Chapter 1
Introduction

Andrew Pilny and Marshall Scott Poole

For many young group researchers, learning about advanced statistical methods can be quite the traumatic experience. Coupled with teaching, professional development, and being theoretical experts in their domain, fine graining the ins and outs of inferential statistics seemed like just another task on a full plate of work. Fortunately, for many of us, there was a rescuer. In 2000, Andy Field published his first book, *Discovering Statistics Using SPPS for Windows*, beginning a series of volumes dedicated to making statistics seem both easy and fun. Clarity was essential for Field, whose volumes always provided relevant examples (usually very humorous), clear screenshots, and example write-ups. Field's volumes were vital for not only learning about statistics, but reducing anxiety and uncertainty the complexities of inferential modeling.

However, the world has changed greatly since then, moving into what is generally referred to as the era of Big Data. Four characteristics generally characterize Big Data (Gandomi & Haider, 2015): (1) *volume* (i.e., bigger size and magnitude), (2) *variety* (i.e., more different types of data), (3) *velocity* (i.e., rate at which data is created), and (4) *complexity* (i.e., complex data structures that require cleaning and integration). But Big Data is not just about *data* per se, it is also about a new way thinking about measurement (King, 2016). For instance, instead surveying groups about their networks, we can now collect their interactions via their cell phones, email, and social media (i.e., trace data). Unfortunately, one of the consequences of Big Data is that many of the methods detailed by Field, which were exclusive variants of the general linear model, are either inappropriate or unsuited for much of the data we have on groups today. For instance, for data on online groups (e.g.,

A. Pilny (✉)
University of Kentucky, Lexington, KY, USA
e-mail: andy.pilny@uky.edu

M.S. Poole
University of Illinois, Urbana, IL, USA
e-mail: mspoole@illinois.edu

A. Pilny, M.S. Poole (eds.), *Group Processes*, Computational Social Sciences,
DOI 10.1007/978-3-319-48941-4_1

coordination in Wikipedia), there can be millions of data points, which can results in nearly every independent variable tested being statistically significant. Likewise, interaction data from group members assumes a type of interdependence that violates many assumptions inherent in linear inference.

To address these issues, many researchers have called upon a paradigmatic change in thinking, largely referred to as *Computational Social Science* (CSS) (Cioffi-Revilla, 2013; Lazer et al., 2009). Computational social science is an interdisciplinary endeavor specifically tailored to handle the complexity of Big Data by merging together social science problems with computer science methods. As Wallach (2016) puts it, CSS can be thought of as research being undertaken by groups of "social minded computer scientists and computationally minded social scientists" (p. 317). The impact of CSS on group research has been especially notable. For instance, the new range of tools and thinking behind CSS has provoked innovative ways of understanding different group dynamics (e.g., Klug & Bagrow, 2016; Shaw & Hill, 2014) and collecting group data (e.g. Madan, Cebrian, Moturu, Farrahi, & Pentland, 2012; Radford et al., 2016).

Although the outlook of CSS is promising for the future of group research, there is a looming problem (e.g., Alvarez, 2016): for all the new work being produced using CSS methodology, there are few explicit avenues available to actually teach these methods. In other words, pedagogy is has taken a back seat to publication. The result is a sort of knowledge concentration or what boyd and Crawford (2012) refer to as a digital divide between the small minority who have access to Big Data and CSS resources and the majority who do not. Indeed, there are few graduate seminars, workshops (often expensive if they do exist), or handbooks that make it easy for the average social scientist to excel at CSS.

What is needed, therefore, is an "Andy Field book" for CSS, a resource to help *demystify* these methods and make it accessible to anyone willing to follow the white rabbit of CSS. To accomplish this goal, a resource would need to do several things. First, it would need to emphasize a *didactic*, rather than an inquiry-laden focus. That is, the primary objective is teaching rather than theory generation or original contribution to research. Second, it would need to be *transparent*, which means that codes and data should be shared and presented in a tutorial fashion. Transparency is vital in an age where we see social science continuing to be criticized for a lack of replication and secrecy regarding data and code. And finally, the resource should be *encouraging*. The spirit behind such an endeavor should reflect a growing notion that the more scholarly use of these methods, the better. As such, opaque and ambiguous language, equations, and procedures should be avoided in order to foster an environment that enables and empowers researchers to carry out a similar analysis.

These three values represent the spirit behind this book. The authors were given a relatively open format to write their chapters as long as it corresponded to a didactic, transparent, and clear avenue for anyone to pick up and take off with. The diversity of these chapters are quite evident: some are longer than others (e.g., Chap. 4: Relational Event Modeling), some introduce needed theoretical introductions (e.g., Chap. 6: Social Sequence Analysis), some use computer code (e.g., Chap. 2:

Response Surface Modeling), some use graphic interface programs (e.g., Chap. 5: Text Mining; Chap. 3: Bayesian networks) and some may not even use data at all (e.g., Chap. 8: Computational simulation).

Although no book on introducing CSS methods will be exhaustive, we aimed to provide the audience with what might interest group researchers the most. For instance, the growth of machine-learning is arguably one of the most dramatic changes in inferential modeling during the last twenty years (Hindman, 2015). Machine-learning models are often better equipped to handle Big Data because they are not dramatically influenced by sample size, often do not make crude normality assumptions, and have clear interpretations that explicitly acknowledge when the model predicts both accurately and inaccurately. As such, we included two chapters that explore different machine learning algorithms, Bayesian networks (Chap. 3) and decision-trees (Chap. 7).

Likewise, there has been a renewed increase in group dynamics that openly acknowledges time and order. In this case, group researchers can begin to seriously consider *dynamic* rather than *static* notions of emergence (Kozlowski, Chao, Grand, Braun, & Kuljanin, 2013). As such, Chap. 4 focuses on group interactions by viewing networks as relational events (i.e., episodic interactions), rather than relational states (i.e., enduring relationships). In this sense, relational event modeling can reveal dominant patterns of interactions by predicting ordered and even time-stamped histories of group interactions. Chapter 6 similarly focuses on time and order, but highlights social sequences of activities. One of the highlighted example of such a technique is that it can determine if group members behave in a synchronized pace (i.e., entrainment), provoking an important inquiry as to whether the emergence of group level properties are related to group performance.

It also important to recognize the new types of data that can be exploited by CSS methods. One example is the growing advent of analyzing text as data. In this sense, Chap. 5 explores text mining procedures and the development of semantic networks represented by co-occurrence relationships between different words and concepts. Sometimes there is not enough data or something was missing from data measurement. Chapter 9 deals with this through computational simulation with empirical data. Finally, sometimes we have enough data on groups with repeated observations that we can run quasi-field experiments. Chapter 2 adapts response surface methodology, a common method in the natural and physical sciences, to group research.

Lastly, as Alvarez (2016) notes, CSS is "developing at a dizzying pace" (p. 25). While researchers are rapidly developing tools to provide unique and sometimes ground-breaking insights into social inquiry, there is a need to pause and give back. Many of the tools used by CSS researchers were not developed individually in a vacuum. We owe a debt of gratitude to those who developed and taught us these methods, and owe it to the next and current generation of CSS researchers to share knowledge on how to use these methods. It can be seen as a sort of methodological "pay-it-forward". This book is one small attempt at such an endeavor.

References

Alvarez, R. M. (2016). *Computational social science: Discovery and prediction*. Cambridge, MA: Cambridge University Press.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society, 15*(5), 662–679.

Cioffi-Revilla, C. (2013). *Introduction to computational social science: Principles and applications*. London: Springer.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144.

Hindman, M. (2015). Building better models prediction, replication, and machine learning in the social sciences. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 48–62.

King, G. (2016). Preface: Big data is not about the data! In R. M. Alvarez (Ed.), *Computational social science: Discovery and prediction* (pp. vii–vi1). Cambridge: Cambridge University Press.

Klug, M., & Bagrow, J. P. (2016). Understanding the group dynamics and success of teams. *Open Science*, *3*(4), 1–11.

Kozlowski, S. W., Chao, G. T., Grand, J. A., Braun, M. T., & Kuljanin, G. (2013). Advancing multilevel research design capturing the dynamics of emergence. *Organizational Research Methods*, *16*(4), 581–615.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., … Gutmann, M. (2009). Life in the network: The coming age of computational social science. *Science*, *323*(5915), 721.

Madan, A., Cebrian, M., Moturu, S., Farrahi, K., & Pentland, A. (2012). Sensing the "health state" of a community. *IEEE Pervasive Computing*, *11*(4), 36–45.

Radford, J., Pilny, A., Reichelmann, A., Keegan, B., Foucault-Welles, B., Hoyde, J., et al. (2016). Volunteer science: An online laboratory for experiments in social psychology. *Social Psychology Quarterly*, *79*(4), 376–396.

Shaw, A., & Hill, B. M. (2014). Laboratories of oligarchy? How the iron law extends to peer production. *Journal of Communication*, *64*(2), 215–238.

Wallach, H. (2016). Computational social science: Towards a collaborative future. In R. M. Alvarez (Ed.), *Computational social science: Discovery and prediction*. (pp. 307–317). Cambridge University Press.