Bias and instability in graph theoretical analyses of neuroimaging data

Mark Drakesmith[1], Karen Caeyenberghs[2], Anirban Dutt[3], Glyn Lewis[4], Anthony S David[3], and Derek K Jones[1]

[1]CUBRIC, Cardiff University, Cardiff, Wales, United Kingdom, [2]Department of Physical therapy and motor rehabilitation, Ghent University, Gent, Belgium, [3]Institute of Psychiatry, Kings College London, London, United Kingdom, [4]Academic Unit of Psychiatry, University of Bristol, Bristol, United Kingdom

Target Audience: Any neuroimager using or considering graph theory as a tool for analysing structural or functional brain connectivity.

Introduction: Graph theory (GT) is a powerful mathematical framework for quantifying topological properties of networks derived from functional and structural neuroimaging [1,2]. However, all connectivity inference methods are subject to false positives (FPs), which inevitably impact on inferred network topology. A common strategy for overcoming FPs is to threshold the matrices to exclude smaller, and potentially spurious, connections from the network [2]. Some studies have explored the reliability and validity of GT metrics with mixed results[3-5]. However, there has yet to be any investigation into: (1) the sensitivity of GT metrics to FPs; (2) the validity of thresholding matrices; and (3) the reliability of statistical comparison of GT metrics.

Method: A 'ground truth' structural connectivity matrix was derived from a human tractography dataset (deterministic tractography with the damped Lucy-Richardson algorithm[6]: 3×3×3mm grid of seed points in white matter, 1mm step size, 45° threshold), from a HARDI acquisition (cardiac-gated EPI sequence, TE=87ms, 60 gradient orientations, 6 unweighted B0s, b-value=1200 smm^{-2}, FOV=96×96mm, 60 slices, voxel-size=1.6×1.6×2.4mm). Tracts termination points were registered to the AAL atlas, creating a 116×116 connectivity matrix. Additional *a priori* anatomical constraints were applied to eliminate spurious streamlines. FPs were randomly introduced into the network over 250 trials. Four common GT metrics were computed from the noisy networks and compared with those of the 'ground truth' network: (1) global efficiency; (2) mean clustering coefficient; (3) mean betweenness; and (4) smallworldness. The effects of thresholding on statistical inference were tested further on a large cohort of tractography data (n=248 subjects, same acquisition parameters as above). Datasets were randomly assigned to two equally sized groups across 100 randomisations. For each randomisation, and for each GT metric, an unpaired *t*-test was performed and the stability of the statistic across thresholds was calculated using the squared-sum derivative of *p*-value across thresholds. GT metrics were compared to two non-GT network metrics: number of edges and number of streamlines. Additional comparisons between a 'healthy' and 'atrophied' group of networks was compared to test stability in the presence of genuine group differences. Short-range connections in the 'atrophied' group were reduced by a random proportion (Gaussian, μ=0.05, σ=0.01).

Results: All FP rates produced significant deviation (p<0.05) from the ground truth in all GT metrics, even with a single FP (fig. 1). Thresholding the network flattens the effect of FPs (fig. 2), but also introduces a large bias (fig. 3). *t*-tests were significantly more unstable across thresholds for GT measures compared to non-GT measures, which varied more smoothly across thresholds (fig. 4). Smallworldness shows the lowest stability, especially at high thresholds. Results for healthy-atrophied comparison show some GT metrics (clustering coefficient and smallworldness) do show effects across a window of thresholds (fig. 5).

Discussion: GT analysis on tractography data is highly sensitive to FPs (we expect similar issues should to apply for functional connectivity). Thresholding effectively flattens the effect of FPs, but at the expense of introducing further bias to GT metrics. Great care is required in pre-processing and analysis of neuroimaging derived networks to ensure errors such as FPs are minimized. Thresholded networks are unlikely to be comparable between different pipelines and acquisition parameters, and in such cases standard statistical inferences on these metrics will be invalid. Statistical comparisons of GT metrics are unstable but genuine group differences can be detected in 'windows of stability' across thresholds. Based on these results, we propose a strategy where permutation tests are performed across thresholds and only clusters of significant results above a critical number of thresholds are considered truly significant effects. This will control for the instability of statistical inference and increase sensitivity to genuine group differences.

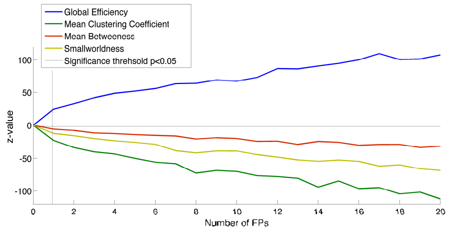

Fig. 1. *z*-statistics for deviation from ground truth for each GT metric.

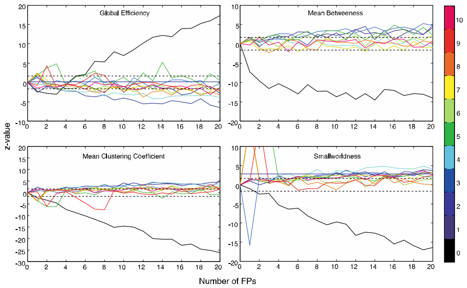

Fig. 2. *z*-statistics for deviation from ground truth across thresholds (threshold indicated by color code). Dotted lines indicate threshold for significance of p<0.05.

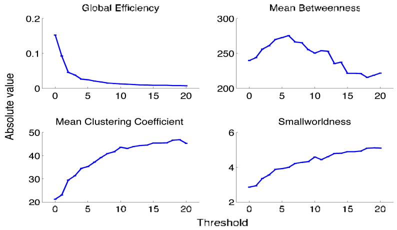

Fig. 3. Threshold-induced bias in GT metrics. Error bars indicate standard error across trials.

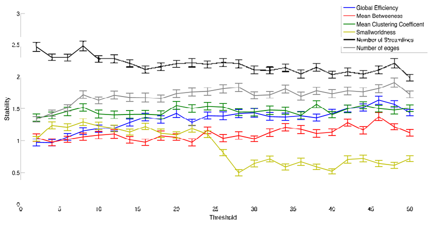

Fig. 4. Stability measures of *t*-statistics across thresholds for GT and non-GT metrics. Error bars indicate standard error across randomizations.

References: [1] Bullmore E & Sporns O. *Nat. Rev. Neurosci*, 2009;10:186-98. [2] Rubinov M & Sporns O. *NeuroImage*. 2010;52:1059–69.[3] Bastiani M. et al. NeuroImage. 2012;62:1732-49. [4] Langer N. et al. *PLoS ONE*. 2013;8:e53199. [5] Braun U. et al. *NeuroImage*. 2012;59:1404–1412. [6] Dell'Acqua F, et al. *NeuroImage*. 2010;49:1446–58.

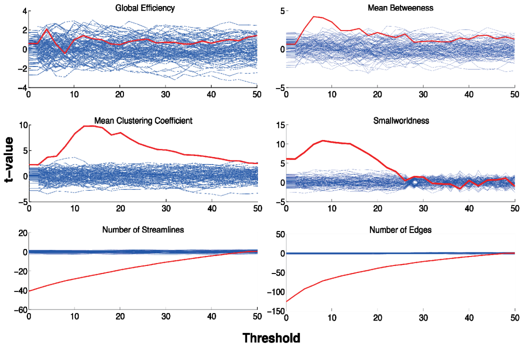

Fig. 5. *t*-values computed between healthy and atrophied networks (red) and random permutations (blue) for GT and non-GT metrics.