From: Joel Levine <Joel.Levine@Dartmouth.EDU>
To: NSSR.65 FIFTH(AbuLughJ)
Date: 7/21/97 6:37pm
Subject: Re: Conference Invitation -Reply -Reply

Actually, I can pull it off the Web myself. To make sure that the whole thing
gets through, please check the last line. The last line should read "Getting our
discipline out into the real world is critical to the health of the discipline."

We Can Count, But What do the Numbers Mean?

Joel H. Levine
Professor of Mathematical Social Sciences
Dartmouth College
Hanover, New Hampshire 03755
joel.levine@dartmouth.edu

Prof. AbuLughod's suggested title carries double content. It is, on the surface, a question about the direction in which quantitative work in sociology will move at the beginning of the new millennium. It is also an invitation to discuss the tension in sociology that characterizes the end of the present millennium, the tension between "numbers" and ..., and what? Like C.P. Snow in his "The Two Cultures and the Scientific Revolution", I'm not sure what label to use for the other side. There are the numbers people, and there are those who are defined by their contrast to the numbers people.

I raise the matter of this tension here, at the beginning, in order to make the key point clear: The key point is that meaning lies in numbers . More precisely, meaning lies in the equations when those equations are closely related to data. In sociology it will be more true than for some of the natural sciences that meaning will lie in the equations.

Would that sociology could be easier. Would that pure mind, rhetoric, and verbal exchange could steer a true course through obstacles that are difficult to penetrate by data, by numbers, and by method. But it is not so. In the words of Emile Durkheim's Rules of Sociological Method, written at the beginning of this century:

> In the present state of the discipline we do not really know the nature of the principal social institutions, such as the state or the family Yet it suffices to glance through the works of sociology to see how it is believed that one is capable, in a few pages or a few sentences, of penetrating to the inmost essence of the most complex phenomena. This means that such theories express, not the facts, which could not be so swiftly fathomed, but the

preconceptions of the author before he began his research.

> Emile Durkheim in the preface to the second edition of The Rules of Sociological Method, 1901, Halls Translation, Free Press, 1982, page 38.

In our era, at the end of the century, much of the quantitative work is concerned with description, significance testing, and extrapolation. There is a balance between the quantitative and verbal communities. But the balance is shifting as both concept formation and theory are becoming quantitative. That which is grandly and vaguely called "sociological meaning" is shifting into the quantitative domain and that which is even more grandly and more vaguely called "sociological theory" is also shifting into the quantitative domain. The shifts are being driven by two things: Serious advances in what are called "mathematical models" but should be called "theory," and rapid decrease in the cost of data describing individuals and events.

Mathematical Models

To understand the advances in models, it is necessary, first to understand that the quantitative community recognizes, within itself, at least three different strategies each of which appropriates the word "model" for different purposes. There is first, and best known, the statistical strategy embedded in the canon of our statistical methods classes. Second, there is the deductive strategy, strongly represented in game theory, decision theory and aspects of rational actor theory. And third, there is the inductive strategy, inductive modeling, strongly represented in network analysis and stratification models.

All three strategies use the word, "model." Within the first, within the statistical strategy, "model" refers to equations like the linear equation and the Gaussian distribution. In Carl Sagan's words, science is skepticism, and in our discipline the statistical doubling of description with significance testing is the strongest barrier among the fragile defenses that separate sociology from pre-scientific speculation and from absorption by ideological agendas. For detecting the presence of relations and for defining criteria of reasonable doubt the statistical disciplines are an intellectual and practical triumph.

By contrast, in both inductive and deductive models a model is theory. The distinction between statistical models and theoretical models is a matter of degree. It lies in the extreme detail with which a theoretical model is compared to data. Outside out side of certain areas of biology, it was never intended that the line and the Gaussian distribution be interpreted precisely, in all their details, as theoretical models. As theory, every part of a theoretical model has meaning.

Let me illustrate by translating the linear equation and the Gaussian distribution into words as if they were theoretical models. To see the difference between one use of a model and the other begin simply, as if you

were once again a bright undergraduate. Take a group of undergraduates, preferably undergraduates with experience in natural science, and show these undergraduates the scatter diagram supporting one of sociology's "big" reliable correlations. Show them something like the correlation between the social economic status of a son and the social economic status of his father. Then talk about these data using the word "line" in your description and watch the undergraduate reaction: They see no line in the scatter diagram of these data and, in detail, they are right. Even with strong correlations like "r=.4," which are strong by sociology standards, even with extreme correlations like "r=.8," which are rare, there is no line in these scatter diagrams. It takes statistics to find the correlation in such data and to verify its presence. There is a positive correlation between the status of the father and the status of the son, absolutely. But a "line"? No.

This level of scrutiny becomes appropriate only when you use statistical models as if they were theories. Perhaps, 50 years ago, sociologists might have believed that they could hack their way through the underbrush of large data bases piece by piece, accumulating "variance explained". They would assemble more complicated models, piece by piece, 15% of the variance explained by one independent variable, 10% more explained by another, 10% more by a third, gradually eliminating uncertainty. And then, presumably, the predictions of sociology would approach certainty as "explanation" approached 100%. That hasn't happened, of course. It was a misreading of these models that would have led to the expectation that it might. And it would be a mistake to become discouraged because it has not.

What would it mean if the line and the bi-variate normal distribution were interpreted as theoretical models? What would it mean if social mobility, status of son compared to status of father, were normally distributed? As theory, the details of these models are alive with meaning, every symbol has meaning for the scientist. Even the "x" in a linear equation is a hypothesis: To write "x" is to say that there is a "space" in which people move. Usually, and surely with the bi-variate normal, to write "x" is to hypothesize that the space in which people move is one dimensional. That is meaning, probably not valid for social mobility, but that is meaning expressed in an equation.

The detail of the Gaussian equation, interpreted as theory, says more. In detail, the equation is a hypothesis about which events are most frequent (those for which the status of the son is equal to the status of the father) and about the rate at which frequencies decrease as a function of the distance between father and son. It says, specifically, that the frequencies decrease as a negative exponential function of the square of the distance between them.

Those are numbers and equations, the square and the exponential, that have meaning about nothing less than the "nature" of social space, or it would have that meaning if the statements were valid. To see the meaning invested in the exponent, 2, consider the square laws of physical space. In physical space the 2 of the inverse square laws goes hand in hand with the "nature" of physical space. The easiest physical example is the attenuation of light, Figure 1.

[Image]

Figure 1

The Attenuation of Illumination Per Unit of Area as a Function of the Distance from the Source

The light falling on a unit square one unit from the source diffuses to illuminate four unit squares at a distance of two units from the source. Therefore it is reasonable to expect that the illumination on a single square will diminish with the square of the distance. "Reasonable" is not enough, of course. It has to be checked. But the point is that the number, 2, in the inverse square laws is rich with meaning. The number is a statement about the dimensionality of the space, three dimensions, and about the process of diffusion. The Gaussian equation with its inverse exponential squared distance "law" would be a theoretical statement about social space - if it were correct.

That's the translation if you use the normal distribution literally, as a social model. "Social space", it says "exists." "Social space", it says, "is one dimensional". And "social mobility," it says, scatters or diffuses in some sort of dependence on the square of the distance. If you means these things literally, as a theory, then the whole package, the equation, the meaning, the interpretations, and the theory - are in jeopardy when the equation is compared to the data.

Fortunately, it fails. "Fortunately" because if it were a good fit we would have to live with the meaning. It fails - not because it is not very good at variance explained. The line and the Gaussian are very good at explaining variance, but they make systematic errors that show that some part or all of the package is wrong. It fails because it systematically predicts that too few people will move, fewer people than are actually found to move when the model is compared to the data. Therefore, all or part of the package - social space, one dimensionality, squared distance - is wrong.

That is what makes a model a theory. But, as I suggested, this is not the end of the story. The current chapter of the story is that some models have begun to work. The disputes among those of us who work with log linear models, and with crossings models, and with special interaction effects continue. The meanings of the equations we write have changed. And the models have begun to work - for some cases, for some data. Some of the current theoretical models work well enough to exhaust the information in the data.

In some cases, with some occupational data and with some social network data, here is what we now know, tentatively:

Fix number 1: The Gaussian model, if it were correct, would mean that social space was one dimensional, at least for the occupational data and the network data to which it applied. That is unlikely. It is reasonable to expect one dimension, status, to dominate the pattern of social distance, but it would be equally surprising if one dimension were sufficient. Changing the "x," to an " ," - which implies coordinates in two or more dimensions, works - better.

Fix number 2: The Gaussian model, if it were correct in two or more

dimensions would include a hypothesis about geometry. Never mind "Euclidean geometry" as it is taught in mathematics. In our science "Euclidean geometry" is a hypothesis about the effect of combining differences in two or more dimensions. It predicts that the strength of the combination will be found, first, by computing the squares of the two differences, second, by adding the two squares together, and then, third, by taking the square root of the result. (That is the Pythagorean theorem, translated). That is a truly bizarre model of the way to combine differences in two or more dimensions. It is all the more bizarre because it actually works - in physical space. Imagine the experience of intellectual triumph that must have accompanied that discovery, 2,500 years ago.

It would also be bizarre if differences in two or more dimensions combined this way in our own science: Imagine comparing my social economic status to someone else's by computing the square of the difference between our years of education, by computing the square of the difference between our incomes, by adding the two squares together, and then by taking the square root of the result. Possible, but not obvious. An alternative would be to simply add the differences in the separate claims to status. Translated, back to the math, that means hypothesizing a city block metric, not a Euclidean metric, for social space. And that model works - better.

Fix number 3: The Gaussian model uses squared distance in the equation and it predicts too few people "staying" in the status of origin. Changing the equation from squared distance to simple distance, without the square, creates a different prediction. It changes the shape of the curves from the shape in Figure 2a to the shape in Figure 2b.

[Image]

Figure 2a Figure 2b
Back to Back Exponential Decay Functions of the Distance from the Center.

Figure 2a shows back to back negative exponential functions of the squared distance. Figure 2b shows back to back negative exponential functions of the simple distance.

Simple distance predicts greater immobility (at the center) relative to other events (off center). If it works, then the explanation of this power is a theoretical problem, a real theoretical problem: It suggests to me that we have to understand something about the diffusion of information in social networks. For mobility it may mean that job searches are, somehow, one dimensional even though the occupational system itself is multi-dimensional. Whatever the explanation, the result, that "1" referring to the first power of distance, is a statement about social process that has to be explained because it works - better.

And finally, the clincher: Join these fixes in one equation and you have a hypothesis that fits the data well enough, in some cases, to exhaust the

information present in the data. The statement is restricted: With some problems in social mobility and with some work in social networks the models match the data as well as it would be matched by going back to the source and replicating the data.

_____ deg. _____

So what? Why should sociology at large be changed by advances in stratification and networks? In part, of course, stratification and networks are central to the discipline. And in part, the mathematics of these models offers simplicity: The mathematics is simpler than verbal gymnastics, and no more difficult than the mathematics of contemporaneous statistical models used for policy and forecasting.

More important, sociology at large is affected because when one piece undergoes serious change, there are consequences. It is already the case that in some areas of stratification the language of theory is mathematics. Words can attach intuition to the equations. Words can help communicate. But the first language, in some areas, is mathematics. Question: Is social space, is social organization divided and partitioned into classes or is it, by contrast, "continuous" (stratified but undivided)? Question: Is ownership of the means of production critical to social status? The primary language for such research is mathematics. In this area discussion debate is waste unless the arguments are put in falsifiable form, as models, and tested in detail against the data.

Data

At least two pieces of the discipline are changing, models are one piece, the cost of data is another. The cost of data is not just book keeping, whether it costs 10, 100, or 1,000 dollars per subject. Cost insinuates itself into the way we think. In principle, at least, sociology is about relations among people and relations among social facts. But sociometric data that describe relations among people in real world organizations is far more expensive that survey data that describe personal attributes of a random sample of individuals.

In social psychology and network analysis much of the work uses using small groups. Much of the research uses small networks, often face-to-face groups, often set pieces like the venerable "bank wiring room" data first analyzed in the early 1930's by George Homans. I ask you a question: Why would anyone want to analyze the data recording petty battles of obscure individuals in odd work groups, children's camps and college dormitories? I don't think anyone cares about these things per se . Part of the interest in small groups was social concern with large events during the 1930's, as fascism grew in Germany. Social psychologists asked, what were the bases of democracy? What were the prerequisites of acceptable human organization? What tendencies led to authoritarianism?

Even in the 1930's it was never clear that small groups were the right experimental animal, but it was worth a try. Other psychologists and sociologists responding to the same moral urgency created work like The Authoritarian Personality, looking into the human mind. Other sociologists

created work like Union Democracy, looking at labor unions as organizations created for laudable purpose which, nevertheless, (in some cases) tended toward authoritarianism. And some social psychologists looked at small groups, like summer camps and school classrooms that seemed capable (in come cases) of ending up like Goulding's Lord of the Flies. Perhaps what ailed these small groups was the "model" for what ailed the larger world.

And, in any case, it would have been difficult at that time to use anything larger than small groups. The data were too expensive. It took the support of the Harvard Business School and a promising graduate student, George Homans, to collect data for the fourteen men of the bank wiring room. To this day there is a premium on detailed sociometric data, who does what to whom, in detail, and in time series, for real world groups, large or small. Too expensive.

Now, sixty-five years later, our options are different. Models, discussed earlier, are changing the nature of theory and "the Web" is changing our access to data. Today on line data bases present any scholar, indeed anyone, a report of "Who does what to Whom?" on a world scale. We have hundreds of paid observers called reporters. that's what they call themselves, but we know they are working for us: Toiling in the interest of our science, collecting the data on who does what to whom and on the nature of the relation - daily, on a large scale, with multiple observations. Citation indices give us access to the developing structure of science, literature, and politics. Newspaper abstracts update the daily activities of our subjects. On-line intelligence reports connect events and their players on a world scale. Library indices connect world problems, international organizations, lobbyists, politicians, intellectuals, and business.

Consider the arithmetic of these new data bases: A master index of world biographies contains about 6 million entries. That is an upper bound on the population that "counts" and it includes massive double counting. If the President of the United States has a biography in Who's Who in America, and in Who's Who in the East - or in Washington, or in Politics, or in any other register, - all of these are counted as separate entries among the total of 6 million. That pares the number of people who "count" considerably below 6 million. And, at risk of offending a few million poets, baseball players, and movie stars who are included among the 6 million, I'm willing to pare them from the total as well.

This leaves perhaps one million people, perhaps one hundred thousand, as the number of people who appear in these archives, about one million to one hundred thousand people who "act" on either a world scale or local stages. This is not a large number by present technical standards.

These data, including both "the Web" itself and a much larger array of electronic archives are an experimental animal waiting for our discipline to adopt it - if we dare. No need for a sociologists to pull someone aside for an interview and ask "how do you feel about so-and so?" All we need do is read as the major players signal who is in, who is out. Or, when signals fail, all we need do is read the budgets to see who is supporting whom, check the movement of armies to see who is the ally of whom, and check the obituaries for the definitive evidence of negative social affect.

The Result

I suggest that these two together, better theory and data on a new scale, can or must change the discipline. When I say they "can" change the discipline I mean to suggest that there is an opportunity, now, to do sociology as our predecessors might have wished to do it from the beginning. Parts of our statistical methodology , parts of our main-line methods of data collection were justified as brilliant improvisations, working around methodological problems and data problems that no longer constrain us.

When I say these changes "must" change the discipline, the challenge is this: Good science requires jeopardy. Sociology can go a long way collecting data, analyzing it, classifying it, comparing and contrasting it - and presenting the result to professional audiences. But good science must take chances. It must look at people and issues that are important, using names and making predictions, saying things that matter to people outside of the science. It must make statements in full public view and - here is the risky part - it must take the chance of being wrong, also in full public view.

If we to make foolish statements about Hasulak and Taylor (two denizens of the bank wiring room), few people outside of the profession would even catch the reference. There is little risk with such data.

By contrast, if we were to make a statement about institution building and coalitions in Russia, or in Central Africa, or Yugoslavia there would be jeopardy. If the statement were foolish, then there would be obvious pressure to reformulate the theories behind the statement. If the statement were prescient, then the larger world would take note. Getting our discipline out into the real world is critical to the health of the discipline.