Introduction to Computation Social Science

Mark S. Handcock

Department of Statistics
University of California - Los Angeles

SICSS-UCLA, June 20, 2023

Why Computational Social Science?

- Traditionally, science has followed the mode of theory and experiment.
- The approach in a *computational science* is to gain insight via the analysis of mathematical models implemented on computers.
  - e.g. computational biology, computational finance, GIS
- Why the need for a "new science"?
  - Pre-computing, the sciences were dominated by mathematical models
  - Because they were required to be tractable, they were compromised
  - And selective

Why Computational Social Science?

- Until the end of the 20th century, empirical research in the social sciences was centered around scarcity
- Traditionally, data was time consuming and costly to produce, store and analyze
  - probability sample surveys taken every few years
  - end-of-period official statistics
  - in-depth studies of particular places, people, or events.

Why is methodology for the social sciences so challenging?

- Measuring social processes is much more difficult
  - large variation and in so many ways
  - difficult to measure and measure accurately
  - "Hawthorne effect" on measures
- Questions often posed in terms of latent variables
  - structure simple given latent variables
  - simple structure of latent variables
  - latent variables often easier to interpret
- Understanding the structure of social relations has been the focus of the social sciences
  - complex dependencies
  - models require a strong theoretical component

Going back to a root: Paul Lasarsfeld

(1901-1976; U.S. 1933-)

- sociologist, social psychologist
- set the direction of empirical analysis in Sociology
- a founder of mathematical sociology
- combined methodological innovation with solving substantive questions

Source: Bardwell Press

Going back to a root: Paul Lasarsfeld

- The founder of Columbia University's *Bureau of Applied Social Research*
  - the prototype of the university-based social research organization
- funded by external sources: the impact of the radio, market research
  - e.g. sample surveys of the gratifications which listeners found in soap operas
  - e.g. self-selection of audiences
  - multivariate cross-tabs of survey data

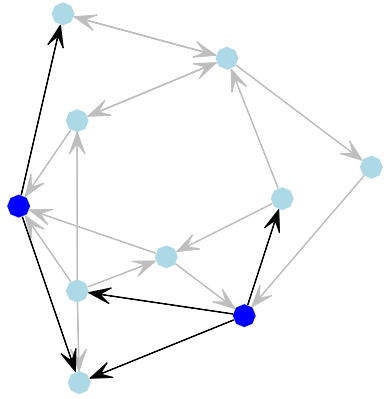
Methodology from substance

- a fundamental interest in the effects of causes: causality
- panel studies to get at cause
- How to study personal influence via media?
  - opinion leaders and followers?
  - standard surveys failed

"Snowball" sampling

- Robert Merton: "name the people who influenced you"
- A follow up interview of the influential people
- used in a panel survey of women in a Midwestern town in 1945
  $\Rightarrow$ Katz and Lasarsfeld (1955)
- Martin Trow: sociology graduate student, 1954.
  - understand the support for anti-democratic popular movements.
  - survey of the political behaviors of men in Bennington, Vermont.
  - Why is there support for Senator McCarthy?

Martin Trow, McCarthy and more



- snowball sample over the friendship networks of men
- James Coleman: In 1957, used to study the influence patterns among physicians.
- Leo Goodman (1961): Statistical analysis of snowball sampling for relations
- Patrick Biernacki and Dan Waldorf (1981): used to study individual characteristics
- Steve Thompson (2002): network sampling designs

tically apathetic people, leaving us with relatively few respondents whose responses we could use to explore more complicated processes of political communication and opinion formation.[4]

We also decided to interview only residents of Bennington Township, so as to increase the concentration of interviews within the community which comprised the actual framework and boundary for most of the networks of social relations we identified, and thus more heavily "saturate" the networks that we did follow out.

The sampling procedure developed to meet our various interests and practical problems was as follows:

1. We first secured lists of the employees of nine business firms located in the area. The largest of these employed some 140 men, the smallest about ten. These firms were chosen arbitrarily; they included the two largest local firms (both of them owned by outside interests), while the smaller firms were chosen more or less at random. The lists of employees, secured from the Village Clerk's office, included all personnel, managerial office workers, as well as craftsmen and operatives.

2. We also secured lists of the local self-employed

---

4. The final sample includes seven women who were named in the course of the "snowball" part of the sample selection. This comprises one per cent of the total sample, and does not materially affect any of the analysis.

farmers, the membership list of the Bennington Merchants Association, which includes substantially all the local merchants and small businessmen, and also a list of the male teachers and administrators of the local high school. Thus, we initially included most of the major occupational groups in the sample by plan; various other groups, such as the free professionals, store clerks and service personnel, casual laborers and farm laborers, entered the sample through the "snowball" technique.

3. With these lists in hand, a concentrated effort was made to interview all the men on each of these lists. One important by-product here was that by so doing we included in the sample specific social units,--i.e., the several firms chosen,--whose (male) members were substantially all interviewed. We thus had several small "universes" which could be used in the analysis in various ways (especially, but not only, in connection with the analysis of the networks of friendships among workmates and associates). All in all, of the total of 771 interviews collected, 359 of them were of men whose names were secured from these arbitrarily chosen lists of employees and occupational groups.

4. As the men on these lists were interviewed, they were asked to name "the three people apart from their immediate families, whom they see or get together with most often outside of working hours." Almost 80% of the respondents named three such friends, and another ten per cent named one or two friends,--

only 12% named no friends at all. The respondents were also asked to name people whose opinions on local, national, and foreign affairs they thought "especially reliable and worth listening to." All the men so named who were not already on our chosen lists and who lived in the area were added to our file of prospective respondents. And as some of these were interviewed, they in turn named others not yet in our files, who were then added as "prospective respondents."

5. We then tried to interview out along the networks thus built up in snowball fashion as far as was practicable. As the relational networks grew wider and more complex, the practical problems of maintaining our card catalogue system of completed interviews, prospective interviews, interviews assigned and "working," interviews refused or postponed, and the correlative problems of arranging for interviews, assigning and scheduling interviews and interviewers, etc. absorbed all time and energy. Thus it was not possible during the field period to decide on any rational grounds which networks should be pursued and which dropped. As a general rule, our first priority were the names on our initial lists; then we followed out along the networks on a "first named, first interviewed" basis. As a result, some networks were developed much further than others,--this quite apart from the fact that some ramified much more widely, while others quickly turned back into groups we had already interviewed. In any case, limits

of time and funds finally called for an arbitrary halt to the field work, with many "prospective respondents" in our files still uninterviewed.

The resulting sample, while not designed to be representative of any specific population, nevertheless includes representatives of all the important occupational groups, from casual laborers on one end to free professionals, plant executives and prosperous businessmen on the other. While the census occupational categories differ sharply from the ones used in the Study, and while the census breakdowns are given only for Bennington Village, whereas a third of our sample come from the outlying rural areas and smaller settled places in the area, a rough comparison is possible by combining occupational categories both in the Census figures and in the Study categories. Such a comparison shows that about 70% of the adult males living in Bennington Village in 1950 were in manual worker or lower white collar jobs such as clerical or service jobs. This compares with 52% of the members of the sample in these same broad categories, in 1954. There is little question that as compared with the distribution of occupations, incomes and education in Bennington Village, our study "oversamples" (and this by conscious design) the upper income, better educated middle class.[5]

---

5. It is interesting however, that when the Bennington

It is of course impossible to draw a representative sample by following out along lines of interpersonal relationships as we did, even if we had wanted to. However, our mode of analysis, which leans heavily on internal comparisons between different groups and categories, rather than attempting to characterize the population as a whole, precludes the necessity for a representative sample.[6]

Nevertheless, there remains a possibility that our "snowball" sample may accentuate the degree of homogeneity in the views of different groups. Preliminary analysis of the "relational" data suggests that there was very little if any homophily within educational or class categories on the attitudes that are at the center of our focus. Since the friends people named were not much more likely to share their views on McCarthy and political tolerance than people in the same socioeconomic status chosen by chance, then the "snowball" technique should not have seriously accentuated the homo-

---

sample was compared with a representative sample of the national population (as reported by Janowitz and Marvick in the Public Opinion Quarterly, Vol. XVII [Summer, 1953] in their article "Authoritarianism and Political Behavior"), the distributions of age and education in the two samples were extremely close, while the difference in the proportion of manual workers in the two samples was only the difference between 51% in the national sample and 40% in the Bennington sample.

6. See Lipset, Trow and Coleman, op. cit., pp. 425-426, on the logic of "internal analysis" in the study of social systems.

Social science methodology in the late 20th century

Handbook of Sociology $\Rightarrow$ Raftery (2008)

- Analysis and modeling of cross-tabulated data
  - Log-linear models: Social mobility
  - Latent class models:
    - model structure via latent categorical variables
    - finite mixtures of distributions
- Analysis and modeling of (social) survey data
  $\Rightarrow$ Blau and Duncan 1967
  - Regression models: Occupational status
  - Structural equation models (SEM) (Blalock 1961; Bollen 1989)
    - path analysis
    - model structure via latent continuous variables
    - factor analysis
  - Event history models (Tuma, Hannan, 1976)
    - Births, marriages, jobs
  - Logistic/probit regression
  - Hierarchical Linear Models (Bryk and Raudenbush 1992)
    - education, demography, fertility decline

Other areas:

- Economics
  - often encapsulating economic theory
  - utility based ideas
    - discrete choice models
  - models for longitudinal data
  - complex sampling designs
    - endogenous sampling
    - choice-based sampling
- Social Psychometrics
  - data analytic approaches (Gifi 1990)
  - decomposing multivariate structure
  - visual and geometric analysis
    - correspondence analysis
    - canonical analysis
    - principal components analysis

Why Computational Social Science?

- Until the end of the 20th century, empirical research in the social sciences was centered around scarcity
- Traditionally, data was time consuming and costly to produce, store and analyze
  - probability sample surveys taken every few years
  - end-of-period official statistics
  - in-depth studies of particular places, people, or events.
- That has changed: massive new and automatically collected data
- these data often provide detailed information about a large number of individuals in the population and how they relate to each other
  - volume
  - velocity
  - variety
  - exhaustivity
  - indexicality
  - relationality

big data: "It's different this time..."

- scales
  - very detailed information on individual units in the population
  - e.g., spatial, temporal, types of variables
  - the individuals are interacting: much relational information
- sensor data
  - ubiquitous spatial, temporal, types
  - e.g., embedded systems, social network updates (Facebook posts, tweets, blog posts, etc.), cellphones, apps
- types of data
  - many more types on the same units
  - diversity of structures and dimensionality
  - e.g., images, audio, video, text, geographic location markers

Beyond the nomenclature ...

- social science research could/will likely be transformed by data at a scale, form and detail that is novel
- social science research could/will likely be transformed by novel complex models implemented on computers.
- These present opportunities and challenges for the social sciences:
  - theory: e.g., new social structures developing, societal level theories
  - design: e.g., experimental opportunities, the effects of causes
  - analysis: e.g., scale, detail and form of data, informing models, testing theories
- New data can enable testing of old/new theories
- New models can more realistically represent social systems

Why Computational Social Science?

- Lazer et al (2009) redefines CSS: collection and analysis of data at a scale and in details leading to new insights into [social systems].
  - e.g. the web, sensors, governments, and commercial enterprises producing detailed information on large numbers of individuals.
- Term motivated by the changes in the type and scope of data for social research
- This is a data-analytic view, rather than a model-analytic view.

Back to Lasarsfeld ...

- Scientific Goals
  - Systematic empirical analysis of media behavior
  - Causal relations among events
  - Effects of causes: Persuasion techniques
- UCLA NewsScape Dataset Source: Joo, Steen, Li and Zhu 2015
  - Broadcasted TV news from 2005
  - 250,000 hours
  - US, UK, Russia, France, ...
  - More than 30 Outlets
  - Multimodal - Video/Audio/Text

Causal Analysis

Many social science questions are about mechanisms that underlie social behavior and social structure: questions about cause

- analysis of the "effects of causes"
  - Dorn (1956): "How would the study be conducted if it were possible to do it by controlled experimentation?"
  - counterfactual approach (Neyman-Holland-Rubin)
  - formal representations via graphical models (Pearl 2000)
- analysis of the "causes of effects"
  - in the social sciences often based on multivariate observational data
  - usually require strong theories to specify causal structure
  - SEM, path analysis: estimate strengths within prespecified structure
- In the social sciences causes and effects are usually complex.

What has changed over the last 30 years?

The growth of Political Methodology

- Many phases, but self-sufficiency in 1990s and innovation beyond statistics (2000s onward)
  - ecological inference, applied Bayesian modeling, ...
- the use of experiments in political inquiry
  - the assessment of causal questions
  - Rapid rise e.g. *Journal of Experimental Political Science*
  - voter mobilization, advertising
- concomitant sophistication of statistical methodology (King et al)
  - enhancements of standard methods: propensity score, kernel balancing
  - causality in networks of social relations

Other advances from the Lasarsfeld root?

Modeling complex dependencies:

- social phenomena with a spatial dimension
  - residential segregation (Massey and Denton 1993)
  - Stephen Matthews (1992+)
- social network phenomena
  - statistical network models (Snijders 1990+)
  - application to social epidemiology (Morris 1989+)
- microsimulation models
  - William McPhee 1962: interaction and influence in voting
  - Hanneman et al 1995: state legitimacy and imperialist policy
- agent-based modeling (1990s onward)
  - Robert Alexrod
  - Kathleen Carley, Michael Macy, et al

Connections with other advances

- Data science: "how to learn from data"
  - 50 Years of Data Science by David Donoho
  - Going back to John Tukey, 1962
- Machine Learning: "How to teach an AI to learn from data?"
  - classification, natural language processing
  - text, image and video analysis
  - often focuses on prediction
- Statistics: "How can a person learn from data?"
  - design of studies
  - representations
  - usually focuses on explanation

Bottom Line

The good:

- Advances in computing, methodology, software and data present a fantastic opportunity for social scientists
- Computational social scientists should strive to use new tools to improve their science

The bad:

- The social construction of data: "Data is!" verses "Data are made!"
- Often data are inappropriately *reified*
- Often the veracity, fidelity, generalizability is over looked
- Societal issues may impact the use of Big data
    - e.g., privacy, confidentiality, ethical perspective

The future of computational social sciences...

- will be what it has always been:
  increasingly sophisticated mathematical representations
  of social phenomena.
- These representations will be increasing:
  - realistic
  - essential
  - incorporate stochastic elements
- In support of these advances will be statistical methodology
  - to link the measured world to these mathematical abstractions,
  - with the goal of providing scientific insight.
- Social science research will change the mathematical sciences
  - new ideas to meet the challenges of social science questions
- Find a *community of scholars*, hopefully diverse, to both nurture others and yourself.

Data Studies: Data through a humanities lens

- *Data experiences*:
  - e.g., Netflix recommender system, social media feeds, hook cycles of apps
- *Data ethics*: data related issues of equality, fairness, accountability, transparency, rights, entitlements, bias and laws
- *Data justice*: changing society and data practices to enact social justice
- *Data activism*: using data to prevent harms to society