Demographics and Dynamics of Mechanical Turk* Workers

Djellel Difallah
djellel@nyu.edu
New York University

Elena Filatova
efilatova@citytech.cuny.edu
City University of New York

Panos Ipeirotis
panos@nyu.edu
New York University

* In our work we analyze the Mechanical Turk (MTruk) crowdsourcing platform. However, the conclusions and experiments are crowdsoucing platform independent.

Why Studying MTurk Population?

Crowdsourcing platforms are a source of human subjects for various tasks (i.e., data collection, surveys). Thus, it is important to know:
- whether MTurk has enough workers who can potentially participate in the experiments;
- whether the set of MTurk workers is diverse;
- whether for the experiments search for participants who have a specific characteristic (e.g., skill, location, etc.), ensure that there are enough workers with this specific characteristic who can potentially participate in the experiment;
- what is the dynamics of the MTurk population.

Data Collection Experiment

Our MTurk task is to answer a short survey: (a) Gender, (b) Year of Birth, (c) Marital Status, (d) Household Size, (e) Household Income, (f) Location (City, Country) §
- One survey is posted every 15 minutes;
- Survey can be completed by a single MTurk worker and takes on average 30 seconds;
- Each MTurk worker can participate in our survey once every 30 days;
- For every submitted survey response, we pay 5 US cents.
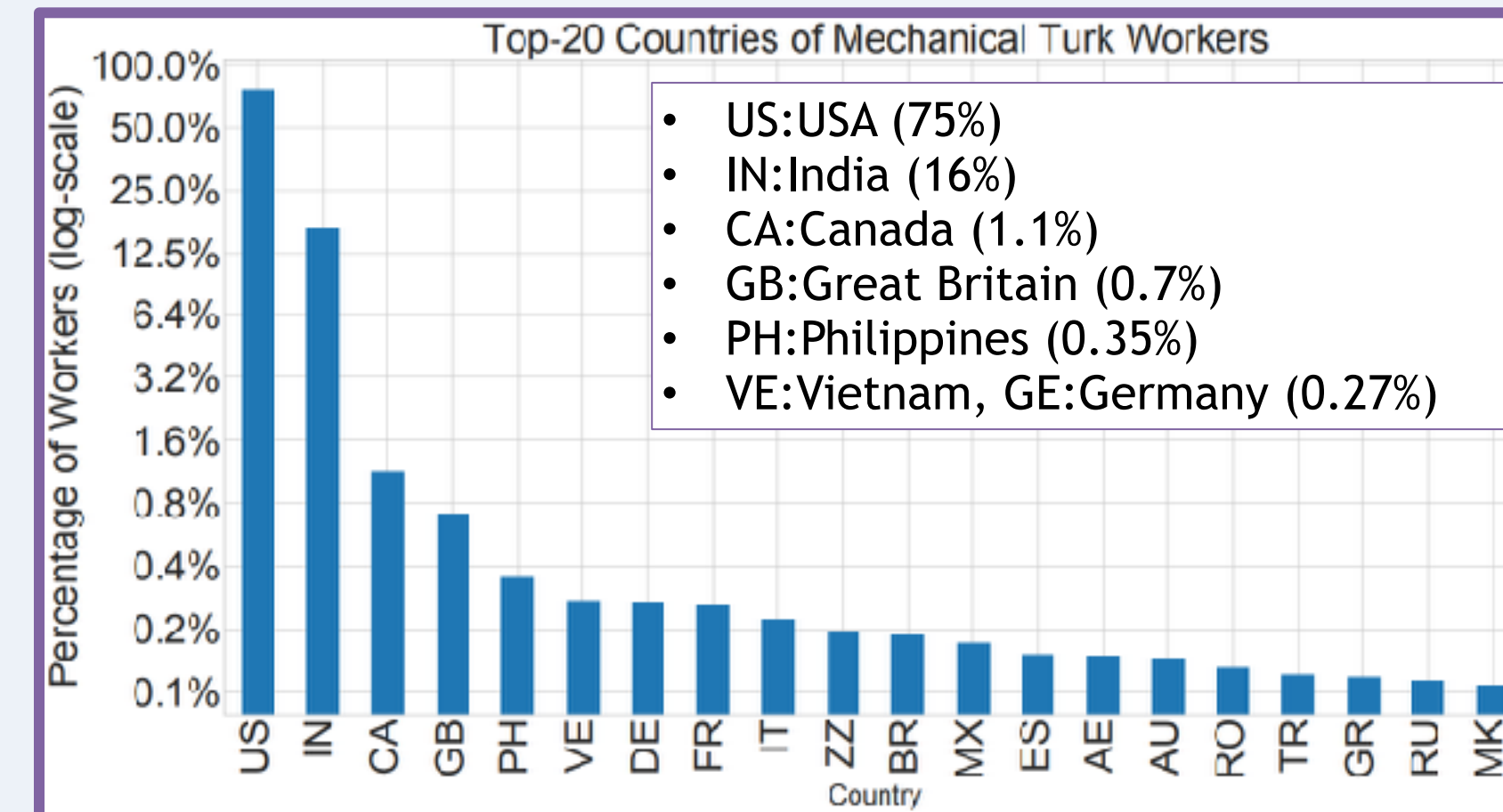
Collected data (as of July 31, 20):
- 84,511 responses (between March 26, 2015 and July 31, 2017);
- Submitted by 39,461 unique workers.

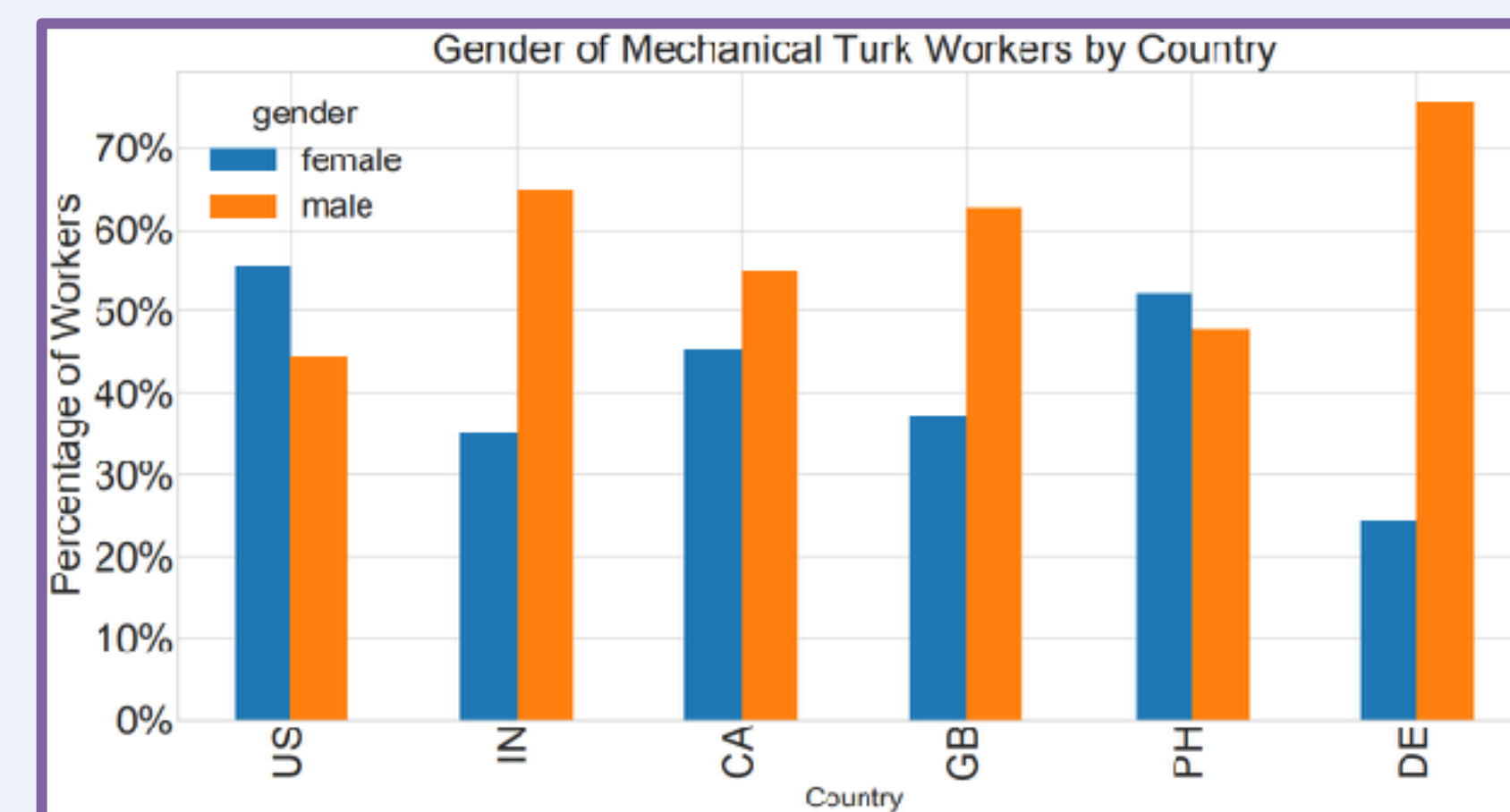Our data collection experiment is still running

§ We also use geolocation tools to verify the location of each participating worker.
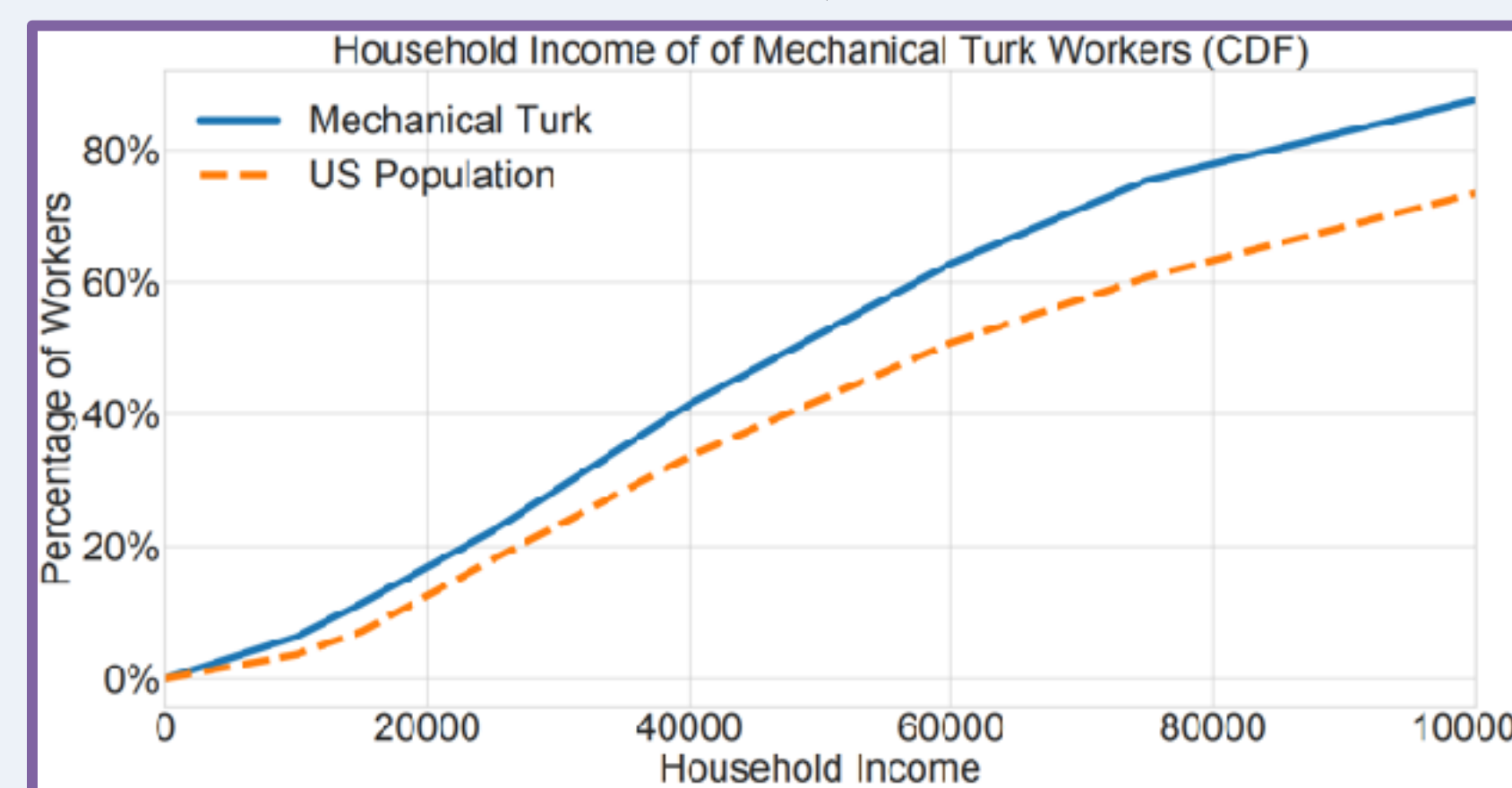
MTurk Demographics Overview

Top-20 countries of origin for MTurk workers:

- US:USA (75%)
- IN:India (16%)
- CA:Canada (1.1%)
- GB:Great Britain (0.7%)
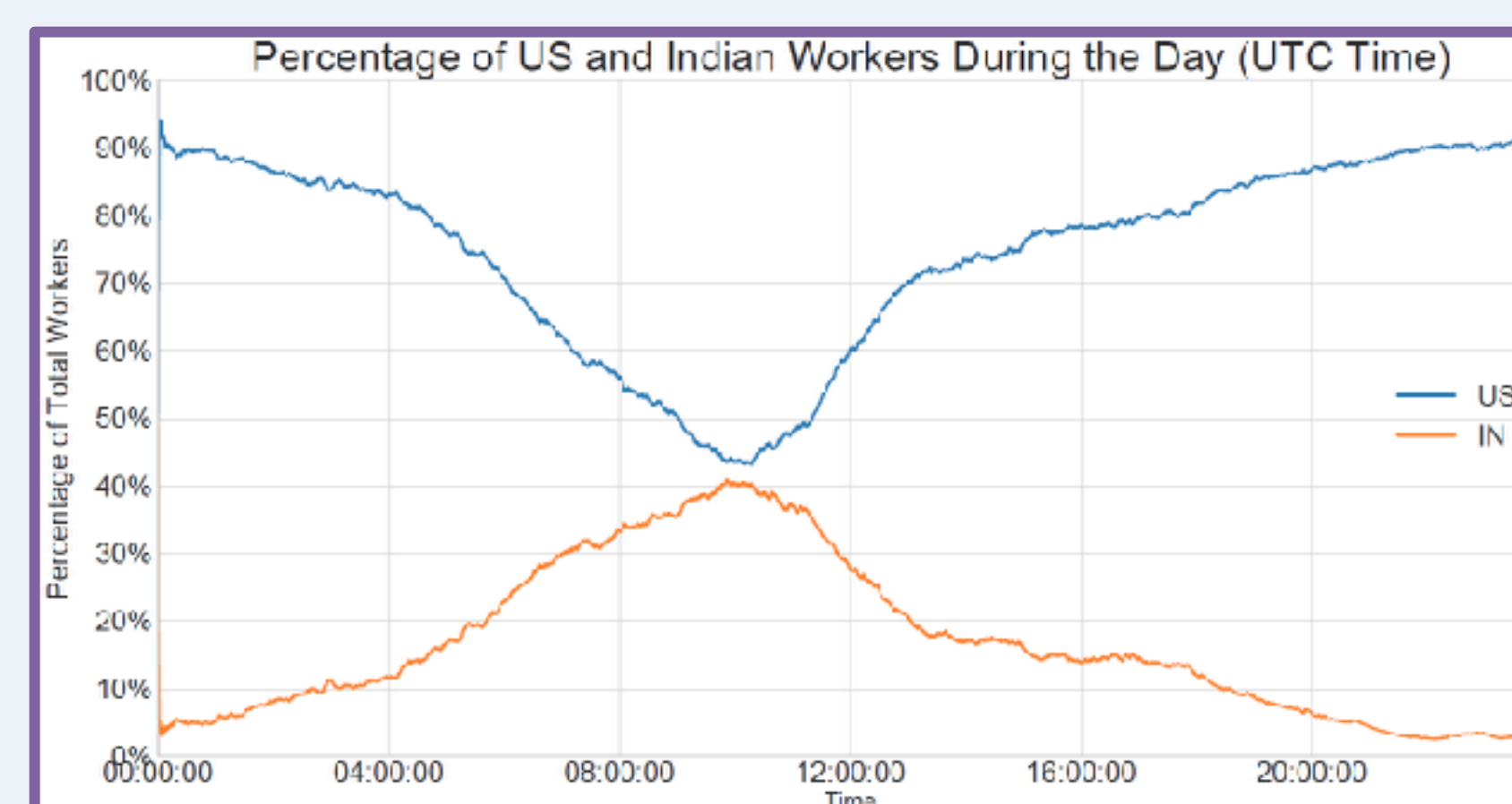- PH:Philippines (0.35%)
- VE:Vietnam, GE:Germany (0.27%)

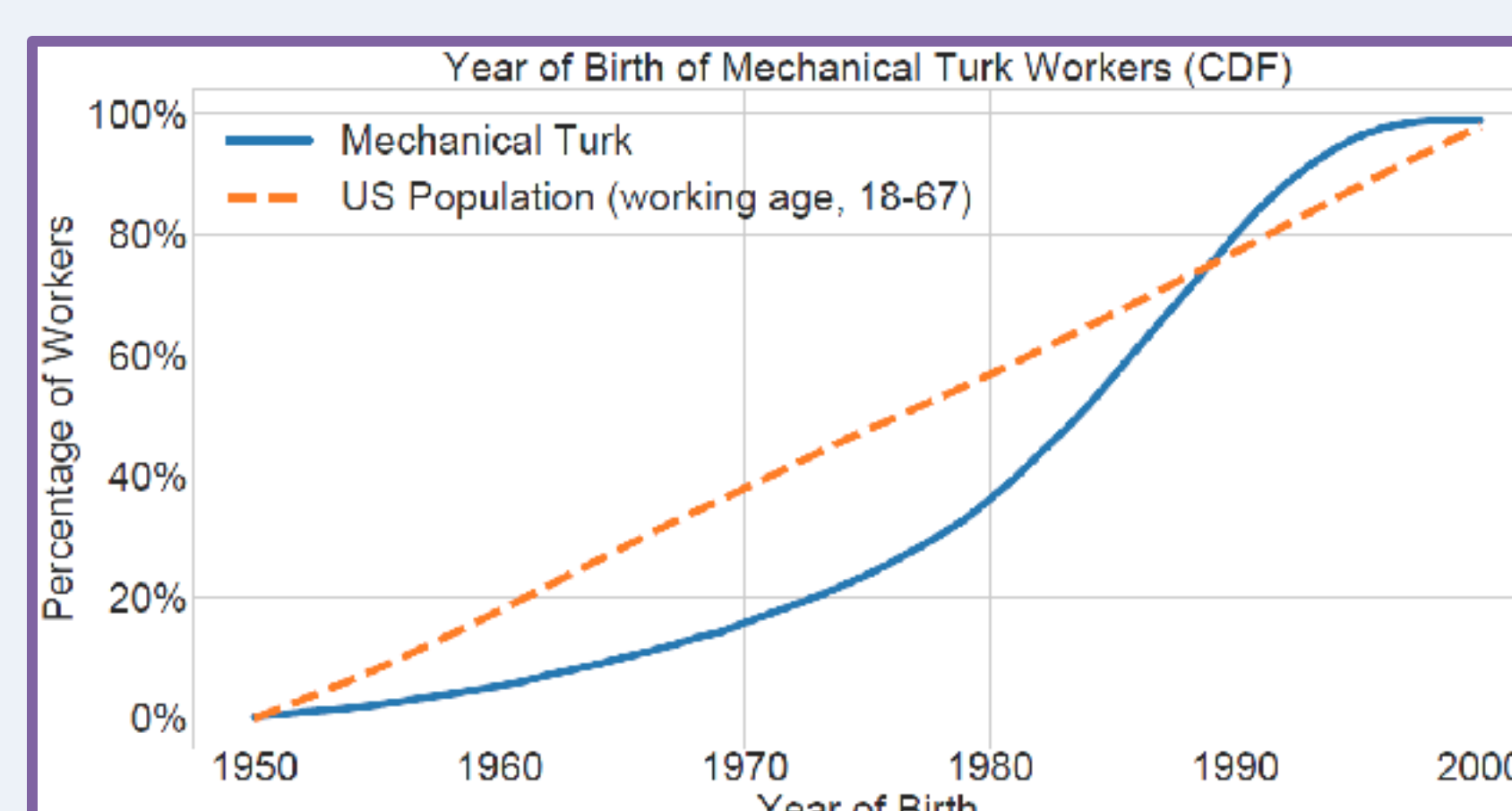Gender distribution (across 6 top countries)

Household Income in $ (US vs. General MTurk)
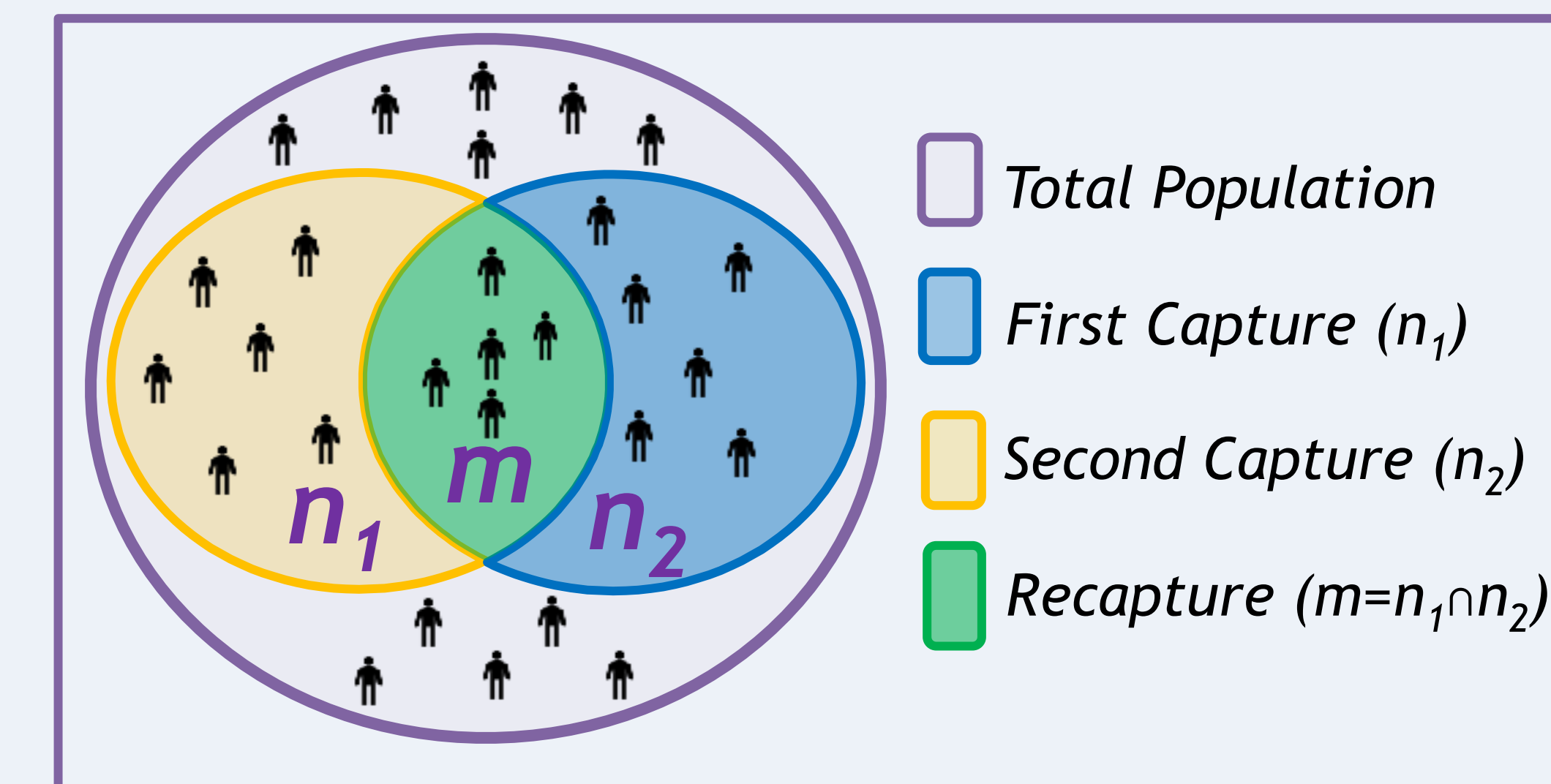

US vs. Indian Workers During the Day (UTC Time)

Year of Birth of MTurk Workers

MTurk Demographics Dynamics

Capture-Recapture model to approximate the population size ecology modelling.

- Total Population
- First Capture (n_1)
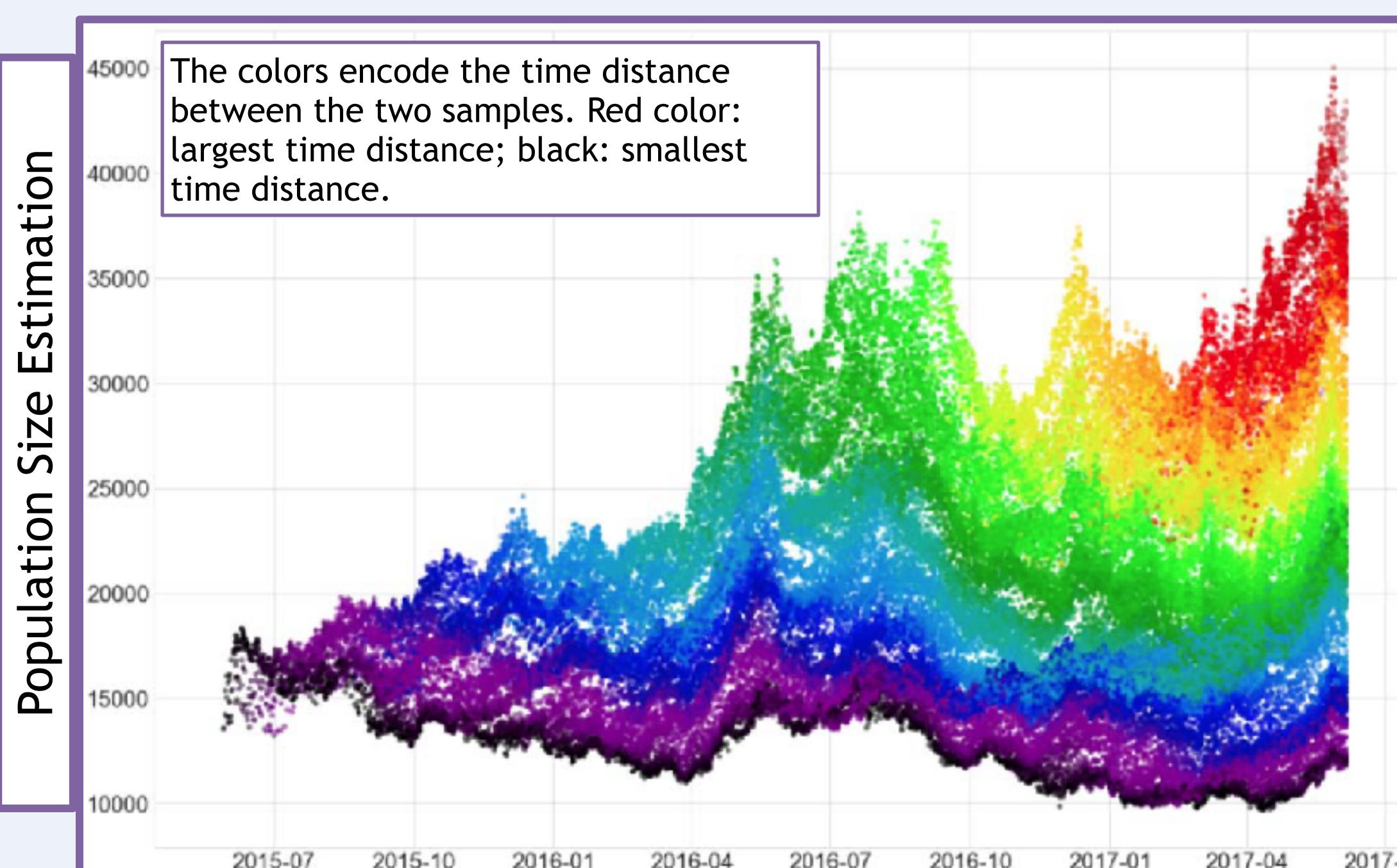- Second Capture (n_2)
- Recapture ($m=n_1 \cap n_2$)

In our scenario, we analyze 30-day periods:
- the capture phase is a 30-day period, we consider "captured" the n_1 workers that participate in the survey during that period;
- the recapture phase is a different 30-day period, out of the n_2 workers "captured" during that period, we consider as "recaptured" the m workers that participate in both surveys.

Assumptions for simple model
(1) Closed population: No MTurk worker joins or leaves the platform.
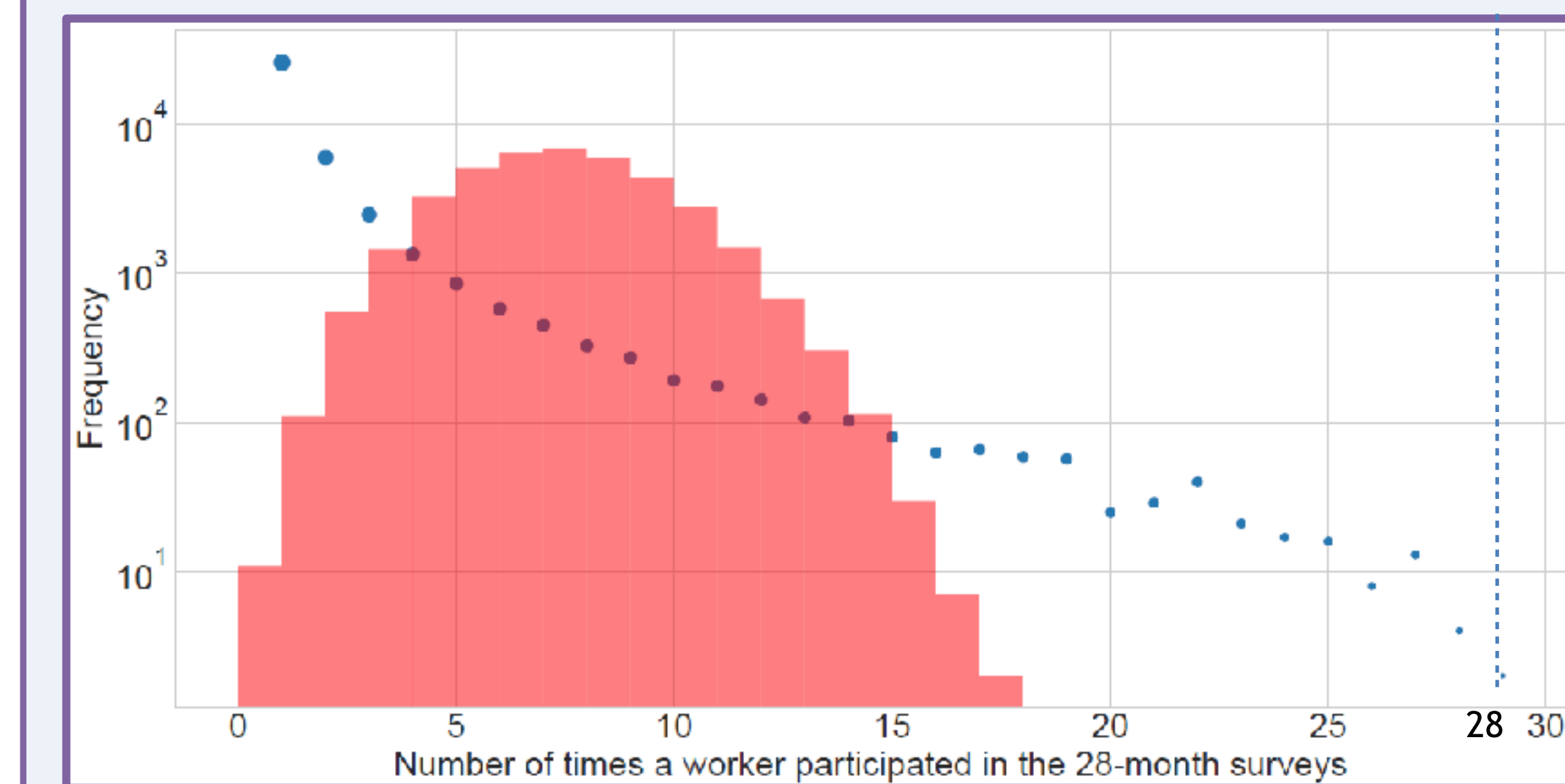(2) Equal catchability: The probability of participating in the survey is equal across all MTurk workers.

Estimation of MTurk Worker Population Size
(function of the start time of the recapture period)

The colors encode the time distance between the two samples. Red color: largest time distance; black: smallest time distance.

Start Date of Recapture Period

Estimated life time of MTurk worker ≈ 12 – 18 month

MTurk Demographics Dynamics

Propensity of Participation

If we assume equal catchability, the distribution of frequency of participation should follow a binomial distribution (p=0.25, n=28). The observed frequency of participation has different behavior.

Hypothesis: The workers have different "propensities" of participation in our survey

Modeling Propensity with Multiple Captures

- Propensity: Our model assumes that the propensity of worker *i* to *participate in our survey* follows a beta distribution B(α, β)
- Super-population: Given an overall set of *N* workers we sample from the population *n* times (28 in our case)
- Probability of seeing a worker *K* times follows a Beta-Binomial Distribution:

$$f(k|n, \alpha, \beta) = \binom{n}{k} \frac{B(k + \alpha, n - k + \beta)}{B(\alpha, \beta)}$$

- We observe S sampled workers (k≥1)

$$S = N \cdot \sum_{k=1}^{n} f(k|n, \alpha, \beta) = N\big(1 - f(0|n, \alpha, \beta)\big)$$

- S= 39,461 distinct workers (k≥1)
- Using MLE estimation: α = 0.29, β= 20.9
- Number of sampling times n = 28
- Prob. of not capturing: f(0|n,α,β) = 0.7793

Estimated size of MTurk (with propensities)

$$N = \frac{S}{1 - f(0|n, \alpha, \beta)} = \frac{39461}{1 - 0.7793} \approx 178{,}800 \text{ workers}$$