May 2025

README for "The Social Integration of International Migrants: Evidence from the Networks of Syrians in Germany" by Michael Bailey, Drew Johnston, Martin Koenen, Theresa Kuchler, Dominic Russel, and Johannes Stroebel

The data used in this paper is described in two places: in Section 1, we describe the data from Facebook on which we conduct most of our analyses, while in Section 4 and Appendix 13 we describe aggregated data that we use to explore regional correlates of integration outcomes. We include in this replication package the R and Stata code that we use to produce the figures and tables in the paper using these and other data.

Given the sensitive nature of the data used in the paper, we are unable to share this data as a part of this replication package. However, replication is possible for any researcher who obtains authorized access to the data directly from Facebook. Researchers interested in working with the de-identified confidential Facebook data can send inquiries to Michael Bailey (mcbailey@meta.com). Facebook works with outside researchers on a case-by-case basis through formal collaborations or internships after a satisfactory data use agreement has been agreed. In the past years, such collaborations were initiated with researchers at a number of universities, including LSE, Stanford, MIT, Harvard, NYU, Northwestern, Duke, and Boston College. The onboarding process may take several months, and researchers should expect to be subject to extensive background checks. The authors will assist with reasonable replication attempts for two years following publication.

In addition to the regression files and logs that are included in the replication package, some data preparation files are integrated into the Facebook codebase; researchers who have obtained internal access through Facebook will find these pipelines in the following folders:

- dataswarm-pipelines/tasks/bi/migration_academic/
- dataswarm-pipelines/tasks /measurementsystems/migration_academic/

Facebook is unable to provide a "stable" version of the raw data that does not change over time. As a result, each re-run of these data construction pipelines will produce a slightly different sample. However, the effects of this change in the sample on final estimates are likely to be negligible.

Steps for internal analysis replication

To produce our analyses using individual-level Facebook data, a researcher can run the following file from a Facebook-provided devserver:

- **dataswarm-pipelines/tasks/bi/migration_academic/run_all.sh**

This file runs the following files:

1. **dataswarm-pipelines/tasks/bi/migration_academic/make_loc_crosswalks.py**

   This step uses shapefiles to build a crosswalk between different geographies (e.g., city and NUTS3).

2. **dataswarm-pipelines/tasks/bi/migration_academic/make_eu_migrant_spine.py**

   This step gathers basic information about users' home region predictions.

3. **dataswarm-pipelines/tasks/measurementsystemts/migration_academic/enrich_posts.py**

   This step gathers information about the language used in users' public posts.

4. **dataswarm-pipelines/tasks/bi/migration_academic/build_paper_tables.py**

   This step performs the bulk of the preprocessing necessary to gather our analysis data sets.

5. **dataswarm-pipelines/tasks/bi/migration_academic/run_all_but_hs.do, which runs Stata files to produce many of our results. A version of these files is provided publicly as a part of this repository in the folder "internal_analysis/code"**
    - **sy_regressions.do**
    - **native_regressions.do**
    - **other_t5_regressions.do**

6. **dataswarm-pipelines/tasks/bi/migration_academic/run_all.R, which runs R files to produce many of our results. A version of these files is provided publicly as a part of this repository in the folder "internal_analysis/code"**
    - **migration_fb_vs_true.R**
    - **individual_summary_stats.R**
    - **integration_over_time.R**
    - **regional_integration.R**
    - **movers_plots_ntiles.R**
    - **movers_plots_levels.R**
    - **movers_plots_language.R**
    - **friendship_initiation.R**
    - **survey_analysis.R**

7. **dataswarm-pipelines/tasks/bi/migration_academic/high_school.do, which produces our results in Section 6. A version of this file is provided publicly as a part of this repository in the folder "internal_analysis/code"**

A version of the files in steps 5-7 are included in the replication package, but they cannot be run as the data used in them is proprietary.

Steps for regional analysis replication

To produce our analyses at the regional level (inside the folder "regional analysis"), after gaining access to the Facebook data a researcher can do the following:

1. Download the replication package and install all necessary software listed below
2. In Stata, set the working directory to the main folder directory and run "run_all.do".
3. In R, set the working directory to the main folder directory and run "run_all.R"

Again, the files for these steps are included in the replication package (in the "regional analysis/code") folder, but they cannot be run as the data used in them is proprietary.

Software Dependencies (Internal Analysis)

In the internal analysis code, we use the following packages on R 4.1.2:

- tidyverse (2.0.0)
- ggplot2 (3.5.1)
- gridExtra (2.3)
- weights (1.0.4)
- RColorBrewer (1.1-3)
- stringr (1.5.1)
- Hmisc (5.1-2)
- dineq (0.1.0)
- lfe (3.0-0)
- stargazer (5.2.3)
- gt (0.10.1)
- data.table
- dtplyr (1.3.1)
- binsreg (1.0)
- ggpubr (0.6.0)

- fixest (0.12.0)

This software will be pre-installed on the server used to run the analysis code.

For the Stata portion of the code, we use Stata version 16.0 and the following packages:

- ftools (version 2.31.3)
- reghdfe (version 5.7.3)
- binscatter (version 7.02)
- xml_tab (version 3.35)
- estout (version 3.23)
- winsor2 (version 1.1)

Software Dependencies (Regional Analysis)

In the regional analysis, we use the following packages on R 4.4.2:

- tidyverse (version 2.0.0)
- sf (version 1.0-19)
- ggplot2 (version 3.5.2)
- RColorBrewer (version 1.1-3)
- haven (version 2.5.4)
- corrplot (version 0.95)
- oiplot (version 0.2)
    - Note that this package is not hosted on CRAN, and can be downloaded from: https://github.com/OpportunityInsights/oiplot/
- ggrepel (version 0.9.6)

For the Stata portion of the code, we use Stata version 18.0 and the following packages:

- renvars (version 2.4.0)
- eststo (version 1.1.0)
- esttab (version 2.0.6)
- ivreg2 (version 4.1.11)
- binscatter (version 7.02)

Data Sources (Public Data for Regional Analysis)

- ./replication_regional_data/raw/geo/Zuordnung-Agenturen.xlsx
    - File containg address information about job centers
    - Used to create crosswalk to assign job centers to counties
    - Downloaded from Bundesagentur für Arbeit (Federal Employment Agency) (click here)
- ./replication_regional_data/raw/apprenticeships/analyse-d-0-*.xls
    - A series of files with information on apprenticeships
    - Downloaded from Bundesagentur für Arbeit (Federal Employment Agency): Bewerber und Berufsausbildungsstellen: Analysedaten (click here)
- ./replication_regional_data/raw/crime/ *
    - A series of files with crime reports at county level
    - Naming convention differs across years
    - Downloaded from Bundeskriminalamt: Polizeiliche Kriminalstatistik (click here)
- ./replication_regional_data/raw/economy/vgrdl_r2b1_bs2019.xlsx
    - County-level GDP per capita
    - Downloaded from Statistische Ämter des Bundes und der Länder (Federal and state statistical offices) (click here)
- ./replication_regional_data/raw/economy/vgrdl_r2b2_bs2019_2.xlsx
    - County-level average income and average hourly wage
    - Downloaded from Statistische Ämter des Bundes und der Länder (Federal and state statistical offices) (click here)
- ./replication_regional_data/raw/elections/european_elections2019.csv
    - European elections results 2019
    - At county level

    - Downloaded from Bundeswahlleiter
- ./replication_regional_data/raw/elections/ew14_kerg.csv
    - European elections results 2014
    - At county level
    - Downloaded from Bundeswahlleiter
- ./replication_regional_data/raw/integration_courses/*
    - Various files containing information integration courses: number of courses, number of participants, number of eligible people
    - Downloaded from Federal Office for Migration and Refugees: Integrationskursgeschäftsstatistik (click here)
- ./replication_regional_data/raw/labor_market_stats/eu-heft-d-0-*.xlsx
    - Various labor market statistics where each file corresponds to a quarter
    - At county level
    - Downloaded from Bundesagentur für Arbeit (Federal Employment Agency): Beschäftigte nach Staatsangehörigkeiten (click here)
- ./replication_regional_data/raw/migration_matrix/*
    - Migration matrices across counties over
    - acquired from Statistisches Bundesamt (Kreiswanderungsmatrizen)
- ./replication_regional_data/raw/population/12411-0018.csv
    - Population by age and gender over time at the county level
    - Downloaded from German Statistical Office: 12411-0018 (click here)
- ./replication_regional_data/raw/population/12521-0041.csv
    - Population by country of origin at the county level
    - Downloaded from German Statistical Office: 12521-0041 (click here)
- ./replication_regional_data/raw/population/migration_integration_regionen/ migration_integration_regionen_daten.csv
    - Population by migration status at the county level
    - Downloaded from German Statistical Office: migration_integration_regionen_daten.csv (click here)
- ./replication_regional_data/raw/geo/VG250_GEM_data.dta
    - Attribute table of shapefile of German counties (VG250_GEM)
    - Downloaded from Federal Agency for Cartography and Geodesy: vg250_12-31.utm32s.shape.kompakt
- ./replication_regional_data/raw/proasyl_groups/MitmachenKarte.xlsx
    - List of ProAsyl groups
    - Not public; received by email from contact at ProAsyl
- ./replication_regional_data/raw/religion/2000S-1022.csv
    - Downloaded from Zensus-Datenbank (click here)
- ./replication_regional_data/integration_sports_clubs/stützpunktvereine.xlsx
    - List of sports clubs being part of “Integration durch Sport” initiative (integration sports clubs)
    - Downloaded from German Olympic Sports Confederation: Integration durch Sport (click here)
- ./replication_regional_data/raw/teachers/315840_Koenen.xlsx
    - County-level teacher unemployment by type of teacher
    - Not public; acquired from: Service-Haus der Bundesagentur fuer Arbeit
- ./replication_regional_data/raw/urban_rural/Ein- und Zweifamilienhäuser_Kreise_2017.xlsx
    - County-level data on fraction of houses that are designed for one or two familes only as of 2017
    - Downloaded from Thünen-Landatlas (by Bundesministerium für Ernährung und Landwirtschaft) (click here)
- ./replication_regional_data/raw/urban_rural/Wohnfläche_Kreise_2017.xlsx
    - County-level data on average living space (Wohnfläche) per person as of 2017
    - Downloaded from Thünen-Landatlas (by Bundesministerium für Ernährung und Landwirtschaft) (click here)
- ./replication_regional_data/raw/urban_rural/Wohnungsleerstand_Kreise_2011.xlsx
    - County-level data on percent of flats that are empty as of 2011
    - Downloaded from Thünen-Landatlas (by Bundesministerium für Ernährung und Landwirtschaft) (click here)
- ./replication_regional_data/raw/urban_rural/LTE_Kreise_2019.xlsx
    - County-level data on LTE coverage as of 2019
    - Downloaded from Thünen-Landatlas (by Bundesministerium für Ernährung und Landwirtschaft) (click here)
- ./replication_regional_data/raw/geo/cw_krs_plz.csv
    - Result of intersection of shapefile of German counties (VG250_GEM, see above) and postal codes

    - Intersection conducted in QGIS
    - Shapefile of postal codes downloaded from ArcGIS HUB (click here)
- ./replication_regional_data/raw/geo/cw_krs_plz_points.csv
    - Assignment of postal codes to German counties based on which county the centroid of postal code is located in.
    - Analysis conducted in QGIS
- ./replication_regional_data/raw/map_shapefiles/*
    - Shape files for NUTS regions in Europe
    - Downloaded from © EuroGraphics (click here)
- ./replication_regional_data/raw/european_social_survey/*
    - NUTS1 level responses to the European Social Survey, used in Appendix L
    - Download from European Social Survey website (click here)
- ./replication_regional_data/raw/european_values_survey/ZA7504_EVS-Trend_sensitive_V2.zip
    - NUTS3 level responses to the European Values Study, used in Appendix L
    - Restricted access data acquired from European Values Study containing respondents' NUTS3
- ./replication_regional_data/raw/world_values_survey/outIntegrated_values_surveys_1981-2021.dta
    - Includes NUTS2 level responses to European Values Study, used in Appendix L
    - Downloaded from European Values Study website (click here)