README for the Reproducibility Package for Diffusion in social networks: experimental evidence on information sharing vs persuasion

Overview

The folder contains 6 subfolders:
- "0. Label raw data": it contains the raw data files and the do-files used for initial labeling and preparation. The output consists of three datasets used as inputs in subsequent folders.
- "1 Data cleaning (pre-treatment)": it contains the files to produce the baseline dataset. The output produced are contained in the folder "1. Cleaned data".
- "2. Randomization": it contains the codes used to randomize the treatment into the treatment arms. The output produced is the dataset "village_seed7557.dta".
- "3. Matching A_B": it contains the codes to match every 1 primary participants to 3 secondary participants within each village. The output produced is the dataset "matching_A_B.dta".
- "4. Merge": it contains the codes to merge the baseline information with randomization and matching_A_B information. The output produced is the dataset "baseline_dataset.dta".
- "5. Data cleaning (post-treatment)": it contains the files to produce the endline dataset. The output produced are contained in the folder "1. Cleaned data".
- "Replication": it contains the datasets and codes to replicate the results in the paper.

Description of the raw datasets and initial data cleaning

Each dataset was de-identified by the researcher in charge of the field experiment at the time of download from the Kobo platform. Below, we provide details on the labelling/cleaning steps applied for each dataset:

Baseline dataset: "FI_Baseline_Survey_Kanpur_Data.xls". This is one unique dataset for the baseline survey. The data was de-identified and Kobo's metadata and system-generated fields, and enumerator identity information were removed at the time of the data was transferred from the Kobo database. The variable names have been labelled in Hindi.

Village dataset: "FI_Village_Survey_Dataset.xls". This is one unique dataset covering information on villages. The data was de-identified and Kobo's metadata and system-generated fields were removed at the time of the data was transferred from the Kobo database. The variable names have been manually labelled in English. Some value labels (specifically for the variables "C9_1st", "C9_2nd", "C9_3rd") have been entered manually in English.

Endline datasets: "FI_Endline_Survey_Control_Data.xls". There are five datasets for the endline survey, one of each treatment arm. The data was de-identified and Kobo's metadata, system-generated fields, and enumerator identity information were removed at the time of the data was transferred from the Kobo database. The variable names have been labelled in Hindi.

No further labeling or corrections were applied directly to these raw files; instead, subsequent labeling and construction of analysis variables were performed through Stata do-files included in the package.

Data Availability

Source	Data file	Link	Year	Access
Authors' data collection	FI_Baseline_Survey_Kanpur_Data.xlsx	-	2025	Open
Authors' data collection	FI_Endline_Survey_C_Data.xlsx	-	2025	Open
Authors' data collection	FI_Endline_Survey_Control_Data.xlsx	-	2025	Open
Authors' data collection	FI_Endline_Survey_IA_Data.xlsx	-	2025	Open
Authors' data collection	FI_Endline_Survey_IAP_Data.xlsx	-	2025	Open
Authors' data collection	FI_Endline_Survey_IP_Data.xlsx	-	2025	Open
Authors' data collection	FI_Village_Survey_Dataset.xlsx	-	2025	Open
Authors' data collection	date_signup_final.dta	-	2025	Open

Due to budget constraints, the data collection was conducted using the free version of KoboToolbox. This version does not store raw data on Kobo servers beyond manual download. The downloaded raw files were initially stored on a work computer owned by one of the coauthors' former employers. After the coauthor left that job, the computer was reclaimed and later wiped by the employer before being reassigned.

Although the original raw download files are no longer available, the coauthor retained all the necessary files for empirical analysis, including de-identified versions of the raw data and all intermediate and final datasets used in the paper. These are fully included in this replication package.

Statement about Rights

We certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.
We certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package. Appropriate permissions are documented in the LICENSE.txt file.

Instructions for Replicators

To reproduce the results in the paper, the replicator must open the folder "Replication", then:
1) Open the master dofile "main.do"
2) Insert their directory in lines 17
3) Run the code
The tables will be reproduced following the order in the paper.

Software Requirements

Stata

Memory, Runtime, and Storage Requirements

12 hours