Social Connectedness Index
Data Notes

Background and Construction of Social Connectedness Index

The *Social Connectedness Index* uses an anonymized snapshot of active Facebook users and their friendship networks to measure the intensity of social connectedness between locations. Users are assigned to locations based on their information and activity on Facebook, including the stated city on their Facebook profile, and device and connection information.

Formally, the *Social Connectedness Index* between two locations *i* and *j* is defined as:

$$Social\ Connectedness\ Index_{i,j} = \frac{FB\_Connections_{i,j}}{FB\_Users_i * FB\_Users_j}$$

Here, FB_Users_i and FB_Users_j are the number of Facebook users in locations *i* and *j*, and $FB\_Connections_{i,j}$ is the total number of Facebook friendship connections between individuals in the two locations.[1]

The publicly available measures of the *Social Connectedness Index* are scaled within each dataset to have a maximum value of 1,000,000,000 and a minimum value of 1. As a result, the public release version of the $Social\ Connectedness\ Index_{i,j}$ measures the **relative probability of a Facebook friendship** link between a given Facebook user in location *i* and a given user in location *j*. Put differently, if this measure is twice as large, a given Facebook user in location *i* is about twice as likely to be connected with a given Facebook user in location *j*. We also add a small amount of random noise and round the *Social Connectedness Index* to the nearest integer, to ensure that no single individual or friendship link can be identified from the data.

This measure was first proposed and analyzed in Bailey, Cao, Kuchler, Stroebel, and Wong (2018), which explored the social connectedness across U.S. counties.

Data Included For Distribution

This folder includes the *Social Connectedness Index* calculated for different geographical areas as of October 2021. Each dataset includes every (symmetric) *i* to *j* and *j* to *i* location pair, including links of each location to itself.

Each dataset has three columns:

user_loc	First Location
fr_loc	Second Location
scaled_sci	Scaled SCI as described above

[1] When geographies *i* and *j* are the same, we make a small adjustment to the index. Instead of dividing through in the denominator by $FB\_Users_i * FB\_Users_i$, we divide through by $FB\_Users_i * (FB_{Users_i} - 1)$, since users cannot be friends with themselves in our data.

The datasets included contain the *Social Connectedness Index* for the following areas:

- **Countries – Countries.** Each row is a country – country pair. Countries are denoted by their ISO2 codes. Excludes certain countries, for example countries where Facebook is banned or countries with few active users.

  This dataset was first introduced, described, and analyzed in Bailey, Gupta, Hillenbrand, Kuchler, Richmond and Stroebel (2021).

- **US Counties – US Counties.** Each row is a US county – US county pair. Counties are denoted by their 5-digit FIPS code. Excludes counties with few active users.

  This dataset was first introduced, described, and analyzed in Bailey, Cao, Kuchler, Stroebel, and Wong (2018).

- **US ZCTA to US ZCTA.** Each row is a US ZCTA (zip code tabulation) – US ZCTA pair. Excludes ZCTAs with few active users. Due to the large size of this data set, it is split across ten files. The files are split according to the first digit of the zip code for location *i*. For instance, all pairs in which the first zip code begins with the digit 7 will be contained in file 7.

  This dataset was first introduced, described, and analyzed in Bailey, Farrell, Kuchler, and Stroebel (2020).

- **US Counties - Countries.** Each row is a US county – country pair. Counties are denoted by their 5-digit FIPS code, countries are denoted by ISO2 code. Excludes counties and countries.

  This dataset was first introduced, described, and analyzed in Bailey, Cao, Kuchler, Stroebel, and Wong (2018).

- **GADM/NUTS – GADM/NUTS.** There are two files built on the Database of Global Administrative Areas (GADM, version 2.8) and the European Nomenclature of Territorial Units for Statistics (NUTS 2016) areas. Excludes regions with few active users.

  - **GADM1_NUTS2**: European countries are divided into their NUTS2 regions (e.g., 12 provinces in the Netherlands). Countries outside of Europe are divided into their GADM level 1 regions (e.g., states in the US). Countries with a population less than 1 million are not divided. Each row is a pair of these areas.
  - **GADM1_NUTS3_Counties**: European countries are divided into their NUTS3 regions (e.g., 40 regions in the Netherlands). The United States, Canada, and some countries in South Asia (Bangladesh, India, Nepal, Pakistan, and Sri Lanka) are divided into their GADM level 2 regions (e.g., US counties). Other countries are usually divided into their GADM level 1 region. Countries with a population less than 1 million are not divided. Each row is a pair of these areas.

  A separate set of files (gadm1_nuts2_levels and gadm1_nuts3_counties_levels) provide the levels of each of the keys in the GADM/NUTS files.

  Shape files for NUTS-level data and GADM-level data can be downloaded from:

  https://gadm.org/old_versions.html

  https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units

  The GADM/NUTS-level dataset was first introduced, described, and analyzed in Bailey, Kuchler, Johnston, Russel, State, and Stroebel (2020).

Other Resources

Some of our academic partners on this project have put together a set of resources for individuals interested in working with these files, including code and shapefiles that easily map the social connectedness of different locations, and code that explores the relationship between the *Social Connectedness Index* and the spread of COVID-19 (see Kuchler, Russel, and Stroebel, 2020). These files can be found at: https://github.com/social-connectedness-index/example-scripts

References + BibTex

- Bailey, M., Cao, R., Kuchler, T., Stroebel, J., & Wong, A. (2018). Social Connectedness: Measurement, Determinants, and Effects. *Journal of Economic Perspectives*, 32(3), 259-80.

```
@article{bailey2018social,
  title={Social connectedness: Measurement, determinants, and effects},
  author={Bailey, Michael and Cao, Rachel and Kuchler, Theresa and Stroebel, Johannes and Wong, Arlene},
  journal={Journal of Economic Perspectives},
  volume={32},
  number={3},
  pages={259--80},
  year={2018}
      }
```

- Bailey, M., Farrell, P., Kuchler, T., & Stroebel, J. (2020). Social Connectedness in Urban Areas. *Journal of Urban Economics*, 118, 103264

```
@article{bailey2020urban,
  title={Social connectedness in urban areas},
  author={Bailey, Michael and Farrell, Patrick and Kuchler, Theresa and Stroebel, Johannes},
  year={2020},
  journal={Journal of Urban Economics},
  volume={118},
  pages={103264},
  year={2020}
}
```

- Bailey, M., Gupta, A., Hillenbrand, S., Kuchler, T., Richmond, R. J., & Stroebel, J. (2021). International Trade and Social Connectedness. *Journal of International Economics*, 129, 103418

```
@article{bailey2021international,
  title={International trade and social connectedness},
  author={Bailey, Michael and Gupta, Abhinav and Hillenbrand, Sebastian and Kuchler, Theresa and Richmond,
Robert and Stroebel, Johannes},
  journal={Journal of International Economics},
  volume={129},
  pages={103418},
  year={2021},
  publisher={Elsevier}
}
```

- Bailey, M., Kuchler, T., Johnston, D., Russel, D., State, B., & Stroebel, J. (2020). The Determinants of Social Connectedness in Europe. *SocInfo2020*

```
@InProceedings{bailey2020determinants,
  title={The Determinants of Social Connectedness in Europe},
  author={Bailey, Michael and Johnston, Drew and Kuchler, Theresa and Russel, Dominic and State, Bogdan and
Stroebel, Johannes},
  booktitle={Social Informatics},
  publisher={Springer International Publishing},
  year={2020},
}
```

- Kuchler, T., Russel, D., & Stroebel, J. (2021). The Geographic Spread of COVID-19 Correlates with the Structure of Social Networks as Measured by Facebook, *Journal of Urban Economics*, 103314

```
@article{kuchler2021geographic,
  title = {The geographic spread of COVID-19 correlates with the structure of social networks as measured
  by Facebook},
  journal = {Journal of Urban Economics},
  pages = {103314},
  year = {2021}
}
```

Frequently Asked Questions

Is the SCI data available for other time periods?

At this stage, the SCI data is not available for other time periods beyond the data listed on HDX. However, we believe that the underlying object captured by the SCI—the geographic structure of networks between regions—is highly stable over time. As a result, social connectedness as measured today is likely to predict interactions over other time horizons. To give two examples.

(1) Bailey, Gupta, Hillenbrand, Kuchler, Richmond and Stroebel (2021) show that social connectedness as measured today predicts trade flows in the 1980s as well as it predicts trade flows today.

(2) Kuchler, Li, Peng, Stroebel and Zhou (2021) show that social connectedness as measured today predicts mutual fund investments in the 2000s equally well as it predicts mutual fund investments today.

More generally, we believe that when studying any observed time-series variation in the SCI—in particular over periods of only a few years—the signal of true changes in connectedness across geographies would be extremely hard to disentangle from changes in Facebook penetration across the two regions over time.

Has the definition of the SCI changed over time?

In the first related academic research paper, Bailey, Cao, Kuchler, Stroebel, and Wong (2018) defined the SCI to be a scaled version of $FB\_Connections_{i,j}$, while it defined a second measure, "relative probability of friendship," as a scaled version of $FB\_Connections_{i,j}/(FB\_Users_i * FB\_Users_j)$.

Since the release of that paper, we found this second object to have much wider usefulness. In the public data release, we therefore decided to only release the second object, and call this object the "*Social Connectedness Index*." We apologize if this causes some confusion.

Can the SCI be aggregated to higher levels of geographic aggregation?

Yes it can, using the formula introduced in equation 4 in Bailey, Gupta, Hillenbrand, Kuchler, Richmond and Stroebel (2021). We reproduce the relevant part below:

In Section 2, we related the volume of exports from country i to country$_j$ to the probability that a representative Facebook user in country i is friends with a representative Facebook user in country$_j$, given by $SCI_{i,j}$. This measure of social connectedness is identical to a population-weighted average of the social connectedness across the regions in countries i and j. Formally, let us index the regions in each country i by $r_i \in R(i)$, let $Friendships_{r_i,r_j}$ count the total number of Facebook friendship links between individuals in regions r_i and r_j, let Pop_{r_i} denote the total (Facebook) population in region r_i, and let $PopShare_{r_i}$ denote the share of that population in region r_i in country i: $\sum_{r_i \in R(i)} PopShare_{r_i} = 1$. Then:

$$
\begin{aligned}
SCI_{i,j} = \frac{Friendships_{i,j}}{Pop_i \times Pop_j} &= \frac{\sum_{r_i \in R(i)} \sum_{r_j \in R(j)} Friendships_{r_i,r_j}}{\left(\sum_{r_i \in R(i)} Pop_{r_i}\right) \times \left(\sum_{r_j \in R(i)} Pop_{r_j}\right)} \\
&= \sum_{r_i \in R(i)} \sum_{r_j \in R(j)} \frac{Pop_{r_i}}{\sum_{r_i \in R(i)} Pop_{r_i}} \frac{Pop_{r_j}}{\sum_{r_j \in R(j)} Pop_{r_j}} \frac{Friendships_{r_i,r_j}}{Pop_{r_i} \times Pop_{r_j}} \\
&= \sum_{r_i \in R(i)} \sum_{r_j \in R(j)} PopShare_{r_i} \times PopShare_{r_j} \times SCI_{r_i,r_j}
\end{aligned}
\tag{4}
$$