12-1-2011

Theoretical approach to legitimizing collaboratively constructed knowledge: a content analysis of Wikipedia science articles based on accidental collaboration

James Patrick Hutchinson
University of Northern Colorado

UNIVERSITY OF NORTHERN COLORADO

Greeley, Colorado

The Graduate School

A THEORETICAL APPROACH TO LEGITIMIZING COLLABORATIVELY CONSTRUCTED KNOWLEDGE: A CONTENT ANALYSIS OF WIKIPEDIA SCIENCE ARTICLES BASED ON ACCIDENTAL COLLABORATION

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

James Patrick Hutchinson

College of Education and Behavioral Sciences
Department of Educational Technology

December 2011

ABSTRACT

Hutchinson, James Patrick. *A Theoretical Approach To Legitimizing Collaboratively Constructed Knowledge: A Content Analysis of Wikipedia Science Articles Based on Accidental Collaboration*. Published Doctor of Philosophy dissertation, University of Northern Colorado, 2011.

This study involved an analysis of 147 Wikipedia science articles using content and social network analysis to explore authorial relationships between articles and test a theoretical approach to using accidental collaboration as a tool to legitimize collaboratively constructed knowledge. Contrary to Wikipedia's tagline of "anyone can edit," this study found that articles had a small number of prolific contributors and that these contributors had educational background and edit history suggesting they were knowledgeable about the topics to which they contributed. Results also showed that articles found via accidental collaboration tended to be scientific in nature and often had direct subject matter relationships to their corresponding seed article. Taken together, these results suggest that Wikipedia science articles are at least partially written by knowledgeable individuals. Implications include rethinking how Wikipedia is used by teachers and students; its potential as a tool for developing critical literacy and 21st century skills; and the need for continued research to further explore the issues of legitimacy and reliability of Wikipedia in various subject areas. Due to the limitations of this study, generalizations beyond the science articles studied cannot be made.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the help and assistance of a number of individuals. Most importantly, the support of my entire committee was invaluable and, in particular, my research advisor Dr. James Gall whose guidance, support, insightful comments and editorial feedback was crucial. Dr. Mia Williams' discussions and suggestions regarding critical literacy added an additional demission to the study that helped extend its practical implications.

The data collection and processing that was necessary for this study would not have been possible without the help of a few key individuals who volunteered their time to lend the technical expertise that I lacked. I appreciate the help of the toolserver.org user *Betacommand* who ran the SQL query which provided the raw Wikipedia data needed. Also, my former student and technological guru, Mr. Caleb Jares, facilitated the writing of the query request and wrote the C# program script that performed the key function of identifying accidental collaborators.

Wikipedia itself would not exist if not for the efforts of millions of volunteer contributors around the world and the creative vision of Jimmy Wales. Their efforts make possible research such as this which will hopefully provide feedback that can help to further improve Wikipedia. We can only hope that it will continue to grow and improve and one day become a library of everything containing the sum of human knowledge.

Finally, and perhaps most importantly, I would like to express my thanks and appreciation to my family for their continued support and patience over the years as I pursued my own educational dream. May they find the encouragement to pursue their own dreams and the determination to realize them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Humans seem to have an innate desire and need to transmit knowledge to future generations. Such knowledge transmission clearly had evolutionary benefits as well. Types and forms of stone tools, for example, demonstrate the impact of culture and shared knowledge. Although the emergence of language cannot be exactly determined, it clearly coincided with a long period of technological and cognitive development of early man (Renfrew, Frith & Malafouris, 2008). Spoken and written communication further facilitated our technological advancement.

Starting with the early Greeks, the encyclopedia emerged as a modern form of culture and knowledge transmission and has served as an important tool for the collecting and archiving of knowledge allowing future generations the opportunity to build on prior developments rather than continual rediscovery. The digital age has rapidly accelerated our ability to create, record and share knowledge as well as offer new opportunities for collaboratively constructing knowledge. Wikipedia is a unique approach that relies on crowd-sourcing knowledge, but while hugely popular it remains to be seen if this approach can result in a legitimate source of authoritative knowledge or will degenerate into a form of cultural tribalism (Arazy, Nov, Patterson, & Yeo, 2011) over who owns the truth.

The encyclopedia has largely been taken for granted and not greatly studied (Kafker, 1981). Nevertheless, the encyclopedia has come to represent the pinnacle of general knowledge transmission and it has become common for school age children and adults to pick up a volume when looking for information on a topic. Dating back to at least the ancient Greeks, the encyclopedia has gone through a number of changes culminating in the modern, multi-volume, alphabetically organized sets we see today such as the English language *Encyclopaedia Britannica* or *The World Book Encyclopedia*. Venerable print encyclopedias such as these are now being challenged by digital encyclopedias that rely on the efforts of unnamed volunteers to add, edit and update content. Currently, the most well-know example is Wikipedia which, since its initial release in 2001, has grown to over 3.7 million articles in English and over 20 million articles in over 280 languages.

The popularity of Wikipedia has also grown and currently (as of October, 2011) ranks fifth in overall global web traffic ("Alexa Top 500 Global Sites," n.d.). Web users looking for information on any topic will likely come across a Wikipedia article fairly quickly. However, the open approach to editing content and even creating new articles, a process in which anyone can edit nearly any page (some pages are locked at times for various reasons), has resulted in a steady stream of criticism regarding quality, accuracy, authority of its authors, susceptibility to vandalism, and overall legitimacy as a reliable reference tool.

Despite a growing body of research suggesting that Wikipedia content is generally credible (Chesney, 2006) and not significantly more error-prone than print encyclopedias (Arazy et al., 2011; Chesney, 2006; Giles, 2005; Magnus, 2006;

Rajagopalan et al., 2010; Rector, 2008; Rosenzweig, 2006), no encyclopedia is ever going to be completely free of errors, but digital encyclopedias have the potential to respond much more quickly when mistakes are found. Shortly after publication of the *Nature* study (Giles, 2005) it was reported that all the identified errors were fixed (Snow, 2006). Conversely, an interesting example of the persistence of outright false information in a print encyclopedia is the story of the so-called Piltdown Man, or Dawson's Dawn Man, reportedly found by Charles Dawson between 1908 and 1912. Dawson claimed the skull was an example of a heretofore unknown missing link in human evolution that contained a mix of modern human and primate features. The discovery was widely reported at the time and accounts of what was later proven to be a hoax remained in such venerable resources as the *Encyclopaedia Britannica* until as recently as 1949 – or nearly 40 years after the initial report (Collison, 1966; "Glacial Epoch," 1949; "Sources and authorities for English history," 1949). Interestingly, accounts of the hoax are now included in both *Britannica* ("Piltdown man," 2002) and Wikipedia ("Piltdown Man," n.d.). In a somewhat ironic passage referring to the Piltdown Man, the 1922 version of the *Encyclopaedia Britannica* stated,

> A vast amount of writing has accumulated since 1912 with reference to this remarkable skull, but most of this literature is irrelevant and misleading, as the authors have not seen the material about which they write and have no adequate realization of the true state of affairs ("Anthropology," 1922).

As a tertiary source, encyclopedias in general could be called "irrelevant and misleading" but for the fact that their authors are trusted as having seen or studied first hand the material about which they write. In other words, encyclopedias are accepted as legitimate sources of information largely because they have shown themselves to be useful and

accurate over time and have developed a level of trust in their authors, editors, creation and publication. The example of the Piltdown Man, however, should cast some doubt over the tendency toward unfailing belief in the printed word and encyclopedic knowledge in particular. Of course, such extreme examples are rare.

One of the more important differences between traditional encyclopedias, such as *Britannica*, and a collaborative, digital encyclopedia such as Wikipedia is the issue of authorship. Modern encyclopedias exercise great control over the editorial process and use highly qualified and vetted authors that results in generally accurate and authoritative information and is largely the reason they have become well accepted and trusted sources, but this process also ensures a fairly slow development of content (Cross, 2006). Following this tradition, Wikipedia also began using only expert authors. Originally called *Nupedia*, its articles were to be written by qualified and vetted authors and subjected to a high level of oversight. This ultimately proved to be a failure and Wikipedia, as it came to be called, achieved very rapid evolution and expansion by allowing anyone to generate and edit articles – a change that opened the door to criticism over the lack of authority and quality control and contributed to the departure of co-founder Larry Sanger (Sanger, 2004) and his later develop of *Citizendium*, a wiki-based encyclopedia that requires contributors to use their real name and employs a high degree of oversight similar to *Nupedia*'s original intent (Rosenzweig, 2006). A few highly publicized incidents such as the claim that former *USA Today* Editor John Seigenthaler Sr. was connected with the assassinations of President John F. Kennedy and Senator Robert F. Kennedy (Helm, 2005; Seigenthaler, 2005; Survey, 2006) helped fuel criticism and increase awareness of the issue among the larger public. Despite these concerns,

anecdotal evidence suggests modern users of Wikipedia generally find the content to be accurate, in-depth and usable, suggesting the model of self-governance and collaboratively constructed information is, to some extent, effective. Nevertheless, the question of authorship and article quality or overall legitimacy will undoubtedly remain as long as Wikipedia continues to operate as an open platform.

These issues, coupled with Wikipedia's ease of access and frequent use by students, which could also apply to web content in general, has caused some concern among educators who feel it is not an appropriate educational tool – particularly for students who may lack sufficient background knowledge and sophistication to discern between accurate and inaccurate information. According to the American Library Association (1989), the emergence of the information age has created new challenges for educators and society as a whole. Prior to the Internet, there was less need to teach students how to determine if information was legitimate. Printed materials, which are subjected to an editorial process and peer review, were generally considered reliable sources of information. The rapid growth of the Internet, however, has created new issues. Web content does not go through the editorial process to which books, magazines and newspapers are subjected, nor is it reviewed and filtered by librarians or teachers before being accessible to students.

In a rather forward-looking move, the American Library Association's Presidential Committee on Information Literacy was commissioned in 1987 with the goal of educating an information literate public. According to the goals of the committee, "to be information literate, a person must be able to recognize when information is needed and have the ability to locate, evaluate, and use effectively the needed information"

(American Library Association, 1989). It was not long before the Internet and the availability of web-based content gave new urgency these words.

The rapid growth of web accessible information and the need to be able to efficiently find it gave rise to companies such as *Google* and their "mission to organize a seemingly infinite amount of information on the web" (Google, n.d.). Pringle (2009) noted "the Net is an astonishing boon to humanity, gathering up and concentrating information and ideas that were once scattered so broadly around the world that hardly anyone could profit from them." However, the process of gathering up, concentrating and organizing content simply assists in location and tells one nothing about whether or not such content is legitimate or accurate. Jimmy Wales, founder of Wikipedia, had a different goal – "a world in which every single person is given free access to the sum of all human knowledge" (as quoted in Lih, 2009). Although not specifically addressed in Wales' comment, the "sum of human knowledge" would necessarily, one would assume, need to be legitimate and reliable information. Early efforts to use qualified and vetted authors were unsuccessful (Lih, 2009; Rosenzweig, 2006) and in order to accomplish their goal, Wikipedia adopted an open editing process that allowed anyone to participate. While this decision proved to be highly successful with Wikipedia growing from just a few hundred articles in 2001 to over 3.7 million by 2011 and 50 times the size of the next largest English language encyclopedia ("Wikipedia: Size comparisons," n.d.) it also gave rise to concerns over the accuracy, authority, and overall legitimacy of the content.

As an educator and library media specialist, I initially had my own concerns over student use of Wikipedia, but, as I watched it grow and found myself using it more and more, I realized that students needed to learn to determine the legitimacy of Wikipedia

content, and web content in general, for themselves – particularly because it was clear they were using it more and more as well. For years, I have observed that students' approach to web-based content, including Wikipedia, often paralleled Freire's (2000) *oppressed* in that they saw information as external and disconnected from themselves, the words of apparent experts that could not, should not, be questioned. This is undoubtedly due, in part, to the banking model (Fieire, 2000) of education that has as its focus the filling of students' heads with facts of the world for later withdrawal – often in the form of a test of their memory and retrieval skills. The analytical and critical approach to learning has often been overlooked. However, as Temple (2005) points out, "only those whose critical faculties have been nurtured, through dialogue about the issues that matter in their lives, develop critical consciousness" (p. 16).

Wikipedia actually offers a unique opportunity to teach students to doubt, question, analyze and explore the legitimacy of apparent factual claims and encourage their development of critical literacy and critical consciousness. Drawing on an idea presented by Harouni (2009), I asked students to select an article in Wikipedia about which they felt they already knew something or considered themselves an expert and then read the article taking note of anything they found that they did not agree with or trust. They then had to verify whether or not this suspect information in Wikipedia was correct. I recall one student who was reading an article on the Denver Broncos football team and felt the information regarding the Broncos only having two NFL Hall of Fame members was surely wrong. In order to verify his suspicion he went to the source of the information – the National Football Hall of Fame. He discovered, much to his dismay, that at that time (early 2011) the Denver Broncos did in fact have only two Hall of Fame

members. Others discovered that the origin of the Australian Shepherd is convoluted and may have little to do with Australia or that the manner of Hitler's death is in dispute and relies somewhat on whose testimony you chose to believe. This type of research was played out over and over as students identified suspicious information, at least to them, and then went through the process of verifying it. The results were illuminating. Students who had generally taken information, web-based or not, at face value were developing skepticism and becoming more analytical. During our debriefings, I asked students what they discovered and most students commented that they were surprised to find that "Wikipedia is usually right" and wondered why they had been repeatedly told by teachers that it was not reliable. Others noted that while the information was not wrong, it was often incomplete or had simplified a more complex issue, such as the origin of the Australian Shepherd, into a sentence or two that obscured a deeper issue. Perhaps due to years of indoctrination by former teachers on the evils of Wikipedia, a few students continued to maintain that Wikipedia was often wrong and full of errors. Further questioning, however, showed that these students tended to hold on to misconceptions or were unsuccessful in finding alternative sources of information and chose to simply believe themselves correct – a common trait among middle school students. While such vignettes are interesting, they do not provide teachers and students the assurances they need regarding the overall legitimacy of Wikipedia and other web based content nor do they fully develop the skills necessary for an information literate populace.

It is also important to remember that Wikipedia is only one example, albeit large and popular, of collaboratively constructed knowledge. Wikis exist all over the web for a variety of purposes and teachers are finding the collaborative nature of the wiki a

powerful educational tool that supports the development of 21st Century Skills including communication, collaboration, problem solving, critical thinking, knowledge construction, and participation in a global community (International Society for Technology in Education, 2007; American Association of School Librarians, 2009). In my own experience working with teachers, wikis have proven to be a unique educational tool. In one instance, students in a geographical information systems class partnered with staff at a nearby state park to help eradicate noxious weeds. The students used hand-held GPS units to mark the coordinates of the weeds around the park. These data points were then shared with park staff using a wiki. The collaborative nature of the wiki allowed all students to contribute to a single shared database that could be accessed by park staff in order to plan and carry out weed control measures. Furthermore, the wiki is not a static, single use product but a living document that can be added to each year while preserving data from prior years. As this collection of data grows, both students and park staff can perform different types of analyses depending on their information needs. For example, students and park staff can use the data to track patterns of weed populations over time to discern if there are migration patterns or if control measures have been ineffective with weeds returning each year to the same areas. This information can be used to inform decisions about future control measures or assist in tracking down the source of a problem.

Wikis are also used to share information on any number of individual topics or projects. Software projects often offer some sort of online documentation for users and the wiki is a perfect tool for both developing the documentation and providing access to

the content. *Open Office*,[1] for example, is a popular, community supported and free software project offering users a tool for creating documents, spreadsheets and presentations. As a free product that is only available by downloading from the project home page, users do not receive any printed documentation. As an alternative, the *Open Office* developers have provided a wiki with extensive information on installing and using the software.[2]

As online, collaboratively developed and shared knowledge becomes more common, it is in our best interest to understand how users interact and collaborate in an open format and how consumers of that information can make decisions about the legitimacy of the content. The purpose of this dissertation is to examine a set of science articles in Wikipedia in order to explore patterns of authorship, and, given the collaborative nature of Wikipedia, co-authorship in particular, in article construction and to determine to what extent, if any, these patterns or profiles can be used to offer some assurance of legitimacy to users of Wikipedia. This dissertation seeks to answer the following research questions:

Q1 What is the profile of contributions to select science articles in Wikipedia?

Q2 What is the profile of a prolific contributor to select science articles in Wikipedia?

Q3 Do prolific contributors to select science articles in Wikipedia contribute to multiple articles?

Q4 What types of articles cluster around select science articles based on accidental collaboration and what conclusions can be drawn?

Q5 What do network maps of article clusters based on accidental collaboration say about the legitimacy of the content?

---

[1] http://www.openoffice.org
[2] http://wiki.services.openoffice.org/wiki/Main_Page

Definitions

Accidental Collaboration. The concept of collaboration generally refers to two or more individuals purposefully working together to a common end. It is, however, possible for two or more people to work together to a common end without consciously intending to do so. In the context of this study, this type of collaboration is considered accidental. Figure 1 shows a graphical representation of how this applies to articles in Wikipedia.

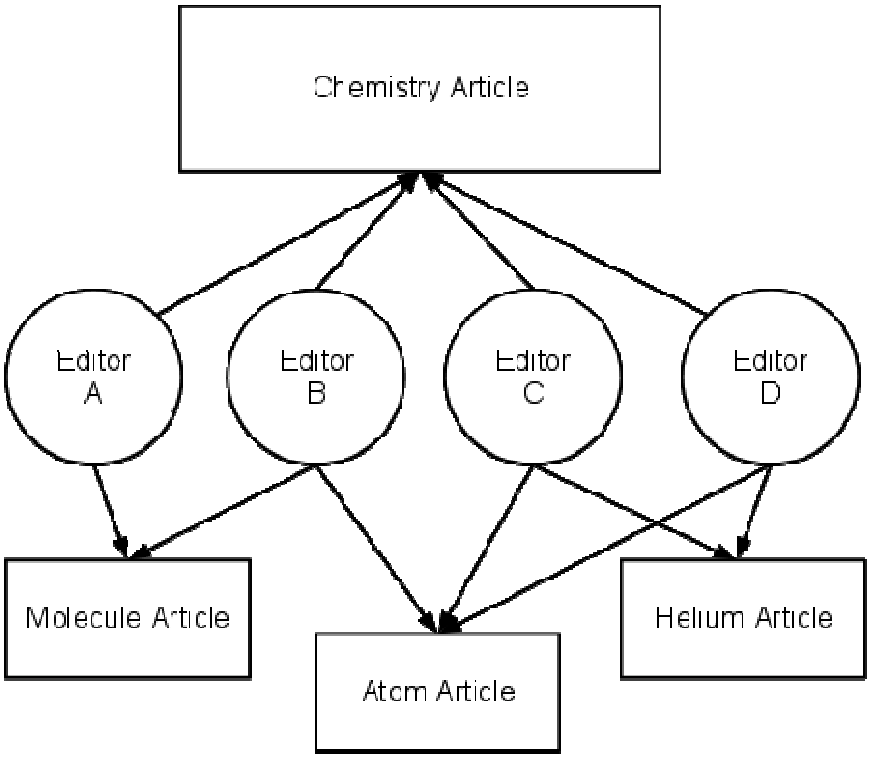


Figure 1. Graphical representation of accidental collaboration showing contributors to the Chemistry article also contributing to other articles.

Co-contributor. When two or more individuals contribute to the same article, they are considered co-contributors (even if they did not work directly together or their contributions occurred at different times).

Contributor. An individual member of the Wikipedia community who contributes content to articles, fixes errors, repairs vandalism or otherwise assists in the maintenance of content. Other researchers and users of Wikipedia have also used the terms author and editor interchangeably when referring to contributors.

Edit. In Wikipedia, a change can range from a single character to paragraphs of text. Regardless of the amount of content added or removed, each time a contributor saves a change or set of changes, this is considered one edit. Edits are tracked in Wikipedia and marked with a time stamp, the name of the user making the edit or IP address for anonymous (i.e. unregistered) users, and a brief description of the nature of the edit. An edit can be either adding, modifying or removing information.

Edit Frequency. Refers to the number of times a contributor to a Wikipedia article or page makes an edit. Each edit adds one to the frequency count regardless of the amount of content added, changed or removed.

Portal. In Wikipedia, "the idea of a portal is to help readers and/or editors navigate their way through Wikipedia topic areas through pages similar to the Main Page. In essence, portals are useful entry-points to Wikipedia content." ("Wikipedia: Portal," n.d.).

Prolific Contributor. For the purpose of this study, any contributor with more than 10 edits to a sampled article was considered a prolific contributor to that article. This was an arbitrary cutoff but due to the presence of hundreds of unique contributors to many of

the articles it was necessary to limit the study to contributors who showed repeated interest in an article.

Seed Article. This study used a selection of 180 science articles. These initial science articles are called seed articles as they constitute a starting point for the analysis of additional articles found using the accidental collaboration process.

Wikipedia. Defines itself as "a free, web-based, collaborative, multilingual encyclopedia project supported by the non-profit WikiMedia Foundation. Its 20 million articles (over 3.78 million in English) have been written collaboratively by volunteers around the world, and almost all of its articles can be edited by anyone with access to the site" ("Wikipedia," n.d.).

Wikipedia Article. Each individual topic of encyclopedic content in Wikipedia is assigned its own page and unique URL and can be considered to be an article in the same sense as each write-up in a print encyclopedia is considered an article on that topic.

CHAPTER II

REVIEW OF LITERATURE

If our understanding of and general knowledge about the world is primarily preserved and passed on via the written and published word, then it is in our best interest to understand how such knowledge is collected, archived, revised and disseminated; how it has been done in the past; and, perhaps most importantly, how will be done now and into the future as print publications slowly give way to electronic forms. The encyclopedia is a well established and respected medium for archiving and sharing general knowledge. Although an understanding of the history and purpose of the encyclopedia is an encyclopedic undertaking itself, a brief history of compiled knowledge is warranted before we can begin to explore the future of knowledge.

History of the Encyclopedia

According to the *Encyclopaedia Britannica* (2002), the term "encyclopedia" comes from the Greek words *enkyklios paideia* meaning well-rounded or general education, or the circle of learning (Kister, 1994; Kogan, 1958), and the modern encyclopedia is a realization of this implied intent (Collison, 1966) – a book or collection of volumes that "contains information on all branches of knowledge" ("Encyclopedia," 2002), or, as Thoreau (1910) put it, "an abstract of human knowledge" (p. 195). In his

Naturalis Historia (79 CE), Pliny the Elder used these words to describe the content of his work as containing the circle of Greek learning (Kogan, 1958; Stockwell, 2000). Stockwell contends that it was not until 1531 when these two words were combined in the term "encyclopedia" by Sir Thomas Elyot in his *Bok of the Governour*, or, according to Kister (1994) in the title of the Latin work *Encyclopaedia: seu, Orbis Disciplinarium, tam Sacrarum quam Prophanum Epistemon* published in 1559 by Paul Scalich. Despite their long history, dating back at least to the fourth century B.C. (see Collison, 1966 for an extensive chronology), and importance, Thorndike suggested they are "the most important monuments of the history of science and civilization" (1924, as cited in Kafker, 1981), the encyclopedia has not been greatly studied (Kafker, 1981).

Nevertheless, the encyclopedia has a rich history dating back to the ancient Greeks. Collison (1966) considered Plato to be the father of the encyclopedia. Although Plato never wrote an encyclopedia himself, he was the founder of the Academy of Athens and was also uncle and mentor to Speusippos who did compile an encyclopedia based on the teaching of Plato to use in his own teaching. One of the earliest known attempts at creating a vast compendium of knowledge is the *Naturalis Historia* of Pliny the Elder (77 C.E.). His thirty-seven books attempted to cover the known natural world and included over 2,500 chapters on topics such as "geography, physiology, zoology, botany, and medicine" (Kister, 1994, p. 5), and, similar to the modern encyclopedia, compiled information from two thousand works and over four hundred authors (Kogan, 1958; Lih, 2009). The Chinese *T'ai P'ing Yu Tan*, published in the tenth century, is generally considered the first modern encyclopedia (Kogan, 1958). The first work to be titled "*Cyclopaedia*" was compiled in 1541 by Ringelberg (Kogan, 1958). The father of the

modern encyclopedia, however, is probably Ephraim Chambers who published the two volume *Cyclopaedia: or, An Universal Dictionary of Arts and Sciences* in London in 1728 which introduced now common elements such as alphabetical arrangement and included a system of cross-references (Kogan, 1958; Lih, 2009). The most comprehensive early encyclopedia was undoubtedly Diderot's much larger, eventually comprising 28 volumes, French *Encyclopédie* published between 1751 and 1772. Originally intended as a translation of Chamber's *Cyclopaedia*, it abandoned the impartial and objective (Kister, 1994) point of view and focus on sharing general knowledge of earlier (and later) encyclopedic efforts, and instead presented its own point of view and even commentary on the state of France and Europe which resulted in attempts at censorship, confiscation by police, orders to have copies burned, and Diderot eventually having to work in secret in order to finish (Kogan, 1958). The first truly comprehensive English language work is generally considered to be *The Encyclopaedia Britannica* originally published in weekly installments beginning in 1768 ("Encyclopaedia," 2002; Kister, 1994; Kogan, 1958; Lih, 2009) and repeatedly in fourteen subsequent editions – the most recent of which was published in 2002. Of itself, *The Encyclopaedia Britannica* claims that it has "evolved into the largest and most comprehensive general encyclopaedia in the English language ("Encyclopaedia," 2002).

Despite their attempt at being a general work of knowledge for laypeople ("Encyclopedia," 2002) and "accessible, both physically and intellectually, to students and other users in as fair, accurate, and precise a manner as possible" (Kister, 1994, p. 3), the encyclopedia has not been readily accessible to average users due to its rather large size and expense (Kogan, 1958). In 1938, H. G. Wells, in arguing for a *world*

encyclopedia pointed out that encyclopedias had largely been reserved for only an elite minority. Even today, users generally have to visit a local public or school library to use an up-to-date encyclopedia. While newer encyclopedias, such as *The World Book Encyclopedia* first published in 1917, attempted to be more family oriented, using stiffer glossy pages and color illustrations, the encyclopedia has never become a common addition to home libraries (Lih, 2009). Furthermore, due to continually evolving content, anyone who manages to purchase an encyclopedia will also find their expensive investment increasingly out of date; a problem which likely limits the number of non-institutional owners.

In the very early days of the computer revolution, the idea of an easy to use, electronic encyclopedia appeared. In his book *World Brain* (1938), H. G. Wells pointed out that

> many people now are coming to recognize that our contemporary encyclopaedias are still in the coach-and-horse phase of development, rather than in the phase of the automobile and the aeroplane. Encyclopaedic enterprise has not kept pace with material progress. These observers realize that the modern facilities of transport, radio, photographic reproduction and so forth are rendering practicable a much more fully succinct and accessible assembly of facts and ideas than was ever possible before. (p. 84)

Although Wells did not specifically mention an electronic encyclopedia, shortly thereafter, Vannevar Bush (1945) proposed what may well have been the precursor to hypertext and digital content. In laying out the foundation of his *Memex,* Bush focused on the power of “associative indexing... whereby any item may be caused at will to select immediately and automatically another.” Ultimately, he envisioned that the *Memex* would give rise to “wholly new forms of encyclopedias.”

While the *Memex* never saw the light of day, the advent of the personal computer did give rise to new forms of encyclopedias stored on optical media. In 1993 Microsoft Corporation released *Encarta* on CD-ROM. While not overly impressive, copies were often included for free in the purchase of new computer, it was often sufficient for home users (Lih, 2009). For the first time, average home users had ready access to encyclopedic content. Microsoft continued to improve its product and *Britannica* released their own electronic version in 1994 – for $995 (Lih, 2009). The rapid growth of the Internet, however, began to undermine the usefulness of CD-ROM-based encyclopedias – especially because all major players were moving toward online, subscription-based content. Seekers of information, however, found that a quick search of the Internet was becoming an effective tool for finding information and was cheaper and even faster than loading a CD-ROM or setting up a subscription. Unfortunately, such ease of access was putting users at odds with credible and legitimate information. The Internet may have become the ultimate realization of Bush's *Memex*, but instead of being deliberately filled with the collected works of humanity it was largely a playground in which anyone could post anything at any time without any sort of editorial or peer oversight. By 2000, the Internet was a wellspring of information but with increasingly divergent and competing purposes. However, in 2001, the advent of Wikipedia began to change the landscape of information seeking on the Internet.

Wikipedia

In his speech to the Royal Institution of Great Britain in 1936, H. G. Wells presented his argument for a "World Encyclopedia" and encourage his learned audience to take up the mantel. He envisioned a world-wide collaboration:

> On the assumption that the World Encyclopaedia is based on a world-wide organization he [the specialist and the super-intellectual] will be – if he is a worker of any standing – a corresponding associate of the Encyclopaedia organization. He will be able to criticize the presentation of his subject, to suggest amendments and re-statements. (Wells, 1938, p. 24)

The publishing world, however, was just not capable of keeping such a vast work "alive and up to date" (Wells, 1938). But 60 years later, the Internet would provide precisely the right combination of speed and access to allow a true world encyclopedia – "a world in which every single person is given free access to the sum of all human knowledge" (Jimmy Wales, founder of Wikipedia, as quoted in Lih, 2009) – an encyclopedia in which everyone, not just super-intellectuals, can not only suggest amendments but write and publish them instantaneously. McLuhan (1964) made a similar prediction regarding the nature of knowing and the collaborative construction of knowledge.

> Rapidly, we approach the final phase of the extensions of man – the technological simulation of consciousness, when the creative process of knowing will be collectively and corporately extended to the whole of human society, much as we have already extended our senses and our nerves by the various media. (McLuhan, 1964, p. 3)

Wikipedia might be considered a necessary outcome of technological progression. Individuals such as Bush, McLuhan, and Wells all hinted at various capabilities that have combined in the form of a large, collaborative collection of human understanding. One

wonders if Wales had not begun Wikipedia if someone else eventually would have begun something similar.

Most people know Wikipedia by what it is today – a vast, free, online encyclopedia freely accessible and editable by anyone (see figure 2). However, that is not how it started. In his book *The Wikipedia Revolution: How a Bunch of Nobodies Created the World's Greatest Encyclopedia*, Lih (2009) details how this came to be. According to Lih, Wikipedia began as a very tightly controlled project called *Nupedia*. Unlike its successor, *Nupedia* had a very convoluted process of article development. While the initial project did rely on volunteers from the start, in order to maintain integrity, authors and editors had to be carefully vetted and either hold a doctorate or otherwise be a recognized expert in their field, and each article would go through a lengthy seven-step process to ensure integrity. The process, however, proved to be too time consuming with only tens of articles produced in the first year (Rosenzweig, 2006; Lih, 2009).

Figure 2. Partial screenshot of Wikipedia main page showing the tag lines *free encyclopedia* and *anyone can edit* as well as the total number of articles in English on October 17, 2010.

Wikipedia was made possible largely due the work of Cunningham (Leuf & Cunningham, 2001) who developed the idea and implementation of *wiki* software which he called the wikiwikiweb from the Hawaiian word *wiki* meaning fast (Kane & Fichman, 2009; Lih, 2009). Simply put, a wiki is a website that can be edited by anyone (Kane &

Fichman, 2009), in the case of one that does not require registration, or only by members of a particular wiki or community. The initial iteration of wikiwikiweb was released in March of 1995.[3] Wales and co-developer Sanger eventually became aware of the wiki software and in an attempt to accelerate the slow pace of article development on *Nupedia* set up a variation of the original wiki software called *UseModWiki* which ran on a web server in January 2001. Although it generated interest, it also was criticized for its open editing process that was counter to the initial intent of *Nupedia* and a week later it was moved to wikipedia.com to continue the experiment. At that time it was still seen as part of the *Nupedia* project and articles developed there were to eventually be moved to *Nupedia* (Lih, 2009).

While ultimately a failure, the founding principles of *Nupedia* survived and ultimately gave rise to what is easily the world's largest encyclopedia (Rosenzweig, 2006). *Nupedia* took its name from the GNU Manifesto written by Richard Stallman in 1985. The manifesto laid out the ground work for the free software movement which had at its core the idea of freedom, that software users had the freedom to examine, modify and redistribute software to suit their needs. An important element of the GNU manifesto was that users not only had the right to redistribute software, they had the obligation to share back their changes and could not restrict the rights of future users to also examine, modify and redistribute (Stallman, 1985). These principals are at the core of Wikipedia which encourages users to modify and redistribute content.

[3] The initial site is still hosted at http://c2.com/cgi/wiki?WikiWikiWeb, and additional information about Cunningham's wiki can be found at http://c2.com/cgi/wiki.

Structure of Wikipedia

> Wikipedia is a free encyclopedia, written collaboratively by the people who use it. It is a special type of website designed to make collaboration easy, called a wiki. Many people are constantly improving Wikipedia, making thousands of changes per hour. All of these changes are recorded in article histories and recent changes. ("Wikipedia: Introduction," n.d.) (see figure 3)

Wikipedia is openly editable by anyone by clicking the *edit* link (see figure 4) on the top of most pages. In an attempt to limit the amount of vandalism on some of the more abused articles, Wikipedia now includes a lock feature that prevents anonymous edits. Most of these pages can still be edited by registered users; although some are only editable by administrators or Wikipedia staff ("Wikipedia: Protection Policy," n.d.).

Figure 3. Screenshot of a Wikipedia page.

Figure 4. Screenshot of a Wikipedia edit page for an article.

Popularity of Wikipedia

Wikipedia has grown to be one of the most popular sites on the web. Worldwide, according to Alexa statistics (“Alexa Top 500 Global Sites,” n.d.), Wikipedia is (as of October, 2011) the fifth most popular site on the web. Over the past six months, Wikipedia has ranked as high as fifth and as low as eighth, and as search and social media tools continue to grow will undoubtedly continue to trade positions with other popular sites as the interests of Internet users are continually in flux. However, it is likely that Wikipedia will continue to be a highly trafficked website and its popularity is continuing to grow showing a 4% increase in Internet users visiting the site in the past three months (“Alexa Top 500 Global Sites,” n.d.). As evidenced by its high ranking in

global traffic, it is not surprising that current growth is relatively low as a large percentage of Internet users are already visiting Wikipedia. Additionally, Wikipedia's article count continues to grow as well and currently contains over 3.7 million articles in English alone and 20 million, as of November 2011, in all languages combined ("Wikipedia:Size comparisons," n.d.). Similar to traffic patterns, article growth rates have fallen off in the past couple of years after exponential growth between 2005 and 2010 when it grew from approximately 500,000 to over 3 million ("History of Wikipedia," n.d.). This is likely due to the decreasing number of potential topics yet to be included.

Research on Wikipedia

Despite its popularity, Wikipedia receives a steady stream of criticism regarding its overall reliability and credibility (Emigh and Herring, 2005; Giles, 2005; Rector, 2008; Rosenzweig, 2006). Not surprisingly, former *Britannica* editor-in-chief Robert McHenry has been a vocal critic focusing on the open editing process that ensures constant change but no guarantee of improvement and places more importance on being free than it does on being reliable. He states, somewhat humorously,

> The user who visits Wikipedia to learn about some subject, to confirm some matter of fact, is rather in the position of a visitor to a public restroom. It may be obviously dirty, so that he knows to exercise great care, or it may seem fairly clean, so that he may be lulled into a false sense of security. What he certainly does not know is who has used the facilities before him. (McHenry, 2004)

One of the most widely reported events that called Wikipedia into question was the creation of a biography linking former USA Today Editor John Seigenthaler Sr. with the assassinations of President John F. Kennedy and Senator Robert F. Kennedy (Helm, 2005; Survey, 2006). Seigenthaler (2005) himself denounced the entry stating "I have no

idea whose sick mind conceived the false, malicious ‘biography’ that appeared under my name for 132 days on Wikipedia, the popular, online, free encyclopedia whose authors are unknown and virtually untraceable.” Wikipedia does not ignore such concerns and criticisms and even maintains an article on its own reliability (“Reliability of Wikipedia,” n.d.).

However, Wikipedia has achieved its phenomenal growth primarily because it opened up its editorial process to anyone and it now has approximately 3.7 million articles in English written by anonymous authors compared to Encyclopaedia Britannica's 65,000 articles in print or 120,000 articles online (Berinstein, 2006) written by their 4,800 worldwide, paid contributors (according to Tom Panelas, director of corporate communications at Britannica as quoted in Berinstein, 2006).

Despite criticisms, there have been a number of studies suggesting that Wikipedia is fairly reliable. The oft cited study in *Nature* (Giles, 2005), for example, found errors in both *Britannica* and Wikipedia. Their review of 42 science articles by content experts found only eight serious errors, defined as misrepresentations of important concepts, which were evenly split among both Wikipedia and *Britannica*. The study also found 162 factual errors or misleading statements in the Wikipedia articles and 123 in *Britannica* or an average of four in each Wikipedia article and three for *Britannica* - a difference they described as “not particularly great” (Giles, 2005). However, Internet skeptic and author of *The Shallows: What the Internet is Doing to Our Brains*, Carr (2006) noted that a more in-depth review of the study showed that it “probably exaggerated Wikipedia's overall quality considerably.” Furthermore, after conducting his own review of the study, Carr summed it up thusly:

> If you were to state the conclusion of the Nature survey accurately, then, the most you could say is something like this: “If you only look at scientific topics, if you ignore the structure and clarity of the writing, and if you treat all inaccuracies as equivalent, then you would still find that Wikipedia has about 32% more errors and omissions than Encyclopedia Britannica.” That's hardly a ringing endorsement.

Fortunately, other studies of Wikipedia have been conducted. With respect to perceived credibility, Chesney (2006) studied the perceptions of subject experts and non-experts on a variety of Wikipedia articles. A total of 258 academics (defined as research fellows, research assistants and doctoral students) were surveyed (with a 21 percent completion rate) and randomly given either an article in their own area of expertise or a random article and asked to review and assess the credibility of the article, the authors and Wikipedia in general. While both groups did not differ in their assessments of author and site level credibility, there was a significant difference in perceived credibility of articles with the subject experts rating articles more credible than the non-expert, random assignment group – suggesting a high level of accuracy in Wikipedia (Chesney, 2006). It was noted, however, that experts found errors in 13 percent of the articles which is consistent with the findings of others (Giles, 2005; Rector, 2008). Rosenzweig (2006) also found slightly more errors in Wikipedia than comparable reference works but also pointed out they were minor. Rector (2008) found that Wikipedia was less accurate than other sources (80% accuracy compared to 96% in *Britannica*). In other words, while errors persist in Wikipedia and in more traditional encyclopedias, such as *Britannica*, there is still a fairly high degree of accuracy and perceived credibility in Wikipedia. Precisely why non-experts felt articles were less credible (Chesney, 2006) was not directly addressed; although it is possible that non-experts lack sufficient background to

accurately judge an article. However, because it is reasonable to expect that many users of Wikipedia would be non-experts, providing a means by which such users can judge the legitimacy of content would be beneficial.

Magnus (2006) conducted a similar study in which copies of articles of similar depth in both *Britannica* and Wikipedia were given to experts for a blind review. The study used a small sample of three articles on somewhat obscure topics: Rawls' theory of justice, Husserl and phenomenology, and bioethics. Experts differed in their evaluations of the articles. The Wikipedia article on bioethics was called bizarre and not written by someone in the field. However, a reader of Husserl called the Wikipedia entry his favorite adding that it was how an encyclopedia article should be written. Magnus (2006) noted that variability in the quality of Wikipedia articles "should come as no surprise, since Wikipedia entries rely on contributors. Different entries will attract contributors" (p. 4). Others (Halavais, 2004 as cited in Read, 2006; Magnus, 2008) have attempted to track the longevity of errors they inserted themselves with varying results. It should be noted that intentionally inserting errors in Wikipedia is considered vandalism and discouraged (Kane & Fichman, 2009). Magnus (2006) pointed out that Wikipedia articles change over time and evaluations of old articles do not inform us about the content of newer versions. He suggested we need ways of evaluating changes in Wikipedia over time.

A time-based approach to evaluating the accuracy of Wikipedia was conducted by Luyt, Aaron, Thian & Hong (2008) who focused on the age of edits. For their study, the authors selected the same 42 articles used in Giles (2005). The earlier study included information on the exact errors that reviewers found which allowed Luyt et al. (2008) to pinpoint the versions of the Wikipedia articles where the errors were introduced. This

was accomplished using the history feature of Wikipedia that preserves every edit with a time and date stamp as well as the name of the user or IP address responsible for the edit. They referred to this process as assigning blame, and tracked the longevity of each error in terms of total number of edits between the introduction of the error and its removal, and the overall amount of time in days between the introduction of the error and the time of the review in Giles (2005). The purpose of the study was to test Cross' (2006) theory that older information that has withstood the test of time would be more accurate and that errors would be attributable to more recent edits that have not had the opportunity to be fully scrutinized. Luyt et al. (2008) found no support for this theory instead finding that at least 20 percent of errors could be attributed to the initial edit that began the article which they termed a "first-mover" effect. They concluded that attempts to validate Wikipedia content based on the age of the surviving edits would be unable to accurately account for this first-mover effect. The implication for Wikipedia and its users is that metrics such as edit age and article maturity are not going to be usable as a tool to measure accuracy or legitimize Wikipedia content.

Researchers have also attempted to evaluate the verifiability of Wikipedia articles by looking at citations. Luyt and Tan (2010) randomly sampled 50 history articles from Wikipedia and compared the citations in those articles with citations from articles in the *Journal of World History* (JWH). In the 50 Wikipedia articles they found a total of 508 citations of 480 distinct references. The 18 articles from JWH, by comparison, contained 1,877 citations of 1,351 distinct references. When comparing the types of references cited, they found 62 percent of Wikipedia citations were of Internet sources compared to 1.2 percent for JWH. Such results, they suggest, indicate that Wikipedia is reliant on low-

level, non-academic sources of information. Whether or not such comparisons are fair is another issue. Scholarly journals exist for an entirely different purpose than encyclopedias, and attempts by Wikipedia to add supporting evidence should be encouraged. Furthermore, scholarly journals tend to focus on original research, which is held to a high standard and expected citation practices. Reporting of original research is specifically prohibited in Wikipedia ("Wikipedia: No original research," n.d.) as it is primarily focused on providing information on general knowledge for laypeople similar to printed encyclopedias.

Other approaches to evaluating Wikipedia, and wikis in general (such as those used in business or the classroom), focus on measuring and evaluating editor contributions. Arazy et al. (2010) proposed a new set of algorithms to calculate authorship in wikis. They pointed out that previous methods to calculate author contributions tended to be flawed due to their focus on basic metrics automatically tracked by wikis such as the number of page edits for each unique contributor – *WikiDashboard*[4] being one such tool. Other attempts focused on evaluating a user's contribution by comparing a current version to a previous one for a particular user's contributions with the sum of all contributions providing a measure of a user's overall effort (Hess, Kerr and Rickards, 2006 as cited in Arazy et al.). Still other approaches mirror efforts currently under investigation by Wikipedia such as measuring the longevity of edits (Adler, de Alfaro, Pye & Raman, 2008; Cross, 2006, Luyt et al., 2008) which is similar to a color-coding scheme currently being explored (Claburn, 2009; Cross, 2006; Leggett, 2009), and the use of a rating system to calculate a user's reputation and, by

[4] http://wikidashboard.appspot.com/

extension, their overall level of contribution (Sabel, 2007). Key differences exist between Sabel's approach and the one currently being explored by Wikipedia ("Wikipedia: Article feedback tool," n.d.). Sabel's (2007) approach proposes weighting the similarity of page versions and assigning an *adoption coefficient* which can then be used as part of a reputation system which could function as a measure of overall contributions and reliability. Wikipedia's implementation, part of an overall strategic plan ("Strategic Plan/Movement Priorities," n.d.), has readers rate articles on four criteria: trustworthy, objective, complete, and well-written. There is also a box for readers to check if they are "highly knowledgeable about this topic." It is interesting, however, that Wikipedia does not view this feedback tool as a measure of quality or accuracy. Of the tool, Wikipedia states,

> The current version of the tool represents a starting point. The Wikimedia Foundation wants to encourage direct reader engagement as a good way to quickly elicit qualitative feedback and to make more readers aware that they can directly improve Wikipedia. We hope that this tool will help the readers in the Wikipedia community become active editors. ("Wikipedia: Article feedback tool," n.d.)

Less knowledgeable users, however, are likely to view an article with a high rating as a more trustworthy or objective article than one with a lower rating regardless of the overall intent. Furthermore, it is unclear how the Wikipedia article feedback tool would account for vandalism or the inevitable changes in articles over time.

Contrary to these approaches, Arazy et al. (2010) propose a new approach for calculating editor contributions to wikis by first breaking edits types into five categories (add, improve navigation, delete, proofread, and adding links) and measuring contributions in each category. They focused on the quantity of contributions and not the quality which they considered quite difficult to measure. They also suggested longevity

could be used as a quality measure because the evolution of wiki pages should involve the removal of low quality content while allowing high quality content to remain. Similar to Luyt et al. (2008), Arazy et al. (2010) failed to find support for this premise. Precisely why errors tend to linger has not been addressed. However, it is possible that errors not addressed within a certain amount of time tend to gain a certain level of legitimacy and may be overlooked by all but the most diligent and knowledgeable editors.

To test their approach, Arazy et al. (2010) compared their algorithms against nine randomly selected and human scored articles in Wikipedia. They found a high level of correspondence between their algorithms and human scores. The results were then used to create visualizations of editor contributions across the five categories. This resulted in several different glyphs showing relative percentage of contributions for editors and are intended to be included on the corresponding article page. These were then user tested to determine their effectiveness. They note, however, that this is contrary to the collaborative and unattributed nature of wikis, but see potential application in classroom or research settings as a way to increase motivation and participation. Teachers using wikis as class projects could also benefit from having a way to evaluate the work of individual members of a group. The value of such visualizations in Wikipedia itself are uncertain because knowing which users contributed in which way does not help us to know if those users are knowledgeable or credible. Glyphs or similar visualizations, however, could potentially be used to provide a form of feedback on articles and how they are related to other articles via the patterns of the contributors. Algorithms such as those developed by Arazy et al. (2010) could prove useful in calculating and visualizing such relationships. Knowing who the major contributors are to individual articles may

also be useful in evaluating content if one could track and measure their contributions across Wikipedia. Similar to Raymond's (1998) comment regarding Open Source software development, that "given enough eyeballs, all bugs are shallow," with enough editors, Wikipedia articles are potentially more credible and accurate. A glyph similar to the one suggested by Arazy et al. (2010) could be used by visitors to Wikipedia to easily visualize if an article was mostly written by many editors or just a few and if the edit history of those editors supports an authoritative background or not.

Other studies have focused on comparisons between Wikipedia articles and professionally maintained information stores. In their study of the accuracy of cancer information on Wikipedia, Rajagopalan et al. (2010) chose 10 articles on types of cancer to compare with the information on a professionally maintained database, the National Cancer Institute's Physician Data Query (PDQ) cancer database. With respect to Wikipedia they found that errors were rare (less than 2%). The Wikipedia articles were also found to be less readable than those on the PDQ database. Interestingly, this readability was measured using the Flesch-Kincaid grade-level scale which found a grade level score of 9.6 for the PDQ database and 14.1 for Wikipedia (higher numbers are considered less readable). This could also be interpreted as meaning that the Wikipedia articles were written at a higher level, as would be assumed from more knowledgeable authors. They also found no significant difference between the depth of coverage of Wikipedia articles compared to the PDQ database.

More recently, Arazy et al. (2011) attempted to measure how several factors, cognitive diversity, group member orientation (administrative or content), and task conflict, interact and what effect they have on the quality of information in Wikipedia.

The study used a stratified sampling approach that randomly selected 15-17 articles from six of Wikipedia's top-level categories: culture, art and religion; math, science, and technology; geography and places; people and self; society; and history and events. They sampled a total of 96 articles using Wikipedia's random article feature. A unique aspect of the study was the focus on cognitive diversity. They argued that deep-level diversity, which relates to education, expertise and knowledge,

> can enhance groups' performance, especially when the task is cognitively complex and requires multiple perspectives or entails creativity, since cognitive diversity increases the variety of perspectives brought to a problem, creates opportunities for knowledge sharing and leads to greater creativity. (p. 76)

When looking at diversity on a per article level, they found a very high level which suggests very little overlap in the activity of the contributors outside the current article. Article quality was measured using independent ratings by senior librarians at a large North American university followed by a negotiated consensus to arrive at a rating. Although article quality was not the primary focus of the study, rather the extent to which group characteristics influenced quality, they nevertheless found that article quality was moderately high scoring 4.4 on a 7 point scale.

Despite indications that Wikipedia is an often accurate and credible resource, concern over who writes the articles continues. The notion of authorship is deeply ingrained in the process of writing, citation and our overall judgement of authority and credibility. In major publication style guidelines, such as the American Psychological Association (APA) style, the Modern Language Association of America (MLA) style, The Chicago Manual of Style (CMOS) and others, prominence is placed on the author of a work. Such citations follow an author, year (APA, 2001), or author, page numbers

(MLA, 2008) format, but what is consistent is the focus on the author. Early encyclopedias, such as the *Naturalis Historia* of Pliny the Elder (77 C.E.) also considered the author as primary. Pliny referenced 473 mostly Greek authors in his 2,493 articles (Stockwell, 2000).

The role of authorship, however, has historically not been a constant. As Foucault (1984) points out in his essay "What is an Author," the importance of knowing the author of a text has changed over time. Text that we would now tend to classify as literary were at one time accepted and passed along without concern over knowing the author, while in the middle ages, scientific texts were generally only accepted as true when attributed to their author. The modern approach has more or less reversed the importance of Foucault's *author function*. Modern scientific discourse places little emphasis on the author while we place great importance on the author of literary texts. There is, for example, some debate over the true author or co-authorship of Shakespeare's works (Foster, 1999; Vickers, 2004) even though knowing the name of the author will not change the nature of those texts but could, if we can prove that it was not Shakespeare, change how they are received. Conversely, finding out that Einstein did not develop the Theory of Relativity would likely have little impact on the nature of that discovery and its use and importance in various scientific fields though it might change our perceptions of Einstein. Interestingly, Foucault does make exception for the few individuals who have essentially made certain discourses possible – what Foucault called "founders of discursivity" (p. 114). Foucault identifies Freud and Marx as examples of individuals who not only wrote their own works but also opened the door to endless possible discourse such as Freudian psychology or Marxism. That, too, may be changing as

Einstein's Theory of Relativity is now so widely accepted and intertwined in various scientific fields that it is often referred to as simply relativity, without reference to Einstein. For example, "another prediction of general relativity is that time should appear slower near a massive body like earth" (Hawking, 1988, p. 32). More recent conversations on Communism and Socialism rarely reference Marx unless it is to point out discrepancies between modern implementations and Marx' original intents.

Modern encyclopedias, however, continue to place traditional importance on the author. Both the *Encyclopaedia Britannica* and the *World Book Encyclopedia* give bylines to authors of articles. Of its contributors, the *Encyclopaedia Britannica* states,

> To meet these challenges and opportunities, Britannica has done what we have always done throughout our 240-year history: sought the very best minds in the world to help us. In the past, they had names like Albert Einstein, Sigmund Freud, Marie Curie, Bertrand Russell, T.H. Huxley, and George Bernard Shaw, all of whom were Britannica contributors in their day. (Encyclopædia Britannica Board of Editors, 2010)

Wikipedia, conversely, takes the opposite approach and relies not on the credibility and recognition of its authors but on citation and the verifiability of its content ("Wikipedia: Verifiability," n.d.) as well as an informal form of peer review inherent in socially constructed knowledge or the wisdom of the crowds (Arazy, Morgan, & Patterson, 2006; Surowiecki, 2005). The extent to which it is achieving that goal is debatable, but the shift in focus is not without merit. Foucault (1984) argued that while authorship was regarded as essential to "truth" in the middle ages, in the seventeenth and eighteenth centuries "scientific discourses began to be received for themselves, in the anonymity of an established or always redemonstrable truth... and not the reference to the individual who produced them" (p. 109). In other words, scientific discussions generally

exist separate from the author. Whether or not various areas of Wikipedia should be treated differently based their author function is another discussion. The current state of Wikipedia ensures we may never know the name, background, credentials, etc. of the true authors of each and every article. However, it may be possible to develop profiles of authors and articles through a process known as social network analysis.

Social Network Analysis

Social network analysis (SNA) is a research methodology with the primary goal of identifying patterns of social relationships based on the connections of actors to each other (Scott, 1991; Wasserman & Faust, 1997). Haythornthwaite (1996) described SNA as "an approach and set of techniques for the study of information exchange" (p. 323). The focus is on the "patterns of relationships between actors" and resources that can include actual goods and services as well as less tangible items such as information. Furthermore, according to Haythornthwaite (1996), the process is empirical and focuses on observable relationships, the networks, between the actors. Additionally, de Laat, Lally, Lipponen, & Simons (2007) suggested that SNA can help in "identifying patterns of relationship between people who are part of a social network" and "assist us in the analysis of these patterns by illuminating the 'flow' of information and/or other resources that are exchanged among participants" (p. 89). Only after an examination of these relationships are they grouped according to the strength of their connections to other regions of the network (Monge, 1987). Actors can also be members of more than one network based on their relationships. The patterns that develop help us understand with whom individuals interact and how they exchange information. Although developed well

before the advent of computers and computer networks, SNA researchers are increasingly looking at ways to understand online networks. Wellman (2001) has suggested we start to consider computer networks, which often serve to connect people, as social networks.

The field of SNA is well established and varied in its application. It has been used a wide variety of studies of human interactions in areas such as: studies of children and adolescents (Sijtsema, et al., 2010; Kobus and Henry, 2010; Van Cleemput, 2010; Witvliet, Van Lier, Cuijpers, & Koot, 2010; Fujisawa, Kutsukake, & Hasegawa, 2009; Pearson, et al., 2006; Ennett, et al., 2006; Xu, Farver, Schwartz, & Chang, 2004; Ryan, 2001; Thompson, 1996), mourning (Rubin, 1990), school leadership, reform and hiring practices (DiRamio, Theroux, & Guarino, 2009; Pitts and Spillane, 2009; Maroulis and Gomez, 2008; Penuel, Sussex, Korbak, & Hoadley, 2006), counseling research (Koehly and Shivy, 1998), online teaching and learning and computer-mediated environments (Wang, 2010; Jahng, Nielsen, & Chan, 2010; Chai and Tan, 2009; Shen, Nuankhieo, Huang, Amelung, & Laffey, 2008; Zhu, 2006) and many others.

While SNA has been used in a variety of fields, including those dealing with online communities and knowledge construction, few studies have been conducted dealing with the relationships between authors and the information they contribute. Instead, they tend to focus on the relationships between the actors. Jarkko, Ahlberg, & Dillon (2010), for example, used SNA, among other approaches in a multi-dimensional study, to explore patterns and relationships in cumulative knowledge building in online networks focusing primarily on the analysis of the content of the messages between actors. Subjects in the study were participants in a cumulative knowledge building

process in an Environment and School Initiatives (ENSI) project between 2000 and 2005. The project made use of *Knowledge Forum*[5] software which was described as:

> an open and flexible collaborative environment for knowledge building developed at the University of Toronto. When knowledge is constructed collaboratively, a shared workspace is used into which every member of the community may contribute messages (also called 'notes'). Messages may consist of text, diagrams or images. When a message is closed, an icon of it, with the title and the name of the author, is displayed. It is possible to open other people's messages and construct 'build-on-messages', and by doing so, develop the ideas of the original writer, possibly in ways that the original writer could not imagine. (Jarkko, et al., 2010, p. 366-67)

Knowledge Forum has some similarities to Wikipedia. First, it consists of actors, the writers, who generate and share ideas and information in a shared space. Unlike Wikipedia's anonymous users, in a *Knowledge Forum*, the actors are generally known. However, in both situations, the actors are able to communicate with each other (Wikipedia users interact on article discussion pages and user talk pages) during the process of building new knowledge (*Knowledge Forum*) and detailing known information and facts (Wikipedia). In the case of the *Knowledge Forum*, Jarkko, et al. (2010) focused on the messages exchanged between actors, specifically the "build-on structure of the knowledge building network" which was used to examine the relationships between who is replying to whom and frequency of interactions, and network analysis was used to "visualize the patterns and centralization of interactions and relations between nodes." Nodes were described as geodesic distances between actors and the software arranges nodes with similar sets of geodesic distances spatially close to each other.

Manca, Delfino, & Mazzoni (2009) focused on interaction patterns in educational web forums claiming that an "analysis of the communication flows that occur in

[5] http://www.knowledgeforum.com/

educational web forum may significantly help researchers and tutors to understand the nature and quality of learning processes" (p. 189). They pointed out that SNA of computer-mediated content generally focused on server log files, termed traditional structural coding, which, they argued, tended to ignore the complexities of communication patterns, for example, messages posted to all users of the forum are traditionally seen as a single communiqué, ignoring the relationship between all users. Instead, they proposed a combined semantic and structural analysis which allowed them to connect a significantly larger number of postings.

Similarly, de Laat et al. (2007) focused on relationships between actors in computer-supported collaborative learning by looking at relational data where the "unit of analysis... is not the individual, but the interaction that occurs between members of the network" (p. 89). Nevertheless, their focus was on the participation of members and how their levels of participation changed over time.

Based on an examination of the literature to date, it appears that SNA, as it has been applied to the study of online communities, online knowledge building, networked learning and computer-supported collaborative learning, has been primarily focused on the interactions between actors, their discourse, and how this information can be used to better understand the learning and teaching process as well the collaborative construction of knowledge. What is missing, is an examination of how SNA can provide insight into collaborative knowledge construction when the actors are largely unknown and inaccessible. In particular, using an SNA approach to provide a method by which we can visualize the contributions of unknown authors and draw conclusions about motivations, their level of content knowledge, the authority of the information they share, and the

overall legitimacy of the content. In other words, we may be able develop an approach using SNA that can be used to generate profiles of unknown contributors to a collaborative project and use those profiles to inform us regarding the legitimacy of the content.

In Wikipedia the anonymous authors of articles could be considered actors from an SNA perspective. The relationships between authors and articles form their own nodes and geodesic distances. An article on Quantum Mechanics, for example, is spatially closer to an article on Einstein's Theory of Relativity, by virtue of being from a related field of science, than it is to an article on Biology, a different area of science, or an article on Einstein himself (biography) or any other non-science article. Authors can also be spatially related to other authors and articles. Authors of the same article are very closely related. We can reasonably expect that many authors will contribute to more than one article and that their contributions are also spatially related (Korfiatis, Poulos, & Bokos, 2006). An author who contributes regularly to an article on Quantum Mechanics may have a strong background in science and may contribute to other articles related to science. Contributions to similar articles would be closer together spatially than article contributions in disparate fields. An author's pattern of contributions and their spatial relatedness may be used to make inferences about an author's level of knowledge and, by extension, the overall authority and legitimacy of an article.

In studies using SNA, the actors are generally known or knowable to some degree or, in other words, researchers usually have access to actors and are able to question them directly or indirectly. What has been less studied is networks and relationships between actors who are largely anonymous and known only by pseudonyms and indirectly

through the information they exchange. This type of information exchange, however, is increasingly common in online environments (for example, consumer related websites; technical, social and support group forums; and blogs and wikis). How and why we decide to trust information or not in such sites is a much larger social issue. From an SNA perspective, however, what is interesting is what we can learn about unknown actors based on the activities within the network or networks in which they participate. However, a review of the literature has not shown an established method for measuring distances between authors based on the articles to which they contribute or between articles based on authorial connections. Korfiatis et al. (2006) did attempt to evaluate authority quantitatively in Wikipedia. They mathematically calculated the degree of centrality for both articles and contributors. With respect to contributors, the degree of centrality is "a degree index of the adjacent connections between the contributor and others who edit the article" (p. 257). Furthermore,

> Contributors can be either connected (belong to the same article) or interconnected (common contributions on two or more articles in the same domain). In an article domain of high credibility it is expected that more interrelations will be found, since the contributors may contribute content to more than one article, thus depicting their common interest. Therefore, the more affiliated a contributor becomes with a domain, the more interested he/she is in the article; thus representing knowledge of the domain. (p. 256)

The authors suggest further research is needed to better define interconnectedness and the organization of topics as well as account for contributors who participate in more than one domain.

Sociological Studies of the Internet

The rise of the Internet has given individuals new social outlets and researchers a wealth of new opportunities for sociological studies. Early studies tended to fall within several sub-categories: issues of access and inequality, social capital, political participation, economic institutions, and cultural participation and diversity (DiMaggio, Hargittai, Neuman, & Robinson, 2001).

Although early implementations of the Internet were used for scientific and military communication in the late 1960s, it was not until the advent of graphical interfaces on personal computers in the mid-1990s that the Internet came to be more widely used by the general public (Abbate, 1999). A common concern at the time was that access to the Internet and access to technology in general would result in a digital divide among those that had and those that did not have access. Despite these concerns, Anderson, Bikson, Law & Mitchell (1995) suggested that the Internet could actually reduce inequality by reducing barriers to information access and making it easier for low-income individuals to gain knowledge previously inaccessible and enable them to compete for better jobs. In 1996, President Clinton established the Technology Literacy Challenge Fund that was intended, in part, to help provide equal access to technology in schools (Cuban, 2001). An interesting example of increased access is MIT's Open CourseWare program that provides lecture notes, exams and videos for 2,000 MIT classes.[6] Wikipedia was also intended to have a similar impact and provide free access to the sum of human knowledge (Lih, 2009).

[6] http://ocw.mit.edu/index.htm

Other researchers saw computer networks as inherently social and therefore able to provide opportunities to build social capital (Lin, 2001; Wellman, 2001; Wellman et al., 1996) or as a measure of recognition based performance (Okoli & Oh, 2007). Cummings, Heeks & Huysman (2006) have argued that social capital has a positive influence on knowledge sharing in online networks. Today, social networking sites like *Facebook*[7] provide an extreme example of the computer network as a social network.

Computer Networks as Social Networks

Prior to the rise of Wikipedia and popularity of *Facebook*, Wellman et al. (1996) suggested that computer networks that serve to link people should be considered as social networks. They referred to these as computer-supported social networks (CSSNs) and stated that "members of virtual communities want to link globally with kindred souls for companionship, information and social support from their homes and workstations" (p. 214). Wellman (2001) further expanded this concept by suggesting that computer networks are social networks in that they link people, organizations, and knowledge. The concept of using computers to link people and knowledge is not new and was one of the primary goals of the early Advanced Research Projects Agency Network (ARPANET) developed under funding from the Defense Advanced Research Projects Agency (DARPA). Originally designed as a defense project linking universities and research laboratories, the project is generally considered the precursor to the modern Internet (Leiner, et al., 2003). However, it was not until the early 1990s that this network of networks (Craven & Wellman, 1973, cited in Wellman, et al., 1996) became open to the

[7] http://www.facebook.com

public at large (Wellman, et al., 1996). As the popularity of the public Internet grew throughout the 1990s its was largely used as a tool to collect and share static information. This mostly text then text and graphics evolved in the early 2000s as the concept of a next generation web began to develop.

> The Web will be understood not as screenfuls of text and graphics but as a transport mechanism, the ether through which interactivity happens. It will still appear on your computer screen, transformed by video and other dynamic media made possible by the speedy connection technologies now coming down the pike. The Web will also appear, in different guises, on your TV set (interactive content woven seamlessly into programming and commercials), your car dashboard (maps, Yellow Pages, and other traveler info), your cell phone (news, stock quotes, flight updates), hand-held game machines (linking players with competitors over the Net), and maybe even your microwave (automatically finding cooking times for products). (DiNucci, 1999, p. 32)

The idea of the "ether through which interactivity happens" was an important element of what Burners-Lee called the "read/write web" and later the concept of the semantic web (Berners-Lee, Hendler, & Lassila, 2001). Over the past decade, the Internet has increasingly become a forum for both knowledge construction and social interaction. The ability to both read content on the web as well as contribute content has become commonplace. Cunningham's development of wiki software contributed to this trend and ultimately to Wikipedia, the largest current example of socially constructed knowledge.

As a vast community of knowledge creators, the contributors to Wikipedia have not only created the world's largest encyclopedia, much more comprehensive than Thoreau's (1910) abstract of human knowledge, but also a laboratory for the study of computer-supported social networks (Kane & Fichman, 2009). Additionally, Wikipedia offers an interesting opportunity to build on the work of Milgram (1967) and his Small World Problem which posited that two randomly selected people could be connected via some limited number of intermediaries or mutual acquaintances. Travers and Milgram

(1969) conducted a real world study of the Small World Problem using chain letters. In that study they randomly selected participants in Nebraska and Boston and asked them to forward a letter to a target individual in Boston by sending it on to someone they knew on a first-name basis who they felt would be closer, geographically or socially, to the target. The average number of intermediaries for the randomly selected Boston and Nebraska participants was 4.4 and 5.7 respectively. The authors concluded that the average number of intermediaries was somewhat greater than five and that other research suggested this number was quite stable. Travers and Milgram (1969) also found that letters tended to converge and pass through a small number of common individuals who they referred to as "sociometric stars." More recent studies of this phenomenon using computerized social networks such as *Facebook* and user generated content sites such as *YouTube* and Wikipedia have also been conducted (see Shu & Chuang, 2011). The next step is to explore the theoretical applicability of Milgram's (1967) Small World Problem and degrees of separation to socially-constructed knowledge. As an extensive example of a computer-supported social networks, Wikipedia offers a unique opportunity to explore relationships between articles and contributors.

Some applications of this theory have already been applied to Wikipedia in the form of games including Wikipedia's own *Six Degrees of Wikipedia*[8] which collects user discovered connections between intuitively remote articles. Connections between randomly selected articles are considered potentially uninteresting and users are encouraged to think before adding such article connections. Similar games include *The*

[8] http://en.Wikipedia.org/wiki/Wikipedia:Six_degrees_of_Wikipedia

Wiki Game[9] which is a real time multiplayer game in which two articles are chosen and players click links in the source article to try and find their way in the fewest number of clicks to the target article, and *Wikipedia Maze*[10] which is played similarly but uses predetermined puzzles which users solve and earn points.

Content Analysis

Content analysis is a research methodology that provides an established procedure for studying texts and other meaningful matter and making inferences about relationships between the content and its surrounding environment (Krippendorff, 2004). While content analysis has a long history, dating back to the beginning of conscious use of symbols and voice (Krippendorff, 2004), the explosion of digital content over the past decade makes it particularly useful today (Weare & Lin, 2000).

Krippendorff (2004) outlines several steps in conducting content analysis. The first step is to define a population of texts or messages that are the focus of the research questions and to sample from this population. Next, the researcher must identify the unit or units of analysis. Krippendorff (2004) defines three types of units: sampling units, recording/coding units, and context units. Sampling units are identified for selective inclusion in an analysis. In more traditional content analysis they could be issues of a newspaper, movies of a particular genre, or a selection of textbooks. Krippendorff (2004) identifies two necessary criteria for establishing sampling units:

1. Connections across sampling units, if they exist, do not bias the analysis.
2. All relevant information is contained in individual sampling units.

[9] http://www.thewikigame.com/
[10] http://Wikipediamaze.com/

Recording/coding units are identified for separate description, transcription, recording or coding. The recording units selected derive from the nature of the analysis and the research questions one wishes to address. Krippendorff (2004) points out that "ingenious definitions of recording units can open the door to many interesting content analyses" (p. 101). Finally, context units are textual matter that set limits on the information to be considered in the description of recording units. In other words, what context is necessary for the recording unit to achieve meaning? Words, for example, often require the context of a sentence or entire paragraph to make their meaning clear. In the case of a newspaper article, the entire newspaper could be the context unit (Weare & Lin, 2000) and, by extension, an encyclopedia could be the context unit for an encyclopedia article.

According to Weare & Lin (2000) the most important element of content analysis is the creation of categories by which messages can be validly and reliably sorted. Krippendorff (2004) refers to the organization of these categories as well as their system of measurement as data languages. A well defined data language allows for the interpretation of coding units and facilitates statistical analysis. The final steps of the content analysis involved the collection and coding of data and the analysis of the data (Weare & Lin, 2000).

Summary

Wikipedia, the free encyclopedia that anyone can edit, has become immensely popular, and currently ranks fifth overall in total web traffic ("Alexa Top 500 Global Sites," n.d.). It is home to over three and a half million articles in English. Will Richardson, educator and proponent of digital tools in education, noted that Wikipedia's

goal is “collecting the sum of human knowledge” (as quoted in Crovitz and Smoot, 2009). Jimmy Wales has made similar comments envisioning “a world in which every single person is given free access to the sum of all human knowledge” (as quoted in Lih, 2009). Whether or not this goal will be attained remains to be seen; however, Wikipedia’s unique method of development, where anyone, anywhere, at any time can make changes to articles, has raised questions of author credibility and overall article legitimacy. In order to remain relevant, up to date, accurate and ultimately useful to seekers of knowledge and information, Wikipedia must continually evolve to address concerns while remaining true to the principles which have fostered its success.

Initially, Wikipedia relied on its army of volunteer editors to police articles and address errors and outright vandalism. Eventually, policies such as Verifiability (“Wikipedia: Verifiability,” n.d.) evolved and articles are now using citations to provide a higher level of credibility. *WikiTrust* is a newer idea in which software would “assign a color code to newly edited text using an algorithm that calculates author reputation from the lifespan of their past contributions” (Leggett, 2009). In other words, it is an attempt to judge the credibility of authors based on the longevity of their contributions across Wikipedia and use that as a way to measure overall article quality. The idea is similar to Cross (2006) who proposed color coding text based on how long it has survived. However, as noted above, other research (Arazy et al., 2010; Luyt et al., 2008) has suggested that longevity of content is not a reliable method of establishing credibility.

While Wikipedia can and should explore any number of unique features to help users judge the legitimacy of articles, under its current “anyone can edit” iteration we cannot know anything about the contributors to articles. The purpose of this study is to

explore a method for profiling contributors' actions across Wikipedia in relation to selected articles with the ultimate goal of developing a theoretical framework for establishing the legitimacy of article content and providing users a tool for independently making judgements as well as to offer guidance to educators regarding how to best address Wikipedia content in their classes. Color coded text might provide a visual clue to help users spot vandalism, but it cannot tell us whether or not a contributor is an authoritative expert and knowledgeable about the subject matter, and, as noted above, studies have found that content longevity is also unreliable. Well meaning, but less knowledgeable users, may contribute misleading or inaccurate information to articles and if their contributions are not challenged or removed they could achieve a level of trust that would not be color coded by the algorithm. While color coding could become an invaluable feature of Wikipedia, it is unlikely to address all concerns regarding article legitimacy and contributor credibility.

This dissertation proposes another approach to article analysis by exploring the accidental relationships, or cognitive diversity (Arazy et al., 2011), between the contributors of selected articles. In order to have some faith in the accuracy, credibility and overall legitimacy of an article, it would be useful to know something about the contributors to that article. However, Wikipedia, by its very nature, obfuscates the nature of contributors by displaying a single, most recent version of each article which is the combined effort of all contributors to date. Previous version of all articles are also preserved but also are the combined efforts of all contributors at that time. There is no easy way to separate out the efforts of individual contributors. While in theory the contributions of a single contributor could be extracted and pasted together, the process

would be time consuming and would not provide a true representation of that contributor's contributions. Each Wikipedia article is the result of the combined efforts of multiple contributors – regardless of whether or not they are actually engaged in a purposeful collaboration in the traditional sense. As a result, each contributor is influenced by the work of the previous contributors. What an individual would have written on their own is not necessarily the same as what they ultimately wrote after reading the additions of others. In essence, with the possible exception of the initiator of an article, their contributions are tainted by what they experience when interacting with an article. Therefore, the work of any one contributor is, in effect, contaminated by the work of everyone else and cannot be considered an accurate representation of that individual's knowledge of the subject. It is entirely plausible that a reader of an article spots something they feel is inaccurate but lacks the knowledge to know for sure and proceeds to conduct some casual research on the topic (such as suggested in Harouni, 2009). After becoming better informed they may make changes to the article in question. In other words, they only made the contribution because of what they read and not because they were knowledgeable about the topic and intending to contribute. As such, it is not possible to determine cause and effect relationships regarding the contributions of individuals to a single article and, by extension, make judgements about their level of authority on a subject. However, it may be possible to observe patterns in contributions to multiple articles in a contributor's history and make inferences about their background and interests.

Rather than attempt to profile individual contributors to an article, this researcher sought to explore an alternative approach by tracking the combined efforts of multiple

contributors. In order to do this, it was necessary to develop an approach to data collection and simplification that was both reasonable and likely to produce results that would be able to help answer the research questions. The method ultimately employed in this study involved selecting prolific contributors to an article, downloading the entire contribution history for each, and then combining these histories and looking for occurrences of what this study will refer to as *accidental collaboration*. Accidental collaboration, as defined in this dissertation, refers to occurrences of two or more contributors to a purposefully selected article also contributing to another article. For example, if contributors A and B are identified as having contributed to article X (an article purposely selected for this study) also contributed to article Y then they are said to be accidental collaborators on article Y. While they are also essentially accidental collaborators on article X, in that it is unlikely they consciously intended to work together, it was the articles outside the originally sampled selections that were of interest as this process provided a method for filtering a potentially large number of articles down to a manageable number while also attempting to extract meaning from contributors' actions. Arazy et al. (2011) used a similar approach to measure the cognitive diversity of contributors to an article (i.e. the degree to which contributors efforts across Wikipedia) do not overlap. Conversely, this study seeks examples of cognitive "similarity" to use as a proxy for measuring the collective knowledge and background of contributors.

Therefore, by looking at the relationships between prolific contributors of a selected article and the full range of other articles to which they have contributed, this dissertation provides a theoretical approach for making judgements about the knowledge of contributors and, by extension, their credibility as authors of a particular subject. For

example, a prolific contributor to an article on quantum mechanics may in fact be quite knowledgeable about the topic and would, by extension, likely be knowledgeable about other related science topics and may also contribute to some of those. If we see such contributions, we might be more inclined to see such a contributor as authoritative and credible. Conversely, if we see that a contributor contributes to no other science articles or only to articles outside the area of science we might have reason to be more suspect of their contributions to the quantum mechanics article. Of course, other alternative explanations exist. The author, for example, may be a professor of quantum mechanics who simply has no interest in contributing to Wikipedia but makes an exception for this article as he knows his students tend to refer to it. However, Wikipedia articles tend to have dozens if not hundreds of contributors and their combined history of contributions, and accidental collaborations in particular, offers a potential tool for profiling and article level analysis heretofore unexplored. While it would be possible to list and categorize all the contributions of individual editors, and even attempt to quantify the quality of those contributions, Wikipedia content is socially or communally constructed and it make sense to explore the community of contributors rather than the individuals. Furthermore, while profiling individual contributors to a Wikipedia article might prove interesting and allow for some level of independent evaluation of an author's level of expertise, ultimately, we are most interested in the overall legitimacy of the article in question and that is a function of all the contributors involved in its creation. By looking at the combined efforts of an article's contributors across Wikipedia we can generate a sort of network map linking the article in question to other articles via the strength or frequency of their accidental collaborations.

Based on the cumulative efforts across Wikipedia of an article's co-contributors, we can begin to explore patterns previously hidden. Our article on quantum mechanics, for example, could be closely linked to other articles related to physics based on accidental collaborations, or it could be networked to any number of random articles depending on the contributions of the most prolific contributors. In other words, by focusing on the combined efforts of multiple contributors across Wikipedia, we can generate an article level analysis that may be useful in making judgements about article legitimacy. Due to the anonymous nature of Wikipedia contributors we can never fully know an author's motivations, credentials, level of background knowledge etc. However, we may be able to generate a profile of authors and articles that helps us to further assess the legitimacy of Wikipedia content. This dissertation proposes to test such a system.

It is important to note, however, that this dissertation will only focus on descriptive profiling of articles based on the accidental collaboration of contributors on other articles. No attempt will be made to infer anything about article quality. This dissertation makes a distinction between quality, which by definition would need to be assessed via comparison to an established reference, and legitimacy, which is used here to mean that article content is reasonably expected to be free of serious errors or fallacies and that its major contributor are reasonably knowledgeable about the subject matter to which they are contributing. Studies of article quality based on accidental collaboration would be an appropriate follow-up to the current study.

CHAPTER III

METHODOLOGY

Research Questions

The purpose of this study was to answer several questions about the nature of Wikipedia content by exploring relationships between an article and those to which it was related via accidental collaborations. The analysis of article contribution networks was expected to offer opportunities to explore the legitimacy of content, offer practical guidance to users, and inform future studies of article quality or legitimacy. Specifically, the research seeks to answer the follow questions:

Q1 What is the profile of contributions to select science articles?

Q2 What is the profile of a prolific contributor to select science articles in Wikipedia?

Q3 Do prolific contributors to select science articles in Wikipedia contribute to multiple articles?

Q4 What types of articles cluster around select science articles based on accidental collaboration and what conclusions can be drawn?

Q5 What do network maps of article clusters based on accidental collaboration say about the legitimacy of the content?

Materials

This study used content analysis of Wikipedia articles and contributors focusing on the frequency of edits to identify prolific contributors to select science articles and collecting edit frequency data from other articles to which they had contributed in order to discover if there were any patterns or relationships in Wikipedia contributorship that could be used to answer the research questions.

Wikipedia. "A free, web-based, collaborative, multilingual encyclopedia project supported by the non-profit WikiMedia Foundation. Its 20 million articles (over 3.78 million in English) have been written collaboratively by volunteers around the world, and almost all of its articles can be edited by anyone with access to the site" ("Wikipedia," n.d.).

WikiMedia Contributors Tool. "A tool written by the user Duesentrieb for showing the version history of a page, with options for sorting, filtering, grouping, and different output formats. May be used to copy a version history on a wiki page" ("User: Duesentrieb/Contributors," n.d.). It allowed for the ranking of contributors to an article by their total number of edits. No attempt was made to analyze the quality of individual edits as that was beyond the scope of this study. Based on an initial informal exploration, it was determined that users with more than 10 edits to an article would be considered a prolific contributor.

SQL Query. In order to identify the entire history of article edits for each prolific contributor to selected articles and extract the article titles and frequency of edits a query of the SQL database which houses this information was submitted. This was done

through the Query Service of the WikiMedia Toolserver which can be found at https://wiki.toolserver.org/view/DBQ.

Article Networks. Using procedures drawn from the field of social network analysis, article network tables for selected articles in the sample were created. These network tables were generated using a programming script written in C# that combined user data for each selected article into a single file and identified examples of accidental collaboration. Compiled data were output as a text file. These text files were opened using Microsoft Excel and summarized using a pivot table. Each summarized article network table connected an originally sampled seed article to other articles and included a count of the contributors who accidentally collaborated on the subsequent articles and a total edit count. The resulting tables provided a summary of the activities of the most prolific contributors allowing for an analysis of article clustering.

Research Design

Content Analysis

The research methodology used for this dissertation was content analysis. Content analysis provides an established, empirical methodology for making reasonable decisions regarding sampling, analyzing and coding data. Krippendorff (2004) defines content analysis as "a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use" (p. 18). The sample for this dissertation was a purposeful selection of science articles in Wikipedia which were further analyzed by looking at the most prolific contributors to those articles and their actions across Wikipedia. Due to the potentially volatile nature of articles in Wikipedia, it

was the original intent of this study to download and store on optical media a static copy of the content of all the articles in the sample, the contributor statistics, and all the related articles in order to preserve a snapshot of Wikipedia and allow the analysis to proceed irrespective of any changes to the live version of Wikipedia. However, this proved to be unrealistic given the size of the data sample which consisted of 180 science articles, over 1,000 unique contributors and the millions of articles to which they, as a group, contributed. It also proved to be unnecessary as the method of data extraction that was ultimately employed was done on a mirror copy of Wikipedia hosted externally and took place over a very short period of time. Furthermore, as demonstrated by Luyt et al. (2008), Wikipedia provides tools for accessing historical data. As long as Wikipedia remains an active website, future researchers would be able to access versions of the articles used in this study as they were represented at the time of this study.

The study was largely quantitative in nature as it focused on numerical data in terms of the frequency of edits by individual contributors to an article and the frequency of edits by those same contributors to other articles in Wikipedia and focusing, in particular, on articles demonstrating examples of accidental collaboration. Because content analysis is often descriptive in nature (Krippendorff, 2004; White & Marsh, 2006), some aspects of this dissertation are also qualitative such as the categorizing of articles as relating to a particular field or subject area and data regarding the background and level of knowledge of the identified contributors. This approach is also inline with recent recommendations of information systems researchers who suggest studies of Wikipedia should seek a balance between quantitative and qualitative interpretations in

order to better understand the nature of collaboration on Wikipedia (Kane & Fichman, 2009).

Using a query of the SQL database housing a mirror copy of the entire content of Wikipedia, the most prolific contributors of each article were selected. The actual number of prolific contributors varied with some articles having no contributors with more than 10 edits and some with more than 60 contributors. Therefore, the number of prolific contributors for each article varied and was is dependent on the nature of contributions to each selected article. More mature or well-developed articles, for example, would be expected to have a greater number of contributors. For this study, all contributors with more than 10 edits to a particular article were considered prolific. Articles that had only a small number of prolific contributors were handled the same as articles with a larger number; however, articles with a small number of prolific contributors were not expected to produce examples of accidental collaboration and so not all articles from the initial sample were used in the final analysis. Furthermore, articles that produced zero or only one prolific contributor were removed prior to the final analysis of accidental collaboration (because at least two contributors were needed before such examples could exist). For each of these identified prolific contributors to an article, an analysis of their overall combined contributions to articles across Wikipedia was conducted. This was done for each article that had at least two prolific contributors. Finally, a random sub-sample of the identified contributors was qualitatively analyzed in order to see what could be learned about their self-reported background and level of education. A random sub-sample of articles producing examples of accidental collaboration were also examined to create network tables.

The data for the dissertation consisted of text files for each originally sampled science article (listing each of the contributors by username and their total number of edits on that article) and separate text files for all of the identified users (listing their entire edit history to date consisting of the title of every article edited followed by the number of edits). If a contributor occurred in more than one of the originally sampled articles, the corresponding user data were only downloaded and stored once. After the data collection was completed for each of the most prolific contributors to each of the selected articles, it was analyzed to look for examples of accidental collaboration and this was compiled into article network tables. These tables constitute the essential data for this study. Content analysis is preferred as it is descriptive in nature and can be used to analyze patterns (Krippendorff, 2004). The current study is similar in some respects to citation studies (see Gall et al., 2010; Neale, Dailey, & Abrams, 2010) or those measuring the verifiability of sources (Rector, 2008), but accidental collaboration is unique to newer online outlets.

Procedure

Sampling Rationale

Krippendorff (2004) distinguishes between sampling units that are equally informative and those that are unequally informative. Random samples of articles across the whole of Wikipedia are not likely to be equally informative due to the wide variability in quantity, quality and number of contributors and edits within each article, age or maturity level of the article, as well as the differences between articles in different fields or those relating to popular culture. When sampling units are unequally

informative, random sampling is not the preferred method (Krippendorff, 2004; White & Marsh, 2006). Additionally, samples from across Wikipedia are not likely to yield answers to the above research questions because the focus is on the potential patterns and relationships that emerge when looking at the activities of contributors. A random sample, by its very nature, would not be expected to produce recognizable patterns. Furthermore, the extent to which the interconnectedness between articles and contributors can be used to make decisions about the legitimacy of information and authority of the contributors is an important element of this study. Highly disparate articles are not expected to offer an opportunity to explore such patterns. Conversely, the smaller the sampling pool, the greater the probability of observing interconnectedness and making meaningful observations. Therefore, following the recommendations of White & Marsh (2006), a purposeful sampling approach was used.

Due to its vast size, studies of Wikipedia must be fairly focused. Giles (2005) randomly selected 42 science articles from Wikipedia. Luyt et al. (2008) focused on the same 42 articles as Giles (2005). Luty and Tan (2010) selected 50 history articles from the approximately 250 articles in the special history section for their study. Arazy et al. (2010) chose only nine articles to have scored by human readers in order to measure the reliability of their algorithm for calculating editor contributions. In a cross-cultural study of articles in four different languages, Hara et al. (2010) selected 30 articles in each of the four languages chosen for the study and focused their analysis on the user talk pages. Rajagopalan et al. (2010) chose only 10 cancer specific articles to compare with a professionally maintained database.

In order to achieve an informative sample, it made sense to focus on a specific section or subsection of Wikipedia. If we hope to uncover relationships between articles based on being authored, at least in part, by the same contributors, then it makes sense to look at articles that have some level of similarity. If there are indeed contributorial relationships between articles, then a study of those relationships may reveal a tool for making judgements about the authorship and legitimacy of content. If such a relationship does not exist in topically similar articles then they are even less likely to appear in a random sample among millions of articles. Additionally, by focusing on articles that have a degree of similarity in content (such as science articles) and completeness (those identified by Wikipedia as such), then we can reasonably expect any article in this subset to be equally informative to other articles. This allows for a more in-depth analysis of a random sub-selection to be used to generalize to the larger sample. This approach was necessary due to the large quantity of data collected.

Following the models of Giles (2005) and Luyt and Tan (2010), this dissertation focused on a selection of Wikipedia articles from one of the special sections that Wikipedia maintains. Wikipedia maintains a list of portals that they believe are particularly useful, attractive, and well-maintained ("Wikipedia: Featured portals," n.d.). This currently includes 149 featured portals of which the science portal is one. The list of articles for the special section on science is further delineated into several sub-topics: formal sciences, physical sciences, life sciences, social and behavioral sciences, applied sciences, and related topics (figure 5 shows a composite screenshot of the entire list of articles as they appeared on the Wikipedia Science article on May 3, 2011 which was the date of data collection). In order to ensure the most informative data, this dissertation

collected data from all articles in these subsections. This resulted in a sample of 180 articles which are listed in Appendix A along with their corresponding URLs. Some articles in the original list found on Wikipedia included a parallel introduction article designed to make the content more accessible and less technical. Due to the overall similarity with the parent article, these were excluded. Introductory articles were present for the following topics: evolutionary biology, genetics and quantum mechanics.

Focusing on articles from this specific portal of science articles is deliberate as the research is attempting to answer whether or not similar articles exhibit similar relationships based on who authors them. While this does limit the applicability of the results to a fairly small section of Wikipedia, the purpose of this dissertation is not to make generalizations about the whole of Wikipedia, but to test whether or not cross-article authorship and examples of accidental collaboration can be useful in describing article legitimacy and by extension suggest future directions for studies of Wikipedia. Similar studies would be needed to test the extent to which such techniques could be used in other content areas or across Wikipedia as a whole.

Part of a series on **Science**

Formal sciences [hide]

Mathematics
Mathematical logic
Computer science
Mathematical statistics

Physical sciences [hide]

Physics
Applied physics · Atomic physics
Computational physics
Condensed matter physics
Experimental physics · Mechanics
Nuclear physics
Particle physics · Plasma physics
Quantum mechanics (introduction)
Solid mechanics · Theoretical physics
Thermodynamics · Entropy
General relativity · M-theory
Special relativity

Chemistry
Acid-base reaction theories · Alchemy
Analytical chemistry · Astrochemistry
Biochemistry · Crystallography
Environmental chemistry · Food science
Geochemistry · Green chemistry
Inorganic chemistry · Materials science
Molecular physics · Nuclear chemistry
Organic chemistry · Photochemistry
Physical chemistry · Radiochemistry
Solid-state chemistry · Stereochemistry
Supramolecular chemistry
Surface science · Theoretical chemistry

Astronomy
Astrophysics · Cosmology
Galactic astronomy · Planetary geology
Planetary science · Stellar astronomy

Earth sciences
Atmospheric sciences · Ecology
Environmental science · Geodesy
Geology · Geomorphology
Geophysics · Glaciology · Hydrology
Limnology · Mineralogy · Oceanography
Paleoclimatology · Palynology
Physical geography · Soil science
Space science

Life sciences [hide]

Biology
Anatomy · Astrobiology · Biochemistry
Biogeography · Biological engineering · Biophysics
Behavioral neuroscience · Biotechnology
Botany · Cell biology ·
Conservation biology · Cryobiology
Developmental biology
Ecology · Ethnobiology
Evolutionary biology (introduction)
Genetics (introduction)
Gerontology · Immunology · Limnology
Marine biology · Microbiology
Molecular biology · Neuroscience
Paleontology · Parasitology · Physiology
Radiobiology · Soil biology
Systematics · Theoretical biology
Toxicology · Zoology

Social and Behavioural sciences [hide]

Anthropology · Archaeology
Criminology · Demography
Economics · Geography
History · Linguistics
Political science · Psychology
Sociology

Applied sciences [hide]

Engineering
Agricultural · Aerospace · Biomedical
Chemical · Civil · Computer
Electrical · Fire protection · Genetic
Industrial · Mechanical · Military
Mining · Nuclear · Operations research
Robotics · Software

Healthcare sciences
Biological engineering · Dentistry
Epidemiology · Health care · Medicine
Nursing · Pharmacy · Social work
Veterinary medicine

Related topics [hide]

Interdisciplinarity
Applied physics · Artificial intelligence
Bioethics · Bioinformatics · Biogeography
Biomedical engineering · Biostatistics
Cognitive science · Computational linguistics
Cultural studies · Cybernetics
Environmental studies · Ethnic studies
Evolutionary psychology · Forestry
Health · Library science · Logic
Mathematical biology · Mathematical physics
Scientific modelling · Neural engineering
Neuroscience · Political economy
Science and technology studies
Science studies · Semiotics · Sociobiology
Systems theory · Transdisciplinarity
Urban planning

Scientific method
History of science
Philosophy of science
Science policy
Humanities
Fringe science
Pseudoscience

v · d · e

Figure 5. Composite screenshot showing the list of science articles appearing on Wikipedia May 3, 2011.

Recording/Coding

Due to the often subjective nature of content analysis, more traditional applications, such as those exploring newspaper articles or propaganda, were often faced

with difficult decisions regarding recording and coding in order to ensure replicability of the research. Web-based content adds its own challenges and benefits (McMillan, 2000; Weare & Lin, 2000; White & Marsh, 2006). The ephemeral nature of web-based, dynamic content practically guarantees that future researchers will not have access to the exact same body of material. However, its existence in a digital form does allow for some ease in archiving. Wikipedia has built-in features that address these concerns. First, every single change made to an article in Wikipedia is documented with a date and time stamp as well as the name of the user, or the IP address for an unregistered user, who made the change and often a statement concerning the reason for the change. Furthermore, the free and open nature of Wikipedia also allows one to take a snapshot of articles, or the entirely of Wikipedia if one was so inclined, at any given point in time. This feature of Wikipedia was put to interesting use by Luyt et al. (2008) who tracked errors identified by Giles (2005) back to their original source. McMillan (2000) found that there was a great deal of variability in the amount of time spent recording and collecting data in web-based studies, ranging from as quick as two days to as long as five months. The current study involved downloading and archiving a snapshot of article edits and contributor histories for all articles used at a given point in time. Regardless of the amount of time needed to code and analyze the data, the collected data will not change and future researchers interested in reviewing the data have access to sections of it as the data tables for a sub-sample of articles used are included in Appendix D. Due to the quantity of raw data, however, only the processed and simplified information is included. If needed, the raw data could be re-extracted via another SQL query. If the precise data used in this study was of interest the query could be limited to edits on or before May 3, 2011. Of

course, the content of the live version of Wikipedia will undoubtedly have changed as that is the nature of the site and its content. This study attempts to evaluate a process that can be used with an ever-changing Wikipedia, and future studies testing the persistence of the findings in this study would be interesting, but the current study relied on a stable snapshot of the data for the purpose of testing and analysis of that tool.

Data Extraction

Wikipedia offers a number of features that allow for analysis of article development, contributions by users, and various statistical data such as number of edits made by the top 10% of active users and page view statistics. Among user-related tools is the *contributors* ("User: Duesentrieb/Contributors," n.d.) tool which provides data on the number of edits per contributor for any selected article and displays rankings from most edits to least. This tool was initially used to test a data collection method using two of the selected articles (Quantum Mechanics and General Relativity) to identify the most prolific contributors (those with more than 10 edits). After identifying a few of the top contributors, their user pages and edit histories were examined in order to get a sense of how prolific they were with respect to the whole of Wikipedia and the types of articles to which they contributed. Initial observations suggested that these prolific contributors also tended to be prolific contributors across Wikipedia and generally contributed to a wide range of articles including examples of other science articles. However, manually listing prolific contributors and compiling a list of their entire edit history would have been extremely time consuming and potentially error prone. Publically accessible data were only available via formatted HTML web pages which contained an excessive amount of

extraneous content that would have to have been manually filtered. Therefore, in order to collect data for each user of each selected article, it was necessary to explore an automated process.

The approach taken here involved an automated query of the data stored in Wikipedia*'s* SQL database. For the 180 articles that were selected (as described above) the query extracted from each article the most prolific contributors and their entire history of edits across Wikipedia were cataloged. For each identified user, this resulted in a list of all article titles to which they contributed and the number of times each article was edited. The process by which this was carried out is described below.

MySQL Query

Wikipedia runs on the *MediaWiki* software developed by the Wikimedia Foundation. The first version of the software was developed to meet the needs of Wikipedia in 2002 ("MediaWiki," n.d.). *MediaWiki* is designed to work with an SQL relational database which is used to maintain a variety of data related to Wikipedia content including page titles, revision histories, summaries of changes, the names of users making changes or associated IP addresses for unregistered users, timestamps of changes, as well as the content of the articles themselves. Theoretically, anyone wanting access to this data could simply run a query of the database and request the pertinent data. In practice, however, this is not easily carried out. Wikipedia is designed as an end-user product and the database is secure and not directly accessible. Furthermore, due to the live nature of Wikipedia and the potential for content to change during the process of collecting user data it is preferable to use a static copy of Wikipedia that is updated

regularly. For those wishing to conduct research on Wikipedia, a mirror copy of the Wikipedia database is currently hosted on the *Toolserver.org* website. The website also facilitates research on Wikipedia by providing opportunities to upload tools and scripts which can be run on the mirror copy. Several of these tools were used in the course of this study. According to the site,

> The Wikimedia Toolserver is a collaborative platform providing Unix hosting for various software tools written and used by Wikimedia editors. The service is operated by Wikimedia Deutschland e.V. with assistance from the Wikimedia Foundation. It consists of thirteen servers... The contents of the live databases are replicated in three clusters: S1 (English Wikipedia), S2 (some major languages), S3 (all others), with varying degree of delay (often referred to as replag) ("Toolserver," n.d.).

One of the services provided by the *Toolserver.org* site is a database Query Service. The data required to conduct this dissertation consisted of 180 article titles, usernames of the most prolific contributors to the articles based on their number of edits, and a historical list of all other articles edited by each identified user in each of the originally selected 180 science articles and their respective edit counts for those articles. These data are stored in a MySQL database and mirrored on the *Toolserver.org* website. The Query Service provided an opportunity for this researcher to request data from the MySQL database. In order to do this, the researcher created an account on the Query Service site and submitted a query request.

On May 3, 2011, a database query request titled "Selecting top contributors with 10 or more edits from a list of science articles in the English Wikipedia and count all contributor edits for all articles they have edited across Wikipedia" was submitted. The query request, written by the researcher, was described as follows:

> This analysis is part of a graduate research project. The data will be used to explore possible patterns in the connections between articles based on how often and in what ways multiple contributors overlap in various articles. The source

> articles for the analysis are those listed as “Part of a series on Science” listed on http://en.Wikipedia.org/wiki/Science and consists of about 200 articles. I can provide a list of all article URLs if necessary or if helpful for clarity.
>
> Data from the SQL tables will be read into UCINET and used to weight edits and create an activity map based on articles most frequently edited by overlapping contributors. This will be done for each of the selected science articles and the resulting “maps” will contain the other articles the top contributors contributed to and showing the strength of the relationships.

A more detailed version of the query request and follow-up comments is included in Appendix B. Some differences in the wording of the original query request and the procedures ultimately selected for this study existed – most notably the reference to the UCINET software that ultimately was not used. Such differences were not important to this study as this phase of the process was merely the data collection. The query was accepted by the *Toolserver.org* user *Betacommand* on May 3, 2011 and results of the query were posted in the form of two zip files on May 4, 2011. The entire query required approximately one hour to complete. It is unlikely that any major changes would have occurred to articles during such a short period of time and an analysis of articles showed an average time between edits of several days. Additionally, the mirror copy on the *Toolserver.org* website is not live but is updated regularly. In the unlikely event that any of the articles did have changes and that those changes were reflected during the time the query was running the resulting differences would not have any discernable consequence on this study as it is focused on the edit histories of a large number of contributors *at the time of the data collection*. Therefore, it was assumed that the data did not contain any variations of consequence and the analysis proceeded as if the extraction represented a snapshot of Wikipedia content at the time of the query. The raw data consisted of 180 text files of prolific contributors to each of these articles and 1,061 text files of user data

consisting of their entire contribution history at the time of the query. These files were stored on a computer in separate folders called article_info and user_info respectively. Original copies of these data files were backed up in several locations. Because some manual manipulation of these text files was necessary before analysis could proceed, copies of the data files were placed in a separate location and modified. This process is described below. Backup copies of the original data were maintained in the event that the modified data became corrupted so that the analysis could be restarted. This was ultimately unnecessary as no such issues arose.

Data Cleaning

Upon receiving the data, it was necessary to inspect it to ensure integrity and conformity to the requirements. Several minor issues became apparent and were addressed. The first of these was to deal with the existence of *bots* in the data. According to Wikipedia,

> bots are automated or semi-automated tools that carry out repetitive and mundane tasks in order to maintain the 3,675,967 [as of July 18, 2011] articles of the English Wikipedia. Bots are able to make edits very rapidly and can disrupt Wikipedia if they are incorrectly designed or operated. For these reasons a bot policy has been developed. (“Wikipedia: Bots,” n.d.)

User data for 11 different bots was included as the SQL query did not distinguish between regular users and bots. Due to the existence of the bots and their extensive number of edits (*ClueBot*, for example had over 250,000 edits and *SmackBot* had more than 1.7 million edits) it was necessary to exclude them from all data analysis. Furthermore, because bots are computerized scripts and not human actors they are not the focus of this study. Additionally, Wikipedia makes use of some computerized program

scripts in order to facilitate maintenance of article content and related data. One such script was the *conversion script* which was used several years ago to migrate data. The conversion script was considered to be a contributor to articles in the same way bots and human users are. As a result, most article_info text files also contained edit counts for this script. Again, because the script is not a human actor this data also had to be removed from the files. In order to remove both the bot users and program scripts from the original article data files, each article_info text file was manually opened and visually scanned for the existence of any of these bots or scripts. If found, the bots or scripts and their corresponding article edit counts were deleted and the changed file was saved. During this process, 11 bots and two scripts were identified and removed from the corresponding article files. Nearly all article files had to be manipulated in this manner.

During this process, it also became apparent that there were a couple of inconsistencies between the kind of data that was requested via the query and what was actually received. The first of these concerned the criteria used to identify prolific contributors to the 180 selected science articles. Based on the initial manual exploration, this researcher determined that any user with 10 or more edits constituted a prolific contributor. However, the actual query was conducted by selecting users with more than 10 edits (the query script used the mathematical greater than sign which excluded an equal to option). This choice of language by the *Toolserver.org* user *Betacommand*, who wrote and ran the query, resulted in the exclusion of users with exactly 10 edits. As the initial decision to focus on users with at least 10 edits was an arbitrary decision, receiving data for users with at least 11 edits was deemed acceptable. Furthermore, the goal of the query was to collect edit histories for a selection of Wikipedia contributors. Even with

this minor change, the query identified over 1,000 unique contributors which was deemed more than sufficient for the study.

The third issue involved the naming conventions for some of the articles in Wikipedia and resulted in two different but related problems. In order to automate the query and focus on the specific articles selected, the titles of the articles as listed in the query request were used instead of the actual URLs. When providing a list of articles to the query service, the names of articles as presented on the Wikipedia Science article page[11] as it existed on May 3, 2011 were used. However, some article titles did not precisely match the actual URL for the same article. For example, the title given for the article on "agricultural engineering" was simply titled "agricultural" and was listed under the broader heading of "engineering" but the corresponding URL for the articled contained the title "agricultural_engineering." The query proceeded by requesting data for article titles and not article URLs and in this example for the article "agricultural". However, there was no actual corresponding article titled simply "agricultural" (only a dummy page that redirects to a separate article titled "agriculture"). As such, the query did not return any data for that particular article. This type of query error occurred 13 times. Somewhat similarly, the second type of error occurred when the article titled used for the query did not refer to the exact article in the Science subsection but to a different article. For example, the article on "software engineering" was titled simply "software" and there was a corresponding article on "software" which was a separate and unrelated article. Similarly, a few articles ended up directing to an older article that had been merged with another, such as the article on "stellar astronomy" which had been merged

[11] http://en.Wikipedia.org/wiki/Science

with the article on “astronomy” in 2007, While it still existed in the database, it lacked contributors with more than 10 edits which was likely due to the relative immaturity of the article at the time. Because articles existed, albeit different articles from those intended, this resulted in user data being collected for what was essentially the wrong article (an article that existed outside the selected science articles) and not part of this study. Although most of these errors did result in the collection of user data, the articles were outside the intended sample and so were rejected. This type of error resulted in an additional 13 articles being rejected. As a result, out of the original 180 articles selected for this study, 154 (85.6%) of the original sample, were identified as useable. This was still much greater than the number of articles used in prior studies such as Giles (2005); Luyt et al. (2008); Lute and Tan (2010); Arazy et al. (2010); Hara et al. (2010); and Rajagopalan et al. (2010).

Finally, a further seven articles (Atomic physics, Behavioral sciences, Computational linguistics, Galactic astronomy, Interdisciplinarity, Systematics, and Theoretical chemistry) did not produce any prolific contributors. These articles were not zero data articles, because they did exist, or mistakenly sampled like those described above, but simply did not have any editors, excluding bots and scripts, meeting the cutoff of more than 10 edits to be included. These articles were also excluded from further analysis because they did not contain any data to analyze. As a result, 147 articles (81.7%) of the original sample, were ultimately useable for the study.

The incorrectly sampled articles discussed above also led to issues with the user_info data. The 13 articles identified as outside the scope of this study and mistakenly sampled (as described above) nevertheless resulted in the collection of user data for those

users identified as prolific contributors to the articles. This did not pose a problem for the analysis of accidental collaborators described below, but ultimately posed some problems during the analysis of contributors.

Over one thousand unique users were identified as prolific contributors to the 180 articles originally sampled. Although only 147 articles were used in the analysis of accidental collaborators, it was not easy to identify which of the over one thousand users contributed solely to an excluded article and should also be removed from this analysis as well. For the analysis of the 147 articles, this was not really a concern because the process for analyzing these articles only focused on the contributors identified in the respective data files and any user who did not contribute to at least one of these final 147 articles was simply ignored. However, it was possible that some of these users could have ended up being sampled as part of the qualitative analysis of self-reported user background described below as well as to the descriptive profiling of contributors. It was therefore necessary to devise a method for identifying users who were ultimately not part of this study by virtue of not contributing to any of the articles that were part of the final analysis.

In order to identify users that should be excluded, a file containing all contributors to the remaining 147 articles was created by combining the data in each article data file into a master file using a command line concatenate function. This file was then loaded into an Excel spreadsheet and sorted. Because some users contributed to more than one of the 147 articles, there were numerous duplicates. This was not unexpected, but until this problem was encountered it was not considered useful information. However, once recognized, this proved to be additional information that could be used to create

contributor profiles and was included and is described in the results section. Using the "remove duplicates" function in Excel, duplicates were removed leaving behind a list of users that had contributed to at least one of the 147 articles. This resulted in a final count of 974 contributors remaining for analysis. There were, however, 1,048 (excluding the bots and scripts) contributor data files and it was necessary to identify which of these should be removed before processing. This was accomplished by using the *vlookup* function in Excel. All 1,048 contributors were entered into a second column and then compared against the list of 974 contributors to search for names of those no longer included. Once identified, their corresponding contributor data files were removed and the analysis of the remaining contributors proceeded.

Finally, some inconsistencies in user files were encountered. For an unknown reason, a number of users were sampled who did not appear to meet the criteria for sampling. This criteria was that they had contributed more than 10 edits to one of the initial 180 science articles. A total of 61 users, most of whom were unregistered IP users, who did not have more than 10 edits to one of the science articles were identified. It is unknown why this happened. One possible explanation is that these initially IP only users eventually registered an account and were somehow linked in the SQL database. Regardless of the reason, they existed as separate users but had an insufficient number of edits to qualify for analysis. These users were also removed prior to the analysis of contributors. Once this process was completed, a total of 913 users remained and were used in the various analyses.

Analysis of Contributors

In order to provide descriptive statistics on the remaining 913 contributors to the 147 articles used in this study, a small programming script written in PHP was used to count the total number of articles each contributor had edited as well as sum all of the edit counts to each article. This provided a method for averaging total article counts and total edit counts among all contributors as well as calculate standard deviation and minimum and maximum edit counts. These data are included in the results in order to better understand the degree to which contributors identified as prolific for the purpose of this study and analysis of the sample articles were also prolific across Wikipedia.

Due to the existence of often large numbers of articles edited only a few times and many just once, it was decided to exclude these low edit count articles and recalculate the data for comparison purposes. This was done by using a small programming script written in Bash that parsed each of the 913 user files and looked for articles with an edit count of greater than five. These were saved while all others were deleted. These trimmed data files were then recombined in the same way as the original files were done above and descriptive statistics recalculated. While excluding a large number of low edit count articles will naturally raise overall averages, this information is provided in order to get a better sense of the contributions of the identified users with respect to the articles in which they have demonstrated a more vested interest. Wikipedia provides some encouragement to users to increase their overall edit count. Excluding these potentially more trivial edits provides a more realistic view of a user's efforts. However, both sets of data are included for the purpose of comparison.

Qualitative Analysis of Wikipedia Contributors

In order to develop an understanding of Wikipedia contributors and determine what, if anything, can be directly learned about them, it was decided to collect qualitative data on a random sub-selection of contributors identified during the process of data collection for this study. For each article selected for this study, contributors with more than 10 edits were identified and the entire edit history of those contributors was collected. Not counting bots, automated scripts, and users removed due to not having contributed to the final 147 articles or otherwise excluded as described above, this resulted in 913 unique contributors out of the original sample of 1,048. Due to the large number of contributors and the time needed to manually inspect each user's page, it was determined that a 10% random sample would be sufficient to build a profile of the contributors. In order to select a sample of these contributors, their usernames were read into a spreadsheet, one per row, and alphabetized. Using a random number generator from the *random.org* website, 105 random numbers between 1 and 1048 were selected. The entire sample of 1048 users was used at this stage as it was not yet known that some users would ultimately be rejected. Of the 105 users randomly selected, 4 were ultimately rejected. As such, it was determined that the 105 users originally identified during this process constituted a fair and untainted sample. The 4 users were removed and the remaining 101 users were analyzed.

The next step involved manually visiting the user pages of the selected contributors. Each page was reviewed to examine what types of information users elected to share. If a user pages existed, it was examined. If a user's page existed but was blank or contained little information, the edit history of the user page was examined to

determine if it contained user content at some time in the past and if so it was also examined. Several categories of data were specifically collected: educational background and degree, professional background, areas of interest, edit count, date of most recent update and additional comments of interest. Additionally, the top five articles based on edit count were also extracted from the downloaded user data files and added. Once data were collected, each user was examined for evidence of a scientific background. This was based on the judgement of the researcher but was primarily determined by whether or not they noted a college degree or a profession in a scientific field. This categorization method was selected because one of the key interests of this research was to attempt to discern if contributors to science articles were knowledgeable and qualified to write on these topics. The primary method for exploring this was to filter the edit history of contributors to selected science articles by looking for occurrences of accidental collaboration and then categorizing the related articles to see if there was a tendency for contributors to the selected science articles to also contribute to other science articles. The rationale being that anyone making significant contributions to multiple science articles might also be more knowledgeable about these topics as compared to someone who contributes only rarely. These primary findings can be further supported or thrown in to greater doubt by also collecting some qualitative data on the reported background of a selection of these contributors. Appendix C contains a partial table of the raw data for the contributors identified as having a scientific background.

Analysis of Articles

Similar to the analysis of contributors, the 147 articles used in this study were also analyzed for the purpose of providing descriptive statistics on their contribution profile. Using a PHP program script similar to the one described above, the total number of prolific contributors was counted and their combined edit counts summed. These data are also included in the results in order to provide a sense of article construction patterns and maturity.

Additionally, a random sub-selection of approximately 10% (15 articles) of the articles ultimately identified as useable in the study were also selected and analyzed by looking at the revision history statistics automatically provided on the view history page for each article. The revision history statistics tool is another automated query service provided by *toolserver.org* and includes statistics on the total number of revisions (edits), number of minor edits, number of IP edits (generally unregistered or anonymous editors), date of the first edit, date of the most recent edit, average time between edits, number of edits in the last day, week, month and year, number of users (contributors), average edits per user, and the number of edits made by the top 10% of active users. The tool also provides graphs of edit activity by year and month as well as graphs showing changes in article size over time. The most interesting of these statistics were the total number of edits, the number of IP edits, first and most recent edits (reported month and year), the average time between edits, the total number of users (contributors), and the number of edits by the top 10% of active users. IP edits are a rough estimation of vandalism. While not all IP users (those who have not created an account) are vandals, these are anonymous, unregistered users who have not, or had not at the time of the edits, taken an

active interest in Wikipedia. The number of edits contributed by the 10% of active users provides a complementary measurement to IP edits as these users have clearly taken an active role in article development. Furthermore, while not precisely related to the prolific contributors identified in this study, these users are nevertheless the top or most active contributors to each article. No attempt was made to determine the relationship between prolific contributors identified for this study and the top 10% identified by the tool. However it is reasonable to assume that there is substantial overlap. Due to the wide variation in the total number of edits among the different articles, it made sense to look at the percentage of edits attributable to these top users. The *toolserver.org* revision statistics tool included data on the number and percentage of edits by the top 10% of active users.

For each of the randomly selected articles, the revision history statistics tool was queried and a screenshot of the data was taken on Aug 28, 2011 for later compiling and analysis. This data, is reported in the results. The random sub-sample was selected by first ordering the 147 articles alphabetically, numbering them from 1 to 147 and then using the random number generator hosted on *random.org* to select 15 numbers between 1 and 147. The corresponding articles where then identified. These articles were:

Archaeology
Bioethics
Biomedical Engineering
Cell Biology
Cryobiology
Geography
Geomorphology
Linguistics
Logic
Mechanics
Neuroscience
Nuclear Chemistry

Operations Research
Philosophy of Science
Urban Planning

Identification of Accidental Collaborators

The most important data for this study consisted of the articles identified as having accidental collaborators. For the remaining 147 articles used in this study, the next step involved analyzing the contributions to each one and identifying all other articles edited by the most prolific contributors (those with more than 10 edits in the selected science article) to look for occurrences of accidental collaboration. Once again, due to the large amount of data and difficulty in manually combining, counting and extracting the pertinent information, this process was accomplished by using a program script written in the C# programming language. This process is described below.

Additionally, when examining the text files of users' edit history, it became clear that users often had contributed to hundreds and sometimes thousands of articles and often with only one or two edits. The user *Vsmith*, for example, contributed over 80,000 edits to over 12,000 unique articles. However, this works out to an average of only 6.3 edits per article suggesting a large number low edit articles. In fact, only 1636 articles, or 12.8%, had 10 or more edits, and *Vsmith* had contributed more than 100 edits to only 94 articles. Similarly, the user *Rjwilmsi* contributed edits to over 388,000 unique articles with 82.13% of those articles having only one edit. On his user page, *Rjwilmsi* described himself as a "stickler for grammar" and it is likely that the vast majority of these edits concerned punctuation corrections. As a result, users who only edited an article a few times were assumed to have little interest in that article beyond random maintenance or

possible reversion of vandalism. In fact, Wikipedia has a number of features that actively encourage “drive-by editing” or editing merely for the purpose of increasing ones edit count. Furthermore, prolific contributors to the selected science articles were those with more than 10 edits. For the sake of consistency it was decided that including articles from a users’ edit history with very low edit frequency could unnecessarily skew results. Therefore, when examining a contributor’s history of edits, it was decided to look only at articles that had more than five edits. While this was a somewhat arbitrary decision, five edits was chosen as the limiter in order to provide an opportunity to simplify the data and remove potentially more trivial edits. Additionally, a frequency analysis of all article contributions by all users showed a very large number of articles (70.2%) had only one edit, and there was a steep drop-off between two to five edits. For edit counts of six or more the distribution flattened out suggesting this was a useful cutoff. This is likely due to specialization among prolific Wikipedia contributors. In other words, users tend to focus their efforts around articles in which they have a particular interest which results in a higher total edit count for those articles compared to articles in which they have little interest and merely happen upon as they spend time in Wikipedia or articles they edit for reasons unrelated to content, such as fixing grammar. In the case of *Vsmith* this resulted in the exclusion of 9,864 articles and 386,476 articles for *Rjwilmsi*, but still preserved the possibility of finding occurrences of accidental collaboration. Given the large number of articles found using this process, the exclusion of low edit count articles did not appear to constrain the results and in fact a higher cutoff might have been prudent.

In order to extract article titles with an edit count of more than five from each prolific contributor to a selected article and look for occurrences of accidental

collaboration, a programming script was written that automated the concatenation of user files for each seed article and the filtering of low edit count articles. For each of the 147 seed science articles, the script output a text file containing only the usernames, article titles and edit frequency for each article that was edited by two or more contributors of the original seed science article. Not all articles resulted in occurrences of accidental collaboration. Out of the 147 articles processed only 117 produced examples of accidental collaboration.

Analysis of Article Network Tables

The initial goal of this study was to provide an analysis of each science article in the sample. After filtering the raw data to extract only examples of accidental collaboration over 10,000 articles were identified among 117 seed articles (30 articles did not produce any examples of accidental collaboration). Manually categorizing these articles according to whether or not they related to a scientific field, a non-scientific but academic field or a non-academic topic proved to be a daunting task. Therefore it was decided to use an approximately 10% sample (12 articles) of the 117 articles identified as having accidental collaborators. This process is similar to the one described above. For each of these 12 articles, a data table was constructed listing the title of each article exhibiting accidental collaboration, excluding reflexive references to the originally article (i.e. the article on Chemistry did not include data for the contributors to the Chemistry article, but, for illustrative purposes, this reflexive reference is included for the Cell Biology example below), the total number of accidental contributors (AC) to each of the articles, and the total number of edits to the articles by the contributors. Table 1 shows a

sample table for the article on Cell Biology (which was not one of the 12 analyzed). For the seed articles in the sub-sample, these examples of accidental collaboration were later categorized based on their relationship to a general subject area.

Table 1.

Example of an article network table (Cell Biology) showing examples of accidental collaboration.

Article title	Count of AC	Sum of Edits
Cell_biology	4	59
Alzheimer's_disease	2	28
Biology	2	19
Biotechnology	2	16
Cell_(biology)	2	33
Cell_division	2	14
DNA	2	55
Drosophila_melanogaster	2	33
Eukaryote	2	12
Life	2	20
Pope	2	33
Privatization	2	26
Protein	2	31
Satellite	2	22
Average	**2.14**	**28.64**
Standard Deviation	**0.53**	**14.00**

It was originally planned that this data would then be analyzed using software capable of graphically displaying the articles using the frequency of accidental collaboration and total edit counts as variables. However, no such software could be located and graphing tools such as those included in Microsoft Excel tended to see the data as linear. Initial attempts to use these graphing tools showed that there was no logical reason for placing one article ahead of another on a linear graph. The only options

were to use either an alphabetical order approach or to base relationships on the frequency count of articles or edits, but each approach produced a liner graph that implied relationships that did not exist nor did this approach facilitate the reading of the data. Another approach that was considered was a quadrant style graph that placed the originally sampled article in the center and the remaining articles showing accidental collaboration placed in one of four quadrants (upper left, upper right, lower left, and lower right) corresponding to their similarity or differences with the originally sampled article (respectively: articles in the original sample of science articles; other science articles outside the sample; non-scientific but related topics, such as a biography on Albert Einstein; and entirely unrelated to the science sample). However, it was determined that this approach was overly subjective and ultimately provided little useful information, and, as with a liner graph, did not greatly facilitate the reading or making sense of the data.

Ultimately, it was determined that using a simple data table, such as the example in Table 1, was the most informative. The relative size of each table gives a visual representation as to the number of articles producing examples of accidental collaboration as well a how often these occurred and a sense of the scope of the editing by the collaborators. Tables for each of the 12 randomly sampled articles that produced examples of accidental collaboration are included in Appendix D including the corresponding subject categories to which articles were assigned. Finally, a summative data table was created for the 12 seed articles analyzed showing the percentage of articles matching one of three general categories: science related article, non-science but academic article, non-academic article.

The process of categorizing articles was somewhat challenging. The 12 randomly selected articles used for this phase of the study produced over 4,800 unique articles (excluding duplicates) using the accidental collaboration process. There was also a very large difference in the number of articles found for each of the 12 seed articles. Limnology, for example, produced only one example of accidental collaboration and Quantum Mechanics produced 3,547 unique articles. Chemistry and Geography also produced large numbers of articles (2,199 and 481 respectively). Because this sub-sample still resulted in a very large number of articles to manually classify, it was decided to look at only articles with higher numbers of accidental collaborators when they existed. Therefore, for the articles on Chemistry and Quantum Mechanics, only the 280 and 472 articles (respectively) with four or more accidental collaborators were categorized. For the Geography article, only the 85 articles with three or more accidental collaborators were categorized. For the remaining nine seed articles all subsequent articles were categorized. This resulted in 809 articles (excluding duplicates) for categorization.

With respect to these articles, many were clearly scientific in nature such as those relating to chemical elements, compounds and processes, physics, space, subatomic processes, biology, zoology, geology, etc. Others, such as articles on countries, or those relating to geographic features such as rivers, mountains or volcanoes were more difficult. One of the originally sample scientific articles was one simply titled "geography" and geography is generally considered a branch of earth science. However, is an article on New York or South Africa more about the land and climate and therefore geographic or about the people and customs so more of a cultural article or is about the history of the place? The article on Antarctica, for example, included subsections on

geography, geology and biodiversity as well as history. Rather than attempt to make these distinctions, any article that was generally related to a physicality defined by its location on the Earth was considered “geography”. Likewise, articles relating to peoples or events from the past were considered “history” as were articles dealing with objects such as the calculator, clock or electric guitar. While such articles did deal with scientific elements such as sound reproduction or mechanics, in general these articles concerned their historical development. Any article generally about a tenant of a major religion, deities, or topics generally inspired by religions beliefs, such as intelligent design, were categorized as “religion”. Articles related to societal beliefs, actions, and norms were categorized as “sociology”. Any article about a person either living or dead was categorized as “biography” but those related to individuals noted for their achievements in a scientific field were categorized “biography – scientist” in order to later include these as examples of articles relating to science. These categories accounted for the vast majority of the articles in the list. The remaining were categorized according to the closest fitting general topic such as art, entertainment, literature, mythology, sports, etc. Although such categorizations are subjective, it is believed that the criteria used were consistent with generally accepted norms. The complete list of articles related to their respective seed article including how they were categorized is included in Appendix D. Descriptive statistics relating to the number of articles identified as showing accidental collaboration per each originally sampled article were also calculated.

Limitations

As with all research, certain decisions were made in the interest of practicality, time, or cost restraints, and these decisions limit the overall generalizability of the results. For this dissertation, the following limitations have been identified.

First, while it is of interest to this researcher, the current study will not attempt to answer questions of overall article quality. By definition, questions of quality should be addressed via established methods involving comparisons with known authoritative sources or other external benchmarks which is outside the scope of this research. It is hoped that future research will attempt to establish a relationship between article profiles that are explored in this study and overall article quality (see Arazy et al., 2011 for a related approach) but this dissertation will only focus on exploring a procedure for generating article profiles. The study does take a first step, however, by developing a theoretical framework for assessing the legitimacy of collaboratively constructed information.

Furthermore, this dissertation makes no distinction between the type or quality of article edits. As described above, each saved change to an article is counted as an edit, but each edit could involve deletion or addition of large or small amounts of text or even minor punctuation changes. Edits can also be considered vandalism or the reversion of vandalism. There are obvious differences between a change that involves the addition of a substantial amount of text and one that simply reverts vandalism or corrects spelling or punctuation. This study will only look at the quantity of edits contributed by prolific contributors without regard to the actual type or quality of the edit. By looking at the most prolific contributors, it is assumed that the distribution of edits among the various

types is reflective of the typical article and normally distributed among prolific contributors. In other words, it is assumed that prolific contributors to an article make a variety of edit types but by virtue of being a prolific contributor they have nevertheless contributed significantly to the article. Anecdotal evidence suggests approximately half of all edits to well established articles are made by the top 10% of contributors. To test this assumption, data were also included for a random sub-sample of articles from the original sample. Again, future research may be necessary to test whether or not quality of edits has any impact on the generation of article profiles.

Finally, this dissertation is also focused on a specialized subsection or subculture of Wikipedia. Vast differences in the quality, content, and overall age or maturity level of articles in Wikipedia prevents any sort of generalization of findings beyond the population of science articles from which the sample was drawn. It is assumed that the quality of content in articles in a specialized science section will differ from that in other sections of Wikipedia such as those dealing with popular television episodes or biographies of actors or musicians. It is nevertheless likely that the theoretical framework studied here can be applied to other areas of Wikipedia and further research would be needed to explore this.

CHAPTER IV

RESULTS

This section includes a description of the results of the study with respect to the research questions. Compiled data for this study are contained in the various appendices and referenced here.

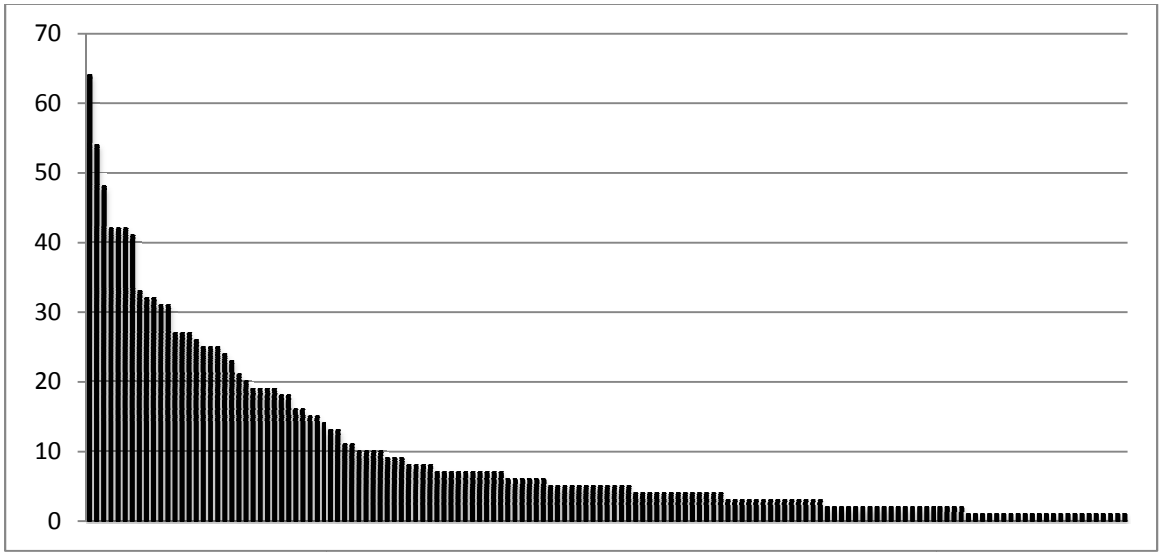
Q1 What is the Profile of Contributions to Select Science Articles in Wikipedia?

The first question concerned the profile of contributions to the selected science articles. For the 147 articles that were ultimately used in this study, there was an average of 9.66 prolific editors (s.d. = 11.85). The distribution of prolific editors by article is shown in Figure 6. The largest number of prolific contributors to a single article was 64 (Pseudoscience). It was also noted that 109 articles (74%) had 10 or fewer prolific editors, and there were a total of 23 articles with only one prolific editor that was not a bot or automated script.

The 147 articles had an average of 454.65 edits (s.d. = 606.87) from prolific contributors. The distribution of the sum total of edits by article is shown in Figure 7. The fewest edits was 11 from just one editor (Condensed Matter Physics). The greatest number of edits was 3,554 from 42 editors (Biology). Because editors differed in their

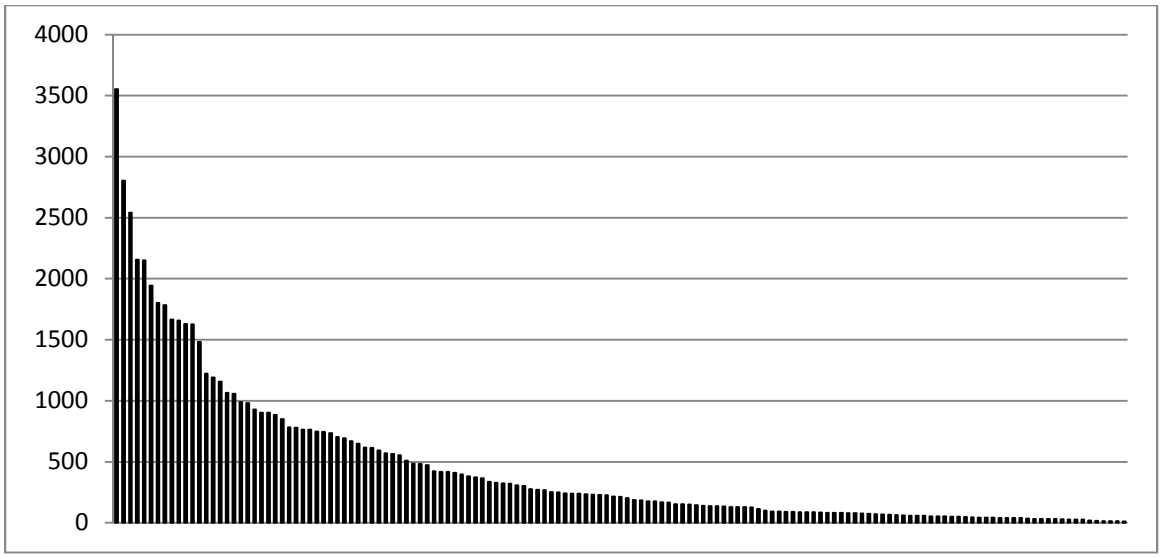
productivity, it was also important to look at the average number of edits per editor per article. This average was 46.57 (s.d. = 28.80). The distribution of the number of edits by editor for each article is shown in Figure 8.

As one might expect, there was a significant positive relationship between the total number of edits to an article and the number of editors, $r(147) = 0.855$, $p < .0001$, indicating that as the number of editors to an article increases so does the total number of edits. However, this is not a perfect correlation suggesting that there is variability in the productivity of various editors, but 73% of the variability in edit counts is explained by the number of editors.



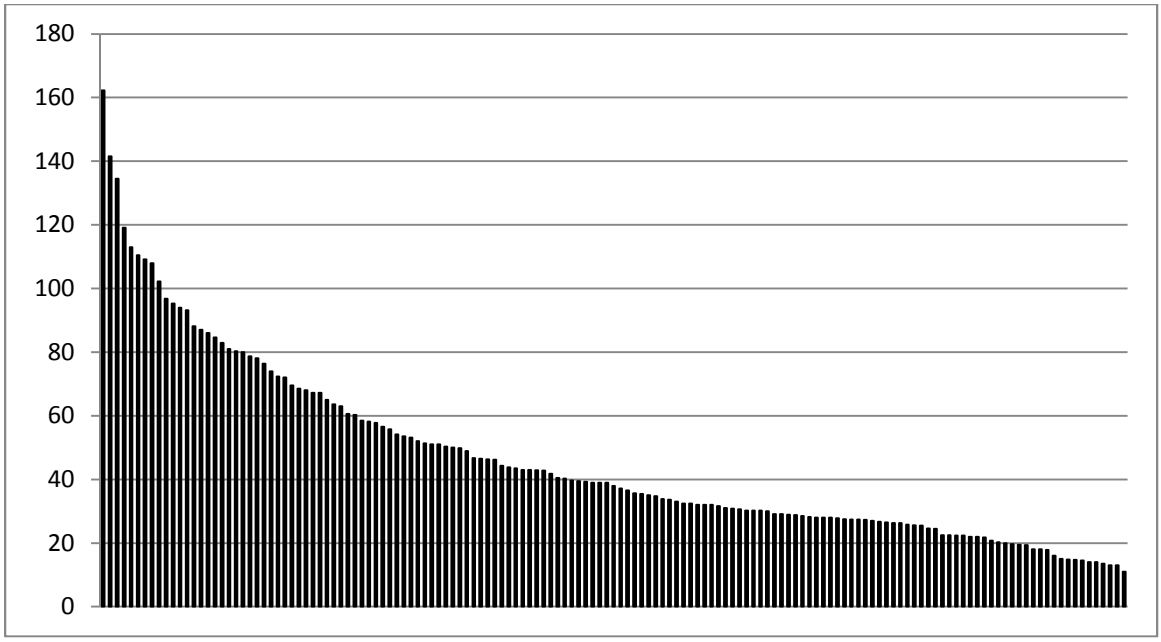
Note. Articles ordered left to right by descending number of editors.

Figure 6. Distribution of number of prolific editors by article.



Note. Articles ordered left to right by descending number of total edits.

Figure 7. Distribution of the sum total of edits by article.



Note. Articles ordered left to right by descending number of edits per editor per article.

Figure 8. Distribution of the number of edits per editor for each article.

A sub-sample of articles was identified in order to further examine data on the edit history of articles. Of the 147 original articles, 15 articles (or roughly 10%) were

randomly selected and examined using a revision history statistics tool hosted on the *toolserver.org* website. Data for the sampled articles are presented in Table 2.

Among the sub-sample of articles, the most infrequently and frequently edited articles (Cryobiology and Archaeology) had 263 and 3,120 total edits respectively. The overall average number of total edits was 1,225.93 (sd = 910.19) with an average of 501.80 IP edits (sd = 346.74). Cryobiology had the fewest number of editors with 120 and Geography the most with 1,616. Total number of editors averaged 698.67 (sd = 445.90). The percentage of edits made by the top 10 percent of active editors ranged from a low of 25.55% (Nuclear Chemistry) to 52.29% (Cryobiology). The overall average percentage of edits by these top editors was 40.54% (sd = 7.13%). Articles were generally mature with the oldest article (Logic) begun in February 2001 and the newest article (Nuclear Chemistry) in June 2003 (Wikipedia began in January 2001). The articles had an average age of 3,449.93 (sd = 244.17) days or almost nine and half years.

Based on this data, a typical science article in Wikipedia is one that has had around 700 different editors. They would have made about 1,200 edits. The top 10% of active users would likely have contributed 40% of the edits. Similarly, 40% of the edits would have come from anonymous IP editors. The article would be mature (approximately 9.5 years old), which is not surprising because science tends to be a traditional encyclopedia topic and would have attracted Wikipedia editors early in its history.

One example (see Figure 9) of a typical article is *Neuroscience* (“Neuroscience,” n.d.). Examples of some of the edit changes that occurred to this article are outlined below.

Table 2.

Summary data for the random sub-sample of 15 articles.

Article Title	**Total Edits**	**# of IP edits**	**First Edit**	**Most Recent Edit**	**Maturity (days)**	**ATBE**	**# of users**	**Top 10% Edits**
Archaeology	3,120	1,236	11/3/01	8/24/11	3,581	1.15	1,533	43.13%
Bioethics	1,145	504	3/5/03	7/24/11	3,063	2.68	594	41.48%
Biomedical Engineering	1,219	601	12/17/01	8/19/11	3,532	2.9	650	38.62%
Cell Biology	756	320	10/26/01	8/20/11	3,585	4.75	477	32.23%
Cryobiology	263	105	5/1/03	8/17/11	3,030	11.52	120	52.29%
Geography	2,922	1,191	11/4/01	8/11/11	3,567	1.22	1,616	38.30%
Geomorphology	480	135	2/13/02	7/22/11	3,446	7.18	261	41.25%
Linguistics	2,283	604	11/14/01	8/10/11	3,556	1.56	963	51.90%
Logic	2,225	741	2/14/01	8/14/11	3,833	1.72	1,115	46.20%
Mechanics	652	267	8/15/01	8/25/11	3,662	5.62	402	33.13%
Neuroscience	1,172	374	11/22/01	8/10/11	3,548	3.08	633	40.61%
Nuclear Chemistry	274	114	6/8/03	7/18/11	2,962	10.81	210	25.55%
Operations Research	777	313	3/9/02	8/24/11	3,455	4.45	448	37.65%
Philosophy of Science	1,330	444	1/29/02	8/23/11	3,493	2.63	634	46.57%
Urban Planning	1,421	578	3/21/02	8/17/11	3,436	2.42	824	39.18%
Average	***1,335.93***	***501.80***			***3,449.93***	***4.25***	***698.67***	***40.54%***
Standard Deviation	***910.19***	***346.74***			***244.17***	***3.28***	***445.90***	***7.13%***

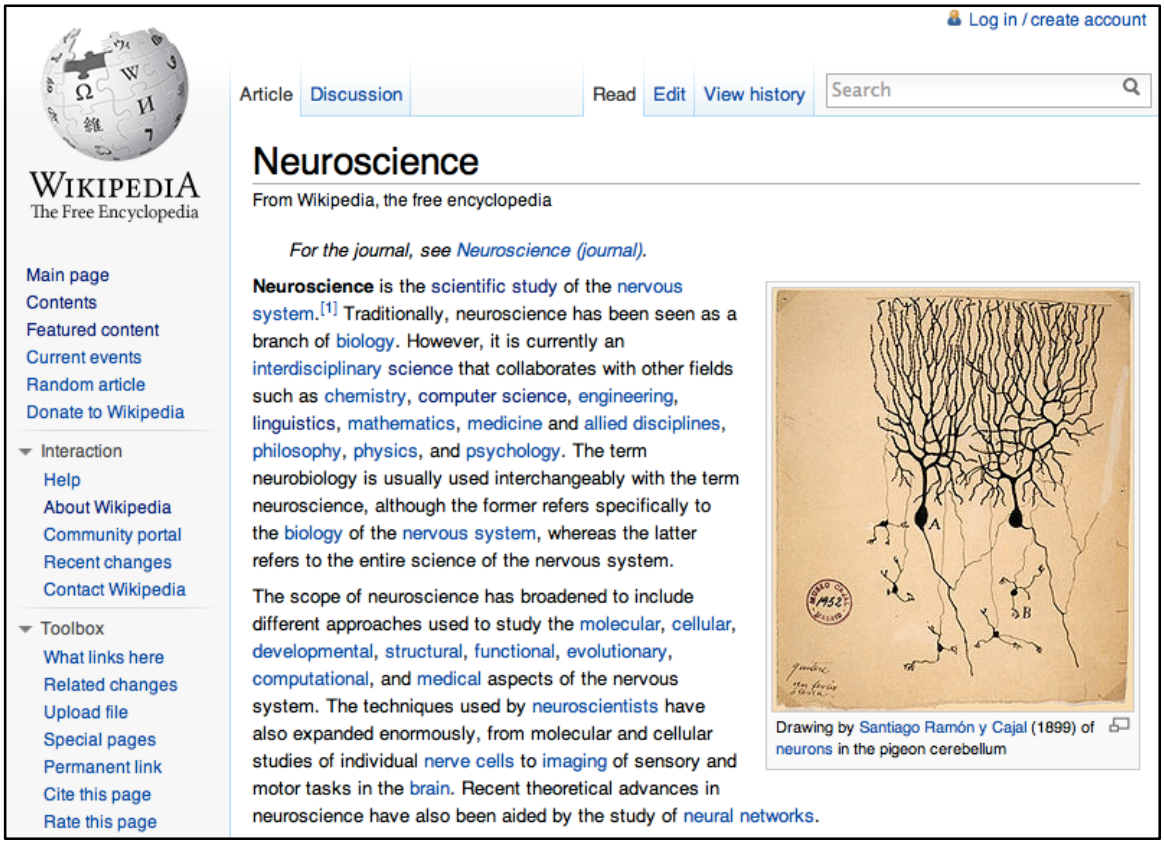


Figure 9. Screen capture of the beginning section of the article on "Neuroscience".

On April 9, 2010 at 18:33, an anonymous user identified by IP address made changes to this paragraph:

> Given the increasing number of scientists who study the nervous system, several prominent neuroscience organizations have been formed to provide a forum to all neuroscientists and educators. For example, the International Brain Research Organization was founded in 1960, the European Brain and Behaviour Society in 1968, and the Society for Neuroscience in 1969.

The modified paragraph included inserted text and read:

> Given the increasing number of scientists who study the nervous system, several prominent neuroscience *its super fun but super confusing.*

The change disrupted a useful section of the article by inserting grammatically poor and generally useless text. This type of change is considered vandalism in that it degrades an article by replacing, disrupting or removing valuable content and replacing it with unhelpful or even blank content. At 18:33, during the same minute the vandalism was recorded, the bot *ClueBot NG* reverted the change back to the previous version.

In another example, on March 10, 2011, the user *Azzurro2882* added additional content to the explanation of the computational neuroscience branch of study noting that the change provided the more common meaning of computational neuroscience (added text is in italics).

> Computational neuroscience is the study of brain function in terms of the information processing properties of the structures that make up the nervous system. *Computational neuroscience can also refer to the use of computer simulations and theoretical models to study the function of the nervous system*.

The added text remained unchanged until at least November 3, 2011, when it was last checked, and was apparently accepted by the community.

It is important to note that this article, like many in the study, is mature in that it was originally begun in September 2001. Much of the recent changes are minor ones that clarify or simplify text, add or update links, revert vandalism or perform other general article maintenance. In order to see major changes to content it is necessary to compare fairly old versions. The original text of the article as it existed September 22, 2001, was a single paragraph of text of 189 words. The article was an automated conversion from *Nupedia*.

> A field of study which deals with the structure, development, function, chemistry, pharmacology and pathology of the central or peripheral nervous system. The biological study of the brain is an interdisciplinary field, which involves many levels of study, from the molecular level through the cellular level (individual

> neurons), the level of relatively small assemblies of neurons like cortical columns, that of larger subsystems like that which subserves visual perception, up to large systems like the cerebral cortex or the cerebellum, and at the highest level the nervous system as a whole. At this highest level the field largely merges with cognitive neuroscience, a discipline first populated mostly by cognitive psychologists, currently becoming a dynamic specialty of its own. Thus, the concern of neuroscience includes such diverse topics as the operation of neurotransmitters, how genes contribute to the embryonic development of the nervous system and to learning, the operation of relatively simpler neural structures of other organisms like marine snails, and the structure and functioning of complex neural circuits in perceiving, remembering, and speaking. Closely related and overlapping fields, besides cognitive neuroscience, include neurology, psychopharmacology, aphasiology, neurolinguistics, and several others.

From the time of its creation until June of 2004 the article underwent a number of minor changes including the addition of hyperlinks to other topics in Wikipedia, external links, some restructuring of the text into several paragraphs and bulleted lists, etc. On June 19, 2004, the user *Sootymangabey* added the "Fields within Neuroscience" section including definitions of the four main areas of study. The most recent version of the article contains over 2,700 words of content addressing history, relationships with medicine, a list of major branches of the field, future directions and a section on education and outreach. In addition, there are 22 citations, a list of 28 books for further reading, 13 external links to various organizations related to neuroscience, five images and a table of contents.

Q2 What is the Profile of a Prolific Contributor to Select Science Articles in Wikipedia?

The second question concerned the profile of prolific contributors to the selected science articles. Profiles were developed by manually viewing user pages of a random sub-sample of 101 users. These user pages provided varying levels of information including 24 pages (23.76%) that were either non-existent, or contained no information.

The remaining 77 pages (76.24%) contained content of some sort. Only 49 users (48.51%) provided enough personal information to clarify their educational background or profession. Of these 49 users, 43 (42.57%) provided information suggesting they had a scientific background (such as *Anlance* who reported having a Ph.D. in physics or *Osborne* who reported a profession of museum curator with a degree in botany). With respect to educational background, five reported holding a bachelor's degree, five were graduate students, one held a masters degree and 11 held a Ph.D. There were also three users reporting they held a medical degree. Finally, four users simply stated they were college students.

Additionally, the top five articles edited were noted for each of the 43 users who had a reported scientific background. Of these, 34 had at least one article in their top five that had a direct relationship to their noted area of expertise. For example, user *CBM*, who listed a background in mathematical logic had the following top five articles: Godel's incompleteness theorems, first order logic, exponentiation, mathematical logic, and computability theory. User *DO11.10*, who claimed a background in immunology, had the following top five articles: poliomyelitis, immune system, Vitamin D, smallpox, and Han van Meegeren (an art forger). User *Iulus Ascanius*, who claimed to have a Ph.D. and a professional scholarly interest in psychometrics and a non-professional scholarly interest in several topics in geography, had the following top five articles: test (student assessment); item response theory; Traverse City, Michigan; Waterton, Wisconsin; and psychometric software. However, not all users followed this pattern. User *Ahoerstemeier*, for example, who claimed to have a diploma in physics and amateur interest in several topics including Thailand, had the following top five articles: index of Thailand related

articles, Thailand, Bangkok, wiki, and Wikipedia. User *Favonian*, who claimed to be a mathematician by education, a software architect by profession and an amateur historian, had the following top five articles: Battle of Hastings, Louis Pasteur, Ali, 2009, and Leif Ericson. The complete list is included in Appendix C.

Additionally, it was noted that some of the 913 users contributed to more than one of the 147 originally selected science articles. Of these contributors, 202 contributed to two or more of the articles and four users contributed to 10 or more. User *Vsmith* contributed to 38 of the articles which was the most.

An interesting issue was noticed in the review of the user pages. There was an undercurrent of frustration among some members. For example, user *Brews ohare,* banned for one year allegedly for being disruptive, left this message on his user page:

> Wikipedia is amazingly successful in producing a variety of articles that, while not authoritative, often contain a lot of interesting information the reader can use to expand their knowledge of a subject. It can be fun to contribute to WP, fun to learn from others, and fun to put together an entertaining and useful article. It also can be very exasperating if the editors contributing to an article you want to work on are not interested in these pursuits, but think of WP as on-line scrimmage, or as a mirror for preening, or as an encyclopedia intended to fit their personal criteria. ("User: Brews ohare," n.d.)

With respect to expert retention, user *Iulus Ascanius*, a self-identified scientist and Ph.D. who apparently withdrew from active participation in Wikipedia, noted on a historical version of his user page, dated July 16, 2008:

> Given the level of dysfunction that has come to prevail on Wikipedia, the most appropriate course for a principled scientist is to withdraw from the project. The bureaucracy should either take corrective steps to fix this situation, or else suffer the eventual loss of huge amounts of valuable talent and volunteered resources. ("User: Iulus_Ascanius," n.d.)

Another user, *JWSchmidt*, pointed out similar concerns with specific mention of the article on Scientology (a controversial religious topic).

> I think that the scientology-related articles provide a good example of one of the serious problems facing Wikipedia. It is instructive to look at the earliest versions of the scientology article, the earliest edits of the original version of the article, and how the scientology-related articles have evolved during the past five years. I would love to obtain some information about how the original version of the Wikipedia scientology article was produced.....I'm guessing it may have originated in nupedia. In any case, it was a fairly scholarly and NPOV [neutral point of view] article and it is sad to see how the article went down hill when it came under the influence of Wikipedians. It is clear that from its earliest moments scientology came under attack by an army of determined opponents. Some open questions to ponder: which is the more interesting phenomenon, scientology or anti-scientology? Will it ever be possible for Wikipedia to produce scholarly and NPOV articles about topics such as scientology or is Wikipedia perpetually doomed to suffer from its editors' biases? (“User: JWSchmidt,” n.d.)

Another user, *Christopher Thomas*, suggested Wikipedians needed to follow one basic rule of collaboration.

> The most important aspect of participating in Wikipedia is being able to work with others. All other qualifications, including expertise or knowledge in a field, are secondary to that one. There will come a time when you disagree with community consensus, and you know that you are right, and it's about something important. The correct thing to do is to respect community consensus anyways. (“User: Christopher Thomas,” n.d.)

This issue is explored further in the discussion section.

Q3 Do Prolific Contributors to Select Science Articles in Wikipedia Contribute to Multiple Articles?

In order to answer the third question, it was necessary to examine the edit history of all prolific contributors to the selected 147 seed articles in order to extract descriptive statistics on their patterns of contribution across Wikipedia. Due to anomalies in the data,

61 editors with fewer than 11 edits to one of the seed articles were excluded from this analysis. They were mostly anonymous IP editors and should not have been included in the original query data set. Furthermore, one editor with over 490,000 edits to over 388,000 articles (nearly two and four times the next closest editor respectively) was excluded from two of the distribution graphs (Figures 10 and 11) due to scaling effects that made the graph difficult to read.

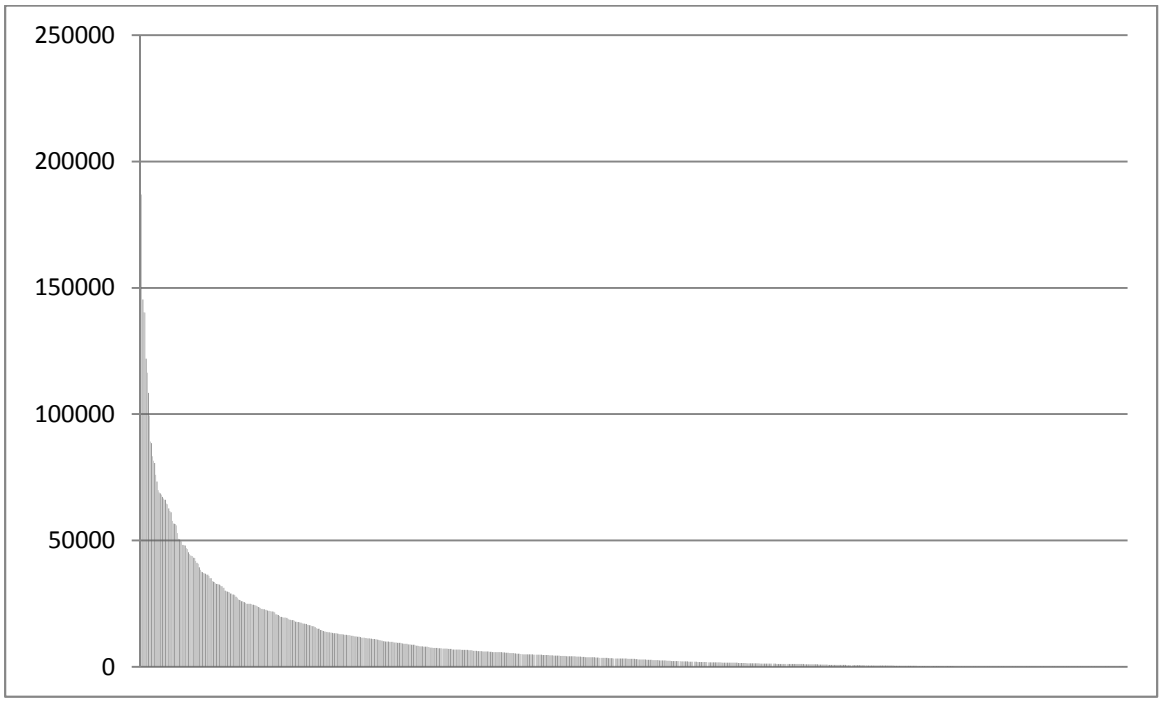
For the 913 contributors identified, the average number of unique articles edited was 4,208 (s.d. = 15,795). The median number of articles edited was 749. The distribution of articles edited was highly positively skewed and leptokurtic (skewness = 16.95, kurtosis = 378.95). The editors contributed an average of 10,572 total edits (s.d. = 25,431) resulting in 2.51 edits per article on average. The median number of total edits was 3,168. This distribution was also highly positively skewed and leptokurtic (skewness = 9.64, kurtosis = 151.62). There was also a significant positive relationship between the number of articles a user edited and their total edit count, $r(913) = 0.93$, $p < .0001$, accounting for 86.5% of the variability. The extremely high number of total articles and total edits suggests a high level of activity. The degree to which these data are impacted by low edit count articles is discussed below.

Table 3.

Summary statistics for editors.

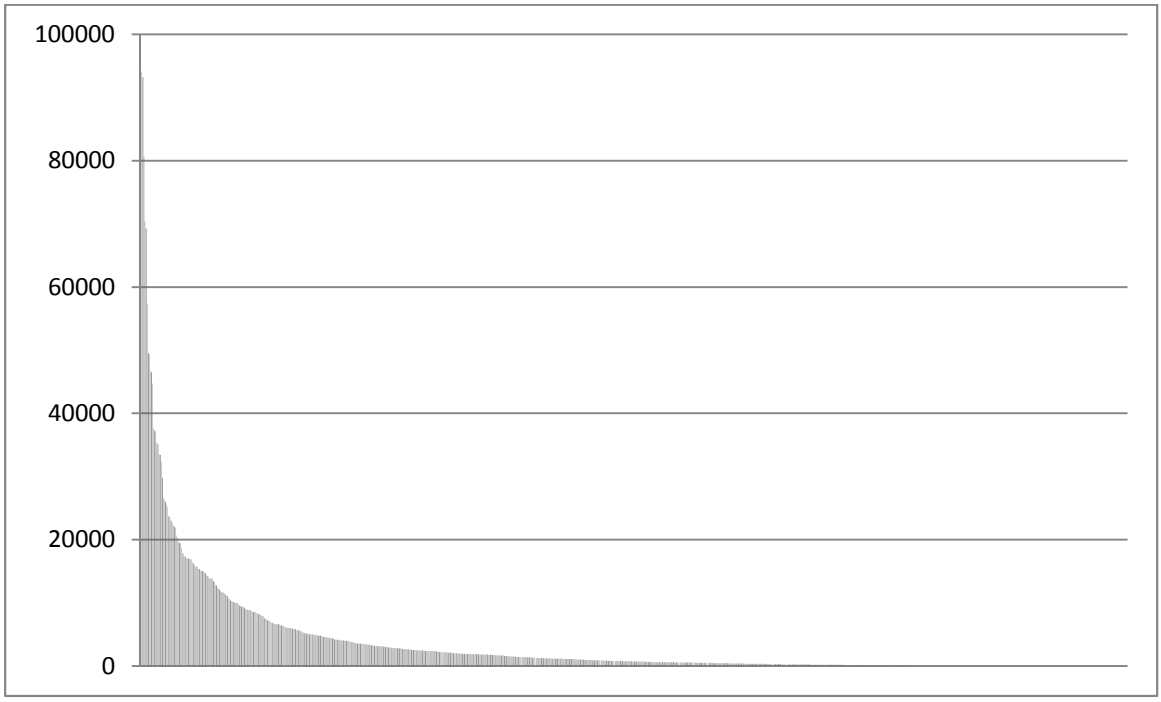
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Articles	913	4,208	15,795	3,841,452	1	388,208
Edits	913	10,572	25,431	9,652,619	12	492,330

The profile of a typical prolific contributor is someone who has contributed to approximately 4,000 unique articles with 10,000 edits (or 2.5 edits per article). Summary data is presented in Table 3.



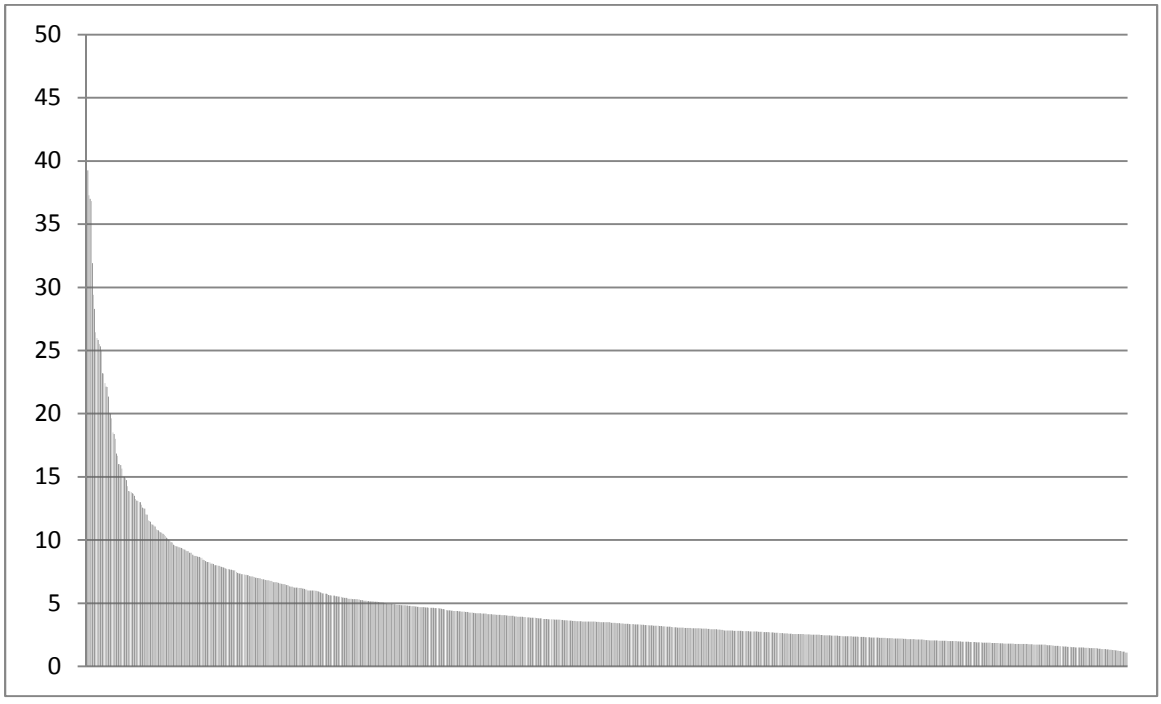
Note. Editors ordered left to right by descending total number of edits.

Figure 10. Distribution of total number of edits by each of the editors.



Note. Editors ordered left to right by descending total number of articles.

Figure 11. Distribution of article counts for each of the editors.



Note. Editors ordered left to right by descending average number of edits per article.

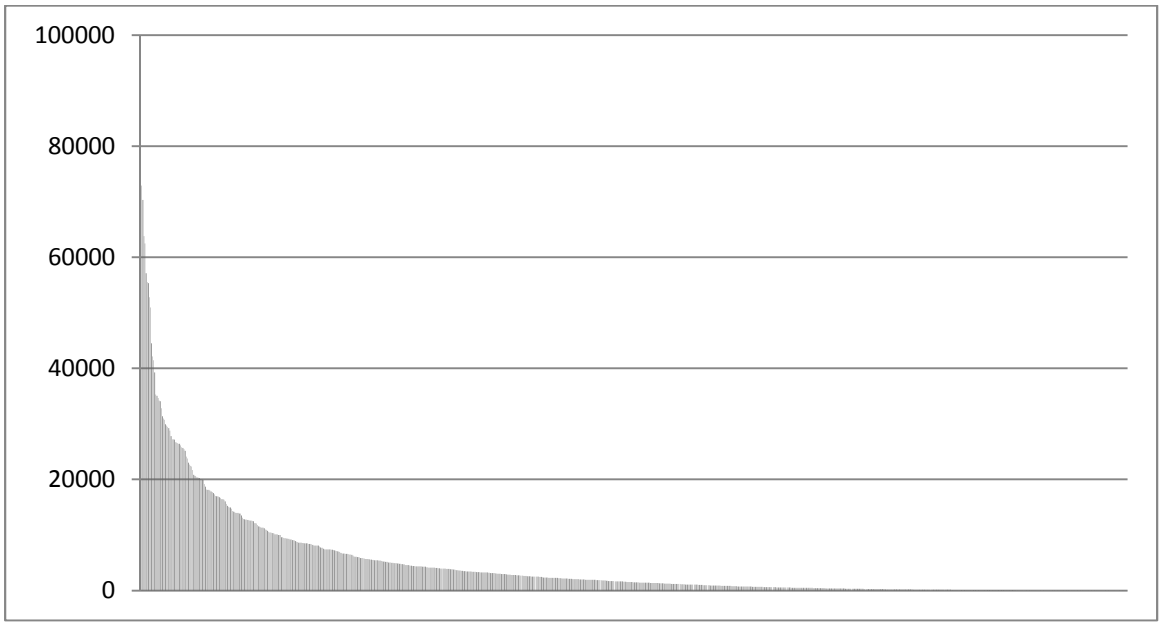
Figure 12. Distribution of average edits per article for each of the editors.

An examination of edit histories showed that editors often had a very large number of articles with very few edits thus unduly biasing those calculated averages. Furthermore, an examination of edit frequencies (Table 5) showed that 70.2% of all articles received only one edit. Another 23.4% had two through five edits. Those with more than five edits made up the remaining 6.4%. In the analysis of accidental collaboration, it was decided to exclude articles with five or fewer edits in order to avoid a large number of trivial hits. Applying that same criteria here had a dramatic effect on the results. When looking only at articles that were edited more than five times, the average number of articles per editor dropped to 268.24 (sd = 532.71) with an average edit count of 4,983.51 (sd = 9,373.16). This resulted in an average of 18.58 edits per article. The very large standard deviations suggest there is still wide variability in the actions of prolific contributors to the selected science articles. There was still a significant positive relationship between the number of articles a user edited and their total edit count, $r(913) = 0.92$, $p < .0001$ accounting for 85% of the variability in edit counts. Note that the removal of low edit count articles resulted in the deletion of 93.63% of the total number of articles edited. In other words, only 6.37% of articles were edited more than five times. However, these articles accounted for 47.09% of the sum total of edits.

Table 4.

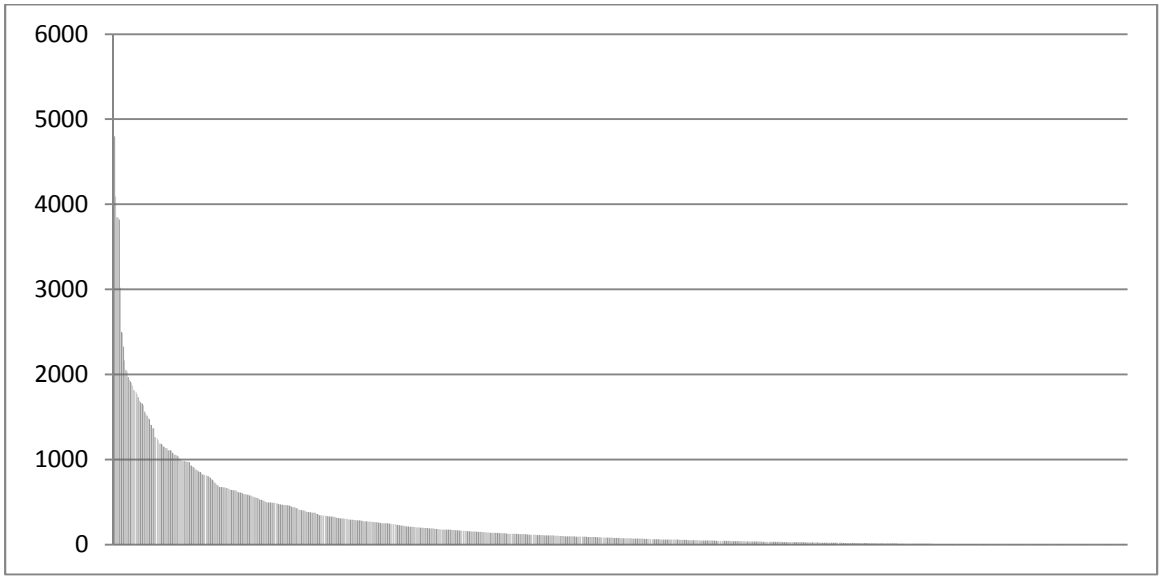
Summary statistics for editors after removing low edit count articles.

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Articles	913	268.24	532.71	244,633	1	6,444
Edits	913	4,984	9,373	4,544,961	11	95,205



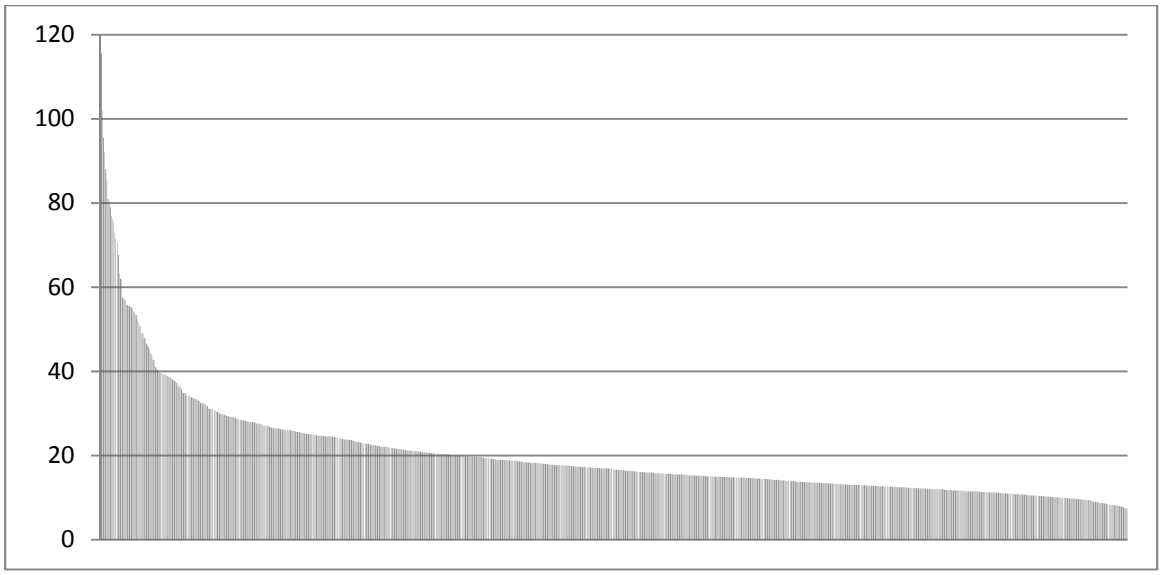
Note. Editors ordered left to right by descending total number of edits.

Figure 13. Distribution of edit counts after removing low edit counts.



Note. Editors ordered left to right by descending total number of articles.

Figure 14. Distribution of article count per editor after removing low article counts.



Note. Editors ordered left to right by descending average number of edits per article.

Figure 15. Distribution of edits per article per user after removing low edit counts.

Table 5.

Frequency of article edit counts for all 913 editors (truncated for easy reading). Edits of 6 or more accounted for a cumulative total percentage of 6.37% of articles.

Edit Count	Frequency	Percent	Cumulative Percent
1	2697817	70.2	70.2
2	524271	13.6	83.9
3	203029	5.3	89.2
4	106321	2.8	91.9
5	65375	1.7	93.6
6	43513	1.1	94.8
7	31360	0.8	95.6
8	23498	0.6	96.2
9	18206	0.5	96.7
10	14799	0.4	97.1

One somewhat typical contributor is *Looie496*. According to the user page, *Looie496* is a neuroscientist specializing in learning and memory with a focus on the hippocampus. Among more recent contributions is a series of 18 edits to the

Consciousness article (“Consciousness,” n.d.) between September 14 and September 21, 2011. Several of the changes were minor such as changing the order of names to a first initial last name format or adding links. One of the more substantial contributions was the addition of a new section titled Anosognosia and added on September 16, 2011. The added content is included below.

> One of the most striking disorders of consciousness goes by the name anosognosia, a Greek-derived term meaning unawareness of disease. This is a condition in which patients are disabled in some way, most commonly as a result of a stroke, but either misunderstand the nature of the problem or deny that there is anything wrong with them. The most frequently occurring form is seen in people who have experienced a stroke damaging the parietal lobe in the right hemisphere of the brain, giving rise to a syndrome known as hemispatial neglect, characterized by an inability to direct action or attention toward objects located to the right with respect to their bodies. Patients with hemispatial neglect are often paralyzed on the right side of the body, but sometimes deny being unable to move. When questioned about the obvious problem, the patient may avoid giving a direct answer, or may give an explanation that doesn't make sense. Patients with hemispatial neglect may also fail to recognize paralyzed parts of their bodies: one frequently mentioned case is of a man who repeatedly tried to throw his own paralyzed right leg out of the bed he was lying in, and when asked what he was doing, complained that somebody had put a dead leg into the bed with him. An even more striking type of anosognosia is Anton–Babinski syndrome, a rarely occurring condition in which patients become blind but claim to be able to see normally, and persist in this claim in spite of all evidence to the contrary.

Although it is a more recent addition, the new text was still part of the article when last visited on November 3, 2011. *Looie496*’s earliest contributions are a series of about 30 edits to articles on Hippocampus and Hippocampus Anatomy. These changes included fixing spelling errors, adding references and images, and rearranging sections for easier reading. Interestingly, one of the earliest edits involved what *Looie496* referred to as fixing “a couple of minor errors.” One of these changes involved information relating to the shape of the hippocampus and its subsequent naming. The new text added on April 15, 2008 read:

> In rodents, where it has been studied most extensively, the hippocampus is shaped something like a banana. In humans, it has a curved and convoluted shape that reminded early anatomists of a seahorse. (“Hippocampus,” n.d.)

The most recent version of this article has been significantly modified since the above addition was made. The content added by *Looie496* no longer exists in its original form, but its influence is still felt. For example, the most recent version from November 3, 2011 includes a section devoted to the history of the name and begins:

> The earliest description of the ridge running along the floor of the temporal horn of the lateral ventricle comes from the Venetian anatomist Julius Caesar Aranzi (1587), who initially likened it to a seahorse, using the Latin: hippocampus… or alternatively to a silkworm. (“Hippocampus,” n.d.)

These types of additions and subsequent changes are illustrative of the overall focus and intent of Wikipedia. Users with varying backgrounds add or change content with the implied intent of improving an article. Their additions over time can morph or be subsequently removed. It would appear that *Looie496* is knowledgeable about topics such as consciousness and the hippocampus and has remained an active contributor to these articles.

Q4 What Types of Articles Cluster Around Select Science Articles Based on Accidental Collaboration and What Conclusions can be Drawn?

Of unique interest to this researcher were the types of articles that would surface when looking at accidental collaborations. What was unexpected, however, was the sheer volume of articles that would be identified. Once the entire list of articles identified by looking at accidental collaborators was combined into one file, sorted and duplicates

removed, over 10,000 unique articles were identified. Instead of attempting to categorize such a large number of articles, a random sub-sample of approximately 10% (12 articles as only 117 of the original 147 seed articles produced examples of accidental collaboration) was again selected. The 12 articles sampled were (listed alphabetically):

Chemistry
Epidemiology
Evolutionary psychology
Forestry
Geography
Geomorphology
Hydrology
Limnology
Quantum mechanics
Social work
Soil science
Zoology

This sample still resulted in over 4,800 articles with accidental collaborators. Articles such as Chemistry and Quantum Mechanics, which produced 2,199 and 3,547 articles respectively, were pared down by looking at the articles with higher numbers of accidental collaborators (in this case four or more which resulted in 280 and 472 articles respectively). A similar process was used with the article Geography using a criteria of three or more accidental collaborators leading to 85 articles. This resulted in a total of 809 unique articles needing to be categorized. Table 6 shows the various categories identified for all 809 articles.

Table 6.

Categories identified for the 809 articles showing accidental collaboration.

Category	**Total**
Art	2
Biography	43
Biography - Scientist	34
Communication	5
Culture	15
Economics	3
Entertainment	5
Geography	145
History	68
Language	2
Literature	6
Mythology	7
Philosophy	4
Religion	34
Science	421
Sociology	12
Sports	2
Wikipedia Navigation Page	1
Grand Total	**809**

Of these, 421 (52.04%) were categorized as relating to a field of science. Additionally, 145 (17.92%) of the articles were categorized as relating to the field of geography, 68 (8.41%) on history, 12 (1.48%) on sociology, and three (0.37%) on economics (which were all considered social and behavioral sciences in the original seed articles). The data also included 34 (4.20%) articles on biographies of scientists which were scientific in nature as articles tended to discuss their contributions to the sciences. Outside of science, other academic articles included 43 additional biographies, 34 on religion, 15 articles on culture, seven on mythology, six on literature, five on

communication, four on philosophy, and two on each of art and language. The remaining 8 articles (0.99%) were on a variety of other topics such as entertainment and sports. As a result, out of the 809 articles identified, 683 (84.43%) were related in some way to a scientific field. In total, 801 (99.01%) of the articles identified were of an academic nature. Summary data is included in Table 7.

Table 7.

Summary of article categories for those identified via the accidental collaboration process.

Category	**# of Articles**	**% of Total**
Biography-Scientists	34	4.20%
Economics	3	0.37%
Geography	145	17.92%
History	68	8.41%
Science	421	52.04%
Sociology	12	1.48%
All Sciences Combined	683	84.43%
Other Academic	118	14.59%
All Science & Academic	801	99.01%
All Others	8	0.99%
Total of all Articles	809	

Individual article summaries are included in Table 8. From the 12 articles in the sub-sample, accidental collaboration was found 6,423 times in 4,849 unique articles. All these examples of accidental collaboration were the result of 99 different users (the 12 articles had 117 unique users as some users appeared in more than one article). With respect to these 12 articles, the average number of articles identified using the accidental collaboration criteria was 535 (s.d. = 1,137) and the average number of accidental collaborators per article was 2.14 (sd = 0.65). These articles had an average edit count of

67.82 (sd = 27.76) from prolific contributors. The fewest number of accidental collaborators was two and this occurred with 4,225 (65.78%) articles including duplicates when they appeared in separate seed articles. The largest number of accidental collaborators was 12 which occurred with only one article (Physics).

Table 8.

Summary data for the 12 articles in the random sub-sample

Article Title	# of Original Contributors	# of Articles Showing AC	Avg # of Contributors	Std Dev	Avg # of Edits	Std Dev of Edits
Chemistry	30	2,199	2.56	1.00	62.77	68.35
Epidemiology	5	12	2.00	0.00	94.33	99.75
Evolutionary psychology	15	35	2.06	0.24	49.94	44.91
Forestry	7	12	2.08	0.29	68.92	65.26
Geography	10	481	2.20	0.45	82.23	119.67
Geomorphology	5	104	2.11	0.31	81.84	67.97
Hydrology	3	16	2.00	0.00	85.94	81.90
Limnology	2	1	2.00	-	60.00	-
Quantum mechanics	33	3,547	2.57	0.90	42.75	45.35
Social work	10	2	2.00	0.00	38.00	35.36
Soil science	2	2	2.00	0.00	91.50	10.61
Zoology	5	12	2.00	0.00	47.75	31.38
Total	**127**	**6,423**				
Average	**10.58**	**535.25**	**2.13**	**-**	**67.16**	**-**
Standard Deviation	**10.49**	**1137.10**	**0.21**	**-**	**19.82**	**-**

Because data were readily available, it was also of interest to look at accidental collaboration on a broader scale. By combining the Wikipedia edit histories (both excluding articles with five or fewer edits and including all edits) of the 913 users identified from the original seed science articles, a master file of approximately 244,000

and 3.8 million articles respectively was created and examples of accidental collaboration were found using a frequency feature in a popular statistical software package. Table 9 shows the top 50 articles identified in each case.

Table 9.

List of the top 50 articles contributed to by looking at all 913 editors both excluding five or fewer edits and including all edits.

All Editors Combined (excluding five or fewer edits)			All Editors Combined (including all edits)	
Article Title	**Count of Editors**		**Article Title**	**Count of Editors**
Evolution	80		Evolution	190
Wikipedia	72		Physics	180
Physics	71		Albert_Einstein	175
Pseudoscience	66		Wikipedia	173
Albert_Einstein	66		Science	171
Mathematics	66		Global_warming	165
Biology	60		George_W._Bush	155
George_W._Bush	58		United_States	151
Psychology	58		Human	149
Jesus	57		Jesus	147
Creationism	57		Biology	145
Human	54		Adolf_Hitler	140
Global_warming	53		Mathematics	139
Science	53		Earth	137
History	52		Intelligent_design	136
Intelligent_design	51		Psychology	136
Energy	50		Atheism	135
Earth	49		Creationism	126
Atheism	48		Christianity	124
Economics	47		Pseudoscience	124
United_States	47		Scientific_method	123
Quantum_mechanics	47		Big_Bang	122
Scientific_method	46		Quantum_mechanics	122
Chemistry	46		Energy	119
Adolf_Hitler	45		Cat	118
Charles_Darwin	44		Water	118
Astronomy	44		Barack_Obama	117
Ecology	43		Philosophy	117
Medicine	42		Death	116
Racism	41		History	116

All Editors Combined (excluding five or fewer edits)			All Editors Combined (including all edits)	
Article Title	**Count of Editors**		**Article Title**	**Count of Editors**
Water	41		Astronomy	114
Sun	41		Charles_Darwin	114
Homeopathy	40		World_War_II	114
God	40		Ecology	113
Christianity	40		God	113
Isaac_Newton	39		Economics	112
The_Holocaust	38		World_War_I	112
Muhammad	38		Universe	111
Hurricane_Katrina	38		Sun	110
Death	38		France	110
Black_hole	38		Isaac_Newton	110
List_of_topics_characterized_as_pseudoscience	37		London	110
Big_Bang	37		Moon	109
George_Washington	37		Chemistry	108
Democracy	37		Speed_of_light	108
Islam	36		Astrology	107
Tsunami	36		Medicine	107
Communism	36		Michael_Jackson	107
Cat	35		Slavery	107
London	34		Black_hole	106

It is interesting to note that while the number of contributors naturally increased when including all low edit counts, there were only minor changes in the articles found in the two lists. This suggests that even when contributors only added a few edits to an article, they were still adding to the same articles as those doing more substantial work. In other words, all these contributors who were identified via being a prolific contributor to one of the original seed articles, tended to contribute to a variety of similar articles. Note that

many of these articles are scientific in nature and those that are not are often still major topics of academic concern.

Q5 What do Network Maps of Article Clusters Based on Accidental Collaboration Say About the Legitimacy of the Content?

Each of the articles which were identified using the accidental collaboration criteria (for the 12 articles discussed above) are examined individually with respect to their relationship to a generally accepted field of science, a non-scientific but generally academic field (such as what one might expect to find in a traditional encyclopedia), or outside either of these two classifications. Articles relating to entertainment, pop culture, sports, etc. may be found in some traditional encyclopedias but are here considered non-academic in order to differentiate between fields of study that would generally require a level of education in order to be considered knowledgeable and other topics which one might gain knowledge through passive actions such as watching a movie or being a fan of a sports team. Table 10 shows a summary of the findings for each of the 12 seed articles with the complete data tables included in Appendix D. The data shows the percentage of articles matching a science category, a generally academic category or other category. Three articles (Hydrology, Limnology, and Soil Science) showed 100% science categories. Two others (Geography and Geomorphology) had a very high percentage (97.65% and 99.04% respectively) of science related articles which was interesting given the larger number of articles (85 and 104 respectively) included. Social Work had the lowest percentage of science articles (50%) but this included just one out of a total of two articles. Epidemiology showed the second lowest percentage of science category articles

(66.67%) which also included a low number (12) of articles overall. Overall, articles in the sub-sample showed an average of 81.98% (sd = 17.37) of science related categories in articles demonstrating accidental collaboration. A further 15.85% articles (sd = 16.41) on average were non-scientific but still generally academic in nature such as those relating to fields that typically require an educational background. Overall, an average of 97.83% of the articles found via accidental collaboration in the sub-sample were scientific or academic in nature suggesting that some of the prolific editors of the seed articles also tended to be prolific contributors to other articles of a scientific or academic nature. This also suggests these contributors have a level of academic knowledge consistent with what would be expected of knowledgeable authors contributing to an encyclopedia.

Three of these articles (Chemistry and Evolutionary Psychology and Hydrology) are examined in more detail. The Chemistry article had 30 prolific contributors. They contributed to 207,044 articles (including duplicates) or 17,400 (including duplicates) excluding articles with five or fewer edits. A total of 2,199 articles were edited by two or more of these 30 contributors (which is the definition of accidental collaboration as used in this study). Of these, 280 had four or more accidental collaborators and were manually categorized into several areas. Science articles accounted for 81.43% (228 articles) of the total. Many of these articles dealt with topics relating to chemical elements, planets, chemical compounds, and popular topics in science such as global warming. The remaining non-scientific articles fell into general academic categories such history, religion and culture. In total, all 280 articles categorized fell into a generally academic category.

Table 10.

Summary of articles showing percentage matching science, academic and non-academic categories.

Article Title	%Science	%Non-Science Academic	%Non-Academic
Chemistry	81.43	18.57	0
Epidemiology	66.67	25	8.33
Evolutionary Psychology	68.57	31.43	0
Forestry	83.33	8.33	8.33
Geography	97.65	2.35	0
Geomorphology	99.04	0.96	0
Hydrology	100	0	0
Limnology	100	0	0
Quantum Mechanics	70.34	28.6	1.06
Social Work	50	50	0
Soil Science	100	0	0
Zoology	66.67	25	8.33

Evolutionary Psychology had 15 prolific contributors who contributed to 23,067 articles (including duplicates) or 2,728 (including duplicates) excluding articles with five or fewer edits. A total of 35 of these articles were contributed to by two or more of these 15 contributors. Science articles accounted for 68.57% (24 articles) of the total. These articles were on a variety of topics such as evolution, evolution controversy and a biography of Charles Darwin, medical conditions such as autism and Asperger syndrome, and other issues relevant to psychology such as gender identity and biographies of psychologist Steven Pinker and evolutionary biologist Stephen Jay Gould. The remaining non-scientific articles were primarily related to religion and culture. All 35 articles were considered to be of a generally academic nature.

At the lower end of the spectrum, the Hydrology article had only three prolific contributors who nevertheless contributed to 16,052 articles (including duplicates) or

2,868 (including duplicates) excluding articles with five or fewer edits. This larger number of articles for only three contributors is likely due to the fact that the user *Vsmith*, previously identified as a very prolific user, was among them. A total of 16 articles were identified as having two accidental contributors (none existed with all three contributors). All 16 (100%) of these articles fell into the scientific categories of science and geography. Articles that were geographic in nature included flood and floodplain as well as river, surface runoff and physical geography. Articles dealing with more traditional scientific topics included climate change, ecosystem, evaporation and water cycle. As the study of hydrology deals with water and its movement and distribution on the Earth and interaction with the environment it is interesting that all articles were topically related in some way to the hydrology seed article.

The majority of the 12 articles followed similar patterns of relationship between the topic of the seed article and those identified via the accidental collaboration process. Forestry, for example, included a number of articles on trees or tree related topics (see Appendix D for a complete list of articles). Only one article (a Wikipedia navigation page) fell outside the general topic of trees. Some of the other 12 seed articles, such as epidemiology, quantum mechanics, and zoology, also contained a small percentage of articles found via the accidental collaboration process that fell outside scientific or generally academic areas but these comprised a very small percentage (0.99%) of the total number of articles. These included articles on space hopper (a toy), a Wikipedia navigation page, sports and entertainment.

CHAPTER V

DISCUSSION

The results of this study led to a number of major findings: selected science articles tend to have a small number of prolific contributors, approximately 40% of these prolific contributors have a self-reported background in a scientific field, prolific contributors tend to be highly active across Wikipedia and contribute to numerous articles, with respect to accidental collaboration these contributors contribute to a high percentage of related science articles, and these articles tend to cluster topically around a seed article. Each of these findings is further discussed below as it relates to the various research questions addressed in this study.

Major Findings

A review of the literature showed that patterns of contributions to articles have not been fully explored with respect to Wikipedia. Studies to date (Arazy et al., 2010; Jarkko et al., 2010) have tended to focus on collaborative knowledge building in controlled environments, such as classrooms, where contributors knew each other and were contributing to contrived projects. There appears to have been little research to date focusing on contributions to a project such as Wikipedia where contributors do not know each other, are not from a particular group (such as students in a class), and are

contributing to a form of social knowledge. Contrary to popular belief and Wikipedia's own tagline, "anyone can edit," the results of this study showed that articles tended to have a small number of prolific contributors (9.66 on average). While an analysis of a sub-selection of articles did show that articles have around 700 unique contributors on average the vast majority have contributed 10 or fewer edits. This suggests that, when excluding minor contributors and vandalism, the structure, content and overall development of science articles is carried out by a small group of individuals who, based on the results to the other research questions, appear to have a background in a related field and are qualified to write on such topics. The extent to which this may be true across Wikipedia is unknown as this study focused only on a selection of science articles. While further research is needed to explore this issue with respect to other fields, it would not be surprising to discover that other academic-oriented articles experience a similar pattern of contribution. It is interesting to note, however, that more recent research (Arazy et al., 2011) that sampled from a broader selection of articles found a much small number of editors and edits, averaging 49.2 (s.d. = 70.4) and 90.9 (s.d. = 125) respectively. This may be due, in part, to the relative immaturity of the articles (they were from 2006 and roughly two years old compared to 9.5 years in the current study) or may be reflective of differences between science articles and those from divergent categories.

This research also focused on developing a better understanding of the major contributors to the selected articles and the extent to which they are qualified to contribute scientific knowledge to an encyclopedia. This process relied upon anonymous user-reported information and is therefore not directly verifiable. However, the examination of articles found via the accidental collaboration process does offer support

for these findings and suggests that Wikipedia contributors who chose to disclose personal information were honest with respect to their educational and professional backgrounds.

Although not all users analyzed chose to disclose personal information, those that did were found to have scientific backgrounds that were often closely related to topics to which they provided major contributions (based on edit count). Previous research that studied article accuracy and credibility (Arazy et al., 2011; Chesney, 2006; Giles, 2005; Magnus, 2006; Rajagopalan et al., 2010; Rector, 2008; Rosenzweig, 2006) found that Wikipedia content was generally accurate, though sometimes not to the same degree as more traditional reference materials, and the findings of this study offer at least a partial explanation. The majority of an article may be written by a small number of individuals, and those contributors are likely knowledgeable about the topic to which they contribute; this helps explain the general level of quality observed in past studies.

There are, of course, examples of users claiming false credentials (Read, 2007) and it is likely that many still exist that have not been exposed. Wikipedia takes the position that the verifiability of content is more important than a user's credentials. In other words, if something is verifiably accurate then it does not matter who wrote the content. Wikipedia, however, has struggled with this issue and Wales himself has proposed that users have an option to verify their credentials noting that it would help to strengthen the culture of mutual trust and discourage false claims (Wales, n.d.). This proposal was not accepted and Wikipedia currently does not have a credential policy ("Wikipedia: There is no credential policy," n.d.). Results of the current study showed that those who did claim to have a background in a particular area of science also tended

to contribute heavily in that area. It is likely that those who contribute heavily to a certain field or topic eventually choose to build a profile of credentials to support their work. It is not necessarily easy to identify those who have false credentials if the edits they contribute are useful and verifiable. However, it is reasonable to assume that those who continually add questionable content would eventually be called out or banned. Contributors who manage to build a substantial body of contributions are likely knowledgeable and would benefit little from claiming false credentials (at least within the framework of Wikipedia).

This research also helps to further understand the actions of prolific editors. This study began by looking at contributors to a science article with more than 10 edits, but having contributed more than 10 edits to a single article does not tell us much about the contributors. It was therefore important to know the extent to which they contributed across Wikipedia to build profiles and, by extension, understand their level of expertise. Results showed that contributors were on average very prolific having contributed more than 10,000 edits to more than 4,000 articles. These results were highly skewed. Of the 913 contributors, 53 had contributed to 10 or fewer articles and 714 to more than 100. At the upper end, 89 had contributed to more than 10,000 articles. However, it was also discovered that most of these articles had received a very low number of edits. When ignoring articles with five or fewer edits, the averages dropped to 268 articles and 4,900 edits. While articles receiving six or more edits accounted for just over 6% of the total number of articles, they are also responsible for over 47% of the edits. This is still a great deal of effort on the part of these contributors suggesting that in addition to their

contributions to the selected science articles they also contributed to a vast array of additional articles.

It is interesting to note that according to Wikipedia statistics, users with more than 5,935 edits (as of October 23, 2011) are among the 8,000 top contributors ("Wikipedia: List of Wikipedians by number of edits," n.d.). In other words, only a very small percentage (approximately 0.05%) of the 15 million registered users ("Wikipedia: Wikipedians," n.d.) have contributed more than 6,000 edits. Although data for this study was collected in May, 2011, there were 324 contributors at that time with more than 6,000 edits suggesting that approximately one-third of the contributors identified in this study were among the top contributors across Wikipedia. Taken together, these results suggest that while Wikipedia is an open platform and invites anyone to edit, only a relatively small number of people contribute substantially.

One aspect of contributors that was not examined in this study was their tendency toward an administrative orientation or a content-level orientation. Other researchers (Arazy et al., 2011) have suggested that contributors who spread their activity across Wikipedia have an administration orientation while those who tend to focus on a particular topic or topics are content oriented. It is likely that the prolific contributors identified in the current study were from a mix of administrative and content orientations.

What remains unknown is a clear picture of who really edits Wikipedia. It is clear from this study that there are a small number of prolific contributors with respect to science articles and these individuals tend to contribute broadly and extensively across Wikipedia. To date, there appear to have been few studies focusing on the knowledge level of Wikipedia contributors. The most comprehensive, perhaps, is a survey conducted

by the United Nations University (Glott, Schmidt, & Ghosh, 2010) that collected responses from 54,000 Wikipedia contributors and found that with respect to scientific fields 70-90% of the contributors (depending on broad themes defined in the study) had self-reported expertise. Arazy et al. (2011) also examined interactions of contributors and found a high level of cognitive diversity among contributors (i.e. there was a low degree of overlap between contributors across Wikipedia as a whole). This is perhaps due to the method used to measure diversity which looked for what was essentially accidental collaboration, though they did not use the term, across all articles edited by the group and apparently including contributors with low edit counts. Korfiatis et al. (2006) attempted to measure relationships between contributors suggesting that higher levels of interconnected contributors (similar to low cognitive diversity and accidental collaboration) would equate to greater authority. The current study used a modified approach that excluded low edit count contributors (only focusing on those with more than 10 edits) and found a different degree of overlap which could be construed as a lower level of cognitive diversity or a higher degree of interconnection. Theories to date are competing (is high cognitive diversity or high interconnection more indicative of authority and legitimacy?) and more research is needed with respect to how to measure cognitive diversity and similar concepts in Wikipedia and other forms of socially-constructed knowledge and its impact on article quality and overall legitimacy.

The current study also sought to uncover the types of articles that would cluster around the selected science articles. It was thought that the accidental collaboration process would have a tendency to show a higher percentage of articles with a scientific focus based on the assumption that prolific contributors to science articles would be

likely to contribute to other science articles. Results of this study supported this assumption showing nearly 85% of articles categorized relating to a scientific field. The implications of these findings are interesting and informative to those wishing to use Wikipedia as a reliable source of encyclopedic content. While there have been previous studies that sought to measure the quality and reliability of Wikipedia content (Arazy et al., 2011; Chesney, 2006; Giles, 2005; Magnus, 2006; Rajagopalan et al., 2010; Rector, 2008; Rosenzweig, 2006), these studies focused on expert review of individual articles, and while they generally supported the overall accuracy of the articles, they do not help us to understand the nature of the contributions or the contributors. To date, there does not appear to have been studies that focused on examining the collaborative efforts of contributors. As such, this study adds a unique perspective to the literature regarding the overall legitimacy of Wikipedia content. Although this study made no attempt to directly measure the quality of articles, it is likely that studies which have found articles to be accurate and reliable did so because at least some of the major contributors to those articles are knowledgeable about the topic helping to ensure that articles are subjected to an informal type of peer review.

Perhaps the most telling results of this study relate to topical relationships between seed articles and those which were identified via the accidental collaboration process. When looking at the articles identified via their relation to the originally sampled article, some interesting patterns emerged. The article on Forestry, for example, contained numerous articles identified via accidental collaboration that were related to forest science, forestry practices, trees, and universities offering degree programs in forestry. Conversely, the article on Chemistry produced a huge number of additional

articles on chemical compounds, elements, chemical processes, and atomic and sub-atomic particles and very few articles on unrelated topics. Many of these unrelated topics were still loosely connected to chemistry such as those on plants and foods which contain various chemical compounds.

This type of specialization was observed in most of the selected articles and was, in fact, a rather interesting outcome of this study. Given the large number of articles in Wikipedia and the rather extensive edit history for many of the contributors, one would expect to see a great deal of randomness in the type of articles to which a small number of individuals would tend to contribute if they appeared at all. It is particularly interesting that not only did rather ordered relationships appear, but that they also tended to show specialized relationships to the seed article (such as Chemistry, Evolutionary Psychology, Forestry, and Hydrology discussed earlier). By way of comparison, the article on Quantum Mechanics produced an extremely large number of articles (3,547) with two or more accidental collaborators. This is not surprising given the large number of prolific contributors (33). It is also not surprising that many of the articles were from a wide range of largely varied topics (waffle, Victorian era, unicorn, tennis, Starbucks, muffin, Malcolm X, Google, Bigfoot, William Shakespeare, taco, and Star Trek to name a few with two or three accidental collaborators which were not among those categorized). However, when looking at the articles with the greatest number of accidental collaborators this randomness tended to disappear. In this case, the top 10 articles based on the number of accidental collaborators where: physics, Albert Einstein, black hole, time, biology, calculus, energy, golden ratio, introduction to quantum mechanics, and mathematics. All of these articles have a fairly direct and obvious relationship to quantum

mechanics in much the same way the Forestry article showed relationships with tree related topics.

These rather startling relationships suggest that prolific contributors to a selected science article may also have a tendency to contribute to many related topics. This would be a natural expectation of individuals with a specialized education: zoologists writing on various animals, botanists writing on plants, and geologists writing about rocks. Given that such a tendency was observed in the articles that were examined more closely, it would be valuable to know the extent to which these relationships are found throughout Wikipedia. This study does not allow for such generalizations to be made, but does provide encouragement regarding the possibility. Further research in this area would be enlightening.

The degree to which the process of accidental collaboration can be practically used to profile articles is uncertain. The process relied on looking at the duplication of effort among a group of prolific editors and ignoring all their other contributions except those showing duplication. While low edit count articles were excluded in order to prevent a large number of potentially trivial hits, the process also did not put any weight on articles with high edit counts. An alternative approach that was only minimally explored here was to look at the types of articles to which each of these prolific editors contributed their greatest amount of effort. It is reasonable to assume that individuals will contribute a greater proportion of effort to articles related to their education and professional backgrounds. The results of this study, to the extent that this was explored, tended to support this assumption. Likewise, the types of articles (and by extension the effort of the contributors) identified via accidental collaboration also supported a strong

tendency toward scientific content. However, if one wanted to explore individual efforts or the collaborative actions of a group, the edit data needed are currently not accessible to users for making such decisions. Studies such as this one may help us to understand the degree to which we can generalize about Wikipedia content and possibly about other forms of socially-constructed knowledge, but this type of content is in continual flux. Ultimately, it will be up to individual users to make decisions about the legitimacy of such content. Providers of this type of content, such as Wikipedia, may wish to explore opportunities to make some form of aggregated edit data available to users.

What remains to be explored is the extent to which patterns discovered here would repeat in other content areas. If it were observed that the prolific contributors to history articles tended to contributed broadly to history and contributors to literary topics tended to contribute broadly to other literary topics, and so on, then there would be added support for using an examination of collaborative effort as a proxy for contributor knowledge. It is also possible that certain disciplines, such as science, tend to exhibit these patterns more than others. Further research is needed to explore this.

Practical Implications

K-12 and Higher Education

Recently, Maehre (2009) has argued against the trend toward banning Wikipedia as a resource, suggesting that it should be considered a useful resource for students in higher education courses – particularly entry-level courses. His arguments largely center on the pedagogical implications of focusing on the research process instead of the ultimate product. He suggests that,

> within an exploration of the dichotomy of student as learner vs. student as producer, is that 1) instruction of information literacy must be done holistically, encompassing a protracted exploration of a large set of concepts throughout a semester, that 2) instructors of *all* introductory courses have a responsibility for taking part in this, and that 3) not only would this be defined as including an open-door policy for a wide range of sources, with students being responsible for finding the best content within them, but that this approach, by removing barriers separating "good" from "bad" sources, is a particularly valuable *tool* for teaching the information literacy skills of evaluation and critical thinking, which are, of course, at the core of higher learning. (p. 231)

Information literacy and critical thinking skills are also at the core of k-12 education as evidenced, at least in part, by recent national technology and information literacy standards (International Society for Technology in education, 2007; American Association of School Librarians, 2009). As such, teachers at all levels may wish to reevaluate their stance on Wikipedia.

Educators who are still concerned about student use of Wikipedia were likely influenced by early reports of inaccuracies, such as a widely reported false biography (Helm, 2005; Seigenthaler, 2005; Survey, 2006), and have not tracked its development over the years. In light of the growing call for greater integration of 21st century skills and information literacy instruction, it is likely time for a new dialog on what constitutes research at the various grade levels with a particular focus on student inquiry and the research process.

The results of the current study offer some reassurance that students will not be inundated with low quality, inaccurate information if they elect to use Wikipedia. However, the nature of this content is that it is in continual flux and students cannot become complacent but must continually reevaluate content. This provides teachers an opportunity for a greater focus on the research and information collection process and the

development of information literacy and 21st century skills; “a centerpiece around which to teach searching and critical reading skills, as well as evaluation of a resource’s content” (Bennington, 2008, p. 47). This approach is no longer simply theoretical. Some educators (Harouni, 2009; Patch, 2010; Pennell, 2007) have begun to tap Wikipedia’s potential as a tool for teaching critical literacy and are actively developing and delivering lessons build around Wikipedia. Patch (2010) argues that lessons involving Wikipedia help “make students smarter consumers of online information and more responsible researchers” (p. 281). Selber (2004) goes beyond Wikipedia and suggests that in order to teach critical literacy we must encourage students to examine not just websites but also how various technologies are used to persuade and control us.

For the foreseeable future, Wikipedia will likely remain a popular resource for information seekers. Rather than discourage its use, teachers and librarians may wish to consider ways in which students could use the built-in tools within Wikipedia to more deeply explore the authorship of articles and use such data to help them make decisions about the likely legitimacy of content. No tool currently exists to allow for an easy examination of a contributor’s edit history, but article histories and discussion pages, as well as user pages, offer students valuable insight into how knowledge is collaboratively constructed and the various debates and decisions which affect how that knowledge is presented. As discovered in this study, more extensive tools that provide greater access to contributors’ top articles or the set of articles that link to a current article via accidental collaboration could be very useful to users of Wikipedia. Those who write and maintain the various tools on the *toolserver.org* website might wish to considered developing one.

A common criticism of Wikipedia is that “anyone can edit it” with the implication being that everyone does, including people without the proper knowledge to do so. This study tended to show that science articles were majorly contributed to and reviewed by a fairly small group of individuals who had a tendency to write on other related topics suggesting probable backgrounds in the related fields. This should give us some reassurance that some articles in Wikipedia are written, edited and reviewed by at least a percentage of people likely qualified to do so.

Research Implications

The process of identifying examples of accidental collaboration was technical and relied on access to the SQL database and would not be an easy or practical approach for average users. Wikipedia is exploring features that can help users to evaluate the legitimacy of content. It may be possible in the future for a similar feature to be added. In the meantime, it may be useful for sites like *toolserver.org* to host tools that allow users to see, for example, the top 10 articles to which a given user has contributed. Used in conjunction with the *contributors* tool, a user could get a pretty good sense of the probable background of the major contributors to an article.

During the course of this study, Wikipedia began to test a new feature called the article feedback tool which, according to Wikipedia, is an “experimental feature that allows any readers of an article (whether they're editors or not) to quickly and easily assess the sourcing, completeness, neutrality, and readability of a Wikipedia article on a five-point scale” (“Wikipedia: Article Feedback Tool,” n.d.). This assessment, however, is based on the input of readers of the page. While an interesting feature, it is unlikely to

help uses assess the legitimacy of content. People tend to read an article in order to get information on a topic about which they know very little. While they may rate a page as easy to read, individuals with limited knowledge are not necessarily qualified to rate an article as complete or neutral though more astute readers may be able to discern these elements. Nevertheless, such ratings do little to help readers evaluate whether or not an article is written by knowledgeable people and contains largely accurate information. In fact, less astute readers may be inclined to equate high marks on the article feedback to high legitimacy. Wikipedia itself does not seem to view this as a tool to rate authority but a way to engage readers and provide information to editors to help them improve articles.

Contributors

Wikipedia is a treasure-trove of data related to online collaboration and socially constructed content that while relatively easy to access, though not directly, it is not readily useable to consumers. One unexpected outcome of this study was a greater understanding of the shear size of Wikipedia and the amount of effort that has gone into its creation and its continued development. This study looked at a very small sample of Wikipedia articles and at a very small percentage of contributors to those articles. This nevertheless resulted in nearly a thousand unique contributors with combined contributions to over 3.8 million articles and over 9.6 million edits. Unfortunately, social aspects of Wikipedia actually encourage elevating ones edit count. Contributors can, for example, earn “stars” that they can post on their user page with different stars available for different actions or number of edits. Although immaterial and of little if any economic value, these rewards are a form of social capital and likely equate to a level of respect

among frequent contributors (Okoli & Oh, 2007). In fact, Huysman (2004) has suggested that a high degree of social capital among members is needed if online networks are to be effective. However, the desire to acquire this social capital most likely also contributes to trivial editing in order to boost one's edit count. The previously identified user *Vsmith*, for example, has over 80,000 edits with the vast majority of those edits spread among 12,000 different articles. However, the top 50 articles, which account for less than 0.4% of his total article count, account for over 10% of total effort (based on edit count). It is also likely that those changes to frequently edited articles are more substantial than those to the thousands of single edit articles. In other words, while *Vsmith* is a highly prolific contributor and undoubtedly a useful member of the community, he has invested heavily in a relatively small number of articles and potentially trivial contributions to a very large number of articles. Such trivial edits are still useful in the overall development of articles, but it is reasonable to conclude that articles to which a user contributes most heavily are those in which they take a greater interest and therefore would likely be adding significant content as well as critically reviewing the contributions of others. According to his user page, *Vsmith* is a geologist with an master's degree and a former high school science teacher with an interest in "almost anything scientific" ("User: Vsmith," n.d.). This undoubtedly accounts for his large number of contributions to articles such as acid rain, mineral, tropical rainforest, weathering, erosion, volcano, plate tectonics, global warming, deforestation, and water pollution (the top 10 articles in his list of contributions). Such patterns were observed repeatedly in the contributors that were studied in depth.

The focus on accidental collaborators was an attempt to pare down the very large numbers of articles edited by a selection of users into a more manageable number while also attempting to find those articles in which users were more likely than not to have a vested interest. The results of this study showed a very high relationship between a seed article and the selection of additional articles, found via the accidental collaboration process, also relating to scientific topics, and that these articles had a tendency to be more closely related to the seed article. Due to the large amounts of data collected in this study, it was not practical to attempt to categorize all articles found via the accidental collaboration process. However, the results suggest patterns but further research is needed to determine if these patterns are found in a larger number of articles and across other fields. We might expect, for example, that historical articles would tend to be closely related to other history articles. The extent to which this may be true is currently unknown but this study does suggest a possible approach to addressing these questions in numerous other fields.

The results of this study also provide additional support for earlier findings regarding the credibility or legitimacy of Wikipedia content. While limited in scope to a small number of science articles, the results do provide some evidence that prolific contributors to these articles are likely knowledgeable about these topics. One of the major concerns regarding Wikipedia is that anyone can contribute and that these contributors may not be qualified to write encyclopedic articles on complex topics. These results suggest that this may not be an accurate representation of the contributors to science articles. Again, further research is needed to determine if these findings can generalize to a larger percentage of Wikipedia content.

With respect to further studies of Wikipedia content using the process of accidental collaboration or similar approaches, it is worth considering what effect research with a more editor-centric focus, or tools that might be developed that focus more attention of the actions of contributors, could have on the motivations of editors. If future researchers or visitors to Wikipedia were to place more importance on the types of articles to which editors most heavily focus, would those editors attempt to "game the system" by adding edits, even more trivial ones, to articles that would be seen as building their credibility or would they be more inclined to more evenly distribute edits to give the appearance of a broader background? It is possible that the approach used here to measure editor contributions and, by extension, their level of perceived credibility, was informative largely because editors have generally worked anonymously in that their individual efforts tended to be ignored by users of the site.

Accidental Collaboration

The process of paring down the historical contributions of prolific contributors to the selected science articles was an attempt to explore an approach to article analysis with the ultimate goal of providing consumers of information on Wikipedia a potential tool for making judgements about the legitimacy of an article's content. An examination of the contributions of Wikipedia editors showed a tendency toward higher edit counts for some articles over others. Manual examination of the edit histories of several users suggested there was a form of specialization that occurs. Many of the contributors identified in this study tended to have high edit counts for articles related to scientific topics suggesting some background and interest in science. Indeed, of the 101 contributors randomly

selected and their user pages explored, most of those who chose to provide background information had a scientific background. This provides support for the idea that prolific contributors to science articles likely have a scientific background and would therefore be expected to contribute to other scientific articles as well, and, in fact, the vast majority of articles identified via the accidental collaboration process were also scientific in nature.

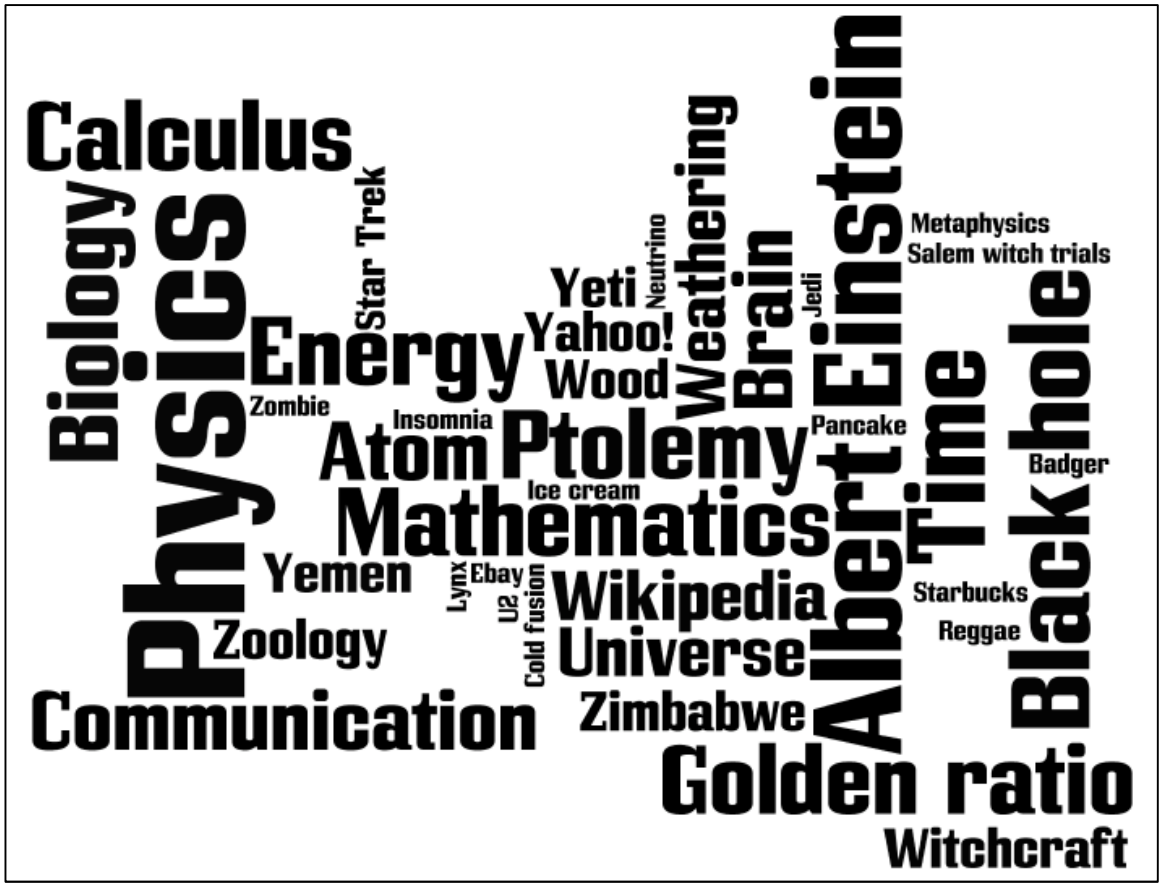


Figure 16. Sample visualization of article titles showing strength of accidental collaboration.

One aspect of the current study that proved to not be possible was the desire to created visualization of articles related to the seed articles via accidental collaboration. No software was found that could graphically display the various articles found and

weight them according to the strength of accidental collaboration. Currently available tools such as the one available on *www.worldle.net* require an extensive amount of manual manipulation to achieve desired results. The example in Figure 16 shows one possible form of visualization based on *Wordle*. For illustrative purposes, selected data were pulled from the Quantum Mechanics article and fed into *Wordle*. Other similar tools exist, but nothing was found that could automatically process a seed article and the related articles and create a visualization.

Arazy et al. (2010) examined the use of visualizations, which they called glyphs, to help explain contributions to articles. Korfiatis et al. (2006) presented some of their results on the centrality of contributors as a measure of authority in a visual format. While both studies were able to generate graphical representations of contributions, the process is still fairly technical and not currently usable by visitors to Wikipedia. These studies, as well as the current research, offer insight into how the problem of visualization may be approached in the future. Taken together, these studies may provide the tools necessary to clearly describe contributions to articles and the level of authority of its authors.

Computer Networks as Social Networks

Another interesting and unexpected outcome was support for viewing Wikipedia as a computer supported social network (CSSN). The primary function of Wikipedia is to provide an online database of encyclopedic content and not to facilitate the creation of social networks. However, because the content is socially constructed, social networks are formed. To some extent, user talk pages and article discussion pages do facilitate

social interaction, but it is generally for the purpose of hashing out the details of article content. From a social perspective, articles did demonstrate associations similar to those found in Milgram's (1969) Small World Experiment. While no attempt was made to count the number of intermediaries between selected articles, there were nevertheless articles, somewhat random in nature, that were linked by the actions of just two individuals such as contributors to the Chemistry article also contributing to an article on Adolf Hitler or contributors to the Zoology article also contributing to an article on the ham and cheese sandwich. Additionally, as in Milgram's study, was the existence of "sociometric stars" who accounted for a disproportionate number of article links. Numerous articles were linked by the actions of a few contributors such as *Vsmith* and all the articles related to the 12 seed articles examined in detail were the work of 99 unique contributors.

Retention of Experts

One of the benefits of including qualitative data in a study of Wikipedia, as encouraged by Kane and Fichman (2009), is the opportunity to develop a better sense of how articles are developed and the various contentious issues that could arise. Arazy et al. (2011) argued that:

> Even though cognitive diversity in online groups could potentially be high, many communities suffer from "cultural tribalism" in which people sample a large number of communities and migrate to the ones in which they hear what they want to hear, resulting in low cognitive diversity. Thus, cognitive diversity in online communities is only temporary and usually diminishes over time, resulting in dysfunctional communities. (p. 76)

The current study uncovered one possible example of this “cultural tribalism” with respect to retaining expert contributors. Based on some of the comments collected from user pages, there is an apparent undercurrent of frustration among subject level experts who often feel that their expert judgement is marginalized by the collective will of those who have staked a claim to an article. Wikipedia is aware of the issue.

> The issue of how to attract and retain specialists, given the anarchic and often frustrating nature of Wikipedia, is one that many Wikipedians feel needs to be addressed. Based on the thousands of articles needing expert attention, there is clearly a project need to encourage their participation and for the community to accommodate them. Some expert editors have withdrawn because of discontent with Wikipedia's policies and processes. No study has been undertaken to determine whether such a withdrawal has occurred in numbers significant enough to be problematic. Nevertheless, the perception alone may be sufficient to cause concern that material in Wikipedia is not written to a high standard of accuracy or completeness because of a lack of participation by subject matter experts. (“Wikipedia: Expert retention,” n.d.)

Efforts have been made by the community to address the issue but have so far been unsuccessful. The are competing essays both in support of and against the importance of credentials in Wikipedia, and efforts to create policies to ignore credentials or allow users to verify credentials have both failed including an effort put forth by Wales (n.d.).

While contributors need not be experts, those with an advanced degree or professional experience in the field they are contributing to, in order to contribute valuable content, experts often possess the depth of knowledge necessary to know how competing theories in a field are viewed or those that are generally given more credence in the academic community. According to Sanger (2004), co-founder of Wikipedia who left partly due to the politics of the project,

> To attact (sic) and retain the participation of experts, there would have to be little patience for those who do not understand or agree with Wikipedia’s mission… A *less* tolerant attitude toward disruption would make the project more polite,

> welcoming, and indeed *open* to the vast majority of intelligent, well-meaning people on the Internet. As it is, there are far fewer genuine experts involved in the project than there could and should be.

Summary

One of the overarching motivations for this study was to test an approach for making decisions about the legitimacy of content on Wikipedia and, by extension, any form of socially-constructed knowledge. Wikipedia was chosen as the focus of this study because it is extremely popular, has a large user base, and is the subject of much debate among teachers, librarians and students. Furthermore, it maintains edit histories for every article and every user that greatly facilitated data collection.

One of the major criticisms of Wikipedia is its lack of authority. There is no way to really know the names or backgrounds of any the various contributors to the articles. One approach to teasing out their probable backgrounds is to look at the types of articles to which an editor has most often contributed – the assumption being individuals will mostly contribute more heavily to topics about which they are knowledgeable – and the type of background and education they choose to report on their user page. However, unlike a traditional encyclopedia, articles in Wikipedia are not written by individuals but by groups, and often very large groups, of people who have all contributed differently at different times and often over the course of a long period of time. To attempt to pinpoint which individuals are mostly responsible for the content is simply not practical or very possible. Furthermore, as a collaborative effort, each article is the sum of their work at a given point in time. The status of an article today is not necessarily its status in the future. It therefore made sense to look at the combined actions of major contributors. To that

end, an alternative approach was developed for this study that examined the types of articles collaboratively contributed to by the major contributors to an initial seed article. This process was termed Accidental Collaboration. This was based, in part, on the assumption that people, taken as group, would tend to spend their greatest efforts on topics about which they had substantial background knowledge. The extent to which multiple individuals would overlap, and the types of articles on which they overlapped, could be used as a proxy tool for assessing the level of expertise and background possessed by these major contributors to the seed articles. It was believed that contributors to the selected science articles would also contribute to a number of similarly related science articles suggesting a probable background on that topic. A review of a selection of user pages also showed users tended to identify areas of interest that parallel their background and that they also tended to contribute more heavily to articles in those areas.

It was also of interest to try to answer questions about the level of user participation in Wikipedia and the extent to which contributors define themselves on their user pages. Because all users in this study were identified by having contributed substantially to science articles, an attempt was made to see if these contributors also made an effort to claim scientific expertise in their user pages. Because users could claim any level of education and professional experience they want, it is not possible to know for certain if they are in fact as they describe. A separate argument could be made regarding social capital in online environments and one's likelihood of providing accurate and truthful information in order to build and maintain social capital.

The primary intent of this research was to attempt to learn something about the patterns of contributions to Wikipedia articles and to determine if relationships among the contribution patterns of the various contributors existed and how they could be used to help make decisions regarding the legitimacy of Wikipedia content. Another related objective was to learn something about the contributors to Wikipedia and explore what types of information they would share about themselves and examine if such information suggested whether or not they possessed the required background and knowledge necessary to contribute to scientific articles. The results of the study may be used to help inform users of Wikipedia as well as how teachers and librarians might approach the topic of Wikipedia with their students. Ultimately, one would hope, this information can be used to better educate students with respect to using content found on Wikipedia, but also how to better address web content in general and to be literate users of information in the digital age.

Conclusion

For the foreseeable future, Wikipedia will likely remain a popular website as well as a source of contention for students, teachers and anyone else concerned with access to reliable and legitimate information. Results of this study showed that a selection of science articles were substantially written by a small group of people who appear knowledgeable about those topics. Such findings cannot be generalized to other areas of Wikipedia until they are examined through additional research. Methods explored in this study, such as a review of a contributor's most edited articles or articles linked to a selected article via accidental collaboration, could be employed by users of Wikipedia to

help them evaluate the legitimacy of the content. Unfortunately, no tools currently exist to allow users to easily examine these things. Until similar tools are developed, users of Wikipedia may wish to develop habits of deeper exploration even if that means only examining the user page of a few prolific editors to an article and the types of articles to which they have recently contributed.

REFERENCES

Abbate, J. (1999). *Inventing the internet*. Cambridge, MA: MIT Press.

Alexa Top 500 Global Sites. (n.d.). *Alexa the Web Information Company*. Retrieved October 14, 2011, from http://www.alexa.com/topsites.

American Association of School Librarians. (2009). *Standards for the 21st-century learner in action*. Chicago: American Association of School Librarians.

American Library Association. (1989, January 10). Presidential Committee on Information Literacy: Final Report. American Library Association. Retrieved April 3, 2011, from http://www.ala.org/ala/mgrps/divs/acrl/publications/whitepapers/presidential.cfm.

American Psychological Association. (2001). *Publication manual of the American Psychological Association*. Washington, DC: American Psychological Association.

Anderson, R. H., Bikson, T. K., Law, S. A., & Mitchell, B. M. (1995). *Universal access to e-mail: Feasibility and societal implications*. Santa Monica, CA: RAND.

Anthropology. (1922). In *Encyclopaedia Britannica - a dictionary of arts, sciences, literature and general information*. Retrieved April 8, 2011, from Google Books: http://books.google.com/ebooks?id=7mgNAQAAMAAJ.

Arazy, O., Morgan, W., & Patterson, R. (2006). Wisdom of the crowds: Decentralized knowledge construction in Wikipedia. In V. Ramesh and A. Sinha (eds.), *Proceedings of the 16th Workshop on Information Technologies and Systems*. Milwaukee: WITS, 79-84.

Arazy, O., Nov, O., Patterson, R., & Yeo, L. (2011). Information quality in Wikipedia: The effects of group composition and task conflict. *Journal of Management Information Systems*, 27(4), 71-98.

Arazy, O., Stroulia, E., Ruecker, S., Arias, C., Fiorentino, C., Ganev, V., & Yau, T. (2010). Recognizing contributions in wikis: Authorship categories, algorithms, and visualizations. *Journal of the American Society for Information Science and Technology, 61*(6), 1166-1179.

Bennington, A. (2008). Dissecting the web through Wikipedia. *American Libraries, 39*(7), 46-48.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American,* May 17, 2001.

Bush, V. (1945). As we may think. *The Atlantic, 176*(7), 1010.

Carr, N. (2006, February 16). Nature's flawed study of Wikipedia's quality. *Rough Type: Nicholas Carr's Blog*. Retrieved August 5, 2011, from http://www.roughtype.com/archives/2006/02/community_and_h.php.

Chai, C. S., & Tan, S. C. (2009). Professional development of teachers for computer-supported collaborative learning: A knowledge-building approach. *Teachers College Record, 111*(5), 1296-1327.

Chesney, T. (2006). An empirical examination of Wikipedia's credibility. *First Monday 11*(11), Retrieved Sept 19, 2010, from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1413/1331.

Claburn, T. (2009). Wikipedia considers coloring untested text. *Information Week.* Retrieved Oct 6, 2010 from http://www.informationweek.com/news/internet/security/showArticle.jhtml?articleID=219500669.

Collison, R. (1966). *Encyclopaedias: Their history throughout the ages*. New York: Hafner Publishing Company.

Consciousness. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved September 30, 2011, from http://en.Wikipedia.org/wiki/Consciousness.

Craven, P., & Wellman, B. (1973). The network city. *Sociological Inquiry, 43*(3-4), 57-88.

Cross, T. (2006). Puppy smoothies: Improving the reliability of open, collaborative wikis. *First Monday 11*(9). Retrieved Oct 10, 2010 from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1400/1318.

Cuban, L. (2001). *Oversold and underused: Computers in the classroom*. Cambridge, MA: Harvard University Press.

Cummings, S., Heeks, R., & Huysman, M. (2006). Knowledge and learning in online networks in development: A social-capital perspective. *Development in Practice, 16*(6), 570-586.

de Laat, M., Lally, V., Lipponen, L., & Simons, R. (2007). Investigating patterns of interaction in networked learning and computer-supported collaborative learning:

A role for social network analysis. *International Journal of Computer-Supported Collaborative Learning, 2*(1), 87-103.

DiMaggio, P., Hargittai, E., Neuman, W. R., & Robinson, J. P. (2001). Social implications of the internet. *Annual Review of Sociology, 27*, 307-36.

DiNucci, D. (1999). Fragmented future. *Print, 53*(4), 32.

DiRamio, D., Theroux, R., & Guarino, A. J. (2009). Faculty hiring at top-ranked higher education administration programs: An examination using social network analysis. *Innovative Higher Education, 34*(3), 149-159.

Emigh, W., & Herring, S. C. (2005). Collaborative authoring on the web: A genre analysis of online encyclopedias. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*. Washington, D.C.: IEEE Computer Society.

Encyclopædia Britannica Board of Editors. (n.d.). Encyclopædia Britannica, Inc. Corporate Site. Retrieved September 21, 2010, from http://corporate.britannica.com/board/index.html.

Encyclopaedia. (2002). In *The New Encyclopaedia Britannica*. Chicago: Encyclopaedia Britannica, Inc.

Ennett, S. T., Bauman, K. E., Hussong, A., Faris, R., Foshee, V. A., Cai, L., & DuRant, R. H. (2006). The peer context of adolescent substance use: Findings from social network analysis. *Journal of Research on Adolescence, 16*(2), 159-186.

Foster, D. W. (1999). The Claremont Shakespeare authorship clinic: How severe are the problems? *Computers and the Humanities, 32*(6), 489-508.

Foucault, M., & Rabinow, P. (1984). *The Foucault Reader*. New York: Pantheon Books.

Freire, P. (2000). *Pedagogy of the Oppressed*. New York: Continuum.

Fujisawa, K. K., Kutsukake, N., & Hasegawa, T. (2009). Social network analyses of positive and negative relationships among Japanese preschool classmates. *International Journal of Behavioral Development, 33*(3), 193-201.

Gall, J., Ku, H., Gurney, K., Tseng, H., Yeh, H., & Chen, Q. (2010). Citations of ETR&D and related journals, 1990-2004. *Educational Technology, Research and Development, 58*(3), 343-351.

Giles, J. (2005). Internet encyclopedias go head to head. Nature, 438, 900-901.

Glacial Epoch. (1949). In *The Encyclopaedia Britannica*. Chicago: Encyclopaedia Britannica, Inc.

Glott, R., Schmidt, P., & Ghosh, R. (2010). Analysis of Wikipedia survey data: Quality of Wikipedia content. *United Nations University – MERIT*. Retrieved October 17, 2011 from http://www.Wikipediastudy.org/.

Google. (n.d.). Google History. Google. Retrieved April 04, 2011, from http://www.google.com/corporate/history.html.

Hara, N., Shachaf, P. & Hew, K. F. (2010). Cross-cultural analysis of the Wikipedia community. *Journal of the American Society for Information Science and Technology, 61*(10), 2097–2108.

Harouni, H. (2009). High school research and critical literacy: Social studies with and despite Wikipedia. *Harvard Educational Review, 79(*3), 473-494.

Hawking, S. W. (1988). *A brief history of time: From the big bang to black holes*. Toronto: Bantam Books.

Haythornthwaite, C. (1996). Social network analysis: An approach and technique for the study of information exchange. *Library & Information Science Research; 18*(4), 323-42.

Helm, B. (2005, December 14). Wikipedia: A work in progress. *Business Week*. Retrieved Sept 26, 2010 from http://www.businessweek.com/technology/content/dec2005/tc20051214_441708.htm?chan=db.

Hippocampus. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved September 30, 2011, from http://en.Wikipedia.org/wiki/Hippocampus.

History of Wikipedia. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved August 29, 2011, from http://en.Wikipedia.org/wiki/History_of_Wikipedia.

Huysman, M. (2004). Design requirements for knowledge-sharing tools: A need for social capital analysis. In M. Huysman and V. Wulf (eds.), *Social Capital and Information Technology*, Cambridge: MIT Press

International Society for Technology in Education. (2007). *National Educational Technology Standards for Students*. Retrieved April 18, 2011, from http://www.iste.org/standards/nets-for-students/nets-student-standards-2007.aspx.

Jahng, N., Nielsen, W. S., & Chan, E. K. H. (2010). Collaborative learning in an online course: A comparison of communication patterns in small and whole group activities. *Journal of Distance Education, 24*(2), 39-58.

Jarkko, M., Ahlberg, M., & Dillon, P. (2010). The dynamics of an online knowledge building community: A 5-years longitudinal study. *British Journal of Educational Technology, 41*(3), 365-387.

Kafker, F. A. (1981). *Notable encyclopedias of the seventeenth and eighteenth centuries: Nine predecessors of the encyclopédie*. Oxford: The Voltaire Foundation.

Kane, G. C., & Fichman, R. G. (2009). The shoemaker's children: Using wikis to improve IS research, teaching, and publication. *MIS Quarterly, 33*(1), 1–22.

Kister, K. F. (1994). *Kister's best encyclopedias: a comparative guide to general and specialized encyclopedias* (2nd ed.). Phoenix: Oryx Press.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Thousand Oaks: Sage Publications.

Kobayashi, T., Ideda, K., & Miyata, K. (2006). Social capital online: Collective use of the internet and reciprocity as lubricants of democracy. *Information, Communication & Society, 9*(5), 582-611.

Kobus, K., & Henry, D. B. (2010). Interplay of network position and peer substance use in early adolescent cigarette, alcohol, and marijuana use. *Journal of Early Adolescence, 30*(2), 225-245.

Koehly, L. M., & Shivy, V. A. (1998). Social network analysis: A new methodology for counseling research. *Journal of Counseling Psychology, 45*(1), 3-17.

Korfiatis, N. Th., Poulos, M., & Bokos, G. (2006) Evaluating authoritative sources using social networks: An insight from Wikipedia. *Online Information Review, 30*(3), 252-262.

Leuf, B., & Cunningham, W. (2001). *The Wiki Way: Collaboration and Sharing on the Internet*, Reading, MA: Addison-Wesley.

Kogan, H. (1958). *The great EB; the story of the Encyclopaedia Britannica*. Chicago: University of Chicago Press.

Leggett, H. (2009). Wikipedia to color code untrustworthy text. *Wired Science*. Retrieved Feb 7, 2011 from http://www.wired.com/wiredscience/2009/08/wikitrust/.

Leiner, B. M., Cerf, V. G., Clark, D. D., Kahn, R. E., Kleinrock, L., Lynch, D. C., Postel, J., Roberts, L. G., & Wolff, S. (2003). A Brief History of the Internet. *The Internet Society*. retrieved June 17, 2011 from http://www.isoc.org/internet/history/brief.shtml.

Lih, A. (2009). *The Wikipedia revolution: How a bunch of nobodies created the world's greatest encyclopedia*. New York: Hyperion.

Luyt, B., Aaron, T. C. H., Thian, L. H., & Hong, C. K. (2008). Improving Wikipedia's accuracy: Is edit age a solution? *Journal of the American Society for Information Science and Technology, 59*(2), 318–330.

Luyt, B. and Tan, D. (2010), Improving Wikipedia's credibility: References and citations in a sample of history articles. *Journal of the American Society for Information Science and Technology, 61*(4): 715–722.

Magnus, P. D. (2006). Epistemology and the Wikipedia. Presented at the North American Computing and Philosophy Conference in Troy, New York. Retrieved September 18, 2011 from http://dspace.sunyconnect.suny.edu/handle/1951/42589.

Magnus, P.D. (2008). Early response to false claims in Wikipedia. *First Monday, 13*(9). Retrieved September 18, 2011 from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2115/2027.

Manca, S., Delfino, M., & Mazzoni, E. (2009). Coding procedures to analyze interaction patterns in educational web forums. *Journal of Computer Assisted Learning, 25*(2), 189-200.

McHenry, R. (2004, November 15). The Faith-Based Encyclopedia. *Ideas in Action*. Retrieved October 22, 2010 from http://www.ideasinactiontv.com/tcs_daily/2004/11/the-faith-based-encyclopedia.html.

McLuhan, M. (1964). *Understanding media; the extensions of man,*. New York: McGraw-Hill.

McMillan, S. J. (2000). The microscope and the moving target: The challenge of applying content analysis to the world wide web. *Journalism and Mass Communication Quarterly, 77*(1), 80-98.

MediaWiki. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved April 19, 2011, from http://en.Wikipedia.org/wiki/MediaWiki.

Modern Language Association of America. (2008). *The MLA style manual*. New York: Modern Language Association of America.

Neale, A., Dailey, R., & Abrams, J.. (2010). Analysis of citations to biomedical articles affected by scientific misconduct. *Science and Engineering Ethics, 16*(2), 251-261.

Neuroscience. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved September 15, 2011, from http://en.Wikipedia.org/wiki/Neuroscience.

Okoli, C., & Oh, W. (2007). Investigating recognition-based performance in an open content community: A social capital perspective. *Information & Management, 44*(3), 240-252.

Patch, P. (2010). Meeting student writers where they are: Using Wikipedia to teach responsible scholarship. *Teaching English in the Two-Year College, 37*(3), 278-285.

Pearson, M., Sweeting, H., West, P., Young, R., Gordon, J., & Turner, K. (2006). Adolescent substance use in different social and peer contexts: A social network analysis. *Drugs: Education, Prevention and Policy, 13*(6), 519-536.

Pennell, M. (2007). "Russia is not in Rhode Island": Wikitravel in the digital writing classroom. *Pedagogy: Critical Approaches to Teaching Literature, Language, Composition, and Culture, 8*(1), 75-90.

Penuel, W. R., Sussex, W., Korbak, C., & Hoadley, C. (2006). Investigating the potential of using social network analysis in education evaluation. *American Journal of Evaluation, 27*(4), 437-451.

Piltdown man. (2002). In *The New Encyclopaedia Britannica*. Chicago: Encyclopaedia Britannica, Inc.

Piltdown Man. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved October 10, 2011, from http://en.wikipedia.org/wiki/Piltdown_man

Pitts, V. M., & Spillane, J. P. (2009). Using social network methods to study school leadership. *International Journal of Research & Method in Education, 32*(2), 185-207.

Pringle, H. (2009, February 27). Archaeology Magazine Blog - Beyond Stone and Bone » Is Google Making Archaeologists Smarter? *Archaeology Magazine*. Retrieved September 15, 2011, from http://archaeology.org/blog/?p=332.

Rajagopalan, M. S., Khanna, V., Stott, M., Leiter, Y., Showalter, T. N., Dicker, A., & Lawrence, Y. R. (2010). Accuracy of cancer information on the internet: A comparison of a wiki with a professionally maintained database. *Journal of Clinical Oncology, 28*(15).

Raymond, E. S. (2001). *The cathedral and the bazaar: Musings on linux and open source by an accidental revolutionary*. Beijing: O'Reilly.

Read, B. (2006). Can Wikipedia ever make the grade?. *Chronicle of Higher Education, 53*(10), A31-A36.

Read, B. (2008). Wikipedia fights bogus credentials. *Chronicle of Higher Education*. Retrieved October 18, 2011, from http://chronicle.com/blogs/wiredcampus/Wikipedia-fights-bogus-credentials/2888.

Rector, L. H. (2008). Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. *Reference Services Review, 36*(1), 7-22.

Reliability of Wikipedia. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved July 29, 2011, from http://en.Wikipedia.org/wiki/Credibility_of_Wikipedia.

Renfrew, C., Frith, C., & Malafouris, L. (2008). Introduction. The sapient mind: Archaeology meets neuroscience. *Philosophical Transactions of the Royal Society B, 363*, 1935-1938.

Rosenzweig, R. (2006). Can history be open source? Wikipedia and the future of the past. *Journal of American History, 93*(1), 117-46.

Rubin, N. (1990). Social networks and mourning: A comparative approach. Omega: *Journal of Death and Dying, 21*(2), 113-127.

Ryan, A. M. (2001). The peer group as a context for the development of young adolescent motivation and achievement. *Child Development, 72*(4), 1135-1150.

Sanger, L. (2004, December 31). Why Wikipedia must jettison its anti-elitism. *Larry Sanger Blog*. Retrieved October 27, 2011, from http://larrysanger.org/2004/12/why-Wikipedia-must-jettison-its-anti-elitism/.

Scott, J. (1991). *Social network analysis: A handbook*. London: Sage.

Seigenthaler, J. (2005, November 29). A false Wikipedia biography. *U.S. & World - USATODAY.com*. Retrieved July 15, 2011, from http://www.usatoday.com/news/opinion/editorials/2005-11-29-Wikipedia-edit_x.htm.

Selber, S. (2004). *Multiliteracies for a Digital Age*. Carbondale: Southern Illinois University Press.

Shen, D., Nuankhieo, P., Huang, X., Amelung, C., & Laffey, J. (2008). Using social network analysis to understand sense of community in an online learning environment. *Journal of Educational Computing Research, 39*(1), 17-36.

Shu, W., Chuang, Y. H. (2011). The perceived benefits of six-degree-separation social networks. *Internet Research, 21*(1), 26-45.

Sijtsema, J. J., Ojanen, T., Veenstra, R., Lindenberg, S., Hawley, P. H., & Little, T. D. (2010). Forms and functions of aggression in adolescent friendship selection and influence: A longitudinal social network analysis. *Social Development, 19*(3), 515-534.

Snow, M. (2006). Wikipedia: Wikipedia signpost/2006-01-30/errors remedied. *Wikipedia, the Free Encyclopedia*. Retrieved August 4, 2011, from http://en.Wikipedia.org/wiki/Wikipedia:Wikipedia_Signpost/2006-01-30/Errors_remedied.

Sources and authorities for English history. (1949). *The Encyclopaedia Britannica*. Chicago: Encyclopaedia Britannica, Inc.

Stallman, R. M. (1985). *The Gnu manifesto*. Retrieved August 18, 2010 from http://www.gnu.org/gnu/manifesto.html.

Strategic Plan/Movement Priorities. (n.d.). *Strategic Planning*. Retrieved June 20, 2011, from http://strategy.wikimedia.org/wiki/Strategic_Plan/Movement_Priorities.

Surowiecki, J. (2005). *The Wisdom of Crowds*. New York: Random House.

Survey: The wiki principle. (2006, April 22). *The Economist, 379*(8474), 10.

Temple, C. (2005). Critical thinking and critical literacy. *Thinking Classroom, 6*(2), 15-20.

Thompson, L. (1996). School ties: A social network analysis of friendships in a multilingual kindergarten. *European Early Childhood Education Research Journal, 4*(1), 49-69.

Thoreau, H. D. (1910). *Walden*. New York: Thomas Y. Crowell & Company. Retrieved Oct 22, 2010 from http://books.google.com/books?id=yiQ3AAAAIAAJ.

Toolserver. (n.d.). *Meta-Wiki, a Wikimedia Project Coordination Wiki*. Retrieved July 16, 2011, from http://meta.wikimedia.org/wiki/Toolserver.

User: Brews ohare. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved September 18, 2011, from http://en.Wikipedia.org/wiki/User:Brews_ohare.

User: Christopher Thomas. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved September 9, 2011, from http://en.Wikipedia.org/wiki/User:Christopher_Thomas.

User: Duesentrieb/Contributors. (n.d.). *Meta-Wiki, a Wikimedia Project Coordination Wiki*. Retrieved Feb 18, 2011, from http://meta.wikimedia.org/wiki/User:Duesentrieb/Contributors.

User: Iulus Ascanius. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved September 9, 2011, from http://en.Wikipedia.org/wiki/User:Iulus_Ascanius.

User: JWSchmidt. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved September 9, 2011, from http://en.Wikipedia.org/wiki/User:JWSchmidt.

User: Looie496. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved September 15, 2011, from http://en.Wikipedia.org/wiki/User:Looie496.

User: Vsmith. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved September 9, 2011, from http://en.Wikipedia.org/wiki/User:Vsmith.

Van Cleemput, K. (2010). “I'll see you on IM, text, or call you”: A social network approach of adolescents' use of communication media. *Bulletin of Science, Technology & Society, 30*(2), 75-85.

Vickers, B. (2004). *Shakespeare, co-author: a historical study of the five collaborative plays*. Oxford: Oxford University Press.

Wales, J. (n.d.). User:Jimbo Wales/Credential Verification. *Wikipedia, the Free Encyclopedia*. Retrieved October 18, 2011, from http://en.Wikipedia.org/wiki/User:Jimbo_Wales/Credential_Verification.

Wang, L. (2010). How social network position relates to knowledge building in online learning communities. *Frontiers of Education in China, 5*(1), 4-25.

Wasserman, S., & Faust, K. (1997). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

Weare, C., & Lin, W. Y. (2000) Content analysis of the world wide web: Opportunities and challenges. *Social Science Computer Review, 18*(4), 272 –292.

Wellman, B., Salaff, J., Dimitrova, D., Garton, L., Gulia, M., & Haythornthwaite, C. (1996). Computer networks as social networks: Collaborative work, telework, and virtual community. *Annual Review of Sociology, 22*, 213-38.

Wellman, B. (2001). Computer networks as social network. *Science, 293*(5537), 2031-34.

Wells, H. G. (1938). *World brain*. Garden City, N. Y.: Doubleday, Doran & Co., Inc.

Wikipedia. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved October 31, 2011, from http://en.Wikipedia.org/wiki/Wikipedia.

Wikipedia: Article Feedback Tool. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved April 18, 2011, from http://en.Wikipedia.org/wiki/Wikipedia:Article_Feedback_Tool.

Wikipedia: Bots. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved April 19, 2011, from http://en.Wikipedia.org/wiki/Wikipedia:BOT.

Wikipedia: Expert retention. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved October 12, 2011, from http://en.wikipedia.org/wiki/Wikipedia:Expert_retention

Wikipedia: Featured portals. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved April 19, 2011, from http://en.Wikipedia.org/wiki/Wikipedia:Featured_portals.

Wikipedia: Introduction. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved August 29, 2011, from http://en.Wikipedia.org/wiki/Wikipedia:Introduction.

Wikipedia: List of Wikipedians by number of edits. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved October 23, 2011, from http://en.Wikipedia.org/wiki/Wikipedia:List_of_Wikipedians_by_number_of_edits.

Wikipedia: No original research. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved April 18, 2011, from http://en.Wikipedia.org/wiki/Wikipedia:No_original_research.

Wikipedia: Portal. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved August 29, 2011, http://en.Wikipedia.org/wiki/Wikipedia:Portal

Wikipedia: Protection policy. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved April 19, 2011, from http://en.Wikipedia.org/wiki/Wikipedia:Protection_policy.

Wikipedia: Size comparisons. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved September 29, 2011, from http://en.Wikipedia.org/wiki/Wikipedia:Size_comparisons.

Wikipedia: There is no credential policy. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved October 18, 2011, from http://en.Wikipedia.org/wiki/Wikipedia:There_is_no_credential_policy.

Wikipedia: Verifiability. (n.d.). *Wikipedia, the free encyclopedia*. Retrieved Sept 24, 2010 from http://en.Wikipedia.org/wiki/Wikipedia:Verifiability.

Wikipedia: Wikipedians. (n.d.). *Wikipedia, the Free Encyclopedia*. Retrieved October 23, 2011, from http://en.Wikipedia.org/wiki/Wikipedia:Wikipedians.

Witvliet, M., Van Lier, P. A. C., Cuijpers, P., & Koot, H. M. (2010). Change and stability in childhood clique membership, isolation from cliques, and associated child characteristics. *Journal of Clinical Child and Adolescent Psychology, 39*(1), 12-24.

Xu, Y., Farver, J. A. M., Schwartz, D., & Chang, L. (2004). Social networks and aggressive behaviour in chinese children. *International Journal of Behavioral Development, 28*(5), 401-410.

Zhu, E. (2006). Interaction and cognitive engagement: An analysis of four asynchronous online discussions. Instructional Science: *An International Journal of Learning and Cognition, 34*(6), 451-480.

APPENDIX A

COMPLETE LIST OF 180 ORIGINALLY SELECTED SCIENCE ARTICLES

Complete list of articles sampled in this study including the relevant URL at the time of the study. Articles marked as "excluded" returned no useable data due to differences in the naming scheme for the article title and corresponding URL or because the title referred to a different article outside the intended sample. Articles marked "no prolific contributors" did not have any contributors with more than ten edits and contained no data. Given the relatively small number of unusable articles these were simply discarded from the study.

Article Title	Article URL	Comments
Acid-base reaction theories	http://en.wikipedia.org/wiki/Acid-base_reaction	Excluded
Aerospace	http://en.wikipedia.org/wiki/Aerospace_engineering	Excluded
Agricultural	http://en.wikipedia.org/wiki/Agricultural_engineering	Excluded
Alchemy	http://en.wikipedia.org/wiki/Alchemy	
Analytical chemistry	http://en.wikipedia.org/wiki/Analytical_chemistry	
Anatomy	http://en.wikipedia.org/wiki/Anatomy	
Anthropology	http://en.wikipedia.org/wiki/Anthropology	
Applied physics	http://en.wikipedia.org/wiki/Applied_physics	
Applied sciences	http://en.wikipedia.org/wiki/Applied_science	Excluded
Archaeology	http://en.wikipedia.org/wiki/Archaeology	
Artificial intelligence	http://en.wikipedia.org/wiki/Artificial_intelligence	
Astrobiology	http://en.wikipedia.org/wiki/Astrobiology	
Astrochemistry	http://en.wikipedia.org/wiki/Astrochemistry	
Astronomy	http://en.wikipedia.org/wiki/Astronomy	
Astrophysics	http://en.wikipedia.org/wiki/Astrophysics	
Atmospheric sciences	http://en.wikipedia.org/wiki/Atmospheric_sciences	
Atomic physics	http://en.wikipedia.org/wiki/Atomic_physics	No Prolific Contributors
Behavioral neuroscience	http://en.wikipedia.org/wiki/Behavioral_neuroscience	
Behavioral sciences	http://en.wikipedia.org/wiki/Behavioural_sciences	No Prolific Contributors
Biochemistry	http://en.wikipedia.org/wiki/Biochemistry	
Bioethics	http://en.wikipedia.org/wiki/Bioethics	
Biogeography	http://en.wikipedia.org/wiki/Biogeography	
Bioinformatics	http://en.wikipedia.org/wiki/Bioinformatics	
Biological engineering	http://en.wikipedia.org/wiki/Biological_engineering	
Biology	http://en.wikipedia.org/wiki/Biology	
Biomedical	http://en.wikipedia.org/wiki/Biomedical_engineering	Excluded
Biomedical engineering	http://en.wikipedia.org/wiki/Biomedical_engineering	
Biophysics	http://en.wikipedia.org/wiki/Biophysics	
Biostatistics	http://en.wikipedia.org/wiki/Biostatistics	
Biotechnology	http://en.wikipedia.org/wiki/Biotechnology	
Botany	http://en.wikipedia.org/wiki/Botany	
Cell biology	http://en.wikipedia.org/wiki/Cell_biology	
Chemical	http://en.wikipedia.org/wiki/Chemical_engineering	Excluded
Chemistry	http://en.wikipedia.org/wiki/Chemistry	
Civil	http://en.wikipedia.org/wiki/Civil_engineering	Excluded
Cognitive science	http://en.wikipedia.org/wiki/Cognitive_science	
Computational linguistics	http://en.wikipedia.org/wiki/Computational_linguistics	No Prolific Contributors
Computational physics	http://en.wikipedia.org/wiki/Computational_physics	
Computer	http://en.wikipedia.org/wiki/Computer_engineering	Excluded

Computer science	http://en.wikipedia.org/wiki/Theoretical_computer_science	Excluded
Condensed matter physics	http://en.wikipedia.org/wiki/Condensed_matter_physics	
Conservation biology	http://en.wikipedia.org/wiki/Conservation_biology	
Cosmology	http://en.wikipedia.org/wiki/Cosmology	
Criminology	http://en.wikipedia.org/wiki/Criminology	
Cryobiology	http://en.wikipedia.org/wiki/Cryobiology	
Crystallography	http://en.wikipedia.org/wiki/Crystallography	
Cultural studies	http://en.wikipedia.org/wiki/Cultural_studies	
Cybernetics	http://en.wikipedia.org/wiki/Cybernetics	
Demography	http://en.wikipedia.org/wiki/Demography	
Dentistry	http://en.wikipedia.org/wiki/Dentistry	
Developmental biology	http://en.wikipedia.org/wiki/Developmental_biology	
Earth sciences	http://en.wikipedia.org/wiki/Earth_science	Excluded
Ecology	http://en.wikipedia.org/wiki/Ecology	
Economics	http://en.wikipedia.org/wiki/Economics	
Electrical	http://en.wikipedia.org/wiki/Electrical_engineering	Excluded
Engineering	http://en.wikipedia.org/wiki/Engineering	
Entropy	http://en.wikipedia.org/wiki/Entropy	
Environmental chemistry	http://en.wikipedia.org/wiki/Environmental_chemistry	
Environmental science	http://en.wikipedia.org/wiki/Environmental_science	
Environmental studies	http://en.wikipedia.org/wiki/Environmental_studies	
Epidemiology	http://en.wikipedia.org/wiki/Epidemiology	
Ethnic studies	http://en.wikipedia.org/wiki/Ethnic_studies	
Ethnobiology	http://en.wikipedia.org/wiki/Ethnobiology	
Evolutionary biology	http://en.wikipedia.org/wiki/Evolutionary_biology	
Evolutionary psychology	http://en.wikipedia.org/wiki/Evolutionary_psychology	
Experimental physics	http://en.wikipedia.org/wiki/Experimental_physics	
Fire protection	http://en.wikipedia.org/wiki/Fire_protection_engineering	Excluded
Food science	http://en.wikipedia.org/wiki/Food_science	
Forestry	http://en.wikipedia.org/wiki/Forestry	
Formal sciences	http://en.wikipedia.org/wiki/Formal_sciences	
Fringe science	http://en.wikipedia.org/wiki/Fringe_science	
Galactic astronomy	http://en.wikipedia.org/wiki/Galactic_astronomy	
General relativity	http://en.wikipedia.org/wiki/General_relativity	
Genetic	http://en.wikipedia.org/wiki/Genetic_engineering	Excluded
Genetics	http://en.wikipedia.org/wiki/Genetics	
Geochemistry	http://en.wikipedia.org/wiki/Geochemistry	
Geodesy	http://en.wikipedia.org/wiki/Geodesy	
Geography	http://en.wikipedia.org/wiki/Geography	
Geology	http://en.wikipedia.org/wiki/Geology	
Geomorphology	http://en.wikipedia.org/wiki/Geomorphology	
Geophysics	http://en.wikipedia.org/wiki/Geophysics	
Gerontology	http://en.wikipedia.org/wiki/Gerontology	
Glaciology	http://en.wikipedia.org/wiki/Glaciology	
Green chemistry	http://en.wikipedia.org/wiki/Green_chemistry	
Health	http://en.wikipedia.org/wiki/Health	
Health care	http://en.wikipedia.org/wiki/Health_care	
Health sciences	http://en.wikipedia.org/wiki/Health_science	Excluded
History	http://en.wikipedia.org/wiki/History	
History of science	http://en.wikipedia.org/wiki/History_of_science	
Humanities	http://en.wikipedia.org/wiki/Humanities	
Hydrology	http://en.wikipedia.org/wiki/Hydrology	
Immunology	http://en.wikipedia.org/wiki/Immunology	

Industrial	http://en.wikipedia.org/wiki/Industrial_engineering	Excluded
Inorganic chemistry	http://en.wikipedia.org/wiki/Inorganic_chemistry	
Interdisciplinarity	http://en.wikipedia.org/wiki/Interdisciplinarity	
Library science	http://en.wikipedia.org/wiki/Library_science	
Life sciences	http://en.wikipedia.org/wiki/Life_sciences	Excluded
Limnology	http://en.wikipedia.org/wiki/Limnology	
Linguistics	http://en.wikipedia.org/wiki/Linguistics	
Logic	http://en.wikipedia.org/wiki/Logic	
M-theory	http://en.wikipedia.org/wiki/M-theory	
Marine biology	http://en.wikipedia.org/wiki/Marine_biology	
Materials science	http://en.wikipedia.org/wiki/Materials_science	
Mathematical biology	http://en.wikipedia.org/wiki/Mathematical_biology	Excluded
Mathematical logic	http://en.wikipedia.org/wiki/Mathematical_logic	
Mathematical physics	http://en.wikipedia.org/wiki/Mathematical_physics	
Mathematical statistics	http://en.wikipedia.org/wiki/Mathematical_statistics	
Mathematics	http://en.wikipedia.org/wiki/Mathematics	
Mechanical	http://en.wikipedia.org/wiki/Mechanical_engineering	Excluded
Mechanics	http://en.wikipedia.org/wiki/Mechanics	
Medicine	http://en.wikipedia.org/wiki/Medicine	
Microbiology	http://en.wikipedia.org/wiki/Microbiology	
Military	http://en.wikipedia.org/wiki/Military_engineer	Excluded
Mineralogy	http://en.wikipedia.org/wiki/Mineralogy	
Mining	http://en.wikipedia.org/wiki/Mining_engineering	Excluded
Molecular biology	http://en.wikipedia.org/wiki/Molecular_biology	
Molecular physics	http://en.wikipedia.org/wiki/Molecular_physics	
Neural engineering	http://en.wikipedia.org/wiki/Neural_engineering	
Neuroscience	http://en.wikipedia.org/wiki/Neuroscience	
Nuclear	http://en.wikipedia.org/wiki/Nuclear_engineering	Excluded
Nuclear chemistry	http://en.wikipedia.org/wiki/Nuclear_chemistry	
Nursing	http://en.wikipedia.org/wiki/Nursing	
Oceanography	http://en.wikipedia.org/wiki/Oceanography	
Operations research	http://en.wikipedia.org/wiki/Operations_research	
Organic chemistry	http://en.wikipedia.org/wiki/Organic_chemistry	
Paleoclimatology	http://en.wikipedia.org/wiki/Paleoclimatology	
Paleontology	http://en.wikipedia.org/wiki/Paleontology	
Palynology	http://en.wikipedia.org/wiki/Palynology	
Parasitology	http://en.wikipedia.org/wiki/Parasitology	
Particle physics	http://en.wikipedia.org/wiki/Particle_physics	
Pharmacy	http://en.wikipedia.org/wiki/Pharmacy	
Philosophy of science	http://en.wikipedia.org/wiki/Philosophy_of_science	
Photochemistry	http://en.wikipedia.org/wiki/Photochemistry	
Physical chemistry	http://en.wikipedia.org/wiki/Physical_chemistry	
Physical geography	http://en.wikipedia.org/wiki/Physical_geography	
Physical sciences	http://en.wikipedia.org/wiki/Physical_science	Excluded
Physics	http://en.wikipedia.org/wiki/Physics	
Physiology	http://en.wikipedia.org/wiki/Physiology	
Planetary geology	http://en.wikipedia.org/wiki/Planetary_geology	
Planetary science	http://en.wikipedia.org/wiki/Planetary_science	
Plasma physics	http://en.wikipedia.org/wiki/Plasma_(physics)	Excluded
Political economy	http://en.wikipedia.org/wiki/Political_economy	
Political science	http://en.wikipedia.org/wiki/Political_science	
Pseudoscience	http://en.wikipedia.org/wiki/Pseudoscience	
Psychology	http://en.wikipedia.org/wiki/Psychology	
Quantum mechanics	http://en.wikipedia.org/wiki/Quantum_mechanics	
Radiobiology	http://en.wikipedia.org/wiki/Radiobiology	
Radiochemistry	http://en.wikipedia.org/wiki/Radiochemistry	

Science and technology studies	http://en.wikipedia.org/wiki/Science_and_technology_studies	
Science policy	http://en.wikipedia.org/wiki/Science_policy	
Science studies	http://en.wikipedia.org/wiki/Science_studies	
Scientific method	http://en.wikipedia.org/wiki/Scientific_method	
Scientific modelling	http://en.wikipedia.org/wiki/Scientific_modelling	
Semiotics	http://en.wikipedia.org/wiki/Semiotics	
Social	http://en.wikipedia.org/wiki/Social_sciences	Excluded
Social work	http://en.wikipedia.org/wiki/Social_work	
Sociobiology	http://en.wikipedia.org/wiki/Sociobiology	
Sociology	http://en.wikipedia.org/wiki/Sociology	
Software	http://en.wikipedia.org/wiki/Software_engineering	Excluded
Soil biology	http://en.wikipedia.org/wiki/Soil_biology	
Soil science	http://en.wikipedia.org/wiki/Soil_science	
Solid mechanics	http://en.wikipedia.org/wiki/Solid_mechanics	
Solid-state chemistry	http://en.wikipedia.org/wiki/Solid-state_chemistry	
Space science	http://en.wikipedia.org/wiki/Space_science	
Special relativity	http://en.wikipedia.org/wiki/Special_relativity	
Stellar astronomy	http://en.wikipedia.org/wiki/Star	Excluded
Stereochemistry	http://en.wikipedia.org/wiki/Stereochemistry	
Supramolecular chemistry	http://en.wikipedia.org/wiki/Supramolecular_chemistry	
Surface science	http://en.wikipedia.org/wiki/Surface_science	
Systematics	http://en.wikipedia.org/wiki/Systematics	
Systems theory	http://en.wikipedia.org/wiki/Systems_theory	
Theoretical biology	http://en.wikipedia.org/wiki/Theoretical_biology	
Theoretical chemistry	http://en.wikipedia.org/wiki/Theoretical_chemistry	
Theoretical physics	http://en.wikipedia.org/wiki/Theoretical_physics	
Thermodynamics	http://en.wikipedia.org/wiki/Thermodynamics	
Toxicology	http://en.wikipedia.org/wiki/Toxicology	
Transdisciplinarity	http://en.wikipedia.org/wiki/Transdisciplinarity	
Urban planning	http://en.wikipedia.org/wiki/Urban_planning	
Veterinary medicine	http://en.wikipedia.org/wiki/Veterinary_medicine	
Zoology	http://en.wikipedia.org/wiki/Zoology	

APPENDIX B

TRANSCRIPT OF THE ORIGINAL TOOLSERVER.ORG QUERY REQUEST

Below is an abridged version of the query request. It includes the actual request as well as most of the follow up comments. This is included to help clarify the process used to obtain data for this study on Wikipedia.

--Begin transcript--

DBQ-140

Selecting top contributors with 10 or more edits from a list of science articles in the English Wikipedia and count all contributor edits for all articles they have edited across Wikipedia.

Details

Environment:

Data will be fed into the UCINET (http://www.analytictech.com/ucinet/) and SAS (http://www.sas.com/technologies/analytics/statistics/stat/index.html) or similar software packages for analysis. An sql dump of the query is preferred.

Any logical ordering of the tables would be alright. What I have in mind is this: Table for each article, with columns USER and EDIT_COUNT. A row would be added for each user that has 10+ edits and is not a bot. The user name and their total edits on that particular article would be added to the table. Table for each USER, with columns ARTICLE_NAME and EDIT_COUNT. A row would be added for each article the user has edited. The article name and the number of times they have edited that particular article would be added to the table. Since there could be possible thousands of tables, a possible table-of-tables would be nice to organize them. Maybe the table would have two columns. TABLE_TYPE (user, article, or original science article) and TABLE_NAME (just the name of the table). If you find any other approach easier or more logical, feel free to adjust as necessary.

Participants:

Betacommand
Hoo man
Jim Hutchinson
and Platonides

Description:

This analysis is part of a graduate research project. The data will be used to explore possible patterns in the connections between articles based on how often and in what ways multiple contributors overlap in various articles. The source articles for the analysis are those listed as "Part of a series on Science" listed on http://en.wikipedia.org/wiki/Science and consists of about 200 articles. I can provide a list of all article URLs if necessary or if helpful for clarity.

Data from the SQL tables will be read into UCINET and used to weight edits and create an activity map based on articles most frequently edited by overlapping contributors. This will be done for each of the selected science articles and the resulting "maps" will contain the other articles the top contributors contributed to and showing the strength of the relationships.

Activity

Betacommand added a comment - 03 May 2011 20:38:56

> Can you provide a clean list of all titles?

Platonides added a comment - 03 May 2011 21:53:24

> Please specify what you want.
> (I include my guesses below)
>
> For which articles?
> (NS main articles linked from http://en.wikipedia.org/wiki/Template:Science)
>
> What users do you want?
> (users with more than 10 edits to a single one of those articles)
>
> What data for each user?
> (list of all articles in main namespace they have edited, with number of edits to each one)

Jim Hutchinson added a comment - 03 May 2011 22:47:37

> Pardon my noob-ness as I adjust to this work flow. Thanks for the clarification questions. Hopefully this will help.
>
> 1. Start with these articles:
>
>   [NOTE TO READER: the list of 180 articles was originally included here]

2. Identify the contributors of each article with 10 or more edits (contributors tool in "article history" on each page will show this. For example, http://toolserver.org/~daniel/WikiSense/Contributors.php?wikilang=en&wikifam=.wikipedia.org&grouped=on&page=Science

3. For each identified contributor in step 2, list all articles they have edited with an edit count for each. For example, http://en.wikipedia.org/wiki/Special:Contributions/Vsmith lists all their contributions. These articles would need to be identified and then a sum of their edit counts for each article. A very basic example for users A, B, C, and D is at http://goo.gl/VIWd6

4. Repeat for each article listed in step 1.

Let me know if there is anything else. Thanks again.

Betacommand added a comment - 03 May 2011 23:48:41

Running

Betacommand added a comment - 04 May 2011 03:24:13

see
http://toolserver.org/~betacommand/articleinfo.zip and
http://toolserver.org/~betacommand/userinfo.zip

Jim Hutchinson added a comment - 04 May 2011 04:38:27

Wow. That was fast. Thank you very much. It will take me a while to work through this, but a quick question based on a quick look through the data. The userinfo data seems to include all article edits for each user which is precisely what I was looking for. However, I do need to differentiate the activities of each contributor based on their activities in one of the original science articles (i.e. their appearance as a contributor in a particular science article). Is there a way to reconstruct just the contributors to each individual article, such as those contributing to quantum mechanics or astrophysics (obviously some will appear in multiple articles as well) in order to do a per article analysis? Perhaps that is the data included in the articleinfo files. Thanks again. This is greatly appreciate.

Jim Hutchinson added a comment - 04 May 2011 04:58:34

One more question for clarification. Was the query run on the live version of Wikipedia or a mirror copy updated on some schedule? It looks like the query may have taken several hours. If this was run on the live Wikipedia then it's possible that some
changes in edit counts could have occurred if a contributor made any edits during the time the query was running. Given the relatively short time it's unlikely there would have been any huge changes, but I will need to explain the details in my final writeup. Thanks.

Betacommand added a comment - 04 May 2011 14:22:49

The queries ran for less than an hour on the toolserver replicated copy of the database. In the articleinfo.zip there is a file that lists all of the page ids/article name pairs. each of the text files with a number refers to the page id of the relevant article, and lists all users who have made 10 or more edits to that given article.

Jim Hutchinson added a comment - 04 May 2011 14:45:28

Thanks again. I see that file now. Looks great. Sorry for all the questions, but one more. In the articleinfo files there are a dozen or so with no data. For example, one of them is 46771 which is the article "Agricultural" which actually maps to http://en.wikipedia.org/wiki/Agricultural_engineering and looks like it would have returned 2 users with more than 10 edits. I think I can probably just exclude these articles from the final analysis, but I'm just wondering what might explain it as I will probably have to explain why they are excluded. On a separate note, I would like to include a "special thanks" section in the final write up. Clearly I wouldn't be able to continue without the help of people here and Betacommand in particular. Please let me know if it would be okay to include your screen name and/or real name if you wish to share. There is probably a way to send a private message or email in response if anyone wishes to avoid posting personal information publicly. I will not include anything unless I receive explicit permission. Thanks again.

Betacommand added a comment - 04 May 2011 15:15:03

I just used what was listed before the URL so you listed Agricultural URL and I just stripped out the URL in order to get the page title. That is probably the cause for that issue. You can just use my screen name is fine.

--End Transcript--

APPENDIX C

DATA COMPILED FROM USER PAGES FOR 43 USERS IDENTIFIED WITH A REPORTED SCIENCE BACKGROUND

Details collected from 43 of the 101 users whose *Wikipedia* userpages were scraped for data related to their reported level of education, background, interests and expertise. This list comprises those users who had a reported background in a scientific field based on the information they chose to share. The usernames may or may not be reflective of an individual's actual name. All information included below was made public by the user.

Username	Top 5 Articles Based on Edit Count	Profession	Background	Noted Areas of Interest	Degree
Ahoerstemeier	Index of Thailand related articles, Thailand, Bangkok, Wiki, Wikipedia	programmer	physics	Physics, Particle physics, cosmology, astronomy, astrophysics, (and unrelated to education, spaceflight, history, biology, geology, geography, Thailand)	BS
Alan Au	Social network, Bioinformatics, Toolbar, Mercer Island Washington, Information science	wikipedian	bioinformatician		graduate student
AndreasJS	Ptolemaida, Greece, Diabetes Mellitus, Ancient Greek phonology	professor, medical scientist	biochemistry		PhD
Anlace	Overpopulation, Sonoma County California, Noise pollution, Richardson Bay, Fairfield Osborn Preserve	physicist	physics		PhD
Arpingstone	Emirates (airline), Boeing 747, Tillandsia, British Airways, Swindon	aerospace	aerospace		
Biophysik	Amyloid, Amylin, Biophysics, Lipid bilayer, Nuclear magnetic resonance	scientist	biophysics	biochemistry, chemistry, physics	PhD
Brews ohare	Centrifugal force (rotating reference frame), Speed of light, Pythagorean theorem, Matter, Maxwell's equations	research scientist, professor emeritus, published author	electrical engineering	device physics, circuit design, solid state physics	PhD
CBM	Godel's incompleteness theorems, First order logic, Exponentiation, Mathematical logic, Computability theory	mathematical logic	mathematical logic	mathematical logic	

Christopher Thomas	Black hole, Quasar, Antimatter, Time travel, Wormhole	student	physics	physics	PhD student
Cquan	University of Rochester, Biomedical engineering, Biotechnology, Stem cell controversy, Tissue engineering	patent law, engineering, research and development	biomedical engineering, chemical engineering	biotechnology, bioengineering	BS and law student
CYD	Quantum mechanics, Richard Wagner, EPR paradox, Physics, Emacs		physics	physics	
Dicklyon	Golden ratio, Mouse (computing), List of inventors, Logarithm, Pixel	research engineer		photography, photometry, color, electronics, signal processing	
DO11.10	Poliomyelitis, Immune system, Vitamin D, Smallpox, Han van Meegeren	research fellow, scientist	immunology		PhD
DoctorW	Psychology, List of Cornell University alumni, Positive psychology, List of psychologists, developmental psychology	psychologist	developmental psychology	computer science, engineering, philosophy of science, late 16th century Korean Confucianism, Buddhism, ballroom dance, guitar, east Asian thought and religions	PhD
Dozenist	Dental caries, Tooth (human), Tooth development, Maxillary central incisor, Dentistry	dentist		music, poetry, novels, teeth	DDS (assumed)
Elekhh	Architecture of Denmark, Lists of national parks of Indonesia, Architecture, Architectural design competition, Sydney Opera House	architect		cities, culture, demographics, ecology, Australia, Europe, Southeast Asia, philosophy, maps	MS
Enormousdude	Energy, Force, Book of Mormon, Magnetic field, Archaeology and the book of Mormon	scientist	atomic/plasma physics	thermonuclear plasma physics, x-ray lasers and shock waves to neutron stars, gravitation	PhD
Favonian	Battle of Hastings, Louis Pasteur, Ali, 2009, Leif Ericson	software architect	mathematics	mathematics, history	
Fnlayson	Lockheed Martin F22 Raptor, Boeing 747, Boeing 777, Lockheed Martin F35 Fighting Falcon, Boeing 787 Dreamliner	engineer	aerospace	aerospace, space flight, aviation, aviation history	
GregBenson	Sea level change, Geologic modeling, Paleoclimatology, Sequence Stratigraphy, Orbital Forcing	geologist	earth science		graduate degree - unspecified

Iridium77	University of Warwick, Israeli West Bank Barrier, Polyethylene, Methylaluminoxane, Chemistry			chemistry, computers	
Iulus Ascanius	Test (student assessment), Item response theory, Traverse City Michigan, Waterton Wisconsin, Psychometric software		psychometrics	psychometrics	PhD
JabberWok	Lists of Jews, Military history of Jewish Americans, History of South Africa, Barry Gurary		physics		PhD student
Jmh649	Obesity, Attention deficit hyperactivity disorder, Gout, Trancendental Meditation, Dengue fever	physician	medicine	preventative medicine, popular science	MD
Joelmills	Dog health, List of dog diseases, Rabies, Canine parvovirus, Lymphoma in animals	veterinarian	veterinary medicine	Any and all articles dealing with veterinary medicine	veterinary
Jvbishop	Oxfordian theory of Shakespeare authorship, Pythagoras, Cell theory, Biology, French Revolution	scientist		biology, natural sciences, fossils,linguistics, paleontology, physics	
JWSchmidt	Francis Crick, James D. Watson, RuBisCO, Influenza A virus subtype H5N1, Hedgehog signaling pathway		chemistry		
Laurascudder	Adrian van der Donck, Hockaday School, Cleopatra VII, Augustas, St. Mark's School (Texas)		physics		graduate student
LeadSongDog	World War I, Alzheimer's disease, List of accidents and incidents involving commercial aircraft, 2009 flu pandemic, Trans fat	electronic engineer	physicist		
Lumos3	New Age, Vitamin C, Hemel Hempstead, Industrial Revolution, Solar energy	Business Systems Analyst and professional Business Facilitator	mechanical engineering	science, engineering, counter culture, self exploration, health, computing, geography, history	
Methcub	Biology, Anorexia nervosa, Evolutionary history of life, Kingsteignton, Snail		computer science		yes - unspecified
Mets501	Polar coordinate system, Trigonometry, 0.999…, Factorization, Violin	student	physics	math, science, physics, computers, aviation, classical music	student

Michael Hardy	List of statistics articles, List of trigonometric identities, Index of religion related articles, Normal distribution, Pythagorean theorem	statistician	mathematics		
Nick Green	Gordon Pask, Viable system model, No Doppelgangers, Self-organization, Variety (cybernetics)	cybernetician			
Osborne	History of phycology, Algae, Fucus, Ulster Museum, Ascophyllum nodosum	museum curator	botany	algae, lichens, birds	BA
Paul EJ King	Chemistry, Prince Albert Catholic School Division, History of chemistry, Quantitative trait locus, Multifactorial inheritance	high school teacher			B. Sc., B. Ed.
Peterlewis	Natural History (Pliny), Dolaucothi Gold Mines, Hushing, Forensic engineering, Pliny the Elder	forensic engineer			
Phmoreno	Kondratiev wave, Productivity, Second Industrial Revolution, Paper machine, Mass production		chemical engineering, information technology	historical economics, energy, natural resources	
Selket	Christine O'Donnell, Larry Darby, Sleep, Brain, Vestibule (architecture)	neuroscientist, computer programmer	neuroscientist		
Silly rabbit	Spinor, Exterior algebra, Hilbert transform, Cartan connection, Circle	mathematician			
Sunray	Sustainability, Vancouver, Community, Consensus decision making, I Ching		social sciences		
Tim Starling	List of compounds, List of topics characterized as pseudoscience, History of Australia, Australia, Semiconductor	wikimedia system administrator	physics		BS
Tryptofish	People for the Ethical Treatment of Animals, Atheism, Aquascraping, Crucifixion, Religion	scientist	biochemistry	eclectic	PhD

APPENDIX D

COMPLETE ARTICLE CATEGORIZATIONS FOR 12 SEED ARTICLES SELECTED

Chemistry

Article Title	Count of Accidental Collaborators	Sum of Edit Count	Category
Physics	11	345	Science
Energy	10	652	Science
Nitrogen	9	288	Science
Science	9	242	Science
Chemical_element	8	204	Science
Molecule	8	181	Science
Periodic_table	8	420	Science
Silver	8	275	Science
Sun	7	181	Science
Atom	7	312	Science
Biology	7	246	Science
Chemical_reaction	7	303	Science
Helium	7	315	Science
Hydrogen	7	210	Science
Isaac_Newton	7	287	Biography - Scientist
Organic_chemistry	7	231	Science
Oxygen	7	234	Science
Planet	7	124	Science
Tsunami	7	378	Science
Water	7	345	Science
Alcohol	6	171	Science
Amazon_Rainforest	6	194	Geography
Arsenic	6	184	Science
Asia	6	203	Geography
Astronomy	6	177	Science
Carbon_dioxide	6	310	Science
Chemical_substance	6	243	Science
Copper	6	305	Science
Covalent_bond	6	267	Science
Evolution	6	175	Science
Francium	6	82	Science
Gold	6	283	Science
Gravitation	6	186	Science
Human	6	114	Science
Industrial_Revolution	6	493	History
Jupiter	6	209	Science
Magnesium	6	228	Science
Mathematics	6	152	Science
Mercury_(element)	6	260	Science
Moon	6	198	Science
Potassium	6	229	Science
Radon	6	120	Science
Saudi_Arabia	6	104	Geography
Silicon	6	188	Science
Star	6	150	Science
Sulfur	6	342	Science
Thermodynamics	6	147	Science
Titanium	6	209	Science
Uranium	6	215	Science

Volcano	6	421	Science
Xenon	6	134	Science
2004_Indian_Ocean_earthquake_and_ tsunami	5	128	History
Adolf_Hitler	5	178	Biography
Africa	5	276	Geography
Aluminium	5	353	Science
Amazon_River	5	80	Geography
Anabolic_steroid	5	216	Science
Ancient_Egypt	5	59	History
Antarctica	5	235	Geography
Antisemitism	5	257	Culture
Benzene	5	131	Science
Boron	5	143	Science
Brain	5	128	Science
Canada	5	185	Geography
Carbon	5	234	Science
Cat	5	189	Science
Chemical_bond	5	177	Science
Chimpanzee	5	122	Science
Chlorine	5	216	Science
Christianity	5	175	Religion
Cocaine	5	127	Science
Coffee	5	417	Science
Dog	5	182	Science
Earth	5	317	Science
Engineering	5	109	Science
Europe	5	194	Geography
Galaxy	5	108	Science
Heat	5	185	Science
History	5	104	History
Iodine	5	149	Science
Iran	5	154	Geography
Iraq	5	156	Geography
Iron	5	224	Science
Islam	5	297	Religion
Jerusalem	5	53	Geography
Jihad	5	70	Religion
Life	5	125	Science
Light	5	148	Science
Lithium	5	252	Science
Mercury_(planet)	5	98	Science
Middle_Ages	5	93	History
Milky_Way	5	157	Science
Muhammad	5	207	Religion
Noble_gas	5	245	Science
Norway	5	88	Geography
Nuclear_power	5	153	Science
Pig	5	142	Science
Platinum	5	160	Science
Plutonium	5	125	Science
Properties_of_water	5	373	Science
Protein	5	97	Science
Quantum_mechanics	5	93	Science

Racism	5	178	Sociology
Religion	5	56	Religion
Saturn	5	158	Science
Scientific_method	5	109	Science
September_11_attacks	5	113	History
Slavery	5	69	History
Sodium	5	180	Science
Solar_energy	5	183	Science
Solar_System	5	163	Science
Spain	5	117	Geography
Tropical_cyclone	5	115	Science
United_States	5	288	Geography
World_War_II	5	195	History
Zinc	5	200	Science
Acid	4	202	Science
Afghanistan	4	56	Geography
Albert_Einstein	4	239	Biography - Scientist
Alchemy	4	107	Science
Ammonia	4	115	Science
Animal	4	110	Science
Argon	4	154	Science
Atheism	4	36	Religion
Atomic_theory	4	147	Science
Aztec	4	167	History
Banana	4	109	Science
Bangladesh	4	95	Geography
Barium	4	82	Science
Bat	4	102	Science
Bear	4	72	Science
Benjamin_Franklin	4	166	Biography - Scientist
Beryllium	4	96	Science
Bible	4	60	Religion
Big_Bang	4	87	Science
Bill_Clinton	4	162	Biography
Biofuel	4	152	Science
Bird	4	132	Science
Bismuth	4	68	Science
Black_hole	4	134	Science
Blood	4	95	Science
Bohr_model	4	65	Science
Brazil	4	124	Geography
Bromine	4	100	Science
Buddhism	4	91	Religion
Caesium	4	74	Science
Caffeine	4	358	Science
Calcium	4	174	Science
Cancer	4	74	Science
Charles_Darwin	4	105	Biography - Scientist
Cheese	4	194	Culture
Chemical_formula	4	172	Science
Cherokee	4	75	History
Chile	4	97	Geography
Christopher_Columbus	4	84	Biography
Chromatography	4	264	Science

Chromium	4	126	Science
Circulatory_system	4	38	Science
Clock	4	35	History
Coal	4	280	Science
Cobalt	4	162	Science
Denmark	4	59	Geography
Diamond	4	372	Science
Dinosaur	4	97	Science
Distillation	4	111	Science
Earthquake	4	269	Science
Egypt	4	111	Geography
Electron	4	79	Science
Electron_configuration	4	124	Science
Elephant	4	78	Science
Enzyme	4	220	Science
Fire	4	59	Science
Fixed-wing_aircraft	4	56	Science
Fluorine	4	134	Science
Force	4	66	Science
France	4	166	Geography
Gallium	4	58	Science
Geography	4	167	Geography
George_Orwell	4	79	Biography
George_W._Bush	4	104	Biography
Germanium	4	99	Science
Germany	4	136	Geography
Glacier	4	159	Science
Global_warming	4	310	Science
Glycerol	4	69	Science
God	4	238	Religion
Gorilla	4	65	Science
Greece	4	109	Geography
Guitar	4	74	History
Hades	4	89	Mythology
History_of_chemistry	4	107	Science
History_of_China	4	64	History
Humanism	4	67	Philosophy
Iceland	4	55	Geography
India	4	183	Geography
Intelligent_design	4	43	Religion
Ireland	4	81	Geography
Israel	4	142	Geography
Japan	4	179	Geography
Jellyfish	4	85	Science
Jesus	4	169	Religion
Judaism	4	70	Religion
Jyllands-Posten_Muhammad_cartoons_ controversy	4	86	History
Kenya	4	91	Geography
Krypton	4	173	Science
Ku_Klux_Klan	4	136	Culture
Lead	4	158	Science
Liger	4	54	Science
Louisiana_Purchase	4	132	History

Lysergic_acid_diethylamide	4	72	Science
Malcolm_X	4	64	Biography
March_2006	4	53	History
Mars	4	137	Science
Martin_Luther_King,_Jr.	4	143	Biography
MDMA	4	202	Science
Metabolism	4	75	Science
Metal	4	119	Science
Michael_Jackson	4	72	Biography
Middle_East	4	33	Geography
Mississippi_River	4	114	Geography
Molybdenum	4	63	Science
Mountain	4	113	Geography
Music	4	66	Art
Natural_gas	4	89	Science
Nazism	4	160	History
Neon	4	245	Science
Neptune	4	91	Science
New_York	4	47	Geography
New_Zealand	4	613	Geography
Nickel	4	174	Science
Nicolaus_Copernicus	4	44	Biography - Scientist
Nigger	4	124	Culture
Nile	4	80	Geography
Obesity	4	65	Science
Orbital_hybridisation	4	85	Science
Osmosis	4	131	Science
Ozone	4	144	Science
Pacific_Ocean	4	79	Geography
Paris	4	59	Geography
Peru	4	162	Geography
Philosophy	4	158	Philosophy
Phosphorus	4	152	Science
Photosynthesis	4	199	Science
Pie	4	95	Culture
Pluto	4	86	Science
Poland	4	85	Geography
Pollution	4	209	Science
Proton	4	46	Science
Robot	4	56	Science
Roman_Empire	4	61	History
Russia	4	104	Geography
Salt_(chemistry)	4	114	Science
Samurai	4	57	History
Scandium	4	68	Science
Scotland	4	303	Geography
Selenium	4	62	Science
Singapore	4	198	Geography
Sodium_chloride	4	81	Science
South_Africa	4	88	Geography
South_America	4	174	Geography
Space	4	67	Science
Strontium	4	68	Science
Sudan	4	82	Geography

Sugar	4	140	Science
Sweden	4	113	Geography
Tetrahydrocannabinol	4	68	Science
Texas	4	90	Geography
The_Holocaust	4	191	History
Tin	4	112	Science
Tornado	4	63	Science
Treaty_of_Paris_(1783)	4	70	History
Tungsten	4	90	Science
Ultraviolet	4	136	Science
Universe	4	162	Science
Venezuela	4	111	Geography
Washington,_D.C.	4	42	Geography
Wiki	4	130	Communication
Wikipedia	4	529	History
Wind_power	4	85	Science
Yttrium	4	74	Science

Epidemiology

Article title	Count of Accidental Collaborators	Sum of Edit Count	Category
Deaths_in_2010	2	288	History
Headache	2	134	Science
Kashrut	2	156	Religion
Miscarriage	2	28	Science
Motor_neurone_disease	2	88	Science
Muscular_dystrophy	2	53	Science
Nicotine	2	23	Science
Old_Testament	2	28	Religion
Parkinson's_disease	2	278	Science
Psychosurgery	2	15	Science
Sexually_transmitted_disease	2	22	Science
Space_hopper	2	19	Entertainment

Evolutionary Psychology

Article Title	Count of Accidental Collaborators	Sum of Edit Count	Category
Ashkenazi_Jews	2	17	Religion
Aspartame	2	28	Science
Asperger_syndrome	2	53	Science
Atheism	2	16	Religion
Autism	2	143	Science
Charles_Darwin	2	83	Biography - Scientist
Chocolate	2	55	Science
Creation_evolution_controversy	2	33	Science
Doomsday_argument	2	15	Science
Evolutionary_psychology_controversy	3	109	Science
Gender_identity	2	18	Science
Global_warming	2	54	Science
Health	2	29	Science
Higher_criticism	2	16	Literature
Historicity_of_Jesus	2	17	Religion
Human_evolution	2	16	Science
Human_gastrointestinal_tract	2	33	Science
Kevin_Smith	2	25	Biography
Language	2	94	History
Martin_Luther_King,_Jr._Day	2	29	Culture
Massage	2	32	Culture
Matriarchy	2	42	Culture
Medical_cannabis	2	116	Science
Memory	2	35	Science
Nonviolence	2	26	Sociology
Origin_of_language	2	22	History
Race_(classification_of_humans)	2	82	Science
Race_and_intelligence	2	112	Science
Richard_Dawkins	2	219	Biography - Scientist
Satanic_ritual_abuse_in_The_ Netherlands	2	23	Sociology
Soy_milk	2	24	Science
Sparta	2	16	Geography
Stephen_Jay_Gould	3	41	Biography - Scientist
Steven_Pinker	2	47	Biography - Scientist
The_God_Delusion	2	28	Religion

Forestry

Article Title	Count of Accidental Collaborators	Sum of Edit Count	Category
Plantation	3	69	History
Castanea_sativa	2	27	Science
Clearcutting	2	61	Science
Clearfelling	2	41	Science
Coffee	2	67	Science
List_of_forestry_universities_and_colleges	2	52	Science
Logging	2	60	Science
Scots_Pine	2	65	Science
Selection_cutting	2	17	Science
Tree	2	263	Science
Tree_planting	2	14	Science
#NAME?	2	91	Wikipedia Navigation Page

Geography

Article Title	Count of Accidental Collaborators	Sum of Edit Count	Category
Antarctica	4	222	Geography
Biology	4	203	Science
Chemistry	4	110	Science
Glacier	4	195	Science
Grand_Canyon	4	252	Geography
Jupiter	4	160	Science
Mississippi_River	4	114	Geography
Nitrogen	4	157	Science
Physics	4	125	Science
Silver	4	127	Science
Tsunami	4	333	Science
2004_Indian_Ocean_earthquake_and_ tsunami	3	98	History
Aluminium	3	200	Science
Amazon_River	3	61	Geography
Archaeology	3	54	Science
Arsenic	3	70	Science
Astronomy	3	129	Science
Beryllium	3	60	Science
Boron	3	70	Science
Canada	3	61	Geography
Carbon	3	143	Science
Chimpanzee	3	76	Science
Copper	3	180	Science
Cougar	3	67	Science
Crater_lake	3	48	Science
Earth	3	242	Science
Energy	3	72	Science
Engineering	3	91	Science
Europe	3	86	Geography
Forensic_science	3	63	Science
Galaxy	3	81	Science
Geographic_information_system	3	35	Science
Global_warming	3	273	Science
Himalayas	3	137	Geography
History_of_geography	3	58	Geography
Human	3	83	Science
Iron	3	120	Science
Light	3	91	Science
Magnesium	3	119	Science
Mathematics	3	64	Science
Mercury_(element)	3	108	Science
Milky_Way	3	136	Science
Moon	3	153	Science
Mountain	3	104	Geography
Niagara_Falls	3	151	Geography
Norway	3	66	Geography
Nuclear_power	3	110	Science
Oceanography	3	169	Science
Oxygen	3	121	Science

Pacific_Ocean	3	66	Geography
Paris	3	49	Geography
Photosynthesis	3	189	Science
Planet	3	81	Science
Platinum	3	99	Science
Plutonium	3	74	Science
Potassium	3	138	Science
Radon	3	44	Science
River	3	109	Geography
Robert_Boyle	3	98	Biography - Scientist
Rocky_Mountains	3	130	Geography
Sahara	3	156	Geography
Saturn	3	118	Science
Science	3	120	Science
Sodium	3	86	Science
Solar_energy	3	145	Science
Solar_System	3	144	Science
Star	3	100	Science
Sulfur	3	173	Science
Sweden	3	96	Geography
Titanium	3	112	Science
Tropical_cyclone	3	90	Science
Tungsten	3	52	Science
Ultraviolet	3	106	Science
United_States	3	141	Geography
Universe	3	138	Science
Virginia	3	35	Geography
Volcano	3	371	Science
Washington_(state)	3	23	Geography
Washington,_D.C.	3	230	Geography
Water	3	252	Science
Wikipedia	3	149	History
Wind_power	3	72	Science
X-ray	3	42	Science
Zinc	3	107	Science
Zoology	3	35	Science

Geomorphology

Article Title	Count of Accidental Collaborators	Sum of Edit Count	Category
Deforestation	3	229	Science
Drainage_basin	3	87	Science
Erosion	3	268	Science
Flood	3	39	Geography
Floodplain	3	44	Geography
List_of_important_publications_in_geology	3	29	Science
Puyehue-Cordón_Caulle	3	31	Geography
River	3	134	Geography
Water_cycle	3	264	Science
Water_resources	3	67	Science
Weathering	3	257	Science
Abyssal_plain	2	28	Science
Age_of_the_Earth	2	129	Science
Amazon_River	2	46	Geography
Andean_Volcanic_Belt	2	25	Science
Arctic	2	31	Geography
Asthenosphere	2	46	Science
Attribution_of_recent_climate_change	2	21	Science
Beringia	2	45	Geography
Biology	2	135	Science
Canadian_Shield	2	84	Science
Cataclysmic_pole_shift_hypothesis	2	41	Science
Catastrophism	2	35	Science
Cenozoic	2	52	Science
Clay	2	83	Science
Climate_change	2	110	Science
Continent	2	41	Geography
Continental_crust	2	42	Science
Continental_drift	2	133	Science
Crater_lake	2	31	Science
Crust_(geology)	2	103	Science
Crystal	2	173	Science
Current_sea_level_rise	2	72	Science
Deposition_(geology)	2	29	Science
Drag_(physics)	2	19	Science
Duluth,_Minnesota	2	14	Geography
Earth	2	178	Science
Ecosystem	2	92	Science
Evaporation	2	72	Science
Everglades	2	18	Geography
Expanding_Earth	2	58	Science
Fault_(geology)	2	94	Science
Geodynamics	2	14	Science
Geologic_time_scale	2	66	Science
Geology	2	212	Science
Geology_of_the_Rocky_Mountains	2	15	Science
Geophysics	2	19	Science
Global_cooling	2	77	Science
Global_warming	2	242	Science
Global_warming_controversy	2	86	Science
Granite	2	135	Science

History_of_the_Rove_Formation	2	29	Science
Human_Rights_Watch	2	18	Sociology
Hydrology	2	54	Science
Ice_age	2	140	Science
Igneous_rock	2	215	Science
Inner_core	2	147	Science
Lake_Superior	2	17	Geography
Lithosphere	2	92	Science
Magma	2	109	Science
Mantle_(geology)	2	141	Science
Mauna_Kea	2	23	Geography
Meander	2	54	Geography
Medieval_Warm_Period	2	25	Science
Metamorphic_rock	2	146	Science
Mid-ocean_ridge	2	63	Science
Midwestern_United_States	2	33	Geography
Mineral	2	256	Science
Missouri_River	2	68	Geography
Mountain	2	32	Geography
Nevado_del_Ruiz	2	62	Geography
Oldest_dated_rocks	2	17	Science
Olivine	2	48	Science
Orogeny	2	58	Science
Orthoclase	2	27	Science
Outer_core	2	64	Science
Pangaea	2	135	Geography
Patagonia	2	36	Geography
Physical_geography	2	33	Geography
Plate_tectonics	2	266	Science
Post-glacial_rebound	2	17	Science
Precambrian	2	81	Science
Quartzite	2	43	Science
Quaternary	2	41	Science
Radiometric_dating	2	52	Science
River_delta	2	99	Geography
Sand	2	96	Science
Sea_level	2	24	Geography
Seafloor_spreading	2	50	Science
Sediment	2	65	Science
Sedimentary_rock	2	152	Science
Seismology	2	39	Science
Shield_volcano	2	88	Science
Snowball_Earth	2	28	Science
Solar_variation	2	26	Science
Structure_of_the_Earth	2	154	Science
Subduction	2	46	Science
Supervolcano	2	41	Science
Surface_runoff	2	30	Geography
Valley	2	15	Geography
Volcano	2	240	Science
Volcanology	2	43	Science
Water	2	169	Science
Wood	2	69	Science

Hydrology

Article Title	Count of Accidental Collaborators	Sum of Edit Count	Category
Attribution_of_recent_climate_change	2	21	Science
Crater_lake	2	31	Science
Deforestation	2	222	Science
Drainage_basin	2	78	Science
Ecosystem	2	92	Science
Erosion	2	252	Science
Evaporation	2	72	Science
Flood	2	30	Geography
Floodplain	2	35	Geography
Geomorphology	2	48	Science
Geostatistics	2	15	Science
Physical_geography	2	33	Geography
River	2	103	Geography
Surface_runoff	2	30	Geography
Water_cycle	2	253	Science
Water_resources	2	60	Science

Limnology

Article Title	Count of Accidental Collaborators	Sum of Edit Count	Category
Eutrophication	2	60	Geography

Quantum Mechanics

Article Title	Count of Accidental Collaborators	Sum of Edit Count	Category
Physics	12	422	Science
Albert_Einstein	9	399	Biography - Scientist
Black_hole	8	167	Science
Time	8	185	Science
Biology	7	198	Science
Calculus	7	145	Science
Energy	7	85	Science
Golden_ratio	7	184	Science
Introduction_to_quantum_mechanics	7	354	Science
Mathematics	7	214	Science
Ptolemy	7	132	Biography - Scientist
Atom	6	363	Science
Brain	6	118	Science
Communication	6	97	Communication
Denmark	6	181	Geography
Electromagnetic_spectrum	6	153	Science
Extraterrestrial_life	6	101	Science
Galaxy	6	102	Science
Geometry	6	74	Science
Gravitation	6	229	Science
Ireland	6	99	Geography
Iron	6	113	Science
Isaac_Newton	6	155	Biography - Scientist
Law	6	87	Sociology
Liger	6	101	Science
Light	6	128	Science
Logarithm	6	101	Science
Magnesium	6	118	Science
Michael_Faraday	6	50	Biography - Scientist
Muhammad	6	84	Religion
Newton's_laws_of_motion	6	97	Science
Norway	6	170	Geography
Ocean	6	180	Geography
Philippines	6	107	Geography
Pi	6	116	Science
Potassium	6	136	Science
Prime_number	6	166	Science
Squirrel	6	78	Science
Thailand	6	113	Geography
United_States	6	161	Geography
Volcano	6	329	Science
Water	6	221	Science
X-ray	6	90	Science
0_(number)	5	65	Science
Abraham	5	82	Religion
Air_pollution	5	223	Science
Alessandro_Volta	5	75	Biography - Scientist
Alfred_Wegener	5	219	Biography - Scientist
Aluminium	5	197	Science
Amazon_River	5	71	Geography

Antarctica	5	174	Geography
Asia	5	59	Geography
Astrology	5	70	Science
Belgium	5	111	Geography
Biomass	5	91	Science
Blaise_Pascal	5	106	Biography - Scientist
Boston	5	43	Geography
Capitalism	5	48	Economics
Carbon_dioxide	5	176	Science
Chernobyl_disaster	5	61	History
Circulatory_system	5	103	Science
Cleopatra_VII	5	162	Biography
Comet	5	59	Science
Copper	5	162	Science
Cubism	5	40	Art
Death	5	98	Science
Dominican_Republic	5	148	Geography
Dubai	5	123	Geography
Ecology	5	120	Science
Encyclopedia	5	77	History
EPR_paradox	5	128	Science
Fibonacci_number	5	278	Science
Finland	5	89	Geography
Flood	5	89	Geography
Force	5	129	Science
Forensic_science	5	91	Science
Francis_Bacon	5	158	Biography - Scientist
George_W._Bush	5	161	Biography
Germany	5	85	Geography
Global_warming	5	253	Science
Greece	5	132	Geography
Haiti	5	100	Geography
Hedgehog	5	49	Science
Hera	5	118	Mythology
Honduras	5	74	Geography
Human	5	91	Science
Ice_hockey	5	60	Sports
Incandescent_light_bulb	5	70	Science
Jesus	5	78	Religion
Johannes_Kepler	5	127	Biography - Scientist
Kansas	5	53	Geography
Kinetic_energy	5	104	Science
Large_Hadron_Collider	5	120	Science
Lebanon	5	114	Geography
London	5	115	Geography
Louisiana_Purchase	5	88	History
Machu_Picchu	5	111	Geography
Mass-energy_equivalence	5	132	Science
Max_Planck	5	70	Biography - Scientist
Metal	5	136	Science
Milky_Way	5	127	Science
Mushroom	5	73	Science
Mythology	5	102	Mythology
Neon	5	114	Science

Nuclear_power	5	137	Science
Number	5	69	Science
Odyssey	5	189	Literature
Organism	5	97	Science
Paul_the_Apostle	5	66	Religion
Pencil	5	48	History
Periodic_table	5	164	Science
Philosophy	5	90	Philosophy
Plate_tectonics	5	308	Science
Protestant_Reformation	5	77	Religion
Quadratic_equation	5	285	Science
Racism	5	108	Sociology
Richard_Feynman	5	134	Biography - Scientist
Robert_Hooke	5	136	Biography - Scientist
Robot	5	96	Science
Rubik's_Cube	5	64	History
Saint_Peter	5	110	Religion
Samurai	5	92	History
Schrödinger's_cat	5	192	Science
Science	5	186	Science
Scotland	5	77	Geography
Space	5	109	Science
Speed_of_light	5	217	Science
Stephen_Hawking	5	57	Biography - Scientist
String_theory	5	66	Science
Sulfur	5	185	Science
Switzerland	5	131	Geography
Texas	5	129	Geography
Theory	5	147	Science
Theory_of_relativity	5	210	Science
Thomas_Becket	5	83	Biography
Treaty_of_Versailles	5	62	History
Triangle	5	132	Science
Trigonometry	5	175	Science
Tropical_cyclone	5	110	Science
Truth	5	60	Philosophy
Uganda	5	78	Geography
Universe	5	161	Science
Weathering	5	266	Science
Wikipedia	5	121	History
World	5	100	Philosophy
1991	4	114	History
1992	4	111	History
1993	4	103	History
1995	4	105	History
1906_San_Francisco_earthquake	4	91	History
Achilles	4	101	Mythology
Adolf_Hitler	4	148	Biography
Advertising	4	72	Communication
Alexander_Graham_Bell	4	36	Biography - Scientist
Alphabet	4	46	Language
Andorra	4	33	Geography
Angola	4	51	Geography
Anne_Frank	4	95	Biography

Antoine_Lavoisier	4	55	Biography - Scientist
Apollo_11	4	40	Science
Archimedes	4	78	Biography - Scientist
Ares	4	200	Mythology
Argentina	4	54	Geography
Argon	4	81	Science
Aristotle	4	240	Biography - Scientist
Arsène_Wenger	4	59	Biography
Astronomy	4	113	Science
Athens	4	57	Geography
Atlantic_Ocean	4	83	Geography
Atmosphere_of_Earth	4	205	Science
Australia	4	44	Geography
Austria	4	75	Geography
Aztec	4	115	History
Bangladesh	4	62	Geography
Barcelona	4	47	Geography
Batman	4	62	Entertainment
Beaver	4	62	Science
Benjamin_Franklin	4	33	Biography - Scientist
Berlin_Wall	4	57	History
Bicycle	4	41	History
Bill_Nye	4	44	Biography - Scientist
Billie_Holiday	4	45	Biography
Binary_numeral_system	4	49	Science
Biofuel	4	58	Science
Black_Death	4	53	History
Blood	4	97	Science
British_Columbia	4	40	Geography
Bronze_Age	4	56	History
Buoyancy	4	45	Science
C._S._Lewis	4	119	Biography
Calculator	4	41	History
Caligula	4	86	Biography
Carbohydrate	4	61	Science
Carbon	4	110	Science
Celebrity	4	59	Entertainment
Cell_(biology)	4	166	Science
Charlemagne	4	81	Biography
Charles_Babbage	4	52	Biography - Scientist
Charles_Dickens	4	70	Biography
Cheetah	4	46	Science
Chelsea_F.C.	4	84	Sports
Chemical_element	4	106	Science
Chemistry	4	80	Science
Child_abuse	4	64	Sociology
Child_labour	4	62	Sociology
Chimpanzee	4	54	Science
China	4	69	Geography
Chiranjeevi	4	32	Biography
Christian	4	89	Religion
Christopher_Columbus	4	50	Biography
Church_of_Scientology	4	66	Religion
Circle	4	62	Science

Clock	4	52	History
Cloud	4	63	Science
Clownfish	4	31	Science
Coal	4	192	Science
Common_cold	4	42	Science
Conservation_of_energy	4	63	Science
Cougar	4	54	Science
Creationism	4	50	Religion
Crystal	4	194	Science
Daniel	4	78	Religion
Dark_matter	4	81	Science
Deforestation	4	266	Science
Democratic_Republic_of_the_Congo	4	61	Geography
Density	4	170	Science
Diabetes_mellitus	4	48	Science
Diana,_Princess_of_Wales	4	78	Biography
Dictionary	4	60	Language
Dog	4	60	Science
Domestic_violence	4	37	Sociology
Dust_Bowl	4	65	History
E_(mathematical_constant)	4	73	Science
Earth	4	181	Science
Ecosystem	4	113	Science
Ecuador	4	107	Geography
Egypt	4	95	Geography
El_Salvador	4	100	Geography
Electric_guitar	4	104	History
Electricity	4	44	Science
Electron	4	72	Science
Emily	4	68	Culture
Emmanuel_Adebayor	4	52	Biography
Ergonomics	4	51	Science
Ernest_Rutherford	4	76	Biography - Scientist
Escherichia_coli	4	30	Science
Ethiopia	4	105	Geography
Euclid	4	89	Biography - Scientist
European_Union	4	115	Economics
Evolution	4	93	Science
Exponentiation	4	31	Science
Fibonacci	4	49	Science
Fire	4	123	Science
Fixed-wing_aircraft	4	71	Science
Florence	4	49	Geography
France	4	67	Geography
Gangster	4	52	Sociology
General_relativity	4	55	Science
Genetic_engineering	4	77	Science
Genius	4	63	Science
George_Washington	4	28	Biography
Geothermal_energy	4	68	Science
Giovanni_da_Verrazzano	4	110	Biography
Global_Positioning_System	4	68	Science
Google_Search	4	63	History
Grand_Canyon	4	248	Geography

Great_Barrier_Reef	4	197	Geography
Great_Britain	4	49	Geography
Great_Depression_in_the_United_States	4	41	History
Greenland	4	62	Geography
Guatemala	4	89	Geography
Guitar	4	57	History
Halle_Berry	4	46	Biography
Hamlet	4	41	Literature
Hard_disk_drive	4	51	Science
Hat	4	53	Culture
Health	4	68	Science
Henri_Matisse	4	30	Biography
Henry_VIII_of_England	4	73	Biography
Hephaestus	4	129	Mythology
Himalayas	4	95	Geography
Hippie	4	86	Culture
History	4	114	History
History_of_the_United_States	4	47	History
History_of_Wikipedia	4	49	History
Human_evolution	4	109	Science
Human_height	4	29	Science
Hurricane_Katrina	4	140	History
Hydropower	4	49	Science
Iceland	4	66	Geography
Igneous_rock	4	221	Science
Imperialism	4	144	Economics
Industrial_Revolution	4	82	History
Inertia	4	91	Science
Infinity	4	107	Science
Internal_combustion_engine	4	63	Science
Internet_slang	4	25	Sociology
Interpretations_of_quantum_mechanics	4	136	Science
IP_address	4	44	Science
IPod_Touch	4	44	History
Italy	4	101	Geography
Jack_the_Ripper	4	76	History
James_Bond	4	40	Entertainment
James_I_of_England	4	54	Biography
James_Madison	4	66	Biography
Jehovah's_Witnesses	4	91	Religion
Jellyfish	4	69	Science
John_Dalton	4	47	Biography - Scientist
John_F._Kennedy_assassination	4	77	History
John_the_Baptist	4	37	Religion
Karl_Marx	4	95	Biography
Kazakhstan	4	54	Geography
Kidney	4	30	Science
King	4	72	Culture
Kitten	4	26	Science
Korea	4	87	Geography
Latin	4	43	History
Leonidas_I	4	92	Science
Lever	4	66	Science
Life	4	113	Science

Lightning	4	49	Science
Linear_algebra	4	36	Science
List_of_unsolved_problems_in_physics	4	68	Science
Liver	4	34	Science
Louis_Pasteur	4	190	Biography - Scientist
Louis_Riel	4	75	Biography
Mali	4	45	Geography
Manhattan_Project	4	122	Science
Marine_biology	4	181	Science
Marketing	4	69	Communication
Mass	4	147	Science
Massachusetts	4	55	Geography
Matter	4	76	Science
Mauritius	4	78	Geography
Maxwell's_equations	4	97	Science
Mediterranean_Sea	4	104	Geography
Mercury_(element)	4	102	Science
Microwave_oven	4	32	History
Millard_Fillmore	4	57	Biography
Mississippi_River	4	100	Geography
Mobile_phone	4	73	History
Moby-Dick	4	27	Literature
Momentum	4	82	Science
Moose	4	99	Science
Mormonism	4	61	Religion
Mount_Etna	4	106	Geography
Mount_Kilimanjaro	4	52	Geography
Mount_Pinatubo	4	95	Geography
Mount_St._Helens	4	165	Geography
Mountain	4	49	Geography
Natural_resource	4	88	Science
Nero	4	135	Biography
Nevada	4	42	Geography
New_York	4	51	Geography
Newton's_law_of_universal_gravitation	4	70	Science
Niagara_Falls	4	171	Geography
Nicholas_II_of_Russia	4	57	Biography
Niels_Bohr	4	82	Science
Nikola_Tesla	4	32	Biography - Scientist
Nitrogen	4	113	Science
Nitrogen_cycle	4	138	Science
Nuclear_energy	4	66	Science
Oil_spill	4	72	Science
Oliver_Cromwell	4	97	Biography
Ontario	4	31	Geography
Operating_system	4	29	Science
Orange_(fruit)	4	80	Science
Oscar_Wilde	4	57	Biography
Paris	4	84	Geography
Pearl_Harbor	4	76	History
Pennsylvania	4	42	Geography
Pepsi	4	90	History
Philadelphia	4	45	Geography
Physical_attractiveness	4	31	Science

Piano	4	35	History
Plasma_(physics)	4	47	Science
Platinum	4	88	Science
Plutonium	4	65	Science
Poland	4	101	Geography
Polymer	4	40	Science
Polynomial	4	60	Science
Portugal	4	102	Geography
Potential_energy	4	95	Science
Properties_of_water	4	146	Science
Purple	4	75	Science
Pythagoras	4	128	Biography - Scientist
Pythagorean_theorem	4	116	Science
Rainforest	4	211	Geography
Rastafari_movement	4	132	Religion
René_Descartes	4	145	Biography
Rice	4	40	Science
Richard_Dawkins	4	61	Biography - Scientist
Robert_Boyle	4	96	Biography - Scientist
Rocky_Mountains	4	121	Geography
Roman_Empire	4	91	History
Roman_mythology	4	66	Mythology
Rome	4	87	Geography
Romeo_and_Juliet	4	78	Literature
Ronaldo	4	61	Biography
Sahara	4	161	Geography
San_Diego	4	47	Geography
San_Francisco	4	53	Geography
Satanism	4	39	Religion
Saturn	4	102	Science
Scientific_revolution	4	39	Science
Sean_Combs	4	47	Biography
Shia_Islam	4	72	Religion
Siberian_tiger	4	92	Science
Siege_of_Yorktown	4	43	History
Silk_Road	4	45	History
Silver	4	113	Science
Simón_Bolívar	4	43	Biography
Slavery	4	80	History
Snow_leopard	4	45	Science
Sock	4	45	History
Socrates	4	135	Biography
Sodium	4	77	Science
Solar_energy	4	138	Science
Solar_power	4	76	Science
South_America	4	35	Geography
South_Korea	4	69	Geography
Special_relativity	4	170	Science
Sphinx	4	43	Religion
Square_root	4	110	Science
Star	4	88	Science
Statistics	4	65	Science
Stephen_King	4	51	Biography
Suicide	4	85	Sociology

Sunflower	4	41	Science
Supernova	4	85	Science
Sweden	4	103	Geography
Sydney	4	62	Geography
Syria	4	65	Geography
Taco_Bell	4	60	Culture
Technology	4	87	Science
Ted_Kennedy	4	42	Biography
Teddy_bear	4	36	History
Terracotta_Army	4	96	History
Testicle	4	51	Science
The_Lord_of_the_Rings	4	57	Literature
The_New_York_Times	4	88	Communication
The_Star-Spangled_Banner	4	38	Culture
Tidal_power	4	71	Science
Trail_of_Tears	4	136	History
Trigonometric_functions	4	68	Science
Trojan_War	4	299	History
Tropical_rainforest	4	281	Geography
Tsunami	4	171	Science
Ultimate_fate_of_the_universe	4	58	Science
Ultraviolet	4	104	Science
Uncertainty_principle	4	81	Science
Uranium	4	103	Science
Uruguay	4	70	Geography
Vatican_City	4	78	Geography
Vlad_III_the_Impaler	4	58	Biography
War	4	156	Sociology
Water_cycle	4	127	Science
Water_pollution	4	274	Science
Wave	4	46	Science
Weather	4	145	Science
Web_2.0	4	55	Science
Whale	4	75	Science
Wicca	4	28	Religion
William_Harvey	4	125	Biography - Scientist
Wind_power	4	94	Science
Witchcraft	4	45	Religion
Wood	4	95	Science
Yahoo!	4	85	History
Yemen	4	42	Geography
Yeti	4	45	Science
Zimbabwe	4	87	Geography
Zoology	4	30	Science

Social Work

Article Title	Count of Accidental Collaborators	Sum of Edit Count	Category
Iraq_War	2	63	History
Medical_social_work	2	13	Science

Soil Science

Article Title	Count of Accidental Collaborators	Sum of Edit Count	Category
Fertilizer	2	84	Science
List_of_universities_with_soil_science_ curriculum	2	99	Science

Zoology

Article Title	Count of Accidental Collaborators	Sum of Edit Count	Category
Alligator	2	135	Science
American_Alligator	2	32	Science
Cat	2	70	Science
Crocodile	2	70	Science
Evolution	2	25	Science
Expelled:_No_Intelligence_Allowed	2	37	Entertainment
Guava	2	29	Science
Ham_and_cheese_sandwich	2	34	Culture
Johns_Hopkins_University	2	24	History
Robert_H._Goddard	2	40	Biography - Scientist
Scoville_scale	2	36	Science
Typewriter	2	41	History