ORIGINAL ARTICLE

Interview

An Interview with John M. Abowd

Ian Schmutte[1] | Lars Vilhuber[2]

[1] University of Georgia

[2] Cornell University

Correspondence
Ian Schmutte, Department of Economics,
Athens, GA, USA
Email: schmutte@uga.edu

Funding information
No funding.

The final article can be found at
http://doi.org/10.1111/insr.12489.

John M. Abowd is the Chief Scientist and Associate Director for Research and Methodology, U.S. Census Bureau. He completed his A.B. in Economics at Notre Dame in 1973 and his Ph.D. in Economics at University of Chicago in 1977 under Arnold Zellner. During his academic career, John has held faculty positions at Princeton, the University of Chicago, and, since 1987 at Cornell University where he is the Edmund Ezra Day Professor Emeritus of Economics, Statistics and Data Science. John was trained as a statistician and labor economist, and his economic research has focused on the rigorous empirical evaluation of labor market institutions. In the late 1990s, he began working with the Census Bureau on projects that would end up leveraging administrative and survey records into official statistical products. Through that work, he has developed a research agenda focused on issues necessary to generate those products, including data privacy, synthetic data, total error analysis, data linkage, missing data problems, among others.

1 | INTRODUCTION

John Abowd (JA): (singing) “*I am just a poor boy though my story’s seldom told. I have squandered my existence for a pocket full of promises.*” No... “*pocket full of crumbles. Such are promises.*” There we go. That was almost right. (laughter)

Lars Vilhuber (LV): *Why don’t we start with that?*

JA: The one I did yesterday was "These are all my trials and tribulations...." Of course, they always tease me that I can never get the lyrics right.

LV: *Well, this is us, Ian Schmutte and Lars Vilhuber, attempting to conduct an interview with John Abowd, whom we have known for a long time. Usually we wouldn’t do this on video, but it’s still 2021. So we can’t accompany this with our habitual dram.*

Ian Schmutte (IS): *John, would you mind telling us something about your youth and how you grew up?*

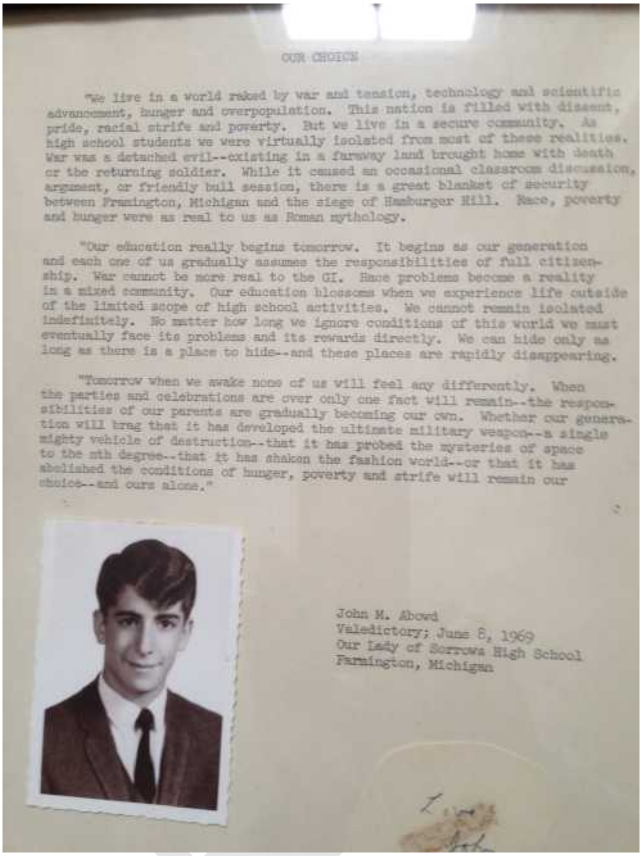
JA: I’m the oldest of 12. I grew up with an automotive engineer father and a 100 percent, non-market economy mother, who had plenty to do. I told Janet, my wife, and my children that before the first one of them was born, my lifetime diapers-changed already exceeded theirs. It was going to exceed theirs forever. My mother spread the tasks out, and she did not discriminate by sex. So all of us learned to cook. All of us learned to clean, all of us learned to fold diapers, wash the old cloth ones. She was an early adopter of paper diapers. We were a relatively insular family. When I was young, it must have been third grade, we moved from the city of Detroit out to the suburbs – Farmington before there were Hills.

The key thing about my youth was, while it didn’t seem particularly regimented to us, there were regular times for everything. So everybody was ready to go to school at the same time, because we all got on exactly the same bus which stopped outside our house and picked up as many of us as there were to pick up and dropped us all off at the same school, which went first grade through high school. That didn’t end until I got my driver’s license at age 16. My mother personally enrolled me in driver’s education classes when I was 15 and a half – she drove me to the Department of Motor Vehicles to get my license so that it was issued on my 16th birthday on December 22. I was driving the family station wagon as the chauffeur from that day forward, a job that passed to the oldest child still living at home until the youngest, my sister Paula, graduated from high school. She’s 16 years younger than I am. And yes, I did ding the car on the first day.

The Abowd family, ca. 1976. John is 2nd from left, top row.

IS: *Were you always drawn to economics?*

JA: "No," is the answer to that question. I had no idea which subjects I was really good at and which ones I was just relatively good at. I went to a small Catholic parish high school, taught primarily by dedicated Dominican sisters and longtime lay faculty, and so had very limited course offerings. I took all the same course offerings as everybody else in my class. The one science class that freshmen took, I took. The one science class that sophomores took, I took. The math classes that we all took, I took. We had regular study times at home. It was not difficult to keep up, and it was not difficult to get way ahead.



OUR CHOICE

"We live in a world raked by war and tension, technology and scientific advancement, hunger and overpopulation. This nation is filled with dissent, pride, racial strife and poverty. But we live in a secure community. As high school students we were virtually isolated from most of these realities. War was a detached evil--existing in a faraway land brought home with death or the returning soldier. While it caused an occasional classroom discussion, argument, or friendly bull session, there is a great blanket of security between Framington, Michigan and the siege of Hamburger Hill. Race, poverty and hunger were as real to us as Roman mythology.

"Our education really begins tomorrow. It begins as our generation and each one of us gradually assumes the responsibilities of full citizenship. War cannot be more real to the GI. Race problems become a reality in a mixed community. Our education blossoms when we experience life outside of the limited scope of high school activities. We cannot remain isolated indefinitely. No matter how long we ignore conditions of this world we must eventually face its problems and its rewards directly. We can hide only as long as there is a place to hide--and these places are rapidly disappearing.

"Tomorrow when we awake none of us will feel any differently. When the parties and celebrations are over only one fact will remain--the responsibilities of our parents are gradually becoming our own. Whether our generation will brag that it has developed the ultimate military weapon--a single mighty vehicle of destruction--that it has probed the mysteries of space to the nth degree--that it has shaken the fashion world--or that it has abolished the conditions of hunger, poverty and strife will remain our choice--and ours alone."

John M. Abowd
Valedictory; June 8, 1969
Our Lady of Sorrows High School
Farmington, Michigan

John and his valedictory address in 1969.

But we were all encouraged to do things besides just the classes. My dad was a frugal guy. He wanted us to do sports, but he had done the calculation that his brood of – I'm not the tallest, but I'm close to the tallest *[ed. note: John is approximately 5'7"]* – relatively short guys and genuinely petite girls, were probably not going to be basketball stars. The only thing we did was essentially any high value added extracurricular activity – so I did high school debate. We were very good at that. There were teams from all the big suburban high schools, which had professional debate coaches, and professional forensic coaches. We beat most of them too, so when I got into economics, I watched people like Larry Summers have a go at their opponents in a policy debate. He was a "second negative", which, if you don't know anything about debate, is the one that gets to do all the cleanup of taking down the other team's argument.

2 | UNDERGRADUATE YEARS

IS: *So how did you end up majoring in economics?*

JA: I made the mistake of telling my mother I was going to go to law school before going to college. But when I got to Notre Dame – they still have this organization, but it was particularly salient when I was there – you didn't declare a major or even a college in your freshman year. So I took honors math, but I didn't take the chemistry course that all the pre-meds took – in my year, people who wanted to be wealthy still were in pre-med, not pre-business or business – that changed about a decade and a half later. In any event, I didn't experience the quintessential Notre Dame freshmen course, which was taught for at least forty years by Emil T. Hoffman.[1] Everybody around me, about half of them were pre-med, and they were all taking this. So I got to focus on the writing classes, which were mandatory, and I took a biology class, which was fine. Very well taught, completely uninspiring, so there was basically no chance that I was going to go into the sciences.

I took Principles of Economics in my sophomore year, and I don't think I ever got an answer wrong in an undergraduate economics class at Notre Dame. It's not because they weren't challenging, it was because they were, except for Principles, taught in a small class format. They were taught by people who had been teaching the subject long enough to be good at it, but not so long as to be out-of-date. More importantly, they were taught before the Notre Dame economics department was captured by "heterodox" economists. Yeah, and so it was just inspiring.

In the class ahead of me was Joe Hotz.[2] We knew each other – not particularly well, but we knew each other – as undergraduates. When he decided to go off to Chicago for graduate school, I was watching when he came back after deciding to transfer to Wisconsin. We had a long talk, and he told me what it was like to go to Chicago. The famous Don McCloskey[3] speech, "now look to your left; look to your right. Neither of those guys will be here at the end of the year." You had a 40% cumulative chance of passing the core exam, and you got three tries. Some pretty famous people in economics – James Heckman[4] and Sandy Grossman[5] leap right to mind – encountered that exam, as well.

I ignored Joe's advice and went to Chicago anyway. It wasn't an unreasonable thing to do. I applied to Harvard, MIT, and Yale, I didn't get in. In the end, I had to choose between Chicago and Penn. Those are two pretty good programs, and at the time, the lead econometrician at Penn was Mark Nerlove[6] and the lead econometricians at Chicago were Arnold Zellner[7] and Henri Theil. So I basically looked at what each of them did, and thought that I would enjoy doing econometrics with Zellner more than I'd enjoy doing it with Nerlove, which is kind of ironic, because Nerlove was at the time doing labor economics, and Arnold was primarily doing Bayesian econometrics system estimators. But it was the right choice, given my choice set. At that point, I had no further desire to go to law school. So that's how it happened.

IS: *While you were at Notre Dame you were involved in evaluating the negative income tax experiments. Can you talk about that?*

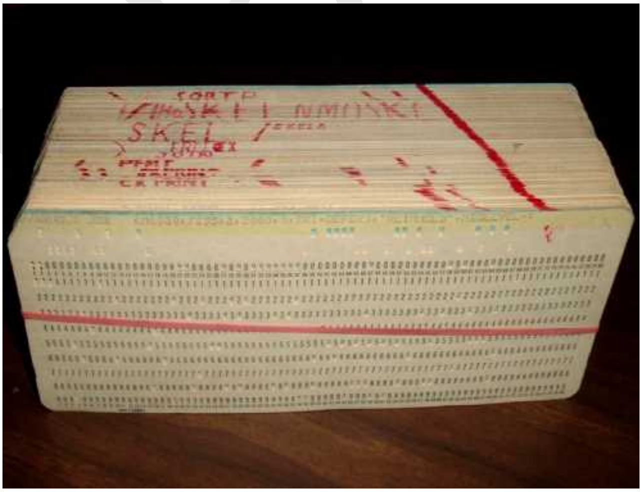
JA: Sure, I can tell that story. I had a National Merit Scholarship to go to college, and that basically leveled the price of Notre Dame and the University of Michigan. They were going to cost exactly the same. Those were my two options. So the National Merit Scholarship covered most of my part. We were expected to work during the summer. The spending allowance was whatever you earned, including whatever you earned at school, so I had a series of jobs at school too.

LV: *And the negative income tax experiment was one of those?*

JA: Yes. I worked before my freshman year, between my freshman and sophomore year, and then between my sophomore and junior years. I worked summer jobs at a Ford dealership that my uncle owned. But I didn't come home for the summer of my junior year. I got a job as a research assistant in the Gary Income Maintenance Experiment [11, 10]. So the economists in our audience will possibly remember that there were a whole series of these. The best known is the New Jersey one, because it was the first one launched. The longest running one was the Seattle-Denver one. The one in Gary is where I met Gary Burtless,[8] who was a Yale undergraduate at the time.

The baseline survey had been taken. These things had huge budgets, but the budgets included the transfer payments. So if I'm remembering right, the single year budget for the Gary income maintenance experiment was about $17 million. But that included the money for the transfer payments to the participants. They were designed as proper randomized control trials. But even before the era of IRBs,[9] the ethical rules underlying them said that if somebody was entitled to more than their treatment effect, they got the max of what they were entitled to under the existing welfare system in their treatment. So, many of the analysis plans were complicated by that factor. But there was no analysis yet because at the baseline, no treatment had been applied.

Our job was to load the questionnaire from the baseline survey and code it all up. Our tool was SPSS – the real SPSS back in the day. The first 15 characters in the 80-column key punch card were the keyword, starting in column 16 were the keyword specific parameters, and columns 73 to 80 were the optional – but if you didn't use them, you were a fool – sequential numberings for your physical card deck. I would program an IBM card punch so that I could type the SPSS commands.



A deck of punched cards comprising a computer program.
Photo by Arnold Reinhold - CC BY-SA 3.0,
`https://commons.wikimedia.org/w/index.php?curid=16041053`

We had a competition to see who could deliver the thickest SPSS printout to his or her supervisor who could then deposit it at the central facility. I don't know whether I won or not, but I do know that Frank Jones buried a 20 dollar bill in the pages of one of my SPSS printouts and said, "I'm betting nobody ever gets this." So you know, there's only so much data cleaning that you can reasonably do. When we were all done doing it, then we printed a codebook. In those days, a codebook was printed on 11 by 14 fanfold, bound up, and stuck on a shelf.

So that's what I did. It was a great job. I had a great time.

John and Janet at her graduation from Notre Dame, 1974.

3 | GRADUATE STUDIES

IS: *Do you think that early experience with data collection and processing shaped your career or any of your approach in grad school?*

JA: Oh, yeah. I knew I didn't want to do theory. I speculate that I might not have been very good at it. But that's neither here nor there. As much as we kid each other about the steps we took when we did statistical programming – doing statistical programming, and being the first one to dive into a data set, especially one being generated by a controlled experiment, that's an exhilarating experience. I think a lot of people got hooked on empirical economics from that experience, trying to glean new knowledge. Our techniques are both more sophisticated and more careful about what the null hypothesis is now than they were perhaps when I entered the profession, but you know ... we're kidding ourselves if we think we're not doing adaptive data analysis. We're basically always doing adaptive data analysis, and we're learning from it. It's an appropriate thing to do in almost every circumstance, it just has to be appropriately documented so that you can understand what is a reasonable and unreasonable conclusion to draw from that particular data analysis.

IS: *So you already knew that you wanted to study econometrics, when you were making your decision about graduate school? Was that unusual at the time?*

JA: So even you, Ian, who entered the profession several decades later, had a very different experience and recruitment to graduate school. To say that the University of Chicago did not recruit graduate students is to trivialize their active non-recruitment of graduate students. They would admit just about anyone who met a set of reasonable criteria and bought option value. That's what they bought, and they made it crystal clear when you showed up.

I took the first year econometrics sequence at the same time as most of the others. So I got to know Arnold Zellner, who taught in the sequence, and I got to know his RA at the time, Walter Vandaele. Walter graduated at the end of my first year, so Arnold had an opening and invited me to take it. So I became Zellner's TA/RA.[10] It was mostly RA work, except that when he was teaching, I was in charge of the problem sets. That was fine for the first two quarters. The third course in the sequence was Bayesian econometrics. I think that modern statisticians are not introduced to this the way Arnold introduced his students to it. I can tell you that I finally had to learn how to change variables underneath the integral like there was no tomorrow. Teaching people how to answer the problems in Arnold's textbook – which never had a second edition, but it didn't really need one – And it didn't have published answers either – but the answers of the TA became extremely useful for the people taking that class.

Agnes and Arnold Zellner at John and Janet's wedding (1979)

Another thing that I think most people who didn't know Arnold Zellner well didn't appreciate is he never wasn't working. He would go home and have dinner, sit down at his desk after dinner, and open his correspondence from applied mathematicians all around the world, all of whom he had sent particular indefinite integrals, asking them, did they have any tricks for doing this integral? He'd look at their tricks and try them. You had to be extremely judicious in your time management when you went to talk to him, because you got the first half hour of his attention. Then he would go to the board and start showing you the latest integrals he was working on, basically working his way through them again and making sure he wasn't missing something and have you watch him too. That would go on until he quit at five o'clock. So if you went at 2:30, you were there until five o'clock, and if you went at 4:30 you were there until five o'clock.

LV: *Growing up as one of twelve siblings certainly involved a lot of management and organization of different people doing different things. Did that play a role in getting you interested in labor organization?*

JA: Ah, no, I can't honestly say that it did. But, you know, I was at Chicago for my PhD in the heyday of Gary Becker[11] and family economics. All of his famous papers were either just being written or were appearing. I was on a fellowship sponsored by T.W. Schultz, the father of modern agricultural economics and human capital theory.[12] As a condition of that fellowship you had to go to one workshop from your very first day of graduate school. So I was in Gary Becker's workshop. The graduate students – I wasn't the only one in this boat – dutifully sat in the second row, and the faculty and advanced graduate students sat around the table.

That was one of the famous Chicago workshops. Not only were you expected to read the paper in advance, but if you had not read the paper in advance, you were just as likely to get called out by Becker as the speaker. So it basically started with Gary asking the first question about some random paragraph in the paper and then proceeding to a discussion. In my entire time at Chicago, in graduate school and on the faculty in the business school, about 11 years, I only saw one person actually give a paper in Gary Becker's workshop. It was Angus Deaton.[13] He showed up, had asked ahead that there be an overhead projector and a screen. He proceeded to give his paper – it was one of his papers on the econometrics of consumption equation systems. And he was permitted to do that. I was just stunned. I'd never seen it before. He was certainly not the first famous person to give a talk in Gary Becker's workshop, but he was the first seated editor of Econometrica to give a paper there. And I'm sure that's why Gary let him present. Not because Gary cared, but because some of the younger assistant professors, one of whom was James Heckman, probably said "You really shouldn't set this guy off!"

IS: *Who were your primary influences, and what was the perception of econometrics and labor at Chicago at that time.*

JA: I never stress enough that Jim Heckman, and Eddie Lazear[14] were assistant professors, when I was a graduate student. They were at the start of their careers. My active choice of Arnold [Zellner] occurred before I went to graduate school and nothing happened in my first year to dissuade me from that. The salient alternatives at the time were some of the statisticians in the business school. I didn't know many of the statisticians in the statistics department, and at the time, the statistics department and the business school didn't have as close an association as they do now.

Now I credit Zellner with my awareness that it was partially our responsibility to speak a version of the statistics language understandable to statisticians and econometricians all the time. I used to jokingly refer to it as translating. I wasn't always true to it, I have to say. When I first got into doing the work with Francis Kramarz[15] modeling earnings in linked employer-employee data, I was speaking entirely the language that econometricians speak, and almost completely unaware of the parallel development in Agricultural Statistics that I should have known about. For many years, including when he was president of the American Statistical Association, Arnold organized a twice-yearly meeting of Bayesian statisticians that still occurs. His graduate student, that would be me at the time, was the scribe for those. The scribe was expected to produce a detailed transcript of the discussion, which basically amounts to the referee's report and the editor's report and the author's reply for every single paper that was presented.

Some really famous papers were presented there for the first time! The EM algorithm [6], Rubin's work on multiple imputation, all of Ed Leamer's early work.[16] It was considered the forum where statisticians presented their Bayesian work. That's where I first met Steve Fienberg.[17] They considered themselves a close knit, scientific group that strongly believed that the common ground of statistics was not a war between Bayesians and frequentists, but a method of doing statistical reasoning that could accommodate randomness, wherever you needed to put it in order to construct the argument.

4 | EARLY CAREER

LV: *Once you finished your PhD at Chicago, what were the next steps?*

JA: That's an interesting story. I think. I did not intend to go on the job market in my third year of graduate school at Chicago, but when your thesis supervisor says you're ready for the job market, you don't really have a choice. I took Arnold's advice, and I cobbled up a job market paper. This will really speak to young economists: in that era, it was not unusual to cobble up a job market paper and head off to give your first job talk with no practice whatsoever.

So I headed off to give my first job talk at – well, you know, you might as well have a throwaway – MIT. It was in the era when it was expected if you were on the faculty of some place like that to make sure that the graduate student coming to interview understood their position in the world! Bob Solow[18] and Paul Samuelson[19] were both at lunch. Lunch was in the faculty club. I was offered beer or wine with lunch, but I knew better. So I passed the first test. I took the questions at lunch with reasonable aplomb. I presented my work with reasonable coherence.

My first in-person interview at my first fly out at MIT was with Michael Piore.[20] And the first words out of his mouth were "I hate Chicago economists." And the conversation didn't go downhill from there, because it had no place to go from there. So I didn't get an offer from MIT. But I wasn't particularly surprised by that. In that era, it was unusual for MIT to hire a Chicago PhD. They mostly hired Harvard PhDs and Harvard mostly hired MIT PhDs, and the occasional Princeton or Stanford person. Just a few years later, that disappeared, but there was still a pretty vitriolic war going on in macroeconomics in 1976. I wasn't participating in it, but it was hard not to be collateral damage in that war.

I did have a fly-out that went pretty well at Princeton. At Princeton, the Chicago economist on the faculty was Gregory Chow.[21] He opened the workshop with "This is not Chicago. At Princeton, we present our papers," and so I presented my paper and I was reasonably well received. I got an offer. I think Orley Ashenfelter[22] saw in me the possibility to rescue me from econometrics and get me to do real labor economics. That was his plan. The competing offer was from Carnegie Mellon. It was an attractive offer, too – considerably higher salary – Princeton famously bragged about having the lowest assistant professor salary in that era. I think it was Gale Johnson,[23] at Chicago, who told me "What are you worried about the salary for? That's not important in your first job." It was, of course, completely correct advice. The chance to work with Orley was exactly what I should have been evaluating and that is what I did.

John and Orley Ashenfelter, at a somewhat older stage of their lives

So I went there and he [Ashenfelter] was on sabbatical the first year I was there. I joined a cohort of assistant professors that included Roger Gordon,[24] Mark Gersovitz,[25] and Claudia Goldin.[26] We all arrived without our PhDs done, because that was absolutely standard operating procedure in the late 60s and early 70s in economics. The market was so hot that you didn't have to have your PhD done. You went out when your advisor said to, you were connected through a network.

So all of us were in the same boat. At the time, Princeton had a Dean of Faculty whose name was Aaron Lemonick.[27] I'm never gonna forget that name. There was a rule known as Lemonick's Rule. You had one year to finish your PhD, or your assistant professor appointment was converted to a term appointment, and then you had to leave. So I diligently worked on my PhD, just like Roger Gordon and the others did. We all finished up in the summer between our first year at Princeton and our second year. I had my thesis advisor Arnold Zellner sign off on the draft in late August. The others went back to their institutions and defended theirs in the late summer, but Chicago doesn't meet in August – the next quarter was October. So I scheduled my thesis defense for December, which was when Arnold wanted me to do it. You do your thesis defense as a live lecture in Chicago, I didn't think any more of it.

And then Lemonick fired me.

John at his University of Chicago office (c. 1978)

Orley got back, and said, “What happened?” I said, “Well, I finished my thesis just like these other guys, but I couldn’t defend it until December, and had I known that that was going to be determinative, I would have done something else. But oh, by the way, I now have an offer from Chicago. So really, you don’t need to feel bad for me.” So I went to the business school at Chicago [in 1978] and never looked back. You know, Orley and I collaborated, we wrote a nice paper during that second year, and I met Hank Farber,[28] and we collaborated and have been close professional friends ever since.

But I did become a “Princeton labor economist”, in marked distinction to “Chicago labor economist”. That’s entirely the influence of Orley and later, David Card.[29] They did have a very different way of addressing the same basic questions. If you were interested in doing family economics with Orley, that was fine, go ahead and do it. But he was interested in doing labor economics the way it was done at Princeton in the heyday of the Industrial Relations section,[30] which I include his reign as the head of that. It was about institutional labor economics done with proper economic toolkit and proper econometric rigor. I have enjoyed that my entire career and I still enjoy that when I get a chance to do it, which isn’t very often these days.

John Abowd and David Card, ca. 2014.

LV: *So to some extent Orley impacted you, even though you didn't actually overlap that much with him?*

JA: Oh, yeah, you know, I spent the first half of my professional career with about half the people I encountered thinking that I was his student, not Arnold's. But I mean, it's true, I didn't do my PhD at Princeton, but I was his student every bit as much as I was Arnold's student. They were quite complementary.

IS: *Would you say that you got your approach to research methodology from your training at Chicago, and the set of economic questions that you ultimately ended up being interested in from your time at Princeton, or is that too reductive?*

JA: Well, it's an oversimplification. But you know, good hypotheses are oversimplifications. I will say that I actively avoided certain subjects in economics. They were distinguished by two features: too much in the family economics tradition of Gary, or something that I knew Jim Heckman was actively working on.

5 | CORNELL

LV: *So how long did you then stay at Chicago?*

JA: I came up for tenure as an associate professor, in 85-86, so I took leave. Chicago didn't have sabbaticals, but if you could fund it, you could take a leave. So I took leave and went to MIT as a visiting faculty member. I taught the graduate labor sequence with Hank Farber that year. We worked on some of our early papers together that year as well. That year I was denied tenure at Chicago, which I have to say was not a surprise. So I went on the academic labor market again that year and was offered the position at Cornell and after I had accepted it, Orley invited me to spend another year at Princeton. That was in the era when "visiting from nowhere assistant professors" –

or associate professors in my case – became a thing you could do. So I believe I was a “visiting from nowhere” Associate Professor.

Anyway, I had a term appointment for one year in the Industrial Relations section. I’ll never know, because I’ve never asked him whether he did it on purpose, or he really did it by accident - he forgot to tell the department that I was coming, so they “forgot” to assign me any teaching. Well, I’m pretty sure that the department didn’t forget that! If you’re not on the list, you’re not going to get assigned. So I had a year of research, and it was extremely productive. It was sort of a transition back into the style of labor economics that Orley had introduced me to - Hank and David – David was back there at the time. David and I kept working on our stuff together.

And so the most curious thing about that year at Princeton was that I learned the rules at Cornell. Of course, I had been offered and felt I had accepted a tenured position at Cornell, and I had, but Cornell doesn’t actually do the tenure paperwork at the time they offer – it’s still true. So one day in the middle of March, I think it was, in my year at Princeton. I started getting telephone calls at Princeton: “Congratulations, congratulations, congratulations!” They had put my tenure paperwork through, and the [Cornell] trustees had approved it. The dean at the time had the letter officially offering the tenure, which I hadn’t even known to wait for. But it was coming my way.

John at the Edmund Ezra Day memorial in the Cornell Botanical Gardens, 2019.

IS: *What was the environment like at Cornell when you got there?*

JA: So the thing I can’t stress enough is that Cornell and I were an incredibly good fit and both prospered from my being there. It was, as I’ve described it for most of my career, the best school that would have me, and one that was actively engaged in being as good as it could be in all the disciplines that I cared about. In my time at Cornell, it went through four different periods of trying to have a major effort to improve economics. I actively participated in the first three and actively declined to participate in the fourth one, but they all had the same theme – basically, recognizing that the social sciences were a multi disciplinary operation, and that, unusually for a major university, Cornell had been set up that way from the beginning. So it had economists all over the place, had statisticians all over the place, sociologists all over the place, behavioral and social psychologists and neural cognitive psychologists all over the place, and it was good for the departments and the colleges that they had their primary appointments in. It was not as good for Cornell's professional profile in those disciplines. It made it hard for Cornell to demonstrate its actual excellence in most of those fields – not that that was disputed in the particular areas where those social scientists had their primary appointments.

Labor economics is one of the best examples of it. Nobody in economics disputed that there was a really strong group in labor economics at Cornell. Either you didn't know they weren't a separate department, or you thought that was the economics department, if you were in the economics profession. That was fine before having an internationally prominent economics department became something that a major research university just had to have. All of the efforts that Cornell made, eventually culminating in a super-department *[ed. note: in July 2021]*, with real authority, were aimed at the external world far more than the internal world and they were aimed at recognizing that Economics, like many of the social sciences, depended on both internal research support and external research support, and if you're going to have external research support, you need the right institutions. When I arrived at Cornell, it didn't have them for social science research and when I left to go to the Census Bureau, it did. I'd like to think I helped get Cornell from A to B. But there were lots of other people who contributed as well.

6 | FRANCE AND LINKED EMPLOYER-EMPLOYEE DATA

IS: *Your work with Francis Kramarz, David Margolis*[31] *and others in developing statistical methods to analyze integrated employer-employee data is very well-known among economists. But could you talk about that earlier work and how you got started in that line of research to begin with?*

John and his family in Paris in 1991.

JA: I've been dying to tell those stories. So let's start from the very beginning. When I arrived at Cornell as a tenured Associate Professor, Cornell gave me three years of credit towards my first sabbatical. So I was eligible for sabbatical in 1991-92, and knew it in time to start making the arrangements in 1990. I started talking to Janet, our kids were young, and she said, "Well, my first choice would be any French-speaking European city." So I made contact with Alain Monfort[32] at INSEE,[33] and someone at the ILO [International Labor Organization] in Geneva. I ultimately got connected to somebody at HEC, which is a French business school. Even before anything was determined, I signed up for intensive French at Cornell – the all-day summer course. Margolis remembers this, because he was already my student, but more particularly my colleagues remember, because when we broke for lunch – and I occasionally had lunch with them during that class – I didn't stop speaking French, which they found more than a little annoying. I did it on purpose. Because I knew that Janet wanted the immersion language experience for our children. She understood that immersion is the only way that you actually master a language, as she had when she lived in France as an undergraduate. She wanted to start teaching French. So she was planning to enroll herself in the *Alliance Française*, of course, while she was in Paris, and we were planning the immersion experience, right? Our kids in French language schools, us working at a French speaking institutions.

I met Alain Monfort in November of 1990. At that point, I'd got my intensive French training and had hired a tutor for Margolis and me. I went alone to Paris to see if I could manage. I could manage tolerably well in a little hotel that didn't have an English speaking clerk. Monfort and I spoke a combination of French and English, his English at the time wasn't spectacular, but it was certainly better than my French. I met Francis [Kramarz] then, too, but only briefly. It was Monfort's idea that this young game theorist who was coming into the research department at INSEE might profitably pair up with an American economist, and that it would probably make both of our research skill sets better. So Monfort made the marriage, and I went back to Cornell with the invitation to visit INSEE on fellowship.

The next time I saw Francis was in August of 1991. He had just gotten back from his honeymoon. I went into his INSEE office, and I sat down. I think we were mostly speaking French, but it was taking me a while to get acclimated. Anyway, he was flipping through some of this famous 11 by 14 fanfold that we discussed earlier. I said, “Whatcha doing?” He said, “I’m checking the derivatives from my asymmetric cost function dynamic factor demand model.” I said, “Well, that’s nice. Did you program them numerically?” He said “No, no, they’re right.” That was so French, I had no idea how French that answer was at the time. So anyway, I ask him, “well, what’s your data”? And he starts describing the linked employer employee data that we would eventually use, and what he had done with it. What he had done was create a super noisy wage measure so that he could put a wage into his dynamic factor demand model and fit the thing.

So I suggested there might be some other interesting things that we could do with the data, and asked “are you interested”? And he was! Immediately! So he and I both dove into understanding how they were constructed, and how the linkages worked, and each of the linkage components, and we built that core database that we used for the work that became “High Wage Workers and High Wage Firms” [3] [ed.note: published in Econometrica in 1999].

David Margolis was using the data as well for the main essay of this thesis—his job market paper— which was a wage bargaining model. Towards the end of my stay we brought him actively into our project. He had met a French woman and was desperate for ways to spend more time in France. He came back to Cornell, finished up his PhD, and then went back to Paris primarily to work on “High Wage Workers and High Wage Firms” as our collaborator in place. Over the coming years I went to France for multiple week periods, several times a year, supported by teaching MBA courses at HEC, to finish that project. I’ll leave you to characterize the influence of that paper on the economics profession.[34]

David Margolis at the Abowd apartment in Paris, ca. 1992.

LV: *We have a special issue forthcoming with the Journal of Econometrics to point to for that.*

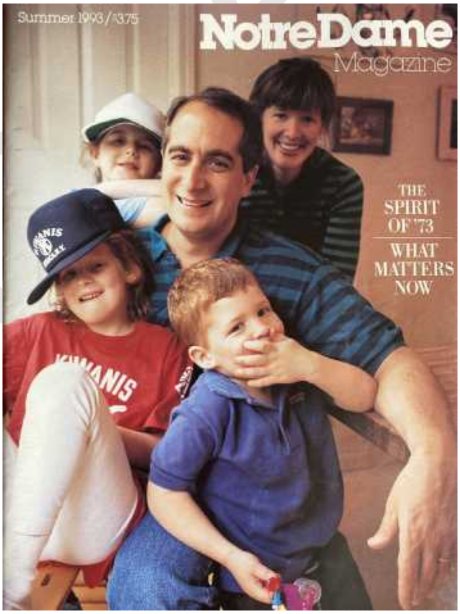
JA: I will tell you that its influence isn't for the reasons that you might think. Labor economists thought, I think, that it was a clever trick, that there was something interesting there. But macroeconomists just took to it like there's no tomorrow. If you want your citations in economics to skyrocket, you need to have macroeconomists reading your work. Labor economists often contribute to that, so they rise. But it was Dale Mortensen,[35] who said, "I need a theory that explains these facts," that we had found. That really made it something that people paid attention to.

I don't think Francis and David [Margolis] and I ever thought that we had made a major theoretical contribution to either econometrics or labor economics. But we had certainly put our fingers on the problem, which is that there's an independent effect of where you work on your compensation. If your model waves its hand to explain that away – if you're comfortable with that, that's on you. I'm not comfortable with that, because I don't think that it can be waved away, I think it is a real labor market phenomenon. It has yielded only gradually to further theoretical and empirical assault, and so it remains something that economics writ large – meaning the multidisciplinary approach to economics that is now the canon – has to be able to explain. How is it that your employer history, as opposed to your personal history, have profound permanent effects on your lifetime earnings? That is an extremely important component of heterogeneity, statistically, in labor market outcomes, and not just in the United States. Although, interestingly, it is dampened in France, and probably by institutions in the French labor market that dampen the ability of that firm effect to play its way all the way through your lifetime earnings.

But for whatever reason – I don't pretend to have a complete theoretical explanation for this either – that's a real fact. It calls on new generations of economists and sociologists now to meet that challenge, they have to keep their statistical skills current. So that's been very gratifying to see, to have a paper that demonstrates something that is extremely difficult to reconcile with the one-price world of neoclassical economics, and that is a persistent phenomenon that other people find interesting and worthy of theoretical and statistical investigation.

LV: *As far as one can tell, most economists picked up on the fixed effects version of the model from that paper. But having been exposed to some of this from its earliest days, you didn't actually see that as the primary or right way to do these kinds of things.*

JA: That's right. I actually think that the JBES paper that models it as a fully random process is the most important contribution that we make statistically [4]. Except for the initial conditions, we captured everything in there that you could expect a statistical model to try to capture, and so the basic empirical evidence can be handled that way. That said, that's still not a particularly elegant implementation of Bayesian estimation. Further refinement would really be important. The point that I have tried to make about the fixed effect estimator in that framework is that from a classical frequentist point of view, the fixed effect estimator is a proper estimator of the underlying phenomenon, and it's statistical properties can be demonstrated using the same tools that we use for other sample based estimators. The finite window problem is a well known problem in repeated observation data sets that can be addressed. So I don't like to use the term biased without context here. I'm happy with the term nonstructural and there's no question that it is a moment estimator of particular moments, but it is a proper moment estimator of those moments and its statistical properties, from a classical point of view, are known and can be used.



John and Janet Abowd and their children on the cover of the Notre Dame Alumni Magazine in 1993.

What I think the random effects interpretation points to, both in the work that I've done and in the work others have done, is that with any of these large scale, network-like data structures and relational-database-like data structures, classical sampling statistics is not the right way to think about the problem. The right way to think about the problem is from a primitive data generating mechanism that explains how the tables are created and how the tables relate to one another, from first principles describing the randomness at every component. That's now, I would say, standard operating procedure in Bayesian economics, and Bayesian econometrics - Bayesian statistics (I don't distinguish between Bayesian econometrics and Bayesian statistics). It's not really standard operating procedure yet in some of the methodologies that economists are particularly fond of, especially regression discontinuity and regression kink design, so unifying those approaches – and I think there's some good work being done there – is an important contribution to the science here.

7 | EARLY WORK WITH CENSUS AND THE LEHD PROGRAM

IS: *Looking at your CV, it is clear that your career started with a focus on analyzing labor markets and questions coming out of labor economics and you have sort of pivoted to focus on production and dissemination of official statistics. It is actually hard to identify a moment of transition. In fact, you can see elements of both all along the way. But is there a phase or a point in time where you think your focus started to change? How might you characterize that transition?*

JA: Well, it was induced more than you realize. I would have been perfectly happy to keep collaborating with Francis Kramarz for the rest of my career, working on French, and then I imagined it would have been German, and probably Swedish or Norwegian [administrative linked employer-employee] data. There were very limited barriers to entry for me to that sort of data since I had already born the principal entry cost. But an Associate Director at the Census Bureau, Nancy Gordon,[36] was also aware of my work with linked employer-employee data. She had been mentoring Julia Lane,[37] who was trying to do similar things in the US. The two of them recognized that it was going to take a major effort by the statistical agencies in the US to get anything like that off the ground. So they recruited me to take a sabbatical at the Census Bureau [in 1998-1999], and start the Longitudinal Employer Household Dynamics (LEHD) program, which did not yet have that name.

I didn't realize it at the time, but that was an active choice to go into official statistics. It just looked to me like an active choice to get the keys to another "candy store" of data. I sort of knew better; that there were no free lunches. I don't think I immediately internalized the lesson that you can organize all these data, but you can't keep using them unless you generate products that can be labeled statistical products from the agency whose infrastructure is enabling this. But I definitely internalized that lesson early enough for it to have a very strong effect on the rest of my career.

I think that that was one of the most important decisions that I made professionally. That was Nancy Gordon's mentoring. She understood, and put the force of her position behind her understanding, that surveys were dead. What I mean is that the notion that you should propose an expensive purpose-built survey as the main way to collect information was dead. Yes, you should have purpose-built surveys. Yes, official statistical agencies should be their primary sponsors, and their toolkit should include how to manage them throughout their lifecycles. But the information age was clearly dawning in the 1990s, and if you were paying any attention at all, you knew the era of big data was coming. So if that was going to be the way the world worked, you're going to have to figure out how to reuse those bits. Nancy understood that in a way that distinguished her from her colleagues, and distinguished her I think, from many people still active in the official statistics community who haven't moved on.

IS: *I have often heard you describe LEHD in the context of a 21st century statistical system, and certainly a much broader view of the role and conception of official statistics. So I wonder if you could comment on the transition from the LEHD as an academic project to an official statistics project, in those early years at the Census.*

JA: I think, to be fair, that the 21st century statistical system tagline came along pretty early in LEHD history. Julia Lane, John Haltiwanger, Pat Doyle, and others, organized a conference in 1998 on linked employer-employee data, and invited a wide range of people who had done such things from around the world to give papers. Francis and I gave a presentation and prepared a paper for the conference volume [1]. During the social hour party, Julia and Nancy Gordon cornered me and said, "Would you entertain coming to the Census Bureau for a year on your next sabbatical, which hopefully is soon, and help us get something like this started in the US?" At the time they had a demonstration project going that was linking lots of things from the state of Maryland. She and Jim Spletzer[38] and some others wrote papers using those data - that was their contribution to the conference volume [12]. We started talking about it, and I said yes – not then but shortly thereafter – and it started probably the most intensive professional collaboration of my career.

We went on several tracks, but the main track was during the first year. By September of 1998, we had an IPA in place.[39] I wasn't on sabbatical I don't think until the next year. I was still teaching part-time. We had a team that included a lot of familiar names – Ron Jarmin[40] was on it, Julia was on it, I believe Nancy Bates was on it, if she wasn't on it at that point, she was shortly thereafter – that was attempting to negotiate the Memoranda of Understanding (MOUs) necessary to get W2s.[41] The original design for LEHD was to use W2 data. I think it's – at least the abstract is a matter of public record – the proposal that we wrote for NSF – described a plan to use the Continuous Work History sample,[42] which I know Ian is pretty familiar with.

So that grant got spectacular reviews at NSF, as reported to me by the program manager, some of the best he'd ever seen. We got all the money we asked for from NSF, which was, in the end, trivial compared to the commitment that Census ended up making, but it was a major investment. In addition, the Sloan Foundation and NIH kicked in, so we had a fair amount of grant support for the beginning of it. We also had a decent commitment of internal staff – B.K. Atrostic became the first program manager. So what happened was that the infrastructure to get a research team in place proceeded. We got permission to hire four economists, one of them is participating in this interview [Lars Vilhuber]. Another one is Martha Stinson[43] – actually Martha was an RA, so were Paul Lengermann[44] and Karen Conneely.[45] And four economists, Lars might have to help me recall.

LV: *Kevin McKinney, me, Roberto Pedace, and Kristin Sandusky.*[46]

JA: So that team got in place and they started using the assets that we had. Those were the linkages of the Survey of Income and Program Participation to W2 data, the linkage of Current Population Survey to W2 data. Even in the early 2000s, there was a plan afoot to link the American Community Survey, which of course is now a fait accompli. What there wasn't was a properly executed set of MOUs. In 1999 the IRS started its routine (every three years) safeguard review of the Census Bureau's curation of federal tax data, and they had two projects in the crosshairs – the expansion of the RDC network, which was using business data almost exclusively at that time, and business data in that era, and still, is thoroughly commingled Census Bureau collection and IRS tax collection. And then LEHD writ large, meaning record linkage projects involving tax records.

That safeguard review included long-standing projects at the Census Bureau and the Social Security Administration, that use the link of the SIPP to the – I've been saying euphemistically the W2 data, but there's various versions of the W2 data, this one's called the Detailed Earnings Record, which is essentially your complete earnings history back to the beginning of the Social Security system. The Census Bureau had been sort of dabbling in using it. Mostly it was used to try to see how well labor income was reported on the SIPP and CPS. But over at Social Security, it was the bread and butter tool of the Office of the Chief Actuary in estimating the various Trust Fund projections for Social Security, with a small army of economists and statisticians and actuaries (those are actually three distinct professions, they have substantial overlap) that used those data routinely.

There was a well executed MOU between the Social Security Administration and the Census Bureau that covered, essentially, the provision of Social Security numbers to SSA, the extraction of the Detailed Earning Records associated with those Social Security numbers, and then the conveyance of those data back to the Census Bureau. At the same time, a copy of the public use Survey of Income and Program Participation lived at the SSA, and a crosswalk between the public use IDs and the IDS necessary to do the linking also lived at SSA under an arrangement that looked like an RDC but didn't have the RDC legal infrastructure surrounding it.

IRS asserted that what SSA did with the DER was their business because they had co-equal statutory authority to use them. What the Census Bureau did with the SIPP was its business, because it was authorized under Title 13 Chapter 5. But as soon as the two met each other, IRS asserted joint statutory custody over the linkage, and required the Census Bureau to justify the linkage in the statute and to get IRS approval for the MOU.

That safeguard review was among the most contentious things that I have witnessed in the federal government. I believe the reason is that both sides, three sides if you count SSA, were correct, but their policy positions in support of their statutes were mutually inconsistent, and each claimed veto privileges. It isn't the end of the world at the Census Bureau if IRS vetoes linking to W2 data for Title 13 purposes. That veto is easily challenged because the statute certainly does permit it. However, that meant that SSA couldn't use the linked data, and that was catastrophic from the point of view of the Office of the Chief Actuary and the research economists at SSA.

They desperately needed a compromise, and LEHD was basically held ransom. I was in Paris – I was in Francis' apartment – when the letter from the Commissioner of the IRS arrived at the Census Bureau *denying* permission to do the LEHD linkage to W2 data. So – I believe if you went back and interviewed every senior executive at the Census Bureau for the last four decades, you would not be able to find a case of an actual denial. That's not what's supposed to happen. Things never getting signed, that's another matter. That is still not uncommon. However, an actual denial – that was an unprecedented move.

Now we needed a plan B for LEHD. What was needed – and this is basically what the IRS repeatedly pointed out – was that the regulations need to actually match the data elements that you asked for, and that was not the case. SSA has no stake in that – they don't have to have a regulation that permits them to use any element of the W2. They're the ingesting agency, co-statutory custodians with the IRS. But Census did, and that was well-established in policy even in the early 2000s, because that was the period in which the Business Register was fully redesigned, and the regulations supporting all of the business tax data that come to the Census Bureau in support of the Business Register, and ultimately in support of economic censuses and surveys, are all properly delineated in 6103j,[47] which is the salient regulation. But the elements of the W2 had never been properly delineated. If you're doing this correctly, those Treasury regulations have to be revised almost every year, because they actually call out the line and the box on the form.

Lars probably remembers this only too vividly – LEHD became the support laboratory for getting the Treasury regulations rewritten and delivering the salient documentation. There was a preliminary publication of revised Treasury regulations in 2001 or 2002 – relatively early in this process – and once the temporary regulations were in place, then the data were released back to the Census Bureau. In the interim, they had actually all been boxed up and driven up to SSA and dropped off at SSA headquarters in a box. I think Gary Benedetto[48] did the drop off. I know Martha Stinson did the pickup when they were liberated.

Meanwhile, back at the ranch, we launched a major redesign of LEHD. It had two features: It was going to use state unemployment insurance wage records, which are quarterly. It was going to use the universe rather than a one in 100 sample, which opened the possibility to produce local labor market statistics in a manner that to be frank was never anticipated in the original NSF proposal. But that's where the 21st century statistical agency moniker comes from.

IS: *So the QWI – clearly it's there in the name – the Quarterly Workforce Indicators – were developed after it was clear that you'd have to follow this Plan B and use Unemployment Insurance wage records. Were there plans for a public use product out of the original LEHD design?*

JA: At the time the NSF grant was written, consistent with the predecessor to the Data Stewardship Executive Policy Committee, nobody had any idea how to produce a public use product that looked like anything recognizable as such. I mean, we knew this when we started just looking at the surveys linked to the SIPP and the CPS. The detail was great, and on top of that, something that modern critics of our disclosure avoidance systems consistently ignore, other agencies have the linking keys. I don't just mean they have your birthday, I mean, they have the actual earnings field that you want to put on the public-use data files. So you have to do something serious and strong. That's been well known for a very long time, it just gets routinely denied by some members of the data user community.

In any event, what we designed into the Quarterly Workforce Indicators was a whole bunch of local labor market statistics that were developed in collaboration with the state labor market experts. This is the area where the labor economists really rule – that all three of the senior scientists, Julia, John Haltiwanger,[49] and myself – have contributed to that literature, both in job and in worker flow measures. So we pretty rapidly ginned up a very large set of statistics. The state partners had plenty of ideas, things that they would have done in a heartbeat, had they been able to link demographic information. But the Census Bureau was the only show in town for linking demographic information, and really ...

LV: *Technology, if I just may weigh in there, played a role in this as well. They might have wanted to compute some of these statistics, but at least at the time, with maybe the budgets to technology in place. I remember that we shipped to California, a hard drive so that they could house the statistics we sent back – not the original microdata – because they had no capability to actually do the computations in house.*

JA: They couldn't manage the computations. We could do the computations because one of my terms and conditions for going to the Census Bureau in 1998 was a state of the art computational environment that could support our work. We had the first DEC Alpha[50] that was deployed at the Census Bureau. The IT guys were spending up post-Census 2000 money, and bought seven or eight. LEHD got one. I think Lars might have been one of the people who was running test programs on it, as we attempted to confirm that nobody's 64 bit software worked.

LV: *We helped develop STATA's 64 bit version for Unix.*

JA: We wrote a test environment for the 64 bit STATA, and SAS wasn't as constrained as STATA was, if I remember right. So the compilers just had to run on the Alpha. The second machine we had [ed. note: a Sun SPARC64 system] also was a nice piece of hardware, and it's kind of regressed to the mean. So that basically consumed, essentially the first six years of the program – recruiting state partners. When we launched I think we had four of the five biggest states plus Maryland. Then Julia [Lane] went on a recruiting campaign, and got it up to 12 or so before she moved on. Jeremy Wu[51] and Rob Sienkiewicz[52] basically finished the job and got it up to 50 states plus DC. [53] Where it's been hovering. You know, it is a maze of memoranda of understanding, one for each major supplier. There's one for each of the 50 state labor market agencies, not all of which are currently in force because they expire, and different rules in Commerce either do or do not allow you to continue on the expired one while you're doing the replacement. There's one with SSA, of course, and there's one with the Office of Personnel Management.

So the statistics that were developed, are still in production, they've been continuously improved and refined. They are reasonably widely used in that community. OnTheMap,[54] which came later, essentially opened the door to a much wider set of uses outside of labor market uses. That really was a Field of Dreams thing. I don't think anybody, with the possible exception of Jeremy Wu, who came from the transportation statistics community, saw the extensive user base that that was going to create. The application to Federal Emergency Management was Chip Walker. Once you have an integration framework, so you've defined the minimal geographic cell, integrating lots of other things becomes, if not straightforward, at least a project that experienced engineers who understand the infrastructure can create specs for and do. So OnTheMap was really what opened the door to very different uses than we ever envisioned in the early 2000s.

8 | STATISTICAL DISCLOSURE LIMITATION AND DIFFERENTIAL PRIVACY

IS: *So were the issues with traditional methods of statistical disclosure limitation for protecting the statistical products that came out of LEHD immediately apparent?*

JA: Of course. We did a suppression study [2], and we showed that 80 to 90 percent of all available statistics were going to be suppressed even before applying complementary suppression, right? Suppression was just not a viable alternative. Of course, no user that I have ever encountered thinks that noise infusion is a good idea, until it is compared to the available alternative, which is suppression, and not "you can have the data without noise infusion." If you can have the data without noise infusion, we wouldn't be having this conversation. So, you know, I think many of our early conferences consisted of me explaining yet again, how the noise infusion system worked, me taking questions from many of the labor market expert users in the various states about what must be being distorted by this noise infusion system, and then the crew at LEHD going back into the data and saying "No, that's not the cause. The cause is bad input data. It survived to the output because the disclosure avoidance system wasn't masking the defects in the data, it was masking the data."

The major source of concern was that we were not able, with the statistical tools available in the early to mid 2000s, to design a system that was completely consistent with the Quarterly Census of Employment and Wages.[55] Starting with the microdata now and using sophisticated statistical modeling, we could – but it's not clear that we should, because the BLS still uses suppression on those data, and the Census Bureau does not. In fact, there are certain cells that are always published in the QWI, and they're noise infused, but they can be used as alternative frames for constructing employment statistics, local employment statistics. You can either use employment, which is sometimes suppressed (although not that often) but always noise infused, or you can use quarterly payroll, which is never suppressed, but also always noise infused. Academic researchers, especially crews out of California, Berkeley and other places, caught on quicker than others that if you understood the structure of the Quarterly Workforce Indicators, you could use them as a large-scale longitudinal database and do very sophisticated statistical analyses. That was not the target use case. The target use case was One Stop Labor Market Information offices.[56] They needed the demographics.

To this day, I think the whole operation should be jointly sponsored by the Census Bureau and the Bureau of Labor Statistics. Multiple times an agreement in the form of a memorandum of understanding that would have accomplished this end was within sight. The closest we ever came was when the person who had to sign it at the Census Bureau, Chet Bowie[57] had signed it. The person who had to sign it at BLS, Phil Rones, didn't sign it before he retired. When he was replaced and Chet was replaced, their two replacements – I'm not gonna name them – let it die. The next time – multiple tries after that – the next time I was involved, was when Erica Groshen was the Commissioner.[58] We had terms written out, but once again it never got to the signature stage.

I find it extremely frustrating that the major statistical agencies have so much trouble collaborating to the point where we get what can only be reasonably described as demonstration projects going, and we hold them up as major collaborations. We finally get some traction on collaboration when there's an emergency. The pandemic got all the statistical agencies to sign on to the Pulse Surveys that Census was doing, if they wanted questions fast.[59] And got changes to the monthly surveys, because you didn't have a choice – face to face interview wasn't an option. So you know, you might not have time to run overlapping mode tests. You're going to have to run with it and use whatever artifacts you have to try to adjust for that. There was enormous creativity at all the statistical agencies – but we're sort of off the subject now.

There's no change in the law, the Census Bureau can't deliver the confidential microdata that enhances the UI wage records to the BLS. It has to be done under the joint custody of Census and BLS, and in an MOU, and in appropriate, secure computational space. No, I didn't write that law and I didn't write CIPSEA.[60] But neither agency denies that those are the facts. It has to be done that way. So let's get creative and do it that way.

9 | MODERN DISCLOSURE AVOIDANCE AND BERTINORO

LV: *One of the things that I've described as a motivating factor for many of us [coming to LEHD] was this idea of not just unlocking the technological potential of these databases, but also the information potential of these kinds of databases, while balancing that with the stringent data confidentiality requirements of national statistical agencies. Was that part of your motivation to get involved in statistical disclosure limitation and privacy research? Or did that come in afterwards as sort of an add-on?*

JA: So I don't want to describe it as an add-on. The first thing we tried to do was enable a data infrastructure that you could build things on. And there, I think that the NSF proposal that provided a substantial portion of the early funding recognized that these data had to be created because they were an asset to the statistical agency. Recognizing them as an asset to the statistical agency obligated us to produce public-use products. So the LEHD team fully internalized the need to generate public-use products. But then we were faced with a set of disclosure limitation problems that no one had figured out a good solution to – at least not as far as any of the associate directors of the census in the late 1990s were concerned. Their view was that it was impossible to publish anything detailed and useful from these data because the toolkit for disclosure limitation wasn't sufficiently developed. So—Lars, you participated in this early research, so you know—we investigated a number of ways to do it.

Clearly, I think looking back, the synthetic data approach was the most promising of them, however, it was way too immature to use in the early 2000s. We recognized that input noise infusion methods required the least amount of retooling, but those methods were really not particularly well developed. There's really only one scientific paper that that is based on [9]. But that paper only tackled the problem of a single magnitude measurement, not of repeated applications of the same disclosure limitation system. The LEHD noise infusion system we developed [5] led the way for the systems that are now pretty common at the Census Bureau and common in other national statistical offices. It wasn't perfect, as you guys know, only too well. But it was, as my colleague John Eltinge likes to say, way better than the 80 percent solution. That is what you have to be prepared to implement in a statistical agency. You can't let perfect get in the way of better. So yeah, it virtually eliminated the need for suppression, and it provided confidentiality protection in a more principled way than any other tool that was available in the early 2000s except suppression. So we went through and we addressed magnitudes, differences in magnitudes, ratios of magnitudes, and the other ways that those data were actually going to be tabulated and went through the litany of special cases, so that noise infusion could be applied. I think that system is still being used in largely the form that we left it when we moved on to other things. It got a differentially private residential side when the residential data started to come in [13].

But the thing we did was we made all the decisions for the users, we decided how accurate was accurate enough. We maintained the non-transparency that disclosure limitation systems have had, since their invention by Ivan Fellegi[61] in 1972, from a statistical point of view. That non-transparency has not served us well. But in a world where nothing was transparent, it also didn't make the LEHD implementation stand out. Yeah, of course, it's not transparent. That's just the way these things are. So we decided how much noise was too much noise and how much noise was just enough to get the job done.

To sell noise infusion to the data user community at the time, we had to show them how much data would be suppressed with even the most generous suppression rules. Between 80 to 90 percent of the data they were interested in would be suppressed even in the best case, so there was no way to make a usable table with other available methods. This is the same problem that agencies that insist on using suppression as their primary disclosure limitation tool are currently facing, including products in our own Economics Directorate that haven't adopted noise infusion.

LV: *One of the ways I tend to think of it, it's easy to sell that we're giving you either nothing or 20 percent, or 80 percent. Which one do you choose? The alternative is nothing. That's a lot easier to get the user base to accept than some of the other situations that one encounters these days.*

JA: And it was the right argument.

LV: *You have mentioned different types of disclosure avoidance systems. Let me get back to where you were part of a group of people that started to look outside of statistical agencies and toward computer science for ideas. Let me just ask what happened in 2005, or before 2005. That seems now, in retrospect, a watershed period for many of the subsequent developments in formal privacy.*

JA: It was a phase change. Stephen Fienberg was aware of the work that we were doing at the Census Bureau. We'd made a couple of publications of our disclosure limitation methods before 2005, including an early working paper on the system for the residential side of OnTheMap. It didn't work very well, but it was something. And so Fienberg knew that more than just the usual disclosure avoidance experts were at work on things in at least the Census Bureau. He also had a decent collection, at that time, of his former students or current students who had been working on disclosure avoidance issues.

Steve Fienberg had interest and talents in many different parts of statistics and he had worked extensively on what are easily recognizable now as modern disclosure avoidance methods, output noise infusion in particular. I don't know how he met Cynthia Dwork[62] – but he was already co-directing the Statistics and Computer Science Laboratory at CMU and taking doctoral students coming from both disciplines. So I think that her early work got to him that way - I never asked him. It's too late now.[63] But I know he had read early versions of Cynthia's work with Kobbi (Nissim), Adam (Smith), and Frank (McSherry). So he organised a week long retreat in Bertinoro, Italy – co-organized it with Cynthia and Alan Karr.

Panorama of Bertinoro, Italy. By Baccolini via Wikipedia. CC BY-SA 4.0.

Cynthia invited a gaggle of computer scientists, who were, to be fair, mostly cryptographers who had moved into safe data publication. And then Steve invited a pretty wide assortment of American and European disclosure limitation (in Europe, disclosure control), experts, and me. You know, I was not recognized as a disclosure limitation expert at that time – it's possible I'm still not – but in any event...

We made presentations about our work. I made a presentation about the way we had protected confidentiality in OnTheMap before the formal privacy routines. There were talks on synthetic data methods and Jerry Reiter[64] was there, Sesa Slavković[65] was there – people who are now pretty prominent on the statistics side – and many of the people who are now prominent on the CS side. One of the things I can't remember is, besides the four main authors – who else was there from computer science. I don't know that a participant list has been preserved. Bertinoro was pretty isolated. But the social events, dinner and Steve's extensive collection of wine, and after-dinner drinks were appreciated by all and then we did a little traveling in Italy afterwards.

LV: *One of the outcomes of Bertinoro was the founding of the Journal of Privacy and Confidentiality.[66] How did that come about? And what do you think of how it's progressed along those lines of its mission to sort of bring all those folks together? All those various threads and various disciplines?*

JA: It started as Steve's brainchild. Along with Cynthia and Alan Karr, they created it. They were relying on Chris Skinner, myself and Kobbi Nissim to be the initial editors. But essentially, there wasn't any way to drum up content, unless the editors went out and found the papers and solicited them and had them reviewed. So Steve ended up being the primary editor for the first seven or eight years. You know, there were submissions, but a trickle. What he did was get people to write journal versions of their conference papers, and got contributors from law and social sciences and traditional statistics, right? Because it isn't the Journal of (FORMAL) Privacy, it really has consistently been multidisciplinary. But it's also consistently been pretty well edited. Some of the papers that have appeared in that journal are now the citation leaders in modern disclosure avoidance. That was Steve. When he passed away, that was really a moment of challenge to figure out how to keep that going.

Since it has a decent stream now, it gets — to be fair — higher quality submissions in formal privacy than in other areas. But we solicit in other areas and we are mindful of how we could make a contribution to the public policy and statistical issues that are at the heart of our multi-disciplinary approach. I couldn't solicit papers from economists — there were only a handful working in the area. Most of them had gotten into it like we got into it — "Rock, hard place, figure something out."

LV: *What made the conference actually take place – why Bertinoro?*

JA: He wanted us where we didn't have any distractions. Evidently, this is a venue that is routinely used by Europeans for this purpose. It's a converted monastery. It's easy to walk up on top of the hill. They had basically a residential conference facility, and it came with the standard Italian amenity — which is Italy — where French people go for food holidays.

There was no Rosetta Stone. There was no common language between the computer scientists and statisticians. I will be honest, I struggled to understand the arguments that Cynthia was making, and her colleagues struggled to understand what we meant when we said disclosure limitation if we didn't have a set of defining principles that generated the methods. Everybody recognized that except for primary and complementary suppression, the rest was based on much looser mathematical fundamentals.

Now, Steve Fienberg had been working in this area and one of his graduate students

Adrian Dobra,[67] had produced a thesis based on some of the algebraic geometry methods that they had developed [7, 8]. Basically, if I give you a set of marginals, can you enumerate all possible tables that are consistent with this set of marginals, including the possibility that there are none, or multi dimensional ones? Steve saw that as closely related to the work that cryptographers were doing, because you were trying to say, within the space of all contingency tables, conditional on these marginals, there is either a tiny, or non-existent, or uncountably large number of high dimensional tables that they are consistent with. So, being able to bury your table in that large number of consistent complete tables was one of his mathematical formalizations. Computationally, I guess it must still remain very difficult, because I've never seen any production system that implemented it, and it's not a useful tool to have in that case. But it is a natural generalization of complementary suppression. How many complete tables are consistent with this set of marginals, interpreting the marginals as a suppression mechanism. Though, the suppression experts who were there - Larry Cox[68] the king of the suppression experts who essentially did the fundamental math after Fellegi's basic theorems – they were already trying to figure out ways to replace suppressions with safe statistics that could be placed in those cells. Once again, that's not so distant from the database reconstruction notions of how to think about blatantly non-private systems that we might be able to make privacy-protected.

So there was enough common ground for us to see the benefit of each other's work, but there was nothing even remotely resembling a usable statistical product. To be fair to the computer scientists, that wasn't what they were trying to do. They were trying to get the basics right. To be fair to the disclosure limitation experts, if you don't have a feasible alternative, then you have to use what is available when the next batch of tables rolls around, and the next batch of microdata rolls around. So the problem was, as it often remains now, overconstrained. Still, we had a lot of time to talk and exchange and absorb.

I came back and started working with Johannes Gehrke[69] at Cornell as a part of the Cornell RDC, Lars as well, and Johannes' postdoc and graduate student at that time. The postdoc was Dan Kifer,[70] and the graduate student was Ashwin Machanavajjhala.[71] They were in his database group, the privacy-preserving part of it. But Dan and Ashwin were thoroughly committed to the privacy preserving data publication part. So the five of us did a reading group and then set out to determine if we could build something feasible that we could actually call formally private. That led to the technical paper on how to protect the residential side of OnTheMap with what we called at the time probabilistic differential privacy, and what would now be called approximate differential privacy [13]. And it was actually implemented.

Ashwin wrote a super-compact C++ program that did the critical calculation. I tried to rewrite it in SAS several times, and then Ashwin, I think, actually tried to rewrite it in SAS. He talked to some programming experts, and they said, "Don't ever try to write a binary search in SAS again." So then Lars figured out how to efficiently export the information needed, do the calculation, and pass it back. Before it was particularly straightforward to write a SAS proc that did that.

The core OnTheMap group – Matt Graham and Heath Hayward, and a group of contractors – implemented our method outside the Quarterly Workforce Indicators production system, so it was its own code base. It was what we claimed it was. It was formally private. I venture that no more than a handful of people in the world understand how it works. Almost no one understands that privacy was applied at the tract level, not the block level. It's modeled data below the tract level. But, you know, modeled to a use case that you should be able to draw arbitrary geographic areas of approximately the size of a tract and get reliable results. The statistics should be accurate in terms of distance and direction, from the relevant focal points, which we put at the center of every tract.

IS: *How difficult was it to get approved at Census? Was it difficult to bring people along with this new concept?*

JA: It was dead easy. There was a straightforward technical memo. A Disclosure Review Board (DRB) review and approval, you know. They were very open to getting help from outside experts. Disclosure avoidance people are not the most popular folks - they prevent data from being released! However, that attitude is definitely changing. I have heard people make correct, cogent, arguments that they believe about the necessity to do the disclosure avoidance properly, and to be transparent about it, and to document what it does to the inferences on the data so that the users are aware. That's a multi-decade, cultural transition, and it's not going to be done when I leave the Census Bureau, but it's gotten a not-inconsiderable push in the right direction. It started with OnTheMap, actually, which started with the noise infusion system for QWI, which was easy to do in terms of the DRB because it was based on work that researchers at the Census Bureau, including Laura,[72] had originally developed [9]. The basic idea was that you had a key that identified the entity and that key was tied to that entity's disclosure avoidance random numbers, and it stayed tied to those random numbers in perpetuity, so that the protections survived reuse of the data.

LV: *So that all directly then led to the 2020 census, right?*

JA: Aaaahh, Yeah. (Laughter)

10 | TIME AS CHIEF SCIENTIST OF THE U.S. CENSUS BUREAU

LV: *At what point did the idea of moving to the Census as a chief scientist become an option? Maybe for some of the readers who won't know some intricacies, the chief scientist position as such didn't actually always exist.*

JA: Well it did, actually, it got relabeled. Here's the history. There used to be a Directorate called Methodology and Standards. Its last career civil servant head was Cynthia Clark.[73] Cynthia Clark was one of the three associate directors who were the internal sponsors of LEHD and the other two were Nancy Gordon and Nick Knickerbocker. So Nancy Gordon was in charge of demographic programs, Knickerbocker economic, and Cynthia methodology and standards.

When he became Director of the Census Bureau, Robert Groves and Tom Messenbourg wanted to have a such a Directorate again, and renamed it Research and Methodology from Methodology and Standards. Groves and Mesenbourg really wanted an IPA from academia to head it–somebody with a very senior status in the profession. So they persuaded Bob's longtime Michigan colleague, Rod Little, to be the first one. Rod came in, and in the original IPA model for the Chief Scientist and Associate Director, he was going to be the dean-like figure. The administration would be done by the assistant director, who was Ron Jarmin. When they went to recruit a second one, the issue of relocating to DC became a serious constraint. But Tom Louis was eager to do the job and didn't have to relocate to DC because he was on the Johns Hopkins faculty and ready for a career transition himself. When he left the job, he also went emeritus at Hopkins. So there's the second Chief Scientist, and Ron stayed as the assistant director. He did all the things that Senior Executive Service career employees in the federal government are expected to be extremely adroit at doing – talking about HR topics, budget topics, managing the multiple roles and responsibilities that an agency that cannot officially matrix but are unofficially matrix in many ways, to keep the place functioning. Those are all the things that Ron did. When Tom was ready to leave I came in. Basically I volunteered. I told Ron, when he told me they were looking for a replacement, I said, "I'm interested if you want to do that," and that was a done deal in one day. He went up and talked to John Thompson and Nancy Potok[74] and the process started. I originally intended to do it on the IPA model and go back to Cornell and finish my career as a tenured professor. But that was not the way that cookie crumbled.

John and Janet in Washington DC in 2016.

The IPA started in June of 2016. Tom [Louis] had told them he could only serve until January 2016. So with John Thompson and Nancy Potok's permission, I moved down on April 1, and I started attending the meetings that were the Associate Director for Research and Methodology's responsibility. That was not uncommon. I was not the only person scheduled to do a job who was in the room and hadn't yet been fully installed.

IS: *What were your tasks meant to be In that role, outside of designing new systems for privacy protection and responding to the occasional federal audit?*

JA: Those things were not the top of my to-do list. Thompson and Potok put them at the top of my to do list.

IS: *When they brought you in, were they thinking formal privacy, initially?*

JA: Oh, no, they wanted more things like LEHD everywhere.

11 | THE 2020 CENSUS OF POPULATION AND HOUSING

IS: *We have mostly stayed away from politically sensitive topics, but there have been a few significant controversies at the Census Bureau that required your involvement over the last few years, and that you must not have anticipated when you took the job.*

JA: We can talk about it a little bit, but not about any ongoing 2020 Census activities or litigation. It has been a very difficult period for all of the career civil servants at all of the statistical agencies. But, when it's being conducted, the decennial census is the most visible statistical program in the country, and in the same way that the rules of the political landscape have been rewritten in the last six years, there was a serious attempt to rewrite the rules of how the Census is conducted. It wasn't successful, but it was extremely aggressive.

IS: *How do you think social scientists and other data users will need to adjust their research activities to account for the disclosure avoidance system?*

JA: The researchers are already adjusting their research activities – they're working with private companies, and the private companies are restricting their ability to compute and publish far more than the Census Bureau does. Transparent disclosure avoidance methods will allow public-use data to be used in a statistically correct manner. That was the impetus for the quality standards that the Census Bureau now has with respect to direct estimates from survey data – you want to publish them in a way that allows them to be used properly. The point was made decades ago in opinion polling – you almost never see an opinion poll that doesn't contain a margin of error. Now, whether that reflects all the uncertainty in the opinion poll or not is a much more interesting question, but the idea that just because you labeled it a census as a full enumeration, there's no uncertainty in the point estimates, is just foolish.

Uncertainty doesn't come just from disclosure avoidance. In fact, that's not even the major source of uncertainty. In a face-to-face, or direct report, population census, coverage error is the probably the biggest problem. The techniques available for measuring coverage error have many of the same statistical limitations that the original data collection had. So you want to have multiple measures, and you want to be cognizant of what they're telling you. Our demographers and internal statisticians have been smart about this for decades. But the communication problem occurs when the actual enumeration is published with precision down to the units column—its accuracy isn't anywhere near that because of the many sources of uncertainty. There's a dissonance that is hard to message.

IS: *We are both aware of some of the discussions about the implications of differential privacy for social science research in the work of demographers and so on. Would a silver lining of these debates be that differential privacy applied to the decennial census, and to other Census Bureau products, will force social science researchers to take the various sources of uncertainty in data products more seriously?*

JA: I hope so, and I hope that in doing so they don't think that the disclosure avoidance error is the major source of uncertainty, because that's certainly not the case. Demographers are trained to take imperfect population data and develop improved estimates of age pyramids and population growth. That's the inherent skill set of that discipline.

LV: *The forecast for future population is always uncertain, the contribution of the disclosue avoidance noise to that is likely to be quite small.*

JA: I don't understand why as a discipline, quantifying uncertainty isn't a higher priority, but I have this discussion regularly with demographers at the Census Bureau, and I can confirm that it is not a priority. Acknowledging uncertainty happens all the time, usually in the form of alternative scenario projections. But as much as they say, explicitly, each point in that multiple assumption set of scenarios is considered equally likely, when a less sophisticated person looks at it, they look at the middle and treat the extremes as a range. Really, the range is much wider than that, because each one of those points had uncertainty around it, and they're all considered equally likely.

IS: *I guess, one one last policy related question. This is just a very easy question. But how did you end up at ϵ – was it 19.8 – as the final privacy loss budget for the Decennial Census disclosure avoidance system?*

JA: Properly calculated it was 17.44 at $\delta = 1/10^{10}$. But you really should use the ρ parameter (2.63) [16]. In the end, the 2020 Census Disclosure Avoidance System used a much more nuanced version of differential privacy, called zero-concentrated differential privacy (zCDP), than the original framework of ϵ-differential privacy. zCDP uses the entire distribution of the privacy-loss random variable to reason about confidentiality protection, not just the worst case. Getting there involved balancing the interests of many different groups but prioritizing the redistricting use case. Redistricting is challenging because the final geographic units—the new voting districts—cannot be specified in advance. They have to be drawn using a basic geographic primitive—for decades now census blocks—to deliver approximately equal population voting districts within pre-specified, politically defined geographic areas. These voting districts must conform to the anti-discrimination provisions of Section 2 of the 1965 Voting Rights Act. The catch is that location protection requires uncertainty in the block populations. That uncertainty must fade rapidly as the voting districts are aggregated into pre-defined political districts. This requirement, along with the requirement that large racial and ethnic minorities had to have the chance to form voting districts in which candidates from those minorities were in principle electable, meant that sufficient privacy-loss budget had to be allocated to racial and ethnic data and to the populations of political entities but relatively little needed to be allocated to the block-level statistics themselves.

12 | RESEARCH CHALLENGES FOR CENSUS

IS: *So that is sort of related to my next question. What are the most important research challenges facing the Census Bureau, or the developers and users of official statistics?*

JA: So it is the same one that was true in 1998. It's now so big an issue that it not only can't be ignored, you can't say you're thinking about it, you actually have to be doing something about it. We have to reduce the burden on respondents, and reuse information as much as possible. You cannot generate every statistic with its own custom instrument. That's still a much more difficult transition than most people realize.

The core set of methods, that the 350 or so mathematical statisticians at the Census Bureau know, are primarily related to finite population sampling theory: How do you draw the samples? How do you compute the estimator? How do you compute the variance of the estimator? And how do you detect failures in your assumptions? An incredibly valuable toolkit. We have sampling estimation and variance estimation branches all over the Census Bureau implementing the solutions for each of our 100 or so surveys. Yes, they use some common software. Yes, they use administrative records. Yes, they use third party data. But the use of administrative records inside the Census Bureau – the big use is to build the frames, which is a very traditional use. And to be frank, very well understood by the mathematical statisticians who do it and run or work for the various parts of the Census Bureau that assemble and maintain those.

But those frames don't talk to each other. That's Ron Jarmin's big project for the first part of this decade – to get those frames to talk to each other. It isn't just that they don't talk to each other. They don't even have a metadata language to use to describe how they want to talk to each other! The way that people think about how to do a new product remains: "Let me get access to all these specialized data assets in one place, and write a custom integration, estimation and data quality application.

We don't have a software stack that says: "You can see all the assets! Here are each of the tables, here's a primary key, the secondary keys. Here's the ones that have been linked exactly, here's the ones that have been linked according to probabilistic methods. Here are the tools you need to assess the impact of those probabilistic methods." The Census Bureau just doesn't have it.

Which doesn't mean that there aren't lots of innovative products — there are. There's very interesting work going on using AI and machine learning to improve autocoders. Autocoders have been around for eons. There's a huge one in the Demographics Directorate that is essentially every verbatim response to a race and ethnicity question that the Census Bureau has ever encountered since it started asking for such verbatims. I didn't know this, so you might as well learn it since I did. We didn't start asking for verbatims for whites and blacks until the 2020 Census. So they have this huge database of verbatims that are used to generate the complex coding schemes that get you all the detailed race and ethnicity codes. We have elaborate ones that do autocoding of NAICS, autocoding of occupations, and natural language processing from paper data capture. It wasn't until very recently that we started doing products that integrated some of this into the actual statistical output. There's a very nice state-level retail trade series that was just released and some other experimental products that are beginning to do serious multiple source integration.

That's the way we're going to live in this century. The 21st century statistical systems are going to reuse burden by harvesting things we couldn't harvest in 2000 because they didn't exist. The problem now is that we haven't figured out how to combine those data harvesting systems with our survey systems to produce statistics that we understand the properties of. But we're a lot closer. That really has to be enabled. So one of the things I have tried to do on my watch is to enable the computing environments that allow this to happen. We now have a very well established policy and computing environment for doing web scraping that respects all the regulations you expect a statistical agency to respect, but allows you to set up a web scraping project fully supervised by our policy office and fully capable of going into production once you have worked out the case. That didn't exist five years ago. Cloud computing environments didn't exist at Census five years ago. Those things happened simultaneously. The biggest push for cloud computing environments was to get the internet self-response instrument for the 2020 Census in the cloud. That had deadlines that you couldn't just push back a couple months. Those ATOs [Authority to Operate] showed up within days of the systems being turned on.

LV: *Those challenges don't sound unique to the US context. To what extent are you learning from others* *statistical agencies facing similar challenges? In particular, the integration of administrative data?*

JA: I haven't been as active in the international environment as I intended to be because of things that happened in the first four years of my appointment. But I'm aware of the large suite of experimental products that Stats Netherlands has. Stats Canada has also developed a framework, I think they have fewer products. We also have a framework, it's well-disseminated within the Bureau, and it's understood by the teams that are trying to use it, as deliberately in between a one-off research project, and a production project. It's an experimental data series or data product. The Household Pulse Survey and the Small Business Pulse Survey are such products. They explicitly do not meet Census Bureau official product quality standards, but they don't hide that fact. They lay it out, and they're in continuous improvement mode.

By the way, the Research and Methodology Directorate enabled the Household Pulse Survey, because I strongly encouraged our Center that does behavioral science methods, CBSM, to have a look at Qualtrics.[75] And they did, and I said, "As far as I can tell, you can prototype in Qualtrics much faster than the other things you're using." That was quickly confirmed. Qualtrics was a small enough company – it has since been bought – that they cooperated in getting an ATO, which meant that we could put Title-13 protected data into the frame: very extensive contact histories that contain emails and phone numbers. That's what enabled the Household Pulse Survey, because that frame could be spun up. Now we needed a lot of help from Qualtrics and from experts at the Census Bureau like Jason Fields, Jeff Sissons and Jenny Hunter Child in implementing standard surveys to get it to run, but it was enabled by having that contract in place.

That wasn't a problem for the Business Pulse Survey, because to run the 2017 Economic Census, they had switched to all electronic responses. So they had an electronic contact history, validated emails. It still has problems if you run on a long term because it needs ways to refresh the frame, which is being addressed. So, there isn't unwillingness to do these kinds of things, it just sometimes takes shocks of enormous magnitude to move the system to actually do it.

The shock that moved the 2020 Census to do modern disclosure avoidance was the demonstration that database reconstruction wasn't just a mathematical possibility identified by some computer scientists in the early 2000s. It was very real, and we were very alarmed.

13 | REFLECTIONS ON THE U.S. STATISTICAL SYSTEM

IS: *You mentioned that the biggest challenge facing statistical agencies is this issue of burden, a "burden budget" in some sense. Then there's a privacy-loss budget to go around. Once again, it kind of seems like those are things that are a little bit easier to think about allocating in the context of an integrated statistical system, which we don't have in the United States and didn't sound like you're particularly sanguine about the prospects for one.*

JA: Slim and none.

IS: *I know there have been some policy changes that would seem to support going in the direction of* *a more unified national statistical system. I wonder what you think about those policies, and then what would a second-best environment look like for the kinds of changes you think need to happen?*

JA: Sounds like you're talking about the National Secure Data Service? What I will point out is that the original *Commission* that Katharine Abraham chaired,[76] and the CNStat panel, the consensus panel that published the two volume series on modernizing federal statistics by using multiple sources, contained considerable overlap [15, 14]. They all got to their consensus by acknowledging that there were many nontraditional sources of data that were available, in principle, to enhance our statistical system at either no or very reduced reuse burden. Also, by acknowledging that there was an enormous privacy problem, which both the commission and the CNStat consensus panel explicitly acknowledged and prioritized in terms of where the research has to happen in order to get this to work.

In its modern instantiation, only the "There's lots of data sources that people want to use," seems to have survived into the public policy discussion. Like all these other problems are just going to go away. It survives in the public policy forum, primarily because open government and open data are already the law, and relatively easy to defend on a variety of grounds. But mostly, they're only useful for data that don't involve human beings.

As soon as the data involve human beings, every single agency that has the authority to collect it has its own confidentiality and privacy statutes. You can wave your hands and you can say, well, we'll fix that or that doesn't really matter. But in the federal statistical system, when people talk about data sharing, they mean "Census and IRS sharing their data with everybody else." That's what they mean. The only way to enable that is a statute that explicitly permits tiered access to those data. Said statute does not exist, and said clauses are not a part of any statute that I've seen. So it's very important to modernize the statistical infrastructure, and the statistical community gets about one new law every 20 years, the Foundations Act of 2018 - Foundations of Evidence-Based Policymaking Act – that's our latest act!

Right now, whether Congress goes on to create a National Secure Data Service or not, within that act, Title III definitely strengthens statistical agencies' ability to share data, definitely provides proper guidance in the form of law about what constitutes appropriate behavior by a statistical agency and what constitutes inappropriate behavior by the department that it may or may not be housed in. Those were really important advances, and they haven't nearly been codified in regulation the way they need to be.

14 | LOOKING FORWARD

IS: *What are some of the research projects – inside census, outside census, your own work, other people's work – that you're the most excited about right now?*

JA: In terms of other people's work, the effort to try to estimate prices and quantities jointly for price indices, which Ron Jarmin is leading, is, I think, likely to be a very important innovation. Not just because we can stop doing price indices by Matthew Shapiro's formula, which is: the Census

Bureau collects the numerator, the BLS collects the denominator, and the BEA does long division ... which belongs to him, not me. Inside the Census Bureau, I'm quite pleased with the enormous research effort that modernizing our disclosure avoidance systems has produced. What is extremely gratifying, though, is to watch as people are developing new products, they've basically been told that they have to use formal privacy methods or they're not going to get them cleared for release, and they have done so. So, as far as I can see, the phase change has happened.

Most of what complicates finishing it is trying to figure out how to handle the existing products, which – the two that are most concerning, the American Community Survey and Economic Census – they're just extremely tough nuts to crack. If you take the workflow and try to make the workflow formally private, that's simply going to fail, and has failed multiple times now. So a fundamental re-engineering has to happen. That happens in those products, but not every cycle. So, you know, patience is not my long suit, but there isn't, in this case, any substitute for it. It takes some patience, and as those products are reengineered, and especially as they're reengineered to use multiple sources, that's when we can intervene with modern disclosure avoidance methods.

In the intermediate term, the things I'm most excited about are moving from research to production uses of large scale statistical modeling as a replacement for direct estimation. I am the third chief scientist, not the first, who has tried to encourage that at the Census Bureau. We have enormous expertise. This is an area where there are plenty of statisticians and data science experts who understand how to do this and why you would want to. So it's going to be done for parts of the 2020 Census. Some of it is likely to be done for the American Community Survey. This comes a lot closer to the kind of win-win that we like in innovation because you basically get a product that has much better statistical properties, but no increased burden. You're supplementing the survey data with extra sources from which you get covariance and, sometimes, direct responses. You're supplementing the survey data by exploiting temporal, spatial and multivariate correlation in the underlying data similar to the way it's done in other products you see routinely supplied by data aggregators, who don't have to document their methods quite as well. That's what they're doing. They're modeling the low-level data. If you dig, you can see that that's what they did. There is really no reluctance to use modeled data in the federal statistical system. The national accounts are modeled data. But the tradition that guides most of the statisticians who work in the survey parts of the federal system strongly favors the direct estimate, even when the evidence is overwhelming that the generic use case would be better served by the modeled estimate.

IS: *That brings us to another question: what kinds of training do you think future official statisticians, or current official statisticians who will continue to work in the field should be focusing on.*

JA: The obvious one is standard methods in data science, which have their own vocabulary that I am somewhat reluctant to use because I wasn't trained in that vocabulary.

IS: *What are the components of the data scientists toolkit that you have in mind?*

JA: The ability to use modern programming languages to structure data you ingest in unstructured form. What I mean by unstructured form is that it doesn't come in rows and columns with a dictionary that defines the contents of each column—the plan that tells you where the rows came from. That's clearly one of the most essential tools. I believe that the term of art there is "wrangling". I learned that as ETL [Extract, Transform, and Load].

LV: *Does that need to be a statistician's skill set, or is the skill in developing the ability to collaborate with a data science expert who knows how to do the wrangling?*

JA: We're basically just adding data science skills to the descriptions of lots of positions. OPM [U.S. Office of Personnel Management] is supporting that by having a list of core competencies that you need to have to claim that you're a data scientist and qualify for one of these positions. But you could get them in a statistics program or in a data science program, and probably in many computer science programs.

It's not that we label it data science. It's that the way the survey lifecycle works at the Census Bureau, the data science toolkit doesn't come into play until much later in the processing, when you try to fix all the things that the formal sampling theory didn't deliver in the realized data quite the way the theory said it ought to have, and that's no secret. People have been doing that for decades as well. So that's the obvious one—add data science tools to the survey processing toolkit.

The far less obvious one is we don't hire enough people trained in applied mathematics and optimization. I have tried to find people willing to work for the Census Bureau who come out of such programs—Cornell has one, lots of the major universities have them. Their graduates are just scarfed up by Google, Apple, Facebook, and Alibaba. They're in incredible demand, because they can put together the software stack that you need to implement the things that survey statisticians or economists can prototype but they can't do at scale. They weren't trained to do at scale and they shouldn't be trained to do them at scale. We need whole units inside the Census Bureau whose job it is to do that at scale – to build the full software stack and allow the disciplinary specialists to exploit their comparative advantage, which is not the comparative advantage of the applied mathematician and optimization specialist. If I hadn't had a few of those specialists to call on, we would not have been able to do the 2020 census disclosure avoidance system.

IS: *You were trained as an economist, do you think that economists are properly trained to be working with or in the production of statistics?*

JA: The economists and the statisticians, the research economists, research statisticians, research demographers, research sociologists, research information science specialists that we hire at the Census Bureau are just like the ones that are hired in academia or in business. They come from their graduate education properly tooled, but then they show up in an environment that is not as enabling as it could be. You can have Python in all of your computing environments at the Census Bureau. Not that Python's the only thing you might want to use. But if you want to spin up a large scale computational facility, we only have a few experts who can help you do that. They're overcommitted and constrained by the rules. You know, I listened in amazement when Pat Bajari came to the Census Bureau three or four years ago, to offer us collaboration with Amazon and his team of economists. They know the statistics. But Pat was joined at the hip by a software engineer, who was his peer, and her job was to build the stack. They just need one working instance. They had one working project, and she scaled it to be able to run 100,000 hedonic regressions a day. With, you know, no army of economists selecting all the right hand side variables and properly trimming them up.

We live in a multidisciplinary research world now. If your team doesn't have specialists in the right proportions in the computational aspects as well as specialists in the subject matter areas, you're not going to get nearly as far as you could. I remember watching this at Cornell in the early 2000s – the biologists and the optimization specialists were already studying how to model protein folding, and they understood it to be a complex optimization problem for which there was not at the time good software. That toolkit did revolutionize biology. Similar toolkits are going to revolutionize the statistical agencies. They're going to tap non-human residuals of our activity, and build reliable indices from them that we can use to measure our well-being and other things with more granularity, more timeliness, and their own set of statistical quality issues.

Just a footnote, when we do that, we will probably stop publishing population numbers down to the units column, because the accuracy of the underlying data does not support that precision. Just FYI.

LV: *A real softball: where do you see yourself in 10 years?*

JA: Oh, I'm in agreement with my life partner and wife that I will not do this job much longer, and that I will have control of my schedule and therefore of our leisure activity back again soon. But I don't know whether I'll go back to the university or do something different. I'm gonna write a memoir. I've got some nice stories to tell. I can't tell them as a civil servant, but I can tell them afterwards.

I got a little nostalgic a while ago. I sent Xiao-Li Meng a paper he had been awaiting for the Harvard Data Science Review on the day it was due, June 30. He wrote back, "Thank you so much. You've got it in just as I'm about to start my sabbatical tomorrow, July 1," which would have been when my next sabbatical at Cornell would also have started, if I were still a professor at Cornell. So I want to figure out some substitute for that. It's hard to imagine retiring! That's not sort of a thing that I ever envisioned within the definition of "labor supply goes to zero." But I am going to take back control of my calendar.

LV: *Yeah, I happen to know that part of Xiao-Li's sabbatical involves going fishing. I consider that to be the ultimate control of your calendar.*

JA: That does sound pretty attractive, but I haven't been skiing in a while, and I wouldn't mind doing that.

LV: *I think that's an excellent note to wrap up the interview.*

Acknowledgements

We thank Janet Abowd for providing us with many pictures from John's life, only some of which found their way into this article.

Notes

| Notes for section 2

[1] Emil T. Hofman was Professor of Chemistry at Notre Dame who taught chemistry to over 32,000 students at Notre Dame. see https://news.nd.edu/news/emil-hofman/ (accessed 8 October 2021).

[2] V. Joseph Hotz is professor of economics, as of 2021 at Duke University. He obtained his Ph.D. in Economics at the University of Wisconsin-Madison in 1980.

[3] Deirdre McCloskey is a professor of economics, history, English, and communication oat the University of Illinois at Chicago. See https://www.deirdremccloskey.com/main/vita.php, accessed 8 October 2021.

[4] James Heckman is an economist, still at the University of Chicago in 2021. He won the Nobel Memorial Prize in Economics in 2000.

[5] See https://en.wikipedia.org/wiki/Sanford_J._Grossman (accessed 8 October 2021).

[6] See https://en.wikipedia.org/wiki/Marc_Nerlove (accessed 8 October 2021).

[7] Arnold Zellner was an economist and statistician, and a professor of economics at the University of Chicago from 1966 until his death in 2010. He was a fellow of various societies, including the American Statistical Association, which he was also the President of in 1991.

[8] See https://www.brookings.edu/experts/gary-burtless/ (accessed 8 October 2021).

[9] Institutional Review Boards, also called "ethics boards" in other countries. IRBs were formally introduced in the United States in 1974 with the signing of the National Research Act.

| Notes for section 3

[10] TA = Teaching assistant, RA = research assistant.

[11] Gary Becker was a professor of economics at the University of Chicago from 1970 until his death. He was elected a Fellow of the American Statistical Association in 1965, and won the Nobel Memorial Prize in Economics in 1992. Gary died in 2014.

[12] Schultz received the Nobel Prize in economics in 1979 for his contributions, see https://www.nobelprize.org/prizes/economic-sciences/1979/schultz/facts/ (accessed 8 October 2021.

[13] Deaton won the Nobel Prize in Economics in 2015 "for his analysis of consumption, poverty, and welfare." See https://www.nobelprize.org/prizes/economic-sciences/2015/deaton/facts/ (accessed 8 October 2021).

[14] Lazear pioneered the field of personnel economics and was founder of the Society of Labor Economists. Eddie passed away in 2020. See https://www.hoover.org/profiles/edward-paul-lazear (accessed 8 October 2021)

[15] https://faculty.crest.fr/fkramarz/

[16] https://www.anderson.ucla.edu/faculty/edward.leamer/

[17] https://www.cmu.edu/news/stories/archives/2016/december/obituary-fienberg.html

| Notes for section 4

[18] https://www.nobelprize.org/prizes/economic-sciences/1987/solow/facts/

[19] https://www.nobelprize.org/prizes/economic-sciences/1970/samuelson/facts/

[20] https://economics.mit.edu/faculty/mpiore/brief

[21] http://www.princeton.edu/~gchow/

[22] http://lapa.princeton.edu/peopledetail.php?ID=289

[23] https://en.wikipedia.org/wiki/D._Gale_Johnson

[24] https://economics.ucsd.edu/faculty-and-research/faculty-profiles/gordon.html

[25] http://sites.google.com/site/gerzoo00/home

[26] https://scholar.harvard.edu/goldin/home

[27] https://en.wikipedia.org/wiki/Aaron_Lemonick

[28] https://irs.princeton.edu/people/henry-farber

[29] https://www.nobelprize.org/prizes/economic-sciences/2021/card/facts/

[30] https://irs.princeton.edu/

Notes for section 6

[31] https://www.parisschoolofeconomics.eu/en/margolis-david/

[32] Alain Monfort is a French econometrician, fellow of the Econometric Society. Together with Christian Gouriéroux, created CREST, originally the research center of the French national statistical agency INSEE, as well as co-author of the econometrics textbook *Statistique et modèles économétriques*.

[33] INSEE is the French national statistical agency

[34] As of 2021, Abowd, Kramarz, and Margolis [3] has about 3,200 citations on Google Scholar.

[35] https://www.nobelprize.org/prizes/economic-sciences/2010/mortensen/facts/

Notes for section 7

[36] Nancy Gordon was associate director at the U.S. Census Bureau from 1995 until she retired in 2012. See https://en.wikipedia.org/wiki/Nancy_Gordon.

[37] As of 2021, Julia Lane is a professor at the NYU Wagner Graduate School of Public Service, see https://wagner.nyu.edu/community/faculty/julia-lane.

[38] James Spletzer is an economist, at that time at the Bureau of Labor Statistics, and as of 2021 with the Census Bureau's LEHD program.

[39] IPA refers to an assignment (secondment) under the Intergovernmental Personnel Act (IPA) Mobility Program. In one variant, an academic remains affiliated with his or her home institution, but works for an agency of the U.S. federal government, which in turn re-imburses the home institution.

[40] Ron Jarmin is the acting director of the U.S. Census Bureau as of October 2021, and has previously held several other lead positions in the Census Bureau.

[41] W-2 is the form used in the United States by firms to report labor earnings and taxes withheld annually to the Internal Revenue Service, the national tax collection agency.

[42] The grant proposal's abstract can be viewed at https://www.nsf.gov/awardsearch/showAward?AWD_ID=9978093&HistoricalAwards=false (accessed 4 October 2021).

[43] As of 2021, Martha Stinson is a senior economist at the U.S. Census Bureau.

[44] As of 2021, Paul Lengermann is an assistant director at the Federal Reserve Board in Washington, DC.

[45] As of 2021, Karen Conneely is an Associate Professor in the Department of Human Genetics at the Emory University School of Medicine.

[46] As of 2021, Kevin McKinney and Kristin Sandusky are a senior economists at the U.S. Census Bureau. Roberto Pedace is professor of economics at Scripps College, CA, USA.

[47] 26 U.S. Code § 6103-J, see https://www.law.cornell.edu/uscode/text/26/6103.

[48] As of October 2021, Gary Benedetto is Assistant Center Chief in the Center for Enterprise Dissemination, Disclosure Avoidance at the U.S. Census Bureau

[49] Haltiwanger is Professor of Economics at University of Maryland and, since 1987, a Research Associate in the Center for Economic Studies at Census. See http://econweb.umd.edu/~haltiwan/.

[50] https://en.wikipedia.org/wiki/DEC_Alpha

[51] Jeremy Wu was the program manager for LEHD from 2004 to 2010.

[52] Robert Sienkiewicz is currently Chief of the Center for Enterprise Dissemination at Census.

[53] New Hampshire joined the LED Partnership in 2010, marking the first time all 50 states, the District of Columbia, Puerto Rico and the US Virgin Islands had signed MOUs with the Census Bureau. https://lehd.ces.census.gov/announcements.html#121310 (accessed 4 October 2021).

[54] https://onthemap.ces.census.gov

Notes for section 8

[55] The QCEW is conducted independently by the Bureau of Labor Statistics. See https://www.bls.gov/cew/.

[56] https://www.dol.gov/general/topic/training/onestop

[57] Chet Bowie was director of the Census Bureau's Demographic Surveys Division until 2005.

[58] Erica Groshen was the BLS Commissioner from 2013 to 2017.

[59] The Household Pulse Survey has been running since April 2020, in collaboration with 13 other federal agencies. https://www.census.gov/programs-surveys/household-pulse-survey.html (accessed 4 October 2021).

[60] CIPSEA is the Confidential Information Protection and Statistical Efficiency Act, a federal law enacted in 2002 (P.L. 107–347 and 44 U.S.C. § 101), and was re-authorized in the Foundations for Evidence-Based Policymaking Act of 2018-2019 (P.L. 115–435).

Notes for section 9

[61] Ivan Fellegi was the Chief Statistician of Canada from 1985 to 2008. https://en.wikipedia.org/wiki/Ivan_Fellegi.

[62] As of 2021, Cynthia Dwork is Professor of Computer Science at Harvard University, see https://datascience.harvard.edu/people/cynthia-dwork.

[63] Steve Fienberg passed away in 2016 after a long illness.

[64] As of 2021, Jerry Reiter is professor of statistical science at Duke University, see http://www2.stat.duke.edu/~jerry/ for more details.

[65] As of 2021, Aleksandra Slavković is professor of statistics at Penn State University, see http://personal.psu.edu/abs12/ for more details.

[66] https://journalprivacyconfidentiality.org/

[67] Adrian Dobra graduated from Carnegie-Mellon in 2002, and as of 2021, is professor in the department of Statistics at the University of Washington.

[68] Lawrence H. Cox was a statistician, and the former assistant director of official statistics at the National Institute of Statistical Sciences (NISS). Larry passed away in 2016.

[69] As of 2021, Johannes Gehrke is the director of Microsoft Research. At the time, he was Professor of Computer Science at Cornell.

[70] As of 2021, Daniel Kifer is a professor in the Department of Computer Science & Engineering at Penn State University.

[71] As of 2021, Ashwin Machanavajjhala is an associate professor in the Department of Computer Science, Duke University, and a co-founder of Tumult Labs, a firm specializing in deploying differentially private data release systems.

[72] Laura McKenna, formerly Zayatz, was then head of the DRB.

Notes for section 10

[73] Cynthia Clark was associate director for methodology and standards from 1996 to 2004.

[74] Nancy Potok was Deputy Director of the Census Bureau from around 2012 to 2017, when she became the Chief Statistician of the United States.

Notes for section 12

[75] Qualtrics is a private-sector company specializing in online surveys. https://qualtrics.com

Notes for section 13

[76] The U.S. Commission on Evidence-Based Policymaking was created in 2016 and delivered a final report to Congress in September 2017. It was chaired by Katherine Abraham. For more information, see https://en.wikipedia.org/wiki/U.S._Commission_on_Evidence-Based_Policymaking#External_links.

References

[1] John M Abowd, Hampton Finer, and Frances Kramarz. Individual and firm heterogeneity in compensation: An analysis of matched longitudinal employer-employee data for the state of washington. In *The Creation and Analysis of Employer-Employee Matched Data*. Emerald Group Publishing Limited, 1999.

[2] John M Abowd, R Kaj Gittings, Kevin L McKinney, Bryce Stephens, Lars Vilhuber, and Simon D Woodcock. Dynamically consistent noise infusion and partially synthetic data as confidentiality protection measures for related time series. *US Census Bureau Center for Economic Studies Paper No. CES-WP-12-13*, 2012.

[3] John M. Abowd, Francis Kramarz, and David N. Margolis. High wage workers and high wage firms. *Econometrica*, 67(2):251–333, 1999. URL: http://www.jstor.org/stable/2999586, https://doi.org/10.1111/1468-0262.00020.

[4] John M. Abowd, Kevin L. McKinney, and Ian M. Schmutte. Modeling endogenous mobility in earnings determination. *Journal of Business & Economic Statistics*, 37(3):405–418, 2019. URL: https://dx.doi.org/10.1080/07350015.2017.1356727, arXiv:https://dx.doi.org/10.1080/07350015.2017.1356727, https://doi.org/10.1080/07350015.2017.1356727.

[5] John M Abowd, Bryce E Stephens, Lars Vilhuber, Fredrik Andersson, Kevin L McKinney, Marc Roemer, and Simon Woodcock. 5. the lehd infrastructure files and the creation of the quarterly workforce indicators. In *Producer dynamics*, pages 149–234. University of Chicago Press, 2009.

[6] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[7] Adrian Dobra. *Statistical Tools for Disclosure Limitation in Multiway Contingency Tables*. Ph.D. Thesis, Department of Statistics, Carnegie-Mellon, Pittsburgh, PA, USA, 2002.

[8] Adrian Dobra, Stephen Fienberg, and Mario Trottini. Assessing the risk of disclosure of confidential categorical data. In J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M Smith, and M. West, editors, *Bayesian Statistics 7*, pages 125–144. Oxford University Press, 2003.

[9] Timothy Evans, Laura Zayatz, and John Slanta. Using noise for disclosure limitation of establishment tabular data. In *Proceedings of the Annual Research Conference, US Bureau of the Census, Washington, DC*, volume 20233, pages 65–86, 1996.

[10] Kenneth C. Kehrer, John F. McDonald, and Robert A. Moffitt. Final Report of the Gary Income Maintenance Experiment: Labor Supply. Mathematica Policy Research Reports 51df25f673f04a369a8883ba4, Mathematica Policy Research, 1979. URL: https://ideas.repec.org/p/mpr/mprres/51df25f673f04a369a8883ba4bc00caf.html.

[11] Terence F. Kelly and Leslie Singer. The Gary Income Maintenance Experiment: Plans and Progress. *The American Economic Review*, 61(2):30–38, 1971. Publisher: American Economic Association. URL: https://www.jstor.org/stable/1816971.

[12] Julia Lane, Javier Miranda, James Spletzer, and Simon Burgess. The effect of worker reallocation on the earnings distribution: Longitudinal evidence from linked data. In *The Creation and Analysis of Employer-Employee Matched Data*. Emerald Group Publishing Limited, 1999.

[13] Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. In *2008 IEEE 24th international conference on data engineering*, pages 277–286. IEEE, 2008.

[14] National Academies of Sciences, Engineering, and Medicine. *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. The National Academies Press, Washington, DC, 2017. URL: https://www.nap.edu/catalog/24893/federal-statistics-multiple-data-sources-and-privacy-protection-next-steps, https://doi.org/10.17226/24893.

[15] National Academies of Sciences, Engineering, and Medicine. *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. The National Academies Press, Washington, DC, 2017. URL: https://www.nap.edu/catalog/24652/innovations-in-federal-statistics-combining-data-sources-while-protecting-privacy, https://doi.org/10.17226/24652.

[16] U.S. Census Bureau. Disclosure avoidance for the 2020 Census: An introduction, 2021. URL: https://www2.census.gov/library/publications/decennial/2020/2020-census-disclosure-avoidance-handbook.pdf.