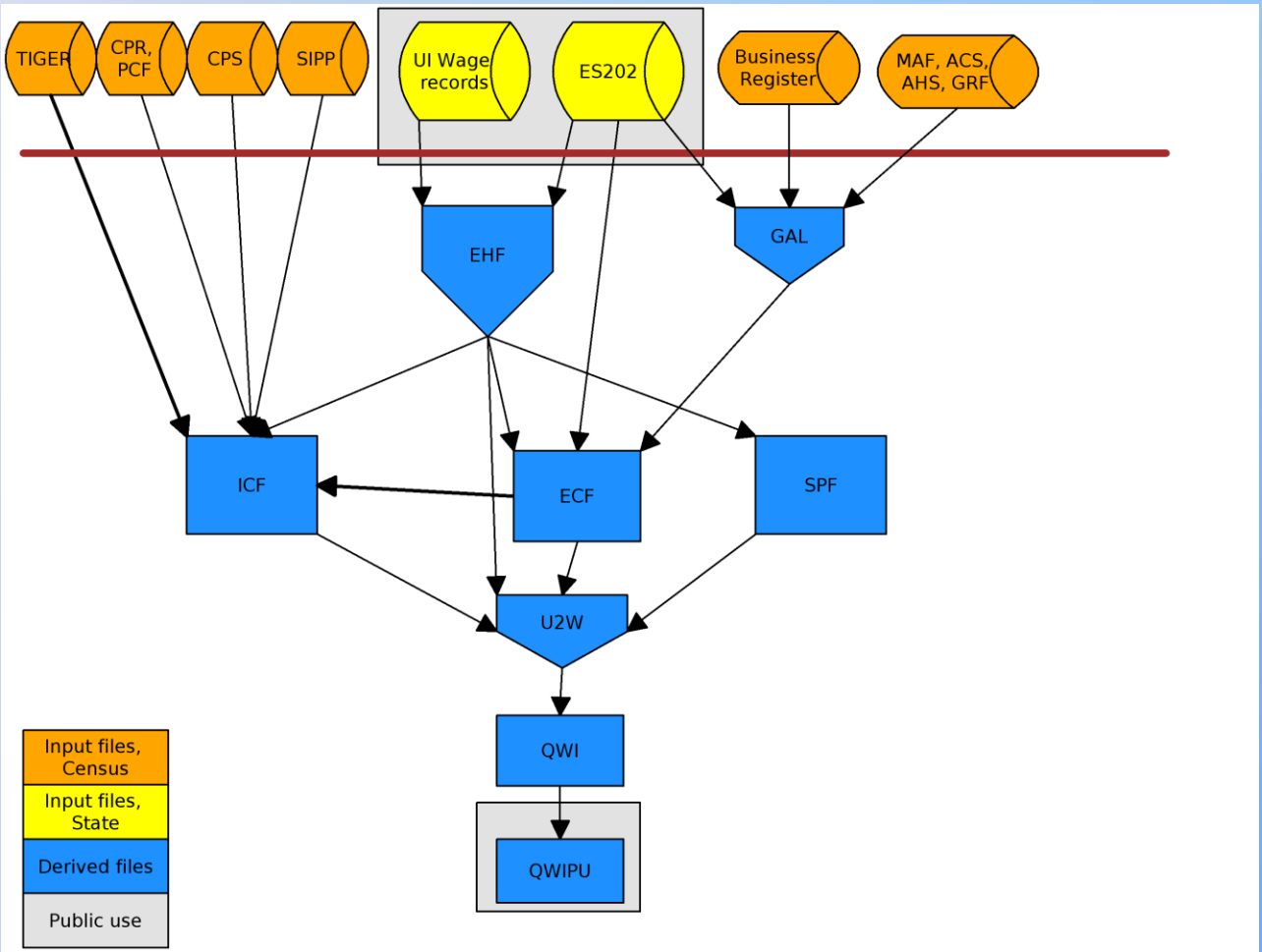
INFO 7470/ECON 7400/ILRLE 7400
Understanding LEHD
Sources and Structure

John M. Abowd and Lars Vilhuber

February 2013, 2016

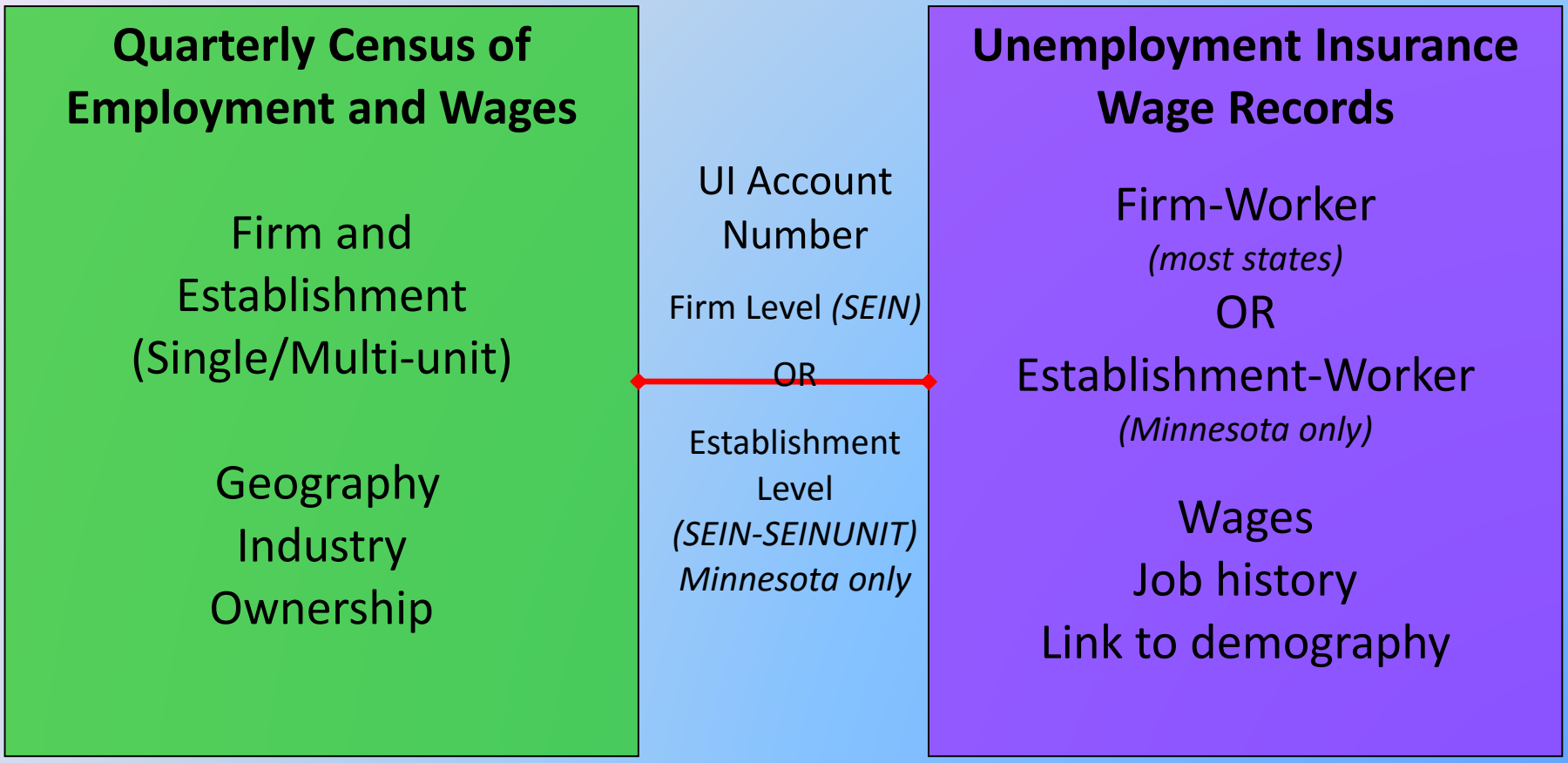
Data Flow View of the LEHD Infrastructure



QWI Production Process: Key Stages

Process	Name	Description
EHF	Employment History File	Longitudinal employment history for individuals
GAL	Geocoded Address List	Assigns coordinates to establishments
ECF	Employer Characteristics File	Longitudinal establishment-level data: employment, NAICS, geography, ownership Imputes when necessary; Maintains “fuzz factors”; calculates weights
SPF	Successor-Predecessor File	Identifies successor-predecessor relationships based on employment flows
ICF	Individual Characteristics File	Individual level demographic data Imputes when necessary
U2W	Unit-to-Worker	Multiply-imputes establishment to jobs at multi-establishment firms
QWI	Quarterly Workforce Indicators	Generate final estimates, apply weighting, confidentiality protections *(QWIPU = QWI Public Use file)*

LEHD Processing: Merging QCEW and UI Data



LEHD Processing: Successor-Predecessor

- Adjustments to account for administrative changes to firms
  - mergers, divestitures, etc.
- Transitions may be identified through:
  - report on the QCEW
    - firm level and establishment level
  - finding large employment flow from the individual wage records
    - firm level only
- Individual job history at predecessor is concatenated with job history for same PIK at successor for purposes of calculating QWI measures
  - Important CAVEAT for RDC work: this does NOT generate a new firm identifier – researchers need to apply similar logic to their research extracts

LEHD Processing: Unit-to-Worker Impute

- Necessary to impute establishment to a job when not available
  - Currently only Minnesota reports establishments on wage data
- Individual job histories are assembled
- Establishment (multiply) imputed to longitudinal job, with the following predictors:
  - Proximity of residence to establishment
  - Size of establishment
- Establishment history (allowing for predecessors) must be consistent with individual job history

LEHD Processing: Weighting

- QWI B is benchmarked against QCEW Mon1 employment
- Firm-level weights (within bounds) are applied to adjust employment towards Mon1 employment
- Secondary weights are applied to match statewide private-only employment
- Weights are calculated at ECF stage, applied at QWI

LEHD Processing: QCEW-QWI Differences

- Sub-state adjustments are not currently applied to QWI data
- While state employment totals should be quite close, sub-state estimates will display deviations from benchmark
  - County, industry employment totals, or smaller cells
- These differences can come from any of a number of QWI-specific processing steps
    - Specific differences observed in the data may also result from an interaction of several sources of deviation

Causes of Differences: Measure Definition

- B and Month 1 do not capture exactly the same universe
  - An individual may count towards either one of the measures, but not towards the other
- Differences generally minor, but may be noticeable in some industries with particular seasonal patterns
  - e.g., education, agriculture

Causes of Differences: BLS Data Editing

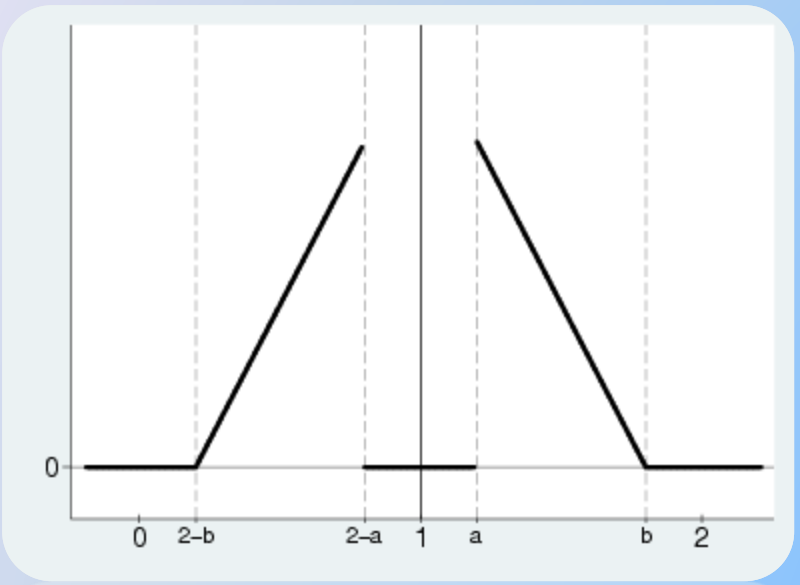
- LEHD data receipts
  - Before 2004 LEHD received BLS edited data from state partners
  - Since 2004 LEHD does not receive BLS edited data from states (CIPSEA)
- BLS QCEW file may be edited/different from that which LEHD receives
  - Completeness
  - Imputed employment
  - Industry/geography changes
- Statewide totals are close (<1% off)
- LEHD QA will periodically note BLS QCEW data inconsistent with internal LEHD QCEW micro-data

Causes of Differences: Noise Infusion ("Fuzzing")

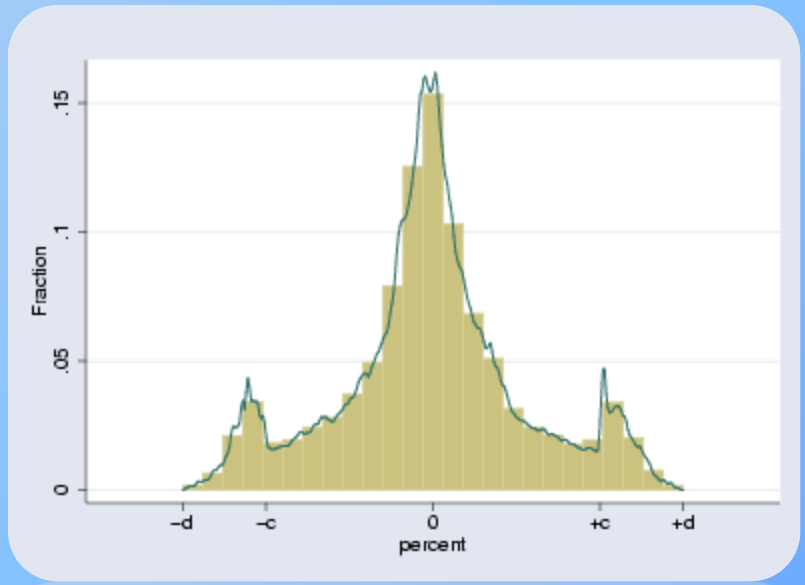
- Why infuse noise into data?
  - Reduce the amount of cell suppression while preserving confidentiality and analytic validity
- Properties of noise
  - Every data item is distorted by a minimum amount
  - For a given workplace, data are always distorted in the same direction, by the same percentage in every period and release of QWI's
  - When aggregated, the effects of the distortion cancel out for the vast majority of the estimates

Noise infusion in QWI

Theoretical distribution



Empirical distribution of noise



Source: [Abowd, Gittings, McKinney, Stephens, Vilhuber, Woodcock (2012)]

Causes of Differences: Noise Infusion ("Fuzzing")

- QWI statistics are flagged when the value is significantly distorted (Status flag 9)
- See infrastructure document, section 6, for more details
- More about this later in the lecture on *Confidentiality Protection*

Causes of Differences: UI Wage Data Reporting

- Firm may fail to report wage records
  - QCEW still reported or imputed
- Firm may report wage records and QCEW records on different account numbers
  - Successor/predecessor mistiming
  - Public sector issues
- PIK (SSN) miscoding prevents linking wage records to same longitudinal job (Abowd & Vilhuber, 2005)

Causes of Differences: Industry Assignment

- Most establishments are assigned based on the reported NAICS_AUX
- For earlier years in the data series, the reported SIC code is probabilistically mapped to the current NAICS codes
  - Imputes may also be used for transitions between 1997, 2002, and 2007 NAICS
- LDB data are used for NAICS back-coding purposes when the file has been provided by state
- Variations in algorithms between LEHD and BLS may result in differences
  - NAICS sector 55 (management of companies) displays particular issues during SIC-NAICS transition

Causes of Differences: Geographic Coding

- LEHD performs own geo-coding of addresses
  - Generates lat-long for distance measures, allows custom geography
- Address data are processed along with address data from other sources
- Results may differ from BLS assignments
  - Marginal shift over county line
  - Significant relocation
- Effort currently underway to reengineer LEHD geographic assignment to improve results

Causes of Differences: Multiple Worksites (U2W)

- QCEW can report Mon1 by building directly from establishment (with geo/industry info)
- LEHD "No transfer" assumption a single job spell to be reported to the same establishment
  - Job spell – PIK-SEIN relationship that does not contain four consecutive quarters with zero earnings.
- A change in firm structure can make it impossible to replicate counts given constraint
  - Long-term differences may result from new, large establishments appearing without predecessor

Causes of Differences: Successor-Predecessor

- QCEW can, again, build up estimates directly from establishment
  - Does not matter for month1 purpose if predecessor existed
- LEHD must have information from previous and following quarters for range of measures
  - If a new firm appears, and that firm does not have a predecessor (with same employees), jobs at the new firm will not count towards primary LEHD *B* employment in that quarter

Data Irregularity: Missing UI Records

- Impact of large firm that fails to report UI wage data (or reports late) in 2009Q2.

	M EmpTotal	B Emp	E EmpEnd	F EmpS
2009 Q1	-	-	⬇	⬇
2009 Q2	⬇	⬇	⬇	⬇
2009 Q3	-	⬇	-	⬇
2009 Q4	-	-	-	-

Data Irregularity: Spike in Wage Records

- Impact of large firm that displays unusual spike for only 2009Q2

  e.g., back pay for a court settlement

	M EmpTotal	B Emp	E EmpEnd	F EmpS
2009 Q1	▬	▬	▬	▬
2009 Q2	⬆	▬	▬	▬
2009 Q3	▬	▬	▬	▬
2009 Q4	▬	▬	▬	▬

Data Irregularity: Unidentified Succ-Pred

- Firm reported under account X in 2009Q1, account Y in 2009Q2 (same geography, industry, job count);
  - Transition not identified in LEHD processing

	M EmpTotal	B Emp	E EmpEnd	F EmpS
2009 Q1	-	-	⬇	⬇
2009 Q2	-	⬇	-	⬇
2009 Q3	-	-	-	-
2009 Q4	-	-	-	-

Data Irregularity: Missing UI Records

- Impact of large firm that fails to report UI wage data (or reports late) in 2009Q2.
  - Assume most workers had stable employment before and after

	A HirA	H HirN	S Sep	FA HirAS	H3 HirNS	FS SepS
2009 Q1	-	-	⬆	-	-	⬆
2009 Q2	-	-	-	-	-	-
2009 Q3	⬆	-	-	-	-	-
2009 Q4	-	-	-	⬆	-	-

Data Irregularity: Spike in Wage Records

- Impact of large firm that displays unusual spike for only 2009Q2

  e.g., back pay for a court settlement for fired workers

	A HirA	H HirN	S Sep	FA HirAS	H3 HirNS	FS SepS
2009 Q1	-	-	-	-	-	-
2009 Q2	⬆	⬆/- *	⬆	-	-	-
2009 Q3	-	-	-	-	-	-
2009 Q4	-	-	-	-	-	-

* If workers had not received earnings in the past year, they would be new hires

Data Irregularity: Unidentified Successor-Predecessor

- Firm reported under account X in 2009Q1, account Y in 2009Q2 (same geography, industry, job count)
  - Transition not identified in LEHD processing
  - Assume most workers had stable employment before and after

	A HirA	H HirN	S Sep	FA HirAS	H3 HirNS	FS SepS
2009 Q1	-	-	⬆	-	-	⬆
2009 Q2	⬆	-	-	-	-	-
2009 Q3	-	-	-	⬆	-	-
2009 Q4	-	-	-	-	-	-

Firm Job Flows: Be Careful about Aggregation

- Note that for categories like age and sex, the published net job flows for the subcategories will sum to the margin
- But for gross Job Creation and gross Job Destruction this is not true
- (Job Creation for men) + (Job Creation for women) does not equal (total Job Creation)
  - For example, a job could be created at a firm and filled by a woman, while another job *at the same firm* is destroyed, previously filled by a man

	Men	Women	Total
Job Creation	0	1	0
Job Destruction	1	0	0
Net Job Flows	-1	+1	0

Overview: Summary

- The QWI are developed by incorporating data from a broad variety of sources
- Differences in data sources, construction, and imputation procedures may cause employment estimates that do not match other sources