

**The Reality of Using Digital By-Product
Data in Social Science Analysis
--A Case Study of Wikipedia**

Zeyi He

Submitted for the degree of Doctor of Philosophy

University of York

Department of Sociology

September 2011

Abstract

In response to a methodological challenge in social science research, especially linked to studies of online phenomena including Web 2.0 applications, this thesis proposes a new methodology that deploys digital by-product data. Digital by-product data is the data created by an internet operating system to back-up content including 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 by digital by-product data. This thesis seeks to demonstrate that using digital by-product data is an important opportunity to help social scientists overcome various bottlenecks such as the deficiency of data and the limitations of analysis and possible risks of bias when using existing research methodology. Proposals relating to the new methodology are based on a discussion and analysis of the current data environment of social science research, the online environment and existing research methodology found within the digital science field.

The experimental aspect of the thesis uses digital by-product data to explore online phenomena, and to evaluate the utility of applying such a methodology more generally. 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 our case study. The thesis uses the digital by-product data that is generated by Wikipedia to analyse its collaborative mode in which millions of participants work together to provide an online encyclopaedia. The research is constructed in such a way that three related issues 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 this, 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 then analyse the administrative system as well as its policy to deal with editing conflicts. Finally, the results of this work are displayed in different ways, including the use of mathematical equations, metrics and visualization.

The thesis demonstrates that using digital by-product data provides a series of benefits to resolve the contemporary methodological challenge in the field and extends the capabilities of social scientists to investigate online phenomena. The thesis also provides practical lessons to guide investigators to help them to avoid the mistakes and problems that were encountered by the author of this thesis. Through studying an actual social phenomenon, the objective of this research is to evaluate the possibility and feasibility of using a new methodology, which

makes use of a neglected data resource to improve the engagement of social science with the world of the web. Such an evaluation can help scholars interested in using digital by-product data in their studies and also can provide some innovative ideas for social scientists in a new information age.

Contents

Chapter 1

	Methodological challenges for social science research in the internet age.....	1
1.1	Internet-based sampling methods	4
1.1.1	Applying internet-mediated methods in studies of the internet.....	5
1.1.2	The email-based survey	6
1.1.3	The web-based survey	7
1.2	Methodological limitations of sampling methods	9
1.2.1	Deficiency of data.....	10
1.2.2	Analytical ineffectiveness	12
1.3	New attempts of using existing data—using digital by-product data to explore online phenomena.....	14
1.3.1	A missed opportunity	15
1.3.2	Digital by-product data.....	17
1.3.3	Why apply digital by-product data in social science research?.....	19
1.3.4	Pioneering attempts	21
1.4	An innovative method to approach a new resource.....	24
1.4.1	What is data mining?.....	25
1.4.2	What is the process of data mining in previous scientific studies?	27
1.4.3	Implementation strategy	30

Chapter 2

	Using digital by-product data resources: Wikipedia as a case study	32
2.1	What data resource does Wikipedia provide?	33
2.1.1	Digitalized daily life—a research environment established by data	33
2.1.2	Wikipedia is a good example of Web 2.0 applications	36
2.1.3	Data from Wikipedia	38
2.2	What is Wikipedia?	40
2.2.1	Wikipedia’s History.....	40
2.2.2	Wikipedia’s policies	43
2.3	Debates surrounding Wikipedia	45
2.3.1	Wikipedia’s reliability	46
2.3.2	Wikipedia as a new innovation.....	49
2.3.3	Disseminating the Wikipedia spirit	50
2.4	Related work.....	52
2.4.1	Technical experiments based on Wikipedia and its data	53
2.4.2	Exploring Wikipedia’s mechanisms and incentives	54

Chapter 3

Assessing the development of Wikipedia	61
3.1 Introduction	62
3.2 Methods and databases	68
3.3 Descriptive development	70
3.3.1 The quantity of articles and participants.....	70
3.3.2 The quality of articles.....	75
3.4 Assessing the development of Wikipedia.....	79
3.4.1 Histogram	80
3.4.2 Pareto Distribution and Matching Analysis.....	82
3.4.3 The Maximum Likelihood of estimating the k parameter	85
3.4.4 Summarize the distribution.....	88
3.5 Conclusion.....	89

Chapter 4

The shift of participation -Who is the most important participant.....	91
4.1 Introduction	93
4.1.1 Debates about the participants of Wikipedia.....	93
4.1.2 Literature review	94
4.2 Method.....	98
4.3 Analysis	98
4.3.1 Growth and fluctuation of Wikipedia.....	99
4.3.2 Hypothesis of influence of participants	103
4.3.3 Edits made by different groups.....	105
4.3.4 Changes of proportion in edits by different groups	110
4.3.5 Influence of administrators	114
4.4 Conclusion.....	119

Chapter 5

Visualizing mass-authoring collaboration in articles.....	123
5.1 Introduction	124
5.1.1 Editing behaviours and the difficulties of reading and understanding them	124
5.1.2 What is visualization	126
5.1.3 Advantages of using visualization.....	128
5.2 Methods	131
5.3 Visualization of mass collaboration	133
5.3.1 How visualization displays individual articles	134
5.3.2 Baseline pattern	136
5.3.3 New way to collaborate	142

5.4	Conclusion.....	148
Chapter 6		
	Visualizing and assessing the administration in conflict-protection situations: A case study of fully-protected Wikipedia articles	151
6.1	Introduction	152
6.1.1	Fully-protected articles.....	154
6.1.2	Related literature	156
6.2	Methods.....	157
6.2.1	Data set.....	158
6.3	Visualizing full-protection in articles.....	160
6.3.1	Visualizing dynamics of full-protection during a selected period.....	160
6.3.2	Clarifying damaging behaviours and the reasons for full-protection	162
6.3.3	Visualizing damaging activities that cause editing-wars.....	166
6.3.4	Visualizing vandalism	173
6.3.5	Attempting to visualize damaging activities causing violation and sock puppetry	181
6.4	Assessing the administration system of Wikipedia	187
6.4.1	Assessing the function and effectiveness of full-protection	188
6.4.2	Assessing encouragement through the administration system	191
6.4.3	Assessing the influence of administrators in fully-protected articles.....	193
6.5	Conclusion.....	195
6.5.1	Power of the administration system represented in fully-protected articles..	195
6.5.2	Advantages and limitations of visualization.....	198
Chapter 7		
	Conclusion	204
7.1	The methodological challenge and the initial study	206
7.2	Overview of empirical works	208
7.3	Benefits of using digital by-product data	213
7.3.1	Value of using digital by-product data compared to sample data.....	213
7.3.2	Advantages of applying digital by-product data when exploring internet phenomena.....	215
7.3.3	Problems and limitations	216
7.3.4	The working flow of applying digital by-product data in social science.....	218
7.4	Summary and reflection	220

List of Figures

Figure 1-1 the work flow from data to knowledge (adopted from Figure 1.4 in definition (Han and Kamber, 2001).....	28
Figure 3-1 the individual components extracted from Wikipedia’s database	70
Figure 3-2 the number of total articles on Wikipedia from Jan, 2001 to Dec, 2007	72
Figure 3-3 the increase of new articles each month on Wikipedia.....	73
Figure 3-4 the number of participants each month on Wikipedia	74
Figure 3-5 the number of new participants monthly on Wikipedia.....	75
Figure 3-6 Average number of edits per article each month from Jan, 2001 to Dec, 2007....	77
Figure 3-7 the average number of participants per article monthly.....	78
Figure 3-8 the frequency of editor’s participation by year	81
Figure 3-9 Pareto Distribution.....	84
Figure 3-10 Change of key variable in Wikipedia model by year.....	87
Figure 4-1 Growth of edits per month in Wikipedia	100
Figure 4-2 Growth of unique participants per month in Wikipedia	102
Figure 4-3 Average number of edits per participant in Wikipedia by month.....	103
Figure 4-4 Number of participant grouped by the number of edits monthly.....	106
Figure 4-5 Number of total edits made by participants in different editing groups	108
Figure 4-6 Average number of edits per participants in different editing groups	109
Figure 4-7 Percentage of edits made by participants in different editing groups	112
Figure 4-8 Percentage of the number of participants in different editing groups.....	113
Figure 4-9 Number of edits made by administrators per month.....	116
Figure 4-10 Average number of edits per participant in admin and.....	117
Figure 4-11 Percentage of edits made by administrators against total edits.....	118
Figure 5-1 Screenshot of the history page related to a Wikipedia article.....	132
Figure 5-2 Visualizing result of the editing history by History Flow tool	133
Figure 5-3 Visualizing the article on “York”	135
Figure 5-4 Visualizing the article on the “European Union”	138
Figure 5-5 Highlighting the baseline pattern in visualization of the article “European Union”	139
Figure 5-6 Highlighting contribution of Cantus A in visualization of the article “European Union”	140

Figure 5-7 Highlighting the contribution of Tobias Conradi in visualization of the article “European Union”	141
Figure 5-8 Blocking and featuring in the same article	143
Figure 5-9 Visualizing the massive deletion in the featured article on “DNA”	145
Figure 5-10 Visualizing the massive deletion in the featured article on “Cell nucleus”	146
Figure 5-11 Visualizing the massive deletion in the featured article on ‘Archaea’	147
Figure 5-12 Visualizing the massive deletion in the featured article “On the origin of species”	148
Figure 6-1 Fully-protected webpages in different categories	159
Figure 6-2 Dynamic changes of the full-protection status of Wikipedia articles	161
Figure 6-3 Reasons to fully protect articles on Wikipedia	163
Figure 6-4 Reasons for full-protection	164
Figure 6-5 History flow of edits in the article on the ‘Battle of Pressburg’	168
Figure 6-6 History flow of edits in the article on ‘Levi Leipheimer’	169
Figure 6-7 History flow of edits in the article on the ‘Southern Baptist Convention.....	170
Figure 6-8 Example of “attack” behaviour in the article on ‘Arpan Sharme’	172
Figure 6-9 Massive deletions in the ‘computer science’ article	174
Figure 6-10 Massive deletions in the ‘Afghanistan’ article	175
Figure 6-11 Massive deletions in the ‘Thrikkunnathu seminary’ article.....	176
Figure 6-12 Massive deletions in the ‘Banqiao station’ article	177
Figure 6-13 Massive replacement in the ‘Zipporah’ article	179
Figure 6-14 Massive replacement in the ‘Day & age tour’ article	180
Figure 6-15 Violation of the ‘Arpan Sharma’ article	183
Figure 6-16 Violation in “Levi Leipheimer” article.....	184
Figure 6-17 Sock puppetry in the ‘David Bradford’ article	186
Figure 6-18 Distribution of full and semi protections	189
Figure 6-19 Average number of conversations in protecting period and general edit period	192
Figure 7-1 Proposed work flow from data to knowledge.....	219

List of Tables

Table 2-1. Data categorization	35
Table 2-2 Features of various web 2.0 applications	37
Table 3-1 Different types of data collecting from the Wikipedia database	69
Table 3-2 the comparison of the number of edits by individual participants on 2002-2007... ..	82
Table 6-1 the relationship between full-protection reasons and damaging behaviours.....	165
Table 6-2 Descriptive statistics of the number of conversation in fully-protected articles ...	191
Table 6-3 Thirty seven cases of fully-protected articles.....	203

Acknowledgements

I would like to give my appreciation and thanks to all persons who generously offered their talent and enthusiasm to help me throughout the four years of PhD life.

I owe my love and thanks to my dearest grandfather, Mr. S. Fan who encouraged me even as he suffered in hospital. I could not have taken this challenge without his trust and support. 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. I would not be able to face all the condemnation and distrust bravely yet still believe in myself without their hortative education and unconditional support. I also have many thanks to Jing 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 give my grateful thanks to my super PhD supervisor Professor Roger Burrows, for inspiring me with the distinctive and sparkling research opinions, for supporting me to introduce the novel quantitative methodology to my sociological research, for guiding me to apply critical and precise attitude into the details of my PhD work. His steady enthusiasm and persistence in sociology research have been a great value to my academic life. I could not have survived or finished my PhD work without his understanding, encouragement, inspiration, criticism and unconditional support. Also, I like to thank Professor Andrew Webster for his patient help in the final stage of my PhD, especially after Roger left York. His strict, careful and meticulous attitude toward research has greatly helped me improve the quality of this thesis in the final few months.

Another important person to thank is my internal examiner, Professor Mike Savage, he offered very valued and useful suggestions to correct my thesis and improve it. His appoint of view enlightened me in the digital data application field and helped me understand the sociological methodology. Thanks to my external examiner, Dr Steven Muncer, who has provided very useful advice for my thesis correction and helped me to speed up the entire process.

Furthermore, thanks to Mr. Brian Loader, who used this spotlighted idea to inspirit me at the beginning of my PhD and offered a very warm welcome to me to the University of York. Thanks to my second supervisor, Dr. Emma Uppriand, who kept reminding me that “you will be fine” faithfully with a smile every moment when I felt frustrated and hopeless.

Third, I would like to give my thanks to Bihui 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 life. I want to thank her for all her help to proofread, to carefully pick the appropriate expression, to solve many problems in content and structure. From the wee hours in the morning to many nights staying up with me, her company and support have been with me and motivated me to carry on. Despite being in a different time zone and location, she has reached out to help me perfect this thesis.

Forth, I give sincere thanks to all technical or professional specialists: Dr. Bo Wang from the Mathematical Department; Dr. Kan and Dr. Rose from the Computer Science Department; Mr. Eric, the execution chair of Wikimedia Foundation; Jimmy Wales, the founder of Wikipedia. Please allow me to give my sincerest thanks to all the technical advice and support given by my colleagues and the staff from National Grid Server and White Rose Grid Server in the past four years: they are Graham Lewis, Gillian Sinclair, Louis Rose, Mark Hewitt, Peter Halls, and Tom Rryan.

Fifth, I want to thanks all my friends who gave me inspiration in my work through either discussions or offering comments after reading, and those who looked after me through the tough time both mentally and physically. They are Dan Mercea, Xinya Zhu, Claire Schofield, Wen Wang, Jing Cui, Joanna Rossiter, Andrew Eggleston, Janette Cranney, and Dr. Charles Russ.

Author's declaration

I hereby certify that this thesis is my own work and has not been published elsewhere.

Signed:

Dated:

Chapter 1

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 thesis 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 'societies' (Hine, 2005, Stanczak, 2007). Internet-mediated sampling methods have expanded 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 thesis.

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.

1.1 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.

1.1.1 Applying internet-mediated methods in studies of the internet

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.

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 participant in surveys using the internet as the primary medium (Braithwaite et al., 2003, Schonlau et al., 2002).

1.1.2 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 a 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.

1.1.3 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.

1.2 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.

1.2.1 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.

1.2.2 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 participate 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 month¹. 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 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”.

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

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

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.

1.3.1 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 “poke” behaviours. Moreover, all systematic information such as user names, passwords, login times and logout times also need to be stored. Every internet application has a huge database to support it which contains abundant information.

More importantly, such technologies also improve data transparency and enforce the sharing mechanisms that enable particular data to be selected, modified and analysed easily by any group or individual. Therefore, much of the internet-related data have become easier to gather and use. Based on the digitisation of data and the availability for access and transmission at a relatively low cost, many commercial companies are quick to adopt such resources to enrich their internet applications (Tapscott and Williams, 2006). We suggest that the masses of readily available data on the internet are invisible to, or overlooked by, many social scientists, and the potential for utilizing such missed opportunities could be immense.

Although such exiting digital data collected from the internet has been used widely in science to test and evaluate new techniques (Bellomi and Bonato, 2005, Capocci et al., 2008, Chan et al., 2008, Cucerzan, 2007, Gabrilovich and Markovitch, 2009, Holloway et al., 2007, Strube and Ponzetto, 2006, Yan et al., 2009), we believe it is possible to use them in social science research as well, especially in areas that focus on observing newly emerged internet phenomena. Traditionally, social scientists regard the subject of studies as individual cases for data collection and analysis. Social science research is largely dominated by a research approach and a studying perspective designed for a particular purpose, rather than to understand the underlying phenomena. Using such exiting digital data can extend the vision of research in practical terms. Additionally, social science research has a propensity for an over reliance on sampling studies. Such studies are useful for exploratory research, revelatory research and theory generation, but are insufficient to examine and generalize the online phenomena and integrate such observations to develop theories about the internet-based society.

Hence, it becomes apparent that traditional sampling methods are not suitable to describe online phenomena, even if they use the internet as a media to communicate, because they only include a relatively small number of response cases. In order to understand internet phenomena, studies need to access more data to describe complicated situations and trends, and this could be best achieved by accessing a large database with diverse data (Barbier and Liu, 2011). In the following section, we will argue that scholars should creatively use free resources on the internet to collect such diverse data that provides an overall description of online processes and records individual activities.

Although the methodological challenges in social science may be more appreciable in studies of internet applications, there are missed opportunities to overcome this. We believe that the existing digital data introduced above is one example of such missed opportunities. Furthermore, the technical environment supporting internet platforms offer alternative opportunities to transact and establish datasets for social science research. One opportunity lies in the possibility of visualizing the online society through “computer assisted analysis” (Smith and Kollock, 1999). Along with this technology, social scientists can also benefit from previous experiences of using existing digital data in scientific fields, which will allow them to obtain and analyse data effectively and efficiently. As we have already discussed, this existing digital data as a by-product of internet technical processes can offer a new opportunity for social scientists, but such attempts also require a certain degree of familiarity with those same processes. In the next section, we will introduce some studies using the existing by-product data along with a summary of the available technical tools (Avital et al., 2007, Chan et al., 2008, Heer and Hellerstein, 2009). We propose that the plethora of existing

digital data could be used as a primary resource for social scientists to respond to the methodological challenges of the internet age. Such data can be gained through increased processing capabilities. In addition, the accuracy of data presentation creates new opportunities to advance social science research (Barbier and Liu, 2011).

1.3.2 Digital by-product data

In the introduction above, we introduced an existing data resource which is another opportunity for social scientists. This section defines two features of such a data type and names it digital by-product data. Additionally, this section also emphasizes that the opportunities provided by ICTs include abundant digital by-product data and related open-source software (Barbier and Liu, 2011). Under the present circumstances, this study boldly suggests “digital by-product data” as a new concept to describe such data, and clarifies its features. Using such data is considered as the second attempt of using ICTs to respond to the methodological challenges caused by internet phenomena.

Here we use the concept of digital by-product data, which is defined not only by its nature as a digital format but also its origin as a by-product from an internet operating or backup system. For instance, when we login or out of our Google email box, the system will save the exact time of we did so, the duration of the browsing session and the number of emails we read and replied to. Such data are collected for the purposes of system operation, and the record is a by-product of the email service. However, this data can be used by social scientists as a means to analyse the behaviour of email users. Technically, any web page has a database which automatically stores every action such as the transfer and creation of information. As Smith and Kollock (1999 p.196) said, “Online spaces become self-documenting natural settings”. Most internet users understand that every website they see on screen has a relevant database to store all content, hyperlinks, users, page view and other information. Organizations use the database management systems to control the creation and maintenance of a website. Generally, this data not only includes the behaviour of millions of website users, such as when they registered and when they logged-in; but also contains the tens of millions of messages they post online (Coderre et al., 2004). The latter may become a treasure trove of information for longitudinal research in social sciences, especially in socio-linguistic studies (Nocera, 2002). The most important point is that such readily established databases offer a possible source of information for social scientists, even though they were not designed for that purpose. We term such data as “digital by-product data”.

Social scientists have largely overlooked the potential and value of using digital by-product data. Internet and related computing technologies generate digital by-product data while carrying out their normal technical functions. These data record human behaviours and the

interactions between people in online communities, and thereby display a comprehensive view of their “online life” (Freeman, 2000, Rheingold, 2000). However, despite an awareness of the existence of such data, social scientists have often overlooked its value. Additionally, since digital by-product data are created effortlessly through technical processes, they usually do not receive due attention for utilization.

Previous studies have referred to digital by-product data by different names, such as transactional data (Savage and Burrows, 2009); internet data (Heer and Hellerstein, 2009) and data from social networks (Barbier and Liu, 2011). Generally, such data have several features. First, they are produced by internet technical processes, and are dependent on the internet as a medium for their dissemination (Barbier and Liu, 2011). Their appearance and maintenance are closely associated with ICTs. Second, such data are unbiased records of internet behaviours. Their generation does not depend on the individuals’ responses, instead they are created by the computing system as an automatic record of an individuals’ behaviour (Capocci et al., 2008). Objectively speaking, it is in contrast to the self-reported data in sampling methods (Eagle et al., 2009, Shaffer et al., 2010). Third, the generation of such data was not specifically aimed to contribute to academic research or similar purposes. On the contrary, these data are created and maintained by computing processes for technical usage. Fourth, most of this data cannot be analysed or investigated as isolated units. In order to successfully analyse, scholars need to assimilate the data that address the same issues or behaviours and study the conglomeration of the data as an integrated subject.

On the one hand, such data are not defined or selected by the researchers as they are in more traditional sampling methods. When using digital by-product data, depending on the availability of the technology and the openness of the personal data source, we may not be able to access the complete set of data to study certain internet behaviours. The restriction of data availability depends on the data source rather than the choice of the researcher. Since such data are objective records of the internet behaviours, their use could prevent the potential bias caused by using self-motivated and self-reported data.

Digital by-product data must not be confused with other data in digital format collected using internet technologies. The former are generated through technical processes and as a by-product of personal or social communications that were carried out without any specific purposes. The latter is obtained through careful design to achieve a certain research goal. In other words, “digital by-product data” are only side products which may not be useful for any research, but “digital data” are produced by research or commercial institutes with a particular aim.

Digital by-product data are not the same as a shared institutional dataset. The shared institutional dataset is another important digital data source which is collected by particular institutions and shared internally for academic research purposes, and some specific examples include the NHS database and the national census. The existence of such databases and the growing acceptance of their use by the general academia have conveniently provided resources for social science research, especially when information can be shared across many disciplines in a digitalized format. However, institutional data are collected for a specific purpose by a particular institute, and this underscores the principal difference between such data and digital by-product data. Because the collection of shared institutional data is made with a certain intention, in order to use it, users must understand the existing data structure and content, and authorization for access to the data may not be free of charge. Therefore, compared to digital by-product data, such shared institutional data could be more difficult to attain due to its high cost of collection and maintenance. Moreover, some of the shared institutional datasets are self-reported data as well, although may not have been subject to sample selection.

1.3.3 Why apply digital by-product data in social science research?

Following our description of the features and definitions of digital by-product data, we now explain why we should utilize such data in social science research. While social scientists are challenged to observe and analyse online interactions to explore a new format of society, internet technologies provide a mechanism to produce and share a massive amount of digital by-product data. Such data avoids the possible skewing effects inherent to the sampling methods. In other words, the internet on the one hand challenges social science by bringing various new social phenomena as subjects for studies; on the other it facilitates data collection and enables the expression of the emerging complicated discourses in social science research. Taking this into consideration, it seems most suitable to use the resources provided by the internet to study phenomena that occurs on the internet.

The most compelling reason for social scientists to start considering digital by-product data over traditional sampling methods is that the former offers better information without incurring possible biases by selecting samples and collecting responses. We have addressed that social science researches may experience limitations in the accuracy, breadth and diversity of collected data, because of a reliance on self-reported data inherent with traditional sampling methods. Using the internet as a medium to apply sampling methods may still pose observational limitations due to the reliance on self-reported data which are affected by self-motivation and the accuracy of human memory (Eagle et al., 2009). On the other hand, these problems are avoided if digital by-product data are use. Based on the forth-mentioned

advantage of digital by-product data, such methods could avoid possible bias and improve the quantity and quality of information that scholars can get from the internet.

Digital by-product data offer an accurate, integrated and comprehensive dataset for internet studies that is beyond any database generated by current methods. Because it is ordered as a by-product, such data avoid skewing the research result by offering an entire genuine record of people's online behaviours. Digital by-product data thus covers various categories of online behaviours and content without being affected or limited by sample selection, self-reporting mechanisms, individual memory or preference. An additional advantage of such data is the diversity of information. Digital by-product data provide detailed information on who, where, when and even how, which helps researchers gain a panoramic view of the internet society. Summarizing the points made above, the possibility of using digital by-product data in social science research has been approved. However, when we evaluate the suitability and practicality of certain methodologies to address particular research objectives, we always need to take into consideration the features that are associated with the particular research methodology.

In practice, the ease of which it is possible to launch a methodology is an important factor in measuring its practicality. As previously mentioned, digital by-product data can be accessed from multiple channels. Some internet applications have already sorted their data into different categories and published them online. Examples include Wikipedia (Ahn et al., 2005, Bellomi and Bonato, 2005, Denoyer and Gallinari, 2006) and Facebook (Paul, 2010, Price, 2010). For applications without already established databases, scholars could gather digital by-product information from individual webpages and aggregate them. Additionally, it is easy to extract shared information from websites such as Facebook and YouTube (Duffy, 2008), but privacy issue should be considered before implementing experimental procedures (Jones, 1994). Furthermore, some technical enthusiasts also share the databases they extracted from raw resource of digital by-product data (Ayers et al., 2008, Price, 2010). In general, there are several possible means of obtaining digital by-product data from the online environment.

In summary, digital by-product data provides several benefits for research. The most important is that digital by-product data, unlike the data from sampling methods, offers the opportunity to avoid potential biases. Interestingly, some scholars have even suggested using established by-product data to validate the responses of online participants (Ellison et al., 2006, Schwarzer et al., 1999). They emphasized that the limitation of surveys lies in that they only collect a self-assessment from participants, who can report their behaviours as they like. Furthermore, some social science researchers have solely relied on digital by-product data in their research and attained impressive results, which we will introduce next.

1.3.4 Pioneering attempts

In recent years more and more social scientists have begun to realize and value the possibility and advantages of using such types of digital by-product data. Social science studies that only use data supplied by the internet are burgeoning. Taking advantage of the easy to acquire and process features of internet data, researches started to explore within different disciplines and topics (Newman et al., 2006, Sanderson and Fisher, 1994). We will introduce some case studies to further discuss what digital by-product data are in specific contexts. Internet technology has supplied a massive amount of data (Janetzko, 2001, Sanderson and Fisher, 1994). This resource provides a fundamentally new opportunity for analysis in different fields. In an attempt to classify the range of research topics, we realize that studies which make use of internet technologies encompass a wide variety of topics. For example, in sociology, social networks can be reformatted based on online friendship data or contacting history (Newman et al., 2006; Lewis et al., 2008); in geography, scholars map geo-coded internet information (Crutcher and Zook, 2009, Goodchild, 2007, Zook and Graham, 2007, Zook et al., 2011); in marketing personalised and localised advertising has been studied (Haddadi et al., 2011); in psychological research, people's emotional expression can be tested and identified through textual analysis (Chmiel et al., 2011); in media studies, gossip and rumour have been identified to clarify its integrating and disseminating process (He, 2011; Dutton, 2006); and in criminology, both online criminality can be traced and offline crimes can be identified and localised by forum discussion (Casey, 2011; Fafinski 2010).

Social networks are the classic topic in social science and new digital by-product data accessed from the internet can describe such network more clearly and in a massive scope. With new datasets, millions of nodes, edges or cases can be traced through an interconnected online network (Newman et al., 2006). More importantly, these datasets also captures the links in the personal network. Therefore, the established database is able to reveal true human behaviours and interactions in online networks (Lewis et al., 2008). It could offer fascinating snapshots of people's actual behaviours, irrespective of what they claim their actions have been (Fisher et al., 2006).

Internet usage will record people's IP addresses along with behaviours, such as log in, opening a link, posting on a blog, uploading a video etc. All such behaviours can be localized the person who did the act based on their IP address (Lakhina, 2003). Therefore, social scientists can use such information to identify a particular community or a certain type of behaviour. For instance, the geographic distributions of all participants in Wikipedia and all users of Facebook have been drawn along a time line. Although such information is only displayed as a simple way to describe geographic location, these results can assist with the

public awareness of usage and popularity of internet applications. More impressively, the development and application of internet IP addresses goes well beyond their mapping and simple analysis. Now researchers are equipped with the ability to collect information through social and internet participation thereby providing detailed information such as a distribution of professions and ratio of personal assets for Facebook users. Researchers in the commercial arena have taken a step ahead of other research areas as the application of such data is especially useful in marketing focus research. This is directly driven by the commercial incentive to hunt for lucrative ways to attract more customers; on the other hand, it is also partially attributed to the fact that the owner and operator of the internet applications are the exact same companies that create and own the data. In comparison to others potential researchers, they undoubtedly have the most convenient conditions of use. A classic example of the use of such data is the recommendation system utilized by many online retailers. Regardless of what browser the customers use to make their online purchases, the system record and store all information on purchases. This “hidden” data can generate the commonly seen, “the costumers who bought this product also viewed/bought”.

Existing digital by-product data resource contains many unstructured text streams which may be valuable to social scientists (Erickson and Herring, 2005, Welser et al., 2007). An obvious opportunity for social scientists lies in analysing the unstructured online content to address peoples’ emotions. Researchers have conducted psychological studies on web users using information collected from personal profile pages, blogs, comments and interactions on internet fora. The main method of analysis is to define various key words which could identify different emotions based on traditional psychological theories, and makes use of machine learning technologies, followed by semantic technology to categorize these words into positive and negative emotions (Chmiel et al., 2011, Thelwall et al., 2012). This method can be used to discover and define the emotional environment on different discussion platforms.

Textual data are stored in content resource of different internet applications, which includes specific information such as participants present at events and details of those events through natural-language process and/or the extraction of keywords or cluster concepts (Fisher, 2007). Besides emotion studies, some scholars have used content analysis to characterize the form and functions of blogs or to trace the trend of topic changes (Herring et al., 2005, Herring et al., 2004). Other studies have analysed the profiles of social networks to address the issue of privacy protection (Hinduja and Patchin, 2008).

Using textual data obtained from the internet, such as forum discussions (Howard, 2002), email triage (Neustaedter et al., 2005) and social network profiles (Ellison et al., 2007), media

scholars can explore online interaction and communication dynamics. Researchers can investigate questions such as the manner of speaking, content change, the dynamics of research discussions, changes in communication and influence of the media etc. (He, 2011). For instance, scholars chose to research rumour spread via the internet as a research topic and conducted studies to investigate the dynamics of the creation, alteration and dissemination of rumour by examining the content of online forum and blogs (Bai and He, 2010). Although such attempts still make use of the most basic stored data from internet, these researches require a certain technical background to integrate both the textual data and numerical data. For social scientists, these by-product data are fascinating, rich, attractive and under-explored (Ellison et al., 2007).

Criminology is still at a relatively early stage of exploring digital by-product data, as demonstrated by the fact that most studies still limit its use to the analysis and sorting phase of research (Fafinski 2010). In fact, the usefulness of the internet for the study of criminology should be more appreciated, because it not only provides a new research topic of “cybercrime”, but also supplies a great number of new possibilities for both new and traditional research topics (Williams, M. 2006). Through the linguistics analysis of online communication, many key words could be used to identify potential or actual criminal events. This theory has been proposed as an important possibility of development in criminology, although the practice of which has a long way to go.

There are many ways to classify studies based on digital by-product data resource collected from the internet directly. By analysing the application of such data in research, we discover that there is more than one type of data that can be utilised by social scientists. While some use public data such as hyperlinks and web IP addresses as their primary data (Halavais, 2000), others choose to use personal data for their analysis (Lampe et al., 2007). Public digital by-product data could be gathered without injecting extra research context (Adamic and Adar, 2005, Park and Thelwall, 2003), as hyperlinks only include geographic and categorical information. For instance, scholars examined hyperlinks from 4000 selected websites to explore the role of geographic borders (Halavais, 2000). Aside from hyperlinks, personal information is another important digital by-product data source for social scientists, such as personal profiles on social networks (Lampe et al., 2007).

Digital by-product data can be obtained from traditional web 1.0 applications, such as internet media and email list, however more data can be obtained from web 2.0 applications, which include blogs, social networks (Facebook and MySpace), multi-media sharing communities (YouTube) and wiki-platforms. Different researchers can extract different information based on their direction and interest. For example, by studying Facebook, geographers have mapped

the distribution of geographic locations of users, sociologies have identified the social networks that exist between users and social psychologists explored the emotional expressions from users' profiles. This illustrates the fact that internet applications can provide a rich body of digital by-product data for research, but also more importantly that the same data can be robustly reused and applied in different areas of academic research.

Additionally, from the approach of research, especially in terms of the technical tools used, social scholars have used alternative means to gain digital by-product data in order to initiate analysis. In general, two approaches are employed to convert raw data obtained from online open sources: one is to use existing tools (Banks, 2001, Brunn and Dodge, 2001, Park and Thelwall, 2003), and the other is to design computer data programmes by the researchers themselves (Adler and Alfaro, 2007, Almeida et al., 2007). In fact, we discovered that in most of the previous studies, the research was never done independently by one social scientist. They are the fruit of either a research group or collaboration between computer scientists and social scientists, or social scientists adopting existing tools to analyse data. These intellectual hubs integrate the diverse expertise of different researchers and marry the application of technology and social scientific ideas and perspectives to achieve their investigations.

As for the second round of attempts to study online society, pioneering studies demonstrated different ways that they used to generate and apply digital by-product data. From their studies, we can summarize the types of used data by understanding the features. However, what is more important for us is to summarize the value and perhaps even the feasibility of the research methodology with regards to such data resource. We aim to assess the advantages and disadvantages of such a methodology, especially in terms of the step-by-step procedure of applying it in internet studies. Due to the lack of summaries and explanations in the process of using such a methodology, another task is to evaluate the associated limitations by using digital by-product data resource in a real study and observe the process.

1.4 An innovative method to approach a new resource

For the newly emerged social science topics provided by ICTs, researchers took an active part in innovating existing research methodologies and exploring the use of new data resources (Caplan, 2003, Illingworth, 2001, Wright, 2005). Initially social scientists attempted to reuse traditional sampling methods but use the internet as a medium to explore the ICT supported society (Best et al., 2001, Hartford et al., 2007). This first attempt in research is facilitated by a convenient information communication platform supported by ICTs and made use of the internet as a communication tool (Chen and Hinton, 1999, Coderre et al., 2004, Couper, 2000, Couper and Miller, 2008, Hampton, 1999, Schmidt, 1997). However such attempts are still limited in the quantity and quality of the information available (Savage and Burrows, 2007,

2009, Webber, 2009). The recognition of these limitations prompted some social scientists to start to explore other features of ICTs such as digitalized information and automatic data storage. Social scientists initiated the second attempt by using new type of data collected from the internet directly.

Using digital by-product data to conduct social science research is an innovative methodology, extracting digital by-product data from the target resource can be a complicated process. In this section, we propose to highlight data mining, which has been applied widely in scientific research to extract and utilize digital by-product data. In order to choose the right process to obtain appropriate datasets, it is necessary to consider the complexity, the structure and the comprehension of extracting such data. It should be noted that data mining is only one possible process to obtain and use digital by-product data; nonetheless it is a rigorously tested and widely used method. In order to address how to use data mining to approach the digital by-product data resource, this section focuses on issues such as the feasibility, usefulness, efficiency and scalability of data mining for the discovery of knowledge hidden in large datasets. Our discussions focus on what data mining is; why it is important; what the function of data mining is; and how to implement it in practice.

1.4.1 What is data mining?

Digital by-product data is abundant both quantitatively and qualitatively because it is produced by internet operating systems automatically. Thus, in order to apply such data, we need to learn how to collect it and integrate it as a large dataset. Large data-set utilization is an important practical field across disciplines and organizations including the social sciences, government, public organisations and commercial institutes. Experts or amateurs alike from around the world construct an information network that contains data which could be easily accessed in a digital format. Such platform provides different type of large dataset for different applications. This has been demonstrated in disciplines such as astronomy, physics, geology, history, archaeology and ocean sciences, which have traditionally relied on the building up and sharing of datasets within fields (Avital et al., 2007). Scholars in biology and economics have shown the proof-of-principle of working together to build large-scale datasets and using them as a public asset for the entire field (Baitaluk et al., 2006, Hine, 2008). Basically, we can learn how to use a large dataset from previous scientific studies. This method is data mining.

Data mining is the process often referred to as the extraction of relevant information from selected data (Han and Kamber, 2001, Hand et al., 2001). Many sources regard data mining as a synonym for another popular scholarly concept “knowledge discovery in databases” (KDD) (Han and Kamber, 2001 p.5). The KDD process is similar to that of traditional miners

extracting precious metals or gold from rocks or sand. Similarly, an analogy of data mining would be sieving through all the sands from the riverbed to find small pieces of gold. In other words, data mining, as a pioneering field, is to extract meaningful information that is not readily apparent or easily discovered from a dataset (Barbier and Liu, 2011, Hand et al., 2001, Larose, 2004).

As a multidisciplinary field, data mining encompasses work from areas such as database technology, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, knowledge acquisition and data visualization (Han and Kamber, 2001). Some literature narrowed its definition to a single step: the convenient extraction of patterns representing knowledge which is stored in large databases or data warehouses (Hand et al., 1999, Inmon, 2002). However, the more widely accepted definition of data mining is the entire process of discovering interesting knowledge from a large data store, including the building up of a data storage, selecting and transforming useful data from the warehouse, mining information and representing knowledge (Han and Kamber, 2001). In this work, we have chosen to adopt this latter definition.

In fact, data mining as a classical methodology that is commonly used in scientific research has been continuously improved and widely applied. Aided by advances in computational and machine learning technologies, the data mining process has undergone much evolution from the 1960s to now (Han and Kamber, 2001). With the establishment of query languages such as Structured Query Language (SQL) in the 1970s, data mining software provided more user-friendly interfaces to facilitate data collection and functional performances. From the mid-1980s, the wide adoption of computational technologies has oriented data mining to a more application focused process, in which advanced data models could be fitted to the database management system and basic analysis could be carried out along with the data cleaning process (Han and Kamber, 2001). Based on the understanding of internet and web languages, the territory of data mining has further expanded to web-based data mining since the 1990s, facilitated by the eXtensibleMarkup Language (XML) based database system which includes features for both data storage and data analysis (Han and Kamber, 2001).

The above introductions are intended to provide an overall impression of what data mining is, and how it developed as a technology. However, more important to us, as social scientists, is to find out how to using data mining in the process of digital by-product data utilization. In order to achieve this goal, we will begin our discussion by introducing how scientists use data mining to extract relevant information.

The major reason that data mining has been discussed frequently in recent years is due to the wide availability of large amount of data and the demand for converting such data into useful

and interesting knowledge (Han and Kamber, 2001). Most of the time, such data contain not only digital by-product data but also other types of data produced by computational processes. These other types of data include biological networks (Baitaluk et al., 2006), brain images (Kremer et al., 1996), and astronomical geography (Fotheringham et al., 2000). All of these are the result of two technical improvements, namely computational techniques for integrating and formatting data and internet technologies for storing and sharing information. Thus, ICTs provide an immense body of datasets, which offer an opportunity for discovering hidden or latent knowledge (Barbier and Liu, 2011).

The large body of data in combination with the complicated content and the intricate structure of such data demand the careful extraction of pertinent information to address specific research questions. In order to construct relevant information repositories, data mining is required to extract useful information and depict certain patterns from different data stores. This process is difficult to accomplish by any single execution but data mining is able to turn a complicated data repository into interesting and comprehensive observations.

1.4.2 What is the process of data mining in previous scientific studies²?

After addressing the definition of data mining in the last part, it is clear that the application of data mining is used widely in scientific fields. The reason that data mining was introduced within in the application of scientific research endeavours is because that although some social scientists have begun using digital by-product data and data mining in their research, there is a lack of clear description, explanation and procedure of the methodology. On the contrary, such a methodology is comprehensively summarized in scientific fields, and has been repeatedly tested and summarised into a systematic approach. Therefore, this part introduces the general process involved in data mining from a scientific perspective (Barbier and Liu, 2011, Han and Kamber, 2001, Hand et al., 2001, Larose, 2004, Witten and Frank, 2005). Although the purpose of this section is to propose the use of data mining after digital by-product data are acquired, we have to admit that the utilization of digital by-product data in social science research is less common than it is in scientific applications. Therefore, in order to master the features of digital by-product data and apply it to social science research, we will summarize its work flow from scientific research experiences.

The process of data mining can be divided into four steps: the first is importing raw data from the original resource, followed by data cleaning and data integration; the second is selecting interesting and useful data and transforming them from the data warehouse, and this process is considered data mining in the narrow definition (Han and Kamber, 2001); the third is

²The working process described here is based on Han & Kamber (2001); Hand et al. (2001); Larose (2004) and Witten & Frank (2005). Although this is an important process in our proposed method, it is not the primary focus for this thesis. Thus, we only present it as a simple introduction.

analysing the selected data and presenting the analysed results and any associated patterns; the final step is turning the pattern into understanding and comprehensive knowledge through evaluation and presentation. We present the step-by-step procedures of data mining in

Figure 1-1. This process is also examined and tested in the empirical work of this thesis, in which we aim to provide an introduction of data mining process to social scientists.

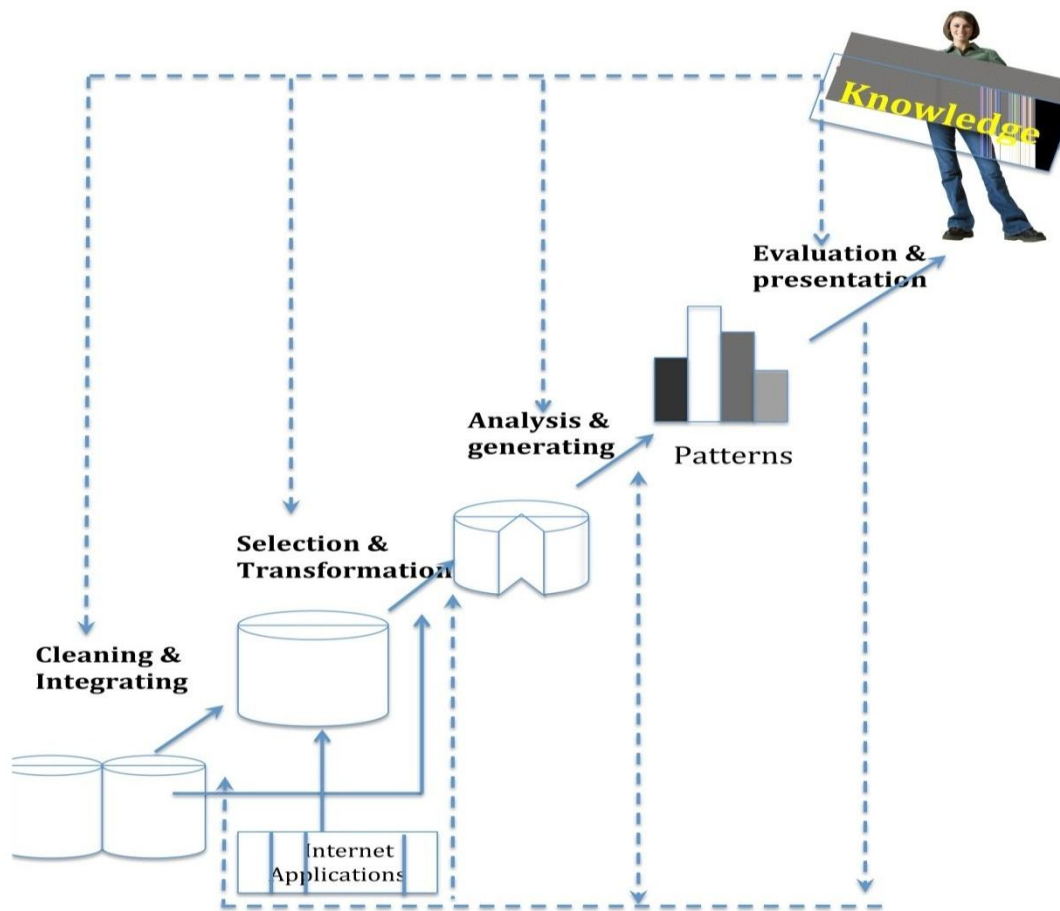


Figure 1-1 the work flow from data to knowledge (adopted from Figure 1.4 in definition (Han and Kamber, 2001)

First, scholars face a heap of raw data, which require cleaning and integration processes to sort and store them into a warehouse by category. The data warehouse is, “a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management’s

decision making process” (Inmon, 2002). Generally, the raw data have to be filtered to remove erroneous and inconsistent entries, a process of which known as “cleaning”. Subsequently, the cleaned data are labelled and any potential connections between data are marked. This way the multiple data resources become organisational and operational. This step is named “data integration”. The establishment of a data warehouse is the basic data mining step to prepare for all subsequent steps, but it does not have to be limited by any specific purpose of data analysis. In other words, the construction of an organized and well-sourced data warehouse could be used for many different researches across disciplines.

Secondly, scholars have to carry out the selection and transformation step to retrieve relevant data from the warehouse, bearing in mind the particular research plan and goal, a process of which known as “data selection”. Data transformation refers to the process where data are transformed or consolidated into sensible aggregates. In fact, this step may merge data from multiple organized data warehouses. There are a number of issues to consider during this process, such as the data format and the data scale. For example, some daily sales data have to be compressed to monthly or annual data using computing processes, in order to generate data that could be easily represented by patterns.

In the third step, data can be characterized to a certain pattern by analysing and generating process, such as formulating the mathematical model. A large body of selected data has the potential to generate hundreds and thousands of patterns or rules but only a small fraction would be of actual interest to scholars. The pattern is able to provide an understandable result with a certain degree of certainty, a definitive conclusion or a confirmation to validate a hypothesis. Namely, the generated pattern offers knowledge to validate existing data through new data.

It is worth pointing out that in data mining the interesting pattern may not be predictable based on the raw data. Thus, interesting patterns have to be sought and explored by data analysis. It is not always certain which part of the data might provide the interesting pattern and whether some potential patterns could describe a structure related to a part of the data. This brings us to two main points in the pattern generation step: that the exploratory pattern is not predictable or designable; and scholars should be prepared for any possibility.

After data cleaning and integration, data selection and transformation, data analysis and generation, the final step is to evaluate useful information and present it in an appropriate way. This step requires two parts: the first is to evaluate the outcomes from data analysis and generation; then the outcomes should be represented in a suitable manner. For instance, scientific studies generally adopt graphical techniques, such as tables, charts, cross tabulations,

curves and matrices to present their discoveries. Particularly, for social science, we need to select an appropriate way to assist readers in understanding our analysis.

1.4.3 Implementation strategy

Following the points addressed above, we came to an important realisation that there is a tentative solution which liberates us from the imminent methodological challenge found when researching newly emerged ICT-related topics. Our entire proposal is based on the assumption that the chosen methodology can work in social science fields effectively and meets the qualitative and quantitative requirements of data for social science research. However, the previous studies utilising digital by-product data did not provide a description of the entire process of approaching and analysing data. Therefore, we have to observe, examine, evaluate and test the proposed methodology by ourselves, before encouraging more social scientists to accept our proposal, convincing social scholars to apply or combine this new methodology with sampling methods.

From a methodological point of view, we need to consider whether the method is practical when it is used to treat a realistic problem, and also evaluate the effectiveness and efficiency of the application. Furthermore, we want to ascertain that such a methodology with origins in scientific areas can also be replicated in conducting social science research. If all of the above could be proved, the significance of our research lies in that it will revolutionise the research of the internet by introducing a novel methodology. Not only will this revolution equip social scientists with advanced technologies to understand and study social phenomena, but it will also define the possibility of collaboration between scientists and social scientist. When faced with the challenge concomitant with technological innovations, such an innovative methodology explores the potential of using current technology to adapt to the evolved society.

In order to address the aims outlined above, the thesis develops as follows. First of all, we identify a series of practical questions, such as what preparations are needed before using digital by-product data, whether the data mining process is suitable in social science to extract and use digital by-product data, how the new social science research angles could be explored using the newly available datasets, etc. Generally speaking, there is a lack of experience in using such novel methodology in both technical practice and preparation.

Therefore, we propose an integrated plan which aims to examine the possibility of using such methodology in social science research and also to evaluate its advantages and limitations during the experimental process. In this study, we plan to observe an online phenomenon where human behaviours and interactions have been changed or even formed by the internet environment. More importantly, we will utilise this new methodology as the only method in

this study to examine the possibility of whether it can be used in social science and the effect of using it to carry out social science research.

In the design of the research proposal, there are two questions that demand our attention: the first is how to design and approach the methods to focus on digital by-product data resources; the second is whether such a methodology could achieve the same level of comprehensiveness compared to traditional sampling methodologies. In other words, the first focuses on the practical issue, whereas the second concentrates on the theoretical perspective. To answer these questions, we need to investigate an appropriate subject using digital by-product data from multiple sources.

Based on the research plan of examining the new methodology, we propose to use Wikipedia as a case study to explore the newly emerged phenomenon of people working together to create and share knowledge (Tapscott and Williams, 2006). Such a choice meets our requirement of using the available resource of digital by-product data, which also represents an interesting social science research topic. Then we will concentrate on the exploration of this research topic using the approach of digital by-product data utilization. Moreover, we intend to use three representational means at the end of our data mining process to deliver our findings: namely, mathematical equations, graphics and visualization. Through this, we will be able to evaluate the possibility of using data mining in social science research, and also explore the potential of the different representational means.

This chapter points out that social scientists face an emerging methodological challenge brought by ICTs. They have responded to this challenge in two ways: using internet-mediated sampling methods whose data limitations and analytical ineffectiveness have been described. The second attempt of using digital by-product data is at the initial exploring stage. We discuss the possible advantages, review the previous pioneering works and design our plan to evaluate its possibility and practicality. Through the next chapter, we carry on designing our implementation plan and outline Wikipedia as our case study to evaluate the proposed methodology.

Chapter 2

Using digital by-product data resources: Wikipedia as a case study

In chapter one, we discussed previous attempts at finding appropriate and effective methodologies in current social science research in regard to their application in internet studies especially Web 2.0 applications. We further pointed out that contemporary social scientists might have overlooked opportunities of using existing data resources that are themselves produced by internet applications. Based on this, we introduced some pioneers' works on using digital by-product data as a second attempt to extricate social scientists from their methodological predicament. Although some pioneers have applied digital by-product data, their works have not yet offered a systematic discussion on the methodological process or provided any evaluation of such a method. In the final section of chapter one, we pointed out that using a case study with abundant digital by-product data in our research would allow us not only to understand the process of applying such a methodology to social science research but also further evaluate our proposed method by comparing the results with the outcomes generated by sampling methods.

Therefore, in order to demonstrate how digital by-product data can be used in social science studies, and examine whether using it can achieve the expected research outcomes, we need to construct an experimental research plan. Specifically, our research needs to derive its data from a web 2.0 application that is able to offer digital by-product data resources. We decided to use Wikipedia. There are two purposes to this chapter, first to address why we chose Wikipedia as a case study and second to introduce Wikipedia and its related studies in order to assist readers to understand our case better and to be able to follow our later three later empirical chapters.

There are two reasons why Wikipedia is selected as our case study. On the one hand, studies of Wikipedia reflect many newly emerged social science issues such as mash-up and

knowledge-sharing/crowd-sourcing, and involves numerous fundamental questions about how the internet is changing our lives, such as the dissolution of copyright, e-democracy, mass collaboration and e-economics. Based on this, we think that researching Wikipedia provides a fruitful site to explore new forms of data and methods, and in addition, has a large set of digital by-product data with reasonable accessibility. The data stored in this database met our requirements to conduct social science research using only digital by-product data both in content and structure. More importantly, compared to other internet applications, Wikipedia's database contains extra types of data because it does not only depend on traditional Web 2.0 functions, it contains unique functions to maintain its sharing and collaborative mechanism, which we will explain later.

In this chapter, we introduce the technological environment from which Wikipedia arose and the factors that prompted social scientists to study it. We then elaborate on the establishment of Wikipedia and current academic discussions on the progress and development of Wikipedia. Finally, we will review the past efforts to explore Wikipedia, leading to our own unique study in the same field.

2.1 What data resource does Wikipedia provide?

The continuous development of ICTs has filled our life with a variety of data. Such data can not only supply commerce with a rich resource to improve customer services and reduce cost, but can also be used in research across many disciplines. The digital by-product data we introduced in the first chapter are only one of many types of existing data. In this section, we will first discuss the current digital data environment from a practical perspective. At the same time, we will also categorize the various types of data and point out how digital by-product data differs from other data resources. Following such a categorization, the format and meanings of the data we drew from Wikipedia are also introduced.

2.1.1 Digitalized daily life—a research environment established by data

With the development and popularity of the internet, our daily life is digitalized and recorded as data automatically. These new types of data are constantly emerging. Such data are no longer limited to those that need to be extracted through selecting and integrating processes, where data generation and data collection are both subjected to personal prejudices (Ruppert and Savage, 2009, Forthcoming); such data are no longer limited to those that are stored in databases of commercial or government-related organizations; and are no longer limited to those that are displayed and disseminated in traditional formats such as a table or set of metrics (Savage et al., 2010). In other words, the formation of these data is more complicated, the information delivered more abundant and the expressive formats more diverse. Such

wealth of data could be generated from an accumulation of events at every moment of our life, actions from every social movement and every piece of text we sent to others or received from others. For instance, a day starts with an alarm which is stored in a mobile database; our calls, texts stored in the server of a mobile telecommunication company; the songs we listened to store in a music device's database; all purchases we make are stored in suppliers' database and our online lives are as individual behaviours stored in different online databases (Ruppert and Savage, 2009). This life, full of data, has been defined as heralding a "new social life of data" (Beer and Burrows, Forthcoming).

In such a data-rich environment (Zinovyev, 2011), data are not merely a resource in research, the analysis and discussion of which can further measure our life. Our preferences, friends, social life, cultural experience and even political orientation can be annotated by and interpreted from diverse data (Beer and Burrows, 2010, Law, 2009, Ritzer and Jurgenson, 2010). The assembly of such individual datasets into a data resource could help us to measure the developmental trend of the society and social movements (Law et al., 2011). Depending on the type, characteristic and the data archives, such data resources can be classified according to its application within research. In order to explain how we could measure our life from a social science perspective within an environment with richly accumulated data, this part will categorize data types mainly in terms of their accessibility and applicability.

First of all, based on the accessibility of data archives, the privacy of the content and whether the data collection violates the individual's will—we can divide online archives into public and private ones. Specifically, public archives are designed to be publicly accessible, the content is not associated with any private matters and furthermore the usage of such data does not engender ethical issues, such as the Wikipedia database we use in this thesis. If the archive does not fit any of the above descriptions, we classify it as a private one—such as Facebook messages, or online bank statements. The differences between public archives and private archives can be summarized into the following three features: accessibility; ethical concerns; and privacy.

From another perspective, based on the purpose of data generation, the transmission mode, and whether there is a fixed group of recipients, we divide data into by-product data and intentionally produced data. The former are not established to serve a transmission purpose, the generation of such data is not intended to elicit any particular response from audiences and there are usually no fixed recipients, such as the Wikipedia edit history, a purchasing history in Amazon or an email-send/receive history. However, intentionally produced data are generated with the specific purpose of dissemination. In the process of data generation, they are further modified and edited, and finally delivered to certain audiences in order to achieve

particular responses. For instance, a post on a Facebook wall is created to disseminate personal information to other friends. Similarly, blogs are written to deliver certain information with an expectation of feedback from readers.

Through these classifications, we can provide a clear demonstration of the data usage situation based on their accessibility and ethical considerations. As shown in Table 2-1, we divide data based on accessibility and relevant ethical issue, and through such a classification, we hope to discover how different types of data and their corresponding archives are used. Table 2-1 not only summarizes our classification of existing digitalized data, but also introduces examples of various data used for current research (Lampe et al., 2007).

Data in category B are rather easy to use in traditional internet research and consequently are the most frequently used because they are widely transmitted and intended to be noticed. Data in category C should not be used due to ethical issues and legal constraints or inaccessibility due to technical restrictions. Type D data are the most debatable kind, since they are intended for dissemination by the involved participants, yet they contain sensitive information that may also cause ethical and privacy issues. Examples include Facebook profiles, uploaded live photo albums etc. Finally, category A type data are the data that we focus our discussions on in this study. On the one hand, they belong to public archives, which is available to anyone with sufficient computing power at their disposal. On the other hand, they are by-product data, which are objective, diverse and abundant. We thus proposed this data as a valuable and effective resource for social science studies.

	By-product data	Intentionally produced data
Public archives	A. <i>e.g. Wikipedia editing record</i>	B. <i>e.g. Blogs</i>
Private archives	C. <i>e.g. Record of Music download</i>	D. <i>e.g. Facebook profile</i>

Table 2-1. Data categorization

The reason we place an emphasis on Type A data is that such data are often overlooked by academic researchers, as the system that produces data generates them automatically and does so for a technical purpose, and organizations who own such data do not intend to disseminate data to the public. Yet by-product data is accessible given appropriate techniques, and so such data can be used to extend our ability to explore our society, especially on internet issues.

We believe that Type A data are the most valuable data and are easiest to use in research, because such data are not only accessible but also are objective in the sense of being ‘given’. The most important advantage of digital by-product data is that it does not include any information with intentional “edits”, which avoid any “adding” information or “selecting” bias. In chapter one, we argued that all sampling data are affected more or less by intentional edit in the process of collecting, formatting and transferring, and the personal bias might occur at any step of the process to affect the accuracy and completeness of information to be used in academic research. Therefore, we propose that the value and significance of digital by-product data cannot be compared with that of other forms of digital data, which is often information collected with a certain purpose and shared within an organisation.

Type A data are particularly suitable for use in academic research, due to the fact that they contain large amount of information, which are easily procurable and most of all do not involve any private information. More importantly, Wikipedia, as the owner of such data, has already opened access to the public (excluding personal information), to encourage both amateur and scholarly involvement.

2.1.2 Wikipedia is a good example of Web 2.0 applications

In Web 2.0 applications, functions of editing and uploading empower users to create products for their own use, which is the big difference between Web 1.0 and Web 2.0. Web 2.0 applications provide platforms for producing applications and creating content by users and such a production system is user-oriented, whereas Web 1.0 applications only provide products and contents for users to read and retrieve. Therefore, Web 2.0 applications contain more digital by-product data about users’ actions and their movements, which is more useful in evaluating the reality of using digital by-product data in online research.

It is worth noting that Wikipedia, as one type of Web 2.0 applications has its uniqueness deferring from other Web 2.0 applications, including blogs, tagging and RSS etc. (Anderson, 2007). In the table below, we address several important functions of Web 2.0 and discuss whether they have similar characteristics. Wikipedia has the history trace and reversion functions which are not offered in other Web 2.0 applications. The reversion function is one of the important functions to maintain the quality of collaborative knowledge, which offers convenience to participants.

	Web 2.0 type	Edit / upload	Social communication	History trace	Reverse action	Product
Wikipedia	Wikis	Yes	Yes	Yes	Yes	Single product required collaboration
Blog	Blogs / Audio Blogging	Yes	Yes	Yes/no deletion	No, only deletion	Multi-combined content depending on individuals
del.icio.us	Tagging / Social bookmarking	Yes	Yes	Yes/no deletion	No, only delete their own marking	Multi-combined content depending on individuals
YouTube	Multi-Media Sharing	Yes	Yes	Yes/no deletion	No, only delete their own uploads	Multi-combined content depending on individuals
RSS	RSS and syndication	Yes	Yes	Yes/no deletion	Dependent on system	Multi-combined content depending on individuals

Table 2-2 Features of various web 2.0 applications

We have argued that Wikipedia is a single-goal web 2.0 application. Consequently, Wikipedia provides some distinct functions to accomplish this goal. Looking at Table 2-2 it is noticeable that this reversion function is designed for Wikipedia to assist participants in collaborating in a single-goal project, which means that Wikipedia only provides the integrated and interconnected content persistently under the same structure. For instance, in Wikipedia, all articles follow the same structure to present standard entries for readers; and all participants contribute and collaborate with each other with an awareness of the same end goal of establishing and maintaining a free online encyclopaedia. Their contributions are consistent even they are from different participants. However, in other applications, such as Facebook and YouTube, users contribute their content or information separately for individual purposes and so are formed by an individual's will.

Wikipedia can generate more digital by-product data about collaborative behaviours because of its extra functions comparing to other web 2.0 applications. It can be said that the reason why Wikipedia can provide behaviour data of a large crowd of participants is thanks to its collaborative feature and a sharing culture.

2.1.3 Data from Wikipedia

Following our macro analysis of the possible data structure within Wikipedia, we will direct our effort to the introduction of the features and types of data in this section. Wikipedia provides a very friendly data platform for users, by periodically sorting digital by-product data and providing them for public download.

Currently, Wikipedia has 18 million articles in over two hundred languages, including 3.6 million articles in English according to official statistics³. Wikipedia represents an effective and efficient system to organize volunteers' edits, whilst its data storage systems are comprehensive, recording high-quality data due to its establishment of manually-derived structures to manage complex people-driving content (Chan et al., 2008, Cucerzan, 2007). Moreover, all data are created or reserved to assist with some functions of Wikipedia. Most of the available data set has been created by Wikipedia as a side-product of the editing process as well as for archive and back up reasons (Chan et al., 2008, Cucerzan, 2007). Wikipedia provides data dumps of this content, which are also made public for academic research (Hu et al., 2007, Kittur et al., 2007b, Ortega, 2009).

In fact, information from Wikipedia's data include every edit record, such as IP addresses from individual webpages, hyperlinks, images and templates, textual information and so on. Additionally, Wikipedia also has other data resources to record all participation behaviours including: user ID; username and the time and content of any edits. These data help Wikipedia offer a "page history", a record of "recent change" and a "discussion page". "Page history" contains: (A) previous versions with time of edit; (B) the differences between the saved version and the previous one, which shows what has been deleted and what has been inserted; (C) who made the edit (all anonymous participants recorded as the IP address used when the edit was made); (D) any comment or explanation editors might have left to describe the change. Wikipedia also offers a "recent change" link to list the latest edit. Both the "page history" and "recent change" tab are able to help participants track updating activities in a particular article since their last view.

Wikipedia keeps all such digital by-product data for two purposes. First, as an online system, Wikipedia needs to store all information for backup and system security (Barrett, 2008). An "article" is a basic entry in Wikipedia, which consists of a hypertext document with hyperlinks to other pages within or outside Wikipedia. For easy organization, Wikipedia identifies its articles according to the category system which are defined and described by voluntary participants. Articles, categories and hyperlinks serve to combine essential information in Wikipedia (Ayers et al., 2008 p.89). Based on wiki technologies, editing in

³ <http://en.wikipedia.org/Wikipedia>. retrieved 2011-05-27

Wikipedia is an easy job, using a few simple mark-ups that can be translated into Hyper Text Markup Language (HTML) in the background (Barrett, 2008).

Second, in order to provide the open-editing platform, Wikipedia allows its users to track, change and reverse any edits through its systematic entries (Mihalcea, 2007, Yan et al., 2009). Because repetitious changes are implicit to the system, most Wiki-type online applications have storage systems to save all archives that record all previous edits of a page and make it simple to return to a previous version. A wiki-type platform is a simple content management system, which enables all readers to add and modify the content of the website without using a programming language (Ayers et al., 2008, Barrett, 2008). Therefore, scholars (Viegas et al., 2004) point out that, “If the ease of adding a contribution is a distinguishing feature of a wiki, so too, paradoxically, is the ease of removing contribution of others by reverting an edit”(p.576). Although conserving this dynamic and large data requires much more computing power, this storage system ensures that vandalism can be reversed immediately and prevent permanent harm (Suh et al., 2008). Collection of such data relies on the Wikipedia operating system– MediaWiki, which is a server-side technology (Ayers et al., 2008, Barrett, 2008). It allows participants to make instant updates to a web page via a web interface.

An analysis of a database relies on the format of its data set. In Wikipedia, all digital by-product data, including its content and other information, are formatted in XML⁴ (Ayers et al., 2008, Leuf and Cunningham, 2001). These collections can be used in a large variety of XML information retrieval or machine learning tasks like categorization, clustering or structure mapping (Denoyer and Gallinari, 2006). However, if social scientists want to use this data for social research, the XML dumps have to be filtered and converted by standard programs. XML dumps only store data under certain formats, and are not of much use for any sort of direct analysis. Fortunately, many volunteers offer their tools to produce XML dumps as well as SQL⁵ statements, which is able to insert data directly into a database.

Wikipedia collects its content into XML dumps every three weeks but the Wikipedia sites with more edits such as the English site requires a long time to compress the edit information and create split stub dumps according to namespaces⁶. Based on this process, Wikipedia XML dumps include page content, page-to-page links, categories and user editing records. Under its privacy policy, Wikipedia does not provide “user data” including password, e-mail address,

⁴ XML (Extensible Markup Language) is a set of textual data format to encoding documents. It is widely used in web services or data storage.

⁵ SQL (Structured Query Language) is a database computer language designed for data control and querying. The majority of database software are based on the SQLlanguage, such as Oracle and MySQL, etc.

⁶ http://en.wikipedia.org/wiki/Wikipedia_database

watching lists, and 'deleted pages'⁷. The ingenious design of the database allows researchers to incorporate the data into their research directly and without ethical concerns.

Wikipedia data format and dynamic update system shows the difficulties of selecting and integrating useful information from a large database. There are two extra steps before we are able to analyse the data. First, the data have to be re-formatted in order to generate query and carry out analysis; secondly, a dynamic system means that in an entire analysis process, timeline can be a unique indication to mark the different behaviours and records. In order to make use of these data, we need to follow a data mining procedure as introduced in chapter one, changing its format and structure, making it easier to read, find and analyse. Our experienced difficulties and solutions in this process will be introduced and discussed in detail later.

It is undeniable that Wikipedia is able to provide well organized digital by-product data but this is not the only reason we decided to use it as our case study. Another important reason is that, Wikipedia and Web 2.0 applications it represents can provide fascinating topics for social science research. We will now introduce what Wikipedia is and what kind of internet innovation it represents.

2.2 What is Wikipedia?

Wikipedia provides a rich resource of digital by-product data. The whole database with a stable structure, is updated constantly and is easy to operate. These features are important methodological considerations; however, they are not the only reason why we selected Wikipedia as a case study. Wikipedia provides scholars what they crave the most in academic research, the possibility of discovering new trends and new phenomenon, and the opportunity to evaluating and examining previous arguments. In this section, we highlight the uniqueness and novelty of Wikipedia, by introducing its history, basic policies and current operation system. From this introduction and the previous literature reviewed in the third and fourth chapters, we show the audience a research topic that is controversial yet immensely valuable for academic research – Wikipedia.

2.2.1 *Wikipedia's History*

The establishment and development of Wikipedia has been unique a fact acknowledged and praised by many academics. Furthermore, it gave birth to the so-called “wiki” model of collaboration. As a web 2.0 application, the idea of Wikipedia borrows from traditional encyclopaedias, yet it successfully differs from them. This difference and the success of Wikipedia might give us new insights on knowledge building, sharing culture and user-

⁷ http://meta.wikimedia.org/wiki/Data_dumps

centred economics. This section is dedicated to introducing the historical aspects of Wikipedia to compare and contrast it with traditional encyclopaedias, and furthermore how Wikipedia is a form of encyclopaedia which is based on having expertise and professionals in relevant areas.

Wikipedia is an online encyclopaedia edited by millions of collaborative volunteers, available under a GNU Free Documentation License⁸. It was launched by Jimmy Wales and Larry Sanger in January 2001, originally acting as a “feeder” project for Nupedia⁹. Wikipedia not only provides friendly digital by-product datasets; it also exemplifies an outstanding instance of a web 2.0 application. This section will introduce the history of Wikipedia, from which we can see the difference between an idealized professionally written encyclopaedia and the publicly edited Wikipedia.

The idea of establishing an online encyclopaedia edited by volunteer experts emerged and is based on the idea of sharing knowledge (Sanger, 2005). The plan was to build an online encyclopaedia using a scholarly collaboration model, with professional participants and a formal peer-review process. This idea was launched as Nupedia in March, 2000 (N/A, 2001). The only difference between Nupedia and scholarly collaborations was that professional participants voluntarily contributed their knowledge without remuneration and the ownership of content was according to the sharing copyright chapter of the GNU Free Documentation License (Sanger, 2005, Tabb, 2008, Timothy, 2005).

However, one year later Nupedia failed because it did not have enough participants. Because of such a restrictive peer-review process (N/A, 2001), it only had two complete articles published (Timothy, 2005). The failure of Nupedia has been attributed to two reasons (Sanger, 2005). First, Nupedia required that participants should be experienced experts in their field or at the very least PhD students, which are few and far between, even in an online environment (Sanger, 2005). Secondly, published articles in Nupedia had to go through an official editing process, which included a critical peer-review by readers followed by comments made by editors (N/A, 2010, Sanger, 2005). The entire system was too complicated by far. And even editors who were professional in this area; found it hard to comment on all reviews by anonymous readers (Sanger, 2005, Timothy, 2005).

To break these two barriers, the founders of Nupedia started a similar project, but with more open and free-style — Wikipedia — in which everyone could access the edit system without

⁸ The GNU Free Documentation License is a copy-left license for free documentation, designed by the Free Software Foundation (FSF) for the GNU Project. It gives readers the rights to copy, redistribute, and modify a work and requires all copies and derivatives to be made available under the same license.

⁹ Sanger, Larry (2001-01-10). "Let's Make a Wiki". Internet archive.is archived from the original on 2003-04-14.<http://web.archive.org/web/20030414014355/http://www.nupedia.com/pipermail/nupedia-l/2001-January/000676.html>. Retrieved 2009-01-26.

any formal educational requirements (Marks, 2007). Meanwhile, the articles are able to be published immediately after editing. The bureaucratic structure of Nupedia disappeared, replaced by this new collaborative model. Thus, the mass collaboration model of “online encyclopaedia” has been established in Wikipedia (Marks, 2007, Sanger, 2006, Sanger, 2005).

The description of how Wikipedia came about can give us a clear idea how the birth of Wikipedia is the result of an innovative way of writing and editing traditional encyclopaedia material. How to discuss these innovative and creative features has become a heated topic that is attracting the attention of many social scientists. Two differences should be clarified. First, we need to address the differences between online encyclopaedia and traditional encyclopaedia, i.e. between Nupedia, Wikipedia, and paper encyclopaedias. This difference, which is based on innovation supported by technical advancement, directly gave rise to the ability of Wikipedia to redefine the meaning and function of encyclopaedia in the information era. We also need to differentiate Nupedia from Wikipedia. In this section we discuss the difference between the internet-based encyclopaedia and traditional encyclopaedia. In the following section, the difference between Nupedia and Wikipedia will be discussed through the reliability arguments surrounding Wikipedia.

First, as an important collection of reference works, an encyclopaedia provides a summary of human knowledge in comprehensible terms (Ayers et al., 2008, Featherstone and Venn, 2006). In doing so, it must be capable of transmitting all classical knowledge from older to newer generations (Diderot as quoted in Hunt, 2007). Moreover, it can be a forum for new scientific issues, a platform for the advancement of knowledge, and a handbook of new discoveries. Finally, in the process of editing the encyclopaedia, authoritarian control has been resisted as the target of censorship (Ayers et al., 2008 P.34). The general way to create an encyclopaedia is to let senior editors consolidate work from disciplined editors who then write entries following instruction (Steinberg, 1951, p.6 in Featherstone and Venn, 2006).

The new encyclopaedia might be a digital notion, a fusion between dictionary, media and archive. “Dictionary” is the function inherited from tradition. The traditional form of the encyclopaedia cannot compete with the potential of internet technologies to change material quickly and easily as real-world events develop (Berners-Lee and Fischetti, 1999). It is conceived that the Internet could provide hypertext crossing each text which is easy to move by copy-and-paste, which will probably change the way to categorize entries. Featherstone and Venn (2006) claimed that a new internet-based encyclopaedia could be built to combine the manual of knowledge function with media-like roles, akin to a newspaper in spreading news to the public (Wikipedia has a news page with a similar function).

A traditional encyclopaedia emerged as a finished project and published volume for searching for items. In contrast, new internet-based forms have an ability to store the whole editing procedure: the process of drafting, researching, discarding and the recovery of “rubbish”, because there is no cost of publication (Ayers et al., 2008 p.36). Wikipedia is able to record a history of knowledge transitions or how knowledge definitions have been developed within the encyclopaedia. Moreover, an internet-based encyclopaedia offers a community for authors to discuss their edits and, even collect reader views and feedback from connected forums. Therefore, an internet-based encyclopaedia does not just retain its essential function as a collection of knowledge, but also changes its format and social responsibility for enlightenment according to culture and technological development. In the following section, we discuss possible formats for new encyclopaedias.

From the introduction and discussion above, it is obvious that there are two distinct differences between a traditional encyclopaedia and an “online encyclopaedia”. Firstly, and most importantly, an online encyclopaedia is free to use but a normal published encyclopaedia is not. The main reason for this is that the ideal ‘online encyclopaedia’ encourages volunteers to edit rather than a paid workforce, potentially motivated by academic reputation (Benkler, 2006, Leuf and Cunningham, 2001). The participants in this online encyclopaedia are unpaid volunteers, whereas traditional experts for editing encyclopaedia might be motivated by monetary benefit and academic reputations (Ayers et al., 2008, Magnus, 2006). This feature could reshape the traditional way of contributing wisdom and creating value; at least shifting the trend away from the exchange of labour and money in traditional society (Benkler, 2006).

Second, an encyclopaedia is a printed publication, which means that its contents cannot be changed once it is published except by publishing a new edition. In contrast, the contents of Wikipedia are produced by a dynamic system that allows knowledge and information to be easily updated. In this respect, some studies believe, an internet-based encyclopaedia has some features in common with news reports (Lih, 2004). Last but not least, a traditional encyclopaedia has one feature in common with internet-based encyclopaedia, as both arguably set out to collect all branches of knowledge and information into ordered categories (Gabrilovich and Markovitch, 2007, 2009). The function of continued reinterpreting knowledge and spreading of social news enables the online encyclopaedia such as Wikipedia to witness and record changes in society, which is undoubtedly a step forward. In summary, as Wikipedia is an online encyclopaedias, the study of it can help us to better grasp how ICTs transformed current society.

2.2.2 Wikipedia’s policies

Wikipedia's policies are the reason it maintains a strong sharing platform as a popular web application. To understand Wikipedia's policies, we need to understand what the primary goal of Wikipedia is and what these policies are working towards.

Wikipedia is an online arena designed to create a free encyclopaedia in multiple languages for the entire world to access and use. The word "Wikipedia" literally suggests an online encyclopaedia that is established in a quick way by its participants. Since its establishment in 2001 by the Wikimedia Foundation (Ayers et al., 2008, Lih, 2009b, Reagle, 2010), it has become an increasingly popular online resource for people searching for information, at the top of Google's ranking system (Woodson, 2007). At present, it is the largest, most popular and dominant general reference platform currently available on the internet (Tancer, 2007). In order to create a worldwide internet platform that is easy to use and free to access for volunteers, Wikipedia established a series of policies, some of which are widely accepted by convention while others are written down as rules. By summarising these policies, we hope audiences can further appreciate Wikipedia's uniqueness.

Wikipedia has been shaped by the engagement of millions of volunteer participants. With the development of information and communication technologies, it has become a significant example of mass collaboration, providing information shared among its millions of members. Wikipedia allows its content to be created, edited, corrected and even deleted by the public without any special requirements (Ayers et al., 2008). The English Wikipedia passed the 1 billion edit milestone on 16th April 2010, and as of May 2011, it had 3,646,518 articles consisting of approximately 1 billion words¹⁰. It is a significant example for any exploration of mass collaboration because it attracts a large number of users (Reagle, 2010). By May 2011, more than 1.4 million people had registered accounts on the English Wikipedia, among whom there were approximately 145,156 active participants who had made edits in 30 days from April, 2011 to May, 2011. The English version of Wikipedia has a large group of people with equal access to this online community. In this study, the English Wikipedia constitutes the subject of mass collaboration.

1,790 administrators have been elected to maintain Wikipedia by giving them privileged authority. They carry extra technical power—generally the ability to block other participants or block articles from being edited. The nominee to become an administrator has to answer five initial questions and be voted by other registered users within seven days¹¹. Above these are the "bureaucrats", who have the power to appoint administrators¹². They are generally selected by "stewards"—employees of the Wikimedia Foundation. So far, there are only

¹⁰ All data in this part come from website: <http://en.wikipedia.org/wiki/Special:Statistics>

¹¹ Policy details come from <http://en.wikipedia.org/wiki/Wikipedia:Administrators>

¹² Policy details come from <http://en.wikipedia.org/wiki/Wikipedia:Bureaucrats>

about 37 stewards appointed by a seven-person Wikipedia Foundation Board. Their autocratic power covers different Wikipedia language versions¹³. It has always been a topic of debate how much influences each of these roles play in the development of Wikipedia.

Another basic rule by convention is that the minority obey the majority. Because Wikipedia is a public platform that allows mass participation and equal editing rights, it is important to have set rules on how to make decisions on direction and how to make judgement when disagreement occurs. Wikipedia uses the “consensus-dominated” mode to replace the traditional “editor-in-chief” in ruling when dealing with conflicts in edits. Such decisions may determine the composition of a certain article, the content that it covers, and the accuracy of the vocabulary usage etc. In Wikipedia, the primary way to decide about editorial process is consensus, which refers the neutrality and verifiability to make decision. Generally, the editorial conflicts and arguments can be discussed in the relevant discussion page, where the involved participants and readers can reach a consensus to resolve problems. How to realize a truly consensus dominated participation platform for millions of users is the other question that attracts many researchers’ attention.

The third important principle of Wikipedia is in regards to its collaborative policies. Wikipedia provides a series of explanations and guidance to help users participate and collaborate; emphasising equal editing rights and consensus dominated decision making processes. The large number of participants, editors as well as viewers makes Wikipedia an excellent expression of “mass collaboration” – an important internet phenomenon (Tapscott and Williams, 2006).

In summary, Wikipedia’s policy framework involves three principles: detailed description of division of labour; consensus dominated decision making; and mass collaboration with millions of participants. The establishment and implementation of each of these policies are supported by many guidelines, the study of which give social scientists more space to explore the internet sharing model and new organization mode. Therefore, we once again showed that Wikipedia as a valuable research subject, supporting our choice of Wikipedia as a case study in addition to our methodological considerations.

2.3 Debates surrounding Wikipedia

As we mentioned above, a completely open, free and unlimited edit mode, as represented by Wikipedia, differs remarkably from that found in traditional encyclopaedias. The differences have attracted much doubt and concerns for Wikipedia’s reliability (Magnus, 2009, Sanger, 2009, Wray, 2009) and quality (Anthony et al., 2005, Hu et al., 2007). On the other hand,

¹³ Policy details come from <http://en.wikipedia.org/wiki/Wikipedia:Stewards>

some scholars believe with an advent of the internet age, the success story of Wikipedia indicates that the traditional top-to-bottom management mechanism in the knowledge industry is outdated (Beschastnikh et al., 2008, He, 2010a). This change needs to be appreciated and studied. In this section, we will discuss relevant opinions by these two groups of scholars. Through such a discussion, we want to demonstrate that in addition to providing a more complete database from a methodological point of view, academic research based on Wikipedia can also provide an interesting topic for exploring new phenomenon from a practical research perspective.

2.3.1 *Wikipedia's reliability*

Many studies have questioned whether Wikipedia can provide a standard quality of knowledge to the public (Ebner and Zechner, 2006, Magnus, 2009, Waters, 2007). Some scholars have attempted in different ways to evaluate the credibility and reliability of articles in Wikipedia (Ortega et al., 2008, Stvilia et al., 2005a, Wilkinson and Huberman, 2007a), whereas others researches focus on how to judge the reliability of Wikipedia from epistemological perspective (Fallis, 2008, Magnus, 2006, Tollefsen, 2009). The former literature will be discussed when we formulate our empirical works and the latter literature will be addressed below to explain what strong and essential oppositions and disagreements Wikipedia has encountered by scholars.

With the increasing popularity of Wikipedia among students and experts, some people have questioned the quality of its content (Denning et al., 2005). More and more teachers have announced they were strongly against their students regarding Wikipedia as a reliable source to be cited (Cohen, 2007, Waters, 2007). Besides, more and more professors have found incorrect definitions and explanations in Wikipedia which were edited by unprofessional volunteers (Chesney, 2006). However, sometimes Wikipedia is claimed to be a useful teaching resource (Noveck, 2007).

There were previously a series of works to examine Wikipedia and to offer epistemological theories on Wikipedia. We chose them as primary literature to address Wikipedia's reliability. Fallis (2008) believes its contents can be trusted because they offer "testimony" to the public. According to this hypothesis, the final measure of the worth of Wikipedia is equal to the question of whether Wikipedia can be recognized as a testimony (Magnus, 2008, Magnus, 2006, 2009). A majority of studies that concentrated on the quality of Wikipedia show a suspicion of the veracity and trustworthiness of its content. They offer four reasons to explain why Wikipedia is not trustworthy as an online encyclopaedia: authority; responsibility; persistence of content; and consequence-free.

Firstly the authority of Wikipedia is what was questioned in the earliest debates. The open edit system which is regarded as one that discourages expertise and specialists to join in is what sustains Wikipedia (Sanger, 2009). In the open-edit policy, Wikipedia allows any individual to edit, amend and even delete content from articles, which provides the equal right for all participants, without considering the professional knowledge background. Analysts argue that this mode of cooperation does not require the participant to have the specific knowledge to participate in the respective edit of the article, and such a problem will affect the quality of Wikipedia overall (Wray, 2009). This seeming “disregard of authority” is thought to cause the following problems: certain content that is written by an expert could be easily altered or even deleted by non-professionals. In extreme cases, when professionals and non-professionals are involved in an ‘edit-war’, the experts are more likely to disengage from making reversions as they have no time for it, and therefore fail to incorporate their opinion in the article (Sanger, 2009). These worries and doubts to a large extent are due to the lack of confidence in Wikipedia’s authority, as compared to the authority system in traditional encyclopaedias (Wray, 2009). More importantly, on Wikipedia people who have corresponding knowledge generally do not obtain the deserved respect and authority to “guard the door” (Sanger, 2009).

Secondly, as articles are delivered to their audience as a finished product that does not contain any links between one edit and the corresponding editor, such an expressive manner renders no direct responsibility of the editor for their own edited content. Although we can trace editors’ user name and the whole history of the edit under the same user name, it is still difficult to judge his or her knowledge contribution. The anonymous edit policy of Wikipedia (Wikipedia:Policies_and_guidelines) is accused of potentially exacerbating the situation of low responsibility in edits, because even if editors are tracked down to their user name, there are no real consequence. Thus, Wikipedia’s quality might be questioned because there is no certain mechanism to hold participants responsible for their edits (Sanger, 2009).

Thirdly, researchers also believe that because Wikipedia’s content is dynamic, it is difficult to deliver to the audience articles with guaranteed quality (Magnus, 2009, Tollefsen, 2009, Wray, 2009). The information editing process is dynamic; a fact which means even correct information or knowledge is changeable without authorization. Readers cannot trust content even if they previously approved the same articles (Magnus, 2009). This causes the accuracy in articles to fluctuate, a process named “doxastic instability” by Tollefsen (2009). With dynamic content, any individual correcting action by another can generate disorderly and unsystematic statements in sense.

Fourth, there is no contribution and profit exchanging relationship in Wikipedia's edit participation (Lipsch, 2009) to encourage participants and maintain motivation. In common sense term, contribution to Wikipedia cannot redeem any social profit, including: money, reputation, recognition and promotion. In comparison to making contributions to other academic publications, contributors of Wikipedia have been concerned that they cannot obtain any reputation from their works, nor the property of their work, which means "an invisible hand cannot ensure quality in Wikipedia" (Wray, 2009).

In conclusion, Wikipedia has been questioned as to its quality in four different ways as discussed above, which can be summarized in terms of two main criticisms. First, scholars question the incentive system in Wikipedia (Javanmardi and Lopes, 2010, Lipsch, 2009, Sanger, 2009, Wray, 2009). According to the traditional collaborative system, participants can be awarded by monetary and reputation benefit through their contribution. Meanwhile, the participants will also be punished if their contributions do not achieve a certain standard. In Wikipedia, many studies believe that this 'reliable' incentive system is not functioning anymore. Therefore, they came to the conclusion Wikipedia is not reliable (Magnus, 2009, Tollefsen, 2009, Wray, 2009).

Additionally, the organizational system of Wikipedia including the recruitment of editors and resolving conflicts has been cast into doubt (Lam and Riedl, 2011, Sanger, 2009). People who have been recruited into the traditional encyclopaedia editing process should be recognized as experts in the relevant editing area, which Wikipedia's open participation system cannot guarantee. On the other hand, Wikipedia may not have the effective and reasonable policy to solve editing conflicts when editors have different views in the same article (Sanger, 2009). Due to both of these traditional solutions not working in Wikipedia, the contents resulting from conflicts could be questioned (Sanger, 2009, Waters, 2007, Wray, 2009).

In fact, all arguments and questions can be overturned if Wikipedia can be found to operate a totally different mechanism which is immeasurable by theories and experiences from a conventional encyclopaedia. These doubts are based on the acceptance and recognition of scientific research and the function model of traditional encyclopaedia. As a social science scholar with a neutral attitude, we ask ourselves whether it is possible that Wikipedia offers a new collaboration model. Furthermore, although there are many debates about Wikipedia, our argument is that, "what is rational is actual and what is actual is rational" (Hegel cited in Fackenheim, 1970P.690, Hegel and Dyde, 2008). Given that the Nupedia, which relied wholly on a traditional cooperation mode, versus Wikipedia, which has broken tradition completely but maintained successful development for over ten years, should we not focus more on its innovation rather than judging and criticizing it to traditional standards?

2.3.2 *Wikipedia as a new innovation*

Although the content and quality of Wikipedia are subjected to questioning from an epistemological perspective, in other research, it is not only regarded as a qualified resource in the spreading of knowledge, but it also represents a novel contribution mechanism considering its achievement both qualitatively and quantitatively.

In fact, scientists who have compared the quality of articles between Wikipedia and the e-encyclopaedia Britannica have approved Wikipedia's quality. The latter has been established via traditional scholarly collaborative means. In a comparative analysis of the current contents of both, results showed their accuracy to be very similar (Giles, 2005). This comparison supported the claim that Wikipedia is at least as reliable as Britannica. The '*Science*' magazine even said "99.8% of Wikipedia's articles are error-free and brilliantly written" (Giles, 2005).

Entries in Wikipedia can be continuously revised and improved, which is regarded by many scholars as feature. Wikipedia has been identified as an up-to-date information resource (Dee, 2007), which is able to expand to all real-time knowledge and news. This idea that Wikipedia is a constantly updating information system is also proved by studies in quantitative analysis fields (Lih, 2004). The open participation system in Wikipedia has been cited as a big advantage for both participants (Lih, 2009a) and audiences; and has been described as the top destination for new information seekers (Woodson, 2007).

Additionally, many studies have used Wikipedia to either observe the application of internet technologies (Denoyer and Gallinari, 2006, Gabrilovich and Markovitch, 2007) or explore new innovation (Lih, 2004, Tollefsen, 2009) by using its digital by-product data. Through this research, the quality of Wikipedia articles has been approved from a quantitative perspective. Although they cannot thoroughly refute the questioning and doubt about Wikipedia articles coming from an epistemological perspective, they at least offer other possibilities: Wikipedia may have established a new cooperative mode that differs from previous ones, in which a total egalitarian platform for editing helps to ensure the quality of the articles (Reagle, 2010, Tapscott and Williams, 2006). Studies believe that it brought a new type of economics beyond the traditional "paid and gain" process (Benkler, 2006).

From a macro point of view, Wikipedia as a case study has brought many discussions to different fields, and scholars try to use various definitions to interpret the new functioning modes of Wikipedia. In the case of "Prosumption" (Beer and Burrows, 2010, Ritzer and Jurgenson, 2010), Wikipedia has been discussed as "peer-to-peer produce" (Benkler, 2006). In the case of a large number of participants, Wikipedia has been defined as mass collaboration (Tapscott and Williams, 2006), especially for its "long tail" model which can be

attributed to 'Web 2.0' (O'Reilly, 2005). Although these discussions did not reach any consensus about Wikipedia's mode, these studies themselves already proved evidence that a new collaboration and sharing model has been created in Wikipedia, and this model may change our understanding and judgement about productive systems in online society (Hippel, 2006).

From a micro point of view, scholars are still discussing the quality of articles in Wikipedia based on the common sense that Wikipedia provides a new model. Although there are debates about whether Wikipedia provides an acceptable level of knowledge to the public (Chesney, 2006, Waters, 2007), scholars still believe that Wikipedia provides an innovative process of a massive group of people working together (Chesney, 2006, Ortega et al., 2008). On such a basis, researchers turned their attention to finding out what administration model Wikipedia employs to manage the participants and their edits (Hu et al., 2007, Ortega et al., 2008, Stvilia et al., 2005a). From the microscopic view, the collaborative mode of Wikipedia is regarded as an innovation for two main reasons: participants in Wikipedia collaborate to write information on the internet; second, participation in Wikipedia is unlimited and is based on dynamic access (Ponzetto and Strube, 2007). We summarize them into a model where there is a large body of participants collaborating on a common project.

The success of Wikipedia has, more than once, been attributed to its establishment of a contribution model. In such a model, all participants have reached a consensus on basic understanding. Not only does this consensus give them the basis to abandon their differences and work together, but it also becomes the direction to guide their contribution (Hippel, 2006). First, every person related to Wikipedia makes efforts or at least understands that Wikipedia is acting with the single-goal of establishing an online encyclopaedia, because the founders of Wikipedia put this aim on the top of every page (Winer, 2008). Second, the public accept the fact that the founders established Wikipedia with the single goal of producing a free online encyclopaedia (Orlowski, 2005, Woodson, 2007). The founder of Wikipedia addressed that the purpose of establishing Wikipedia is to create the online encyclopaedia for free access (Wales, 2005a, b). Third, the communities within Wikipedia are operated to improve the "online encyclopaedia"¹⁴. Therefore, the majority of editors bear this purpose in mind as they contribute to the establishment of the online encyclopaedia. Meanwhile, readers are aware of this single goal when they read Wikipedia, whether they can convince themselves to trust its reliability or not.

2.3.3 *Disseminating the Wikipedia spirit*

¹⁴ Many chapters ordered by country state they work to assist the "online encyclopedia"—Wikipedia. Their official views can be found on the Wikimedia link web.

In addition to the epistemological studies questioning the validity of Wikipedia and the explorative studies investigating the innovations of Wikipedia from both macroscopic and microscopic views, there are many other studies that attempt to directly emulate Wikipedia's "innovative model".

Because of Wikipedia's popularity, many studies have attempted to apply Wikipedia's model to other fields, such as business (O'Reilly, 2005, Prahalad and Ramaswamy, 2010), medical communities (Boulos et al., 2006, McLean et al., 2007) and education (Brown and Adler, 2008, Duffy and Bruns, 2006, Ebner and Zechner, 2006). Generally, some of them use Wikipedia as a primary source to discuss the possibilities of applying its collaborative mode in different fields (Braun and Schmidt, 2007, Tredinnick, 2006). We will now consider if it is possible to transfer the successful principle of Wikipedia to other areas, such as business studies, the medical information community and the learning environment.

The popularity of Wikipedia as a means of sharing information encouraged experts to discover how to introduce wikis into business (Tredinnick, 2006). Experts tried to introduce Wikipedia-like features into their firms—collaborating and sharing information into other domains (McKnight and Chervany, 2001), such as customer service and system of organization. In business research, some have analysed not only the influence of Wikipedia—'wiki-spirit', but also have investigated 'collective wisdom'—building upon an initial core value by public contribution, based on Wiki technologies (O'Reilly, 2005, Tredinnick, 2006). Studies recommend commercial services to provide the "customer-centricity" platform based on Wiki technology in order to hear feedbacks from customers (Prahalad and Ramaswamy, 2010). Another suggests building up the sharing-information between businessman and customer, or supplier and demander to reduce the cost of cooperation and communication (Majchrzak et al., 2006, Wagner and Majchrzak, 2006).

There are many examples to support the application of the Wikipedia's model in the medical field. Wikipedia's technological and organisational format is already expanding into many medical areas. Many Wikis provide a communication platform for medical researchers and doctors, such as Wiki Surgery (<http://wikisurgery.com>), Healtheva (<http://www.healtheva.com>) – which is open access to every registered user – and Sermo (<http://sermo.com>) which is accessible by people with medical credentials. Moreover, these wiki applications also offer communication between experts and interested members of the wider population. Wiki Flu (<http://fluwikie.com>) offers expertise to help people make appropriate preparations for an avian influenza pandemic.

In education studies, the appearance of Wikipedia is the signature for many educational experts who thought they had found a new way to improve the efficiency of learning for

students (Brown and Adler, 2008, Duffy and Bruns, 2006, Ebner and Zechner, 2006). However, students do not always have enough interests to engage in this “experiments” on Wiki platform, because of the boundary between teachers and students, the difficulties of contributing on the Web, the lack of motivation and so on (Duffy, 2008, McLean et al., 2007). In fact, due to the blurriness of studies of motivation of participation on Wikipedia (Forte and Bruckman, 2008, Riehle, 2006), it is considerably difficult to create a Wikipedia-like platform where students have similar motivation to share their knowledge and collaborate on the content of courses (Duffy, 2008, Duffy and Bruns, 2006, Ebner and Zechner, 2006). However, such applications did not consider the feasibility of adopting Wikipedia features, instability of students’ motivation.

What we have covered in this section and our intention of studying Wikipedia, together with our empirical works in later chapters are not necessarily clearly related. Our descriptions of these studies have tried to argue that the new model of Wikipedia has already been widely accepted by applied sciences, and relevant experiences abound. However these experiments must be based on a deep understanding of Wikipedia’s collaborative model. If there are any deviations from the true model, such misunderstandings could tarnish these experiments completely. Therefore, we argue that as long as Wikipedia has been proven to bring a new mode of participation because of its establishment and development in the last decade, the first job for the Academe is to clarify the organizational and collaborative model operating there.

In this chapter we introduced the academic discussions involving Wikipedia and the new participation model it represents. These discussions fall into three categories: the debates concerning the current function of Wikipedia and the quality of articles; the discussion and investigation of the innovativeness of Wikipedia, macroscopically and microscopically; and the recycling of Wikipedia’s innovative model. In fact, despite that there is still much debate on the quality and operational model of Wikipedia and the application of that model to other areas, we have reasons to believe that overall the discussion and investigations of the collaborative model hidden in Wikipedia is academically meaningful – this will be the primary purpose of our empirical work.

2.4 Related work

We will now introduce how different researchers have used different data resources to complete their studies. Just as we have chosen Wikipedia as our case study because its digital by-product data are comprehensive and safe to use; many scientists have used Wikipedia as their working database. Although their research itself does not help ours directly, their methods of data processing are clearly instructive to our research. Such is another remarkable

benefit that scientific research based on Wikipedia's digital by-product data can offer valuable experiences to teach us how to extract data and how to generate them for our research purposes. The use of digital data attracts ethical concerns; therefore how much digital data should be used and how to use it is still the focus for discussions. On the contrary, Wikipedia is a digital by-product resource that does not involve any private information, and thus can be used safely in academic researches without infringing privacy issues.

2.4.1 Technical experiments based on Wikipedia and its data

As we discussed above, Wikipedia generously makes its entire database available to the public, which for academics opens up a large source of digital by-product data (Gabrilovich and Markovitch, 2007, Kittur et al., 2007a, Kittur and Kraut, 2010, Strube and Ponzetto, 2006). Wikipedia, as much of literature has pointed out, is, "A huge mine of information about words and concepts" (Milne et al., 2006) and has attracted many scholars to use its data in different ways with specific techniques. Some technical researchers offer their fresh tools to aid Wikipedia with semantic analysis (Volkel et al., 2006), whereas others use the digital by-product data offered by Wikipedia to examine new technologies (Budanitsky and Hirst, 2006, Gabrilovich and Markovitch, 2007, Strube and Ponzetto, 2006).

The data resource from Wikipedia, a well-structured and large size database, has been used widely to examine a variety of new software and semantic approaches with promising results (Ahn et al., 2005, Bunescu and Pasca, 2006). For instance, Strube and Ponzetto (2006) presented their tool "WikiRelate" on computing semantic relatedness by using Wikipedia data. These experiments are based on the belief that natural machine language can help to categorize web content with semantic relatedness more quickly and accurately. With a similar purpose, (Gabrilovich and Markovitch, 2006, 2009) have also proposed Explicit Semantic Analysis, as a new approach to compute the semantic relatedness of natural language texts from Wikipedia data. Many scientists have invented new tools or methods in which they have used Wikipedia data to evaluate this process which provides one example of how scientists use digital by-product data in their studies (Cucerzan, 2007, Fachry et al., 2007, Gabrilovich and Markovitch, 2006, 2009, Muchnik et al., 2007, Ponzetto and Strube, 2007).

In fact, in addition to scholars who use Wikipedia data to examine their own research, others believe that their research has a unique contribution to the maintenance and development of Wikipedia. Although Wikipedia and its categorization system were both established and are developed by collaborative human effort, computer scientists and program technicians have also attempted to apply their tools to help categorize and organize content. Some even attempt to reformat Wikipedia editing techniques to allow semantic recognition (Milne et al., 2006, Volkel et al., 2006). On the other hand, scholars have discussed some features of

Wikipedia to address possibilities to use it in challenging technologies (Gabrilovich and Markovitch, 2006, 2009, Strube and Ponzetto, 2006).

Research that aims to understand Wikipedia within a purely scientific realm is not directly relevant to this thesis. However, they do review and summarize some features of Wikipedia's data resource. First, Wikipedia provides entries on a large number of named entities and specific concepts, which can be located by semantic relatedness (Milne et al., 2006, Milne and Witten, 2008). Second, the attraction of Wikipedia as a resource lies in its database size, which can meet technical requirements and scalability issues. Many technical experiments need to test a great number of web data, which can be difficult to find (Milne and Witten, 2008, Volkel et al., 2006). Third, Wikipedia's categories are based on a user-supplied tagging system that enables the participant to categorize any Wikipedia entries' content (Capocci et al., 2008, Priedhorsky et al., 2007). Compared to systematically engineered categories, these include more unordered sub-categories (Strube and Ponzetto, 2006).

2.4.2 Exploring Wikipedia's mechanisms and incentives

In addition to purely scientific research, some scientists and social scientists with scientific backgrounds have begun to use scientific methods for treating data in order to conduct research on social scientific topics using Wikipedia data. This research has varied forms and themes, but they have in common a social scientific angle and make use of scientific methods to process Wikipedia's digital by-product data. We introduce them here in order to understand the process of studying Wikipedia with different resource which will inspire our own study.

As we discussed before, Wikipedia not only offers quantitative data but also generously provides all editing content for download (Braendle, 2005, Emigh and Herring, 2005). In large part due to this feature much research has emerged based on the content analysis of this data. (Emigh and Herring, 2005) measured formality and informality in 15 entries, comparing it to a hard print encyclopaedia that shares its content online but retains traditional editing rules. They claim that Wikipedia almost matches the standard set in print. More importantly, Wikipedia has been found to become more standardized and formal if more control is exercised over the contribution. Using similar methods, Pfeil et al (2006) discussed whether contributions can be affected if users come from different cultural backgrounds through content analysis in four different language versions. They revealed that cultural differences observed in the off-line world have been brought into Wikipedia contributions by millions of amateur authors. Focusing on 450 articles on the German Wikipedia, Brandle (2005) examined the quality of Wikipedia based on content analysis. The results suggest that the higher interest (the number of edits and unique participants) and relevance (the Google ranking) varies, the better its content quality. Although limited in their samples, their studies

demonstrate the possibility to use digital by-product context data from Wikipedia for specific academic purposes (Braendle, 2005, Emigh and Herring, 2005, Pfeil et al., 2006).

Other major studies have also assessed the quality of Wikipedia's articles, proposing a number of different methodological approaches. Lih (2004) analysed the difference in quality of Wikipedia articles before and after they had been cited by the press with metrics, however no justification was put forward. (Giles, 2005) published his paper in *Nature*, claiming Wikipedia was on equal standard with the *Encyclopaedia Britannica*, another good-reputation online encyclopaedia based on the contribution of professional authors. After coming up with a set of metrics related to article quality, (Stvilia et al., 2005a, Stvilia et al., 2005b) also assessed the quality of Wikipedia. But the way of solely relying on metrics to test the quality of articles has been questioned because invisible factors may have been ignored, for example, the popularity of articles can affect the number of viewers, and more contributions and corrections from viewers may result (Wilkinson and Huberman, 2007a). These factors are able to affect an article's popularity, namely the possibility of polishing articles by viewers.

Wikipedia has also been examined as a complex and dynamic system. Some claim that its development depends on the capacity to provide benefits that outweigh the costs of participation; depending on studies on the cost of online participation (Moreland and Levine, 1984, 2001, Uzzi, 1999). Meanwhile, other experts believe that the development of interpersonal fellowship, companionship and affiliation has helped to establish and maintain online social structures (Walther et al., 1994) like Wikipedia through many sorts of benefits (Riehle, 2006, Walther et al., 1994); it is a platform for individual people to access information and quickly share and discuss their ideas with others (Hoadley and Enyedy, 1999). Collecting contributions from individuals is another good angle to build up a picture of notable achievements, such as software development and political action (Butler, 2001, Ogan, 1993, Oh and Jeon, 2007). These researchers seem to believe that the social structure can provide a variety of benefits for their participants, encouraging them to become deeply involved and enabling the whole network to flourish. Some of these benefits are based on the fact that participants' contributions get explicitly recognized as a 'featured article' (Riehle, 2006). Those studies may be able to introduce Wikipedia's incentive system.

Viegas and his colleagues (2004) discussed several collaboration patterns by investigating the dynamics of Wikipedia, through which they introduced a new tool to visualize the dynamics of the editing process. One of the contributions of this research is the development of a new exploratory tool to simultaneously show broad trends and outline abnormal episodes through scanning normal individual editing behaviours. This research also examines the thriving

model of Wikipedia, through which they claim we can better understand the mechanisms for reaching the consensus described here which may apply in other contexts.

To study individual behaviour in Wikipedia, researchers not only focus on participation benefits, but also on the self-selection of authors, as well as engagement and retention in the editing process. Ciffolilli (2003) makes use of team and club organizational theory—transaction cost economics — to address the self-recruitment process in Wikipedia. He states that a principal reason for the success of Wikipedia is due to reducing the transaction costs of submitting contributions, which is also an important feature in mass collaboration. Moreover, (Ciffolilli, 2003) argues that accumulated editing records can be counted as a source of authority just as much as reputation. The lineal relationship between the number of participant's contributions and the growth in reputation of individual editors on Wikipedia was established in the study. However, the author has not provided enough evidence and experimental analysis to support this claim.

Although Wikipedia has been recognized as a collaborative work, there are debates about who actually edits on it, or who contributes most (Kittur et al., 2007a, Voss, 2005). Generally, scholars are suspicious of whether Wikipedia's information is really created by millions of volunteers or just a few elite members (Kittur et al., 2007a, Kittur et al., 2008, Kittur et al., 2007b, Lih, 2004). Jimmy Wales, one of Wikipedia's founding members, has noted that the majority of Wikipedia's articles are edited by a small cabal of interested participants¹⁵. Additionally (Ortega et al., 2008) claim that a “core” group of authors is responsible for the majority of total contributions to Wikipedia by using Lorenz curves and Gini coefficients. They suggest Wikipedia might need to focus on this small group to improve the efficiency of its projects (Ortega and Barahona, 2007). However, Kittur et al (2007b) offers a different suggestion, claiming that the effective way to reduce conflict and article errors is to increase the number of users rather than depending on the same ‘core’ people. Instead of answering the question, “Who edits Wikipedia?”. His paper concentrates on formulating the conflict cost in the Wikipedia editing process.

Nevertheless, other scholars have used more specific ways of analysing this topic (Priedhorsky et al., 2007). Swartz (2004) used a specific measure to count the number of letters in each edit, rather than just calculating the number of edits. He then argued that less frequent participants were providing a much larger proportion of Wikipedia's content. Although this result only relied on the final version of each article, it provided another angle from which to define who the most active participants were. Similarly, Anthony et al.(2005) noticed that two main groups may be responsible for most content by formulating the

¹⁵ Documents come from Wikipedia archives

percentage of aggregated content. They claim that registered participants with a high interest in obtaining reputation have created a lot of content, and anonymous participants only sporadically made high quality contributions. Apparently, all authors have their own understanding of, “Who edits Wikipedia”. However, as a dynamic system, the answer will always change along with the transitions in the wider phenomena of mass collaboration, which is still worth investigating further.

When some scholars have concentrated on identifying who edits Wikipedia, other social scientists tend to concentrate on the hierarchical system built between millions of participants (Antin and Cheshire, 2010, Forte and Bruckman, 2008, Reagle, 2007). Regarding to the policy of open edit in Wikipedia, anyone is able to edit, delete and revise articles in Wikipedia without any authorial permission, by which Wikipedia intends to make a friendly environment for participants and potential participants by eliminating authorial control. In this system, anyone wanting to take on the administration responsibility can nominate him or herself on the specific wiki page by making a statement to that effect. After going through the Q&A section, a nominator can enter into the election process by voting for anyone whether a registered or anonymous user. More importantly, Wikipedia officially suggests that contributions made by nominated individuals should not be a significant factor in these elections. If the nominated person gains the half percentage of the vote from people who posted their opinions there, then s/he will be made an administrator with the authority to block articles or users. Based on this process, Wikipedia has been studied as a self-governing process (Beschastnikh et al., 2008, Kriplean et al., 2008), with systematic self-selective recruitment for administration committees (Ciffolilli, 2003). This self-selected and self-governing process have also been approved by leading members of the Wikipedia community: “The welcoming committee is a self-selected group of people who say they will help with welcoming new users” (Riehle, 2006). Self-selective governance where people can volunteer to take responsibility rather than follow pay-and-gain economic rules (Benkler, 2006), may be able to eliminate hierarchy.

Others argue that Wikipedia was established with an integrated hierarchy (Butler et al., 2008). Kittur and Kraut (2008) addressed the importance of a small influential group who created articles in Wikipedia at the beginning. They believe that early participants act as leaders by implicitly building, “The direction, scope and structure of an article (Kittur and Kraut, 2008); while at the same time set a framework of collaboration. Therefore, the small group of early participants took a leadership role in creating Wikipedia as it is today. Moreover, Wiki platforms, not only breed a peer-based non-hierarchical community, but also create hierarchical systems by providing a series of facilitation, support and management policies and rules (Butler et al., 2008). Clearly, whether Wikipedia is hierarchical or not is debatable

after identifying its administrative commitments, power transfer mechanisms, internal policies and rules.

In hierarchical systems, research has further revealed the importance of Wikipedia's administrators. The administration role presents an 'elite' participant, who has a strong editorial record, more involved in communication issues and taking part in committees trusted with more power than the average participant (Kittur et al., 2007b). Based on this definition, Kittur et al. (2007b) argues that these administrative members took a "leading role" in the early stage of Wikipedia by creating enough content to attract readers and participants, building up guidelines to help new members and establishing procedures to promote and reward good contributions. The significance of the administration role led Burke and Kraut (2008b) to explore how elites pass the peer review process and gain power in an administrative role. This research constructed a model to explain how people who select themselves for election can pass the process and be approved for an administration role. Burke and Kraut (2008a) suggest this model can be applied as an automatic tool to select administrative participants according to the correlation between edit history, varied experience, participants' interaction and helping with chores. Due to the duties of such participants, many studies in this area also cross into coordinated research into Wikipedia, which try to explore ordination, negotiation and organization within the Wikipedia community.

Some scholars visualized Wikipedia by using quantitative analysis. (Voss, 2005) addressed power distribution within Wikipedia by looking at the distribution of edits per article and participant. However, Voss offers a demonstration of quantitative data to measure Wikipedia as a dynamic system. He provides many graphs to describe the growth of Wikipedia, the percentage of different contributions, average edits per distinct author, average number of edits per distinct article and so on. However, most of the analysis is based on German language Wikipedia entries instead of the larger English language Wikipedia. His work has been noted as a demonstration of measuring Wikipedia in quantitative terms by related scholars. With similar methods, Wilkinson and Huberman (2007a) argued that high-quality articles are distinguished by a considerable increase in the number of edits, the number of unique participants and the intensity of cooperation amongst participants. This possibility was tested again in another paper by the same authors (Wilkinson and Huberman, 2007b). In this new paper, they specifically compared "featured articles" (those given pride of place each day on the Wikipedia homepage, judged on quality) and common articles on the number of edits present on each. In this way, they assessed the value of cooperation in Wikipedia and claimed that it is a successful collaborative effort, because the number of discrete edits directly increased the quality of an article. Compared with their previous study, an improvement made

in this paper was to use many matrices instead of only one. Concentrating on the same trend, Almeida et al. (2007) argued that the value of each editor's contribution decreases as Wikipedia gets larger, whereas Arazy et al. (2006) indicated that increasing size and diversity of the author-base improves content quality. Although authors here offer their results from the analysis of by-product data, such research does not provide explicit illustration of Wikipedia's developing shape. The reason for this is that Wikipedia is far too complicated to be described by simply one or a single series of distributions.

Wikipedia has also been investigated specifically on how its collaborative mechanism works on finalized documents, such as articles judged to be completed. With a specific proprietary technical tool called "history flow", Viegas et al. (2004) made the collaborative editing process of individual articles visible through pictures. Through this innovation, we can easily see what content has been edited and when it was revised or deleted ordered along a time line. They introduced a set of data analysis tools to explore the design and governance of online collaborative social networks. Additionally, Viegas et al. (2007a) improved their tool to scrutinize the coordination in the "Talk Page", from which authors found that conversation could formalize Wikipedia's direction and editing policies.

Although all works surrounding Wikipedia contribute to a developing partial understanding (Viegas et al., 2007b, Wilkinson and Huberman, 2007a), we lack systematic research into such a complex online community and semantic editing system, especially as a developing example of mass collaboration. However, studies of Wikipedia demonstrate the efficiency and possibility of using Wikipedia by-product data for academic purposes, as a majority have adapted quantitative data into useful graphs and metrics (Almeida et al., 2007, Kittur et al., 2007b, Lih, 2004, Viegas et al., 2007b, Voss, 2005). A number of studies are interested in research and have implications for the design of other online collaborative and consensus communities in the Wikipedia model (Boulos et al., 2006, Prahalad and Ramaswamy, 2010).

All studies of Wikipedia have demonstrated different angles to explore interesting topics based on the analysis of digital by-product data. Understanding such studies provides two lessons for us. First, these studies pointed out the general direction for us to understand Wikipedia and all discussions surrounding it, which offer many interesting questions for our empirical work. Second, and more importantly, these studies provided the process and the means of applying digital by-product data to accomplish different research questions, which can also be used for our own research. In this chapter, we introduced why we chose Wikipedia as the case study in this thesis and what we want to explore by using its data resources. In order to explain why we choose Wikipedia in detail, we elaborated on two details. First, Wikipedia has a wealth of digital by-product data that are in a well-structured,

easily accessible format. Second, Wikipedia as a debatable and innovative phenomenon could stimulate the motivation and passion of researchers to explore it. To address the second reason, we have described the history of establishing and developing Wikipedia and discussed the research about Wikipedia's features.

Such discussions about Wikipedia from both epistemological and practical perspectives we hope not only assist readers in understanding why we choose Wikipedia as a case study, but also demonstrate that the key reason for studying Wikipedia is to explore its participation model. More importantly, from the understanding of Wikipedia's collaborative model we will be able to discuss whether Wikipedia is an innovation in the internet age or a high-risk information resource with an unreliable participating system. These questions will be addressed in the following three empirical chapters.

Finally, how to use digital by-product data to explore the participation and collaborative model in Wikipedia is the vital thing we want to learn from previous literatures. In the last section, we listed many different studies that utilize scientific methods to make use of the digital by-product data of Wikipedia for different research purposes. The methodological approach is similar to that which we proposed to use in this thesis. Based on these studies, we want to clarify how to use such a data resource to address our objective.

In the next chapter, we will construct the participation pattern of Wikipedia from a macro point of view using our proposed method. This will give a more direct impression about how people participated in maintaining Wikipedia based on the edit record from 2001 to 2007. On the other hand, this chapter will also demonstrate how we, as social scientists, can establish patterns and generate mathematical models to describe participation on Wikipedia.

Chapter 3

Assessing the development of Wikipedia

The previous chapters stated that our research stemmed from the appraisal of internet applications in social science, based on our analysis of digital by-product data. We also introduced our case study – Wikipedia, in order to explore mass collaboration, a mode of working together which has already demonstrated impressive productivity. Guided by our research aim and experimental plan, we begin our exploration from this chapter and onwards. This part of the research mainly focuses on using digital by-product data to observe and describe the development of Wikipedia, and provide further exploration and investigation of the mass collaboration behind the scenes of Wikipedia. At the same time, we examine the process of conducting our research and we hope that such an examination can give us more evidence to support our proposed methodology. Through the observation of our own research process, we further discuss the issue of whether using scientific methods to assimilate digital by-product data can address the methodological issues raised earlier in the thesis.

In order to investigate whether mass collaboration exists in Wikipedia and what form it takes, this chapter is divided into two sub-topics. The first sub-topic aims to describe the development of Wikipedia by examining several salient variables and the second intends to formulate Wikipedia's developmental trend based on the analysis results from the first sub-topic. To make the reader familiar with our analysis, we must introduce the data selected for the purpose; the software used for analysis; and the process of using such software. Based on the preliminary analysis of results, we offer the proposal that there is a pattern in the growth of Wikipedia's editing contributions. To test our hypothesis, a number of statistical analyses are carried out, and we model the developmental trend of Wikipedia through mathematical equations.

In addition to the practical contributions in describing and exploring Wikipedia, this chapter also bears methodological significance, in accomplishing the collection and analysis of data

and matching the data to a classical and well established distribution model — Pareto distribution. Furthermore, from the calculation of the parameter, we go on to construct a participation model in Wikipedia. We hope to demonstrate the proof-of-principle that as a social scientist, data mining can be used to extract useful information from digital by-product data and so test our hypotheses. Furthermore, we use the “equation” as a representational approach to data mining together with relevant graphics to explain our research findings.

3.1 Introduction

Wikipedia has been identified as a prototype of mass collaboration with impressive achievements in producing knowledge in much of the literature (Reagle, 2010, Tapscott and Williams, 2006) whereas other scholars have raised serious doubts about the quality of its product (Agarwal, 2009, Fallis, 2008, Forte and Bruckman, 2008, Magnus, 2008, Magnus, 2009). In fact, the debates around Wikipedia include the belief held by some scholars that Wikipedia has developed in a positive way (Voss, 2005), and this trend is brought by the unique participation model of Wikipedia (Reagle, 2010, Viegas et al., 2007a). Such participation is defined as “mass collaboration” (He, 2010b, Tapscott and Williams, 2006), which in essence is a democratic model involving many participants (Hippel, 2006). However, on the other hand, some scholars do not agree that Wikipedia represents a successful mode of producing and sharing knowledge and some even have serious doubts about the knowledge products created by Wikipedia; therefore, they would rather discuss the problems and shortcomings of Wikipedia (Magnus, 2009, Sanger, 2009, Tollefsen, 2009). After we have closely examined and criticised these debates, the main task of this chapter is to discuss whether Wikipedia has created a steadily developing participation model, and construct a description of this model.

For those researchers who believe that Wikipedia provides a new mode of participation, such a new model is regarded to have two important characteristics: open participation without restriction (Lih, 2009b, Reagle, 2010) and the positive correlation between the quality of products and the quantity of participants (Lih, 2004, Voss, 2005). A working mode that is created and defined by these characteristics is regarded as ‘mass collaboration’ (Tapscott and Williams, 2006). Such a collaboration represents a new productive force, which abandons the traditional hierarchical organisation and instead chooses to focus on gathering individuals’ contributions in a decentred organizational system (Hippel, 2006). The results of mass collaboration have been mainly positive and are comparable to knowledge produced by professionals from many fields (Antin and Cheshire, 2010, Beschastnikh et al., 2008, He, 2010b).

This new model of collaboration has gained wide attention within both academia and industry alike, and is thought to have revolutionary potential and innovative values in certain areas (Reagle, 2010, Tapscott and Williams, 2006). From the macro point of view, voluntary participation in Wikipedia's knowledge production will likely have an unnerving impact on financial systems and intellectual property which are traditionally based on monetary exchange and accumulation of knowledge by professionals (Benkler, 2006). Its open mode of operation has been given due recognition despite its lack of intended expertise and management (Tapscott and Williams, 2006). From the micro point of view, it defies the glorified authority in the production of contemporary knowledge by allowing amateurs to contribute to the integration of knowledge from different areas (Surowiecki, 2004), while the entire process of consuming is driven by a logic of "self-production" and "self-consumption" (Beer and Burrows, 2010, Hippel, 2006, Ritzer and Jurgenson, 2010). Undeniably, the development of such a collaborative model and its increasing popularity has encouraged wide enthusiasm in the social scientific area, with an ensuing abundance of both discussion as well as theoretical and practical exploration.

Although some proponents have explored Wikipedia from both macro and micro levels, other scholars have stated their concern and doubts about it; opposing the recognition and acceptance of the underlying collaborative mode by mainstream academics (Sanger, 2009, Tollefsen, 2009, Wray, 2009). There is an apparent logical relationship: this argued if Wikipedia does not have a steady development trend, its operational model may not be worthy of discussion. In other words, the positive development of Wikipedia both in terms of quantity and quality determines whether it has viable and unique participation model. Our task, therefore, becomes more specific: to consider whether Wikipedia has such a developing trend, and if so, whether Wikipedia has a stable model of collaboration, as suggested by some researchers (Viegas et al., 2007b, Voss, 2005).

With this realisation in mind, we are able to decide on the first question that demands our attention and discussion: whether there is a stable and positive development trend, and hidden in this trend whether there are a pair of 'invisible hands' which represents a unique participation mode that allows Wikipedia to continue to develop. First of all, we believe that in order to discover whether there has been a steady improvement to Wikipedia, we need to study the quantitative and qualitative changes of relevant variables. In other words, Wikipedia must have a regular and stable developing trend as a whole, to prove its potential and significance in representing a new mode of participation. Secondly, we need to consider that if Wikipedia indeed has such a development trend, what exactly is the so-called 'mass collaboration' model of participation?

Apparently, the research into the development of Wikipedia spans a wide spectrum of topics, such as evaluating the quality of articles (Hu et al., 2007), comparative studies between the collaborative and administrative management (Gabrilovich and Markovitch, 2007), and the assessment of data interpretation collected from hyperlinks and citations (McGuinness et al., 2006). However, along with the mathematical changes in the number of edits and participants, more and more researchers have started carrying out quantitative measurements and qualitative evaluations of the related sub-objects, such as articles, participants and edits (Lih, 2004 2007, Voss, 2005). Such studies not only describe the rise of articles and edits quantitatively, but also provide an optimistic outlook for the appearance of Wikipedia, the model of its existence and development, the means and forms of communication facilitated by Wikipedia, and emphasise the underlying value and significance of its collaborative mechanism. Despite the superficial differences between these studies and ours, the method of applying pure quantitative analysis based on the existing data resource of Wikipedia will be discussed late in this chapter.

In addition to investigating the quantitative development of Wikipedia, we also hope to use some quantitative methods to prove that the quality of the articles in Wikipedia is constantly improving. The chief aim of these attempts is to show the reliability of articles in Wikipedia through literature research and content analysis (Voss, 2005). Voss examined some variables such as the number of articles; division of language-specific sites; growth of the site; editing behaviour of authors; size of articles, and other formal elements of the site (Voss, 2005). Other experts have performed a genre analysis on Wikipedia and its web community supported by the popular wiki-technology—*Everything2*, concluding that the style of the content was shaped by the socio-technical processes of article creation (Emigh and Herring, 2005). Yet other papers have concentrated on the process of collaboration (Bryant et al., 2005); the categories of articles (Halavais and Lackan, 2008); the content of articles (Braendle, 2005); and the dynamics of Wikipedia (Okolia and Ohb, 2007).

Based on these studies, we discovered that the articles of Wikipedia are the subject most often focused on by researchers who try to measure its development. As the essential product in Wikipedia, articles have received a systematic evaluation which is based solely on the editing history database (Lih, 2004). Lih proposed that the total number of edits (rigour rate) and the total number of unique editors (diversity) could denote the “level of good standing” of articles. Under such operationalized estimation, high quality articles have been identified with the median values of these features. Additionally, he also argued that citations of Wikipedia articles by other media could contribute to the improvement of article quality by driving more public attention towards them. It can be explained that higher citation rates mean that articles attract more potential viewers or editors from other online sources. Lih was the first

researcher to create a model for the systematic measurement of the quality of articles. His research inspired us to study the relationship between the quantity of articles and the quality of Wikipedia. Starting from his example, in this chapter, we also regarded articles as an important variable to measure whether Wikipedia has indeed developed gradually.

Moreover, the 'edit' has been concerned as another significant factor to measure Wikipedia quantitatively. Voss (2005) intended to find some regular patterns on the development and edit processes of Wikipedia based on editing changes. First Voss used the calculation of the total number of words, internal links, articles, and the number of participants, and especially the number of edits to show Wikipedia's growth. More importantly, he recorded the average number of edits per minute and then compared the percentage of anonymous edits to the total number of edits in the different language versions of Wikipedia. Finally, he tried to establish the relationship between the number of ingoing and outgoing links for each article. His work covered most Wikipedia factors which are supported with digital by-product datasets, and his experience enlightens and equips us with the knowledge of which data could be used for specific research.

Another frequently examined facet of Wikipedia is the change in participants. Voss formulated the number of active participants according to their edits in one month, which can also describe the development of Wikipedia (Voss, 2005). More interestingly, the number of distinct participants per Wikipedia article can be defined as "diversity" in this collaborative process of producing knowledge (Lih, 2004). In many studies, the number of participants has been analysed as one of the important features to formulate a Wikipedia model. But different definitions of which participant is associated to the number of edits per participant can make a huge difference to the conclusion (Javanmardi and Lopes, 2010, Zachte, 2009).

In addition to using only one variable to examine the development of Wikipedia, many scholars also use different variables together to study the special features of Wikipedia. Zeng et al. (2006) evaluated the trustworthiness of Wikipedia articles by utilizing the revision history database. Based on a computing analysis, they offered a simplified hypothesis that, "trustworthiness of the revised version depends on the trustworthiness of the previous version, the author of the last revision, and the amount of text involved in the last revision". In order to test this hypothesis, they adopted the model of the Dynamic Bayesian Network (DBN). To extrapolate the equation of DBN, they identified four Wikipedia participation types: administrators; registered editors; anonymous editors; and blocked participants. Based on the analysis from this model, they concluded that the mean trustworthiness in featured articles¹⁶ is higher than that in cleaned-up articles, but is just slightly higher than that in normal articles.

¹⁶ According to Wikipedia, featured articles are considered to be the highest-quality articles Wikipedia has to offer,

In an attempt to determine the quality of edits in Wikipedia, Adler and Alfaro (2007) arbitrarily defined the survival time of edits as the criterion for the quality of edits, and relied solely on this criterion to evaluate the reputation of editors in Wikipedia. The relationship between authors and edits is established to be that authors earn their reputations from long-surviving edits and lose reputation from reverted or undone edits. Adler and Alfaro propose a chronological method, which in essence is to measure the reputation of an individual editor by examining the time span (text survival) and the number of total revisions (edit survival). Additionally, they applied the same method to the French and Italian versions of Wikipedia to validate the accuracy and reliability of this evaluation model. Through a series of experimental analyses, they claimed that, “Of the short-lived edits performed by low-reputation users, fully 66% were judged bad” (p. 268).

Hu et al. (2007) used two variables, the article quality and the author authority, and the relationship between them to construct a model. Their study provided the hypothesis that, “good contributors not only author but also review a considerable amount of good quality content”. However, their following analysis of correlation cannot mark the distinction of authority between registered users and anonymous users, which disproved the hypothesis. This case indicates that a hypothesis is not always approved by data analysis even is established according to common sense and previous conclusions (Hu et al., 2007).

While some literature focuses on measuring articles, other researchers turn their attention specifically to the Discussion Page; which is relevant to every article page and records every interaction between participants. In other words, studies of articles and edits focus on the process of producing knowledge whereas the research of discussion page concentrate on the communication and social network interaction between participants. Stvilia et al. (2005b) selected 834 article discussion pages to examine the quality of Wikipedia. They attempted to measure how the collaborative mode of Wikipedia could maintain and develop the quality of content by allowing interactions between participants as recorded in the Discussion Page. Their results revealed that, taken together, all of the participatory mechanism, edit processes and community-based recognitions are able to maintain and improve the quality of information in article pages. Another study, organized by the IBM Research Group (Viegas et al., 2007a), also focused on the editing coordination found in Wikipedia Discussion Pages. Through the investigation of 25 Discussion pages, authors found that the information mechanism, including written guidelines and reference policy enforcement, played a crucial role in fostering and reforming participants' behaviour.

As shown in the literature review, many researchers have based their studies on the analysis of the current Wikipedia database, and collected a number of variables to rebuilt their interests (Hu et al., 2007, Viegas et al., 2007a, Viegas et al., 2007b). The selection of variables differs in each study and depends on the focus of the research plan. For example, although investigating the same editing variable, some researchers chose to examine the quality of articles under the assumption of “more edits, better quality” (Lih, 2004); while others judge the editors’ reputations by the time their edits lasted (Adler and Alfaro, 2007).

It is through logical assumptions or observation of the facts that these scholars were able to depict the relationships between variables, and construct a reasonable mathematical model from these relationships. The proposed models need to be tested by applying them to a real data source, and checked to see whether they contain any regular pattern. The generation of regular patterns would illustrate the collaborative mode of Wikipedia if it is matched, and bring significant contributions to similar studies on relevant internet applications. These explorative works can discover the underlying value and the future development trends of collaborative work based on internet platforms. However, sometimes such models based on arbitrary assumptions may not be suitable for the subject of analysis in practice, because the original assumptions could be biased or erroneous. Some authors also could discover unexpected knowledge through the process of data mining and testing assumptions (Hu et al., 2007).

Although the research topics of the above studies do not appear similar and neither do they share similarities in terms of methods selection, they demonstrate an analogous procedure in conducting research. Such a procedure could be roughly divided into four steps. First of all, they selected certain basic data from Wikipedia; such as: edits (Voss, 2005); hyperlink (Fachry et al., 2007); participants (Lih, 2004); or articles (Hu et al., 2007). Following this selection, they made some assumptions based on the connection between their selected data and specific research questions. Some examples of their assumptions include that the proportion of each topic covered in various edits reflects the focus of Wikipedia (Halavais and Lackan, 2008) and the number of articles could predict Wikipedia’s growth (Voss, 2005). The third step involved discovering the intrinsic mathematical relations between the variables and constructing a mathematical model based on presumed assumptions. Finally, they fit the originally selected data back into the mathematical model for analysis; thereby accomplishing the assessment. This common procedure introduced a conceptual framework for constructing and analysing a complex and multi-variable model based on certain social phenomena.

In the process of our investigation into Wikipedia, we hope to use digital by-product data to produce some meaningful analysis and results. Based on the discussion and following

inspiration, we propose to use similar procedures to look at digital by-product data on our study. We do attempt to use an understandable representation to answer our research question and display the outcome from data analysis.

3.2 Methods and databases

In Wikipedia, the metadata of articles have been investigated thoroughly and used to measure many social science concepts like reputation (Adler and Alfaro, 2007), trustworthiness (Zeng et al., 2006) and general measurement (Voss, 2005). However, other metadata based on the participants' information has not been explored effectively. Such a situation offers us great opportunity to gain interesting knowledge based on participation data sets.

On the other hand, we need to clarify how millions of registered or anonymous participants create Wikipedia with individual edits. It is important to discover any possible correlation between participants and their associated edits. The insight of how participants edit in Wikipedia from a macro point of view could offer us a general sense of possible "mass collaboration". If mass collaboration does indeed exist, it would be intimately linked to the number of participants in the project and the output of each individual.

With such reasoning, we decided to adopt the research methods in previous literature to describe and assess Wikipedia. Our research differs from theirs in the sense that although we are also analysing data related to the article, such as edits and participants; our attention is mainly focused on participants because we are interested in the model of how people collaborate and regularly participate on Wikipedia.

The database of article namespace has been selected as the subject in our study. On Wikipedia, designers classify different namespaces (information pages by subjects) to store and exhibit varying content. Articles contributed by different participants are the primary content of Wikipedia, and are similar to entries in a traditional encyclopaedia. However, hidden behind the scenes of the finished articles are other namespaces to record additional information, including summary of edits, which memorize every development step of the knowledge production process.

The database of articles is comprised of three parts: "revision", "page" and "text", which are similar to categories in a library to store different types of books. "Revision" stores single edits as individual units and those data related to single edits, including; who edited, where they edited, when it was edited, and a summary of the edit content. "Page" stores webpages as individual cases and related information, such as the article version at a particular time point, which contains the time of creating each version, title of its webpage, basic contents, hyperlinks etc. "Text" stores all text in individual webpages, but it does not include content,

only the name to identify content for the tracing of respective texts. Among these three databases, page storage has the largest size but revision storage contains all information related to participants and their edits. Therefore, the revision data set has been selected to formulate this part of our study.

Revision Table	Rev_id	The system ID for each revision
	Rev_page	The page ID which this revision comes from, same as the one in page table
	Rev_text_id	The pointer to text content, it is possible for different revisions to share the same text as an admin could do the reverse action to restore a previous version
	Rev_comment	An editor's edit summary (the editor's comment on revision)
	Rev_user	The participant ID of who made the edit
	Rev_user_text	The participant's user name or IP address of who made the edit
	Rev_timestamp	The time point of the submission of the edit
	Page Table	Page_id
	Page_namespace	Namespace defines the category to which this page belongs
	Page_title	The title of page stored as text

Table 3-1 Different types of data collecting from the Wikipedia database¹⁷

There are four types of information involved in every participating-behaviour as shown in the figure below (Figure 3-3). In terms of the identity of the user, only registered users will be

¹⁷ Information adopted from www.mediawiki.org

recorded according to the names they chose, while the anonymous users will be identified by IP addresses. In order to clarify participation from a macroscopic point of view, we decided to use the entire revision data, excluding the summary of contents.

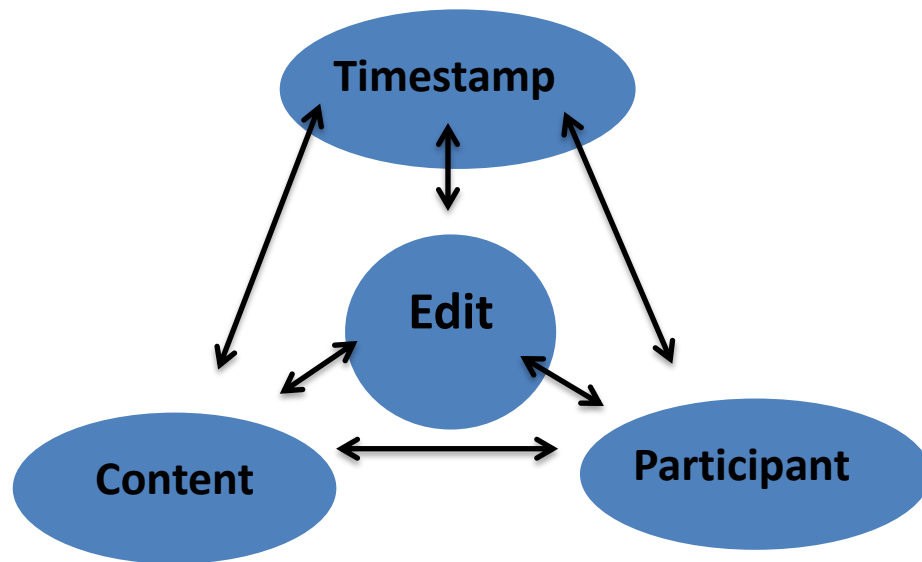


Figure 3-1 the individual components extracted from Wikipedia's database
[Originally in colour]

Regarding the requirement of a highly comprehensive database as we discussed in the beginning of this chapter, the stub-meta-current data is selected as our original database. All data is obtained and extracted from the English Wikipedia metadata, which includes all edit records of revisions from 2002 to 2007, and excludes any detailed editing content. Specific information such as editors, timestamps (the time the edit was submitted), interlink, storing namespace and summary (the editors' personal descriptive summary for the edit) was also collected.

3.3 Descriptive development

As we discovered from previous literature, digital by-product data was first utilized to provide a descriptive analysis of Wikipedia. Drawing on this work, we want to question whether Wikipedia has a stable and regular developing trend that is due to a systematic mode of participation. Therefore, we will discuss the change of Wikipedia in quantitative and qualitative terms to unravel this developing trend.

3.3.1 *The quantity of articles and participants*

The rapid increase of participants and the number of articles in Wikipedia has become a popular topic in the news press. By focusing on the amount of articles produced by millions of volunteers, the media not only advocate Wikipedia's development¹⁸¹⁹²⁰²¹, but also question the possible wane of both participation and production²²²³. Such intensive concentration on discovering the value of articles as the essence of Wikipedia also occurs in academia. Many studies attest to the increase of the number of articles created since the launch of Wikipedia (Lih, 2004, Voss, 2005, Wilkinson and Huberman, 2007a, b). With such a quantitative rise, Wikipedia has been regarded as a successful platform for gathering knowledge (Giles, 2005).

Moreover, as an increasingly productive mechanism, Wikipedia mainly relies on its volunteers to contribute continually. The remarkable difference between Wikipedia and traditional organizational means of producing knowledge is that Wikipedia has millions of participants who voluntarily collaborate, which is its primary labour force of production. This indicates that the number of participants is another fundamental variable to denote the current status of Wikipedia. Thus, when the number of participants was found to be falling, it became a considerable concern for both the Wikipedia organization and the public (Barnett, 2009).

Following such literature, we will measure in quantitative terms the development of Wikipedia in its articles and participants. Most of our results are presented in scatter charts, many of which share a number of descriptive characteristics as the X-axis is defined as the time line and is graded monthly in all graphs. To visualize the amount of articles or participants in each month, the data is displayed as spots. All graphs are produced by statistical software and generated in a coloured format, therefore we recommend viewing them in colour if possible.

Figure 3-2 reveals that the number of articles and the amount of new articles each month increased from 2001 to 2007 respectively. It clearly shows that the total number of articles has kept steadily increasing since 2001, which can illustrate the positive development of Wikipedia. Although the number of newly created articles showed little variation since 2006, the overall change of new articles by month has increased. Wikipedia can put its trust in its enthusiastic volunteers based on the following figures (Figure 3-2 and Figure 3-3), despite a stagnant period in mid-2006 of new participant recruitment.

¹⁸<http://www.dailymail.co.uk/news/article-1208941/Free-edit-Wikipedia-appoints-volunteer-editors-vet-changes-articles-living-people.html>

¹⁹ <http://www.techshout.com/internet/2006/03/wikipedia-puts-up-1-millionth-english-language-article/>

²⁰ <http://www.theawl.com/2011/05/wikipedia-and-the-death-of-the-expert>

²¹ <http://www.smh.com.au/news/web/wikipedia-cracks-twomillionth-mark/2007/09/13/1189276859147.html>

²² <http://www.newscientist.com/article/mg20327206.000-wikipedias-quality-under-threat-by-territorial-editors.html>

²³ <http://www.telegraph.co.uk/technology/wikipedia/6020775/Wikipedia-growth-slowing-as-it-reaches-3-million-articles.html>

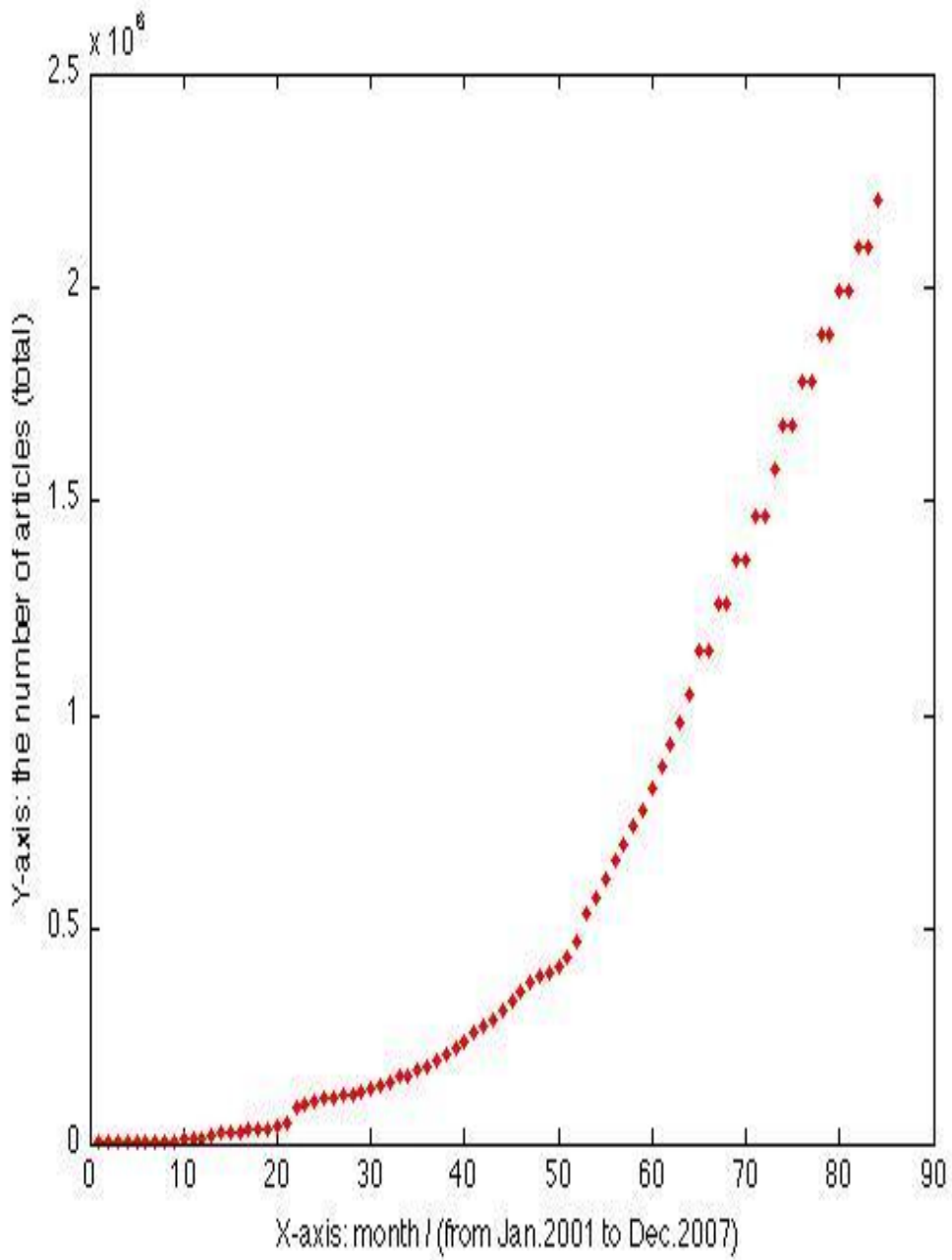


Figure 3-2 the number of total articles on Wikipedia from Jan, 2001 to Dec, 2007

[Originally in colour]

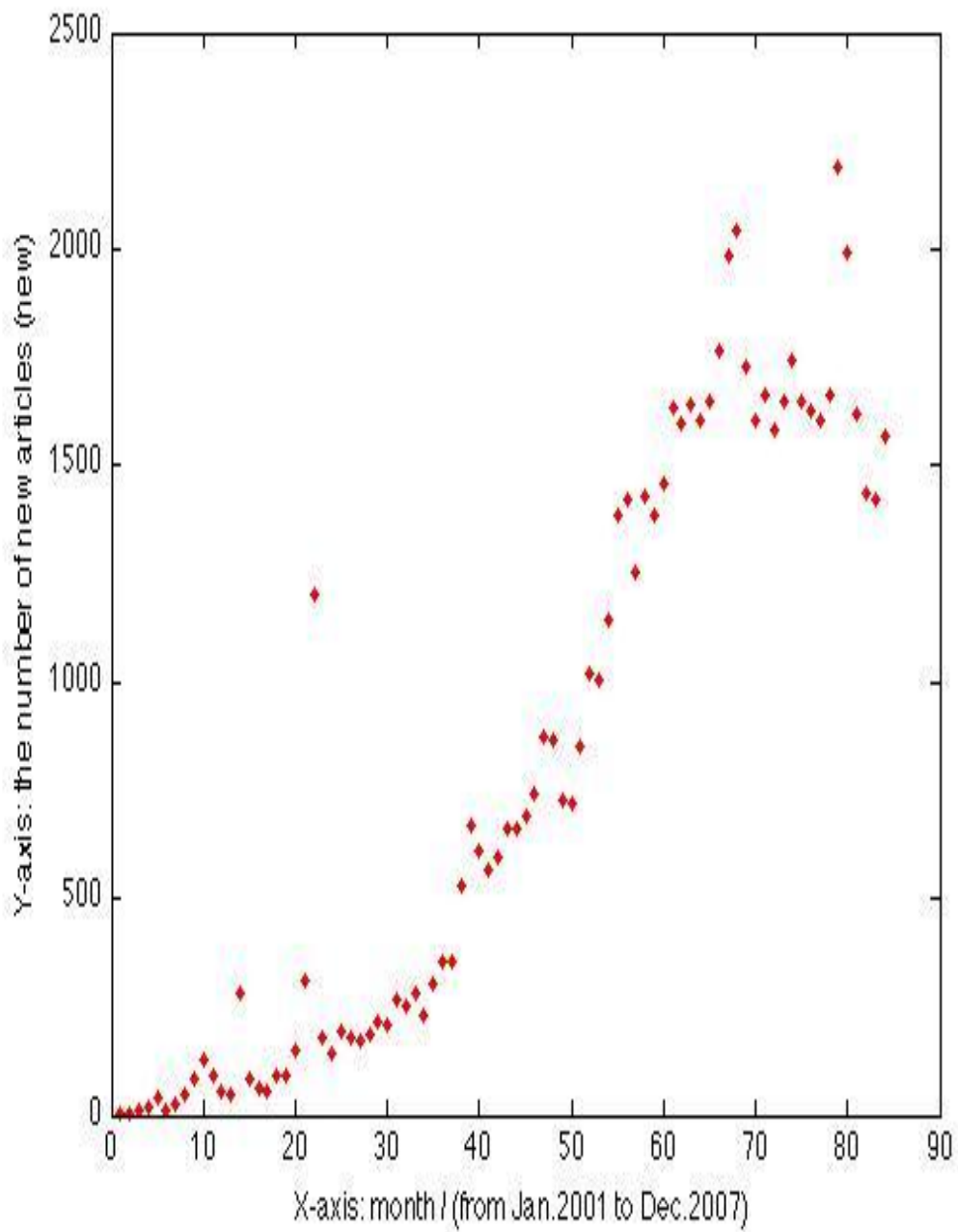


Figure 3-3 the increase of new articles each month on Wikipedia
 from Jan, 2001 to Dec, 2007
 [Originally in colour]

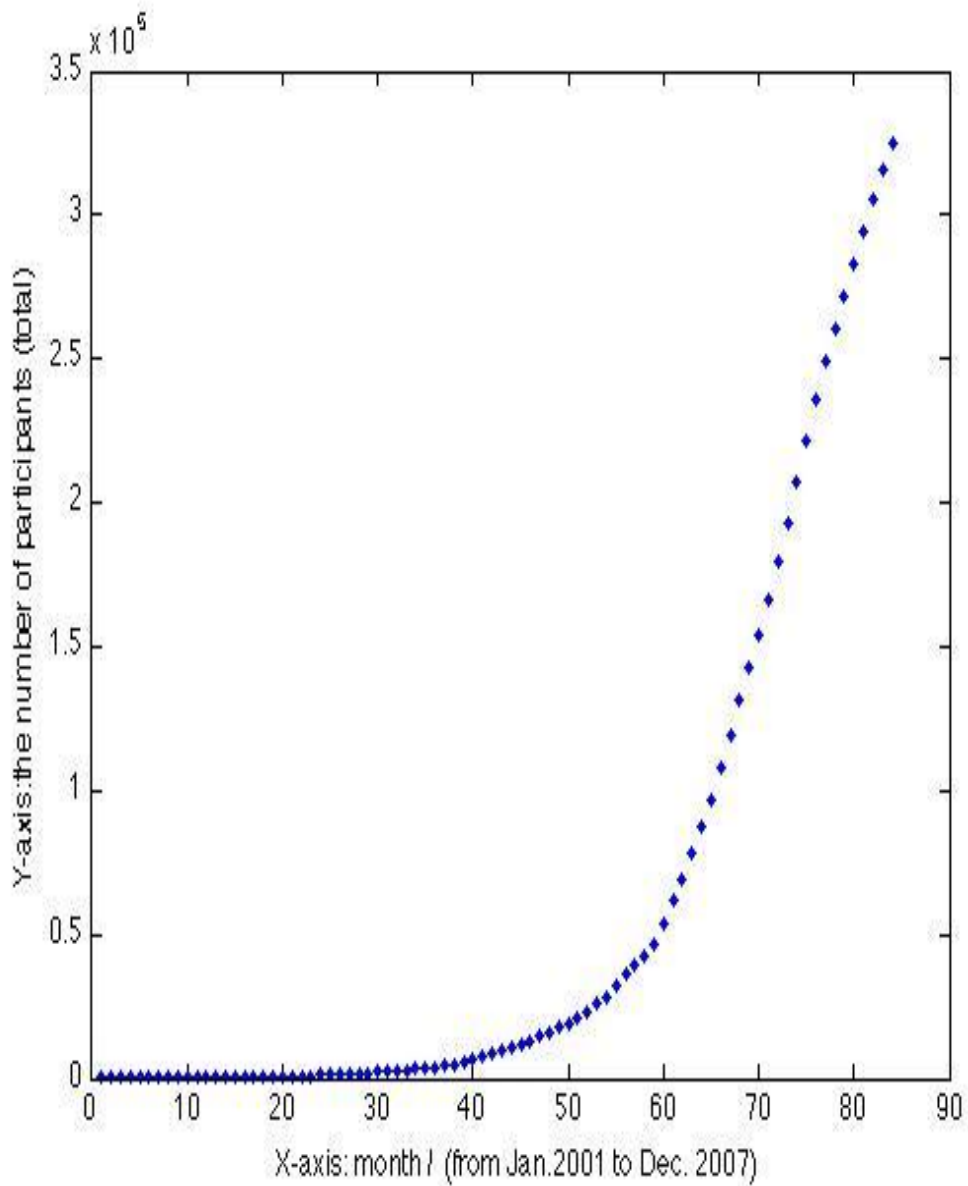


Figure 3-4 the number of participants each month on Wikipedia

from Jan, 2001 to Dec, 2007

[Originally in colour]

(This chart displays the amount of participants in Wikipedia each month, in which the “participants” have been identified as users which have edited at least ten times since their registration, excluding all anonymous editors with IP address recorded in the system.)

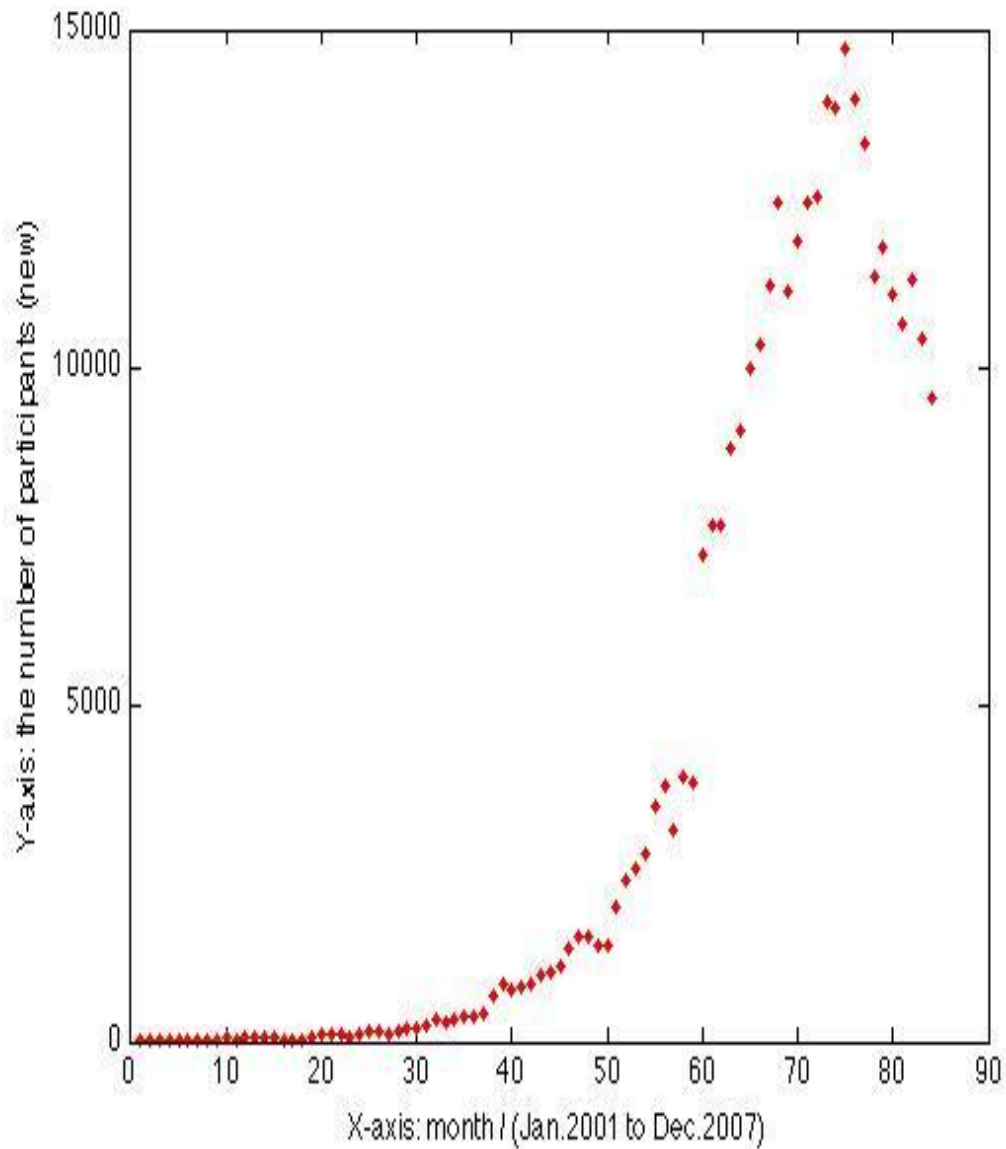


Figure 3-5 the number of new participants monthly on Wikipedia
from Jan, 2001 to Dec, 2007

[Originally in colour]

(This chart represents the number of new participants who made at least ten edits since they registered.)

3.3.2 *The quality of articles*

To measure the quality of articles in Wikipedia, some literature offers a measuring method based on the quantity of both the number of edits and the number of authors (Lih, 2004). The

former represents the amount of effort made and the latter indicates the diversity of the contribution. In our study, the basic unit for investigation is the average number of edits per article and the average number of authors per article per month. Such analysis is based on the assumption that the quality of an article is dependent on the number of edits that an individual article has received and the number of participants who were involved in editing the article.

In fact, the first part of the assumption presumes that more edits mean deeper contribution and effort was injected into the production of articles (Lih, 2004). The second part of the assumption is that the greater number of different authors not only means diversity of content (Lih, 2004), but also implies the possible number of times the articles were viewed (Hu et al., 2007). Therefore, it is assumed that more participants can bring more diversity of contribution and more views and corrections. We thus can draw lessons from previous literature about the positive correlation between the increase of edits and quality of article; and the increase of participants and quality of articles (Hu et al., 2007, Kittur et al., 2008, Lih, 2004).

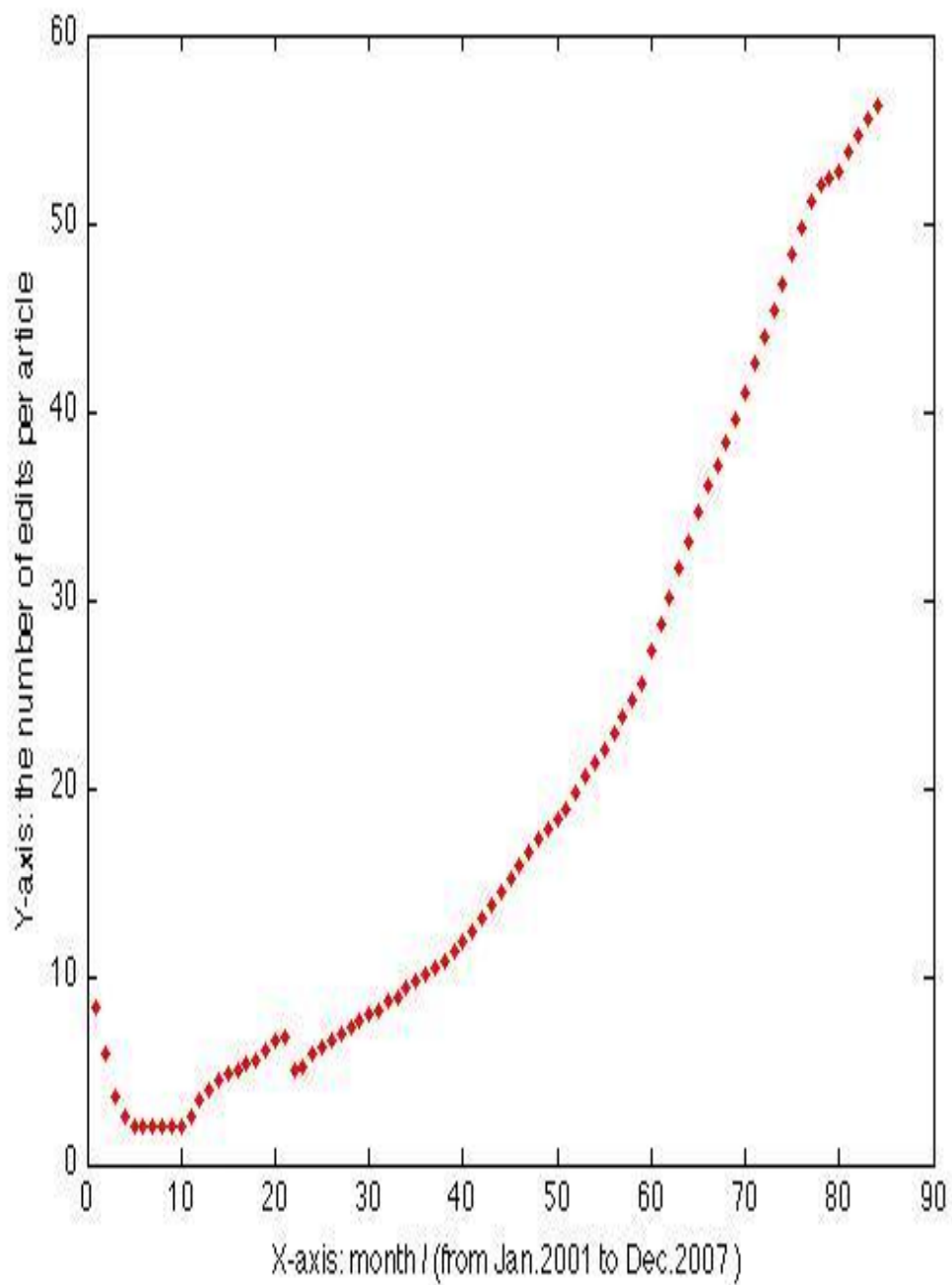


Figure 3-6 Average number of edits per article each month from Jan, 2001 to Dec, 2007

[Originally in colour]

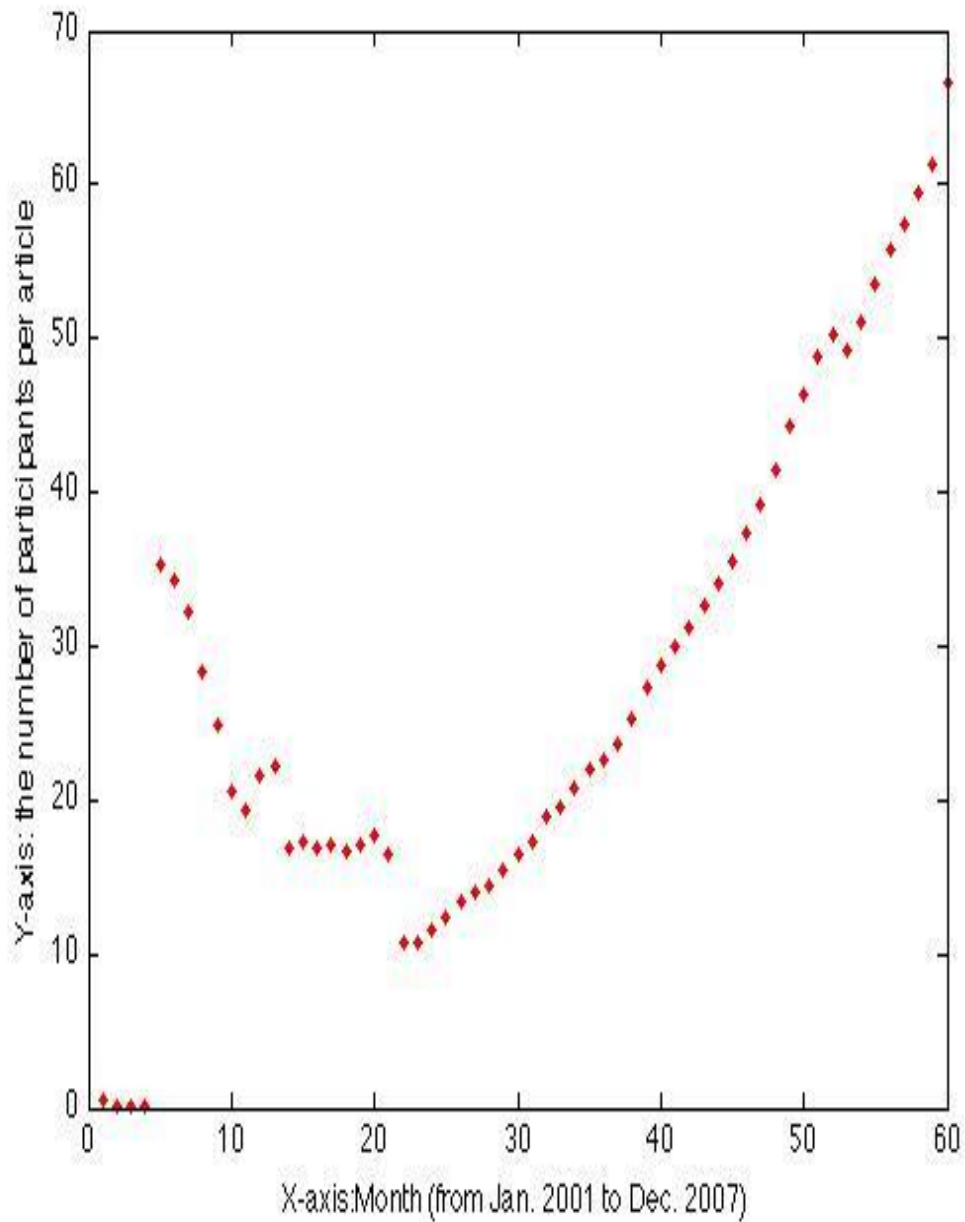


Figure 3-7 the average number of participants per article monthly
from Jan, 2001 to Dec, 2007

[Originally in colour]

This section outlines some benchmarks in depicting the quantitative changes of Wikipedia. We first assumed that the growth of Wikipedia was directly correlated to some basic variables. From a quantitative perspective, we linked the expansion of Wikipedia to the number of

engaged participants and the number of articles. The engaged participants are an indispensable component of Wikipedia and make up its mass collaboration model. The articles are the direct product of such collaboration. We contended that the changes in these two variables explicitly reflect the quantitative changes of Wikipedia as a product of knowledge assimilation, both in the process of its production and the final results. Through graphical depictions, we demonstrate that both variables are increasing, and such inclines reveal the growth and expansion of Wikipedia quantitatively. From another perspective, we also discuss the changes of Wikipedia in qualitative terms. In this process, we also made an assumption that the development of Wikipedia is expressed as the improvement of the article quality. The quality of articles is mainly affected by two variables, the number of edits in each article and the number of participants in each article. We made the assumption that the quality of an article is determined by the number of times that it is edited, and the number of editors involved in the article. The number of participants involved in the editing could influence the number of times that an article is reviewed to a certain extent, and the latter is also an important factor that has an impact on the quality of the article. The graphics generated from our digital by-product data illustrates that the average number of edits per article and the average number of participants per article are both on the rise overall, despite slight fluctuations at the early stage of Wikipedia's development. Such trends confirm the improvements in terms of the quality of Wikipedia and also reveal an optimistic model for the development of mass collaboration.

3.4 Assessing the development of Wikipedia

To summarize the points made above, we argue that the manifestation of the growth of Wikipedia is most apparent in its articles and participants. Through descriptive measurements, we qualitatively and quantitatively demonstrated that the development of articles correlates with the growth of Wikipedia. In the next section, we hope to use mathematical modelling to formulate a model of such a developmental trend.

In this step, we hope to discover how mass collaboration works in Wikipedia. If the above graphs verify the existence of mass collaboration because of the steady increases in the massive number of participants and the gross number of collaborative products, i.e. articles, the following research aims to assess the mass collaboration via analysing the relationship between edits and participants.

To provide a suitable description of mass collaboration on Wikipedia, we assume the relationship between participants and their contribution can be formulated as a stable pattern. There are two reasons for such an assumption, the visible and stable increase on the number of participant and the number of edits; and the ratio of participant to articles and the ratio of edits to articles were relatively stable (Figure 3-6 and Figure 3-7). Therefore, we start to

generate a pattern from the trend we previously discussed. In order to model how individual contributions accumulate and integrate into the final product of Wikipedia through every edits, we offer the following testable hypothesis:

There is a certain model in Wikipedia of the annual change of participation that is based on the distribution of edits by individual participants.

The correlation between participants and the content they contributed has been formulated into some specific distribution patterns in web 2.0 applications, which is normally expressed as the majority of participants being responsible for or linked to a minority of content (Brown and Adler, 2008, O'Reilly, 2005, Shirky, 2005). Following their theories of correlation between participants and edits, we constructed histograms to examine them, verify previous findings, and establish a new model.

3.4.1 Histogram

In order to do so, we first ask this question: what is the distribution of contributions made by individual participants in Wikipedia each year? It is first necessary to outline the participation situation in Wikipedia by plotting the changes in participation, which comprises edits from individual participants. Making use of the individual edit records from Wikipedia's original database, we note the number of edits made by individual participants. The list of individual participants and the number of edits they have made annually will describe the participation situation year on year.

Figure 3-8 including six histograms from 2002 to 2007 shows the frequency of editors' participation against their number of edits. It is clear that similar patterns are seen from year to year. Markedly, editors who produce a large number of edits only make up a small percentage of all participants, whereas those who only edit a few times are in the majority. It should also be noted that the total number of edits by individual participants (shown on the X axis) rapidly increases from year to year and the frequency of editing (shown on the Y axis) in the same contribution area increases from 700 in 2002 to more than 20000 in 2007. In particular, the number of participants who have produced a certain amount of edits has increased from 2002 to 2007, which means that the number of participants who edit most frequently and who edit less than ten times has risen quantitatively.

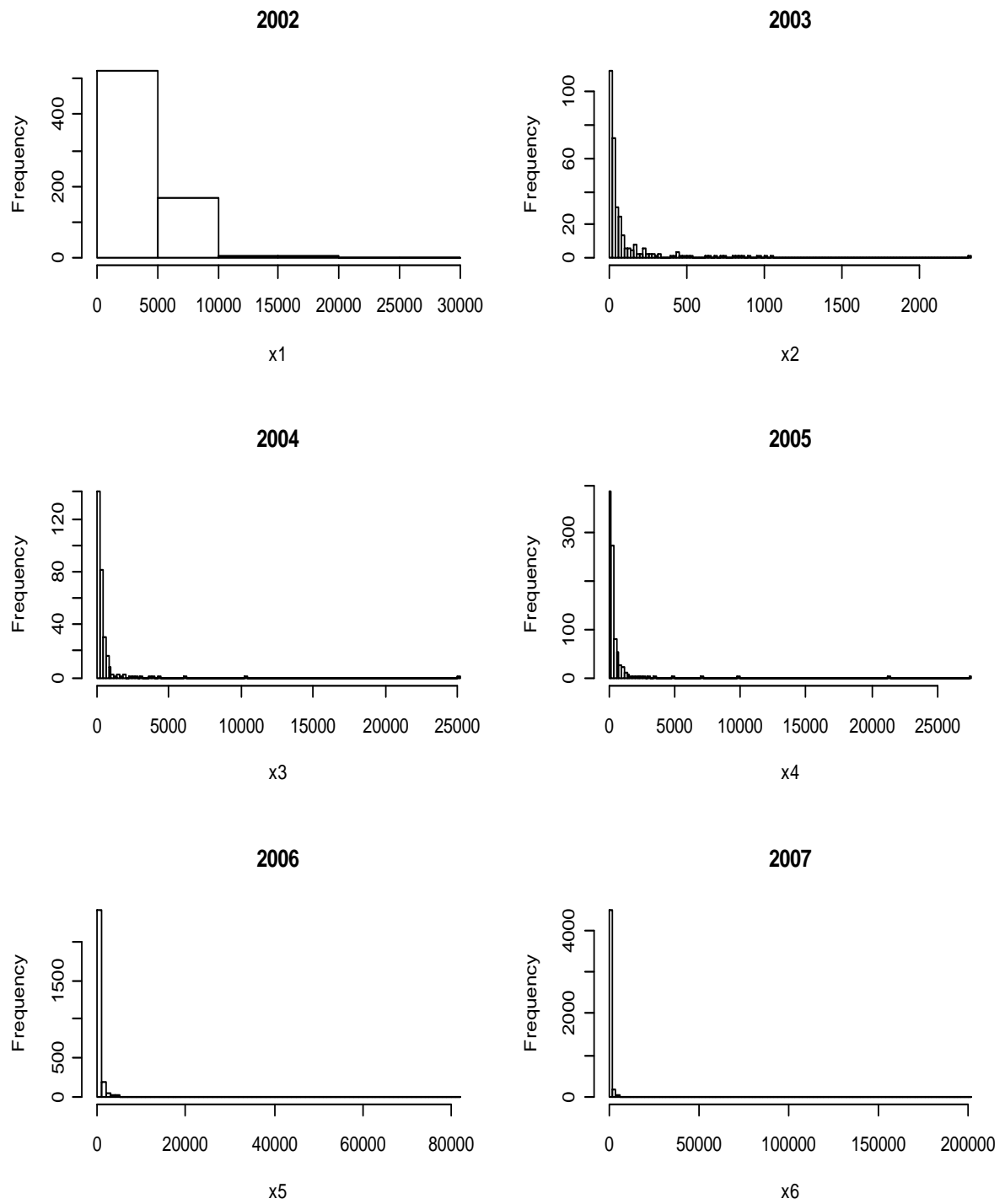


Figure 3-8 the frequency of editor's participation by year

The histograms for each year follow a similar pattern: a majority of participants only edit a few times and a few participants edit more than ten-thousand times in total. For instance, in 2002, more than 61% of the participants only edited once, and there is only one participant who edited 2287 times totally. Similarly, 66.9% of the total participants (155 thousand) only

edited once, and three people (less than 1%) have more than one million edits in 2007. Because all histograms from 2002 to 2007 show a similar trend, we assume that they have a similar distribution type, which is explained in Table 3-2:

Weighted average (edits) yearly	Percentiles							Max edits by individual	Min edits by individual
	5 %	10 %	25 %	50 %	75 %	90 %	95 %		
2002	1	1	1	1	4	15.4	55.1	2287	1 (61%)
2003	1	1	1	1	3	15	49	2340	1 (62.5%)
2004	1	1	1	1	3	15	43	25156	1 (59.8%)
2005	1	1	1	1	2	11	35	27477	1 (62.3%)
2006	1	1	1	1	2	7	22	81198	1 (66.2%)
2007	1	1	1	1	2	7	21	200112	1 (66.9%)

Table 3-2 the comparison of the number of edits by individual participants on 2002-2007

From Table 3-2, it can be seen that the change in the number of edits by individual participants in Wikipedia from 2002 to 2007 suggests a linear trend. The minimum number of edits by an individual is one, which is understandable. However, the percentage of people who only edit once increased from 2002 to 2007, the only exception being 2003. Meanwhile, the maximum number of edits rises from 2287 in 2002 to more than 20 thousand in 2007. The percentiles from 2002 to 2007 also suggest that the histogram may take up a “long tail” shape. As the average number of edits in 95% of total participants is decreased and the maximum edits made by individual is increased, we can imagine that the line between the point of 95% and the point of 100% (peak point) becomes longer and longer each year. We assume that the change in distribution of editors making amount of edits yearly can be modelled. Additionally, we could discover the development trend over the past year and this may enable us to predict the developing situation in the following years.

3.4.2 Pareto Distribution and Matching Analysis

In the previous section, we discussed that the distribution of edits by distinct individual participants shows that a majority of people only edit a few times, and a minority of participants edit the majority of content. This nature of the distribution is reminiscent of the

expression of the “size-power” relationship²⁴, which is also described as the Pareto Distribution. Since all the histograms meet the criteria of the Pareto Distribution in shape, we assume that the participation situation in Wikipedia is one type of the Pareto Distribution.

Pareto Distribution is one of the distributions encountered in economics and other realms of inquiry to exhibit the size-power relationship (Newman, 2005). It is sometimes expressed as the Pareto principle, which is the famous economic theory known as the “80-20 rule”. This suggests that 20% of the population owns 80% of the wealth (Koch, 1999). Similar situations are also observed in online communities where a minority of participants are responsible for a majority of the content, and concerns over such distribution have been voiced by numerous scholars (Fisher et al., 2006, Shirky, 2005, Whittaker et al., 1998). Many studies of Wikipedia have mentioned the “size-power” relationship, otherwise known as the power law (Capocci et al., 2008, Kittur et al., 2007a, Voss, 2005), the power-law distribution (Panciera et al., 2009), and the power-law relationship (Priedhorsky et al., 2007, Royal and Kapila, 2009). However, none of these studies made use of Wikipedia’s data to test whether such a distribution exists on Wikipedia; nor did they use an equation to model the existing distribution to accurately describe some key features of Wikipedia. Models that depict the characteristics of Wikipedia can provide a clear conceptual framework of the mass collaboration in Wikipedia and therefore the construction of models is an important goal of our research.

²⁴ “size-power” relationship focuses on the interaction between the size of participants and the contribution or influence by them. The classic “size-power” claim is that in any society, 80% of social wealth always belongs to 20% of people. The “size-power” is normally described by Pareto Distribution.

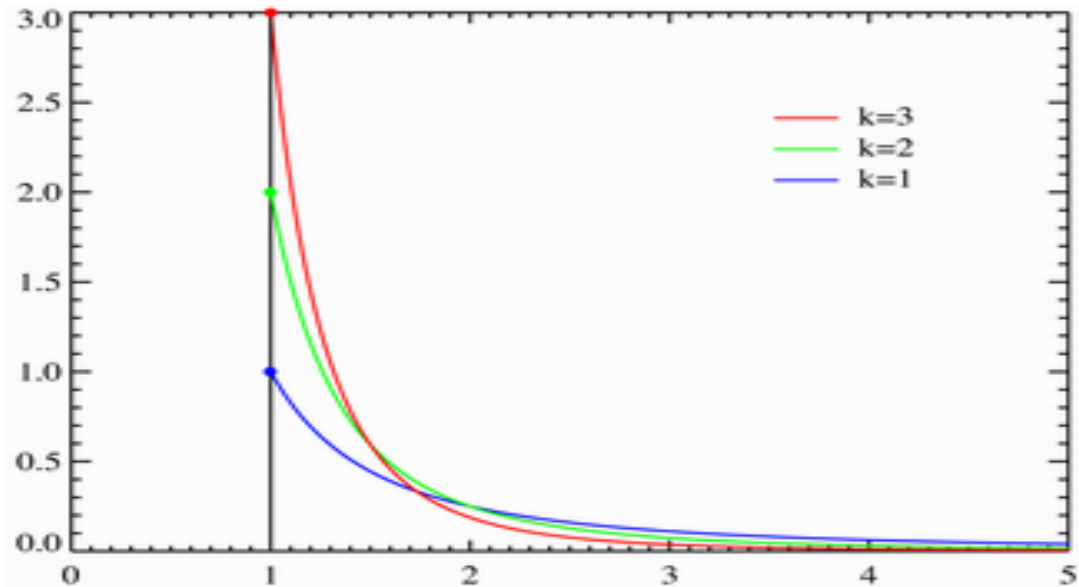


Figure 3-9 Pareto Distribution

[Originally in colour]

In

Figure 3-9, the Pareto Distribution shows different shapes according to the shape parameter k when $x_m=1$. This distribution has the probability density function that the "probability" or fraction of the population $f(x)$ that owns a small amount of wealth per person (x) is rather high, and such probabilities decrease steadily as personal wealth owned by different members of the population increases. This pattern is not limited to describing only the distribution of wealth or income, but it is also suitable for many other situations where an equilibrium is found in the distribution from the 'small' to the 'large'.

Most importantly, the family of Pareto distributions is parameterized by two quantities, x_m and k . When this distribution is used to model the distribution of wealth, the parameter k is called the Pareto index. Figure 3-9 above shows the shape of the distributions with different k parameter values. Simply, k has a main effect on the shape of the distributions by changing the scale of log-spacing of orders statistics²⁵.

²⁵ On the Pareto Distribution, if X is a random variable within a Pareto distribution, then the probability that X is greater than x is given by:

$$\Pr(X > x) = \left(\frac{x_m}{x} \right)^k$$

Following this, the probability density function can be represented as:

$$f_x(x) = k \frac{x_m^k}{x^{k+1}} \text{ for } x > x_m$$

According to the similar patterns observed between our histograms in Wikipedia and the Pareto distribution, it suggest that the statistical results of who is top of the editor list in Wikipedia based on the number of edits can be expressed more simply by the Pareto principle, or the "80-20 rule", which states that, in many cases, 80% of an effect comes from 20% of the causes. Moreover, it is obvious that the k parameter value could be changed from year to year in our histogram. Estimating the k parameter is vital to modelling the participation situation on Wikipedia. The change of k parameter from 2002 to 2007 also reveals the pattern of change in the Wikipedia size-power model.

3.4.3 *The Maximum Likelihood of estimating the k parameter*

As we discussed in the previous section, in order to model the distribution of edits by individual participants in Wikipedia, the crucial issue is to estimate k . In this part, we try to formulate the k parameter through the Maximum Likelihood approach.

According to the histogram formulated from the database, we denote that ‘ x ’ (the number of edits relevant to a particular individual) is distributed as:

$$x_i \sim P(x; k) = \frac{k \bullet x_m}{x^{k+1}}$$

From the comparison of the number of edits by individual participants from 2002 to 2007 in Figure 3-8, it is obvious that $x_m=1$. Once the mathematical model of ‘ x ’ (the number of edits relevant to a particular individual) is set, we apply the Maximum Likelihood approach to estimate the unknown parameter k . As there is only one unknown parameter ‘ k ’ in the model, estimating the value of ‘ k ’ is the crucial work here.

Maximum Likelihood Estimation (MLE) is a classic statistical method, which is used to fit a statistical model to data, and providing estimates for the model's parameters. In this study, we have a statistical distribution—Pareto Distribution and some data from Wikipedia, therefore our aim is to estimate the model’s parameter k to establish this model. The process of model formulation is provided below, and is only for the scholars who are of interest.

For all $x \geq x_m$, where x_m is the (necessarily positive) minimum possible value of X , and k is a positive parameter.

The joint probability density function (PDF) from $x_1 \dots x_n$ is as follows:

$$P(x_1 \dots x_n) = P(x_1; k) \dots P(x_n; k)$$

$$= \frac{k}{x_1^{k+1}} \cdot \frac{k}{x_2^{k+1}} \cdot \dots \cdot \frac{k}{x_n^{k+1}}$$

Thus, the likelihood is set as:

$$L(k) = P(x_1 \dots x_n)$$

Then, for computational convenience, we take the logarithm for the likelihood function and obtain the function as follows:

$$\ell(k) = \log L(k)$$

$$= \log \frac{k}{x_1^{k+1}} + \dots + \log \frac{k}{x_n^{k+1}}$$

$$= [\log k - (k+1) \log x_1] + [\log k - (k+1) \log x_2] + \dots + [\log k - (k+1) \log x_n]$$

To estimate the unknown parameter 'k', we maximise the likelihood function following the MLE;

$$\ell'(k) = \left[\frac{1}{k} - \log x_1 \right] + \left[\frac{1}{k} - \log x_2 \right] + \dots + \left[\frac{1}{k} - \log x_n \right]$$

$$= \frac{n}{k} - [\log x_1 + \log x_2 + \dots + \log x_n]$$

$$= 0;$$

We can compute the 'k' by:

$$k = \frac{n}{(\log x_1 + \log x_2 + \dots + \log x_{n-1} + \log x_n)}$$

Following the formulation process above, the k value for each year from 2002 to 2007 is calculated and plotted as shown in Figure 3-10. It can be seen that there is a linear relationship.

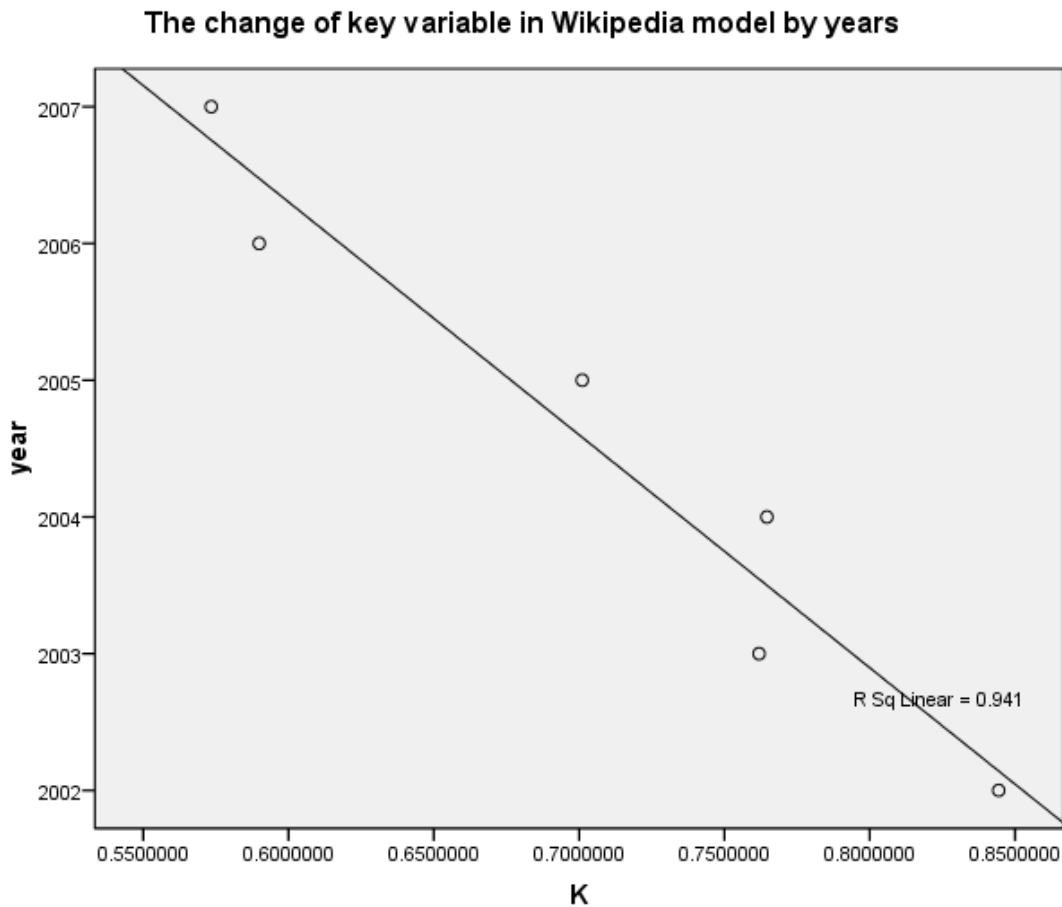


Figure 3-10 Change of key variable in Wikipedia model by year

The results can be summarized by the allocation of contributions among individual participants. It is quite difficult to formulate a sound model, because there are only six years of data. However, as Wikipedia was only launched in 2001, there is not as much data as we have hoped to analyse. The result here is the best we can achieve based on a limited database.

This graph suggests that there is a linear relationship underlying these yearly changes. Meanwhile, this trend has become more significant along with the development of Wikipedia. This idea is expressed simply with “ k ” as a parameter in distribution. According to the database of Wikipedia’s record system, we may be able to generate a mathematic model

written as a linear array of symbols from 2002 to 2007, which may enable us to represent the change of the percentage of editors who have edited at the lowest rate, from 1 to 5 times in total.

3.4.4 Summarize the distribution

Our analysis is designed to accommodate the original histograms which have a shape of the Pareto Distribution to estimate the tail exponent by fitting the data from Wikipedia's record of edits. The method of Maximum Likelihood is able to estimate an extreme-value approximation in statistics. Our result can model the 'long tail' participation pattern on Wikipedia, which represents the size-power relationship in the mass collaboration process.

By doing this, we discovered the relationship between the population of individual participants and the number of their edits. Such a relationship establishes the size-power model to describe Wikipedia. Meanwhile, the Pareto Distribution suggests the possibility that there is a certain level of invariance in the participation of mass collaboration. In this model, the distribution of editors suitably coincides with the true distribution of the participation of Wikipedia if and only if the distribution is of the Pareto form and the variable "k" has a linear regression in the equation from 2002 to 2007.

Through modelling the participation pattern of individuals, the result shows that there is a *regular rule in the change of participation pattern each year based on the distribution of individual edits*, which supports our original hypothesis.

On the one hand, our research into the participation in Wikipedia produces two findings. First, this participation has a similar shape to that of the Pareto Distribution, with the majority of participants only contributing a few edits and a few participants contributing a great number of edits. Second, the maximum number of edits from individual participants has grown yearly and both the number of actively-editing participants and the number of least-actively—editing participants increased from year to year. Namely, the quantity of participants and the number of edits from individual participants show an increasing trend from 2002 to 2007.

Furthermore, the developing pattern in Wikipedia has been fitted to a linear relationship, and a mathematical equation has also been established to model this pattern specifically. In this modelling process, we formulated the variables in the editing process and presented Maximum Likelihood (ML) estimates for the parameter used in modelling the "long tail" through the understandable description. We also used the estimates to calculate optimal chip scaling factors for a data set consisting of thousands of variables.

This analysis demonstrates the possibility of establishing models to automatic detect and predict participation on Wikipedia based only on digital by-product data. Specifically, social

scientists are able to observe and formulate complex phenomena such as the edits/participants distribution to predict what trend Wikipedia will follow in the future using quantitative models. The approach of building such a model can also give us a better understanding of how metrics correlate with the investigating matters, which could lead to the creation and improvement of new tools and new policies. These modelling techniques may have wider applications for complicated online topics to simplify their trends and changes on a macro level.

3.5 Conclusion

This chapter aims to explore the development of Wikipedia through both descriptive graphics and mathematical modelling. The revision sets from 2001 to 2007 in the sub-meta-current data dumps are selected as our quantitative database and the entire data analysis process is comprised of three steps: extracting data from original resources; selecting relevant data sets; and building new databases. First, our study of the development of Wikipedia is supported by quantitative and qualitative analyses, which were guided by theories from previous researches (Hu et al., 2007, Lih, 2004, Voss, 2005). Generally, articles and participants have been considered as the two primary features to measure the development of Wikipedia. Based on the quantitative changes of articles and participants from 2001 to 2007, we conclude that the rapid growth in numbers of both articles and participants signified that Wikipedia maintains a stable developing trend.

Moreover, the quality of the products on Wikipedia is indicated by the number of edits per article and the number of participants per article. We assumed that the number of edits per article can represent how much effort has been put into improving the quality of the article. On the other hand, the number of participants per article illustrates the involved human labour and possible correction following edits. By showing a rise in the average number of edits per article and average number of participants per article, we demonstrated the exponential growth of Wikipedia in qualitative terms.

Then this chapter tested the hypothesis of how Wikipedia developed by mass collaboration through the discussion of the frequency of participation and edits. The distribution of the individual participation each year suggests that we could use a mathematic equation to explore the change of participation. With available models such as the Pareto Distribution, we found that the participation model in Wikipedia followed the Pareto Distribution. However, further examination should take place to test this model on the history-version data in the future.

Through the empirical studies in this chapter, it is possible to summarize three features of mass collaboration on Wikipedia. First, the total number of articles and the total number of participants has increased from 2001 to 2007, which suggests that Wikipedia has developed quantitatively shown by the number of involved people and their products. Second, the number of participants per article and the amount of edits per article on average not only demonstrates the amount of effort involved in the production of the article, but also implicate the number of possible views and corrections. Third and most importantly, statistical analysis suggests that the participation situation from 2002 to 2007 has a linear relationship, and can be represented by a mathematical equation created originally for this chapter. Based on this, we assume that the participation situation of each year in the English Wikipedia followed this equation. The prediction of following years thus can be formulated. Additionally, in the analysis of the model, we also found that the participation situation in each year may be explained as a majority of participants having contributed a few edits whilst a few participants have edited the majority of articles.

Although we have come to the mathematical model of describing the participation pattern in Wikipedia, it is necessary to emphasise the limitation of this analysis. In fact, this study only uses the data from the current-meta database as an original resource, excluding deleted or changed edits data, which is invisible in the current Wikipedia. This selection of data avoids the risk of an overwhelming database but on the other hand misses some information from the history record. In addition, we only have six years of data provided by Wikipedia to establish the participation pattern. Therefore, the result of this study only offers tentative explanations to the process of how Wikipedia established and developed through mass collaboration. The equation we build here should be tested further if we have access to more data in the future.

The general analyses and descriptions of the several basic variables of Wikipedia and the connections between them only represent our first attempt to understand Wikipedia. In the following chapter, we will specifically investigate the influence of individual participants in the process of mass collaboration on Wikipedia, based on the number of edits contributed by each participant. Through grouping the participants together according to the number of their edits, we also investigate who dominates Wikipedia's development within a certain period, and what proportion of participants influence the development of Wikipedia the most. From such observations, we could further mature the model of mass collaboration, and clarify whether such a collaborative model of Wikipedia exists and how it can be improved.

Chapter 4

The shift of participation -Who is the most important participant

In the previous chapter, we investigated the developmental trends of Wikipedia, which indicate that participation in Wikipedia is systematic and predictable. To clarify its development, we also introduced an equation to demonstrate the distribution of individual participants. The participation pattern in Wikipedia was illustrated by a Pareto distribution, which provided a general overview of the frequency of individual contributions into Wikipedia under the mass collaboration model. However, this result only provided a macro view of mass collaboration by offering a description of its regularity. To understand more about mass collaboration, for instance, the different types of participants, the editing behaviours and so on, we have to use digital by-product data to formulate more detail of the collaborative process.

Wikipedia is a collaborative platform to which millions of participants can contribute freely. Based on the policy announced on Wikipedia's community page, every participant shares an equal right to direct the encyclopaedia and edit it²⁶. Specifically, the policy claims that, "If a rule prevents you from improving or maintaining Wikipedia, ignore it"²⁷ and all the rules have to be changed by consensus. It could not be denied that the openness of Wikipedia's edit policy has allowed the number of articles in Wikipedia to grow exponentially since its establishment. However, there are a number of questions remaining, including: how open is

²⁶ http://en.wikipedia.org/wiki/Wikipedia:Five_pillars

²⁷ http://en.wikipedia.org/wiki/Wikipedia:What_%22Ignore_all_rules%22_means

Wikipedia?; Is Wikipedia a democratic system where all participants have equal rights to make a decision?; If so, do different types of participants share equal rights?; And if not, which is the most important group of participants? These series of questions outlined above illustrate our research interests. The chief aim of this chapter is to answer these questions as thoroughly as we can.

This chapter explores mass collaboration by dividing participants into different groups according to the number of their edits. The working question addressed in this chapter is who the primary participant really is among the more than five million registered users. By identifying the primary participants on Wikipedia, we are able to answer whether Wikipedia relies on mass collaboration for its function or whether Wikipedia is just another product co-authored by a small group of participants. By identifying the most important participants in Wikipedia, we may also be able to illustrate the decision-making process that occur on Wikipedia, which will help us to understand how millions of people can work together to generate value.

From a methodological perspective, the previous chapter used digital by-product data to establish the equation for modelling the development trend of Wikipedia, which is a classic method used by scientists and technicians in similar studies. However, as we delve deeper into particular aspects of mass collaboration, exact replication of scientific methods becomes insufficient to address our questions. We will use digital by-product data in order to accomplish our goal in this chapter. We have opted to follow the methodological approach of previous research on Wikipedia. The empirical work of this chapter mainly aims to explore the participation mechanism of mass collaboration and to verify the findings of previous studies. In doing so, we will use graphical representations to discuss our research questions, by which we continue to examine the reality of applying digital by-product data in social research.

First, this chapter introduces what we aim to discover and what possibility there is for using such data. Second, the related literature is discussed, along with our empirical work, and from the comparisons we propose a series of hypotheses to predict the shift of participation. The findings suggest that during a specific period there has been a slight shift among the participants who dominate the content of Wikipedia, which we plan to test again when more data resources become available. Third, we introduce the database used in this chapter and related numerical details. More importantly, we discuss how we divided participants in different groups by accounting for their edits and administration status. By providing the hypothesis that such a shift in participation exists, the fourth part of this chapter examines two related hypotheses of the shift specifically in the general edit-divided groups and admin-

divided groups. Finally, we summarise the shift hypothesis and discuss who dominates Wikipedia based on the number of edits. Additionally, we discuss the advantages and disadvantages of applying graphic expression as a primary methodology based on digital data.

4.1 Introduction

In studies of Wikipedia, there has been a mixture of approval and disbelief at the same time. First we discuss the doubts about the quality of articles produced by the large number of participants on Wikipedia. Such doubts mainly originate from the analysis of some statistical data, followed by pessimistic predictions of the quality of Wikipedia and its future. Secondly, this section also discusses the academic discourse and data analyses which aim to investigate whether the decrease of participants in Wikipedia will indeed have a negative impact on the quality of its articles. Through the discussion of previous literature, we will not only explore who are the most important participants in mass collaboration, but also will learn how to evaluate such issues by using digital by-product data.

4.1.1 Debates about the participants of Wikipedia

In early 2009, a heated debate broke out within mass media, blogs, the Wikipedia organization itself, and related statisticians about the departure of a large number of editors, and whether such a widespread dissipation would have an impact on the sites quality of content. Certain media casted considerably bad light during such debate: “Volunteers log off as Wikipedia ages” (Angwin and Fowler, 2009); “Report claims Wikipedia losing editors in droves” (Edwards, 2009); “Report: Wikipedia losing volunteers” (Whitney, 2009); “Wikipedia Editors are leaving Wikipedia” (Agarwal, 2009), and “Wikipedia goes down” (N/A, 2010).

These news or comments were based on the statistical report from a Spanish doctoral thesis by Ortega (2009). This study used Wikipedia’s online database and analysed the complete history of edits made to Wikipedia by registered users in the top ten Wikipedia language versions including English Wikipedia. This study claimed that the English Wikipedia had lost its editors rapidly, suffering a net loss of 4900 in the first three months of 2008 and over ten times that number in the same period in 2009 (Angwin and Fowler, 2009, Ortega, 2009). Such continuing decline has been considered a signal of a crisis in Wikipedia’s development (N/A, 2010), while some even asserted Wikipedia was in decay because its mechanisms were getting older (Angwin and Fowler, 2009).

Such decline has been explained and untangled by different scholars and specialists. Ortega suggested this downward spiral of editors leaving Wikipedia would continue and eventually harm Wikipedia as a mass collaborative outlet (Barnett, 2009). This quantitative decline has

been verified by many other studies. However Wikipedia has responded that this claim of decline only rests on Ortega's definition of a participant on Wikipedia. Wikipedia defines participants as people that must have edited at least five times in total, but Ortega's "participant" considered those that have edited only once. However, even if the definitions differ greatly, they still cannot deny the fact that the number of at least one-time participants in Wikipedia has decreased since 2007 (Zachte, 2009).

In addition, Wikipedia's founder Jimmy Wales announced Ortega's database was inaccurate, as the number of editors in English Wikipedia has stabilized when interviewed by Lomas (Lomas, 2009). He further pointed out that the majority of articles and edits on Wikipedia were contributed by a small group of people. Jimmy Wales said during this interview that, "There is a vast majority of Wikipedia where the entry was started by one person, really heavily edited by one more, and two or three more have added some comments or critiques and changed some spelling or something, so that it does tend to be small group collaboration" (Lomas, 2009). Thus he summarized that Wikipedia, in fact, was edited by only about a hundred people instead of millions. This statement might imply that even though the number of one-time editors in Wikipedia has slightly dropped; it would not affect the development of Wikipedia as a whole.

Graphs in chapter three indicate that there was indeed a "slight decline" in the number of new articles and the number of new participants, although it has been explained as numbers that "declined slightly and have now stabilized" (Edwards, 2009). Therefore, the debates triggered by the media raise two further important questions. The first is whether the number of participants of Wikipedia is indeed in decline, and if such decline is confirmed then to what extent has the effect been in quantitative terms since Wikipedia was established in 2001. This leads us to ponder the second question, that is – who is, after all, the real contributor to Wikipedia? In other words, who are the important participants in Wikipedia according to the number of their edits?

4.1.2 Literature review

In the past decade, Wikipedia has experienced a very fast growth rate, which has been discussed as a result of "mass collaboration" (Tapscott and Williams, 2006) by millions of volunteers. Researchers have not only recognized the overall importance of the working model of mass collaboration, but have also investigated mass collaboration from different angles (Benkler, 2006, Surowiecki, 2004, Tapscott and Williams, 2006). Specifically, comparison studies between Wikipedia's collaborative model and that of existing open source software have been carried out. Such investigations have confirmed from all perspectives including the quantity, diversity and quality that Wikipedia has created a new mass

collaboration mode, where the number of contributions correlates with the quality of the collaborative products (Anthony et al., 2005, Arazy et al., 2006).

In fact, the definition of mass collaboration as a mode of working together on a single project, has even started in open source software research, in which studies suggested that most of the products were created by a small number of experts who are also active in the communication process (Lakhani and Hippel, 2003, Mockus et al., 2002). The point of “mass collaboration” (Tapscott and Williams, 2006) or “wisdom of crowd” (Surowiecki, 2004) has been considered controversial by experienced Wikipedia participants. We found that the quality of mass collaboration was defined by the number of participants in some studies (Kittur et al., 2008, Kittur and Kraut, 2010, Kittur et al., 2009). Such correlation has been further noted as the primary characteristic of Wikipedia – mass collaboration, and also referred to as “the wisdom of mobs”, and “swarm intelligence”, which describes the process whereby millions of individual users each make contributions and out of this emerges a coherent body of work (Swartz, 2004). The central debate about whether Wikipedia is created by a small group or millions of participants.

This freedom and collaborative mode of Wikipedia has generated intense scholarly interest, because it not only demonstrates that Wikipedia has created a new means of working together (Surowiecki, 2004), but more significantly, it indicates that this model may diminish the necessity of authority and expertise (Hippel, 2006). If all of the participants are working at the same level on the hierarchy, and the number of participants directly defines the outcome of the project, this would overthrow the traditional organizational structure, which relies on authority and specialists for quality control. If such hypothesis is true, mass collaboration may produce a new model of working organization where the quantity of participants could directly affect the quality of products.

Some studies have addressed their assessment of whether the quality of products from mass collaboration depends on the number of participants (O'Reilly, 2005). Initial studies indicated that certain features of mass collaboration, such as size, diversity and aggregation trend, may be associated with, or even affected by the number of participants and edits. This theory has been supported by studies using metrics. Others have argued that the number of participants and edits in an article is directly associated with the quality of the article (Lih, 2004, Stvilia et al., 2008). These researchers simply provided a correlation between the number of participants and the quality of the product, and therefore deduced that Wikipedia is dependent on a mass collaboration which sustains itself by all participants (Olleros, 2008).

However these discussions cannot lead to the conclusion that the number of participants may directly reflect the number of contributions. The former is the determinant of the quantity of

productive activities as discussed by previous literature, whereas the latter is the number of people involved as the component of collaboration. In other words, mass collaboration could be executed in two modes: the first is that only a limited number of people participate, yet the output from such a small group is still significant; alternatively, there could be a large participating group, but each individual only contributes a little, yet the collective contribution would also be significant. In this light, previous research does not define which model of function could represent Wikipedia's operation.

The issue of whether Wikipedia has been mostly based on contributions by a small group of people or by a massive crowd also affects some fundamental policies and strategies within Wikipedia. For instance, Jimmy Wales, the founder of Wikipedia, claimed on his blog that the majority of the total contributions to Wikipedia came from a small group of participants, citing the statistics from December 2004, in which 2.5% of the registered participants on the site made half of the edits (Wales, 2005b). Based on this result, Wikipedia has developed tools and features to meet the demand of this small group of people (Reagle, 2010).

However, Wikipedia also claimed that all participants are equal on the platform (Ayer, 2008, p.57). According to Wikipedia's policy, the openness and freedom of participation is the primary factor which facilitated its speedy development (Almeida et al., 2007, Arazy et al., 2006, Reagle, 2010). Its strategy of offering everyone including anonymous participants the equal position to edit has been questioned by many studies (Denning et al., 2005, Sanger, 2009) but has also been confirmed as its most useful and valuable mechanism by others (Anthony, 2005; Surowiecky, 2005). In fact, these studies on the one hand advocate the wisdom of an equal editing policy, on the other hand, they imply every participant in this mechanism should have equal right to affect the final product, which is the basic principle of mass collaboration. Furthermore, such domination could establish mass collaboration on Wikipedia, which could be replicated in other mass collaborative web 2.0 applications.

However, this conception has been disputed by empirical studies of Wikipedia with statistical description of the power law, which means a minority of participants creates a majority of content. According to the description in chapter three, the power-law distribution, also named the Pareto distribution, has been addressed by some related statistical studies. Voss (2004) presented many growth figures of Wikipedia features such as articles, words, links, bytes and users, and suggested that the number of unique authors per articles followed a Pareto distribution. The Pareto distribution has been cited again in other studies (Kittur et al., 2007a, Olleros, 2008, Swartz, 2004). The proposal that a small group leads Wikipedia has been supported by the "Inequality of Contribution" study of (Ortega et al., 2008) based on statistical analysis. They claimed that the general ratio between the number of participants

and their contributions is less than 10% of the total number of participants was responsible for more than 90% of the total amount of contributions. This result came from the analysis of the top ten language versions of Wikipedia including the English one. This study includes the data collected on 30th November 2006. Based on this result, Ortega and his colleagues suggested that it is possible to identify a “core” group of participants that require the majority of attention from software or tools designed for Wikipedia (Ortega et al., 2008). Similarly, through exploring who dominates Wikipedia’s edits, it is possible to determine which style of mass collaboration Wikipedia assumes.

The statement that “mass collaboration” is dominated by a small group of participants has been argued by (Kittur et al., 2007a). Firstly, their study proved that a Pareto distribution existed in the participation pattern of Wikipedia. However, the primary argument from their study was that, “A closer look revealed a major shift in the distribution of work”. This study claimed that the small group of participants who contributed the majority of the edits had been replaced by an increasing number of participants who each made only a small number of edits. Furthermore, they also claimed the shift could be explained as the outcome of a marked growth of low-edit participants rather than as an effect of high-edit participants leaving or reducing their activities.

However, there are many methodological issues, which might affect the accuracy and consideration of Kittur et al.’s study. The test database was the history dumps generated on 2nd July 2006, which could be considered small, especially given Wikipedia’s growth, although it was new at that time. In his work, Kittur claimed that there was a shift in 2004 as to who contributed the majority of content in Wikipedia from the “elites” to the “crowd” contributors. But it is debatable to reach a conclusion about a shift having occurred in 2004 on the basis of an analysis of only six years of data culminating in the middle of 2006. To claim, further, that if the trend existed in 2004 then the shifting trend of crowd’s contribution would be maintained in the future is also problematical.

In conclusion, the participation pattern of Wikipedia has been described in two different ways. Some theoretical studies emphasize its open and free mechanism in Wikipedia and thus suggest that the mass collaboration pattern in Wikipedia has developed because every participant has equal influence to the content. However, other empirical works based on statistical results argue that mass collaboration of Wikipedia follows the power law, in which a majority of contents are created by a small portion of participants. From our own analysis in chapter three, the power-law has been supported by the generated Pareto distribution. However, the third statement is addressed by Kittur (Kittur et al., 2007a), which suggests that although the most of Wikipedia’s content was previously contributed by a small group, a

change occurred alongside the development of Wikipedia. He indicated that there was a shift from the obvious power-law (a minority of participants responsible for majority of content) to the wisdom of the crowd, in which the majority of participants increased their proportion of edits compared to previous periods. But his argument was based on only six years of data, which lacks enough information to fully convince.

From the introduction of previous literature, we have attempted to clarify that this chapter focuses on examining whether the distribution of edits per participants changed alongside the development of Wikipedia. Is it possible that Wikipedia has already become a more decentralized system where a majority of content is contributed by a majority of participants instead of the established power law? If such prediction could be validated, we could further prove that the development of Wikipedia is based on both the manifestations of mass collaboration, which would enable us to extend our knowledge on mass collaboration, and even the entire concept of web 2.0 applications.

4.2 Method

The data used in this chapter were generated from the free data dump “stub-meta-history” on 11th Oct 2010. The data is 13.8GB with 13% compression. This data set was imported into the Oracle 11 database management system to process and analyse. The data set consists of information about every edit, including the edit time, the user ID of the editor and the summary of content. In other words, we are able to inspect people’s edit behaviours over time.

The original data set includes 359,407,803 cases in total, among which 272,286,668 edits were created by 3,884,256 distinct registered participants and 87,121,135 edits were made anonymously. The first edit was made at 20:08 16th Jan 2001 and the last edit at 19:17, 11th Oct 2010.

For the definition of administration in Wikipedia, we use the user status data dump downloaded from Wikipedia. From this data set, we collected lists of 1769 administrators and 660 bots for analysis. The administration list shows the editing behaviours of administrators. A bot is an automatic technical tool authorized by administrators and some senior participants for maintenance of Wikipedia by reversing deletion and preventing vandalism.

4.3 Analysis

In order to evaluate who dominates Wikipedia in terms of participation, we will generate some graphical descriptions. The first section will introduce the growth and fluctuation of Wikipedia to illustrate that the change of contributions in the development of Wikipedia is dynamic and rapid. In the second section, we hypothesize who may be the important

participants. Following our hypothesis, the third section illustrates the changes in the number of edits made by different participant groups categorized by their number of edits. These results provide clues about who might dominate Wikipedia, and also address the question of whether Wikipedia is a process of mass collaboration or just a specific collaborative product made by a small group of elites. Finally, the fourth section discusses whether the administration group has specific influence on editing Wikipedia.

4.3.1 Growth and fluctuation of Wikipedia

The number of articles on Wikipedia experienced a considerable growth, which is considered a great achievement for collaborative participation. However, in addition to the increasing number of articles, there are factors in Wikipedia that have changed in the last decade. These changes represent the trend of Wikipedia's development and have important implications for the future direction of the site. Wikipedia has seen overall increases in its edits since it was launched in 2001.

Figure 4-1 illustrates that the number of edits per month has been maintained at a steady level since it was first launched. From 2004 to 2007, this number experienced a sharp increase, followed by fluctuations at 5,000,000 edits per month. Based on this trend in the change of number of edits in Wikipedia, we can predict two possible outcomes of development. One is that the number of edits will follow its current fluctuation; another is that the number of edits may undergo an intermittent increase as was the case between 2004 and 2007.

Growth of edits per month in Wikipedia

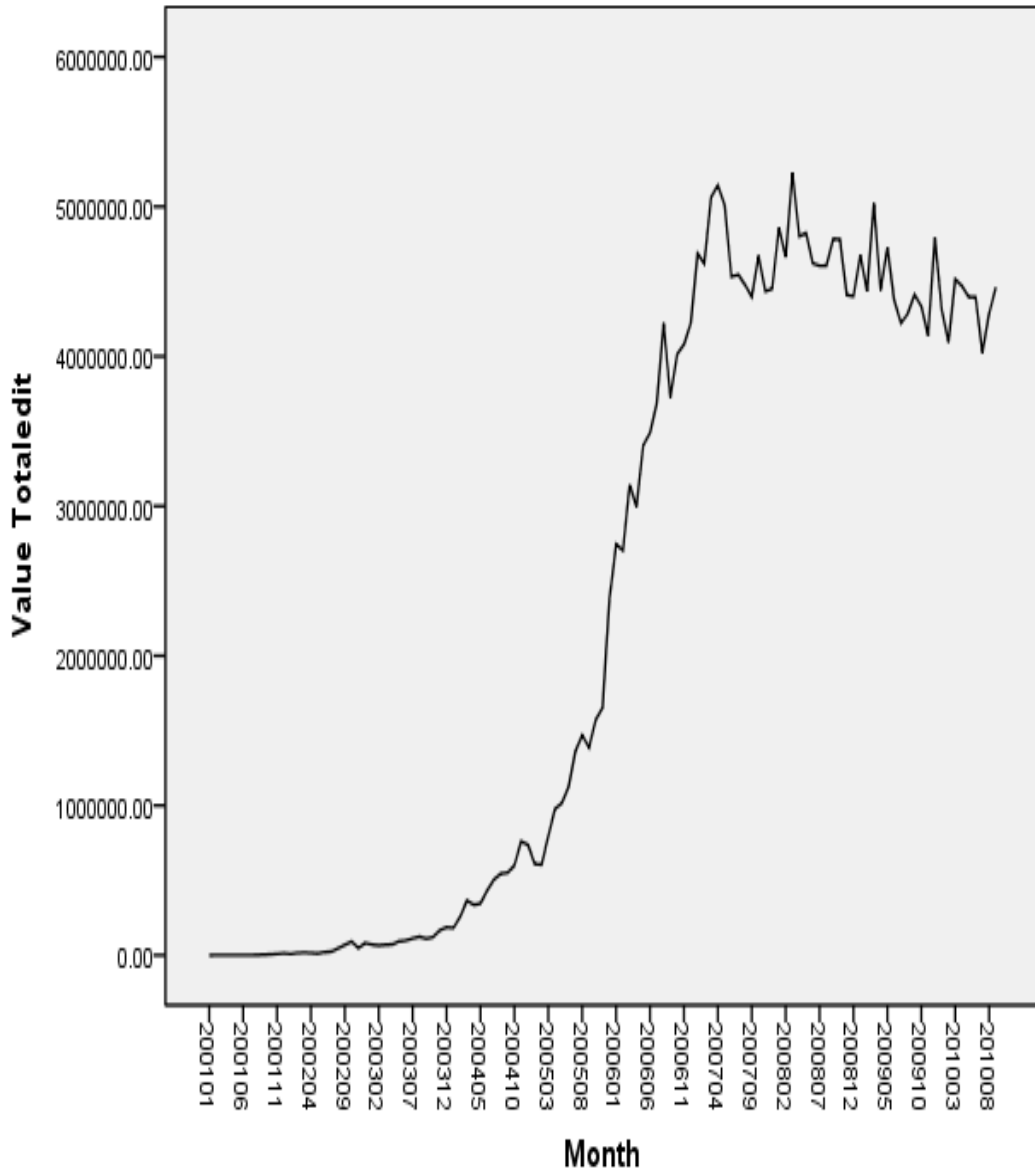


Figure 4-1 Growth of edits per month in Wikipedia

The number of active participants monthly is shown in Figure 4-2. According to our records, Wikipedia currently has millions of registered participants, but not everyone contributes all the time. The change of number of participants has a similar appearance as the change of number of edits monthly. After a steady period, the number of active participants who edited articles increased suddenly from 2004 to 2007. After this jump in active participation, the number of active participants fluctuated until mid-August, 2010.

The shift from a stable horizontal growth to a sharp rise is not an exponential model. However, it suggests that Wikipedia experienced considerable development in the period between 2004 and 2007, which will help to illuminate our hypothesis.

Based on the two observations above, we found that changes in the total number of edits per month and the total number of active participants have a similar shape, which could be described by the invariance-sharp raise-invariance model. It suggests that Wikipedia has attracted an increasing number of participations quantitatively within the specific period.

Figure 4-3 shows that the average number of edits per participant rose for a short period of time and fell quickly at the beginning of Wikipedia's development. More specifically, the sudden rise occurred between May 2002 and January 2003. The average number of edits per participants peaked at 150, which means that there were on average 150 edits made by one participant at that particular time point. However, such a number remained at a stable level from January 2006. This curve not only shows the change of edits per participant, but also illustrates that the rate of edits per participants keep the certain value from September 2006 approximately.

This section provides a general description of growth in Wikipedia observed from a number of perspectives, including the number of edits, the number of participants and the average edits per participant. According to these graphs, we found both the number of edits and participants rose sharply and then tended to remain unchanged from 2006 onwards. The average number of edits per participant had fluctuated at the end of 2002 to the beginning of 2003, then the entire line changed downwards gradually, which might suggest that the most active and productive period of participation may have been the year between 2002 to 2003 since the unit of edits per person was significantly higher. Based on these general descriptions of numbers, we propose some hypotheses in the next section to discuss the influence of participants at different editing levels.

Population growth of unique users per month in Wikipedia

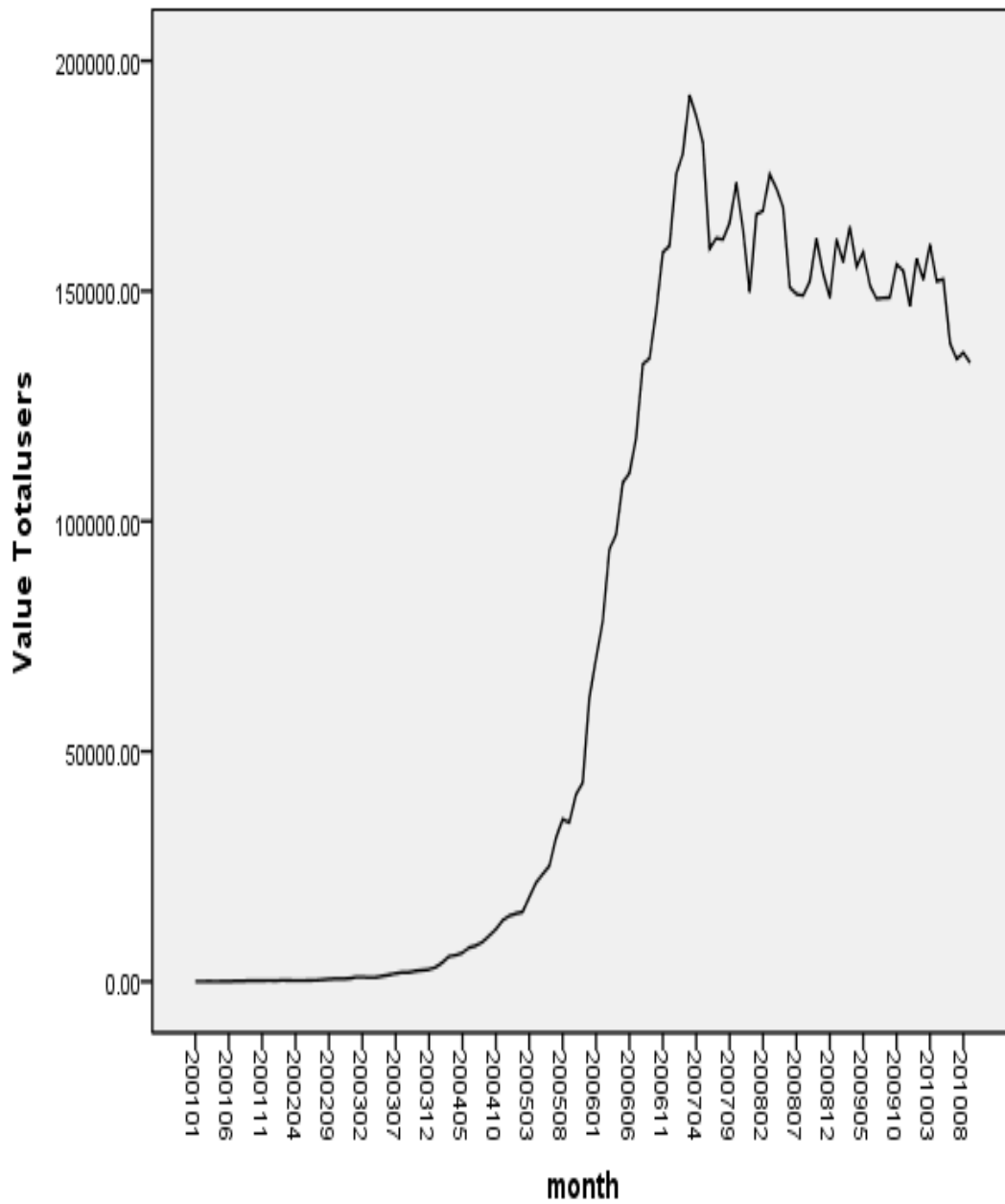


Figure 4-2 Growth of unique participants per month in Wikipedia

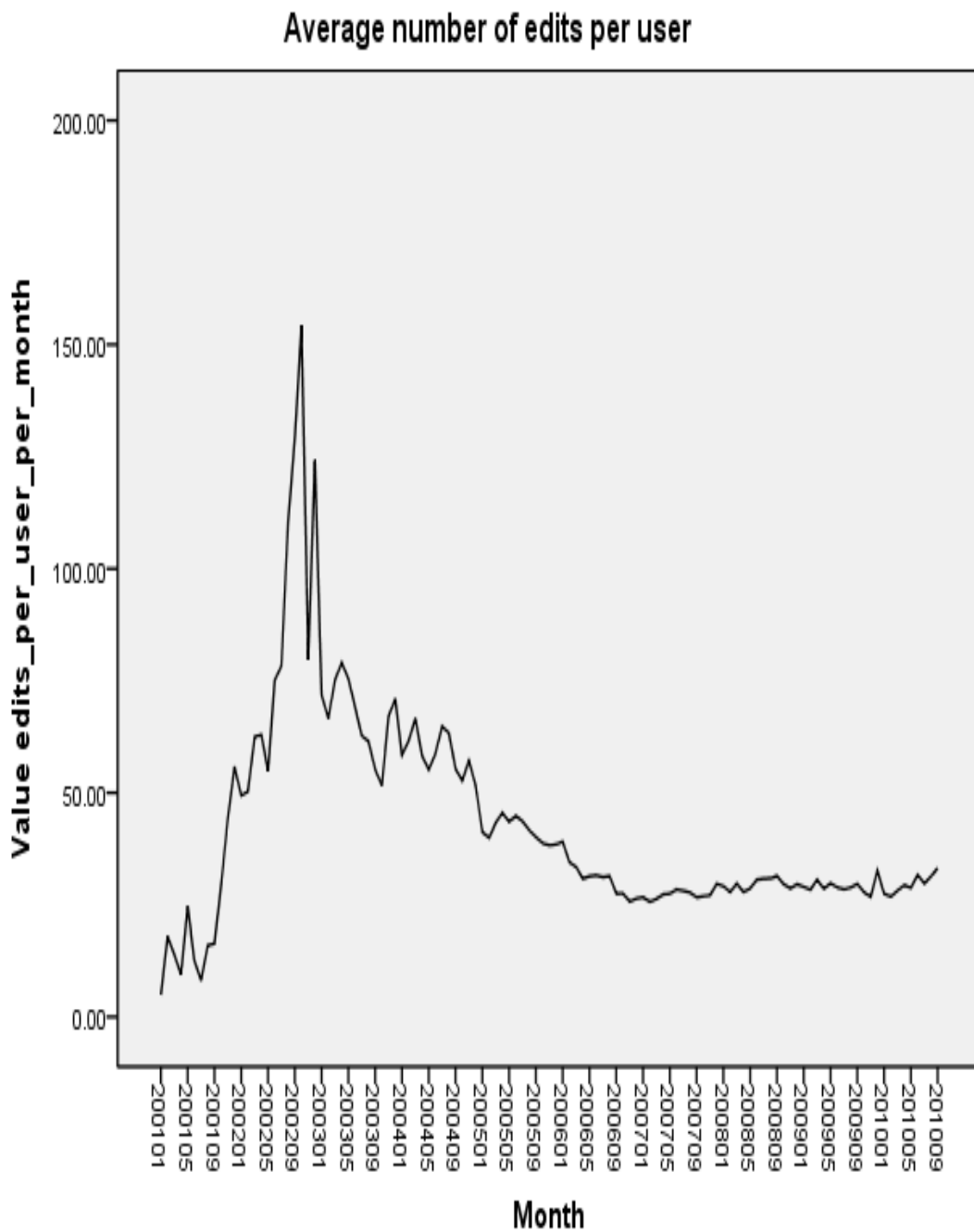


Figure 4-3 Average number of edits per participant in Wikipedia by month

4.3.2 Hypothesis of influence of participants

The last section illustrated that Wikipedia has been able to attract an impressive number of participants who have created accounts and have edited at least once. However, the fact that a massive number of people can contribute easily does not mean that everyone made an equal

contribution to Wikipedia. Similarly, the fact that anyone can modify or change the articles does not mean that all authors have similar power to affect Wikipedia as in a democratic system. In this section, we will examine the argument of Kittur et al. (2007a) to explore which mass collaborative pattern Wikipedia follows and whether it has changed during Wikipedia's development.

In order to examine how equal or unequal the contributions and the participations from different participants are, we proposed three possibilities to characterize mass collaboration. One possibility is that Wikipedia is indeed produced by a massive "nobody". Technically, this possibility would show the Pareto distribution has changed or is changing along with Wikipedia's continuing growth. In this case, Wikipedia would provide an important case study of how internet-based platforms eliminate privilege and hierarchical systems in order to make millions of participants collaborate towards one goal. The second possibility is that the content of Wikipedia is generated by a small group of active participants without authority, represented by a continual Pareto distribution. In this hypothesis, Wikipedia is a semi-democratic organization, because the great majority of participants in Wikipedia depend on their active participation instead of administration title. In other words, Wikipedia has mainly been produced by a small group of volunteers.

The two possibilities above were sketched out in previous studies but if the second possibility existed, we propose to further examine the position of the administrators of Wikipedia, and, in particular, to test for the existence of a hierarchical structure. A third possibility is that Wikipedia is still a web-based encyclopaedia dominated by a small group of authorized administrators, who have a privileged technical power in comparison with normal participants. If the third hypothesis is proven then Wikipedia, surprisingly, will not be a verification of the concept of mass collaboration. It could be that there is no mass collaboration in current internet-based platform, despite Wikipedia championing itself as a freely collaborative community.

In order to specify these possibilities against the measurable factors of Wikipedia, we summarize the process of deduction as follows:

If the first possibility is true, then Wikipedia is an entirely decentred mass collaboration;

If the second possibility is true, then Wikipedia is an elite-dominated mass collaboration;

If the third possibility is true, then Wikipedia is an administrator-dominated working system.

4.3.3 *Edits made by different groups*

To categorise participants, we have used the number of edits to assign them to a particular group. Following the detection of a Pareto distribution, we named participants with the highest number of edits as “elites” and participants with lower number of edits as “crowds”. It is important to separate the “elites” and “crowds” in our analysis in order to examine mass collaboration in Wikipedia. Therefore, we propose to use the number of edits to define groups based on the basic calculation of edits as follows. We divided participants in Wikipedia into five groups in decreasing order of editing activity:

- Group 1: Participants who made more than 15,000 (15K+) edits;
- Group 2: Participants who made between 1000 (1K) and 14999 (<15K) edits;
- Group 3: Participants who made between 100 and 999 (<1K) edits;
- Group 4: Participants who made between 5 and 99 edits;
- Group 5: Participants who made between 1 and 4 edits;

The numerical criteria defining the groups are based on our basic calculation about the number of edits made by the most active participants. The ‘15K’ is the average number of edits of the top 100 participants in 2007 and the ‘1K’ is the average number edits of the top 100 participants in 2004. The ‘100’ is the average edits of the top 100 participants in 2002. The ‘5’ is the standard Wikipedia statistics offer to define the ‘Wikipedian’, active participant.

Figure 4-4 shows the growth in the number of participants in the different groups. The curve illustrates that group 1 and 2 represented by the green and blue lines, share a similar shape, which kept the same value from 2001 to 2004, and increased sharply until the middle of 2006, after which the number of participants fluctuated and decreased gradually. It is noted that the period of growth takes place between 2004-2006, which is the same period as an increase in the number of total participants takes place. The number of participants in group 3, represented as a brown line, rose gradually from 2004 to 2006 when compared to group 1 and group 2. The number of participants in group 3 has also fluctuated, and decreased slightly from 2006. The number of participants in group 4 (purple line) has a shape generally similar to that of the group 3 but the gradient is much more gentle. Finally, the number of participants in group 5 (yellow line) has increased only fractionally. That suggests that from the outset the number of those who have made the largest number of edits has always been small. Conversely, the number of participants with fewer than 100 edits totally rose sharply and comprises the majority of participants.

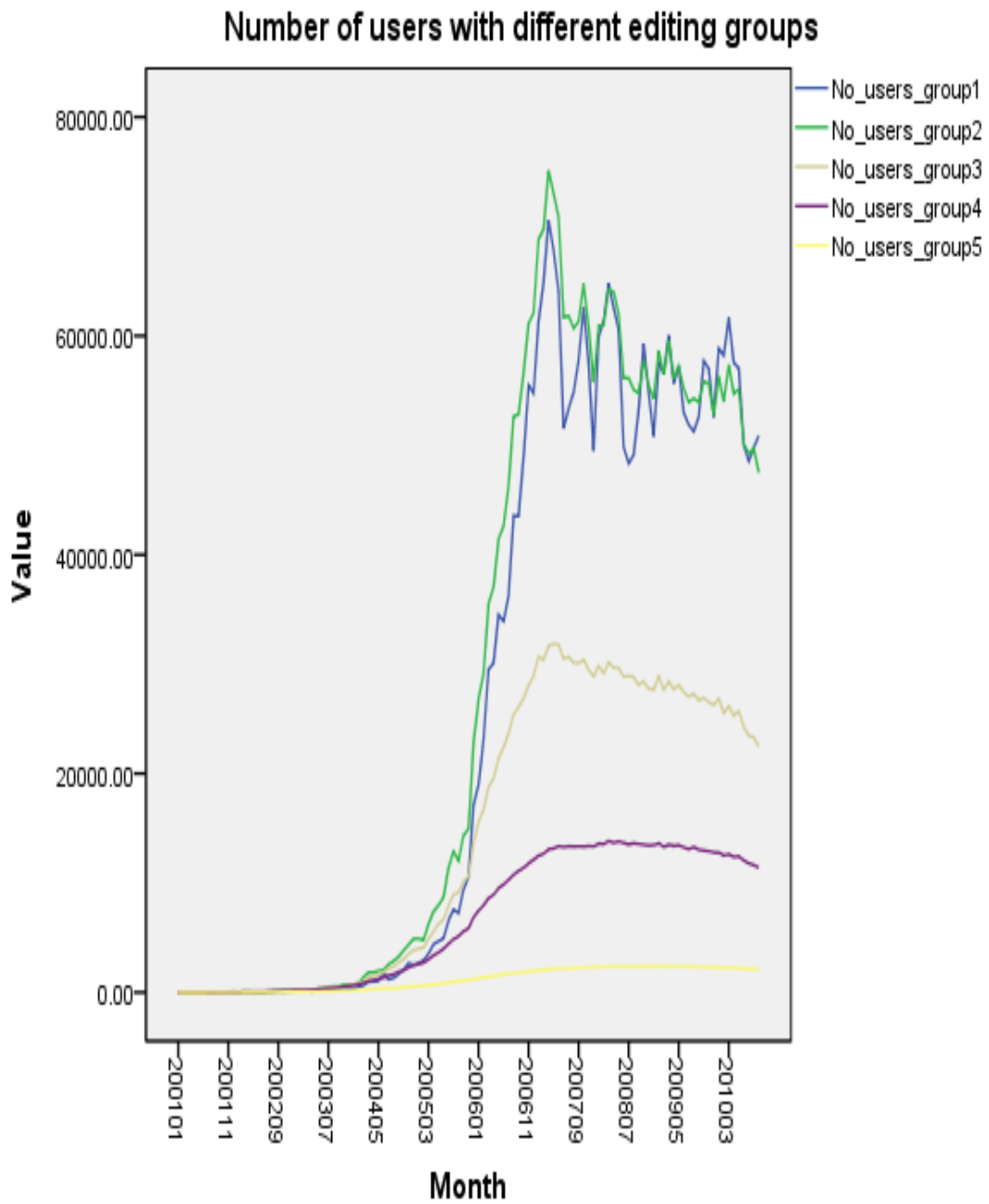


Figure 4-4 Number of participant grouped by the number of edits monthly

[Originally in colour]

We now turn to discuss the number of edits made by participants in different editing levels based on Figure 4-5. It is interesting to note that the order of edits made by participants in different groups is the reverse of the order of the number of participants in the different groups. Specifically, group 5 which had more than 15,000 edits in total, has provided the highest number of edits, although it has the smallest number of participants. In fact, the shape of the number of edits made by participants differs from that of participants. The curves representing the number of edits have fluctuated much more than the lines of participants. However, when combining these two figures, it becomes clear that though the majority of participants are in group 1, the group who made fewer than 5 edits, have the lowest number of contributions in terms of absolute value; whereas the smallest number of participants, those in the group 5 who made more than 15,000 edits contributed at the highest level compared to other groups. It is also noted that group 4 and 5 share an almost similar shape until the end of 2006. After that, group 4 declined gradually whereas the number in group 5 increased with some fluctuation.

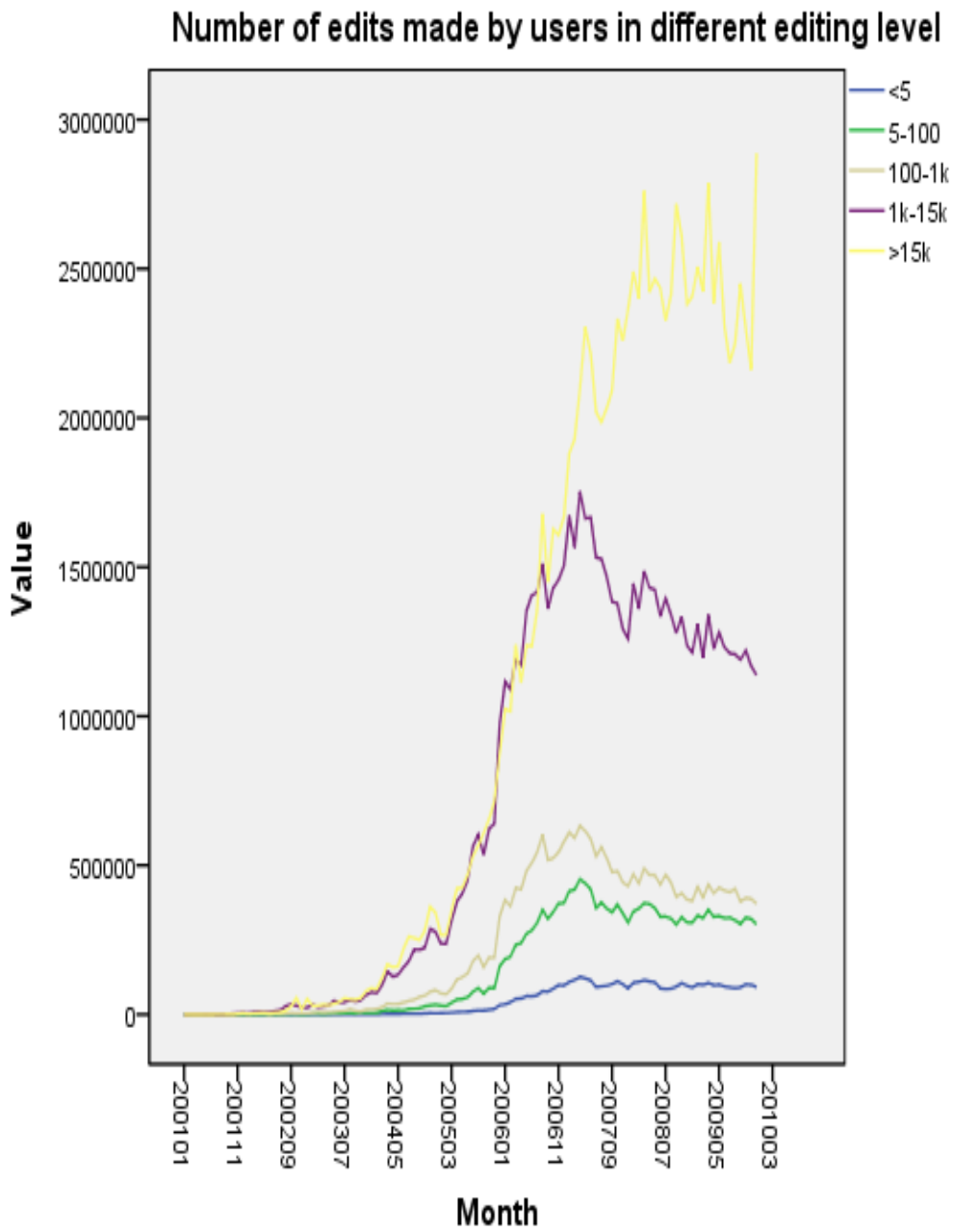


Figure 4-5 Number of total edits made by participants in different editing groups

[Originally in colour]

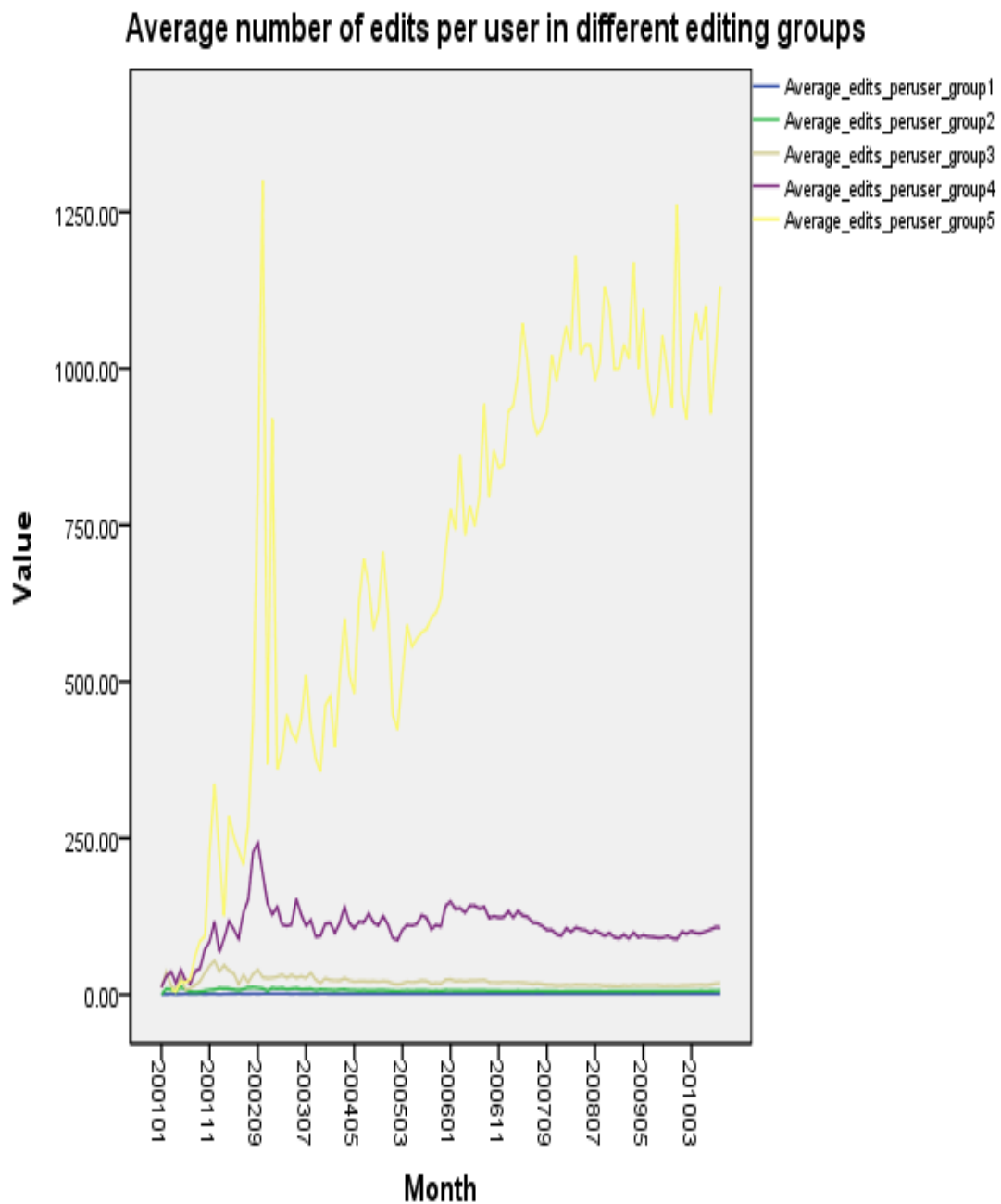


Figure 4-6 Average number of edits per participants in different editing groups

[Originally in colour]

To further analyse the number of edits used to define the groups, we calculated the number of edits per participant in different groups. Figure 4-6 shows that the average number of edits per participant in different editing groups, in which the average number of edits in the group 5 (elites participants with more than 15000 edits) has the highest average number of edits. More

importantly, the peak point of approximately 1250 edits per participant has been reached twice during Wikipedia's development in 2002 and in 2009. It suggests that high-contribution elites might have had two active periods in which their contribution rate increased. Another important finding is that the average number of edits per participant in group 1, 2, and 3 is very similar and there were no marked changes over the last decade.

In this section, we have discussed the changes in the number of edits, participants and average number of edits per participant in different groups as defined by the number of edits per individual. We found that the small group of people have edited more than the groups including a large number of participants. However, the absolute value of edits in different groups could not define who had contributed the majority of the content in Wikipedia. In the next section, we calculate the percentage of edits made monthly by different groups, which will allow us to specifically explore who have made the major contributions in Wikipedia and also to examine whether there was any change in editing practices during the development of Wikipedia.

4.3.4 Changes of proportion in edits by different groups

This section will examine the proportion of edits produced by the different participation groups. Following the study of Kittur et al. (2007a), we can test whether the proportion of edits has changed from 2001 to 2010. Another important point is their claim that this "shift" occurred in 2004 (Kittur et al., 2007a).

In Figure 4-7, it is clear that group 5 made approximately 40% of the total edits in the latter part of 2002. More interestingly, group 4 also contributed quite a high percentage from March, 2001 to the latter part of 2002; whereas prior to that, group 3 contributed the majority of edits. Another important issue is that group 4 has edited almost 40% from 2003 to 2006, which was roughly equal to the proportion made by group 5. However, this balance of equal contribution was broken from the end of 2006, when group 5 started making the absolute majority of contributions among the five groups and the percentages made by other groups have remained stable up until 2010.

Figure 4-7 shows that groups divided by their number of total edits have been responsible for the majority of edits in order of: group 3; group 4; and group 5 – which made around half of all edits during Wikipedia's development. This observation illustrates that the development of Wikipedia has always been led by a small group of participants. In other words, a small group of top-level editors has always been responsible for the majority of content on Wikipedia. However, along with the development of Wikipedia, the standard of top-level edits has changed from hundreds to millions. Following this assumption, we can deduce that if the

participants with the top-level edits have maintained their contribution in both quality and quantity, the quality and quantity of Wikipedia should not have weakened.

Although in the specific period under review, group 4 and group 5 share an almost equal percentage of edits, we still can claim that Wikipedia has been produced by a small group of active participants because the total number of participants in those groups comprises fewer than 20% of the total number of participants. However, there is no evidence of any marked shift in the proportion of the overall total of edits which these two groups contributed from the middle of 2002 to the end of 2010.

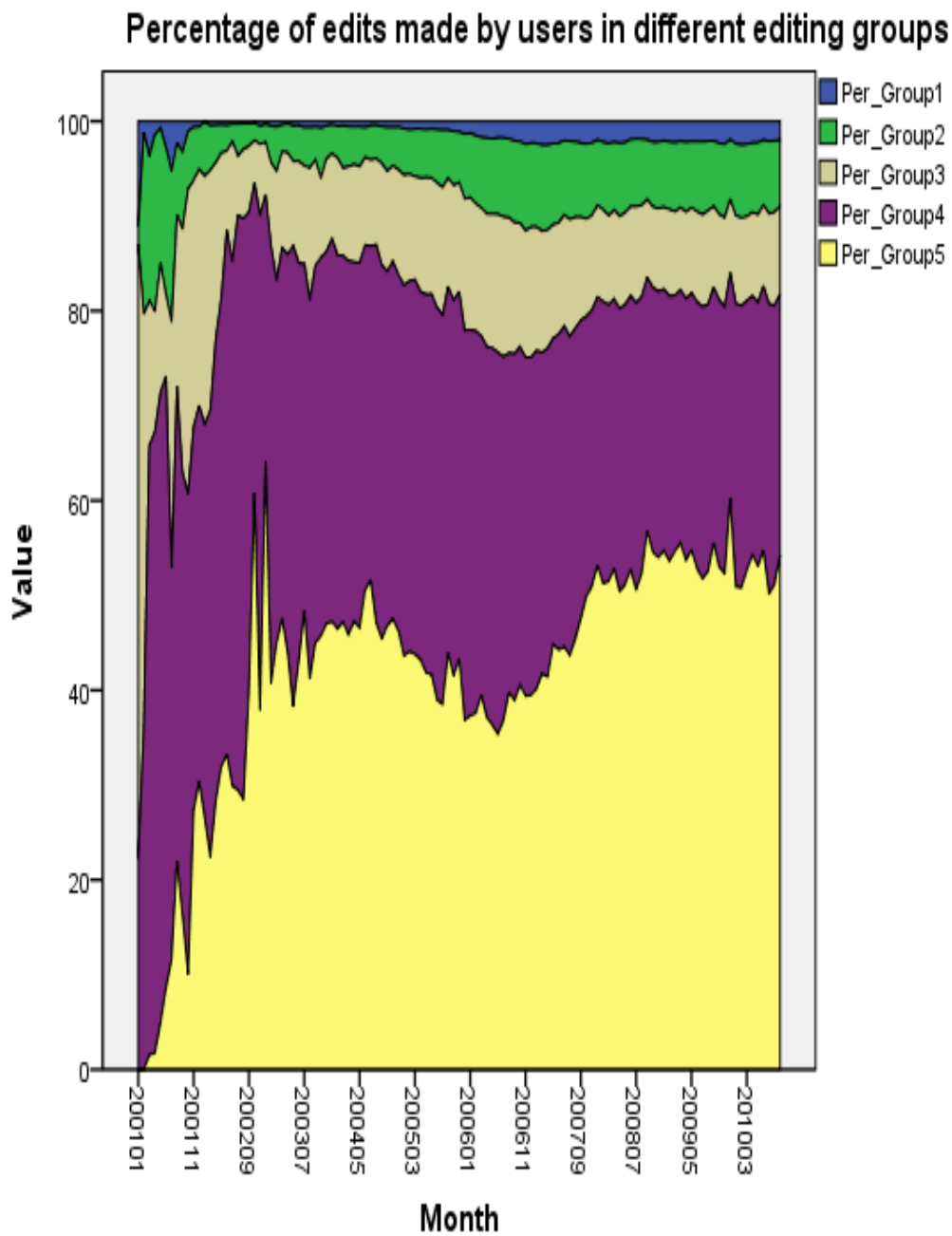


Figure 4-7 Percentage of edits made by participants in different editing groups

[Originally in colour]

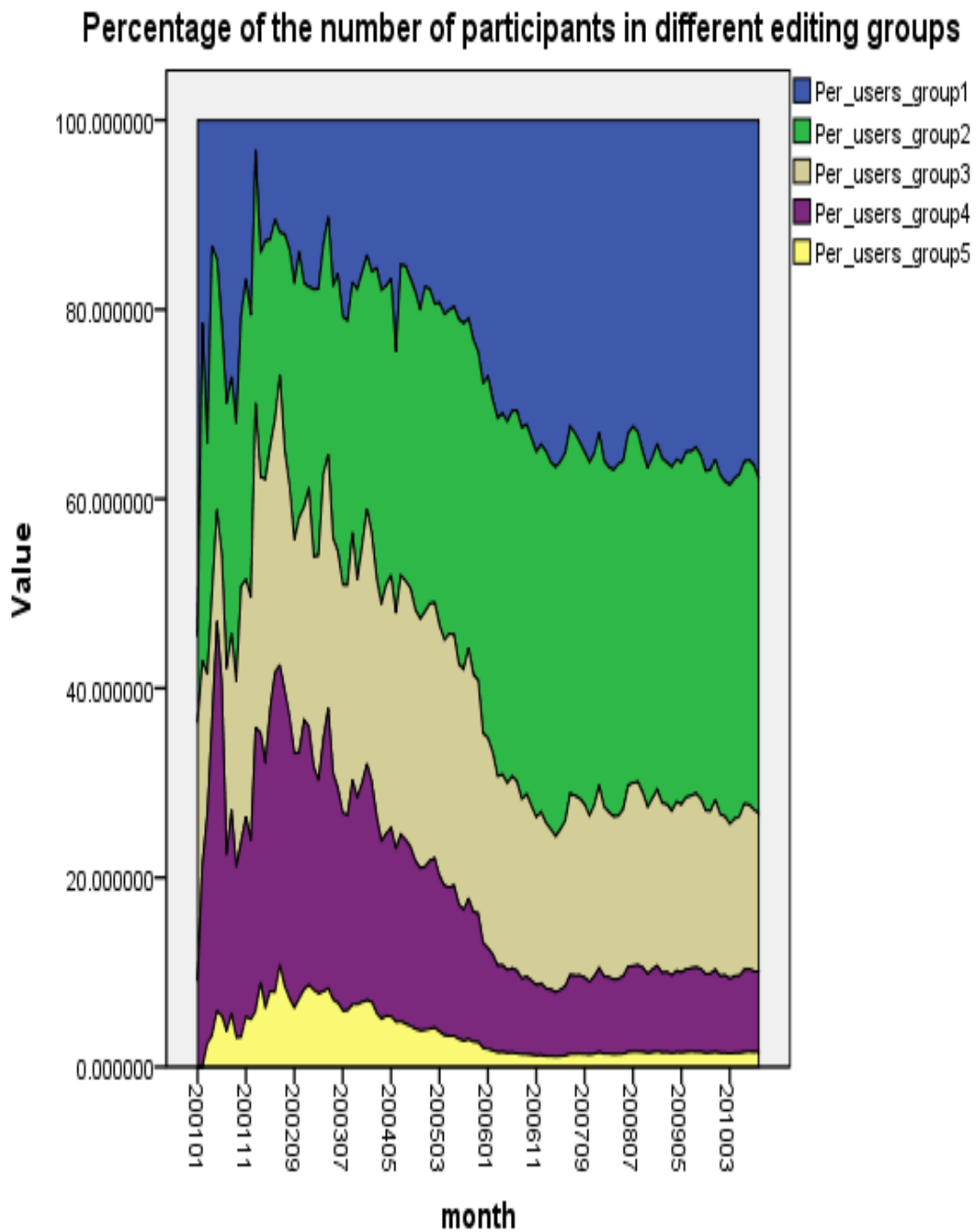


Figure 4-8 Percentage of the number of participants in different editing groups

[Originally in colour]

As showed above, Figure 4-8 discusses the change in the percentage of participants in different editing groups. It clearly indicates that the percentage of participants with different editing levels has fluctuated since Wikipedia was founded, up until 2006. After that the

percentage of groups 1 and 2 has increased to almost 40% each, while group 5 keeps less than 3% of the total number of participants. It does not only show that the participants with higher edits only take a few percentage of participants, also illustrates that participation in Wikipedia may have stabilised in 2006 and all groups have remained unchanged since then.

Figure 4-7 and Figure 4-8 demonstrate that Wikipedia still follows the Pareto distribution, in which a majority of content is only produced by a small group of participants. The results from both the number of edits and the number of participants suggest that Wikipedia has maintained a continual Pareto distribution over the last decade. Therefore, we propose that Wikipedia might follow either the elites-dominated system or administrators-dominated system. However, this result by itself cannot provide an answer as to whether Wikipedia is produced by privileged participants-administrators or just unauthorized participant “elites” who have come to dominate the semi-democratic system. This issue will be discussed in the next section.

4.3.5 Influence of administrators

The previous section introduced the proportion of edits and participants respectively in different groups which were identified by their edit level, whereas this section will examine the percentage of the amount of edits and the number of people categorised between administrator groups and normal participants group, which is defined based on their authority. With this issue explored we will then be able to discuss whether Wikipedia is an authority-oriented system or a non-authority-oriented system. Although in the last section we clarified that Wikipedia has been produced by a small group of participants, we cannot define whether administrators were included in this small group of “elites”. Therefore, we have divided the participants into administrators and non-administrators to test whether administrators contributed more to Wikipedia.

First of all, we need to clarify who the administrators of Wikipedia are. As we introduced in chapter two, administrators are selected by an agreed process, which involves: self-nomination or nomination by others; providing a self-statement; a poll; and empowering them with an administration title. Administrators are elected by the consensus of more than 80% approval in the community. Wikipedia claims it is a non-hierarchical system in which the status of administration is granted only for technical reasons since the server cannot entitle everyone with similar technical rights. However, administrators do indeed have some privileges which general participants do not have, such as blocking or unblocking articles, and

blocking or unblocking editing rights for particular participants²⁸. Because of this, administrators could be considered particular participants with specific rights.

Unlike the “elites” with higher number of edits, administrators may not have a higher quantity of contributions, because the number of edits is not the primary factor of their election²⁹. In other words, administrators are not required to be “elites” with a large number of edits to stand for election. In particular, we will first explore the number of edits made by administrators. It is clear that the contribution of administrators has increased sharply since the end of 2006 and fluctuated afterward.

Administrators may affect participation in Wikipedia in different ways due to their ability to control others’ edits. As we can see Figure 4-9, the average number of edits per administrator is higher than the average number made by normal participants. Generally, the average number of edits per person made by normal participants has remained steady between 50 to 100, whereas the number of edits per person made by administrators rose and remained between 500 to 600. In other words, the number of edits per person made by administrators is ten times that made by the crowd. This contrast repeatedly manifests the Pareto distribution of Wikipedia where the administrator group representing a small group of elites contributed the most.

²⁸ The details from the administration policy of Wikipedia at <http://en.wikipedia.org/wiki/Wikipedia:Administration>

²⁹ In the explanation of guiding to requests for administrators nominees, the number of edits is important but not identified as a necessary factor of success, from http://en.wikipedia.org/wiki/Wikipedia:Guide_to_requests_for_adminship

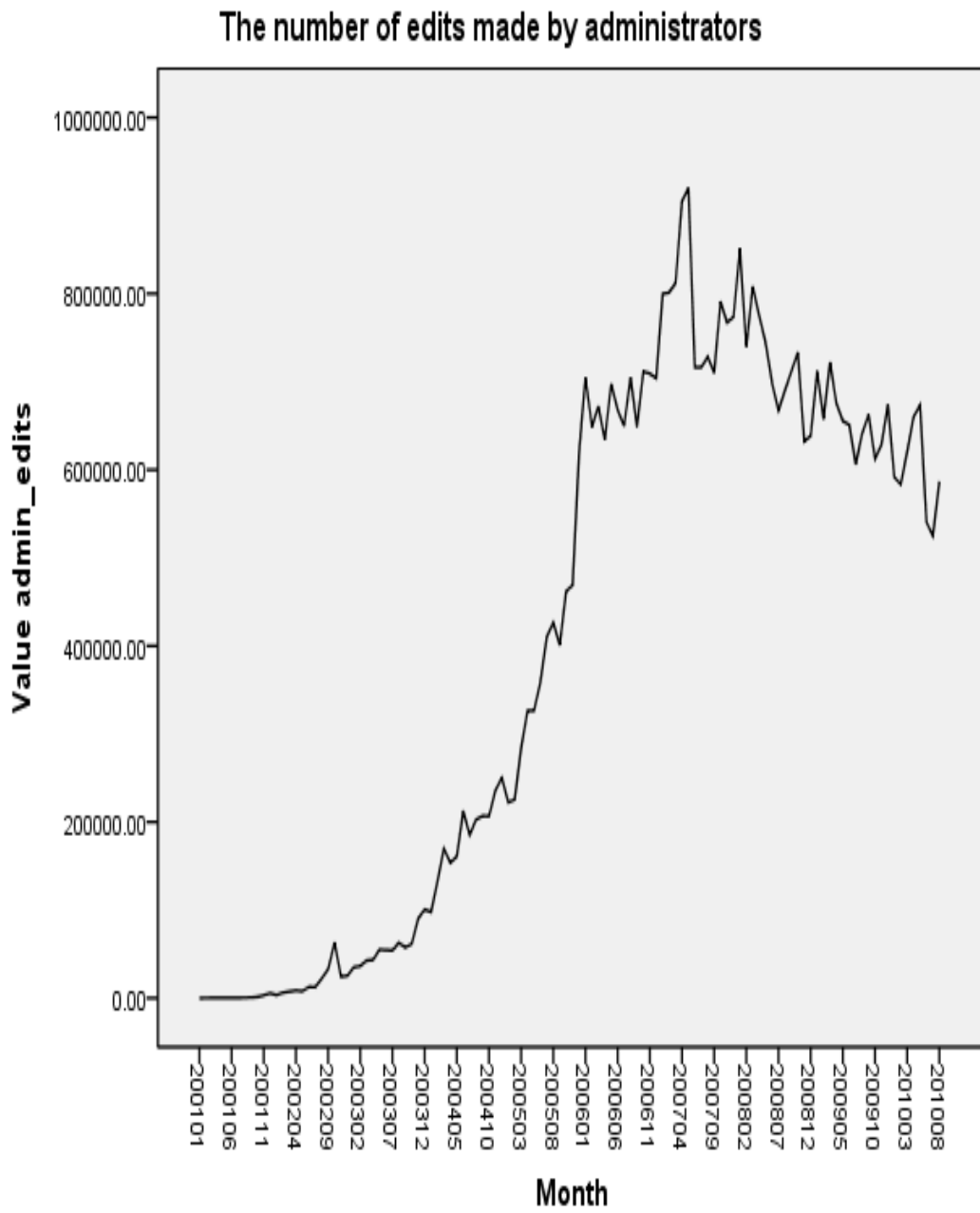


Figure 4-9 Number of edits made by administrators per month

Average edits per participants in admin and normal user groups respectively

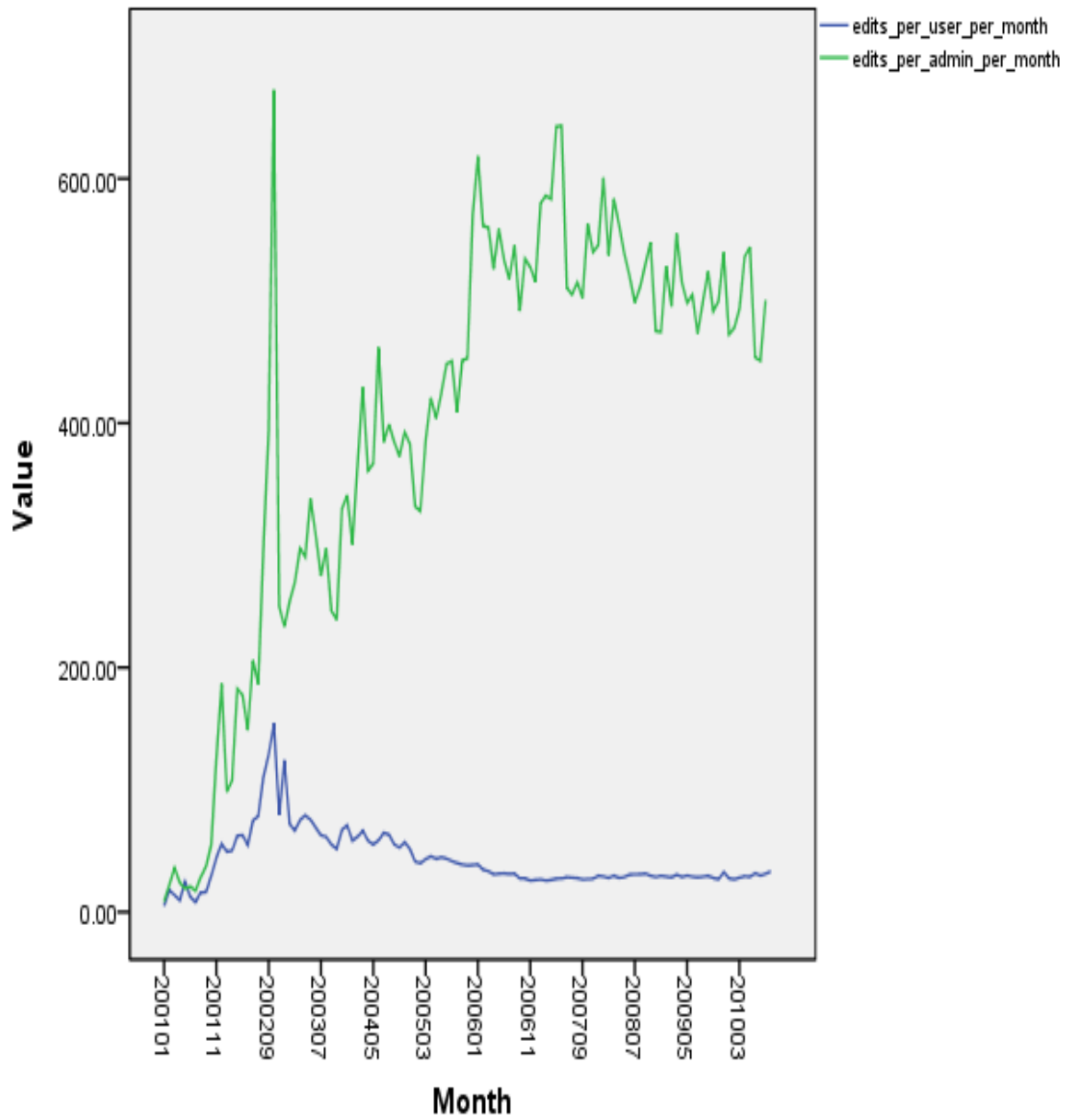


Figure 4-10 Average number of edits per participant in admin and normal participant groups respectively

[Originally in colour]

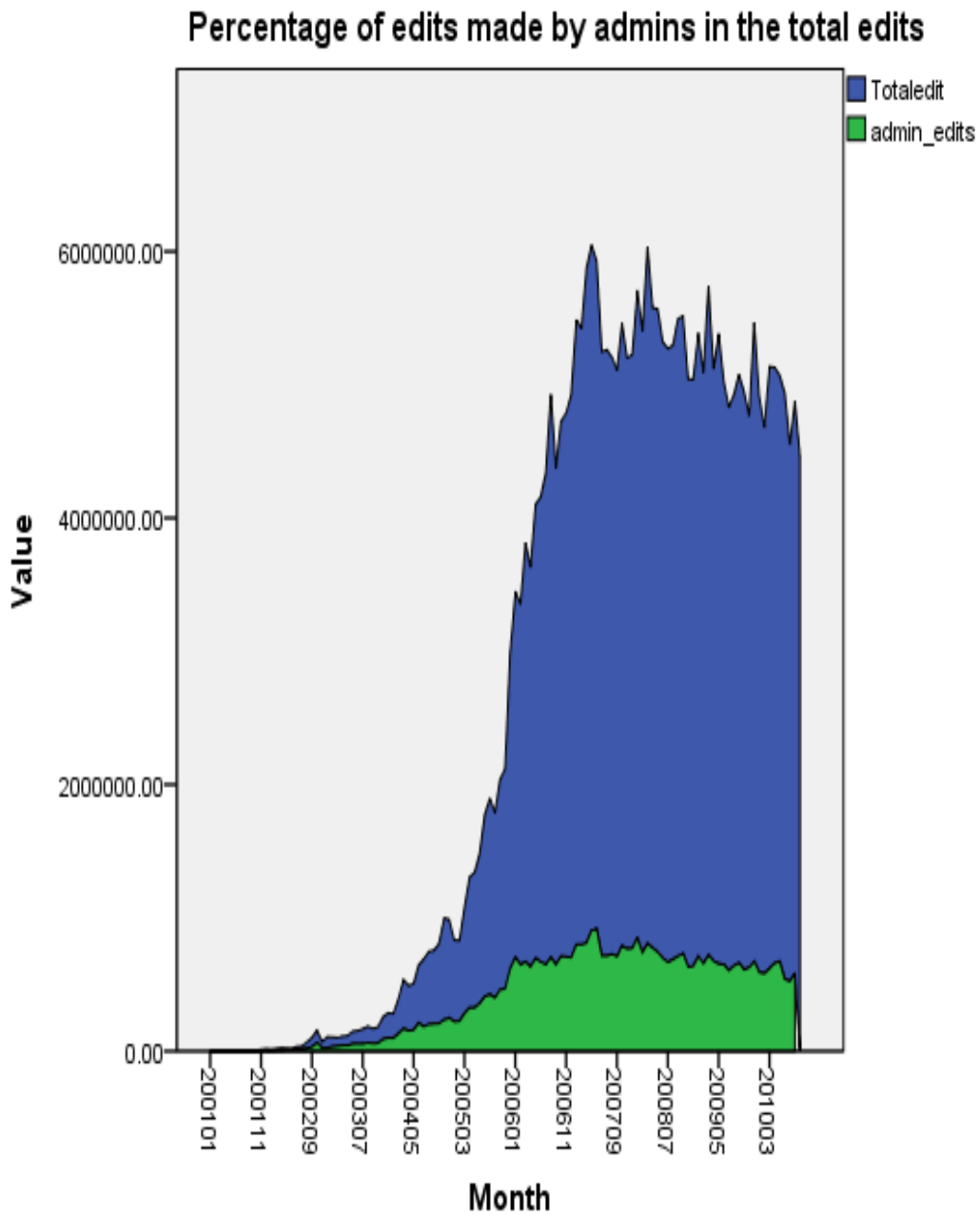


Figure 4-11 Percentage of edits made by administrators against total edits

[Originally in colour]

Although the average number of edits per administrator is higher than that of normal participants, the percentage of edits made by administrators did not count bigger in the total number of edits in Figure 4-12. It shows that administrators' contributions do not dominate

the total number of edits, unlike those of the “elites” with high-edits which we highlighted in the previous section.

This section has discovered the administrators’ influence on Wikipedia by calculating their edits and the percentage of total edits. There are two findings: first, the quantity of edits produced by administrators is much higher than the average amount of edits made by unauthorized participants, which suggests that administrators are one of the important parts of the “small group of elites” who contribute the greater part of content to Wikipedia. Second, the number of edits produced by administrators actually provided only a small portion of total edits to Wikipedia, because there are only over 1790 administrators (as of data) compared with millions of participants. The fact that administrators have made only a small proportion of edits also demonstrates that they have not dominated the edits in the way that our hypothesis postulated. Therefore, the hypothesis of an administrator-dominated system is disproved.

Thus, by combining the examinations discussed in the previous sub-sections within this section, our analysis indicates that the mass collaboration pattern in Wikipedia has semi-democratic characteristics, in which non-authorized elites with high-level edits dominate a majority of content in terms of the quantity of edits.

4.4 Conclusion

The success of Wikipedia has attracted intense academic attention; especially since a large body of diverse volunteers generates its contents in a seemingly loose and uncontrolled organizational system. The question of who has made the majority of contributions to Wikipedia has been the subject of heated discussions within and outside Wikipedia’s community. Some believe that the secret of Wikipedia’s success is the result of the cohesion of administrators and their higher contributions, and accordingly they have claimed that Wikipedia is a hierarchical system like other off-line multi-author productions (Beschastnikh et al., 2008, Burke and Kraut, 2008b). Others have claimed that Wikipedia is an innovative system which provides equal rights to all participants to affect content, and their opinions supported the “wisdom of crowds” in terms of practice (Lih, 2009b, Surowiecki, 2004). Between these polar opposites, is the view of Kittur et al. (2007a) who advanced a third view, that the mass collaboration pattern in Wikipedia has shifted from manifesting the contribution of smaller portion of elites to the wisdom of crowds.

In contributing to this academic discussion this chapter offered descriptive figures to indicate the quantitative changes in articles, edits, and participants. Secondly, we proposed three possibilities of mass collaboration which might be represented in our findings; including a

democratic system in which everyone has equal authority; a semi-democratic system which is dominated by un-authorized elites, and a hierarchical system dominated by elected administrators.

In order to evaluate the proposed possibilities, we divided participants in two ways: the first were divided into five groups according to the quantity of their edits; and the second were divided into two groups based on their privilege status. Based on these definitions, we compared the proportion of edits produced by each group to determine their domination of Wikipedia's content in quantitative terms.

We have demonstrated that the overall change of edits was not due to the growth of low-edit participant groups, but instead was driven by an increase in elite participants' activity according to the rise of average edits per person in different groups. In other words, the elites with a high number of edits were responsible for approximately half of the total edits from 2006 because of their active and continuously increasing contribution compared to unchanged level of edits by other groups. It was also illustrated that participation in Wikipedia did not shift from a small group of elites to a massive contribution from crowds, as suggested by Kittur et al (2007a). On the contrary, after a slight shift between activities in groups 4 and 5, the elites with the number of edits still continued to produce around 50% of edits from the middle of 2007. These results suggest that elites with the highest number of edits were not only the early pioneers who selected and refined Wikipedia's direction, but also were the mainstay of maintaining its development.

Although Wikipedia did have a team of administrators to maintain the organization, those administrators did not dominate the content. It can be assumed that Wikipedia is directed by the group who make a majority of these contributions. Our analysis demonstrates that the administrators generally are high-edit participants but their edits did not count for a considerable percentage of the total edits. Therefore, it was found that administrators are unable to affect Wikipedia's content in a meaningful sense. It is noted that our study did not deny that administrators influenced maintaining Wikipedia in terms of organization, such as preventing vandalism. Our argument is that administrators are not the major contribution group that actually works on content.

This chapter suggests that Wikipedia is produced by a small group of participants regardless of their administrative status. A collaborative process like Wikipedia is different from that of traditional technological products in that a certain number of authorized members could be empowered to manage the entire system. Wikipedia has been dominated by a small group of participants with high-edits, but such a group is dynamic and changes along with the activities of volunteers.

We also argued that such a collaborative system is much looser than the traditional multi-author system in which knowledgeable persons have been selected to be responsible for edits according to their knowledge and related experience. Wikipedia offers a self-nominated system where participants could work at editing more to become one of the dominating groups. Based on our studies, we clarified that the administrator group does not dominate the process of mass collaboration on Wikipedia, which suggests that being an administrator might not be a decisive factor to consider when participants want to make their contributions more significant.

Our judgement of the dominating participants of Wikipedia is based on the examination and comparison of the amount of edits and the proportion of total edits to which they had contributed. There are two main reasons why we chose to investigate the amount of edits. First, this is the way that previous researchers have employed to simplify the collaborative model of Wikipedia to a type of production activity based on the unit of edits. This simplifies the process of analysis and facilitates direct calculation when there is a large database. Secondly, our research regards Wikipedia as a new cooperative model and an example of mass collaboration based on the discussion in chapter three. In fact, mass collaboration is the process of people working together under a certain organizational system. In the traditional organization system, we take the working hour of individuals as the basic evaluating method to calculate their contribution. Here, in Wikipedia, the basic calculating unit of their contribution is the “edits” they make. Therefore, although edits might not be only way to evaluate participants’ contribution in Wikipedia, we regard it as the basic unit of evaluating contribution.

From a methodological perspective, this chapter contributed to the issue of applying digital by-product data to explore the influence of participants when divided by their edits. Generally, this chapter attempted to use digital by-product data to define their cases instead of using social data such as gender and age collected by surveys. We used the quantitative term of what they did to identify what they are and what influence they have. This chapter demonstrates that digital data can be used to describe a macro view of participating behaviours dynamically.

Besides defining distinct groups, this chapter demonstrated the effectiveness of graphic presentations in the data mining process. The thousands and hundreds of cases have been formulated into several figures to illustrate changes alongside Wikipedia’s development. The emphasis was put on the dynamics found in the time line. This method offers considerable advantages for exploring a huge number of data in spatio-temporal terms.

The findings of this chapter led us to ponder another question, what is the role and function exactly for administrators within Wikipedia's mass collaboration model. As we have shown, Wikipedia is a semi-democratic system in which the majority of contributions come from the elite groups, so we wonder what the value of these administrators is in ensuring mass collaboration. This question will be addressed in the next chapter.

Chapter 5

Visualizing mass-authoring collaboration in articles

The previous empirical chapter examined the possibility of domination by particular groups in the collaborative participation of Wikipedia. We discovered that the most influential participants in Wikipedia were those who were without the privileged power of administrators, but had a large number of recorded edits. By comparing the quantity of edits made by different types of participants, we concluded that Wikipedia is operating a semi-democratic system; the content of which is dominated by a population of elites without the authority vested in administrators to change content. Such a conclusion was drawn because a small group of high-edit-record elites contributed to more than half of the contents on Wikipedia since 2006, and there was less influence from administrators with regards to the quantity of edits. In fact, we discovered that the group of administrators did not have a considerable percentage of total edits so as to dominate the direction of content. Thus, we demonstrated that Wikipedia is mainly produced by a small group of non-administration participants with high levels of enthusiasm and a high edit-record.

In the previous two chapters, we focused on a broad and general analysis of Wikipedia to explore mass collaboration. This allows us to understand the development and organization mechanism of Wikipedia; how each article in Wikipedia comes about with the collaboration of millions of participants awaits further research. Therefore in this chapter, we will utilize visualization to study the establishment and development of individual articles.

As a product created by millions of participants, we cannot assume that collaborative activities on Wikipedia only show one pattern. Therefore, we attempt to describe such patterns of collaboration by visualizing the edit histories of individual pages. Every individual page could be treated as a single dataset of mass collaboration for visualization. In order to understand how mass collaboration works on individual pages, all investigations will be based

on digital by-product data, in contrast to transactional data from traditional sample surveys. This approach will demonstrate how digital by-product data could be used by social science researchers taking a micro-view. Because the histories of edits in Wikipedia are generated in an electronic format, special tools are required to extract and analyse that data. Also, because Wikipedia has millions of webpages including articles, discussion pages, user pages and so on, this study will pick out samples from different type of articles to generate patterns of mass collaboration.

5.1 Introduction

We argue that the best way to extract patterns from a large amount of data with regular patterns and in chronological order is visualization. First, we clarify what editing behaviours are in the context of the establishment and development of individual articles, and demonstrate the possibility of visualising this data. We then introduce what visualization is and its wide application in scientific research. Finally, we introduce the advantages of applying visualization and the detailed procedures involved.

5.1.1 Editing behaviours and the difficulties of reading and understanding them

As we previously introduced, editing behaviours are recorded and visible for us in the digital by-product data format. However, these datasets may be huge with many intricate connections, which make it difficult for researchers to extract useful information using simple data processing approaches. For studies aiming to gain a microscopic view of Wikipedia, researchers wish to observe more clearly and directly how Wikipedia developed, how each article came about, what each participant brought to the site and similar information. However, as the generation of digital by-product data is automatic and without any particular purpose, the data that we wish to acquire will not simply appear, and even if links appear, there will be no straightforward answer to those what and how questions without some in-depth analysis of the data. For example, in looking at the million editing records for the article ‘York’, each contains the time of edit, the participant and the content of that edit. Yet the simple conglomeration of edit records still cannot answer the seemingly simple questions of how the article “York” came into existence, and how the editing of the article took place. Using multi-dimensional digital by-product data, we want to visualize how people produce an article collaboratively. Although the platform based on Wiki-technique can store all individual edits in a chronological order, it still remains a challenge for us to use and present the data appropriately.

The three main challenges of using large databases to answer descriptive questions are: First, the amount of data defies answering any academic questions by its sheer scale. In fact, we can hardly present all relevant information in a single graph with a single dimension. With regard to the scale of production in Wikipedia, the English Wikipedia has seen 122,387 edits every 10 minutes from 2004 to 2006. Such a large scale of collaboration is hard to scrutinize and understand in simple terms. Secondly, the intricate data structure prevents scrutiny of all the relationship between data at the same time. Some of the data may have direct or indirect connection with others, and understanding such relationships can provide further insight and enrich meaning. For example, Wikipedia records all editing behaviour in chronological order. However, the data from the record do not readily reveal that some of these edits are for the same article, and some edits are made by the same participant. Some edits may not be written by one participant, they may be generated by authors who share a common interest. This extra information hidden in the database can serve as beneficial supplementary explanations for us to understand the database. However, realistically, it can be a challenge to express this supplementary information together with the originally findings about editing behaviours simultaneously. Third, the behaviours are dynamic, which make it difficult to clearly demonstrate the changes and the multi-dimensional connections using simple textual descriptions or line graphs. To illustrate the point, with regard to the frequency of article edits, on average an individual article has more than 17 edits³⁰, and most featured articles have ten-thousand edits in total. An article's editing history includes so many individual edits which prevents easy perception of a clear pattern. Particularly, edits created by different participants with varying lengths can increase the difficulty of using statistical modelling.

We want to specifically narrow down our research to observe mass collaboration in Wikipedia articles. It is generally difficult to describe how people edit one article with equal rights in Wikipedia, as articles in Wikipedia mostly appear as a finished textual product, which provides no details of how much each author contributed to the article's content. The structure and the content of the article, the editing participants, the time of editing and the number of edits all comprise an intricate dataset, which prevents a simple description by textual or graphical explanation.

Therefore, the only possible way to see the evolution of an article produced by thousands of participants over time is to represent all editing behaviours with a timeline. In doing so, we need to identify all edits and the respective authors in the same image according to the time of edit. Therefore, we explored the presentational form of data mining– which is visualization. In order to describe how people collaborate in producing an individual article and evaluate the

³⁰ http://meta.wikimedia.org/wiki/Wikimedia_in_figures_-_Wikipedia

article based on edits, the most direct and efficient way is to use colour-coding graphics to describe the information flow.

5.1.2 What is visualization

Visualization is a vague term without any agreed definition in academia. It is often used to analyse large databases and represent the interesting information behind it. This method has been applied widely in scientific research since the 1960s (Gallagher, 1995a). In the early days, visualization was used in the field of computer graphics and engineering design, in order to make, “Complex states of behaviours comprehensible to the human eye” (Gallagher, 1995c). This method has been described as the process in which, “Images and signals may be captured from cameras or sensors, transformed by image processing, and presented pictorially on hard or soft copy output” (McCormick et al., 1987b).

In particular, the definition of visualization in science research emphasizes two issues: function of display and a solution to a problem. First, visualization is an improved solution of communication (Shephard and Schroeder, 1995), and a comprehensive way to display outputs (Gallagher, 1995c). The goal of using visualization in scientific analyses is to help interpret the results of the information generated from studies. Scientists and engineers believe that visualization can depict their intricate results in an appropriate manner, without tampering with the accurate analysis produced (Shephard and Schroeder, 1995).

Second, visualization is more than merely displaying results in scientific research; it also plays an important role as part of the analysing process during the research. In computer science and engineering, visualization has been concerned as an equally essential component as modelling and analysis in engineering design. It is clear that visualization using computer graphics shares an equal position as modelling and analysis in scientific research. Just like contributing new equations in order to contribute to research, scientists now could also improve visualization to enhance their research performance. Generally, visualization itself has been regarded as a science to be concentrated on and developed into different representation and techniques (Gallagher, 1995b); such as surface rendering algorithms (Kaufman and Sobierajski, 1995), and animation design (Pepke, 1995).

The visualization in traditional scientific areas mainly relies on calculations and simulations by computers to generate three-dimensional products in architectural, meteorological, medical and biological fields. Visualization represents a single, unified collection of computer graphic techniques for displaying scientific behaviours. The reference work of visualization in scientific research includes how to mine visible data; how to use computing techniques to visualize data and the different software required to support them (Gallagher, 1995a). In its

application, visualization uses several means including searching, data mining and exploratory data analysis, analysing and modelling of data in order to extract abstract data into concrete visual representations and displays (Workshop, 1997).

Depending on the technical basis, the available data and the expressive terms, visualization can be categorized differently. As we have mentioned earlier, there is no unified definition of visualization as a whole, not to mention its various categories and subsets. Therefore, we will emphasize two subsets of visualization, which will be the focus of this thesis.

The first subset of visualization we will introduce is called information visualization, which is the study of, “The visual representation of large-scale collections of non-numerical information, such as files and lines of code in software systems, library and bibliographic databases, networks of relations on the internet and so forth” (Friendly. M, 2008). This method is mainly used to generate descriptions and summaries for large datasets. In order to represent the information in an appropriate format, this method is designed for the human eye to distinguish and read. Through visual impression and gross categorization through suitable colours and bandwidths, audiences can readily detect interesting pieces of information from an entire visualization (Sack, 2000). In comparison to other categorizations, this way of classifying the visualization subject mainly depends on whether it contains non-numerical information.

Another way to divide visualization is to focus on the subject, which is the focus of flow visualization. This method relies on fluid dynamics to generate visible patterns from a data flow. In reality, this method was frequently used in environmental studies to observe and tract the movement of water, air or smoke. Specifically speaking, it uses different intensities of colour to represent differences in speed and quality. Furthermore, the colour changes can also reflect changes over time. As the method matured, flow visualization was widely applied to describe the dynamics of information transactions. The information can be treated like water or air which could be represented by colours to describe the amount, speed and source. Additionally, the way that colours can represent the dynamic flow on the time axis will be useful in visualizing mass-authored text.

In fact, visualization has received attention from the general academia as, “A method for seeing the unseen” (McCormick et al., 1987a) in the 1980s. This definition provided a much wider space for social scientists to apply it at the beginning stages of application. Social scientists identify visualization methodology as including all visible images, such as photographs, videos, movies, computer graphics and so on. In social science, visualization acts as a methodology of representing phenomena, which is not like the method of displaying

procedure and results of research. The former one is more intuitionistic, and is visualized and obtained directly, whereas the latter entails computation, analysis, animation and requires accuracy and precision, which cannot be obtained directly.

A variety of creative and innovative visualization techniques have emerged more recently that enable social scientists to see and explore online phenomena, especially among sharing communities and internet social networks (Brandes, U. and Wagner, D, 2001). In contrast to providing software to scientific studies, many technicians and scientists found that more and more visualization software can provide social scientists with a comprehensive slew of analytical options, which is otherwise generally unavailable to more conventional social scientific studies. In addition, the increasing popularity of the internet, as we mentioned before, has also promoted the accessibility of many digital by-product data resources, allowing more and more social scientists to utilize an unprecedented range of visualization techniques and software. We argue that visualization methodology has many advantages which can encourage social scientists to improve their productivity in the internet age. Visualization has become an essential component in many scientific fields including biology and medical research. Biologists study cells and generate 3D confocal microscopy datasets, radiologists identify and quantify diseases from MRI and CT scans, and neuroscientists detect regional metabolic brain activity from PET and MEG scans. Through our introduction above, we argued that visualization enhances researchers' ability to study, diagnosis, and monitor and explain the complex systems in nature or the human body.

The application of visualization in the areas of science, geography and psychology has opened the way for the application of visualization in social science. (Orford et al., 1998) claimed the vital reason that both disciplines, "as mixed social science and science subjects" have advantages in terms of using computing technology when compared to many other subjects. This reports indicated that visualization has been considered a computing-skills-derived methodology, whose application and speed of adoption have a close association with certain disciplines, especially those with a scientific perspective. Furthermore, the unpopularity of using visualization in politics, economics and sociology has been explained by such disciplines having a strong and solid, "combination of traditional resistance to graphic techniques mixed with a relatively lower level of computer literacy" (Orford et al., 1998). Disciplines in social science that has an experience of using visualization generally have two features: closer links to the natural sciences, and a tradition of graphical representation (Orford et al., 1999).

5.1.3 Advantages of using visualization

Visualization has been discussed as a good opportunity to help research understand computation better (McCormick et al., 1987a), although many publications using visualization methods have appeared earlier. Based on such applications, visualization has been summarized as the tool to visualize the interactive process or active networks and is generally preferred over the comprehensiveness of large scale multi-dimensional information (Brandes and Wagner, 2004). With the benefit of visualization, it was utilized as a primary method by many scientific disciplines, such as molecular modelling, medical image, brain structure and function, geosciences, space exploration, astrophysics, computational fluid dynamics and finite element analysis (McCormick et al., 1987a). The reason that visualization has become widespread in scientific research and became one of its primary methods is because it offers many benefits to scientific development.

First, visualization provides a more effective way to present results. We explained that visualization is a method that incorporates human vision with computer technology. In doing so, computations can systematically synthesize the data, information and knowledge and effectively communicate to researchers, allowing the perception and identification of useful information by the human visual system. With the assistance of effective visual interfaces, we can discover the hidden characteristics, features, patterns and trends among large datasets by “quickly glancing” at such data. In our current informatics-society, changes and improvements in the interactions between human perception and data resources augments our ability to understand the world, and also positively affect our decision making process.

Second, wider application of visualization encourages many disciplines to develop visualization tools together, which creates an integrated set of portable tools (Fotheringham et al., 2000). The development and application of visualization tools exploded during this period. This is mainly due to two reasons. First, the early pioneers in the scientific community laid a firm foundation for the practices and theories in visualization, which will be conducive to the later development of this method. Additionally, the prevailing prevalence of the internet; rapid development in commerce and defence areas; urgent demand for internet data warehouses; and a huge curiosity in online interactions from the public have all provided further incentive for the further development of visualization. In this case, many institutes and organizations provide a platform for sharing data resource and visualization tools in specific communities. Additionally, much visualization software is designed for application in a range of different projects (Upson et al., 1989). All in all, these all provided a nurturing technical environment and supportive culture for visualization applications.

Third, visualization can help to improve the efficiency and productivity of the users through numerous expressive means, both microscopically and macroscopically. It has been illustrated

that visualization can enhance scientific productivity to a certain extent, as it can help scientists understand the problem faster and solve it more quickly (DeFanti et al., 1989). On the other hand, visualization, “enable users (e.g. in the commercial and the defence sectors) to get information fast, make sense of it and reach decisions in a relatively short time” (Gershon et al., 1998). For example, in the context of studying a large body of text, the advantage of enhancing effectiveness by visualization particularly stands out. Representing a bulk of text in a visual form should allow audiences to have an immediate impression and grasp the information represented, and furthermore they can locate any particular text piece that they are interested in. This is made possible because the means of visualization can express both an overall description as well as the exact detail. Functions such as browsing or zooming allow audiences to have an accurate and quicker understanding of textual documents in both its contents and history.

The other reason why visualization facilitates the comprehension process is that there are many visualization tools which are good at representing information with complicated dimensions and complex relationship (Gershon et al., 1998). The data that we see today are often multi-dimensional and with complex relationships. In dealing with multi-dimensional data structures, an effective approach for us is to use nodes to represent entities and lines to represent connections. However, if the data structure has additional layers of complexity, such as nodes from different hierarchies also having interconnections, simple linking may cause confusion in visualization. Multidimensional visualization can reconcile the limitations of this single dimensional graph through changing the colour, time point, height and other coordinates in the 3-D visualization to represent complicated relationships.

However, it is worth noting that the advantages of visualization reach far beyond to what we have mentioned so far. From a non-technical point of view, visualization has increased communication across disciplines; broken the bottle neck of communication between science and social science; and allowed the public to better understand research discoveries. These contributions have also helped the rapid and widespread application of visualization. This popularity has appeared not only in scientific areas, but with the development and maturation of internet technologies, it is becoming more widely used in the social sciences (McCormick et al., 1987a).

In such a situation, our research aims to discover a series of questions about the application of visualization in social science, such as whether visualization is useful; does it have any irreplaceable function for social science; how to apply visualization to correspond to specific social scientific topics, what are the steps to begin visualization in social science research; is that similar to the process in the natural sciences; what crucial problems or difficulties are

there for social scientists to apply visualization, and so on. In this thesis, all of the above questions will be explored by an experimental study applying visualization and other methods based on digital by-product data, and the Wikipedia datasets.

5.2 Methods

In this chapter, we will mainly use visualization to address how multi-authoring articles are established by mass participants after numerous edits. An investigation of the multi-authoring article requires us to first identify the contributions from different authors and secondly to clarify their editing behaviour, which could be achieved using visualization. We could use existing visualization tools to process digital by-product data in order to explore how people edit a particular webpage, how such edits affect the content of this article, and furthermore how debate and argument between participants is reflected in the article content. Normally, such visualizing tools can describe collaborative processes including conflicts and cooperation. They also enable us to assemble information from different domains to offer a sense of the whole. Moreover, such tools offer the opportunity for social scientists to discover and analyse a large amount of data.

The tools for visualizing multi-author articles can represent a clear picture of how articles are created, improved, changed, deleted and reverted along with the timeline and by particular editors. In order to do so, visualization tools colour-code the edits by different participants. This tool also uses a sentence-by-sentence search engine to define how old every sentence is. Based on that, the timeline showing the article development are drawn according to the structure of the article. The visualization of a multi-authored article pictures the sequences of editing through a colour coding scheme, which identifies the link between participant, timestamp and content. Such tools are used in multi-author collaboration systems to identify the dynamics and interactions among the collaborating authors by visualizing the editing history. This analysis is based on the digital by-product database of fully-protected articles with all their respective edit records.

History Flow is open source software created by the IBM Research group of User Experience, which is designed to visualize dynamic and evolving multi-authored documents. This tool was originally produced to explore users' behaviours in collaborating on content, especially on a wiki platform. Technically, this tool is coding content to different authors and their respective contribution in a Java environment. This tool has been shared on IBM's Research Web for free download and use since its creation in 2004. This tool has comprehensive functions, is user-friendly, and provides a mature analysis platform to support the database we obtained from Wiki technology. However, at the same time, this small software that runs in Java relies on distant server to handle data, download partial data to analyse. The data

handling process is often disrupted due to unstable software. Therefore when we used this tool, we made some changes to it in order to improve its performance so that it can better visualize specific editing behaviour data we collected from Wikipedia.

We will use History Flow to focus on databases including meta-data of participants, the recorded time of every edit and text content, all of which Wikipedia digital by-product data can offer. With the aid of such visualization tools, readers are able to understand how articles are produced, changed and improved by mass collaboration. Moreover, these pictures will show most of the editing conflicts present on Wikipedia, including deletion, reversion, and vandalism. In order to realize our research plan, we aim to find an existing visualization tool to investigate the editing process of fully-protected articles.

We use the example of an article edited by three authors to demonstrate how the History Flow tool pictures the collaborative process in an individual article. The graphs below illustrate how History Flow works with a Wikipedia dataset. Figure 5-1 shows a screenshot of the history page in Wikipedia, which records the history of all edits relevant to the particular page. Using History Flow to analyse the history records in Figure 5-2, we are also able to obtain a picture to visualize the editing process. If three people have edited a particular article at different times and each person's edit is saved as one version in the history record, as soon as the edit is saved, History Flow will automatically assign different colours to each editor. This is shown in Figure 5-2: Mary in red, Suzanne in blue and Martin in green.

- 
- [\(cur | prev\)](#)   [09:27, 27 April 2002](#) [Suzanne](#) ([talk](#) | [contribs](#)) (6,513 bytes) (prominent link to PHP Script)
 - [\(cur | prev\)](#)   [08:05, 17 April 2008](#) [Martin](#) ([talk](#) | [contribs](#)) (6,481 bytes) (edit any page *with few exceptions*)
 - [\(cur | prev\)](#)   [18:07, 26 January 2008](#) [Suzanne](#) ([talk](#) | [contribs](#)) m (6,358 bytes)
 - [\(cur | prev\)](#)   [15:28, 26 January 2008](#) [Mary](#) ([talk](#) | [contribs](#)) m (5,340 bytes)

Figure 5-1 Screenshot of the history page related to a Wikipedia article

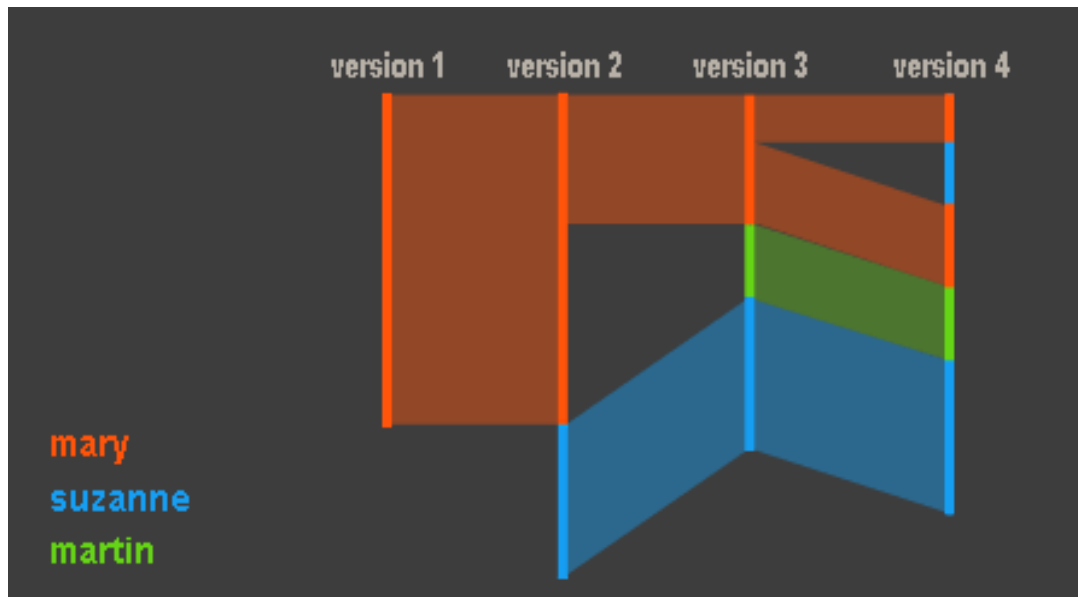


Figure 5-2 Visualizing result of the editing history by History Flow tool

[Originally in colour]

In Figure 5-2, the bold red line represents the first edit by Mary. The length of this red line shows the amount of text in bytes that Mary has written. Suzanne then added more content to the end of Mary's edit. The fact that the start of the blue line follows the end of the red line indicates the sequential edits by Mary and then Suzanne. All of Suzanne's additions and Mary's original work have been saved by Suzanne as Version 2. In Version 3, Martin has deleted some text by Mary (red line became shorter than that in Version 2) and added text. Martin did not edit Suzanne's work as the length of the blue line remained the same. In Version 4 of the edit, Suzanne came back to add a short link in the middle of Mary's text, but did not change the rest of Mary's work.

Moreover, the texts that do not have corresponding text in the previous version are not connected which results in a gap in the visualization result. Such a gap could represent either deletions or insertions. Examples include that in Version 3, Martin (Schwarzer et al.) deleted Mary's work (red) and in Version 4, Suzanne (blue) inserted more content into Mary's (red) work.

5.3 Visualization of mass collaboration

As we discussed above, the visualization of mass-authored collaboration in individual articles can help us understand a complicated working mechanism in Wikipedia based on a large sample of digital by-product data. In this section, we explain how visualization can display mass-authoring collaboration by using one normal article as an example. Then we discuss the

baseline pattern based on random selected open-edit articles and featured articles. The third section summarizes the occurrence and frequency of vandalism in article edits. Through this description, we further argue that infrequent vandalising activities should not raise doubts in the collaboration model of Wikipedia, and that mass collaboration in a free edit environment is fully viable and has potential for development.

5.3.1 How visualization displays individual articles

Each paragraph in an individual article can be linked to the respective participant and a certain time point. Such a connection is the basis of visualization. The following graph shows a record of edit behaviour that we extracted from some downloaded data. The graph records two levels of data structure; the first level being the name, length, IP address and content of the article; and the second level bears information of the time, participant and the content of each edit. As metadata, the link between the two levels of data structure provides us with an additional layer of important information, for example where each edit occurred in respect to the timeline of the articles development.

In order to illustrate how we used the History Flow tool to accomplish the visualization of one article, we used the following one as an example.

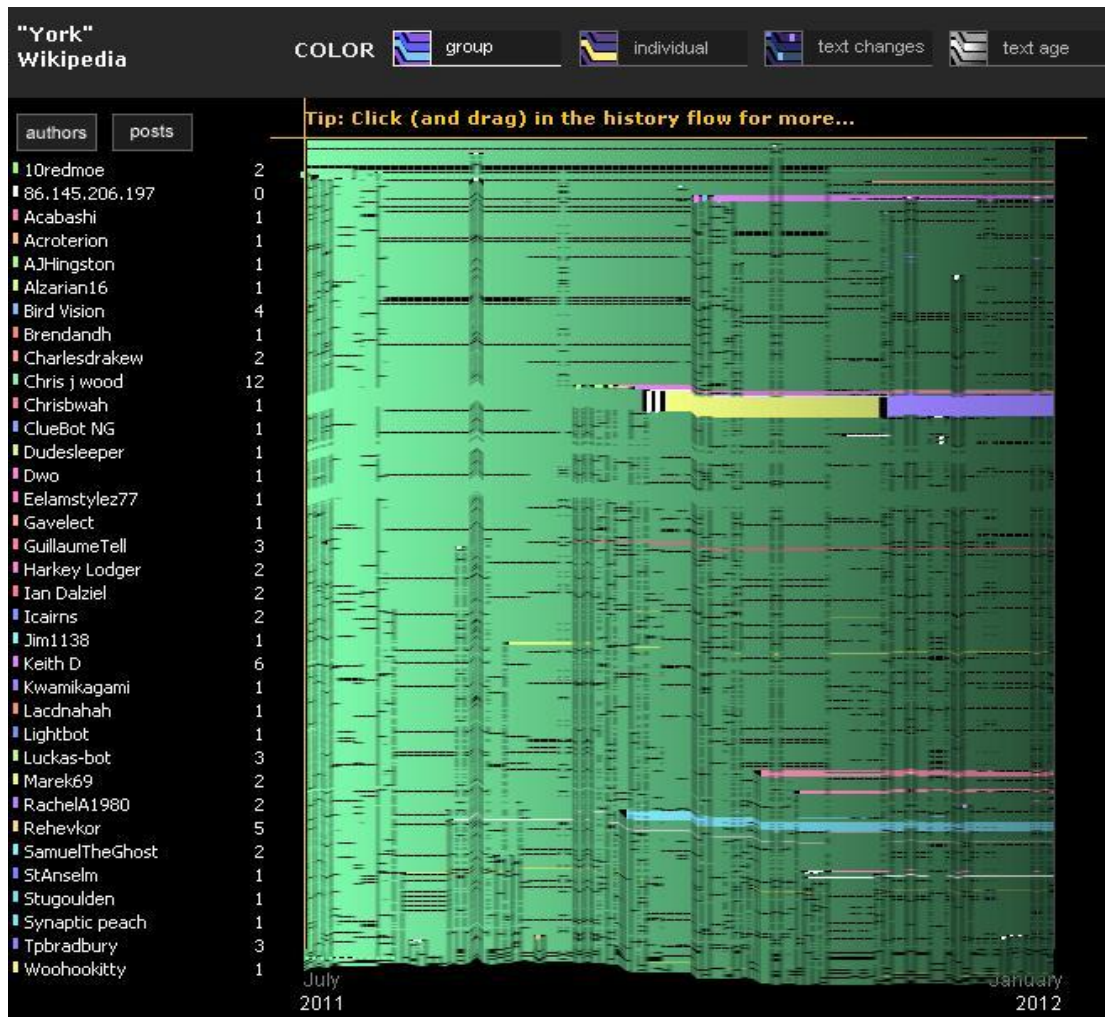


Figure 5-3 Visualizing the article on “York”

[Originally in colour]

Figure 5-3 above shows the visualization of edit flow in the article on York. This article was created on 18 November 2001; the total number of edits is 3204 up until 16th January 2012—the date we collected original data for visualization. There were 924 IP users (non-registered users) who edited this article along with 1209 registered users. It is too difficult to display the entire evaluation of this article in one graph by visualizing 3204 edits produced by 1209 unique registered users and 924 IP users, which is 1.16 days per edit and 2.65 edits per user.

Therefore, with the assistance of the History Flow tool, we can visualize 100 edits of a certain article at a particular time. In Figure 5-3, we visualized the 100 most recent edits since 16th January 2012 and the change and impact on the whole article. From the visualization in this graph, we are able to see the earliest edit was made in July 2011.

The article title “York” and the data source “Wikipedia” are shown in the upper left corner. The user names of each participating editor, the respective colour to represent each participant and the statistics of edits are listed on the left. We observe that Chris J. Wood, an individual participant, has edited this article 12 times in total, as represented in green. Additionally, edits in green are the dominant colour, implicating that this user has contributed most in this article. As we said before, this visualization is able to display all information flow in the process of editing an article including the length of content. Therefore, the majority of the visualization in green means the participant with green colour has contributed a majority of content in quantitative terms. On the other hand, the fact that his/her contribution has been saved and followed by many other participants (the green colour did not change and was not removed in the following edits) suggests the green colour contribution established the baseline and direction of content.

The y-axis shows the entire structure of the article, the highest point representing the first word and the lowest point the last. This visualization approach helps us to exhibit the location of each participant’s edit in the article and identify the impact on the structure of the article. The x-axis shows the chronological order of the 100 edits from left to right. The 101 vertical white lines divide the main area of the chart into 100 columns which represent those 100 edits. This description not only reveals the article content and structure of each edit as individual data, but also collectively shows the dynamic change of the article by placing all edits together.

So far we have introduced how to visualize an individual article and in doing so what information we can present, as well as what patterns and trends we can find. It is clear that using visualization, scholars can display complicated information through a simple diagram. The following sections will discuss some patterns we found from these visualizations.

5.3.2 *Baseline pattern*

In Wikipedia, articles can be categorized according to whether they are open to edit, semi-protected, or fully-protected. Semi-protected and fully-protected articles are two types of articles which have restricted editing rights. Semi-protected articles are allowed to be edited by registered users and fully-protected articles can only be edited by administrators. The reason for such restrictions will be discussed in the next chapter. The rest of the articles follow Wikipedia’s open-edit policy and can be edited by anyone without authority and monitoring. In Wikipedia, more than 90% of the total articles are open. With regard to the topic of various restrictions to article edits, we will focus on that in the next chapter.

Among these open-edit articles, a number of them are regarded as high quality by consensus, those which are relatively comprehensive in content and clear in structure, and are termed ‘featured articles’. Such articles are usually selected by secret ballot among the participant community. In other words, articles representing the fruitful product of Wikipedia’s open and freed collaborative policy, and featured articles are amongst its best quality. In contrast, semi-protected and fully-protected articles are products that may have some faults. Therefore, visualizing open-edit articles which follows the standard editing policy can help us to understand how massive co-authoring works under the open and free editing policy of Wikipedia.

We selected 330 open-edit articles from the English Wikipedia, including 230 open-edit normal articles and 100 open-edited featured articles. All such subjective articles have been randomly selected from Wikipedia’s article list with related complete edit records. We expect to generate a certain pattern of mass collaboration from visualizing such 330 open-edit articles, which theoretically can be examined by all articles in Wikipedia based on digital by-product data.

We extracted a “baseline pattern” in this part of the research which represents the editing process of a single article and will show that such a pattern follows the power-law. The previous two chapters proved the existence of a power relation in Wikipedia from a grand perspective; from this chapter we will microscopically demonstrate the existence of a power relation through a baseline pattern. In other words, in every article of Wikipedia, the majority of participants only edit a small portion of the content while the primary content is contributed by a few participants. This is demonstrated by examining the visible history flow pictures generated using the History Flow tool.

The “baseline pattern” suggests that the person who initially created the article has established a baseline for this article. In other words, the original creator of the article lay out the initial directions, structures and primary topics for each article. All other participants follow this “baseline” to continue the growth and development of the article. It means that articles will demonstrate a baseline as soon as it was created and then the rest of the edits could follow it during development. By investigating the edit history of open-edit articles, we found that the collaborative model in Wikipedia was a “baseline pattern”, which was proven by 93% cases out of the total selected open-edit articles. It should be noted that in the visualized cases, the creator who established the baseline can be either one person or a small group of people.

For instance, we take the article on the ‘European Union’ as an example of baseline pattern. Figure 5-4 shows the edit history of that article. The 100 edits are from December 2004 to January, 2005. From this figure we found that the majority of the content is created by the

participant represented in pink, whereas the other colours representing other participants are sparse. The History Flow tool allows us to easily detect this ‘power distribution’ pattern in a single article.

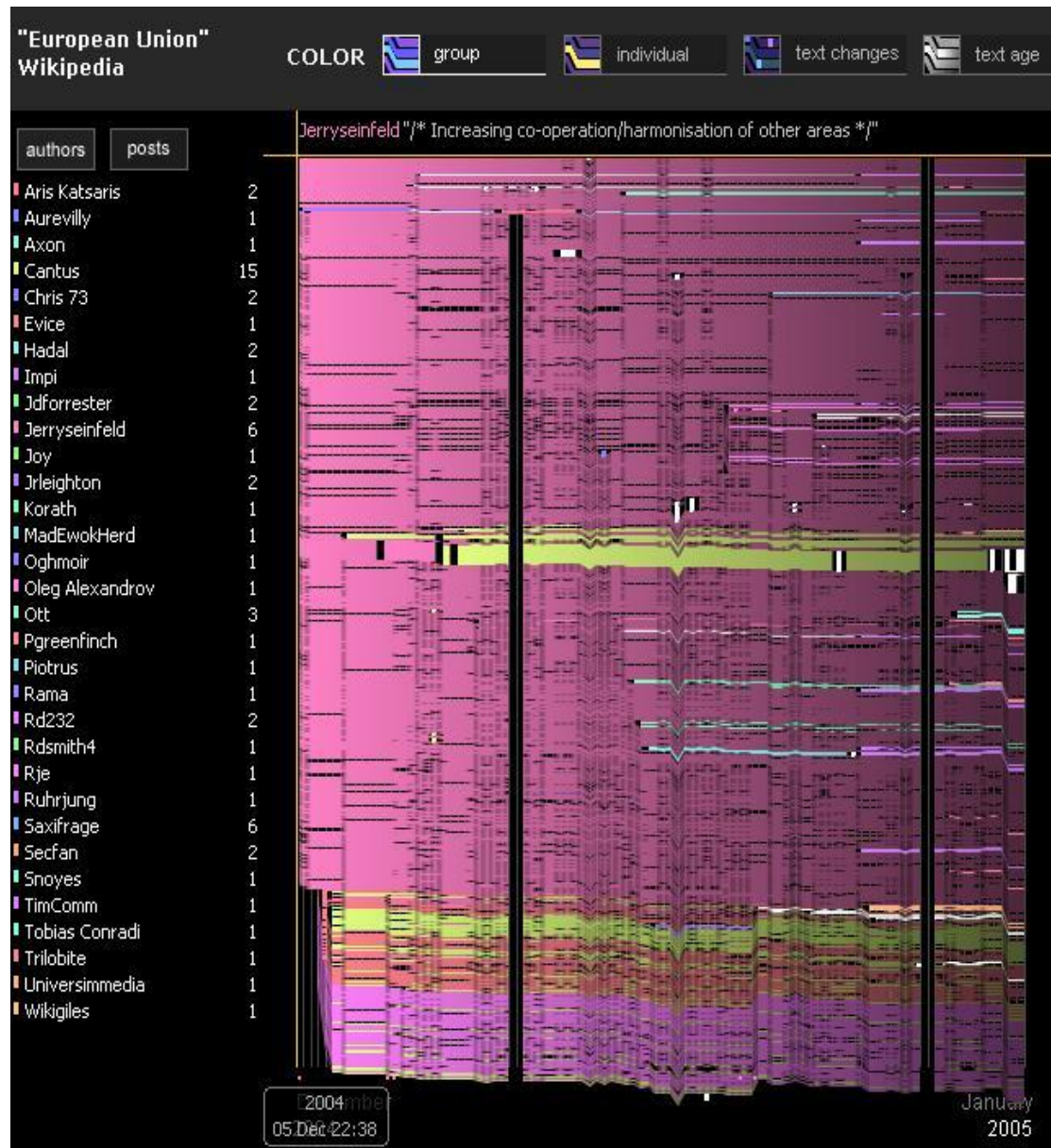


Figure 5-4 Visualizing the article on the “European Union”

[Originally in colour]

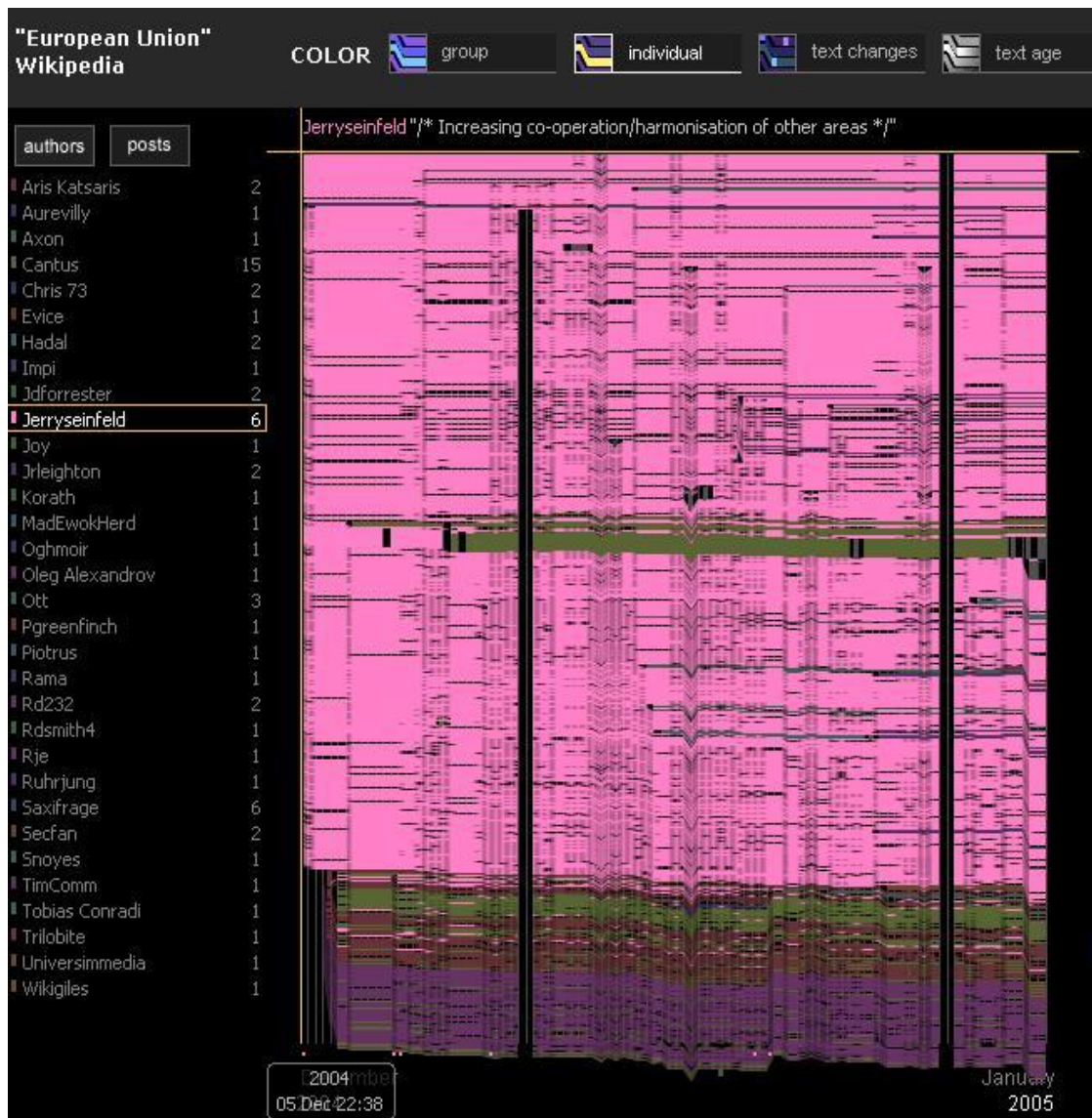


Figure 5-5 Highlighting the baseline pattern in visualization of the article “European Union”

[Originally in colour]

Specifically, in Figure 5-5, we are able to see that the participant named as ‘Jerryseinfeld’ (highlighted) created the majority of content in this article. Around 80% of content in this article has been produced by Jerryseinfeld by only six edits. It is more important to point out that the content of Jerryseinfeld’s edit not only comprise the main body of the article, it has also been preserved over time. This, to a certain extent proves that the process of producing articles in Wikipedia is a collaborative system with a single goal of creating an online encyclopaedia to share knowledge.

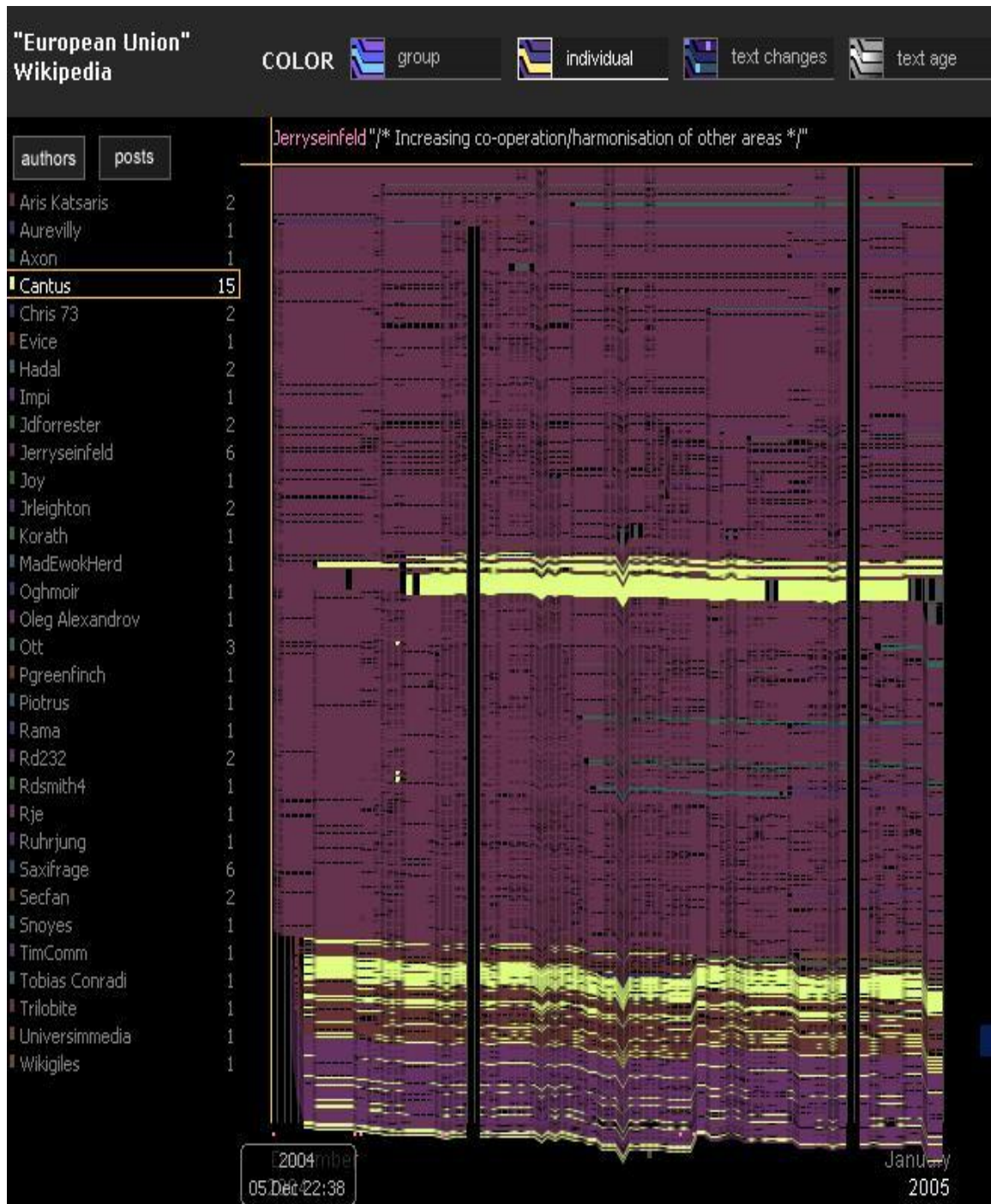


Figure 5-6 Highlighting contribution of Cantus A in visualization of the article “European Union”

[Originally in colour]

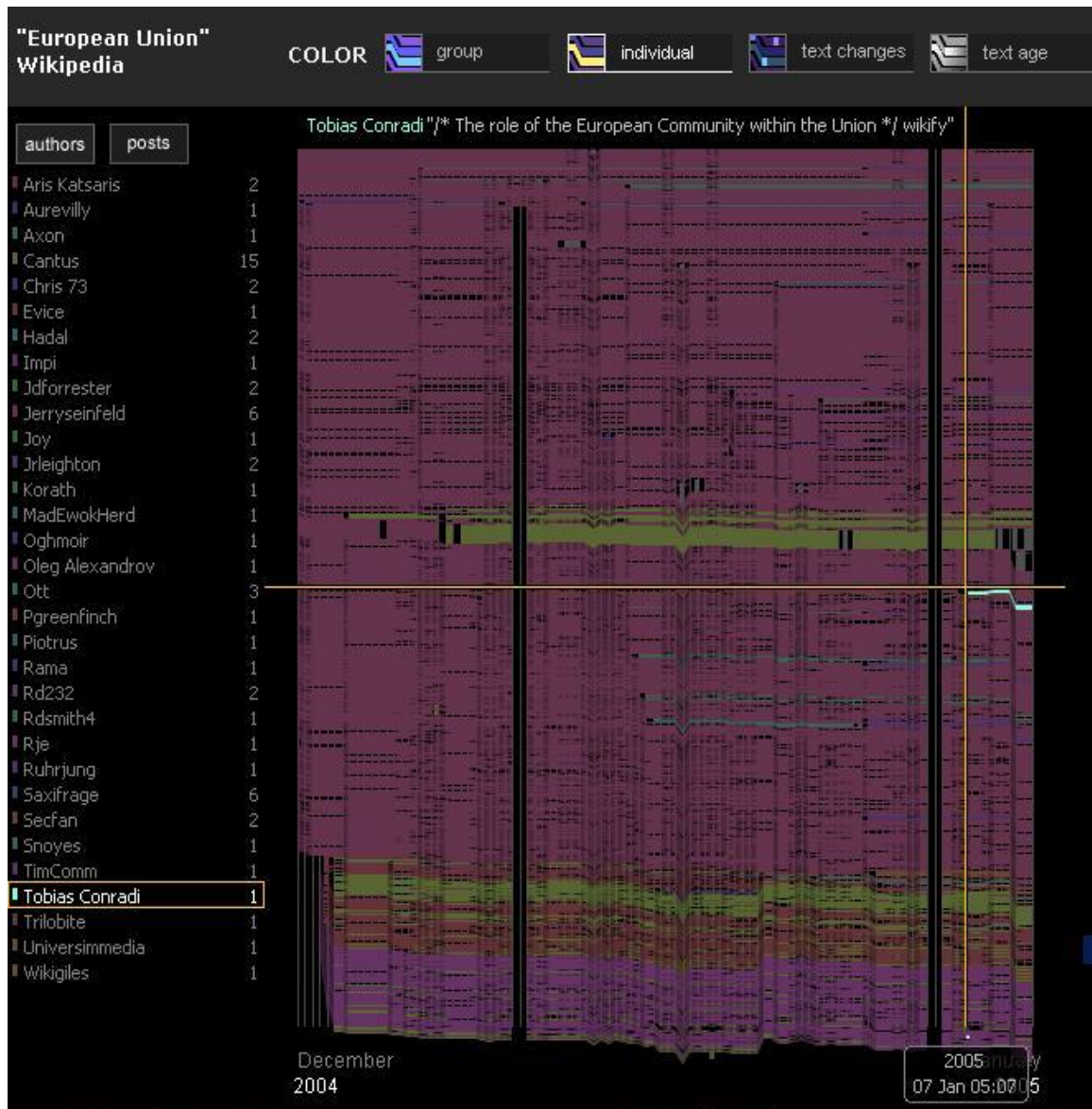


Figure 5-7 Highlighting the contribution of Tobias Conradi in visualization of the article “European Union”

[Originally in colour]

Interestingly, in Figure 5-6, we see the participant “Cantus” has edited 15 times during this period but his/her edits affects less than 10% of the content. This suggests that there is no positive correlation between the contribution to the article and the number of edits a certain participants has. In the fourth figure, ‘Tobias Conradi’ (highlighted in blue) is another participant who only edited once during this period. It is obvious that his/her edit is created on 07 January 2005 and also survives later edits. From the illustrated graphs, we show that

baseline is a phenomenon that is prevalent in Wikipedia articles, where several core authors are responsible for the majority of the content, and define the basis and direction of the article

In addition to the “baseline” creators, we also found many intervening episodes of edits which are represented in the history flow graphs by different colours. These segments represent hundreds of minor corrections of the baseline contributions in order to develop and polish the article. From the visualization we can clearly see the inheritance between edits, i.e. subsequent editors will scrutinise the previous edits and preserve valuable ones. This relationship is one of the expressive terms of collaboration.

In order to further elaborate that baseline creation is a common pattern in most mass-author articles, we examine the existence of a baseline in different types of article. First, following the examples above, we divided the different edit periods of the article on ‘York’ into twelve visualization units, and still found the baseline pattern present to varying degrees. Secondly, we visualized randomly selected 100 articles from featured articles and 100 common articles. All of these articles contain baseline pattern in their visualization.

We argue that the innovation of mass collaboration lies in the fact it provides a platform for minor corrections which could result in a major contribution, and such a collaborative mode differs from the traditional multi-authorship format of collaboration. It is worth noting that mass collaboration gives more opportunities to make minor contributions for people who are interested in the topic and willing to contribute but are limited in their knowledge and participation time. In other words, mass collaboration draws minor contributions from a collectively wise crowd and the phenomenon could be described by the aphorism that ‘a single spark sets the prairie fire’. In terms of knowledge contribution, if everyone can contribute a single spark of their knowledge or information, the result will be a worldwide fire.

The History Flow tool only visualizes the editing process by individual participants in mass collaboration. The entire analysis process will be exemplified using a selected fully-protected article. In this way, digital by-product data could generate visible pictures, from which we can detect the elementary collaboration and conflicts in Wikipedia during the edit process. However, all these visualizations place more focus on the description of the edit process rather than the investigation of conflicts and consensus in great detail. Different visualizing tools can provide alternative possibilities for describing collaboration on Wikipedia.

5.3.3 New way to collaborate

The mass-author collaboration on Wikipedia not only displays a baseline pattern but also suggests an innovation of collaboration in which conflict and collaboration may exist together. This collaboration model is in direct contrast to that of traditional multiple author

collaboration. The traditional model does not tolerate conflicts, as they inflict discordance in cooperation and therefore lead to questions regarding the quality of the product. However, through visualization we discovered that on Wikipedia, conflict and collaboration often co-exist and such co-existence does not negatively influence the quality of the articles.



Figure 5-8 Blocking and featuring in the same article

For instance, as shown in Figure 5-8 above, the article “7 World Trade Centre” describes the famous building in New York City, which has been a featured article on Wikipedia. As we introduced before, featured articles are recognized as high-standard and well-structured articles by consensus. As a featured article, the quality of the article is marked out by the star on the upper-right corner. However at the same time, there is a “lock” symbol next to the star, suggesting that this article is protected from free edit as it is the subject of much debate. As the level of debate varies and the risk of potential damage is thought to differ, different articles receive different levels of protection. In this instance, this article is semi-protected, meaning that only registered participants who have edited at least once can edit this article, and any new participants or IP users without one edit are not allowed to change this article. This kind of protection has a time limit, such as to protect the quality of the article without encroaching on the free to edit characteristics of Wikipedia. Since we will focus on protected articles in the next chapter, we will not elaborate further here. However, as a featured article that also attracts vandalism and large scale editing arguments, it exemplifies the coexistence of conflicts and collaboration in Wikipedia. However as the whole edit process is dynamic, we cannot know the time of occurrences of conflict and collaboration in relation to each other.

Based on the descriptions of the baseline pattern, we can clearly define the expression of different actions and the relationship between them through visualization. For the next step, we hope to observe the relationship between conflicts and collaboration by visualizing the editing records of individual articles. Through randomized sampling of 100 featured articles,

we discover that all visualization contains more than 10 participants in 100 edits. Their contribution to the content of articles varies. For instance, the article on DNA (Figure 5-9), a featured article, contains 100 edits from 14th November 2005 to 17th December 2005. In which 24 participants contributed to the content. Among the participants, the one with the most edits has six edits and the least has none (i.e. their one edit was deleted). Although most participants only edited once, they still contributed to the article both in structure and in content.

Visualization also demonstrated to us another interesting phenomenon – which is massive deletion. Using the information flow which identifies different authors by different colour, the content of the article is marked by such respective colours. However, in the visualization of articles, we found there are some blank bars which suggest the article did not contain any content in that period. As the visualization background is black, all such blank areas as shown as black. We term these black bars ‘massive deletion’, which is caused by participant deleting all content without adding anything. So for a period of time, this article will have had no content. Such an instance is a stereotypical damage phenomenon and apparently damages the quality of articles. However, we found almost every featured article contain a period of such massive deletion. This also proves that conflict and collaboration can coexist on Wikipedia.

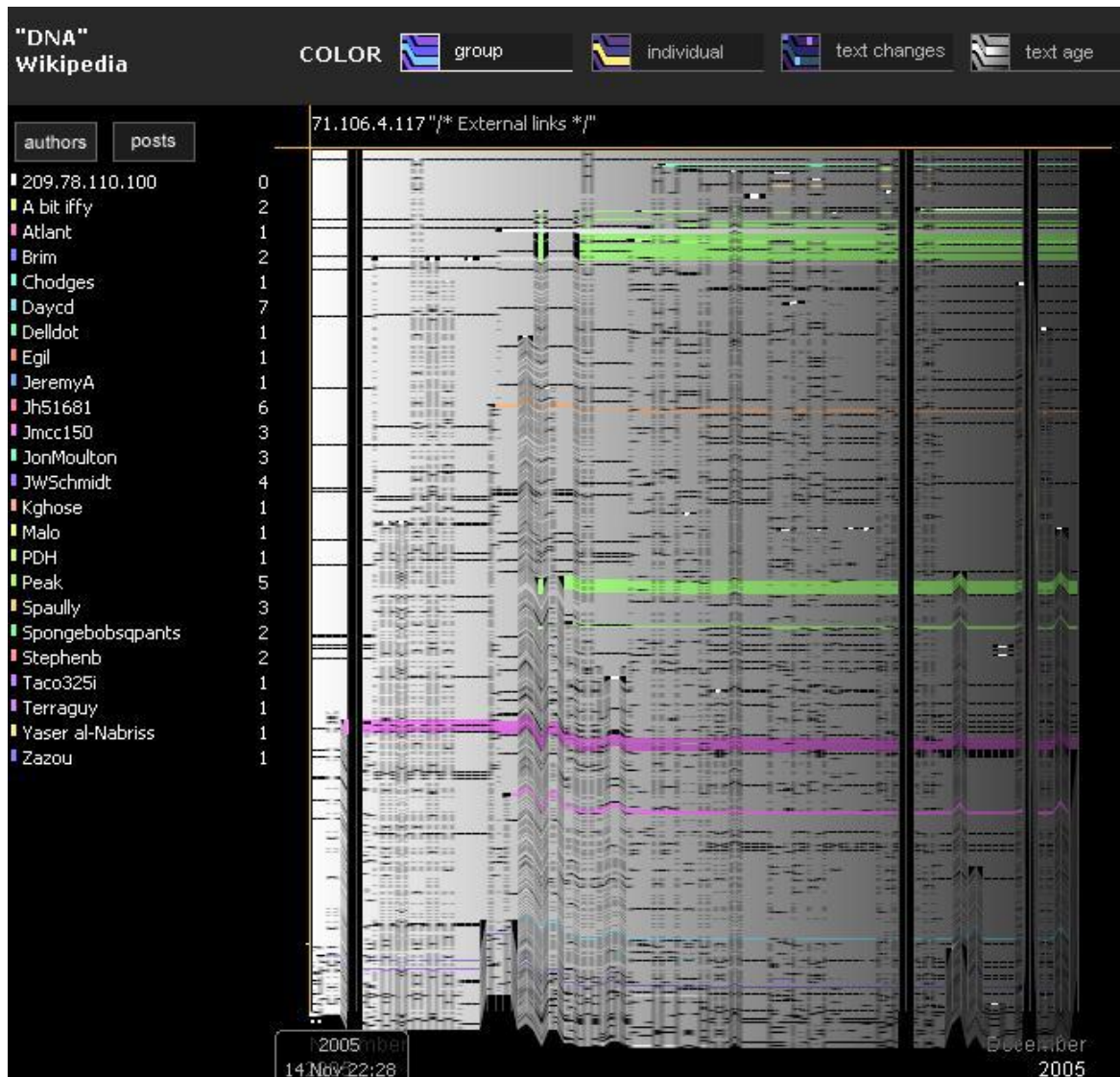


Figure 5-9 Visualizing the massive deletion in the featured article on “DNA”

[Originally in colour]

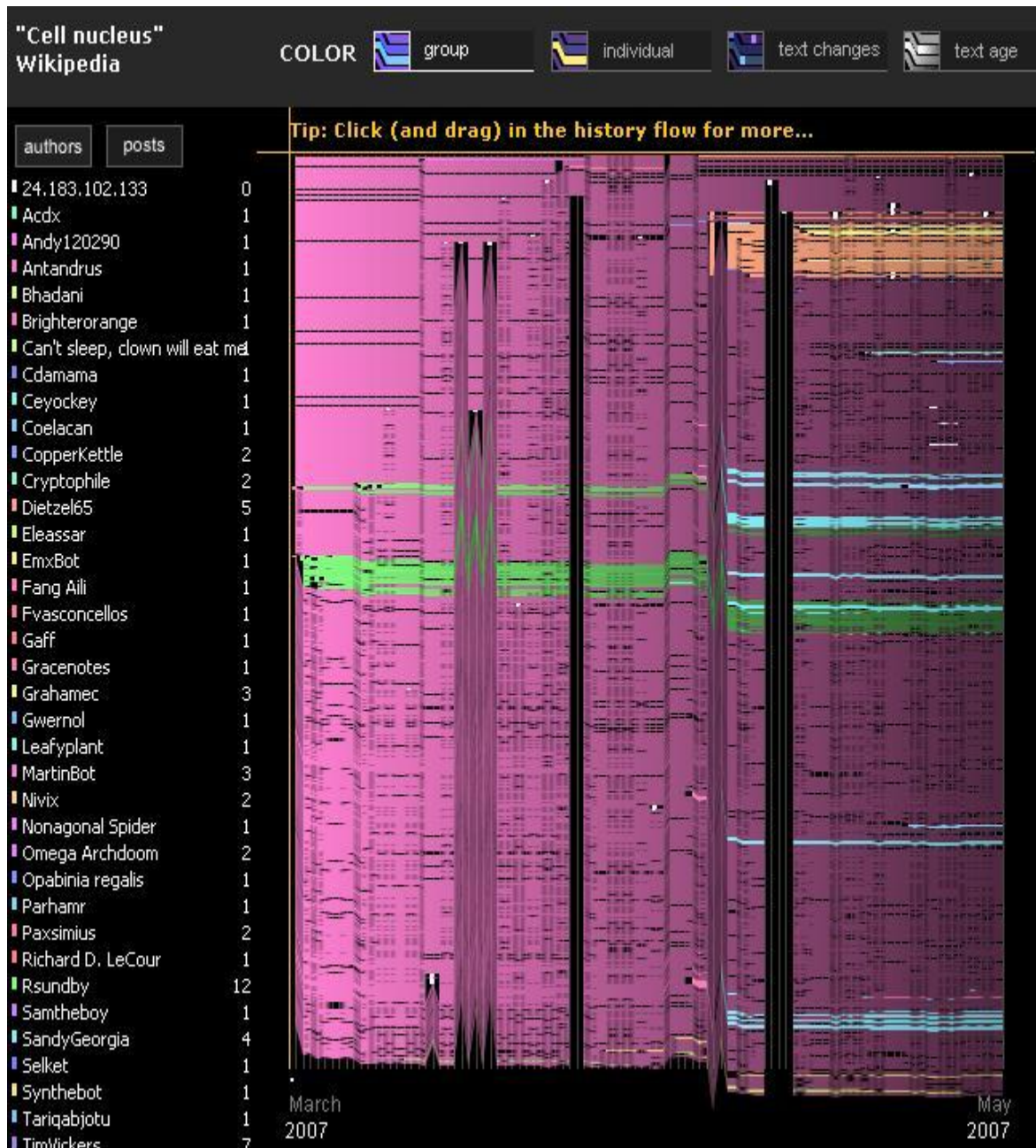


Figure 5-10 Visualizing the massive deletion in the featured article on “Cell nucleus”

[Originally in colour]

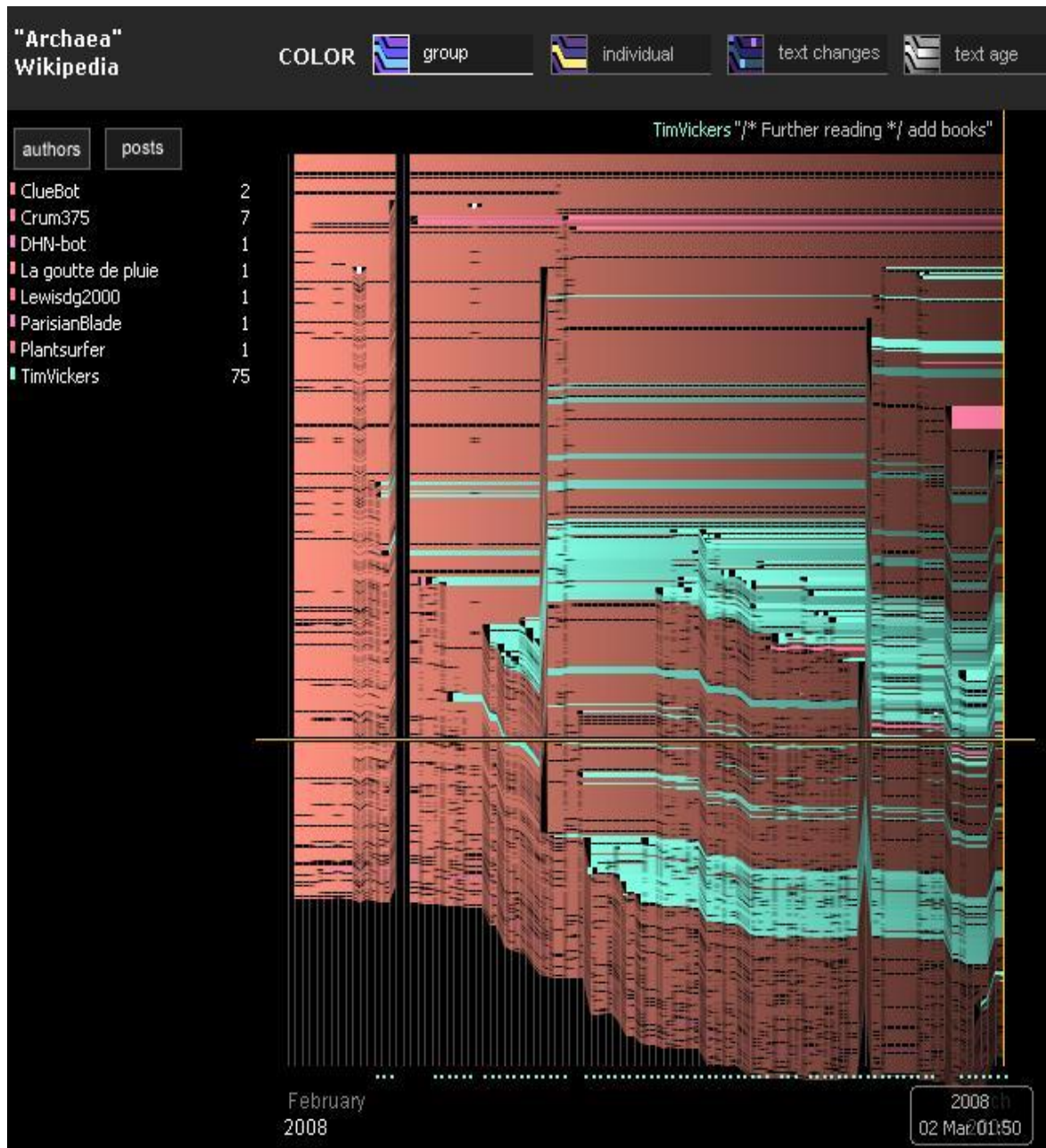


Figure 5-11 Visualizing the massive deletion in the featured article on ‘Archaea’

[Originally in colour]

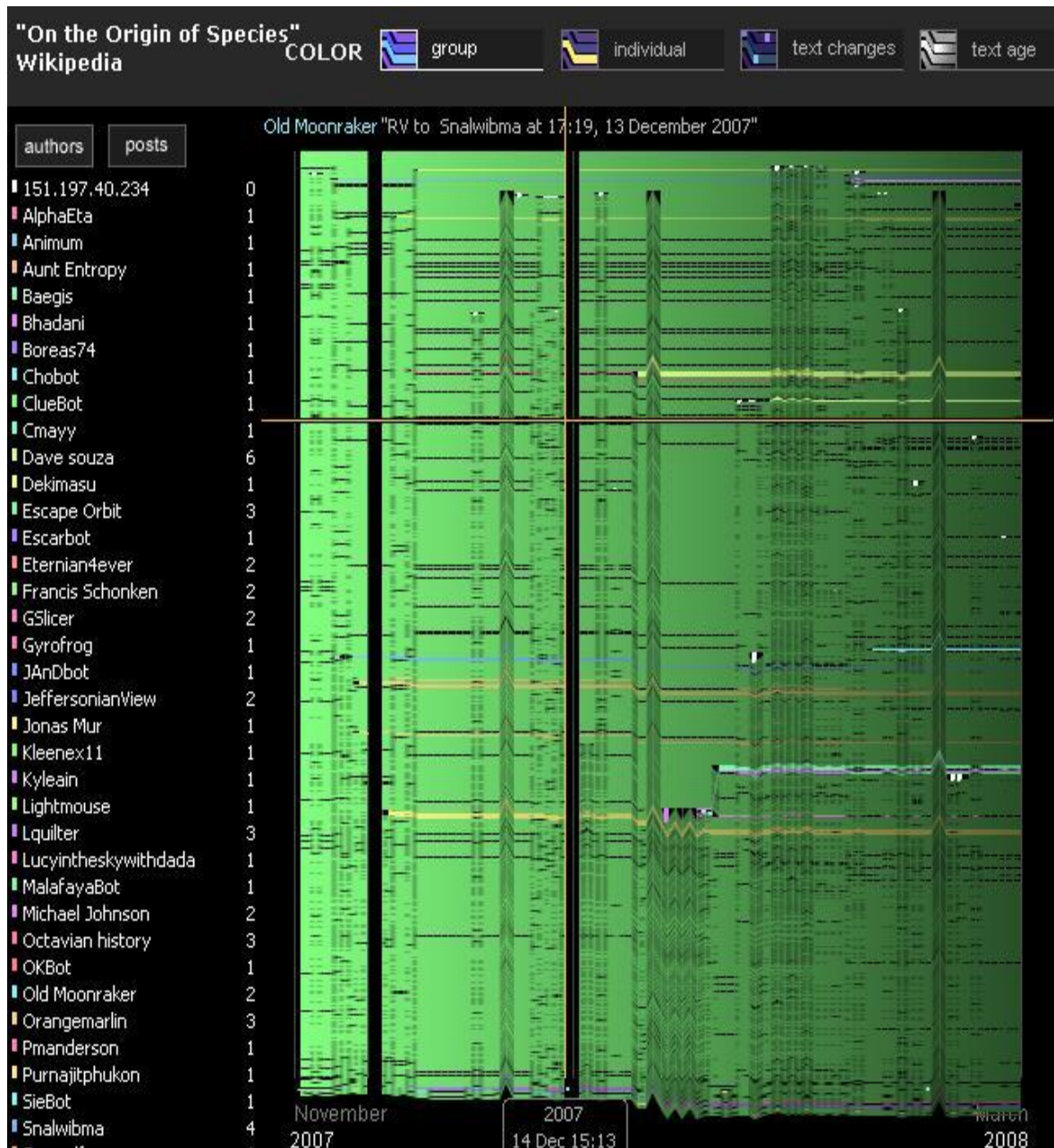


Figure 5-12 Visualizing the massive deletion in the featured article “On the origin of species”

[Originally in colour]

5.4 Conclusion

In this chapter, we used visualization as a new research and expressive approach to process digital by-product data. In order to differentiate from the analysis of mass collaboration in the previous two chapters, this chapter addresses a similar question but from a microscopic view by visualizing the mass-authoring process in individual articles. We have made two

significant achievements in this chapter: through visualizing regular articles of Wikipedia, we discovered two apparent characteristics of mass-authorship collaboration: baseline pattern and the coexistence of conflicts and collaboration; furthermore, we gained first-hand experience of using visualization and demonstrated its applicability in dealing with digital by-product data.

The baseline pattern was extracted from the analysis of randomly selected 300 open-edit common articles and 100 featured articles. This pattern corroborated the notion that the “majority of participants contribute a minority of content and a minority of participants contribute a majority of content” as mentioned in the last two chapters. For individual articles, the contribution from one person or a small group of people usually sets the cornerstone for the article’s structure and content – the baseline work. Most of the other participants only contribute small-scale changes. Such a collaboration model is defined as a “baseline pattern” in this thesis.

The coexistence of conflict and collaboration is generated as another feature of mass-authoring collaboration by visualizing individual featured articles. Wikipedia has been scrutinised in its mass-author collaboration because the entire editing platform is open to the public. One of the main concerns is the potential damage done to the quality of articles –this is addressed by a new model of collaboration in this chapter. Through visualization, we directly observe that featured articles which are regarded as high quality are also susceptible to damage and potential vandalism. However, because such vandalizing activities occur in the editing process, perhaps visualization showed us a better working model of allowing conflict and collaboration. While collaboration puts constraints on the development and influence of conflicts, the free and open collaborative platform allows conflicts to occur.

By experimentation we proved that visualization has unique features when dealing with large amounts of data. The ability to effectively process and present a huge body of data as well as the intricate connections between data using a multidimensional approach allows users to immediately identify relevant information; discover the uniqueness of certain data, and the similarities and dissimilarities of the information. It is this characteristic that enabled us to recognize and define the baseline pattern. Through different shapes, colours and dynamic changes, visualization can present many layers of information; this inherently makes it more capable of revealing patterns compared to textual descriptions or mathematical models. Next, based on our preliminary results, we believe that visualization can be an important research methodology with great potential particularly for descriptive research. Specifically, under the influence of more complex participation means and multidimensional communication manners, the virtual society nourishes an environment for data to become more complex and

diverse. Visualization can simplify the expression of complicated information, facilitate the user to discover phenomenon and identify patterns. This feature makes such a research methodology more straightforward in comparison to traditional statistical methodologies, and more suitable for dealing with larger-scale digital by-product data.

We encourage more social scientists to apply visualization in their studies following our demonstration of its practicality. We used the visualization tool as developed and provided for free to the public - History Flow. In fact, after we decided on our research direction, we actually found many other sharing tools from different internet platforms. The fact that scholars can easily use open source software to visualize digital by-product data should encourage more social scientists to do so.

In the next chapter, we will use visualization to explore a new type of article – those that are fully-protected. As the highest risk articles, visualizing fully-protected articles can answer what function administrators carry out in protecting the quality of articles and maintaining collaboration.

Chapter 6

Visualizing and assessing the administration in conflict-protection situations: A case study of fully-protected Wikipedia articles

The previous empirical chapter examined the possibility of domination by particular groups of the collaborative participation of Wikipedia. We discovered that the most influential participants in Wikipedia were those, who were without the privileged power of administrators, but had a large number of recorded edits. By comparing the quantity of edits made by different types of participants, we concluded that Wikipedia is operating a semi-democratic system, the content of which is dominated by a population of elites without the authority vested in administrators to change content. Such a conclusion was drawn because a small group of high-edit-record elites contributed to more than half of the content on Wikipedia since 2006, and there was less influence from administrators with regards to the quantity of edits. In fact, we discovered that the group of administrators did not have a considerable percentage of total edits so as to dominate the direction of the content. Thus, we demonstrated that Wikipedia is mainly produced by a small group of non-administration participants with high levels of enthusiasm and a high edit-record.

We thus argue that understanding the process of conflict and its resolution can provide important insights into the collaborative model of Wikipedia. As an important part of the Wikipedia system, administrators play various roles in editing content, resolving conflicts, and guiding new participants. In order to incorporate two aspects of our research question,

especially to understand the role of administration in resolving editing conflicts and maintaining participation, this chapter starts with using fully-protected articles as the subject for our observation. This is primarily because fully-protected articles are a major manifestation of the administration system. As we introduced in chapter two, only administrators have the technical authorization to selectively block or unblock articles and users. Secondly, it is worth noting that fully-protected articles only reach the status of full-protection following unresolved debates among participants. Such debates may lead to the onset of editing wars, where the content of the article is constantly changed, and presumably such occurrences could diminish the quality of these articles. As discussed above, fully-protected articles are the explicit manifestation of the technical privileges of administrators, and they are blocked by administrators to prevent the article from potential damage. Such judgment comes from the repeated changes and continuing reversions in a particular article. Thus observing the fully-protected article as an object of study allows us to better measure the function of administrators and explore the onset and resolution of debates during the editing process. This chapter briefly reviews what Wikipedia's administration system is, how administrators work and their principle of resolving editing conflicts, which has been introduced in chapter two when fully-protected articles were introduced along with its policy. Secondly, the editing process in fully-protected articles is visualized by the History Flow tool to allow us to explore conflicts and coordination that characterise mass collaboration on Wikipedia. In this section, we provide descriptive conclusions to identify different editing behaviours visually and to discuss why they cause full-protection. Finally, we analyse the database of fully-protected articles using traditional correlation statistics, which together with visualization allows us to represent the conflicts and coordination in Wikipedia and to assess the role of administrators within this mass collaboration system. From a methodological perspective, this chapter aims to evaluate the application of visualization in the social research process. By doing so, this chapter aims to explore particular damaging behaviour from fully-protected articles in order to examine the function of administrators. Visualization was used as one of many possible methods to answer such a question. Therefore, this chapter answers another question – whether visualization can be readily used in social science inquiries – and assesses the quality of the answer using this particular method. Understanding these findings will help social scientists to know whether, from a methodological perspective, visualization could be applied to research using digital by-product data from a methodological perspective.

6.1 Introduction

The open mode of participation in Wikipedia raises the concern that people holding conflicting knowledge sets or attitudes may use the collaborative system to stage drawn out

arguments. The ensuing chaos³¹ could potentially destroy achievements made by collaboration (Sanger, 2009). In order to explore how Wikipedia resolves such chaos while keeping millions of its active participants, we will attempt to discover the collaboration pattern in high-risk articles³². We will use them as specific examples to learn how Wikipedia resolves conflicts and arguments among participants and how it maintains the quality of collaborative works, and the extent to which administrators influence this process. To help readers understand the series of questions that will be addressed in this chapter, we will introduce the protection policy in Wikipedia using the example of fully-protected articles and related policies. The chapter will then discuss the relationship between the protection policy and the administration system to clarify the function of administrators in fully-protected articles.

Although it was ascertained that administrators did not influence the content in Wikipedia that much, there were still concerns that, as a particular type of participants in Wikipedia with specific privileges, administrators could affect the collaborative system in one way or another. We deduced that collaboration in Wikipedia did not rely on the administration-oriented system. However, in addition to the collaborative editing, arguments and conflicts are abundant as a result of differing opinions and the differing knowledge level of participants (Kittur et al., 2007b, Viegas et al., 2004). To fully investigate the collaborative mode of Wikipedia, exploring such conflicts and subsequent methods of resolution is likely to yield meaningful insights. Therefore, in the following chapter³³ we focus our discussion on how mass collaboration works in Wikipedia by visualizing multi-authoring edits in normal articles and confliction in fully-protected articles.

Wikipedia is known for its egalitarian system, where millions of participants contribute to the content directly without censorship (Ayers et al., 2008, Lih, 2009b). However such a system also attracts criticism with regard to the quality of its articles because of the possibility of causing chaos to the process of article production (Sanger, 2009, Smith, 2009, Tollefsen, 2009). Wikipedia thus far has put in place a particular tool to avoid such chaos, which is only operated by administrators (Ayers et al., 2008, Winer et al.). When administrators or ‘bots’ (intelligent tools that identify obvious damage to the content of Wikipedia) find an

³¹ The Wikipedia model is based on the belief that the contributions from many different sources can lead to good content. But, in practise, we get both good content and chaos. We only discuss the possible chaos caused by debates and arguments in this chapter. However, we do not deny the possibility that after chaos and the following protection, participants can eventually achieve consensus and the quality of articles can be improved as a consequence.

³² High-risk articles refer the articles with many arguments among their participants which caused potential damage to the quality of articles. Therefore, such articles have more risk than average.

³³ This chapter is twice as long as the other empirical chapters because we applied two different ways to answer the research question whilst others only applied one.

inappropriate edit or an overt vandalizing incident, they have the authority to block that article from being edited.

Generally, protection is applied on high-risk pages to prevent editing wars or vandalism whilst retaining the process of mass collaboration. It must be noted that Wikipedia has many different types of page with different functions to support its collaborative system. To prevent damage, it is possible to protect all pages. In fact, Wikipedia offers a variety of protection statuses, including full-protection, create-protection, semi-protection, and move-protection. These four types of protection are applied in different situations. Full-protection prohibits editing by anyone except administrators; create-protection prevents users re-creating a deleted page; semi-protection proscribes editing by IP users and un-autoconfirmed users (unstable or suspicious users); move-protection forbids moving a page to a new title associated with a new web address. According to the protection policies, create-protection and move-protection are considered specific arrangements against the use of excessive computing power in Wikipedia. Technically, recreating deleted pages or moving pages to a new name could exhaust server resources and undermine Wikipedia's data structure.

6.1.1 Fully-protected articles

Fully-protected articles are those that are prohibited from free editing on Wikipedia in order to prevent damage. They are a type of locked article maintained by administrators, which ordinary participants are forbidden from editing, changing or moving. Full-protection is normally set up to be lifted automatically after a certain period or to be downgraded in level by administrators. It should be noted that full-protection of articles is a demonstration of the privilege of administrators, as the creation and elimination of such a status requires the specific authorization of an administrator. Ordinary participants lack the privilege of changing the protection status of articles. Full-protection often appears as a result of controversial discussions in the editing and unsuccessful protection by the semi-protection status.

As discussed above, full-protection is an important example of how the administration system in Wikipedia manages high-risk pages in order to prevent potential damage from mass editing because fully-protected articles can only be labelled—denoting a change in status, and removed by administrators. During the full-protection period, only administrators can change content, whereas users can view and copy but not edit. However, any modification proposal must be raised on the relevant talk page. After consensus has been reached, the proposed change or complete removal of protection status, will be carried out by administrators but not specifically by the administrator who blocked the article. If the article is uncontroversial, or consensus decides to remove full-protection, any administrator can lift the protection.

Furthermore, article-protection in turn plays a crucial role in preventing vandalism. Full-protection as the highest level of protection represents the best choice to exemplify conflicts and control on Wikipedia. As we have discussed, fully-protected articles are of concern as high-risk articles in Wikipedia are more likely to be damaged or vandalised. Thus, the assessment of full-protection will be made by investigating its designed function and real influence on maintaining the quality of articles. This result will test how effective the full-protection system is in the collaboration process that is fundamental to Wikipedia.

According to our research plan, fully-protected articles will be examined based on individual cases, in order to address collaborative process by multiple authors in a single article. In Wikipedia, every article exists as an independent entry for viewing, although some of the articles may overlap with each other in category terms. In this case, collaboration could be performed in a single article. The editing process of individual article can generally be divided to three periods according to a timeline surrounding the protection status. Here, we name them as the: pre-protection period; full-protection period; and post-protection period. Such definitions help us to understand the dynamics of full-protection and focus on different editing behaviours during the different processes. For instance, in the pre-protection period we will focus on the editing activities which damaged the quality of articles, whilst in the full-protection period, we concentrate on administration activities.

In order to gain an insight into how participants debated and argued, and then compromised and reached consensus, we need to investigate the editing and communicating behaviours of the participants in close detail. Such behaviours, include editing by ordinary participants, conflicting opinions of ordinary editors, discussions among participants and last but not least, editing and management activities conducted by administrators.

Based on the classification of the different types of editing behaviour, we can divide these different behaviours in fully-protected articles into categories. The editing behaviour refers to edits from both ordinary participants and administrators and such behaviour directly influences the generation of an article's content. The discussion behaviour describes that of both participants and readers which occurs on the respective article discussion page, and does not pose a direct influence on the article's content. The management behaviour refers to the execution of the privileged power by administrators to prohibit the editing of specific articles or editing by a particular participant. Although administration would not affect the article content which is the product of collaboration, protection itself still hinders the development and improvement of an article's content since any further contributions could be forbidden.

6.1.2 Related literature

In previous studies, many researchers hoped to further understand the operational model of mass collaboration by defining the role and function of the administration system in Wikipedia. As a matter of fact, the organization of Wikipedia is constantly evolving and developing despite the doubts cast by many critics. First of all, the organizational system of Wikipedia is not aimed at controlling data or resources, instead it is designed to encourage and maintain collaboration and cooperation among volunteers (Forte and Bruckman, 2008). Secondly, its organization system has been treated as an important innovation by scholars who believe that Wikipedia provides an alternative model of self-governance in an online community (Benkler, 2006, Viegas et al., 2007b). Furthermore, some researchers hold the view that in Wikipedia, there is still a complete administration system that is similar to traditional forms of administration – reliant on authority to impose sanctions on participants who vandalize or damage knowledge products (Loubser and Besten, 2008). The importance of administrators is also demonstrated by the fact that they are the essential participants in Wikipedia (Panciera et al., 2009) and are selected based on strict criteria introduced in chapter two (Riehle, 2006, Suh et al., 2009).

The debates about Wikipedia mainly fall into the following categories; first, whether Wikipedia adopts a management mechanism that is similar to traditional firms, especially in the control and punishment mechanisms of authority; second, if Wikipedia has such an authority-based administration system, does this system benefit or harm the current collaborative mechanism underlying Wikipedia's model of operation.

A considerable number of scholars on the other hand think that Wikipedia has a unique organizational system with highly refined policies (Forte and Bruckman, 2008, Forte et al., 2009); comprehensive promotion mechanisms (Burke and Kraut, 2008a, Riehle, 2006); increasingly rigorous management mechanisms (Suh et al., 2008) and a complicated participation model (Lih, 2009a, b). These researchers argue that Wikipedia has a management mechanism similar to that of traditional firms; especially those which are project-oriented. The organization in traditional firms refers to the hierarchical management in which administrators are promoted and share higher privilege of both benefit and power. More specifically, they stressed the point that such an administration system currently dominates the entire collaborative process in Wikipedia (Burke and Kraut, 2008a). They also emphasize that the maintenance of the quality of the product relies on such an administration system (Ayers et al., 2008, Lih, 2009a, b). Under similar circumstances, researchers hold widely diverging views of the influence and function of Wikipedia's administration system (Kittur et al., 2007a, Panciera et al., 2009).

Some researchers believe that administration plays a positive role in Wikipedia, and that it is becoming stronger and taking a more dominating role as Wikipedia evolves. Such a mode of management will, it is argued, lead the developmental trend of Wikipedia into a similar pattern to that seen in traditional organizations (Loubser and Besten, 2008). In such a system, there are various rules that regulate participants' behaviours and maintain Wikipedia's development and quality of articles (Forte and Bruckman, 2008), especially around the error detection and correction process (Stvilia et al., 2008). The administrators in Wikipedia have been regarded as leaders in the open resource system, although successful leaders, "are more likely to demonstrate flexibility and to rate as egalitarian" (Reagle, 2007 P.114).

However, there are scholars who believe that Wikipedia should represent a free system, with a decentred administration system and wider democracy in its administration promotion process (Burke and Kraut, 2008a) and even within the entire collaborative system (Benkler, 2006, Hippel, 2006). Having recognized this, some scholars have begun to discuss how the administration system influences and limits collaboration in Wikipedia's development (Lih, 2009b, Suh et al., 2008). The research presented here builds on this existing literature by presenting a novel analysis that explores Wikipedia's administration system. Being inspired by this, we hope to conduct an evaluation of administrators' function in fully-protected articles through the analysis and investigation of digital by-product data,

6.2 Methods

In order to explore the function and influence of the administration system in high-risk articles to discuss how Wikipedia deals with the conflicts and coordination among its participants, we design a series of steps to study its administration system. First, we attempt to graphically address what collaboration, including conflicts and coordination, looks like in the edit process of fully-protected articles. In doing so, we visualize all edit records and the respective participants in fully-protected articles. This type of approach not only shows the collaborative processes at work but also displays conflicts and reversions. After this, we argue that visualization is good for making descriptive statements but it is not sufficient for explorative analysis. Therefore, we adopt correlation statistics to examine the many variables presenting different behaviours in fully-protected articles, through which we are able to assess the effectiveness of full-protection as an important administration activity on Wikipedia; evaluate the function of administrators in executing full-protection; and hence discover the influence of administrators on their engaged edit communities.

This section aims to provide relevant background information for readers before we fully embark on the details of the empirical work. First, we will introduce what data we chose as our database for study. Second, we will review previous research on visualization and try to

provide a categorization according to its application. Third, we will place a specific emphasis on the visualization tool we adopted for the research— History Flow.

6.2.1 Data set

In Wikipedia, every fully-protected webpage displays a “lock” symbol to denote that the article has been protected from editing due to conflicts. Therefore, we collected our data by identifying such “locked” articles. In total, there were 1590 fully-protected web pages in Wikipedia as of 13th, May, 2010. All fully-protected articles from such a data pool have been selected because articles are the main product of Wikipedia, and there were 69 cases in total of such articles. Figure 6-1 shows the distribution of the full-protection status in different webpages including articles.

There are two issues that need to be emphasized about the cases selected in this chapter. First, all cases in this chapter are selected from the record of full-protection in Wikipedia. There are 69 cases of full-protection regardless of the reasons for protection in the English Wikipedia, among which 37 of them were caused by edit-behaviours. We specifically choose these cases for our analysis. It is also important to note that all cases used here were not obtained through sampling but extracted from Wikipedia’s database. Second, the data used in this chapter was selected on the 13 May 2010. Given the dynamics and continuing development of Wikipedia’s system, the status of full-protection in articles is likely to be constantly changing. Therefore, the 37 cases of fully-protected articles have been used to examine the process of protecting articles and the function of the administration system in Wikipedia after exclusion of fully-protected articles caused by technical reasons rather than destructive edit-behaviours. In the following discussion, we will describe how we obtained our final database of these 37 cases in detail –the process is also documented in the appendix table.

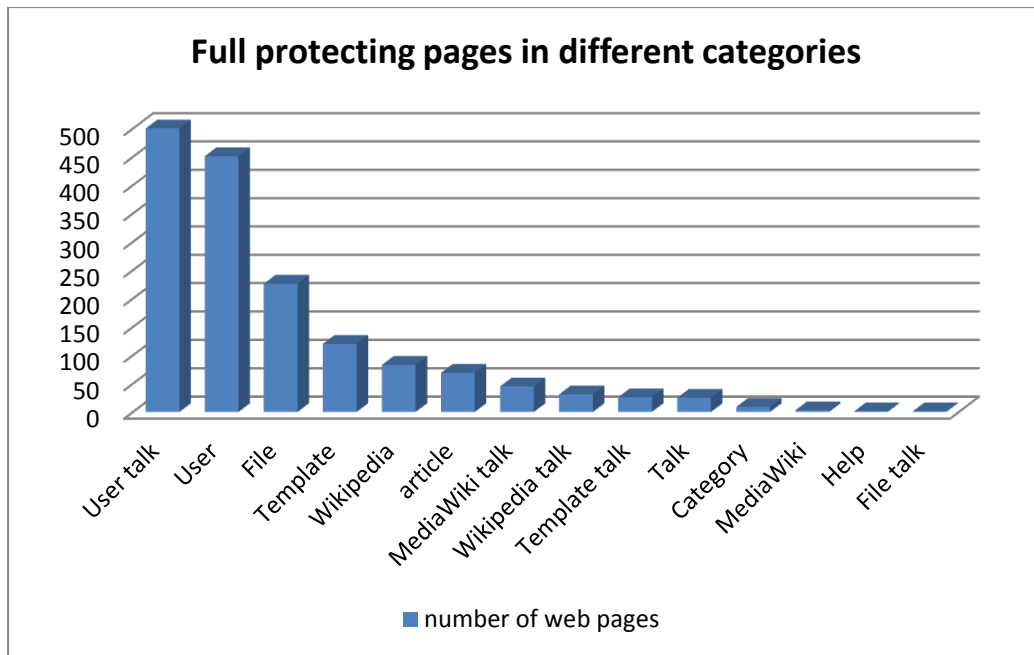


Figure 6-1 Fully-protected webpages in different categories

[Originally in colour]

Wikipedia has 951 fully-protected webpages of user talk and user pages out of 1590 fully-protected pages in total. It could be suggested that vandalism and conflict occur more frequently in personal performances or communication platforms rather than in article pages. Despite the more prevalent vandalism and conflict occurrences in other categories, our main subject will be the 69 fully-protected articles.

In addition to the 69 fully-protected articles collected on 13th May 2010, we also collected the record of associated behaviours, including: edits, discussions and administration. In the visualization section, we will explore all edits including those done in periods of pre-full-protection, full-protection and post-full-protection. However analysis was limited to one hundred edits due to the limitation of the analysis tool. In the assessment section, we chose to examine the database of the current full-protection status, and the previous full-protection records were excluded from the examination unless necessary. The time when full-protection is activated was regarded as the coordinate on the time axis, and we studied cooperative editing before the full-protection, the debates which led to the full-protection, as well as the inhibition of free edits after the full-protection was carried out by the administrators. All investigations of full-protection are simplified by the visualization of the edit time and the respective participants.

6.3 Visualizing full-protection in articles

This section uses visualization to describe fully-protected articles from different angles. Taking full advantage of a complete dataset, we attempt to provide a descriptive analysis of the dynamics of full-protection, collaboration and damaging behaviours in fully-protected articles using visualization.

6.3.1 *Visualizing dynamics of full-protection during a selected period*

As a product of mass collaboration, Wikipedia operates as a complicated system which maintains multiple-entries and a large amount of content. In this collaborative process, full-protection is a type of administrative activity which only focuses on high-risk articles. In order to explore how participants settle their arguments and eventually establish and improve articles, we use fully-protected articles. It is important to note that fully-protected articles vary in their length and frequency of being fully-protected. This is because full-protection can be applied and removed according to the individual situation as decided by administrators. Therefore, the status of articles might change between that of full-protection and of normal open-editing, which illustrates that full-protection is a fully dynamic process.

Unless a platform is open to public input, mass collaboration cannot maintain its vigorous and active changes, which is a significant feature of Wikipedia. In other words, mass collaboration on Wikipedia is not an immobile scene but a temporally and spatially dynamic process. The edits on Wikipedia are extremely dynamic, such that the changes of articles and addition of content can be calculated in seconds. Therefore it is easily conceivable that articles on Wikipedia, which are edited by millions of freelance-like editors without restrictions, are changed almost every second with regard to their content, structure, references etc. Therefore, we argue that the collaborative system of Wikipedia is under constant changes. As an important link in the collaborative process, full-protection is also a fully dynamic process.

As we discussed above, full-protection is a means to prevent damage to the general development of articles which can only be launched by an administrator. Such changes can only be accomplished by just over a thousand administrators. However, full-protection is not an act of free will by the administrator, instead it is an emergency management act based on judgement about an article's status. Therefore, the presence of full-protection is a consequence of the collaborative system. If the latter is dynamic, full-protection should be dynamic as well. Therefore, it is crucial to visualize the dynamics of full-protection. We start with observing the frequency of full-protection changes. The following figure (Figure 6-2) shows the situation of fully-protected articles as of October 2011.

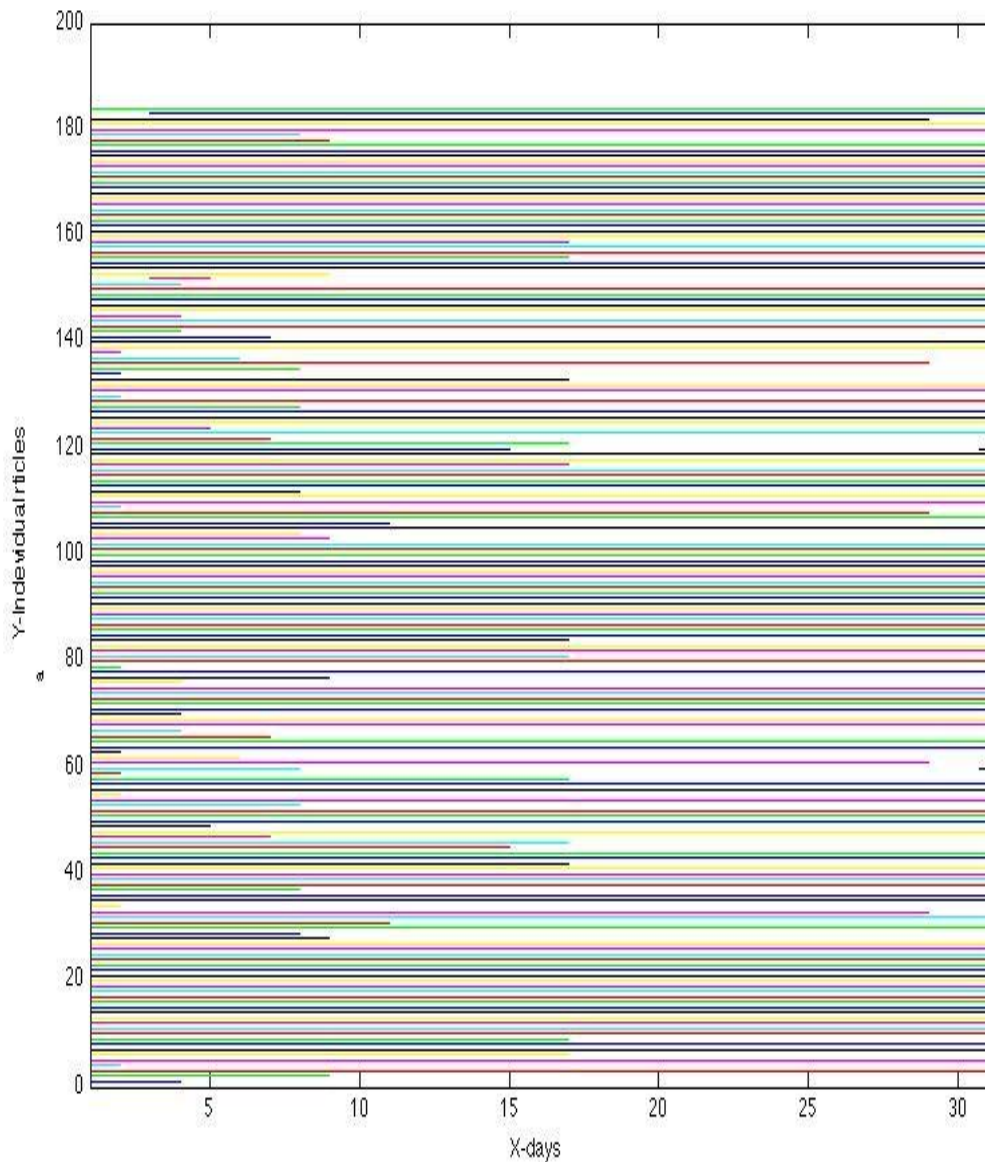


Figure 6-2 Dynamic changes of the full-protection status of Wikipedia articles

[Originally in colour]

illustrates the dynamic changes of fully-protected articles in a one month period, from the 1st of October to the 31st of October 2010. In Figure 6-2, the x-axis shows the date of the month and the y-axis represents the 183 fully-protected articles. Each colour line represents one article, and the length of the line indicates their period of full-protection. Figure 6-2 demonstrates that full-protection is dynamic in both temporal and spatial terms. In order to

focus on changes continuously, the data we analysed here are based on the protection status at a specific time point, which could be different at other times. This result suggests that our analysis only represents the static performance of mass collaboration at a particular time point, although full-protection is a dynamic process.

6.3.2 Clarifying damaging behaviours and the reasons for full-protection

The reason that fully-protected articles become protected is because the edit process may be impeded by continued damaging editing behaviours. In this section, we will mainly rely on the visualization of edits by multiple-authors in fully-protected articles to discover and explore damaging behaviours. From this we hope to address how the administration system of Wikipedia utilizes technical tools to implement protective measures and to ensure the functioning of the collaborative model in Wikipedia, and to improve the quality of articles. The cause of full-protection of articles could be for two reasons: either a technical requirement from the system or from consideration of article quality. With regard to the technical requirement of the system, there are two reasons leading to the full-protection status. The first is that a certain article may have been deleted, and administering full-protection of web pages with that name could prevent it being re-created. Second, some articles may have similar or identical content, but are under different names, and according to Wikipedia's regulatory rules, such articles need to be combined into one. After merging the two articles under one unifying title, the previous article whose title is abandoned would become fully-protected, in order to prevent it from being re-established. Such an instance is often referred to as "redirection".

The second cause of full-protection mainly concerns the maintenance of an article's quality. When damaging edits pose a potential threat to the quality of the respective article, by imposing a negative influence on the quality of the article or interferences in normal edits from other participants, full-protection can also be initiated. Because this type of full-protection is caused by misconduct, which is the type of participation that we would like to focus on, the following analyses of full-protection will be narrowed down and based on these 37 cases.

In order to classify the different causes that lead to full-protection, we only need to investigate the protecting reasons administrators submitted when they labelled "full-protection lock" on articles. On examining these reasons, we discovered that they usually fall into four categories, which are "edit-war", "violation", "vandalism" and "sock puppetry"³⁴. These are also the primary reasons that Wikipedia suggests administrators consider fully-protecting articles.

³⁴ Edit-war is the war in which participants with different opinions fight against each other by reversing edits. Sock puppetry refers to participants using different user name in order to mislead the community on consensus.

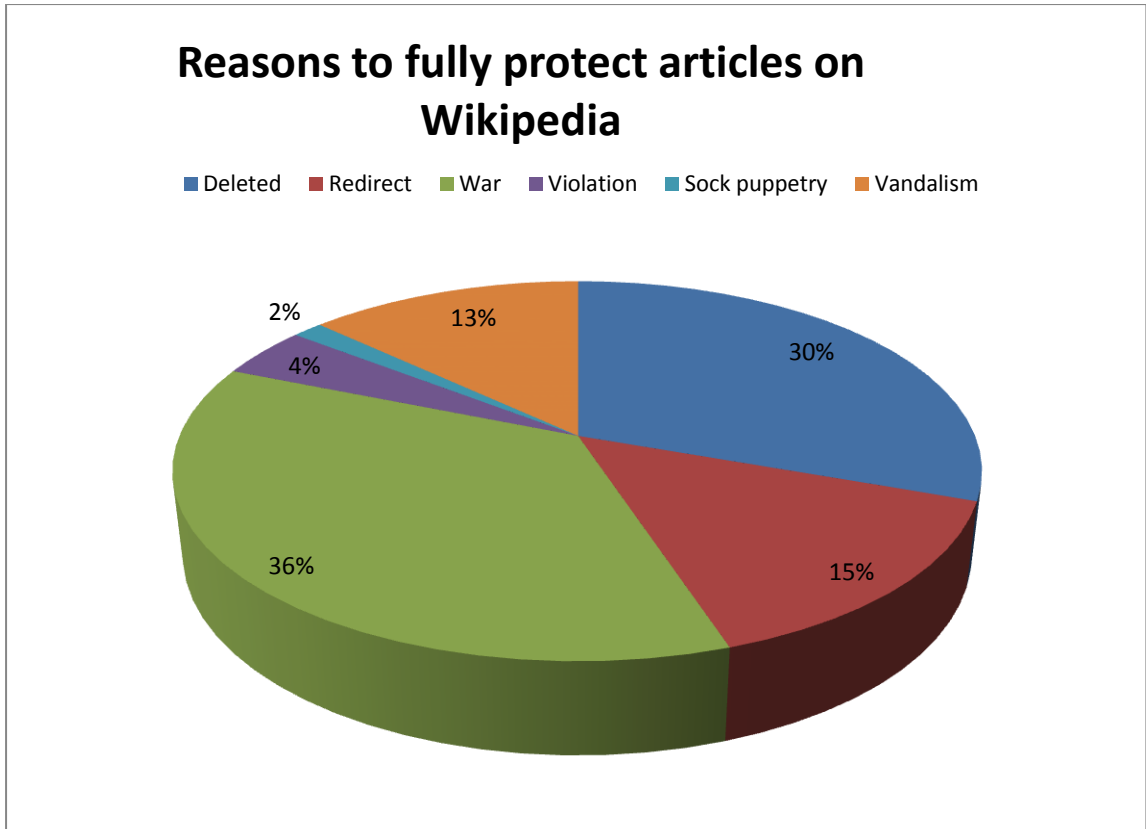


Figure 6-3 Reasons to fully protect articles on Wikipedia

[Originally in colour]

As we have described, although administrators have the authority to label “full-protection” on articles, they need to justify the necessity of doing so. A summary of their reasons should be provided with full-protection. By searching the digital by-product data of Wikipedia, we are able to collect reasons for every single instance of full-protection. In this section, we will introduce the four categories of reasons summarized from the edit history of articles and discussion pages.

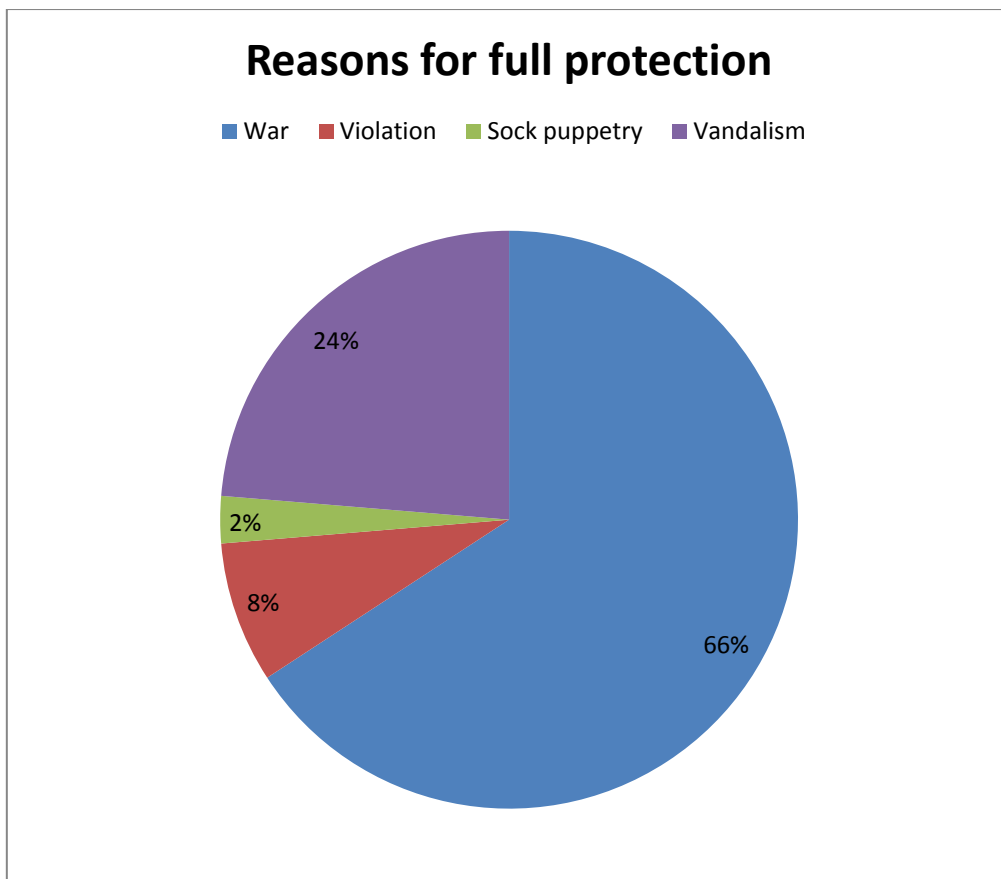


Figure 6-4 Reasons for full-protection

[Originally in colour]

By classifying the reasons administrators gave, we discovered there were four different reasons which could lead to full-protection in a total of 37 cases. As shown in Figure 6-4, these cases were excluded from systematic protection which includes deletion and redirection. The graph also shows that out of the total 37 cases of full-protection, only 24% of the cases are caused by vandalism. This demonstrates that vandalism alone is not the primary reason for instigating protection on Wikipedia. It also suggests that vandalism is not responsible for major damage to Wikipedia. Interestingly, it shows that more than half of the full-protections were caused by an edit-war.

As we have pointed out above, full-protection as an important administration action and the imposition of such a protection status is entirely dependent on the judgement of administrators. Therefore, although Wikipedia provided a categorization of the reasons for full-protection, such reasons only provide a sketchy definition of full-protection. The real causes of full-protection are the various damaging behaviours. In other words, because in the

process of editing an article there may have been particular damaging behaviours, administrators think it necessary to protect the article in order to prevent such behaviours from damaging the quality of the article and the enthusiasm of other editors to contribute. Understanding what damaging behaviours are can help us comprehend the reasons for full-protection. To realize such an aim, we will categorize the damaging behaviours as follows.

From the table below, we are able to see the relationship between the reasons for full-protection provided by administrators and the damaging behaviours in the editing process. The former places more emphasis on the characteristics and potential damage of the damaging activities, whereas the latter focus more on the specific form and manifestation of the acts within articles.




Category of reasons	Reasons of full-protection	Damaging behaviours causing full-protection
Due to waste of technical power	Deleted	
	Redirect	
Due to individual behaviours	Edit-war	Reversion/Attack
	Violation	Reversion/Attack
	Sock Puppetry	
	Vandalisms	Mass deletion/ mass replacement

Table 6-1 the relationship between full-protection reasons and damaging behaviours

Table 6-1 clearly illustrates which damaging behaviours could be linked with reasons of full-protection, and such a relationship will be explored in more details in the next section. What needs to be pointed out here is that the definition of damaging behaviours are decided and based on the specific means of the damaging edits, especially its duration and the impact on the article content. In the visualization process, these behaviours also have shown their unique characteristics. As shown in the above table, we clearly demonstrated what behaviours are linked to reasons of full-protection, and these connections will be introduced in greater details

in the next section. What needs to be pointed out here is that we defined these damaging behaviours based on the specific expression of the edits, especially the lasting period and the influence on article's content. These behaviours show their uniqueness in the visualization process.

6.3.3 Visualizing damaging activities that cause editing-wars

As we have mentioned before, the collaborative model with complete freedom to edit has encouraged millions of participants to join Wikipedia. However the differing opinion and knowledge structure of the participants also has an impact on the robustness and credibility of the knowledge product which results from the open mode of participation. Because of the large population of participants, varied aims of participation, diverse cultural background and different participation means, various forms of damaging behaviours invariably occur as a consequence. In this section, we will visualize the damaging behaviours that result from conflicting opinions among participants and will term them as "argument of content". Depending on the persistence of the argument, the action in edits and the influence on the content, such behaviours can be divided into three distinct aspects: reversion, attack and mass deletion. It needs to be pointed out that there are several other damaging behaviours which we will discuss later, but their occurrence is sparse and cannot be directly identified in the visualization process. Therefore, they were not included in this discussion.

Reversion refers to the instance where two antagonising groups of participants continuously reverse the edit of the opposing group to maintain their own. It mainly comprises a persistent argument without any meaningful improvement of the articles content. Attack defines the action where one or more participants with alternative opinions to the existing article content replaces the current revision with his or their own edit without giving any reason or logical support for their action. Such behaviour is characterized by its abruptness and unexpectedness and does not last long. The potential harm of deletion without a reason includes the negative impact on the quality of an article, causing confusion among participants, and dampening the enthusiasm of other editors.

Reversion means a persistent reverting of the edits from different parties with their conflicting opinions. In other word, reversion represents a repeated argument between two antagonizing sides in the edition of one article. In order to challenge the opinion of the opposing side, the participants could revert or undo each other's edits continuously. For instance, revision A of a certain article is created or supported by a first group of participants, and a second group of participants may disagree with some parts of revision A. The second group creates revision B which changes the parts they disagree with in revision A. If the first group restores revision B

to their previous revision (i.e. revision A), such reversion is continued between revision A and B.

The unresolved argument about content brings constant reversion. Continually changing content is likely to confuse readers who are seeking definitive information on a subject. Because edit wars are rather rare incidents and persistent edit wars are usually fought by particular participants compared to the large number of edits in Wikipedia overall, such wars could be prevented by directly blocking the participants involved. Edit-war is also termed “content dispute” according to Wikipedia’s policy. In this situation, Wikipedia believes that full-protection can force contending parties to discuss their opinions on the talk page in order to reach consensus rather than continually reverting or undoing each other’s edits.

The reversions witnessed in edit-wars can be visualized using History Flow tools. We use another three examples to explain what edit-wars look like. Reversion is an expressive term of edit-war in editing history, and it has the potential to lead to full-protection. The selected examples are: “Battle of Pressburg”³⁵ (Figure 6-5, “Levi Leipheimer”³⁶ (Figure 6-6), “Southern Baptist Convention conservative resurgence”³⁷ (Figure 6-7).

As shown in these graphs, the series of waves represent arguments between different participants, which signify a period of continuous reversions. The repeated reverting of the edit is represented as a regular wave in the visualization results generated by the History Flow tool. By identifying the wave patterns in the edit history flow, we are able to locate when arguments and conflicts occurred during the edit history. We can visualize the edit-war waves in the middle section of Figure 6-5, the later part of Figure 6-6 and Figure 6-7. The density of waves temporally defines the frequency of reversion and the amplitude of the waves shows the amount of text altered in the arguments among participants. For instance, the middle part of Figure 6-5 shows obvious waves including some large waves representing full-blown disputes leading to reversions and some smaller waves representing content reversions on a smaller scale. Meanwhile, Figure 6-6 shows a series of regular waves, which means that particular content in the article was subjected to continuous reversions.

³⁵ Battle of Pressburg, refers to a battle fought east of Vienna on July 4, 907, during which the Bavarian army was defeated by the Hungarians.

³⁶ Levi Leipheimer is an American professional road bicycle racer, who is born on Butte, Montana in October 24 1973.

³⁷ The Southern Baptist Convention experienced an intense struggle for control of the resources and ideological direction of the now sixteen million member denomination.

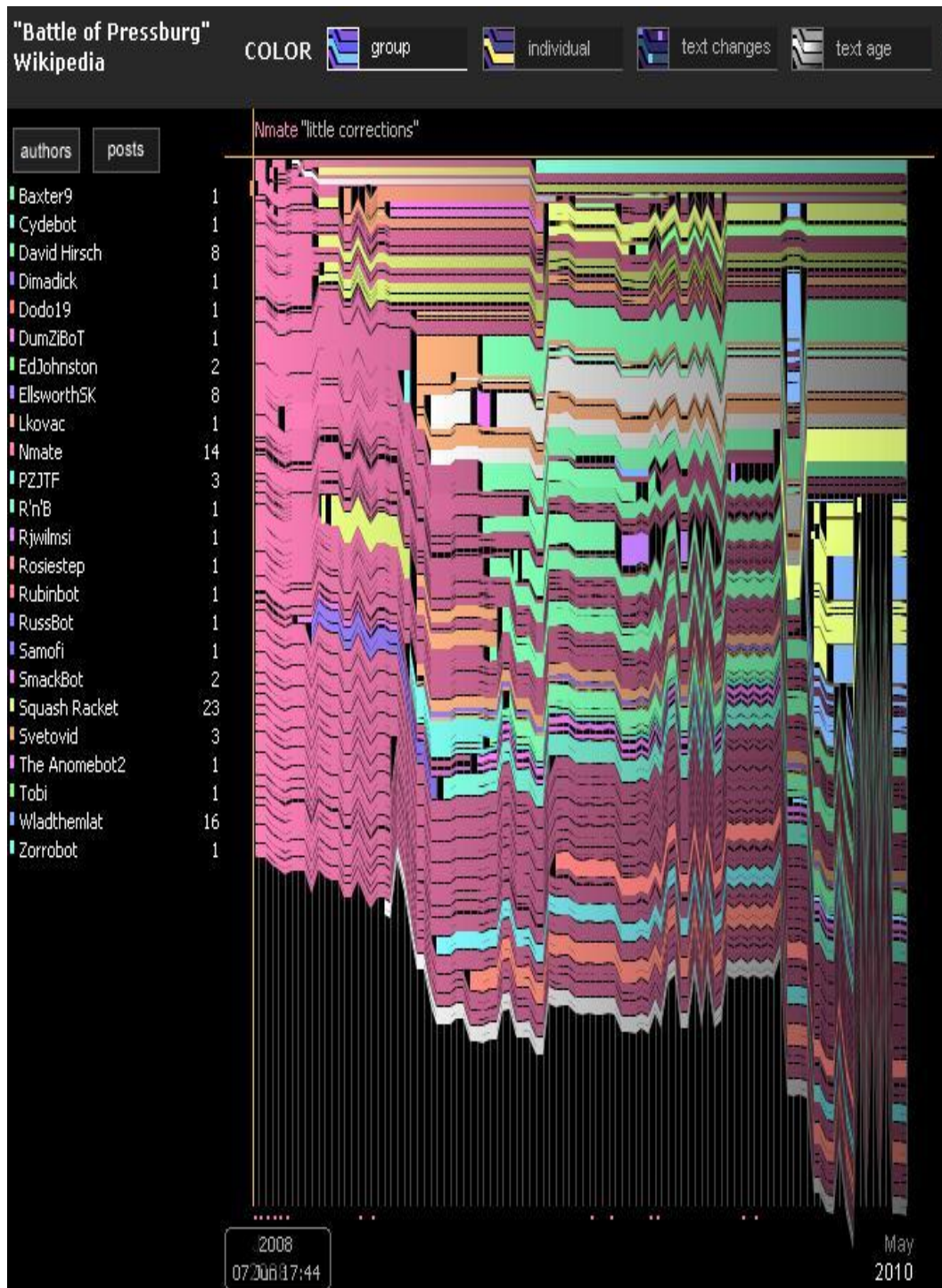


Figure 6-5 History flow of edits in the article on the 'Battle of Pressburg'

[Originally in colour]

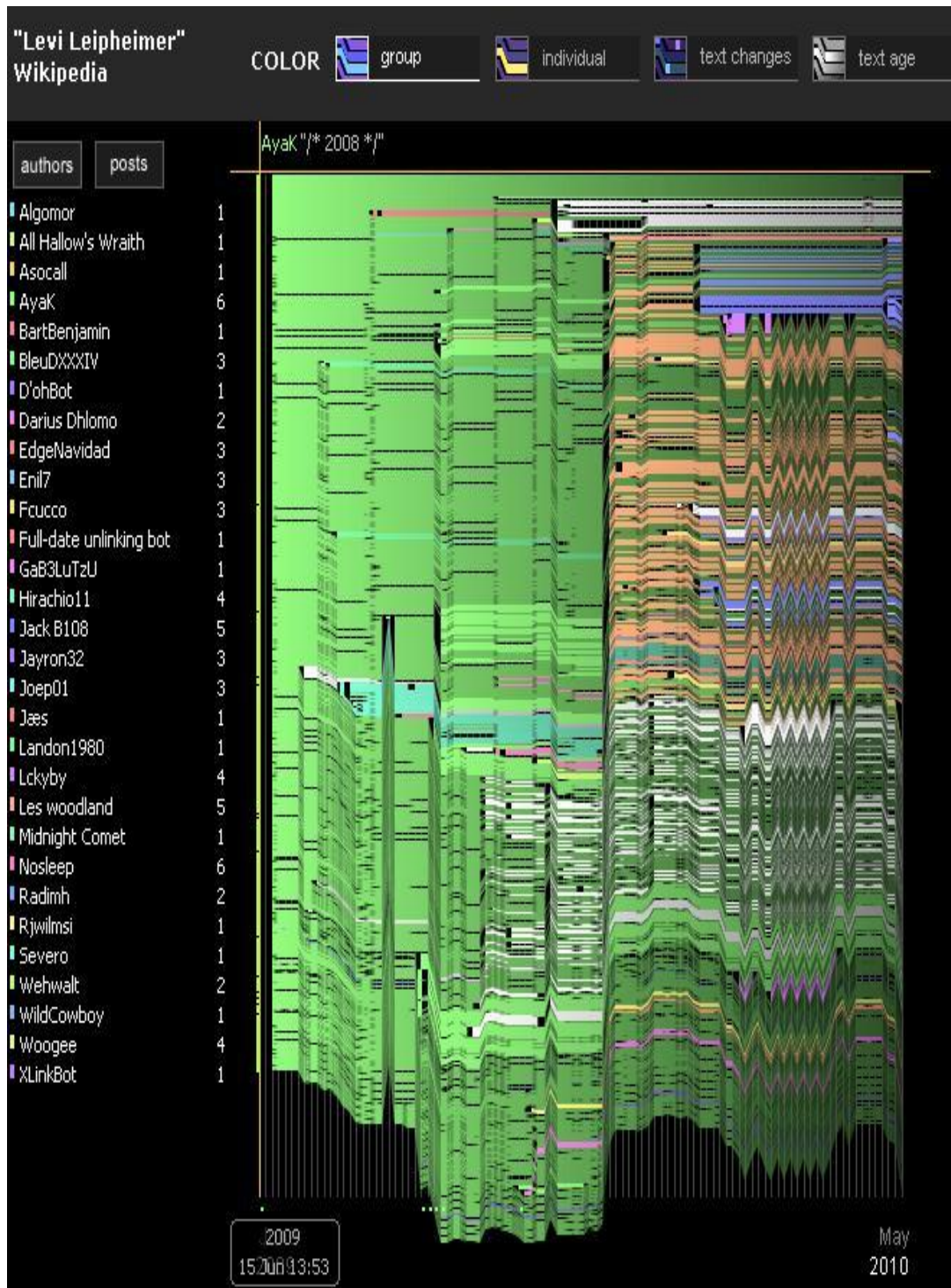


Figure 6-6 History flow of edits in the article on 'Levi Leipheimer'

[Originally in colour]

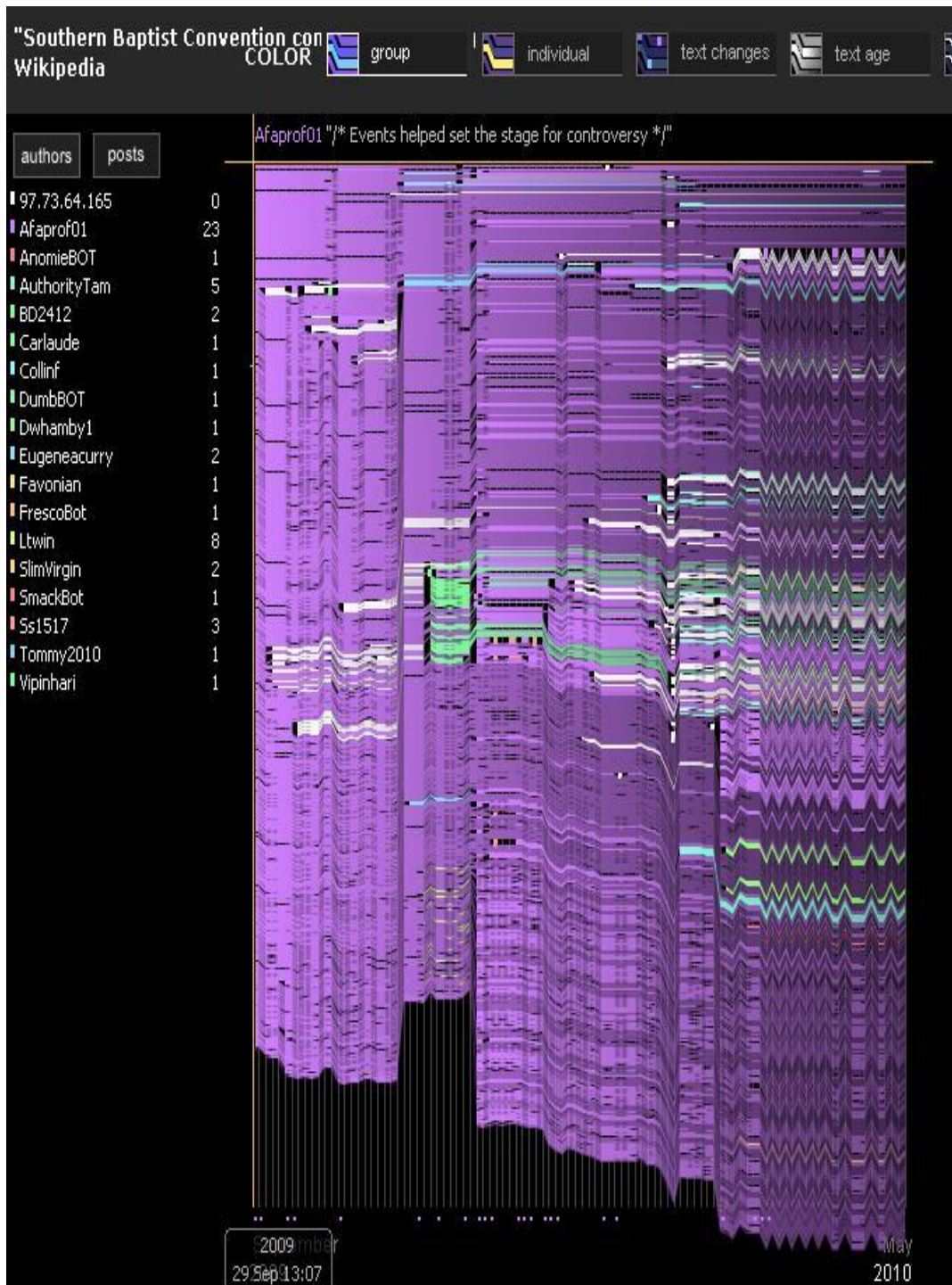


Figure 6-7 History flow of edits in the article on the ‘Southern Baptist Convention

[Originally in colour]

Besides reversion, another expression of argument behaviour is attack, which is characterised by small amount of changes and deletion without giving any specific reason. According to Wikipedia’s data, edit-wars comprise the majority of damaging actions that lead to full-

protection and semi-protection. We visualized how attack actually affected the content of articles, and demonstrated that they indeed interrupted the continual development of articles in Wikipedia. They impeded the process of mass collaboration by imposing a negative influence on reviewers and potential participants.

Attacks as we have termed them here describe the disruptive activities of transient arguments without later counter-attacks. The characteristics of such activity are transient and short-lived, yet if a number of attacks occur within one article, the content of the article could nonetheless be damaged and the cooperative editing process disrupted. For instance, we illustrate below an example of one fully-protected article in which “attacking” edits appeared frequently. As shown in Figure 6-8, there are a number of small irregular patterns occurring abruptly amidst the other edits, for example the interrupting white lines in the top section. These visual marks indicate the content that was added and deleted abruptly and were generally short-lived in the process of article development. Among such marks, a great many of them indicate “attacks”, and they record the instances where participants try to add a relatively small part of the content as compared to the full article. These small editions are insignificant or irrelevant to the article development overall and are therefore eventually removed.

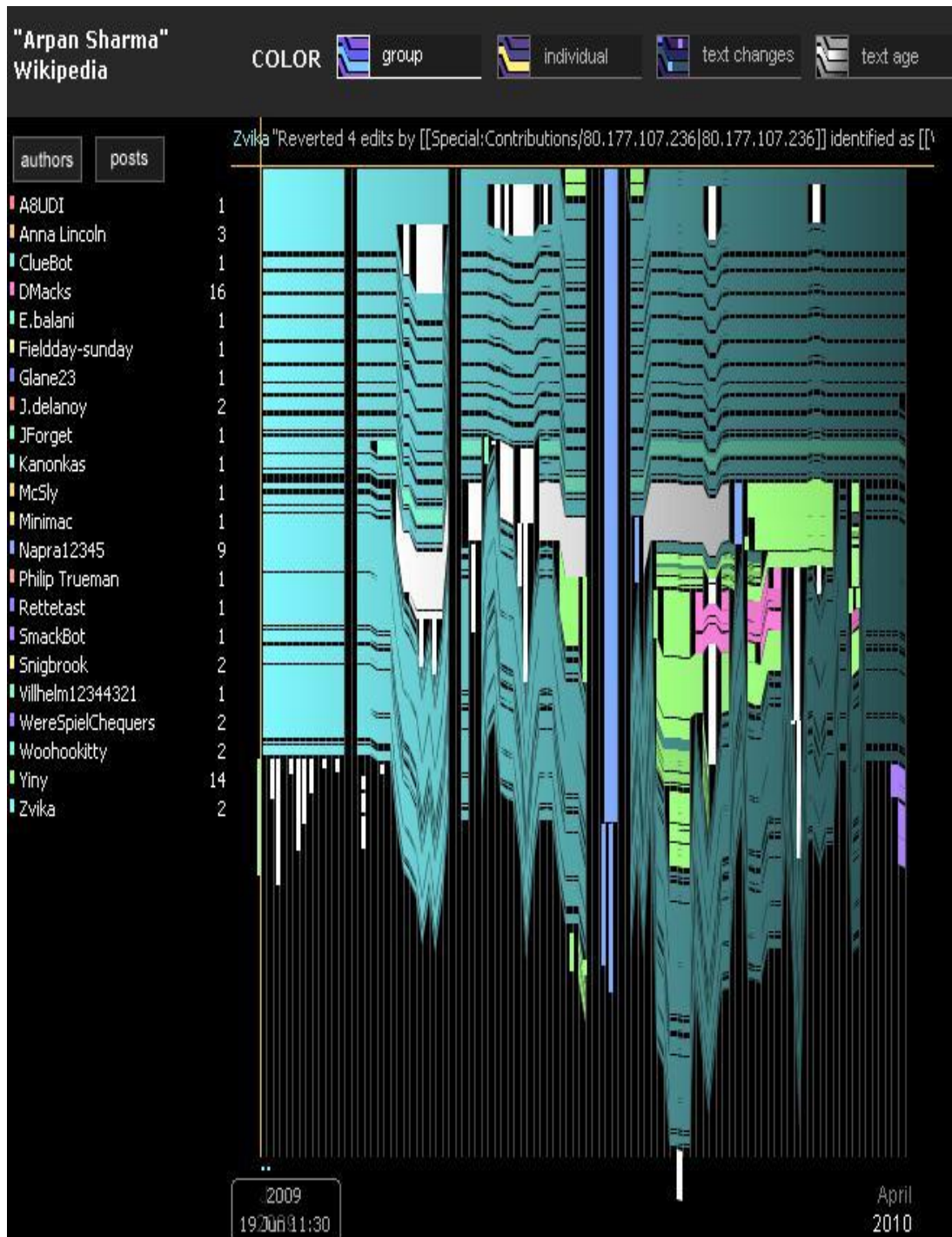


Figure 6-8 Example of “attack” behaviour in the article on ‘Arpan Sharma’

[Originally in colour]

6.3.4 Visualizing vandalism

On the other hand, deliberate vandalism is the most noticeably damaging editing activity and it could present false information to Wikipedia's readers³⁸. Vandalism refers to the activities that are carried out intentionally and have a direct and negative effect on the quality of Wikipedia. Currently, Wikipedia does not have a clear definition of what constitutes vandalism to guide their administrators. In the absence of an official definition, the community of Wikipedia set up their unofficial definition of vandalism which includes bad jokes, nonsense, obscenities, unnecessary humour and page blanking.

Vandalism as another reason that leads administrators to fully-protect articles are recognized when a massive replacement of content or a massive deletion happen. Generally, computing technology can monitor Wikipedia to detect simple vandalism activities such as several lines of "HAHAHAHAH", or the deletion of entire pages without explanation. These activities are easily discernable and could be visualized with very little effort.

We selected three articles representing massive deletion to explain what massive deletion is and how it damages the regular edit process. It is noted that although massive deletions may appear frequently in many Wikipedia articles, only repeated massive deletions may cause full-protection in order to stop such vandalization.

³⁸ Brian Wheeler, BBC News, http://news.bbc.co.uk/1/hi/uk_politics/7921985.stm

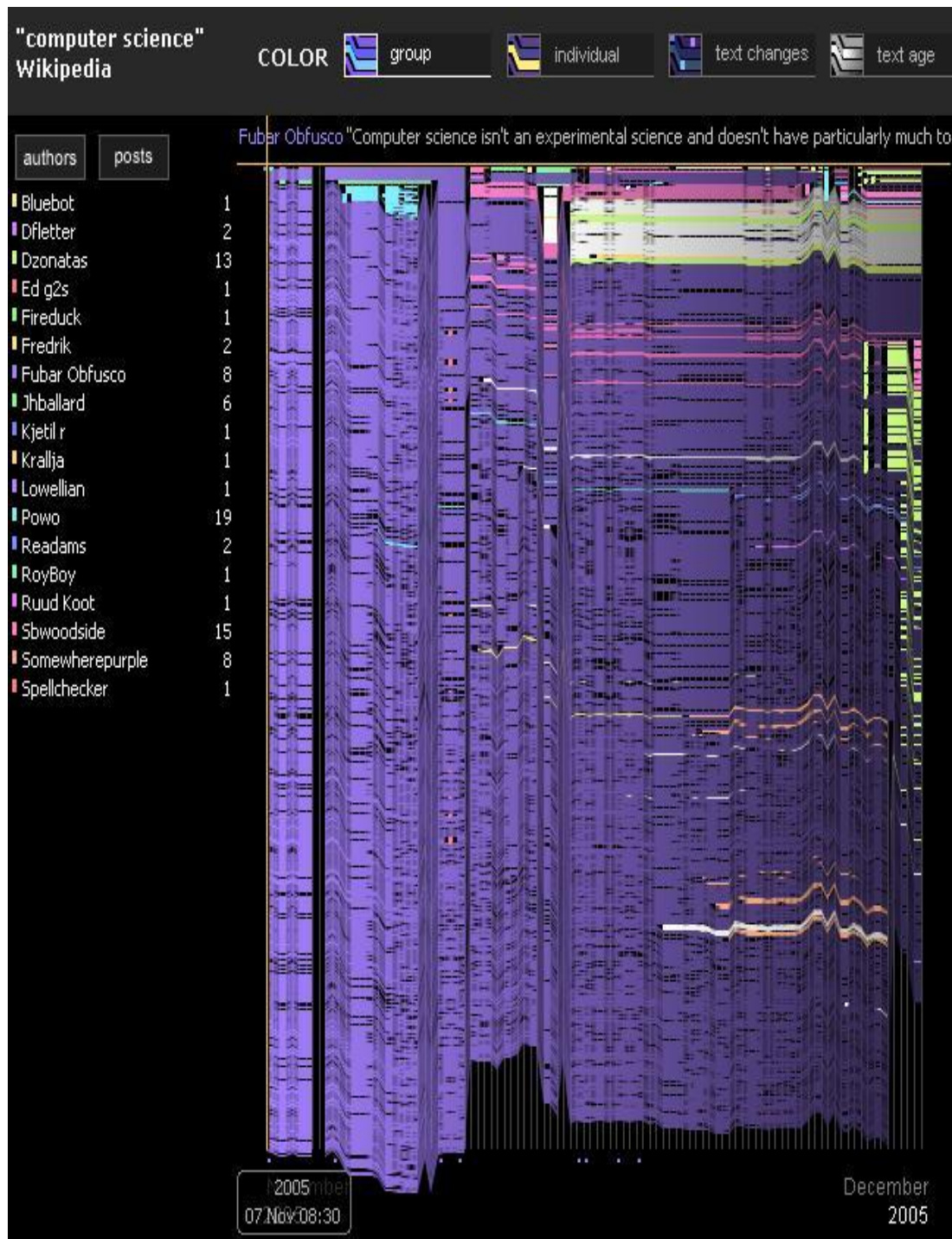


Figure 6-9 Massive deletions in the 'computer science' article

[Originally in colour]

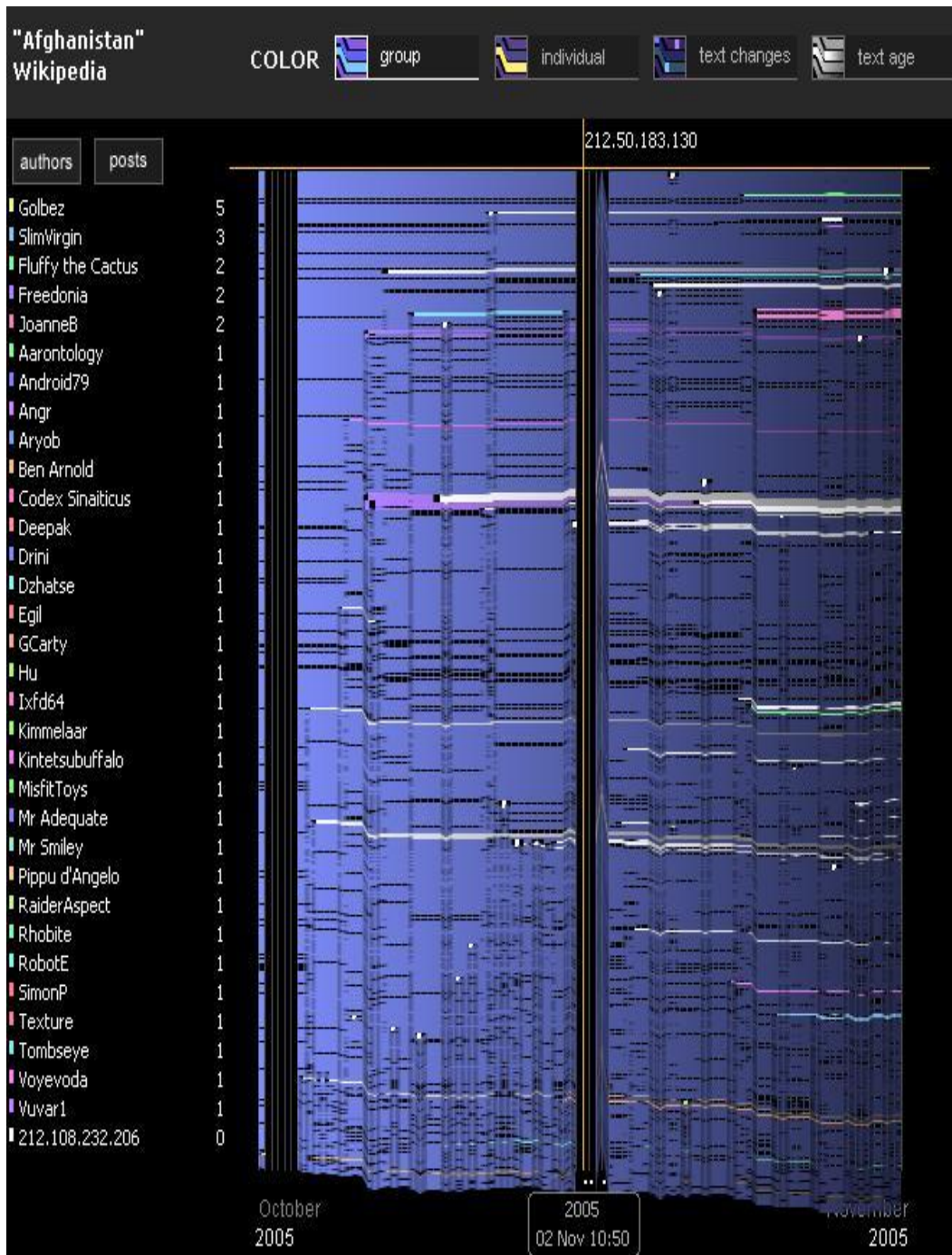


Figure 6-10 Massive deletions in the 'Afghanistan' article

[Originally in colour]

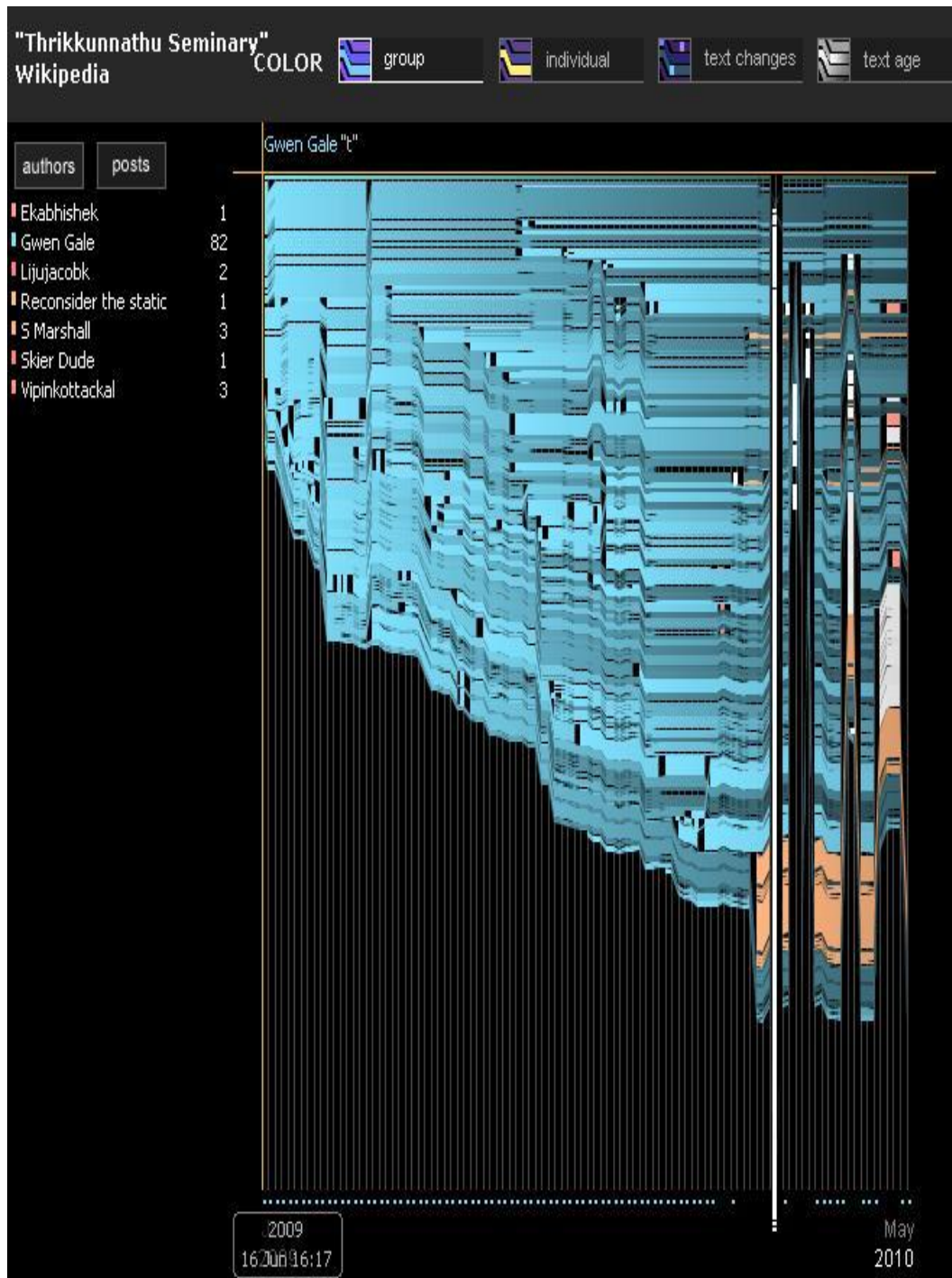


Figure 6-11 Massive deletions in the 'Thrikkunnathu seminary' article

[Originally in colour]

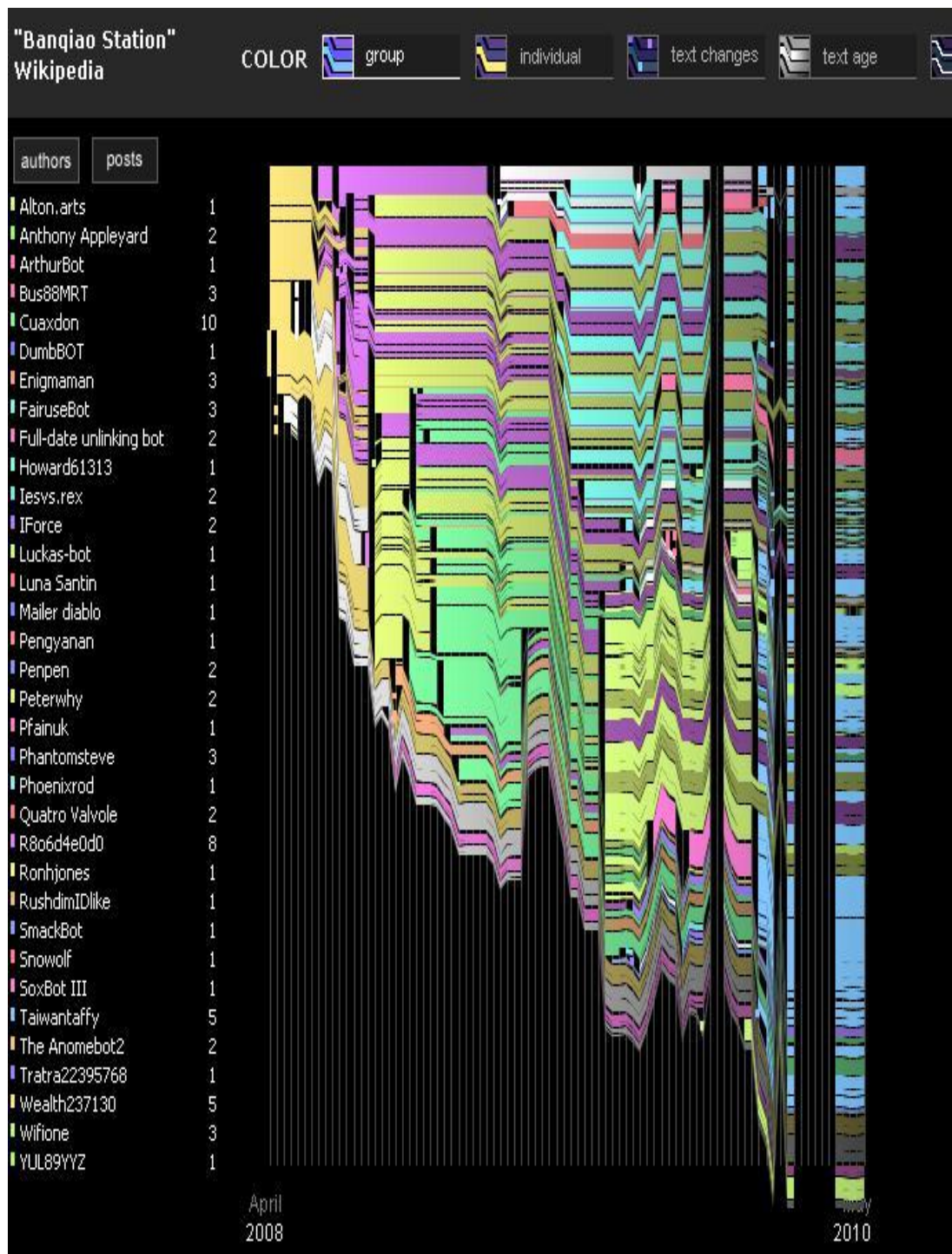


Figure 6-12 Massive deletions in the 'Banqiao station' article

[Originally in colour]

The four figures above show the edit histories of the four fully-protected articles on the topics of “computer science”, “Afghanistan”, “Thrikkunnathu Seminary”³⁹, and “Banqiao station”⁴⁰

³⁹ Thrikkunathu Seminary is an historic former [1] seminary and closed church in the Thrikkunnathu
177

respectively, which present massive deletion. Massive deletion can be seen as an entirely black space in the diagrams.

We are able to see the massive deletions from these articles. The different colours represent editing content from different participants. The small grid divided by the white lines represents individual edits. Therefore, blank grids represent massive deletions, which mean that the content of the article at a specific time point has been completely deleted. In practical terms, the Wikipedia article page will appear empty during such times. From the figure above, we can easily identify the blank grids i.e. period immediately following massive deletion in the edit history, which appears in black. In addition, the total blank period in all six examples only comprises a small proportion of the whole edit history, and in all instances the articles have been restored to a previous edition.

Additionally, massive replacement is another example of vandalism that is shown in the following figures, which are represented by one colour being replaced with another. From Figure 6-13 and Figure 6-14, we discover that the visualization of the edit history is mainly displayed using two colours, and it is the total replacement of the article content rather than editions which defines massive replacement. The subsequent edit completely replaced the previous participation, and led to the possibility of degrading the quality of the article and changed the style of the article set out by the “baseline”. Admittedly, some massive replacement may be beneficial and may enhance articles; however, when massive replacement occurs in articles, it puts the article under the suspicion of administrators, and this unnecessary action can easily affect the quality of articles negatively.

neighbourhood of Aluva, Ernakulam.

⁴⁰ Banqiao Station or Banciao Station (THSR)[6] is a joint-use railway station located in Banqiao District, New Taipei City, Taiwan.

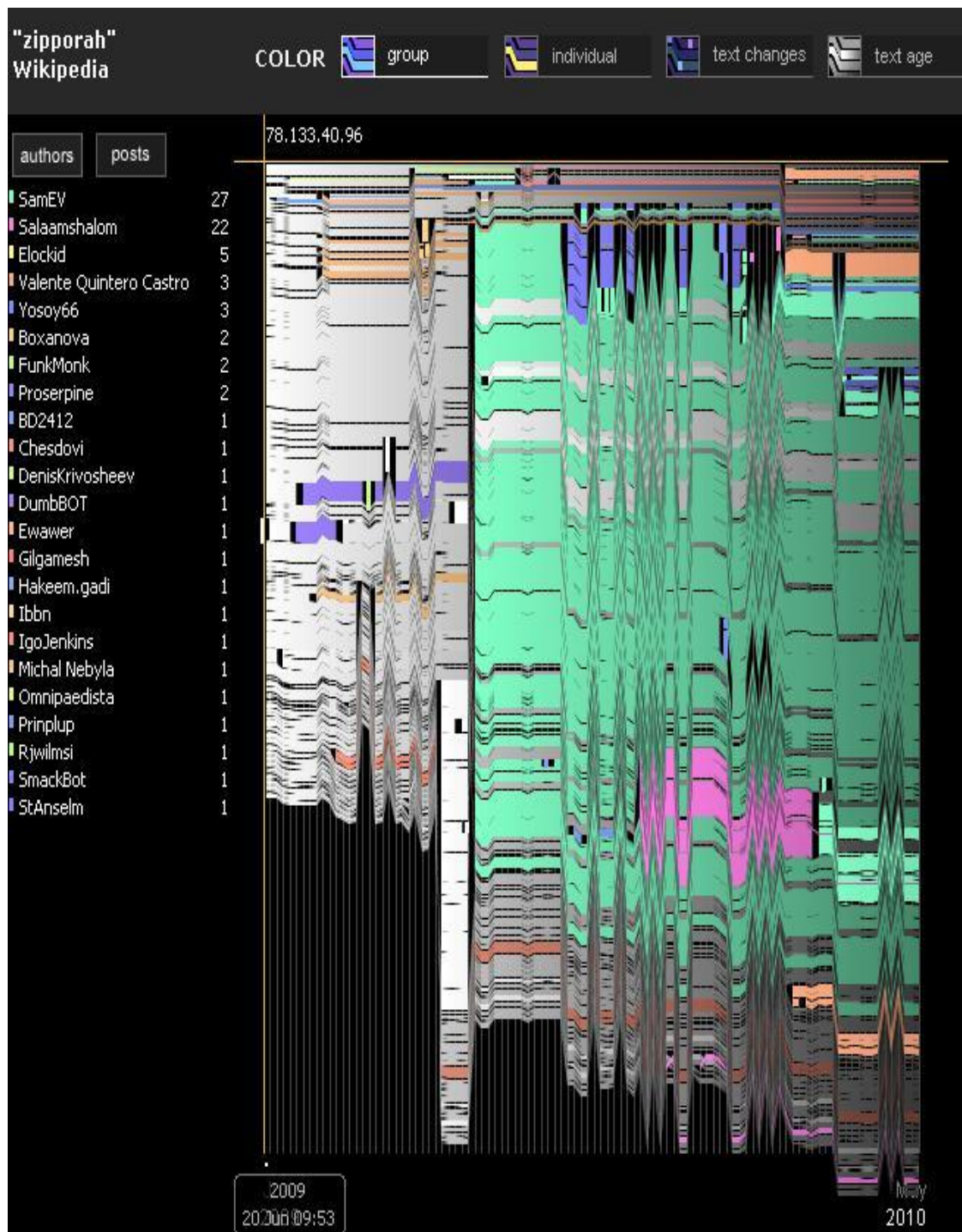


Figure 6-13 Massive replacement in the 'Zipporah' article

[Originally in colour]

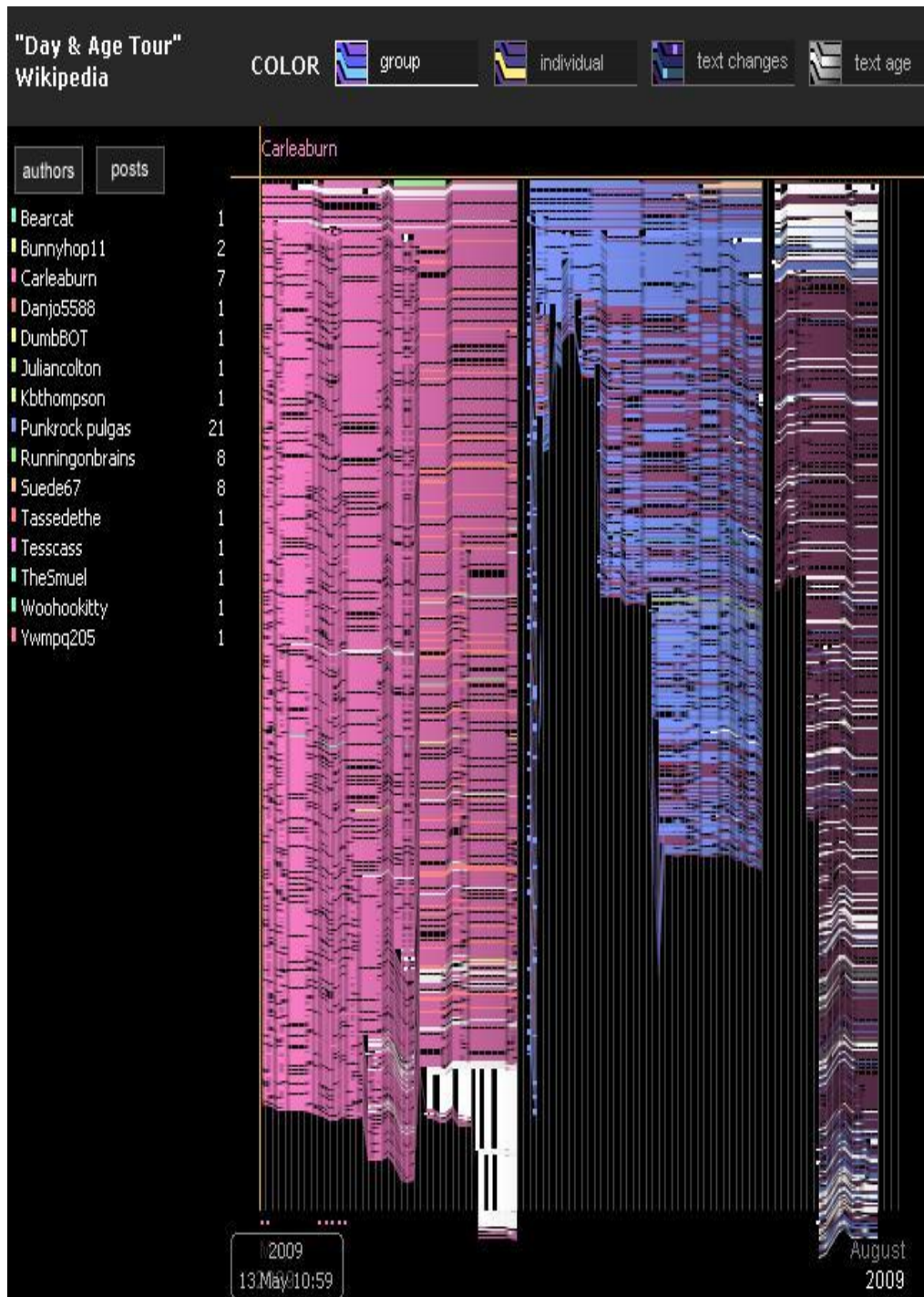


Figure 6-14 Massive replacement in the 'Day & age tour' article

[Originally in colour]

As we have shown, reversion in Wikipedia is an effective method to resolve massive deletions by restoring the blank page to the previous revision. It is interesting to note that the revision before and after massive deletion is almost the same in the six figures shown above,

after observing the length and comparing the diverse colours before and after massive deletion. From these, we found that the most effective way to solve massive deletion is to revert it to the last edition before deletion, which was demonstrated in all of the selected examples as a solution to massive deletion.

However, it should also be noted that another function of reversion is to intentionally obliterate another edit and to preserve the previous edit in an edit-war. In fact, Wikipedia provides a user-friendly operation system that not only facilitates editing for participants but also makes it easy to reverse edits. The convenience of the Wikipedia system encourages more people to participate in the information accumulation of Wikipedia but also may cause some organizational problems such as edit-wars.

6.3.5 Attempting to visualize damaging activities causing violation and sock puppetry

Besides edit-war, which is expressed as disputes of content and vandalism which represent intentions to destroy, there are two additional reasons which administrators frequently provide when they fully-protect articles: violation and sock puppetry. In fact, these reasons were named according to the influence of behaviour and the action of participants.

Firstly, violations comprise copyright violation and defamation of living persons. Although Wikipedia claims that all its information is copyright free, there are still restrictions in the policy regarding the instances which reference or quote contents from other resources. The acts of using other resources without referencing or acknowledgement are considered damaging behaviours on Wikipedia's content and reputation. If such behaviour continues in the editing process, administrators will consider a protection action in order to stop it. The defamation of living persons is another reason to protect articles. Generally, if the article refers to the biography of some living persons, all edits require a high degree of caution and must be from strictly selected materials⁴¹.

However, sometimes edits violate the living person by exposing personal privacy or damaging their reputations. Besides the potential of abusing personal privacy, it may cause legal issues for Wikipedia. In this situation, the content involved would be deleted immediately and the relevant article protected to prevent repeated edits. Moreover, Wikipedia requires a high accuracy in articles that introduce living persons. Contentious information of living people is likely to bring harm, and what is worse is that false statements could injure a person's reputation and the participant could be accused of slandering. For articles on living persons, administrators can fully-protect articles in accordance with the protection policy as soon as they confirm that the edit is malicious or biased.

⁴¹ Information from http://en.wikipedia.org/wiki/Wikipedia:Biographies_of_living_persons

Figure 6-15 and Figure 6-16 demonstrate two examples of violation. However, violation as the reason of full-protection is associated with the content of certain articles but not their edit behaviours. Thus, visualizing violation cannot distinguish between damaging behaviours such as attacks and massive deletions which were previously introduced. Administrators identify that particular content might harm a living person's privacy and the edit behaviour itself is otherwise innocuous. Therefore, we are unable to detect a particular type of damaging behaviour that cause violation by visualizing the edit records in a single article.

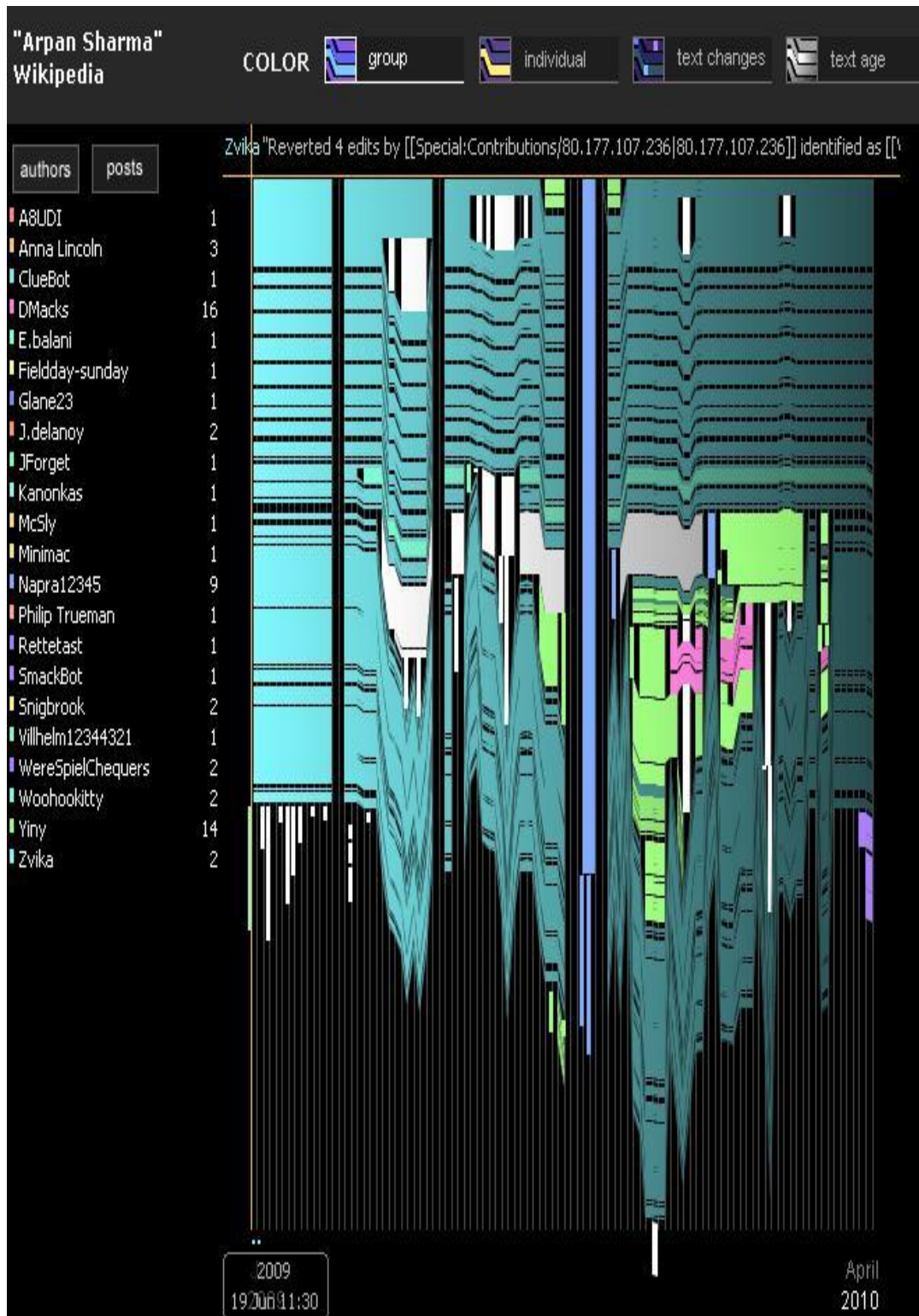


Figure 6-15 Violation of the 'Arpan Sharma' article

[Originally in colour]

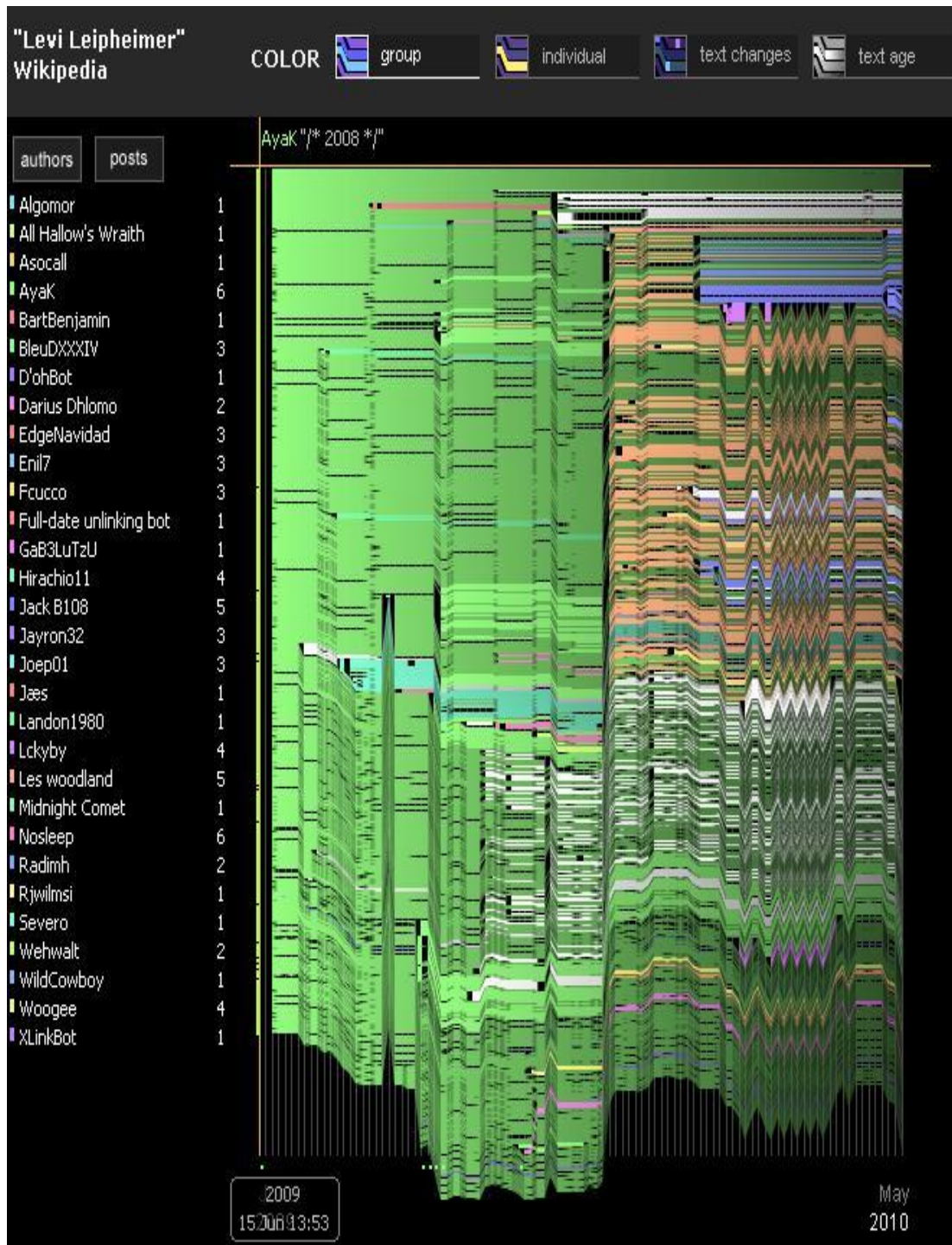


Figure 6-16 Violation in “Levi Leipheimer” article

[Originally in colour]

From the graphs above, we discover that violation could be expressed as either massive deletion as in the case of Figure 6-15 or as reversion shown in the case of Figure 6-16. In

other words, violation is not associated with any special behaviour in the editing process without considering its content and topic. Any basic damaging behaviour could impinge on the right of living persons and thereby lead to full-protection. Thus, from visualization we cannot provide a distinctive description to introduce violation as one of the reasons for full-protection. This is also the same for sock puppetry.

Literally “sock puppet” is a puppet made from a sock, which is manipulated by a person by fitting his hands into the sock and “talking”. In the context of internet (electronic communication), “sock puppetry” could imply false identities. Normally registered users should only edit under their own user account. “Sock puppet” describes the situation where participants disregard the regulation and create other false identifications to edit articles which he/she has already edited with other participants name. The “sock puppet” violation could confuse other participants with online identities and shift the weight of certain opinions to unethically reach consensus. For instance, one participant could create many different “sock puppets” to speak for him/her in a conflict, which would appear as if many editors take on the same side to reach consensus. “Sock puppet” has been used to avoid scrutiny of administration and it has the potential to deceive other participants, mislead edit discussion, distort consensus, stir up controversy and circumvent sanctions, all of which are considered violations. Administrators thus will apply full-protection on persistent “sock puppetry” situations, which cannot be prevented by other administration actions such as blocking participants’ accounts or IP addresses. In fact, generally, sock puppetry instances are dealt with by blocking individual participants. Therefore, sock puppetry has never been the main reason leading to full-protection. However, in order to understand full-protection more comprehensively, we still attempted to visualize the edit behaviours representing sock puppetry in fully-protected articles.

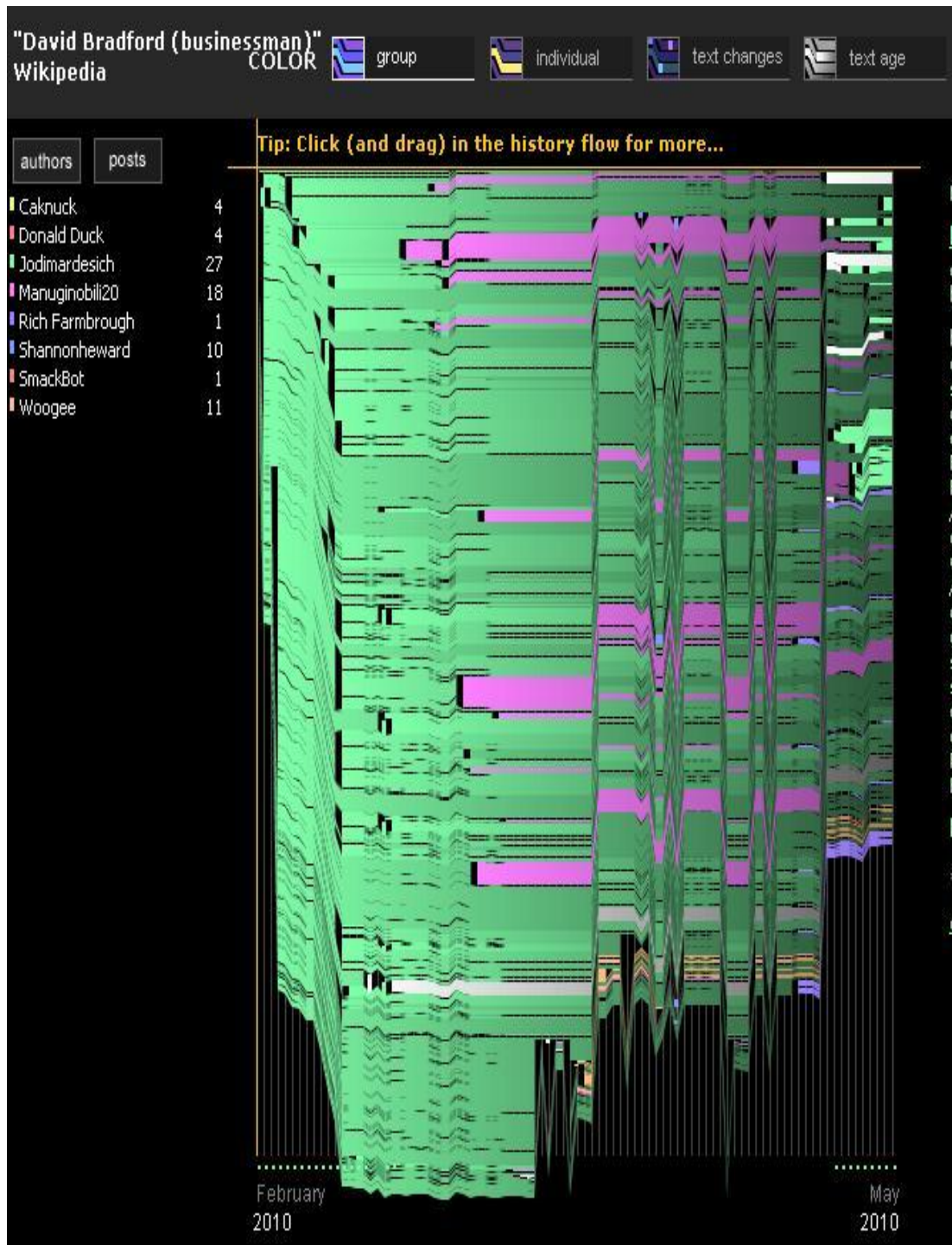


Figure 6-17 Sock puppetry in the 'David Bradford' article

[Originally in colour]

Figure 6-17 shows the fully-protected article “David Bradford” with a protective reason given of “sock puppetry”. With this example, we found that it is difficult to visualize sock puppetry

based on only the edit history record. Although we can colour-code authors according to their edits, we still cannot discern which two or more online participants were created by the same person because the administrator did not provide the full details behind his/her protection. To find out which two or more participants were actually sock puppets, we would need to contact the administrator who carried out the full-protection at the time. However, such data exceeds the basis of digital by-product and therefore was not pursued further. In other words, if we only rely on digital by-product data, especially the database provided by Wikipedia, we are unable to visualize the damaging behaviours in the fully-protected article caused by sock puppetry. Visualization is limited by the accessible information in research. From our description above, we found that not all datasets or questions could be described by visualization. More importantly, we used visualization to accurately and thoroughly describe what fully-protected articles are, what their edit process looks like visually and how various destructive behaviours are represented in the edit process etc. However, we did not answer the question of how the administration system affects the quality of articles by executing full-protection. Furthermore, we even found that visualization alone is insufficient to address such a question.

In the proceeding introduction of visualization, we described it as a method based on the observation of real behaviours, the collection of information from such observations and the deduction of a descriptive conclusion. Therefore, it is no surprise that it can help us to summarise a vivid and specific edit history flow from the intricate record involving millions of editors. Despite this, as visualization is based on analysing actual facts, we were unable to judge and measure the positive influence of protection in maintaining article quality and enhancing Wikipedia beyond our descriptive analysis. In order to further address our questions and accomplish our originally intended evaluation of the function of protection and the related administration system, we need to seek alternative method to analyse our collected data and continue the analysis. Thus, in the next section, we propose a new representational method to analyse digital by-product data of fully-protected articles to explore the function and influence of the administration system, thereby also examining the importance of administrators in Wikipedia's collaborative system.

6.4 Assessing the administration system of Wikipedia

Through visualizing the edit records, the previous section offered a descriptive introduction to the full-protection system and the relevant destructive activities and also describes an attempt at visualizing without particular relevant behaviours. In order to explore the administration system of Wikipedia, we not only plan to describe this system in detail but also want to examine its function and the influence from administrators by investigating its operating

process in fully-protected articles in this chapter. Although Wikipedia provides a platform with equal rights to all participants, full-protection is still one of the very few privileges possessed by administrators as discussed previously. Thus, examining the function of full-protection is vital for the evaluation of Wikipedia's administrative system. As an organizational system, the administration system of Wikipedia is expected to maintain the quality of articles and encourage more participation. Thus, we assume an ideal administration system needs to complete two assignments in fully-protected articles based on the function of full-protection⁴²: terminate the argument and conflicts in fully-protected articles, and to encourage more participants to contribute. In the following section, we will assess these two proposed functions of the administration system through statistical analysis of fully-protected articles.

6.4.1 Assessing the function and effectiveness of full-protection

Does full-protection stop damaging behaviour?

We have compared some variables before and after full-protection to assess whether full-protection can stop the abnormal edits that cause protection. Our assumption is that if there is no further full-protection or semi-protection then the observed full-protection is effective in stopping damaging behaviour. Continued protection could endanger the freedom to edit and the collaborative model in Wikipedia to a certain extent. Therefore, articles that have been fully-protected but entered a protected status again⁴³ suggest that full-protection is ineffective or has a negative effect on the collaboration process.

The data is selected from fully-protected articles in May 2010 and we investigated whether they have had any previous protection or subsequent protection record. From this analysis, we found that 54% of the selected fully-protected articles have had at least one more protecting action according to the history record and 43% of them have had more than one full-protection. This suggests that in more than half of the fully-protected articles, locking the article away from public edit does not reduce the risk that may affect the quality of the article. Our comparative analysis suggests that full-protection cannot effectively prevent further contention that may lead to re-protection.

⁴² As we discussed in the beginning of this chapter, the protection policy in Wikipedia is designed to provide a neutral environment for participants to avoid damaging articles. Wikipedia believe that most participants want to improve Wikipedia rather than damage it. With this in mind, the policy of protection places specific emphasis on consensus and communication, as it is thought that these two factors will influence the decision process of executing or revoking protection. All of these policies are designed to encourage participants to achieve consensus during the protection period. Therefore, we assume the ideal function of full-protection would be terminating arguments on the one hand and increasing communication within the community on the other hand.

⁴³ This situation refers that the articles is put back to un-protected status, but then protected again due to further damage.

Moreover, we found that semi-protection is more likely to be used in contentious articles at the start of contention or before the launch of full-protection (Figure 6-18). In 43% of the total cases, administrators have semi-protected articles from anonymous participants' edits before actually fully-protecting from edits by all participants. It suggests that administrators prefer to start with semi-protection to stop destructive actions until such measures are ineffective and they consider more severe action subsequently protects the quality of articles.

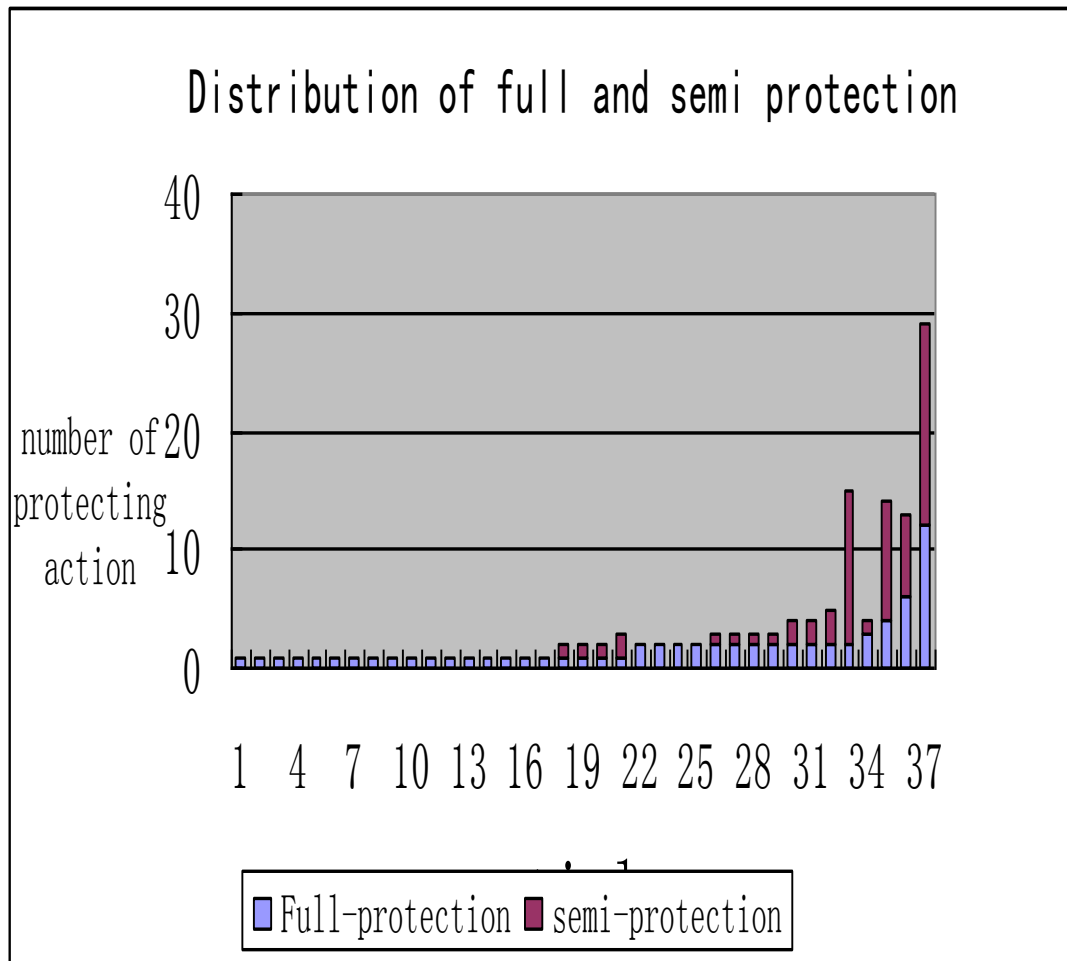


Figure 6-18 Distribution of full and semi protections

[Originally in colour]

On closer inspection of some specific cases, we found that some articles with high levels of protection status are related to hot topics. For example, articles in the database we considered with the most heated debates are “Afghanistan”, “Jewish Internet Defence” and “Led Zeppelin”. Interestingly, two of these topics are related to politics and religion. According to the history record, they have become fully-protected four times and semi-protected ten times.

However, this correlation is not a general occurrence in all protected articles. In other words, there is no correlation between the article content and its continual protection records because p is less than 0.05%.

In summary, we examined the full-protection in Wikipedia by consulting edit history, which suggests that approximately 54% of the fully-protected articles have more than one protection period (i.e. semi-protection and full-protection periods) and around half of them have been fully-protected at least twice. This result suggests that full-protection may be unable to reduce the risk of attracting contention to articles. On the contrary, following the removal of protection, more than half of the articles have been attacked and the quality of the article damaged again. In the next section, we will examine how effective full-protection is in stopping repeated damaging behaviours which triggered the initial protection.

Is full-protection able to stop the same damaging behaviours?

The stated aim of full-protection is to stop continuing arguments and persistent edit-wars by blocking the editing process so as to isolate conflicts. In this section, we intend to observe whether full-protection stops on-going arguments and disputes by analysing the reasons given by administrators when protecting articles. Our assumption is that when administrators have protected articles using the same reason more than once, full-protection is unable to stop repeated damage by conflict.

Our data analysis shows that 43% of the articles have more than one full-protection in its edit record. We then classified these articles according to the reasons administrators gave to justify their protection action. When articles were repeatedly fully-protected for different reasons, it is thought that throughout the article's development, there is no repeated vandalism on the same ground. However, if the articles were protected for the same reason, then it suggests that full-protection is ineffective in its function of stopping the same conflicts.

Based on our selected data, we found that out of the articles that have been protected more than once, only 37.5% of them have alternative reasons for the second and further protections. Thus, in only 37.5% of the cases, full-protection effectively suspended the impairment of article content by similar destructive edits. However, the remaining 62.5% of the articles were protected for the same reason more than once, which suggests that full-protection did not in fact suspend the repeatedly damaging edits.

The results we discovered in this section may suggest that full-protection has little effect in preventing similar damaging edits in order to maintain the quality of articles. Therefore, taking together the first and second result from our analysis, we found that full-protection cannot prevent articles from requiring other protecting actions or being exposed to the same damaging behaviours. However, another important function of full-protection is to encourage

participants to communicate on the discussion page rather than directly present their disputes on articles. Based on this policy, full-protection is designed to stimulate and encourage communications between participants holding different opinions during the editing process. To evaluate this function, we will launch another series of analysis.

6.4.2 *Assessing encouragement through the administration system*

Does full-protection encourage communication?

It is hypothesized that when articles have been fully-protected, the participating editors would be encouraged to communicate on discussion pages to debate and possibly reach a consensus. Wikipedia assumes that “on pages that are experiencing edit war, temporary full-protection can force the parties to discuss their edits on the talk page, where they can reach consensus”⁴⁴. This hypothesis provides the justification of launching full-protection in Wikipedia. Wikipedia believes that it is able to encourage people to engage in conversations aimed at resolving arguments, and prevent malicious vandalising activities. In this section, we compare the average amount of communication in protection-free period with that in protection periods to investigate whether full-protection has any influence on communication on the discussion pages.

To retrieve such data we queried the amount of edits in the talk pages from the digital database provided by Wikipedia. By linking the data of fully-protected article with that of the associated discussion pages, we can calculate whether the number of conversations in the protection period increases compared to that in total regardless of the protection status of each article.

	N	Minimum	Maximum	Mean	Std. Deviation
Conversation_perday_protecttime	35	.0000	8.5000	1.195962E0	2.1834323
conversationperday_overall	34	.0000	1.9649	.169221	.3537277
Valid N (listwise)	34				

Table 6-2 Descriptive statistics of the number of conversation in fully-protected articles

⁴⁴ Quotation from http://en.wikipedia.org/wiki/Wikipedia:Protection_policy

The statistics on the average number of conversations per day Table 6-2 reveal that, during the protection period, the maximum number of conversations is 8.5, whereas during the protection-free period it is 2. More interestingly, the mean number of conversations in the protective period is ten times more than that in general.

Average conversation in protecting period and general edit period

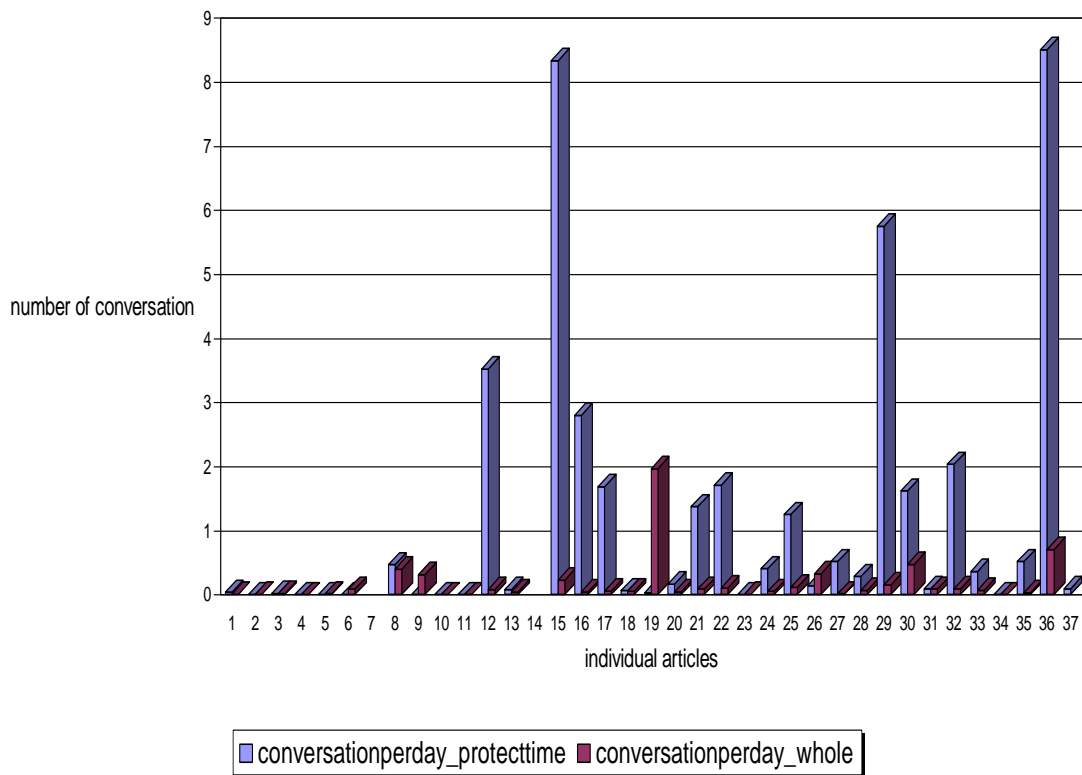


Figure 6-19 Average number of conversations in protecting period and general edit period

[Originally in colour]

Figure 6-19 shows the number of conversations per day during protection periods and those during protection-free periods. The blue bars represent the number of communications per day on the discussion pages during protection periods whereas the red ones show the number of communications in open-edit periods. It is apparent that communication in the full-

protection periods is remarkably more frequent than in open-edit periods. Out of all the fully-protected articles, 64.86% of them experienced significantly increased communications in the respective discussion pages during the protection period. However our hypothesis only holds true when the protection period is sufficiently long for participants to carry out discussions on discussion pages, and there are exceptions in some extreme cases when there are no recorded conversations during a short-term protection. In summary, full-protection indeed encourages conversations.

All in all, assessment in this section shows that full-protection has increased the amount of conversations in related discussion pages by the comparison of the average number of conversations between the fully-protected periods and normal periods. It suggests that participants in Wikipedia are encouraged or forced to communicate to realize their purpose of editing articles, which proves that the full-protection policy is effective in increasing communication among the participating community. Such a finding not only evaluates the function of full-protection in increasing communication, but also demonstrates that Wikipedia to a certain extent is still an administration-oriented system. This is especially true when an article is under an abnormal condition, and administrators still have a leading role in such a process.

6.4.3 Assessing the influence of administrators in fully-protected articles

In the last section, we discussed how full-protection as an administration regulation protected the quality of Wikipedia and improved its development. In this section, we will mainly focus on the execution body of full-protection— the administrators – and discuss their influence. More specifically, we will investigate whether the individual performance and behaviour of administrators will influence full-protection, how they affect full-protection, and whether they affect the functioning of the administration system. In this section we will evaluate the influence of administration by examining the protection action, encouraging communication and the un-protecting action in the process of full-protection.

Since we found that the administration system can encourage communication but cannot terminate the persistence of conflicts in fully-protected articles, we further investigate the influence of administration in full-protection. The administrators play an important and irreplaceable role in the full-protection system for three reasons. First, marking and removing full-protection is a management action that is only permitted by administrators. Second, normally the period of full-protection is decided by and only by the administrator who locked it. Third, only administrators could decide whether the full-protection status needs to be maintained or whether consensus has been reached. Thus, the entire decision-making process may be influenced by the individual biases of administrators. More importantly, this situation

raises another question of what role administrators play in the conflict versus protection process. The answer to this question could provide an insight to the importance of administration in resolving conflicts in mass collaboration.

According to Wikipedia's edit policy, fully-protected articles can only be sealed up and opened by administrators, who may also edit the article during the full-protection period. Other users are only allowed to view and copy the article but have no right to edit. Furthermore, fully-protected articles in Wikipedia can only be labelled and removed by administrators. During the full-protection period, any proposals for modifications should be submitted to the related talk page and will be processed by administrators if consensus has approved it. Normally, the administrator who labelled full-protection can remove it if he/she decides that the article is no longer controversial, or alternatively other administrators can remove the protection if consensus from general participants decides that full-protection is unnecessary. Therefore, administrators of Wikipedia have been empowered with certain specific rights for the maintenance of articles.

On the other hand, administrators also have the general features of normal participants who can edit freely in different articles. Yet at the same time, administrators as the deciding factor for full-protection on Wikipedia directly influence the status of full-protection. In this section, we hope to examine and analyse whether the editing behaviour of administrators will affect the management behaviours of administrators in insisting on full-protection. In order to do so, we first examine the correlation between administrators' editing participation and other variables in full-protection, such as the frequency of edits, amount of conversation, and protecting period etc. Second, we examine the influence of administrators' involvement in the communication of discussion pages during the protection period. The results will show to what extent administration could affect mass collaboration in full-protection, especially in obtaining consensus. Generally, the involvement of administrators might affect the process of achieving consensus and affect the full-protection system eventually. Having addressed the two questions above, we would understand the function and influence of administrators in full-protection.

Our analysis shows that whether the administrator who labelled the particular article for full-protection has participated in editing the article does not have any influence on instigating full-protection. In other words, there is no direct correlation between the execution or removal of full-protection and the previous edition of the particular article by the administrator, which is contrary to our assumption. We originally reasoned that administrators who have edited a certain article are more likely to be aware of and take action against the potential risks surrounding articles that they have edited and are familiar with. Yet, this was not the case.

Additionally, the protecting actions might be implemented by any administrators with or without previous involvement in the protected articles. This is because we could not find any significant relationship between the edits of administrator who protected articles and their editing activities in the article. The absence of such correlations may be caused by specific requests of full-protections by normal users through the request website for page protection⁴⁵.

Meanwhile, the results showed a significant correlation between the number of edits by the administrator who unprotected a specific article and a few variables, including the number of unique editors ($r = .387$, $p < .05$), the amount of total conversations ($r = .365$, $p < .05$), and the total number of views ($r = .880$, $p < .01$). It can be noted that the number of edits is directly proportional to the variables mentioned above. These results might suggest that the involvement of the administrator who unblocked articles from full-protection could affect or be affected by the total contribution and communication and even public interest in the particular article.

6.5 Conclusion

Following a series of detailed analyses, we visualized some fully-protected articles in order to describe all damaging behaviours which led to protection. Meanwhile, the function of administration was also evaluated by testing the influence of administrators on the process of fully-protecting articles. The conclusion not only summarizes all the empirical findings in this chapter based on the analysis of digital by-product data, but it also addresses the lessons learnt about how to apply visualization and statistical analysis using only digital by-product data.

6.5.1 *Power of the administration system represented in fully-protected articles*

The main purpose of this chapter was to determine the role and influence of the administration system in Wikipedia through analysing the editing behaviours in fully-protected articles. In order to carry out the investigation, we first attempted to visualize certain destructive behaviours which caused full-protection in particular articles. We then examined the visualization results to understand how people resolve conflicts and reach consensus.

Through visualization, we illuminated the characteristics of fully-protected articles and their dynamics, the collaborative as well as the damaging behaviours. This description can assist readers in understanding this complicated system and understand the issue of why articles should be protected following certain damaging behaviours. Based on the visualization of the digital by-product data, this chapter firstly discussed the dynamics of full-protection in

⁴⁵ http://en.wikipedia.org/wiki/Wikipedia:Requests_for_page_protection

Wikipedia articles by colour-coding fully-protected periods within a selected month. From the graph obtained, readers are able to view how frequent fully-protected articles changed their original full-protection status in just one month.

Then, this chapter proposed the concept of “baseline” to describe the “power-law” distribution in the editing process of a single article. “Baseline” can be explained as follows: in a single article, the majority of the content is contributed by one or two principle participants. The content they created in the first edition is termed the baseline which includes the significant and essential documents and the baseline points at a specific direction for subsequent edits to follow in both content and structure. From a micro point of view, such a discovery also demonstrates that Wikipedia operates as a collaborative mode in which the majority of the content is produced by a small group of people and most of the participants only edit a minor portion of the content, This conclusion further corroborated our “power-law” discussions in the previous two chapters and illustrated the collaboration on Wikipedia.

The significance of visualizing fully-protected articles lies in that it can describe the various types of damaging editing behaviours that lead administrators to implement full-protection of articles. Through such a visible process, we are able to view the damaging behaviours which caused edit-wars and vandalism. Such diverse destructive activities represent the arguments and conflicts among participants. The damage of such activities to the structure and content of articles can be visualized directly from our analysis, illustrated by the sudden changes to the article structure or the loss of content. However, we also argue that visualization can be limited by the data resource which are available for research because it is difficult to visualize distinct damaging behaviours that cause sock puppetry and violation in articles. These are defined by identifying whether participants use only one name to verify their contribution and by detecting whether the content violates the copyright of other resources or harms living persons, which are beyond our analytic capability due to the limited data we have access to.

Through visualizing the edit records of fully-protected articles, we assembled a comprehensive description of the editing behaviours that lead to full-protection. But this result still did not completely answer our question on how the administration system functions in fully-protected articles. To specifically address this problem, we decided to employ statistical analysis on variables of administration and full-protection. Through this, we hope to gain a thorough understanding of the role and function of the administration system of Wikipedia in the context of full-protection.

First of all, we concentrated our effort to analyse whether full-protection, as an expressive term of administration, realized the originally intended functions, which are to reduce the arguments in edits and ensure the quality of article content. When an article is under full-

protection, any potential improvement and development of the article are inhibited. Although full-protection is designed to stop the debates and arguments of different participants, we found that full-protection could not terminate such conflicts. Over half of the fully-protected articles have entered at least one full-protection period following the initial full-protection, which suggests that similar risks caused by conflicts repeatedly occurred in the same articles even with full-protection administered.

Second, we examined whether full-protection is able to prevent articles from the same damaging behaviours by analysing the reason given by administrators on subsequent protecting actions following the first full-protection. Based on our data analysis, among the articles which have entered full-protection more than twice, 62.5% of the cases were protected for the same reasons on the second occasion. This shows that full-protection not only is unable to prevent articles from similar attacks but also cannot guarantee the article will not require subsequent protection. We acknowledge that such protective measures can preserve the article content in the short terms. However over a long term; it limits the improvement of article content by inhibiting new edits. This suggests that full-protection cannot protect the article from similar destructive editing behaviours; therefore it does not prevent the article from requiring a second full-protection.

Thirdly, full-protection as a forced management measure has also been used by Wikipedia as a tool to encourage debaters to discuss the controversial issues on discussion pages, rather than directly change the article content to express their conflicting opinions. In this section, we also tested whether this function is effectively realised. By investigating the number of edits in discussion pages during the full-protection period, we found that the average number of conversations in the full-protection period is four times higher than that in the period as a whole. We argue that full-protection remarkably increased communication among participants who are engaged in editing high-risk articles.

Fourthly, we analysed the correlation between the personal edit behaviours of administrator and the fully-protected articles labelled by them. In other words, we hope to discover any potential influence from the personal engagement of administrators. However, such a correlation could not be established between the decision by the administrator to fully-protect a certain article and whether or not they have previously edited that article. When administrators carry out full-protection, such an action is not limited to or influenced by their personal edits. This basically demonstrates that when administrators implement their administrative duties, they are not affected by their personal advantage or interest. This discovery is conducive to proving that full-protection is a relatively neutral and fair administration activity.

To summarize, we used a visualization tool to describe what the fully-protected articles were and discussed what damaging behaviours could turn regular articles into high-risk ones. Through the visible results generated by the History Flow tool, readers can easily comprehend why such behaviours would damage the regular structure and content of such articles to eventually necessitate full-protection action. Following this, we used standard statistical methods to analyse digital by-product data to examine the effectiveness and influence of Wikipedia's administration system.

6.5.2 Advantages and limitations of visualization

In this chapter, our initial plan was to use visualization to address another descriptive analysis of full-protection and to examine its function to evaluate the administration system of Wikipedia. However, we found that visualization alone cannot provide enough information. In order to answer our original research question, we have to utilize additional statistical analysis to formulate correlations among several variables. Therefore, in this section we will discuss the advantages and limitations of using visualization from a methodological perspective.

Visualization has long been used in the scientific community. It is certainly true that visualization could offer readers pictorial explanations and provide them with an overview and a grand perspective, making it easy to explain complicated situations. However, such a widely accepted tool in scientific researches has demonstrated both advantages and disadvantages when used in social science research.

On the one hand, visualization allows the simplification of a complicated system and the demonstration of such a system to wider audience. The expressive power of such a method is rarely demonstrated via other means, such as textual explanation and graphical presentation. Through visualization, we can make a spatially and temporally dynamic system visible. We will use two examples to illustrate how other scientists and technicians have used alternative tools to visualize Wikipedia.

It is obvious that visualization has many advantages to deliver accurate explanations of complex systems with a large amount of data, which is also why many social scientists chose such a method to closely study Wikipedia. There are specific examples from other researches such as how participants contributed to an article on Obama⁴⁶, or describe the trend of deletion decisions in articles⁴⁷. In these cases, visualization of data gleaned from Wikipedia answered the question "what is Wikipedia?". Yet we ponder what else visualization can do for research, whether it is only able to describe a situation or whether it can be used to

⁴⁶ The result from such a study can see at <http://vimeo.com/2177573>

⁴⁷ The research detail can be find on <http://notabilia.net/>

complement other analysis, and as social scientists, what attitude we should have towards applying visualization in our study. This section will try to discuss these issues.

Visualization has been conceived as a new method in social scientific studies. In fact, visualization could provide a comprehensive sense of the movement, style and orientation of specific social phenomena, and it has already been used in previous social network studies (Neustaedter et al., 2005, Wasserman and Faust, 1994). Through our analysis of the visualization of Wikipedia using the History Flow Tool and other software, it is clear that visualization is able to provide an insight of systems, social structure and people's behaviours on a macro level.

One of the advantages of visualization is that it is a descriptive and interpretative approach to explain the complicated issues addressed. Such a method allows large amount of quantitative information to be represented pictorially, by transforming numerical and digital data into colours and delineated. Time, space and individual participants are converted to x, y and z axis respectively. More importantly, visualization provides a more accurate description of social phenomena compared to traditional survey methods. With the aid of rigorous techniques, such as mathematical modelling and computer graphic software, social phenomena could be visualized and summarized into several types of graphs. These graphs accurately represent the behaviours and the time points of the action. This macro-description could not be represented easily and clearly by textual explanation.

Instead of collecting data through personal observation, visualization in this chapter is designed to provide an unbiased description using web 2.0 applications. Digital by-product data is used to describe a systematic interaction on the internet platform through visualization. However, the description is only the first step for social scientists to explore social issues. For instance, there are more questions to be addressed regarding full-protection such as whether full-protection is beneficial to Wikipedia's editing process, and how full-protection maintains the quality of articles. These questions require more in-depth analysis rather than mere descriptive images to understand the mass collaboration on Wikipedia.

However, the usage of visualization is often limited to research on descriptive and observational topics by its function. The deduction and conclusion from visualization are based on the observation, comparison and summarization of the graphs generated from the data. For data that cannot be represented by graphs, or that require further analysis and comparison, the limitations of visualization become pronounced.

Visualization only provides a descriptive but not analytical or critical measure of situations. In other words, visualization could help us to understand specific concepts but it does not offer any assessment or argument. The distinguishing feature of visualization is that it not

only enables audiences to pay more attention to the interesting parts of social research, but also helps audiences to understand complicated social interaction or social performance extracted from large amount of information. However, we would be hard pressed to use visualization to provide any further analysis beyond factual recount

To sum up the arguments above, visualization indeed provides advantages when analysing process of online issues from a social science perspective. In the last section, we attempted to use visualization together with other methods to test our hypothesis of the existence of the “baseline”. Using visualization combined with other statistical tools, we concluded that a “baseline” exists in the majority of fully-protected articles. Through this, visualization depicts what baseline looks like in the edit history of each individual article. Additionally, it is important to note that the distinguishing function of visualization is to simplify complicated information especially that gathered from a large dataset. In our case study to determine the “baseline” of editing, we can easily confirm our hypothesis because we simply see it from our visualization results. In this case, it eases the effort of social scientists by visualizing social phenomena in their studies; especially when social interactions need to be addressed using a large amount of digital information.

Based on the previous argument, visualization could be a popular method for social scientists because of the easy access to digital data and the availability of numerous visualization software and technical support. However, we propose that visualization should not represent the only or even primary method for social research. Social scholars, who want to apply visualization, should be aware of the associated technical issues. During the visualization process, another difficulty is that scholars could choose to create their own tools or adopt existing open source software.

From the empirical work presented in this chapter, we discussed how to combine visualization with other analytical methods to clarify complications and approach assessments in social science studies. We proposed that for most research questions related to internet phenomenon, various visualization methods could be utilised to analyse accessible digital by-product data. To determine which method could best deliver information and address the research question directly is a topic for social scientists. Based on the descriptive characteristic of visualization, we recommend that it should be used in combination with other methods which are suitable for further critical analysis and logical reasoning to maximize the advantages of visualization when dealing with large amount of information to assist descriptive research.

Name of article	Discussion ID	Category of content	Protection time	Un-protection time	Reason for protection
David Bradford (businessman)	26401971	People	2010/3/2 1:17	2010/5/20 5:56	sock puppetry
Lock (device)		knowledge	2010/5/11 5:31	2010/5/15 12:11	vandalism
Bully Kutta	2643356	knowledge	2009/11/6 19:09	2010/6/16 19:45	vandalism
Pump It Up Pro	11916491	Music&Arts			vandalism
List of vehicles the United States Marine Corps	12699830	Military			vandalism
Banciao Station	17967769	knowledge	2010/4/30 14:15	2010/5/1 11:40	vandalism
Hungary ? Slovakia relations	20871139	Politics	2010/4/12 1:51	2010/5/12 1:51	vandalism
Afghanistan	16830602	Politics	2010/5/5 0:00	2010/6/2 0:00	vandalism
Computer science	7290	knowledge	2010/5/12 0:00	2010/5/13 0:00	vandalism
M39 Enhanced Marksman Rifle	19581920	Military	2010/3/29 3:17	2010/5/29 16:42	vandalism
Arpan Sharma		People	2010/5/27 1:25	2010/6/30 10:07	violation
Levi Leipheimer	5988736	People	2010/4/30 3:41	2010/5/22 23:55	violation
Bad Boys Blue	4720751	People	2009/10/11 0:00	2010/11/1 0:00	violation
David Mairs (delete)		People			war
Quagmire's Dad	26802587	Music&Arts	2010/5/12 13:42	2010/5/14 13:42	war
Play On Tour	27192295	Music&Arts	2010/5/11 8:23	2010/5/15 6:41	war
Erik Paulsen	19162984	People	2010/4/28 3:41	2010/5/13 1:30	war

Thrikkunnathu Seminary	21463497	knowledge	2010/5/8 15:03	2010/6/8 15:03	war
Jewish Internet Defense Force	18723555	Politics	2010/3/15 22:11	2010/6/15 22:11	war
Oj, svijetla majska zoro	830970	Politics	2010/5/5 9:27	2010/5/12 9:27	war
20th Waffen Grenadier Division of the SS (1st Estonian)	5309342	Military	2010/4/22 18:05	2010/5/22 18:05	war
Arabian Gulf	1276492	knowledge	2010/4/22 17:41	2010/5/22 17:41	war
Zipporah	1918611	Religion	2010/5/11 1:19	2010/6/11 1:19	war
Ahmed Raza Khan Barelvi	2469728	Religion	2010/5/15 18:05	2010/6/15 18:05	war
Noel Gallagher	1096508	Music&Arts	2010/5/9 18:29	2010/5/12 19:59	war
Caucasian Albania	561738	knowledge	2010/5/7 20:37	2010/5/27 20:37	war
Institute for Policy	4850414	knowledge	2010/4/22 18:13	2010/5/26 23:04	war
Pakistan Army	2959565	Military	2010/5/2 0:36	2010/6/2 0:36	war
Ghurid Dynasty	8511255	Politics	2010/5/1 19:33	2010/5/8 19:33	war
Reincarnation research	8609507	Religion	2010/3/27 4:17	2010/5/17 20:08	war
Ronn Torossian	8454038	People	2010/4/6 13:29	2010/4/6 13:29	war
Battle of Pressburg	15370335	Politics	2010/5/11 4:30	2010/6/11 4:30	war
Role of the media in the Yugoslav wars	15827654	Politics	2010/1/25 18:34	2010/5/26 22:32	war

Southern Baptist Convention conservative resurgence	28661579	Religion	2010/5/12 15:53	2010/5/15 15:53	war
History of Georgia (country)	8871921	Politics	2010/3/28 18:52	2010/6/5 19:52	war
Led Zeppelin	17916	Music&Arts	2010/5/9 9:51	2010/5/22 16:46	war
Dusha	17121741	Music&Arts	2010/5/12 7:20	2010/6/12 7:20	war

Table 6-3 Thirty seven cases of fully-protected articles

Chapter 7

Conclusion

This thesis has addressed the challenge that social scientists face caused by the limitation of using traditional sampling methods to explore internet phenomena. In order to clarify the essence of such a methodological challenge brought by new ICTs, we argue that using sampling methods, even when aided by internet-mediated methods, is insufficient. The principle of relying on self-reported data with a limited number of cases should be questioned. Furthermore, we pointed out that such limitations could become more severe in social studies of internet phenomena.

Based on this perspective of the methodological challenge and the possibility of applying new data, this research begins an experimental journey in which we attempt to use digital by-product data to explore a real Web 2.0 application—Wikipedia. From this completed social research which concentrates on internet phenomena, we aim to evaluate the possibility, and the practical influence of applying digital by-product data in the social studies of the internet. The thesis has been divided into two parts: one considers the entire process of using digital by-product data in a social research project, and the other investigates applying such data and evaluates the pros and cons of using this method to meet the methodological challenges as previously discussed.

To begin to use digital by-product data, we had to format and pre-clean the original dataset downloaded from Wikipedia. This is an important step in data mining, as presented at Figure 1-1, which was not discussed in the empirical chapters. In this chapter, we will go through the process of data mining and offer a realistic perspective for social scientists who want to use a large body of digital by-product data in their research.

The study started out with four empirical chapters that investigated the editing and collaborative patterns. We provided a brief overview of what we found from Wikipedia based on its digital by-product data by addressing the following three questions: Does a

collaborative mode exist in Wikipedia? If so, what is it, and how does it work? Through this process, we discovered that the proper function of Wikipedia is maintained by an ingeniously designed semi-administration system, which has its own special model of collaboration and management. Although we could not comprehensively depict such a mass collaboration mode; through our four chapters of empirical work, we nonetheless attempted to analyse the general characteristics of the collaborative system essential to Wikipedia's survival and development. In the end, all these experimental works evaluate the possibility of using digital by-product data.

Following the series of empirical chapters, we sought to address two issues: the first is whether using digital by-product data can resolve methodological demands of data, especially when the traditional methodologies have certain unavoidable problems, for example the effective number of samples and the personal prejudice of responses. The second issue is to measure the pros and cons of using digital by-product data for social scientists from a number of perspectives, such as research expectation, knowledge limitation, and the potential for collaboration across different disciplines. Through such clarifications, there are several essential questions that can be answered in this thesis: does using digital by-product data resolve the limitations and problems of using traditional methods? Can digital by-product data replace traditional sampling methods for exploring online topics? How should social scientists use digital by-product data to maximise research productivity?

In order to answer such questions, this chapter is organised around three themes: the advantages of using digital by-product data in general; the specific benefits of using digital by-product data for internet studies; and finally the problems and limitations of using digital by-product data that we experienced during our empirical work.

Based on the arguments and conclusions made through the entire thesis, in the last section we further explore the data environment and challenges faced by current social scientists. As mentioned in chapter two, we categorise the data types supported by the current internet technologies based on two criteria: privacy and by-product. These two features delineate the possibility and risk of using various data for academic research. Many researchers suggested the possibility of using by-product data, and reminded us that we should not overlook the accumulative characteristic of this kind of data in an array of applications including cultural, consumptive and transactional processes. The contribution of this thesis is that it offers for the first time a comprehensive examination of the possibility of using real digital by-product data for academic research. In this situation, data is not merely a new cultural phenomenon, or simply a technological impact, it implies an opportunity for new use. This opportunity allows academic research to be better integrated with society, and also allows scholars to better

understand the world and keep up with changes in society. The significance and value of this research thus lies in the demonstration of such an opportunity.

7.1 The methodological challenge and the initial study

Our entire research is based on the proposition that there is a limitation to the current research methodology in social science fields in regard to studying the virtual society and internet phenomena, and the consequence of this is the challenge this poses to deploying conventional methodology. Thus, the main focus of the thesis has been, first of all, through the discussions and debates of some social scientists, the presence of the methodological challenge in current social science research is apparent; furthermore, we proposed that such a limitation of traditional sampling method can become even more severe when applied in internet studies.

We explored numerous attempts social scientists have made to use the internet as a medium to contact respondents in order to complete sampling surveys or interviews. While such internet-facilitated sampling studies demonstrate the various possibilities of using the internet in academia, at the same time they manifested the development and progress of internet technologies and the diverse interactions among people online. However, along with the development of the internet and related technologies, we argue that there are more possibilities to collect information and complete research more effectively. In fact, the research on applying digital by-product data has repeatedly appeared in social science publications, although this application has never been systematically discussed or examined. In addition, inspired by the research methodologies used in scientific studies of extracting answers from a large data set, we further propose to use a data mining process to deal with the potentially massive datasets collected directly from internet resources.

In the subsequent chapter, we introduced Wikipedia, which was chosen to explore in exploring internet phenomena in order to test the feasibility of using digital by-product data. There are three reasons why we chose Wikipedia as the case study to evaluate the proposed method from a methodological perspective. First, Wikipedia offered a relatively comprehensive and more structured resource of digital by-product data compared to other internet applications. Second, compared to other web 2.0 applications, Wikipedia provided two specific functions: history trace and reversion. The technical process of operating such functions can generate more types of digital by-product data which can reveal and describe people's behaviours. Third, many previous scientific studies demonstrated how to mine Wikipedia's database, which offered an established process and a clear solution as to how to obtain and use digital by-product data. Fourth, unlike some other resources of digital by-product data involving personal information, which could entail ethical issues, the digital by-

product data sets of Wikipedia have already had removed all private information. Such a database not only reduced risks of infringing on the privacy of participants during the academic process, but also technically reduced the research time by removing steps such as detecting and deleting personal data or obtaining permission to use such data. More importantly, Wikipedia provides a series of heated topics associated with internet applications and interaction within them, such as, “Is it changing the way of life?”, “Is it changing the way of work?” and so on. We not only planned to examine the proposed method by exploring Wikipedia as the primary purpose of launching this research, but also expected to have some initial findings from this experimental study itself. Chapter two introduced Wikipedia and its politics of collaboration and administration, offering a basic awareness of the site before the thesis then proceeded to the subsequent empirical chapters.

Wikipedia has been discussed as an organization that produces knowledge by operating a collaborative mode, in which millions of volunteers work together and achieve consensus if facing conflicts. Although its achievement and the quality of its articles have been questioned from both epistemological and practical perspectives, Wikipedia is still regarded as an excellent demonstration of mass collaboration and continues to receive attention and research interest. Wikipedia has a unique recruiting mechanism and relevant incentive system, through which every participant can edit without obstacles or concerns. The number of participants and the amount of edited content both prove the uniqueness of such an administration system in both qualitative and quantitative terms. Other than the twenty or thirty employees of Wikipedia, the administrators at various levels have joined Wikipedia voluntarily and are elected by a majority of participants in the Wikipedia community. Meanwhile, Wikipedia operates a unique system to maintain the quality of articles and resolve editing arguments among participants, via reversion and protection. Our introduction in Chapter Two not only provided a better understanding of Wikipedia, it drew attention to several relatively important aspects regarding Wikipedia, in preparation for the later empirical study.

Although we attempt to explain Wikipedia and its organization process, all our social science-type questions are not answered only through the analysis of digital by-product data. We need to introduce basic steps of cleaning and integrating the data set before the real analysis can begin. As we have pointed out at the end of Chapter One, the use of digital by-product data relies on data mining as a means to procure useful information and then generate that into patterns in order to solve a problem. However, to get from original to recognizable data, we need to take a few steps beforehand: first, we need to download the data to a certain domain where it can be retrieved easily and stored safely; secondly, the data needs to be transformed into a machine readable format to facilitate recognition by computer programmes; thirdly, the

data need to be categorized according to certain classifications which is analogous to the situation when new books are purchased and they need to be sorted according to the subject or author names before they are put on the shelves in order to easily allow users to find them; finally, according to the category and size of the data, different analysis and visualization tools will be chosen.

It needs to be pointed out that these four steps are described as the “data cleaning and integrating” process in data mining. The consideration of these procedures is scarce in the literature; therefore they have not received their due attention. In responding to this lack, this part of the thesis contains a significant amount of technical information and procedural strategy for data mining, including the choice of different code or graphs. We suggest that informed by this approach, social scientists can fully prepare for this step in order to generate appropriate data set for studies. Also, and more generally, we recommend social scientists take more support and help from scientists and technicians who have plenty of experience in cleaning data and integrating database.

7.2 Overview of empirical works

The principle of this thesis has been to consider the issue of the methodological challenge in studying internet phenomena and to propose a possible methodology of using digital by-product data. However, the process of developing this research method has also allowed us to gain some substantive understanding of Wikipedia itself from a social scientific perspective.

In the original plan, we aimed to complete a descriptive and analytic study of Wikipedia and its collaborative system based on the available digital by-product data resources. In a step-by-step manner, we answered four inter-related questions in our four empirical chapters. These questions arose from the heated debates related to Wikipedia and the innovation it represents, and many previous studies have attempted to explore its organization, just as we planned to do. First, we answered the question of whether Wikipedia has an existing collaborative mode. In other words, we discussed whether there is a regular participation mode based on a system with millions of participants. When we identified the collaborative mode of Wikipedia, we asked the second question – what is this collaborative model. This question was answered in the second part of Chapter Three and in Chapter Four. Specifically, we discussed whether such a collaborative model changes along with the development of Wikipedia, as some literature has suggested. The answer to this question is addressed by exploring the most important participants in Wikipedia. We concluded with a clear idea of the collaborative mode of Wikipedia, while considering the administration function. The third question is how such a collaborative model works in Wikipedia. To answer this question, we used the

visualization tool to examine the collaboration and conflicts in the development process of single articles. In Chapter Five, normal articles were visualized to describe how participants work together and achieve the high-quality content. The fourth question was, what are administrators' functions in such collaboration and conflict? Chapter Six specifically focuses on fully-protected articles to explore the protection policy and to examine the relevant administrators' function and influence. We propose four coherent and interrelated questions regarding the collaborative mode of Wikipedia. Each question is based on the results of the previous empirical work and sets out to ponder and question the results; each of these examines previous academic discoveries by using digital by-product data. The consideration of, and answer to, each question will help readers better understand the characteristics of Wikipedia's collaborative mode. The four empirical chapters can be regarded as a stage-by-stage social science report.

As mentioned in the third chapter, based on previous literature, we discovered that opinions regarding Wikipedia are divided into two camps: one recognises and appreciates Wikipedia itself and its development process, and believes that Wikipedia brings a new participation model; the other camp comes from an understanding of the traditional encyclopaedia, and holds a pessimistic attitude towards the development of Wikipedia. Therefore, we proposed to question whether Wikipedia has an established collaborative mode. In fact, there are three sub-questions that need to be addressed in order to answer these questions. First, is Wikipedia experiencing a positive developing trend? Second, if it is, is there a stable and supportive collaborative mode for the continued development of Wikipedia? Third, if there is such a model, what is it, and how does it work?

For the first question, it is suggested that Wikipedia is expanding quantitatively and improving qualitatively based on our analysis of the data from 2001 to 2007. Such trends convincingly prove the potential of the future development and research value of Wikipedia as an internet application. We then begin to answer our second sub-question. From the data collected from 2001 to 2007, we demonstrate that there is a Pareto distribution in Wikipedia and the constant K varies linearly from year to year. Such a distribution describes the established and stable model for Wikipedia. Based on this statement, we answer the third question of what this mode is by formulating the k parameter by applying the Maximum likelihood method. The significance of our study lies in that we obtained a mathematical model that can predict the changes of edits in Wikipedia. Through these studies, we depicted a systematic and predictable development trend of edits in Wikipedia. It is worth pointing out that this series of descriptions of the development trend of Wikipedia and the establishment of the mathematical model is based on the digital by-product data resource from Wikipedia.

In Chapter Four, we elaborated on the model developed in Chapter Three. Although from a macro point of view, such a mathematical model helps us to argue that the edits in Wikipedia follow the pattern of a minority contributing to a majority of the content whereas the majority only contributes a little, this view of a “minority dominated Wikipedia” is questioned by many researchers. Theoretically, some researchers are convinced that Wikipedia brought forward a meaningful new cooperative mode, where all participants have equal influence on producing knowledge (Hippel, 2006); at the same time, others believe that the “minority-dominated” model is changing gradually as Wikipedia develops (Kittur et al., 2007a).

In order to answer the question of what is the more accurate model we proposed the following as the primary question of Chapter Four: “Who leads the development of Wikipedia?” Previous researches that used edits as a standard to measure individual participation in Wikipedia have inspired us to follow a similar step. Therefore, we make our argument from several approaches, including dividing participants of Wikipedia into different groups according to their number of edits or administrative status, and analysing the absolute change and percentage change of the edits from each individual group from January 2000 to October 2010. Our arguments can be summarised into the two following points: First, the content of Wikipedia is dominated by a minority of participants who maintain a high number of edits; second, the administrators do not dominate the editing of Wikipedia from a quantitative perspective.

There are three main contributions from the research in this chapter. First of all, it confirmed that the participation mode of Wikipedia is still a “minority dominated editing system”, although such a minority is not under the restriction and influence of administrators. This helps us to better understand the importance of Wikipedia’s functional model from a macro point of view. Secondly, the research proves that the development and content of Wikipedia depends on a small proportion of participants, as pointed out by Wikipedia’s founder. Therefore, the fluctuation of the number of the total participants of Wikipedia will not immediately affect the development of Wikipedia. To a certain extent, this conclusion answers the question and doubts of the media towards the development trend of Wikipedia. However, it needs to be pointed out that over the long term, if the number of participants keeps dropping, there will be some impact on those participants with higher-editing records. Thirdly, based on the analysis of the participation records in Wikipedia over a longer period (2000-2010), we definitely refute the suggestions by Kittur et al. (2007a) that there are shifts in the structure of contributing participants.

The combination of results from Chapter Three and Chapter Four illustrate the value of the model developed in the thesis, where the edits in Wikipedia follow the Pareto distribution

pattern – the minority of participants are responsible for the majority of the content, and such content has been dominated by people who have higher edits records but who have not been considerably affected by the privileged administrators.

From the macro point of view, we used statistics and graphical methods to provide a series of descriptions of mass collaboration on Wikipedia. Following that, Chapter Five and Six examined, from a micro point of view, how mass collaboration exists in individual articles. Specifically, how the content of one article is generated through the collaboration via a large body of participants, and how conflicts are resolved in the process. The editing process of Wikipedia is a complicated one. For example, the number of editors for an article can range from ten to over a thousand. Any of them have the right to add, change or delete the previous edits in an article. Such a multi-authored editing system is already difficult to grasp, and an additional challenge for the analysis of such a system comes from the enormous number of participants. Therefore, we used visualization to assist with this analysis, a technique which is often used in scientific studies to handle large and structurally complicated datasets.

Through the overall description provided in chapters three and four, we first proposed the following questions in chapter five: From a microscopic view, how do participants edit one article together? How is mass collaboration demonstrated in this editing process? To address these questions, we chose individual normal articles as our examples and visualize them to illustrate the process of editing and collaboration, through different colours and data structures to link edits with participants. This visualization shows the mass collaboration in a single article within a certain time period. From 330 case studies, we conclude that mass collaboration in an individual article match the pattern of “minority domination”, which we named the “baseline pattern”. This tells us that one or a few participants are responsible for a majority of content in an article and their contributions establish the baseline for such article to direct and guide future participants.

Moreover, from visualizing featured articles which have been identified as high-standard articles in Wikipedia, we argue that Wikipedia provides an innovative participation model where quality and risk co-exist. Traditionally it is believed that in a product with multiple participants, the greater that opinions diverge, the more arguments and lower quality of product is the result. However, as illustrated by the featured articles in Wikipedia, many articles are regarded as high quality collaborative products, and at the same time labelled as high-risk articles which need to be protected from open-edits because of frequent conflict and arguments. Although we cannot identify the sequence of events between conflicts and the appearance of high-quality content, at least we show that the two are not mutually exclusive

in the Wikipedia collaboration model and can co-exist. This discovery provides many interesting views for our research on the mass collaboration in Wikipedia.

Based on the conclusion of chapter four, we raised another question in chapter six: if administrators cannot dominate the edits of Wikipedia in quantitative terms, how does this collaborative mode work, especially what is the function and influence of the administration system in Wikipedia? Based on this question, we chose the fully-protected articles to investigate the function of administrators. According to Wikipedia's policy, the two primary technological privileges of administrators are the ability to protect articles and block participants. In this chapter, we attempt to use visualization to analyse fully-protected articles.

From the visualization, we hoped to understand some relevant questions about fully-protected articles, and the respective function of administrators in the process of full-protection. We visualized all fully-protected articles selected on 13th May 2010 and based our analysis only on the article page. Through visualization, we better understand why these articles were fully-protected, the reasons for administrators to carry out full-protection, and how damaging behaviours affected the content of the articles and obstructed the editing process.

Meanwhile, from our practical experiences, we find that visualization can indeed provide a simple and clear description of the data, yet it offers no analytical results. Although we can provide a series of descriptions of fully-protected articles, we cannot further study the function and influence of administrators in this process. In order to address the research question of this chapter in detail, we used statistical analysis as an addition to examine the correlation between administrators' behaviours and the variables for fully-protected articles. Through the examination of the relationship among the different variables and the influence between them, we discovered that the personal editing behaviour of administrators does not affect the execution of their administrative function. At the same time, we found that in the process of full-protection, administrators do take the leading role, although the length and type of full-protection are determined by some general rules, which are not restricted by the personal behaviours of administrators. From the observation of using digital by-product data for research, we attested that using only digital by-product data can answer a series of questions from a social scientific perspective. Importantly, such experimental research requires a deeper understanding of Wikipedia and its entire innovative and collaborative model. In fact, all these empirical studies are launched to evaluate the proposed method of using digital by-product data in social research as an example. In order to achieve this research goal, we will in the next section, address the advantages and limitations of the method using digital by-product data.

7.3 Benefits of using digital by-product data

In Chapter One, we proposed the notion of a methodological challenge, which can be summarised as a lack of appropriate and effective methods in social science study, especially in studying the internet and the related interactions involved. We expressed our anxiety and concerns as social scientists, towards a professional deficiency of growing powerful data resource that is capable of helping to study society and concurrent social patterns. This lack of research methodological expertise directly affects our ability to understand, analyse and measure some developing aspects of the digital society. Our anxiety and hope for the continued development of social science research led us to propose a new methodology of using digital by-product data; meanwhile we learnt the process of data mining from previous scientific studies. Through experimental studies of Wikipedia, we examined the proposed methodology and evaluated how it resolves the mentioned methodological challenge.

We address the underlying problems of traditional methodology and enumerate the respective advantages of the newly proposed method. Through the description of the characteristics and advantages of the new method, we hope to encourage more social scientists to start using such methods to research online phenomena, as the advantages are clear. Following experimental research on Wikipedia, the benefits of using digital by-product data relate to the possibility of using such data to resolve the limitations of sampling methods and the advantages of using digital by-product data to explore internet phenomena. In addition, we also caution readers of the limitations of their use as the sole research method.

7.3.1 Value of using digital by-product data compared to sample data

To examine the proposed method of using digital by-product data, we carried out a series of empirical studies. Through these studies, we were able to conclude several benefits from using such data. In order to encourage social scientists to use this proposed method, we attempt to underscore the value of using digital by-product data as follows: first of all, the limitations of applying sampling methods to discover online issues are, as Chapter One shows, two-fold: the deficiency of data and analytical ineffectiveness. We will address how digital by-product data can make up for these two weaknesses in data collection.

With new technology maturing every day, more and more social scientists are beginning to adopt internet technologies as a way to find respondents without prior contact, and invite them to participate in questionnaires and interviews. Previous studies showed that launching internet-mediated sampling surveys could contact potential respondents easily and quickly at a considerably lower cost. However, the fact that many potential respondents are invited does not guarantee more participants or more useful data for the research. In other words, the

quantity of invited person cannot increase the quantity of valid responses. However, using digital by-product data can more effectively access and procure effective data, because it depends on the type of by-product and already existing data resources. Furthermore, rapid processing, comprehensive data sets, and low costs are all advantages the new method has when compared to traditional methodologies.

Secondly, digital by-product data, as an already established data structure, can be used and analysed rigorously, and this carries great significance in the verification and repeated experiments for later parts of research. In our research, chapter five and chapter six were based on the same database, in order to maintain continuity and comprehensiveness in the study. When using digital by-product data, we are able to use the exact same dataset, but investigate different variables to study different aspects of a problem. However, if traditional methods were used, we would be hard pressed to reconnect with the same group of participants to collect a second set of data for further analysis.

Thirdly, digital by-product data are not chosen according to specific rules set out by researchers; instead the full collection of data is suitable for study. Using such data avoids the potential bias that may be incurred by selecting sampling objects in social science research. In traditional methodologies, we often need to design the study and choose an adequate sampling group. This is because the costs of collecting data and analysing data need to be taken into consideration. A more prominent advantage of using digital by-product data is that it is based on observations or records of actual facts, rather than self-reported accounts. In previous chapters, we discovered that it is unnecessary to consider sampling cases. Taking the example of chapter six, we can collect all the records of the fully-protected articles as our case for analysis, whilst, if traditional sampling methodologies were to be used, a random selection process or one with designed selection criteria would precede the analysis. It becomes apparent that the application of digital by-product data provides good performance in terms of both the quantity and quality of data in our empirical works.

Finally, in research using internet-mediated sampling methods, self-motivated and self-reported data are the most dubious and questionable. In the first chapter, we point out that the information collected from such methods could be limited or distorted by personal preferences, memory and manner of expression etc., which could lead to analytical ineffectiveness and biases. Our proposed method is independent from the reliance on any information from self-motivated and self-reported resources, and rather depends on an established database for observing behaviours. Such a process of obtaining data guarantees the research has complete and effective quantitative and qualitative datasets, and it avoids artificial selection caused by the prejudices of researchers. Digital by-product data contains

records of the content, number and exact time of participation. This data in comparison to self-reported data which are obtained from recalling facts after incidences is much more reliable and accurate. Additionally, the comprehensive collection of machine collected by-product data eliminates the possibility of exaggerating, diminishing and alternating the information provided in self-motivated and self-reported data.

In summary, there are many advantages of applying the proposed method in response to the methodological challenge social scientists face when using sampling methods. First, this method can make use of digital by-product data resources with abundant information in sharp contrast to the considerably low response rate in sampling methods. Secondly, this method allows data to be examined more than once if scholars need to restart or create different projects using the same database. Thirdly, this method can avoid many possible biases caused by human involvement.

7.3.2 Advantages of applying digital by-product data when exploring internet phenomena

Besides the prominent advantages in comparison to traditional sampling methods, the proposed method also provides new features for exploring internet phenomena. This section addresses two important features of applying digital by-product data, which might be of benefit to social scientists in the internet study fields.

Studies of the internet and related society have a peculiar difficulty in that the speed of change in such an environment is unmatched in traditional society. In other words, it is very difficult to investigate the dynamics of internet society using static information because the speed of establishing and developing such societies or communities is so fast. Consequently, there is a need for social scientists to use relatively new and updated information to conduct studies of online phenomena. Sampling data is unsuitable as it is take relatively long to collect. In contrast, digital by-product data not only provides comparatively recent data, but also updates all changes on the existing data through an established database. This advantage allows social scientists to always possess the most updated information for analysis. It accurately monitors the changes and development of the online society and avoids the problems of inaccuracy and long-drawn-out process of collecting data.

In addition to the advantage of having constantly updated information when using digital by-product data resources, many analysis tools can be found on the internet. These tools are shared by associated institutes or interested amateurs. We emphasise in the first and second chapters that to apply the proposed method, we need certain technical tools to assist us to aggregate and mine data. An advantage of using digital by-product data for studies of internet phenomena is that, many internet applications provide related tools along with their data

resource, and some amateurs also share their analytical tools associated with a particular database, such as Facebook or Wikipedia.

The appearance of these convenient tools is mainly due to three reasons. First, digital by-product data is an open data resource accessible to anyone who is interested and wants to use it for non-commercial purposes. In fact, many people from different backgrounds and with different research interests can share the same database. This unconditional public access provides the possibility of creating adoptable tools for open usage. In a traditionally isolated research environment, each individual researcher has to create and use different databases, and this situation prevents the sharing of research tools. Second, rapidly changing internet phenomena not only draw the attention of academics but also attract extensive public attention. Thus, scholars in public or commercial institutes also want to obtain and discover useful information by extracting digital by-product data. The fact that different fields show their strong and continued interests in understanding the internet— especially Web 2.0 applications— by formulating digital by-product data offers a user-friendly environment for collaboration in creating and producing analytical tools. Third, the internet as a superior media platform provides an opportunity to publish amateur tools, and to peer-review and peer-modify them.

Data resources which are constantly updated by associated technical systems on internet applications and ready for analysis by accessible tools online should encourage more social scientists to use the proposed method in their research on exploring the internet. In fact, these two encouraging factors include updated information and accessible tools accelerate studies using digital by-product data, not only in social scientific areas but also scientific research.

7.3.3 Problems and limitations

Just as all coins have two sides, any research method will have certain limitations and difficulties. In the four empirical chapters, we find that there are three pertinent points which researchers need to consider when using digital by-product data.

First, using digital by-product data in social science studies requires the generation of a series of appropriate variables specifically serving the research purpose. Because digital by-product data occurs alongside real online behaviour, the existing database is much bigger and more complicated than the database actually needs to be. It includes more unnecessary and irrelevant information. Thus, scholars need to use specific tools to discriminate and select valuable data with regard to their research plan. This step is unnecessary in traditional sampling methods, where a database is established exactly following its data requirements. In

other words, to use the proposed method researchers need an extra step in data collection, which would be considered costly in terms of both research time and energy.

Second, the digital by-product dataset sometimes does not include all the required information for the study. As discussed above, the existing database is not designed and created for any research purpose. The included information thus does not necessarily perfectly match the information scholars need for their research. Consequently, scholars might be unable to find certain desired information from the database of digital by-product data. For instance, we were unable to find enough information from the digital by-product data set to identify a particular participant who was responsible for “sock puppetry” in a fully-protected article. Therefore, the visualization of fully-protected articles cannot represent the perfect visual picture to explain what “sock puppetry” is. In fact, such a problem is quite common when using digital by-product data, which could limit the scope of applying this method

Thirdly, even when using the same database, different analytical approaches and expressive means may generate diverse effects, which challenges scholars’ ability to design in appropriate ways. Based on our empirical studies, we also find that there are many options to represent results from analysing digital by-product data. On the one hand, this provides various ways to display the same data from an analysis of digital by-product data from different angles of research; on the other hand, it is challenging for scholars to choose the right representation approach. For instance, chapter six provides two different representational ways to analyse and display the same set of information on fully-protected articles. In fact, when chapter six was initially designed, visualization has been planned as the only way to display the information, and the correlation analysis was included when we found that visualization only provided descriptive results. In the process of applying digital by-product data, we may experience some incomprehensiveness of the research results due to a lack of certain data. These situations are unpredictable before the results are generated, and will to a certain extent increase research time.

It is important to point out that the limitations above are concurrent with the special features of digital by-product data. These features provide many convenient advantages for social science studies, but also bring about some problems and have limitations in the application process. These shortcomings should not overshadow the advantages of such a research methodology nor should they fundamentally deny the prospects of using this method. On the contrary, the limitations and the advantages of this method are caused by the exact same reasons. For example, digital by-product data ensures the objectiveness of the information collected; while it may not meet all of the requirements of the research project, or could incompletely address the research question. Therefore, based on a rational analysis of both the

advantages and disadvantages of using digital by-product data, we still argue that the method proposed in this thesis has unique and considerable advantages for resolving the methodological crisis faced by social scientists, and thus we encourage more social scholars to consider applying this proposed method in their research project, or at least let it be a part of the methodology in their studies with an awareness of its limitations.

7.3.4 The working flow of applying digital by-product data in social science

Having no relevant experience in using data mining, we directly borrowed the work flow of data mining in scientific studies to discuss the process of applying digital by-product data (Figure 1-1). The reason why we chose to use data mining directly is in part due to the fact that such a process can organize and analyse a large amount of data whose structure is often complicated, and at the same time, many scientific scholars generally choose to use data mining for sorting and analysing data from Wikipedia. However, for social scientists, the general process of data mining extracted from scientific studies only provides a relatively simple technical guidance from obtaining data to analysing results. In this section, through our own empirical work, we attempt to summarise the process based on data mining, which is designed specifically for social scientists to treat massive amounts of data in research as shown in Figure 7-1.

Figure 7-1 shows the working flow of the proposed method with digital by-product data extracted from our empirical research. We need to reiterate that this work flow is drawn based on the descriptions on data mining from scientific research and the practical experience that we gained through this thesis. From unnoticed data to the extraction of knowledge and further more widely accepted knowledge, there are four steps. First, we need to define the target. In order to avoid the difficulties and delays we experienced in our research plan due to a lack of preparation as mentioned in chapter six, social scientists have to design the research question while considering the accessibility of the data and the meaning behind that data.

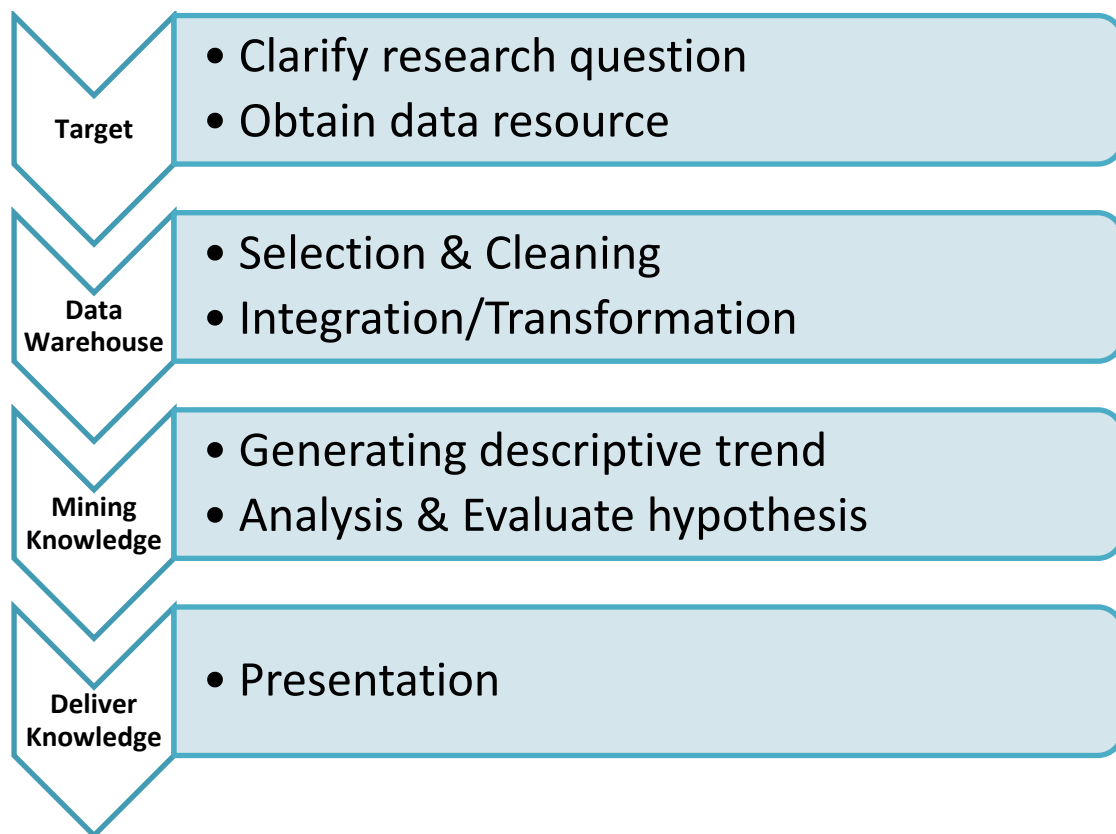


Figure 7-1 Proposed work flow from data to knowledge

Establishing the appropriate data warehouse is the second step in data mining. Building up the data warehouse is a vital and essential part of using digital by-product data, the establishment of which is based on scholars' understanding of the relationship between different data. Therefore, we first need to select the data needed for the research, and exclude irrelevant or erroneous data. Secondly, because data storage is designed for the convenience of storage and read-outs, it is often unsuitable for calculations and edits on a large scale. Therefore, such data need to be reformatted for easier editing. Such a choice depends on the judgements of the scholars, based on their technological background and available hardware resources.

Thirdly, data analysis is the main step of data mining. It is the process of discovering and extracting meaningful information from complicated and multitudes of data and organizing them into a systematic body of knowledge. Digital by-product data is especially suitable for descriptive analysis as they contain rich and comprehensive information on behaviour. Additionally, the interrelationship among the data and the patterns formed can help us to generate a series of hypotheses. This large amount of reliable information can be the most effective means to prove such hypotheses. In actual fact, through this section, we accomplish

the transformation of a large quantity of valueless data into knowledge with epistemological value.

Finally, it is an important step in research to allow the research results to be disseminated and communicated for the public to acknowledge and professionals to assess. Therefore, choosing an adequate representation manner to express the massive complicated-structured and intricately-linked results is key. Internet phenomena and the relevant interactions among participants are complicated and dynamic research subjects; therefore the results often contain huge quantities of data and various relationship patterns. One needs to pay attention to the choice of representation, as it needs to deliver the descriptive terms, but also, express the deeper meaning of the analysis.

From our practical experience, we outline a working flow of using data mining that is easy for social scientists to comprehend and accomplish. In fact, we reorganize the steps in traditional data mining, emphasise the understanding of data, planning the target, and the expressing of the results of the analysis, which are as important as the organization and analysis of results. Such reorganization is based on the lessons learnt from our experience of using digital by-product data.

In conclusion, this thesis contributes to the social sciences in two ways. First, the research discusses possible methodologies that could be employed by social scientists to study internet phenomena, and points out the limitation of the current research methodology. We further propose a possible methodology to solve such a methodological challenge. Second, although it is not the principle purpose, the empirical work not only demonstrates the working process of using digital by-product data, but also summarizes many interesting and important features of the unique collaborative model of Wikipedia.

7.4 Summary and reflection

Contemporary social scientists are facing a brand new social environment brought by new technologies. This new environment is described by various research ideas and social perspectives as “knowing capitalism” (Thrift, 2005), “network society” (Castells, 2000), and “knowledge production” from social networks (Benkler, 2006, Hippel, 2006). These researches have explained, from various angles, changes in society that are brought by new technology and new communication means. Information and communication technologies (ICTs) have not only expanded the communications means of people, but have also revolutionized life styles. Yet innovations in technology go beyond such scenarios. With the application and use of digital devices becoming more and more popular, digital data in a multitude of forms is an important part in every aspect of our lives.

In response to the challenge of observing the new digitalized society, we need to use a novel methodology to collect data for research. Based on a series of practical experiences on using digital by-product data and scrutinizing the outcome of this method, we are optimistic about its use. The move towards an innovative methodology can help us describe and measure newly emerging online phenomena. As Abbott puts it, “More and more things can be measured more and more often” (Abbott, 2000).

From the macro point of view, the combination of different types of archives and data in the post-digital era immensely improves the ability of social scientists to analyse and measure social phenomena. Fundamentally, data provides us with certain variables to better measure our society; interactions between people; economic trends; and various social movements. This thesis reveals just the tip of the iceberg regarding how to utilize different types of data and archives with varied degrees of openness to investigate our society and measure individual phenomena. From this point of view, the internet has moved social science research to a new frontier where data do the talking. While the natural sciences rely on data to discover nature, social sciences depend on data to “measure” society.

Based on the recognized digital innovation supported by ICTs to develop social scientists’ research capability, we first categorized the existing types of data from an angle of scientific utility. In fact, scholars pay special attention to ethical issues in producing and collecting data. From these two perspectives, we classified data into two types, as shown Table 2-1. On the one hand, the availability of online archives determines the accessibility of the data and subsequently the usability of the data for research; on the other, according to the model of data generation and data transmission, we can roughly classify data into by-product data and intentionally produced data. Through the classification of data from the internet given above, we try to depict a research environment based on using and applying such data. The description of this research environment clearly defines the types of data that can be used by scholars; moreover, it provides a comprehensive comparison of each data type to guide scholars. In this section, we recognize the potential of these four types of data in scientific research, yet we also emphasize the need to understand the risks of using certain data that contains private information. More importantly, we discover a type of data, digital by-product data, which not only avoids ethical problems, but also provides a relatively thorough record of behaviours.

Therefore, we propose that using digital by-product data might be an effective and economic solution from the quagmire of limited and biased research data. More importantly, from a micro point of view, in order to examine whether such data can provide a comprehensive and effective research method, we use Wikipedia as a case study to carry out a series of research

investigations. These researches and the examinations of the research process and the research data in one part provide some empirical results, which were concluded in the second part of this chapter; and in another part provide a more thorough introduction and explanation of how to use these data and the working flow of such a process, which were introduced in the third part of this chapter.

We need to point out here that social scientists are facing a bright new trend of digitalism along with the active development of ICTs. Contemporary life is becoming more and more supported by digital devices, and therefore the description of contemporary life begins to be based on data accumulation and data transformation generated and created on different devices. Thus, we propose a new methodology based on such an overflowing resource. Unlike with the theoretical engagement, we elaborate the entire working flow and evaluate the advantages and limitations via applying this method to the real research topic of Wikipedia.

However, our research only concerns the possibility of using such methods and the superficial benefits they could bring. Data are accumulated through everyday life and the best way to harvest them is not a simple problem to solve. As a new academic method, how to treat and continue to use archive data is a question that demands investigation for the long-term and from different angles. The study of the collaborative model in Wikipedia can encourage social scientists to make use of such a data resource with more confidence, and it also provides them with some practical experiences to learn from. Because the spectrum of topics in social sciences is broad and the research angles vary, our research may not cover all individual types or areas. How to use these resources for specific research purposes; how to better data mine; and whether our work flow is suitable for all other social science topics, are questions that require further investigation. Needless to say, as the by-product data produced by a wide application of technologies and their respective products further fills our lives, they also bring immense challenges and valued opportunities.

Bibliography

- ABBOTT, A. 2000. Reflections on the future of sociology. *Contemporary Sociology*, 29 (2), 296-300.
- ADAMIC, L. & ADAR, E. 2005. How to Search a Social Network. *Social Networks*, 27 (3), 187-203.
- ADAMIC, L. & GLANCE, N. 2005. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. *Proceedings of The 3rd International Workshop on Link Discovery*, Chicago, Illinois. ACM, 36-43.
- ADLER, B. T. & ALFARO, L. D. 2007. A content-driven reputation system for the wikipedia. *Proceedings of The 16th international conference on World Wide Web*, Banff, Alberta, Canada. 1242608: ACM, 261-270.
- ADLER, B. T., CHATTERJEE, K., ALFARO, L. D., FAELLA, M., PYE, I. & RAMAN, V. 2008. Assigning trust to Wikipedia content. *Proceedings of The 4th International Symposium on Wikis*, Porto, Portugal. ACM, 1-12.
- AGARWAL, A. 2009. Wikipedia Editors are leaving Wikipedia [Online][Accessed 2011/5/26] <http://www.labnol.org/internet/wikipedia-editors-leaving/11265/>
- AHN, D., JIJKOUN, V., MISHNE, G., MULLER, K., RIJKE, M. D. & SCHLOBACH, S. 2005. Using Wikipedia at the TREC QA Track. *Proceedings of The thirteenth text retrieval conference (TREC 2004)*, Gaithersburg, Maryland, US.
- AINGE, D. J. 1996. Upper Primary Students Constructing and Exploring Three Dimensional Shapes. *Journal of Educational Computing Research*, 14, 345-369.
- ALFARANO, C., ANDRADE, C. E., ANTHONY, K., BAHROOS, N., BAJEC, M., BANTOFT, K., BETEL, D., BOBECHKO, B., BOUTILIER, K., BURGESS, E., BUZADZIJA, K., CAVERO, R., D'ABREO, C., DONALDSON, I., DORAIRAJOO, D., DUMONTIER, M. J., DUMONTIER, M. R., EARLES, V., FARRALL, R., FELDMAN, H., GARDERMAN, E., GONG, Y., GONZAGA, R., GRYSAN, V., GRYZ, E., GU, V., HALDORSEN, E., HALUPA, A., HAW, R., HRVOJIC, A., HURRELL, L., ISSERLIN, R., JACK, F., JUMA, F., KHAN, A., KON, T., KONOPINSKY, S., LE, V., LEE, E., LING, S., MAGIDIN, M., MONIAKIS, J., MONTOJO, J., MOORE, S., MUSKAT, B., NG, I., PARAISO, J. P., PARKER, B., PINTILIE, G., PIRONE, R., SALAMA, J. J., SGRO, S., SHAN, T., SHU, Y., SIEW, J., SKINNER, D., SNYDER, K., STASIUK, R., STRUMPF, D., TUEKAM, B., TAO, S., WANG, Z., WHITE, M., WILLIS, R., WOLTING, C., WONG, S., WRONG, A., XIN, C., YAO, R., YATES, B., ZHANG, S., ZHENG, K., PAWSON, T., OUELLETTE, B. F. F. & HOGUE, C. W. V. 2005. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Research*, 33 (suppl 1), D418-D424.

- ALMEIDA, R. B., MOZAFARI, B. & CHO, J. 2007. On the Evolution of Wikipedia International Conference on Weblogs and Social Media, 2007 March, 26-28, 2007 Boulder, Colorado, USA.
- ANDERSON, E. W., PRESTON, G. A. & SILVA, C. A. T. 2010. Using Python for Signal Processing and Visualization. *Computing in Science and Engineering*, 12 (4), 90-95.
- ANDERSON, P. 2007. What is Web 2.0? Ideas, technologies and implications for education. JISC Technology and Standards Watch. Available: <http://www.jisc.ac.uk/media/documents/techwatch/tsw0701b.pdf>.
- ANDERSON, S. E. & GANSNEDER, B. M. 1995. Using electronic mail surveys and computer-monitored data for studying computer-mediated communication systems. *Social Science Computer Review*, 13 (1), 33-46.
- ANGWIN, J. & FOWLER, G. A. 2009. Volunteers Log Off as Wikipedia Ages. *The Wall Street Journal*, November 25, 2009.
- ANTHONY, D., SMITH, S. W. & WILLIAMSON, T. 2005. Explaining Quality in Internet Collective Goods: Zealots and Good Samaritans in the Case of Wikipedia In Fall 2005 Innovation & Entrepreneurship Seminar, MIT, America.
- ANTIN, J. & CHESHIRE, C. 2010. Readers are not free-riders: reading as a form of participation on wikipedia. Proceedings of the 2010 ACM conference on Computer supported cooperative work, Savannah, Georgia, USA. 1718942: ACM, 127-130.
- ARAZY, O., MORGAN, W. & PATTERSON, R. 2006. Wisdom of the Crowds: Decentralized Knowledge Construction in Wikipedia 16th Annual Workshop on Information Technologies & Systems (WITS), Phoenix, Arizona, USA.
- ARDS, S., CHUNG, C. & MYERS, S. J. 1998. The effects of sample selection bias on racial differences in child abuse reporting. *Child Abuse and Neglect*, 22 (2), 103-115.
- AVITAL, M., SAWYER, S., KRAEMER, K., SAMBAMURTHY, V., LYYTINEN, K. & IACONO, C. S. 2007. Data Rich and Data Poor Scholarship: Where Does IS Research Stand. Proceedings of ICIS 2007 Proceedings, Jeju Island, South Korea.
- AYERS, P., MATTHEWS, C. & YATES, B. 2008. How Wikipedia Works and How you can be a part of it, San Francisco: No Starch Press.
- BACHMANN, I., KAUFHOLD, K., LEWIS, S. C. & ZUNIGA, H. G. D. 2010. News Platform Preference: Advancing the effects of age and media consumption on political participation. *International Journal of Internet Science*, 5 (1), 34-47.
- BADER, G., DONALDSON, I., WOLTING, C., OUELLETTE, B., PAWSON, T. & HOGUE, C. 2001. BIND--The Biomolecular Interaction Network Database. *Nucleic Acids Research*, 29 (1), 242-245.
- BAI, Y. & HE, Z. 2010. He says, She Says: The dynamic rumour process on the internet forum The Asian Conference on Social Sciences, Osaka, Japan.
- BAILLET, S., MOSHER, J. C. & LEAHY, R. M. 2001. Electromagnetic Brain Mapping. *Signal Processing Magazine, IEEE*, 18 (6), 14-30.
- BAINBRIDGE, W. S. 2000. New Technologies for the Social Sciences. In: RENAUD, M. (ed.) *Social Sciences for a Digital World: building infrastructure and databases for the future*. Paris: OECD.

- BAITALUK, M., SEDOVA, M., RAY, A. & GUPTA, A. 2006. Biological Networks: visualization and analysis tool for systems biology. *Nucleic Acids Research*, 34(suppl 2), W466-W471.
- BANE, A. F. & MILHEIM, W. D. 1995. Internet insights: How academics are using the internet. *Computers in Libraries*, 15 (2), 32-36.
- BANKS, M. 2001. *Visual Methods in Social Research*, London: Sage Publications Ltd.
- BARBIER, G. & LIU, H. 2011. Data Mining in Social Media. In: AGGARWAL, C. C. (ed.) *Social Network Data Analytics*. Boston, Dordrecht, London: Springer Science and Business Media.
- BARNETT, E. 2009. Wikipedia's Jimmy Wales denies site is 'losing' thousands of volunteer editors. *Telegraph*, 26 November 2009.
- BARRETT, N. J. 2008. *MediaWiki (Wikipedia and Beyond)*, Sebastopol, CA: O' Reilly Media.
- BATAGELJ, V. & MRVAR, A. 1998. Pajek--Program for large network analysis. *Connections*, 21, 47-57.
- BAYM, N. K. 1995. The Emergence of Online Community. 138-163. In: JONES, S. G. (ed.) *CyberSociety 2.0: Revisiting Computer-Mediated Communication and Community*. Thousand Oaks, Calif; London: Sage Publications.
- BAYM, N. K. & LEDBETTER, A. 2009. Tunes that Bind? . *Information, Communication and Society*, 12 (3), 408-427.
- BAYM, N. K., ZHANG, Y. B. & LIN, M.-C. 2004. Social interactions across media: Interpersonal communication on the internet, telephone and face-to-face. *New Media & Society*, 6 (3), 299-318.
- BEAUMONT, G. 1983. *Introduction to Neuropsychology*, New York: The Guilford Press.
- BEER, D. & BURROWS, R. 2010. Consumption, Prosumption and Participatory Web Cultures. *Journal of Consumer Culture*, 10 (1), 3-12.
- BEER, D. & BURROWS, R. Forthcoming. Popular Culture, Digital Archives and the New Social Life of Data. *Theory Culture & Society*.
- BELL, C. & NEWBY, H. (eds.) 1977. *Doing Sociological Research*, London: George Allen & Unwin.
- BELL, C. & ROBERTS, H. (eds.) 1984. *Social Researching*, London: Routledge & Kegan Paul.
- BELLOMI, F. & BONATO, R. 2005. Network analysis for Wikipedia. *Proceedings of Wikimania 2005—The First International Wikimedia Conference*, Frankfurt am Main, Germany.
- BENKLER, Y. 2002. Coase's penguin, or, Linux and the nature of the firm. *The Yale Law Journal*, 112 (3), 369-446.
- BENKLER, Y. 2006. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*, New Haven, London: Yale University Press.
- BERNERS-LEE, T. & FISCHETTI, M. 1999. *Weaving the Web: the Past, Present and Future of the World Wide Web by its Inventor*, New York: HarperCollins.
- BERNERS-LEE, T., HENDLER, J. & LASSILA, O. 2001. The Semantic Web. *Scientific American*, 17 May 2001.
- BERTRAND, M. & MULLAINATHAN, S. 2001. Do People Mean What They Say? Implications for subjective survey data. MIT Economics Working Paper, 01-04.

- BESCHASTNIKH, I., KRIPLEAN, T. & MCDONALD, D. W. 2008. *Wikipedian Self-Governance in Action: Motivating the Policy Lens* the 2008 AAAI International Conference on Weblogs and Social Media, Seattle, Washington, USA.
- BEST, S. J. & KRUEGER, B. S. 2004. *Internet Data Collection* London: Sage Publications Ltd.
- BEST, S. J., KRUEGER, B. S., HUBBARD, C. & SMITH, A. 2001. An assessment of the generalizability of internet surveys. *Social Science Computer Review*, 19 (2), 131-145.
- BIRD, D. N/A. Wiki, the story behind the world's most controversial encyclopedia. Available: <http://www.subter.com/is/?p=485#comments> [Accessed 20, May 2011].
- BOASE, J. & WELLMAN, B. 2004. Personal relationships: on and off the internet. In: PERLMAN, D. & VANGELISTI, A. L. (eds.) *Handbook of Personal Relations*. Oxford: Blackwell.
- BOULOS, M. N. K., MARAMBA, I. & WHEELER, S. 2006. Wikis, blogs and podcasts: a new generation of Web-based tools for virtual collaborative clinical practice and education. *BMC Medical Education* 2006, 6 (41).
- BOYD, D. M. & ELLISON, N. B. 2007. Social Network Sites: Definition, History and Scholarship. *Journal of Computer Mediated Communication*, 13 (1), 210-230.
- BRAENDLE, A. 2005. Many cooks do not spoil the broth. *Proceedings of Wikimania 2005-The First International Wikimedia Conference*, Frankfurt Germany.
- BRAITHWAITE, D., EMERY, J., DE LUSIGNAN, S. & SUTTON, S. 2003. Using the Internet to conduct surveys of health professionals: a valid alternative? *Family Practice*, 20 (5), 545-551.
- BRANDES, U. & WAGNER, D. 2004. Vision-Analysis and Visualization of Social Networks. 321-340. *Graph Drawing Software*. Springer-Verlag.
- BRAUN, S. & SCHMIDT, A. 2007. Wikis as a Technology Fostering Knowledge Maturing: What we can learn from Wikipedia. *Proceedings of the 7th International Conference on Knowledge Management (I-KNOW 2007)*, Special Track on Integrating Working and Learning in Business (IWL), Graz, Austria.
- BROWN, J. S. & ADLER, R. P. 2008. Minds on fire: Open education, the long tail, and learning 2.0. *EDUCAUSE Review*, 43 (1).
- BRUNN, S. D. & DODGE, M. 2001. Mapping the "Worlds" of the World Wide Web: Restructuring Global Commerce Through Hyperlinks. *American Behavioral Scientist*, 44 (10), 1717-1739.
- BRYANT, S. L., FORTE, A. & BRUCKMAN, A. 2005. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, Sanibel Island, Florida, USA.
- BUDANITSKY, A. 1999. *Lexical Semantic Relatedness and Its Application in Natural Language Processing*. Unpublished thesis. In: Department of Computer Science, University Of Toronto.
- BUDANITSKY, A. & HIRST, G. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32 (1), 13-47.
- BUNESCU, R. & PASCA, M. 2006. Using Encyclopedic Knowledge for Named Entity Disambiguation the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06), April 2006 Trento, Italy.

- BURKE, M. & KRAUT, R. 2008. Mopping up: modeling wikipedia promotion decisions. Proceedings of the 2008 ACM conference on Computer supported cooperative work, San Diego, CA, USA.
- BURKE, M. & KRAUT, R. 2008. Taking up the mop: identifying future wikipedia administrators CHI '08, Florence, Italy.
- BUTLER, B., JOYCE, E. & PIKE, J. 2008. Don't Look Now, But We've Created a Bureaucracy: The Nature and Roles of Policies and Rules in Wikipedia. Proceedings of CHI 2008, Florence, Italy.
- BUTLER, B. S. 2001. Membership Size, Communication Activity, and Sustainability: A Resource-Based Model of Online Social Structures. *Journal of Information Systems Research*, 12 (4), 346-362.
- CAPLAN, S. E. 2003. Preference for Online Social Interaction: A Theory of Problematic Internet Use and Psychosocial Well-Being. *Communication Research*, 30 (6), 625-648.
- CAPOCCI, A., RAO, F. & CALDARELLI, G. 2008. Taxonomy and clustering in collaborative systems: The case of the on-line encyclopedia Wikipedia. *EPL (Europhysics Letters)*, 81 (2), 28006.
- CAPOCCI, A., SERVEDIO, V. D. P., COLAIORI, F., BURIOL, L. S., DONATO, D., LEONARDI, S. & CALDARELLI, G. 2006. Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. *PHYSICAL REVIEW*, E74 (3), 036116.
- CHAN, B., WU, L., TALBOT, J., CAMMARANO, M. & HANRAHAN, P. 2008. Vispedia: Interactive Visual Exploration of Wikipedia Data via Search-Based Integration. *IEEE Transactions on Visualization and Computer Graphics*, 14 (6), 1213-1220.
- CHEN, P. & HINTON, S. M. 1999. Realtime Interviewing Using the World Wide Web. *Sociological Research Online* [Online], 4. Available: <http://www.socresonline.org.uk/4/3/chen.html>.
- CHENG, L. 2010. Visualizing brain images in an undergraduate signal processing course. Proceedings of Bioengineering Conference, Proceedings of the 2010 IEEE 36th Annual Northeast New York, USA. 1-2.
- CHESNEY, T. 2006. An Empirical Examination of Wikipedia's Credibility. *First Monday*, 11 (11), 1-11.
- CHI, E. H.-H., REIDL, J., SHOOP, E., CARLIS, J. V., RETZEL, E. & BARRY, P. 1996. Flexible Information Visualization of Multivariate Data from Biological Sequence Similarity Searches. Proceedings of the 7th IEEE Visualization Conference (VIS'96), San Francisco, CA, USA. 133-140.
- CHMIEL, A., SIENKIEWICZ, J., THELWALL, M., PALTOGLOU, G., BUCKLEY, K., KAPPAS, A. & HOLYST, J. A. 2011. Collective emotions online and their influence on community life. *PLoS ONE*, 6 (7), e22207.
- CHO, H. & LAROSE, R. 1999. Privacy Issues in Internet Surveys. *Social Science Computer Review*, 17 (4), 421-434.
- CHRISTODOULOU, E. G., SAKKALIS, V., TSIARAS, V. & TOLLIS, I. G. 2011. BrainNetVis: An Open-Access Tool to Effectively Quantify and Visualize Brain Networks. *Computational Intelligence and Neuroscience*, 2011 (2011), 10 - 22.
- CIFFOLILLI, A. 2003. Phantom authority, self-selective recruitment and retention of members in virtual communities: The case of Wikipedia *First Monday*, 8 (12).

- CODERRE, F., MATHIEU, A. & ST-LAURENT, N. 2004. Comparison of the quality of qualitative data obtained through telephone, postal and email surveys. *International journal of Market Research*, 46 (3), 347-357.
- COHEN, N. 2007. A History Department Bans Citing Wikipedia as a Research Source. *The New York Times*, February 21, 2007.
- COOMBER, R. 1997. Using the Internet for Survey Research. *Sociological Research Online* [Online], 2. Available: <http://www.socresonline.org.uk/2/2/2.html>.
- COUPER, M. P. 2000. Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, 64 (1), 464-494.
- COUPER, M. P. & MILLER, P. V. 2008. Web Survey Methods Introduction. *Public Opinion Quarterly*, 72 (5), 831-835.
- COUPER, M. P., TRAUGOTT, M. W. & LAMIAS, M. J. 2001. Web survey design and administration. *Public Opinion Quarterly*, 65 (2), 230-253.
- CRUTCHER, M. & ZOOK, M. 2009. Placemarks and waterlines: Tacialized cyberscapes in post-Katrina Google Earth. *Geoforum*, 40 (4), 523-534.
- CUCERZAN, S. 2007. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. *Proceedings of EMNLP 2007: Empirical Methods in Natural Language Processing*, Prague, Czech Republic. 708 - 716.
- DAHLQUIST, K. D., SALOMONIS, N., VRANIZAN, K., LAWLOR, S. C. & CONKLIN, B. R. 2002. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genetics*, 31 (2002), 19-20.
- DALE AM., A., LIU, B., FISCHL, R., BUCKNER, J., BELLIVEAU, J. L. & HALGREN, E. 2000. Dynamic Statistical Parametric Mapping: Combining fMRI and MEG for highresolution imaging of cortical activity. *Nueron*, 26 (2000), 55-67.
- DAMME, C. V., HEPP, M. & SIORPAES, K. 2007. FolksOntology: An intergrated approach for turning Folksonomies into Ontologies In *Bridging the gap between semantic web and web2.0*, Innsbruck, Austria.
- DEE, J. 2007. All the News That's Fit to Print Out. *The New York Times*.
- DEFANTI, T. A., BROWN, M. D. & MCCORMICK, B. H. 1989. Visualization: Expanding Scientific and Engineering Research Opportunities. *Computer*, 22 (8), 12-25.
- DENNING, P., HORNING, J., PARNAS, D. & WEINSTEIN, L. 2005. Wikipedia Risks. *Communication of the ACM*, 48 (12).
- DENOYER, L. & GALLINARI, P. 2006. The Wikipedia XML corpus. *ACM SIGIR Forum*, 40 (1), 64-69.
- DODGE, M., MCDERBY, M. & TURNER, M. 2008. *Geographic Visualization: Concepts, Tools and Applications*: Wiley.
- DUFFY, P. 2008. Engaging the YouTube Google-eyed generation: Strategies for using Web 2.0 in teaching and learning. *Electronic Journal of e-Learning*, 6 (2), 119-129.
- DUFFY, P. & BRUNS, A. 2006. The Use of Blogs, Wikis and RSS in Education: A Conversation of Possibilities. *Proceedings of Online Learning and Teaching Conference 2006*. 31-38.
- DUTTON, W. H. & SHEPHERD, A. 2006. Trust in the Internet as an Experience Technology. *Information, Communication and Society*, 9, 433-451.
- DWYER, C., HILTZ, S. & PASSERINI, K. 2007. Trust and Privacy Concern Within Social Networking Sites: A Comparison of Facebook and MySpace. *Proceedings of*

- 2007 Americas Conference on Information Systems (AMCIS), Keystone, Colorado, USA. 13 - 15.
- EAGLE, N., PENTLAND, A. & LAZER, D. 2009. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106 (36), 15274-15278.
- EASTERBY-SMITH, M., THORPE, R. & JACKSON., P. R. 2008. *Management research*, London: SAGE.
- EBNER, M. & ZECHNER, J. 2006. Why is Wikipedia so Successful? Experiences in Establishing the Principles in Higher Education. *Proceedings of I-KNOW' 06*, 6-8 September, 2006 Graz, Austria.
- EDWARDS, L. 2009. Reports claims Wikipedia losing editors in droves. November, 30, 2009. Available: <http://www.physorg.com/news178787309.html>.
- ELLISON, N., HEINO, R. & GIBBS, J. 2006. Managing Impressions Online: Self-Presentation Processes in the Online Dating Environment. *Journal of Computer-Mediated Communication*, 11 (2), 415-441.
- ELLISON, N., STEINFELD, C. & LAMPE, C. 2007. The Benefits of Facebook "Friends": Social Capital and College Students' Use of Online Social Network Sites. *Journal of Computer-Mediated Communication*, 12 (4), 1143-1168.
- EMIGH, W. & HERRING, S. C. 2005. Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias. *Proceedings of the Thirty-Eighth Hawaii International Conference on System Sciences HICSS-38*, Hawaii, U.S.A.
- ENGEL, K., HASTREITER, P., TOMANDL, B., EBERHARDT, K. & ERTL, T. 2000. Abstract Combining Local and Remote Visualization Techniques for Interactive Volume Rendering in Medical Applications. *Proceedings of VIS '00 Proceedings of the conference on Visualization '00* Los Alamitos.
- ERICKSON, T. & HERRING, S. 2005. Persistent Conversation: A Dialog Between Research and Design. *Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS 2005)*, Big Island, HI, USA. IEEE Press.
- EVANS, A. R. & MATHUR, A. 2005. The value of online surveys. *Internet Research*, 13 (2), 195-219.
- FACHRY, K. N., KAMPS, J., KOOLEN, M. & ZHANG, J. 2007. Using and Detecting Links in Wikipedia 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007 December 17-19, 2007 Dagstuhl Castle, Germany.
- FACKENHEIM, E. L. 1970. On the actuality of the rational and the rationality of the actual. *The Review of Metaphysics*, 23 (4), 690-698.
- FALLIS, D. 2008. Toward an epistemology of Wikipedia. *Journal of the American Society for Information Science and Technology*, 59 (10), 1662-1674.
- FEATHERSTONE, M. 2006. Genealogies of the Global. *Theory, Culture & Society*, 23 (2-3), 387-392.
- FEATHERSTONE, M. & VENN, C. 2006. Problematizing Global Knowledge and the New Encyclopaedia Project: An Introduction. *Theory, Culture & Society*, 23 (2-3), 1-20.
- FEATHERSTONE, M., VENN, C., BISHOP, R. & PHILIPS, J. (eds.) 2006. *Theory Culture & Society*.
- FELT, A. & EVANS, D. 2008. Privacy Protection for Social Networking Platforms Workshop on Web 2.0 Security and Privacy, Oakland, California, USA.

- FINKELSTEIN, L., GABRILOVICH, E., MATIAS, Y., RIVLIN, E., SOLAN, Z., WOLFMAN, G. & RUPPIN, E. 2002. Placing Searching in Context: the concept revisited. *ACM Transactions on Information Systems*, 20 (1), 116-131.
- FISHER, D., BRUSH, A. J., GLEAVE, E. & SMITH, M. A. 2006. Revisiting Whittaker & Sidner's "email overload" Ten Years Later. *Proceedings of the 20th Anniversary Conference on Computer Supported Cooperative Work*, Banff, Alberta, Canada.
- FISHER, D., SMITH, M. & WELSER, H. T. 2006. You Are Who You Talk To: Detecting Roles in Usenet Newsgroups *Proceedings of the 39th Annual Hawaii International Conference on System Sciences - Volume 03*.
- FISHER, P. 2007. The Linear Medical Model of Disability: Mothers of Disabled Babies Resist with Counter-Narratives. *Sociology of Health & Illness*, 29 (1), 66-81.
- FLORIDI, L. 2009. Web 2.0 vs. the Semantic Web: A Philosophical Assessment. *Episteme*, 6 (1), 25-27.
- FOLEY, J., DAM, A. V., FEINER, S. K. & HUGHES, J. F. 1997. *Computer Graphics: Principles and Practice*: Addison-Wesley Publishing Company.
- FORTE, A. & BRUCKMAN, A. 2006. From Wikipedia to the classroom: exploring online publication and learning. *Proceedings of the 7th international conference on Learning sciences*, Bloomington, Indiana. 182 - 188.
- FORTE, A. & BRUCKMAN, A. 2008. Scaling consensus: increasing decentralization in Wikipedia governance *Hawaiian International Conference of Systems Sciences (HICSS)*, Hawaiian, USA.
- FORTE, A., LARCO, V. & BRUCKMAN, A. 2009. Decentralization in Wikipedia Governance. *Journal of Management Information Systems*, 26 (1), 49-72.
- FOTHERINGHAM, A. S., BRUNSDON, C. & CHARLTON, M. 2000. *Quantitative Geography: Perspectives on Spatial Data Analysis*, London: SAGE Publications.
- FRANKEL, M. S. & TEICH, A. 1999. *Anonymous communication on the internet*, London: Taylor & Francis.
- FREEMAN, L. C. 1992. Filling in the blanks: A theory of cognitive categories and the structure of social affiliation. *Social Psychology Quarterly*, 55 (2), 118-127.
- FREEMAN, L. C. 2000. Visualizing Social Networks. *Journal of Social Structure*, 1, 13 - 18.
- FREEMAN, L. C., ROMNEY, A. & FREEMAN, S. 1987. Cognitive structure and information accuracy. *American Anthropologist*, 89 (2), 310-325.
- FRENSCH, P. 1994. Composition during serial learning: A serial position effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20 (2), 423-443.
- GABRILOVICH, E. & MARKOVITCH, S. 2006. Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. *Proceedings of the 21st National Conference on Artificial Intelligence*, Boston, Massachusetts. 1301-1306.
- GABRILOVICH, E. & MARKOVITCH, S. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis *Proceedings of the 20th international joint conference on Artificial intelligence*, Hyderabad, India.
- GABRILOVICH, E. & MARKOVITCH, S. 2009. Wikipedia-based Semantic Interpretation for Natural Language Processing. *Journal of Artificial Intelligence Research* 2009 (34), 443-498.
- GALLAGHER, R. (ed.) 1995. *Computer Visualization: Graphics Techniques for Scientific and Engineering Analysis*: CRC Press.

- GALLAGHER, R. 1995. Scalar Visualization Techniques. In: GALLAGHER, R. (ed.) Computer Visualization. CRC Press.
- GALLAGHER, R. 1995. Scientific Visualization: An Engineering Perspective. In: GALLAGHER, R. (ed.) Computer Visualization. CRC Press.
- GANS, H. J. 1967. The Levittowners, London: Allen Lane-The Penguin Press.
- GERSHON, N., EICK, S. & CARD, S. 1998. Information Visualization. *Interactions*, 5 (2), 9-15.
- GILES, J. 2005. Internet Encyclopaedias Go Head to Head. *Nature*, 438, 900-901 <http://www.jimgiles.net/pdfs/InternetEncyclopaedias.pdf>.
- GOODCHILD, M. F. 2007. Citizens as Seasors: the World of Volunteered Geography. *GeoJournal*, 69 (4).
- GOSLING, S. D., VAZIRE, S., SRIVASTAVA, S. & JOHN, O. P. 2004. Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, 59 (2), 93-104.
- GOUTHRO, L. 2000. Building the World's Biggest Encyclopedia. *PCWorld*, Mar, 10, 2000.
- GRANELLO, D. H. & WHEATON, J. E. 2004. Online data collection: Strategies for research. *Journal of Counseling & Development* 82 (387-393).
- GUTIERREZ, M. A. A., VEXO, F. & THALMANN, D. 2008. *Stepping into Virtual Reality* London, Springer.
- GUTMANN, M. P., ABRAHAMSON, M., ADAMS, M. & ALTMAN, M. 2009. From Preserving the Past to Preserving the Future: The Data-PASS Project and the Challenges of Preserving Digital Social Science Data. *Library Trends*, 57 (3), 315-337.
- HADDADI, H., HUI, P., HENDERSON, T. & BROWN, I. 2011. Targeted Advertising on the Handset: Privacy and Security Challenge. *Pervasive Advertising*. Springer.
- HALAVAIS, A. 2000. National Borders on the World Wide Web. *New Media & Society*, 2 (1), 7-28.
- HALAVAIS, A. & LACKAN, D. 2008. An analysis of topical coverage of Wikipedia. *Journal of Computer Mediated Communication*, 13 (2), 429-440.
- HALFORD, S. & SAVAGE, M. 2010. Reconceptualizing digital social inequality *Information, Communication & Society*, 13 (7), 937-955.
- HALL, C. M., MCMULLEN, S. A. H., HALL, D. L., MCMULLEN, M. J. & PURSEL, B. K. 2008. Perspectives on Visualization and Virtual World Technologies for Multi-sensor Data Fusion Information Fusion, 2008 11th International Conference on Jun 30 - Jul 3 Cologne, Germany.
- HALPIN, B. 1999. Simulation in Sociology. *American Behavioral Scientist*, 42 (10), 1488-1508.
- HALSEY, A. H. 2004. *A History of Sociology in Britain: Science, Literature, and Society*, Oxford: Oxford University Press.
- HAMPTON, K. N. 1999. Computer Assisted Interviewing The Design and Application of Survey Software to the Wired Suburb Project. *Bulleting de Methode Sociologique*, 62 (2), 49-68.
- HAN, J. & KAMBER, M. 2001. *Data Mining: Concepts and Techniques*, San Francisco, London: Morgan Kaufmann Publishers.

- HAND, D. J., KOK, J. N. & BERTHOLD, M. (eds.) 1999. *Advances in intelligent data analysis : third international symposium, IDA-99*, Amsterdam, The Netherlands, August 1999 : proceedings, Berlin; London: Springer.
- HAND, D. J., MANNILA, H. & SMYTH, P. 2001. *Principles of data mining*, Cambridge, Mass. ; London: MIT Press.
- HARGITTAI, E. 2007. Whose Space? Differences among users and non-users of social network sites. *Journal of Computer Mediated Communication*, 13 (1), 276-297.
- HARTFORD, K., CAREY, R. & MENDONCA, J. 2007. Sampling bias in an international internet survey of diversion programs in the criminal justice system. *Evaluation and the health professions*, 30 (1), 35-46.
- HE, Z. 2008. *Wikipedia and Wikipedians* ICS-PG-conference, Leeds, UK.
- HE, Z. 2010. *Online to Offline: Democratic Innovation in the Offline Development of Wikipedia*. Proceedings of 18th Biennial and Silver Anniversary Conference of ITS, Tokyo, Japan.
- HE, Z. 2010. *Visualizing the mass collaboration of Wikipedia by adopting digital by-product data*. Proceedings of CRESC conference, Oxford, UK.
- HEER, J. & HELLERSTEIN, J. M. 2009. Data visualization and social data analysis. *Proc. VLDB Endow.*, 2 (2), 1656-1657.
- HEGEL, G. H. W. & DYDE, S. W. 2008. *Philosophy of Right: Cosimo*.
- HENDRIKS, P. 1999. Why share knowledge? The influence of ICT on the motivation for knowledge sharing. *Knowledge and Process Management*, 6 (2), 91-100.
- HERRING, S. C., KOUPER, I., PAOLILLO, J. C., SCHEIDT, L. A., TYWORTH, M., WELSCH, P., WRIGHT, E. & YU, N. 2005. *Conversations in the Blogosphere: An Analysis "From the Bottom Up"* Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 4 - Volume 04.
- HERRING, S. C., SCHEIDT, L. A., BONUS, S. & WRIGHT, E. 2004. *Bridging the Gap: A Genre Analysis of Weblogs* Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 4 - Volume 4.
- HEYLIGHEN, F. 2006. *Why is Open Access Development so Successful? Stigmergic organization and the economics of information*.
- HINCHLIFFE, S. & WOODWARD, K. 2000. *The Natural and the Social: Uncertainty, Risk, Change*, London: Routledge.
- HINDUJA, S. & PATCHIN, J. W. 2008. *Bullying Beyond the Schoolyard: Preventing and Responding to Cyberbullying*, London: Corwin Press.
- HINE, C. 2005. *Virtual methods: issues in social research on the Internet*, Oxford: Berg.
- HINE, C. 2006. *New infrastructures for knowledge production: understanding E-science*, Hershey, PA ; London: Information Science Pub.
- HINE, C. 2008. *Systematics as cyberscience : computers, change, and continuity in science*, Cambridge, Mass. ; London: MIT.
- HIPPEL, E. V. 2006. *Democratizing innovation*, Cambridge, MA ; London: MIT Press.
- HOADLEY, C. M. & ENYEDY, N. 1999. *Between Information and Communication: Middle Spaces in Computer Media for Learning Computer Support for Collaborative Learning 1999*, Palo Alto, California

- HOLLOWAY, T., BOZICEVIC, M. & BORNER, K. 2007. Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors. *Complexity Wiley Periodicals*, 12 (3), 30 - 40.
- HOWARD, P. N. 2002. Network Ethnography and the Hypermedia Organization: New Media, New Organization, New Methods. *New Media & Society*, 4 (4), 550-574.
- HOX, J. & LEEUW, E. D. 1994. A Comparison of Nonresponse in Mail, Telephone, and Face-to-face Surveys: Applying Multilevel Modeling to Meta-analysis. *Quality & Quantity*, 28 (4), 329-344.
- HU, M., LIM, E.-P., SUN, A., LAUW, H. W. & VUONG, B.-Q. 2007. Measuring article quality in wikipedia: models and evaluation Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, Lisbon, Portugal.
- HU, Z., MELLOR, J., WU, J. & DELISI, C. 2004. VisANT: an online visualization and analysis tool for biological interaction data. *BMC Bioinformatics*, 2004 (5), 17.
- HUNT, L. A. 2007. *Inventing human rights: a history*: W. W. Norton.
- ILIEVA, J., BARON, S. & HEALEY, N. M. 2002. Online surveys in marketing research: pros and cons. *International Journal of Market Research*, 44 (3), 361-376.
- ILLINGWORTH, N. 2001. The Internet Matters: Exploring the Use of the Internet as a Research Tool. *Sociological Research Online*, 6 (2), 30-34.
- INMON, W. H. 2002. *Building the data warehouse*, New York ; Chichester: Wiley.
- IRANI, M. & PELEG, S. 1991. Improving resolution by image registration. *Graphical, Models and Image Processing*, 53 (3), 231-239.
- JANETZKO, D. 2001. Processing Raw Data both the Qualitative and Quantitative Way. *Qualitative and Quantitative Research: Conjunctions and Divergences*, 2 (1), 11-14.
- JAVANMARDI, S. & LOPES, C. 2010. Statistical measure of quality in Wikipedia 1st Workshop on social media analytics (SOMA '10), 25 July 2010 Washington, DC, USA.
- JENKINS, S. P. & LAMBERT, P. J. 1997. Three I's of Poverty Curves, with an Analysis of U.K. Poverty Trends. *Oxford Economic Papers: New Series*, 49 (3), 317-327.
- JOHNSON, K. A. 2001. *Neuroimaging Primer* [Online]. Harvard Medical School. Available: <http://www.med.harvard.edu/AANLIB/hms2.html>.
- JOINSON, A. 2008. Looking At, Looking Up or Keeping Up with People? Motives and Use of Facebook. Proceedings of the Twenty-sixth annual SIGCHI conference on Human factors in Computing Systems, Florence, Italy. ACM Press, 1573 - 1582.
- JONES, R. A. 1994. The Ethics of Research in Cyberspace. *Internet Research* 4(3), 30-35.
- KANEHISA, M., GOTO, S., KAWASHIMA, S. & NAKAYA, A. 2002. The KEGG databases at GenomeNet. *Nucleic Acids Research*, 30 (2002), 42-46.
- KAPLOWITZ, M. D., HADLOCK, T. D. & LEVINE, R. 2004. A comparison of web and mail survey response rates. *Public Opinion Quarterly*, 68 (1), 94-101.
- KAUFMAN, A., YAGEL, R., BAKALASH, R. & SPECTOR, I. 1990. Volume Visualization in Cell Biology. Proceedings of the 1st IEEE conference of Visualization'90 1990, San Francisco, CA, USA.
- KAUFMAN, A. E. & SOBIERAJSKI, L. M. 1995. Continuum Volume Display. In: GALLAGHER, R. (ed.) *Computer Science*.

- KIESLER, S. & SPROULL, L. S. 1986. Response effects in the electronic survey. *Public Opinion Quarterly*, 50, 402-413.
- KINGA, W. R. & JR., P. V. M. 2008. Motivating knowledge sharing through a knowledge management system. *The International Journal of Management Science*, 36 (2008), 131-146.
- KITTUR, A., CHI, E., PENDLETON, B. A., SUH, B. & MYTKOWICS, T. 2007. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie 25th Annual ACM Conference on Human Factors in Computing Systems (CHI 2007), San Jose, CA.
- KITTUR, A., CHI, E. H. & SUH, B. 2008. Crowdsourcing user studies with Mechanical Turk. Proceedings of The twenty-sixth annual SIGCHI conference on Human factors in computing systems, Florence, Italy.
- KITTUR, A. & KRAUT, R. E. 2008. Harnessing the wisdom of crowds in wikipedia: quality through coordination Proceedings of the 2008 ACM conference on Computer supported cooperative work, San Diego, CA, USA.
- KITTUR, A. & KRAUT, R. E. 2010. Beyond Wikipedia: coordination and conflict in online production groups Proceedings of the 2010 ACM conference on Computer supported cooperative work, Savannah, Georgia, USA.
- KITTUR, A., LEE, B. & KRAUT, R. E. 2009. Coordination in collective intelligence: the role of team structure and task interdependence Proceedings of the 27th international conference on Human factors in computing systems, Boston, MA, USA.
- KITTUR, A., SUH, B., PENDLETON, B. A. & CHI, E. H. 2007. He says, she says: conflict and coordination in Wikipedia. Proceedings of the SIGCHI conference on Human Factors in Computing Systems, San Jose, California, USA.
- KLUVER, R., FOOT, K., JANKOWSKI, N. & SCHNEIDER, S. 2007. The internet and national elections: A comparative study of web campaigning, London and New York: Routledge.
- KNUTSON, B., WESTDORP, A., KAISER, E. & HOMMER, D. 2000. FMRI Visualization of Brain Activity during a Monetary Incentive Delay Task. *NeuroImage*, 12 (2000), 20-27.
- KOCH, R. 1999. *The 80/20 Principle: The secret to achieving more with less*, New York: Crown Business.
- KORFIATIS, N., POULOS, M. & BOKOS, G. 2006. Evaluating authoritative sources using social networks: an insight from Wikipedia. *Online Information Review*, 30 (3), 252-262.
- KOSCHAT, M. A. & SWAYNE, D. F. 1996. Interactive Graphical Methods in the Analysis of Customer Panel Data. *Journal of Business & Economic Statistics*, 14, 113-126.
- KOVACS, D. K., ROBINSON, K. L. & DIXON, J. 1995. Scholarly e-conferences on the academic networks: how library and information science professionals use them. *Journal of the American Society for Information Science and Technology*, 46 (4), 244-253.
- KREMER, J. R., MASTRONARDE, D. N. & MCINTOSH, J. R. 1996. Computer Visualization of Three-Dimensional Image Data Using IMOD. *Journal of Structural Biology*, 116 (1996), 71-76.
- KRIPLEAN, T., BESCHASTNIKH, I. & MCDONALD, D. W. 2008. Articulations of wikiwork: uncovering valued work in wikipedia through barnstars. Proceedings

- of the 2008 ACM conference on Computer supported cooperative work, San Diego, CA, USA
- LAKHANI, K. R. & HIPPEL, E. V. 2003. How open source software works: "free" user-to-user assistance. *Research Policy*, 32 (6), 923-943.
- LAM, S. K. & RIEDL, J. 2011. The past, present and future of Wikipedia. *Computer*, 44 (3), 87-90.
- LAMPE, C. & RESNICK, P. 2004. Slash(dot) and burn: distributed moderation in a large online conversation space Proceedings of the SIGCHI conference on Human factors in computing systems, Vienna, Austria.
- LAMPE, C. A. C., ELLISON, N. & STEINFELD, C. 2007. A familiar face(book): profile elements as signals in an online social network Proceedings of the SIGCHI conference on Human factors in computing systems, San Jose, California, USA.
- LAROSE, D. T. 2004. *Discovering Knowledge in Data: An Introduction to Data Mining*: Wiley-Interscience.
- LAW, J. 2009. Seeing Like a Survey. *Cultural Sociology*, 3 (2), 239-259.
- LAW, J., RUPPERT, E. & SAVAGE, M. 2011. *The Double Social Life of Methods*. CRESC Working Paper Series.
- LEEuw, E. D., HOX, J. & SNIJKERS, G. 1995. The Effect of Computer-Assisted Interviewing on Data Quality. *Journal of the Market Research Society*, 37 (4), 325-344.
- LEUF, B. & CUNNINGHAM, W. 2001. *The Wiki Way: Quick Collaboration on the Web*, Boston; London: Addison Wesley.
- LEWIS, K., KAUFMAN, J., GONZALEZ, M., WIMMER, A. & CHRISTAKIS, N. 2008. Tastes, Ties and Time: A New Social Network Dataset Using Facebook.com. *Social Networks*, 30 (2008), 330-342.
- LIH, A. 2004. Wikipedia as Participatory Journalism. Proceedings of 5th International Symposium on Online Journalism, University of Texas at Austin, USA.
- LIH, A. 2009. P.I. Podcast: Interview with Andrew Lih [Online][Accessed 2011/7/8] <http://www.principledinnovation.com/blog/2009/04/21/pi-podcast-interview-with-andrew-lih/>
- LIH, A. 2009. *The Wikipedia Revolution: How a Bunch of Nobodies Created the World's Great Encyclopedia*, New York: Hyperion.
- LIPSCH, M. 2009. National culture and the presence of experts on the online encyclopaedia Wikipedia. Master Unpublished MasterThesis In: Information management.
- LIU, H. 2008. Social Network Profiles as Taster Performances. *Journal of Computer-Mediated Communication*, 13 (1), 252-275.
- LOMAS, N. 2009. Exclusive: Jimmy Wales on what's next for Wikipedia. 2009/11/05. Available: <http://www.silicon.com/technology/networks/2009/11/05/exclusive-jimmy-wales-on-whats-next-for-wikipedia-39626372/> [Accessed 11, July, 2010].
- LOUBSER, M. & BESTEN, M. L. D. 2008. Wikipedia Admins and Templates: the Organizational Capabilities of a Peer Production Effort. Available: <http://ssrn.com/abstract=1116171>.
- MAGNUS, P. D. 2006. Epistemology and the Wikipedia the North American Computing and Philosophy Conference, Troy, New York.

- MAGNUS, P. D. 2008. Early Response to False Claims in Wikipedia. *First Monday* [Online], 13. Available: <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/2115/2027>.
- MAGNUS, P. D. 2009. On Trusting WIKIPEDIA. *Episteme*, 6 (1), 74-77.
- MAJCHRZAK, A., WAGNER, C. & YATES, D. 2006. Corporate wiki users: results of a survey. Proceedings of the 2006 international symposium on Wikis, Odense, Denmark.
- MANFREDA, K. L. & BATAGEJI, Z. 2002. Design of web survey questionnaires: Three basic experiments. *Journal of Computer Mediated Communication*, 7 (3), 149-152.
- MARCUS, B. 2003. Attitudes Towards Personnel Selection Methods: A Partial Replication and Extension in a German Sample. *Applied Psychology*, 52, 515-532.
- MARKS, P. 2007. Interview: Knowledge to the people. *New Scientist*, 44-45 <http://www.newscientist.com/article/mg19325896.300-interview-knowledge-to-the-people.html>.
- MCCORMICK, B. H., DEFANTI, T. A. & BROWN, M. D. 1987. Scientific and Engineering Research Opportunities. *Computer Graphics*, 21 (6), 15-22.
- MCCORMICK, B. H., DEFANTI, T. A. & BROWN, M. D. 1987. Visualization in Scientific Computing: A Synopsis. *IEEE Computer Graphics and Applications*, 7 (7), 61-70.
- MCGUINNESS, D. L., ZENG, H., SILVA, P. P. D., DING, L., NARAYANAN, D. & BHAOWAL, M. 2006. Investigations into Trust for Collaborative Information Repositories: A Wikipedia Case Study WWW200, May 22-26 Edinburgh, UK.
- MCKNIGHT, D. H. & CHERVANY, N. L. 2001. What Trust Means in E-Commerce Customer Relationships: An Interdisciplinary Conceptual Typology. *International Journal of Electronic Commerce* 6(2), 35-59.
- MCLEAN, R., RICHARDS, B. H. & WARDMAN, J. I. 2007. The effect of Web 2.0 on the future of medical practice and education: Darwinian evolution or folksonomic revolution? *The Medical Journal of Australia*, 187 (3), 174-177.
- MELLOR, J., YANAI, I., CLODFELTER, K., MINTSERIS, J. & DELISI, C. 2002. Predictome: A database of putative functional links between proteins. *Nucleic Acids Research*, 30 (2002), 306-309.
- MERING, V. C., HUYNEN, M., JAEGGI, D., SCHMIDT, S., BORK, P. & SNEL, B. 2003. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research*, 31 (2003), 258-261.
- MICHAELIDOU, N. & DIBB, S. 2006. Using email questionnaires for research: Good practice in tackling non-response. *Journal of Targeting, Measurement and Analysis for Marketing*, 14 (4), 289-296.
- MICHALAK, E. & SZABO, A. 1998. Guidelines for Internet Research: An Update. *European Psychologist*, 3 (1), 70-75.
- MIHALCEA, R. 2007. Using Wikipedia for Automatic Word Sense Disambiguation. Proceedings of the North American Chapter of the Association for Computational Linguistics Rochester, New York, USA.
- MILLER, J. P. 1994. Should you get wired? . *Library Journal*, 119 (2), 47-49.

- MILNE, D., MEDELYAN, O. & WITTEN, I. H. 2006. Mining Domain-Specific Thesauri from Wikipedia: A case study Web Intelligence, 2006. IEEE/WIC/ACM International Conference, 18-22 Dec. 2006 Hong Kong.
- MILNE, D. & WITTEN, I. H. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. Proceedings of AAAI 2008 Conference, Chicago, Illinois, USA.
- MOCKUS, A., R.T.FIELDING & J.D.HERBSLEB 2002. Two case studies of Open Source Software Development. ACM Transaction on software engineering and methodology, 11 (3), 309-346.
- MORELAND, R. L. & LEVINE, J. M. 1984. Role Transitions in Small Groups. 181-195. In: ALLEN, V. & VLIERT, E. V. D. (eds.) Role Transitions: Explorations and explanations. New York: Plenum.
- MORELAND, R. L. & LEVINE, J. M. 2001. Socialization in Organizations and Work Groups. 69-112. In: TURNER, M. E. (ed.) Groups at work: Theory and Research. Routledge.
- MUCHNIK, L., ITZHACK, R., SOLOMON, S. & LOUZOUN, Y. 2007. Self-emergence of knowledge trees: Extraction of the Wikipedia hierarchies. Physical Review, E (76), 112-124.
- MUELLER, K., WELSH, T., ZHU, W., MEADE, J. & VOLKOW, N. 2001. BrainMiner: A Visualization Tool for ROI-Based Discovery of Functional Relationships in the Human Brain. Proceedings of Medical Imaging 2001: Physiology and Function from Multidimensional Images, San Diego, CA, USA 481- 485.
- MULBRANDON, C. 2004. Visualizing Economics: Designing a Persuasive Argument. Master of Design in Interaction Design. In: The School of Design.
- N/A. 2001. How to be an editor or peer reviewer for Nupedia [Online]. Web archive. Available: <http://web.archive.org/web/20010410035607/www.nupedia.com/steering.shtml> [Accessed May, 27 2011].
- N/A. 2001. W3C Semantic Web FAQ. Available: <http://www.w3.org/2001/sw/SW-FAQ#isthisresearch>.
- N/A. 2007. Wikipedia cracks two-millionth mark. Sydney Morning Herald.
- N/A. 2010. Wikipedia goes down. telegraph, 24, March, 2010.
- NESS, P. H. V. & KASI, S. V. 2001. Religion and Cognitive Dysfunction in an Elderly Cohort. Journal of Gerontology, 58 (1), 21-29.
- NEUSTAEDTER, C., BRUSH, A., SMITH, M. & FISHER, D. 2005. The Social Network and Relationship Finder: Social Sorting for Email Triage. Conference on E-mail and Anti-Spam.
- NEWBY, G. B. 1993. Virtual reality. Annual Review of Information Science and Technology, 28, 187-229.
- NEWMAN, M. 2005. Power laws, Pareto distributions and Zipf's law. Contemporary Physics, 46 (5), 323-351.
- NEWMAN, M. E. J., BARABASI, A.-L. & WATTS, D. J. 2006. The Structure and Dynamics of Networks: Princeton University Press.
- NOCERA, J. L. A. 2002. Ethnography and Hermeneutics in Cybercultural Research Accessing IRC Virtual Communities. Journal of Computer-Mediated Communication, 7 (2), 0-3.

- NOVECK, B. S. 2007. Wikipedia and the Future of Legal Education. *Journal of Legal Education*, 57 (3), 3-9.
- O'CONNOR, H. & MADGE, C. 2001. Cyber-Mothers: Online Synchronous Interviewing Using Conferencing Software. *Sociological Research Online*, 5 (4).
- OGAN, C. 1993. Listserv communication during the gulf war: What kind of medium is the electronic bulletin board? *Journal of Broadcasting & Electronic Media*, 37 (2), 177-196.
- OH, W. & JEON, S. 2004. Membership Dynamics and Network Stability in the Open-Source Community: The Ising Perspective. *Proceedings of 25th International Conference on Information Systems*, Washington DC, USA.
- OH, W. & JEON, S. 2007. Membership Herding and Network Stability in the Open Source Community: The Ising Perspective. *Management Science*, 53 (7), 1086-1101.
- OKAWA, S. & HONDA, S. 2004. Noise Reduction from MEG data. *Proceedings of SIGE 2004 Annual Conference*. 1431-1435.
- OKOLIA, C. & OHB, W. 2007. Investigating recognition-based performance in an open content community: A social capital perspective *Information & Management*, 44 (3), 240-252.
- OLLEROS, F. X. 2008. Learning to Trust the Crowd: Some Lessons from Wikipedia *Proceedings of the 2008 International MCETECH Conference on e-Technologies*, Montreal (Quebec) Canada.
- O'REILLY, T. 2005. What is Web 2.0-- Design patterns and business models for the next generation of software [Online]. Available: <http://oreilly.com/web2/archive/what-is-web-20.html>.
- ORFORD, S., DORLING, D. & HARRIS, R. 1998. Review of Visualization in the Social Sciences: A State of the Art Survey and Report. *School of Geographical Sciences*, University of Bristol.
- ORFORD, S., HARRIS, R. & DORLING, D. 1999. Geography: Information Visualization in the Social Sciences. *Social Science Computer Review*, 17 (3), 289-304.
- ORLOWSKI, A. 2005. Wikipedia science 31% more cronky than Britannica's. *Music and Media* [Online]. Available: http://www.theregister.co.uk/2005/12/16/wikipedia_britannica_science_comparison/ [Accessed 2005-12-16].
- ORTEGA, F. & BARAHONA, J. M. G. 2007. Quantitative analysis of the wikipedia community of users. *Proceedings of the 2007 international symposium on Wikis*, Montreal, Quebec, Canada. 75-86.
- ORTEGA, F., GONZALEZ-BARAHONA, J. M. & ROBLES, G. 2008. On the Inequality of Contributions to Wikipedia. *Proceedings of HICSS '08 Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences IEEE Computer Society Washington, DC, USA*. 304-304.
- ORTEGA, J. E. F. 2009. Wikipedia: A quantitative analysis. PhD Unpublished Doctoral Dissertation. In: *Ingeniero de Telecomunicación*. Available: <http://libresoft.es/Members/jfelipe/thesis-wkp-quantanalysis>.
- OUTHWAITE, W. & TURNER, S. P. 2007. *The Sage handbook of social science methodology*, London: Los Angeles, SAGE.
- PANCIERA, K., HALFAKER, A. & TERVEEN, L. 2009. Wikipedians are born, not made: a study of power editors on Wikipedia *Proceedings of the ACM 2009 international conference on Supporting group work*, Sanibel Island, Florida, USA.

- PARK, H. W. & THELWALL, M. 2003. Hyperlink Analyses of the World Wide Web: A Review. *Journal of Computer-Mediated Communication*, 8 (4).
- PARKER, L. 1992. Collecting data the e-mail way. *Training and Development*, July, 52-54.
- PATWARDHAN, S., BANERJEE, S. & PEDERSEN, T. 2005. SenseRelate: TargetWord-A generalized framework for word sense disambiguation. *Proceedings of AAAI-05*.
- PAUL, I. 2010. The Facebook Data Torrent Debacle: Q&A. *PCWorld* July 10th 2011, http://www.pcworld.com/article/202167/the_facebook_data_torrent_debacle_qanda.html.
- PELEG, S., KEREN, D. & SCHWEITZER, L. 1987. Improving Image Resolution Using Subpixel Motion. *Pattern Recognition Letters*, 5 (1987), 223-226.
- PEPKE, E. 1995. Animation and the Examination of Behavior Over Time. In: GALLAGHER, R. (ed.) *Computer Visualization*. CRC Press.
- PFEIL, U., ZAPHIRIS, P. & ANG, C. S. 2006. Cultural Differences in Collaborative Authoring of Wikipedia. *Journal of Computer-Mediated Communication*, 12 (1).
- PILLAY, S. 2009. Why Does Visualization Not Work For You?
- PISKORSKI, M. J. & GORBATAI, A. 2010. Testing Coleman's Social Norm Enforcement Mechanism: Evidence from Wikipedia. HBS Working Paper Number: 11-055. Harvard Business School. Available: <http://www.hbs.edu/research/pdf/11-055.pdf>.
- PITKOW, J. & RECKER, M. M. 1995. Using the Web as a survey tool: Results from the second WWW user survey. *Journal of Computer Networks and ISDN Systems*, 27 (6), 809-822.
- PONZETTO, S. P. & STRUBE, M. 2007. Deriving a Large Scale Taxonomy from Wikipedia Association for the Advancement of Artificial Intelligence 07, Vancouver, British Columbia, Canada.
- PRAHALAD, C. K. & RAMASWAMY, V. 2010. *The Future of Competition: Co-Creating Unique Value with Customers*: Harvard Business School Press.
- PRICE, E. 2010. 100M Facebook Profiles Now Available for Download. *PCWorld*, July 10th 2010, http://www.pcworld.com/article/202126/100m_facebook_profiles_now_available_for_download.html.
- PRIEDHORSKY, R., CHEN, J., LAM, S. K., PANCIERA, K., TERVEEN, L. & RIEDL, J. 2007. Creating, destroying, and restoring value in wikipedia *Proceedings of the 2007 international ACM conference on Supporting group work*, Sanibel Island, Florida, USA.
- PRUSINKIEWICZ, P. 1993. Modeling and Visualization of Biological Structures. *Proceedings of Graphics Interface '93*, Toronto Ontario.
- REAGLE, J. M. 2007. Do as I do: authorial leadership in wikipedia *Proceedings of the 2007 international symposium on Wikis*, Montreal, Quebec, Canada.
- REAGLE, J. M. 2010. *Good Faith Collaboration: The Culture of Wikipedia*, Cambridge; Mass: The MIT Press.
- REIPS, U.-D. 2000. The Web Experiment Method: Advantages, Disadvantages and Solutions. 89-118. In: BIRNBAUM, M. H. (ed.) *Psychological Experiments on the Internet*. San Diego, CA: Academic Press.

- RETTIE, R. 2001. An exploration of flow during Internet use. *Internet Research*, 11 (2), 103-113.
- RHEINGOLD, H. 2000. *The virtual community: Homesteading on the electronic frontier*, Cambridge; Mass; London: MIT Press.
- RIEHLE, D. 2006. How and why Wikipedia works: an interview with Angela Beesley, Elisabeth Bauer, and Kizu Naoko Proceedings of the 2006 international symposium on Wikis, Odense, Denmark.
- RITZER, G. & JURGENSON, N. 2010. Production, Consumption, Prosumption: The nature of capitalism in the age of the digital 'prosumer'. *Journal of Consumer Culture*, 10 (1), 13-36.
- ROBB, R. A. 2000. Virtual endoscopy: development and evaluation using the Visible Human Datasets. *Computerized Medical Imaging and Graphics*, 24 (2000), 133-151.
- ROBLER, F., TEJADA, E., FANGMEIER, T., ERTL, T. & KNAUFF, M. 2006. GPU-based Multi-Volume Rendering for the Visualization of Functional Brain Images. Proceedings of SimVis 2006, Magdeburg, Germany. Publishing House, 305-318.
- ROYAL, C. & KAPILA, D. 2009. What's on Wikipedia, and What's Not . . . ? *Soc. Sci. Comput. Rev.*, 27 (1), 138-148.
- RUEDEN, C., ELICEIRI, K. W. & WHITE, J. G. 2004. VisBio: A Computational Tool for Visualization of Multidimensional Biological Image Data. *Traffic* 2004 (5), 411-417.
- RUPPERT, E. & SAVAGE, M. 2009. *New Populations: Scoping Paper on Digital Transactional Data*. CRESC Working Paper Series. Centre for Research on Socio-Cultural Change.
- RUPPERT, E. & SAVAGE, M. Forthcoming. *Transactional Politics*. Sociological Review Monograph Series.
- SACK, W. 2000. Discourse diagrams: Interface design for very large-scale conversations. Proceedings of 34th Hawaiian International Conference on System Sciences, Los Alamitos. IEEE Computer Society Press.
- SALADI, S., PINNAMANENI, P. & MEYER, J. 2001. Texture-based 3D Brain Imaging. Proceedings of 2nd IEEE International Symposium on Bioinformatics & Bioengineering, Bethesda, Maryland.
- SANDERSON, P. M. & FISHER, C. 1994. Exploratory Sequential Data Analysis: Foundations. *Human-Computer Interaction*, 9 (4), 251-317.
- SANGER, L. 2006. *Toward a new compendium of knowledge (longer version)*. Available: <http://www.citizendium.org/essay.html>.
- SANGER, L. M. 2005. *The Early History of Nupedia and Wikipedia: A Memoir*. In: DIBONA, C., COOPER, D. & STONE, M. (eds.) *Open sources 2.0: the continuing evolution*. Sebastopol, CA: O'Reilly Media.
- SANGER, L. M. 2009. *The Fate of Expertise after WIKIPEDIA*. *Episteme*, 6 (1), 52-73.
- SAVAGE, M. 2009. *Contemporary Sociology and the Challenge of Descriptive Assemblage*. *European Journal of Social Theory*, 12 (1), 155-174.
- SAVAGE, M. & BURROWS, R. 2007. *The Coming Crisis of Empirical Sociology*. *Sociology*, 41 (5), 885-899.
- SAVAGE, M. & BURROWS, R. 2009. *Some Further Reflections on the Coming Crisis of Empirical Sociology*. *Sociology*, 43 (4), 762-772.

- SAVAGE, M., RUPPERT, E. & LAW, J. 2010. Digital Devices: nine theses. CRESC Working Paper Series. Centre for Research on Socio-Cultural Change.
- SAX, L. J., GILMARTIN, S. K. & BRYANT, A. N. 2003. Assessing response rates and nonresponse bias in web and paper surveys. *Research in Higher Education*, 44 (4), 409-432.
- SAYER, A. 1992. *Method in Social Science*, London: Routledge.
- SCHILLER, N. 1994. Internet training and support, Academic libraries and computer centers: Who's doing what? *Internet Research*, 4 (2), 35-47.
- SCHMIDT, A. & BRAUN, S. 2007. People Tagging & Ontology Maturing: Towards Collaborative Competence Management 8th International Conference on the Design of Cooperative Systems (COOP '08), Carry-le-Rouet
- SCHMIDT, W. C. 1997. World-Wide Web survey research: benefits, potential problems and solutions. *Behavior Research Methods, Instruments & Computers*, 29 (2), 274-279.
- SCHONLAU, M., FRICKER, R. & ELLIOTT, M. N. 2002. *Conducting Research Surveys via E-mail and the Web*, Santa Monica, CA: RAND.
- SCHWARZER, R., MUELLER, J. & GREENGLASS, E. 1999. Assessment of Perceived General Self-efficacy on the Internet: Data Collection in Cyberspace. *Anxiety, Stress & Coping*, 12 (2), 145-161.
- SCOTT, J. 1988. Trend Report: Social Network Analysis. *Sociology*, 22 (1), 109-127.
- SELM, M. V. & JANKOWSKI, N. W. 2006. Conducting online surveys. *Quality & Quantity*, 40 (3), 435-456.
- SEPULVEDA, V. 2006. Online Survey Research: Benefits and Limitations. *Journal of Counseling Practice, Archives*, 24-27.
- SHAFFER, H. J., PELLER, A. J., LAPLANTE, D. A., NELSON, S. E. & LABRIE, R. A. 2010. Toward a paradigm shift in Internet gambling research: From opinion and self-report to actual behavior. *Addiction Research & Theory*, 18 (3), 270-283.
- SHANNON, P., MARKIEL, A., OZIER, O., BALIGA, N., WANG, J., RAMAGE, D., AMIN, N., SCHWIKOWSKI, B. & IDEKER, T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 2003 (13), 2498-2504.
- SHEEHAN, K. & HOY, M. 1999. Using E-mail to survey internet users in the United States: Methodology and Assessment. *Journal of Computer-Mediated Communication*, 4 (3).
- SHEEHAN, K. B. 2006. E-mail survey response rates: A review. *Journal of Computer-Mediated Communication*, 6 (2).
- SHEPHARD, M. S. & SCHROEDER, W. J. 1995. Analysis Data for Visualization. In: GALLAGHER, R. (ed.) *Computer Visualization*. CRC Press.
- SHIRKY, C. 2005. Power laws, weblogs and inequality. 372. In: LEBKOWSKY, J. & RATCLIFFE, M. (eds.) *Extrane democracy*. Lulu.com.
- SILVERMAN, D. 2004. *Qualitative Research: Theory, Method and Practice*: Sage Publications Ltd; 2nd edition
- SIMSEK, Z. & VEIGA, J. F. 2001. A Primer on Internet Organizational Surveys. *Organizational Research Methods*, 4 (3), 218-235.

- SMETS, K., GOETHALS, B. & VERDONK, B. 2008. Automatic Vandalism Detection in Wikipedia: Towards a Machine Learning Approach. Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium.
- SMITH, C. 1997. Casting the net: Surveying an Internet Population. *Journal of Computer-Mediated Communication*, 3 (1).
- SMITH, D. 2009. Wikipedia Woes: Pending crisis as editors leave in droves. davessmith_at_the lighter side of eu. thunderbolts.info.
- SMITH, K. 2008. Mind-reading with a brain scan. *Nature News*.
- SMITH, M. A. & KOLLOCK, P. 1999. *Communities in Cyberspace*: Routledge.
- SMUELSSON, M., THERNLUND, G. & RINGSTROM, J. 1996. Using the Five Field Map to Describe the Social Network of Children: A Methodological Study. *International Journal of Behavioral Development*, 19 (2), 327-345.
- SPENCE, R. 2006. *Information Visualization: Design for Interaction*: Prentice Hall.
- SPOERRI, A. 2007. What is Popular on Wikipedia and Why? *First Monday*, 12 (4).
- STANCZAK, G. C. 2007. *Visual research methods: Image, Society, and Representation*, Los Angeles; London: Sage Publications.
- STRUBE, M. & PONZETTO, S. P. 2006. WikiRelate! computing semantic relatedness using wikipedia the 21st national conference on Artificial intelligence, Boston, Massachusetts, U.S.A.
- STVILIA, B., TWIDALE, M. B., GASSER, L. & SMITH, L. C. 2005. Information Quality Discussions in Wikipedia International Conference on Knowledge Management-2005, Charlotte, NC USA.
- STVILIA, B., TWIDALE, M. B., SMITH, L. C. & GASSER, L. 2005. Assessing Information Quality of a Community-based Encyclopedia. *Proceedings of the International Conference on Information Quality MIT, Cambridge, Massachusetts, USA*. 442-454.
- STVILIA, B., TWIDALE, M. B., SMITH, L. C. & GASSER, L. 2008. Information Quality Work Organization in Wikipedia. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, 59 (6), 983-1001.
- SUH, B., CHI, E. H., KITTUR, A. & PENDLETON, B. A. 2008. Lifting the veil: improving accountability and social transparency in Wikipedia with wikidashboard Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, Florence, Italy.
- SUH, B., CONVERTINO, G., CHI, E. H. & PIROLI, P. 2009. The Singularity is not near: Slowing Growth of Wikipedia WikiSym '09, Orlando, Florida, U.S.A.
- SUROWIECKI, J. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*: Little, Brown & Company.
- SWARTZ, A. 2004. Who Writes Wikipedia. Aaron Swartz's Raw Thought [Online]. Available from: <http://www.aaronsw.com/weblog/howwriteswikipedia>.
- TABB, K. 2008. Authority and Authorship in a 21st-Century Encyclopaedia and a 'very mysterious Foundation'. E-Sharp, (12: Technology and Humanity).
- TANCER, B. 2007. Look Who's Using Wikipedia [Online][Accessed Nov. 11, 2007] <http://www.time.com/time/business/article/0,8599,1595184,00.html>

- TAPSCOTT, D. & WILLIAMS, A. D. 2006. *Wikinomics: How Mass Collaboration Changes Everything*, New York: Portfolio.
- THELWALL, M., SUD, P. & VIS, F. 2012. Commenting on YouTube videos: From Guatemalan rock to El Big Bang. *Journal of the American Society for Information Science and Technology*, 63 (3), 616-629.
- THRIFT, N. J. 2005. *Knowing capitalism*, London: SAGE Publications.
- TIMOTHY 2005. The early history of Nupedia and Wikipedia: A memoir. slashdot.com.
- TOLLEFSEN, D. P. 2009. WIKIPEDIA and the Epistemology of Testimony. *Episteme*, 6 (1), 8-24.
- TREDINNICK, L. 2006. Web 2.0 and Business: A pointer to the intranets of the future? *Business Information Review*, 23 (4), 228-234.
- TSE, A., TSE, K. C., YIN, C. H., TING, C. B., YI, K. W., YEE, K. P. & HONG, W. C. 1995. Comparing two methods of sending out questionnaires: E-mail versus mail. *Journal of the Market Research Society*, 4 (37), 441-445.
- UPSON, C., THOMAS FAULHABER, J., KAMINS, D., LAIDLAW, D., SCHLEGEL, D., VROOM, J., GURWITZ, R. & DAM, A. V. 1989. The Application Visualization System: A Computational Environment for Scientific Visualization. *IEEE Computer Graphics and Applications*, 9 (4), 30-42.
- UZZI, B. 1999. Embeddedness in the Making of Financial Capital: How Social Relations and Networks Benefit Firms Seeking Financing. *American Sociological Review*, 64 (4), 481-505.
- VALENTINO, D. J., MAZZIOTTA, J. C. & HUANG, H. K. 1989. Visualization of Human Brain Structure-Function Relationships. *Proceedings of IEEE Engineering in Medicine & Biology Society 11th Annual International Conference*, Seattle, WA, USA. 1737-1738.
- VALENZUELA, S., PARK, N. & KEE, K. F. 2009. Is there social capital in a social network site?: Facebook use and college students' life satisfaction, trust and participation. *Journal of Computer Mediated Communication*, 14 (4), 875-901.
- VIEGAS, F. B., WATTENBERG, M. & DAVE, K. 2004. Studying cooperation and conflict between authors with History flow visualizations *Proceedings of the SIGCHI conference on Human factors in computing systems*, Vienna, Austria.
- VIEGAS, F. B., WATTENBERG, M., KRISS, J. & HAM, F. V. 2007. Talk Before You Type: Coordination in Wikipedia. *Proceedings of the 40th Hawaii International Conference on System Sciences*, Hawaii, USA.
- VIEGAS, F. B., WATTENBERG, M. & MCKEON, M. M. 2007. The Hidden Order of Wikipedia the 2nd international conference on Online communities and social computing, Beijing, China.
- VOLKEL, M., KROTZSCH, M., VRANDECIC, D., HALLER, H. & STUDER, R. 2006. *Semantic Wikipedia WWW2006*, Edinburgh, Scotland.
- VOSS, J. 2005. Measuring Wikipedia *Proceedings International Conference of the International Society for Scientometrics and Informetrics: 10th*, Stockholm, Sweden.
- VOSS, J. 2006. Collaborative thesaurus tagging the Wikipedia way. 1. Available: <http://arxiv.org/ftp/cs/papers/0604/0604036.pdf>.
- WAGNER, C. & MAJCHRZAK, A. 2006. Enabling Customer-Centricity Using Wikis and the Wiki Way. *Journal of Management Information Systems*, 23 (3), 17-43.

- WALES, J. 2005. Wikipedia is an encyclopedia. Available: <http://mail.wikipedia.org/pipermail/wikipedia1/2005-March/038102.html> [Accessed 2011, May, 28].
- WALES, J. 2005. Wikipedia, Emergence, and The Wisdom of Crowds [Online]. Available: <http://mail.wikipedia.org/pipermail/wikipedia-1/2005-May/039397.html>.
- WALT, N., ATWOOD, K. & MANN, A. 2008. Does survey medium affect responses. *The Journal of Technology, Learning and Assessment*, 6 (7).
- WALTHER, J. B., ANDERSON, J. F. & PARK, D. W. 1994. Interpersonal Effects in Computer-Mediated Interaction A Meta-Analysis of Social and Antisocial Communication. *Communication Research*, 21 (4), 460-487.
- WASSERMAN, S. & FAUST, K. 1994. *Social Network Analysis: Methods and Applications Structural Analysis in the Social Sciences*: Cambridge University Press.
- WATERS, N. L. 2007. Why You Can't Cite Wikipedia in My Class. *Communications of the ACM*, 50 (9), 15-17.
- WATTENBERG, M., VI GAS, F. B. & HOLLENBACH, K. 2007. Visualizing Activity on Wikipedia with Chromograms *Human-Computer Interaction 2007*, 4663, 272-287.
- WEBBER, R. 2009. Response to 'the coming crisis of empirical sociology': an outline of the research potential of administrative and transactional data. *Sociology*, 43 (1), 169-178.
- WEI, C., MAUST, B., BARRICK, J., CUDDIHY, E. & SPYRIDAKIS, J. H. 2005. Wikis for supporting Distributed Collaborative Writing. *Proceedings of The society for technical communication 52nd annual conference*, 8-11, May, 2005 Seattle, WA.
- WELSER, H. T., GLEAVE, E., FISHER, D. & SMITH, M. 2007. Visualizing the Signatures of Social Roles in Online Discussion Groups. *Journal of Social Structure*, 8 (2).
- WHITNEY, L. 2009. Report: Wikipedia losing volunteers. *CNET News*, November, 23, 2009.
- WHITTAKER, S., TERVEEN, L., HILL, W. & CHERNY, L. 1998. The dynamics of mass interaction *Proceedings of the 1998 ACM conference on Computer supported cooperative work*, Seattle, Washington, United States.
- WILKINSON, D. M. & HUBERMAN, B. A. 2007. Assessing the Value of Cooperation in Wikipedia. *First Monday*, 12 (4).
- WILKINSON, D. M. & HUBERMAN, B. A. 2007. Cooperation and quality in wikipedia. *Proceedings of the 2007 international symposium on Wikis*, Montreal, Quebec, Canada. 157 - 164.
- WINER, D. 2008. What's wrong with Wikipedia. *The Scripting*, 2008-03-20.
- WINER, G. A., COTTRELL, J. E., KAREFILAKI, K. D. & GREGG, V. R. 1996. Images, Words and Questions: Variables that Influence Beliefs about Vision in Children and Adults. *Journal of Experimental Child Psychology*, 63, 499-525.
- WITTEN, I. H. & FRANK, E. 2005. *Data mining : practical machine learning tools and techniques*, San Francisco, Calif. Morgan Kaufmann ; Oxford: Elsevier Science.
- WOODSON, A. 2007. Wikipedia remains go-to site for online news. *REUTERS US*, 2007-07-08.
- WOOLGAR, S. 2002. *Virtual society?: technology, cyberbole, reality*, Oxford: Oxford University Press.

- WORKSHOP, E.-A. 1997. Euro-American Workshop on Visualization of Information and Data [Online]. Available: <http://dl.acm.org/citation.cfm?id=271268>.
- WRAY, K. B. 2009. The Epistemic Cultures of Science and WIKIPEDIA: A Comparison. *Episteme*, 6 (1), 38-51.
- WRIGHT, K. B. 2005. Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire Authoring Software Packages, and Web Survey Services. *Journal of Computer-Mediated Communication*, 10 (3), 11.
- YAN, Y., OKAZAKI, N., MATSUO, Y., YANG, Z. & ISHIZUKA, M. 2009. Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web. Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Suntec, Singapore. 1021-1029.
- ZACHTE, E. 2009. New editors are joining English Wikipedia in droves? *Infodisiac*.
- ZAGORCHEV, L., GOSHTASBY, A. & SATTER, M. 2006. Flow visualization for qualitative assessment of brain shift. In: CLEARY, K. R. & GALLOWAY, R. L., eds. *Proceedings of Medical Image 2006: Visualization, Image-Guided Procedures and Display*, San Diego, CA, USA. International Society for Optical Engineering; 1999.
- ZENG, H., ALHOSSAINI, M. A., DING, L., FIKES, R. & MCGUINNESS, D. L. 2006. Computing trust from revision history Proceedings of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services, Markham, Ontario, Canada.
- ZHANG, Y. 1999. Using the Internet for Survey Research: A case study. *Journal of the American Society for Information Science*, 51 (1), 57-68.
- ZINOVYEV, A. 2011. Data Visualization in Political and Social Sciences. In: BADIE, B., BERG-SCHLOSSER, D. & MORLINO, L. A. (eds.) *International Encyclopedia of Political Science*. SAGE.
- ZOOK, M. & GRAHAM, M. 2007. The Creative Reconstruction of the Internet: Google and the Privatization of Cyberspace and Digiplace. *Geoforum*, 38, 1322-1343.
- ZOOK, M., GRAHAM, M. & SHELTON, T. 2011. Analyzing Global Cyberscapes: Mapping Geo-coded Internet Information iConference 2011, Seattle, WA, USA.