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 any time, 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. Therefore, the main objective of this research is to study how data is shared and collaborated across networks and to model this problem in order to get clearer and more complete results. In addition to this, it is also investigated whether some restrictions and protections on open source data are effective in protecting the data and how these affect the data sharing process in terms of speed, scope, etc.

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, conclusions were drawn and the hypothesis was verified. That is, the addition of protection measures to a data source affects the speed of data sharing, and the more stringent the protection measures are, the slower the data sharing will be. 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. But 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.

Translated with www.DeepL.com/Translator (free version)