Computing Scalable Multivariate Glocal Invariants of Large Brain-Graphs (with a Web-service to boot)

JOHNS HOPKINS UNIVERSITY

Disa Mhembere, William Gray Roncal, Daniel Sussman, Rex Jung, Sephira Ryman, R. Jacob Vogelstein, Carey Priebe, Joshua T. Vogelstein, Randal Burns

disa@jhu.edu

Motivation

Graphs have quickly emerged as a leading abstraction for the representation of data that can potentially encode multivariate relationships between data points. "Connectomics" models the brain as a graph [1]; vertices correspond to neurons (or collections thereof) and edges correspond to structural or functional connections between them. Computing their invariants may enable neuropsychiatric diagnosis by partially extracting biologically-relevant characterizations of the human brain. Magnetic Resonance-derived connectome graphs (connectomes) are of scale $O(10^6)$ vertices and $O(10^8)$ edges. Their scale makes computing invariants (graph statistics) extremely challenging as their size and order can far surpass the bounds of current software tools [2, 3] available for analysis.

To address these computational deficiencies we present an open-source package and complementary Web-services to compute six high-accuracy "glocal" invariants. We define multivariate glocal graph invariants as features of the graph that simultaneously capture various local and global topological properties.

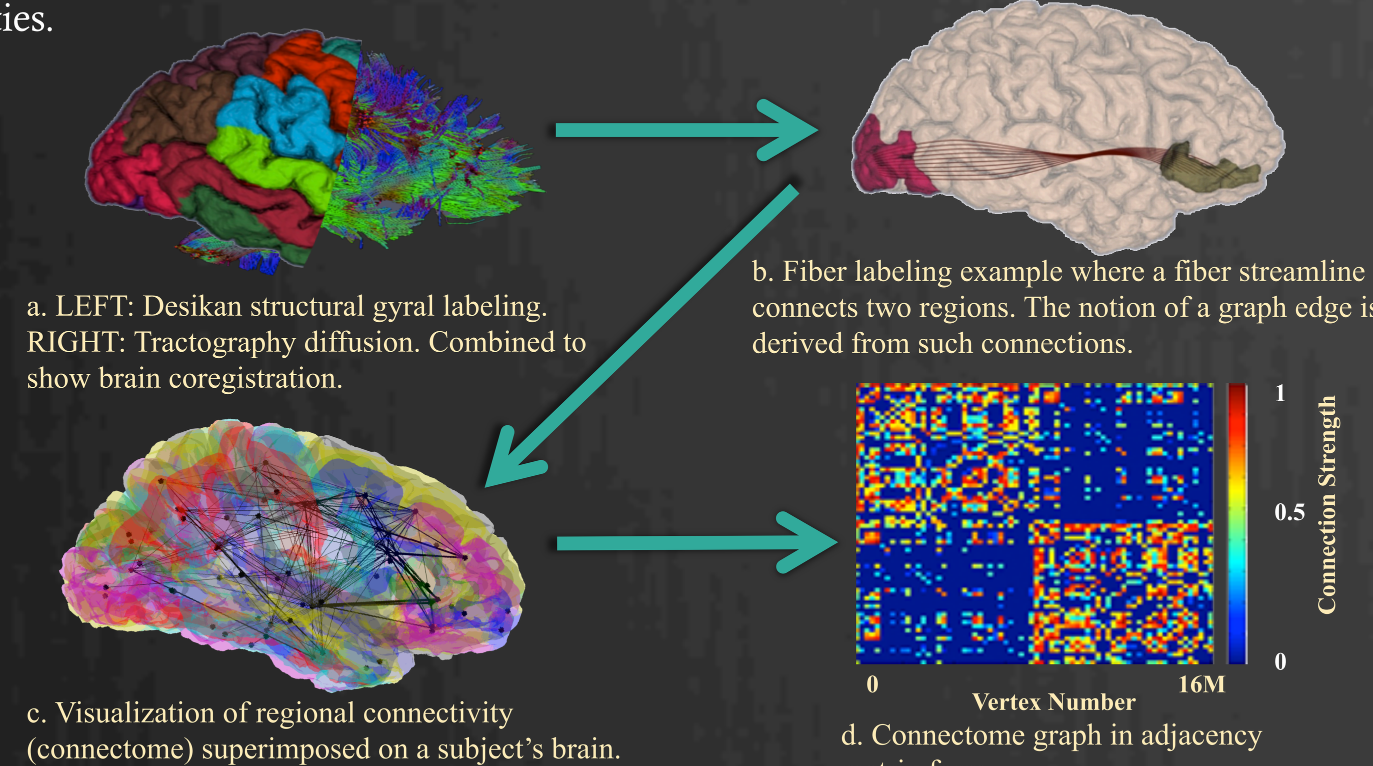

Figure 1. Illustration of the last phase of a connectome estimation pipeline that produces graphs on which invariants are computed [4].

The Invariants

The invariants computed are chosen for their effectiveness in characterizing network connectivity [5]. For all equations below let G be a graph with vertices $v \in n$, *the set of all vertices.*

- Latent Position-k Matrix [6], $LP-k \in^{k\times n}$: via Lanczos' algorithm for spectral decomposition
- Degree Vector, $Deg \in [n]^n$: $Deg(v,G) = A\mathbf{1}, where \mathbf{1}$ *is an n-dimensional vector of ones*
- Number of Local 3-Cliques Vector [7], $NL-3 \in [\binom{n}{3}]^n$: $NL-3(v,G) \approx \frac{1}{2}\sum_{k=1}^{K} \lambda_k^3(G) x_{vk}^2(G)$

where, $_{vk}(G)$ *is the* v^{th} *entry of the* k^{th} *eigenvector of* A, *and* $\lambda_k(G)$ *is the* k^{th} *eigenvalue,* K *is the number of eigen-pairs used in the approximation*

- Local Clustering Coefficient [8], $CC \in [\binom{n}{3}]^n$:

$$CC(v,G) = 2\times NL\text{-}3(v,G)/(Deg(v,G)(Deg(v,G)-1))$$

- Local Scan Statistic-1 [5], $SS-i \in [n]^n$: via edge counting. We count the number of edges in the induced sub-graph within the *i-hop neighborhood* i.e. $size\,(\Omega(N_i[v;G]))$

Mobile scan it!

There are currently no known integrated packages for computing such invariants on brain-graphs of this size.

We provide downloadable **open-source code** for computing invariants on your own machine, in addition to publicly-accessible **Web-services** that run on our data-intensive cluster.

Figure 2. Mobile Code for invariant web-services and packaged code download.

Methodology

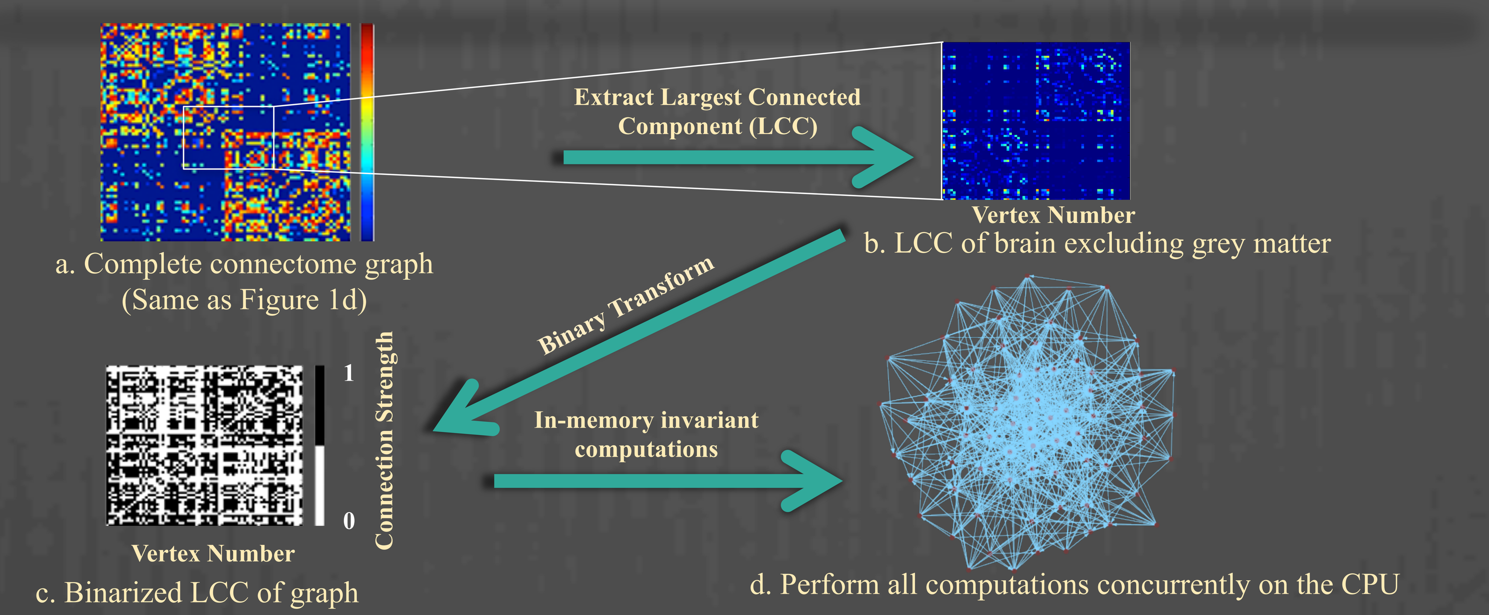

Figure 3. Pipeline showing data transformation phases where we extract the LCC [6] prior to brain-graph binarization and invariant computation.

Software Stack

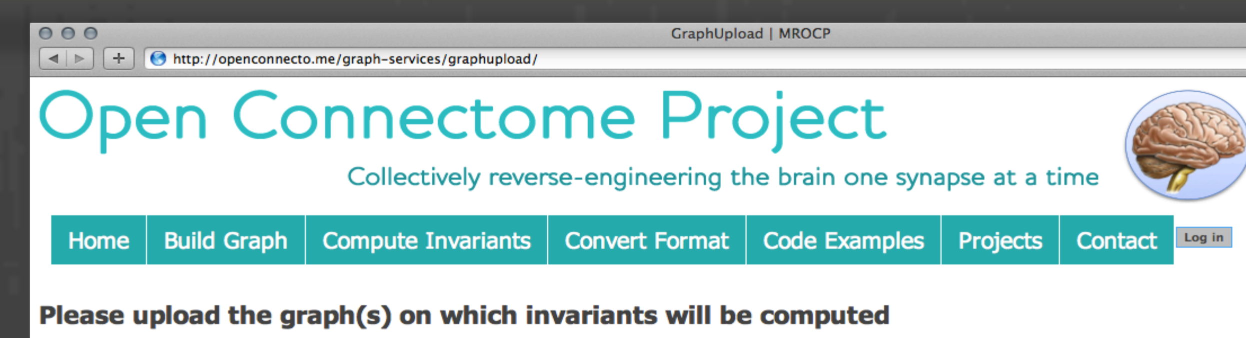

Figure 4. Screenshot of Web-services for graph building and invariant computation on graphs in accepted formats as described online.

- We developed an accurate and efficient method to compute multivariate invariants on *any graph*, including voxel-wise brain-graphs.
- We created Web-services for the public to do so as well, found at: http://openconnecto.me/graph-services/
    - Batch (multiple graphs) and programmatic interfaces are also supported.
- You can also download our Python 2.7 modules at: https://github.com/openconnectome/MR-connectome/tree/stand-alone

Experimental Results

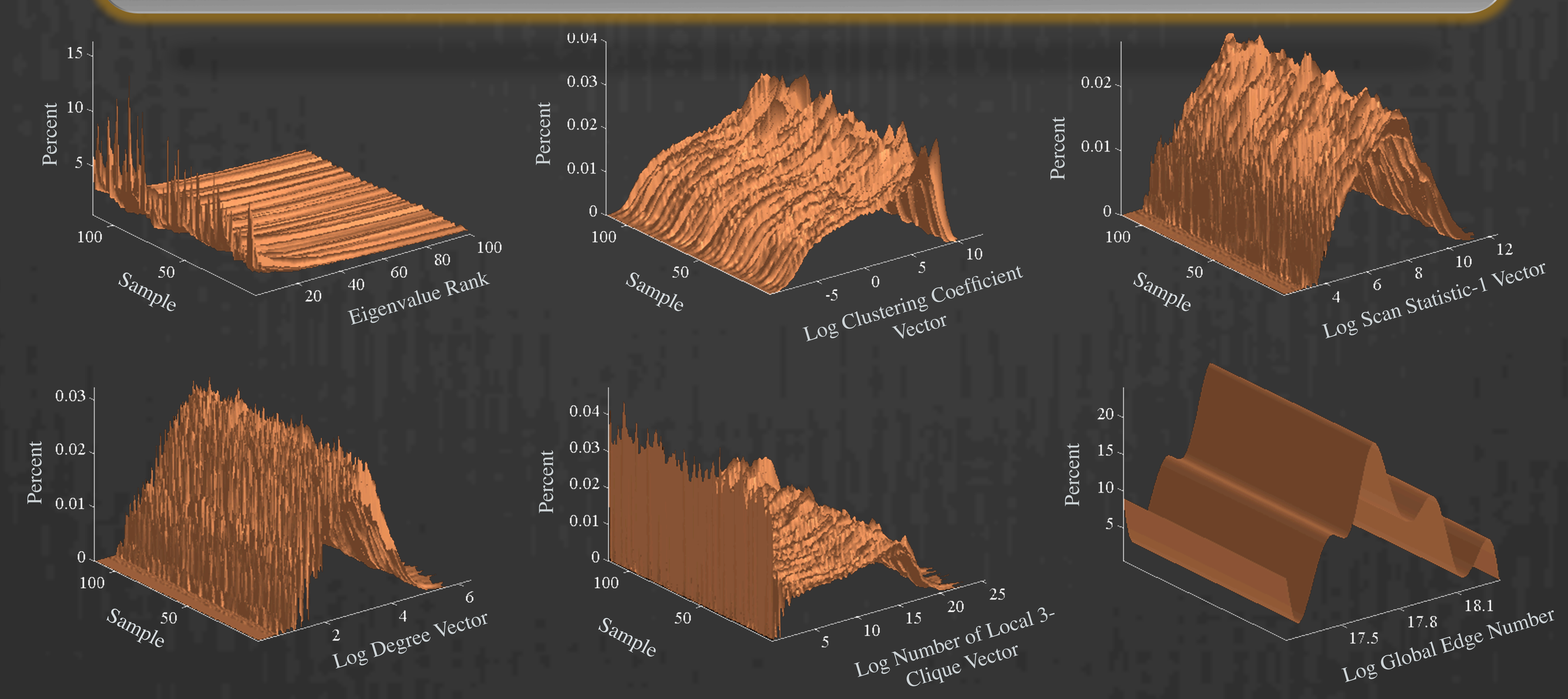

Figure 5. Multivariate invariants computed on 120 subjects. Note that NL-3 uses an approximation algorithm only utilizing the top-k eigenvalue and eigenvector pairs. At 100 eigen-pairs we obtain ~91% accuracy on graphs with $O(10^6)$ nodes and $O(10^8)$ edges. All other unrelated invariants (i.e. LP-k, SS-1, Deg) are exact.

Performance

Invariant	Time Complexity	Mean time per vertex at 1 core
Clustering Coefficient Vector	O(n + k)	0.59 (± 2.4) μs
Number of Local 3-Cliques Vector	O(n + k)	48.51 (± 0.9) μs
Degree Vector	O(n)	66.64 (± 1.8) μs
Scan Statistic-1 Vector	O(nm)	0.47 (± 0.2) ms
Latent Position-100 Matrix	O(100(m+n))	45.41 (± 0.1) μs

Table 1. Asymptotic "Big O" analysis of algorithms used and experimental time measurements using a single 8 core, 2.4 GHz processor Linux server with 16 GB of RAM.

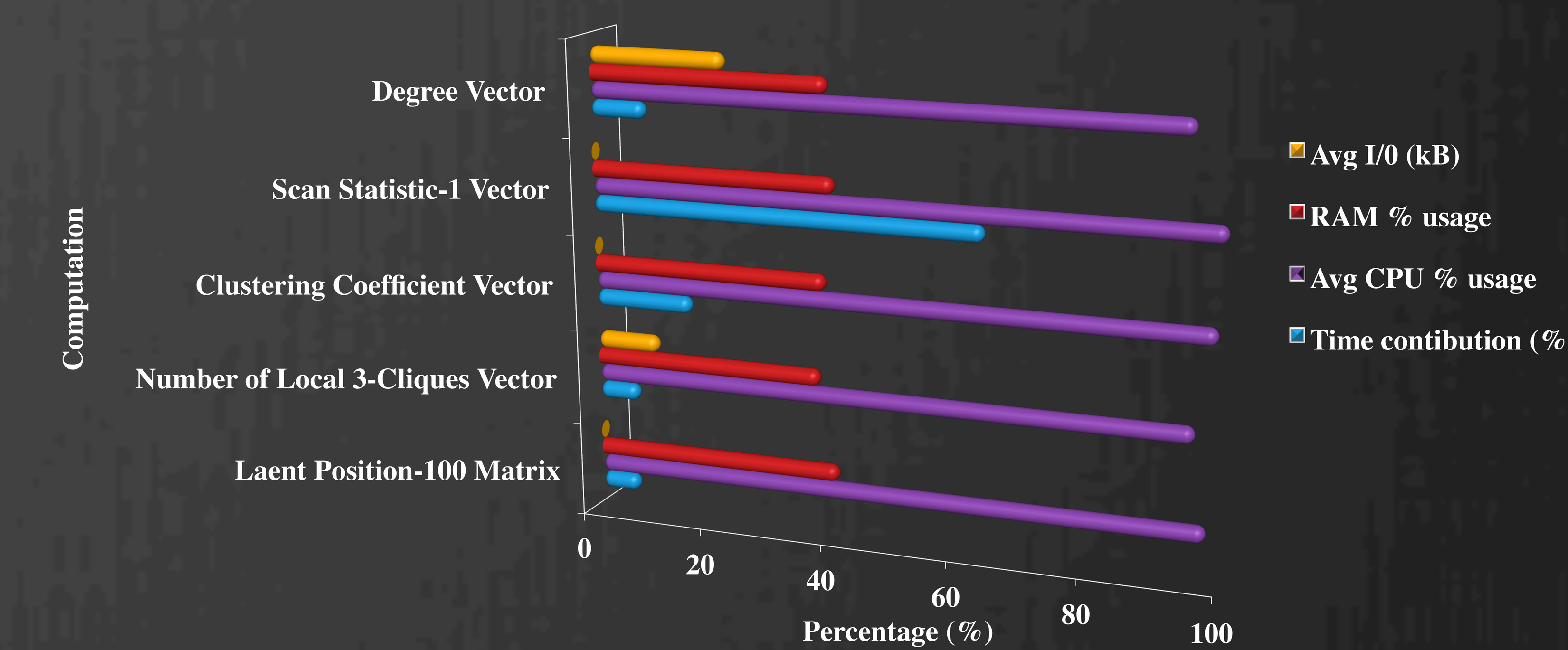

Figure 6. Performance of each invariant when computed serially and independently of any non-dependent invariants on an 8 core, 2.4 GHz processor Linux server with 16 GB of RAM. Total compute time for all invariants is ~3.7 h per graph, per core.

Future Work

As extensions to this work, we hope to:

- Develop fully parallelized implementations of all invariants.
- Generate the entire graph from raw diffusion/function MRI data.
- Add additional invariants such as Graph Diameter and Scan Statistic-2.
- Support a wider range of graph data formats for Web-Services.

References and Acknowledgements

1. Iturria-Medina et al (2007). Neuroimage 'Characterizing brain anatomical connections using diffusion weighted MRI and graph theory'.
2. Rubinov, M et al (2010). Complex network measures of brain connectivity: uses and interpretations.
3. Hagberg, A et al (2010). Exploring network structure, dynamics, and function using networkx.
4. Gray, William R., et al (2012). "Magnetic resonance connectome automated pipeline: an overview." Pulse, IEEE 3.
5. Henry PAO et (2011). 'Statistical Inference on Random Graphs: Comparative Power Analyses via Monte Carlo'.
6. Eric Jones et al (2001). 'SciPy: Open source scientific tools for Python'. http://www.scipy.org.
7. Charalampos E. Tsourakakis (2008).'Fast Counting of Triangles in Large Real Networks: Algorithms and Laws'.
8. Jari Saramäki et al (2007). 'Generalizations of the clustering coefficient to weighted complex networks'.

This work was supported by the National Institutes of Health (NIBIB 1RO1EB016411-01) and the National Science Foundation (OCI-1244820).