Unlocking the Information in Integrated Social Data

John M. Abowd

Cornell University

US Census Bureau

CREST and NBER

May 2002

This paper was prepared for the New Zealand Ministry of Labor Conference on Data Integration and Linked Employer-Employee Data held March 21-22, 2002. I would like to thank my coauthors Patrick Corbel, Rob Creecy, Bruno Crépon, John Haltiwanger, Francis Kramarz, Julia Lane, Paul Lengermann, David Margolis, Kevin McKinney, Ron Prevost, Sébastien Roux, Kristin Sandusky, Martha Stinson, Lars Vilhuber, and Simon Woodcock from whom I borrowed shamelessly for this presentation. I would also like to thank my Cornell, Census, and CREST colleagues Fredrik Andersson, Gary Benedetto, Chet Bowie, Bahattin Buyuksahin, Karen Conneely, Pat Doyle, Ron Ehrenberg, Nancy Gordon, Cheryl Grim, George Jakubson, Frederick Knickerbocker, Guy LaRoque, Cyr Linonis, Charles McCulloch, Nicole Nestoriak, Sébastien Perez-Duarte, Jean-Marc Robin, Marc Roemer, Bryce Stephens, Marty Wells, Bill Winkler, and Laura Zayatz, each of whom contributed substantially to this work. Funding was provided by the US Census Bureau, INSEE/CREST, the National Science Foundation (SES-9978093 to Cornell University), the National Institute on Aging, the Sloan Foundation, and the Russell Sage Foundation.

1. Introduction

Modern national and research statistical systems acquire their information from three related sources: censuses, periodic surveys, and administrative records. In this talk, I want to focus on benefits of using these three sources in a more integrated fashion. In particular, I want to reflect on the benefits to social science, twenty-first century statistical systems, and policy analysis that accrue from careful creation and analysis of integrated employer-employee data. When Francis Kramarz and I (1999) first reviewed the creation and use of such data, there were already more than 100 scientific studies using data from 18 different countries. Recent years have seen an explosion of such data for three related reasons. First, the unique ability of these data to address important open questions in a variety of social sciences has put academic researchers at the forefront of their creation. Second, continual improvements in information technology have permitted both improved confidentiality protection and better analysis methods. Third, the marginal cost of creating integrated data is orders of magnitude lower than the marginal cost of censuses and surveys, even after properly accounting for the extra costs associated with enhanced privacy protections.

Why do I put advancing scientific interests at the top of my list? Like many of the statistical innovations of the past (*e.g.,* national income accounts) work on integrated employer-employee data has been driven by academic social scientists addressing substantive open questions. My examples will come from the study of labor markets but others (see Davis and Haltiwanger 1999) consider applications to macroeconomics and industrial organization. Demand and supply is sharp tool that works well in many markets. In the labor market, however, the tool has consistently delivered predictions far sharper than the actual data support. Longitudinally-integrated employer-employee data have permitted the first clean empirical decomposition of the sources of this remarkable labor market heterogeneity—and it is this scientific analysis that has stimulated much of the demand to create such data for additional countries and time periods.

The second item on my list is the creation of twenty-first century statistical systems. Such systems, whether used for official national statistics or for research purposes, will inevitably involve substantial integration of information from multiple sources. The use of sophisticated information technologies and statistical matching methods already pervades national statistical systems. So, what do I mean by a twenty-first century system? The most precise definition is a system in which, *by design*, the information used to produce the desired estimates is collected from the lowest marginal cost sources then integrated using high-powered information technologies and formal probability models. The important distinction with historical systems is the improvement that comes with designing a system to run with integrated information rather than integrating related data *ex post*. My example here will be the Employment Dynamics Estimates system under development at the US Census Bureau (see LEHD Program 2002).

As a part of the discussion of twenty-first century statistical systems, I also want to consider the role that social science researchers can play in the protection of the confidentiality of these data. The research community has, perhaps, not paid sufficient attention to the possibilities inherent in cooperative development of confidentiality protection systems (but see Doyle, Lane, Theeuwes, and Zayatz, 2001 for an important advance in this area). Confidentiality protection has very important implications for the way integrated data are produced and used. When such systems are developed with minimal input from subject-matter specialists, avoidable limitations in the final data product often occur. Because the privacy and confidentiality protection issues associated with integrated employer-employee data have been stressed by all of the agencies that have created them, it is important that the same scientists who helped create the data participate in the design of the confidentiality protection system. This participation will help insure that the fundamental insights that these data permit are preserved in the statistics produced by the confidentiality-protected system (see Abowd and Woodcock, 2001).

The final area where I will argue that integrated social data unlock important information is the analysis of the effects of policy changes on the target community or market. Once again, my examples will come from the labor market but should clearly illustrate how the same ideas can apply to areas like health care and public finance. The innovation permitted by the use of integrated employer-employee data is the possibility of distinguishing more cleanly the effects of incentives on each side of the market. Researchers in France using INSEE's rich archive of integrated data have analyzed the effects of French labor laws on the way firms adjust employment. I will use these analyses (see Abowd, Corbel and Kramarz, 1999; Kramarz and Philippon, 2001; and Crépon and Kramarz, forthcoming) to show how the separate incentives on individuals and employers, which are a part of French labor laws, drive the way in which businesses react to changes in the economy.

2. Untangling the separate influence of people and employers

People bring history and many unobserved skills to the labor market, which appears to recognize and reward these talents. Employers try many strategies to motivate, retain and otherwise compensate employees. To the external observer, it is often hard to tell the difference. Conceptually, longitudinally-integrated employer-employee data permit one to disentangle the separate influence of individual and business heterogeneity as they affect labor market outcomes.

Francis Kramarz, David Margolis and I (1999) first demonstrated this property in work that we did using an integrated system that we developed at INSEE.

High-wage workers or high-wage firms?

We started with the observation that the measurable characteristics of workers explain about 40% of wage variation (Rosen, 1986; Willis 1986). It is important to remember that this base explanatory power (40%) refers to a comprehensive scientific sample of the labor market taken at a point in time. For restricted or specialized samples, the explanatory power of worker characteristics is typically much higher but only because the range of wage variation is much lower in such samples.

How important is the identity of the individual? The importance of "who you are" (unmeasured characteristics of the worker) can be partially captured when we observe the same person at several points in time. Using data constructed for this purpose, one can account for an additional 30% of wage rate variation, beyond the 40% attributed to individual characteristics (Willis, 1986).

How important is the identity of the employer? The importance of "who you work for" (unmeasured characteristics of the employer) can be partially captured when we observe multiple employees of different firms at the same point in time. Using data constructed for this purpose, one can account for an additional 30% of wage rate variation, once again, beyond the 40% attributed to individual characteristics (Groshen, 1991).

What happens when we try to simultaneously account for individual and employer heterogeneity? First, we encounter the fundamental measurement problem that Kramarz, Margolis and I analyzed—namely, in order to untangle the effects associated with these two distinct sources of variation, the data must have a longitudinal dimension for both the employers and the employees. Prior to the creation of longitudinally-integrated employer-employee data, there was no way to address this question properly. We now know, however, that when we simultaneously control for both individual and business heterogeneity, we can explain about 90% of the wage rate variation (see Abowd, Creecy and Kramarz, 2002, for a summary).

And the winner is … *a dead heat*. About half of the unexplained wage rate variation (beyond the variation due to observable characteristics of the employee) is due to "who you are." This individual heterogeneity is the proper basis for distinguishing between high-wage and low-wage workers because the effect is portable—it stays with the worker in movements from firm to firm.

About half of the unexplained wage rate variation (again, beyond the variation due to observable characteristics of the employee) is due to "who you work for." This employer heterogeneity is the proper basis for distinguishing between high-paying and low-paying firms because it stays with the firm regardless of the current pool of employees.

Most remarkably, there is no correlation between the individual (employee) and business (employer) effects when measured at the observation level (an individual-employer-year data point). Significant correlation between the component of pay associated with individual heterogeneity and the part associated with employer heterogeneity does not emerge until the data are aggregated to meaningful economic units like firms, industries or markets. This is the subject of my next example.

What are the high-paying and low-paying sectors?

To make effective use of the decomposition of wage rates into individual and employer effects we have to choose economically meaningful levels of comparison. Although individual and employer components of wage rates are not correlated at the observational level, they are positively correlated when aggregated in economically meaningful ways. Person and firm effects are positively correlated across industries or firm sizes. (see Abowd, Kramarz and Margolis 1999; Abowd, Kramarz, Lengermann, and Roux 2002). Even more striking than the positive correlation between the average person effect and the average firm effect in an industry is our ability decompose the measured wage rate differences between industries into the part that is due to "who you are" and the part that is due to "who you work for" (controlling for the effect of other measurable characteristics).

Tables 1 and 2, which are based on results from Abowd, Kramarz, Lengermann, and Roux (2002), show the big winners (Table 1) and big losers (Table 2) in the inter-industry wage rate decomposition. The first two columns of each table show the standard identifier and name of the industry. The third column is the percentage difference of the industry average wage rate from the average wage rate in the economy as a whole, based on data from a variety of American states from the mid-1980s to the late 1990s and adjusted for observable characteristics of the workers including sex, labor force experience, and location of the job. The column, labeled "Who You Are," is the percentage difference of the average person in the industry from the economy-wide average person. The fifth column, labeled "Who You Work For," is the percentage difference of the average firm in the industry from the economy-wide average firm.[1] All averages are based on individual-year observations.

Interpreting a few of the rows of the winners and losers tables will illustrate the importance of the decomposition. The biggest overall winner is the security, commodity, and broker services industry. This industry paid 79% more than the economy-wide average,

[1] The actual statistical analysis is conducted in logarithms. Tables 1 and 2 convert these logarithms to percentages for clarity of exposition. The logarithmic decomposition of the column labeled "Different from Average" into the columns labeled "Who You Are" and "Who You Work For" is exact (up to sampling error). The percentages shown in the last two columns of the table do not, therefore, add up to the one shown in the third column.

controlling for observable characteristics of the employees. This differential was accomplished by employing individuals who would have earned 32% more than average in any job and paying them an additional 37% more than average. This financial service industry has the largest positive differential associated with employing high-wage workers. By contrast, consider the bituminous coal mining industry. This industry paid 40% more than average over the same period; however, that differential consisted of employing individuals who would have earned 26% less than average in any job and paying those individuals 93% more than average. This coal mining industry has the largest positive differential associated with high-paying employers.

The biggest negative inter-industry wage differential is in eating and drinking places, where wages are 43% below the economy-wide average. In this industry, the wage differential can be decomposed into employing individuals who would have earned 10% less than average at any job and paying those individuals 37% less than average. The restaurant and bar industry has the biggest negative wage rate differential associated with "who you work for." In contrast, the private household sector pays 36% below the economy-wide average. In this industry, the differential is attributed to employing individuals who would have earned 25% less than average in any job and paying these people 16% less than average. The private household sector has the largest negative wage differential associated with "who you are."

I hope that these two examples have shown you that there is substantial scientific advantage to considering labor market questions using integrated employer-employee data. I'll turn now to the development of twenty-first century statistical systems.

2. Building statistical systems for the 21st century

As I mentioned in the introduction, a twenty-first century statistical system is one that is designed to produce the desired estimates by systematically integrating information from multiple sources using the lowest marginal cost source, formal probability modeling for the integration, and comprehensive confidentiality and privacy protection. Since 1998, I have been associated with the US Census Bureau's Longitudinal Employer-Household Dynamics Program, which is attempting to build just such a system by integrating information from state-level unemployment insurance system records and the Bureau's own economic and demographic surveys and censuses (see Abowd, Lane and Prevost, 2000). The program has several projects and I am going to describe two of them.

As a part of its statistical mission, the US Census Bureau supplies about three-fourths of all of the data used to produce the American National Income and Product Accounts, which are the responsibility of the Bureau of Economic Analysis, a sister agency in the Department of Commerce. In support of this data collection effort, the Bureau, through its Economic Censuses and Surveys, collects information about the inputs and outputs of most sectors of the American economy. In virtually every sector, the most important input is labor services. The Bureau finds it monumentally expensive to gather detailed information about the labor input. In addition, the Bureau of Labor Statistics, part of the Department of Labor, has primary statistical responsibility for direct measurements of the US labor force and associated statistics. From a scientific viewpoint, it is quite advantageous to have both input and output measures at the same level of economic activity—ideally, the business establishment. Direct collection of detailed information by the Census Bureau on the labor service component of inputs isn't feasible for the reasons I just gave. The LEHD Program uses its integrated employer-employee data to provide just this information at the business establishment level and without additional data collection costs. I'll describe this human capital project below.

The LEHD Program has also developed a system of Employment Dynamics Estimates (EDE) from its integrated data. Using quarterly unemployment insurance information, and working closely with state-level experts on these data, the Bureau's new EDE system produces quarterly information on job creation, job destructions, accessions, separations, new hires, recalls, and earnings. All of the new estimates, which cover all employment in the unemployment insurance reporting system, can be produced by detailed geography and industry and by age and sex. The US Census Bureau could undertake such a project because the information integrated from the two sides of the labor market is available for use under a specially-created confidentiality protocol that physically separates the identifiers protected by American privacy laws from the identifiers used for the data integration.

Integrating labor market information from both the individual and the employer

Why should we care about measuring the detailed characteristics of an employer's work force; isn't knowing the employment and payroll sufficient? Modern economic methods for accounting for productivity growth in the economy (see, for example, Jorgenson, Gollop and Fraumeni, 1987) emphasize that growth due to substitutions among different types of labor cannot be measured without knowing the productive characteristics of the work force at multiple points in time. If we want to know whether part of the explanation for increased productivity in American business was the employment of more skilled workers, we must measure those skills at the same level of analysis at which we measure productivity. The LEHD Program human capital project attempts to do this by measuring the skill level of every employee of businesses covered by its integrated system (see Abowd, Lengermann and McKinney, 2002, for basic details and Abowd, Haltiwanger, Lane and Sandusky 2001 for the application to productivity measurement). These data permit one to ask the question: Did American firms up-skill in the 1990s, which I illustrate below.

What is human capital? Following Gary Becker's (1964) analysis, human capital is all the productive characteristics that employees bring to the job. Wage rates are proportional to general human capital—compensation for factors that are employed by many firms in the economy. An individual's measured human capital is, then, proportional to the "portable part" of the employee's wage rate (the part due to observable and unobservable individual characteristics) because it is the portable part of the wage rate that reflects the compensation the individual could get from any employer—the "who you are" part of wages. As illustrated in the section on high-wage workers and high-wage employers above, longitudinally-integrated employer-employee data provide the minimum information required to estimate the portable part of an individual's wage rate. The LEHD Program has applied these principles to its integrated data to create a detailed measure of the amount of human capital employed by American businesses at two points in time, corresponding to the 1992 and 1997 Economic Censuses.

American businesses did up-skill in the 1990s according to the LEHD Program's human capital measure. Figures 1-4 (from Abowd, Lengermann and McKinney 2002) illustrate this up-skilling for the retail trade sector and its component industries. A similar picture holds for every major sector of the US economy between 1992 and 1997. For each of these graphs, the horizontal axis is the decile of the economy-wide human capital distribution, as it existed in first quarter 1992. The lowest decile is on the left; the highest is on the right. The vertical axis measures the average proportion of employees in that human capital decile employed at businesses in the indicated industry. Since the 1992 economy is the reference, an industry that employs human capital in a manner that is identical to the economy as a whole will have a perfectly flat profile at 10%, which would mean that it uses workers in each human capital decile in the same proportion as the economy. The figures are all drawn to the same scale so that visual comparisons among them are appropriate.

Figure 1 shows that in 1992 all the 2-digit industries in the retail trade sector had a very low human capital profile that was concentrated in the first and second deciles with very little use at the top decile. Figure 2 shows that by 1997, the retail trade sector had a much flatter profile. Thus, the retail trade sector up-skilled between 1992 and 1997 by employing substantially more human capital—individual in the upper deciles of the human capital distribution. The up-skilling was accomplished by exits of the low-skill intensive firms (compare Figure 3 to Figures 1 and 2), entry of more skill-intensive new firms (Figure 4 to Figures 1 and 2) and up-skilling of the continuers—firms present in both 1992 and 1997 (not shown in a separate figure).

In both 1992 and 1997, the financial services sector is a much more skill-intensive industry than the average industry or the retail trade sector illustrated above. Although the financial services sector involves substantial use of the highest human capital individuals among all 2-digit industries, there was still substantial up-skilling in this sector. The up-skilling of financial services was accomplished by exits of lower-skill firms, entrance of higher-skilled firms and up-skilling of continuing firms, just as in the retail trade sector (not shown in separate figures).

The capacity to measure both the level and change in the entire distribution of human capital employed by American businesses provides the first detailed description of the work force at the level of the business itself that can be followed over time. Thus, the integrated data open up the possibility of detailed productivity analysis at the level of the business itself.

Integrating detailed demographic, geographic and industry variation into the data

The LEHD Program's Employment Dynamics Estimates system provides information about movements of jobs and workers that are more detailed than any other American system. No national statistical system can survive without providing information to the general public. The EDE system was designed to use the integrated employer-employee data to measure worker and job flows and to characterize these flows by demographic, geographic and industrial classifications (see LEHD Program, 2002). I'm going to illustrate how this works using data from the state of California.

The first step in the process is to combine the administrative records of several agencies. The EDE system is based on unemployment insurance records for the participating states. Demographic and worker place of residence information is integrated from the Census Bureau's personal characteristics and place of residence files, which are based on administrative records from a variety of federal agencies. Employer information comes from the state's employment security reporting system. Work history information comes from the unemployment wage record reporting system. The data are integrated using the individual and business identifiers from the wage record reporting (after appropriate editing and privacy protection).

To make the flows, the EDE uses a snapshots of the work force at the boundary of each quarter. The system defines point-in-time employment and full-quarter employment using these snapshots; then, it calculates accessions, separations, job creations, and job destructions using these reference snapshots and the conventional definitions (see Davis and Haltiwanger, 1992; Davis, Haltiwanger and Schuh, 1996; Abowd, Corbel and Kramarz, 1999; and Burgess, Lane and Stevens 2001). All flows and related earnings data are summarized at the business level from the underlying work history microdata. At this level, the geographic and industrial classifications are integrated and the business-level microdata are confidentiality protected using as system that was purpose-built for the EDE. Finally, the system produces worker and job flows by county/industry, sex and age group.

Figure 5 shows an example of worker and job flows for 14-18 year-olds in the state of California. Four series are plotted. Accessions are the number of 14-18 year-olds employed in the indicated quarter who were not employed in the previous quarter. The accessions were either newly hired or recalled into a job with reportable earnings in the unemployment insurance system during the indicated quarter. Separations are the number of 14-18 year-olds employed in the indicated quarter who are not employed in the subsequent quarter. The separations consist of persons who were laid off, terminated, retired, or quit during the indicated quarter. Job creations are the sum over all firms of the increase in total employment of 14-18 year-olds when employment in that age category increases. Job destructions are the sum over all firms of the absolute value of the decrease in total employment of 14-18 year-olds when employment in that category decreases.[2] There is considerable interest in this age group because it forms the lowest target age group of the American Workforce Investment Act. The estimates are produced by sex, eight age categories, every county of participating states, and every major industry group. Future releases will include detailed industry data and county by industry data for each sex and age group.

[2] The definitions of job creations and destructions provided here are simplifications of the actual formulas, which follow the conventions established in Davis and Haltiwanger (1992). That is, the actual job creation and destruction estimates in each firm for each age and sex category are computed using the growth rate of employment in the indicated category (positive growth rates contribute to creations; negative growth rates contribute to destructions) and the average employment in the age and sex category over the quarter.

Protecting privacy and confidentiality while allowing scientific analysis

Integrated employer-employee data present very challenging confidentiality protection issues (see Abowd and Woodcock, 2001, in particular and other essays in Doyle, Lane, Theeuwes, and Zayatz, 2001, for other challenging confidentiality protection issues). The EDE system that I just discussed has its own confidentiality protection system that was designed to permit the release of very detailed estimates without the use of cell suppression, a widely adopted method that would have resulted in the suppression of most of the demographic detail in the EDE system (LEHD Program 2002).

I'm going to discuss another example of confidentiality protection of this type of data—the one that Simon Woodcock and I prototyped for the INSEE data. It is reasonable to ask why a pair of economists would concern themselves with such issues when they already have access to the confidential microdata. The answer is that there is a perpetual tension between the vigor of scientific inquiry, which requires that multiple research teams have access to the same data, and the realities of confidential microdata, which require access in tightly controlled environments. Subject matter specialists are very good judges of the usefulness of a confidentiality protection system because they understand how the underlying data are used in their fields of specialization. This is why I strongly suggest that other social scientists also become more familiar with these methods.

It is a very difficult problem to provide general confidentiality protection of microdata and to preserve their analytic properties. By developing new confidentiality protection methods, we are trying to increase the research access to the data. By developing systems that build directly on the underlying linked microdata, without destroying any of the original confidential data, we hope to provide a layered path from confidentiality-protected public use data to confidential microdata that is as analytically useful as resources permit. The new techniques are based on masking the microdata using the predictive distribution of the confidential data (see Rubin, 1993; Feinberg 1994; and Kennickell, 1997). These techniques preserve many of the analytic features of the integrated data, including moments and complicated multi-factor nonlinear relations, because they are implemented using massive amounts of the information in the underlying, integrated, microdata.

My example concerns the INSEE linked employer-employee data housed at the Cornell Restricted Access Data Center. In this application, we masked individual, employer, and work history characteristics (birth date, education, wage rate, days paid, sales, employment, capital) based on the posterior predictive distribution conditional on all of the linked longitudinal data. The mask reproduces the analytic properties of the data with excellent resolution. Figure 6 provides just one example. The figure shows the log wage-experience profile for men and women estimated from the confidential microdata and the masked microdata. The underlying statistical model includes full individual and employer heterogeneity (see Abowd, Creecy and Kramarz, 2002, for a description of the actual statistical model, which is a two-factor analysis of covariance with fixed individual, employer, and time-varying characteristic effects). There is no substantial difference between the statistical models estimated on the confidential and masked data.

3. Policy analysis using linked data

My final section concerns the use of integrated employer-employee data to study the effects of labor market policies. My French colleagues have made extensive use of INSEE's integrated data for these purposes. (See, for example, Kramarz and Philippon, 2001, and Crépon and Kramarz, forthcoming.) I'm going to focus on work I did with Patrick Corbel and Kramarz (see Abowd, Corbel and Kramarz, 1999). In all of these examples, particular features of French labor law are used to predict the economic incidence of policy changes. Then, the integrated microdata are used to measure the magnitude of those changes. Thus, the data are used to directly measure the effects of the policy on the targeted individuals and employers.

Who is churning through those French jobs? French labor laws created a distinction between fixed-term and indefinite-term employment in 1982. Fixed-term contracts can run up to 24 months (all extensions included). Fixed-term contracts cannot be used to replace indefinite-term employment. Fixed-term contracts have known hiring and separation costs that are not subject to the legal restrictions and economic costs protecting indefinite-term employment.

That's the law but one can reasonably ask "What actually happened?" More than 90% of employment remains on indefinite-term contracts. However, virtually all labor market adjustment now occurs through the use of fixed-term contracts—70% of all accessions, more than half of all separations. Adjustment occurs primarily through changes in the hiring rate, using the fixed-term contract, not primarily through separations. Employers respond to shocks by adjusting the accession of employees on fixed-term contracts.

Figure 7 illustrates how this happens. The figure is based on integrated personnel flow, personal characteristics, and establishment data for France from 1987 to 1990. The horizontal axis shows the establishment growth rate—the annual rate of change of employment at the establishment. The vertical axis shows the accession and separation rates per 100 employees associated with these changes in employment. The separation rate of workers is roughly constant for establishment growth rates that exceed -15%/year, a rate of shrinkage that is extremely unusual in France. In contrast, the accession rates rise monotonically for growth rates exceeding this same level. Because of the very low incidence of establishment size reductions below -15%/year, the figure illustrates that most employment adjustment occurs through the accession rate and the basic statistics confirm that virtually all accessions are into fixed-duration employment contracts.

4. Conclusion

I hope that I have managed to convey my sense that we are just beginning to realize the potential of integrated employer-employee data for understanding labor markets. But it is not just labor markets that will benefit from more extensive integrated social data. Regional growth and transportation specialists can exploit the same systems that I have discussed in my examples by using the integration of place of work (a characteristic of the employer) and place of residence (a characteristic of the individual) to study mobility patterns and time-of-day populations in urban areas. Industrial economists can study the patterns of births, deaths, and complex recombinations among business using both the mobility of employees and of capital assets to measure the reallocations of economic resources. Social network specialists can use the pattern of connections among individuals and employers to measure the extent to which colleagues and employment histories affect productivity and wages (see Lengermann, 2001). Health care specialists can use such data to untangle the complex relation among demographic characteristics of the household and employer-provided health care options (see Stinson, 2001). The examples I have given demonstrate that the scientific value of social data, the quality of official statistics programs, and the protection of the confidentiality of the underlying microdata can all be enhanced by the active participation of the research community in the development of these integrated systems. We need to turn the keys and open the lock.

References

Abowd, John M., Patrick Corbel and Francis Kramarz, "The Entry and Exit of Workers and the Growth of Employment: An Analysis of French Establishments" *Review of Economics and Statistics*, 81(2), (May 1999): 170-187.

Abowd, John M. and Francis Kramarz, "The Analysis of Labor Markets Using Matched Employer-Employee Data," in O. Ashenfelter and D. Card (eds.) *Handbook of Labor Economics*, Volume 3(B), Chapter 40 (Amsterdam: North Holland, 1999), pp. 2629-2710.

Abowd, John M., Francis Kramarz and David Margolis, "High Wage Workers and High Wage Firms," *Econometrica,* 67 (March 1999): 251-333.

Abowd, John M., Julia I. Lane and Ronald C. Prevost, "Design and Conceptual Issues in Realizing Analytical Enhancements through Data Linkages of Employer and Employee Data" in *Proceedings of the Federal Committee on Statistical Methodology*, November 2000.

Abowd, John M., John Haltiwanger, Julia Lane, and Kristin Sandusky "Within and Between Firm Changes in Human Capital, Technology, and Productivity," December 2001.

Abowd, John M. and Simon D. Woodcock, "Disclosure Limitation in Longitudinal Linked Data," in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.), (Amsterdam: North Holland, 2001), pp. 215-277.

Abowd, John M., Robert Creecy and Francis Kramarz, "Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data," March 2002.

Abowd, John M., Francis Kramarz, Paul Lengermann and Sébastien Roux, "Inter-industry and Firm-size Wage Differentials: New Evidence from Linked Employer-Employee Data" 2002, working paper.

Abowd, John M., Paul Lengermann and Kevin McKinney "Measuring the Human Capital Input for American Businesses," January 2002.

Becker, Gary S. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education,* 1st edition, (New York: Columbia University Press, 1964).

Burgess, Simon, Lane, Julia I., and Stevens, David, "Job Flows, Worker Flows, and Churning*,"* *Journal of Labor Economics*, (2000): 473-502.

Crépon, Bruno and Francis Kramarz, "Employed 40 hours or Not-Employed 39: Lessons from the 1982 Workweek Reduction in France," *Journal of Political Economy*, forthcoming.

Davis, Steven J. and John Haltiwanger "Gross Job Creation and Destruction: Microeconomic Evidence and Macroeconomic Implications," *NBER Macroeconomics Annual 1990*, O. Blanchard and S. Fischer, eds. (Cambridge: MIT Press, 1990), pp. 123-68.

Davis, Steven J. and John Haltiwanger "Gross Job Creation, Gross Job Destruction and Employment Reallocation," *Quarterly Journal of Economics* 107 (1992): 819-63.

Davis, Steven J. and John Haltiwanger, (1999). "Gross Job Creation and Destruction," in O. Ashenfelter and D. Card (eds), *Handbook of Labor Economics*, Volume 3(B), Chapter 41 (Amsterdam: North-Holland, 1999), pp. 2711-2805.

Davis, Steven J., Haltiwanger, John C. and Schuh, Scott, *Job Creation and Destruction*, (Cambridge, MA: MIT Press, 1996).

Doyle, Patricial, Julia Lane, Jules Theeuwes, and Laura Zayatz, "Introduction," in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.), (Amsterdam: North Holland, 2001), pp. 15.

Fienberg, Stephen. E., "A Radical Proposal for the Provision of Micro-data Samples and the Preservation of Confidentiality," Carnegie Mellon University Department of Statistics Technical Report, No. 611, 1994.

Groshen, Erica, "Sources of Intra-Industry Wage Dispersion: How Much Do Employers Matter?" *Quarterly Journal of Economics*, 106, (1991): 869-884.

Jorgenson, Dale W., F. Gollop and Barbara M. Fraumeni, *Productivity and US Economic Growth*. (Cambridge, MA: Harvard University Press,.1987)

Kennickell, Arthur. B. "Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances," SCF Working Paper, 1997.

Kramarz, Francis and Thomas Philippon, "The Impact of Differential Payroll Tax Subsidies on Minimum Wage Employment," *Journal of Public Economics*, 82, (2001): 115-146.

Lengermann, Paul, "Is it Who You Are, Where You Work or With Whom You Work: Reassessing the Relationship between Skill Segregation and Wage Inequality," working paper, November 2001.

Longitudinal Employer-Household Dynamics Program, Employment Dynamics Estimates Project Versions 2.2 and 2.3, US Census Bureau, LEHD Program, Technical Working Paper TP 2002-05 rev1 (May 2002).

Rubin, Donald. B. "Discussion of Statistical Disclosure Limitation," *Journal of Official Statistics* 9 (1993): 461-468.

Rosen, Sherwin, “The Theory of Equalizing Differences” in *Handbook of Labor Economics*, in O. Ashenfelter and R. Layard (eds.), (Amsterdam: North Holland, 1986), pp. 641-692.

Stinson, Marth H, “ Estimating the Relationship between Employer-Provided Health Insurance, Worker Mobility, and Wages,” November 2001 working paper.

Willis, Robert, “Wage Determinants: A Survey” in *Handbook of Labor Economics*, in O. Ashenfelter and R. Layard (eds.), (Amsterdam: North Holland, 1986), pp. 525-602.

The Big Winners

SIC	Sector	Different From Average	Who You Are	Who You Work For
62	Security, commodity, brokers and services	79%	32%	37%
67	Holding and other investments	61%	29%	26%
46	Pipelines, except natural gas	57%	4%	54%
48	Communication	55%	3%	50%
49	Electric, gas and sanitary services	50%	2%	48%
28	Chemicals and allied products	47%	11%	34%
81	Legal services	43%	14%	26%
12	Bituminous coal mining	40%	-26%	93%
61	Credit agencies other than banks	36%	14%	20%
63	Insurance carriers	35%	10%	24%
37	Transportation equipment	30%	-1%	32%
87	Engineering, accounting, research services	29%	9%	18%
38	Instruments and related products	27%	6%	21%

Source: Abowd, Kramarz, Lengermann and Roux (2002)

Table 1

The Big Losers

SIC	Sector	Different From Average	Who You Are	Who You Work For
58	Eating and drinking places	-43%	-10%	-37%
88	Private households	-36%	-25%	-16%
1	Agriculture-crops	-34%	-12%	-26%
79	Amusement and recreation services	-32%	-7%	-28%
72	Personal services	-32%	-11%	-24%
70	Hotel and lodging services	-30%	-14%	-20%
78	Motion pictures	-29%	-7%	-25%
53	General merchandise stores	-26%	-3%	-24%
2	Agriculture-livestock	-25%	-13%	-14%
56	Apparel and accessory stores	-25%	1%	-26%
83	Social services	-24%	-14%	-12%
59	Miscellaneous retail	-24%	-1%	-24%
54	Food stores	-24%	1%	-25%

Source: Abowd, Kramarz, Lengermann and Roux (2002)

Table 2

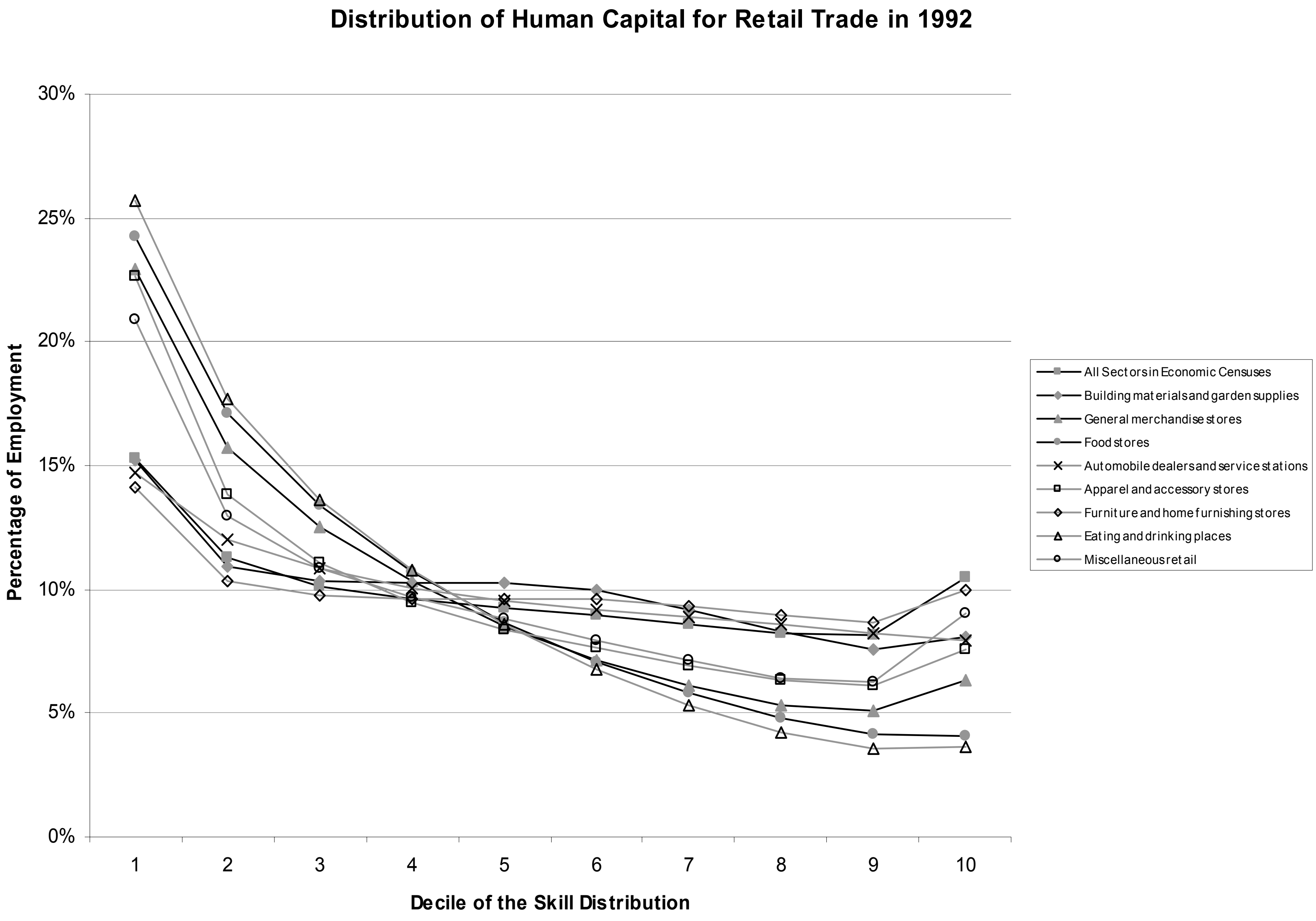


Figure 1

(Source: Abowd, Lengermann and McKinney, 2002)

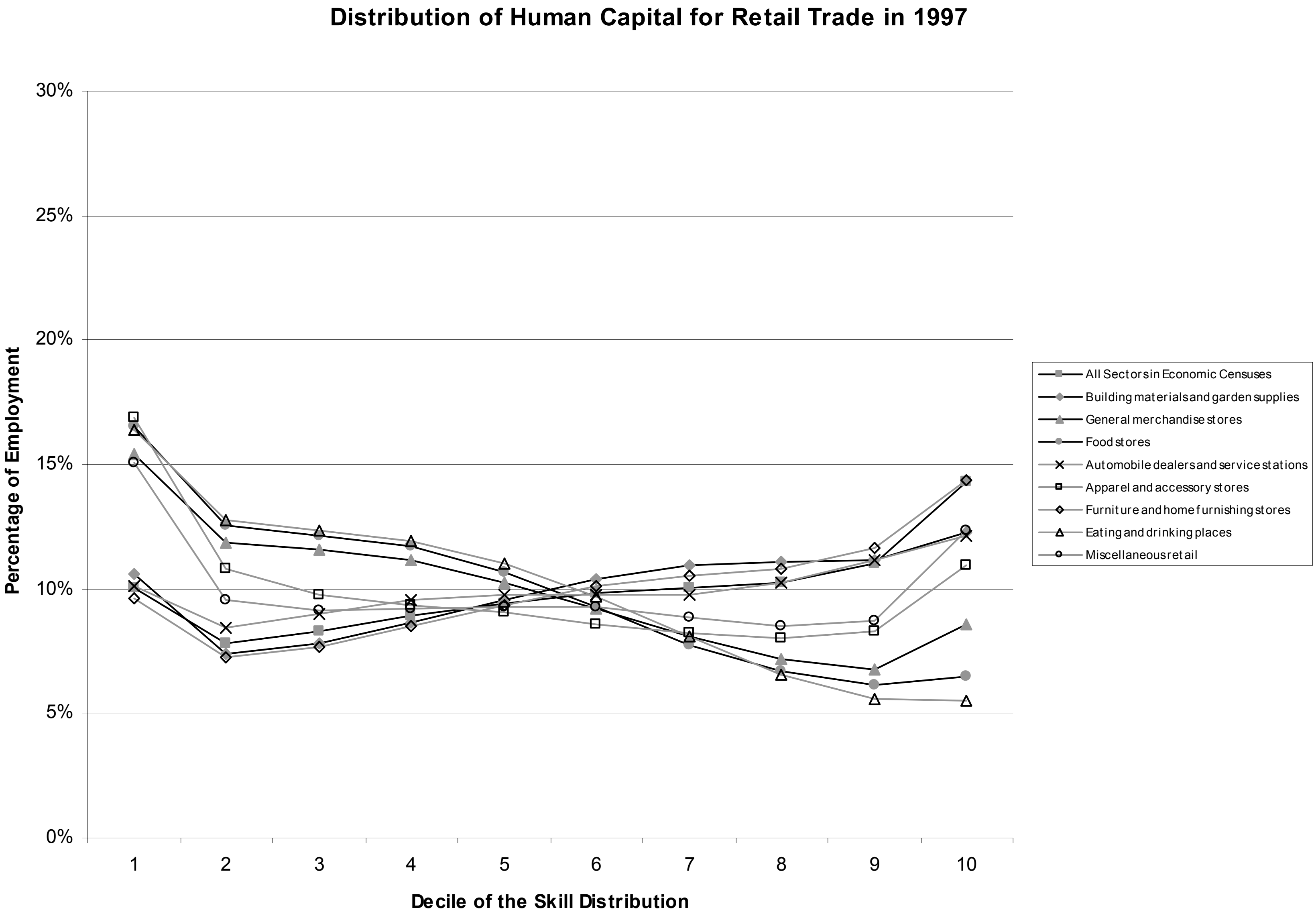


Figure 2
(Source: Abowd, Lengermann and McKinney, 2002)

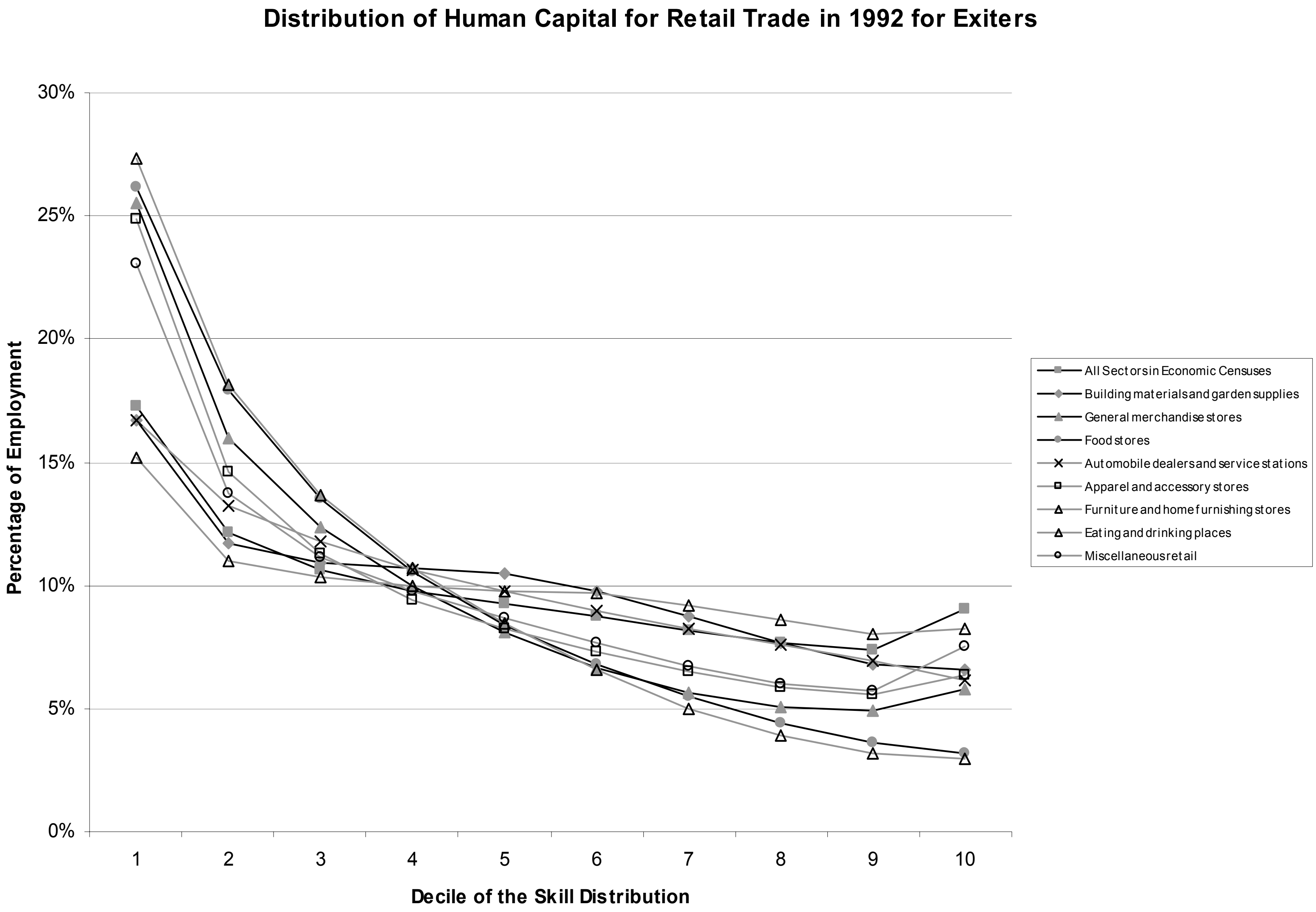


Figure 3
(Source: Abowd, Lengermann and McKinney, 2002)

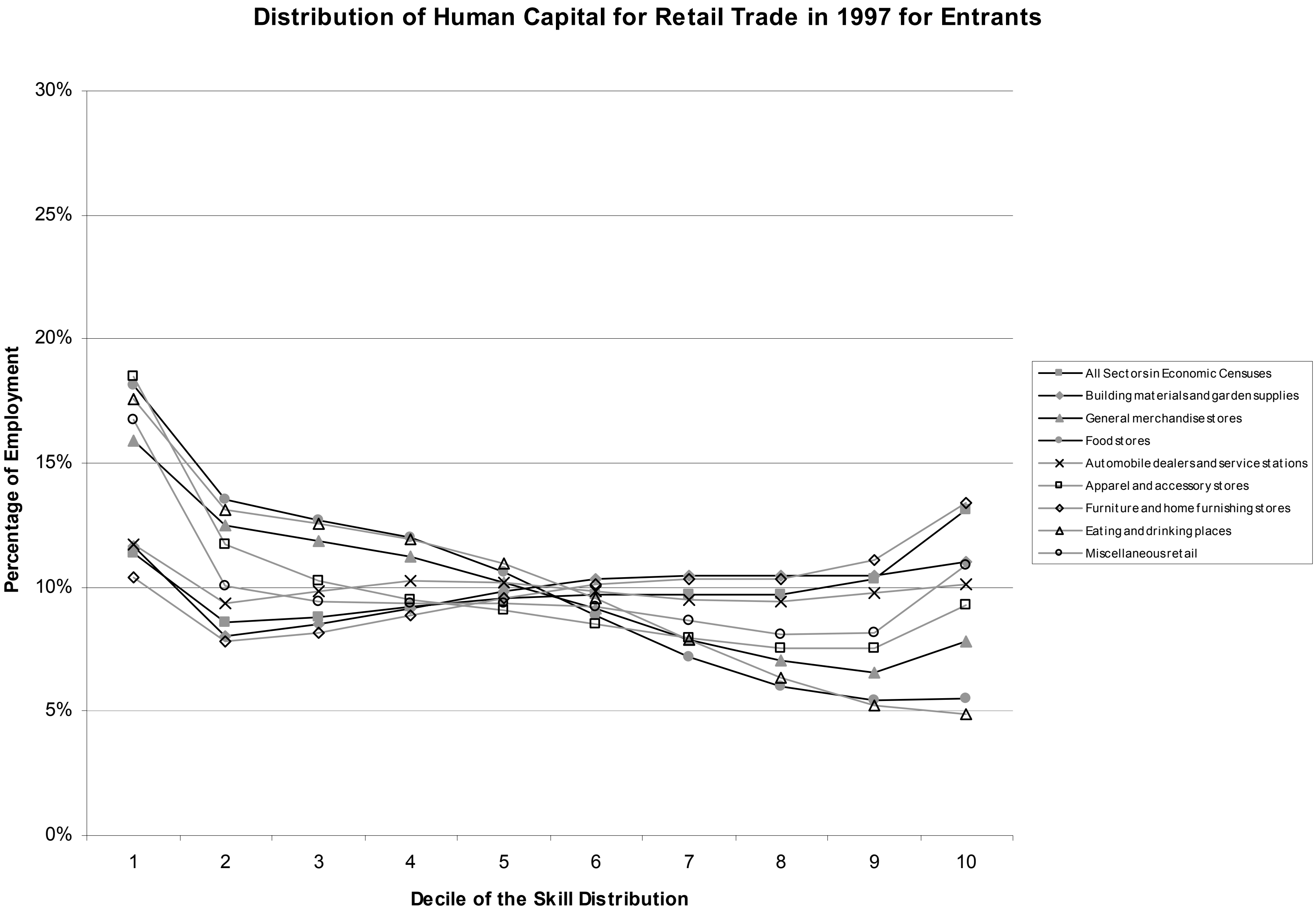


Figure 4
(Source: Abowd, Lengermann and McKinney, 2002)

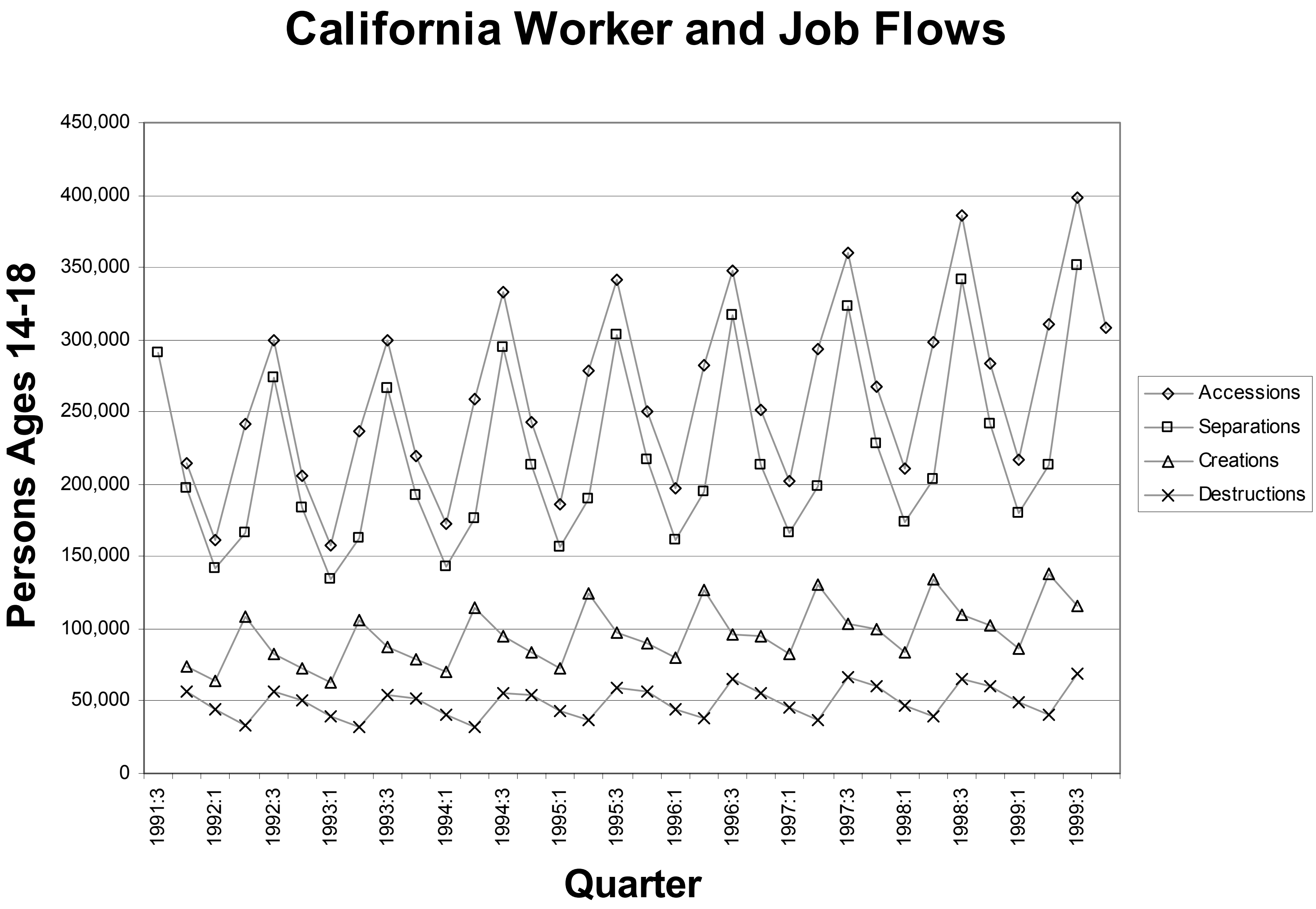
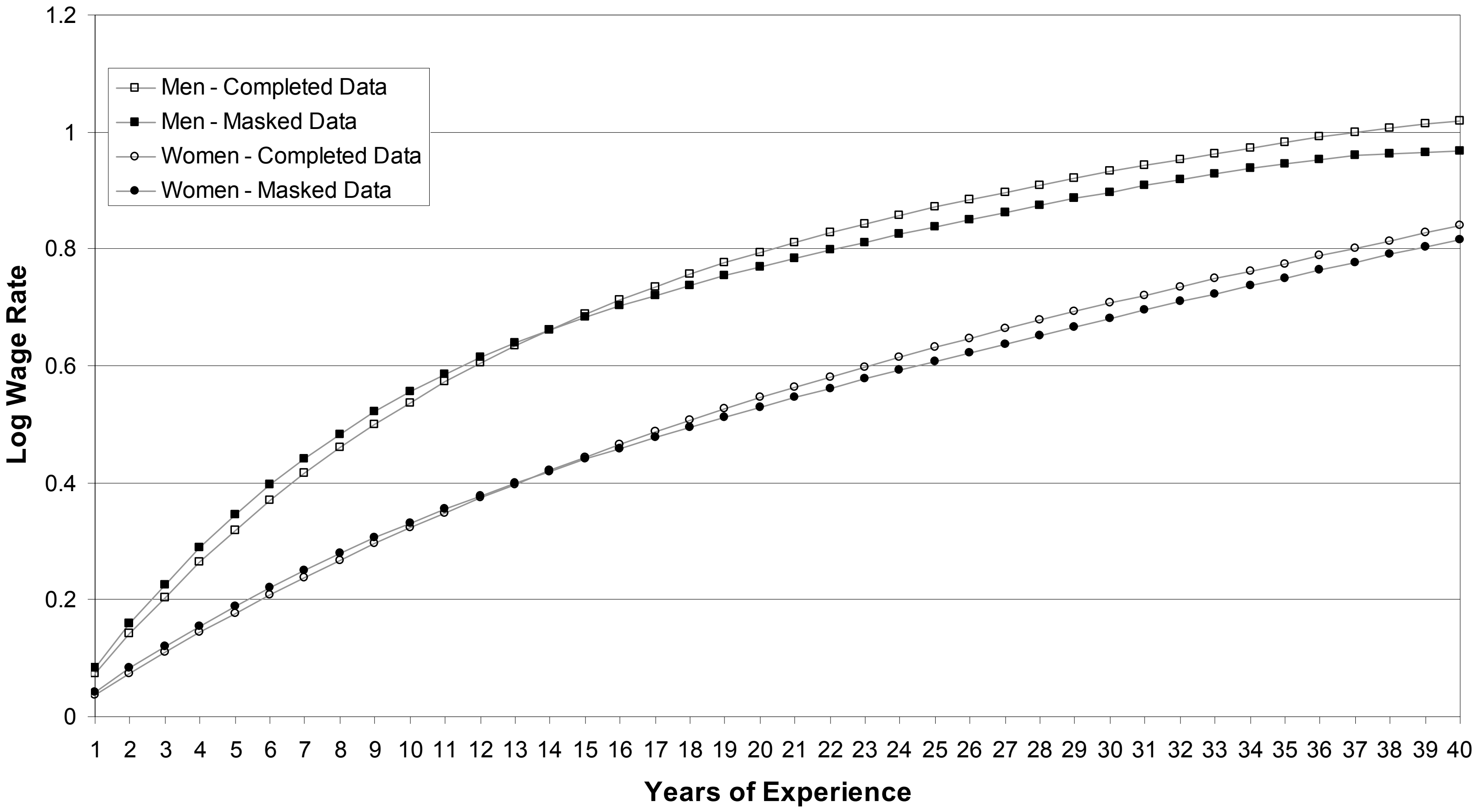


Figure 5
(Source: LEHD Program, 2002)



Experience Profile in Wage Regression With Fixed Worker and Firm Effects

Figure 6
(Source: Abowd and Woodcock, 2001)

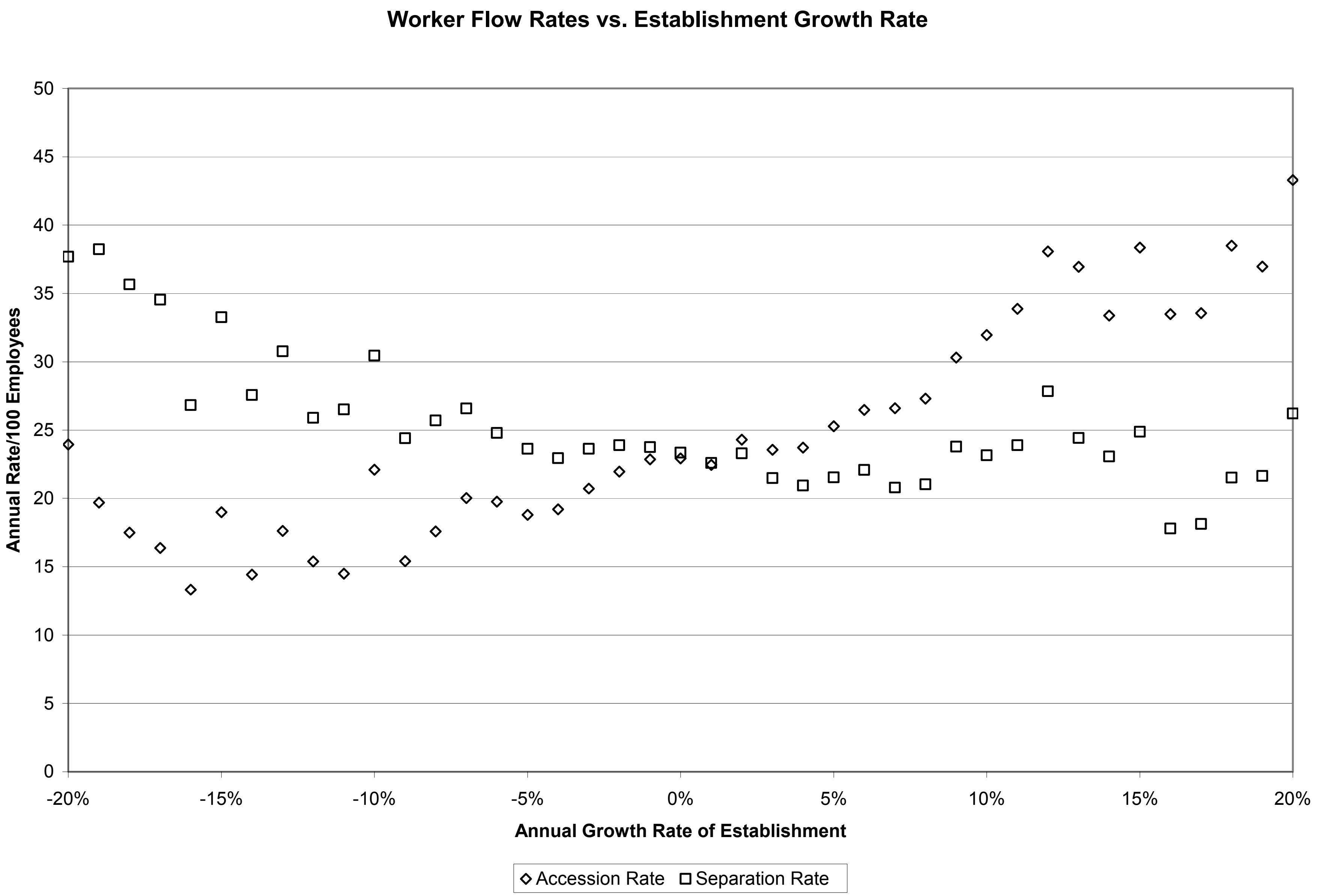


Figure 7
(Source: Abowd, Corbel and Kramarz, 1999)