ISA eSymposium for Sociology

What should Sociologists Know about Big Data?

Cai, Tianji and Zhou, Yisu University of Macau

Abstract

The emergence of big data provides both opportunities and challenges to social scientists, and promotes a paradigmatic shift in sociological thinking. This paper elaborates how the arrival of big data will change the way sociologists approach their research interests, and discuss how social scientists—sociologists in particular—can prepare for the implementation of big data. We argue that social scientists need to learn techniques traditionally used in the computer sciences in order to inform their future research practices. The current training in the social sciences is insufficient and limits our ability to recognize and leverage the rich field of big data, thus prohibiting social scientists from playing a more active role in data-based research and the era of big data. Concurrently, social scientists need to review methodology frameworks that build upon sampling techniques and hypothesis testing, in order to develop a hybrid technique that is applicable to their field—tor example, atheoretical inductive searching followed by a hypothesis-testing deductive procedure.

Key words: Big data, Methodological challenges, Modeling strategy, Hybrid procedure

Introduction

It is the eve of the 'big data' era for social science research. Big data may come from different fields such as business (e.g. online trading records), social media platforms (e.g. Facebook, Twitter) and health care services (e.g. genetic sequencing, medical records). It can come in different formats, including traditional databases, text documents, video, audio or meta information of digital behavior. Big data is also generated at an unprecedented speed (e.g. traffic sensors) and constantly regenerating available information.

Big data promotes new research interests that were previously not feasible. For example, dynamic changes of mood over the day and the week (Golder and Macy 2011), pattern identification in social networks (Sudhahar, Veltri and Cristianini 2015) and genetic contribution for educational attainment (Okbay et al. 2016) to name a few. Big data provides a new angle to study a wide range of human activities with greater precision, but a careful survey of the current analytical landscape shows that big data also poses methodological challenges to mainstream analytical approaches. Because of its size, specialized computational facilities and software packages are needed to store, process and analyze big data (Manovich 2012). From a technical point of view, most social scientists work with data sets that are relatively small in size, ranging from a couple Megabytes to a couple thousand Megabytes, on their own desktops or mobile systems. Conversely, a typical size of big data is above 1 Terabyte (1 Terabyte =1,000 Gigabyte) and data usually are stored on cluster systems (Woodie 2014). The sheer size of data files already exceeds the capacity of most modern day desktop computers. Even with the price of data storage dropping significantly, a conventional statistical package is not capable of handling such large data files for tasks as simple as basic data management or descriptive analysis.

More importantly, big data pushes a paradigmatic shift in sociological thinking. Previous studies documented the transition in the 1970s when the social sciences adopted a scientific paradigm that focused on hypothesis-testing using statistical methods from the fields of science and technology (Kitchin 2014). As a result of quantification, statistical modeling and survey-based data collection became prominent in psychology, economics, sociology and the other social sciences (McFarland, Lewis and Goldberg 2015). While each field operates within its own sets of theories and distinctive focus, the emergence of big data offers opportunities for researchers from different disciplines, such as computer science, the social media industry and marketing, to investigate common questions—such as factors that contribute to certain behaviors (e.g. online posting, purchasing or interactions)—using similar types of data. This abundance of accessible data potentially allows researchers from different fields to study the same behaviors.

As big data changes the way we think about data, it also raises new ethical concerns. For example, private companies or government agencies that collect online posts, purchasing records, communication records and genetic information could threaten individual privacy and public security when linking information across multiple databases. The possibility of honoring promises to 'informed consent' might become diminished as data is collected instantly and involuntarily. The circumstances under which sociologists use such data for research purposes has now become a significant issue. In addition to learning analytical skills and exchanging knowledge, social scientists need to define their own roles in the data collection process.

Big data may provide new insights to aid sociologists in understanding human society, but our ability to utilize such information is currently restricted by various technical and conceptual limitations. Social scientists, and sociologists in particular, will likely experience both opportunities and challenges in this era. This paper discusses how the arrival of big data might change the way sociologists approach their research interests by asking the following questions: What should social scientists know about big data? How can social scientists prepare for and learn to use big data? To understand the current situation and provide suggestions for further research, this paper is organized into three sections. Section 1 reviews the types of big data and the differences between big data and the regular data we have been working with, Section 2 discusses the methodological issues of analyzing big data; and Section 3 provides some discussion and suggestions for social scientists who are interested in big data.

Whatestions for

Although the definition of hbig data the debatable, data scientists and industry experts typically characterize big data with three distinct features: volume, variety, and velocity. The volume of big data usually exceeds the computing capacity that current individual systems can handle. For example, social media website Twitter produces approximately 500 million tweets per day, which roughly equals 45 Gigabytes of data (Oreskovic 2015); online commerce giant eBay handles more than 1 billon transactions per day; and traditional retailers such as Walmart processes over 40 Petabytes (1 Petabytes=1,000,000 Gigabytes) of data per day, which is roughly 2,672 times the amount of information contained within all the books in the US Library of Congress (Rijmenam 2015). Even common surveys can produce large amounts of data. A recent example of the production of big data is the Health and Retirement Study (HRS) conducted by the University of Michigan. The HRS is a longitudinal study of sampled Americans over the age 50, where data has been collected every two years since 1992. The HRS began testing human genetics in 2012, and is now genotyping 2.5 million Single Nucleotide Polymorphisms (SNPs) for each of 12,507 phenotyped individuals from a current sample of

26,000 respondents (Health and Retirement Study 2014). The result is a database with a file size above 30 Gigabytes and 2.5 million genetic variables along with the standard thousands of social demographic measures, far exceeding the processing capacity for most desktop systems.

In terms of types of data, big data may contain varied forms of structured and unstructured information including text documents, transaction records, photos, video, audio, and meta-information of digital behavior such as clicks, mouse movement, scrolls, forms and more. Data can be collected from a traditional research agency (e.g. university) or government branches to fulfill research purposes. But more often, digital information is gathered by untraditional sources such as social media websites, financial institutions (Singh, Bozkaya and Pentland 2015) and sports organizations (Cervone et al. 2014).

The speed of data collection is unprecedented. In conventional sociological surveys, it could take weeks—if not months—for surveyors to collect responses from participants. A typical follow-up survey is usually six months away from the baseline study. Therefore, it is common for traditional studies to use cross-sectional data as a snapshot of social behaviors at one particular moment. But for big data, the data collection mechanism is built into the digital systems that participants use (e.g. social media). Once the infrastructure is ready, it is simple and quick to collect new information at short time intervals. Hence, big data can be collected almost in real-time. For example, Facebook users post an average of 510 comments, 293,000 status updates, and 136,000 photos every minute (Pring 2012). This constantly-regenerating new information is both a blessing and a burden for researchers. The very concept of 'population of interest' thus needs to be reevaluated. Computationally, even if we only study a snapshot from big data, special facilities and packages, and new statistical techniques are needed.

Moreover, in traditional surveys, questions are carefully designed to gauge the respondents are caattitude toward a social issue. Researchers have to confirm the validity of their questions to avoid issues such as social desirability. For big data, instead of asking what people do, it records what people actually do. For example, from analyzing onea, uch as over a period of time, a profile of political orientation can be built to predict onetes, r own roles in thn in schools. Heoutcome could be validated (Sudhahar, Veltri and Cristianini 2015). The ability to predict the outcomes of individuals' behaviors also opens up the possibility of targeted behavioral interventions. With linkages among diverse databases (e.g. purchasing history, medical records and financial information), further measures could be implemented to alter individualsulbehaviors (Brown 2016).

In summary, big data and traditional data differ not only in terms of data size, but also in terms of richness and variation of contentodiversified data mediums, numbers of variables and links can be built across several databases. Big data are also collected by a variety of organizations other than research agencies, and might be curated for purposes other than sociological research. Given these differences from conventional survey data, new data processing and analytical infrastructures will be needed.

How does big data challenge the current way of analyzing data?

Generally, data analysis refers to a process of reducing dimensionality, identifying patterns and relationships, interpreting patterns and relationships, and predicting outcomes. Handling and analyzing big data require innovative solutions for each part of this process.

Since the size of big data is beyond the capacity of traditional individual computer systems, most of the required data management, such as sorting, merging, and subsetting, has to be done interactively on cluster systems. There are solutions such as Hadoop and Hive (McAfee and Brynjolfsson 2012) that use distributed file systems and parallel processes to minimize the burden of large data files. However, these require researchers to move away from familiar territory (e.g. Stata) and learn a new data infrastructure. Training on how to use cluster systems and programming is needed. Social scientists will have to become literate in data processing on the UNIX operating system, data extraction from unstructured records (e.g.

webpages, emails and documents) and programming languages (e.g. SQL, Python and R).

Big data also raises new challenges to conventional statistical approaches. One commonly shared view among social scientists is that big data is not more informative compared to regular survey data (Japec, et al., 2015). In a sense, big data is a 'whole' dataset that contains a significantly high proportion of interesting population and behavior, but despite the raw numbers, there is actually relatively little useable and useful data, for example, in a study of social network using 'friendship' on Facebook, one Fafriendship network on Facebook can be huge in size but with little interaction between most connections. Hence, the question of how to garner useful information is extremely important (e.g. identify clusters in oneion is extremely importantally . New algorithms and statistical methods have been developed to help researchers reduce dimensionality or introduce sparsity on parameters. For example, principal component analysis (PCA) is a traditional dimensionality reduction tool that dates back to Karl Pearson in 1901 (Pearson 1901). PCA projects a covariance matrix of data into a lower dimensional space represented by eigenvalues and eigenvectors. However, it does not work very well on large sparse matrices, which is typical for big data. Although the initial dimensionality could be large, there are many holes (zeros) in the matrix, which creates numerical instability issues. To deal with large matrices with sparsity, penalties (e.g. lasso or elastic net) can be imposed in order to develop modified principal components with sparse loadings (Zou, Hastie and Tibshirani 2006). Variable selection is another example. In general, social scientists are not interested in variable selection because the number of variables involved in analysis is not large, and more importantly, variable selection as a data mining technique is not theoretically justifiable. However, a variable selection procedure is required— suchvariable selection based on False Discovery Rate (FDR), Bonferroni correction, or sparsity—when dealing with a large amount of variables, for instance, 2.5 million SNPs in the HRS dataset (Chun and Keles 2010).

Similar to dimensionality reduction, identifying patterns and correlations for big data also introduces new challenges. The generalized linear model and related techniques have been some of the most important tools in social science for identifying patterns and correlations in survey data. However, such techniques may not be suitable for big data analysis. To begin with, the number of observations is large enough to detect any tiny effect under the classic hypothesis-testing framework that is based on the assumption of sampling distribution. In addition, when the number of variables included in a model is larger than that of observations, the so-called 'curse of dimensionality' appears because the model will saturate the degrees of freedom and no hypothesis testing can be conducted. The HRS dataset can again be used as an example. The HRS contains 2.5 million SNPs for each of the 12,507 individuals in addition to the usual demographic and behavioral variables. Theoretically, each of the SNPs or any combination of SNPs could serve as independent variables (IVs) in an analysis. The model could quickly run out of degrees of freedom when these IVs enter the modeling process. Statistical inference and cross validation based on resampling methods have been proposed to deal with the large number of observations (Yu 2003). Usually, a large number of sub-samples of a reasonable sample size are selected from the dataset, and analysis is done on each of these sub-samples. A variability of estimates is obtained by summarizing estimated parameters across sub-samples (Yu 2003). To cope with the 'curse of dimensionality', a group of techniques such as variable selection or dimensionality reduction can be employed. For example, Yang et al. (2010) extended the Genomic-Relatedness-Based-Restricted-Maximum-Likelihood (GREML) method from an animal study as a dimensionality reduction tool to estimate the contribution of the large amounts of SNPs on a phenotype without variable selection.

Another major difference is the purpose of analysis. For most social science studies, researchers are primarily interested in understanding the underlying mechanisms that generate data, in other words, how findings are justifiable under the current theoretical framework or how a new theoretical framework can be developed from empirical findings. However, for other disciplines, such as the media industry and marketing (both contribute significantly to big data) knowing underlying mechanisms is less important than whether a model works. Their primary interest is to predict the occurrence of certain behaviors and promote behavioral changes, particularly for out-of-sample observations. Hence, prediction is a heavily evaluated component of model fitting. For example, gradient boosting algorithm is a machine-learning technique that has shown considerable improvements in reducing bias and increasing the accuracy of predictions (Schapire 2003). For binary 0-1 classification cases, using a logistic model as classifier, the algorithm works through an iteratively reweighting process that gives more weight to previously misclassified observations at each step of iteration. The final classification is obtained by taking a weighted average of predictions from each step (Schapire 2003). When new data is added, as is the case in big data, predictions could be further improved. Gradient boosting has been implemented in predicting online dating matches—an area in which many sociologists have shown interest—and had shown superior performance compared to other sociological theories based on models (Chen 2013).

Many big data analyses in the media or marketing industries involve solving real-time problems such as price biding, purchase suggesting or traffic controlling (Barlow 2013). Therefore, a dynamic modeling and predicting system is particularly useful. In contrast, social sciences rarely engage in this type of analysis—ngagemodelings are static, post-hoc and theoretically orientated. In the era of big data, social scientists will inevitably adopt analytic techniques from other disciplines, and this may lead to changes on both theoretical perspectives and research frameworks in the social sciences. With more and more big data available in the field, a shift of research focus might take place. Should the social sciences switch from a theory-driven strategy to a purely data-driven one? Or a hybrid strategy that uses new techniques, such as machine learning, as search tools for potential explanations and the development of predictive models? These options are becoming more viable in the social sciences, but new research strategies are still subject to debate. Yet, it is quite certain that change is inevitable.

How can we prepare for big data?

Compared to traditional survey data, big data is an array of various types of data from diverse fields. Analyzing big data requires special skills that are not currently taught systematically in social science graduate programs. We believe that social scientists need to learn from other disciplines to remain competitive and relevant in the big data era. Currently, disciplines such as computer science lead the way in advanced data analysis. As big data begins to proliferate in sociological studies, an existential question for our field is posed: what unique value can we offer? A typical interdisciplinary collaboration between social scientists and researchers from other disciplines on big data seems to comprise a clear division of labor, where social scientists are responsible for theory and others conduct analysis (McFarland, Lewis and Goldberg 2015). We hope that more subsequent collaborations would facilitate a deeper integration of academic fields, although a full adoption of the othertion of lsate onduct and res

Social scientists may still stay in their comfort zone and ignore what is happening outside of their disciplines. But as we see growing interest, both within and outside academia, regarding new possibilities enabled by big data, can we afford to remain oblivious? If we fail to catch up, will we gradually become less attractive to students, the general public, and other science and social science disciplines? Auguste Comte insisted that sociology, as the 'science of society'cimust adopt a scientific methodology used in the natural sciences with rigorous and objective scientific investigation and prediction. Accordingly, the emergence of big data promotes a paradigm shift from a knowledge-driven to a data-driven research (see details in Hey et al., 2009).

The influence of big data has already shown an effect on analytical strategies in sociological studies. Although when compared to other social sciences such as economics, sociology is more inductive and relies more on data-driven methodstive and arise from empirical results—sociologists are also more cautious of being atheoretical. Therefore, many sociologists have suggested the possibility of an 'iterative combination of atheoretical induction and theory-led deduction' (McFarland, Lewis and Goldberg 2015). In a study of the impact of labor force participation on the trajectory of cognitive functioning decline after retirement, the relationship between the two is argued to be influenced by behavioral, social and biological factors. Oneocialis on eduction and is discovered to contribute to both his/her labor force participation and the change of cognitive functioning after retirement. To control the confounding effect of genetic architecture on labor force participation, a genetic propensity score is usually estimated using logistic regression with a limited number of candidate SNPs, and included as a control in sequential models that focus on the relationship between labor force participation and trajectory of cognitive functioning decline. However, a biased or less accurate genetic propensity leads to a biased estimate for the effect of labor force participation. A machine learning algorithm with ensemble methods such as boosting, bagging and blending can be used to improve the accuracy of genetic propensity scores, and in turn produces a less biased effect for labor force participation. In such cases, analysts used both conventional sociological models and machine learning techniques.

We believe sociologists can prepare to work with big data in the following ways. First, we, as social scientists, need to decide what could be learned from other disciplines. Driven by the push by big data, some fields (mainly computer sciences) are clearly pioneers in the big data era. Analyzing big data involves skills that are not currently covered in standard social science curriculums, such as programming, knowledge of cluster systems, methods based on resampling techniques, basic text mining, etc. Even though a formal coverage of all these topics is not necessary, some exposure (e.g. colloquium and workshops) should at least be provided.

Secondly, we need to develop a framework that allows researchers to combine abductive, inductive and deductive approaches to understand a phenomenon (Kitchin 2014). For instance, an inductive procedure is employed to develop a deductive hypothesis, and then be evaluated. It sounds like a reasonable hybrids-data mining reveals some hidden truths that are not suggested by theory, and only those of theoretical potential are further investigated. However, this framework challenges the current practice of analysis that hypothesis testing and statistical inference are abductive/deductive and exploratory analysis is inductive. There are many uncharted areas for this framework, both in theory and practice. For example, in a study of gene and environmental interaction on educational attainment, using a dataset such as the HRS with 2.5 million SNPs, it is not feasible to run a regression model with all possible interactions of SNPs and a key predictor. The regular modeling process involves a variable selection on SNPs using educational attainment first, according to criteria such as FDR or Bonferroni correction. Subsequently, only SNPs that satisfy the criteria, such as those with a p-value less than 10^{-8} (extremely high predictive power) are tested in a separated replication. The significant ones are then selected and a series of interaction models using educational attainment as dependent variables are estimated. However, selected SNPs with supposed high predictive power might only explain less than 1% of variation in educational attainment and might not have any meaningful connections to biochemical and sociological mechanisms. A better modeling strategy that maximizes predictive power and produces theoretically meaningful results is needed.

In addition, we need think about what we can offer and how we can translate our perspective into big data analysis. For example, fundamentally, sociologists examine relationships and connections among individuals, social groups, communities and institutions, and recognize that individuals' experiences (e.g. beliefs, values and expectations) and social interactions are shaped by structural forces. Allowing others to understand the sociological perspective is difficult, if not impossible, and translating it into big data analysis takes extra effort. Instead of taking prediction as an end-of-data analysis, new inquiry that aims to build meaningful and useful theoretical constructs from identified patterns and connections is necessary. This may likely lead to a new division of labor in an interdisciplinary collaboration where social scientists take a more active role in negotiation and interaction with researchers from other fields.

Conclusion

Each day, about 2.5 Exabytes (1 Exabytes=1,000 Petabytes) of data are produced from diverse fields (IBM 2016). An explosion of the amount of data available along with business opportunities is changing the way knowledge is produced. It has far-reaching consequences on individuals, communities and societies. Many argue that the emergence of big data is promoting a paradigm shift across disciplines. At present, researchers from multiple disciplines such as computer science, engineering, the media industry and health services, work on this rich amount of data through the use of different perspectives.

What is the role of social sciences in this process? How can social scientists contribute? The answers to these questions involve defining new directions of inquiry in the big data era. Technical solutions from engineering focus more on practical values than scientific explanations, which social scientists emphasize. Nevertheless, social scientists need to learn techniques from the computer sciences because the current training in social sciences limits our ability to recognize the richness and vastness of big data, and thus prohibits us from playing a more active role in the era of big data. At the same time, we need to review our methodology that builds upon sampling techniques and hypothesis testing, to develop a hybrid model, for example, atheoretical and inductive searching followed by a hypothesis-testing deductive procedure.

References

Barlow, M. (2013). Real-Time Big Data Analytics: Emerging Architecture (1 edition). O'Reilly Media.

Brown, M. S. (2016). Big Data Analytics and The Next President: How Micro targeting Drives Today's Campaigns. Retrieved September 9, 2016, from http://www.forbes.com/sites/metabrown/2016/05/29/big-data-analytics-and-the-next-president-how-microtargeting-drives-todays-campaigns/

Cervone, D., D'Amour, A., Bornn, L., and Goldsberry, K. (2016). A Multiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes. Journal of the American Statistical Association, 111(514), 585–599.

Chen, L. (2013). A social matching system : using implicit and explicit information for personalized recommendation in online dating service (Thesis). Queensland University of Technology. Retrieved from http://eprints.qut.edu.au/64157/

Chun, H., and Kele , S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society. Series B, Statistical Methodology, 72(1), 3–25.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). The Annals of Statistics, 28(2), 337–407.

Golder, S. A., & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and day length across diverse cultures. Science (New York, N.Y.), 333(6051), 1878–1881.

Health and Retirement Study: A Longitudinal Study of Health, Retirement, and Aging. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI. [WWW Document], URL http://hrsonline.isr.umich.edu/ (accessed 09.10.16).

Hey T, Tansley S, and Tolle K (2009) Jim Grey on eScience: A transformed scientific method. In Hey T, Tansley S,Tolle K (eds) The Fourth Paradigm: Data-Intensive Scientific Discovery, Redmond: Microsoft Research, pp.xvii–xxxi.

IBM - What is big data? (2016, September 7). Retrieved September 10, 2016, from https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Usher, A. (2015). Big Data in Survey Research AAPOR Task Force Report. Public Opinion Quarterly, 79(4), 839–880.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data & Society, 1(1), 2053951714528481.

Manovich, L. (2012). Trending: The Promises and the Challenges of Big Social Data. In M. K. Gold (Ed.), Debates in the Digital Humanities (pp. 460–475). University of Minnesota Press. Retrieved from http://minnesota.universitypressscholarship.com/view/10.5749/minnesota/9780816677948.001.0001/upso-9780816677948-chapter-47

McAfee, A., and Brynjolfsson, E. (2012). Big Data: The Management Revolution. Retrieved September 9, 2016, from https://hbr.org/2012/10/big-data-the-management-revolution

McFarland, D. A., Lewis, K., and Goldberg, A. (2015). Sociology in the Era of Big Data: The Ascent of Forensic Social Science. The American Sociologist, 47(1), 12–35.

Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., Benjamin, D. J. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. Nature, 533(7604), 539–542.

Oreskovic, A. (2015). Here's another area where Twitter appears to have stalled: tweets per day. Retrieved September 10, 2016, from http://www.businessinsider.com/twitter-tweets-per-day-appears-to-have-stalled-2015-6

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. Philosophical Magazine Series 6, 2(11), 559–572.

Pring, C. (2012) 100 social media statistics for 2012, The Social Skinny. Retrieved from http://thesocialskinny.com/100-social-media-statistics-for-2012/

Rijmenam, Mark van. (2015). Datafloq - Big Data at Walmart is All About Big Numbers; 40 Petabytes a Day! Retrieved September 10, 2016, from https://datafloq.com/read/big-data-walmart-big-numbers-40-petabytes/1175

Schapire, R. E. (2003). The Boosting Approach to Machine Learning: An Overview. In D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, & B. Yu (Eds.), Nonlinear Estimation and Classification (pp. 149–171). Springer New York. Retrieved from http://link.springer.com/chapter/10.1007/978-0-387-21579-2_9

Singh, V. K., Bozkaya, B., and Pentland, A. (2015). Money Walks: Implicit Mobility Behavior and Financial Well-Being. PLOS ONE, 10(8), e0136628.

Sudhahar, S., Veltri, G. A., and Cristianini, N. (2015). Automated analysis of the US presidential elections using Big Data and network analysis. Big Data & Society, 2(1).

Woodie (2014) "What Is Big Data" Question Finally Settled? (2014, October 29). Retrieved September 9, 2016, from https://www.datanami.com/2014/10/29/big-data-question-finally-settled/

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. Nature Genetics, 42(7), 565–569.

Yu, C. H. (2003). Resampling methods: Concepts, applications, and justification. Practical Assessment Research and Evaluation, 8(19). Retrieved from http://pareonline.net/getvn.asp?v=8&n=19

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse Principal Component Analysis. Journal of Computational and Graphical Statistics, 15(2), 265–286.

Cai, Tianji is an assistant professor in the department of Sociology at University of Macau. His research interests center on two areas: social mechanism of how biological and social factors influence behaviors, and developing quantitative research methods. Specifically, he is interested in the integration of sociology with biological factors in the studies of sociological issues such as social and health behavior, stratification, and social network. The current project in research methods focuses on the issues of sampling weights in multilevel/ longitudinal models, for example, evaluating the effects of ignoring or incorporating sampling weights on the estimation of multilevel models under various sampling designs, and developing methods to test the informativeness of the sampling weights. tjcai@umac.mo

Zhou, Yisu is an assistant professor in the Faculty of Education at University of Macau. He is interested in quantitative social sciences in general. His research expertise is educational policy in greater China region. His past research projects use large-scale assessment data both internationally and domestically, covering issue such as social environment of learning, teacher education and labor market, and social stratification in schools. He is currently working on social segregation in schools in Chinese societies.