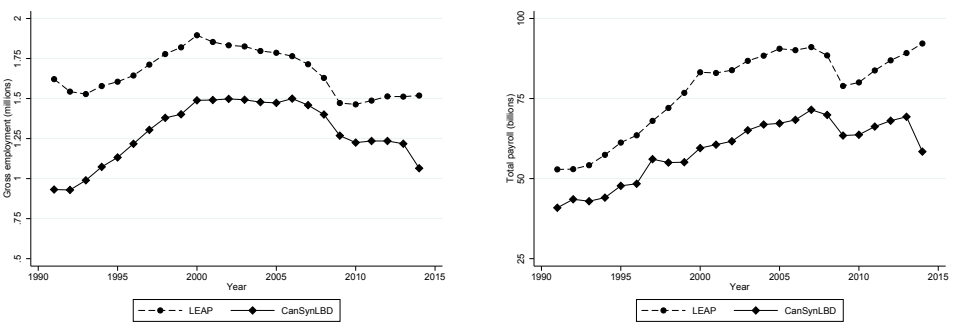
Online Appendix

"Applying Data Synthesis for Longitudinal Business Data across Three Countries"

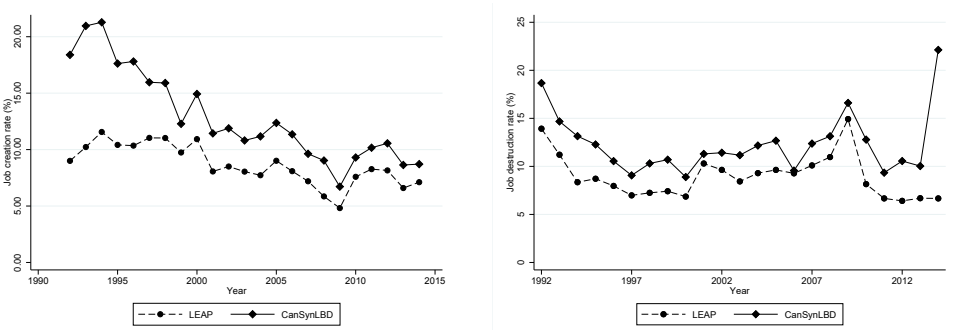
M. Jahangir Alam, Benoit Dostie, Jörg Drechsler, Lars Vilhuber

A. Figures for the Manufacturing Sector in Canada



(a) Gross employment level by year (b) Total payroll

Figure 8: Entity characteristics for the manufacturing sector in Canada by year.



(a) Job creation rates (b) Job destruction rates

Figure 9: Dynamics of job flows for the manufacturing sector in Canada by year.

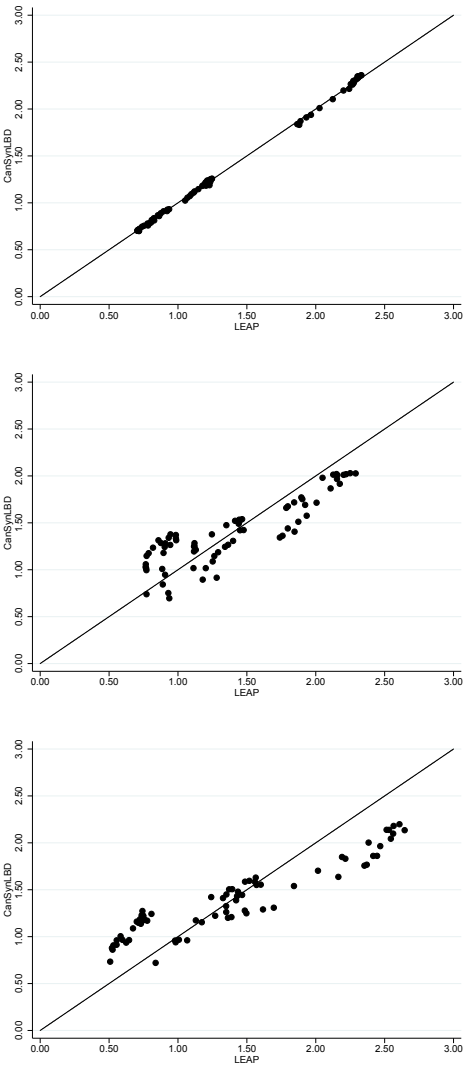


Figure 10: Share of entities (upper panel), share of employment (middle panel), and share of payroll (lower panel) by year and industry for the Canadian manufacturing sector.

Table 4: Detailed results for pMSE estimation by sector and country

Independent Variables	**Canada**		**Germany**
Sector:	Manufacturing	Private	All
Ln ALU	0.158	0.7138	-0.2895
	(0.0039)	(0.001)	(0.0033)
Ln Pay	0.0039	-0.4426	0.2584
	(0.0037)	(0.001)	(0.0028)
Age 3-4	0.0392	0.0972	-0.0987
	(0.0078)	(0.0017)	(0.007)
Age 5-7	-0.0382	0.0477	-0.0973
	(0.0073)	(0.0016)	(0.0066)
Age 8-12	-0.1258	-0.0263	-0.1172
	(0.0071)	(0.0015)	(0.0063)
Age 13 or more	-0.219	-0.1024	-0.1487
	(0.0074)	(0.0016)	(0.0059)
N	2243011	34638723	2121956
pseudo R-sq	0.0112	0.0318	0.0038
pMSE	0.0041	0.0121	0.0013

Note: See Equation 1 for estimation method. An observation is a entity-year in the combined database of each country-sector combination. All specifications include time and industry fixed effects. Standard errors are in parentheses.

B. Appendix Tables

B.1. pMSE

B.2. Regression analysis tables

Table 5: Regression coefficients (OLS) for LEAP

Independent Variables	**LEAP**		**CanSynLBD**	
	Private	Manufacturing	Private	Manufacturing
AR(1) Coefficient	0.2031***	0.2481***	0.3970***	0.4405***
	(0.0001)	(0.0005)	(0.0002)	(0.0007)
Ln Pay	0.7847***	0.7300***	0.5481***	0.5228***
	(0.0001)	(0.0005)	(0.0002)	(0.0006)
Age 3-4	-0.1202***	-0.1717***	-0.1223***	-0.2340***
	(0.0003)	(0.0014)	(0.0004)	(0.0016)
Age 5-7	-0.1260***	-0.1891***	-0.1235***	-0.2507***
	(0.0003)	(0.0014)	(0.0004)	(0.0016)
Age 8-12	-0.1268***	-0.1973***	-0.1169***	-0.2551***
	(0.0003)	(0.0013)	(0.0004)	(0.0016)
Age 13 or more	-0.1246***	-0.1992***	-0.1101***	-0.2577***
	(0.0003)	(0.0014)	(0.0004)	(0.0017)
N	15708195	1015293	13573225	959764
R^2	0.9696	0.9743	0.9444	0.9523

Note: In all specifications, we include both year and industry fixed effects. Standard errors are in parentheses. ***, **, and * indicate statistically significant coefficients at 1%, 5%, and 10% percent levels, respectively.

Table 6: Regression coefficients (OLS) for GLBD

Independent Variables	**GLBD**	**GSynLBD**
AR(1) Coefficient	0.4430***	0.4143***
	(0.0007)	(0.0008)
Ln Pay	0.4629***	0.5143***
	(0.0006)	(0.0007)
Age 3-4	-0.0695***	-0.0642***
	(0.0017)	(0.0016)
Age 5-7	-0.1066***	-0.0891***
	(0.0017)	(0.0016)
Age 8-12	-0.1324***	-0.1109***
	(0.0017)	(0.0016)
Age 13 or more	-0.1880***	-0.1600***
	(0.0016)	(0.0015)
N	848871	966084
R^2	0.9167	0.8968

Note: In all specifications, we include both year and industry fixed effects. Standard errors are in parentheses. ***, **, and * indicate statistically significant coefficients at 1%, 5%, and 10% percent levels, respectively.

Table 7: Regression coefficients (Dynamic) for LEAP

Independent Variables	**LEAP**		**CanSynLBD**	
	Private	Manufacturing	Private	Manufacturing
AR(1) Coefficient	0.0805***	0.1189***	0.5722***	0.5425***
	(0.0003)	(0.0018)	(0.0024)	(0.0084)
Ln Pay	0.8991***	0.8523***	0.4101***	0.4302***
	(0.0002)	(0.0015)	(0.0018)	(0.0067)
Age 3-4	-0.0450***	-0.0797***	-0.2075***	-0.2972***
	(0.0002)	(0.0014)	(0.0010)	(0.0051)
Age 5-7	-0.0438***	-0.0860***	-0.2129***	-0.3162***
	(0.0002)	(0.0015)	(0.0011)	(0.0059)
Age 8-12	-0.0418***	-0.0923***	-0.2187***	-0.3294***
	(0.0003)	(0.0017)	(0.0013)	(0.0070)
Age 13 or more	-0.0379***	-0.0898***	-0.2318***	-0.3414***
	(0.0003)	(0.0019)	(0.0015)	(0.0080)
N	15708195	1015293	13573225	959764
m2	-14.5000	-2.2200	-27.5400	-9.4400
Sargan test	6.9e+04	4.6e+03	1.5e+04	1.5e+03
df of Sargan Test	252.0000	252.0000	252.0000	252.0000
P value of Sargan test	0.0000	0.0000	0.0000	0.0000

Note: In this table, $m2$ is the Arellano-Bond test for zero autocorrelation in first-differenced errors for order two. Standard errors are in parentheses. ***, **, and * indicate statistically significant coefficients at 1%, 5%, and 10% percent levels, respectively.

Table 8: Regression coefficients (Dynamic) for GLBD

Independent Variables	**GLBD**	**GSynLBD**
AR(1) Coefficient	0.0489***	0.6999***
	(0.0051)	(0.0057)
Ln Pay	0.7559***	0.2916***
	(0.0035)	(0.0042)
Age 3-4	-0.0070***	-0.1026***
	(0.0012)	(0.0015)
Age 5-7	-0.0233***	-0.1386***
	(0.0014)	(0.0017)
Age 8-12	-0.0473***	-0.1694***
	(0.0015)	(0.0018)
Age 13 or more	-0.1084***	-0.2183***
	(0.0015)	(0.0018)
N	848871	966084
m2	-2.5100	-4.1300
Sargan test	3.6e+03	2.0e+03
df of Sargan Test	495.0000	495.0000
P value of Sargan test	0.0000	0.0000

Note: In this table, $m2$ is the Arellano-Bond test for zero autocorrelation in first-differenced errors for order two. Standard errors are in parentheses. ***, **, and * indicate statistically significant coefficients at 1%, 5%, and 10% percent levels, respectively.

Table 9: Regression coefficients (Dynamic - system GMM) for LEAP

Independent Variables	**LEAP**		**CanSynLBD**	
	Private	Manufacturing	Private	Manufacturing
AR(1) Coefficient	0.0978***	0.1614***	0.5111***	0.5780***
	(0.0002)	(0.0014)	(0.0008)	(0.0041)
Ln Pay	0.8854***	0.8161***	0.4562***	0.4022***
	(0.0002)	(0.0012)	(0.0006)	(0.0033)
Age 3-4	-0.0555***	-0.1097***	-0.1828***	-0.3177***
	(0.0002)	(0.0012)	(0.0004)	(0.0028)
Age 5-7	-0.0558***	-0.1201***	-0.1860***	-0.3408***
	(0.0002)	(0.0013)	(0.0005)	(0.0031)
Age 8-12	-0.0548***	-0.1298***	-0.1875***	-0.3583***
	(0.0002)	(0.0014)	(0.0005)	(0.0036)
Age 13 or more	-0.0524***	-0.1317***	-0.1943***	-0.3747***
	(0.0002)	(0.0016)	(0.0006)	(0.0041)
N	15708195	1015293	13573225	959764
m2	-11.4300	1.3900	-41.6000	-7.6700
Sargan test	7.7e+04	6.3e+03	1.8e+04	1.7e+03
df of Sargan Test	274.0000	274.0000	274.0000	274.0000
P value of Sargan test	0.0000	0.0000	0.0000	0.0000

Note: An observation is an entity-year. In this table, $m2$ is the Arellano-Bond test for zero autocorrelation in first-differenced errors for order two. Standard errors are in parentheses. ***, **, and * indicate statistically significant coefficients at 1%, 5%, and 10% percent levels, respectively.

Table 10: Regression coefficients (Dynamic - system GMM) for GLBD

Independent Variables	**GLBD**	**GSynLBD**
AR(1) Coefficient	0.1883***	0.6140***
	(0.0021)	(0.0027)
Ln Pay	0.6599***	0.3553***
	(0.0014)	(0.0020)
Age 3-4	-0.0292***	-0.0934***
	(0.0011)	(0.0013)
Age 5-7	-0.0512***	-0.1266***
	(0.0011)	(0.0014)
Age 8-12	-0.0791***	-0.1545***
	(0.0011)	(0.0015)
Age 13 or more	-0.1400***	-0.2012***
	(0.0011)	(0.0015)
N	848871	966084
m2	19.4900	-8.8300
Sargan test	4.5e+03	2.8e+03
df of Sargan Test	526.0000	526.0000
P value of Sargan test	0.0000	0.0000

Note: An observation is an entity-year. In this table, $m2$ is the Arellano-Bond test for zero autocorrelation in first-differenced errors for order two. Standard errors are in parentheses. ***, **, and * indicate statistically significant coefficients at 1%, 5%, and 10% percent levels, respectively.

Table 11: Regression coefficients (Dynamic - system GMM with MA(1)) for LEAP

Independent Variables	**LEAP**		**CanSynLBD**	
	Private	Manufacturing	Private	Manufacturing
AR(1) Coefficient	0.2005***	0.2821***	0.4850***	0.5737***
	(0.0007)	(0.0040)	(0.0012)	(0.0059)
Ln Pay	0.8044***	0.7135***	0.4760***	0.4056***
	(0.0005)	(0.0034)	(0.0009)	(0.0046)
Age 3-4	-0.1245***	-0.2033***	-0.1716***	-0.3158***
	(0.0005)	(0.0032)	(0.0006)	(0.0037)
Age 5-7	-0.1328***	-0.2264***	-0.1733***	-0.3389***
	(0.0005)	(0.0035)	(0.0006)	(0.0043)
Age 8-12	-0.1383***	-0.2454***	-0.1731***	-0.3560***
	(0.0006)	(0.0039)	(0.0007)	(0.0051)
Age 13 or more	-0.1441***	-0.2586***	-0.1774***	-0.3717***
	(0.0006)	(0.0042)	(0.0008)	(0.0058)
N	15708195	1015293	13573225	959764
m2	8.2000	7.0600	-40.0300	-6.6400
Sargan test	2.8e+04	2.3e+03	1.7e+04	1.3e+03
df of Sargan Test	251.0000	251.0000	251.0000	251.0000
P value of Sargan test	0.0000	0.0000	0.0000	0.0000

Note: An observation is a firm and a year. In this table, $m2$ is the Arellano-Bond test for zero autocorrelation in first-differenced errors for order two. $LEAP$ is the Longitudinal Employment Analysis Program and $CanSynLBD$ is the Canadian synthetic database based on LEAP. In this table, we use 2015 vintage of LEAP and drop last year observation of each firm. Standard errors are in parentheses. ***, **, and * indicate statistically significant coefficients at 1%, 5%, and 10% percent levels, respectively.

Table 12: Regression coefficients (Dynamic - system GMM with MA(1)) for GLBD

Independent Variables	**GLBD**	**GSynLBD**
AR(1) Coefficient	0.3701***	0.5268***
	(0.0060)	(0.0048)
Ln Pay	0.5349***	0.4202***
	(0.0041)	(0.0036)
Age 3-4	-0.0594***	-0.0831***
	(0.0015)	(0.0013)
Age 5-7	-0.0922***	-0.1105***
	(0.0018)	(0.0015)
Age 8-12	-0.1252***	-0.1351***
	(0.0019)	(0.0016)
Age 13 or more	-0.1850***	-0.1802***
	(0.0019)	(0.0017)
N	848871	966084
m2	19.0300	-11.6900
Sargan test	3.1e+03	2.5e+03
df of Sargan Test	494.0000	494.0000
P value of Sargan test	0.0000	0.0000

Note: An observation is a firm and a year. In this table, $m2$ is the Arellano-Bond test for zero autocorrelation in first-differenced errors for order two. Standard errors are in parentheses. ***, **, and * indicate statistically significant coefficients at 1%, 5%, and 10% percent levels, respectively.

C. Canada: Synthesized Observations

Table 13: Synthesized observations

Category	# of Observations (millions)	Percentage
Synthesized	22.01	93.35
Not synthesized	1.57	6.65
Total	23.58	100.00

Note: Not synthesized industries are NAICS 4481, 4482, 4483, 4511, 4513, 4841, 4842, 5241, and 5242. These industries are not converging for each time of implementation We drop industries, from the synthesized industries, which have less than ten observations in a given year. We do not synthesize the public sector (NAICS 61, 62, and 91).

D. Confidentiality assessment

Table 14: Observed entity births given synthetic births for LEAP.

First (Birth) Year		% of Births over NAICS		
Synthetic	Actual	Minimum	Mean	Maximum
1991	1991	0.00	27.69	83.02
1992	1992	0.00	3.37	11.11
1993	1993	0.00	3.79	33.33
1994	1994	0.00	3.73	33.33
1995	1995	0.00	3.86	20.00
1996	1996	0.00	4.25	33.33
1997	1997	0.00	4.10	16.94
1998	1998	0.00	4.41	25.00
1999	1999	0.00	4.23	33.33
2000	2000	0.00	3.41	25.00
2001	2001	0.00	2.73	22.22
2002	2002	0.00	2.65	25.00
2003	2003	0.00	2.22	10.00
2004	2004	0.00	2.60	17.86
2005	2005	0.00	2.71	20.00
2006	2006	0.00	2.83	50.00
2007	2007	0.00	2.90	33.33
2008	2008	0.00	2.38	20.00
2009	2009	0.00	2.47	50.00
2010	2010	0.00	2.12	33.33
2011	2011	0.00	2.65	50.00
2012	2012	0.00	2.41	20.00
2013	2013	0.00	2.48	25.00
2014	2014	0.00	2.23	20.00
2015	2015	0.00	2.15	33.33

Table 15: Observed entity births given synthetic births (GLBD)

Birth Year		% of Births over NAICS		
Synthetic	Actual	Minimum	Mean	Maximum
1976	1976	18.34	19.77	21.20
1977	1977	1.35	1.55	1.75
1978	1978	0.97	1.50	2.02
1979	1979	1.99	2.05	2.11
1980	1980	1.15	1.61	2.07
1981	1981	0.76	1.28	1.80
1982	1982	1.29	1.39	1.48
1983	1983	1.54	1.57	1.61
1984	1984	0.99	1.03	1.07
1985	1985	0.83	1.56	2.28
1986	1986	1.36	1.79	2.21
1987	1987	1.99	2.00	2.02
1988	1988	1.18	1.49	1.81
1989	1989	1.65	1.84	2.03
1990	1990	2.44	2.79	3.14
1991	1991	7.59	9.17	10.75
1992	1992	5.19	8.81	12.42
1993	1993	3.20	3.40	3.60
1994	1994	3.50	3.93	4.35
1995	1995	2.86	3.26	3.65
1996	1996	1.89	2.62	3.35
1997	1997	3.46	3.96	4.45
1998	1998	3.58	3.68	3.78
1999	1999	5.56	5.78	6.00
2000	2000	3.19	3.64	4.10
2001	2001	3.26	3.59	3.93
2002	2002	2.04	3.00	3.97
2003	2003	2.13	3.17	4.20
2004	2004	2.57	3.24	3.91
2005	2005	1.66	2.54	3.41
2006	2006	2.15	3.06	3.97
2007	2007	2.17	2.90	3.62
2008	2008	2.37	2.42	2.47