# DATA MANAGEMENT PLAN

This data management plan is part of the proposal “Collaborative Research: Concurrent Windows into Time-Series of Graphs for Network Change Analysis and Cyber-Threat Detection”. Data management for this project will be conducted at the Institute for Data-Intensive Engineering and Science. This is also where all shared data, computational, and analysis resources will be hosted. IDIES has 15 years of experience hosting publicly accessible scientific databases and data-intensive Web-services. The best practices of the Institute will be adopted and customized for the needs of graph as described below.

## Expected data:

*Time-Series of Graphs*: The project will build a repository of time-series graph datasets as described in the proposal. These will represent Internet usage, social networks, recommendation engines, physical systems, and human interaction data. These will be stored and served in multiple formats. We will publish a series of snapshots in standard formats, such as Problem Based Benchmark Suite (PBBS) and MatrixMarket. No standard formats support continuous time. We will make the full data available in time-optimized edge-list CSV files and provide libraries to interpret read and interpret data in Python and C++.
*Publications:* Pre-print versions of the papers and the ASCII data files required to generate the publication figures will be preserved.
*Documentation and Metadata:* Source code documentation will be collocated and released with software, governed by the same open source license, and hosted on public repositories. This is the current practice of the collaborators. Documentation for datasets are treated similarly. They will be managed in version control public repositories. They will be available as links that are collocated with the Web-services that provide access to data. This is already standard practice of the collaborators.

## Data Integrity:

Our systems make several efforts to ensure the integrity of the data and to prevent malicious or unintended modification. We regularly crawl our data checking content against pre-computed checksums to protect against corruption from hardware and software errors.

## Security, Privacy, and Embargo:

The privacy of datasets is implemented with enterprise user authentication and secure Web-services (*https://*). The combination allows us to tightly manage access control to the data. We also determine when to make data public, and when to release data from embargo.

## Roles and Responsibilities:

The PI and CoPI’s will have decision-making authority over all data management. They will devote a meeting to draft an overall data management policy during the first three months, following the broad principles listed above. PI Randal Burns will serve as the Data Management officer with oversight of backup, archival, capacity, and data integrity. This plan will be reviewed yearly and policy revisions communicated to system administrators.

## Storage and Backup During the Project:

All datasets will be backed up on site and archived off-site. IDIES provides this capability to data tenants of their systems. IDIES also participates in a mirroring relationships with the University of Illinois at Chicago through the Open Cloud Consortium, using a dedicated 100 Gbps link to Internet2.

## Long-Term Archival and Preservation:

IDIES has preserved all scientific data that it has ingested. The first resources developed were the Sloan Digital Sky Survey (sdss.org). Multiple strategies have been used to maintain these data. We have partnered with Google to store all image data on an ongoing basis. The development of new data-intensive clusters (GrayWulf 2006, Data-Scope 2012, and SciServer 2018) has provisioned a small fraction of resources to maintain the entirety of our previously collected data. Our preservation ethic includes preserving the function and semantics of the data long after project operation completes. We will do so by defining archival packages that include algorithms and methods as well as data. We recognize sustainable preservation as one of the most challenging aspects of developing data resources and note our commitment and experience in defining strategies to fund and maintain resource-sharing beyond project lifetimes.

## Data Sharing and Dissemination:

This project focuses on making public datasets broadly available, leveraging the team's expertise in large-scale storage architectures and data processing to host scientific datasets that are unmanageable by prevailing approaches. This process makes data a community resource, greatly enhancing their utility. All public datasets referenced in this project will be available through Web-services. We will publicize our datasets through outreach activities conducted by IDIES, which include museum exhibitions, the National Science Fair, and community programs.

## Ownership, Copyright, Intellectual Property:

Our project will open-source all products under permissive licenses. Software will be released under the Apache 2.0 license and data under the Creative Commons CC-By license. Software and data will be a community resource, unrestricted for non-commercial use. We retain copyright privileges and reserve licensing and approval rights for commercial uses of data. Users of the data reserve rights to their algorithms and analysis techniques.