Ph.D. Preliminary Meeting: Gregory Kiar

October 19th, 2017

Reliable, Robust, and Accessible Big Data Neuroscience

With an explosion in data collection across nearly all scientific disciplines, we are entering the big-data era of scientific exploration. From the Sloan Digital Sky Survey [1] in cosmology, to the UK Bio-Bank [2] and many others in Neuroscience, initiatives are being launched to federate this data-tsunami, many with particular emphasis and commitment towards open data-sharing. As access to data has increased drastically, it has become apparent that a lack of reproducibility in data analysis methods or collection paradigms is becoming a plague in many current disciplines of science [3]. Unless addressed, our analyses will continue to fail to do justice to the extraordinary datasets we possess. Computational sciences in particular have an opportunity to develop more rigorous practices and provide tools which lower the barrier for producing reproducible analyses.

In the first portion of my Ph.D. I propose to develop and release an open-source platform which enables researchers to reproduce and distribute computational pipelines across a variety of environments both locally and in the cloud, empowering researchers to create, evaluate, and share more reliable, and therefore more impactful scientific tools, datasets, and discoveries. This platform will be called Clowdr. The second portion of my Ph.D. will be the development and analysis of provenance graphs which record pipeline executions, enabling system-level evaluation of the stability and robustness of tools. The third and final portion of my Ph.D. will be the application of the above methodologies towards generating a neuroinformatics pipeline and evaluating their efficacy for producing generalizable scientific results.

Part I: Clowdr

Clowdr will provide three modes of operation: prototype, parallelize, and publish. The prototype mode will enable users to develop and run their tool on local resources by leveraging virtual containers such as Docker [4] or Singularity [5]. While pipeline accuracy may be evaluated in a manner suitable for a given domain, pipeline execution records will be recorded using Reprozip [6] and later analyzed.

Once tools are deemed stable and ready to be deployed across large cohorts of data, the parallelize mode will allow users to distribute these pipelines on high performance computing (HPC) environments. Initially, this will be done using the Amazon Web Services Batch service, which acts as a computing cluster in the cloud; this will be later extended to operate within the Compute Canada HPC environment through CBRAIN [7]. This integration will accelerate the adoption of new tools and datasets into CBRAIN, and increase the accessibility of this platform through the existing web interface and user-base. Integrations with data management services such as LORIS [8] and Datalad will also increase the accessibility of the datasets they expose.

Finally, the publish mode will launch a web server which provides a website and the RESTful API, Apine, which will be developed leveraging the existing CARMIN [9] standard, for users to interrogate or share the execution information of their tool, as well as the derived data results. Clowdr will be the only known tool which integrates the development, evaluation, deployment, and distribution/publication of tools, datasets, and analyses all together, and in the cloud.

Part II: Provenance Graph Evaluation

Clowdr will provide a foundation for the second stage of my PhD: the terascale evaluation of neuroimaging pipelines in the cloud. As tools launched through the above infrastructure will be executed in a controlled environment, provenance traces of commands run and files written or accessed will be created for each execution. Here we consider a *provenance trace* to be a record of all operations performed by the machine during the evaluation of a given tool. Recording and evaluating provenance enables sources of instability across parameter settings, data, or computational infrastructures to be identified, isolated, and corrected. A recent study in which cortical surfaces estimated with FreeSurfer [11] were shown to be significantly different when generated with insignificantly different input data [12] is one example for which this approach will be valuable. It is important to note that as the data processed by a given pipeline will not be stored in the graph, direct quality control will not be performed but rather inter-execution differences.

Representing these provenance traces as graph objects, wherein a node is a system command and an edge is an input parameter or file, we are able to apply graph statistics to analyze properties of these traces and compare similar traces to one another. As pipeline execution flow is sequential, the graphs will be directed and acyclic. As there are many features which can be recorded such as file size, time, network speed, etc., the graphs will be richly attributed. The resulting provenance graphs will be henceforth referred to as Richly Attributed Directed Acyclic Record (RADAR) graphs.

As no known studies explicitly explore the properties of RADAR graphs or apply graph statistics to execution provenance records, we will consider previous work spanning portions of this space. The features explored in [13] have previously been used to identify differences in source-code for Malware detection, and provide insight into statistics that may be useful in the context of software. The OddBall algorithm [14] has been developed by identifying common characteristics of induced subgraphs from arbitrary weighted networks for identifying potential nodes of interest, and demonstrates a potentially valuable methodology for learning graph properties of typical pipelines. Additionally, four types of downstream effect modifications have been categorized based on directed acyclic graphs [15], which will be explored in the context evaluating RADAR graphs. I will unify approaches across various domains of graph statistics and propose a methodology for comparing RADAR graphs and those produced by other provenance-based initiatives such as NIDM [16].

The methodology developed above will be tested against two classes of familiar pipelines in neuroimaging: those believed to be unstable empirically (e.g. FreeSurfer [11], FSL’s FNIRT [17]), and those believed to be stable (e.g. ndmg [18] , FSL’s FLIRT [17]). Executions will be evaluated through quality control metrics and manual inspection, receiving a score on a spectrum from “unstable” to “stable.” The RADAR processing methodology will be employed and iterated upon until there is agreement between the reported statistic and manual inspections. These graph statistics will then be tested against new tools.

Part III: Neuroinformatics Tool Development

The third and final portion of my Ph.D. will demonstrate the efficacy of Clowdr and the statistical analysis of RADAR graphs through the principled development and refinement of a neuroinformatics experiment leveraging open-access data and based on the joint optimization of robustness, reproducibility, and stability. This will ultimately serve as a demonstration of the utility of the methods developed and provide a scientific basis for the potential impact of this principled approach and methodology for tool development in neuroscience and beyond.

As developing a novel neuroimaging pipeline from scratch is an extraordinary amount of work, an existing tool (or tools) to be selected later will be evaluated, the sources of instability corrected, and the changes contributed back to the original project. I will then attempt to replicate and generalize an impactful study produced using the tool before and after the modifications, to quantify the scientific impact of the core changes. This will effectively demonstrate the proposed methodology, encourage developers to continue using this approach when generating new analyses, and have an immediate impact on existing neuroimagers and their research.

Justification of Academic Background

Throughout my Master's degree I developed a structural connectome estimation pipeline from multimodal MRI (M3RI) scans of the human brain. I packaged this tool using containerization engines such as Docker and Singularity, documented it with the Boutiques descriptive command-line framework, and deployed it across a variety of computational infrastructures including Amazon Web Services, CBRAIN, and OpenNeuro. The packaging and deployment of this pipeline afforded me considerable experience in a variety of high performance computing environments, and required that I learn to evaluate pipeline performance across infrastructures and datasets and compare the quality of produced derivatives.

The pipeline I developed was optimized with respect to the reliability and stability of resulting connectomes, and provided the user with a variety of intermediate quality control figures to enable proof-reading of the results and ensuring their accuracy qualitatively (as no ground-truth exists, quantitative accuracy is not achievable). The connectomes produced were also evaluated through a variety of graph statistical measures that operated on either an entire graph, its nodes, weighted edges, binary edges, or the largest connected component of the connectome.

During my tenure at Johns Hopkins I also taught several courses on applications of graph statistics, including Introduction to Connectomics, Statistical Connectomics, The Art of Data Science, and Neuro Data Design which was a design class for undergraduate engineering and computer science students who were tasked with developing and validating properties of their own neuroimaging processing pipelines that operated on Electron Microscopy, CLARITY, M3RI, or EEG data. I have significant experience in applying graph theoretical measures to biologically derived networks.

I believe that my experience in pipeline development, the evaluation and execution of reliable high performance computing systems, and applications of graph statistics to real-world graphs will be invaluable assets towards me completing the project described and outlined above.

References

[1] York DG. Astronomical J. 2000 Sep;120(3):1579.
[2] Sudlow C. PLoS med. 2015 Mar;12(3):e1001779.
[3] Baker M. Nature News. 2016;533(7604):452.
[4] Merkel D. Linux Journal. 2014;2014(239):2.
[5] Kurtzer GM. PloS one. 2017; 12(5):e0177459.
[6] Chirigati F. ICMD 2016 (pp. 2085-2088). ACM.
[7] Sherif T. Frontiers in neuroinf. 2014;8.
[8] Das S. Frontiers in neuroinf. 2011;5.
[9] Glatard T. Fronteirs in neuroinf. Aug 2015.
[10] Harary F. Mat. di Palermo. 1960 May;9(2):161-8.
[11] Fischl B. Neuroim. 2012 Aug 15;62(2):774-81.
[12] Lewis L. OHBM 2017 Proceedings. 2017 Jun 25.
[13] Dullien T. SSTIC. 2005 Jun;5(1):3.
[14] Akoglu L. PAKDD 2010. vol 6119.
[15] VanderWeele TJ. Epidem. 2007 Sep;18(5):561-8.
[16] Ghosh S. OHBM 2017 Proceedings. 2017 Jun 25.
[17] Jenkinson M. Neuroim. 2012 Aug;62(2):782-90.
[18] Kiar et al. bioRxiv. 2017 Jan 1:188706.