ABM-based open source data sharing and collaboration model

By

Zixiang Ma

Student ID: 2056647

Supervisor: Prof. Christopher Baber

A thesis submitted to the University of Birmingham

For the degree of MSc in Advanced Computer Science

School of Computer Science

University of Birmingham

Birmingham, UK

August 2020

Table of Contents

List of Figures

Abstract

With the rise of artificial intelligence, mobile Internet and the Internet of Things, the amount of data generated every day has exploded, which has also brought unlimited imagination and commercial application value to the development of society. At the same time, with the development of big data and data storage technology, more and more data can be stored online, and can be shared and exchanged on a global scale. On the one hand, this provides the possibility of collaboration between different people, companies or scientific research institutions around the world, but on the other hand, unlimited data sharing also has potential dangers: the private data or copyright data of the data source may be malicious Steal, security cannot be guaranteed. So the real question is how do we share as much data as possible to allow collaboration while maintaining control over who can access that data?

The project aims to establish a model to study this social issue: how the unlimited sharing of all forms of data can support collaboration and the impact of this sharing on the protection of data sources, which may involve data privacy, copyright and security. This project uses NetLogo, a programming modeling environment, to simulate and model this problem, so as to study the method of data sharing and how different agents collaborate through data. On this basis, some restrictions and protections are added to the model, and the impact of different degrees of restriction measures on data sharing is simulated by changing the values of these variables. Finally, through the analysis and reasoning of model images and data in different situations, the model with the best experimental effect is obtained, that is, the data can be shared quickly and widely, and privacy and security have been sufficiently protected.

Keywords

Data sharing; Data collaboration; Data protection; Privacy protection; NetLogo; Modeling

Acknowledgements

First of all, I would like to express my heartfelt thanks to my supervisor, Professor Christopher Baber, because he gave me a lot of help and support during the project and solved many of my doubts. Without his guidance and advice, this study would not have been completed. Secondly, I would also like to thank my parents for their continued encouragement, support and patience. They supported my dream, sent me abroad to continue my studies, and gave me tremendous financial support. Without them, it would be impossible for me to write this paper here. They are the main source of inspiration and creativity in my life. Finally, I would like to thank my friends and my girlfriend. This year, due to the COVID-19 epidemic, everyone is going through a very difficult period, especially during the lockdown. Without their spiritual support, I might not be able to survive these miserable months.

I especially want to thank every teacher and staff of the School of Computer Science. They provided me with challenging courses and comfortable laboratories so that I could learn a lot of useful knowledge and complete my projects efficiently and quickly. Finally, I would like to express my gratitude to the University of Birmingham. I am very honored to offer me the opportunity to study for a master's degree here. This will be a beautiful period in my life worth remembering.

CHAPTER I: Overview

Introduction

This chapter is the first part of this report and gives a brief overview of the whole report. It mainly includes the project problem to be solved, the background knowledge needed to understand the background of the project, the motivation for choosing the project, a clear and detailed description of the project goals and an overview of the solution structure.

The purpose of the work described in this report is to establish a model that simulates data sharing and collaboration, and to study the impact of some of the model's restrictions on data sharing and security. This method of controlling variables helps to get a perfect data sharing and collaboration model.

The second chapter of the thesis is the literature review part, which introduces the previous work in the academic literature related to this project and provides proof for the project, proving that the problem is a practical and important issue. The third chapter is the model realization part. This part introduces in detail the definition of the model structure and goals, the variables contained in the model, the functions implemented by the model, and subsequent improvements to the model. Readers can understand how the model does not even run the model by themselves. use. The fourth chapter is the model testing part, which introduces in detail the adjustment and selection of model variables, the results of model operation and the sample results obtained. The fifth chapter is the conclusions and shortcomings. This part introduces the results of the model operation in the form of graphs, and compares, analyzes and infers the results, so as to obtain convincing conclusions and proves the results before the model is established. Hypothesis. In addition, it also explained the current shortcomings of the project and the work to be done in the future.

1.1 Background

In the network age, the storage and dissemination of data has become a breeze. Data sharing is the foundation of the development of the digital economy and the digital transformation of social governance [1]. The world is connected through huge network data, and through sharing data for all-round and diversified collaboration [2]. This is somewhat similar to a concept that was introduced in the 1980s: Computer-Supported Cooperative Work (CSCW) [3]. The concept was first proposed by developers and researchers from different fields after analyzing future trends in computing, and it covers a wide range of areas, including software development, scientific research, product design, daily office work and group work [4]. Although the Internet had only recently emerged, scientists and researchers were already aware of the benefits of computer-supported computing for human work, and especially for human-human collaboration [5]. The reason is that by using the computing power of computers, people can perform more complex scientific research, and by using the communication and transmission power of the Internet, people can be in touch with each other anytime and anywhere [6]. By combining the two, people can work together on

the same problem, propose different theories and discuss them from thousands of miles away [7]. This is similar to the World Physics Congress at the beginning of the last century, which brought together leading researchers from around the world each year, where people were able to share information and collaborate to solve many problems, thus advancing society. Only as society has progressed, it's now much easier for people to do the same thing again. To this end, Professor Ronald M. Baecker, co-founder of the University of Toronto, has published a book detailing Computer-Supported Cooperative Work (CSCW) theory and how it can help people collaborate with each other [8]. It can be argued that this concept is a prototype of the current theory of information sharing and collaboration, and shows the importance of it.

For example, the European Union attaches great importance to network data sharing, and specifically proposes the "European Data Strategy" guidelines for this purpose. Among them, it is proposed to strengthen the top-level design of cross-departmental data use, improve the level of public sector data sharing, and use legal measures to promote corporate data sharing. In addition, there are research institutions such as health and medical institutions, document management institutions, etc., which have established the basic principles of international health and medical data sharing, scientific research data open sharing guidelines and other data sharing and collaboration guidelines [9]. People all over the world can use data to cooperate or conduct research together without meeting, and the era of big data has also been deepened in different fields. In addition, this year's COVID-19 epidemic also proved the importance of data sharing and collaboration. After the outbreak of the epidemic, China notified the World Health Organization of the epidemic situation and provided the virus sample and structure test report, so that other countries can take precautions and prepare for the epidemic early. After the domestic epidemic was brought under control, China submitted a hundred-page epidemic report and response measures to the World Health Organization, such as blocking cities and streets and isolating suspected patients. Facts have proved that after some European countries (the United Kingdom, Serbia, etc.) have adopted corresponding measures, the epidemic has indeed been effectively controlled, and this is due to the data and experience shared by China. This also proves the importance and necessity of data sharing and collaboration in today's globalization.

But it is worth noting that no matter what field or type of data, it should be shared and collaborated under the premise of reasonableness, legality, and compliance. With the rapid progress of information technology, the leakage and transactions of personal information data are causing great troubles to people's lives. The importance of personal information protection has become more prominent, and so is data. However, due to imperfect privacy protection key technologies, imperfect privacy protection laws and regulations, insufficient privacy protection awareness, and insufficient industry self-discipline, in recent years, Internet technology-based crimes that violate data security and personal privacy have shown explosive growth. For this reason, the most important issue for organizations or institutions in various countries or regions is information and data security, and they have taken actions one after another [10]. For example, the European Union, although it is difficult to compare Europe's competitiveness with China and the United States in the Internet industry, its role and role in the Internet field should not be underestimated and

ignored. With the entry into force of the “General Data Protection Regulation” (GDPR) [4] in May 2018 and the huge US$5 billion fine imposed on Google in July, Europe has been involved in data governance, super network platform antitrust, personal information and privacy protection and other current network governance issues. A series of measures taken in key areas shocked the world. As the Internet Governance Forum (IGF) has been held in Europe for four consecutive years, together with the World Summit on the Information Society in Geneva (WSIS), the European Internet Governance Dialogue (EuroDIG), and the ICANN CEO position held by Europeans, they all show that Europe The unique and multi-level endowments and abilities have already demonstrated Europe's gradual occupation of the global network governance system. The above examples of the "General Data Protection Regulation" (GDPR) cannot be ignored, the GDPR is still being constantly improved, which also illustrates the importance of open source data security and privacy.

1.2 Motivation

This project will use NetLogo (NetLogo will be introduced in detail in the second part) modeling methods to simulate the sharing and collaboration of open source data, and explore the impact of restrictions and protection measures on data sharing. The project is driven by the following considerations:

- The sharing of information and data is a major trend in today's society. The research on data sharing helps to understand the storage and dissemination of network information, and to understand how cross-country and cross-regional cooperation is achieved. This will also help me in the future. Learn how to better share and collaborate with others in your learning career.

- People’s reliance on the Internet is increasing, and more and more users tend to upload their personal data information (text messages, photos, videos, etc.) to the Internet for backup. Whether it is public or private, there are Risk of stealing and leakage. This project helps to understand the relevant standards and methods of data protection, and on this basis attempts to propose more secure methods.

- This project will also illustrate the impact of data protection on data dissemination, especially on speed and scope. Because we cannot blindly pursue stringent safety measures, we must also consider the practicality of these measures, and we must not lose sight of one another.

- There is not much work related to modeling research on pure data sharing and collaboration, especially when security is considered. Few studies have studied this issue by establishing simulation models. Therefore, the project analyzes and studies this problem from a modeling perspective, and encourages researchers to explore other ways to simulate the problem.

1.3 Research question

From the above background and a study of the relevant literature, we can obtain that the main research question is: How can we have a way of sharing information that can protect the privacy of individual information? In addition, there are a number of sub-questions that need to be research.

- How do constraints affect information sharing?
- How do constraints minimize risk of losing private information?

Details of exactly how these questions arose, what the research needs to include, and exactly how to work on them will be described in the literature review section.

1.4 Methodology

The realization of this project is divided into two parts. The first part is the research method to determine the problem. After browsing some relevant documents, we can learn that building models is a fast and accurate method for research on social issues. Of course, there are ways other than modeling that you can use to study and answer the above questions. So why hasn't this project taken a different approach? For example we can take a questionnaire or a social survey, we can interview a lot of people who have been involved in information sharing, whether they are researchers, company employees, or random passers-by, at least they have downloaded or uploaded file data on the web, so even a mostly random crowd can be considered to have been involved in information sharing. We could survey or interview them to get their views on data sharing and collaboration. Or we could invite them to conduct an experiment in which people share information freely, and then we could study their information sharing process. All of these are possible ways of studying the problem, and all of them have merit, but they are not as good for this project as the modeling approach, because they have more obvious drawbacks. Because even if we interviewed people, the results we got may not be usable when we put them together. Respondents to the questions we asked may have disliked or felt insecure, so they refused to answer or gave the wrong answers. The number of questions we were interested in was high, and interviewees may have given up halfway through the interview. Halfway through our experiment a participant would have to drop out and the experiment would be forced to abort. All these uncertainties will directly hinder our study.

Therefore using a modeling approach is definitely the best because we can create an information sharing network with all the different kinds of constraints that we want to see, which is impossible to implement or unavailable in the real world. So the advantage of modeling is that you can create many versions of the problem to compare, which is something that interviews or observations or even experiments can't do, such as COVID-19, which is currently being looked at all over the world, and the results or studies that we can see in terms of predicting the number of infections, predicting peak infections, etc., are all derived from modeling. Because every country is different in terms of population, geography, infection status, and we can create many different versions of the model to study these issues quickly and efficiently. So that explains why modeling is a good idea, not just because we can't recruit at this particular time, but because it's better, more actionable and more accurate than any other method that we think makes sense.

Therefore, this project will use modeling methods to simulate this problem and conduct research. The second part is the modeling process and realization. The modeling method of this project is as follows:

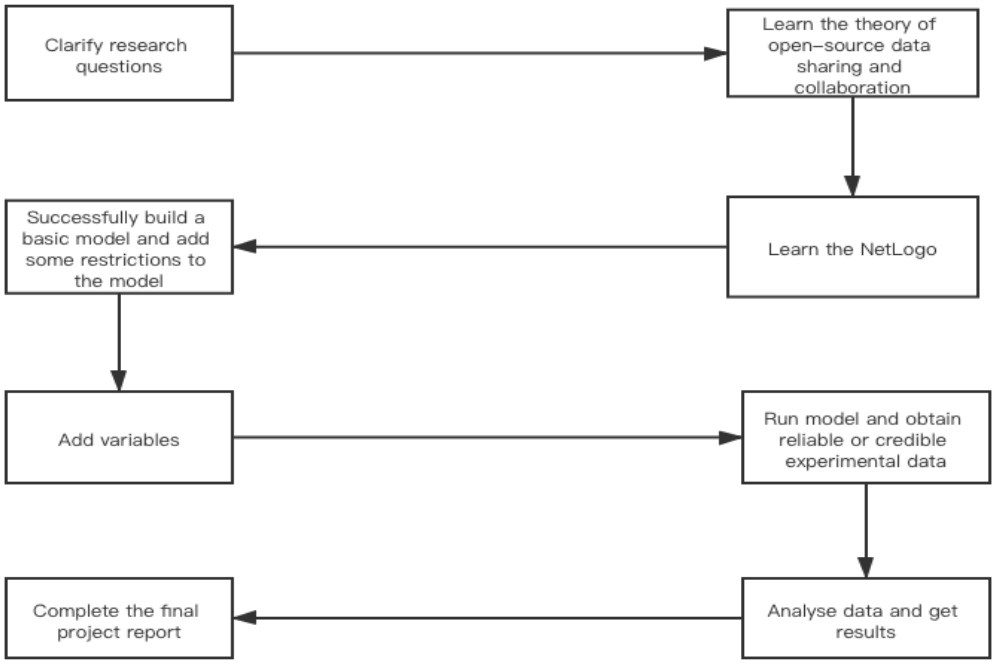

Figure 1. Flowchart of the project.

The method and tools of model realization will be explained in detail in the following chapters.

CHAPTER II: Literature Review

Introduction

This chapter will introduce the collaboration and protection of open source data related to the project, the NetLogo simulation programming modeling environment, and related research on the reference .nlogo model related to this topic. Section 1.1 discusses the current status of open source data collaboration and data privacy protection, as well as several collaboration and setting protection methods discussed in the relevant literature. Section 1.2 respectively introduced the NetLogo-based modeling environment and some models for reference and learning. For this project, it is mainly an information dissemination model and a computer network virus model.

2.1 Collaborate and Protect Open Source Data

With the continuous advancement of Internet technology and information processing technology, we have obtained massive amounts of data today, whether it is on paper or on the Internet [11]. At the same time, with the development of network sharing technology and data storage technology, data can be stored online, and shared and collaborated on a global scale in the form of open source, which also provides convenience for cooperation between different people or companies [12]. But everything should have a limit or a restriction, just like the law, otherwise it means potential danger. The private data or copyright data of the data source may be maliciously stolen, and the security of the data source cannot be guaranteed [13].

The subject of this project research is based on this background, that is, how do we ensure that while sharing and collaborating to the greatest extent possible, we can also protect the security and privacy of data sources. In other words, it is to prevent some malicious people or criminals from reconstructing the source data while making the data open source, and then use this to profit or create confusion.

2.2 Data Collaboration Methods

Data sharing and collaboration are an essential part of the development of today's society, because people in today's society are in the digital world. It is no exaggeration to say that the world is now made up of a lot of data, and of course we are also in it [14]. Therefore, whether you admit it or not, the sharing and collaboration of data is actually ongoing. Weather forecasts, car advertisements, and even WhatsApp and Skype messages are all manifestations of data sharing. However, unlike the data that ordinary people come into contact with every day, the data used for research is usually more and more comprehensive. For example, national medical institutions or health departments require comprehensive, high-quality data from a large number of different groups of people for medical care research on daily people [15]. These data are usually collected through general surveys or

online records of medical treatment information, which means that collecting these comprehensive data specifically for research purposes may be expensive, let alone some scientific research institutions or non-profit organizations. Therefore, the use and collaboration of open source data must comply with ethics, so as to maximize the value of the hard-collected data. There are usually two ways of data collaboration, direct use of data for collaboration

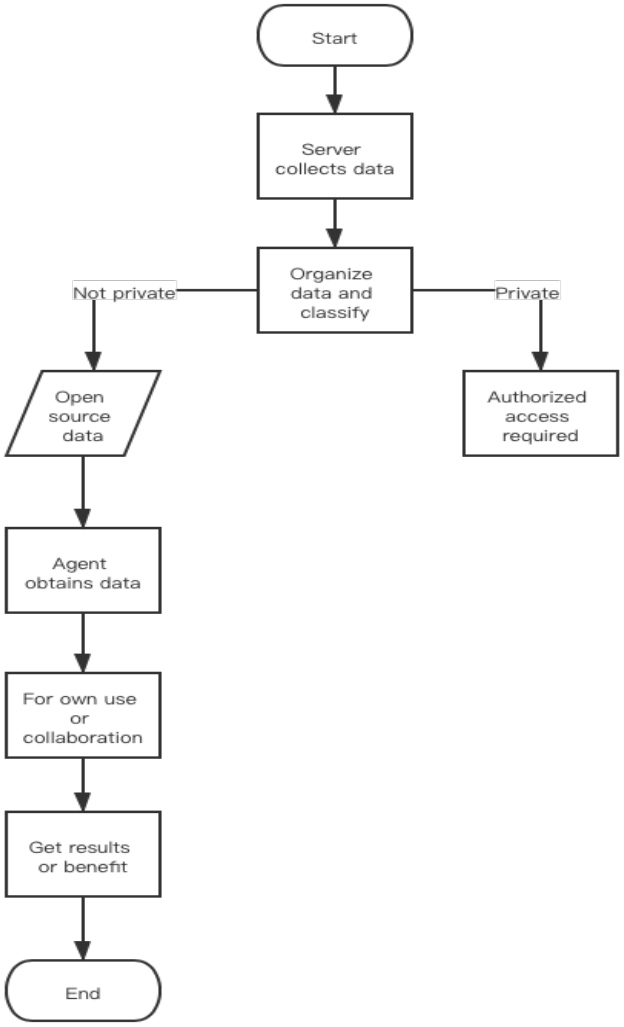

Figure 2. Flowchart of the first method.

and data-based models for collaboration.

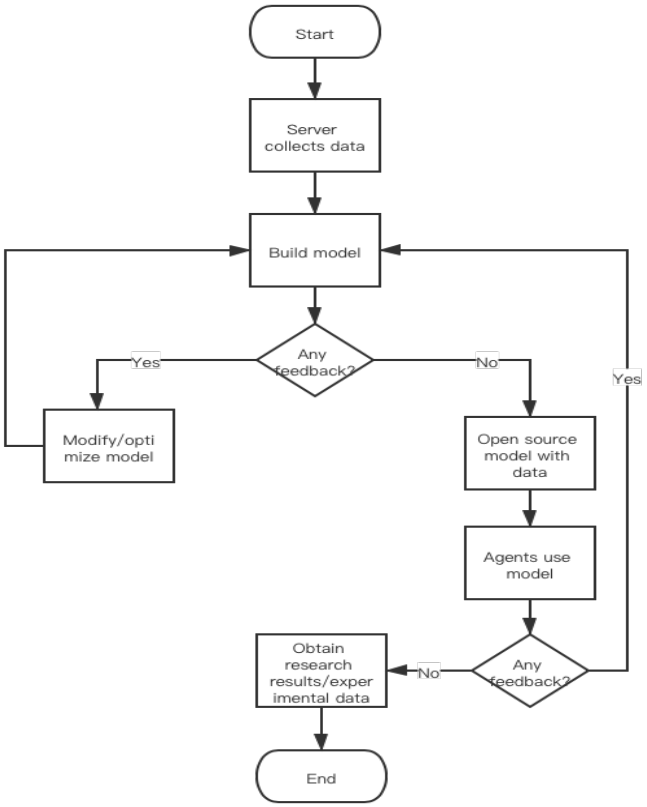

Figure 3. Flowchart of the second method.

2.3 Models of Collaboration Methods

For example, the above-mentioned public health care data information. In order to collect data effectively, researchers usually collect information in the process of providing services, generating bills and insurance claims [16]. This is the most common form of data sharing and collaboration, that is, to obtain the required source data directly from different channels for free or paid.

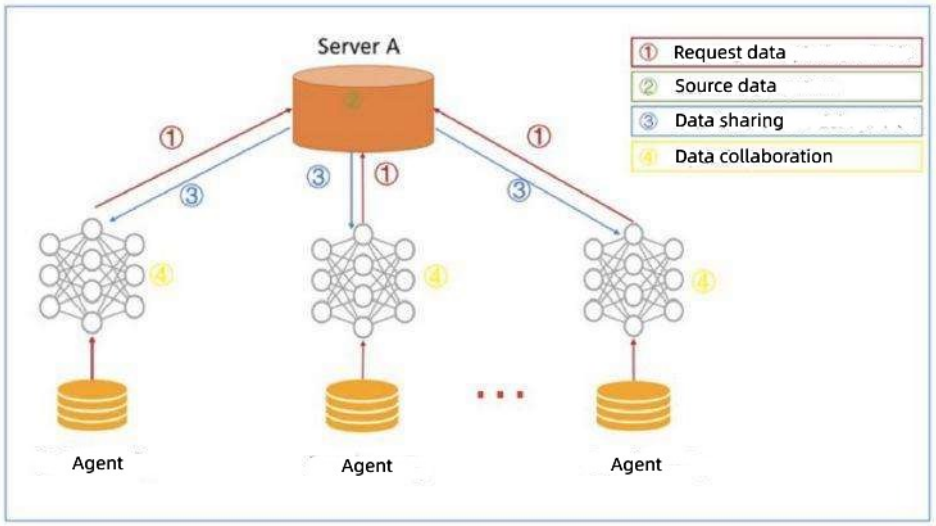

Figure 4. Prototype of the first method.

In addition, we can also upload the collected data to the Internet as part of data analysis and model construction. Therefore, data is shared in the form of models, and different learners or researchers can study or collaborate together, and share research results or make recommendations for modification.

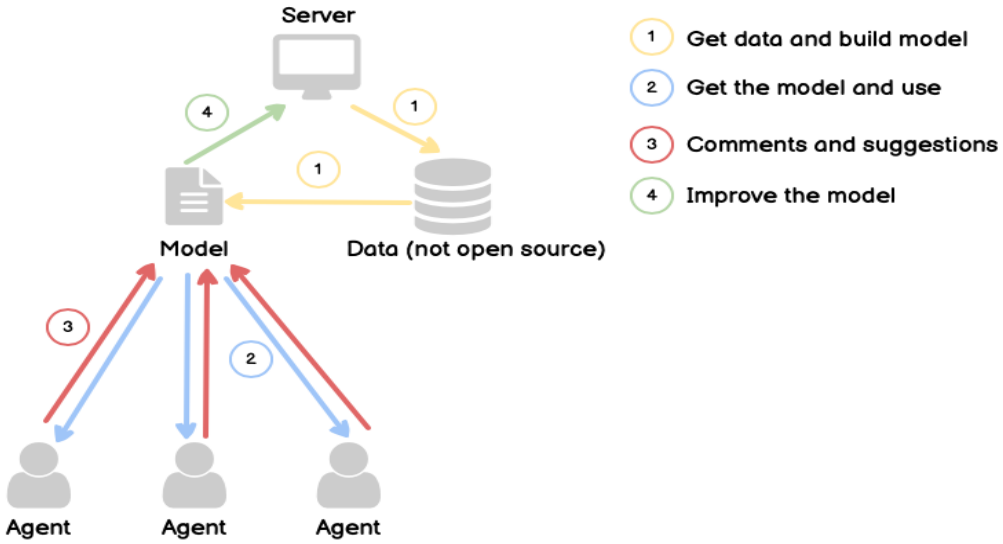

Figure 5. Prototype of the second method.

This is another form of data collaboration.

2.4 Data Protection Methods

With the rapid development of big data and social networks, user privacy data in social networks is facing a huge risk of leakage, including data from companies and some scientific research institutions [17]. Therefore, we urgently need to establish a complete and effective social network privacy data protection method. When the Internet and social

software first appeared, it might be easy, but now, as people get better at using and researching the Internet, the tasks that people have to deal with have become more and more complex, and this is also true in real life. Therefore, the standard of privacy data protection must be higher and higher, and cannot be set in stone, so that whether the processing or use of data is legal can be assessed in real time [18]. Especially in today's faster and faster globalization process, data sharing and collaboration can easily cross national borders and play a key role in the global digital economy. Therefore, organizations such as the European Union have enacted data protection laws, privacy protection laws and other laws to protect the private data of individuals or companies. For data on the Internet, we first need to discuss data protection in an open environment, because most information, articles or data on the Internet are in open form by default. Such as Twitter, Kaggle and Google Scholar. Privacy and security protection at this time are usually based on the role of the person trying to access the resource to develop policies. This method is designed to allow the resource provider to authenticate the user, identify the role the user plays and determine whether the user should be allowed to access the specified resource.

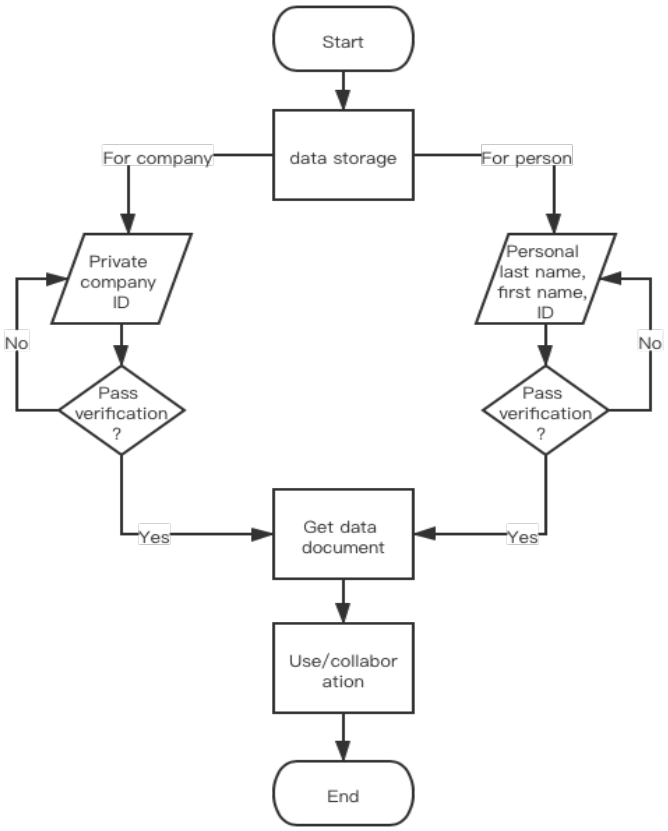

Figure 6. Flowchart of protection method.

This is also the most common security strategy in the current network, similar to the *GOLD project*'s security strategy of middleware [25].

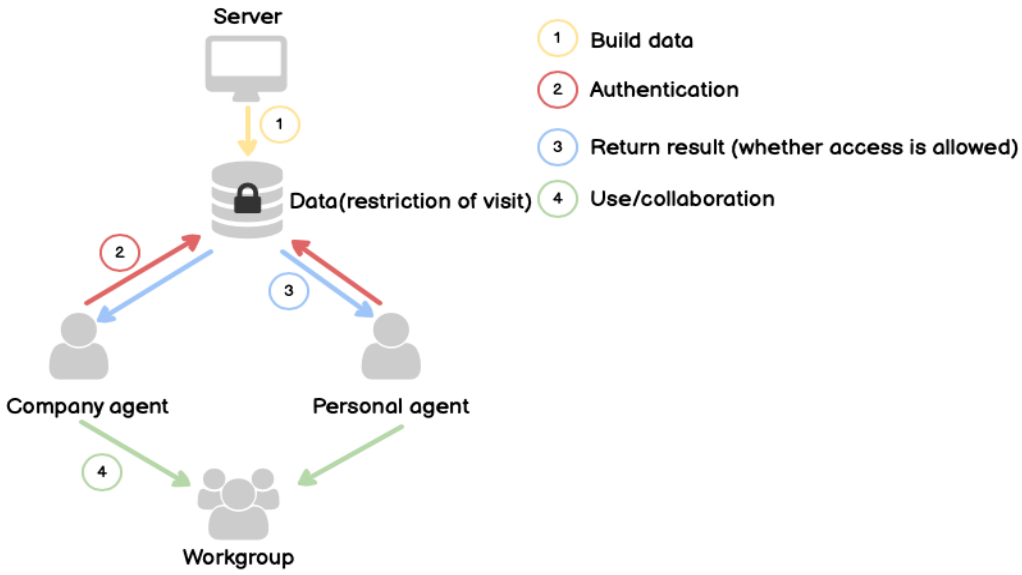

Figure 7. Prototype of protection method.

In addition, there are some more commonly used privacy protection methods, such as anonymity protection and encryption protection. Without exception, these methods can effectively solve the problem of privacy leakage in social networks. However, these methods cannot eliminate the risk of source data being inferred, analyzed and reconstructed by criminals, because these protection measures only protect the data source, but when the data is obtained by different entities or agents, these measures will become invalid. In other words, these measures can prevent undisclosed data from being stolen, but open source shared data cannot be effectively protected. This is also a problem that must be faced after data is open source, and it is also a potential risk. In addition to establishing a data sharing and collaboration model, the project will also try to solve this problem, that is, how to prevent illegal acquisition and reconstruction of data (for example, online advertising agencies, data mining companies, online pharmacies, mortgage companies)? From this we can get a hypothesis: imposing restrictions/protection measures on data will affect the speed and amount of data sharing. This is the work to be done in design and modeling. From the above literature review, we can build a list of requirements for the models, i.e. if we are to build models to study these problems, we need to be clear about what the model needs to contain and how it should work so that we can build the model quickly and accurately. The following is a list of what the model needs to contain:

1. Participants: also referred to as subjects or agents, are the basic units of this model. Because the main question of the model is "How can we have a way of sharing information that can protect the privacy of individual information? ", and the process of information sharing cannot be separated from human participation, because the ultimate purpose of information is to be used by people. And privacy also requires a relevant carrier that cannot stand on its own. Therefore the model needs to include participants and be of many different identities, such as individuals, organizations, companies, etc.

2. Information/Data: The model to be built for this project is about information sharing,

so information is essential. In addition, one of the sub-questions is to investigate how do constraints affect information sharing?Therefore, information is one of the essential elements of the model. Information is just like the books that we usually read, which can be divided into paper version and electronic version. Through reviewing the literature, we can see that most of the current studies on information sharing mainly focus on network information, so this modeling will also use network information.

3. The sharing process: the sharing process is an important part of the model and this process is also the medium that connects the different participants to each other. We mentioned above that this modeling will use network data, which means that this sharing process also takes place between networks. It is easy to see from everyday use that information sharing on the network is uncertain, as each subject can be both a sharer and a recipient, and may even be a third-party organization like Google Scholar that receives data uploaded by users while sharing this data in real time to visitors. Thus individual-to-individual connections are uncertain and connections are randomly generated. This modeling will also use randomly created shared connections to simulate this reality.

4. constraints: The main problem and the two sub-problems we posed earlier both include the condition constraints. Because the reality is that there are restrictions and protections on the sharing of information in a network, it is impossible to share all information without restrictions, then there is no privacy and people's security may be threatened. Therefore, the model should also consider the actual situation and include some constraints in order to establish the information sharing basis so that it can be called a simulation model. Since the model has both an ordinary participant and a 'spy (illegal participant)', the constraints set by the model also contain two kinds of constraints, one for the process of obtaining information by the ordinary participant and the other for the process of the spy trying to steal the information.

The above four points are the basic elements of the model to be built in this project, how to implement these elements and how these elements work will be described in detail in the model implementation section. The next part will introduce the modeling tools and learning models used in this project.

2.5 NetLogo Modeling Environment

NetLogo is a programmable modeling environment for simulating natural and social phenomena. It is also a free downloadable agent-based software package [19].

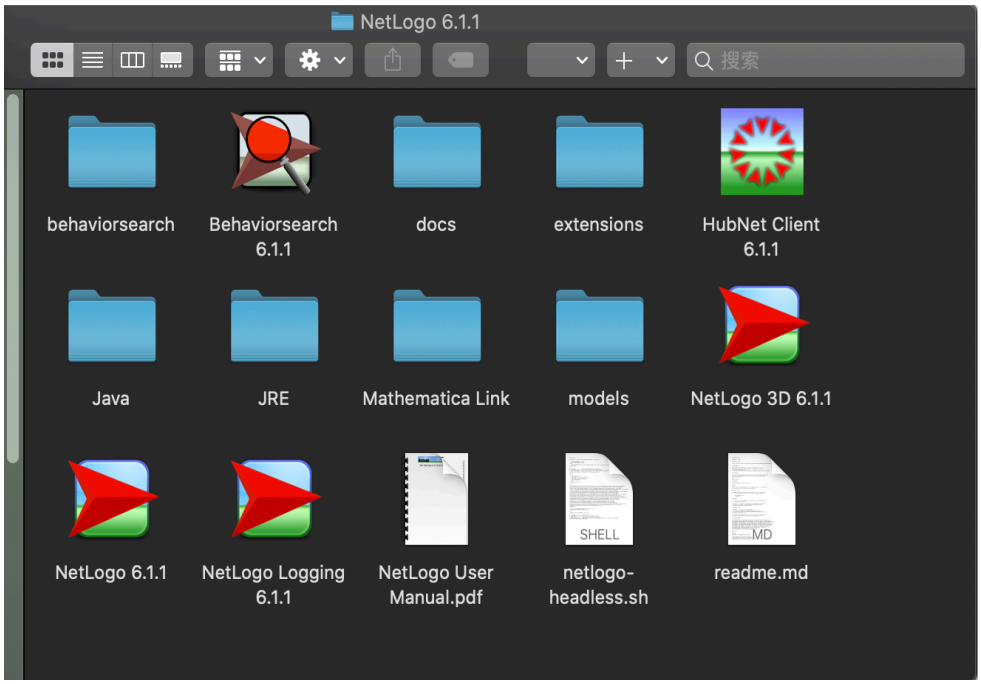

Figure 8. NetLogo download package.

The software package is the Center for Connected Learning and Computer Modeling (CCL) of Northwestern University under the guidance of Uri Wilensky. Created [20]. NetLogo is a functional programming language, which also means that many language sentences are almost read as sentences, which enables even unskilled and untrained users to understand and learn it through examples [21].

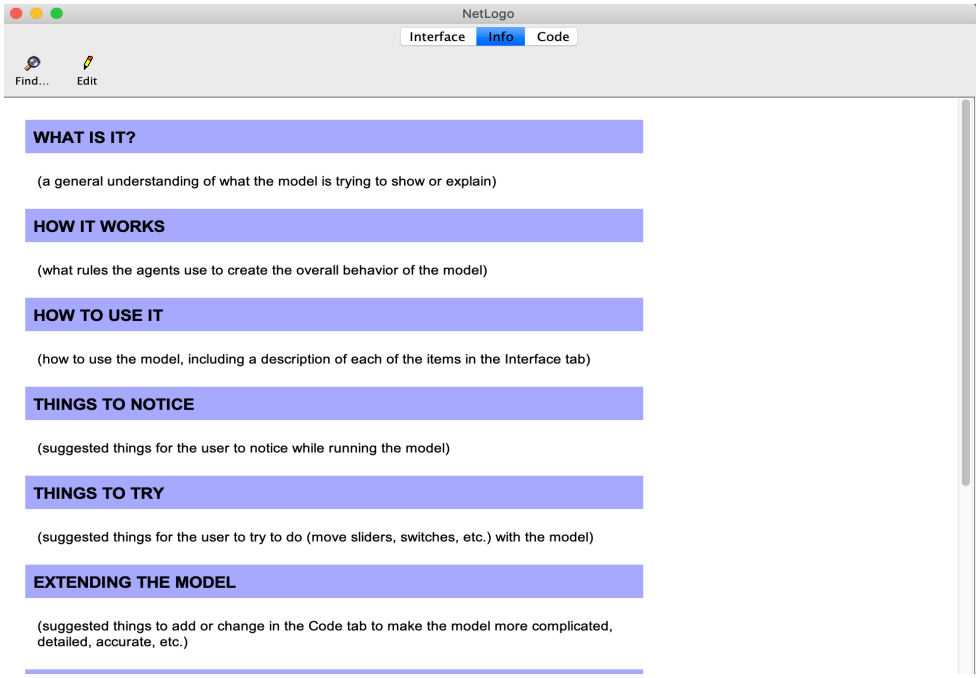

Figure 9. NetLogo information page.

The "turtle" represents the agent, and the "patch" represents a given point in the simulation space. Both of these attributes can have multiple attributes that the user can define, such as age, color, and location [22].

NetLogo is particularly suitable for modeling complex systems developed over time. Modelers can provide instructions to hundreds or thousands of independently operating "agents". This makes it possible to explore the connection between the individual's micro-behavior and the macro-pattern emerging from their interaction. At the same time, NetLogo allows students to open simulations and "play" with them, exploring their behavior under various conditions [23].

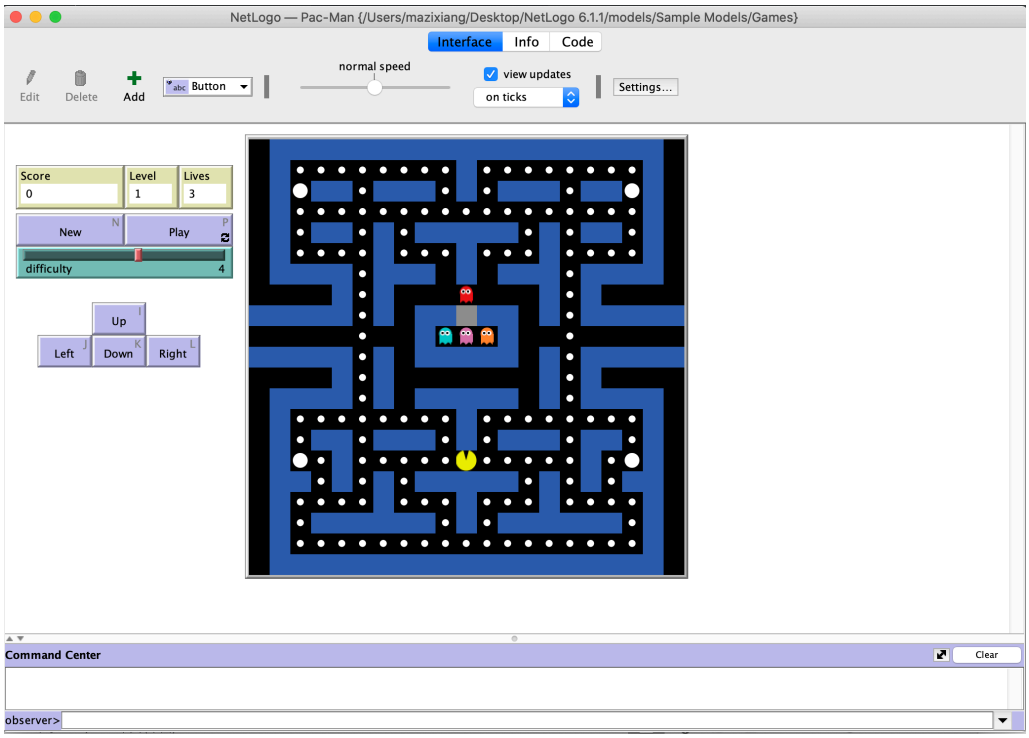

Figure 10. NetLogo running page.

It is also a creative environment that allows students, teachers and course developers to create their own models. NetLogo is simple enough for students and teachers. Although NetLogo may be slower than other tools, it is very easy to use. It supports automatic drawing agents in 2D or 3D form [24]. It provides the possibility of simple user interface construction and provides Many examples and HOWTOs have been added to make it a platform suitable for beginners, but it is enough to be a powerful tool for researchers in many fields. Finally, NetLogo has a lot of documentation and tutorials. The NetLogo homepage (http://ccl.northwestern.edu/netlogo/) includes a download area, model page, sample downloadable extension, user manual, FAQ, and links to various resources. It also comes with a "model library", which contains a large number of pre-written simulations that can be used and modified.

Figure 11. NetLogo official website.

These simulations address the content areas of natural sciences and social sciences, including biology and medicine, physics and chemistry, mathematics and computer science, and economics and social psychology. The above advantages of NetLogo perfectly fit the needs of this project, which is the reason why other modeling tools such as Matlab and Netica were not chosen for this project. As a result, NetLogo has been widely used by scholars in the fields of computer and ``hard'' science from elementary school students to society [25].

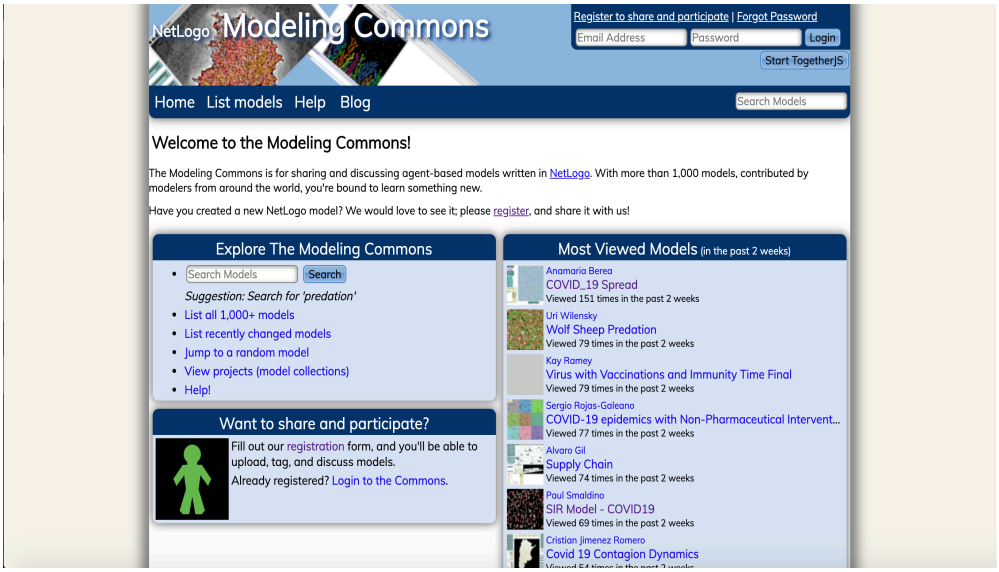

Figure 12. NetLogo community website.

The online community page contains models built by a wide range of representatives from all levels of these people.

NetLogo was first released in 1999, and the latest version (as of this writing) is version

6.1.1 (September 26, 2019). NetLogo can actually run on any of the most popular platforms today-Microsoft Windows, Mac OS X and Linux. The environment used in this project is version 6.1.1 under Mac OS 10.15.

2.6 Related Model Learning

In addition to learning Netlogo's official user manual and sample programs, this project also learned and referred to two existing models and their documentation, namely: 'Virus on a Network' [16] and 'An Agent-Based Model of Crowd Evacuation: Combining Individual, Social and Technological Aspects' [17]. Because information sharing (especially when considering gossip) can be modelled as a spread that operates a little like viral infection in real life and on computer networks, with some agents being able to respond to bits of information. The 'Virus on a Network' model demonstrates the spread of a virus through a network.

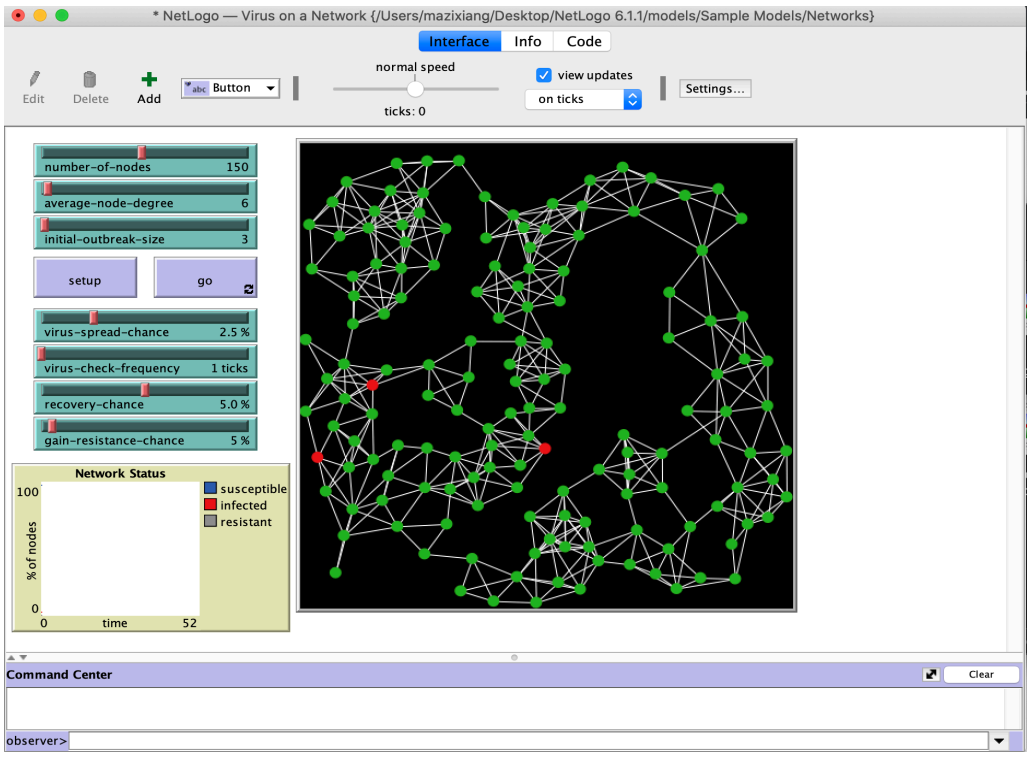

Figure 13. Virus on a Network model.

Although the model is somewhat abstract, one interpretation is that each node represents a computer, and we are modeling the progress of a computer virus (or worm) through this network. Each node may be in one of three states: susceptible, infected, or resistant. In the academic literature such a model is sometimes referred to as an SIR model for epidemics.

Each time step (tick), each infected node (colored red) attempts to infect all of its neighbors. Susceptible neighbors (colored green) will be infected with a probability given

by the VIRUS-SPREAD-CHANCE slider. This might correspond to the probability that someone on the susceptible system actually executes the infected email attachment. Resistant nodes (colored gray) cannot be infected. This might correspond to up-to-date antivirus software and security patches that make a computer immune to this particular virus. Infected nodes are not immediately aware that they are infected. Only every so often (determined by the VIRUS-CHECK-FREQUENCY slider) do the nodes check whether they are infected by a virus. This might correspond to a regularly scheduled virus-scan procedure, or simply a human noticing something fishy about how the computer is behaving. When the virus has been detected, there is a probability that the virus will be removed (determined by the RECOVERY-CHANCE slider). If a node does recover, there is some probability that it will become resistant to this virus in the future (given by the GAIN-RESISTANCE-CHANCE slider). When a node becomes resistant, the links between it and its neighbors are darkened, since they are no longer possible vectors for spreading the virus. This is similar to the spy node in the "open-source data sharing with privacy / security protection" model established in this project. The difference is that Resistant node represents a node that has a probability of not being infected, that is, unilaterally rejects the spread of information from neighboring nodes. This is also one of the independent variables of the model. The purpose is to study the influence of different resistance probabilities on the time required for the virus to spread to all nodes. The spy node means the node that wants to obtain data as much as possible, that is, the data source itself has access restrictions, and the spy node wants to try to break through or bypass (such as accessing secondary agents) these restrictions to obtain as much data as possible. Therefore, the degree of restriction imposed by the data source is one of the independent variables of the model. The purpose is to study the influence of different degree of restriction on the time it takes for the spy node to obtain the data.

The 'An Agent-Based Model of Crowd Evacuation: Combining Individual, Social and Technological Aspects' model addresses a challenge by combining individual,

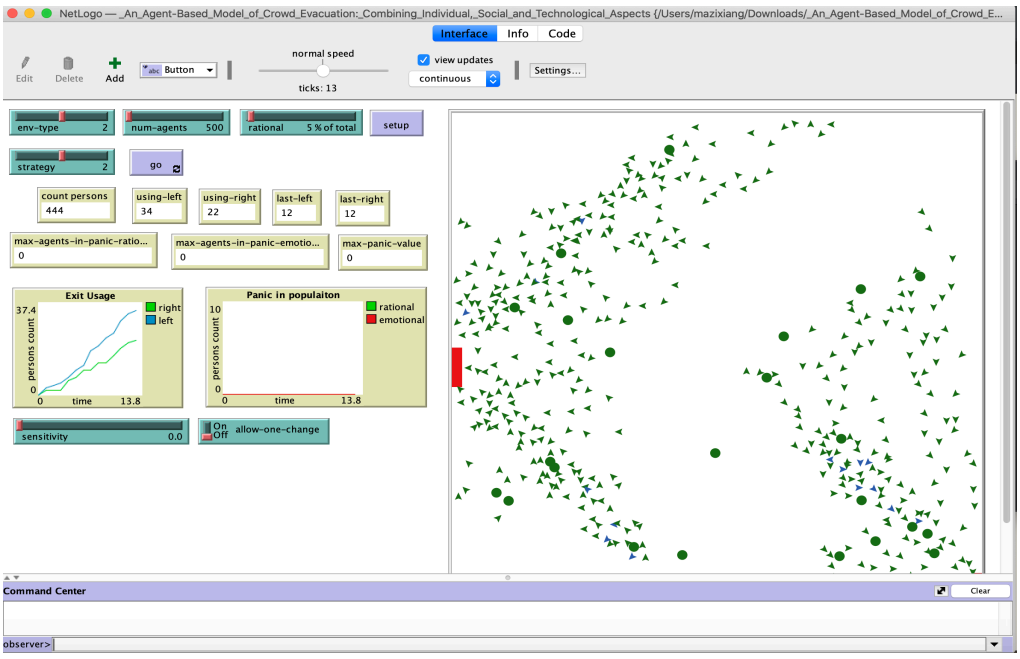

Figure 14. An Agent-Based Model of Crowd Evacuation.

social and technological models of people during the evacuation process, that is, evacuation modeling and simulation are usually used to analyze complex scenarios when the sex is high, analyze various possible results as the situation develops. Combining different aspect categories into a unified modeling space, while concentrating all these aspects on a common body-based modeling framework and a grid-based hypothetical environment. By simulating these models, we can gain insight into the effectiveness of several interesting evacuation plans. The basic setting of the model is that when an agent exits from one of the two exits, our model explores the relationship between panic (or non-panic) and decision-making. In addition, some attempts can be made, such as using different types of environments, different population densities and percentages of rational subjects, and three exit behavior strategies.

One of the above two models is an official open source model and the other is a model created for community users. It not only includes some methods of using NetLogo, but also involves methods of model expansion, and has a certain connection with the theme of this project. Therefore, I chose the above two models as the learning and reference in the early stage of modeling.

In addition to these two models, this project also referenced the information sharing model and the gossip model, which was done to increase understanding of the virus model. Many people use virus models to study information dissemination or gossip and use them as a basis to build their own models. So virus-spread is a good basis for a model, and I will use a virus-spread model as a basis to build my own model. This is the basis for my own model.

To conclude, this project will study how we can build an open source data sharing model and add protections to achieve data collaboration and protect the privacy and security of individuals and companies. And the impact of varying degrees of restriction measures and different numbers of agents on the rate of information dissemination. The documentation field mainly focuses on data sharing and privacy protection. Some of the methods listed have different options for projects with different research content. In terms of models, although there are a few information dissemination and virus dissemination models for beginners to learn, the models for data collaboration and protection are extremely limited, and there are no existing similar models for reference. Therefore, during the course of the project, I need to build my own model and make continuous modifications to achieve satisfactory results.

CHAPTER III: Model Implementation

Introduction

This section is the main part of the report, which describes the implementation process of the model, consisting of goals, tools, framework building, functional implementation, and adding variables. It is important to note that the NetLogo modeling tool is more than a mere stack of components and images, the dynamic model is implemented by code, so the following section will be accompanied by a presentation of some of the key code so that the reader can better understand the model building process and the implementation of each step.

3.1 Description of Objectives

The goal of the model is to simulate the process of data sharing and collaboration and to add protections to the data source to investigate if the protections are effective and if they have an impact on data sharing. To accomplish these goals, the model is divided into two versions, a restricted version and an unrestricted version. The unrestricted version enables data sharing, but that's about it. The restricted version has the option to add protection to the data source as well as to the spy node, and has a line graph that can be used to further investigate the impact of added protection. Therefore, we need to give the assumption before the model is built that applying protection to the data source can be effective in protecting the data, but it can also have an impact on the speed of data sharing. This hypothesis will be demonstrated in the results section of the running test.

3.2 Tool

As explained in the literature review section, NetLogo is the modeling tool of choice because of its clear and concise pages and powerful modeling capabilities, as well as its official repository and open source community for beginners to use. In addition, a review of most of the literature on modeling shows that Matlab is used more often for modeling mathematical problems, mechanical engineering, aerospace, and deep learning. This is due to the fact that Matlab is an excellent tool for scientific computation, as all variables are matrix objects, and it uses matrix operations instead of loops, which makes it fast. In addition, Matlab has a simple syntax and is the closest scientific computing language to a general-purpose language. It also supports various language extensions such as python, c, cuda, and so on. However, the disadvantages of Matlab are also obvious, in terms of functionality, its loops are very slow, and the model to be built in this case is designed to simulate real-time changes in data sharing, which requires constant looping of the code (repetitive running of the go button), so using Matlab is very inappropriate and will slow down the running of the model. In addition, Matlab is strictly a software, the whole installation needs 10~20GB, and the running core (similar interpreter) is several hundred meters, which makes the developed program less portable. NetLogo is more like a

simulation modeling environment, you only need to download some basic configuration files (107MB) from the official website to use it, and there are no high requirements for the operating system, version, CPU performance and GPU performance. It can be said that after getting a .nlogo model file, the reader can be up and running in a few minutes, saving a lot of time on software download, installation and configuration. Also, NetLogo is very suitable for solving social problems and phenomena, and its own model library has many models based on real social problems, which is very suitable for the theme of this project. In summary, NetLogo's simulation modeling environment will be used for this modeling.

For flowcharting, Microsoft Office Visio was chosen, a software from Microsoft that helps IT and business professionals easily visualize, analyze, and communicate complex information. Visio is the standard software used by most papers, journals, academic journals, etc., so this project will follow that standard.

3.3 Functional Implementation

This section describes the implementation of the model base functionality. The unrestricted version of the model is a data sharing process, while the restricted version of the model involves more variables, so this section mainly introduces the implementation of the unrestricted version of the model.

First of all, let's make it clear that data sharing is a process, and it is a process that needs to be done by many subjects, so first we need to create many subjects, called agents in NetLogo. We represent the agents by creating a scalable number of nodes, distinguished by different colors.

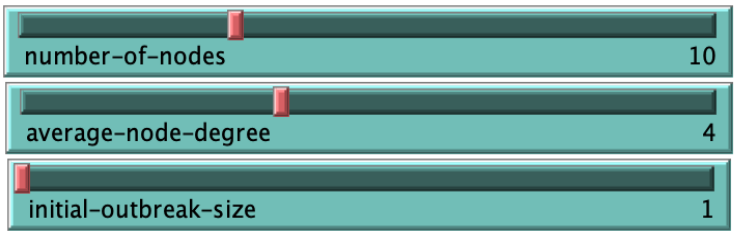

Figure 15. Number of nodes variable.

As shown above, 'number-of-nodes' is used to control the total number of nodes included in the model, including data source nodes and spy nodes. 'average-node-degree' indicates the degree between the nodes and is used to control the total number of connections in the model. 'initial-outbreak-size' is used to control the number of data source nodes, the default case is 1, i.e. all the data fetched by the nodes comes from one data source.

The blue color represents the data source nodes and the normal nodes that have already acquired the data, which can be distinguished by the dynamics of the model. In addition, the white color represents the nodes that are waiting for data to be acquired and the yellow color represents the spy nodes.

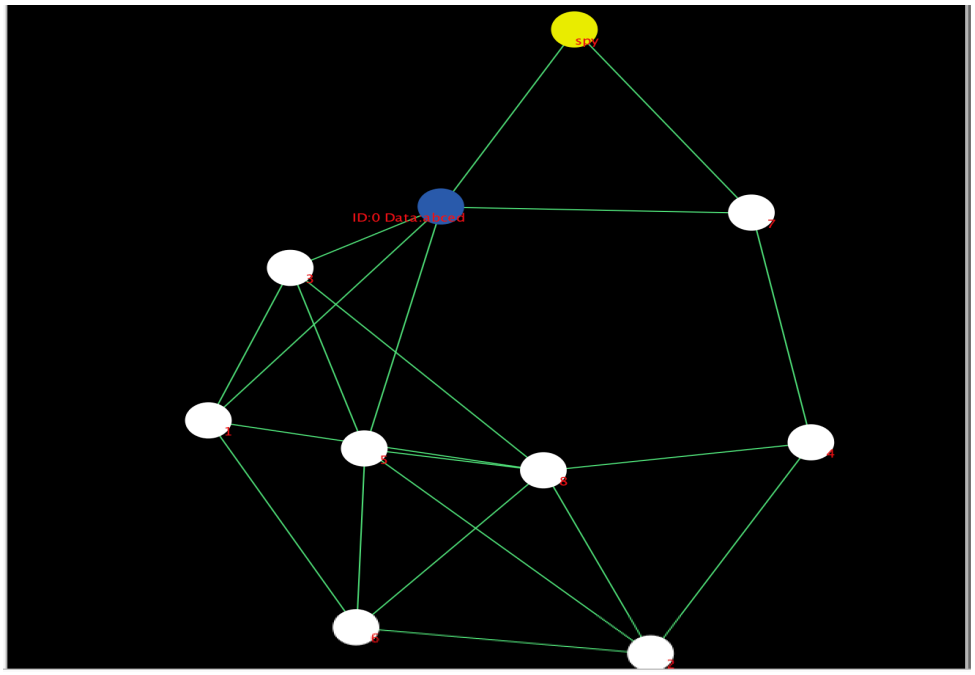

Figure 16. The initialized model.

These nodes are recreated each time we initialize them with the 'setup' button, and have a new location to ensure that the dynamic model is random and can be run repeatedly. Once the nodes are created, we need to create a connection between them. If a connection exists between two nodes, it means that they can communicate with each other and transfer data between them. In order to better simulate the process of data sharing, whether connections are created between nodes or not is random, because in reality, it is impossible for every subject in the network to establish connections with all other subjects that are in the same network at the same time, so the creation of connections in the model is also random. In addition, the model also considers the case of secondary, tertiary agents. This is because in real life, the way we get data is usually not by directly accessing the relevant source database, but by accessing the institution or organization that owns the data. For example, if we want to find out the weather conditions for the coming week, we usually get the information from weather forecasts or website searches, and it is almost impossible and impossible to obtain data directly from the Met Office. Nodes that do not have a direct connection to the data source cannot obtain data directly from the data source, but only from the secondary agent to which they are connected, and therefore only partially. The settings related to the data obtained by the nodes are described in the next section 'Set variables'. We then set the data that the data source node owns, i.e., for sharing, and number all the nodes for subsequent observation and analysis.

Finally we add run logic to the model. The first step is to click the 'setup' button to initialize the model and then click the 'go' button to run the model. The first run of the model will start at the data source node and will attempt to share data with other nodes directly connected to it. This process may or may not work, because we set the probability of success for this process, and in real life, we are not 100% sure about the success of data

acquisition and sharing. NetLogo can be set to run in a certain mode, if the 'go' button is set to repeat mode, then After a click, the model will start running and will continue to run repeatedly until the termination condition is reached. On the second run, the model will start with a node that already contains data and repeat the above operation of sharing data to other nodes directly connected to it, also containing the data source node. Again, this process may or may not work, and there may or may not be any nodes with data. The exact run time of the model is affected by the total number of nodes and other variables described below. The model will stop running if all nodes have data (spy nodes may return empty-handed), regardless of the number of nodes. A run-down model would look like the following.

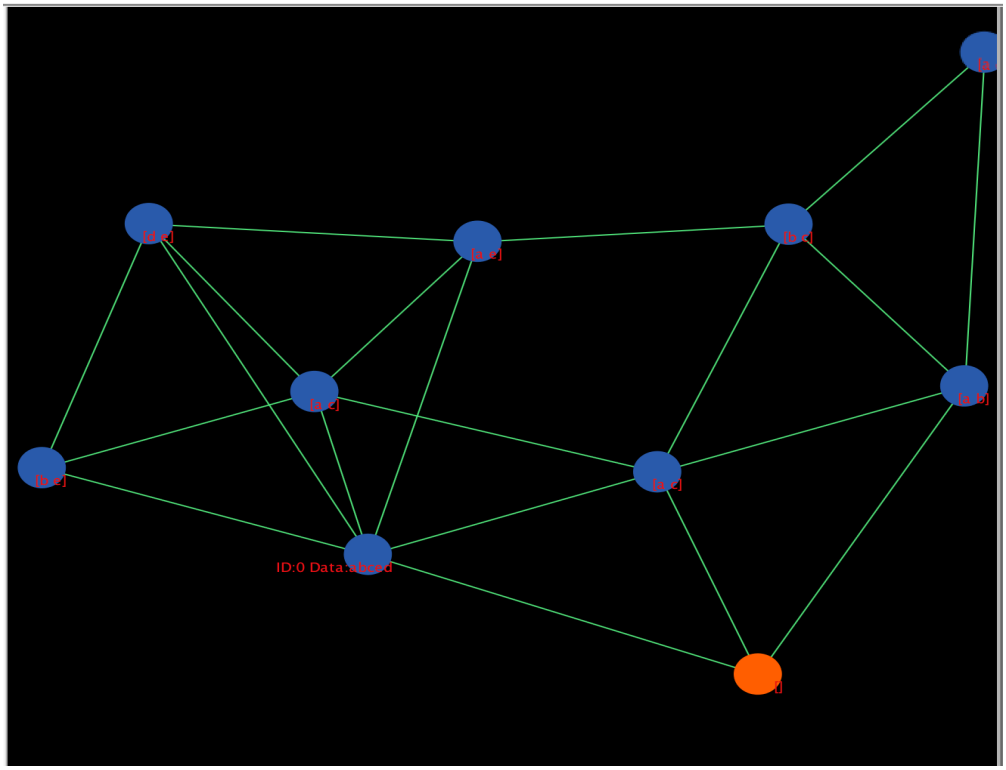

Figure 17. Completed runs of the model.

This is an introduction to the basic functional implementation of the model, which focuses on the simulation of data sharing and collaboration processes. Restricted versions of the model are described in detail below.

3.4 Set Variables

This section complements the functional part of the model, as simply implementing the data sharing and collaboration process is only the unrestricted version of the model. Therefore, we also need to set up a restricted version of the model to facilitate the next step, so we need to add some variables to the original model to implement the function of adding restrictions.

Figure 18. Protection variables.

As shown in the figure above, we added the limit-tolerance and protection-tolerance variables to the model. The Iimit-tolerance is used to impose restrictions on the data source, i.e., after an agent requests access, the Iimit-tolerance variable applies protection to the sharing process of the data source data as a way to secure the data and initially filter out some insecure access sources. This is similar to the firewall settings on Windiws systems, which can be optionally turned off or on, and can perform basic protection functions. However, unlike the Iimit-tolerance variable, which can change its value, i.e., the protection of data sharing is divided into degrees, the larger the value of Iimit-tolerance means that the more tolerant the protection is, the more relaxed the restrictions are, and the easier it is for agents to gain access to the data. Conversely, when we lower the value of Iimit-tolerance, it also means that the data source is more scrutinizing the access request and each agent is less likely to get access to the data, so the protection the data receives is enhanced in disguise. As explained earlier, NetLogo's interactive components are implemented with bound code, and the slider is no exception. Therefore, we can consider the implementation of this feature in the form of a random number, i.e., each time the model executes the 'go' command, each node will be randomized with a value between 0 and 100 to obtain a random number. We then compare this random number with the value of Iimit-tolerance we set, and when the random number is less than the value of Iimit-tolerance, we consider the node to be within the tolerance range of the data source, indicating that the node has gained access to the data. If the random number is greater than the value of Iimit-tolerance, we consider the node to be outside the tolerance range of the data source, marking it as unauthorized and unable to access the data. The next time the 'go' command is executed, the node that did not obtain the data is re-randomized, and the process is repeated thereafter until all nodes have obtained the data.

The number-of-data-received variable is set on top of the Iimit-tolerance variable and represents the amount of data that each data node can acquire (except spy nodes), in order to more recreate the real situation of data sharing and improve the accuracy of the model. Because the reality is that when we share and collaborate on data on the network, we often need only a portion of the data from a large database of data sources, and rarely use all of the data at once. Google Scholar, for example, is a free searchable Google web application for academic articles that includes the vast majority of the world's published academic journals, articles, papers, books, and abstracts, and is an easy way to search a wide range of academic literature. When we access information through Google Docs, we usually search for the literature we want by keyword search, it is impossible to access the Google Docs database directly and then find them one by one, which is like a fantasy. Other data such as images, weather forecasts, etc. are also accessed by accessing one part of the data

source to get the information you want. So considering the practical factors, it is necessary to add number-of-data-received variables to the model as part of the simulation. However, it should be noted that due to the uncertainty of the agents represented by each data node, it is almost impossible that some different agents will want to access the exact same data. Therefore, although we set the amount of data that can be fetched by the nodes, the content of the data fetched by each node is random and may vary or be duplicated, depending on the total amount of data from the data source. This also provides a more realistic simulation of data sharing and collaboration.

The protection-tolerance variable is set because of the presence of spy nodes, so this variable indicates how restrictive the data source is with respect to spy nodes. Since the spy node is independent of other different nodes and data source nodes, it represents uncertainty. The higher the value of protection-tolerance, the weaker the data source is against spies or criminals, and the more likely it is that a spy will succeed in breaking through the data source's protection. Conversely, a smaller value of protection-tolerance means that the data source is less tolerant of criminals, and these spy nodes will be denied access to the data if they are not careful enough to steal it. But it is worth noting that, even if the spy node lucky to break through the protection mechanism of the data source, is not necessarily to get the data, because the actual situation, the server or agent will be important data or data privacy involved in setting more stringent protection measures. This means that it is very difficult for a spy to get all the data of a server. To simulate this situation, we set a private variable for the spy node, denoted by a random number. Each time the model executes the 'go' command, the spy node will represent the total amount of data stolen by the spy node as a random number in the range of 0 - the total amount of data held by the data source (including both 0 and all data). This simulates the uncertainty of the data stealing process. The reality is that attacks on servers and data sources do not always succeed, even at the risk of committing a crime and being arrested, because people are becoming more aware of data security and data protection in recent years.

In addition, in reality, we may also encounter situations in which we receive sales advertisements, insurance advertisements or scam messages from salespeople, insurance companies or criminals who do not have direct access to our personal information and to whom we have never provided it. This is because these salespeople, insurance companies or criminals get some of our information, such as name, age, etc., from other organizations or platforms, and some other information, such as address, phone number, etc., from some platforms. It is through this kind of data reconstruction that these annoying organizations or individuals have access to all of our personal information that we have made public. In fact, this is no different from committing a crime, as it is illegal to obtain privacy and harass people. Illegal reconstruction of data is an impossible to prevent, because the service provider or data source can only authenticate and limit the first level of agents, but there is no direct link to the second, third or lower level of agents. Therefore, it is not possible to effectively protect and restrict the subsequent data dissemination process. The model also attempts to solve this problem, and we distinguish between two cases when setting up a spy node: the spy node is directly connected to the data source; the spy node disguises itself as a Level 2 or Level 3 agent and is not directly connected to the data source. The first case, as described above, is controlled by the protection-tolerance variable. The second case is

the one we are currently discussing, where the source data is potentially reconstructed by a low-level agent. X represents the total amount of data (excluding duplicates) held by all nodes directly connected to the spy node, which means that the spy node does not always succeed in Reconstructing the source data, as the data used by its higher-level agents may not be all the data, depending on our total number of nodes, number-of-data-received values and other data nodes each time random results. By running the model multiple times we find that with a total number of nodes of 10 and number-of-data-received of 3, the probability of the spy node acquiring different amounts of data is shown in the following figure.

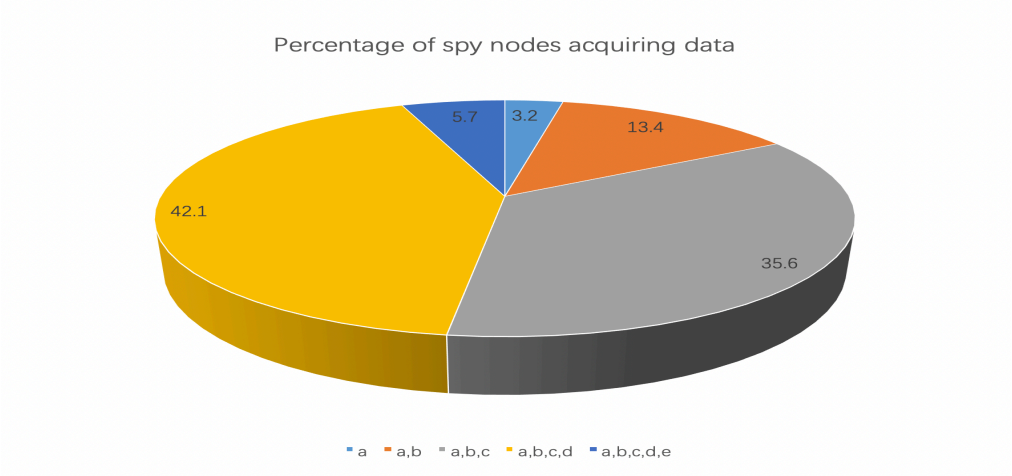

Figure 19. Probability of successful data acquisition by spy nodes.

From the above diagram, we can easily see that the probability of the spy node successfully reconstructing the data (getting all five data) is very low, which means that the protection measures we have set up are very effective, and to some extent, they solve the problem of the source data being maliciously reconstructed, which further protects the security of the data.

Finally, to facilitate further study of the model later, we added line diagrams to the model. The horizontal axis of the line graph shows the runtime of the model in 'tick'. It should be noted that in NetLogo, if the model is dynamic, it is represented as '1 tick' every time the model is executed, and we usually use 'tick' to represent the runtime of the model. This way, if we set up a log when a runtime error occurs, we can see exactly which 'tick' the error occurred at, which makes it easy to modify and debug the code. The vertical axis of the line graph represents the percentage of the total number of nodes, in %. Thus the line graph shows the variation in the number of different types of nodes as a function of time/number of runs. The line graph can provide us with data for quantitative analysis in the subsequent hypothesis validation section to get more accurate conclusions.

At this point, all the features of the model including variables have been added and functionally divided into two versions. The next section will provide a detailed description of the testing and running process of the model.

CHAPTER IV: Model Testing and Running

Introduction

This section is the run and test section of the model. The main objective is to run and test the built model to check if the pre-designed functionality is implemented and to adjust the variables to run the model multiple times to ensure the accuracy, stability and scalability of the model. This is in addition to verifying the proposed hypothesis that different levels of protection can affect the data sharing and collaboration process while protecting the data source. By running the model multiple times many sample data can be obtained, which can be reasoned and analyzed and the results obtained can be used to validate the hypothesis. Finally, the results of the runs and data analysis are used to determine whether the model can be presented as the final outcome of the project.

4.1 Unit Test

This section is a unit test, the main purpose of which is to test the model against the various components such as code, information labels, sliders, buttons, etc. to ensure that there are no errors or defects and that the model runs smoothly.

The first part is the testing of code modules, this part uses the conventional code testing method, i.e. white box testing, to divide the code by functional modules and test them one by one to ensure that they can run independently. Although NetLogo's language style and statements are different from those of a normal object-oriented programming language, the basic data structures such as conditions, judgments, loops, arrays, lists, etc. still exist, so they conform to the testing methods of a normal programming language. The methods used in this test include statement override, which creates test cases so that each executable statement in the program can be executed once.

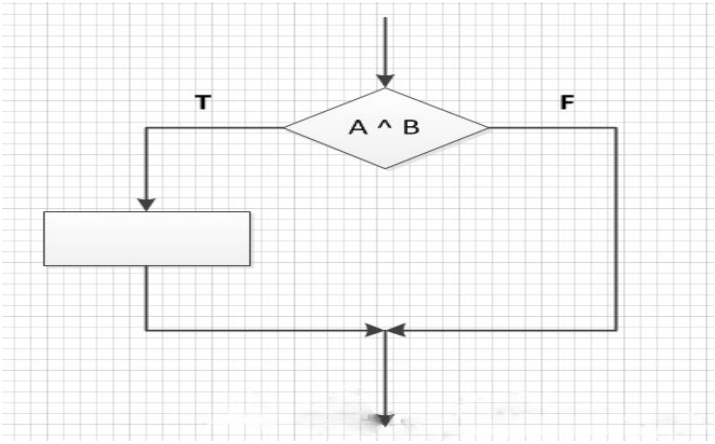

Figure 20. The first test method.

Judgment coverage method, that is, for judgment statements, the use cases are

designed so that the result of the judgment statement is both True and False.

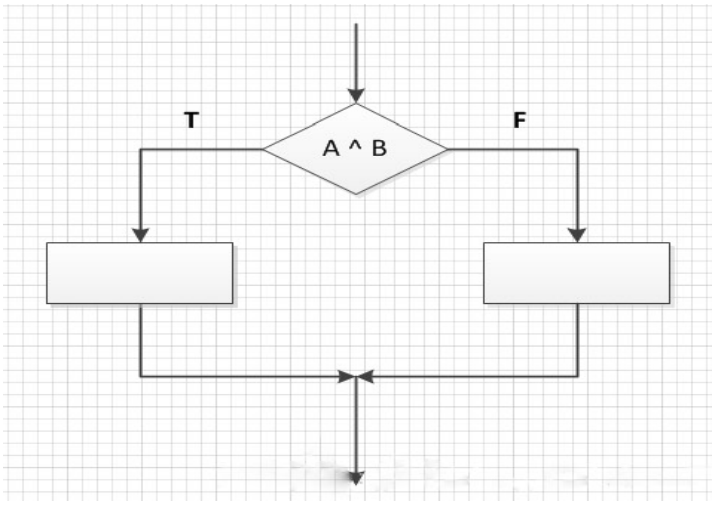

Figure 21. The second test method.

Conditional override, that is, each conditional expression true and false should be taken once when designing the use case without considering the calculation result of the judgment statement.

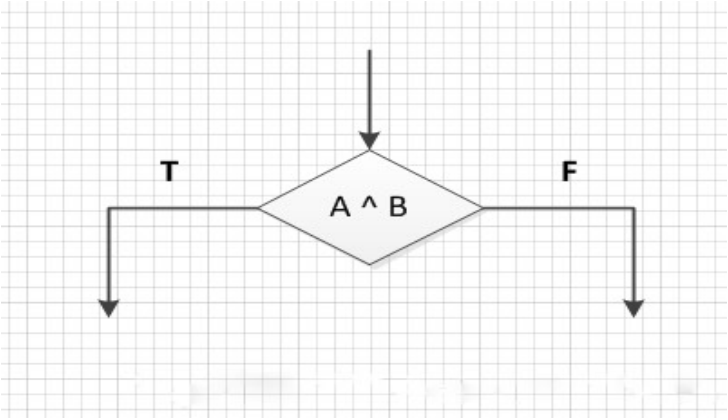

Figure 22. The third test method.

Judgment condition override, that is, when designing test cases, make all possible results of each conditional expression in the judgment statement appear at least once, and all possible results of the judgment statement itself appear at least once.

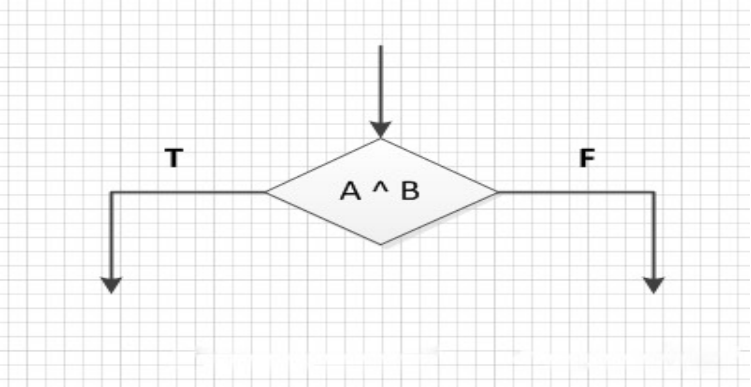

Figure 23. The fourth test method.

As a result of these tests, none of the code modules showed any functional problems. Therefore, we believe that the code part of the model is capable of implementing the pre-designed functions. The next step is to test the sliders, buttons, switches, etc. of the interactive interface. The unit test phase is to test the interactive functionality of these interactive tools, the functionality will be tested in the next phase. After confirming that the slider can be slid freely, the button can be clicked and the switch can be turned off, the next phase of testing is carried out.

4.2 Functional Test

This section is the functional testing section, which focuses on the basic functionality of the model, the functionality of the interactive components of the user interface. The established functions of the model are: setting up different kinds of nodes (data sources, common agents, spy nodes), setting up data types and quantities, building data sharing networks, simulating data collaboration, and applying restrictive protection measures to data sources. This part of the test is mainly through repeatedly running the model to test, because the construction of the data sharing network is random, each node every time to obtain the data is also random, so through a lot of repeated runs of the model can verify the model to see if there are functional omissions and defects. As for the interactive components, since the interactive components in NetLogo need to be bound to the code, the testing of the interactive components, such as the step size of the slider, whether the button works, etc., is carried out in the same way as the unit testing in the previous section, i.e., syntax checking and running tests on the code. After the tests, the model completes its intended functionality and the interactive component is ready to be used.

In addition, since we have added some variables (restrictions, protections) to the model during the model implementation phase, we have also tested the functionality and interaction of these variables to ensure that the model can be further investigated based on the original functionality. Finally, the line graphs were also tested, as the data from the line graphs are needed for the later validation of the hypothesis, so testing the functionality of the line graphs can ensure the accuracy of the data and make the conclusions obtained from the analysis more realistic. The test method is the same as unit testing, because the line graph is also bound to the code, as the program runs in real time changes.

4.3 Integration Test

This section is integration testing, which, as the name implies, is the complete testing of the entire model at the system level to ensure that the model is running as expected with accurate results and data that can be used for inferential analysis. The 'setup' button is first tested to see if it can initialize the model. Since the model is dynamic, i.e. it changes over time, we need to initialize the model at the end of the run to make it easier for the next run. The 'go' button is then tested, as it is divided into single and repeated executions (until the program runs to a pre-set end condition).

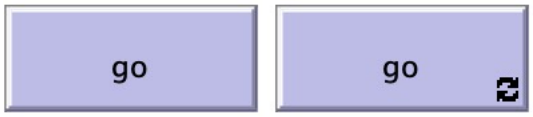

Figure 24. The ‘go’ button.

Normally we would set the 'go' button to repeat when building a model, as this gives us a quicker view of the model and a clear trend of the variables over time on a line graph. But if you want to see the model changes frame by frame or want to record the run data, you need to select the 'go' button for a single run. This test phase therefore tests both cases and ensures that both buttons are available.

Next the model was run multiple times, as the networks and nodes generated by the model are random each time, so the purpose of this step was to check if there were any extremes or conditions that were not taken into account during the programming phase that could lead to distortion of the model. With the assurance that the model will not be distorted or has a very low probability of occurring, the model can now be run multiple times and the data recorded for analysis and hypothesis validation.

4.4 Results of Model Runs

After the model has been tested and ensured that there are no problems affecting the operation, the model can be run multiple times to check that the results are as expected and objective, and to ensure the stability and scalability of the model. Since the model is divided into two versions, we run them separately and record the results, then compare them.

The unrestricted version is first run and tested, unrestricted means that the data source does not have any firewalls or protections, so any agent can access the data at any time and any place, i.e., a fully open source state.

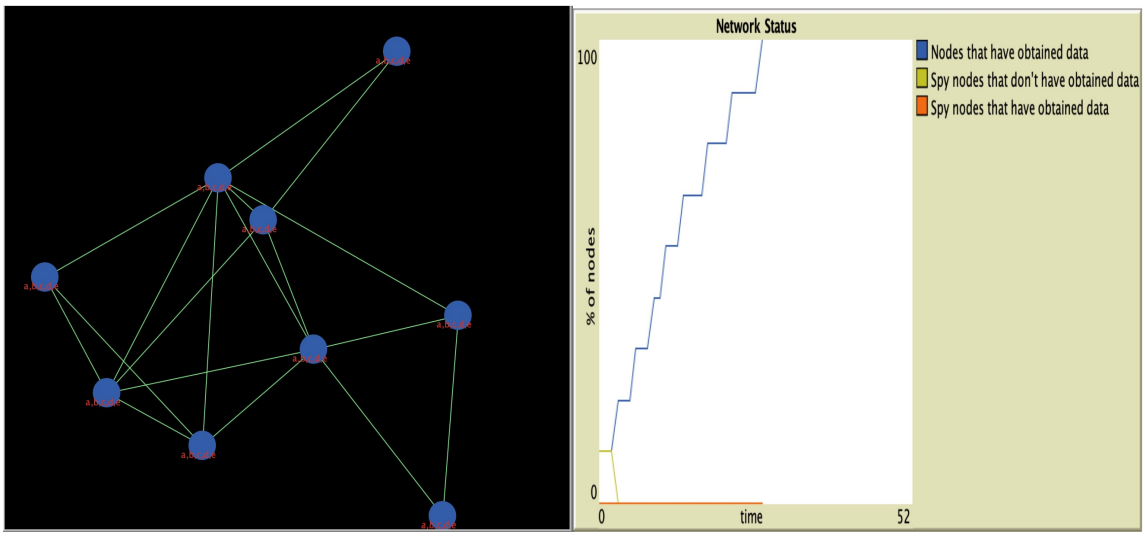

Figure 25. Overall model operation results.

As can be seen from the diagram, in the fully open source state, all the agents can get all the data as long as they request from the data source, and then marking the spy node is meaningless because everyone can get the data they want and the existence of the spy is unnecessary. So we can see that each node ends up being the same color as the data source node and has the same data as the data source. In addition to this, we can also see that the time it took for all nodes to get data was very short, taking only 26 ticks (number of model runs) to achieve full coverage, as there were no restrictions and agents did not have to request access and wait for feedback. The unrestricted version of the model also achieves the goal of data sharing and collaboration.

The unrestricted version only achieves data sharing under ideal conditions without considering potential dangers and is therefore imperfect; it is the model with restrictions that is the focus of the study. Through multiple runs of the model we can see that after the introduction of restrictions and spy nodes, the model is still able to achieve data sharing, and the results of each complete run are random, in line with the characteristics of data sharing in a network.

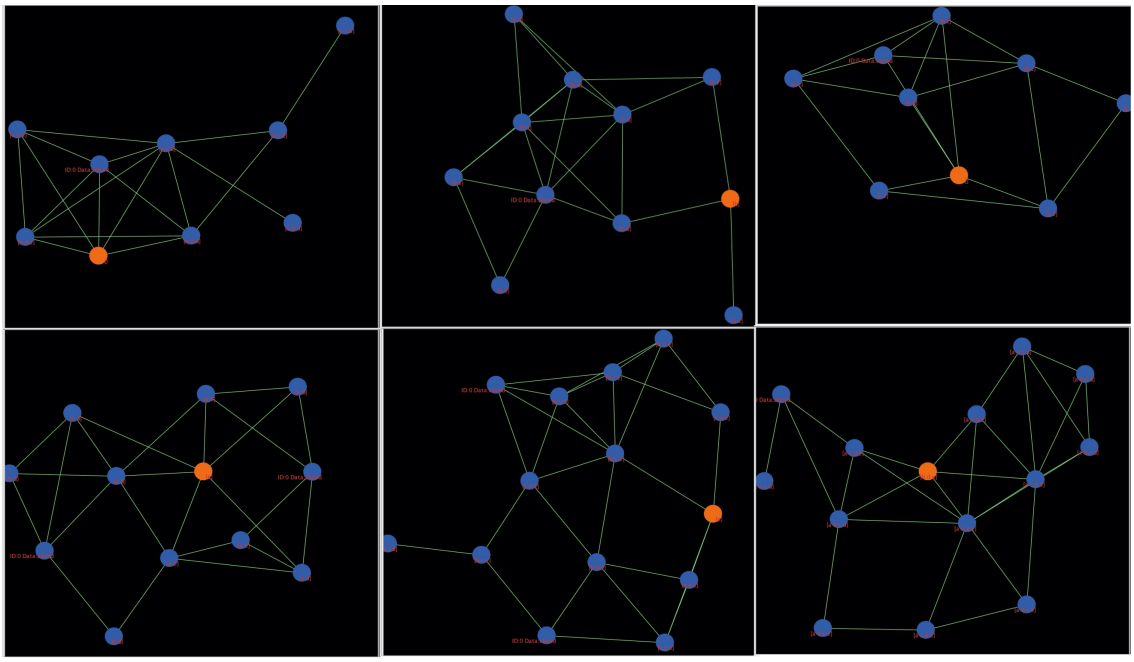

Figure 26. Results of 6 model runs.

It is worth noting that the restricted version of the model has several variables that can be adjusted, such as the total number of nodes, the amount of data each node can respond to, the level of restrictions, etc. We can adjust some or all of these variables before each run in order to increase the randomness of the model and obtain different results for each run. As shown in the figure above, each model is different, some models only have access to two data per node (randomly), some have a higher total number of nodes, and some have stricter restrictions on data sharing, the effects of these different variables can be seen in the subsequent line graphs. It is worth noting that because of the restrictions introduced, we want to protect the security and privacy of the data, which is why the spy node exists; it will try to steal the data from the data source because it is not authorized to access it. Therefore, we designed the spy node to be a different color than all other nodes, so that it is easy to distinguish (in practice, the spy will try to disguise itself to prevent detection) and so that its rules for accessing data are independent of those of other nodes (the protection-tolerance slider does the job), which makes it more representative of the uncertainty of 'spies'.

The model interface can be used to study the effect of restrictions on the extent of data sharing, while the line graph can be used to study the effect of restrictions on the speed of data sharing.

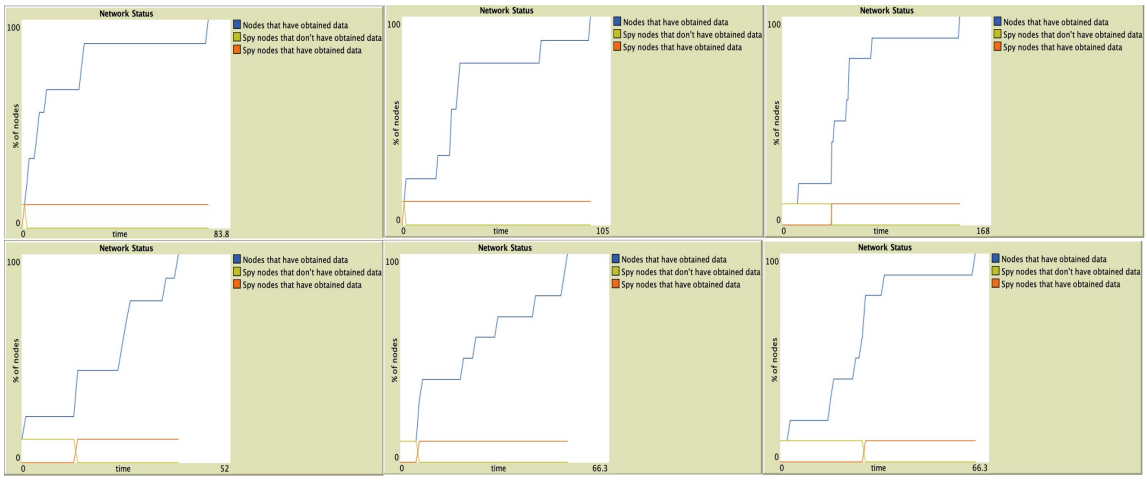

Figure 27. Results of 6 line charts.

As we can see from the above figure, the time taken by all nodes to obtain the data increases significantly after the introduction of restrictions and the impact of different levels of restrictions is different. For example, with the same 10 nodes, when the limit-tolerance is 2, the average time taken by all nodes to obtain data is 148 ticks, but when the limit-tolerance is 4, the average time taken by all nodes to obtain data is 65 ticks, the greater the probability that each node gets data on each tick run, and the easier it is to get data. The tighter the restriction means that each data point may take several or even dozens of ticks to obtain the data, even if some data nodes are common and have no malicious agents. For spy nodes, we can see that under the unrestricted version, the spy node gets the data almost as soon as it starts propagating, spending on average only 5 ticks and all. However, under the restricted version, the time spent by the spy node to steal the data is significantly higher. Also in the case of 10 nodes, the average time spent by the spy node to acquire data is 58 ticks at protection-tolerance 4 and 133 ticks at protection-tolerance 2. The protections are very effective.

From the diagram below it is easy to see that when the spy node is not directly connected to the data source node, we only need to consider the effect of limit-tolerance. the larger the limit-tolerance, the less time it takes for the data to propagate to all nodes. When the spy node is directly connected to the data source, we fix the value of limit-tolerance to 2, and find that the larger the value of protection-tolerance, the less time it takes for the data to propagate to all the nodes. All of the above conclusions are also consistent with the original hypothesis.

Figure 28. The impact of different levels of restrictions.

Also, the amount and content of data that the spy node is able to retrieve each time is completely random and is not controlled by the data-of-number-received variable. In other words, even if a spy node gets by with a high level of protection, it may return empty-handed, or with only a small amount of data, making its efforts futile or unproductive.

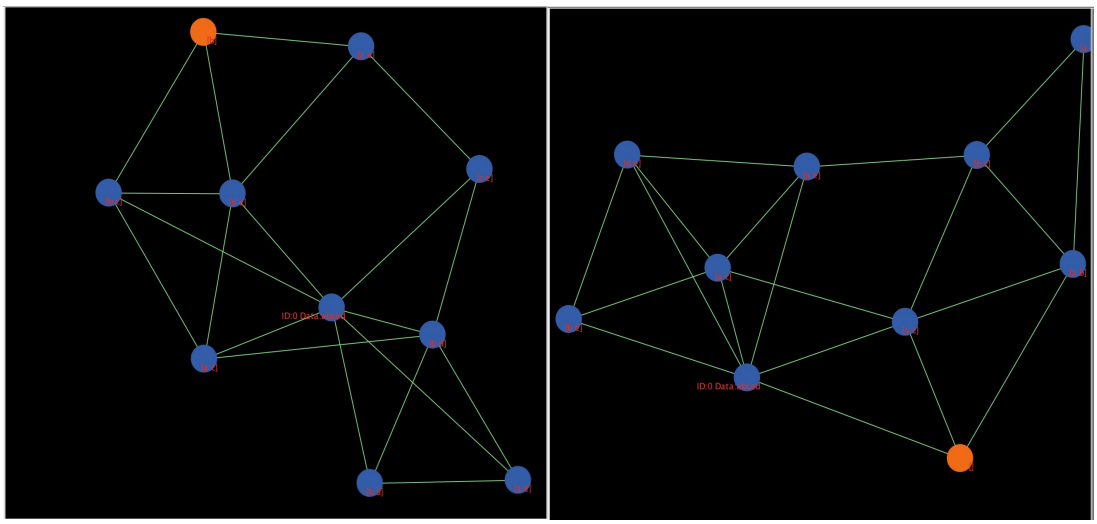

Figure 29. The amount of data a spy node may acquire.

This is also to simulate a real world situation where a data source or database may have more layers of protection, and simply breaking through one layer may still not help.

In summary, we believe that the model largely fulfills its stated function and can be used to model data sharing and collaboration, and to study the impact of data protection measures. Likewise, our hypothesis that imposing different levels of protection measures on data sources is effective in protecting data security and privacy, but also has an impact on the time spent on the process of data sharing.

CHAPTER V: Conclusion

As technology continues to advance and society evolves, there is a huge amount of data being generated every day, whether it's on paper or online. How to properly handle and use this data is an issue that cannot be ignored, and open source data sharing and collaboration between data is certainly a good way to do so. But open source means unlimited, that is, anyone can access the data at anytime, anywhere, for any purpose, then the consequence is that the data may be maliciously stolen, security is not guaranteed or the user's privacy is compromised. This is a noteworthy issue.

So the main question that this project wants to examine is how can we have a way of sharing information that can protect the privacy of individual information? It is worthwhile to examine how do constraints affect information sharing and how do constraints minimize risk of losing private information? Data and knowledge of information sharing, collaboration, and privacy theories were not enough to complete this project, so I conducted another literature review related work before beginning my research. By reading the literature on information sharing and collaboration, computer-supported cooperative work, and data protection, I learned about the origins, principles, and methods of the concepts of data sharing and collaboration, as well as the policies and methods related to data protection, which built a theoretical foundation for this research. Likewise, from reviewing the literature, we learned that modeling is the most effective method for studying social problems, and that information sharing is widely present in our society, so this project chose the research method of modeling to study these problems.

The project uses NetLogo to model the problem to be investigated, NetLogo is a programmable modeling environment used to simulate natural and social phenomena. It has a simple user interface, simple interaction, and a very simple yet powerful programming syntax, making it particularly suitable for modeling social problems and complex systems. The models created are well tuned and visualized, and variables can be added to experiment with the models. In addition, NetLogo has an official library and repository, as well as a user-created community for learning from each other and solving problems. This makes NetLogo the perfect tool for a beginner like me, who is new to modeling.

The model built for this project is a data sharing and collaboration model. It simulates the randomness and uncertainty among users of the network by having an indeterminate number of nodes and connecting them together with a random number of links, i.e. whether each node is connected to other nodes or not is random. Then by setting up data source nodes, common agent nodes to distinguish between the input and output of data, and each agent to obtain the amount of data and data content are random, so that it will complete the data sharing simulation process. In order to further study the effect of restrictions and protection measures on data sharing, spy nodes and some variables, i.e., protection measures, are added to the model, and their sizes are continuously adjusted to simulate different levels of protection measures.

Finally, by analyzing the results of multiple runs of the model and the resulting data, the hypothesis was verified and the previously raised sub-questions is solved. That is, the constraint will have an impact on the process of information sharing, which will result in slowing down the information sharing. However, as long as the model is run long enough, then all agents still have access to the data and therefore have little impact on the extent of data sharing. In addition to this, we can minimize the risk of privacy data theft by imposing different constraints on different visitors based on their own characteristics. The stricter the protection measures, the greater the impact on the spy node, the less data the spy node gets and the longer it takes to get the data, which means that imposing certain restrictions on the data from the data source is effective in protecting the data. It also proves the importance of data protection.

Limitations and Future Work

Although this research was carefully implemented, there were some identified limitations. The line graph part of the model currently shows the number of all nodes acquiring data and the number of spy nodes over time, which is clear for us to observe the process of data sharing. The problem, however, is that in real life, we usually do not request access to all the data of a data source or server in general, but rather to some of the data in a database in a targeted and purposeful way. Therefore, if a line graph can show the changes in the amount of data shared for one or some particular data in real time, it will be more relevant to the real situation and the results obtained from the inferential analysis will be more convincing.

In addition to this, the representation of the relationships between the model nodes has shortcomings. The connections between the nodes in the current model are directionless, but in fact, all data transfers in the network are directional, i.e., from the data source to the visitor or from the server to the client. The simulation of this process should therefore also have a directional relationship between the nodes, i.e., using lines connected with arrows, which would give a clearer representation of the direction of data transmission, although the model already makes use of different colors to distinguish between data sources and ordinary nodes.

As for future work, more variables can be added to the model to simulate more scenarios based on the optimization of the above problem. For example, several other data protection methods mentioned in the literature, some other forms in which data could be maliciously stolen, and by combining different methods to simulate and study the model where data protection works best. Finally, read more of the literature and try other modeling tools or approaches.

Reference

[1] Sukumar, Sreenivas R. and Ferrell, Regina K. (2013). "'Big Data' collaboration: Exploring, recording and sharing enterprise knowledge". *Information Services & Use,* 33(3-4), 257–270.

[2] Meneely, A., & Williams, L. "Secure open source collaboration". *Proceedings of the 16th ACM Conference on Computer and Communications Security* - CCS '09. 2009.

[3] Edwards, Paul N., Matthew S. Mayernik, Archer L. Batcheller, Geoffrey C. Bowker, and Christine L. Borgman. "Science Friction: Data, Metadata, and Collaboration." *Social Studies of Science* 41, no. 5 October 2011, 667–90.

[4] Voigt, Paul, and Axel von dem Bussche. "The EU General Data Protection Regulation (GDPR)" 2017

[5] Jona, Grudian. "Computer-supported cooperative work(CSCW): history and focus. " *Computer*, 27(5), 19–26, May 1994.

[6] Neale, D. C., Carroll, J. M., & Rosson, Mary. Beth "Evaluating computer-supported cooperative work." on *ACM Conference on Computer Supported Cooperative Work - CSCW '04*. 2004.

[7] Ronald M. B. "Readings in Groupware and Computer-Supported Cooperative Work: Assisting Human-Human Collaboration." *Elsevier,* 1993.

[8] T. D. Palmer and N. A. Fields, "Computer supported cooperative work," in Computer, vol. 27, no. 5, pp. 15-17, May 1994.

[9] Judith S. Olson, Stuart K. Card, Thomas K. Landauer, Gary M. Olson, Thomas Malone and John Leggett. "Computer-supported co-operative work: research issues for the 90s." *Behaviour & Information Technology*, 12:2, 115-129, 1993.

[10] Acquisti, A. and Gross, R. "Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook." *Lecture Notes in Computer Science*, 36–58. 2006.

[11] Mesmer-Magnus, J. R. and DeChurch, L. A. "Information sharing and team performance: A meta-analysis." *Journal of Applied Psychology*, 94(2), 535–546. 2009.

[12] Smaldino, Paul. "How to Translate a Verbal Theory into a Formal Model." *MetaArXiv*, 26 May 2020.

[13] Ross, T. R., Ng, D., Brown, J. S., Pardee, R., Hornbrook, M. C., Hart, G., & Steiner, J. F. "The HMO Research Network Virtual Data Warehouse: A Public Data Model to Support Collaboration". *EGEMS (Washington, DC)*, *2*(1), 1049. 2014

[14] George Danezis, Josep Domingo-Ferrer, Marit Hansen, Jaap-Henk Hoepman, Daniel Le Métayer, Rodica Tirtea, Stefan Schiffner. "Privacy and Data Protection by Design - from policy to engineering". *European Union Agency for Network and Information Security (ENISA)*. 2014

[15] Sandra Wachter; Brent Mittelstadt, "A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI," *Columbia Business Law Review* 2019, no. 2 (2019): 494-620

[16] Gaudou, B., Lang, C., Marilleau, N., Savin, G., Rey Coyrehourcq, S., & Nicod, J.-M. (2017). "NetLogo, an Open Simulation Environment". *Agent-Based Spatial Simulation with NetLogo,* Volume 2, 1–36.

[17] Musaeus, L. H., & Musaeus, P. (2019). "Computational Thinking in the Danish High School". *Proceedings of the 50th ACM Technical Symposium on Computer Science*

Education - SIGCSE '19.
[18] Uri Wilensky, William Rand. "An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo". *MIT Press*, 2015.
[19] Sklar, E. "NetLogo, a Multi-agent Simulation Environment". *Artificial Life,* 13(3), 303–311. 2007.
[20] Chiacchio, F., Pennisi, M., Russo, G., Motta, S., & Pappalardo, F. "Agent-Based Modeling of the Immune System: NetLogo, a Promising Framework". *BioMed Research International*, 1-6, 2014.
[21] Albiero F., Fitzek F.H.P., Katz M.D. "Introduction to NetLogo". *Cognitive Wireless Networks. Springer*, Dordrecht. 2007
[22] Wilensky, U. (1999). NetLogo. http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling, Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL.
[23] Stonedahl, F. and Wilensky, U. "NetLogo Virus on a Network model". *http://ccl.northwestern.edu/netlogo/models/VirusonaNetwork*. Center for Connected Learning and Computer-Based Modeling, Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL. 2008.
[24] Kashif, Z. "An Agent-Based Model of Crowd Evacuation: Combining Individual, Social and Technological Aspects" *ACM SIGSIM PADS 2020.* 2020
[25] Perioellis P, Cook N, Hiden H, Conlin A, Hamilton MD, Wu J, Bryans J, Gong X, Zhu F, Wright A. "The GOLD Project: Architecture, Development and Deployment". *National e-Science Centre.* 8, vol1, 2006.

Appendix A: GitLab Repository

To run the experiments which were developed in this research, access the GitLab repository using the below link:

https://git-teaching.cs.bham.ac.uk/mod-msc-proj-2019/zxm947.git

The repository includes 1 NetLogo file.

Please note: the code had better run on NetLogo 6.1.1 or latest vision, it will not run on older versions.

Further details can be found in Read_Me.pdf .