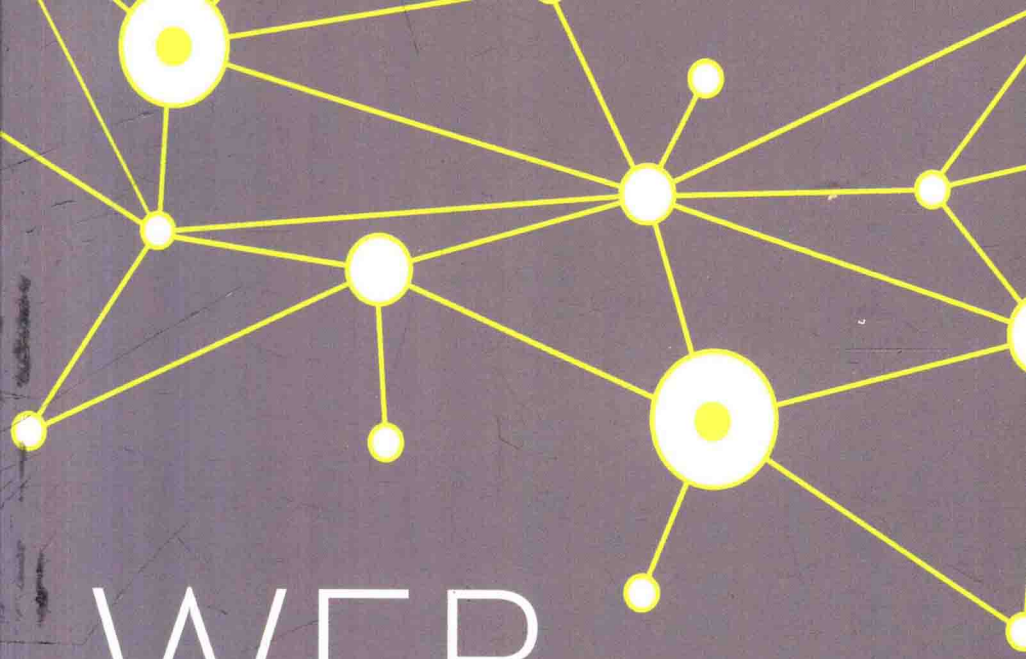

WEB SOCIAL SCIENCE

CONCEPTS, DATA AND TOOLS FOR SOCIAL SCIENTISTS IN THE DIGITAL AGE

ROBERT ACKLAND

WEB SOCIAL SCIENCE

CONCEPTS, DATA AND TOOLS FOR SOCIAL SCIENTISTS IN THE DIGITAL AGE

ROBERT ACKLAND

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd
3 Church Street
#10-04 Samsung Hub
Singapore 049483

First published 2013

Editor: Chris Rojek
Editorial assistant: Martine Jonsrud
Production editor: Katherine Haw
Copyeditor: Richard Leigh
Proofreader: Jonathan Hopkins
Indexer: Cathryn Pritchard
Marketing manager: Alison Borg
Cover design: Lisa Harper
Typeset by: C&M Digitals (P) Ltd, Chennai, India
Printed and bound by CPI Group (UK) Ltd

Library of Congress Control Number: 2012950465

British Library Cataloguing in Publication data

A catalogue record for this book is available from
the British Library

ISBN 978-1-84920-481-1
ISBN 978-1-84920-482-8 (pbk)

WEB SOCIAL SCIENCE

For Kazuko

List of Figures

List of Tables

List of Boxes

About the Author

Robert Ackland is an Associate Professor in the Australian Demographic and Social Research Institute at the Australian National University. He has degrees in economics from the University of Melbourne, Yale University (where he was a Fulbright Scholar) and the ANU, where he gained his PhD in 2001. Prior to commencing his PhD, which was on index number theory in the context of cross-country comparisons of income and poverty, Robert worked as a researcher in the Australian Department of Immigration and an economist in the Policy Research Department at the World Bank, based in Washington, DC, 1995-97. Since 2002 Robert has been working in the fields of network science, computational social science and web science, with a particular focus on quantitative analysis of online social and organisational networks. He has given over 50 academic presentations in this area and his research has appeared in journals such as the *Review of Economics and Statistics*, *Social Networks*, *Computational Economics*, *Social Science Computer Review* and the *Journal of Social Structure*. Robert leads the Virtual Observatory for the Study of Online Networks project (http://voson.anu.edu.au), and teaches on the social science of the Internet, statistics, and online research methods. He has been chief investigator on four Australian Research Council grants, and in 2007 he was both a UK National Centre for e-Social Science Visiting Fellow and James Martin Visiting Fellow at the Oxford Internet Institute. In 2011, he was appointed a member of the Science Council of the Web Foundation's Web Index project.

Preface

This book aims to provide students, researchers and practitioners with the theory and methods for understanding the web as a socially constructed phenomenon that both reflects social, economic and political processes and, in turn, impacts on these processes.[1] Specifically, readers of this book will:

- learn about relevant data, tools and research methods for conducting research using web data;
- gain an understanding of the fundamental changes to society, politics and the economy that have resulted from new information and communication technologies such as the web;
- learn how Internet data are providing new insights into long-standing social science research questions;
- understand how social science can facilitate an understanding of life in the Internet age, and how approaches from other disciplines can augment the social scientist's toolkit.

There are three main motivations for social scientific research into behaviour on the web. First, it can be argued that behaviour on the web is a unique cultural form that deserves to be documented and understood. This is more a perspective taken in virtual ethnography, for example, and it is not central to the present book. Second, it may be the case that some behaviour on the web is similar enough to offline behaviour, that its study can provide new insights into the offline behaviour. This perspective is a strong one in this book. Finally, social science research into web behaviour has been motivated by the need to understand whether certain online behaviour may have effects in the 'real world', that is, that there may be a *social impact* of the web. This motivation is again an important aspect of this book.

Since the early days of the Internet there has been research into its social, political and economic impacts, with contributions from a range of disciplines: media and cultural studies, communications, economics, political science, sociology, law and public policy. So what does this book offer that is new or different? What sets web social science apart from other approaches for studying the web?

---

[1]The inventor of the World Wide Web, Tim Berners-Lee, advocates the use of 'Web' when referring to the proper noun (see http://www.w3.org/People/Berners-Lee/FAQ.html#Spelling), but for convenience, this book uses 'web' throughout.

SHAPING FORCE OR SOCIAL TOOL?

On the face of it, it would appear that the Internet has had a huge influence on the way we live our lives – it would seem that it has transformed the way we work, collaborate, engage in commerce, participate in the political process and interact socially. However, it is useful to keep in mind the words of the social historian Claude S. Fischer:

> Visions of new technologies revolutionizing the way we live are often bold, sweeping, and millenarian. They are exciting to hear; they sell books; they can earn one a good living on the corporate lecture circuit. But their shelf-life is roughly equivalent to that of a 'Big Mac' ... we ought to think more about these technologies as tools people use to pursue their social ends than as forces that control people's actions. (Fischer, 1997: 115)

This is not to say that the Internet has had *no* social impacts, but rather that the real-world impacts of the Internet are likely to be complex and hard to disentangle from other major forces, for example demographic and economic, that are affecting patterns of communication and community.

So if we agree with Fischer's view (in 1997, note) that the Internet has not controlled people's actions but has rather provided tools to allow people to pursue their social ends, the implication is that the Internet can provide new data sources for studying human behaviour. As a tool for communication, the web differs from the telephone in that interactions between people often lead to digital trace data that can be used for research: email repositories, archives of posts on forum sites and blog sites, and profiles in social network sites such as Facebook.

RECOGNISING THE DISCIPLINARY APPROACHES TO STUDYING THE INTERNET

One of the major aims of this book is to identify the disciplinary approaches to studying the web: what does a social science approach to studying the web involve, and how is it different from (or similar to) other disciplinary approaches? What is included under the banner of *social science*? The book identifies social science by the core fields of economics, politics and sociology, with a focus on research that is quantitative and/or grounded in network theory and methods.

When applied physicists look at the web, they see a massive network of web pages and the hyperlinks between them – a source of data for measuring and understanding large-scale network properties such as power laws. For information scientists, the web represents a huge scientific citation network, ripe for the application of bibliometric techniques to understand scholarly

output and impact. Media and communications scholars are interested in the web as a channel for distributing news and opinion; they study the production and consumption of web content in the context of understanding opinion leadership and agenda setting. When social scientists look at the web, they see individuals and organisations interacting socially, economically and politically; often this behaviour can be described as occurring in networks.

All of these disciplines have made important contributions, and the question that needs to be asked is: is there a need for web social science? Do we need to – by using the term *web social science* rather than, say, *Internet research* or *Internet studies* – effectively establish a boundary or demarcation line that will include some disciplines and exclude others? The use of the term *web social science* does not necessarily involve boundary marking: there are a lot of people who study the Internet who do not identify themselves as social scientists, and, similarly, there is a lot of research into the Internet which, while focusing on social aspects of the Internet, is not recognisable to social scientists as being social science. The mere fact that the behaviour being studied is social does not make it sociology.

Why in this era of interdisciplinarity (encouraged by universities and funding councils) should there be a book seeking to promote the social science of the Internet? Is this not simply helping to maintain the disciplinary silos, rather than break them down? Even though the Internet was built by engineers and there has been a lot of influential work by applied physicists and computer scientists, it is important for social scientists to take an active role in the development of new approaches to studying the Internet. Otherwise they may be bound by other disciplines' tools, frameworks and research questions, and thus constrained by the world view of other academic fields.

So, while this book draws from these other disciplines, the primary goal is the advancement of social science. While future advances in our understanding about the digital and real worlds will continue to involve the collaboration of social scientists and other disciplines (as the emerging field of computational social science promises; see, for example, Lazer et al., 2009), social scientists still need to be able to make contributions on their own terms, and this book aims to equip them to do so.

THE ROLE OF NETWORKS

In the above characterisation of how the web is viewed by scholars from physics, information science, media studies and social science, the word *network* appeared several times. This book emphasises quantitative network research methods (in particular, social network analysis). Already hugely influential in disciplines such as applied physics, network science is said to be the science of the twenty-first century. While sociologists have been studying social networks for 40 years, the use of networks in other social

science disciplines is still limited, often being used in a metaphoric or heuristic sense but without formal rigour or empirics.

But to understand a lot of interesting behaviour on the web you have to understand networks. This book reflects the increasing use of formal network concepts and methods to understand the structure of the web from a social science perspective. However, it is not assumed that readers are already familiar with network theory or methods: the book provides an introduction to network theory and methods for social science web research. Further, it should be noted that not all chapters in the book are focused on research involving networks. While just about any interaction between individuals and organisations can be represented as a network, often that behaviour can be more appropriately described and understood without using network concepts.

THE ROLE OF QUANTITATIVE RESEARCH

This book also emphasises the role of quantitative research in the social science of the Internet. A particular focus is on digital trace data (data that are unobtrusively collected from the Internet); however, the book also covers quantitative analysis of obtrusive online data (e.g. online surveys) as well as (to a lesser extent) offline surveys of households and individuals.

CAUSALITY VERSUS CORRELATION

Identifying causal relationships is one of the main challenges for empirical social science. The web offers opportunities to overcome methodological limitations in social science by enabling, for example, natural and field experimentation that can help discern causation and would be impossible to conduct in the real world. But the movement of social life onto the web, and the concomitant realisation that social networks are key to much of the behaviour that is of interest to social scientists, have exposed us to new methodological challenges with regard to identifying causality. The issue of causality arises in several chapters of this book and is one of the core methodological challenges for web social science.

CONSTRUCT VALIDITY OF WEB DATA

Social scientists are interested in the web as a source of data that can potentially provide new insights into long-standing questions in social science. But online relationships and social behaviours may not have quite the same meaning, or dynamics, as they do in the real world. The fact that Facebook decided to use the term 'friending' to describe the act of two people making

a connection on the site does not necessarily mean that Facebook data are appropriate for studying homophily (the idea that people become friends with others who share similar attributes, or 'birds of a feather flock together'), for example. This book emphasises the importance of establishing that there is an appropriate *mapping* from the online to the offline, or, in other words, that web data have *construct validity* for the context in which they are being used.

THE WEB AS RECONFIGURING FORCE

The web has been described as a force that can 'reconfigure' or radically alter the status quo in various areas in the real world. This book considers the arguments and evidence relating to how the web can potentially: enable non-government organisations to leapfrog government and commercial interests in engaging with the public; level the political playing field in favour of minor or fringe political actors; enable protesters to more effectively circumvent government authority; democratise access to scientific expertise; or increase sales diversity as firms profit from the Long Tail.

SAMPLING IS STILL IMPORTANT

Sampling from a target population is one of the core techniques of empirical social science. In a research environment such as the web, where there are potentially millions of observations in the target population, one would think that devising appropriate sampling techniques for digital trace data would be at the forefront of the minds of methodologists. A lot of social science research into the web requires greater understanding about the units of observation (be they websites, Twitter users, Facebook users) that can be garnered via automatic methods alone. But oddly, sampling has not been a large focus of methods development for web research, and one wonders to what extent this has been due to the influence of other disciplines where a research design is not considered exciting or worthwhile unless it is viable at 'data scale'. The fact that numerous web data are in the form of networks (hence exhibiting interdependence) makes sampling less straightforward than in some real-world social science settings where observations can be validly assumed to be independently distributed of one another. However, this book emphasises the fact that sampling is an integral part of web social science.

BRINGING IT ALL TOGETHER

Finally, this book is different in that it aims to bring all these elements into a single text that equips readers with the tools, theory and methods for conducting web social science.

While the primary audience is social scientists, the book will appeal to a range of disciplines. The quantitative/mathematical nature of parts of the book will draw scientists and engineers interested in learning how social scientists are studying the web. The web is now an important tool for organisational marketing and communication and coordination (e.g. non-government organisations building virtual communities for collective action, and corporations wishing to measure their online brand presence). So this book will also be useful for practitioners in many areas.

The author has given over 50 academic presentations in this area since 2002 and a number of times has heard audience members say that they had never thought of the Internet as a research area or source of data, and that the presentation had opened up new possibilities. The major goal of this book is to open eyes both to the Internet as a source of new data for social science research, and to the insights and tools that social science provides for understanding life in the Internet age.

STRUCTURE OF THE BOOK

The book consists of an introductory chapter (Chapter 1), which introduces web social science and the major themes covered in the rest of the book, followed by nine chapters arranged in two parts.

Part I is about how to do web social science. It introduces readers to the methods, tools and data for researching behaviour on the web and for using the Internet as a delivery medium for online research tools (e.g. for studying offline behaviour via the web). It consists of three chapters:

- Chapter 2 is an introduction to online research methods.
- Chapter 3 introduces social media network analysis.
- Chapter 4 is focused on the analysis of hyperlink networks.

Part II provides examples of web social science that draw on the methods, data and tools introduced in Part I. The examples illustrate how social science has been advanced using data from the web and the contribution of empirical social science to greater understanding of the social, economic and political impacts of the Internet.

- Chapter 5 looks at homophily and the closely related topic of social influence, showing the inherent difficulty in identifying these phenomena and how web data are providing new insights.
- Chapter 6 is focused on organisational collective behaviour on the web – in particular, the hyperlinking behaviour of non-government organisations.
- Chapter 7 looks at the potential influence of the web on politics from the perspective of the visibility of political information and the engagement of citizens in the political process, and the tendency of people to cluster

on the basis of political affiliation. The chapter also covers the impact of the Internet on social connectivity more generally.

- Chapter 8 is concerned with government in the digital age, looking at how web data can be used to conduct comparative analysis of the information provision activities of government, the impact of social media on government authority and the potential of virtual worlds for public policy research.
- Chapter 9 looks at how the Internet has transformed production and collaboration, and our ability to study and measure these activities. Internet-enabled peer production and its relation to information public goods is discussed. The chapter also investigates how the web has changed our ability to measure scholarly output, and the distribution of scholarly visibility and authority. Finally, the use of data from virtual worlds for studying social networks and individual performance is covered.
- Chapter 10 is concerned with Internet marketing and commerce, first reviewing the concept of the Long Tail and related empirical evidence, and then looking at both how Internet marketing is leveraging social networks and the potential impact of online recommender systems on sales diversity.

ACKNOWLEDGEMENTS

I would like to thank Heather Crawford, Francisca Borquez and Susannah Sabine for assistance with editing and proofreading various versions of the manuscript. I would also like to thank Paul Vogt for helpful comments on an earlier version of the manuscript, and Chris Rojek, Martine Jonsrud and Katherine Haw from SAGE for helping me to write this book.

I also thank my students in the Social Science of the Internet stream of the Master of Social Research and PhD programme at the Australian National University, who have contributed much to my thinking about this topic, and who have been reading very underdone versions of chapters since 2008.

I would also like to thank colleagues with whom I have written papers or research software, or who have otherwise encouraged and inspired me in this new phase of my career. I wish to especially thank: Sean Batt, Bruce Bimber, Noshir Contractor, Bill Dutton, Rachel Gibson, Mathieu O'Neil, Rob Procter, Jamsheed Shorish and Jonathan Zhu.

Finally, I want to thank Kazuko for the constant encouragement and gentle nudging that helped me to finish this project.

Contents