Description of data (DOI: 10.17634/141304-9)

This document describes the raw data of share links and clicks, and several supplementary lookup tables, from the App Movement platform[1]. These data can be converted to data sets fitted by the network epidemic model in Lee, Garbett & Wilkinson (2017).

1 Share links

The file `<YYYYMMDD>_links.csv` contains the following columns:

id (integer) The identifier of the share link. It does not range from 1 to the total number of share links because some invalid links were removed

share_link_type_id (integer) The type of the share link, which is as follows:

1. movement
2. venue
3. review
4. app

user_id (integer) The identifier of the user.

site_user_id (integer) The identifier of the device associated with the user.

parent_id (integer) The identifier of the type associated with the share link. For example, a record with `share_link_type_id = 1` and `parent_id = 12` means the share link is for movement 12.

created (datetime) Time of the share link being created.

[1] `https://app-movement.com`

2 Share link clicks

The file `<YYYYMMDD>_clicks.csv` contains the following columns:

id (integer) The identifier of the click.

share_link_id (integer) The identifier of the share link clicked, that is, the column `id` in the table described in Section 1

user_id (integer) The identifier of the user.

site_user_id (integer) The identifier of the device associated with the user.

created (datetime) Time of the click being made.

3 Lookup tables for user_id and site_user_id

There have been users creating multiple accounts, that is, different **user_id**'s sharing the same **site_user_id**, likely for increasing the support for a movement. In a similar fashion, there have been users using multiple devices, that is, different **site_user_id**'s sharing the same **user_id**. Efforts have been made to detect such behaviour, resulting in the following tables, which can be used to clean the links and clicks tables described above.

(a) The file `<YYYYMMDD>_bogus_two_ids.csv` contains all combinations of the two columns **user_id** and **site_user_id**, the definitions of which are the same as those in the links and clicks tables.

(b) The file `<YYYYMMDD>_bogus_user_id.csv` contains the following columns:

cluster (integer) The identifier of the cluster of multiple **user_id**'s which belong to the same actual user.

value (integer) The **user_id**'s involved in each cluster.

user_id_interim (integer) The **user_id**, unique for each cluster, that can be used for cleaning.

(c) The file `<YYYYMMDD>_bogus_site_user_id.csv` contains the following columns:

cluster (integer) The identifier of the cluster of multiple **site_user_id**'s which belong to the same actual user.

value (integer) The **site_user_id**'s involved in each cluster.

user_id_interim (integer) The **user_id**, unique for each cluster, that can be used for cleaning.

(d) The file `<YYYYMMDD>_user_id_to_remove.csv` contains the **user_id**'s which are not actual users but test users created by the developers of App Movement.

4 Date stamp

The date stamp that prefixes the files described above is the end date of the data collection, *not* the date of creation of the files or of this document.

References

Lee, C., Garbett, A. & Wilkinson, D. J. (2017), 'A network epidemic model for online community commissioning data', *Statistics and Computing* .
URL: *https://doi.org/10.1007/s11222-017-9770-6*