Monthly Project Report for SIMPLEX DARPA Grant: From RAGs to Riches: Utilizing Richly Attributed Graphs to Reason from Heterogeneous Data

PI Joshua T. Vogelstein[1], Co-I Randal Burn[2], Co-I Carey E. Priebe[3]
[1] Dept Biomedical Engineering, [2] Dept Computer Science, [3] Dept Applied Mathematics and Statistics
Johns Hopkins University

May 31, 2015

Overview

This monthly report lists cumulative progress on our SIMPLEX statement of work. It is organized into Tasks (captial roman numerals), their respective subtasks (captial alphabet), and subsubtasks (black bold headings). The most recent month's progress is highlighted in red. The appendix lists all cumulative deliverables, including manuscripts, code, data, and resulting derivatives.

I Mathematical Framework

I(A) RAG Embedding

Tensor Factorization:

- **May:** While many tensor factorization algorithms exist, all of them must solve the question of: how many factors to keep. We formulate this question as a model selection question, and are developing model selection for tensor factorization. In preliminary work, we have written four manuscripts detailing these methods over the last couple years: http://arxiv.org/abs/1312.7559, http://arxiv.org/abs/1406.6315, and http://arxiv.org/pdf/1406.6319v3.pdf). In May, we have begun to further develop these methods, make the code open source, and transition to more scalable implementations.

II Computational Infrastructure

II(A) Data Management

Dense Spatial Multi-way Arrays:

- **May:** In previous work, we buit a n-way dense spatial database for petascale data (http://arxiv.org/abs/1306.3543). However, for the data types we used previously (serial electron microscopy and array tomography), the data were anisotropic. To visualize that data, our collaborators want to downsample only along the xy dimensions, keeping the z dimension fixed. However, for the new datasets that we will work with for this grant, CLARITY and M^{3}RI, our collaborators desire isotropic downsampling. Thus, we have extended our infrastructure to support multiple types of downsampling, as appropriate for different datasets, including a uniform downsampling. We have already begun using this Web-service to support alignment of CLARITY brains to the Allen Institute for Brain Science's

mouse atlas, which is 25 micron isotropic.

III Datafication

III(A) Data Ingest

Diffusion MRI:

- **May:** To ingest diffusion MRI data into our spatial database and corresponding annotation database, we require to additional object types in our data model. First, a *skeleton* object type, to store tracts. Second, a *region of interest* (ROI) object type, to store anatomical regions. In May, we have implemented the skeleton object type into our RAMON framework. We have also been working with the designers of COINS and the Human Brain Project, to register our data with them, to enable search across datasets.

IV Discovery

IV(A) RAG Construction

RAG Random Walks:

- **May:** We have empirical evidence as well as theoretical results adapted from Rohe, Chatterjee and Yu (2011) showing that the spectrum of the graph Laplacian is robust to noise. In practice, for dimensionality reduction and embedding purposes, we build the graph Laplacian on the near neighbor matrix of a data set rather than on a dense graph, so one would like to know if similar properties hold when instead of simply adding independent entry-wise noise to a kernel matrix, we have a noisy version of the near neighbor matrix. The core idea is to use the fact that the graph Laplacian is related to the commute times of the graph: we want to show that certain kinds of noise do not change the random walks on that graph, and hence do not greatly change the Laplacian eigenmap embedding.

A Summary of Deliverables

I(A) Pre-prints and Publications

N/A

I(B) Codes

N/A

I(C) Data and Data Derivatives

N/A

B Other Personnel

- Eric Bridgeford, BS student in Biomedical Engineering
- Greg Kiar, MS student in Biomedical Engineering
- Kunal Lillaney, PhD student in Computer Science
- Keith Levin, PhD student in Applied Mathematics and Statistics
- Disa Mhembere, PhD Student in Computer Science
- Youngser Park, research staff in Center for Imaging Science
- Da Zheng, PhD student in Computer Science