MRI Graph Analysis and Inference for Connectomics (MIGRAINE)

William Gray Roncal*†, Zachary H. Koterba*, Disa Mhembere†,
Dean M. Kleissas* Joshua T. Vogelstein‡§ Randal Burns† Anita R. Bowles¶
Dimitrios K. Donavos¶ Sephira Ryman‖ Rex E. Jung‖ Lei Wu** Vince Calhoun** and R. Jacob Vogelstein*†
*JHU Applied Physics Laboratory, Laurel Maryland 20723, USA. Email: willgray@jhu.edu
†Johns Hopkins University, 3400 N Charles Street, Baltimore, Maryland 21218, USA
‡Duke University, Durham, NC 27708, USA
§Child Mind Institute, 445 Park Avenue, New York, NY 10022, USA
¶University of Maryland, Center for Advanced Study of Language, 7005 52nd Avenue, College Park, MD 20742, USA
‖University of New Mexico, 1 University Blvd NE Albuquerque, NM 87131
**The Mind Research Network, 1101 Yale Blvd. NE, Albuquerque, New Mexico 87106

***Abstract*—Currently, connectomes (e.g., functional or structural brain graphs) can be estimated in humans at an $\mathcal{O}(1\ mm^3)$ scale using a combination of diffusion tensor imaging (DTI), functional magnetic resonance imaging (fMRI) and structural magnetic resonance imaging (MRI) scans. This manuscript summarizes a novel, scalable implementation of open-source algorithms to rapidly estimate magnetic resonance connectomes, using both anatomical regions of interest (ROIs) and voxel-size vertices. Here we provide an overview of the methods used, demonstrate our implementation, and discuss available user extensions. We conclude with a use case showing the efficacy of the pipeline and example results.**

***Index Terms*—connectomics, magnetic resonance imaging, network theory, pipeline**

I. Introduction

The ability to estimate an individual's connectome, i.e. a description of connectivity in an individual's brain, promises advances in many areas from personalized medicine to learning and education, and even to intelligence analysis [1], [2]. The ability to "classify" an individual's connectome further allows for inferring characteristics of an individual based on the degree to which his or her patterns of brain connectivity align with those observed in cohorts having known properties or outcomes, such as gender, handedness, intelligence, the ability to learn a foreign language, psychological impairments, disease susceptibility, etc. A robust analysis of these properties is on the horizon due to recent efforts to collect large amounts of multimodal magnetic resonance (MR) imaging data [3], [4], but will be hampered by the lack of a robust, reliable pipeline that is shareable across labs and institutions.

The basic approach for estimating structural MR connectomes is fairly well-established in the community. Other MR-scale connectome processing pipelines exist, such as [5], [6], [7], but these methods often lack robustness and repeatability, and in practice are difficult for users to modify, share, or scale; others only provide functional graphs [8]. In this manuscript, we propose an approach to remedy these deficiencies. The primary contribution of our efforts is the creation of a robust, high-throughput pipeline for estimating connectomes, beginning with diffusion MR images and MPRAGE structural data and ending with both small ~ 70 , and big $\sim 10^6$ vertex brain graphs. Further, we have made enhancements to the pipeline that include validation and analysis algorithms (e.g. graph embedding) to enable high-throughput, end-to-end solutions.

II. Framework

A. Initial Algorithms and Pipeline

We begin with an existing pipeline [9] implemented using the Java Image Science Toolkit (JIST) [10], and MIPAV [11]. We improve on this method by offering increased capability to scale to a production cluster environment, new functionality, and tools to incorporate new algorithms and export existing algorithms to new pipelines. as illustrated in Figure 1.

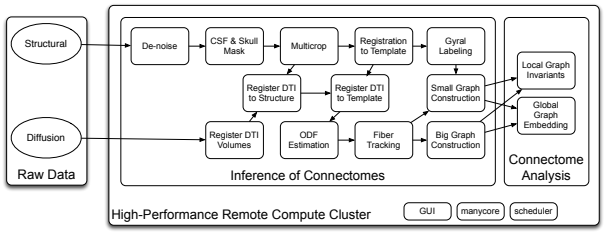

Fig. 1: MIGRAINE Pipeline Overview

B. LONI Processing Framework

The original JIST pipeline, consisting of 22 Java-based modules, have been wrapped and integrated within the LONI pipeline framework [12]. We created a library of tools that enables the importing of JIST-based algorithms into this environment. These tools improve module I/O and communication, and facilitate integration with the LONI environment and scheduler. Other modules (e.g., small and large graph generation, invariant computation) are written in Python. Finally, validation and packaging scripts have been developed to facilitate rapid data analysis. The pipeline is flexible and

can be modified using existing neuroimaging modules already adapted to LONI, such as [13] or custom code, as long as the algorithms are command-line executable. The baseline pipeline provides a robust, rapid reference implementation that has been shown to provide discriminative ability and is easily adapted to other methods.

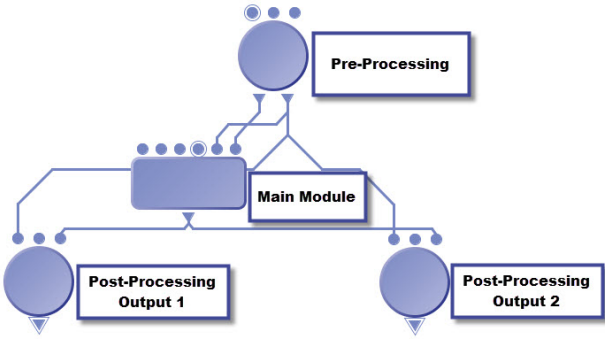

Fig. 2: Example of wrapped pipeline

III. GRAPH GENERATION

Our graph generation and analytic code utilizes the Magnetic Resonance One-Click Pipeline (MROCP), which we introduce here. The MROCP can be utilized as either a module within MIGRAINE (or another connectome pipeline) or as a standalone web-service, available at: http://openconnecto.me/graph-services/ [14]. It is currently the only known integrated tool that builds big (and small) graphs and has the capacity to compute graph analytics with a single click. MROCP's analytic computation is a facet of graph theory and thus germane to any graph/network and not only connectomes.

A. Small Graphs

Small graphs (e.g. 70 vertices and $\binom{70}{2} = 2415$ edges) were computed as detailed in [9] by labeling the structural brain volume with the Desikan atlas [15], and co-registering the structural and diffusion data. We then compute tensors, and utilize tracking algorithms (e.g. FACT) to produce fiber streamlines. Finally, we record an estimate of connectivity between each pair of regions (e.g., the number of times each region pair is connected by a fiber).

B. Big Graphs

To generate big graphs, we begin with the fiber streamlines and region of interest mask created during the small graph estimation process. We apply the mask (e.g. brain mask, gray matter, specific ROIs) to the data; the surviving voxels each become a vertex of the big graph. Next, we build a sparse column compressed graph where each vertex is a single imaged voxel. Edges represent a pair of vertices connected by a single fiber where both vertices are within the bounds of the interest regions defined by the mask. We iterate over each fiber streamline; an edge is recorded between every two vertices that can be reached (i.e. that are connected) by a single fiber.

1) Graph Analytics: Computing multivariate analytics on big graphs is a challenging endeavor due to the computational intensity associated with processing $\mathcal{O}(10^8)$ edge networks. Equivalent computational tasks are thus generally designated to specialized hardware like GPUs, graph processing engines like GraphLab [16] or distributed solutions like MapReduce. The MROCP takes a different approach, computing six high-accuracy multivariate analytics using only in-memory, CPU based algorithms and no preprocessing. This package is able to be easily integrated or deployed in new environments. The analytics currently computed are: Top-k eigenvalues and eigenvectors (TKE) [17], Local number of triangles (NT)[18], Clustering Coefficient (CC), Local Scan Statistic-1 (SS-1) [19], Degree and Edge count. An example of the invariants produced by the pipeline are shown in Figure 3.

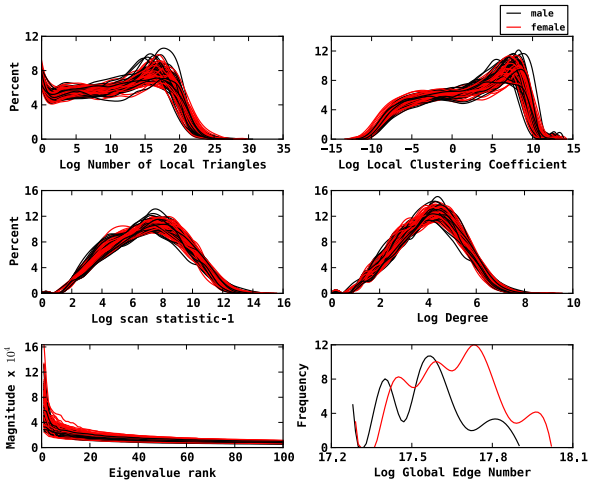

Fig. 3: Analytic computation on 114 subjects' brain graphs partitioned into male (black) and female (red)

We empirically determined that using the graph of the Largest Connected Component (LCC) [20], instead of the entire brain, produces highly representative data that is ideal for use in analytic computations. The LCC discards vertices with edge counts that are zero and near-zero, while maintaining fidelity to edge connectivity (0.7 ± 0.4 mean percent difference between global full graph and LCC number of edges). The LCC thus excludes sparsely connected grey matter voxels that are generally less useful for classification tasks and for statistically inferring dissimilarities in graphs. Subsequently, using the LCC results in a 90% reduction in processing time, and a system-wide memory usage reduction of approximately 46% (as measured on an 8 core, 2.4 GHz, 16GB RAM standalone server). We further benchmarked our approximation algorithms with ground truth graphs and found them highly accurate and within 94% of global expected results for any single analytic.

C. Validation

We demonstrated that our initial implementation of MIGRAINE matched the MRCAP baseline [9] for small graphs.

A variety of tools were developed to compare intermediate products (both quantitatively and qualitatively). Examples include matrix comparison tools and analysis of fiber counts. We subsequently made modifications to preprocessing steps for robustness and also decided to register subjects to a common template (e.g., MNI), instead of to individual subject space as in [9]. A common registration space enables both faster processing and additional analysis techniques (e.g. spatial graph analytics).

To validate that our graphs produce a repeatable signal, we used the KKI Test-Retest-Data [21] to analyze intra- inter-subject reliability, similar to [22]. We were able to show that the MIGRAINE pipeline produced a stable connectivity estimate across multiple scans of the same subject. The results differed by 13% from the MRCAP baseline and produced slightly better subject separability as shown in Figure 4.

Pipeline	Intra-Sub Diff	Inter-Sub Mean Diff	Closest Inter-Sub Diff	# Matches
MRCAP	26032	51584	38451	40/42
MIGRAINE	20378	56126	42663	42/42

Fig. 4: Validation showing improved discrimination relative to MRCAP using the KKI-21 dataset [21].

For all 42 graphs, the most closely related graph (as computed with the Frobenius norm) belonged to the same person, scanned at a different time. The subject comparison is shown in Figure 6. A visualization of the graphs for six test-retest pairs are shown in Figure 5.

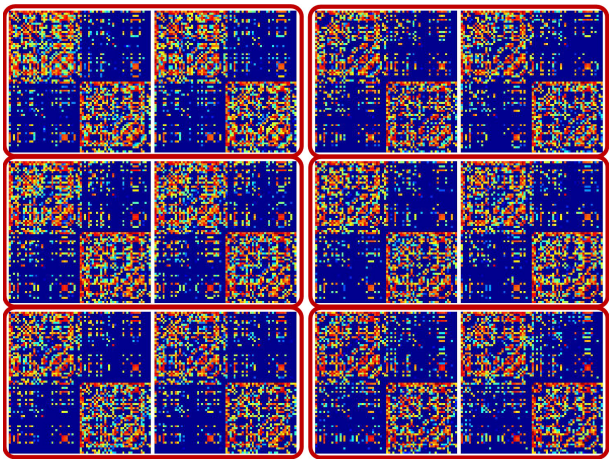

Fig. 5: Six Test-retest graphs. Top (L-R): Male, 25 years old (M25), F26, Middle: M25, F30, Bottom: M38, F61.

IV. Results

We successfully processed subjects from a variety of data corpora (both existing and new) totaling over 1500 subjects from multiple centers and acquisition paradigms.The datasets are described in the acknowledgements section, and we plan to process additional datasets as they become available. All

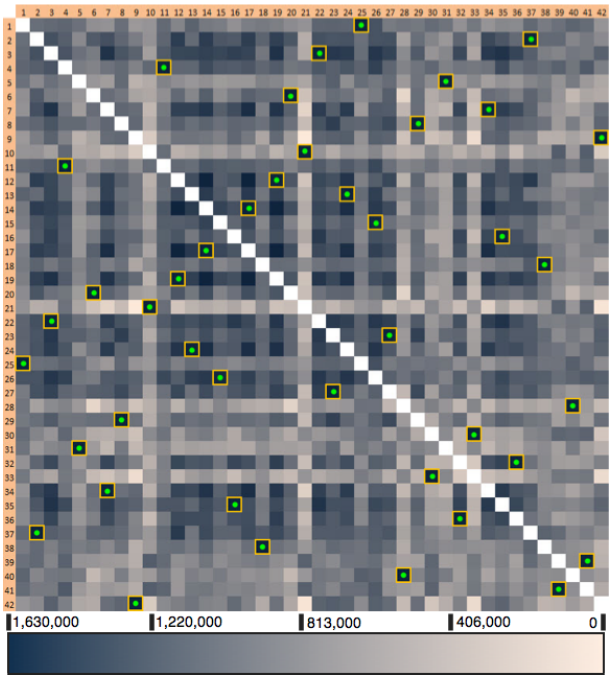

Fig. 6: KKI Test-Retest Data Comparison. Yellow boxes:lowest difference (or highest similarity), Green dots:True pairs, White: Exact Match (self comparison).

of the resulting graphs and analytics are currently being used to develop classifiers, provide new insight into the way brains are wired and to determine which aspects of the network are informative in predicting cognitive properties.

A. Scalability and Benchmarks

The current iteration of our software in the LONI Pipeline results in significant improvements to both scalability and processing time relative to the MRCAP baseline [9], which could produce a small graph in approximately 10 hours on our small cluster (248 concurrent nodes, 1 TB total RAM). On average, the MIGRAINE baseline takes approximately 3 hours/subject to compute small graphs (i.e., the output from MRCAP [9]), an additional 5 hours/subject to produce big graphs, and 3.5 hours/subject for graph invariants, for a total of 11.5h/subject. Much of this improvement is obtained by utilizing a common registration template, allowing for anatomical labels to be computed only once and then reused. Multi-node capabilities only contributes marginally for a single subject (in both pipelines) because the most intensive computations occur serially. However, there are significant efficiencies when evaluating a large number of subjects, and the number of nodes is the limiting factor. Run time for each of our datasets is presented in Figure 7. A univariate measure of total fiber count per subject is shown in Figure 8.

V. Conclusions

MIGRAINE is robust and has been validated on over 1500 subjects from a variety of datasets in a rapid, extensible,

Dataset	#	Time (Hours)				
		Small Graphs	Big Graphs	Invariants	Total	Average per Subject
KKI21	42	2.9	4.7	3.9	11.5	11.2
KKI21 [1K Test]	1000	14.4	23.7	20.0	58.1	11.4
CASL	36	3.1	4.6	3.7	11.3	11.0
NKI - TRT	24	3.2	5.2	3.4	11.8	11.6
MRN 111	111	1.5	6.1	3.0	10.5	9.8
MRN 1313	1313	9.9	37.3	18.3	65.5	10.2

Fig. 7: Total and average run times for each dataset.

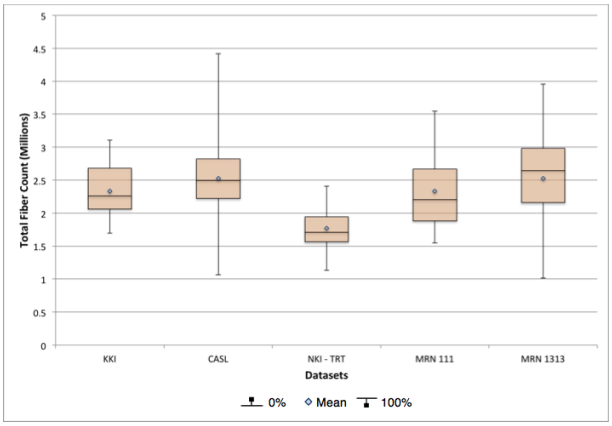

Fig. 8: Box plots for each data set, showing total fiber count for each data corpus.

fully automated framework. In addition to the standard small-graph results, we have demonstrated additional processing capability through the estimation of big graphs and analytics. The pipeline is scalable and has internal validation and packaging scripts that enable efficient analysis. Finally, we provide a demonstration of the classification signal present in our resulting brain graphs.

ACKNOWLEDGMENTS

The authors thank the Image Analysis and Communications Laboratory at Johns Hopkins University, Baltimore, MD.
This work has been supported by the NSA Research Program on Applied Neuroscience and NIH/NINDS 5R01NS056307.

Datasets used for testing MIGRAINE are as follows:

- KKI21: Consists of 21 subjects, each with test-retest scans as described in [21].
- CASL: T1 weighted 3D MPRAGE (1 mm isotropic) whole brain images along with 64 direction non-collinear diffusion weighted images (b =1000 s/mm?) were collected from 36 right-handed participants (16 male, 20 female). Additionally, the dataset included scores on a set of cognitive and perceptual behavioral measures as well as data from a training task focused on word learning.
- MRN111: Consists of 111 DTI and MPRAGE scans collected at the University of New Mexico, and funded through a grant from the John Templeton Foundation entitled "The Neuroscience of Creativity."
- MRN1313: 1313 subjects collected via a data sharing program at MRN, including 1171 healthy controls, 88 schizophenias and 54 neuropsychiatricpatients with various diagnoses (bipolar, OCD, depression, etc).
- NKI - TRT: Consists of 24 paired subjects. Data is available online at http://fcon_1000.projects.nitrc.org/indi/enhanced and described in [23].

REFERENCES

[1] O. Sporns, "Networks of the Brain," *Learning*, no. August, p. 375, 2010.
[2] J. W. Lichtman and J. R. Sanes, "Ome sweet ome: what can the genome tell us about the connectome?" *Current opinion in neurobiology*, vol. 18, no. 3, pp. 346–53, Jun. 2008.
[3] D. C. Van Essen *et al.*, "The Human Connectome Project: a data acquisition perspective." *NeuroImage*, vol. 62, no. 4, pp. 2222–31, Oct. 2012.
[4] M. Mennes, B. B. Biswal, F. X. Castellanos, and M. P. Milham, "Making data sharing work: The FCP/INDI experience." *NeuroImage*, Oct. 2012.
[5] Z. Cui *et al.*, "PANDA: a pipeline toolbox for analyzing brain diffusion images." *Frontiers in human neuroscience*, vol. 7, no. February, p. 42, Jan. 2013.
[6] Y. Cointepas *et al.*, "BrainVISA: Software platform for visualization and analysis of multi-modality brain data," *Neuroimage*, vol. 13, no. 6, pp. 98–98, 2001.
[7] A. Daducci *et al.*, "The connectome mapper: an open-source processing pipeline to map connectomes with MRI." *PloS one*, vol. 7, no. 12, p. e48121, Jan. 2012.
[8] S. Sikka *et al.*, "Towards Automated Analysis of Connectomes: The Configurable Pipeline for the Analysis of Connectomes (C-PAC)," *Neuroinformatics*, 2012.
[9] W. R. Gray *et al.*, "Magnetic Resonance Connectome Automated Pipeline," *Pulse*, no. APRIL, pp. 1–5, 2012.
[10] B. C. Lucas *et al.*, "The Java Image Science Toolkit (JIST) for rapid prototyping and publishing of neuroimaging software." *Neuroinformatics*, vol. 8, no. 1, pp. 5–17, Mar. 2010.
[11] M. McAuliffe *et al.*, "Medical Image Processing, Analysis and Visualization in clinical research," *Proceedings 14th IEEE Symposium on Computer-Based Medical Systems. CBMS 2001*, pp. 381–386, 2001.
[12] I. D. Dinov *et al.*, "Efficient, Distributed and Interactive Neuroimaging Data Analysis Using the LONI Pipeline." *Frontiers in neuroinformatics*, vol. 3, p. 22, Jan. 2009.
[13] M. Jenkinson *et al.*, "FSL," *NeuroImage*, vol. 62, no. 2, pp. 782–790, 2012.
[14] R. Burns *et al.*, "The Open Connectome Project Data Cluster : Scalable Analysis and Vision for High-Throughput Neuroscience Categories and Subject Descriptors."
[15] R. S. Desikan *et al.*, "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest." *NeuroImage*, vol. 31, no. 3, pp. 968–80, Jul. 2006.
[16] J. E. Gonzalez, D. Bickson, and C. Guestrin, "PowerGraph : Distributed Graph-Parallel Computation on Natural Graphs," pp. 17–30.
[17] C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, 1950.
[18] C. E. Tsourakakis, "Fast Counting of Triangles in Large Real Networks : Algorithms and Laws."
[19] H. Pao, G. A. Coppersmith, and C. E. Priebe, "Statistical Inference on Random Graphs : Comparative Power Analyses via Monte Carlo," 2010.
[20] E. Jones, T. Oliphant, P. Peterson, and Others, "SciPy: Open source scientific tools for Python," 2001.
[21] B. A. Landman *et al.*, "Multi-parametric neuroimaging reproducibility: a 3-T resource study." *NeuroImage*, vol. 54, no. 4, pp. 2854–66, Mar. 2011.
[22] W. Gray *et al.*, "Magnetic Resonance Connectome Automated Pipeline and Repeatability Analysis," *Society for Neuroscience Abstract*, 2011.
[23] K. B. Nooner *et al.*, "The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry," *Frontiers in Neuroscience*, vol. 6, no. 152, 2012.