Published in final edited form as:
Med Image Comput Comput Assist Interv. 2012 ; 15(0 3): 305–312.

Test-Retest Reliability of Graph Theory Measures of Structural Brain Connectivity

Emily L. Dennis[1], **Neda Jahanshad**[1], **Arthur W. Toga**[2], **Katie L. McMahon**[3], **Greig I. de Zubicaray**[5], **Nicholas G. Martin**[4], **Margaret J. Wright**[4,5], and **Paul M. Thompson**[1]

[1]Imaging Genetics Center, Laboratory of Neuro Imaging, UCLA, CA, USA

[2]Laboratory of Neuro Imaging, UCLA, CA, USA

[3]Center for Advanced Imaging, Univ. of Queensland, Brisbane, Australia

[4]Queensland Institute of Medical Research, Brisbane, Australia

[5]School of Psychology, University of Queensland, Brisbane, Australia

Abstract

The human connectome has recently become a popular research topic in neuroscience, and many new algorithms have been applied to analyze brain networks. In particular, network topology measures from graph theory have been adapted to analyze network efficiency and 'small-world' properties. While there has been a surge in the number of papers examining connectivity through graph theory, questions remain about its test-retest reliability (TRT). In particular, the reproducibility of structural connectivity measures has not been assessed. We examined the TRT of global connectivity measures generated from graph theory analyses of 17 young adults who underwent two high-angular resolution diffusion (HARDI) scans approximately 3 months apart. Of the measures assessed, modularity had the highest TRT, and it was stable across a range of sparsities (a thresholding parameter used to define which network edges are retained). These reliability measures underline the need to develop network descriptors that are robust to acquisition parameters.

1 Introduction

Graph theory is increasingly used to analyze brain connectivity networks. Graph theory, a branch of mathematics concerned with the description and analysis of graphs, describes the brain as a set of nodes (brain regions) and edges (connections). Information on either structural or functional connectivity may be expressed in connectivity matrices, from which various network properties may be derived, such as clustering, efficiency, or small-world organization. Several of these measures have been shown to change during childhood development [1], and to be heritable [2], associated with specific genetic variants [3, 4] and be altered in various neuropsychiatric disorders [5]. To date, only one study has examined the test-retest reliability (TRT) of these measures for *structural* networks, finding high reliability [6], but they did not examine different network sparsities. Results in the TRT of these measures in *functional* networks have been inconsistent. Low reliability [7], remarkably high reliability [8], and moderate reliability [9, 10] have all been found. To define which connections are present in a network, often a sparsity threshold is applied, to

retain only those connections whose edge strengths exceed a given threshold, or to eliminate "weaker" connections. Telesford et al. [8], found that the reliability did not depend on the network sparsity, while Wang et al. [7], and Braun et al. [10], found it depended heavily on network sparsity and other user-selected network parameters. Zalesky et al. [11] suggested small-worldness or scale-freeness measures can change drastically depending on the scale of the parcellation, but few studies have assessed their reproducibility. As so many papers have been published using network measures, their reproducibility deserves further analysis.

We set out to examine the test-retest reliability of graph theory analyses of brain structural connectivity by scanning 17 young adults twice, over a 3-month interval, using high-angular resolution diffusion imaging (HARDI) at 4-Tesla. Other ongoing studies have assessed how connectivity matrices dependent on the scanner field strength, spatial, angular, and *q*-space resolution [12]. Here we assessed the reliability of commonly used network measures over a wide range of network sparsities, as well as inherently more robust measures integrated over different sparsity ranges.

2 Methods

2.1 Subjects

Our analysis included young adults aged 20–30 scanned twice with both MRI and DTI at 4T. Our analysis included a subset of a much larger cohort who was asked to return for a second scan, to assess reproducibility. Of these, some subjects were filtered out due to artifacts in their raw data or errors in tractography, leaving us with 26 subjects. Of these, 2 were statistical outliers on at least one graph theory metric (>3 SD from group mean), 5 had a large difference in the number of fibers tracked in scan 1 and scan 2 (difference of more than 33% in number of fibers in each scan), and 2 had a much larger interval between scan 1 and scan 2. After these subjects were filtered out, we were left with 17 subjects. Subjects were 12 female, 5 male, 100% Caucasian, mean age: 23.6 years, SD 1.47.

2.2 Scan Acquisition

Whole-brain anatomical and high angular resolution diffusion images (HARDI) were collected with a 4T Bruker Medspec MRI scanner. T1-weighted anatomical images were acquired with an inversion recovery rapid gradient echo sequence. Acquisition parameters were: TI/TR/TE = 700/1500/3.35ms; flip angle = 8 degrees; slice thickness = 0.9mm, with a 256x256 acquisition matrix. Diffusion-weighted images (DWI) were also acquired using single-shot echo planar imaging with a twice-refocused spin echo sequence to reduce eddy-current induced distortions. Acquisition parameters were optimized to provide the best signal-to-noise ratio for estimating diffusion tensors [13]. Imaging parameters were: 23cm FOV, TR/TE 6090/91.7ms, with a 128x128 acquisition matrix. Each 3D volume consisted of 55 2-mm thick axial slices with no gap and 1.79x.l.79 mm^2 in-plane resolution. 105 images were acquired per subject: 11 with no diffusion sensitization (i.e., T2-weighted b_0 images) and 94 diffusion-weighted (DW) images (b = 1159 s/mm^2) with gradient directions evenly distributed on the hemisphere. Scan time for the HARDI scan was 14.2 min. The average scan interval was 101 days, SD 18 days.

2.3 Cortical Extraction and HARDI Tractography

Non-brain regions were automatically removed from each T1-weighted MRI scan, and from a T2-weighted image from the DWI set, using the FSL tool "BET" (FMRIB Software Library, http://fsl.fmrib.ox.ac.uk/fsl/). A trained neuroanatomical expert manually edited the T1-weighted scans to refine the brain extraction. All T1-weighted images were linearly aligned using FSL (with 9 DOF) to a common space [14] with 1*mm* isotropic voxels and a 220x220x220 voxel matrix. Raw diffusion-weighted images were corrected for eddy current distortions using the FSL tool "eddy_correct" (http://fsl.fmrib.ox.ac.uk/fsl/). For each subject, the 11 eddy-corrected images with no diffusion sensitization were averaged, linearly aligned and resampled to a downsampled version of their corresponding T1 image (110x110x110, 2x2x2mm). Averaged b_0 maps were elastically registered to the structural scan to compensate for EPI-induced susceptibility artifacts. 35 cortical labels per hemisphere, as listed in the Desikan-Killiany atlas [15], were automatically extracted from all aligned T1-weighted structural MRI scans using Free Surfer (http://surfer.nmr.mgh.harvard.edu/). As a linear registration is performed by the software, the resulting T1-weighted images and cortical models were aligned to the original T1 input image space and down-sampled using nearest neighbor interpolation (to avoid intermixing of labels) to the space of the DWIs. To ensure tracts would intersect labeled cortical boundaries, labels were dilated with an isotropic box kernel of width 5 voxels.

The transformation matrix from the linear alignment of the mean b_0 image to the T1-weighted volume was applied to each of the 94 gradient directions to properly re-orient the orientation distribution functions (ODFs). At each HARDI voxel, ODFs were computed using the normalized and dimensionless ODF estimator, derived for *q*-ball imaging (QBI) in [16]. We performed a recently proposed method for HARDI tractography [17] on the linearly aligned sets of DWI volumes using these ODFs. Tractography was performed using the Hough transform method as in [18]. Elastic deformations obtained from the EPI distortion correction, mapping the average b_o image to the T1-weighted image, were then applied to the tracts' 3D coordinates for accurate alignment of the anatomy. Each subject's dataset contained 5000–10,000 useable fibers (3D curves). For each subject, a full 70x70 connectivity matrix was created. Each element described the proportion of the total number of fibers in the brain that connected a pair of labels; diagonal elements of the matrix describe the total number of fibers passing through a certain cortical region of interest. As these values were calculated as a proportion - they were normalized to the total number of fibers traced for each individual participant, so that results would not be skewed by raw fiber count.

2.4 Graph Theory Analyses

On the 70x70 matrices generated above, we used the Brain Connectivity Toolbox ([18]; https://sites.google.com/a/brain-connectivity-toolbox.net/bct/Home) to compute five standard measures of global brain connectivity - characteristic path length (CPL), mean clustering coefficient (MCC), global efficiency (EGLOB), small-worldness (SW), and modularity (MOD) [18]. CPL is a measure of the average path length in a network; path length is the minimum *number* of edges that must be traversed to get from one node to another; it does not depend on the physical lengths of the fibers, only their network

topology. MCC is a measure of how many neighbors of a given node are also connected to each other, as a proportion of the total number of connections in the network. EGLOB is inversely related to CPL: networks with a small average CPL are generally more efficient than those with large average CPL. SW represents the balance between network differentiation and network integration, calculated as a ratio of local clustering and characteristic path length of a node relative to the same ratio in a randomized network. We created 10 simulated random networks. MOD is the degree to which a system can be subdivided into smaller networks [19]. Figure 1 visualizes these measures in an example network.

One step in binarized graph theory analyses is selecting a sparsity, which may be considered a thresholding operation on the edge strengths (here, fiber counts). The sparsity can alternatively be defined as the fraction of connections retained from the full network, so setting a sparsity level of 0.2 means that only the top 20% of connections (in this case, greatest numbers of fibers) are retained for calculations. The networks reconstructed at a given density will not be identical for any two people, but should be comparable as healthy people have highly similar white matter pathways, especially for the larger tracts. Selecting a single sparsity level may arbitrarily affect the network measures, so we typically compute measures at multiple sparsities, and integrate them across a range to generate more stable scores. We have previously used the range 0.2–0.3 to calculate and integrate these measures, as that range is biologically plausible [20] and more stable [4]. To determine whether the test-retest reliability varied across different sparsities we calculated these measures across the entire range (0–1 in 0.01 increments) as well as integrated across several smaller ranges (0.1–0.2, 0.2–0.3, 0.3–0.4, and 0.4–0.5, in 0.01 increments). We calculated these measures for the whole brain over these different sparsity ranges, and computed the area under the curve of those 11 data points to derive an integrated score for each measure.

2.5 Test-Retest Reliability Analyses

Test-retest reliability was measured by assessing the ICC (intraclass correlation coefficient) between graph theory measures generated from scan 1 matrices and scan 2 matrices. ICC is calculated according to the following formula:

$$ICC=\frac{MS_{Btwn}-MS_{Win}}{MS_{Btwn}+(k-1)MS_{Win}}$$

Here MS stands for mean squared deviation from the mean - within an individual or between individuals, and *k* stands for the number of scans for each subject (here *k=2*).

3 Results

3.1 Global Results

Test-retest reliability results for the full range of sparsities are shown for the 5 global measures in Figure 2. The very sparse measures (0–0.10) have low reliability, perhaps because different sets of nodes are retained between the first and second scans. Modularity (MOD) was most reliable network measure, with an *r*-value (reproducibility) between 0.35–

0.65 for most of the range. Mean clustering coefficient (MCC) has an *r*-value mostly between 0.2–0.6, besides a sharp dip between sparsities 0.30–0.33. Small-worldness (SW) has an *r*-value that greatly fluctuates between 0.1–0.7, with many dips and peaks. Characteristic path length (CPL) and global efficiency (EGLOB) both are rather unreliable until around a sparsity of 0.25, at which point they both have *r*-values mostly between 0.3–0.6; the data depend heavily on the sparsity: note the sharp dip in reliability between sparsities 0.3–0.31.

Next we assessed how reliable the global measures were when scores were integrated across a range of sparsities, to improve stability. The integral of the measures over the range 0.2–0.3 has been shown to be stable [4] and biologically plausible [20], yet we checked the test-retest reliability of scores integrated over 4 different ranges: 0.1–0.2, 0.2–0.3, 0.3–0.4, and 0.4–0.5. Values above 0.5 were not used because it is not considered biologically plausible [20]. Graph theory scores were integrated across these ranges, not the ICC *r*-values. Results for these test-retest reliability analyses are shown in Figure 3. For the 70x70 matrices, 57% of connections had a reliability of at least 0.30. The majority of the most reliable connections (≥0.70) were connections of the frontal cortex.

4 Discussion

In this paper we examined the test-retest reliability of graph theory measures of network connectivity applied to structural networks derived from HARDI scans. Reliability varied both across measures and across sparsities. Modularity was most reliable, and most stable with respect to sparsity threshold. This makes sense given that modularity has to do with how well the network can be broken into sub-networks – a measure of broad network topology, it may depend less on individual connections.

Characteristic path length and global efficiency are almost inverse measures of each other, except that global efficiency takes into account zeros in the connectivity matrix while characteristic path length does not. This could be responsible for the difference in reliability between characteristic path length and global efficiency. At low sparsities with more ‘0’ entries, networks get fractured. Most measures are vulnerable to this loss of nodes, but especially characteristic path length and global efficiency, as the mean shortest path length changes drastically if a significant portion of nodes is deleted. If networks get fractured differently between scan 1 and scan 2, this could lead to the very low reliability of characteristic path length and global efficiency at low sparsities. Characteristic path length and global efficiency are determined by calculating the path length between each node in a network and every other node in the network, for the shortest paths that exist, and averaging over all of those path lengths. Mean clustering coefficient, however, is determined by calculating for all the nodes connected to a given node, how many of its neighbors are also connected to each other, averaged over the whole network. Characteristic path length traces shortest paths, so if one path changes, many paths may take a different course, which could drastically alter mean shortest path length. For mean clustering coefficient, however, one path loss may reduce a node’s clustering coefficient from 5/6 to 2/3 (for example); when averaged over the whole network this may not be a large net change.

Another factor that could be responsible for the difference in reliability between characteristic path length/global efficiency and mean clustering coefficient may be which paths are trimmed at low sparsities. Long-range paths heavily influence characteristic path length/global efficiency, but the mean clustering coefficient depends more on short paths. If long-range paths are generally trimmed before short-range paths, then the reliability of characteristic path length and global efficiency will drop sooner than that of mean clustering coefficient, as sparsity decreases, and their reliability will be impaired. In support of this, the reliability for characteristic path length, global efficiency, and mean clustering coefficient, are all much closer to each other at the highest sparsity, when all connections are retained.

There was a substantial dip in the reliability of a number of measures when integrated over the range 0.3–0.4. This was due to an increase in both the within- and between-subject variability in these measures. The average percent connectedness of these matrices was 26.5%, with all subjects fully connected at a sparsity of 0.30. The range of sparsities where all subjects are beginning to become fully connected may be associated with some instability in the measures, especially if many more unreliable (weak) connections are added.

5 Conclusion

Here we examined the test-retest reliability for a number of graph theory measures commonly used to assess brain structural connectivity. This depends to some extent on the tractography method, as well as the angular and spatial data resolution (we consider these topics elsewhere). Even so, we minimized several sources of error, using a 4-Tesla high angular resolution (94-direction) protocol, and a Hough method that uses ODFs to compute tracts. We found that modularity had moderately high reliability (mean r=0.58 for integrated analyses), as expected for a measure of general network topology. Mean clustering coefficient had higher reliability than characteristic path length or global efficiency for the lower sparsities, perhaps because networks fracture at lower sparsities. Integrating over a range of sparsities improved the reliability of MCC and SW, while decreasing that of CPL and EGLOB at some sparsities. Selecting an appropriate sparsity range to integrate over, and defining network measures robust to sparsity, deserves further analysis.

Acknowledgments

Supported by grants from the NIH and the NHMRC (Australia).

References

1. Gong G, et al. Age- and gender-related differences in the cortical anatomical network. J. Neuroscience. 2009; 29(50):15684–15693.
2. Dennis, E., et al. SFN 2012. Washington, D.C: 2011 Nov 12–16. Heritability of structural brain connectivity network measures in 188 twins.
3. Brown J, et al. Brain local network interconnectivity loss in aging APOE-4 allele carriers. PNAS. 2011; 108(51):20760–20765. [PubMed: 22106308]
4. Dennis E, et al. Altered structural brain connectivity in healthy carriers of the autism risk gene, CNTNAP2. Brain Connectivity. 2012; 1(6):447–459. [PubMed: 22500773]
5. Van den Heuvel M, et al. Aberrant frontal and temporal cortex network structure in schizophrenia: A graph theoretical analysis. J. Neuroscience. 2010; 30(47):15915–15926.

6. Bassett D, et al. Conserved and variable architecture of human white matter connectivity. NeuroImage. 2011; 54(2):1262–1279. [PubMed: 20850551]
7. Wang J-H, et al. Graph theoretical analysis of functional brain networks: Test-retest evaluation on short- and long-term resting-state functional MRI data. PLoS One. 2011; 6(7):1–22.
8. Telesford Q, et al. Reproducibility of graph metrics in fMRI Networks. Front. Neuroinform. 2010; 4:1–10. [PubMed: 20428515]
9. Deuker L, et al. Reproducibility of graph metrics in human brain functional networks. NeuroImage. 2009; 47(4):1460–1468. [PubMed: 19463959]
10. Braun U, et al. Test-retest reliability of resting-state connectivity network characteristics using fMRI and graph theoretical measures. NeuroImage. 2012; 59(2):1404–1412. [PubMed: 21888983]
11. Zalesky A, et al. Whole-brain anatomical networks: Does the choice of nodes matter? NeuroImage. 2010; 50(3):970–983. [PubMed: 20035887]
12. Zhan, L., et al. ISBI 2012. Barcelona, Spain: 2012 May 2–5. How do Spatial and angular resolution affect brain connectivity maps from Diffusion MRI?.
13. Jones D, et al. Optimal strategies for measuring diffusion in anisotropic systems by magnetic resonance imaging. Magn. Res. Medicine. 1999; 42(3):515–525.
14. Holmes C, et al. Enhancement of MR images using registration for signal averaging. JCAT. 1998; 22(2):324–333.
15. Desikan R, et al. An automated labeling system for subdividing the hu-man cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage. 2006; 31:968–980. [PubMed: 16530430]
16. Aganj I, et al. Reconstruction of the orientation distribution function in single-and multiple-shell q-ball imaging within constant solid angle. Magn. Res. Medicine. 2010; 64(2):554–566.
17. Aganj I, et al. A Hough transform global probabilistic approach to mul-tiple-subject diffusion MRI tractography. Med. Image Anal. 2011; 15(4):414–425. [PubMed: 21376655]
18. Rubinov M, Sporns O. Complex network measures of brain connectivity: uses and interpretations. NeuroImage. 2010; 52:1059–1069. [PubMed: 19819337]
19. Bullmore E, Bassett D. Brain Graphs: Graphical Models of the Human Brain Connectome. Annu. Rev. Clin. Psychol. 2010:1–37. [PubMed: 20192801]
20. Sporns, O. Networks of the Brain. Cambridge: MIT Press; 2011.

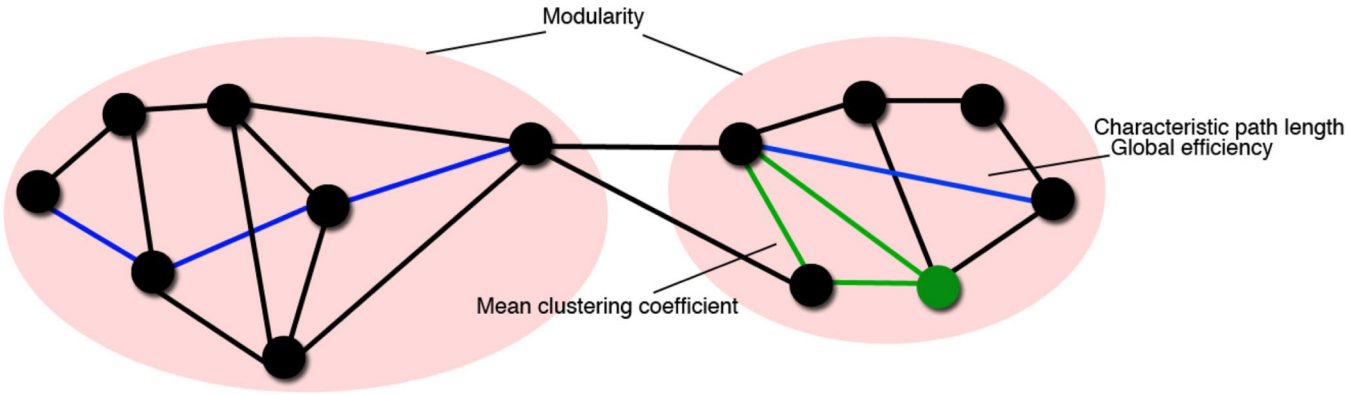

Fig. 1.

Measures **of global connectivity**. Examples show network motifs that serve the basis of each measure. Adapted from the diagram in [18].

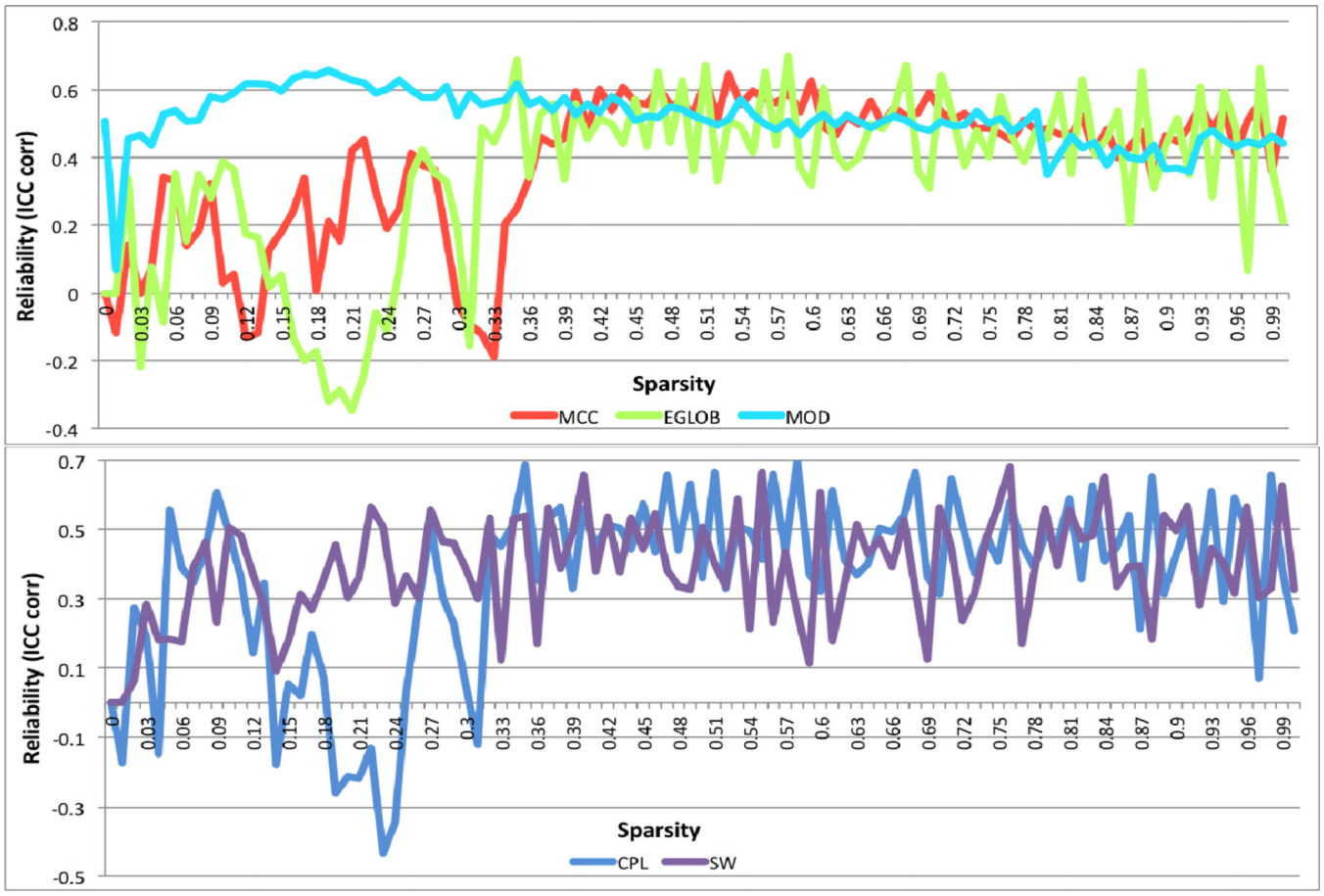

Fig. 2.

Chart **of ICC *r*-values for 5 commonly used network topology measures across the full range of sparsities**. Clearly, the measures, and their reliability, depend on the sparsity threshold: this determines the proportion of connections retained, when sorted by edge strength. Retaining almost all connections (sparsity near 1) may include some that are unreliable, but using a very high threshold (sparsity near zero) may greatly affect which nodes are included in the network at all, promoting instability.

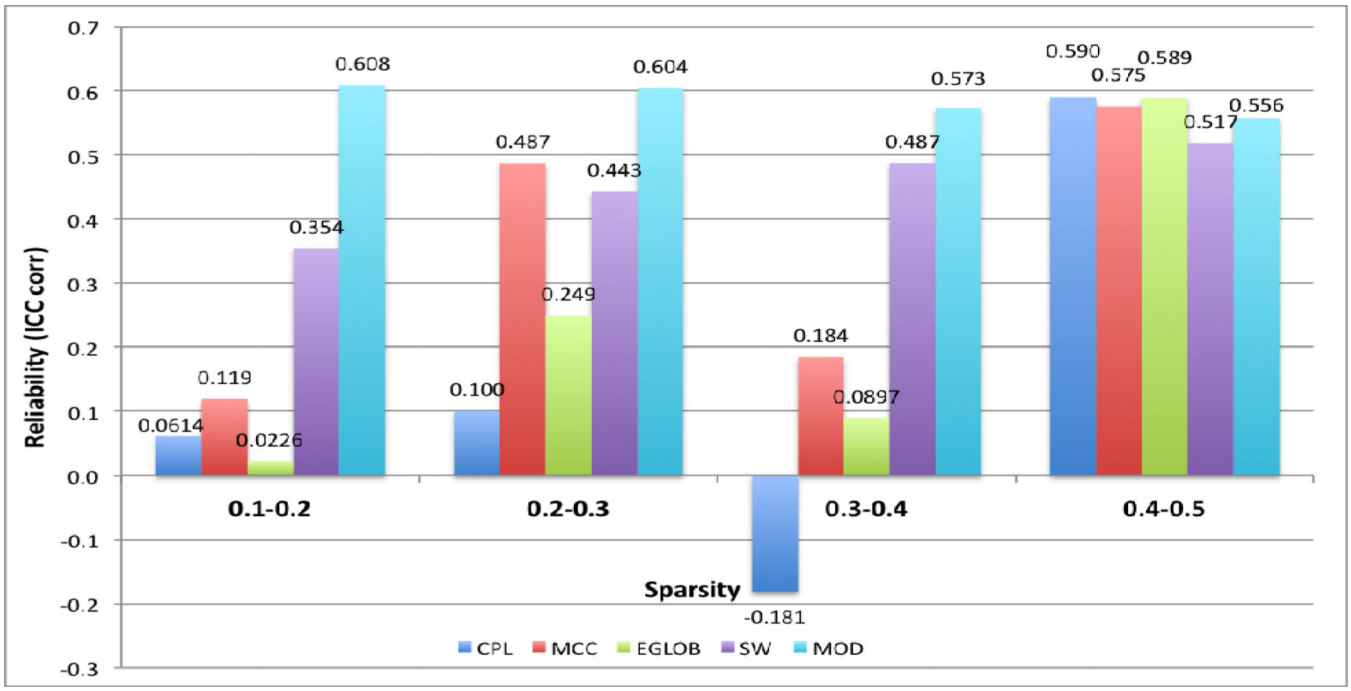

Fig. 3.

ICC *r*-values show the reproducibility of 5 commonly used network measures, when they are integrated, to improve robustness, across 4 different sparsity ranges. MOD is still the most reliable measure. Also MOD is more stable than the other measures with respect to the sparsity, at least over the range examined here.