Title: "Is OTT Video a Substitute for TV? Policy Insights from Cord-Cutting" Readme for data and code replication, 9/15/2021

Software: The data analysis was done using Stata 16, Matlab 2021a, and Python 3. All commands used are either a part of the standard packages or can be installed/loaded using code included in the scripts.

Data: The raw data used in our analysis were obtained from a North American MSO. Our NDA agreement with the MSO prohibits us from directly sharing these data, but we can provide guidance on how to obtain the same or similar data from any MSO.

To this end, the data come from a deep-packet inspection (DPI) platform. Within these platforms, there are a variety of mechanisms that can be used to create customized reports. For one of the most commonly-used DPI platforms, Sandvine, our data can be re-created through two customized reports. The first report would include hourly byte counts for a household (i.e., MAC) for each category of traffic (e.g., Sandvine currently has 11): bulk, email, gaming, miscellaneous, cloud, peer-2-peer, real-time communication, real-time entertainment, social media, tunnel, and web. The second report is an event-level disaggregation of real-time entertainment traffic for each household (i.e., MAC). This report would include a record for each streaming event within a household that tracks duration in seconds, content source (e.g., Netflix), and total bytes transmitted. Certain implementations of the DPI platform can now also report information on the device used to stream the content (e.g., Roku). These two reports can then be merged via MAC-time stamp pairs to create a detailed time series of overall usage of a household, and the breakdown of streaming activities by source at the desired frequency (hour or higher). If these usage data are then complemented by subscription/billing data for each household, an analysis similar to ours can be replicated. To further assist authors interested in acquiring similar data to replicate our analysis, we've included headers from a current implementation of Sandvine for the two reports described above. The two .csv files include file headers with variable names, and immediately below a description of each variable.

Files:

- **01_prep_data.py, 02_aggregate_data.py, 03_id_cord_cutters.py** – Processes the raw data files from the MSO, saves datasets at various levels of aggregation for subsequent analysis, and identifies cord-cutters.
- **04_paper_figures.py, 05_traffic_comp_figure.py, 06_leadin_figures.py** – Generates the following list of figures from the paper using prepared data files.
    - Figure 1
    - Figure 2
    - Figure A1
    - Figure A2
    - Figure A3
    - Figure A4
    - Figure A5
    - Figure A6
    - Figure A7
- **analysis0.do** – Aggregates data file created in 03_id_cord_cutters.py to the household/customer level.
- **analysis1.do** – Contains code to produce all regression and summary statistics tables available in the published paper. Uses data file produced in analysis0.do. Also

contains a description of variables contained in the dataset. Generates the following tables:
    - Table 1
    - Table 2
    - Table 3
    - Table A1
    - Table A2
- **ccrate_3d_ll.m** – generates Figure A2.
- **tmpR.csv** and **tmpB.csv –** Example reports described above.