Digital By-Product Data in Web 2.0

Digital By-Product Data in Web 2.0: Exploring Mass Collaboration of Wikipedia

By

Zeyi He

CAMBRIDGE SCHOLARS PUBLISHING

Digital By-Product Data in Web 2.0:
Exploring Mass Collaboration of Wikipedia,
by Zeyi He

This book first published 2012

Cambridge Scholars Publishing

12 Back Chapman Street, Newcastle upon Tyne, NE6 2XX, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-4174-9, ISBN (13): 978-1-4438-4174-0

In the memory of my lovely Grandpa
and unremitting me

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

This book provides the core methodology of using digital by-product data in web 2.0 environments, and the key process of using this method in an approachable and "user-friendly" style. Via the case study, this book also gives many interesting substantive conclusions about Mass collaboration in Wikipedia.

In response to a methodological challenge in social science research, especially linked to studies of online phenomena including Web 2.0 applications, this book proposes a new methodology that is aided by digital by-product data. Digital by-product data is the data created by the internet operating system to back up content including the browsing history, files downloaded, photos uploaded and so on. With the emergence of Information and Communication Technologies (ICTs), our daily life is becoming digitalized and can be described and even profiled by digital by-product data. This book demonstrates that using digital by-product data is an important opportunity to help social scientists overcome various bottlenecks such as the deficiency of data, the ineffective analysis, and possible risks of bias in the existing research methodology. Proposition about the new methodology is based on the discussion and analysis of the current data environment of social science research, the online environment and the existing data mining and analysis methods in scientific areas.

The experimental research of the book uses digital by-product data to explore online phenomena, and to evaluate the utility of applying such a methodology. After considering the availability of the data resources, the diversity of the data types, the usability of the data, and the research value of the subject, Wikipedia was chosen as the case study. The substantive research of Wikipedia is accomplished by using the digital by-product data that is provided by Wikipedia to analyse its collaborative mode in which millions of participants work together to provide the knowledge-sharing system, the online encyclopaedia. The research is constructed in such a way that three related questions are addressed in a step-by-step manner. We aim to answer whether there is a collaborative model in Wikipedia and if so, what it is and how it works. In the process of answering the questions above, we describe the existing dynamics of mass collaboration; build a model of the collaborative model; explain the approaches and ratio

of contribution by the various participants; and furthermore analyse the administration system as well as its policy to deal with editing conflicts. Finally, the results of these experimental researches are displayed in different ways, including the use of mathematical equation, metrics and visualization.

Using digital by-product data provides a series of benefits to resolve the contemporary methodological challenge in social science and extend the capabilities of social scientists to investigate online phenomena. In a review of our empirical studies, the book also illustrates the working flow of using digital by-product data from a social science perspective, and provides practical lessons to guide investigators in order for them to avoid the mistakes and problems that are encountered by the author. Through studying an actual social phenomenon, this research is dedicated to evaluate the possibility and feasibility of using a new methodology, which makes use of the neglected data resource to improve the productivity of social scientists in their research endeavours.

The author considers the book is the adventure of combining scientific methods with social scientific proposal. Within the book, data mining, data analysis, statistics, visual analytics etc. have been implemented to social research journey. It is clear; when facing contemporary issues, the boundary between disciplines is vague or even vanished. As modern scholars, we should discard existing prejudice to create new methodologies in a new information age.

ACKNOWLEDGEMENTS

I give my thanks to all my family, friends and colleagues for their love and support.

I owe my love and thanks to my dearest grandfather, Mr. S. Fan who encouraged me while he fought against cancer. His trust and support installed courage in me to explore the bigger world.

I have a thankful heart to my parents Xinghua Jiang and Pingping Fan, for offering me a nurturing environment where I have turned out the way I wanted to. With their hortative education and unconditional support, I faced all criticism and distrust bravely yet believed in myself. I also give many thanks to Dr. Kan, who continued to support and encourage me when I was in serious self-doubt. Her support lit the fire for me to fight for the future.

I would like to give my thanks to B. Xu for helping me resolve all language issues with patience and responsibility. Her help and support has been unparalleled and irreplaceable in my last year of PhD and at the time of writing this book. From hours in the morning to many nights staying up with me, her company and support are appreciated.

I would like to give my appreciation and thanks to all senior professionals who generously offered their talent and enthusiasm to help me throughout my research and publish this book.

I give grateful thanks to Professor Roger Burrows, for inspiring me with distinctive and sparkling research ideas, for supporting me to introduce the novel quantitative methodology to my research. Also, I thank Professor Andrew Webster for his patient and advice. His strict, careful and meticulous attitude toward research has greatly helped me. My thanks go to Professor Mike Savage, who offered valued and useful suggestions to correct this work. His appoint of view enlightened me in the digital data application field. My appreciation also goes to Mr. Brian Loader, whose spotlighted idea inspired me at the beginning of my research career and who offered me a warm welcome to the University of York; and Dr. Emma Uppriand, who kept reminding me that “you will be fine” with a smile every time when I felt frustrated and hopeless.

I am indebted to all technical and professional specialists: Dr. Bo Wang from the Mathematical Department; Dr. Rose from the Computer Science Department; Mr. Eric, the execution chair of Wikimedia

Foundation; Jimmy Wales, the founder of Wikipedia, Graham Lewis, Gillian Sinclair, Louis Rose, Mark Hewitt, Peter Halls, and Tom Rryan. I owe my gratitude to all my friends who gave me inspiration through discussions or comments. They are Dan Mercea, Xinya Zhu, Claire Schofield, Wen Wang, Jing Cui, Joanna Rossiter, Andrew Eggleston, Janette Cranney, and Dr. Charles Russ.

CHAPTER ONE

METHODOLOGICAL CHALLENGES FOR SOCIAL SCIENCE RESEARCH IN THE INTERNET AGE

With the rapid advances in information and communication technologies (ICTs), the routine life of individuals is becoming increasingly reliant on such technologies (Boase and Wellman, 2004). Life, in what has been described as "virtual society", has transformed life styles, modes of communication, interaction with others, and generated social and psychological changes (Woolgar, 2002). ICTs therefore unsurprisingly, offer the exploration of particularly valuable and novel research topics related to the virtual society, for example: as a participant platform; associated blurred geographic boundaries; as a new medium for self-expression; improved cost effectiveness, and expanding social networks (Michalak and Szabo, 1998). Social scientists are interested in exploring the multitude of interactions mediated by the internet (Banks, 2001, Couper and Miller, 2008, Hine, 2005, Reips, 2000), and the series of new social issues engendered by the internet (Illingworth, 2001, Schiller, 1994, Wright, 2005). From a social science methodological perspective, the emergence of numerous novel areas of inquiry via the internet has produced new challenges. In response to such challenges social scientists have been, and still are, seeking more efficient and effective methods. Even with such efforts, we discover that the methodological challenge still presents itself unabated and perhaps grows more intense. We suggest that such a methodological challenge is caused by the emerging technical environment and changing society, and is further aggravated by the delayed reaction of social scientists to the changing environment and a lack of effective solutions.

ICTs have been discussed on many fronts. The phrase is used to refer to the emergence of internet social networks (Sohn, 2008), "knowing capitalism" (Thrift, 2005), the democratizing of manufacture (Hippel, 2006), internet-cultural communication (Kluver et al., 2007) and the "virtual society" (Woolgar, 2002). The primary task for this book is to

study how we can respond to the methodological challenge of exploring the internet and so how best to conduct internet-based studies.

Soon after its emergence researchers took advantage of the internet as an updated and convenient means of communication to implement traditional sampling methods (Hine, 2005, Sheehan and Hoy, 1999). The extensive development of ICTs has transformed conventional communication means, both spatially and temporally; extended the dynamics of our lives; and enabled new types of "society" (Hine, 2005, Stanczak, 2007). Internet-mediated sampling methods have expended the range of social research to understand and describe online phenomena, with their low cost, accessibility to respondents and high-tech means of information collection. However, it is argued that internet-mediated sampling methods might exacerbate the negative impact of sampling methods (Savage, 2009, Savage and Burrows, 2007, 2009).

As internet-based communication gained increasing popularity, more social scientists started to use the internet as a means to spread their message; attract participation from respondents; and communicate with one another in order to accomplish their research. Their work continued to follow traditional sampling methods. Although this marriage between these old methods and the internet as a means of communication have adequately made use of some of the advantages of the internet, it is thought that such techniques may be limited from a methodological perspective (Savage and Burrows, 2007, 2009, Smith and Kollock, 1999). These commentators recognize the salient features of traditional research methodologies, yet they also express concern about the limitations of relying solely on such traditional methods, especially when research topics have undergone tremendous socio-scientific developments during the same time period. Undoubtedly, the traditional sampling research methodologies used have an important role in traditional social science research. Most contemporary social science research, especially that which focuses on observing phenomena of the traditional society, such as criminology, women's studies, and communication studies still rely heavily on sample surveys and qualitative interviews (Sayer, 1992, Silverman, 2004). This illustrates that the use of traditional research methods is still the dominant approach in academia (Outhwaite and Turner, 2007). However, in addressing the problems and phenomena that arise from the internet, the limited use of the internet as a communication tool to conduct research inherits the limitations of traditional sampling methods.

In order to explore the notion that conventional sampling methods bear intrinsic and extensive limitations on the investigation of internet phenomena, we reviewed three main methods: email-based survey, web-

based survey and online interview (Hampton, 1999, Pitkow and Recker, 1995, Schmidt, 1997, Schonlau et al., 2002). Then we put forward the argument that these research methodologies suffer from the same problems as general sampling methods. Savage and Burrows (2007) point out that traditional sampling method may have shortcomings when used to deal with online research topics. For instance, some apparent weaknesses of sampling methods include a low response rate and the fact that they ignore new data forms (Smith, 1997, Wright, 2005). Therefore, it becomes obvious that these attempts to use modified sampling methods and utilize the internet as a medium to implement traditional methodologies are ineffective when studying internet phenomena.

Subsequently, researchers identified previously unused data available on the internet in various forms. Facing rapidly changing topics of emerging research in the ICT environment and with various available internet applications, many pioneers began to explore alternative methodologies to retrieve new types of data to conduct social science research (Brunn and Dodge, 2001, Fisher et al., 2006, Park and Thelwall, 2003, Viegas et al., 2004). These pioneers provided a brief discussion of the necessity to support the use of new data forms in social science research. As a result, the possibility arose of using internet data to accomplish social research using new methodological perspectives: this, in fact, will be the main task of this book.

When we try to gain a comprehensive view of the methodological challenges that hinder studies of the internet and internet-based behaviour, we invariably face a problem with causality. While ICTs provided an unprecedented platform of social interaction and as a consequence, novel social patterns, the development of such technologies also facilitates social science research with new data and approaches. First, social science research is changing to encompass internet-related topics with the development and popularity of ICTs. Second, ICTs themselves expand and speed-up forms of communication, which allows social science researchers to get in touch with target research subjects and obtain responses more easily. Third, the digitisation of data as supported by the platform can help social scientists to organise, aggregate and analyse information. Finally, because of the way in which the internet functions in producing and transferring masses of data every second, those data transactions themselves could become a unique and useful source of information for social scientist. The advantages of ICT systems and the internet for social scientist are mainly expressed in two different schools of thought and research explorations. In order to understand these two concepts, we need to clarify the difference between internet-mediated

sampling methods and internet-new-data methods. After the comparison, we will confirm the value and importance of these new methods of research using data, and we will continue to discuss the possibility of using these methods to conduct social science research from both practical and methodological perspectives.

In summary, this chapter discusses strategies in response to the new internet environment. First we investigate the new topics and new social pattern that arose from the development of ICTs, as these changes inevitably lead to challenges in research methodology and ways of thinking for social scientists. Second, after introducing three of the main internet-mediated sampling methods we focus on discussing the problems and the limitations of these research methodologies. Third, we study a new source of data and the possibilities opened up by utilising this type of data. We then introduce some research based on this type of data. Fourth, we discuss the problems and difficulties of this research methodology and its advantages and disadvantages, and how to maximize its advantages. Finally, based on our research question, we develop a blueprint and outline a proposed research methodology.

Internet-based sampling methods

The social impact brought about by ICTs and other technologies associated with the internet has long been recognized, and it has been discussed comprehensively in several influential publications (Featherstone, 2006, Hine, 2006). In studies of the changed society propelled by new technologies, most social scientists chose to adopt a more conventional and conservative approach (Banks, 2001, Hine, 2005). They preferred to use new technologies as tools to make observations and obtain data.

Internet social science research involves the investigation of relationships (Boase and Wellman, 2004), interactions (Baym, 1995) and even societies (Baym et al., 2004). In order to explore early attempts at internet mediated sampling methods, we need to clarify three points: the sampling method applied; related limitations; and any biases of a particular method. These researchers have outlined the attempts made by social scientists in response to the methodological challenge in studying internet phenomenon. Although there are still limitations in the application of these preliminary attempts, as a well-established and widely used research methodology, sampling methods provide a demonstration of how social scientists can respond to the methodological challenges of the information era. This section introduces how ICTs provide social science research with more diverse research topics, and how it transformed

modern living and means of communication. At the same time, we explore how social scientists have made use of the internet as a medium to conduct research on new topics (Hine, 2005). This section introduces two primary sampling methods which use the internet as a medium to conduct research on topics related to the internet. We mainly focus on introducing previous attempts made by social scientists to use the internet as a medium to complete the sampling survey process. Empowered by multimedia functions, the internet provides various means, such as email and websites, to make connections between people, and these are also the two main ways being adapted to recruit sampling respondents (Anderson and Gansneder, 1995, Best et al., 2001, Schonlau et al., 2002, Wright, 2005). We will first discuss the email-based method and then focus on how to use the web to disperse information and invite people to take part in surveys. This section will give an overview of how social scientists have used the internet to develop their sampling methods.

Applying internet-mediated methods in studies of the internet

Based on common knowledge and scholars' experiences, it is clear that traditional sampling methods with telephone or postal-mediated communications are less effective for researching internet-related topics (Best et al., 2001, Coomber, 1997, Evans and Mathur, 2005). This situation has led to the emergence of an altered form of sampling method, which uses the internet as a medium to collect data from surveys or interviews (Best et al., 2001, Wright, 2005). Scholars have started to consider initiating their invitations to participants in surveys using the internet as the primary medium (Braithwaite et al., 2003, Schonlau et al., 2002).

As ICTs gained more and more popularity and recognition among academics, most scholars saw that the internet can be used as a legitimate platform to deploy sampling methods (Hine, 2005, Smith and Kollock, 1999). Internet-based sampling methods have become the main methods for research because it is believed that data can be generated more quickly, less expensively and more easily than using traditional methods, such as paper-based surveys (Coomber, 1997, Evans and Mathur, 2005, Hox and Leeuw, 1994). Generally, internet-based methods exceed traditional ones in producing data for research about the internet for the following three reasons:

When using the internet to distribute surveys and questionnaires, scholars are able to reach a wider audience, at a reduced cost (Evans and

Mathur, 2005, Selm and Jankowski, 2006, Sepulveda, 2006). Under these circumstances, the benefit of using internet-mediated communications is apparent, as it greatly facilitates the delivery of information to potential respondents. In fact, it may give access to respondents that would be impossible to otherwise reach (Wright, 2005). Another obvious advantage of internet-mediated methods is that they lower the cost of the design and distribution of surveys (Evans and Mathur, 2005, Sepulveda, 2006).

Unlike traditional social scientific surveys, internet-mediated methods preserve the anonymity of the participants and therefore it is easier to recruit respondents who may be reluctant to participate using traditional survey techniques (Selm and Jankowski, 2006, Sepulveda, 2006). Internet research normally explores online behaviours, motivations and psychological experiences (Wright, 2005). If online participants are allowed to maintain their internet anonymity or identify themselves by their online personas, they may not experience potential uneasiness that may be a facet of traditional postal surveys or face-to-face interviews (Frankel and Teich, 1999). In other words, online participants may be more willing to respond by hiding their real identity (Schonlau et al., 2002).

Internet-mediated methods can gain access to particular people for studies of internet-related topics (Hine, 2005). Not only may potential respondents be apprehensive about joining traditional mediated surveys or interviews, but scholars launching such surveys might be concerned about how to encourage potential respondents to participate. Several studies have used the internet to study behaviours of the people who use internet-based media as their main social platform (Hine, 2008, O'Connor and Madge, 2001, Riehle, 2006). Researchers believe it is better to use the same communication medium to arouse the subject's motivation to participate (Hine, 2005). For example, if scholars want to explore social networks on Facebook, it would be difficult to use traditional sampling methods to carry out that research.

The email-based survey

The email-based survey has become an important means to facilitate the sampling process in social science (Ilieva et al., 2002, Sheehan, 2006), but it is only used as a replacement to the traditional mail survey in order to contact respondents (Sheehan and Hoy, 1999). Many researchers have carried out surveys via email distribution (Hampton, 1999, Smith and Kollock, 1999). Email-based surveys have significant advantages over more traditional techniques, due to their simplicity, reduced cost, and the

potential for wide dissemination (Ilieva et al., 2002, Michaelidou and Dibb, 2006, Sheehan, 2006).

Technically, an email-based survey has a similar working process to the traditional postal survey, and does not require definite new strategies for social scientists to implement (Coderre et al., 2004, Ilieva et al., 2002). This might be one of the reasons that scholars are attracted to this method (Hine, 2005, Schonlau et al., 2002). Additionally, email-based surveys can considerably reduce the cost of research that arises from distributing questionnaires and conducting interviews, when compared to pen-and-paper surveys and face-to-face interviews (Coderre et al., 2004). As one scholar puts it, "the most attractive aspect of this technology probably remains the elimination of the time and costs associated with the separate step of entering information into a computer for data analysis" (Hampton, 1999 p.50). Researchers believe that the use of email-based questionnaires can avoid the extra costs associated with face-to-face interviews and pen-and-paper surveys, such as travel expenses (Lampe and Resnick, 2004) and labour costs (Sheehan and Hoy, 1999).

Based on internet technologies, the email-based survey has a huge potential in attracting attention from a large number of potential respondents (Anderson and Gansneder, 1995). In most internet communities, email is the major, if not only, manner for scholars to communicate with sampling group(s). In most cases, people involved in online communities only share online contact details such as email address or instant message user names. As a result, the easiest way to invite them to participate in a study is by email (Ellison et al., 2007). Therefore, an email-based survey has an advantage in communicating with people who live in an internet-centred environment. Such a survey is also able to encourage potential participants to respond by sending subsequent reminder emails (Ellison et al., 2007).

Given the astounding growth of its application in the social science, the email-based survey also presents enormous potential for examining online communities, such as Facebook (Ellison et al., 2007), Wikipedia (Majchrzak et al., 2006), and Myspace (Dwyer et al., 2007). Studies have also applied this method to the examination of professional activities (Bane and Milheim, 1995, Kovacs et al., 1995, Schiller, 1994); the exploration of new technologies (Miller, 1994), and investigating the use of internet-related technologies. Unlike the web-based survey, which will be discussed below, the email-based survey is usually used to study a selected small group of samples (Parker, 1992, Tse et al., 1995). Additionally, some studies combine email contacts with web-based surveys to target a particular group of online users, by sending email

invitations which contain the hyperlink to the website-hosted survey (Ellison et al., 2007). To some extent, email-based surveys were the initial attempt to apply internet technologies to social science research (Dwyer et al., 2007). This adventurous attempt has been expanded to other internet-technical formats, as we will discuss next.

The web-based survey

The web-based survey, as another method of conducting surveys, is represented by questionnaire or real-time conversation on webpages (Coomber, 1997, Gosling et al., 2004, Ilieva et al., 2002, Schmidt, 1997). Like the email-based survey, possible advantages of using a web-based survey include the reduced cost of accessing respondents and collecting data (Coomber, 1997, Reips, 2000, Schmidt, 1997). The increased popularity of the internet has also accelerated the use of web-based surveys in social science fields (Manfreda and Batageji, 2002, Schmidt, 1997).

Although web-based interviews require a basic technical understanding of web browsers and databases, there are many readily available tools for researchers to produce such surveys (Schmidt, 1997). In addition, web-based interviews also require a conversational platform supported by technical software (Coomber, 1997). Because designing the supporting software is often too difficult for people without a strong technical background, social scientists normally make use of existing software (Manfreda and Batageji, 2002). Such software provides a platform for communication between interviewees and interviewers via a website, which enables them to complete the interview regardless of the location of each participant. Additional support comes from the database system in the server which records and stores each sentence of the interview in real time, which can benefit subsequent reviews (Chen and Hinton, 1999).

Similar to the email-based survey, the web-based survey enables people to provide responses quickly and easily regardless of geographic distance (Chen and Hinton, 1999). Advances in ICTs offer an inexpensive way to collect sampling data from interviews or questionnaires (Ilieva et al., 2002, Kaplowitz et al., 2004). Thus, web-based surveys have gained immense attention and interest among individuals with little financial or technical support (Couper and Miller, 2008). From the discussion above, it is apparent that the web-based survey shares similar features with the email-based survey, such as: no constraints on geographic distance; effective distribution; and economic efficiency. These salient features have

helped promote the web-based survey to become more accepted and used more widely within social science research.

Moreover, the web-based survey can incorporate software to check responses, re-format results and share the outcomes in scholarly environments (Couper and Miller, 2008), which could theoretically increase efficiency. The automatic checking features of web-based surveys can eliminate missed or invalid responses and transcription errors. This greatly increases the efficiency of data collection and analysis (Ilieva et al., 2002, Schmidt, 1997). For instance, web-based surveys can be designed to exclude any unacceptable responses, which will remarkably reduce the number of meaningless samples. At the very least, they can be designed to allow the manual exclusion of invalid samples before the formulation of results (Couper and Miller, 2008, Couper et al., 2001).

The web-based survey provides a friendly and favourable environment for participant involvement. This is particularly important with a real-time website-based survey. Web environments have been perceived to give respondents a sense of control, which comforts and empowers them during the survey process (Illingworth, 2001, Rettie, 2001, Simsek and Veiga, 2001). This assumption suggests that participants may feel that they are in charge of the conversation, unlike in a conventional setting where they may feel reluctant to speak out. Such a psychological sense of taking control can increase their motivation to complete surveys (Jones, 1994).

Using sampling methods based on ICTs is the first and most primitive step towards improving the methodologies of social science research. As we have suggested, ICTs bring new research topic and methodological challenges for social scientists. For research topics, ICTs provide new phenomena and social incentive mechanisms to study. From a methodological perspective, the internet can play the role of a medium to connect survey publisher and respondent; interviewer and interviewees. In the last two decades, social scientists have developed various methodologies to use the internet as a medium to communicate with potential respondents, as the majority of them have become heavy internet users.

With the steady development of ICTs and the wider popularity of such technologies among social scientists, many limitations and problems of internet-mediated sampling methods are beginning to present themselves. These drawbacks demonstrate the inherent limitations of sampling methods, especially when society and communication media are evolving alongside developing technologies, and internet-mediated sampling methods alone are becoming more and more insufficient to help social scientists observe and study social phenomena. These problems re-

emphasise the methodological challenge, and at the same time point to other areas of potential enquiry.

Methodological limitations of sampling methods

In response to the challenge that arose in an environment filled with new ICTs, sampling methods have produced impressive datasets, and have been used widely in various internet-research fields (Gutmann et al., 2009). However, as these methods mainly observe the activity and thinking of a society through a small population (Outhwaite and Turner, 2007, Sayer, 1992), internet-mediated sampling methods thus have inherent weaknesses. This section will mainly discuss the impact of such weaknesses on the accuracy and reliability of the collected data.

In the twenty-first century, most social scientists still hold the view that the traditional sampling method is one of the most important methodologies for observation, and holds an irreplaceable position for observing contemporary phenomena and monitoring change in society (Halsey, 2004). This method refers to using information collected from a small group of people as a "sample" to represent a view or behaviour trend of the entire society. Not considering new methodologies incurs a number of problems in internet-based sampling surveys and qualitative interviews (Best et al., 2001, Chen and Hinton, 1999, Schonlau et al., 2002, Wright, 2005). In fact, such methods are only a variation of the traditional methods and still belong to sampling methods which are based on self-reported data collected from a small representative population (Eagle et al., 2009, Shaffer et al., 2010). Therefore, they carry all the limitations intrinsic to those traditional sampling methods (Ards et al., 1998). Additionally, the incompatibility between some of the features of internet societies and sampling methods may cause even more severe biases (Best and Krueger, 2004, Best et al., 2001, Hartford et al., 2007). Under such a situation, we argue that the methodological challenge is more pronounced and potentially more serious in internet societal research than that of studies into "real" society.

Sampling methods based on the internet have been used in different areas, including social networks (Dwyer et al., 2007, Ellison et al., 2007), financial capital (Uzzi, 1999) and market research (Ilieva et al., 2002, Michaelidou and Dibb, 2006, Tse et al., 1995). However, the domination of sampling as the primary method to collect national demographic information is challenged in "knowing capitalism" (Thrift, 2005). Meanwhile, complex digital forms of social and cultural data are emerging (Thrift, 2005) and revolutionary products are created and disseminated in

the online environment (Benkler, 2006). Losing their previous confidence and jurisdiction, social scientists now have to face a new methodological crises, whilst being inundated by commercial and by-product information (Savage and Burrows, 2009). We propose that social science fields face a crisis precipitated by a deficiency of data and ineffectiveness of analysis when applying internet-mediated sampling methods. Meanwhile, we argue that the biases in sampling methods in internet research are primarily caused by self-motivated and self-reported data (Eagle et al., 2009, Shaffer et al., 2010). This problem to a certain extent encouraged social scientists to seek new methodologies to respond to the challenge set by ICTs. There are obviously alternative data resources for social scientist to respond to the methodological challenges in observing internet society. We propose that by using these new data resources, researchers may acquire more unbiased and comprehensive results and contribute to the discipline by developing both theoretical and practical innovations.

Deficiency of data

The duty of social science is that researchers should dedicate their energy to increasing the knowledge of human society and the interactions between people by interpreting different types of data. However, the internet-mediated sampling methods social scientists used on topics of internet phenomena are deficient data with a merely acceptable level of quality (Couper and Miller, 2008, Couper et al., 2001, Kaplowitz et al., 2004). At first glance, this data impoverishment presents itself in the form of data deficiency and this is due to the fact that social scientists still use unchanged methods to collect data in a changed social atmosphere.

Researchers are often frustrated by low response rates when collecting data via the internet (Couper and Miller, 2008). Online users have less motivation to fill in questionnaires or to be interviewed than the populations targeted in research that uses more traditional methods (Wright, 2005), although invitations to participate are easier to distribute when using internet-mediated methods. Social scientists have long believed that questioning a (small) group of people can give responses that are representative of the community (Savage and Burrows, 2007). This belief is based on the assumption that it is possible to gather opinions from individuals who are interested in the topic or feel responsible to influence it. Many studies suggest that internet-mediated surveys cannot obtain a response rate similar to other methods, although explanations for this vary (Couper, 2000, Couper and Miller, 2008, Kaplowitz et al., 2004, Walt et al., 2008).

Additionally, it is difficult to obtain valid responses in an internet-based sampling survey. Even if the targeted sampling group responds to an internet-based survey or interview, there remain serious challenges for researchers to obtain a fully completed response. Depending on the complexity and length of a questionnaire, many respondents have less motivation to complete a survey without supervision. It is likely that many respondents will drop out in the middle of the process (Schmidt, 1997). Furthermore, the fact that respondents may answer a question in a variety of ways rather than following a stringent format may give rise to unacceptable responses and damage the validity of the response (Zhang, 1999).

Sampling cases collected through the internet generally lack the precise personal information required to show the social status of the participants. Previous research has indicated that web-based surveys which do not request personal information can have a higher response rate (Kiesler and Sproull, 1986). This presents a dilemma, since if the questionnaire launched online requires basic information then the response rate could be lowered, but if such studies avoid collecting the basic information from respondents, the quality of research could be negatively affected. Particularly, the requirement for respondents to supply their personal details in online surveys may arouse the suspicion of infringement of privacy and raise ethical issues. Participants in surveys could easily feel uncomfortable and insecure when they are asked to provide basic personal information, such as their name, gender and social status. This concern is more serious in web-based surveys than in pen-and-paper surveys and email-based surveys, which explains why requesting personal information could dramatically decrease the rate of responses (Kiesler and Sproull, 1986). Indeed, the most successful internet surveys were conducted anonymously (Sheehan and Hoy, 1999). Although scholars may be able to use email addresses to identify respondents in some way, some studies also showed that email addresses may become out-of-date fairly quickly (Pitkow and Recker, 1995, Smith, 1997). Despite the benefits of using an anonymous survey to increase the response rate, it obviously complicates the matter when researchers try to draw connections between the results and the real social characteristics of the participants.

Although social scientists have begun to use internet technologies to collect data, they are only reinventing the sampling method wheel. The process of data collection is hampered by low response rates (Couper, 2000, Couper and Miller, 2008) and data biases (Ards et al., 1998, Coomber, 1997, Hartford et al., 2007, Sax et al., 2003). The limited data availability and the lack of refined approaches to gather valuable and

credible data force social scientists to seek more raw datasets (Savage and Burrows, 2007, Webber, 2009). However, the limitation is not only at the level of data collection; the ineffective analytical approach presents an additional challenge for social scientists.

Analytical ineffectiveness

The use of sampling methods in internet-based research has limitations in collecting reliable information and validating responses. In addition to the limitations on the quality and diversity of the collected data, this section focuses on the biases that may occur in the process of data analysis using internet-mediated sampling methods.

When using the internet as a communication media, it is fairly difficult for researchers to conduct a second round of research on the same group of respondents. Because of potential changes within the first set of collected data after pre-analysis, social scientists may need to revisit respondents in order to confirm some information or collect additional data. However, such a process is difficult to implement in an internet-based survey because of the respondent's distrust of data collection by web-survey (Cho and Larose, 1999, Schonlau et al., 2002). Both web-based and email-based surveys also face the problem of finding out the respondent's true identity and contacting them if required.

The information collected from internet-mediated sampling methods may lack data about respondents' social status and/or personal identification. In traditional sample surveys, scholars are able to confirm certain information provided by the respondents in subsequent face-to-face communications. However this is more difficult to realise in an internet-mediated sampling survey. The data collected from internet-mediated methods generally provide poor documentation of certain information, as respondents prefer not to provide their personal information such as gender, age and social status etc. (Frankel and Teich, 1999). Although internet-based surveys provide an alternative means to collect data for sampling studies, they raise serious concerns on how to check the credibility of data. The major obstacle of research is how to obtain representative samples from un-identified respondents (Braithwaite et al., 2003).

Furthermore, it has been widely noted that there are apparent biases in the responses of internet-mediated surveys which are caused by differences of participant's motivations (Lampe and Resnick, 2004, Lampe et al., 2007). These studies claimed that people who are more active online are also more likely to participant in a relevant survey; which may in turn

skew the representative sampling group. Such biases are the result of the self-selected mechanism in internet-based surveys which fully depend on the self-motivation of the respondents. In traditional surveys, the selection process is random. In other words, internet-mediated surveys may receive more responses from motivated people who are willing to participate (Ards et al., 1998, Hartford et al., 2007).

Such a research bias could be exacerbated in internet research, especially during studies of particular internet communities. Sampling methods mainly rely on the active participants from respondents in surveys or interviews. This method is dependent on the respondents' self-motivation to cooperate in the data-collecting process. It is effective because people who have the motivation to join in should be interested in the topic; however, it also raises the possibility of bias stemming from such self-motivation. Such biases are thought to be the invariable result of self-motivated actions, as different participants are driven by different motivations. Some scholars worry that in online surveys, "the sample is very diverse, but skewed" (Baym and Ledbetter, 2009). Some studies tried to put their surveys online to attract online users who are associated with a particular community, thereby expanding the sampling domain (Baym and Ledbetter, 2009). However, under such an investigation the respondents have to be heavily involved in the user groups. In general, these are long-term users, who are involved in online events more than average; have considerably more peers within that community; have a strong community spirit; and feel more responsibility to publicize it (Baym and Ledbetter, 2009). Therefore the results of such sample surveys can be skewed toward those who strongly identify themselves within and are emotionally involved with a particular community, or at least more skewed than that obtained from average users (Baym and Ledbetter, 2009).

Last but not the least; sampling surveys present information from a small group of respondents, which is suitable for the traditional society which has clear identification on each class. However, whether these classifications are still effective and true in an internet environment has raised some doubts. It has been called into question whether this limited data could provide representative samples and adequate descriptions of a virtual society. For instance, it is debatable that such research can accurately represent the internet applications used by millions of people. Furthermore, from a qualitative point of view, there is no clear consensus of the participant makeup of the internet; therefore, we cannot claim that certain participants could represent the activity and opinion of the entire population. The organizations and communities of Web 2.0 contain far more participants than any traditional community. Internet digital

technologies have changed physical aspects of human life, enlarged the size of communities and increased interactions (Smith and Kollock, 1999). All of these might facilitate the objectives of social science by providing new research opportunities. For instance, Facebook has more than 250 million active users and the average user has almost 120 friends; Wikipedia has 10,366,177 registered users and YouTube gets 1,586,000 website hits daily, streams 100 million videos a day and has about 63 million unique visitors per month[1]. This technological development could threaten the use of traditional methodologies to conduct quantitative studies, as we cannot possibly develop theories about new Web 2.0 applications by only surveying hundreds of participants to represent the members of a community that includes over 100 million members. There is an obvious hurdle to realistically display online phenomena through traditional sample surveys (Smith and Kollock, 1999).

In summary, there are two inherent weaknesses of the sampling methods typically used in internet-based research. First, internet-mediated methods limit the possibility for scholars to re-collect data from the same group of respondents after first contact. More importantly, scholars may be unable to collect the basic information required to identify respondents, because people may feel uncomfortable and unwilling to offer personal information online (Coomber, 1997, Hampton, 1999, Pitkow and Recker, 1995). Although such a situation is understandable, studies using this collected data may find it difficult to validate the collected responses. Second, both sampling surveys and qualitative interviews generate self-reported data via self-motivated means (Eagle et al., 2009). A number of factors may hamper this process of data collection, such as: the attitude and prejudice of respondents towards the research topic (Bertrand and Mullainathan, 2001, Marcus, 2003); the personality of the scholars (Ards et al., 1998); or even the memory of the subjects (Freeman, 1992, Freeman et al., 1987, Frensch, 1994). Because the inherent limitations of sampling methods could not be reconciled, internet-mediated sampling methods inevitably engender the limitations mentioned above and lead to analysis biases.

Because social science is a discipline that investigates and explores society, the development of such a discipline builds upon the collection of vast information from data resources. Therefore, if the credibility and availability of the data could not meet the expected demand, social science research, particularly that related to observing the online society, could be negatively affected. In spite of the difficulties experienced by social

[1] Information comes from the official press release of these three websites, Facebook, Wikipedia and YouTube.

scientists during the collection of internet-mediated sampling data, some promising opportunities and approaches for data enrichment have emerged recently along with the widespread use of the internet. The lack of awareness of these opportunities and an appreciation of their potential in research has limited the development of new methods and led us to notice the "opportunity".

New attempts of using existing data—using digital by-product data to explore online phenomena

The methodological challenge we reiterate upon refers specifically to the difficulties of capturing online phenomena effectively and efficiently. While the first attempt i.e. using internet-mediated sampling methods in response to this challenge has been shown to have certain limitations, many more researchers realised another possibility – instead of using the internet as a means to communicate, they believed it possible to use the existing features of ICTs which collect and stores data, and then use that stored information in research that studies online phenomena. This marks the second attempt by social scientists to alternatively and creatively tackle the methodological challenge and will be introduced in this section.

A missed opportunity

Social scientists face obstacles in that the data required for fruitful study may be in the possession of others; expensive; not accessible for the public; and, most of the time, practically out of reach due to a lengthy collection process (Avital et al., 2007). Specifically, Avital et al. regard most of the social science fields as "data-poor" because they are, "populated with individuals or small groups of scholars that collect their own data. Datasets are often small due to limited resources and incompatible methods" (Avital et al., 2007 p.6). Following the above analysis, it is not hard to realize that internet-mediated sampling methods are not perfectly suited for research on the internet and related topics, because of the bias caused by the limited sampling number and quality.

In the meantime, internet and related technologies provide an operational platform that facilitates the storage of colossal amount of data and even the construction of comprehensive and robust datasets. In order to ensure stable and visible websites, internet applications collect and store information including content, hyperlinks and personal behaviours automatically. For instance, Facebook needs to backup all uploaded profiles, photos, posts and even the interactions between users, such as