**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.

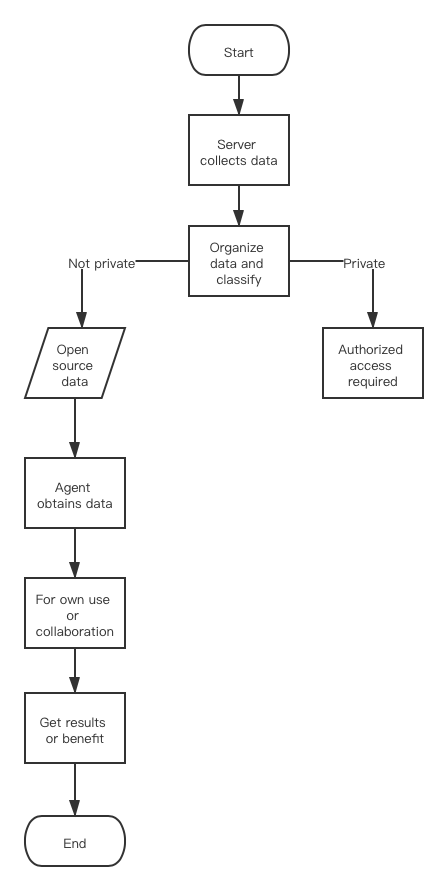
**1.1.1 Collaborate and protect open source data**

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 [1]. 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 [2]. 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 [3].

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.

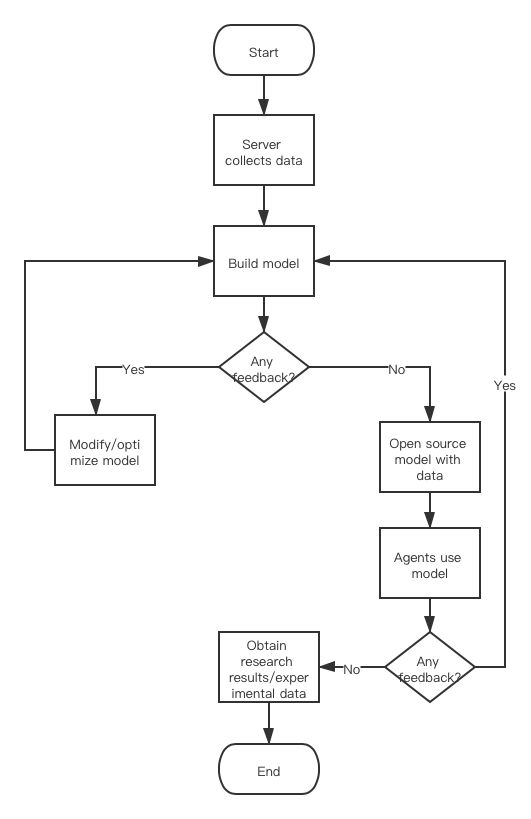
**1.1.2 Data collaboration methods**

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 [4]. 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 [5]. 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



**Fig. 1.** Flowchart of the first method

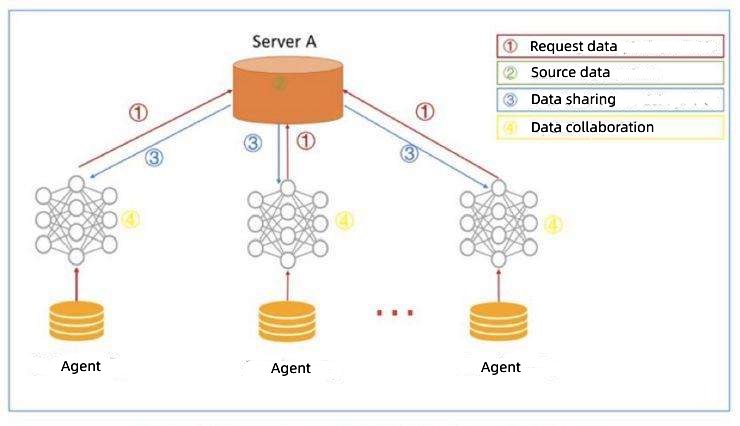
and data-based models for collaboration.



**Fig. 2.** Flowchart of the second method

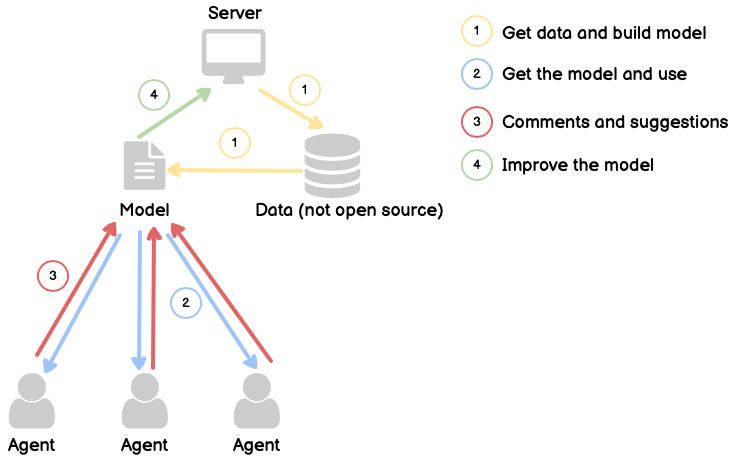
**1.1.3 Models of collaboration methods**

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 [6]. 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.



**Fig. 3.** 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.

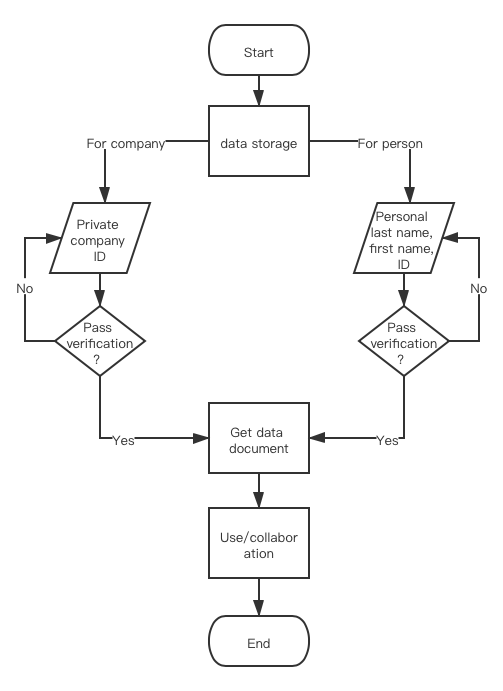


**Fig. 4.** Prototype of the second method

This is another form of data collaboration.

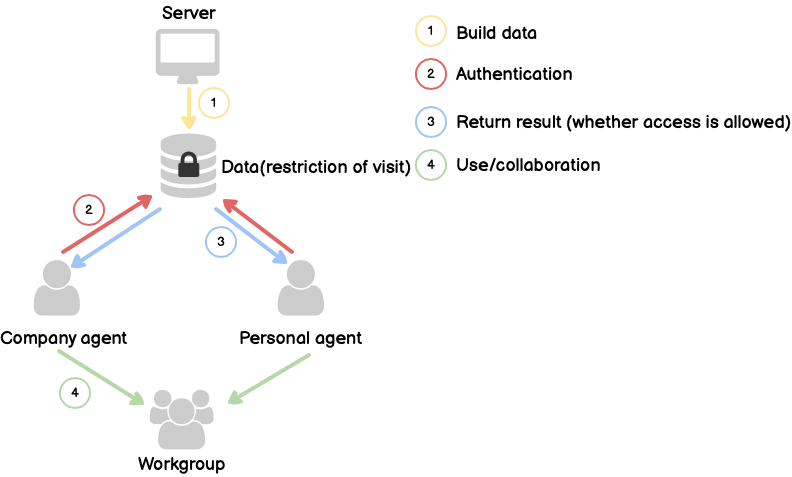
**1.1.4 Data protection methods**

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 [7]. 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 [8]. 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.



**Fig. 5.** 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 [18].

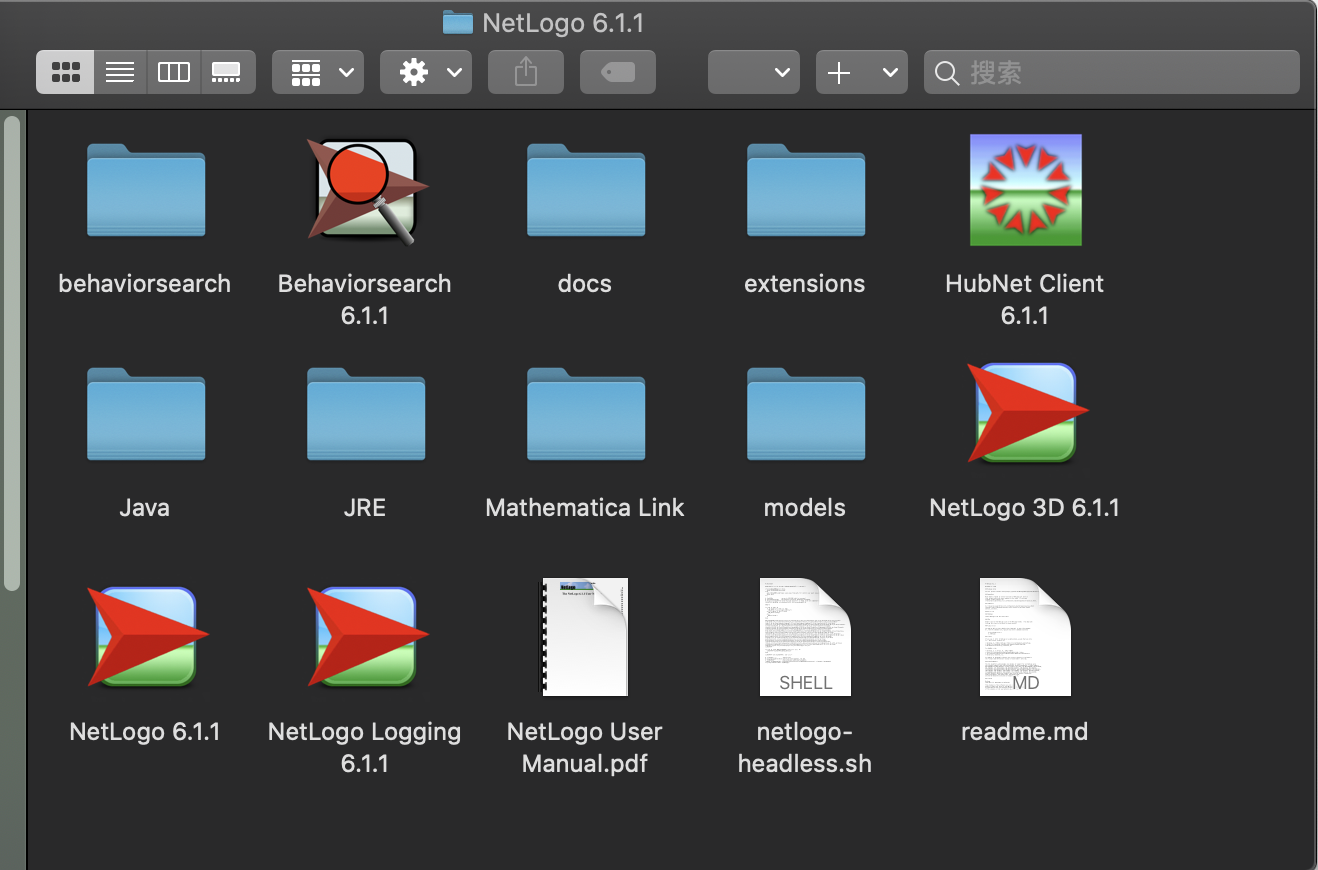


**Fig. 6.** Prototype of protection method

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. The next part will introduce the modeling tools and learning models used in this project.

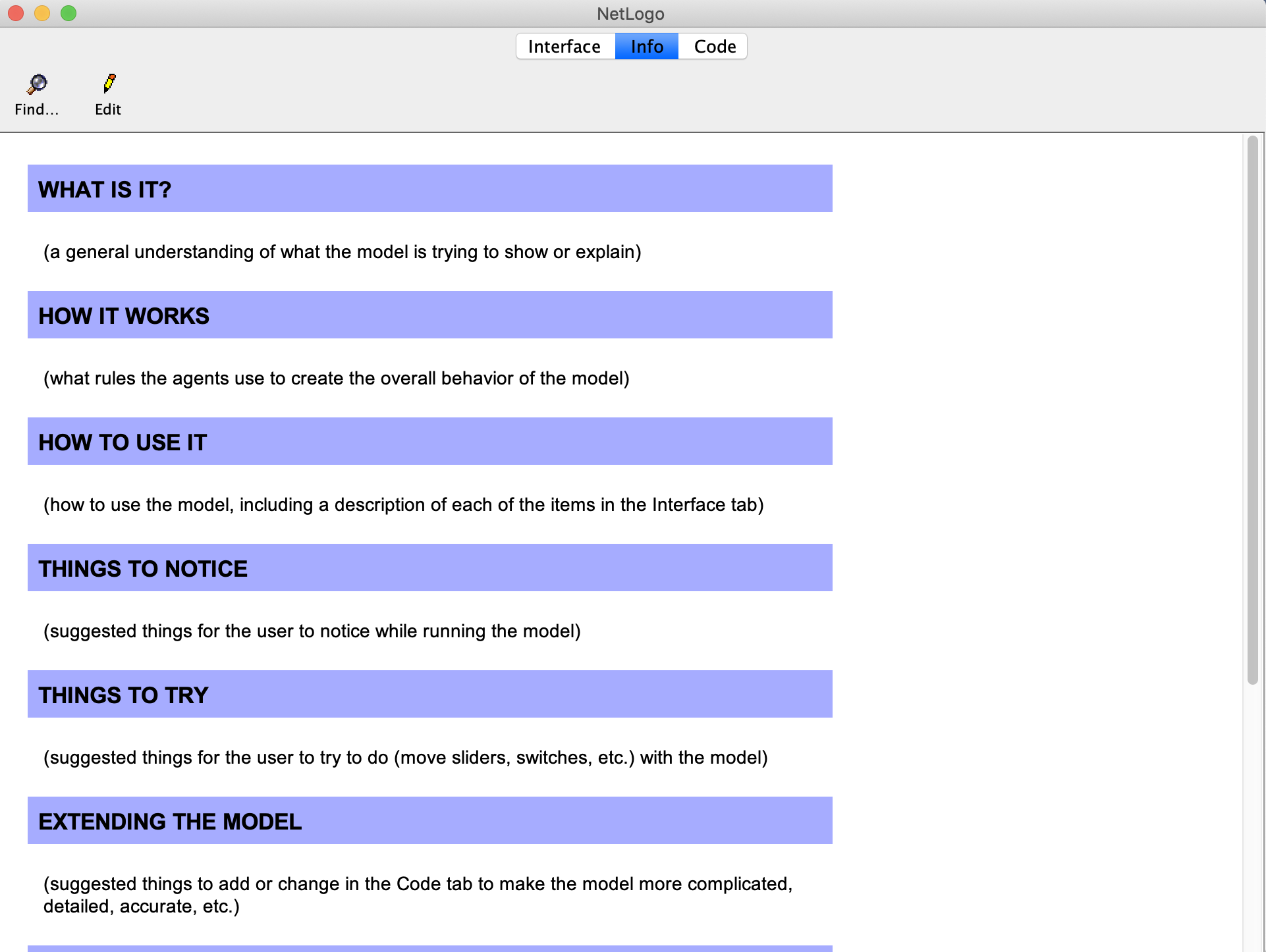
**1.2.1 Netlogo modeling environment**

NetLogo is a programmable modeling environment for simulating natural and social phenomena. It is also a free downloadable agent-based software package [9].



**Fig. 7.** NetLogo download package

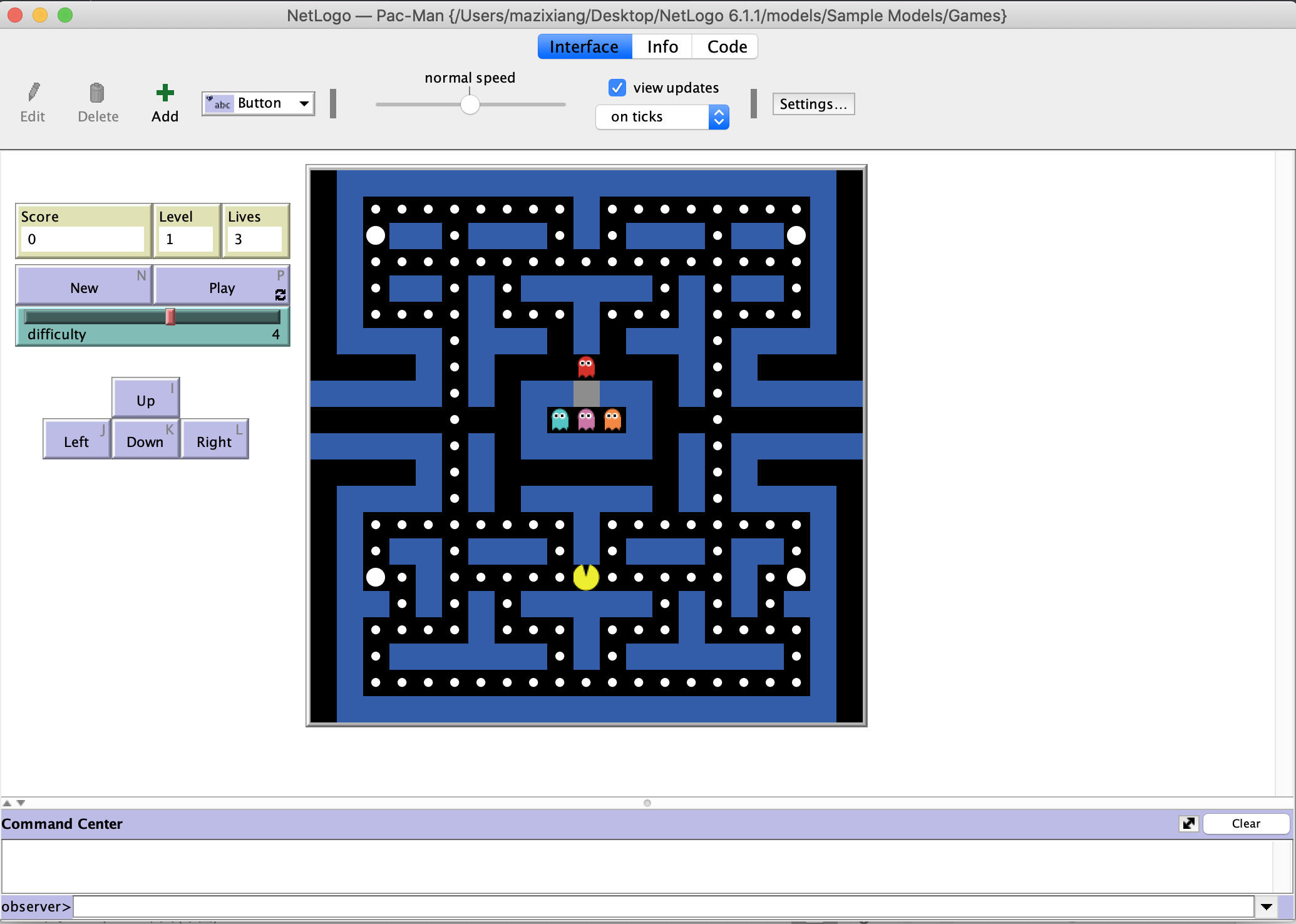
The software package is the Center for Connected Learning and Computer Modeling (CCL) of Northwestern University under the guidance of Uri Wilensky. Created [10]. NetLogo is a functional programming language, which also means that many language sentences are almost read as sentences, which enables even unskilled and untrained users to understand and learn it through examples [11].



**Fig. 8.** NetLogo information page

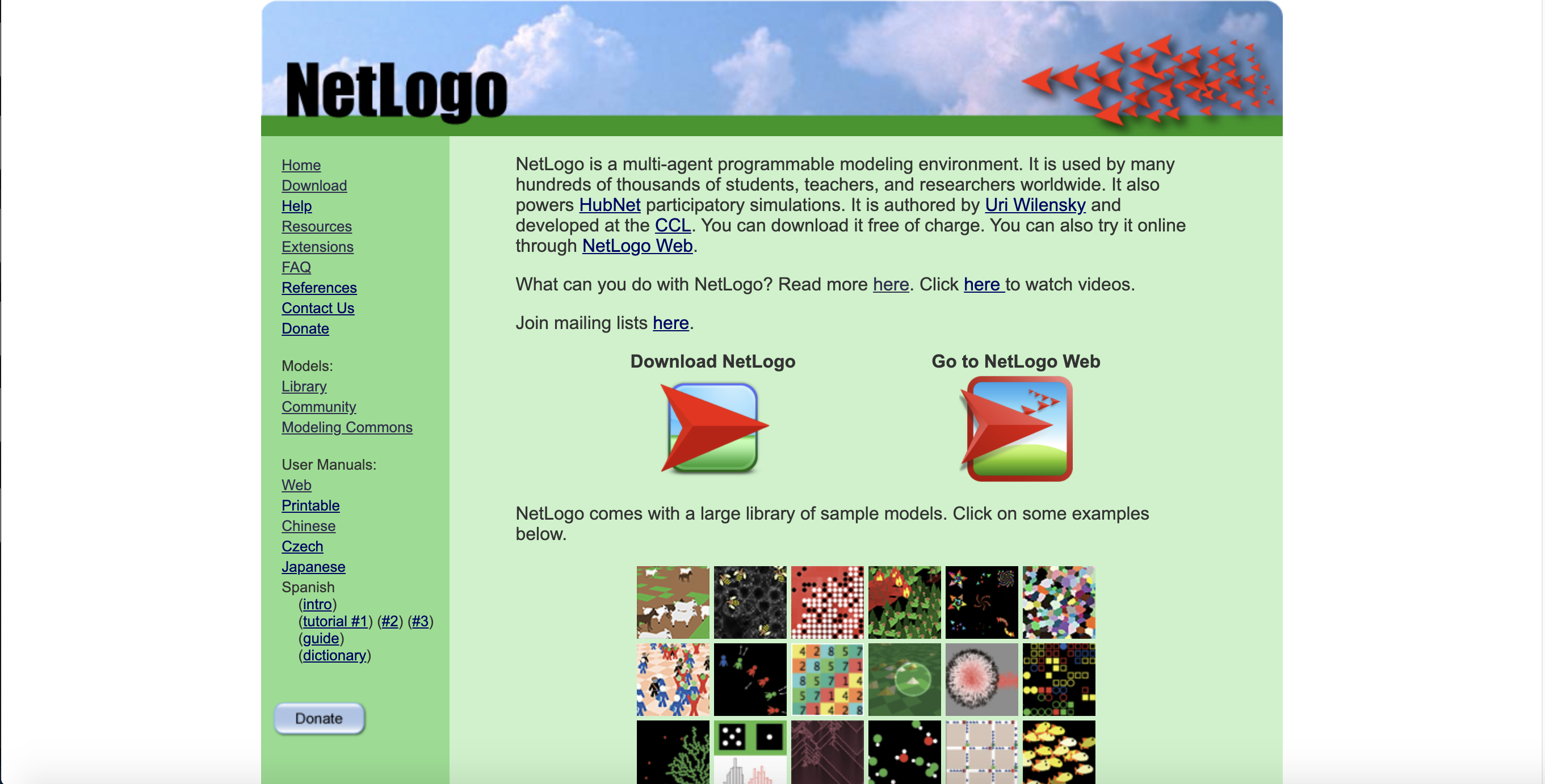
The "turtle" represents the agent, and the "patch" represents a given point in the simulation space. Both of these attributes can have multiple attributes that the user can define, such as age, color, and location [12].

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 [13].



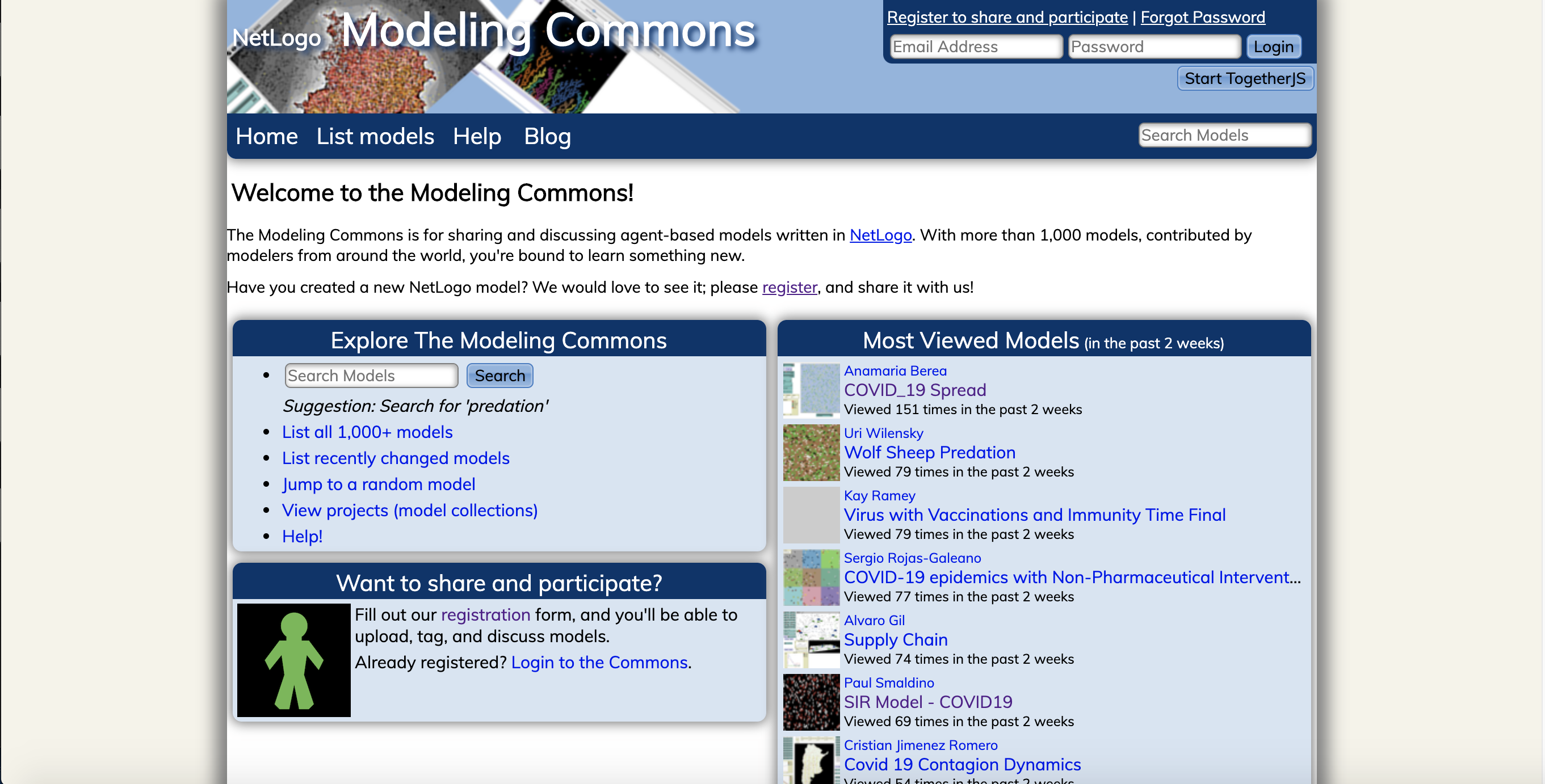
**Fig. 9.** NetLogo running page

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 [14]. 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.



**Fig. 10.** 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 [15].



**Fig. 11.** NetLogo community website

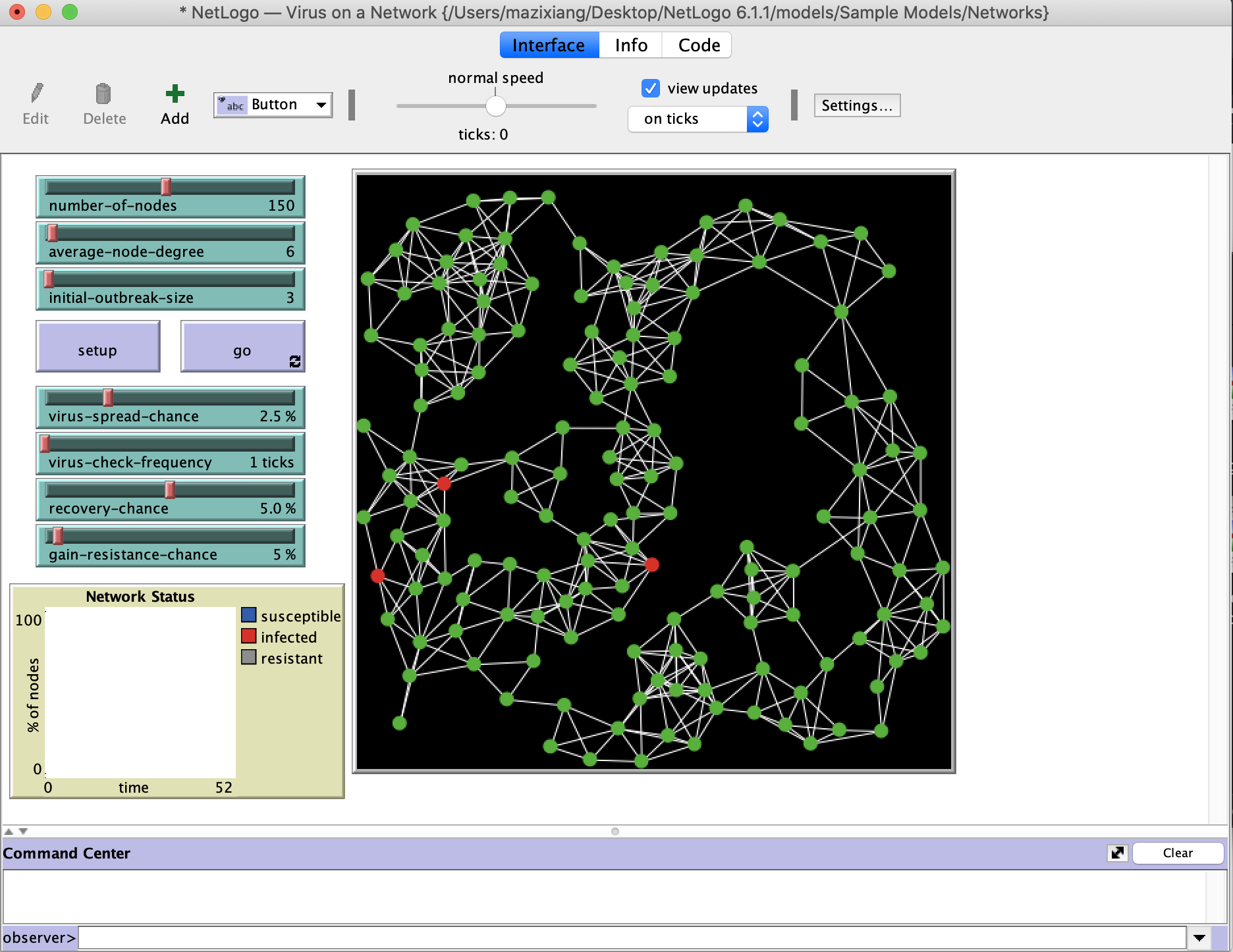
The online community page contains models built by a wide range of representatives from all levels of these people.

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.

**1.2.2 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.

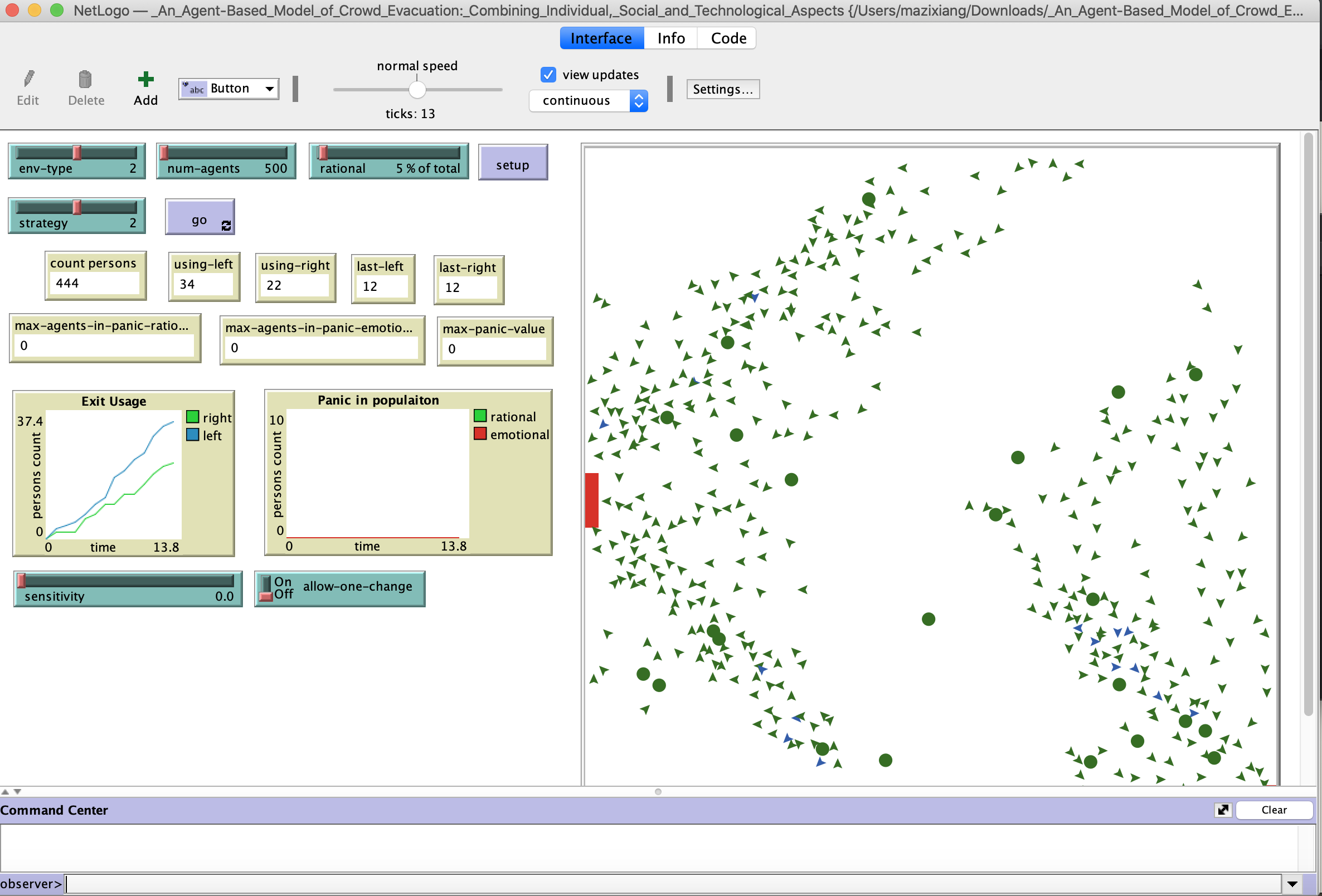


**Fig. 12.** Virus on a Network model

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,



**Fig. 13.** 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.

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.

**Reference**

[1] Sukumar, S. R., & Ferrell, R. 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] Smaldino, Paul. “How to Translate a Verbal Theory into a Formal Model.” *MetaArXiv*, 26 May 2020.

[4] 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

[5] 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.

[6] 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

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

[8] 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

[9] 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.

[10] 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.*

[11] Uri Wilensky, William Rand. “An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo”. *MIT Press*, 2015.

[12] Sklar, E. “NetLogo, a Multi-agent Simulation Environment”. *Artificial Life,* 13(3), 303–311. 2007.

[13] 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.

[14] Albiero F., Fitzek F.H.P., Katz M.D. “Introduction to NetLogo”. *Cognitive Wireless Networks. Springer*, Dordrecht. 2007

[15] 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.

[16] 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.

[17] Kashif, Z. "An Agent-Based Model of Crowd Evacuation: Combining Individual, Social and Technological Aspects" *ACM SIGSIM PADS 2020.* 2020

[18] 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.