# 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

June 29, 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.
- **June:** We are using tensor and matrix factorization techniques for two subequent inference tasks. First, we are computing the population average graph. Theory, simulations, and real data experiments demonstrate that we can significantly improve the current state of the art using these methods. Second, we have developed a community detection technique using such methods. Here, numerical experiments demonstrate a significant inference advantage, with a minimal increase in computational cost.

## 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 $M^3RI$, 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.

- **June:** We have refactored our code such that all projects now have multiple tokens. This is important for both MRI and CLARITY datasets, as both of them might have many tokens per dataset, corresponding to different color channels (in CLARITY), or different modalities (in MRI). We are testing and debugging this code, which will go live shortly.

## 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.
- **June:** We have implemented 3 new RAMON object types: *skeleton*, *point*, and *ROI*, to subserve three different functions. ROI can be used for MRI and CLARITY brains, for storing ROIs, and then eventually build new atlases. Skeletons can be used to store fiber tracks from diffusion MRI, or in electron microscopy and array tomography and CLARITY, to trace microscale processes. Points can be used to store any kind of point, be it a synapse, a cell body, or some other point.

## 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.
- **June:** We are investigating the computational trade-offs associated with computing normalized versus unnormalized adjacency and Laplacian matrices.

## 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

## A Milestones and Deliverables

| Phase | Task | # | Sub-Task | Phase I | | | | | Phase II | | | | Phase III | | | | Responsible |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | GFY | | Milestone | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Q11 | Q12 | Q13 | Agent |
| I | RAG Embedding | I.1.A | Tensor Factorization | ■ | ■ | | | | | | | | | | | | Priebe |
| | | I.1.B | JOFC | | | ■ | ■ | | | | | | | | | | Priebe |
| | | I.1.C | Benchmarking | | | | | ■ | | | | | | | | | Priebe |
| II | FlashRAG | II.1.A | FlashMatrix | | | | | | ■ | ■ | | | | | | | Priebe |
| | | II.1.B | FlashAttributes | | | | | | | ■ | ■ | | | | | | Priebe |
| | | II.1.C | FlashR | | | | | | | | | ■ | | | | | Priebe |
| III | Rag Testing | III.1.A | 1-sample | | | | | | | | | | ■ | ■ | | | Priebe |
| | | III.1.B | 2-sample | | | | | | | | | | | ■ | ■ | | Priebe |
| | | III.1.C | independence | | | | | | | | | | | | | ■ | Priebe |
| I | Data Management | I.2.A | Dense Arrays | ■ | ■ | | | | | | | | | | | | Burns |
| | | I.2.B | Sparse Arrays | | ■ | ■ | | | | | | | | | | | Burns |
| | | I.2.C | Sparse Cutouts | | | | ■ | ■ | | | | | | | | | Burns |
| II | Remote Access | II.1.A | 2D Web Viz | | | | | | ■ | ■ | | | | | | | Burns |
| | | II.1.B | Surface Extraction | | | | | | | ■ | ■ | | | | | | Burns |
| | | II.1.C | 3D Web Viz | | | | | | | ■ | ■ | | | | | | Burns |
| | | II.1.D | Graph Viz | | | | | | | | ■ | ■ | | | | | Burns |
| III | Local Analysis | III.1.A | API | | | | | | | | | | ■ | ■ | | | Burns |
| | | III.1.B | Downloads | | | | | | | | | | | ■ | ■ | | Burns |
| | | III.1C | GPU 3D Viz | | | | | | | | | | | | ■ | ■ | Burns |
| | | III.1.D | Remote Annotation | | | | | | | | | | | | ■ | ■ | Burns |
| I | Data Ingest | I.3.A | diffusion MRI | ■ | ■ | | | | | | | | | | | | Vogelstein |
| | | I.3.B | functional MRI | | ■ | ■ | | | | | | | | | | | Vogelstein |
| | | I.3.C | CLARITY | | | ■ | ■ | | | | | | | | | | Vogelstein |
| | | I.3.D | LFM | | | | ■ | ■ | | | | | | | | | Vogelstein |
| II | Data Register | II.3.A | Human Align | | | | | | ■ | ■ | | | | | | | Vogelstein |
| | | II.3.B | Mouse Align | | | | | | | ■ | ■ | | | | | | Vogelstein |
| | | II.3.C | Multi-Graph-Match | | | | | | | | ■ | ■ | | | | | Vogelstein |
| III | Quality Control | III.3.A | diffusion MRI | | | | | | | | | | ■ | ■ | | | Vogelstein |
| | | III.3.B | functional MRI | | | | | | | | | | | ■ | ■ | | Vogelstein |
| | | III.3.C | CLARITY | | | | | | | | | | | | ■ | ■ | Vogelstein |
| | | III.3.D | LFM | | | | | | | | | | | | ■ | ■ | Vogelstein |
| I | RAG Construct | I.4.A | Random Walks | ■ | ■ | | | | | | | | | | | | Lee |
| | | I.4.B | Graph Sparsitifcation | | ■ | ■ | | | | | | | | | | | Lee |
| | | I.4.C | Benchmarking | | | | ■ | ■ | | | | | | | | | Lee |
| II | RAG Summary Statistics | II.4.A | Moments | | | | | | ■ | ■ | | | | | | | Lee |
| | | II.4.B | Motifs | | | | | | | ■ | ■ | | | | | | Lee |
| | | II.4.C | Modes | | | | | | | | ■ | ■ | | | | | Lee |
| III | RAG Predict | III.4.A | Nearest Neighbor | | | | | | | | | | ■ | ■ | | | Lee |
| | | III.4.B | Network of Networks | | | | | | | | | | | ■ | ■ | | Lee |
| | | III.4.C | Tensor Factorization | | | | | | | | | | | ■ | ■ | | Lee |
| | | III.4.D | Benchmarking | | | | | | | | | | | | ■ | ■ | Lee |
| | Monthly Tech Reports | 5 | | X | XXX | XXX | XXX | XXX | XXX | XXX | XXX | XXX | XXX | XXX | XXX | XXX | Vogelstein |
| | Monthly Financial Reports | 5 | | X | XXX | XXX | XXX | XXX | XXX | XXX | XXX | XXX | XXX | XXX | XXX | XXX | Vogelstein |
| | Quarterly Reports | 5 | | X | X | X | X | X | X | X | X | X | X | X | X | X | Vogelstein |
| | Final Report | 5 | | | | | | X | | | | X | | | | X | Vogelstein |