

Data Management Plan

This data management plan covers the data to be generated by the proposed project in the time period of three years. The data of this research will be acquired, preserved, and shared, where appropriate, in accordance with the NSF policy and guidance regarding data sharing and stewardship. Our data management and sharing practice will also follow the policies of the PIs' institutions on intellectual property, privacy, security, and record retention and management, and will respect publication copyright and confidentiality.

Standards of data and metadata. The proposed research will mainly include observational data, trained graph neural network models, results for evaluation, as well as reports and documents pertaining to outreach and educational activities. Research data will be saved in text files, Word, Portable Document Format (PDF), Rich Text Format (RTF), Python Format (PY), Checkpoint File Format (CKPT), and Excel (XLS) format. Figures will be produced in JPEG and TIFF formats. We will organize each type of data into a folder and offer a README file for every folder that describes how the data are formatted and the programs are executed. We will also provide information of the code's contributors and license for sharing.

Polices for data access and sharing. Data will be made accessible to the research community immediately upon publication. Issues related to intellectual property, privacy, and confidentiality are not expected with the data generated by this project. Both PIs will maintain control of all project data and in control of access to the data stored at IIT and UNT. When publishing our source code and data, we will use one of the following Free Software Foundation licenses:

- GNU General Public License: <http://www.gnu.org/copyleft/gpl.html>
- GNU Lesser General Public License: <http://www.gnu.org/licenses/lgpl.html>

Policies and provisions for data storage, re-use, and re-distribution All project data (including any relevant data not published) will be maintained electronically at each PI's institute. Both PI Ji, PI Wang, and Co-PI Bhowmick have already bought their own servers. All the datasets will be stored on the servers. Moreover, we will use RAID for robust data storage on hard disk drivers. PI Ji will set up a git repository to share and synchronize datasets with PI Wang and Co-PI Bhowmick. Any data obtained will be retained for at least 5 years after the end of this project. Any outreach and programs that make use of these data will be documented and any presentations, flyers, and supplemental information will be archived as part of the data storage procedures for this project. These materials will be retained for at least 5 years after the event has taken place.