

# Estimation and Sampling with Longitudinal Person-Firm Data

Kevin McKinney  
FCSM 2012

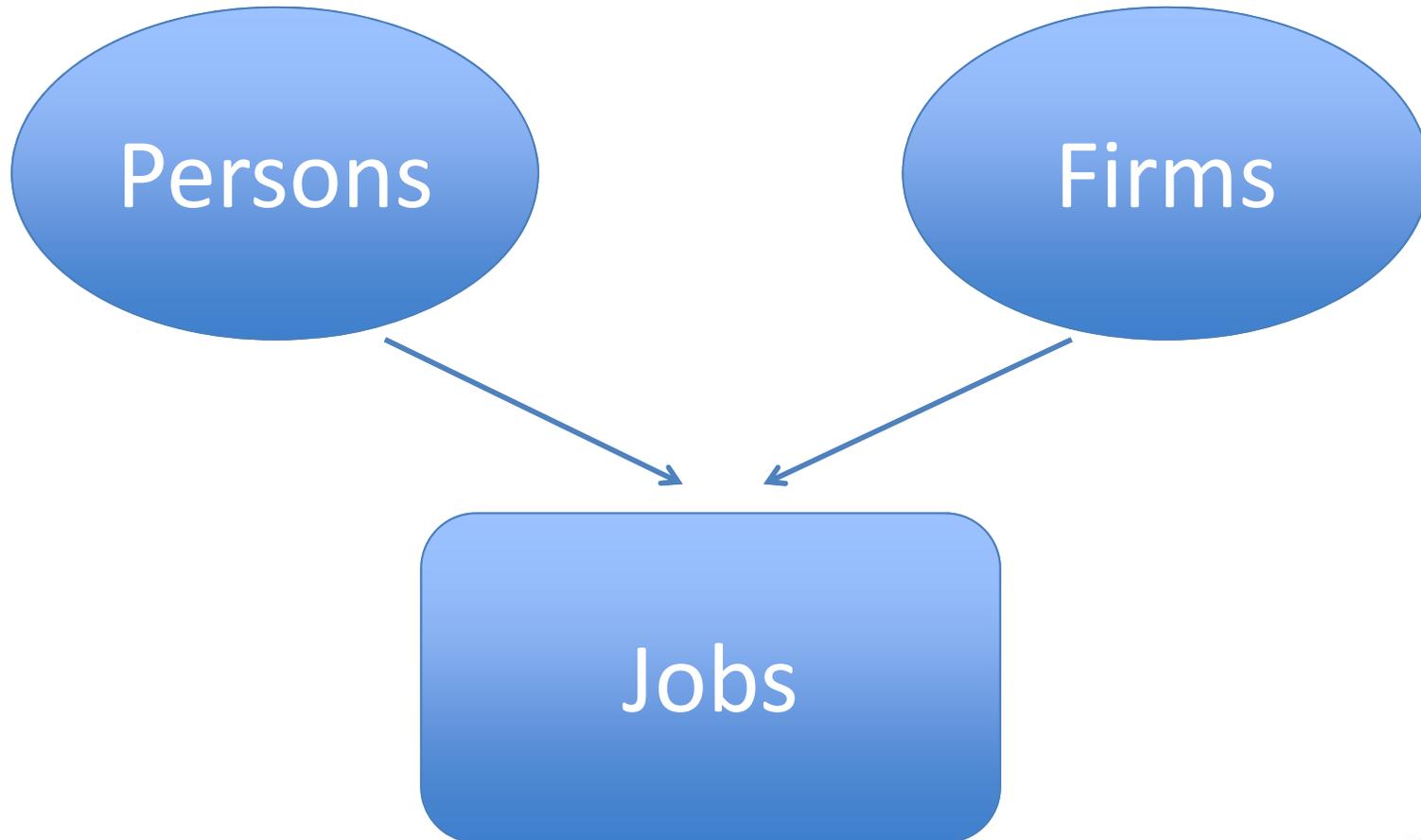
# Overview

- Person-Firm Longitudinal Data
- Simple Random Sampling
- Person Firm Graph
- New Connected Sampling Method
- Results
- Conclusion

# Longitudinal Person-Firm Data

- Demographic and economic data typically contains only one population.
- Multiple populations, if they exist, are nested (i.e. persons within households and establishments within firms).
- Longitudinal person firm data is 3 datasets in one; persons, firms, and jobs.
- Persons, firms and jobs are not nested.
  - Persons can work for more than one firm.
  - A job is by definition associated with both a person and a firm.

# Three Populations in One Dataset



# Simple Random Sampling

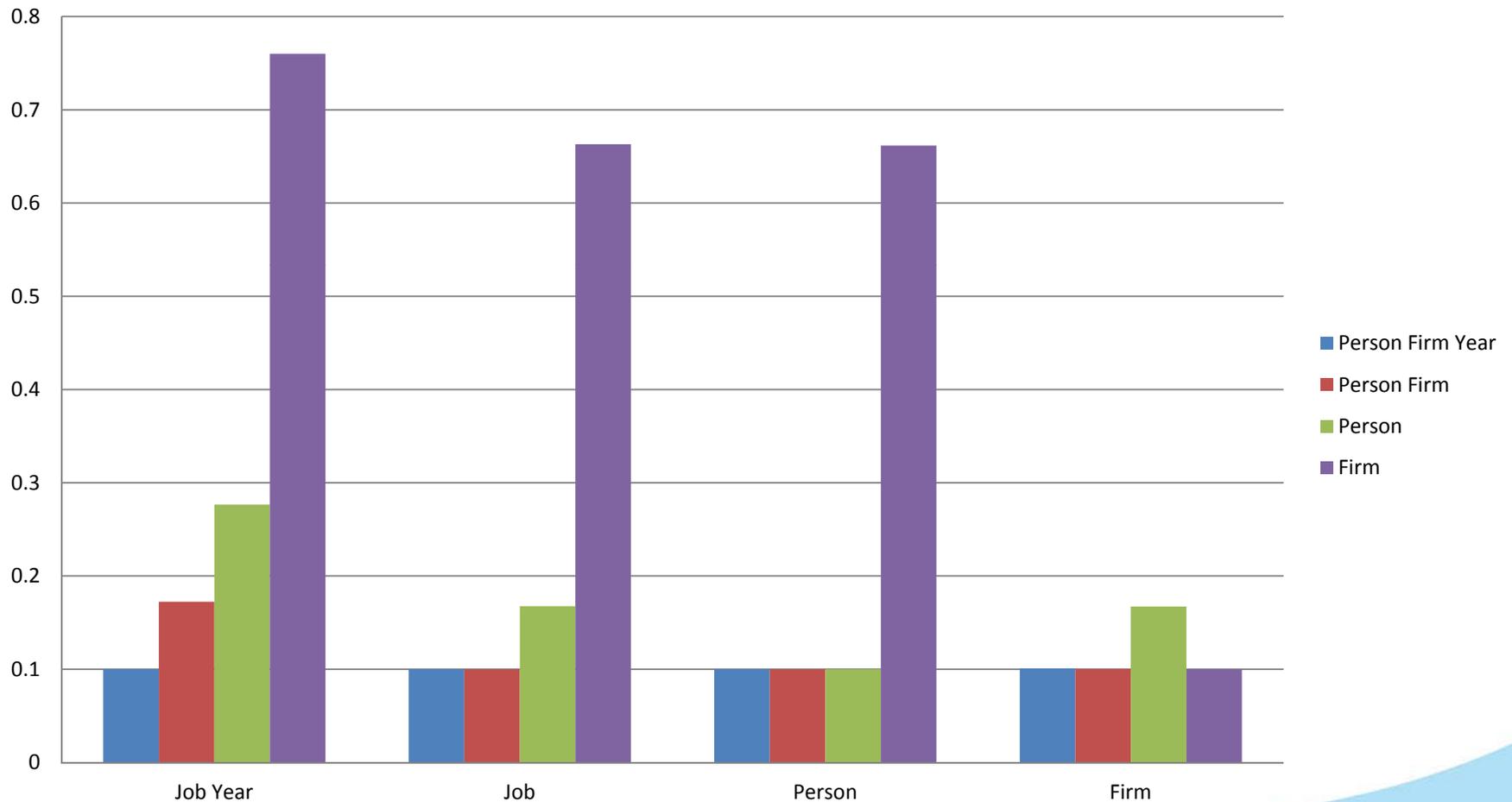
- Sampling longitudinal person-firm data presents some challenges
  - Multiple populations (person, firm, job)
  - Populations observed over time (person, firm, and job histories)
    - Should the history of a sampled unit be preserved?
  - User community employs a wide array of possible estimators (simple statistics such as means as well as model based methods such as linear regression)

# Four Candidate SRS Methods

- Sample randomly from each list frame without replacement.
- Job Year
  - Most basic unit of observation in the data. Select a sample of worker earnings records.
- Job
  - Select a sample of jobs, include all annual earning observations for the selected jobs.
- Person
  - Select a sample of persons and include the complete job history for each selected person.
- Firm
  - Select a sample of firms and include the complete employment and payroll history for the workers at each selected firm.

# Percent of Population(s) Selected

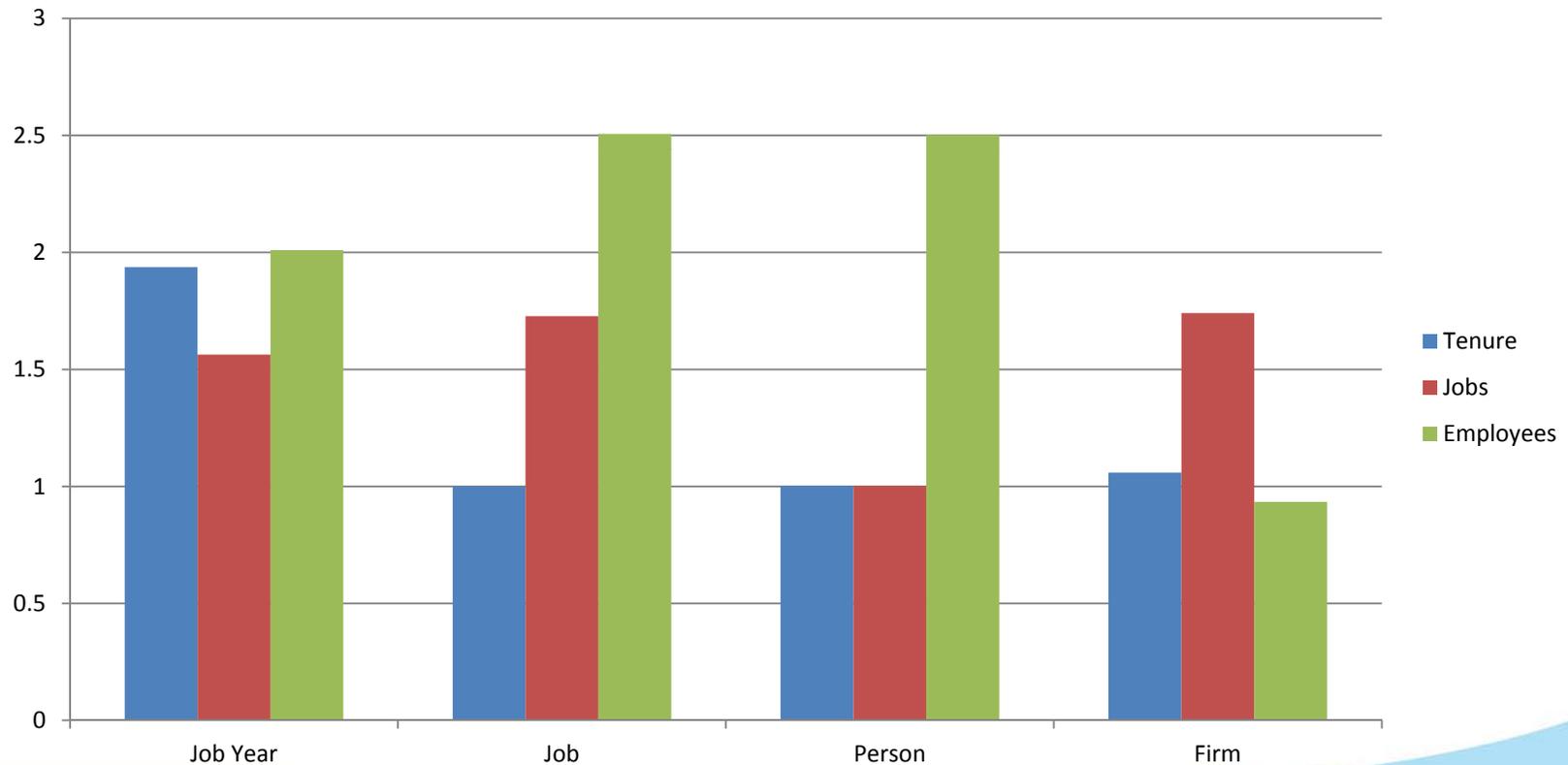
(10% Sampling Rate)



# Population Oversampling

- No method produces a representative sample for all three populations (person, firm, and job).
- Job Year
  - Oversamples jobs, persons, and firms.
- Job
  - Oversamples persons and firms.
- Person
  - Oversamples firms.
- Firm
  - Oversamples persons.

# Population Characteristics of Sampled Units (2% Sampling Rate)



# SRS Summary

- A self-weighting or representative sample for each population would be ideal, but each SRS method produces a biased sample for at least one of the three populations. Magnitude of the effect will depend on the sampling rate and the correlation of tenure (jobs), number of employers (persons), and the number of employees (firms) with the sample characteristic(s) of interest.
- Weights are a useful tool when estimating univariate statistics such as means and standard errors.
- The application of weights to models with multiple populations is often not straightforward.
  - Which weight do you use?
  - Many advanced methods are not designed to use weights.
- No method is a silver bullet, choose the sampling method best suited to your analysis.

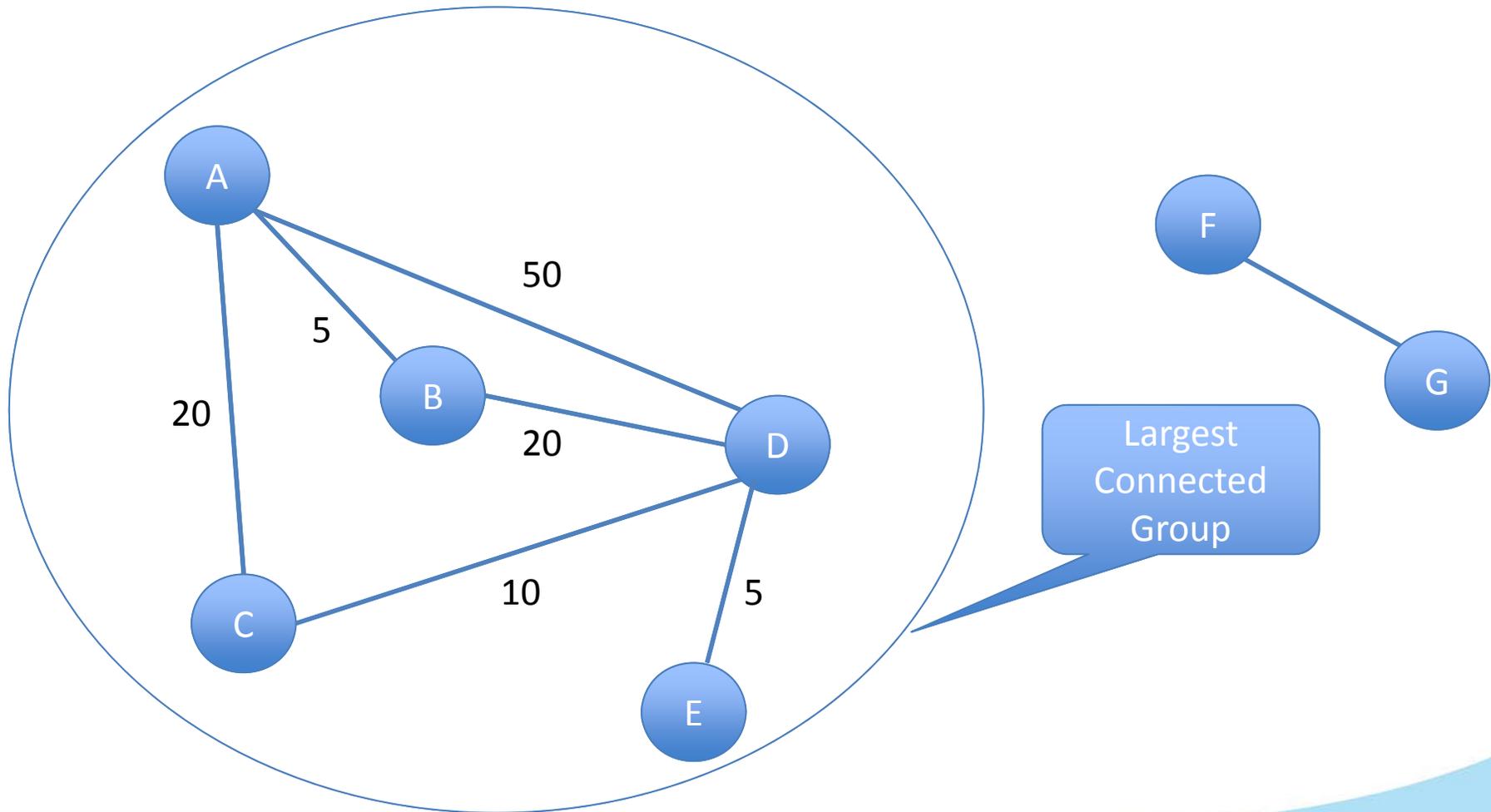
# The Firm Graph

- The firm graph shows the relationships between firms created by employees that change and/or have multiple employers.
- Traditional sampling methods are not designed to preserve the firm graph.
  - The big component.
- Goal: produce a representative connected sub-graph.
  - Enable the estimation of person-firm fixed effect earnings models.
  - While not required, other methods such as mixed effects should benefit from the improved connectivity.

# The Firm Graph

- Projection of person-firm bipartite graph onto the firm nodes.
- $G=(V,E)$
- $V=\text{Firms}$
- $E=\text{Edges}$ 
  - An edge is generated by the movement of a person from firm  $v_m$  to  $v_n$  . Self loops (where  $m=n$ ) add no value when sampling and are thus ignored when creating the graph.

# Firm Graph Example



# New Connected Sampling Method

- Use a modified or “balanced” random walk on the largest connected component of the firm graph.
  - Choose node  $v_0$  at random from  $V$
  - Find all adjacent nodes of  $v_0$
  - Select one of the adjacent nodes based on values in a transition probability matrix  $T$ .
  - Move to the selected node and repeat until you reach the desired sample size.

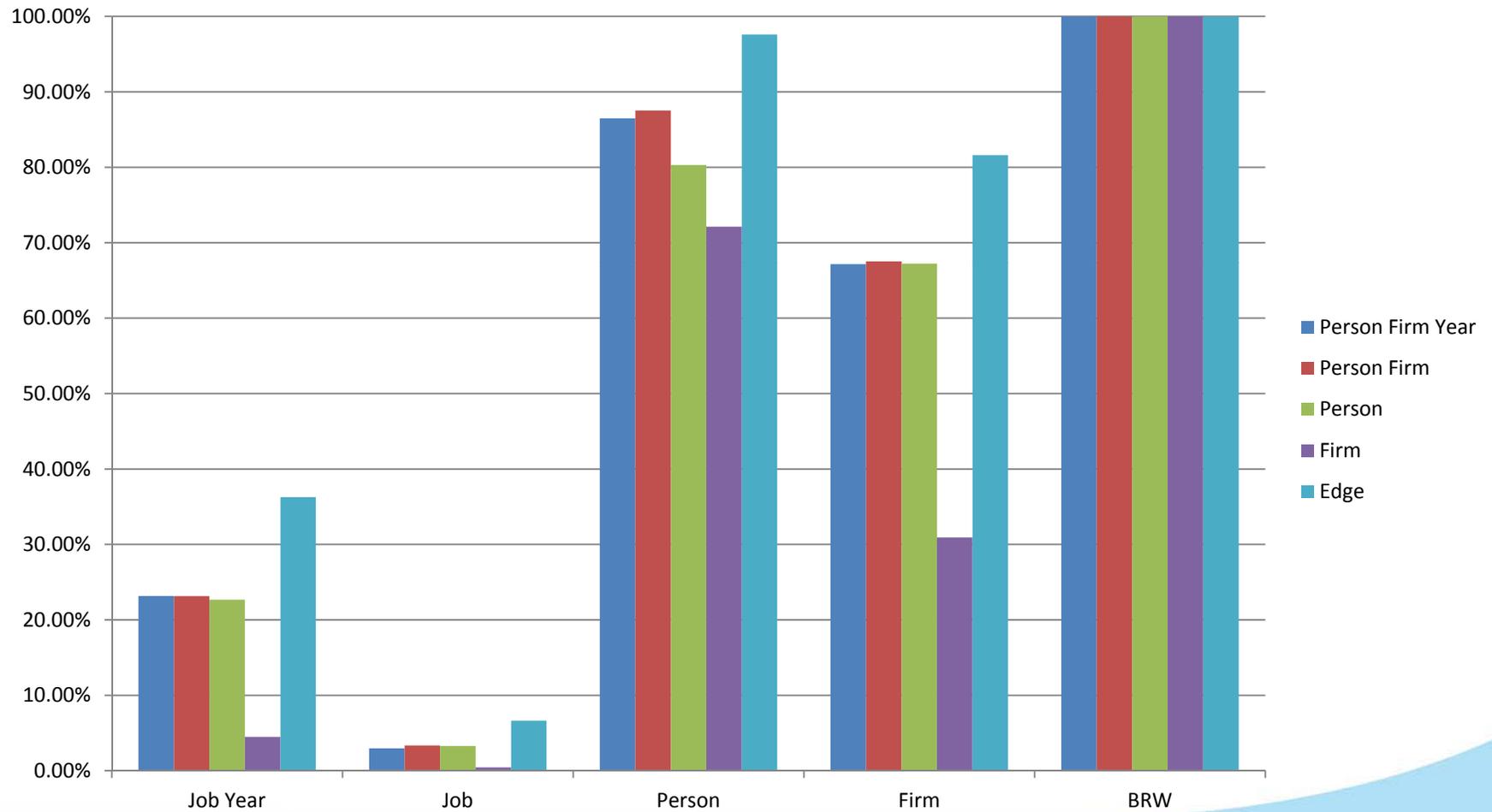
# Solving for the Transition Probability Matrix

- Markov chain theory can be used to show that if a graph is undirected, connected and non-bipartite, a random walk will converge to a steady state  $\pi$ .
- The steady state is characterized by the following equation:  
 $\pi^* T = \pi$ .
- The transition probability matrix for the standard random walk results in a sampling probability that is biased towards high degree nodes:  $\pi_i = \text{deg}(v_i) / (2 * |E|)$
- Looking for the transition probability matrix such that  $\pi_i = 1/|V|$
- Key insight is that when  $T$  is doubly stochastic,  $\pi_i = 1/|V|$
- Start at standard random walk values and use iterative proportional fitting or raking to solve for  $T$ .
- Bound transition probability matrix from zero and one so that the graph remains connected.

# Data

- LEHD core master files.
- Select workers with positive earnings any time during 2002:1 to 2008:4 in 20 contiguous counties along the borders of Idaho, Montana, and Wyoming.
- Workers with more than 28 jobs were dropped from the sample
- Only workers, firms, and jobs in the largest connected group are retained.
- Resulting sample is 1,895,697 job year, 995,664 jobs, 20,618 firms, and 535,194 persons.

# Size of Largest Component (2% Sampling Rate)



# Balanced Random Walk

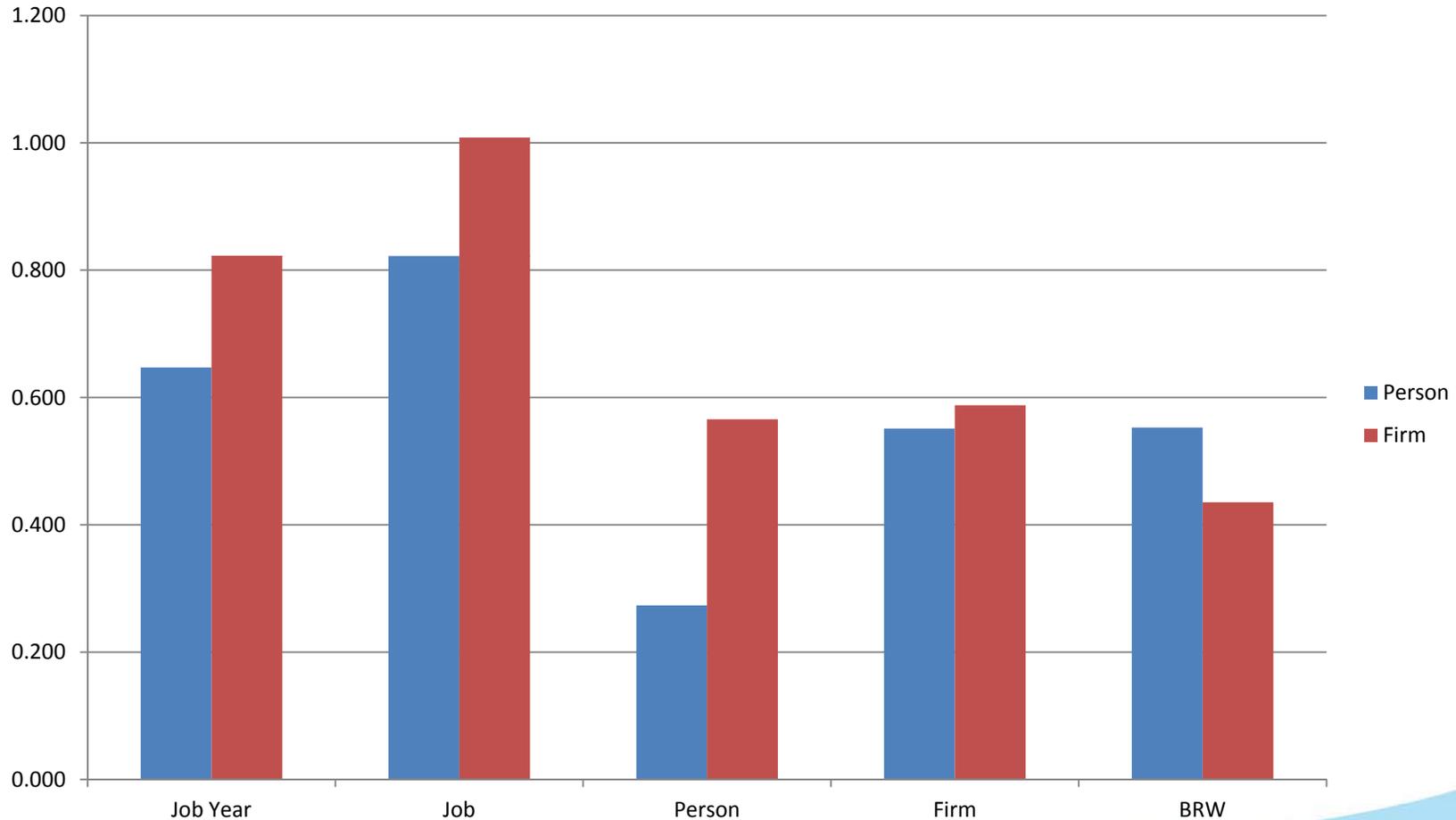
- Sampling bias is similar to SRS firm method.
- All sampled units are in a single connected component.
- For the 10% (2%) sampling rate, over a 70% (320%) increase in the number of edges compared to the SRS firm sampling method.
- Increased sample to sample variability.

# Earnings Model

- $Y = X\beta + Z\mu + \varepsilon$ 
  - $Y$ =vector of log real earnings
  - $X$ =quarters worked dummies, age, and age squared
  - $Z$ =matrix of design effects for person and firm.
- Fixed or mixed effects can be used to estimate  $\mu$ .
- Connected sampling enables fixed effect estimation for the entire sample.
- Mixed effects does not require a connected sample, but may benefit.

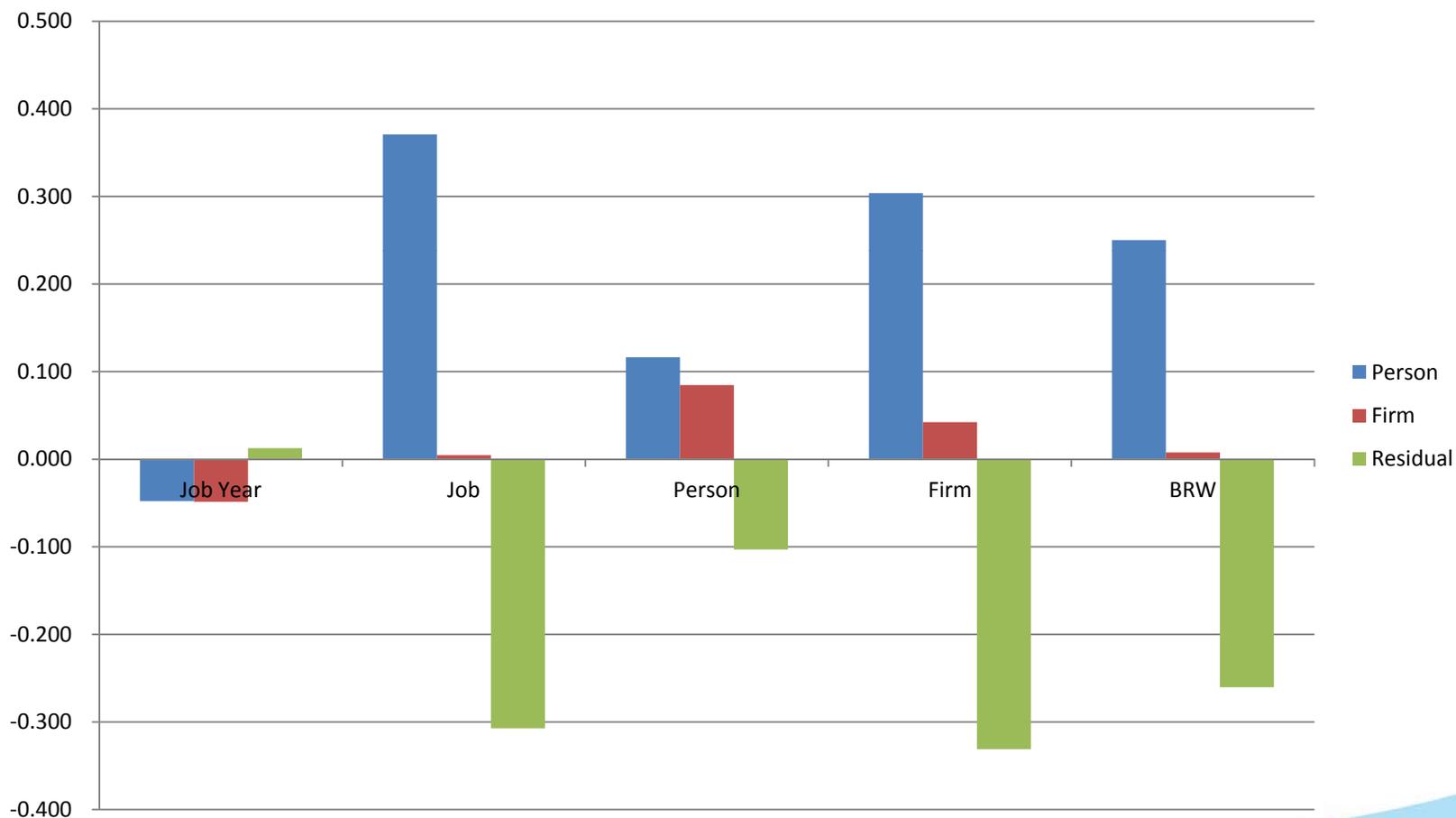
# Fixed Person and Firm Effects

(abs(Estimate-Population Value) for Largest Connected Group)  
(10% Sampling Rate)



# Mixed Effect Variance Components

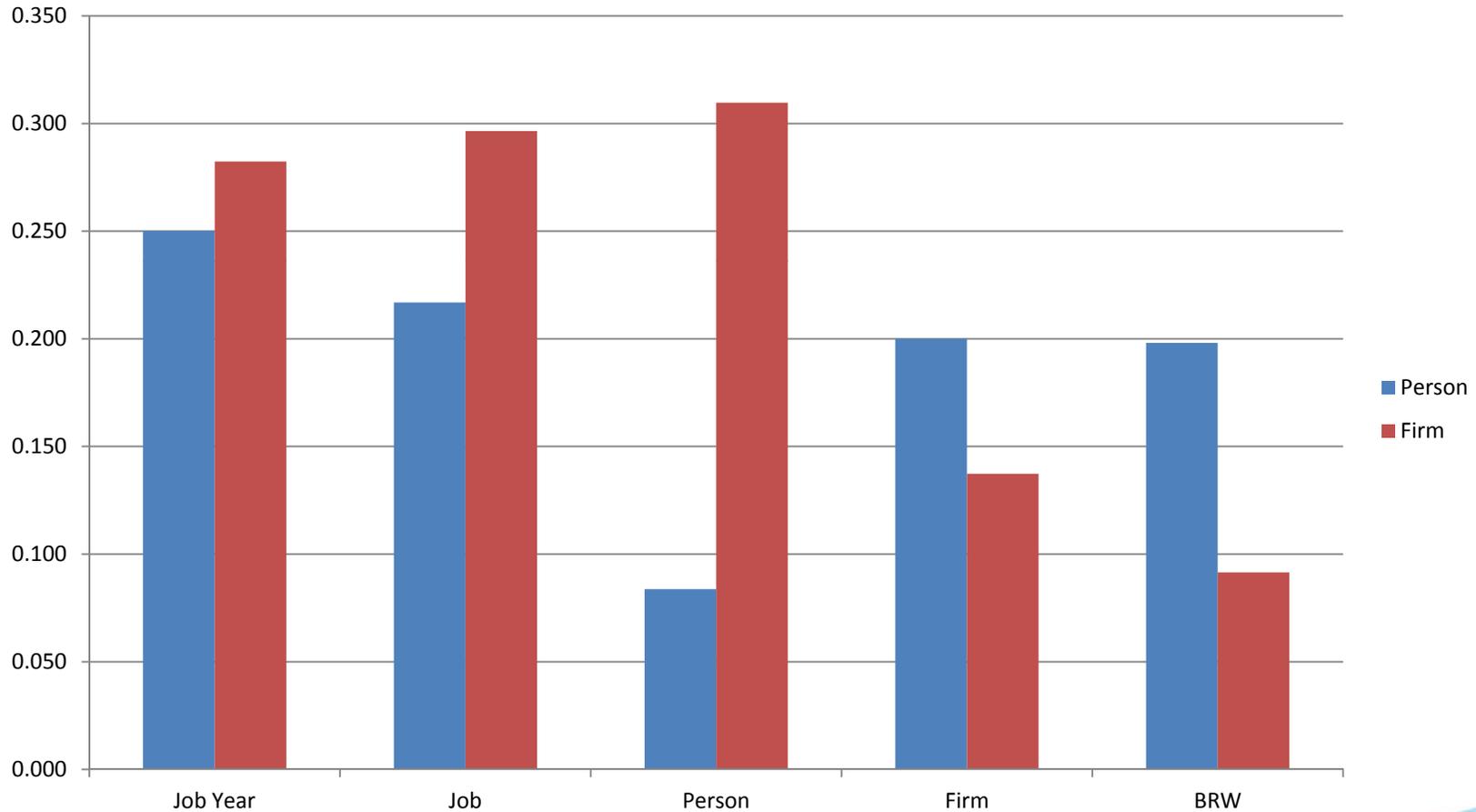
(Estimate-Population Value)  
(2% Sampling Rate)



# Mixed Person and Firm Effects (eBLUP)

(abs(Estimate-Population Value))

(10% Sampling Rate)



# Conclusion

- Sampling bias present for at least one population when using any of the SRS methods.
- Balanced random walk produces a connected sub-graph of the largest component.
- Fixed effect estimates of  $\mu$ .
  - Person sampling and the BRW perform the best (smallest average absolute deviation from the population estimate).
  - Person sampling method performs best for fixed person effects. BRW performs best for fixed firm effects
  - Unlike other methods, BRW enables fixed effect estimation for the entire sample.
- Estimation of variance components for mixed effect model does not show a large benefit from using a connected sample. Job year samples have smallest bias.
- Mixed effect estimates of  $\mu$  (eBLUP).
  - Have smaller bias than fixed effects.
  - Person sampling performs best for realized random person effects. BRW performs best for realized random firm effects
- The person sampling method over-samples firms, allowing for the creation of a large number of edges. Increasing the BRW firm sampling rate and sub-sampling jobs within firms would allow for a fairer comparison of the two methods.