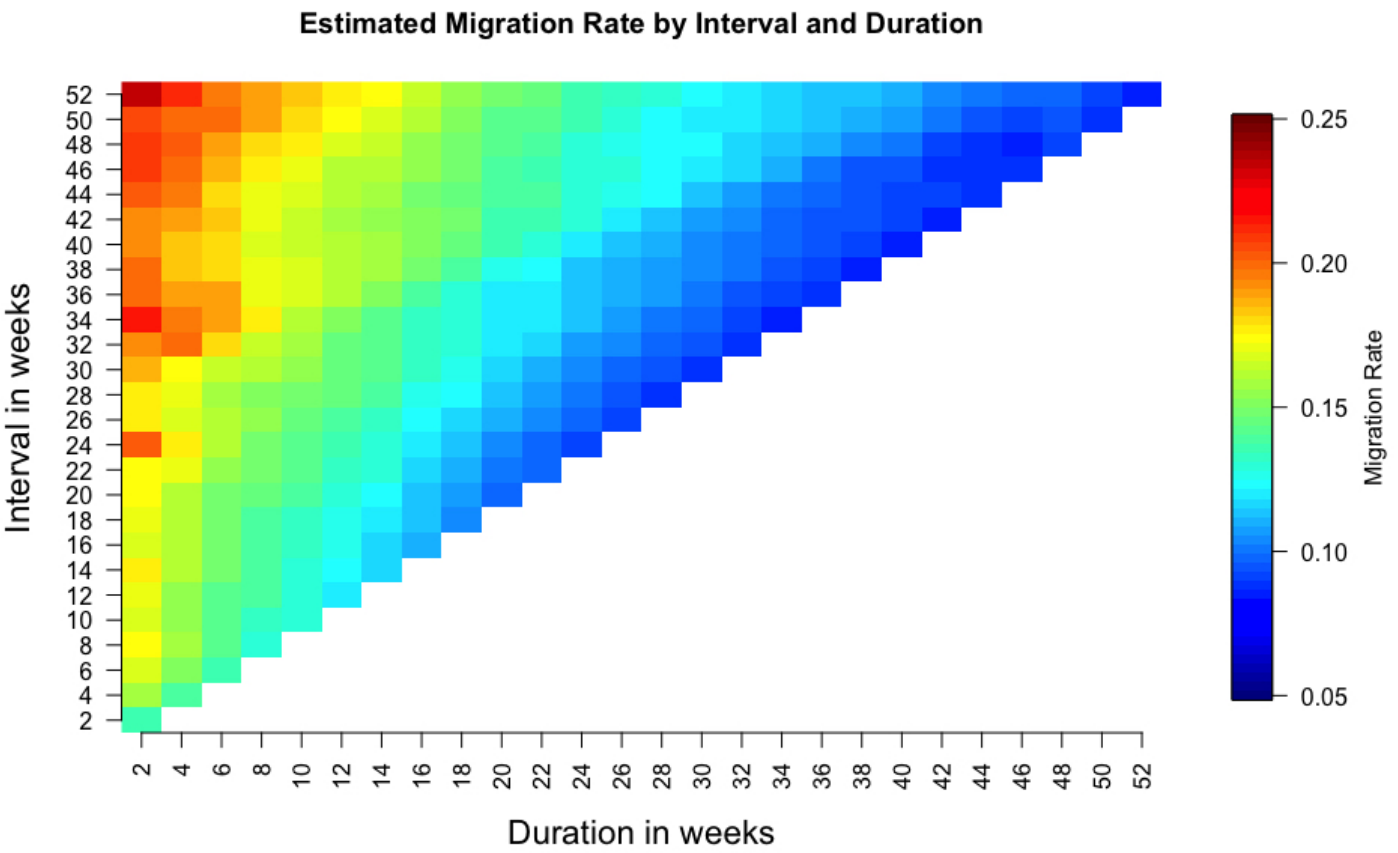
How do estimates of migration change when the 'interval' changes?



Migration rates as a function of ‘interval’ in our sample



Migration rates as a function of both ‘buffer’ and ‘interval’ in our sample



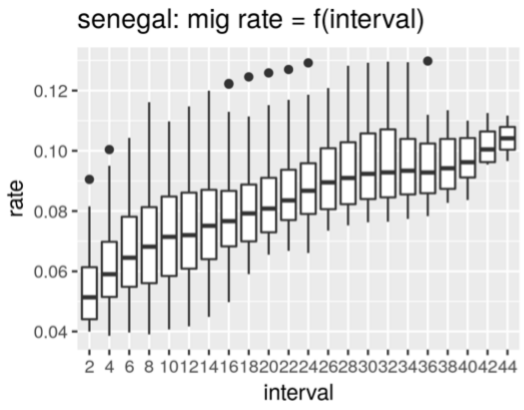
Some reflections

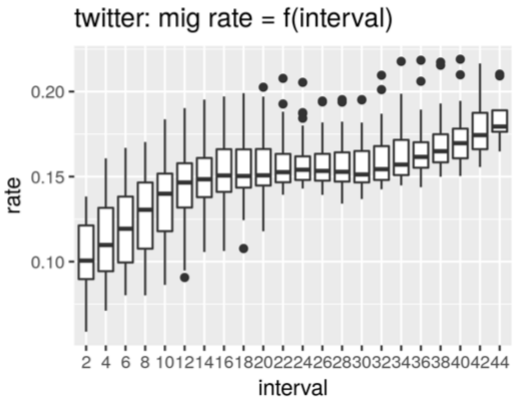
- These data enable us to generate migration histories.
- Example of opportunity to address questions that cannot be answered with traditional data... but no ‘ground truth’ available
- We are expanding our empirical analysis to include cellphone data (Senegal), other social media data (Gowalla) and potentially administrative data (e.g., Sweden)

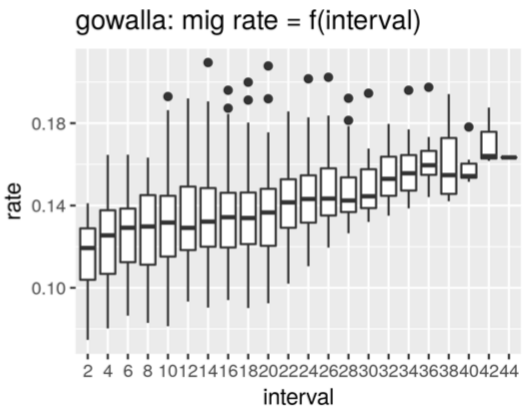
Some reflections

- These data enable us to generate migration histories.
- Example of opportunity to address questions that cannot be answered with traditional data... but no ‘ground truth’ available
- We are expanding our empirical analysis to include cellphone data (Senegal), other social media data (Gowalla) and potentially administrative data (e.g., Sweden)

Regularities across platforms







3. Cultural Assimilation

Key questions

- Can we measure cultural taste by examining online behavior?
- To what extent are interests/preferences of migrants similar or different to the ones of people in the origin or destination countries?

$\Rightarrow$ Develop metrics to evaluate distance in terms of prevalence of interests between immigrants and other populations or sub-populations

Cultural Assimilation in Germany

Table 3. Assimilation score (AS) for different choice of *Target* in Germany using the top 50% distinctly German interests of 2,907 interests. Lines 1–5 of the first column correspond to non-Arab populations. The remaining cells, all with the "A:" prefix, correspond to Arabic-speaking migrants and sub-groups of this population.

Target	AS	Target	AS
Austrian Migrants	.900	A: Men	.648
French Migrants	.803	A: Women	.503
Spanish Migrants	.864	A: Uni. Grad.	.637
Turkish-Sp. Non-Expats	.922	A: Not Uni.	.626
Turkish Speakers	.746	A: <18	.590
A: Arabic-Sp. Migrants	.643	A: 18–24	.665
A: Men, Uni. Grad., 18–24	.677	A: 25–44	.603
A: Men, Uni. Grad., 25–44	.620	A: 45–64	.504
A: Women, Not Uni., 45–64	.461	A: >64	.553

Assimilation in terms of Musical taste in Facebook in the US

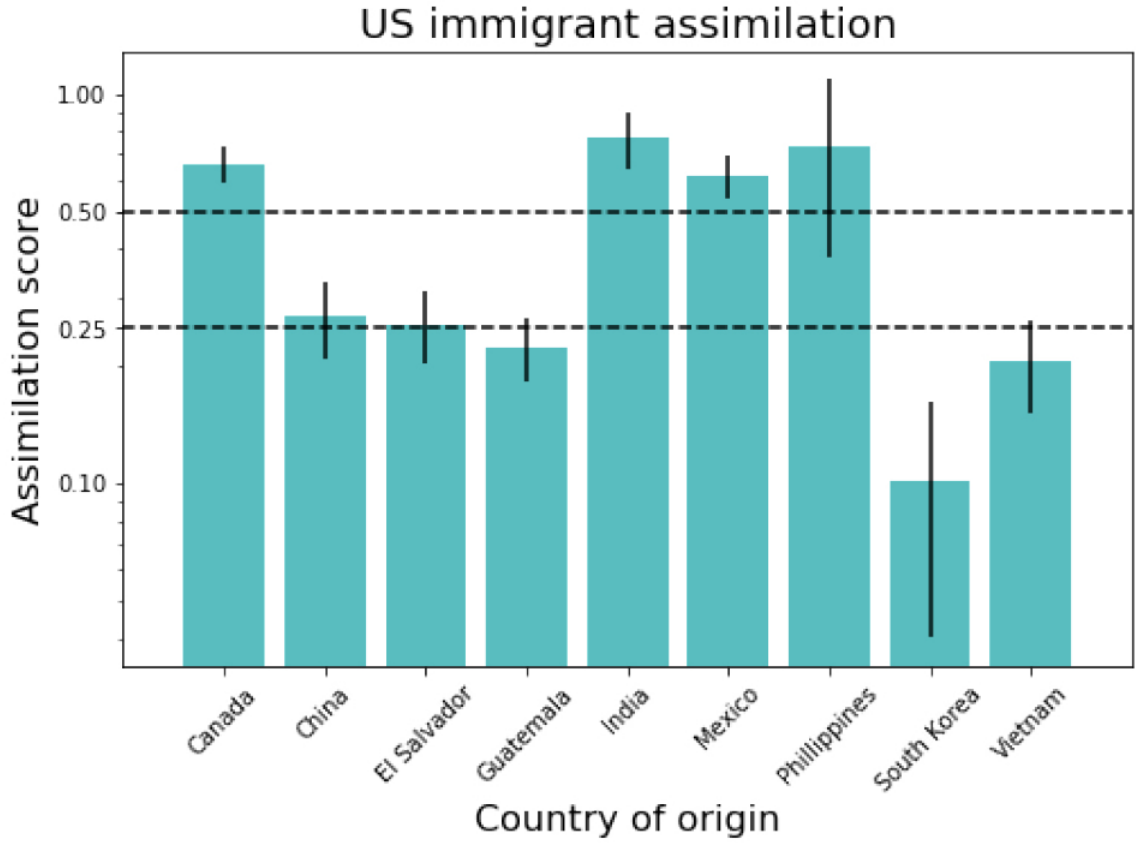
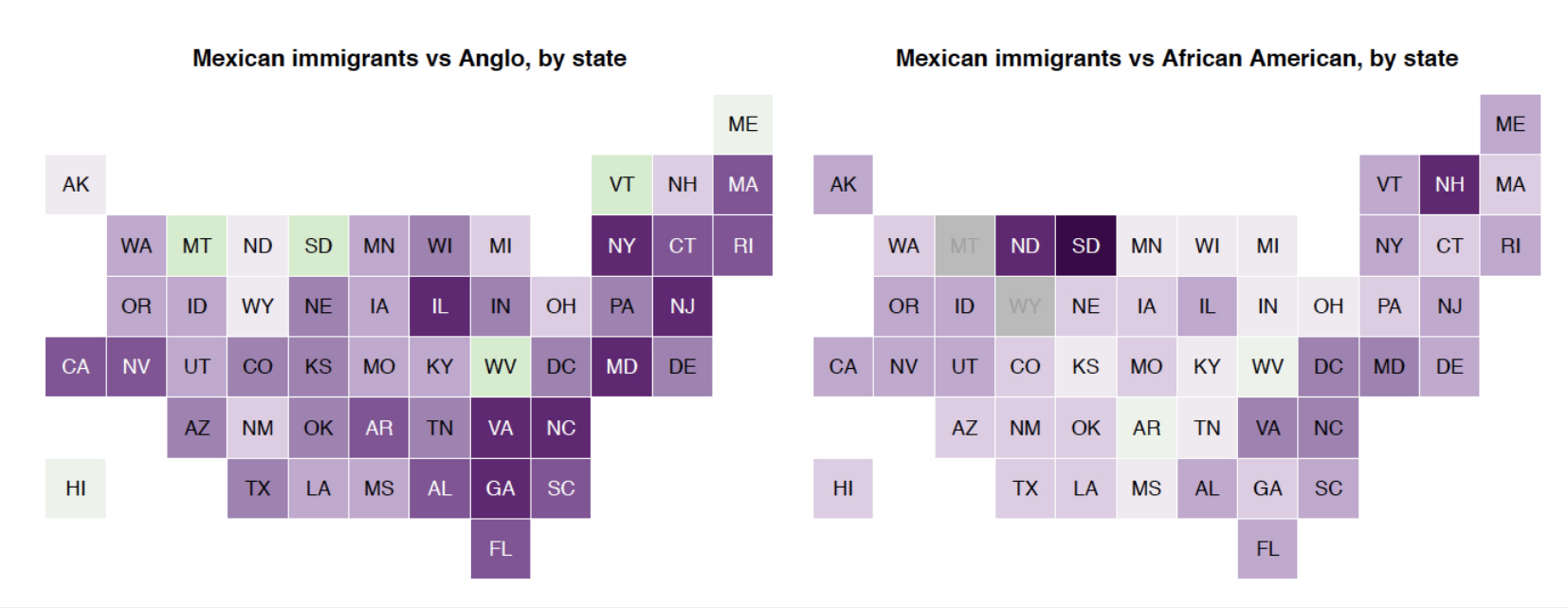


Figure 2: Median assimilation scores for selected immigrant groups. Error bars indicate 95% confidence intervals for the median.

Assimilation in terms of Musical taste in Facebook in the US



Lighter color = higher assimilation

Back to the outline

1. Measuring stocks
2. Harmonizing flows
3. Evaluating cultural assimilation

Some Reflections and Questions for Discussion

- Many additional untapped types of data: Bibliometric data, genealogies, call detail records, Linkedin, new forms of data collection, etc.
- How to build a flexible and resilient system?
- Raw data vs. modeling. What is the perspective of statistical offices?
- What are the incentives and constraints of statistical offices when working with academics in this space?

MAX-PLANCK-INSTITUT
FÜR DEMOGRAFISCHE
FORSCHUNG

MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC
RESEARCH

MAX-PLANCK-INSTITUT FÜR DEMOGRAFISCHE FORSCHUNG

MAX PLANCK INSTITUTE FOR DEMOGRAPHIC RESEARCH

Comments or Questions?

www.demogr.mpg.de