Analyzing and Designing the Open Collaboration of Knowledge Content Creation

Huichen Chou

Doctoral Thesis Series of
Ito Laboratory
Department of Social Informatics
Kyoto University

Abstract

Open collaboration offers unlimited potential. It can bring together people with different backgrounds and skills without the need for physical proximity or for central management. This thesis considers knowledge content, an important social asset, and focuses on two types of open collaboration. The first type of open collaboration involves calling for people with diversity of backgrounds and skills to work together online, while the second uses existing knowledge content to allow users to build their own knowledge content. However, the quality of the knowledge content generated through the former approach is mostly low, and it is not yet fully understood why some forms of collaboration can achieve quality and others do not. The usage of other people's creations in new knowledge content can also be restricted by copyright. With the aim of supporting the continued development of open collaboration on knowledge content, the objective of this thesis is to:

- Deepen the understanding of how high-quality knowledge content is generated from open collaboration, in order to provide a reference to create more good-quality knowledge content.
- Consider the creative process involving the use of knowledge content generated by others and designing a system to support open collabora-

tion and copyright sharing for the creation of new knowledge content.

In this context, this thesis uses Wikipedia articles to support the case for the former type of collaboration, in which people are invited to work together online. We use the creation of teaching materials as an example of the latter type of open collaboration, which allows people to use existing knowledge content to create their own knowledge content. Our work makes three main contributions, as follows:

1. We analyze collaborations by investigating how different teams achieve output of similar quality.
   In Wikipedia, articles at Good Articles (GA) quality level of the same category exhibit large differences in number of editors. We analyze this situation using an approach in which we first employ factor analysis to identify and score editing abilities, and then use these scores to distinguish between editors. The sequence of participation by editors in the work process is generated in order to analyze patterns of collaboration. As a case study, three Wikipedia categories are examined, covering two general topics and a science topic, in order to demonstrate our approach. The results show a characteristic pattern in which editors with strong content-shaping ability are involved in the later stages of the collaboration process, regardless of the size of the team. Editors who perform few editing activities are mainly involved before this stage, and this causes the differences in terms of team sizes. Our results demonstrate that the proposed approach can provide a clearer understanding of how Wikipedia GA are created through open collaboration.

2. We propose a method to discover different collaboration patterns that

create GA for Wikipedia.

Since the process of collaboration required for the creation of quality articles is still unclear, we propose a novel method of identifying the collaboration patterns that lead to high-quality articles. We analyze GA from the same category of Wikipedia with the intention that our findings can be used as a reference for the creation of more high-quality articles for this category. Our method first differentiates between editors based on their activities. Next, we use dynamic time warping (DTW) to cluster the articles based on their file size changes. We then calculate the mean sequence for each cluster to identify the three phases that characterize the evolution of an article: growth, plateau, and decline. We finally examine the composition of active editors for each phase, to identify the overall collaboration patterns. We demonstrate this approach based on GA in three Wikipedia categories and find the main collaboration patterns for each category. Our approach extends existing knowledge in this field by characterizing the different types of collaboration patterns that are used to create GA of similar quality. These patterns can act as a reference for the generation of more GA for these Wikipedia categories.

3. We propose a solution that allows for open collaboration using copyrighted teaching materials.

   The use of existing resources to generate teaching materials can save effort and allow creators to achieve the desired quality more easily. However, although some resources can be used freely for educational purposes, others such as textbooks or online course materials cannot. Hence, a solution that facilitates the usage of copyright-restricted re-

sources for generating teaching materials with copyright sharing is needed. Our work exploits the advantages of blockchain technology and proposes a system for bonding participants with a smart contract. Our scheme securely registers records of multiple authorship and contributions for teaching materials involving the reuse of existing resources. Such records can be used as authorship evidence to claim economic benefits when a material is used. With the smart contract in place, a user interface and a service layer can be added to the system to provide further management functionalities for teaching materials.

These contributions can provide recommendations for how to create more high-quality content through open collaboration and can offer options to content producers who wish to retain their copyright when participating in open collaboration.

Acknowledgements

This thesis and my Ph.D. study would not be possible without the help of many people. Firstly, I would like to thank my supervisor Associate Professor Donghui Lin, for his endless patient in teaching me on research and academic writing as well as guide me through the process of obtaining a Ph.D. of Kyoto University. I would also like to thank Professor Takayuki Ito. Although we met during the later stage of my study, he has been very supportive and giving insightful comments on my research. I also would like to express my gratitude to my adviser Professor Masatoshi Yoshikawa, who has advised me since my master's degree. I always received valuable comments that enriched my work. I also want to thank another adviser Professor Hiroaki Ogata, who always gave advice to broaden my prospective and improve the quality of my research. In addition, I would like to thank Assistant Professor Rafik Hadfi, who helped me in documenting and gave some ideas to my research.

I would like to express my sincere gratitude to the faculty and members at Graduate School of Informatics at Kyoto University for their teaching and support to my study as well as their helps in my living in Japan. I also like to thank my fellow members of Ito laboratory and past members of Ishida

and Matsubara Laboratory for being good friends. I am grateful to you all.

Above all, I would like to express my deepest gratitude to Professor Toru Ishida, my master's degree supervisor, who accepted me into computer science study and laid the foundation of my doctoral research. His endless encouragement and inspiration have supported me much during my study. Finally, I would also like to thank my family and friends who have accompanied me throughout the Ph.D. study journey. I am blessed to have you.

My Ph.D. research is supported by a Grant-in-Aid for Scientific Research (A) (17H00759, 2017–2020), a Grant-in-Aid for Scientific Research (B) (21H03556, 2021-2024, and a Grant-in Aid for Challenging Research (Exploratory) (20K21833, 2020–2023).

Contents

Abstract **i**

Acknowledgements **v**

1 Introduction **1**

1.1 Overview 1

1.2 Objectives 2

1.3 Issues and Approaches 3

1.4 Thesis Outline 10

2 Background **11**

2.1 Introduction 11

2.2 Wikipedia: The Successful Case of Online Open Collaboration of Knowledge Content 13

2.3 Collaboration Studies on Wikipedia 14

2.4 Blockchain Technology: A Potential Solution for Open Collaboration of Copyright Sharing 19

2.5 Blockchain Based Collaboration Systems for Knowledge Content Creation 21

3 Understanding Open Collaboration of Wikipedia Good Articles with Factor Analysis **23**

3.1 Introduction 23
3.2 Related Work 26
3.3 Research Approach 29
3.3.1 Wikipedia Articles and Topic Selection 29
3.3.2 Two Phases Approach 30
3.4 Result Analysis 35
3.4.1 Overview of the Three Categories of Wikipedia . . . 35
3.4.2 Analysis of "US State Parks" GA 36
3.4.3 Analysis of "Children's Book" GA 42
3.4.4 Analysis of "Chemical Compounds and Materials" GA 48
3.5 Discussion 55
3.6 Conclusion 57

4 Discovering the Collaboration Patterns in Good Wikipedia Articles **60**

4.1 Introduction 61
4.2 Related Work 64
4.3 Method 66
4.3.1 Clustering Articles with Version Size Development . 67
4.3.2 Characterizing the Types of Editors 69
4.3.3 Representing Collaboration Patterns 70
4.4 Experiment and Results 71
4.4.1 US State Parks 72
4.4.2 Children's Book 77

4.4.3 Chemical Compounds and Materials 82
4.5 Discussion and Conclusion 87

5 TMchain: A Blockchain-Based Collaboration System for Teaching Materials 91
5.1 Introduction 92
5.2 Related Work 95
5.3 TMchain System Overview 97
5.3.1 System Requirement 97
5.3.2 System Framework 98
5.3.3 Smart Contract Functions 100
5.4 Scenario Implementation 106
5.4.1 Register Original Teaching Material in TMchain . . 106
5.4.2 Registering Teaching Material That Uses Other Materials in TMchain 107
5.5 Evaluation and Discussion 110
5.5.1 Function Evaluation 110
5.5.2 Performance Evaluation 111
5.6 Discussion 112
5.7 Conclusion 118

6 Conclusion 120
6.1 Contributions 120
6.2 Future Direction 123

Publications 125

Bibliography 127

List of Tables

2.1 Statistics of Wikipedia rated articles by quality 13

3.1 Editing activity categories 34
3.2 Statistics of “US state parks” GA 37
3.3 Rotated factor matrix of the “US state parks” category: 14 editing activities 39
3.4 Number of editors in different factor score ranges 39
3.5 Statistics of “children’s Book” GA 43
3.6 Rotated factor matrix of the “children’s book” category: 14 editing activities 44
3.7 Number of editors in different factor score ranges 45
3.8 Statistics of “chemical compounds and materials” GA . . . 49
3.9 Rotated factor matrix of the “chemical compounds and materials” category: 14 editing activities 51
3.10 Number of editors in different factor score ranges 51

List of Figures

1.1 Matrix for the creation of knowledge content via open collaboration. 3
1.2 Example of collaboration to create teaching materials. 7

3.1 Two phases approach to understand GA creation 30
3.2 Three examples of GA collaboration creation process for the US state park Wikipedia category. 40
3.3 Three examples of GA collaboration creation process for the children's books Wikipedia category. 46
3.4 Three examples of collaboration process for GA on chemical compounds and materials of Wikipedia category. 52

4.1 Proposed method for finding collaboration patterns 65
4.2 Clustering results of the US state parks GA. 73
4.3 Collaboration patterns of US state parks GA. 74
4.4 Clustering results of children's book GA. 78
4.5 Collaboration patterns of children's book GA. 79
4.6 Clustering results of chemical compounds and materials. . . 83

4.7 Collaboration patterns of chemical compounds and materials GA. 85

5.1 Teaching material creation process and TMchain system framework. 99
5.2 Smart contract codes of TMchain. 101
5.3 Logical graph data structures of TM information. 102
5.4 Physical data structures of TM information in Blockchain. . 103
5.5 Transaction flows of TMchain 104
5.6 *createMaterial* scenario: Teacher1 registers TM1 to TMchain 107
5.7 *deriveMaterial* scenario: Teacher3 registers TM3 to TMchain 108
5.8 Function runtime in milliseconds 111
5.9 Cost of for collaboratively created teaching materials 112
5.10 Example of collaborative TM management system architecture 116

Chapter 1

Introduction

1.1 Overview

Open collaboration is a relatively novel kind of human enterprise that has become popular since the emergence of internet technology. It largely relies on a system in an online environment to support the collective production of an artifact, and offers new opportunities for people to form ties with others and create things together [Forte and Lampe, 2013]. It provides unlimited potential by allowing volunteers of different background and skills to cooperate without the need for physical or centralized management. [Levine and Prietula, 2014] define open collaboration as "any system of innovation or production that relies on goal-oriented yet loosely coordinated participants who interact to create a product (or service) of economic value, which they make available to contributors and non-contributors alike." This term has been applied to a diverse range of ventures, including the creation of knowledge content. In this research, we consider knowledge to be an important

asset created by mankind, and define it as "facts, information and skills acquired through experience or education" [Dictionary, 2006]. Knowledge content records the knowledge produced by all of humanity and is used to pass on this knowledge from one generation to the next. This thesis considers different types of collaboration involving a variety of workers and the objective of content output, as shown in Figure 1.1. Workers who are participating in an open collaboration on knowledge content may have a wide range of backgrounds and skills, or may have similar backgrounds and skill levels. For the output of open collaboration, there are collaboration for a single output to share among the workers. There is also situation where outputs are independent from each other and shared to allow others to use and to build on, in order to create many more independent outputs.

1.2 Objectives

The primary objective of this thesis is to support the continuing development of knowledge content created through open collaboration, with a focus on the generation of quality output and content ownership such as copyright. We have two main motivations for achieving these goals:

1. We aim to provide a solution for creating more good-quality knowledge content through open collaboration by calling people to work together online.

   In order to achieve this, it is necessary to understand how high-quality knowledge content is generated in this context. Our findings can act as a reference and provide suggestions for collaboration to create good-quality content.

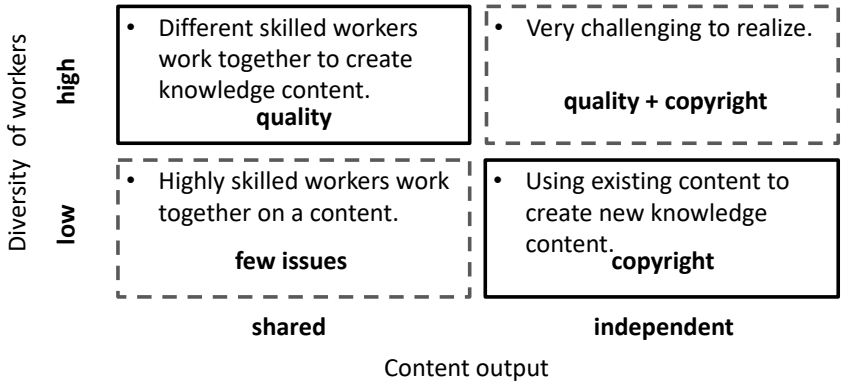


Figure 1.1: Matrix for the creation of knowledge content via open collaboration.

2. We design an alternative system to allow for the sharing of copyrighted collaborative knowledge content.

   Despite the efforts made in the area of Creative Commons [Lessig, 2004], there is a still a great deal of existing knowledge content that cannot be freely reused. The opportunity to collaborate is limited when a creator does not wish to share his/her work freely. There is therefore a need for an alternative system that can support the usage of copyrighted knowledge content via open collaboration.

1.3 Issues and Approaches

Based on the viewpoints described in Section 1.2, we identified the relevant issues and our approach for each goal, as shown in Figure 1.1.The matrix contains four types of collaboration. The upper left of the matrix represents the most common type, but the quality of the output cannot be guaranteed

due to the wide range of skill levels of the participants and the freedom to participate. The lower left of the matrix represents a set of workers with similar skills who create a single output that is shared among the collaborators. This approach is common in the real world for the creation of knowledge content and the free distribution of an otherwise copyrighted "work". In this case, the license is for non-profit usage, which allows other people the right to share, use, and build upon an original work. For copyright-restricted knowledge content that is not covered by this license, prearrangement with the author is required to obtain permission to use his/her creation.

The research objectives of this thesis are achieved through the following approach, by addressing the issues identified above.

1. We design a method to understand of how different teams create good quality output.

   Although open collaboration allows contributors with different skill levels to participate freely, the quality of the output cannot be guaranteed. In order to generate more good-quality knowledge content through open collaboration, it is necessary to understand how such content is generated. Understanding the collaboration process and discovering the collaboration patterns which govern how different teams produce quality knowledge content can provide a reference for the creation of more good-quality content.

   We use Wikipedia as a case study for our research on collaborations in which people work together online. Wikipedia is currently the most widely used knowledge content created via open collaboration by online volunteers [Wikimedia, 2022b]. According to a Wiki quality project Wikipedia articles can be graded into different quality levels

and only around 0.6% are recognized as being the level at Good Articles (GA) or above. The requirement for a GA include being "well written with no obvious mistakes and approaching the quality of a professional encyclopedia." [Wikimedia, 2022a]. This means that the content of a Wikipedia article that does not reach this level of quality may be not reliable.

Previous research has compared the editing activities involved and the formation of teams for the production of Wikipedia articles at different levels of quality. Studies have found that the quality generally improves with more words, more edits [Huberman and Wilkinson, 2007], and more surface edits [Jones, 2008]. Subsequently, studies started to argue that good editors were needed to produce quality work and focused on understanding this process by extracting editors through clustering or by identifying their expertise or reputation [Kane, 2011]. Several studies have addressed the diversity of editors and have introduced various diversity measurements for achieving higher quality [Liu and Ram, 2011, Robert and Romero, 2015]. However, these authors also found that large teams do not necessarily produce better work, and were unable to explain how teams of various sizes could create work of similar quality.

The issue of how teams of different sizes can produce content of the same quality (here, we focus on GA within the same Wikipedia category) remains unclear. An understanding of this will enable us to recommend a form of collaboration that will allow editors to create more good-quality content for a given Wikipedia category. We propose a approaches to deepen our understanding of the collaboration

involved in creating a GA. We first categorize the editors and then observe the sequence of participation by an active editor in the working process, to understand the situation. Inspired by the psychological research of Cattell, Horn and Carroll and their theory of human cognitive traits [Carroll et al., 1993], we apply factor analysis to the editing activities to obtain an editing trait. Then, each editor is scored on each trait to indicate his/her strength in terms of this trait. After that we observe the sequence of different editors engaging in the article creation process, from the initial article to its nomination as a GA, in order to study how a quality article is created. We use three different categories of Wikipedia article to demonstrate our approach: "US state parks", "children's books" and "chemical compounds and materials".

2. We design a method of finding collaboration patterns in order to extract reference collaboration models for making more good-quality content.

   This research extends the above research and proposes a method for revealing the existence of the different collaboration patterns that are involved in creating similar quality output for the Wikipedia category. Our findings can be used as a reference for the creation of more GA for this Wikipedia category. We extend our approach to differentiate editors and track their sequence of participation in the GA creation process. Here, we adopt our previous approach and apply factor analysis to each editor's activities to obtain editing traits. Then we study the involvement of editors in the GA creation process to identify collaboration patterns. The process of creating a GA is represented based on the growth in the article size, and we use the dynamic time warping

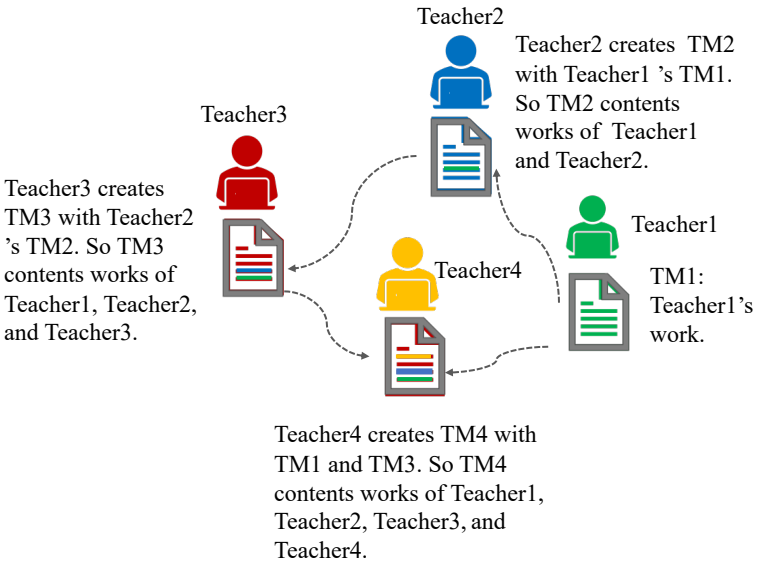


Figure 1.2: Example of collaboration to create teaching materials.

(DTW) method to identify clusters based on the sequence article size development. After that we use DTW Barycenter Averaging (DBA) to obtain the mean sequence for each cluster. The number of clusters determined through a hierarchical approach and the elbow clustering method [Shi et al., 2021]. We then divide the mean sequences into growth (G), decline (D), plateau (P) phases, according to the growth in the article size over time. Lastly, the collaboration patterns are visualized based on the distribution of editor's trait scores in each phase of the mean sequence for each cluster. Again, we use three different categories of Wikipedia article to demonstrate our approach: "US state parks", "children's books" and "chemical compounds and materials".

3. We analyze the creation process in which existing knowledge content is reused and design a solution to allow copyright restricted knowl-

edge content to be reused by sharing copyright.

To achieve this, we use teaching materials as a case study. Open collaboration is often involved and encouraged in the creation of teaching materials as knowledge content. It is common to use existing knowledge resources to save time and effort when creating teaching material [Hilton III and Wiley, 2009]. Open Educational Resources [Hylén, 2021] is an organization that aims to support open collaboration for knowledge content creation in this context, where resources are supported under a Creative Commons license. However, other resources such as textbooks or online courses cannot be shared. Authors who do not wish to donate their work and who require royalties cannot collaborate with each other without some form of prearrangement. There is therefore a need to design an alternative system to support the usage of copyrighted knowledge content in the context of open collaboration. A typical use case arises when teachers create teaching materials to pass on knowledge to students. Open collaboration involves making use of existing resources such as work by other teachers when creating new teaching material [Hilton III and Wiley, 2009] We present a simple example of such a case in Figure 1.2.

Figure 1.2 shows four teachers who are collaborating in the creation of teaching material (TM). Teacher1 has created TM1, and allows other teachers to use it. Teacher2 uses part of TM1 to create TM2, meaning that the authorship of TM2 belongs to both Teacher1 and Teacher2. Next, Teacher3 uses TM2 in his/her TM3. This means that the authorship of TM3 now belongs to Teacher1, Teacher2 and Teacher3. We can therefore view the creation of TM as a collaboration

between different teachers as they share their materials. In addition, if a TM with multiple authors is used to create new material, the existing authorship should be folded into the new material.

In this research, we propose to use a relatively new technology - the blockchain to design a solution for the copyright sharing of collaborative knowledge content. Blockchain has been suggested for use as a smart contract system among participants to support open collaboration. It provides a public ledger with secure, transparent, immutable, and distributed records that allows for decentralized control [Novotny et al., 2018]. There have been many research efforts towards the use of blockchain systems in the context of education. In terms of supporting academic collaboration, research has focused on the collaboration process. There are systems that can support researchers and institutions in sharing their research via open access publications [Günther and Chirita, 2018, Orvium, 2020]. There are also systems that can provide public and transparent tracking of all activities on a research paper, from first submission to revisions, peer reviews, copyright, and changes to the user license [Niya et al., 2019, Mohd Pozi et al., 2018]. However, these systems focus on collaboration on the creation of individual scientific papers or academic publications, and the reuse of existing resources for the collaborative production of knowledge content is not currently supported.

We therefore propose TMchain, a blockchain-based system for open collaboration on teaching material. Our work exploits the advantages of blockchain technology and develops a system that can bind participants via a smart contract; it securely registers records of multiple

authorship and the distribution of contributions to a teaching material that partially reuses existing resources. Such records can be used as authorship evidence to claim economic benefits when a material is used. In this way, open collaboration involving copyright-restricted content can be supported.

1.4 Thesis Outline

This thesis is organized into six chapters. The content of the remaining chapters can be summarized as follows. Chapter 2 presents the background to this thesis, and discusses the quality and copyright issues involved in knowledge content created via open collaboration. Chapter 3 presents a method for understanding the collaboration in the creation of output of similar quality, based on GA within the same Wikipedia category. We examine three categories of Wikipedia articles, and compare the collaboration patterns for teams of different sizes. In Chapter 4, we propose a method of discovering different collaboration patterns that can yield quality output. In Chapter 5, we introduce a blockchain-based system that can support open collaboration involving copyright-restricted work. Our scheme records multiple authorship and the contributions made to a collaboratively created work, and these records can be used to claim authorship. Finally, Chapter 6 concludes the thesis by summarizing our contributions and suggesting possible directions for future research.

Chapter 2

Background

In this chapter, we report the background of the two types of open collaboration of this research. We first report using Wikipedia as a case the open collaboration of calling people to work together on a single output. Then we report the research efforts have been contributed to study the quality issue of this type of collaboration. After that, we focus on another type of open collaboration which is making use of existing knowledge content to create new knowledge content. We use teaching material as a case and explain how blockchain technology can support this type of open collaboration. Finally, we present existing blockchain-based systems that support the open collaboration of creating knowledge content.

2.1 Introduction

This research focuses on two types of open collaboration in creating knowledge content. One type is calling people to work together online to create

knowledge content for a single shared output. This can be realized by allowing people to interact with each other through social media dialogue as well as acting collaboratively to create user generated content. People working collaboratively on online platforms are an important resource for creating social assets nowadays. Such platforms collect people of different ability and knowledge and realize on-demand work forces that can ignore physical constraints. This online collaboration harvests the "wisdom of crowds" and can tackle highly complex tasks such as creating knowledge contents. A prime example is Wikipedia. It has huge participation numbers and creates hundreds of thousands of articles every year. Its importance as a knowledge resource to society now yet its open collaboration faces challenges, especially with regard to the quality of the work, since the ability and the knowledge of the contributors must be expected to differ. Due to the concerns about the quality of Wikipedia content, studies have, over several decades, examined how to facilitate high quality work from the crowd by studying editors and their collaboration. In order to present the current understanding on how open collaboration can achieve higher quality output in the Wikipedia case, we report Wikipedia platform and its article quality assessment in Section 2.2 and discuss related collaboration studies of Wikipedia in Section 2.3.

Another type of open collaboration in creating knowledge content is making use of existing knowledge content into one's content. Currently, there is a lacking of a system to support the authorship sharing of such kind of collaboration. Blockchain technology has been suggested to facilitate open collaboration as well as provide solutions for copyright issues. We report the technology of blockchain in Section 2.4. In addition, blockchain has

Table 2.1: Statistics of Wikipedia rated articles by quality

Quality level	No. of articles
Featured Articles	6,017
Good Articles	36,061
Total	6,460,556

been proposed to support various collaborations under educational context. To provide a more comprehensive viewpoint, we report relevant systems that supports the collaboration of knowledge content in Section 2.5.

2.2 Wikipedia: The Successful Case of Online Open Collaboration of Knowledge Content

Wikipedia is one of the most successful stories for knowledge contents created from open collaboration. Its online platform, for English Wikipedia along, includes over 6 million articles of over 55 million revision pages and involves over 1 billions edits by over 43 million users [Wikimedia, 2022b]. It has a Wikipedia:WikiProject Wikipedia/Assessment which is a WikiProject that focuses on assessing the quality of Wikipedia-related articles[Wikimedia, 2022a]. The project department evaluates the quality of articles with a rating system and gives them banners on their talk page to reflect assessment results. This system helps users recognize a page's quality and the excellent contributions of editors as well as identifying articles that need further work. The rating system consists of the following seven levels written in ascending order of quality: Stub, Start, C-class, B-class, Good Articles (GA), A-class, and Featured Articles (FA). Wikipedia gives a detailed list of the criteria for these levels. A Featured Article is defined as fol-

lows: “A featured article exemplifies the very best work and is distinguished by professional standards of writing, presentation, and sourcing.” GA are “Useful to nearly all readers, with no obvious problems; approaching (but not equaling) the quality of a professional encyclopedia.” B-class: “Readers are not left wanting, although the content may not be complete enough to satisfy a serious student or researcher.” The assessment is done by impartial reviewers. One can consider article reaches GA level can be a reliable source of knowledge content. GA needs to be “well-written, comprehensive in coverage, well-researched with proper verifiable references, neutral in viewpoint, stable without any need to be updated often, compliance with Wikipedia style guidelines, and appropriate images and length [Wikimedia, 2022a]. Yet with the large amount articles available in Wikipedia, the portion of articles that can reach or above GA is minimum as shown in Table 2.1. There are less than 0.6% of articles passed GA and Featured Articles evaluation. This quality issue of Wikipedia articles has been a research interest for decades.

2.3 Collaboration Studies on Wikipedia

The interests of collaborative work of Wikipedia can trace back to year 2005. [Bryant et al., 2005] based on an activity theory and legitimate peripheral participation to study Wikipedia editors. They selected nine active Wikipedia contributors and studied their activities of edits on various articles throughout times and interviewed them. The research found the collaboration of Wikipedia editors differ greatly from the traditional publishing, such as they did not set a quality standard or goal when starting the project and they do not pay attention to labor of a division or a working team.

In addition, in order to research on a large amount of Wikipedia data, the automation annotation also receives research efforts. In 2006, on World Wide Web conference, [Völkel et al., 2006] suggested that the Wikipedia content is difficult interpreted by machine and therefore proposed a Semantic Wikipedia project called "Semantic MediaWiki". The project suggests editors to add markup language to the words, so the markups can provide extensions to enable users to semantically annotating wiki content. [Daxenberger and Gurevych, 2013] found the limitation of adding markup languages and design a model can automatically classify editing categories in Wikipedia reversions. The model uses metadata, textural, language properties and markup. They use supervised machine learning on the annotation of activities between two different versions of the articles. Yet they only receive average accuracy rate of 0.62 on 21 activities. They think it is because of the imbalanced and highly skewed activities category distributions of the training data. Yet this model is still used by other research, such as [Arazy et al., 2015] analyzed the emergent role behaviors of Wikipedia editors. [Yang et al., 2016] improved the model of "Semantic MediaWiki" and increased the test accuracy to 0.643 and then used the model to further identify the role of Wikipedia editors into eight types.

Content quality of collaborative work and in relations to the editor types has also been researched. [Lih, 2004] found the higher number of edits and unique editors can presume higher quality on an article. [Emigh and Herring, 2005] used word counts and letter counts to analysis the content variance and found the formality featured Wikipedia is indistinguishable from the expert created Encyclopedia. They suggest the "good" editors are extremely actives in the system to maintain the article quality. [Stvilia et al.,

2008] have been focused on quality of Wikipedia since 2005. They firstly used grounded approach to compare the creation process of featured articles (FA) and some random articles with content log and discussion pages to get qualitative information of the collaboration. Then in 2008, they added quantitative variables such as the ratio of different user/editor groups (administrator, editor and blot) to present their edit shares. They conclude the quality problem is context sensitive and no single model can assure the quality cross different systems. [Jones, 2008] compared a FA and a non-FA's editorial pattern to get the understanding of quality work creation on details of editing history pages. The non-FAs are selected from featured articles nominees but fails to get to FA status. He uses the data coding directly generated by Wikipedia in the revision history and he acknowledges some of these data somehow cannot represent the activity truthfully. The research result shows both article types have a higher percentage of addition of new materials compared with deletion and rearrange text. Yet non-FAs have fewer surface revisions to meet Wikipedia's policy. The surface revisions are structural guidelines and so on.

[Kittur and Kraut, 2008] also researched the Wikipedia quality. They focused on the coordination of collaboration and suggest adding more contributors seems to improve quality if appropriate coordination is achieved. They firstly studied the pattern on adding contributors and their coordination along the longitudinal life cycle of one FA. They also considered discussion page usage counts throughout the time as explicit coordination parameter and work share of an article as implicit coordination. They found that there are few editors contribute to the large work share and suggest both types of coordination can improve the quality of work at the article formation stage.

Then they further studied six sample articles of different quality levels and studied their quality change throughout different quality stage in relation to both explicit and implicit coordination. The result suggests the implicit coordination through concentration of work is more helpful to the article quality. [Kane, 2011] identifies the quality of Wikipedia articles is related to top contributors in terms of their experience in structuring, formatting and polishing documents but does not consider the other members of the team. [Liu and Ram, 2011] researched the relationship between Wikipedia article quality and editor types.

While most common ways to measure Wikipedia's work pattern are based on activity account and bits or words of contributions. [Geiger and Halfaker, 2013]use time spent contribution of Wikipedia editors to research the collaborative work. Their work confirm that human activity is often not normally distributed but occurs in a burst. In addition, the result also shows top editors appear differently based on different output metrics of edits counts and edit hours. [Robert and Romero, 2015] researched on the quality of 4,317 articles in the WikiProject: Film community. They suggested size and diversity are two key characteristics of crowds and identified their relationship to the performance. They considered the diversity of contributors as their work on different topics; in one article and the whole Wikipedia. [Arazy and Nov, 2010] found the quality of Wikipedia is related to coordination and contribution inequality with structural equation modeling. They found that global inequality (Wikipedia wide) has a significant positive impact on article quality, while the effect of local inequality (article wide) is indirect and is mediated by coordination. Another research stream focuses on editors and suggests the importance of major contributors while others

acknowledge the diversity provided by many minor editors [Kittur et al., 2007, Huberman and Wilkinson, 2007, Kittur et al., 2009]. The impact editor number on article quality has also been studied by comparing Wikipedia articles of different quality levels. All the results suggest adding more editors can improve article quality only in certain conditions, mainly in concentrated editing efforts [Kittur and Kraut, 2008, Kittur et al., 2009, Robert and Romero, 2015].

Collaboration patterns have also been studied. [Liu and Ram, 2011, Yang et al., 2016] again used articles of different quality levels to study what kind of team can achieve higher quality. These works indicated that different collaboration patterns can significantly impact article quality, but they failed to provide a clear model. The temporal changes in collaboration dynamic have been considered recently, and the finding confirms GA experiences a radical change in group behavior just prior to which is not found in non-GA. The changes occur in the level of activity, workload centralization, and a decrease in conflicts. [Zhang et al., 2017] studied the interplay between crowd evaluation and collaborative dynamics in Wikipedia articles. They considered editor group behavior over time and found teams become centralized, increase activities, and focus on the content a few months before the end of the GA nomination period. Our research also finds similar result. We break down the article creation into phases and find that different types of editors are involved in different phases.

2.4 Blockchain Technology: A Potential Solution for Open Collaboration of Copyright Sharing

Before the availability of blockchain, there are existing version control systems supporting collaboration creation. Well-known systems include Git and Wiki. Git is a tool initially for managing collaborative software development. It has also been used in create a documentation collaboratively and has been used for supporting teaching [Niya et al., 2019]. The system acts as a repository to store all the changes when people working on the same set of files. Yet Git has a centralized control mechanism and allows a manager level contributor to control the acceptance and revert a certain work submission to the system. To provide a truthful collaboration record against alteration is not the central focus of the system.

Another well-known system which supports collaboration and provides version change records is the Wiki system. The Wiki is a web-based system with discussions page and log and allows contributors to work collaboratively towards a Wikipedia article. Although it allows and records open collaboration in create a content, its data storage is centralized. Collaboration in Wiki is many authors works together in creating one work. This is different from collaborative educational resource creation that one single creation would be used multiple times. In addition, the works on Wikipedia are under Creative Commons license which do not allow using copyright works in the system [Lessig, 2004].

Blockchain is not a version control system but can act as a public ledger

to record collaboration activities which is more suitable to provide a record of authorships of educational resource collaboration. The record is kept distributed and the management is decentralized with network consensus. Such nature makes the record transparent, immutability and against alteration as well as performs management by the peers in the network without centralized control. Many blockchain based applications act as public ledgers have been proposed, namely for medical record, logistics and Internet of Things as well as for academic publications etc. [Crosby et al., 2016]. Blockchain is a system supports sharing ledger of transactions. It provides higher security, transparency, immutability of a record with decentralized management. It uses a network consensus to ensure all blockchains in the network are legitimate and all the copies in the network are the same. In this way, the system can make sure once a record has been added to the chain it is very difficult to change due to multiple copies of such a record exists in the network [Belotti et al., 2019].

The blockchain system was firstly used as cryptocurrency - Bitcoin. It uses a block to store transaction information when a toke is changing hands. It uses linked blocks to store the list of transaction history record to provide the prove of the existence of such a money. It is a "permissionless" blockchain system which means anyone can join the network to do transactions or participate in verification of transaction with network consensus. Then the usage of blockchain technology was extended with a "smart contract" concept. "Smart contract" is triggered by event or participants enquiries with prior designed computer protocol. Ethereum [Wood et al., 2014] is the most widely known system. It is a publicly distributed computing platform featuring smart con-tact functionality to build decentralized applications by

allowing code on blockchain. User needs to pay Ether – a crypto-token and a cost parameter call "gas" when using the platform. More recent development on blockchain technology is making usage of its distributed shared ledger nature to record immutable transactions among different entity's collaboration. Its nature lends itself to supporting community collaboration and shared outcomes. It applies to the open collaboration of knowledge content.

2.5 Blockchain Based Collaboration Systems for Knowledge Content Creation

Educational resources are important for education. Under open collaboration, many of the resources for education can be used freely based on copyright exemption for educational purpose or under open resources with collaborative commons license [Hylén, 2021]. OER Commons is a public digital library of open educational resources. It allows collaboration among teachers by sharing each other's work and encourage teachers to make use of existing contents – works of other teachers. However, the constrain of using not open education resources, such as textbooks and educational materials provided by online courses, in creating one's own educational resources is unsolved. There are also research efforts in using blockchain system to support collaboration for knowledge content but with focus on support academic publication.

ScienceRoot [Günther and Chirita, 2018] is created in 2017 as the first blockchain-enabled scientific ecosystem. It focuses on tokenization to drive the research process and view itself as a science research marketplace. It creates "Science Token to supports grant funding, publishing, and scien-

tific collaboration. Orvium is established in 2018 [Orvium, 2020]. It is an open source and decentralized platform to manage and support the transparent collaboration in science publication with blockchain. The system allows researchers and institutions to share their work as well as to create open access journal. The system provides a public transparent trace of all the activity pertaining to a research paper from first submission, revisions, accepted and rejected peer reviews, copyright and user license changes. Research papers are stored in a digital object identifier (doi) system with proof stamp to create a hash of the work.

[Niya et al. 2019] also proposed a blockchain-based an incentive publication model called Eureka. Eureka enables authors, referenced/linked author, editors, data providers and reviewers to share the economic reward with digital token 'EKA". The block contents collaboration information and the publication file is stored on the sciencematters.io platform. [Mohd Pozi et al. 2018] considers collaboration writing of scientific publications and uses blockchain system to preserve editing history on the block which can then be used for contribution calculation. The contribution rate of each author is also calculated based on the reversions stored in the blockchain network. It uses a smart asset platform to test its system design and design each block can store max 1024 characters.

These studies have been focused on allowing multiple authors to collaborate to create a single output. They cannot support collaboration on shared output and allow others to merge into a new piece of content or build. There is a large amount of existing knowledge content that cannot be used for open collaboration due to copyright restrictions.

Chapter 3

Understanding Open Collaboration of Wikipedia Good Articles with Factor Analysis

In this chapter, we report the research of understanding how the open collaboration that calls people online to work together can create quality knowledge content. This research aims at understanding how different teams create quality knowledge content with Wikipedia Good Articles (GA). To achieve this goal, we analyze who contributes to the collaborative creation and how they are involved in the collaboration process.

3.1 Introduction

The importance of Wikipedia to our society is beyond question. It is a free source of knowledge created by volunteers working collaboratively over an

online platform. Over the years, the collaboration of Wikipedia has received much research interest as it harvests the "wisdom of crowds" online to produce valuable contents [Kittur et al., 2007, Kittur et al., 2009, Niederer and Van Dijck, 2010]. Although calling upon online volunteers offers unlimited promise, the quality of the work is a concern. In 2007, a Wiki project on quality Wikipedia/Assessment was launched to manually assess the quality of Wikipedia-related articles. It grades Wikipedia articles into Feature Articles, A-class, Good Articles (GA), B-class, C-class, etc. according to their quality levels [Wikipedia 2020]. Among all the Wikipedia articles, only around 0.6% are recognized as GA level [Wikimedia, 2022a].

Due to the concerns about the quality of Wikipedia content, studies have, over several decades, examined how high-quality articles are yielded by open collaboration. Some studies focused on the need for more editors and more editing activities. They found that higher word counts, or surface edits seems to improve quality [Jones, 2008, Blumenstock, 2008]. Others suggest certain types of editors are needed to create high quality work [Klein et al., 2015]. Research also found adding more editors with more diverse backgrounds can also raise quality [Arazy et al., 2011, Robert and Romero, 2015]. Previous studies also researched what type of collaboration would yield better quality [Stvilia et al., 2008, Kittur and Kraut, 2008, Kittur et al., 2009, Ren and Yan, 2017]. Some found more editors with diverse backgrounds might not yield better quality [Kittur and Kraut, 2008, Kittur et al., 2009, Ren and Yan, 2017] and that implicit coordination-work directly on the articles is more important [Kittur and Kraut, 2008]. Some studies report patterns of editor combinations for different quality levels [Liu and Ram, 2011, Ren and Yan, 2017, Lin and Wang, 2020]. Another study focused on

the creation process and found that the editing activity increases and there are fewer editors prior to the quality assessment [Zhang et al., 2017].

This research continues past work on understanding the open collaboration of Wikipedia and focuses on the Good Articles (GA) level. GA is an important quality level and indicates a useful knowledge source. It is also considered as "useful to nearly all readers, with no obvious problems; approaching (but not equaling) the quality of a professional encyclopedia" [Giles 2005]. In addition, literature suggests there are topical differences in creating GA among Wikipedia categories [Pfeil et al., 2006]. The distribution of article categories is also uneven. To research all the articles of a certain quality level together would create bias in the observation and complicate the understanding of the collaboration necessary for creating quality articles [Halavais and Lackaff, 2008, Ren and Yan, 2017].

Here we explore the collaboration yielding Wikipedia GA in different categories. It is intuitive that topics requiring deep knowledge would have different editors involved than general knowledge topics. Therefore, the collaboration can be expected to be different. To fill out the prior study, we assess the collaboration used in creating GA of three particular categories so as to achieve a more granular understanding of the situation. For this, we choose two general topics and one science topic with similar numbers of GA. They are US state parks which is a general topic to which everyone can contribute; children's books which is a topic that also requires no deep knowledge; the science topic, chemical components & materials, requires deep knowledge. These three categories have around 20 GA each, which allows easier comparison. We use the approach first categories editors and then observe the sequence of active editor in the working process at articles

of different editor group sizes.

To distinguish the editors, we manually annotate the editing activities and run factor analysis based on editor's editing activities to obtain editing traits. Then each editor would have scores in each editing trait, and we can category editor's according to their scores on each editing trait. Then the sequence of different editors engaging in the article creation process is plotted from the article's start to its GA nomination passes to illustrate the collaboration. Last the collaboration patterns are reported according small, medium, and large groups of editors in each topic. Although the findings cannot represent the collaboration of Wikipedia GA as a whole. Based on the general topic and science topic cases used in this research, it gives insight of how topics of different knowledge requirement can be created.

3.2 Related Work

How quality articles are created by open collaboration activities on Wikipedia has been studied for decades. Wikipedia's success depends on volunteer editors, each of whom does a little bit of work that incrementally advances the article in coverage size [Kittur et al., 2007]. Yet it is known that Wikipedia articles are largely the result of a small number of editors [Kittur et al., 2007] and the quality of most of article is low. This concern triggered the launching of the "Wikipedia:WikiProject Wikipedia/Assessment" to evaluate the quality of articles with a rating system and place award banners on their talk page to reflect assessment results [Wikimedia, 2022a].

Wikipedia/Assessment consists of seven levels, written here in ascending order of quality: Stub, Start, C-class, B-class, Good Articles (GA), A-class,

and Featured Articles (FA). Wikipedia gives a detailed list of the criteria for these levels. A Featured Article is defined as follows: “A featured article exemplifies the very best work and is distinguished by professional standards of writing, presentation, and sourcing.” GA are “Useful to nearly all readers, with no obvious problems; approaching (but not equaling) the quality of a professional encyclopedia.” B-class: “Readers are not left wanting, although the content may not be complete enough to satisfy a serious student or researcher.” The assessment is done by impartial reviewers. A large body of studies has explored the benefit of this project and tried to find metrics that could quantify article quality.

Early Wikipedia quality and collaboration studies examined the number of editors, edits, and types of editing activities in higher quality articles. They found article quality generally is improved with more words, more edits [Blumenstock, 2008, Klein et al., 2015]. Subsequent studies started to argue that good editors are needed to produce quality work [Li et al., 2014] and focused on using cluster analysis to understand them [Liu and Ram, 2011] or identifying their expertise or reputation [Jones, 2008, Yarovoy et al., 2020]. Several studies addressed editor diversity and introduced various diversity measures for achieving better quality. They concluded large teams might not produce better work [Liu and Ram, 2011, Robert and Romero, 2015, Ren and Yan, 2017].

The methods of collaboration among editors have also been addressed. To attain quality output, editors need to coordinate and various coordination studies have been conducted [Stvilia et al., 2008, Kittur and Kraut, 2008, Kittur et al., 2009]. [Kittur et al., 2009] found that having implicit coordination - editors working together on the article itself is more important

and more editors might not be efficient with regard to creating quality work due to higher coordination costs. [Kane, 2011] found the volume of editing activities was not related to article quality, but the amount of effort spent in shaping articles has a positive impact on article quality. [Klein et al., 2015] studied the collaboration structure and found more editors can yield better article quality in some categories, but more editors per article can also reduce value. [Lin and Wang, 2020] also found that increasing the proportion of core members (people who frequently participate in editing) is likely to yield higher article quality.

There are works that looked for collaboration patterns of different Wikipedia quality levels. [Liu and Ram, 2011] first classified editors based on their editing activities with K-means clustering; they reported five collaboration patterns that yielded different levels of article quality. Unfortunately, they failed to identify a clear collaboration pattern that yielded better quality regardless of editor type. They also provided no clear recommendations that Wikipedia could apply to create quality articles.

[Zhang et al., 2017] researched on Wikipedia quality considered editor group behavior over time and found teams increase activities and focus on the content a few months before the end of the GA nomination period. However, no previous research has demonstrated a granular approach to understanding the collaboration activities in different topics. In this research, we propose an approach that can be generalized and applied to any Wikipedia category by providing a better understanding of open collaboration yielding GA.

3.3 Research Approach

In researching Wikipedia collaboration, subject difficulty would impact the collaboration pattern and the uneven distribution of article coverage in different categories might bias the GA creation pattern. Accordingly, we choose three different topics of different difficulty to test our proposed approach and confirm if an understanding of how open collaboration creates GA can be obtained. In this section, we first give details of the Wikipedia categories that we selected to study and then explain the two phases approach of our study in finding the collaboration pattern of creating. Figure 3.1.

3.3.1 Wikipedia Articles and Topic Selection

For this research, we selected "US state parks," "children's books" and "chemical compounds and materials" to represent two general topics and one science topic. The three categories contain around 20 GA each, which simplifies making comparisons among the three categories. We also avoid categories that might attract editors with commercial motivation or strong opinions, both of which would imply bias on content creation.

We choose "US state parks" - a sub-category under "Geography and places" category. It is considered to be a general topic that people without specific topic knowledge can contribute to and the topic is unlikely to attract strong opinions. We select only "children's books" GA from the "children's books, fairy tales, and nursery rhymes" category. We consider "children's books" to also be a general topic. Although it is literature related, even people with just a weak literature background can contribute. We also choose "chemical

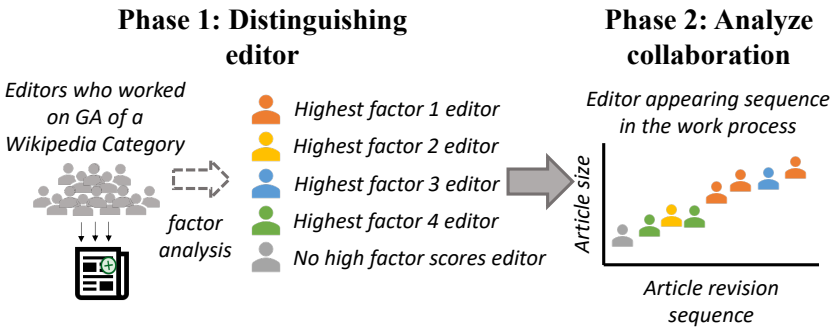


Figure 3.1: Two phases approach to understand GA creation

compounds and materials" from "Natural sciences" category as a science topic that requires editors who have specific knowledge. Our proposal can be easily extended to observe GA creation in other Wikipedia categories.

3.3.2 Two Phases Approach

We use a two-step approach to study collaboration within teams of editors: phase one: "Distinguish editors" and phase two: "Analyze the collaboration". In step one, we find out the difference in editors with the factor scores of factor analysis based on their involvement in the creation of GA in the same Wikipedia category. Then we plot the sequence of their appearance together with article size changes in phase two. We show the overview of our method in Figure 3.1. In the following, we detail our approach.

Distinguishing Editor Type

Editors bring different skills and knowledge to their projects and using such aspects to differentiate the editors helps us understand the source of quality discrepancies [Liu and Ram, 2011]. This, however, is difficult to achieve by

just observing editing activity type. Since each editor is involved in different activities, it is difficult to explain the differences among them from just the counts of their different editing activities. We borrow an approach from the psychology field - the "Cattell-Horn-Carroll model" of human cognitive abilities [Carroll et al., 1993]. The model is used to measure human cognitive abilities by performing factor analysis on the correlation of different data sets such as psychological tests, school marks and competence ratings to produce factors - a taxonomy of cognitive abilities. Factor analysis is a multivariable statistical approach for grouping large numbers of primary features and finding their linear combination that yields a global factor. Factor analysis can explore the possible underlying factor structure. Factors are latent variables that observed variables have similar patterns of responses to [Child 1990].

To determine appropriate editing categories for factor analysis, we examined the history entries of the view and talk pages of several GA and observed the changes in the articles between revisions. Based on [Pfeil et al., 2006]'s defined activity categories, we also followed their approach and used the grounded theory approach and extracted the possible categories of editing activities semantically. Since we also performed the category extraction process several times until saturation was reached, our category construction is a variant of theirs. We discarded the annotations of style/-typography and mark-up language since neither impacts the GA standard, and the US state park GA do not have these activities. This research adds a semantic of link activity with finer granularity of add links in the content, add links in the reference, and add links in the category of the article. We also added a talk page, which is a critical editing activity that represents col-

laboration cost. This yielded the following 14 activity categories as shown in Table 3.1.

We manually annotated these activities from the differences among revision pages in this research. This is because the current automated categorization system offers accuracy of 0.643 in terms of the semantic annotation on Wikipedia articles [Daxenberger and Gurevych, 2013]. We expect with the continuous advancement of automation on semantic annotation, the extension of this research can cover much larger numbers of Wikipedia categories in the future. In addition, our annotation is based on the revision history pages and the record of different activities on each page. If there were two "add information" activities in different parts of the content of one revision history page, we counted them as one activity to reduce the counting ambiguity created by extracting multiple bits of in-formation to describe one single issue.

In addition, we also normalized the editing activity counts based on the article and activity categories. Since the numbers of editors and editing activities exhibit a large variance among articles, we normalized the original data of the edit activity counts by dividing them by the total number of such editing activity of each article. This is the same method used in [Chou et al., 2020]. Accordingly, the variance of factor scores is 1 and the mean of factor scores of each factor is 0. We used IBM SPSS premium software for factor extraction. We firstly confirm our data is suitable for factor analysis with Kaiser-Meyer-Olkin (KMO) and Bartlett's test two parameters [Hill, 2011]; all three categories yielded a KMO Measure of Sampling Adequacy near 1 and a Bartlett's Test of Sphericity of significance close to zero. We used Principal Component Analysis (PCA) for the latent factor extraction and

eigenvalues equal or greater than 1 to determine the number of factors. We also performed Varimax rotation to minimize the number of variables that have high loading on each factor, which helps to simplify the interpretation of the factors [Child, 1990].

To conclude, this phase involves in the following steps:

- Step 1: Collect Good Articles in the same Wikipedia category.
- Step 2: Annotate the editors' editing activity (different types) counts of each editor.
- Step 3: Perform factor analysis on editing activities to obtain editing abilities. We normalize the editing activity counts of the GA of the same Wikipedia category. Then we perform factor analysis to obtain the factors and tag the factors with the different editing abilities.
- Step 4: Use factor analysis, calculate the scores of each editor's editing abilities.

Collaboration Analysis

Previous studies on the collaboration yielding Wikipedia contents proposed many different metrics [Liu and Ram, 2011, Klein et al., 2015]. Our approach to collaboration analysis uses the reversion sequences of editors as they appear in the GA creation process. This research investigated team collaboration with the sequence of editors' appearances in the work process from an article's start until it is accepted as a GA candidate. We rank the editors according to their editing ability scores. We highlight the editors of highest scores of each editing ability in the team and present the rest of the editors as one type of editor – low score editor. We also plot the article

Table 3.1: Editing activity categories

Activity Name	Activity Description
Format	Contribution that alters the appearance or structure of the whole page.
Add information	Additions to topic-related information.
Delete information	Removal of topic-related information.
Clarify Information	Rewording of existing information.
Correct spelling	Correction of spelling.
Correct grammar	Correction of grammar.
Reversion	Reuse of earlier version, which is normally triggered by the undo button on the editing history page.
Fix link	Modification of existing links and changing dead links to correct web address links, including changing the text of link addresses.
Delete link	Removal of existing links.
Vandalism	Entriesactions that damage the page.
Use of talk page	Messages left by editors on it.
Add links to the category	Addition of links in the article's category section.
Add links in content	Addition of them in article's main content.
Add links in reference	Addition of them to the article's reference section.

size over time to give more information for understanding how GA are created. Last, we compare the collaboration of editors for different Wikipedia articles in the same Wikipedia categories and perform analyses to conclude how different-sized teams yield work of similar quality yielding a better understanding of how GA are created.

In phase two, our collaboration analysis method proceeds in the following steps:

- Step 1: Plot the sequences of editors as they appear in the GA rever-

sion process.

Based on the editing ability scores, we identify the highest score editors of each editing ability and plot the reversion sequence of their appearance in the GA creation process.

- Step 2: Plot the editor sequence together with article size changes.
- Step 3: Find the relationship of major editor (highest content-shaping characteristic editor) appearance in the work sequence and article size changes.
- Step 4: Compare collaboration patterns (Step 3) of different teams.

Our method introduces a novel approach to investigate open collaboration that involves in distinguishing the differences among editors and studying their appearance sequence in the quality work creation. In addition, to the best of our knowledge, we are the first research aim to obtain editing ability of editors of Wikipedia and give each editor factor scores based on factor analysis. This method can also be used as an editor reputation system for other future Wikipedia research as well as other open collaboration research in understanding the editor.

3.4 Result Analysis

3.4.1 Overview of the Three Categories of Wikipedia

The data consisted of GA from three categories: "US state parks," "children's books," and "chemical compounds and materials" to demonstrate our proposed approach of studying GA collaboration. The 20 GA of the "US

state parks" of Wikipedia are from the Wikipedia Good Articles list as of April 1, 2018, from the Wikipedia GA category list of "national and state parks, nature reserves, conservation areas, and countryside routes" under "Geography and places." We omitted national parks because its articles are much longer than those of state parks. We omitted national parks because its articles are much longer than those of state parks. We also chose 17 GA (children's books) from the Wikipedia GA category list of "children's books, fairy tales, and nursery rhymes" under "Language and literature" as of December 31, 2018. The category had 43 children's book, fairy tales, and nursery rhymes articles and we choose only "children's books". Note that this category contained a series of books from identical authors: four from American singer Madonna, four from Dr. Seuss, and 12 from Beatrix Potter. The editing patterns and the editors were similar for books from the same authors. To prevent the bias caused by uneven distribution of different books, we selected two books from each from these authors with typical revision patterns to ensure that our data distribution was well balanced. We also used 13 GA of chemical compounds and materials from the Wikipedia GA category list of "chemical compounds and materials" under the "Natural sciences" as of December 31, 2018. We ignored four extra-large GA in these categories because they had over 500 revisions, excessively large in comparison with other articles of this category.

3.4.2 Analysis of "US State Parks" GA

Overview

In this section, we first report the statistics of the 20 GA taken from the "US state park" category as shown in Table 3.2. The overall statistics of article

Table 3.2: Statistics of "US state parks" GA

GA title	Article size (bytes)	Number of activities	Number of editors
Above All State Park	8,166	73	6
Albany Pine Bush	37,899	347	62
Beaver Brook State Park	8,811	56	4
Becket Hill State Park Reserve	6,458	69	7
Brown County State Park	38,444	299	56
Clark State Forest	7,589	150	22
Cloudland Canyon State Park	12,265	257	35
Farm River State Park	8,757	77	7
Haddam Island State Park	9,334	82	4
Haley Farm State Park	9,687	62	5
Hopeville Pond State Park	9,893	55	3
Kayak Point County Park	17,942	77	6
Minneopa State Park	28,623	156	40
Pettigrew State Park	19,425	200	24
Piedmont Park	31,647	462	108
Pomeroy State Park	4,985	45	4
Silver Springs State Fish and Wildlife Area	12,935	137	6
Tualatin River National Wildlife Refuge	21,721	145	12
Vogel State Park	11,973	175	16
White Pines Forest State Park	13,205	148	16

size, numbers of editing activities and editors show large variations. From an observation of collaboration, the smallest team that could create GA in the "US state park" category consists of 4 editors while the largest team have 108 editors. The editing activity counts are proportional to team size. While article size also varies, the variation is much less than that for team size.

Editing Traits: Results

Table 3.3 shows the factor analysis results of the "US state park" GA. F1 to F6 are the six factors extracted as the basic dimensions of editing traits. The numbers in the matrix represent the factor weighting of the editing activities. Numbers closer to 1 indicate a higher loading of the activity in the factor. We use bold format to highlight the factors providing the highest loading for each editing activity. Each factor can be explained according to its higher loading editing activities and the editing traits are represented by factor. Each factor represents the editing traits we extract from the editors of this GA category, they are:

- F1 is content-shaping trait with five activities focusing on content information coverage: format, add information, delete information, add link in content, and add link for reference. Since these activities mainly target content-information enrichment, they are called content-shaping.
- F2 is the copy-editing trait that emphasizes clarifying information, spelling, grammar, and use of talk page. All target writing improvement.
- F3 is indexing trait and links articles to the Wikipedia category index page by adding link in the category.
- F4 is reversion trait.
- F5 is vandalism trait, which covers the activities made to damage articles.
- F6 is link-fixing trait, which is the activity of fixing links.

Both content-shaping trait and copy-editing trait have weights of 0.6 or more

Table 3.3: Rotated factor matrix of the "US state parks" category: 14 editing activities

Editing activity: categories	F1	F 2	F 3	F4	F5	F6
Format	**0.85**	0.40	0.06	-0.01	-0.02	0.14
Add information	**0.90**	0.36	0.10	0.04	-0.01	0.17
Delete information	**0.90**	0.30	0.14	0.08	-0.01	0.16
Clarify information	0.55	**0.71**	-0.08	0.01	-0.02	0.17
Spelling	0.54	**0.69**	0.09	0.09	-0.02	0.09
Grammar	0.38	**0.80**	0.10	0.01	-0.01	0.25
Reversion	0.07	0.04	0.03	**1.00**	-0.01	-0.01
Fix link	0.41	0.32	0.19	-0.02	-0.02	**0.83**
Delete link	**0.66**	0.60	0.09	0.01	-0.01	0.21
Vandalism	-0.02	-0.02	-0.03	-0.01	1.00	-0.01
Use of talk page	**0.63**	0.62	0.00	0.05	-0.02	0.04
Add links to the category	0.17	0.04	**0.97**	0.03	-0.03	0.12
Add links in content	**0.78**	0.28	0.23	0.04	-0.02	0.25
Add links in reference	**0.86**	0.41	0.09	0.06	-0.01	0.16

Table 3.4: Number of editors in different factor score ranges

Editors counts of different factor score ranges	F1	F 2	F 3	F 4	F 5	F6
>6	2	4	3	4	2	2
5 to 6	3	0	0	0	0	0
4 to 5	9	2	0	0	1	3
3 to 4	1	5	6	2	1	2
2 to 3	1	6	7	7	2	11
1 to 2	4	9	28	5	12	19
0 to 1	53	72	73	10	20	59
< 0	370	345	326	415	405	347

for "Delete links" and "Use of talk page" activities. We consider them to be less important in comparison with the other editing activities in representing the trait. So that we did not consider them while naming F1 and F2.

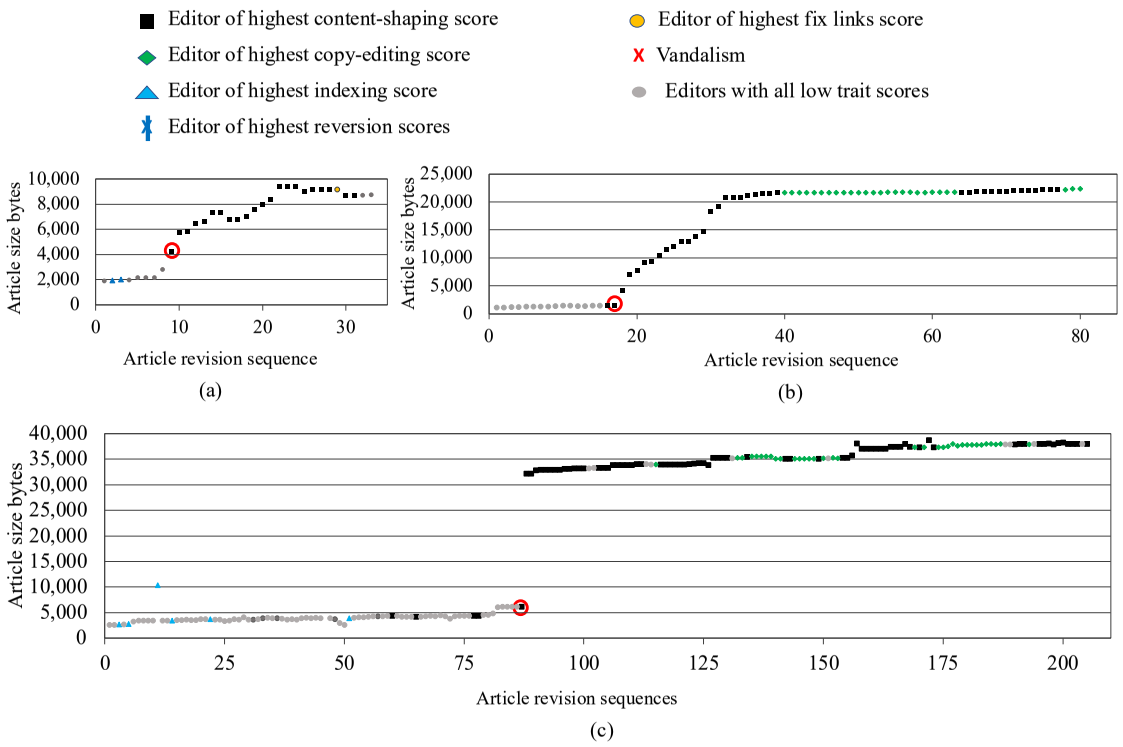


Figure 3.2: Three examples of GA collaboration creation process for the US state park Wikipedia category.

Editor Composition

Table 3.4 shows the normalized factor scores of editors with the global average factor scores of all editors set at 0. From this data, we found that the majority of editors have all editing traits below 0. Only a small number of editors receive high scores in specific editing traits. We identified the highest scored editor of each trait of each article and labeled the remainder as low scored editors. This yields editors' involvement in article creation and a clearer picture of the collaboration can be obtained.

Collaboration Pattern

Three typical examples of GA collaboration creation process from various team sizes for US state park category. (a) Farm River State Park created from 2013-01-22 to 2014-11-20 with 33 revisions, red circle is the 9th revision on 2014-05-23. (b) Tualatin River Nation Wildlife Refuge created from 2006-07-28 to 2009-03-27 with 80 revisions, red circle is the 14th revision on 2009-02-21. (c) Albany Pine Bush created from 2005- 06-12 to 2010-09-23 with 203 revisions, red circle is the 81st revision on 2010-07-19. For Farm River State Park GA in Figure 3.2(a), the whole article creation process took almost two years. The highest content-shaping trait editor only started to work six months before successful GA nomination. This editor also had the highest copy-editing trait score. He/she pretty much finished the work without significant help from others. For the Tualatin River Nation Wildlife Refuge in Figure 3.2(b) the highest content-shaping trait editor only worked for one month and ten days before successful GA nomination. For Albany Pine Bush GA in Figure 3.2(c) the highest content-shaping trait editor only started to work two months, continuously, prior to article evaluation completion. This editor also had the highest copy-editing score of the

team. In the same period, there was an editor with the second highest copy-editing score who helped in GA completion. This editor also had the highest score in link-fixing trait, but due to the significance of the copy-editing trait, we mark this editor as having the highest copy-editing score. We indicate when this editor started to work with a red circle in the figure and report details of the number of revisions and its date.

The three examples in Figure 3.2 indicates the editors with the highest score in content-shaping mainly appeared in the months just prior to GA completion regardless of team size. They worked continuously to increase the article size to pass the evaluation requirements. We also observe that their activities were performed only for a few months regardless the length of the period taken to complete the GA. It can be years before this editor appears. Before this, various editors perform minimal revisions on the articles and article size remained low. They are editors with no dominant editing traits and are the cause of the large differences in team size and activity counts. For the Albany Pine Bush GA, the highest score content-shaping trait editor also had the highest scores for copy-editing, revision, and link-fixing traits.

3.4.3 Analysis of "Children's Book" GA

Overview

In this section, we present the result of the 17 GA we assessed in the Wikipedia category of "children's books". Among the general statistics, we also found large variations in article size, team size with number of editors, and the editing activity counts. As shown in Table 3.5, the smallest team had only one editor while the largest team had 180 editors. The larger

Table 3.5: Statistics of "children's Book" GA

GA title	Article size (bytes)	Number of activities	Number of editors
Crown: An Ode to the Fresh Cut	13,838	94	14
Don't Forget the Bacon!	24,903	133	7
Goldilocks and the Three Bears	18,513	826	180
Horton Hatches the Egg	15,207	479	152
Lucky and Squash	14,661	73	3
Marlon Bundo's a Day in the Life of the Vice President	18,799	255	39
Maurice (Shelley)	15,492	145	5
Puss in Boots	26,018	354	14
Radiant Child: The Story of Young Artist Jean-Michel Basquiat	9,897	32	7
The English Roses	36,662	240	92
The History of the Fairchild	16,144	84	1
The Princess and the Pea	15,135	808	135
The Snowman	13,805	302	3
The Tale of Peter Rabbit	23,279	326	71
The Tale of the Pie and the Patty-Pan	28,330	564	31
Thumbelina	21,317	637	137
Wolf in the Snow	8,140	47	4

teams usually produced more editing activities, but they did not produce the largest GA in this category.

Editing Traits: Results

There are four editing traits found as the basic dimensions for the "children's book" category. We show the factor analysis results in in Table 3.6. Again, numbers closer to 1 indicate activities with higher loading in the factor. We name the factors according to these highest loading editing activities. We use bold format to highlight the editing activity with the highest loading. We

Table 3.6: Rotated factor matrix of the "children's book" category: 14 editing activities

Editing activity: categories	F1	F 2	F 3	F4
Format	**0.88**	0.39	0.17	-0.01
Add information	**0.88**	0.38	0.20	-0.01
Delete information	**0.61**	**0.61**	0.06	-0.01
Clarify information	**0.82**	0.49	0.15	-0.01
Spelling	0.25	**0.87**	0.04	-0.02
Grammar	0.46	**0.67**	0.29	-0.01
Reversion	0.25	0.12	**0.95**	-0.01
Fix link	**0.78**	0.47	0.18	-0.01
Delete link	**0.81**	0.46	0.23	-0.01
Vandalism	-0.02	-0.02	-0.01	**1.00**
Use of talk page	**0.87**	-0.01	0.07	-0.02
Add links to the category	**0.78**	0.42	0.24	-0.01
Add links in content	**0.86**	0.39	0.22	-0.01
Add links in reference	**0.89**	0.36	0.15	-0.01

found the content-shaping trait of children's books also covered link-related activities as well as writing improvement, so only four editing traits were identified. Each factor demonstrates the editing traits we extracted from the editors of this GA category. They are:

- F1 is content-shaping trait that focuses on content-information coverage: format, add and clarify information; fix and delete links; add link in the category, content, and reference; and use talk page. These activities mainly target content-information enrichment, so they are content-shaping.
- F2 is a copy-editing trait that emphasizes spelling and grammar.
- F3 is reversion trait which is the activity to reverse previous editing activity.

Table 3.7: Number of editors in different factor score ranges

Editors counts of different factor score ranges	F1	F 2	F 3	F 4
> 6	7	8	5	3
5 to 6	2	0	1	1
4 to 5	4	2	2	7
3 to 4	2	4	1	3
2 to 3	5	1	1	6
1 to 2	9	13	37	20
0 to 1	74	134	62	50
< 0	817	758	811	830

- F4 is vandalism trait that performs edit activities that damage the articles.

For the editors of children's book GA, we found a large concentration on editing activities for editing trait F1. This indicates the content-shaping trait covers 10 editing activities in GA creation. While there is a copy-editing trait that improves writing, the rest of the traits are vandalism and the activity to fix vandalism with "revision".

Editor Composition

Table 3.7 shows the normalized statistics of the factor scores of editors with the global average factor scores of all editors set at 0. From this data, we found that the composition of editors largely consists of editors with all editing traits below 0. Only a small number (less then 10 %) of editors had high scores in specific editing traits. We then identified the highest scored editor of each trait of each article and labelled the rest as low scored editors.

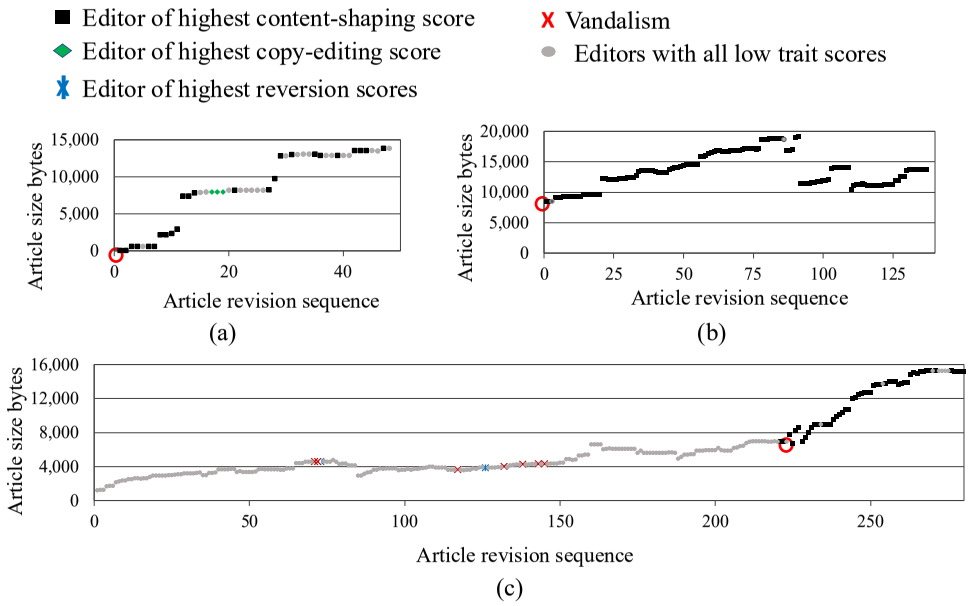


Figure 3.3: Three examples of GA collaboration creation process for the children's books Wikipedia category.

Collaboration Pattern

We again focus on editors with highest scores of each editing trait and plot the editor's sequence in the revision history of article creation. In Figure 3.3, we give three typical examples of GA of different creation time and team size to represent the collaboration that occurred when creating "children's book" GA. (a) Crown: An Ode to the Fresh Cut created from 2018-09-21 to 2018-12-09 with 48 revisions. Red circle is the 1st revision on date 2018-09-21. (b) Snowman created from 2013-01-22 to 2014-11-20 with 137 revisions. Red circle is the 1st revision on date 2013-01-22. (c) Horton Hatches the Egg created from 2004-11-03 to 2013-10-25 with 278 revisions. Red circle is the 219th revision on 2013-06-30.

Figure 3.3(a) is the Crown: An Ode to the Fresh Cut article. The highest content-shaping trait editor started the article and was responsible for increasing the article size up to successful GA nomination. The whole article creation process only took two and half months. Figure 3.3(b) is the creation of the Snowman GA. The article was done by one editor with high scores in all editing traits. Figure 3.3(c) is the creation of Horton Hatches the Egg GA, the high content-shaping trait editor started to work only four months before successful GA nomination even though the whole GA creation took nine years. This high score content-shaping trait editor also had the highest copy-editing trait score. We can also conclude that in the months prior to GA nomination, high score content-shaping trait editors increased the article's size and diligently worked to secure a successful GA nomination. Those editors had little help from the others in finishing the GA. We also found a few children's book GA were created by just one editor.

Similar to the "US state parks," GA with large teams have long creation

periods and high numbers of editors involved. The editors of overall low editing trait scores performed scant editing activities over a long period of time. These low score editors are the cause of the large differences in teams, such as team size, and revision and editing activity counts. Editors with only high copy-editing trait are not so important in creating some GA of this category. In some cases, the high content-shaping trait editors also had the high copy-editing trait score. Again, we see the high content shaping editors mainly involved in the GA creation at the end for large teams. For the others, the high content-shaping editors worked from the start to finish the GA gradually. The involvement from the others is minimum and normally do not increase the article size. In addition, the involvement of high copy-editing editors appears to be less important in compare with the US state park GA. They appeared less often in overall GA creation and were not really involved in the periods prior to GA nomination.

3.4.4 Analysis of "Chemical Compounds and Materials" GA

Overview

In the "chemical compounds and materials" category, we examined 13 GA. Although there was 17 GA in total in the category, we did not include the four articles that had over 500 revision events as they were too large to annotate manually. The general statistics of this category also showed a large variation in team size and number of editors and article size as shown in Table 3.8. The smallest team contained ten editors while the largest team consisted of 231 editors. Again, the larger teams usually produced more editing activities, but they did not produce the largest GA in the category.

Table 3.8: Statistics of "chemical compounds and materials" GA

GA title	Article size (bytes)	Number of activities	Number of editors
Aluminum chloride	11,283	66	17
Benzylpiperazine	30,755	180	44
Boron nitride	38,316	571	136
Calitoxin	11,359	150	12
Compounds of berkelium	18,279	70	12
Copper(I) chloride	10,674	22	10
CS GA	20,689	310	62
Hexamethylbenzene	41,855	494	51
Iron (III) chloride	10,993	98	27
Oxazolidine	21,865	207	15
Silicon nitride	23,964	295	58
2,3,7,8-Tetrachlorodibenzodioxin	23,645	262	68
Zinc oxide	38,214	889	231
Aluminum chloride	11,283	66	17
Benzylpiperazine	30,755	180	44

Editing Traits: Results

In this category, we identified four factors, F1 to F4, as the editing traits. Again, numbers closer to 1 indicate editing activities with higher loading and we highlight them in bold as shown in Table 3.9. We named the factors and explained them below.

There are four editing traits found as the basic dimensions for the "chemical compounds and materials" category. We show the factor analysis results in Table 3.9. Again, numbers closer to 1 indicate activities with higher loading in the factor. We name the factors for these highest loading editing activities. We use bold format to highlight the editing activity with the highest loading. We found the content-shaping trait of children's books also covered link related activities as well as improved writing, so only four editing traits

were identified. They are detailed below. Each factor demonstrates the editing traits we extracted from the editors of this GA category; they are:

- F1 is a content-shaping trait with nine activities that focus on content-information coverage: format, add information, delete information, clarify information, fix link, delete link, add link in the category, add link in content, and add link in reference. These activities mainly target content-information enrichment, so they are content-shaping.
- F2 is a copy-editing trait that emphasizes spelling and grammar.
- F3 is reversion trait which is an activity to reverse editing activities performed by the previous editor.
- F4 is vandalism trait which performs edit activities that damage the article.

The factor analysis results in Table 3.9 show, similar to the "children's book" category, that the content-shaping trait also covers link-related activities as well as improving the writing. In addition, the content-shaping trait also covers link-related activities such as: fix link, delete link, add link so we did not extract any link-related trait as a separate item. The copy-editing activity only covers spelling and grammar and but not clarifying information which is different from the copy-editing activities of "US state park" category. We also identified vandalism and reversion which reverses vandalism activity as editing activities. So, we identify only four editing traits in this Wikipedia category.

Editor Composition

Again, we found most editors in this category have all editing trait scores

Table 3.9: Rotated factor matrix of the "chemical compounds and materials" category: 14 editing activities

Editing activity: categories	F1	F 2	F 3	F4
Format	**0.93**	0.17	0.11	0.01
Add information	**0.93**	0.13	0.12	0.02
Delete information	**0.89**	0.14	0.08	0.01
Clarify information	**0.88**	0.26	0.09	-0.02
Spelling	0.25	**0.71**	0.08	0.01
Grammar	0.07	**0.80**	-0.09	-0.03
Reversion	0.16	0.01	**0.98**	-0.01
Fix link	**0.88**	0.19	0.06	-0.03
Delete link	**0.85**	0.30	0.06	-0.01
Vandalism	-0.02	-0.02	-0.01	**1.00**
Use of talk page	0.52	0.58	0.12	0.00
Add links to the category	**0.92**	0.13	0.08	-0.03
Add links in content	**0.90**	0.23	0.07	-0.02
Add links in reference	**0.87**	0.14	0.03	-0.01

Table 3.10: Number of editors in different factor score ranges

Editors counts of different factor score ranges	F1	F 2	F 3	F 4
>6	5	5	6	6
5 to 6	3	3	0	0
4 to 5	1	3	0	4
3 to 4	5	4	6	1
2 to 3	3	3	2	3
1 to 2	10	18	15	16
0 to 1	101	58	45	24
< 0	606	640	660	680

below 0. Only a few editors received high scores in specific traits (less than 10%) as shown in Table 3.10. We observe the collaboration with focus on the highly scored editors of each trait and do not distinguish overall low editing trait score editors.

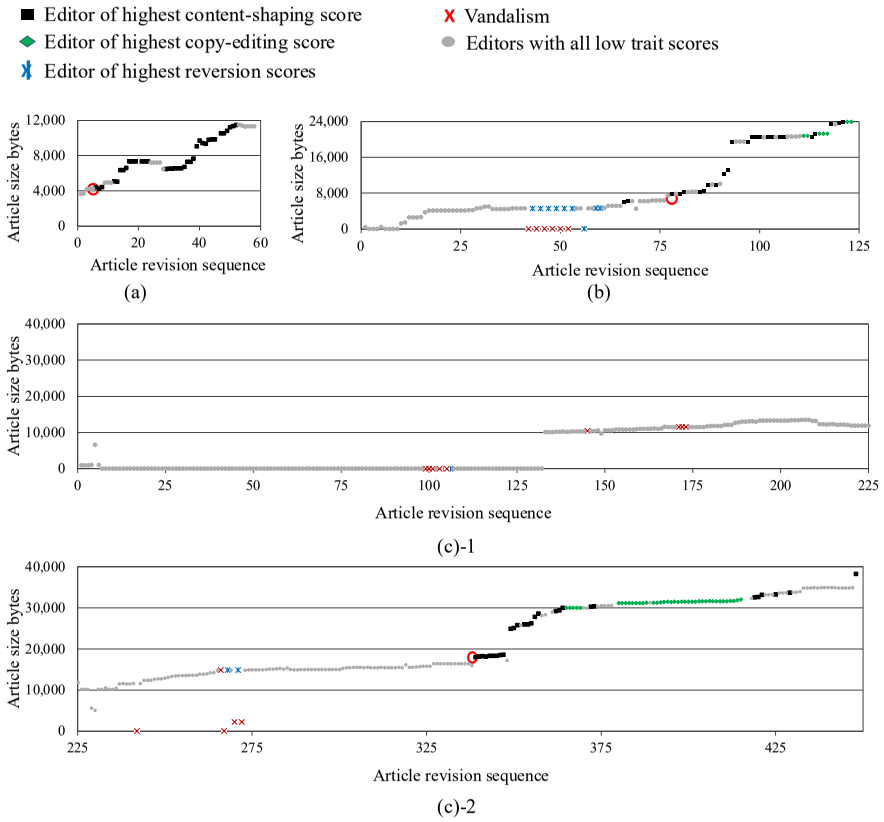


Figure 3.4: Three examples of collaboration process for GA on chemical compounds and materials of Wikipedia category.

Collaboration Pattern

In this section, we report the collaboration characteristics observed in GA for chemical compounds and materials. With focus on the editors with highest factor scores in each editing trait, we plot the editor's sequence of appearance in the revision history of article creation. We again mark the editors with the highest scores in the four editing traits of the team. Editors who have all traits scores all below 0 are tagged as low score editors. We also red circled the revision at which the highest content-shaping editor started to work and show the sequences of active editors as they appear in the GA creation process as well as information of the revision number and date. Figure 3.4 shows three typical examples of different creation periods and team size to demonstrate how GA of "chemical compounds and materials" were created. (a) Calitoxin created from 2014-10-12 to 2014-11-21 with 58 revisions. Red circle is the 6th revision on 2014-10-12. (b) Silicon nitride created from 2005-12-05 to 2009-08-14 with 123 revisions. Red circle is the 78th revision on 2009-04-26. (c) Zinc oxide created from 2004-03-09 to 2009-03-09 with 448 revisions. (c)-1 is the first 225th revision and (c)-2 is from the 226th revision to 448th revision. Red circle is the 339th revision on 2009-01-14.

For Calitoxin, which has a short revision history, the highest content-shaping trait editor started to contribute to the article the same day the article was created. He/she continuously worked on the article and also increased the article size to meet GA nomination requirements. The second highest content-shaping trait editor actually did the last fix before GA nomination although he/she is indicated by the grey circle. The whole article creation process took only one and half months with 58 revisions. As for Silicon

Nitride, the median length creation period example, the highest content-shaping trait editor only started to work a few months before successful GA nomination at the 78th revision; there was a strong copy-editing trait editor as well to help in the same period. This period lasted for less than four months. For Zinc oxide, GA creation took five years with 448 revisions. Due to the long revision history, we split the chart into two for clearer visualization. Despite the long revision history, the high content-shaping trait editor only started to work two months before successful GA nomination. A strong copy-editing trait editor was present in the same period to help.

Although there were also editors with low scores in the same period, the final revision was performed by the strong copy-editing trait editor who increased the article size to GA level. In the years before these two months, editors with weak editing activities and low scores in all editing traits contributed by making over 400 revisions These editors were the cause of the differences in team size and the high number of revisions for these articles.

Similar to the "US state parks" and "children's books," the GA of chemical compounds and materials category had a highest scored content-shaping trait editor who increased the article's size and worked mainly prior to GA nomination to secure its acceptance. Those content-shaping trait editors also gradually worked to increase the article size up which is similar to the "children's book" collaboration pattern. High scored copy-editing trait editors appeared towards the end of GA creation process to improve the writing as seen in the "US state parks" GA.

To conclude our findings on the three Wikipedia categories researched. In general, we found there is burst of article size growth prior to the GA quality level nomination. This mainly is due to a high scored content-shaping

trait editor during this period. We found this trend existed in all three categories regardless of article size and the number of editors involved in article creation.

We also found prior to the appearance of the high scored content shaping editor, GA creation involved editors with all low scored traits. The large variation in team size is caused by the different number of editors with low scored trait of the article. Large teams have a high number of editors with low scores in all editing trait as shown in Figure 3.2-3.4 and small teams have a lower number of these editors. Regarding the variation in article creation period, it is mainly caused by the time when only the low scored editors contribute. In Figure 3.2-3.4, we report the period of the whole GA creation and the date when the high scored content-shaping editors started to work. It shows, for an article, it can be years of only low trait score editors involved until a high scoring content-shaping trait editor gets to work and completes the GA in a few months.

3.5 Discussion

In this research, we chose three Wikipedia categories and assessed the GA in the category to elucidate topical differences in the collaboration activities that yielded high quality articles. We considered the "US state parks" and "children's book" are general topics for which editors do not need to have deep knowledge. We also observed the collaboration of creating chemical compounds and materials GA and considered it is a science topic so contributors would require deep knowledge of the topic.

In general, there exists a burst in article growth prior to the Good Articles

(GA) quality level nomination. This is done by the highest scored content-shaping trait editors, and they work continuously to finish the GA. We found evidence of this phenomenon in all three categories regardless of article size and number of editors. In addition, all low scored trait editors are active mainly at the beginning of GA creation.

As regards the editor traits, there is some similarity as well as differences across these three topics. There are four traits found in all three categories we studied. Content-shaping trait focuses on enriching article content while the copy-editing trait focuses on improve writing. Reversion trait focuses on correcting vandalism and there is also vandalism trait representing editor conduct that damaged the article. We extracted six editing traits from the "US state parks" category but only four editing traits from the children's book and chemical compounds and materials categories. The difference is that the former includes editing activities related to links to other web-sites for article content yielding the indexing trait and link-fixing trait. For the "children's books" and "chemical compounds and materials" categories, the editing activities related to adding link and fixing link are subsumed by the content-shaping trait. Our factor analysis method can clearly identify the different editing traits present in different Wikipedia categories and the editing activities present in each trait.

For the children's book is a topic requires no deep knowledge, it seems not to have editors focuses on editing activities other than content-shaping or copy-editing. For the science topic we study in the research, it also has four editor traits extracted from editors' editing activities with main traits of on content-shaping and copy-editing. Then the reversion trait and vandalism trait are pair up.

For GA collaboration in three categories, we can see that regardless of team size, there appears to be a three-phases development process. The "Initial Phase" involves editors with scant editing activities; these editors have no one dominant editing trait. This phase may not exist if only one editor is active in GA creation. It can also stretch for years and result in different numbers of editors being involved. This is followed by the "Growth Phase" in which a high scored content-shaping trait editor becomes involved and article content is increased by this editor. Last, in the "Completion Phase," the "Growth Phase" editor along (particularly in the children's book category) or sometime with another editor of high copy-editing trait score works together continuously to secure GA acceptance. This last phase normally only takes only a few months.

One potential implication of this finding is to call upon editors with high content-shaping trait scores to work on more articles. This can be achieved by a mechanism that identifies the editors with high content shaping trait scores and incentivize them with an award badge or status in the Wikipedia community to encourage them to contribute.

3.6 Conclusion

Although open collaboration provides unlimited opportunities to society, it also faces some challenges. A good example is Wikipedia. While its importance to modern society is beyond question, the quality issue has long been considered problematic. How quality articles are created by different teams of editors remains unclear. In addition, the literature ignores the variation among topics and the collaboration activities in creating quality articles of

different topics.

This research considers GA is an important quality level and applies an approach that allows an understanding of GA collaboration. The proposed approach first distinguishes editor in terms of editing ability and then identifies their involvement sequence in the GA creation process. The approach can be applied to different Wikipedia categories to gain better understanding of the collaboration needed in creating GA of certain topics. This research chose the GA of two general topics and a science topic from Wikipedia categories to demonstrate the usefulness of our research approach. We first found differences in editing abilities among editors of different topics. All three categories exhibit the four basic abilities of content-shaping, copy-editing, reversion, and vandalism.

We found there is common collaboration pattern in most of the GA examples analyzed in this research. That is, prior to achieving GA nomination, an editor with strong content-shaping ability expands the article and continuously works on it. We also found that GA creation relies only slightly on the sheer number of editors as most who contribute little to the editing activities. This finding agrees with the finding of [Zhang et al. 2017] who found a concentration in workload immediately prior to GA nomination. Our approach can identify the strong content-shaping editors who work continuously during this period. It also discovered some difference in the collaboration process for creating GA in different subcategories. In the "children's book" subcategory, some GA were produced by just one editor. In some of the "chemical compounds and materials" articles, there were more than one high score "content-shaping" ability editors and only the editors who appeared later could guarantee GA acceptance.

The findings from this study can be used to make recommendations in creating GA for the three Wikipedia subcategories examined. That is to secure a high content-shaping scored editor to work continuously to a non-GA before GA evaluation to guarantee GA acceptance.

To conclude, our research uses Wikipedia Good Articles as a case study on understanding how open collaboration can create quality work. We demonstrate that our approach offers more understanding of how quality content is achieved through open collaboration that calls people online to work together. The result of this research has been published in [Chou et al., 2020] and [Chou et al., 2022a].

Chapter 4

Discovering the Collaboration Patterns in Good Wikipedia Articles

In this section, we report our continued effort in providing a solution to create more quality knowledge content from open collaboration. Here we continue to use Wikipedia Good Articles as a case for the quality knowledge content. In this research, we propose a novel method of identifying collaboration patterns that lead to articles of Good Articles quality level in the same Wikipedia category. These patterns can act as a reference for the generation of more GA for these Wikipedia categories.

4.1 Introduction

Online collaboration platforms are attracting considerable attention as the new generation systems for open knowledge production and organization [Dai et al., 2013]. Wikipedia is an example of such platforms and has grown to become the largest free encyclopedia [Giles, 2005]. Volunteer editors possessing different levels of knowledge and expertise are working collaboratively to produce quality articles. However, the quality of Wikipedia articles is not always guaranteed and there is an urgent need for a constant assessment of such quality [Wikimedia, 2022a].

In order to analyze how quality is preserved within an article, earlier studies explored quantities such as the editing counts, file sizes, and number of editors [Kittur et al., 2009, Kane, 2011, Liu and Ram, 2011, Robert and Romero, 2015]. Recent research efforts attempted to look at the interaction patterns of the editors. For instance, [Ren and Yan, 2017] studied a combination of different types of editors for different quality levels. [Chou et al., 2020] studied the order in which different editors work on Good Articles (GA). [Lin and Wang, 2020] investigated how different combinations of editors achieve better quality. [Zhang et al., 2017] focused on the collaboration dynamics such as the number of participants and the concentration of workload over time. Another methodology relies on the quantification of integrated information within the Wikipedia articles [Engel and Malone, 2018] and shows that when applied to the editing activity, such measure correlates with the collective intelligence of groups [Woolley et al., 2015] . Another study uses the same measure to identify the hierarchies of editors that lead to higher interactions and improved article quality [Hadfi and Ito, 2021].

Departing from the same motivations and the need to identify the optimal editing patterns of Wikipedia articles, this research aims at gaining a deeper understanding of the open collaboration model of Wikipedia by proposing a novel approach to identify and analyze the collaboration patterns within one category of good quality articles. We deliberately chose GA as the research subject due to their prevalence as a credible source of knowledge. The extract the collaboration patterns of Wikipedia GA articles can be a reference to create more GA of the same category.

The proposed method first uses two important components identified in our previous work to study the collaboration patterns used in creating GA [Chou et al., 2021a]. These components include the editors themselves and the development of articles as represented by the evolution of article size. It is known that the combination of different editors can impact article quality. It is therefore important to distinguish among editors in studying collaboration patterns. Here, we use the same approach in the previous chapter to apply factor analysis to each editor's editing activities and thus obtain "editing traits." Then, each editor is scored on each trait to indicate his/her strength in that particular editing trait. In addition, the editor's involvement in the article-creation process is an important element. One work [Zhang et al., 2017] studied the editing workload over time and found that the collaboration dynamic changes with the article creation process. In Chapter 3, we also use combination the time series of article sizes in GA as well as the editors with their trait scores. We found that certain types of editors were involved in different phases of GA creation.

Studying the involvement of editors in the GA-creation process can unravel the collaboration patterns used. After finding the collaboration pattern of

each article, the next necessary step is to infer reasons for particular patterns achieving assignment to the GA level in the article's Wikipedia category. Because the article-creation period differs, we propose using the dynamic time warping (DTW) method [Senin, 2008] to identify clusters of articles that share a similar sequence of article sizes at all points of revision in creating the article and, moreover, to obtain the mean sequence of each cluster. The number of clusters is obtained through hierarchical clustering and the Elbow clustering algorithm [Shi et al., 2021]. Finally, we use the DTW Barycenter Averaging algorithm (DBA) [Petitjean et al., 2011] to obtain the averaged time series of each cluster. Finally, to observe the collaboration pattern, we split articles into growth (G), decline (D), and plateau (P) phases by calculating the differences in article size over time. Collaboration is represented using the distribution of the editors' trait scores at each phase over the article's lifetime.

We demonstrate the proposed approach using three Wikipedia categories. We chose the "US state parks" sub-category under the "Geography and places" category along with "children's books" from the "children's books, fairy tales, and nursery rhymes" category to represent two general topics. We chose "chemical compounds and materials" from the "Natural sciences" category to represent a science topic. We followed the category-selection approach of our previous research of Chapter 3 by using GA data as well as the editor-trait data.

4.2 Related Work

The collaboration methods among editors in creating quality articles of Wikipedia have also been the target of several studies. For instance, it was found that in order to obtain quality output, editors need to coordinate their efforts [Kittur and Kraut, 2008, Kittur et al., 2009, Stvilia et al., 2008]. [Kittur et al., 2009] found that more editors might not be more efficient in creating quality work due to coordination costs. [Kane, 2011]found that the volume of editing activities is not related to article quality but to the amount of effort spent in shaping the articles, which has a positive impact on the article quality. [Klein et al., 2015] studied the collaboration structure and found that more editors can yield better article quality in some categories, but more editors per article can also create devalue. [Engel and Malone, 2018] found that large groups of editors having higher variance in editing activity counts produced higher quality. Another study used the same measure to identify the hierarchies of editors that contribute to article quality [Hadfi and Ito, 2021]. [Lin and Wang, 2020] also found that the greater the proportion of core members, defined as the people who frequently participate in editing, the more likely it that higher article quality will be achieved. These studies mainly looked at the collaboration patterns of different Wikipedia quality levels. [Zhang et al., 2017] investigated the behaviors of groups of editors and their dynamics over time. They found that in a short period immediately prior to GA nomination, the editing activities increased and that such activity involved just a few members of the original team. There is therefore a research gap on the analysis of the collaboration patterns in the same Wikipedia quality level. To bridge this research gap, we present a novel approach that finds the collaboration patterns in GA Wikipedia articles.

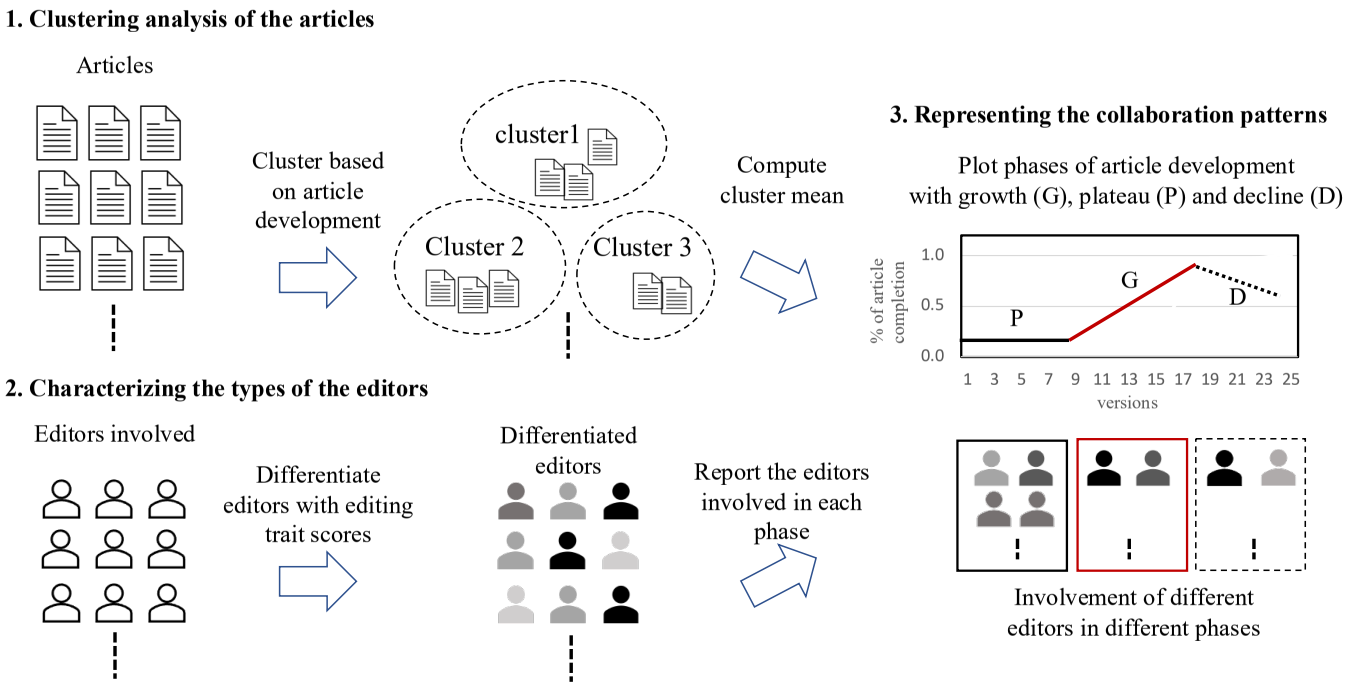


Figure 4.1: Proposed method for finding collaboration patterns

4.3 Method

We use an exploratory approach to find the common patterns in collaborative article creation. We rely on two components to assess and visualize such collaboration: the editors and the evolution of article size. For the editors, it is necessary to distinguish between their types if we are to understand how they collaborate. Here, we adopt the findings of our previous chapter, where editors are given editing-trait scores according to their expertise used in GA editing. For the second component, the evolution of article size, we propose using the percentage of article completion at each revision. These percentages are extracted from the Wikipedia editing history page. Then, we cluster the articles with similar developmental patterns together and use the mean sequence of each cluster to represent the articles gathered in that cluster.

After that, we split the article evolution time series into the phases. For this we perform a number of calculations on the article size time series. We start by calculating the differences between the article size values over time, we then capture the sign of the change as a growth (G), decline (D), or as a plateau (P). Consecutive changes of the same type are compacted. For instance, a sequence of the form "GGGDPPG" is transformed into "GDPG". We assume that the change in the article size is only meaningful if it is beyond 5% of the overall article size. This condition smooths out any internal fluctuation due to vandalism and reversions. We finally represent the collaboration with the distribution of the editor trait scores in each phase of the article development cycle. The overall process of the methods involved in the three parts is shown in Figure 4.1. We provide a detailed explanation of each part below.

4.3.1 Clustering Articles with Version Size Development

To find the collaboration patterns that predominate among the GA of a particular Wikipedia category, we perform cluster analysis and extract the means of the article size time series of each version. This analysis groups articles with similar article size evolution to obtain the mean of each time series cluster. Such an analysis can be used to explain the collaboration patterns behind the creation of GA of the same Wikipedia category. To this end, we extract the data representing the file size of each version in the revision history from when an article is first created until its GA status is granted. A good article is defined by the sequences of its size over discrete revision time steps. That is, the article is represented by an incremental sequence of data points defined as $[S_0, S_1, \ldots, S_i, \ldots, S_n]$, where data point S_i is the size of the article in its ith version. The integer n is the version count up to the moment when the article obtains its GA nomination. The sequence is normalized using percentages, where S_0 is close and/or equal to 0% and S_n is close and/or equal to 100%.

There are three steps in the identification of a common article development cycle: clustering the articles, determining the number of clusters, and extracting the mean value to represent each cluster. Here, we give a detailed explanation of each step.

Step one, clustering the articles.

Since we need to cluster the articles based on a series of data points across time, we rely on algorithms that manipulate and analyze time series data. Time series analysis is a statistical method that analyzes an ordered series of numerical data points. Here, we use dynamic time warping (DTW) to cal-

culate the distance between each series of data points. DTW is a common method that can measure the similarity between two temporal sequences that may vary in length [Senin, 2008]. The method dynamically compares data points in the time series and creates a Euclidean distance matrix to find the minimum distance of the two time series patterns. Then, the set of time series can be clustered according to their similarity. Finally, hierarchical clustering is performed to cluster the series. Hierarchical clustering does not require a predefined number of clusters. The result is a tree-based representation called a dendrogram. The dendrogram shows the height of each link, which indicates the distance between objects or clusters. The main use of distances is to determine the number of clusters.

Step two, determining the number of clusters.

The number of clusters can be determined by the elbow method, which is one of the main methods for determining the optimal number of clusters in a dataset for analysis purposes [Shi et al., 2021]. The elbow method plots the within-cluster variance as a function of the number of clusters and picks the "elbow" of the curve as the number of clusters to use. It is intuitive in that increasing the number of clusters could improve the "fit" of the model (less within-cluster variance). The elbow point is determined once the within-cluster variance becomes stable. In our method, we use the highest distances at each tree depth and then calculate the distance differences between two consecutive tree depths. The difference is used to determine the optimal number of clusters that yields the minimal distance difference. Other approaches use indices that allow comparison of within-cluster distances [Kodinariya and Makwana, 2013].

Step three, calculate the average sequence.

The DTW Barycenter Averaging (DBA) algorithm [Petitjean et al., 2011] is run on the GA of the same cluster. DBA mines the article-development sequential datasets and iteratively calculates a potential arbitrary average sequence. DBA uses the minimized summed and squared distances of DTW distances from the average sequences. The length of the mean sequence of DBA results would be the same length of the longest sequence among the sequences involved. In our case, we run DBA on the normalized article-size sequences defined in subsection 4.3.1. We then use this average sequence to represent the article-size development patterns that exist within the studied GA.

We used the DTAI distance library to perform DTW and hierarchical clustering [20]. After performing the above three steps, we obtained the mean sequence of article development for each cluster. Based on these mean sequences, we can study and analyze the collaboration patterns.

4.3.2 Characterizing the Types of Editors

Editors contributing to Wikipedia articles have different skills and knowledge. Their editing activities can, therefore, be used to differentiate them. However, it is difficult to categorize editors based simply on the counts of their various editing activities. We thus borrow a method from psychology, namely the "Cattell-Horn-Carroll model" (CHC) of human cognition [Carroll et al., 1993]. to obtain the editing traits. The CHC model performs factor analysis on the datasets of psychological tests, school marks, and competence ratings to produce a taxonomy of human cognitive abilities. Similarly, we use editing activities to obtain editing traits. Factor analysis finds the correlation among given data variables that represent the editors

and reduces them into fewer factors. Such factors can subsequently explain the data and any potential observation. The factor scores in CHC can represent an individual's cognitive ability according to each factor. The global mean of the factor scores is often set to zero, so a positive score indicates a stronger trait for a certain cognitive ability while a negative score indicates a weaker trait for that same cognitive ability.

Similarly, we can obtain a taxonomy of editor traits from the editing activities. Based on the factor scores of each editing trait, it is possible to differentiate between the editors. We adopt a categorization of the editing activity as annotated in our previous work [5] and as listed in Table 3.1 of Chapter 3.

4.3.3 Representing Collaboration Patterns

To clarify the collaboration patterns that exist in the GA of a particular Wikipedia category, we split the article evolution time series into phases of growth (G), plateau (P), and decline (D). To identify the phases that a GA article goes through, we perform calculations based on the article size time series. That is, we start by calculating the differences in article size over time, then capture the sign of the change as a growth (G), a decline (D), or a plateau (P). Consecutive changes of the same type are then compacted. For instance, a sequence of the form "GGGDDPPPPP" is transformed to "GDP." The calculation of the article-development phase is used to help in the observation of article development and composition of editors in different phases. In this research, we assume that a change in the article size is only meaningful if it exceeds 5 % of the overall article size in a single revision or continuous revisions [4]. We also smooth out any internal fluctuation

if the size reverts within the next five versions.

After obtaining the phase transition of each cluster, we map the phases to each article to obtain the group of editors involved in each phase. Then, we derive the composition and distribution of each trait score of the editors involved in the same phase.

4.4 Experiment and Results

In this section, we use the proposed method to find the collaboration patterns of GA of the researched Wikipedia category. We used 20 GA of "US state parks," 17 GA of "children's books," and 13 GA of "chemical compounds and materials" in this dataset. We omitted GA of over 400 revisions because those articles are much longer in comparison with other GA of the same category. We report each figure with phase patterns (letters or labels) determined by the evolution of article size (DBA-averaged time series). The composition of editors involved in each phase is shown in the right side of figures. In addition to reporting the average trait scores of the editors involved in each phase, we report the composition of the editors with the distribution of trait scores and visualize them as box plots. The box plot gives information of the data distribution with their minimum, maximum, mean, first quartile (Q1), and third quartile (Q3) of the trait scores of the editors involved in each phase. Q1 and Q3 of the distribution are represented using a distribution box for easier visualization. This is useful for indicating whether a distribution is narrow or wide. In addition, the outliers are also shown with dots in the figures to give more information on the composition of editors in each article-development phase. We mainly

focus on the mean and the plot box's location to indicate the types of editors involved and draw a comparison between their score levels in different article-development phases. We also use use traits scores around or above the value of 2 to indicate higher scores based on observation of the data.

4.4.1 US State Parks

In this section, we report results of using the proposed method to find the collaboration patterns in US state parks GA. Six editing traits of US state parks were extracted from 20 GA:

1. A content-shaping trait with five activities focusing on the coverage of information content: format, information addition, information deletion, added link in content, and added link for reference. Since these activities mainly target information-content enrichment, they are labeled content-shaping trait.
2. A copy-editing trait. This emphasizes any clarifying information, spelling, grammar, and use of a talk page. All of these activities represent writing improvements.
3. An indexing trait of linking articles to the Wikipedia category index page.
4. A reversion trait of reversing (undo) the activities of previous editors.
5. A vandalism trait, which marks the activity of vandalism.
6. A link-fixing trait, which characterizes the activity of fixing broken links.

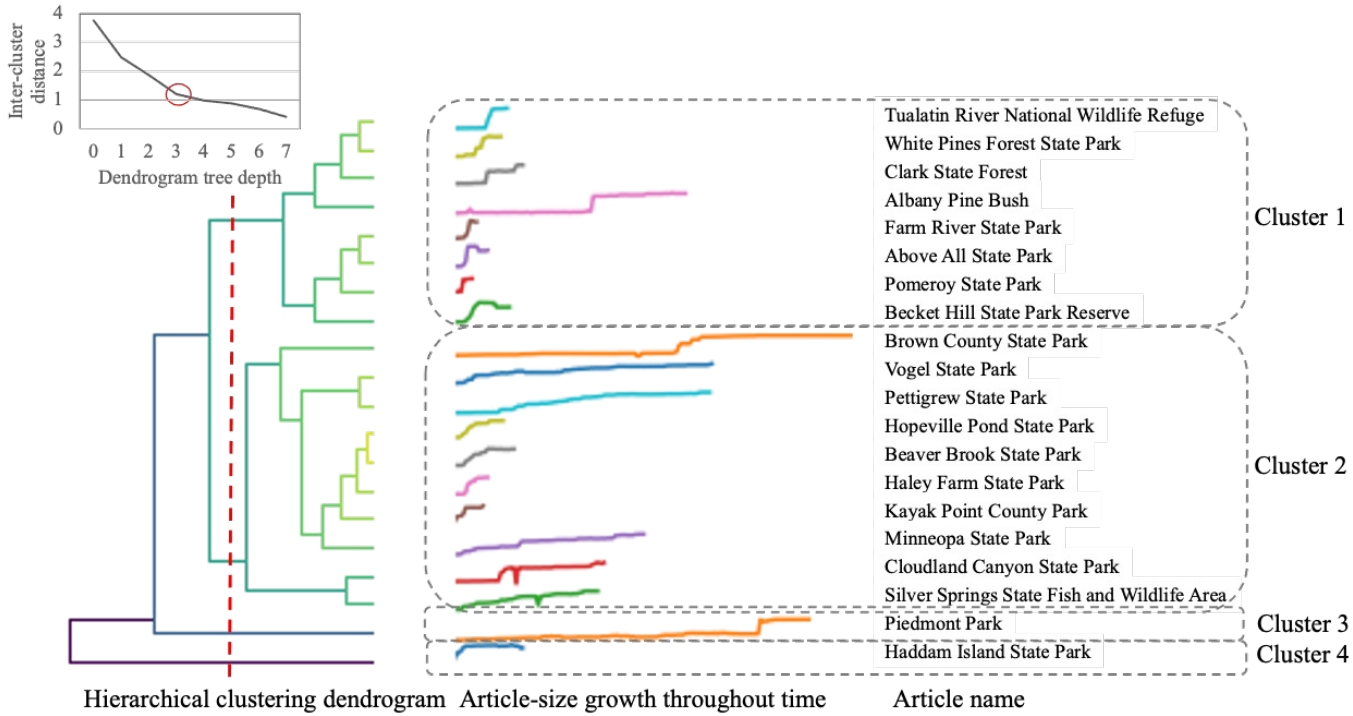
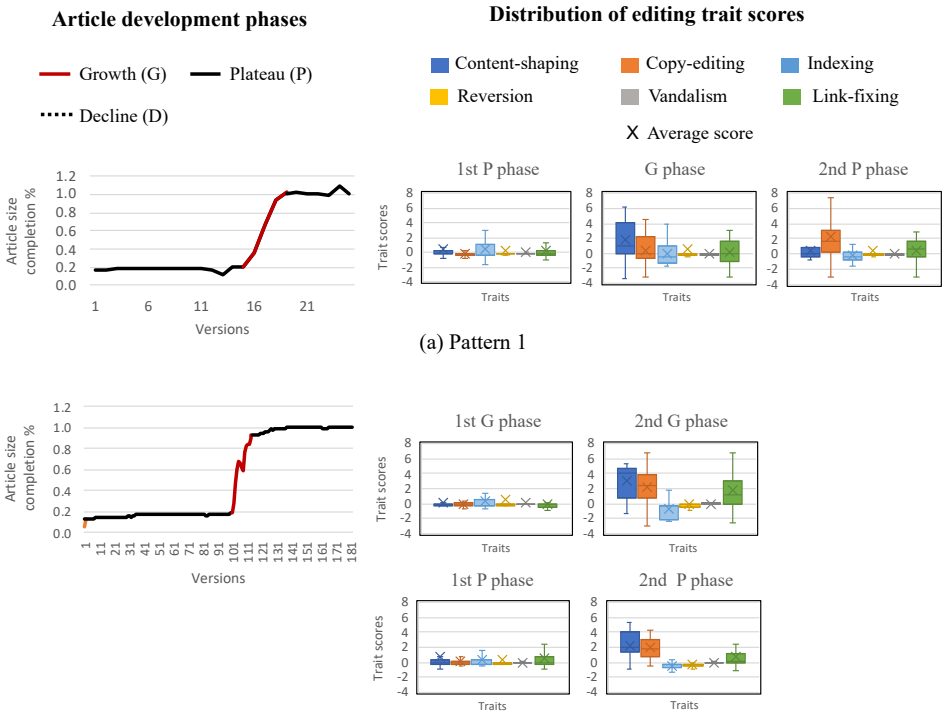


Figure 4.2: Clustering results of the US state parks GA.



(a) Pattern 1

(b) Pattern 2

Figure 4.3: Collaboration patterns of US state parks GA.

Clustering Result

Here we report the clustering result of the article size changes along the time series with version updates in time. The distance of each article is calculated based on DTW, and articles with least distance are clustered together to form a dendrogram as shown in Figure 4.2. The optimal number of clusters identified by the elbow method is 4 at tree depth 3 with inter-cluster distance of 2.22. If we increase the tree depth to 4, the inter-cluster distance is only reduced by 0.48, which we consider insufficient for improving the clustering significantly. Therefore, we determined 4 clusters in the US state parks GA.

Collaboration Pattern Results

In this section, we report the collaboration patterns that exist in the US state parks GA. Four different collaboration patterns were found, and here we report the two main collaboration patterns illustrated in Figure 4.3, omitting the collaboration patterns of clusters 3 and 4 as they consist of only one article each. We found that the distribution of articles in each pattern category is uneven. Patterns 1 and 2 together account for 18 of the 20 articles. As mentioned, Patterns 3 and 4 have only one article each (not reported here), so we consider Patterns 1 and 2 the common collaboration patterns discovered in the creation of the US state parks Wikipedia GA.

Pattern 1, shown in Figure 4.3(a), illustrates the collaboration pattern of cluster 1. We first show a PGP sequence in the article development cycle. In the first P phase, the editors mostly have lower scores across six editing traits in comparison to the later phases. The interquartile range of Q1-Q3 is also close to 0.

In the G phase, the editors exhibited higher trait scores in content-shaping, since half of the quartile box is located above 2, indicating that, in this phase, there are mainly strong content-shaping editors. In the second P phase prior to the GA completion, the highest editing trait is copy-editing, with its box positioned higher than its position in the other two phases. This indicates that there are more high-ranked copy-editing editors involved in the phase, where most work is directed toward finalizing the article. In addition, the results also indicate there are stronger link-fixing editors involved in the G and second P phases prior to finalizing the article.

Figure 4.3(b) illustrates Pattern 2, which is the collaboration pattern of clus-

ter 2. It has a GPGP sequence in its article-development cycle. This pattern is also associated with much larger revision numbers in comparison with Pattern 1. There is a very small G phase at the beginning of the article followed by a long P phase. Similar to Pattern 1, at the beginning of the article's development, in these two phases, there are more editors involved, but most editors have lower scores across all six editing traits. The boxes of all six traits are located close to 0 score. After the first GP phases, the article has a burst in article-size growth in the second G phase. In this phase, there are six editors involved and they exhibit higher trait scores in content-shaping and copy-editing, since half of each box is located above the score of 2. In addition, the results show that the group of editors have, on average, higher link-fixing scores. Prior to article completion, in the second P phase, there were on average only two editors involved, even though the number of revisions was higher. The second P phase appears to mainly involve editors with higher content-shaping and copy-editing scores but with scores lower than the editors involved in the previous G phase.

The above results show that there are two collaboration patterns that can yield GA in the US state parks Wikipedia category. Both patterns show a burst in article-size growth, the G phase in the article-creation process. Prior to this G phrase, the number of editors involved is higher, but generally such editors have scores of around 0 across all six traits. In the burst phase, the editors involved in creating the GA are different from the earlier period and mainly have high content-shaping scores. They also show a burst in article-size increase and a P phase toward the end of the GA-creation process.

The two main collaboration patterns found for the US state parks GA are similar. If 10% of article-size growth is used to determine a G phase, then

both Patterns 1 and 2 have PGP article-development sequences with editors of overall low editing-trait scores involved in the earlier article creation. There are also editors with high content-shaping and copy-editing scores in the later period of article development who complete the GA. The main difference between Pattern 1 and Pattern 2 is the length of revisions.

4.4.2 Children's Book

In this section, we report the findings on GA in the Wikipedia category of children's book. Four editing traits were extracted from 17 GA:

1. A content-shaping trait with five activities focusing on content-information coverage: format, add and clarify information; fix and delete links; add links in the category, content, and reference; and use the talk page. Since these activities mainly target content-information enrichment, they are labeled as a content-shaping trait.
2. A copy-editing trait that emphasizes spelling and grammar.
3. A reversion trait of reversing previous editing activity
4. A vandalism trait of damaging the article.

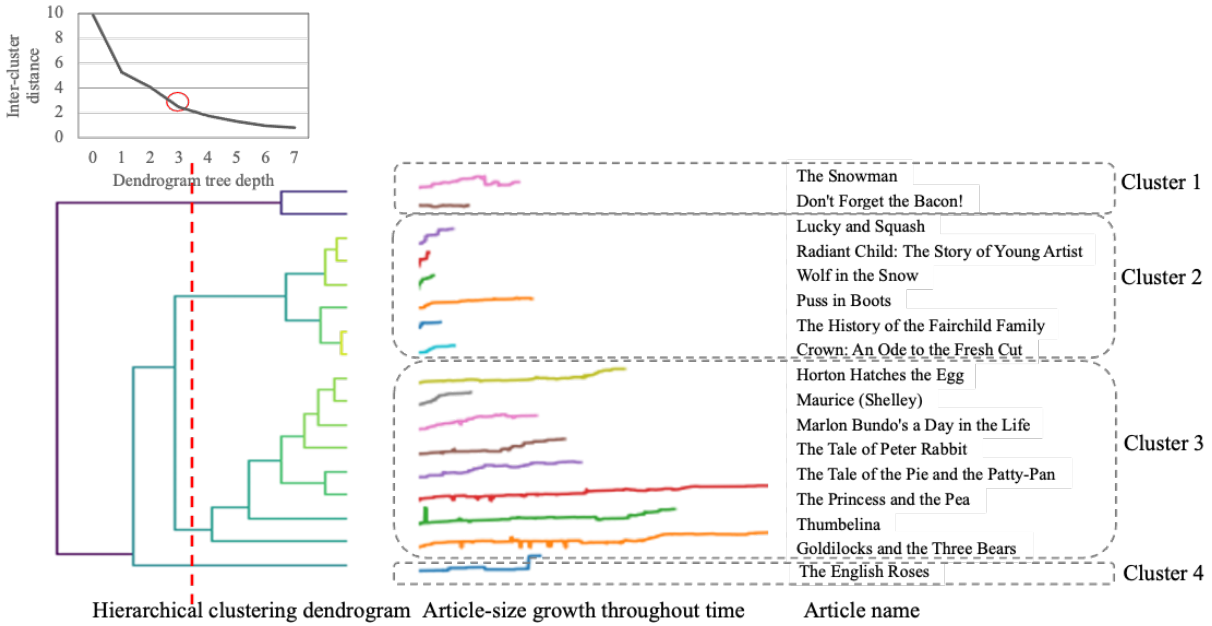
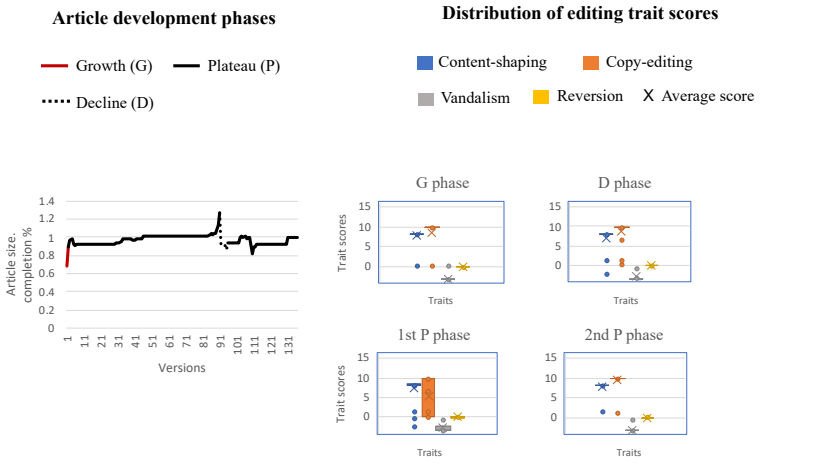
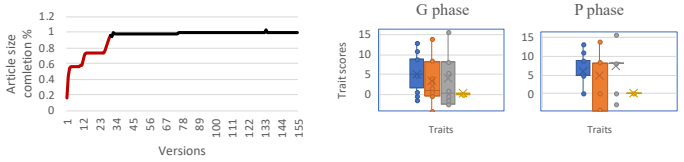


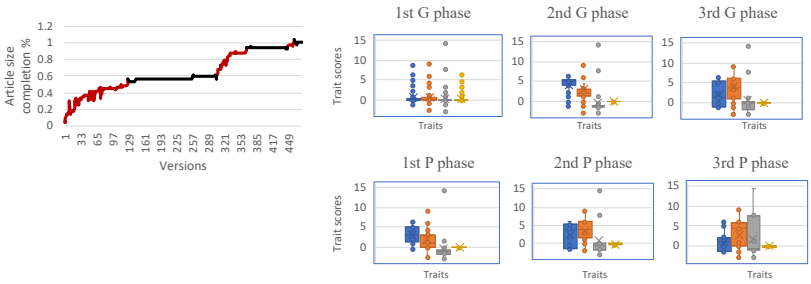
Figure 4.4: Clustering results of children's book GA.



(a) Pattern 1



(b) Pattern 2



(c) Pattern 3

Figure 4.5: Collaboration patterns of children's book GA.

Clustering Result

In this section we report the clustering results of children's book GA. The distance of each article is calculated based on DTW, and articles with least distance are clustered together to form a dendrogram as shown in Figure 4. The optimal number of clusters identified by the elbow method is 4 at tree depth 3 with inter-cluster distance of 2.45 to determine the extraction results of 4 clusters. Cluster 1 consists of only two articles and accounts for 11% of the total GA. Clusters 2 and 3 include the majority of GA at 35% and 47% of GA, respectively. Cluster 4 consists of only 1 article, which is considered an exceptional case and thus reporting of its collaboration is omitted.

Collaboration Pattern Results

In this section, we report the collaboration patterns that exist in children's book GA. There exist four different collaboration patterns. Again, we omit the cluster that involves only one GA and report the main collaboration patterns of the remaining 3 clusters as illustrated in Figure 4.5. We first found that all three collaboration patterns have a G phase to begin the GA. While Patterns 1 and 2 have a long P phase to complete the GA, Pattern 3 has a continuous PGPG sequence in GA development to finish the GA.

As shown in Figure 4.5(a), Pattern 1 is the collaboration pattern of the GA in cluster 1, which has only two articles. It shows a PGDP sequence in the article development cycle. It also shows a similar composition of the editors' involvement throughout the GPDP phases. There are high content-shaping and copy-editing score editors involved in all phases, and only in the D phase does a wider range of copy-editing editors appear. This indicates that the composition of editors is similar regardless of the article development.

In Figure 4.5(b), we show the collaboration Pattern 2 of cluster 2. It shows a GP sequence in the article-development cycle, and the composition of editors is similar throughout the article development. That is, high-scoring content-shaping and copy-editing editors are involved throughout the GA development. There is also a wider range of editors in content-shaping score involved in the G phase than in the P phase, as well as a wider range of editors in copy-editing score involved in the P phase than in the G phase.

Figure 4.5(c) shows the collaboration Pattern 3 of the GA in cluster 3. The GA development goes through GPGPGP sequences. The first G phase has mainly the involvement of editors of low content-shaping and copy-editing trait scores. There is also a wider range of editors in all four trait scores, since many outliers are represented with dots outside of the figure's plot box. In the following second P phase, there are editors with higher content-shaping and copy-editing trait scores. The second G phase and third P phase appear to have a similar editor composition. The second P and third GP phases also have editors with higher copy-editing trait scores. The scores of vandalism and reversion traits are similar throughout the article development.

The above results show that there are three collaboration patterns that can yield GA in the children's book Wikipedia category. Patterns 1 and 3 both have a scattered range of editors starting the article, while Pattern 2 shows that the compositions of editors involved are similar throughout the article development. Yet the two main patterns, Patterns 2 and 3, account for over 80% of the articles. Pattern 2 shows a simple burst of article-size growth and then a long refining period without much change in article size. Pattern 3, on the other hand, shows that the article is completed through gradual

development with a burst of article-size growth followed by a P phase, and then another growth of article size, and so on.

4.4.3 Chemical Compounds and Materials

In this section, we use the proposed method to find the collaboration patterns in the chemical compounds and materials GA. Four editing traits were extracted from 13 GA:

1. A content-shaping trait with nine activities that focus on content-information coverage: format, add information, delete information, clarify information, fix link, delete link, add link in the category, add link in content, and add link in reference. These activities mainly target content-information enrichment, so they are content-shaping
2. A copy-editing trait that emphasizes spelling and grammar.
3. A reversion trait of reversing (undo) the activities of previous editors.
4. A vandalism trait representing the activity of vandalism.

Clustering Result

In this section we report the clustering results of chemical compounds and materials GA. Again, we use the DTW algorithm to calculate the distance of each article, and the articles with least distance are clustered together to form a dendrogram as shown in Figure 4.6. The optimal number of clusters is identified by the elbow method. It is 4 at tree depth 3 with inter-cluster distance of 0.90 to determine the extraction result of 4 clusters. Again, we found an uneven distribution of different collaboration patterns.

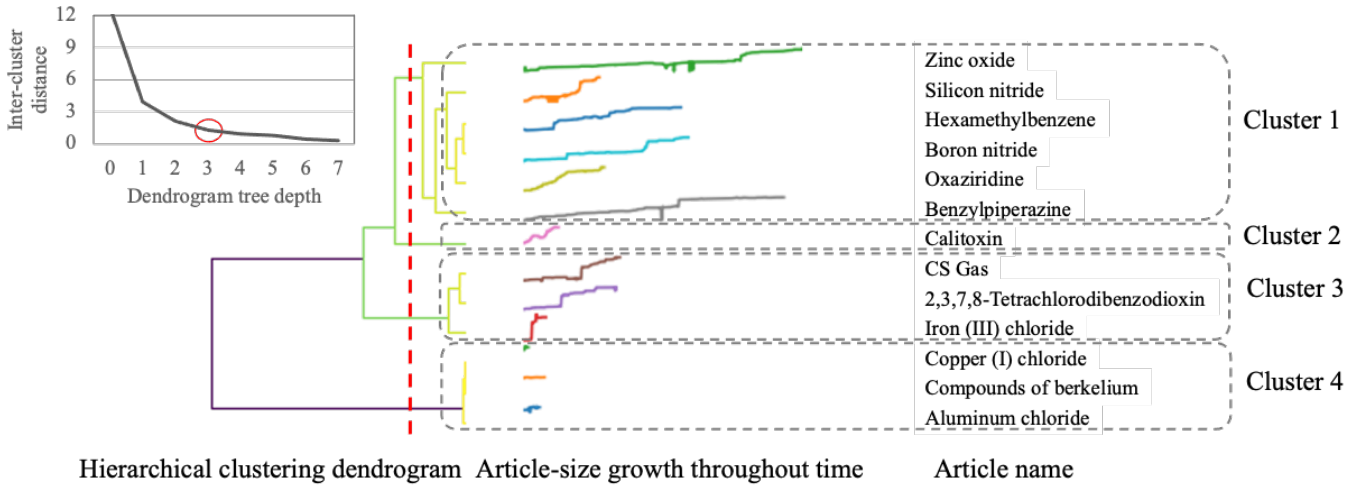


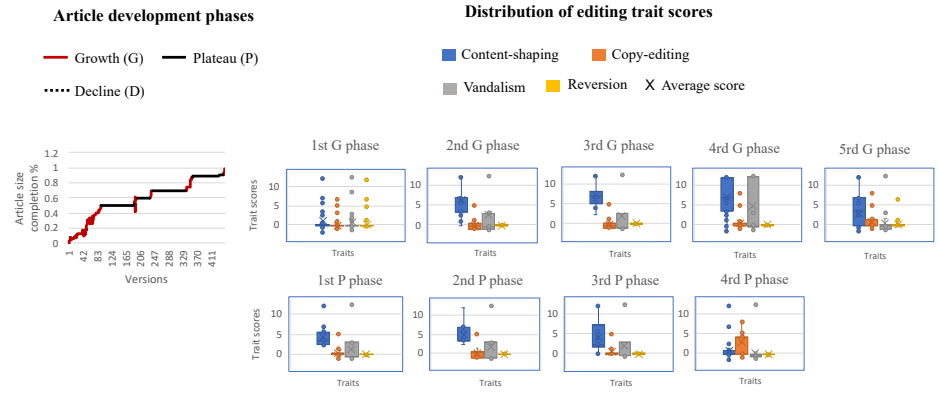
Figure 4.6: Clustering results of chemical compounds and materials.

The extraction results show 4 clusters. Cluster 1 accounts for 46% of the total GA, which represents the largest portion of GA in the category. Cluster 2 consists of only one article, which is viewed as an exceptional case and thus its collaboration pattern is omitted here. Clusters 3 and 4, on the other hand, each accounts for 23% of the GA in the category.

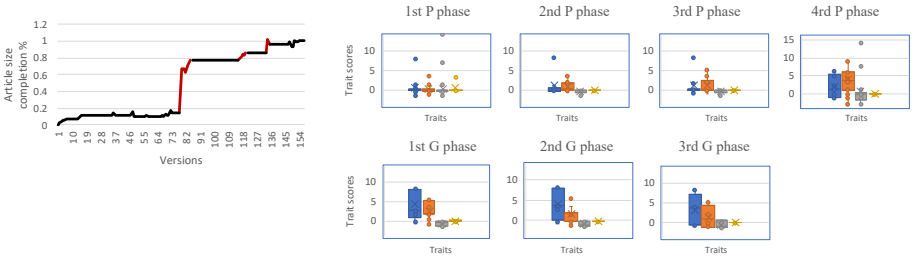
Collaboration Patterns Result

In this section, we report the collaboration patterns that exist in the chemical compounds and materials category. We report the three main collaboration patterns illustrated in Figure 4.7 and omit the collaboration patterns of cluster 2, since it consists of only one article. The three collaboration patterns show much difference in article development. Pattern 1 starts with a G phase and then goes through several PG phases to complete the GA. Pattern 3 starts the article with a P phase and then receives a burst of G phase to complete most of the article. After that, there are small GP phases before the completion of the article. Pattern 4 only has a P phase because the article reaches the complete GA size when it is established in the first version. After that, the article mainly receives refining editing activities.

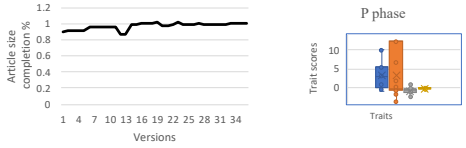
In Figure 4.7(a), we show the collaboration pattern of cluster 1 as Pattern 1. It has a GPGPGPGPG sequence in the article-development cycle. The burst of article size occurs in the first G phase, and then there are smaller G phases with P-phase intervals. In the first G phase, the editors cover a wide range of trait scores. It also appears to have high content-shaping editors involved in all phases. Yet editors in the fourth G phase have the highest average content-shaping editing scores. The editors of the third P phase also shows a higher average copy-editing score. The compositions of editors in the first



(a) Pattern 1



(b) Pattern 3



(c) Pattern 4

Figure 4.7: Collaboration patterns of chemical compounds and materials GA.

P phase, the second GP phase, and the third GP phase are similar with small differences in the content-shaping trait scores. This indicate that stronger content-shaping editors and copy-editing editors are mainly involved in the later stage of the GA development.

Figure 4.7(b) illustrates Pattern 3, which is the collaboration pattern of cluster 3. It starts with a P phase and then goes through a GPGPGP sequence in the article-development cycle. There are fewer article-growth phases compared with Pattern 1. The results show that in the first three P phases, mainly editors of low overall trait scores are involved. There are editors with high content-shaping trait scores involved in the first three G phases, though slightly fewer in the final P phase. The vandalism editors appeared in the beginning, and while the reversion trait scores did not change throughout the article development. This indicates the high-ranked content-shaping editors involved in the articles gradually appear with breaks, which results in P phases in the article development.

Figure 4.7(c) illustrates Pattern 4, which is the collaboration pattern of cluster 4. The article development only has a P phase, and the editors involved in creating the GA include those with high content-shaping scores and those with even higher copy-editing scores.

These findings show that there are different collaboration patterns that yield GA in the same Wikipedia category. Most of the collaboration patterns show scattered editors at the beginning of the article-creation process. Editors with high content-shaping scores were mainly involved prior to the GA nomination, with the help of editors having high copy-editing skills to improve the writing. Vandalism appears more active in the later stage of the GA development when there is more content in the GA. These findings were obtained for all three Wikipedia categories.

Our approach of using the article-development phase advances our understanding of collaboration by elucidating the dynamics of the editor composition changes as well as identifying the specific types of editors involved in

each phase. The collaboration pattern found with our method can be used as a reference for creating GA of the same Wikipedia category.

4.5 Discussion and Conclusion

Although open collaboration provides unlimited opportunities to society, it still faces several challenges. While Wikipedia articles serve as a reliable source of information, the quality of the articles and how they are created are issues still under scrutiny. Our research proposes a novel method to study different types of collaborations that could yield similar output from open collaboration. The method combines factor analysis and time series analysis to study the collaboration patterns of those working on GA-level articles in the same Wikipedia category.

We first apply factor analysis to the editing activities to obtain the editing traits, and then we differentiate the editors based on the scores of such traits. Furthermore, from the sequence of active editors in the GA-creation process, we can gain a better understanding of the collaboration patterns. To extract the collaboration patterns, we use dynamic time warping algorithms to handle the different time durations required in GA creation, and then we use hierarchical clustering to obtain the best possible summarized patterns of the created GA within a Wikipedia category. The collaboration patterns are first defined using the three article-development phases of growth, plateau and decline. Then, we report the composition of editors involved in each phase. Our findings advance the previous knowledge and clarify the increase in editing activities and concentration of workload prior to GA nomination.

Our method revealed different types of collaboration patterns in the creation of GA in three different Wikipedia categories. The collaboration patterns can be used as reference collaboration to create more GA. In addition, there are also different phase patterns in the article development, where each pattern corresponds to a different number of articles. There also exist single-article clusters, which are considered unusual collaboration patterns of GA.

We use a hierarchical clustering approach to reveal different collaboration patterns in creating GA. However, there are different clustering methods that can be applied to obtain the optimal number of clusters. In this research, we focus on proposing a method that can characterize the collaboration done on a Wikipedia GA. Accordingly, we use the most common and widely accepted approach of DTW to handle time series data of different lengths and a hierarchical clustering algorithm to find the clusters and the elbow method for determining the number of clusters. There are also different methods that can be used to determine the number of clusters. Another approach is based on the inter-cluster distance. The collaboration pattern found can be changed according to the number of clusters. The higher number of clusters would provide granularity of the collaboration pattern, but it could also be too complicated for reporting the collaboration pattern. In addition, we use the DBA method to obtain mean sequence results based on the longest time series of each cluster. However, this might stretch some of the shorter sequences of the same cluster and distort the calculation of phases.

Our approach of dividing article development into phases is based on revision history data. In this research, we use 5% of the article to differentiate the G phase and P phase. If we used a different value, we would likely find other article phase sequences. The goal of the proposed method is to

make article development easier given observations of the mean sequence. Based on three Wikipedia categories, we found it might be more practical to use a higher percentage for articles with a higher number of revisions. As can be seen in Figures 4.5(a), 4.7(b) and 4.7(c), a long article-development sequence makes pattern observation difficult. Moreover, there are similar compositions of editors in different continuous GP phases. If these phases were combined, the observations on the editors involved would not change.

Another consideration is the particular time-frame of article development. The study of Chapter 3 found that certain types of editors work intensely during the period prior to GA nomination. In this research's method, such information is not precisely monitored due to the fact that our method does not use the date on which each version is created. Nevertheless, our previous work confirmed using the same dataset in which the high-scoring content-shaping and copy-editing editors are mainly active in the period prior to GA evaluation to secure the GA status.

For the composition of editors in each phase, we use a box plot to reveal the relevant statistical properties; however, other approaches could be used to provide more information on the collaboration patterns. Here, we suggest that our findings on collaboration patterns could be used as a reference to create more GA. However, we acknowledge that, under the open collaboration scheme, this recommendation approach might limit volunteers' involvement if we imposed some restrictions to guide people in how they should contribute. On the other hand, it would be valuable for volunteers to know how their work contributes to creating GA and how different types of editors can productively contribute during the different phases of article development.

As a future research direction, our method could easily be extended to studying other quality levels or categories of Wikipedia articles. Additionally, there is a need for the ability to scrutinize the composition of editor sequences, their information, and the order in which they contribute. Other temporal scales could be considered in analyzing the editing activities as well as the article sizes. Finally, we might consider how different patterns of collaboration interact regardless of article quality. Depending on the findings, this could indicate that there are optimal collaboration patterns that can harness the collective intelligence arising in open crowds of editors.

The finding of this research contributes to advance previous research by characterizing the different types of collaboration patterns that exist in creating quality knowledge content of similar topics. The finding also provides a potential solution for creating more quality knowledge content. That is to use the collaboration pattern of creating good quality knowledge content as a reference to create more quality knowledge content. The preliminary finding of this result was published in [Chou et al., 2021a].

Chapter 5

TMchain: A Blockchain-Based Collaboration System for Teaching Materials

Currently, there is lacking of a system that can support the open collaboration of using existing knowledge content to build into one's content with copyright sharing. A solution that facilitates the usage of copyright-restricted resources is needed. Here we consider teaching material as a case and exploits the advantage of blockchain technology to propose a system to provide records of multiple authorships and contribution distribution of a teaching material that reuses in part, existing resources. Such records can be used as authorship evidence to claim copyright.

5.1 Introduction

Collaboration is common in the education sector. Apart from working collaboratively to write a book or doing research, collaboration in the educational context consists of sharing original creations for others to use, build on and so on [Hilton III and Wiley 2009]. Research also suggests that using and building on existing materials can save effort and yield quality materials more easily [Putnik et al., 2009]. Many resources can be used freely based on the standard copyright exemption for educational purposes or as open resources under the collaborative commons license [Lessig, 2004], such as Open Educational Resources (OER) [Hylén, 2021]. However, others, such as textbooks or online courses cannot. Authors who do not donate their works and want royalty sharing cannot collaborate with each other without some prearrangement. It can be risky for teachers to violate the copyright law when using other people's work in creating teaching materials. This is a particular issue during a pandemic as most of teaching as well as the teaching material goes online. Accidentally redistributing copyrighted content online can have severe consequences.

Thus, an alternative system is needed that can allow the usage of copyright-restricted resources for collaboration in teaching material generation. Sharing copyrights or royalty among the contributors of the resources constituting a teaching material can be a solution. We design TMchain, it provides a full record of multiple authorships and contributions when education resources are used in creating a teaching material and such records support royalty sharing.

TMchain exploits the advantages of the blockchain technology, as

blockchain provides a smart contract among participants. The transaction output yielded by the smart contract is stored in a secure, immutable, and reliably distributed ledger without centralized management. This is suitable for community collaboration [Torres et al., 2017]. It can store authorship and contribution distribution information securely for the individual works involved. This characteristic can facilitate the sharing of other teachers' work and thus support collaboration in creating teaching material.

There are blockchain studies dedicated to the protection of intellectual property by providing reliable records of the collaboration process in creating academic papers and scientific research [Niya et al., 2019, Novotny et al., 2018, Orvium, 2020, Mohd Pozi et al., 2018]. Unfortunately, these studies provide solutions for the creation of single outputs, such as a research paper. They fail to address collaboration in the sharing and reuse of existing materials. In addition, as these studies proposed to store the collaboration history on the blockchain, the memory usage and calculation cost of storing the history is high.

Many blockchain applications also have been developed for the education ecosystem, such as storing certificates issued by different institutions, identifying online education solutions, protecting the intellectual property of educational contents, supporting collaboration between students and teachers in higher education, cryptocurrency payments for education and administration of the educational process etc. [Chen et al., 2018, Fedorova and Skobleva, 2020].

Other papers use blockchain or others system to tackle the distribution of teaching material and make records when the material is distributed to students [Ocheja et al., 2019, Hou et al., 2019, Guo et al., 2020, Chunwijitra

et al., 2016]. Again, collaboration in the reuse of existing materials to create new materials was not considered.

Our proposal, TMchain, uses blockchain technology to tackle the authorship problem of using other teacher's material by storing the authorship of a completed teaching material. The blockchain system first provides a smart contract among teachers who agree with the use of their materials and creates secure records of the use of teaching materials. The transactions on the blockchain provide proofs of a material's authorship as well as recording multiple authorships and contribution distribution of the product when multiple materials are involved. The system uses word processing software to record editing activities involved in reusing existing material and only the authorship information and contribution distribution information of a completed teaching material is stored in the blockchain. In addition, the material files are stored in the network rather than in the blockchain. In this way, we can minimize blockchain cost and enhance support scalability.

We introduce real-world scenario implementations that show our solution has the ability to record the authorship of a teaching material, calculate contribution distribution of multiple authors, and handle the authorship records when the material is updated. We also report the feasibility and effectiveness of system with ether and run time costs. There are three main contributions of this research:

- We propose a system to support the collaboration needed when creating teaching materials that involve existing copyrighted resources.
- We design a novel blockchain-based system with the functions required to track authorship and contribution distribution records to al-

low educational resources to support royalty sharing.

- We implement the smart contract of the proposed system with material creation scenarios on Ethereum Remix-IDE to demonstrate its practical usage potential.

5.2 Related Work

Blockchain technology has been proposed that can record collaboration history [Crosby et al., 2016]. Because it sets participants to commit to a smart contact and stores the transactions in a ledger shared by the participants. Blockchain also has the advantages of providing a higher security, transparency, immutability of a record with decentralized management. The "smart contract" of blockchain is triggered by an event or participant's enquiry via prespecified computer protocol with predefined parties who can join the network to read and transfer data [Ellervee et al., 2017]. It also has a network consensus to support decentralized management of the ledger. The consensus ensures all blockchains in the network are legitimate and supports the existence of multiple copies in the network so no single party can manipulate the record [Belotti et al., 2019]. Many blockchain-based applications that act as public ledgers have been proposed, namely for medical records, logistics and Internet of Things as well as for academic publications [Belotti et al., 2019]. Yet no consideration was made of collaboration that combines existing educational resources into one teaching material like our work.

There are research efforts on using blockchain systems in the context of education collaboration. For supporting academic publication collabora-

tion, research has focused on collaboration for creating individual academic publications [Niya et al., 2019, Novotny et al., 2018, Nizamuddin et al., 2018, Günther and Chirita, 2018, Orvium, 2020, Mohd Pozi et al., 2018]. These systems preserve participant activities to acknowledge the contribution of each party.

Eureka [Niya et al., 2019] is a blockchain-based public network for cooperation publication. It has incentive sharing scheme that enables authors, referenced/linked author, editors, data providers and reviewers to receive the economic reward with digital token "EKA". [Orvium, 2020] is an open source blockchain platform to manage and support collaboration in science publications. The system allows researchers to share their work as well as to create open access journals. The system provides a public transparent trace of all the activities pertaining to a research paper from first submission, revisions, accepted or rejected peer reviews, copyright and user license changes. [Mohd Pozi et al., 2018] considered collaborative writing of scientific publications and preserving editing history in a block which can then be used for contribution calculation.

[Guo et al., 2020] proposed a blockchain-based digital rights management system for recording digital rights of educational resources. Yet each editing history is limited to 1024 characters which cannot support a creation of large documents with figures. ScienceRoot focused on a blockchain-enabled scientific ecosystem which tokenized the research process; it views itself as a science research marketplace that supports grant funding, publishing, and scientific collaboration [Günther and Chirita, 2018]. [Marjit and Kumar, 2020] introduced a solution with IPFS to support the OER to resolve the high cost of centralized storage of these resources in blockchain. In addi-

tion, other smart contracts have used in this research field for controlling access to teaching material [Günther and Chirita, 2018]. Yet none of these works provide multiple authorship recording functions when combining existing educational resources into one teaching material and our proposed system provides an alternative solution.

5.3 TMchain System Overview

We consider the collaboration needed in creating teaching material through the sharing of work. In this section, we describe the collaboration process in which multiple authors participate in developing one teaching material. We then design a system that records authorship and contribution sharing. After that we illustrate the smart contract code provided by the blockchain and transaction flow for registering authorship and sharing financial returns by function calls in the smart contract.

5.3.1 System Requirement

To elucidate system requirements, we first consider how teaching material is created. Research suggests a simple teaching material creation process [Putnik et al., 2009]. The teacher collects resources, then convert them into the teaching material via word processor or content conversion software. Teacher also adds in his/her own content. Last, the teacher exports the final material. The upper part in Figure 5.1 illustrates this process. To incorporate this process into a collaboration system, there are two core requirements. First, the system allows reuse of a resources and records attributing authorship when a material is used.

In addition, there is also a consideration of using blockchain [Nakamoto, 2008]. Due to the transaction recorded on the blockchain is immutable, the chain can grow so long that scalability is eventually degraded. This means the transaction data recorded on the blockchain should be kept to a minimum [Belotti et al., 2019]. Thus, we propose to store only the authorship distribution information of the finished teaching material in the blockchain, not each revision event. The authorship distribution information of the finished material on the blockchain is enough to confirm authorship and acts as evidence. In addition, the revision history recorded in the material itself by the word processing software, such as MS word or GoogleDoc and TM-chain registers the authorship and contribution distribution of a completed teaching material. The material file is stored outside the blockchain, using technology such as IPFS to identify the correct version of material related to the authorship records in the blockchain [Nizamuddin et al., 2018].

5.3.2 System Framework

To extract the required information from the material creation process and denote authorship record in blockchain, our framework has two main parts:

- Extract authorship distribution from editing activities: Existing word processing systems such as MSword and GoogleDoc provide the function of recording editing history and file mergers. So, this requirement can be fulfilled with an addon function to calculate the contribution distribution of a finished teaching material as shown in the middle part of Figure 5.1. This part is outside the blockchain. While there are various methods to calculate the contribution share [**?**], we discuss this in detail in a later section.

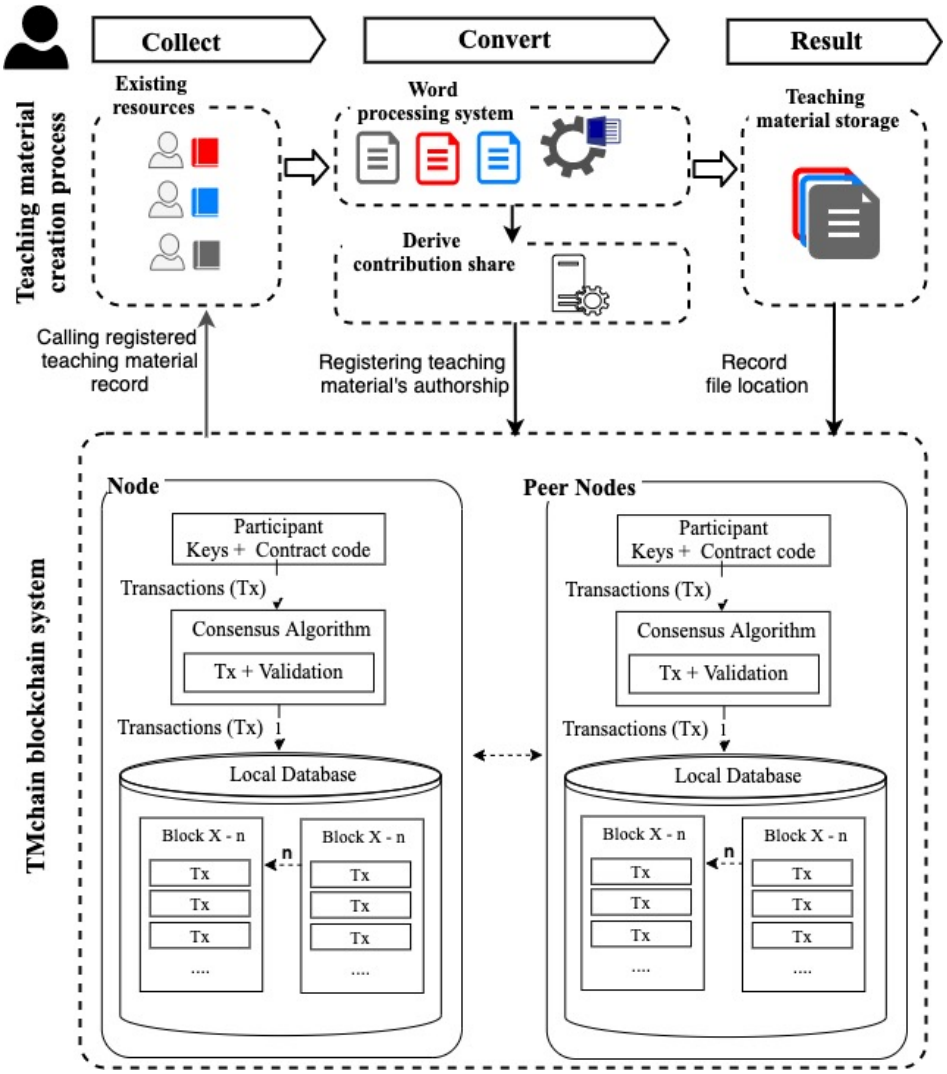


Figure 5.1: Teaching material creation process and TMchain system framework.

- Register authorship and contribution distribution information into blockchain (lower part of Fig.5.1): This is realized by the smart contract function of blockchain. The two pieces of information must be

recorded. First is the authorship of each material created. Second is the contribution distribution information of each material with its authorship in the blockchain system.

The proposed framework works with the teaching material creation process. During the creation process, the teacher starts by collecting existing resources to be used. Once they are selected, their content is extracted and edited to form the teaching material. The editing history recorded by the word processing software is used to create authorship records and contribution distribution of a finished teaching material. Then the information is stored in the blockchain.

5.3.3 Smart Contract Functions

To allow teachers make use of each other's teaching materials, the participants are bound with smart contracts. Entering into a smart contract is taken to mean that the teachers allow their works to be used in the collaboration system as well as truthfully committing to record authorship distribution information of the material they created.

We use Solidity language to create the smart contract named TeachingMaterialManager.sol to govern this agreement where record is the transaction executed by teachers. Each teacher needs to have an account in the system. The contract is specified with two methods (functions) namely *createMaterial* and *deriveMaterial*. *createMaterial* is used to register the authorship of a work and *deriveMaterial* is used to record the authorship.

To extract the required information of collaboratively created teaching materials. This function creates transactions to record the material that incor-

```
## A material registration function to record the material information
on the blockchain. Input: author account ID, name of the material, hash
of the material, return: teaching material ID.

    function createMaterial(string calldata name, string calldata hash)
    public returns (uint){
        uint id = materials.length;
        uint[] memory references;
        bytes8[] memory proportions;
        materials.push(Material(
            id, msg.sender, name, hash, references, proportions));
        materialToAuthor[id] = msg.sender;
        materialToRegisteredTime[id] = block.timestamp;
        ownerMaterialCount[msg.sender]++;
        emit NewMaterial(id, name);
        return id;
    }
```

(a) createMaterial function

```
## Register the record of a teaching material that incorporates other
teachers' materials. Input: author account ID, name of the material,
hash of the material, incorporated material ids and proportions,
return: teaching material ID.

    function deriveMaterial(string calldata name, string calldata hash,
        uint[] calldata references, bytes8[] calldata proportions)
    public returns (uint){
        uint id = materials.length;
        materials.push(Material(
            id, msg.sender, name, hash, references, proportions));
        materialToAuthor[id] = msg.sender;
        ownerMaterialCount[msg.sender]++;
        emit NewMaterial(id, name);
        return id;
    }
```

(b) *deriveMaterial* function

Figure 5.2: Smart contract codes of TMchain.

porates the teaching materials of others in blockchain. We show the code snippets written in Solidity in Figure 5.2. The input of calldata name and calldata hash is the author's account name and account hash registered in the blockchain system.

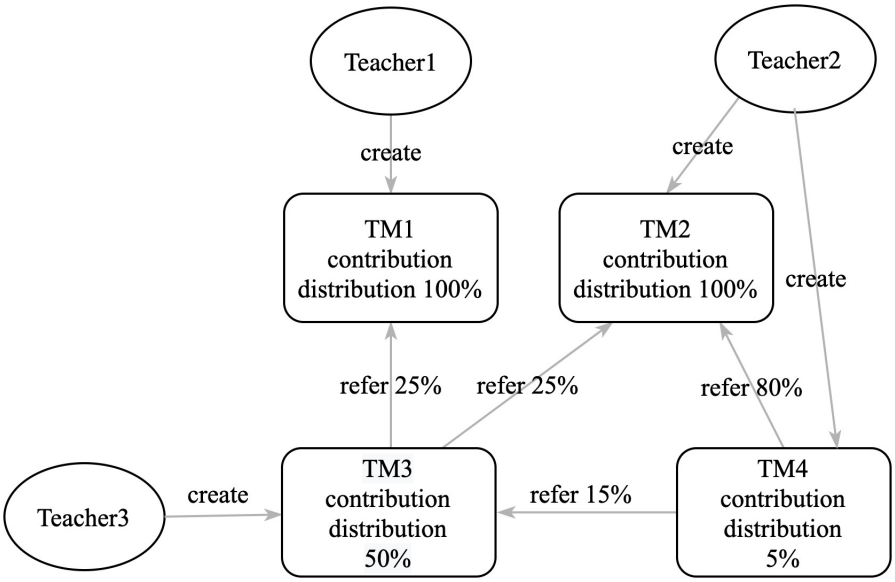


Figure 5.3: Logical graph data structures of TM information.

The data structure of a material contents material id, author information(msg.sender), name of the teaching material, a string hash to identify the material, references for the material id array of used teaching materials. The proportion array in bytes form is for contribution distribution which indicate the proportions of the materials involved in the teaching material. The contribution distribution calculation is performed by *getProportion* function which is outside of *deriveMaterial*. It is also possible to embed the calculation within *deriveMaterial*. The *getProportion* function reads only data from the transaction and calculates the contribution percentage of reused material as input. Then it calculates the remaining portion as the contribution from the teacher creating the collaboratively created teaching material. This proportion share is recorded in blockchain for contribution distribution. In the case of material created using another teaching material that already

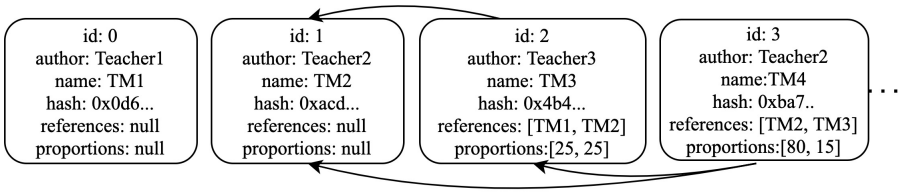
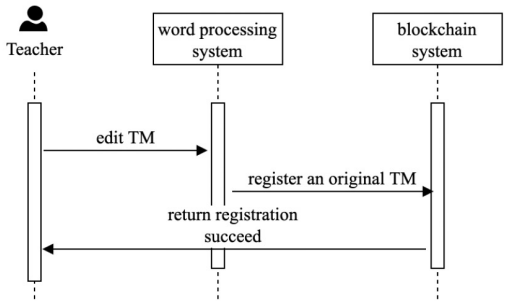


Figure 5.4: Physical data structures of TM information in Blockchain.

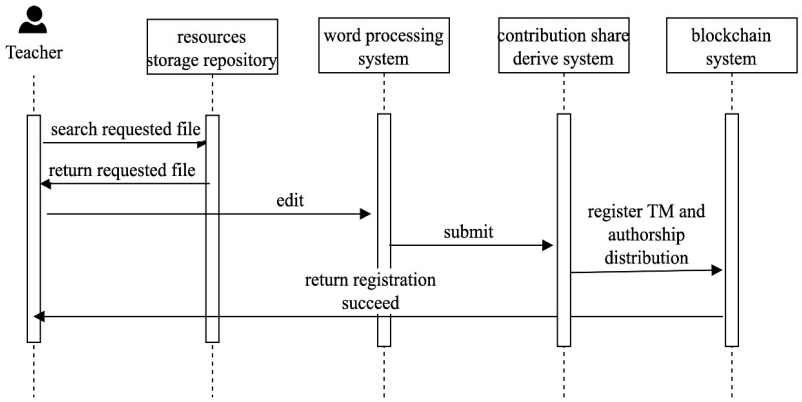
contains other collaboratively created materials, the authorship distribution of the materials used can be called to perform nested calculations.

The blockchain part of our system is a simple application of the Ethereum system. One main difference from other Ethereum systems is the storage of authors and material/graphs as smart contract states. Though graphs usually represent nodes, edges as properties, our system stores the material reference graph as a simple sequential array that contains material information and links to other information. This approach decreases memory usage and calculation costs.

An example of logical data structure representation of teacher and teaching material is shown in Figure 5.3. Teacher1, Teacher2 and Teacher3 created teaching materials (TM) TM1, TM2 and TM3, respectively. The graph structure shows that TM1 is 100% created by Teacher1 and TM2 is 100% created by Teacher2. TM3 is created by Teacher3 but also used TM1 (accounts for 25% contribution share) and TM2 (accounts for 25% contribution distribution). Thus, the contribution distribution of TM3 created by Teacher3 is 50%. TM4 is an updated version of TM2 (accounts for 80% contribution distribution) and TM3 (accounts for 15% of contribution distribution.) So, the originality of TM4 is 5%. The physical data structure



(a) Register an original Teaching Material (TM)



(b) Register collaboratively created Teaching Material (TM)

Figure 5.5: Transaction flows of TMchain

stored in blockchain is shown in Figure 5.4.

Transaction Flows

As we saw in the previous section, the authorship distribution recorded in blockchain mainly involves two functions: *createMaterial* and *deriveMa-*

terial. To make this clearer, we illustrate the general transaction flows of registering an original teaching material and registering a teaching material which has other materials through the use of the two functions. Figure 5.5a shows a typical transaction flow of registering an original material; it calls *createMaterial* in the smart contract. The procedure starts with a teacher using a word processing system to write the material and upon completion the material is registered with blockchain by calling *createMaterial*. Once the authorship of the material is registered on the blockchain system, the system returns log information notifying the author that registration has succeeded.

Figure 5.5b illustrates how the authorship and contribution distribution information are created in the blockchain if the teaching material uses other teaching materials. The teacher first searches for resources that he/she wants to use. The resource storage repository returns the requested resources to the teacher. After that, the teacher uses a word processing system to edit the teaching material and submit the changes. When the final version of the teaching material is confirmed, the editing history held by the word processing system can be used to generate the authorship distribution information based on contribution share calculation. *deriveMaterial* is then called to register multiple authorships and contribution distribution information in blockchain. Once such information is recorded successfully, the blockchain system returns a log message indicating transaction success to the teacher.

The smart contract generates transactions of authorship of an original teaching material as well as the multiple authorships and contribution distribution of a teaching material that uses other materials. The transaction data is held in blockchain. The document file of the teaching material is stored outside blockchain. When a collaboratively created material is used by other

teachers, multiple authorship and contribution distribution information can be extracted from previous records and the latest teacher simply adds on his/her editing activities to create and register the new material.

5.4 Scenario Implementation

In this section, we demonstrate TMchain smart contract implementation with teaching material creation scenarios and test its functionality with Ethereum Remix IDE [Wood et al., 2014, Ethereum, 2021]. We use lecture presentation slides from the "Field based Learning/Problem Based Learning" (FBL/PBL) course of the Design School of Kyoto University. The course has been taught for several years and the teaching material is constantly updated by different teachers.

5.4.1 Register Original Teaching Material in TMchain

In Figure 5.6. we show a scenario of the system registering an original teaching material. The teaching material was first created for "FBL/PBL" course by Teacher1. It was named TM1 by Teacher1 for TMchain registration. Teacher1 was given an account with id account0 with account hash of "0x1fb3e76fA2b83d7F8A53ba74867296c0fcDC6c37" by the TMchain system. In this scenario, it called with *createMateria*l function under TeachingMaterialManager.sol contract. When the registration of TM1 succeeds, the transaction is stored in block 342 with txIndex[0] from account0 (shown as account hash as "from" item in the block).

The transaction data is stored as a binary record under "data". The blockchain transaction log and storage on blockchain is shown in Figure

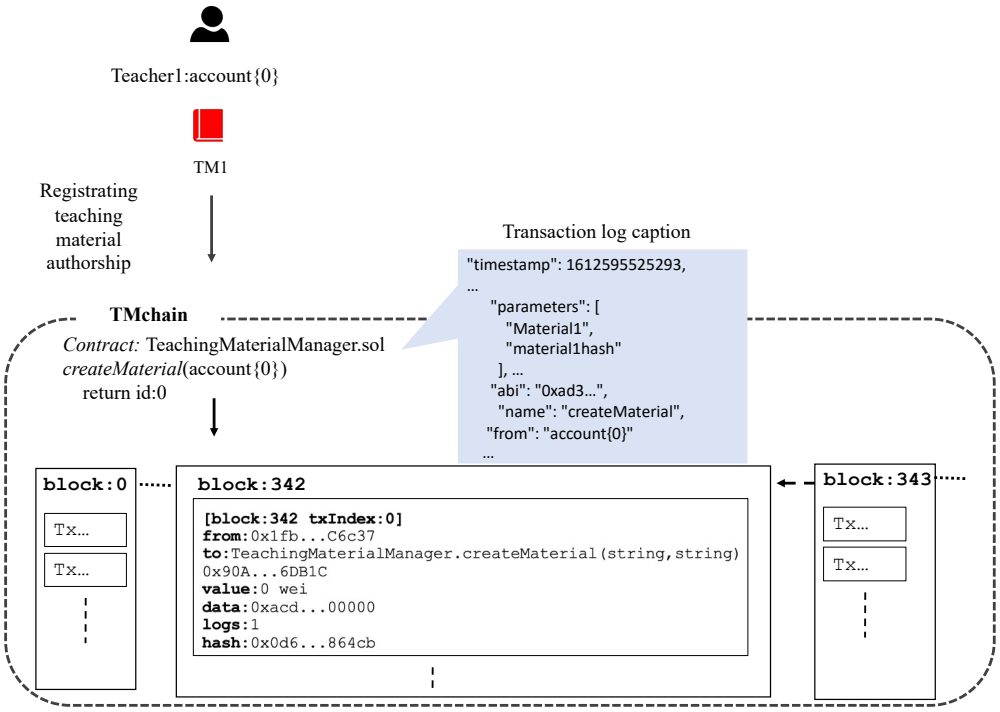


Figure 5.6: *createMaterial* scenario: Teacher1 registers TM1 to TMchain

5.6.

5.4.2 Registering Teaching Material That Uses Other Materials in TMchain

In this section, we demonstrate how a teaching material that uses other materials is registered in TMchain. We count the presentation slides to determine contribution distribution.

In a later semester, Teacher3 is assigned to teach “FBL/PBL” course. She decides to extract four presentation slides from the previous “FBL/PBL”

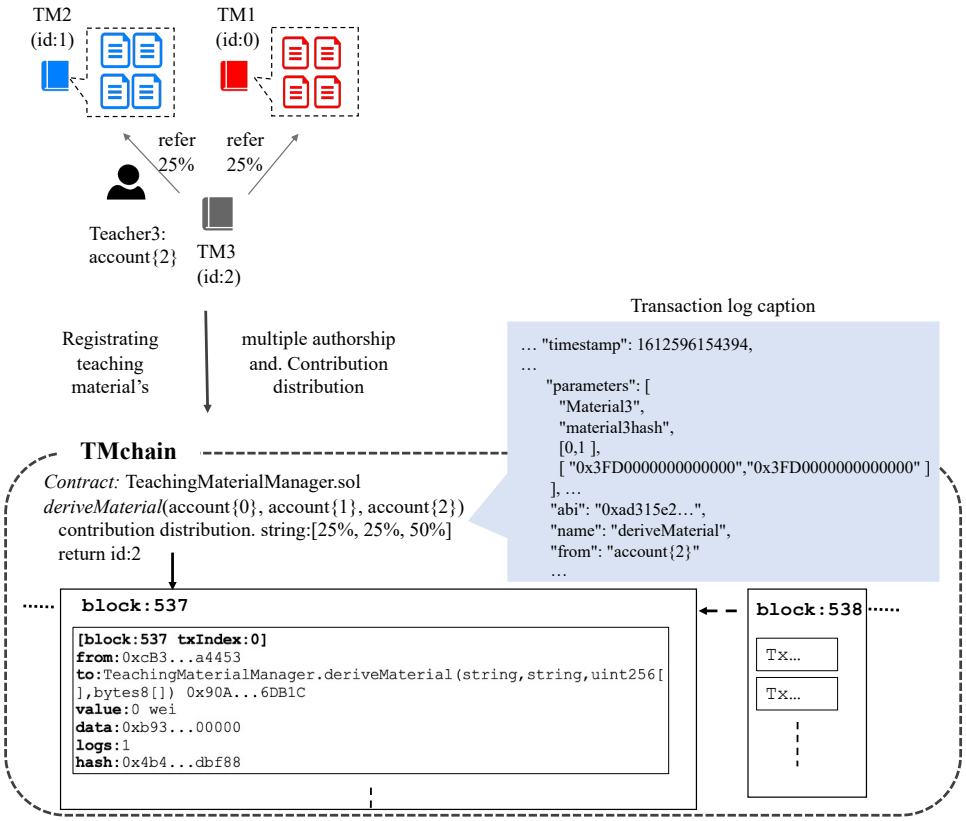


Figure 5.7: *deriveMaterial* scenario: Teacher3 registers TM3 to TMchain

teaching material (TM1) and extract some content from the teaching material (TM2) created by Teacher2 for another course: "Information and Society". To use TM2, Teacher2 needs to register this material in TMchain. The registration of TM2 follows the process shown in Figure 5.6. So, we omit the log and block information here. Teacher3 uses TM1 to create four slides and TM2 to create four slides. Last, she creates 8 slides by herself as shown in Figure 5.7. For TM3, authorship is splitting among Teacher1, Teacher2 and Techer3 with contribution shares of 25%, 25% and 50%, respectively, as shown in Figure 5.7.

To register this teaching material, which has multiple authorship and contribution distribution, *deriveMaterial* is called. The input data are material ids and their proportions. TM1 (material id: 0) accounts for 25% and TM2 (material id:1) accounts for 25%. The transaction log holds materials used with their work id in a list array as [0,1] and contribution distribution is represented in binary 64 format in a list array of [”0x3FD0000000000000”, ”0x3FD0000000000000”]. The rest of the 50% contribution share belongs to Teacher3 (account2) who creates TM3. This transaction was executed by Teacher3 with account hash of “0xcB3420DD4D4573b779517f605646849595Fa4453” as “from” item in the block transaction item and the information is recorded in binary form in “data” item.

This scenario calls *deriveMaterial*. When the registration of TM3 succeeds, the transaction is stored in block 537 from account2 as in Figure 5.7. In block:537, txIndex:0 shows Teacher3’s account hash by “from” item and used TeachingMaterialManager.sol contract with *deriveMaterial*. The data of the transaction is stored in binary form under the “data” item. This transaction runtime is reported to be within 5145 milliseconds. In the case that Teacher2 updates TM2 to yield new teaching material TM4 (in Figure 5.3 and Figure 5.4) through the addition of more presentation slides. This also calls *deriveMaterial* function to register a new material. The transaction process and result are similar to the registration of TM3 shown in Figure 5.7.

5.5 Evaluation and Discussion

In this section we provide evaluation details of feasibility and effectiveness based on TMchain's implementation. In addition, we discuss the practical usage and future research direction of our proposed system.

5.5.1 Function Evaluation

We consider the need for a solution to support teaching material sharing with royalty sharing property. The system is used to support the collaboration in using existing material. The system requirement is to record the authorship of a teaching material as well as multiple authorship and contribution distribution information when there are multiple teaching materials involved. We report its functionality evaluation in this section.

TMchain satisfies the required feature with *createMaterial* and *deriveMaterial* functions. In Section 5.4, we show that the authorship information as well as the contribution distribution information can be stored successfully. Each original teaching material is first registered by calling *createMaterial* to receive a unique ID with timestamped transaction record in blockchain to provide security. Such information is immutable and cannot be altered due to the property of the blockchain technology.

The record of a teaching material that uses multiple materials is created by *deriveMaterial*. *deriveMaterial* is called to establish multiple authorships, calculate contribution distribution, record the result in blockchain. With these two functions, the system records the authorship of teaching materials. The authorship information can be used as evidence to support royalty sharing when collaboration involves the use of copyright restricted material.

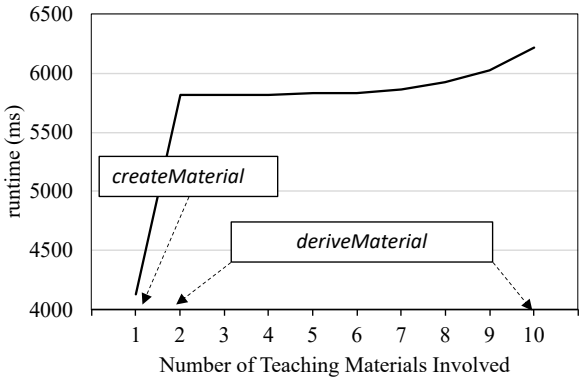


Figure 5.8: Function runtime in milliseconds

5.5.2 Performance Evaluation

The performance evaluation focuses on function call runtime. We *createMaterial* and *deriveMaterial* from create new transaction block to save material information on blockchain but does not include off-chain function runtime. Our smart contract just read the data of contribution distribution generated by the off-chain word processor.

We called *createMaterial* 50 times and determined the average runtime to be 4127 milliseconds. For *deriveMaterial*, we use scenario of create a material that uses the material of previous version. For example, a material version 1 is used in creating material version 2. Then material 2 is used in creating material 3 and so on. We perform *deriveMaterial* function call up to 10 version levels. The average runtime for 50 calls was around 6000 milliseconds when the teaching materials involved did not exceed the 8 version levels as shown in Figure 5.8. We call *createMaterial* with 1 material and call *deriveMaterial* with 2 to 10 version levels involved in collaboratively created teaching materials. The runtime grew exponentially as teaching ma-

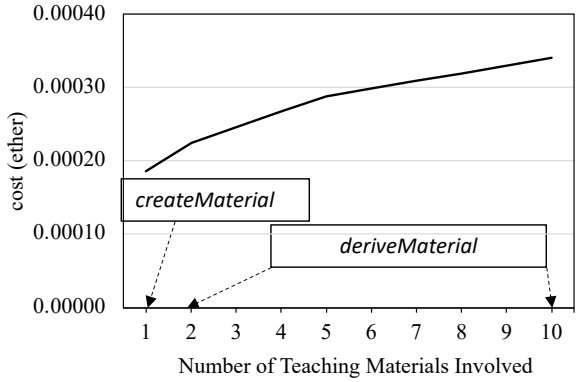


Figure 5.9: Cost of for collaboratively created teaching materials

terial versions increased as teaching materials of previous versions had to be called upon and nested calculations performed. Note that the runtime for calculating contribution distribution has less impact on runtime required. It only took 20 milliseconds when there was only one other teacher's material involved. It took on average less than 100 milliseconds when there were less than eight materials involved and 0.5 seconds with 10 materials involved. Considering it can take days and even weeks to create a teaching material, the time to register the material and/or calculate the authorship distribution on TMchain is insignificant at only several seconds. This runtime makes real world usage possible.

5.6 Discussion

TMchain currently supports only teachers who have accounts and materials for reuse also need to be registered with the system. We adopted the Ethereum Remix-IDE platform environment and code the smart contract with Solidity language. This implementation might be altered if a different

platform is used. The system relies on teachers faithfully using the system. We use existing network consensus of existing blockchain technology to prevent alteration of the transaction records. In its present version it cannot guarantee if a participant is registering material not created by him/ herself. Of course, such actions would leave evidence of the illegal acts.

There is cryptocurrency payment cost involved when using the blockchain system. In the experiment, we assigned the cost to the teacher who registered their teaching material. It costs 0.000185488 ether per transaction on TMchain for calling *createMaterial*. This is the cost for registering a teaching material with one author in the system. For *deriveMaterial* transactions, the ether cost increases linearly with the number of materials involved. It costs 0.00002096 ether for each additional teaching materials as shown in Figure 5.9. It shows the cost of ether with 1 material by calling *createMaterial* and a teaching material with 2 to 10 materials involved by calling *deriveMaterial* function. How to share this cost among stake holders is a future research direction.

Ethereum uses gas to indicate the amount of computational effort required to execute specific operations on the Ethereum network and gas fee is determined by supply and demand between the network's miners and users. In order to represent the practicality of TMchain in the real-world scenario, we calculate the gas cost of ether. In our system, 1,000,000 gas cost is fixed to 9.21147E-12 ether. Yet in real world, ether cost is not fixed. This issue, considering gas price with real Ethereum network when using TMchain, can be part of future research.

In addition, the execution time of our smart contract is influenced by both test environment and contract overhead. In this research, we used a Mac-

Book Air PC with 8GB and 1.6GHz CPU. It is expected that runtime can be improved by using a more powerful computer [Schäffer et al., 2019]. Yet they also found that the runtime reduction saturates at high computing powers. They implemented a simple smart contract using an 8GB and 3GHz CPU and the result was a runtime over 4 seconds. This was reduced to slightly over 3 seconds for both 16GB and 32GB. This means when the computer's memory and CPU power reach a certain level, the smart contract overhead is more influential as regards the runtime.

In the TMchain smart contract, functions are also simple. They only add one array element and few map elements. Yet of the functions *createMaterial* and *deriveMaterial*, *deriveMaterial* is more complex than *createMaterial*. An example can be found in Figure 5.4; id:0 is created by *createMaterial* and id:2 is created by *deriveMaterial*. The average runtime of *createMaterial* with only 1 teaching material involved is 4127 milliseconds as shown in Figure 5.8. Figure 8 also shows that the runtime of *deriveMaterial* with 2 teaching materials is 5816 milliseconds. This means the time required is also influenced by the complexity of the functions in the smart contract. Yet this runtime result shows it only takes a few seconds to register a collaboratively created material with the system. This is still insignificant compared with the whole teaching material creation process.

Here we mainly consider ether cost and runtime in evaluating the practical usage of TMchain. Yet there are additional parameters that must be considered by a full-scale scalability report. They include block size, transaction rate and others [Schäffer et al., 2019]. Future work includes large scale scalability analysis of the system. Contribution allocation is also a research area of interest. Various contribution calculation methods have been suggested

[Torres et al., 2017]. In addition, the scenario here considered the number of slides used in a material, but this is merely for demonstration purposes. The proposed system can adopt any contribution calculation algorithm.

The legal implication of using blockchain records is also a concern. While our proposed system has special provisions aimed at facilitating teacher's trust in blockchain records, their good faith usage of copyrighted teaching materials is assumed. How to legalize the transactions and the status of smart contracts and their consequences is also a future research area. This involves how to support intellectual property right enforcement and so assumes blockchain records will be legitimate evidence in different legal regimes.

Blockchain technology provides an immutable ledger that complicates the alteration of a registered teaching material. In the case of removing some content, it is better to register the material by recompiling the multiple authorship and contribution distribution based on the editing history captured by the word processing tool used. TMchain treats a finished teaching material as a unit and stores its authorship. Our system does not record the editing history of the material, only the authorship distribution of a completed material. Therefore, we rely on the word processing system used to support the tamper resistance of the editing history. If the submitted record of authorship of a completed work is the final version, our system stores the record. Our system uses blockchain technology to assure the alteration resistance of the authorship information of completed teaching materials.

Teaching material files can be stored off-chain in TMchain with a hash to indicate file location. While the hash can be an URL, the author needs to keep the content of the URL unchanged to support the usage of TMchain.

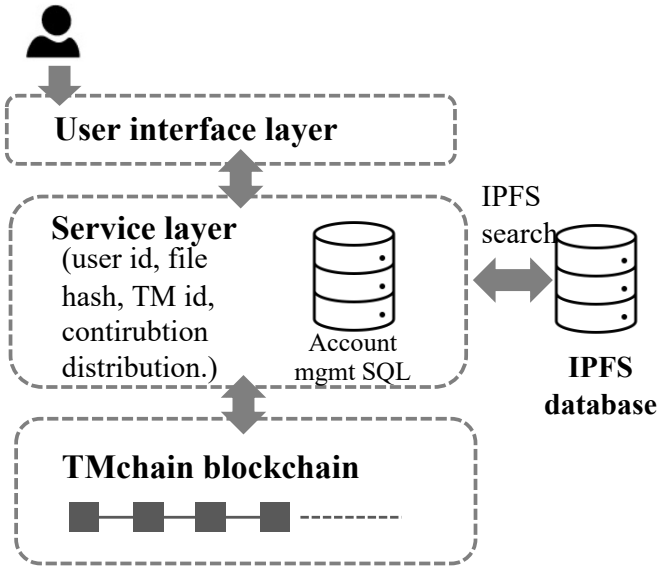


Figure 5.10: Example of collaborative TM management system architecture

It is better to use IPFS to store material files in our system to resist content modification.

In this research we focus on the design of the smart contract to support the collaboration of knowledge content. The public ledger provided by the blockchain system can be regarded as a database to store the record of authorship and contribution distribution of collaborative teaching materials. With the smart contract in place, additional functions can be added on according to the design of the teaching material management system shown in Figure 5.10.

The data in the blockchain can only be added and not modified. To make use of the data store in the blockchain, it is common to add a service layer to extract records, perform calculations etc. Above the service layer there is a user interface to allow users to manage and interact with their records.

To add two layers on a blockchain is common for blockchain based applications [Niya et al., 2019]. Take the cryptocurrency market for example. People do not access cryptocurrency blockchain systems directly, they have cryptocurrency exchanges to manage their cryptocurrency and their keys which are stored in "wallets" - a service offered by exchange platform.

Similarly, to have a complete collaborative teaching material system, many functions can be added to the service layer. For example, a teacher can have an account name and a list of TM he/she created. These data can be stored in a small local database in the service layer. When there is a transaction of a collaborative TM using this teacher's TM, the transaction can be noted into the local database too. In addition, when searching for a TM, it can be searched based on the content and found its IPFS hash [IPFS, 2022]. The found IPFS file name (hash) can be related to the record in the blockchain to get the registration data of the TM (TM id and transaction hash) and then call the authorship data (user id) to generate the authorship information and contribution distribution data for TMchain registration.

The proposed system currently cannot protect false registration. However, there are existing methods to detect similarity between documents which can be used to check if the document has another authorship record in TMchain or other copyright recording system, for example, Non-Fungible Token. The record of a TM in TMchain is immutable, therefore if anyone tries to register a document which he/she did not create, this record can be used against them as copyright infringement. Finally, the TMchain system is impacted by inheritance of existing blockchain technology as well as the implementation platform and system environment limitation. While it took about 10 seconds on average to complete the registration of a collabora-

tively created teaching material in our experiment, we can envisage a faster blockchain system and configuration to improve system performance.

5.7 Conclusion

We consider the type of open collaboration of using existing knowledge content to build into one's knowledge content. We propose a system that records authorship when using other people's knowledge content. Here we use teaching material as a case. Currently, Open Educational Resources Organization organizes a library where teachers can donate their materials to be reused under creative commons to provide open collaboration opportunities. Yet people who demand for copyright sharing when using their teaching material cannot participate. We propose to use blockchain technology to support royalty sharing in collaboration when using another teacher's teaching material. It features to store authorship of an originally created material as well as multiple authorships when various material files are involved in a teaching material.

We design TMchain to support collaboration in the use of existing teaching materials as well as to record attributing authorship. It is a simple Ethereum application that stores the authorship and contribution distribution information when existing resources are used and /or referenced in creating teaching material. The security of storing such records is supported by blockchain. We also consider how to minimize the blockchain cost by storing material files off-chain and utilizing the functionality of existing word processing systems to capture edit history before finalizing the authorship and contribution distribution for blockchain storage. In addition, we also propose to

utilize a distributed file system on the network to store material files. With the core application on blockchain, functions can be added to blockchain service layer to provide user account management and TM usage counts. We implemented TMchain on Ethereum Remix-IDE and used real life scenarios to demonstrate the practicality and effectiveness of TMchain. The research result was published in [Chou et al., 2021b] and [Chou et al., 2022b].

Chapter 6

Conclusion

6.1 Contributions

The goal of this thesis was to support open collaboration on the creation of knowledge content and to enable continuing development in this area. We have considered two different types of collaboration, and our research goal was achieved by identifying the issues arising from these two forms of open collaboration. We used specific case studies to perform our analysis and develop solutions. For the first type of open collaboration, which calls on people with a variety of backgrounds to work together to create a single shared knowledge content output, we used Wikipedia as a case study. For the second type, which involves participants with similar backgrounds who wish to make use of other people's work to create independent knowledge content output, we used teaching materials as a case study. The research was conducted in three parts: analyzing how different teams can create knowledge content of similar quality through open collaboration; ex-

tracting the different types of collaboration patterns in the creation of good-quality knowledge content; and designing an alternative blockchain-based system to provide multiple authorship records for knowledge content created via open collaboration. We acknowledge that the use of Wikipedia and teaching material as case studies cannot cover the full scope of the types of knowledge content created via open collaboration. In addition, there may be other forms of open collaboration that have been neglected. However, our research contributes through our analysis and the design of solutions for knowledge content created via open collaboration within the specific scope of the study. Our contributions can be summarized as follows.

1. We have deepened the current understanding of open collaboration to produce high-quality knowledge content.

   Despite years of research effort, the issue of how open collaboration by teams of different sizes can create quality knowledge content remains unclear. We used Wikipedia articles at the GA level of quality as a case study, and proposed a novel method that sheds light on the problem of creating quality Wikipedia articles. We applied factor analysis to differentiate between editors based on their editing abilities, and then studied the collaboration process in terms of editors' throughput in the creation of GA. To illustrate our method, we analyzed GA in the subcategories of US state parks, children's books and chemical compounds and materials. We found specific collaboration patterns that were used to create GA. The key finding was that an editor with strong content-shaping ability worked continuously for several months to secure GA acceptance. Sometimes another editor with strong copy-editing skills would be recommended to strengthen

the writing quality; at other times, if the content-shaping editor was also strong in copy editing, he or she might also be solely responsible for the writing quality.Years might pass before this editor started to work, and low-quality editors (those with weak editing abilities) performed scant editing activities on the article. These editors were the reason for the differences in the sizes and diversity of teams.

2. We identified the collaboration patterns involved in creating good-quality Wikipedia articles.

   With the same motivations, and based on the need to identify the optimal editing patterns of Wikipedia articles, we proposed another novel approach to exploring the collaboration patterns in the creation of GA articles. We used time series clustering to find the pattern of creation of an article, and then identified the phases through which a GA article passed before being nominated as such and the types of editors involved in each phase. We illustrated our approach using GA in three Wikipedia categories, and found different collaboration patterns for these categories. The results of our study build on previous work by identifying the types of editors involved in the different phases of article development. Our findings can be used as a reference model for recommending how to create more GA in the same Wikipedia category via collaboration. Our method can easily be extended to other Wikipedia articles in different categories.

3. We designed an alternative system to allow for copyrighted sharing of collaborative knowledge content.

   This work addressed the challenges arising from copyright sharing when creating knowledge content through open collaboration. We

considered the case where teachers want to make use of other people's copyrighted work protected to produce teaching materials. We used blockchain technology to design a smart contract for teachers who are willing to share their materials, which facilitates collaboration through the automatic recording of authorship when other teachers' materials are used. The public ledger of the blockchain system can also offer secure authorship records without the need for centralized management. We demonstrated our design using a real-life scenario to exemplify its practical usage. Authorship records held on the blockchain can be used as evidence for claims of multiple authorship for content created through open collaboration.

6.2 Future Direction

In this section, we describe some potential areas for future research. The present research could be extended in the following ways:

1. Open collaboration based on inviting people to work together online: Since different collaboration patterns are involved in this type of approach, we suggest the creation of more quality output by requiring collaboration to follow a particular pattern. However, open collaboration relies on volunteers, and it is not clear whether restrictions should be imposed on the collaboration or what their impact on participation would be. This could be a direction to be considered for future research.

2. Attribution of contributions during open collaboration on knowledge content:

Another research area of interest explored in this work is contribution attribution. Since we focused on simple scenarios, the distribution of contributions could be calculated based on file sizes or numbers of content pages. However, the output from open collaboration may involve different types of media, such as photos, videos, and other forms of data, and a simple calculation based on file size might not be fair to contributors. There is therefore a need for a better method of handling contribution attribution.

3. Open collaboration involving people and artificial intelligence:
   In this research, we focused on collaboration between humans in the creation of knowledge content. However, with the advancements in artificial intelligence (AI) technology, new avenues relating to the capabilities of AI and how to make use of AI to create more desired output could be explored to assist open collaboration.

Publications

Journal

Huichen Chou, Donghui Lin, and Toru Ishida, "Understanding Open Collaboration of Wikipedia Good Articles with Factor Analysis", *Journal of Information and Knowledge Management (JIKM)*. 2022 Sept. (*in press*)

Huichen Chou, Donghui Lin, Takao Nakaguchi, and Toru Ishida, "TM-chain: A Blockchain-Based Collaboration System for Teaching Materials", *Journal of Information Processing. (JIP)*. 2022 May. (*in press*)

Conferences

Huichen Chou, Donghui Lin, Toru Ishida and Naomi Yamashita, "Understanding Open Collaboration of Wikipedia Good Articles." In *International Conference on Human-Computer Interaction (HCII 2020)*, Springer, Cham., pp 29-43, 2020.

Huichen Chou, Takao Nakaguchi, D. Lin, and T. Ishida, "A Blockchain-based Collaboration Framework for Teaching Material Creation." In *International Conference on Human-Computer Interaction (HCII 2021)*,

Springer, Cham., pp. 3-14, 2021.

Huichen Chou, Rafik. Hadfi, Donghui Lin, and Takayuki Ito, "Identifying Collaborative Editing Traits and Phases in Good Wikipedia Articles." In *ACM Collective Intelligence Conference*, 2021.

Bibliography

[Arazy and Nov, 2010] Arazy, O. and Nov, O. (2010). Determinants of wikipedia quality: the roles of global and local contribution inequality. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 233–236.

[Arazy et al., 2011] Arazy, O., Nov, O., Patterson, R., and 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 et al., 2015] Arazy, O., Ortega, F., Nov, O., Yeo, L., and Balila, A. (2015). Functional roles and career paths in wikipedia. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 1092–1105.

[Belotti et al., 2019] Belotti, M., Božić, N., Pujolle, G., and Secci, S. (2019). A vademecum on blockchain technologies: When, which, and how. *IEEE Communications Surveys & Tutorials*, 21(4):3796–3838.

[Blumenstock, 2008] Blumenstock, J. E. (2008). Size matters: word count as a measure of quality on wikipedia. In *Proceedings of the 17th international conference on World Wide Web*, pages 1095–1096.

[Bryant et al., 2005] Bryant, S. L., Forte, A., and Bruckman, A. (2005). Becoming wikipedian: transformation of participation in a collaborative online encyclopedia. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 1–10.

[Carroll et al., 1993] Carroll, J. B. et al. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

[Chen et al., 2018] Chen, G., Xu, B., Lu, M., and Chen, N.-S. (2018). Exploring blockchain technology and its potential applications for education. *Smart Learning Environments*, 5(1):1–10.

[Child, 1990] Child, D. (1990). *The essentials of factor analysis*. Cassell Educational.

[Chou et al., 2021a] Chou, H., Hadfi, R., Lin, D., and Ito, T. (2021a). Identifying collaborative editing traits and phases in good wikipedia articles. In *Collective Intelligence Conference, ACM*, pages 1–4.

[Chou et al., 2022a] Chou, H., Lin, D., and Ishida, T. (2022a). Understanding open collaboration of wikipedia good articles with factor analysis. *Journal of Information and Knowledge Management*.

[Chou et al., 2020] Chou, H., Lin, D., Ishida, T., and Yamashita, N. (2020). Understanding open collaboration of wikipedia good articles. In *International Conference on Human-Computer Interaction*, pages 29–43. Springer.

[Chou et al., 2021b] Chou, H., Lin, D., Nakaguchi, T., and Ishida, T. (2021b). A blockchain-based collaboration framework for teaching ma-

terial creation. In *International Conference on Human-Computer Interaction*, pages 3–14. Springer.

[Chou et al., 2022b] Chou, H., Lin, D., Nakaguchi, T., and Ishida, T. (2022b). TMchain: A blockchain-based collaboration system for teaching materials. *Journal of Information Processing*, 30.

[Chunwijitra et al., 2016] Chunwijitra, S., Junlouchai, C., Laokok, S., Tummarattananont, P., Krairaksa, K., and Wutiwiwatchai, C. (2016). An interoperability framework of open educational resources and massive open online courses for sustainable e-learning platform. *IEICE TRANSACTIONS on Information and Systems*, 99(8):2140–2150.

[Crosby et al., 2016] Crosby, M., Pattanayak, P., Verma, S., Kalyanaraman, V., et al. (2016). Blockchain technology: Beyond bitcoin. *Applied Innovation*, 2(6-10):71.

[Dai et al., 2013] Dai, L., Xu, C., Tian, M., Sang, J., Zou, D., Li, A., Liu, G., Chen, F., Wu, J., Xiao, J., et al. (2013). Community intelligence in knowledge curation: an application to managing scientific nomenclature. *PloS one*, 8(2):e56961.

[Daxenberger and Gurevych, 2013] Daxenberger, J. and Gurevych, I. (2013). Automatically classifying edit categories in wikipedia revisions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 578–589.

[Dictionary, 2006] Dictionary, O. (2006). Oxford dictionary online. *Retrieved November*, 20:2006.

[Ellervee et al., 2017] Ellervee, A., Matulevicius, R., and Mayer, N. (2017). A comprehensive reference model for blockchain-based distributed ledger technology. In *ER Forum/Demos*, pages 306–319.

[Emigh and Herring, 2005] Emigh, W. and 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*, pages 99a–99a. IEEE.

[Engel and Malone, 2018] Engel, D. and Malone, T. W. (2018). Integrated information as a metric for group interaction. *PloS one*, 13(10):e0205335.

[Ethereum, 2021] Ethereum (2021). Remix-ide. https://github.com/ethereum/remix-ide (last accessed: 01.08.2022).

[Fedorova and Skobleva, 2020] Fedorova, E. P. and Skobleva, E. I. (2020). Application of blockchain technology in higher education. *European Journal of Contemporary Education*, 9(3):552–571.

[Forte and Lampe, 2013] Forte, A. and Lampe, C. (2013). Defining, understanding, and supporting open collaboration: Lessons from the literature. *American behavioral scientist*, 57(5):535–547.

[Geiger and Halfaker, 2013] Geiger, R. S. and Halfaker, A. (2013). Using edit sessions to measure participation in wikipedia. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 861–870.

[Giles, 2005] Giles, J. (2005). Special report internet encyclopaedias go head to head. *nature*, 438(15):900–901.

[Günther and Chirita, 2018] Günther, V. and Chirita, A. (2018). " scienceroot" whitepaper. *Scienceroot*.

[Guo et al., 2020] Guo, J., Li, C., Zhang, G., Sun, Y., and Bie, R. (2020). Blockchain-enabled digital rights management for multimedia resources of online education. *Multimedia Tools and Applications*, 79(15):9735–9755.

[Hadfi and Ito, 2021] Hadfi, R. and Ito, T. (2021). Exploring interaction hierarchies in collaborative editing using integrated information. In *Collective Intelligence Conference, ACM*.

[Halavais and Lackaff, 2008] Halavais, A. and Lackaff, D. (2008). An analysis of topical coverage of wikipedia. *Journal of computer-mediated communication*, 13(2):429–440.

[Hill, 2011] Hill, B. D. (2011). *The sequential Kaiser-Meyer-Olkin procedure as an alternative for determining the number of factors in common-factor analysis: A Monte Carlo simulation*. Oklahoma State University.

[Hilton III and Wiley, 2009] Hilton III, J. and Wiley, D. A. (2009). The creation and use of open educational resources in christian higher education. *Christian Higher Education*, 9(1):49–59.

[Hou et al., 2019] Hou, Y., Wang, N., Mei, G., Xu, W., Shao, W., and Liu, Y. (2019). Educational resource sharing platform based on blockchain network. In *2019 Chinese Automation Congress (CAC)*, pages 5491–5494. IEEE.

[Huberman and Wilkinson, 2007] Huberman, B. and Wilkinson, D. (2007). Assessing the value of cooperation in wikipedia. *First Monday*, 12(4).

[Hylén, 2021] Hylén, J. (2021). Open educational resources: Opportunities and challenges.

[IPFS, 2022] IPFS (2022). Search engine for the interplanetary filesystem. https://ipfs-search.com// (last accessed: 03.28.2022).

[Jones, 2008] Jones, J. (2008). Patterns of revision in online writing: A study of wikipedia's featured articles. *Written Communication*, 25(2):262–289.

[Kane, 2011] Kane, G. C. (2011). A multimethod study of information quality in wiki collaboration. *ACM Transactions on Management Information Systems (TMIS)*, 2(1):1–16.

[Kittur et al., 2007] Kittur, A., Chi, E., Pendleton, B. A., Suh, B., and Mytkowicz, T. (2007). Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World wide web*, 1(2):19.

[Kittur and Kraut, 2008] Kittur, A. and Kraut, R. E. (2008). Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 37–46.

[Kittur et al., 2009] Kittur, A., Lee, B., and Kraut, R. E. (2009). Coordination in collective intelligence: the role of team structure and task interdependence. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1495–1504.

[Klein et al., 2015] Klein, M., Maillart, T., and Chuang, J. (2015). The virtuous circle of wikipedia: recursive measures of collaboration structures.

In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1106–1115.

[Kodinariya and Makwana, 2013] Kodinariya, T. M. and Makwana, P. R. (2013). Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95.

[Lessig, 2004] Lessig, L. (2004). The creative commons. *Mont. L. Rev.*, 65:1.

[Levine and Prietula, 2014] Levine, S. S. and Prietula, M. J. (2014). Open collaboration for innovation: Principles and performance. *Organization Science*, 25(5):1414–1433.

[Li et al., 2014] Li, H., Zhao, B., and Fuxman, A. (2014). The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing. In *Proceedings of the 23rd international conference on World wide web*, pages 165–176.

[Lih, 2004] Lih, A. (2004). Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. *Nature*, 3(1):1–31.

[Lin and Wang, 2020] Lin, Y. and Wang, C. (2020). Wisdom of crowds: the effect of participant composition and contribution behavior on wikipedia article quality. *Journal of Knowledge Management*.

[Liu and Ram, 2011] Liu, J. and Ram, S. (2011). Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Transactions on Management Information Systems (TMIS)*, 2(2):1–23.

[Marjit and Kumar, 2020] Marjit, U. and Kumar, P. (2020). Towards a decentralized and distributed framework for open educational resources based on ipfs and blockchain. In *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*, pages 1–6. IEEE.

[Mohd Pozi et al., 2018] Mohd Pozi, M. S., Muruti, G., Abu Bakar, A., Jatowt, A., and Kawai, Y. (2018). Preserving author editing history using blockchain technology. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, pages 165–168.

[Nakamoto, 2008] Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review*, page 21260.

[Niederer and Van Dijck, 2010] Niederer, S. and Van Dijck, J. (2010). Wisdom of the crowd or technicity of content? wikipedia as a sociotechnical system. *New media & society*, 12(8):1368–1387.

[Niya et al., 2019] Niya, S. R., Pelloni, L., Wullschleger, S., Schaufelbühl, A., Bocek, T., Rajendran, L., and Stiller, B. (2019). A blockchain-based scientific publishing platform. In *2019 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, pages 329–336. IEEE.

[Nizamuddin et al., 2018] Nizamuddin, N., Hasan, H. R., and Salah, K. (2018). Ipfs-blockchain-based authenticity of online publications. In *International Conference on Blockchain*, pages 199–212. Springer.

[Novotny et al., 2018] Novotny, P., Zhang, Q., Hull, R., Baset, S., Laredo, J., Vaculin, R., Ford, D. L., and Dillenberger, D. N. (2018). Permissioned

blockchain technologies for academic publishing. *Information Services & Use*, 38(3):159–171.

[Ocheja et al., 2019] Ocheja, P., Flanagan, B., and Ogata, H. (2019). Decentralized e-learning marketplace: Managing authorship and tracking access to learning materials using blockchain. In *International Cognitive Cities Conference*, pages 526–535. Springer.

[Orvium, 2020] Orvium (2020). Orvium: The open source and decentralized framework for managing scholarly publications life cycles and the associated data. http://orvium.io (last accessed: 04.2020).

[Petitjean et al., 2011] Petitjean, F., Ketterlin, A., and Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern recognition*, 44(3):678–693.

[Pfeil et al., 2006] Pfeil, U., Zaphiris, P., and Ang, C. S. (2006). Cultural differences in collaborative authoring of wikipedia. *Journal of Computer-Mediated Communication*, 12(1):88–113.

[Putnik et al., 2009] Putnik, Z., Budimac, Z., and Ivanović, M. (2009). A practical model for conversion of existing teaching resources into learning objects. *MASAUM Journal of Computing*, 12:205–214.

[Ren and Yan, 2017] Ren, R. and Yan, B. (2017). Crowd diversity and performance in wikipedia: The mediating effects of task conflict and communication. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6342–6351.

[Robert and Romero, 2015] Robert, L. and Romero, D. M. (2015). Crowd size, diversity and performance. In *Proceedings of the 33rd Annual ACM*

Conference on Human Factors in Computing Systems, pages 1379–1382.

[Schäffer et al., 2019] Schäffer, M., Angelo, M. d., and Salzer, G. (2019). Performance and scalability of private ethereum blockchains. In *International Conference on Business Process Management*, pages 103–118. Springer.

[Senin, 2008] Senin, P. (2008). Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, 855(1-23):40.

[Shi et al., 2021] Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., and Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP Journal on Wireless Communications and Networking*, 2021(1):1–16.

[Stvilia et al., 2008] Stvilia, B., Twidale, M. B., Smith, L. C., and Gasser, L. (2008). Information quality work organization in wikipedia. *Journal of the American society for information science and technology*, 59(6):983–1001.

[Torres et al., 2017] Torres, J., Jimenez, A., García, S., Peláez, E., and Ochoa, X. (2017). Measuring contribution in collaborative writing: An adaptive nmf topic modelling approach. In *International Conference on eDemocracy & eGovernment (ICEDEG)*, pages 63–70. IEEE.

[Völkel et al., 2006] Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., and Studer, R. (2006). Semantic wikipedia. In *Proceedings of the 15th international conference on World Wide Web*, pages 585–594.

[Wikimedia, 2022a] Wikimedia (2022a). Wikipedia assessment. https://en.wikipedia.org/wiki/ Wikipedia:WikiProject Wikipedia/Assessment (last accessed: 03.2022).

[Wikimedia, 2022b] Wikimedia (2022b). Wikipedia: Statistics. https://en.wikipedia.org/wiki/Wikipedia:Statistics (last accessed: 03.2022).

[Wood et al., 2014] Wood, G. et al. (2014). Ethereum: A secure decentralised generalised transaction ledger. *Ethereum project yellow paper*, 151(2014):1–32.

[Woolley et al., 2015] Woolley, A. W., Aggarwal, I., and Malone, T. W. (2015). Collective intelligence and group performance. *Current Directions in Psychological Science*, 24(6):420–424.

[Yang et al., 2016] Yang, D., Halfaker, A., Kraut, R., and Hovy, E. (2016). Who did what: Editor role identification in wikipedia. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 446–455.

[Yarovoy et al., 2020] Yarovoy, A., Nagar, Y., Minkov, E., and Arazy, O. (2020). Assessing the contribution of subject-matter experts to wikipedia. *ACM Transactions on Social Computing*, 3(4):1–36.

[Zhang et al., 2017] Zhang, A. F., Livneh, D., Budak, C., Robert Jr, L. P., and Romero, D. M. (2017). Crowd development: The interplay between crowd evaluation and collaborative dynamics in wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–21.