Reproducibility of graph metrics of human brain functional networks

Lorena Deuker [a,b], Edward T. Bullmore [b,c,*], Marie Smith [d], Soren Christensen [c], Pradeep J. Nathan [b,c], Brigitte Rockstroh [a], Danielle S. Bassett [b,e,f,*]

[a] Department of Psychology, University of Constance, Constance, Germany
[b] Brain Mapping Unit, Department of Psychiatry, Behavioural and Clinical Neurosciences Institute, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK
[c] GSK Clinical Unit Cambridge, Addenbrooke's Hospital, Cambridge, UK
[d] MRC Cognition and Brain Sciences Unit, Cambridge, UK
[e] Genes Cognition and Psychosis Program, Clinical Brain Disorders Branch, National Institute of Mental Health, NIH, Bethesda, MD, USA
[f] Biological Soft Systems Sector, Department of Physics, University of Cambridge, Cambridge, UK

ARTICLE INFO

ABSTRACT

Graph theory provides many metrics of complex network organization that can be applied to analysis of brain networks derived from neuroimaging data. Here we investigated the test–retest reliability of graph metrics of functional networks derived from magnetoencephalography (MEG) data recorded in two sessions from 16 healthy volunteers who were studied at rest and during performance of the n-back working memory task in each session. For each subject's data at each session, we used a wavelet filter to estimate the mutual information (MI) between each pair of MEG sensors in each of the classical frequency intervals from γ to low δ in the overall range 1–60 Hz. Undirected binary graphs were generated by thresholding the MI matrix and 8 global network metrics were estimated: the clustering coefficient, path length, small-worldness, efficiency, cost-efficiency, assortativity, hierarchy, and synchronizability. Reliability of each graph metric was assessed using the intraclass correlation (ICC). Good reliability was demonstrated for most metrics applied to the n-back data (mean ICC = 0.62). Reliability was greater for metrics in lower frequency networks. Higher frequency γ- and β-band networks were less reliable at a global level but demonstrated high reliability of nodal metrics in frontal and parietal regions. Performance of the n-back task was associated with greater reliability than measurements on resting state data. Task practice was also associated with greater reliability. Collectively these results suggest that graph metrics are sufficiently reliable to be considered for future longitudinal studies of functional brain network changes.

Introduction

The recent application of graph theoretical analysis to human brain time series data, e.g., functional MRI, magnetoencephalography (MEG) and electroencephalography (EEG), provides a complex systems approach to the study of functional brain architecture (Bullmore and Sporns, 2009). This whole brain network approach extends and expands upon the current reductionistic understanding of specific regional functions. Graphs of functional connectivity in the human brain can, for example, be derived from fMRI and MEG/EEG time series by estimating the correlation or coherence (or some other measure of association) between voxels/regions of interest (in fMRI) or between sensors (in EEG/MEG) and then thresholding the resulting association matrix to generate a binary adjacency matrix, which can be drawn as a graph.

Several graph theoretical metrics, such as the clustering coefficient, minimum path length or cost-efficiency, have been applied to topological analysis of brain functional networks, and many of them have been shown to reflect disease and state-related differences between groups. For example, longer minimum path length has been reported in patients with Alzheimer's disease (Stam et al., 2007). Brain functional network configuration changes have also been described in relation to performance of simple tasks (Bassett et al., 2006), acute dopamine antagonist drug challenges (Achard and Bullmore, 2007), and normal ageing (Meunier et al., 2009). These preliminary studies suggest that brain functional network parameters might serve as useful biomarkers for neurocognitive disorders and therapeutics.

However, in assessing the potential utility of network measures as markers of brain function in studies designed to test longitudinal changes or drug treatment-related effects, it is important to consider the reliability of the measurements on repeated testing in the same subjects. Unreliable measures will naturally be less attractive as endpoints in a cross-over trial of drug versus placebo, for example, because they will reduce the statistical power of the experiment to detect a true treatment effect. There are prior reasons to consider that

* Corresponding authors. Brain Mapping Unit, Department of Psychiatry, Behavioural and Clinical Neurosciences Institute, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK.
E-mail addresses: etb23@cam.ac.uk (E.T. Bullmore), dp317@cam.ac.uk (D.S. Bassett).

brain functional network metrics might have acceptable reliability. In particular, several studies have shown that topological properties of functional networks are similar to those of underlying structural networks, and to the extent that functional networks are anatomically constrained they are expected to be reliably measured over the course of several weeks.

To provide a first assessment of the test–retest reliability of graph theoretical metrics of human brain functional networks, we used magnetoencephalography (MEG) to record neurophysiological dynamics at rest and during performance of the n-back working memory test in normal volunteers, each studied in two sessions several weeks apart.

For each MEG dataset, we constructed a set of functional networks operating at different frequency intervals and we estimated multiple graph theoretical metrics of the global topological organization of each network. We then quantified the reliability of each global metric, in each frequency interval and under each task condition, in terms of the intraclass correlation (ICC) between measurements in different sessions on the same subjects. In finer-grained analyses, we estimated the reliabilities of graph theoretical metrics of network organization at the level of individual network nodes (MEG sensors). At both global and nodal scales, we tested the hypothesis that a number of experimental factors, including rehearsal of working memory task performance, might have a significant influence on the reliability of graph theoretical metrics of brain functional networks.

Methods

Subjects, MEG data acquisition and cognitive tasks

At the Cognition and Brain Sciences Unit, Cambridge, MEG data were recorded from 16 healthy subjects twice each while in a resting state and while performing a n-back working memory task. The sessions were 4–6 weeks apart. The data were recorded with a 306-channel Vectorview system (Elekta Neuromag, Helsinki) which combines 204 planar gradiometers and 102 magnetometers. Only planar gradiometer data were considered in this study. Data were sampled at 1000 Hz, then downsampled to 250 Hz. A single epoch that lasted for the duration of the resting state or n-back working memory task was used for analysis. The data were not corrected for eye-blinks, muscle artefacts or cardiac artefacts, but a continuous head position monitoring (cHPI) using signal space separation was employed to minimise the effects of adjusting for movement within and between visits (Taulu et al., 2005).

Resting state MEG was recorded for 2.25 min each with the subjects' eyes being open. The participant was told to relax and try to think of nothing in particular. The n-back task is a test of working memory. In the version used here, it requires a continual working memory response from the subject during continual presentation of incoming stimuli (Winterer et al., 2004). Subjects were presented with numbers 1, 2, 3 and 4, displayed in isolation on a screen. Their task was to indicate via four different buttons the integer that was (a) currently displayed (0-back), (b) displayed in the previous run (1-back) or (c) displayed in the run before the previous run (2-back). Stimuli were presented for 500 ms with an inter-stimulus interval of 1800 ms. There were 6 blocks in total, each block consisted of 14 runs of first the 0-back task, then 14 runs of the 1-back task and lastly 14 runs of the 2-back task, thereby increasing difficulty within each block. The task lasted for approximately 9 min.

MEG data processing

For each individual MEG dataset, the sensor time series were analyzed with a Daubechies 4 discrete wavelet transform (Percival and Walden, 2000). Scales 2 through 7 corresponded roughly to

standard MEG frequency bands γ (31.2–62.5 Hz), β (15.6–31.2 Hz), α (7.8–15.6 Hz), θ (3.4–7.8 Hz), δ (1.7–3.4 Hz) and δ^{-1} (0.8–1.7 Hz). As the wavelet transform requires the number of time points of the time series input to be a power of two, 2^{15} data points were considered for resting state data (corresponding to 2.25 min) and 2^{17} were considered for the n-back data (approximately 9 min).

Frequency dependent functional connectivity was estimated by the mutual information (MI) between wavelet coefficients at each scale for each pair of sensors. This resulted in a $\{204 \times 204\}$ association matrix for each MEG dataset at each wavelet scale. We also estimated the mean MI, over all entries in the association matrix, as a simple measure of the average strength of functional connectivity in each dataset. Mutual information has previously been shown to provide superior sensitivity for estimation of functional connectivity in band-pass filtered time series (David et al., 2004).

For graph theoretical analysis, the association matrix was thresholded to create a binary adjacency matrix where the $\{i, j\}$th element was either 1 (if the MI between sensors i and j was greater than the threshold) or 0 (if it was not). The adjacency matrix can be visualized as an undirected graph where an edge or connection is drawn between each pair of nodes that has mutual information greater than threshold. Threshold values were chosen so that the total number of non-zero entries in the adjacency matrix, also known as the cost or connection density of the graph K, was at the lowest value consistent with all nodes being connected in all networks at each wavelet scale. Thus all networks within the same frequency band and task condition had the same cost or number of edges. We decided against the use of a single common threshold for all frequency bands and task conditions because it would either have led to some networks being disconnected (if the cost was chosen to be very small), which would have made the estimation of some graph metrics impossible, or it would have led to over-densely connected networks (if the cost was chosen so that all networks would be fully connected). Fig. 1 provides a schematic overview of the procedure.

For each network the following graph metrics were calculated: clustering, path length, small-worldness, efficiency, cost-efficiency, synchronizability, assortativity, and hierarchy (see Table 1 for an overview). These metrics can be divided into first-order metrics, which depend on only one graph property, and second-order metrics, which depend on more than one property. Thus clustering coefficient, minimum path length, global efficiency, cost-efficiency and synchronizability were classified as first-order metrics, while assortativity, hierarchy, and small-worldness were classified as second-order. Each of these metrics is briefly defined in more detail below:

Clustering, path length and small-world topology

The clustering coefficient of node v is defined as the ratio of the connected triangles, δ_v to the connected triples τ_v, and therefore the clustering coefficient of a graph, G, can be defined as:

$$C(G) = \frac{1}{|V'|} \sum_{v \in V'} \frac{\delta_v}{\tau_v} \tag{1}$$

where V' is the set of nodes with degree greater than 2 (Schank and Wagner, 2005). In a study that compared healthy individuals with brain tumour patients (Bartolomei et al., 2006), a reduced clustering coefficient was found in functional networks that were derived from MEG data. This provides a clear demonstration of how clustering coefficient can in principle be applied to analysis of human functional neuroimaging data, even though a slightly different estimator of the clustering coefficient than the one described above was used.
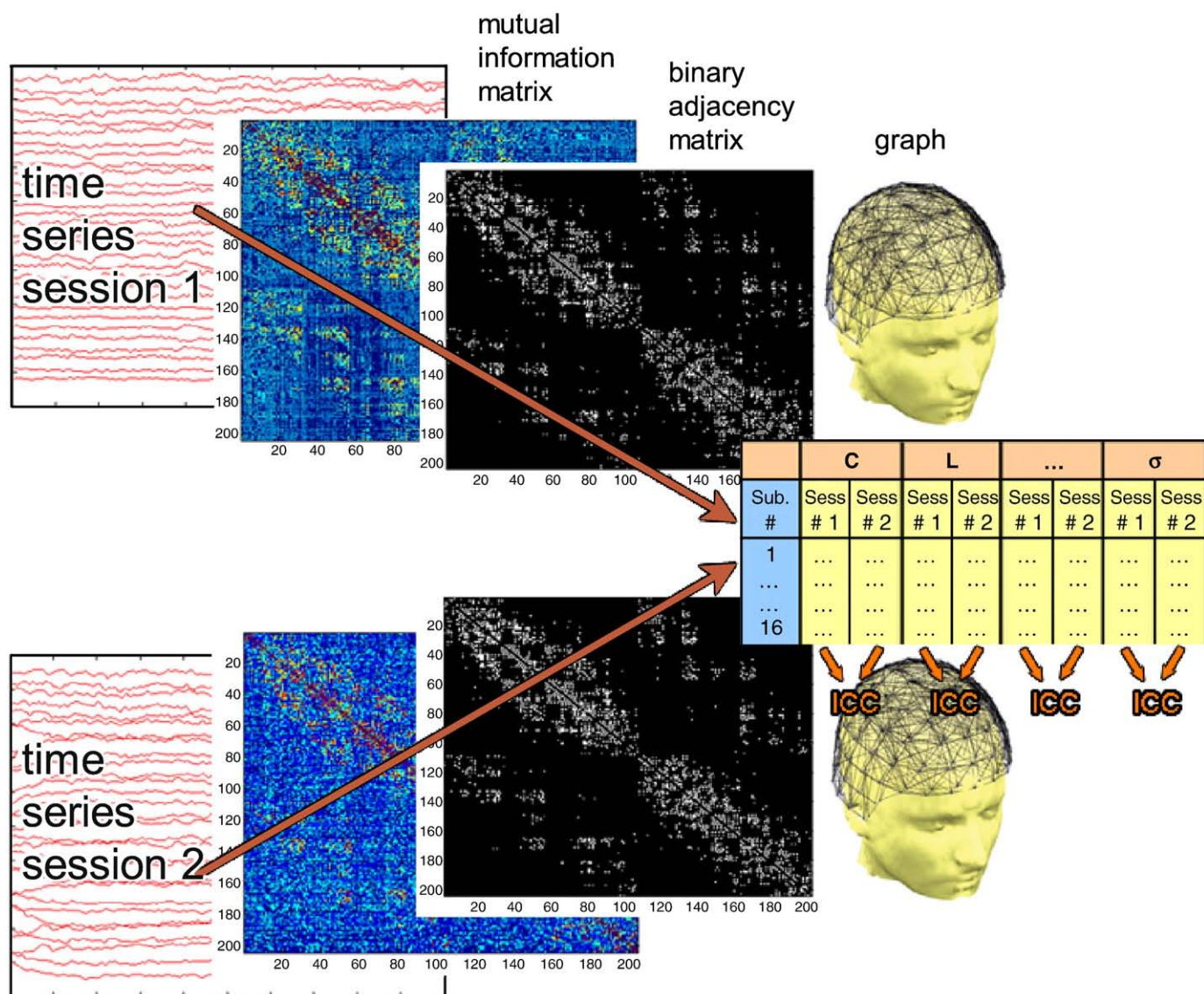
Fig. 1. Time series data from two MEG sessions were used to determine session-specific mutual information matrices, to which a binary threshold was then applied. Based on the binary matrices, graph metrics were calculated for each session. To determine reliability, an intraclass correlation between sessions was calculated for each metric.

Minimum path length L was defined as the average of the shortest paths between each node and every other node in the network (Watts and Strogatz, 1998).

Table 1
Description of graph metrics.

Metric	Description	Reference
Clustering coefficient	Cliquishness of a graph	Watts and Strogatz (1998)
Minimum path length	Mean of shortest paths between all nodes	Watts and Strogatz (1998)
Small-world	High clustering in the presence of small path length	Watts and Strogatz (1998)
Synchronizability	Structural property that enables network to synchronize	Motter et al. (2005); Barahona and Pecora (2002)
Assortativity	Degree correlation between neighbouring nodes	Newman (2002)
Hierarchy	Power law relationship between clustering and degree of the nodes in the network	Ravasz and Barabasi (2003)
Global efficiency	The inverse of path length	Latora and Marchiori (2001)
Cost efficiency	Global efficiency at a given cost minus the cost	Achard and Bullmore (2007)

Overview of the 8 different graph metrics that were used to analyse the networks of functional connectivity.

Small-world topology refers to a network structure with higher clustering coefficient C than a comparable random network, but equally short path length L (Watts and Strogatz, 1998). Using C and L, the small-world value σ could then be computed as

$$\sigma = \frac{C / C_{Ran}}{L / L_{Ran}} \qquad (2)$$

(Humphries et al., 2006), where C_{Ran} and L_{Ran} denote the clustering coefficient and path length, respectively, estimated in comparable random networks. This ratio will be greater than unity for a small-world network. In this study, one representative random network was created for each frequency band in order to compute the small-world value σ. This random network did not preserve the degree distribution of the original networks, but had the same number of edges.

Interestingly, in networks derived from a working memory task, the small-world coefficient has been found to be higher in healthy subjects when compared to schizophrenic patients (Micheloyannis et al., 2006).

Hierarchy, assortativity and synchronizability

A fundamental property of any node is the number of edges connecting it to the rest of the network, also known as the degree k of

each node. In hierarchical networks, it has been shown that nodes with high degree tend to have a small clustering coefficient C, while nodes with smaller degree tend to have higher clustering (Ravasz and Barabasi, 2003). Accordingly, the hierarchical structure of a graph can be quantified in terms of the power law relationship between clustering C and degree k of the nodes in the network:

$$C \sim k^{-\beta}. \tag{3}$$

The hierarchy coefficient β was estimated by fitting a linear regression line to the plot of $\log(C)$ versus $\log(k)$.

The assortativity of a network refers to the degree to which nodes are linked to nodes with a similar degree. Assortativity can thus be defined as the correlation r between the degree of a node and the mean degree of its immediate neighbours (Newman, 2002). Positive values of r indicate assortative networks, while negative values indicate disassortative networks. Technical and biological networks are typically disassortative while social networks tend to be assortative. For example, networks of co-authorship in scientific journals were found to be assortative (Newman, 2004).

Synchronizability, S, refers to structural properties of the network that enable it to synchronize rapidly and is defined as

$$S = \frac{\lambda_2}{\lambda_N} \tag{4}$$

where λ_2 is the second smallest eigenvalue of the Laplacian L of the adjacency matrix, and λ_N is the largest eigenvalue of L (Motter et al., 2005; Barahona and Pecora, 2002). Fully synchronized systems have been found to have S between 0.01 and 0.2, suggesting a threshold value of 0.01 for global synchronization.

Efficiency and cost-efficiency

The efficiency of a node is defined as the inverse of the path length L of that node (Latora and Marchiori, 2001); thus a node with high efficiency will have short minimum path length to all other nodes in the graph. Global efficiency E_{glob} is defined as the mean of the efficiency of all nodes and global cost-efficiency is then simply the global efficiency at a given cost minus the cost, i.e., $(E-K)$, which will typically have a maximum value $\max(E-K)$ greater than zero, at some

cost K_{max} for an economical small-world network (Achard and Bullmore, 2007).

Statistical analysis

The intraclass correlation (ICC) was estimated as a measure of test–retest reliability for average MI and for each graph metric under each task condition and at each frequency scale. The ICC is close to $+1$ if the measurements made in the two MEG recording sessions are consistent on repeated testing for each subject in the sample. We reported ICCs descriptively and also used analysis of variance (ANOVA) models to test hypotheses about effects of experimental and other factors on reliability of the various global network metrics. Additionally, we estimated ICCs at each sensor for local or nodal measures of network organization, mapped these local reliability estimates on renderings of the scalp surface, and used multiple ANOVA testing to identify significant factorial effects on reliability of nodal network metrics.

All statistical comparisons were implemented in Statistica (StatSoft Inc., http://www.statsoft.com/) and all other computations were performed in Matlab (MathWorks Inc., http://www.mathworks.com/).

The scalp plotting program used in this work was adapted for the current use from Delorme, A. (2002) Headplot Matlab Code (CNL/Salk Institute, La Jolla, CA).

Results

Reliability of global metrics on working memory networks

The results of the reliability analysis for global metrics on networks derived from MEG data recorded during the n-back working task are presented in Fig. 2 and Table 2. The intraclass correlations for the 8 different graph metrics, plus average mutual information, ranged from 0.02 to 0.89, with a mean of 0.62 ± 0.20(SD).

Reliability varied considerably over different frequency bands. The lowest mean ICC over all metrics was found in the γ band (0.40 ± 0.10 (SD)), and the highest was found in the α band (0.75 ± 0.07(SD)). Accordingly, an ANOVA model that treated the different frequency bands as repeated measures showed a highly significant effect of frequency band ($F(5,48) = 8.04$, $P = 0.00003$) which is reflected in a
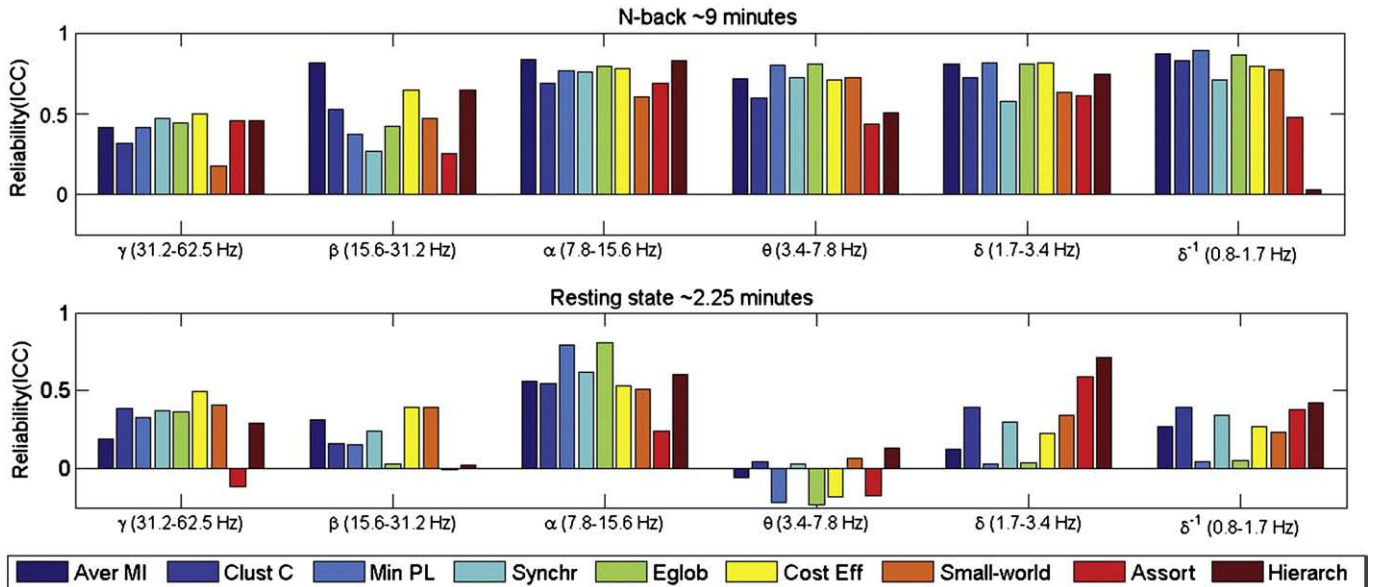

Fig. 2. Intraclass correlation (ICC) coefficients for 9 different metrics over 6 frequency bands. Average mutual information as a non-graph metric is presented first, followed by the first-order metrics and then second-order metrics. Top: During performance of a n-back task, ~9 min. Bottom: For resting state data, ~2.25 min.

Table 2
ICC values for all frequency bands.

	γ	β	α	θ	δ	δ^{-1}
N-back						
Average mutual information (MI)	0.416	0.817*	0.835**	0.714**	0.805*	0.868**
Clustering coefficient (C)	0.315	0.529*	0.686*	0.594*	0.72	0.828**
Minimum path length (L)	0.414	0.37	0.763**	0.8**	0.816**	0.891**
Synchronizability (S)	0.467	0.267	0.755**	0.725*	0.577*	0.708*
Global efficiency (E_{glob})	0.439*	0.419	0.795**	0.804**	0.807**	0.866**
Cost efficiency (CE)	0.501*	0.646*	0.779**	0.708*	0.814**	0.793**
Small-world (σ)	0.173	0.467*	0.6*	0.721*	0.63*	0.769**
Assortativity (r)	0.454*	0.254	0.687*	0.433*	0.609*	0.475*
Hierarchy (β)	0.455*	0.647*	0.827*	0.508*	0.743**	0.027
Resting state						
Average mutual information (MI)	0.187	0.313	0.558	−0.056	0.121	0.267
Clustering coefficient (C)	0.389	0.158	0.546	0.041	0.395	0.391
Minimum path length (L)	0.325	0.15	0.791**	−0.215	0.031	0.041
Synchronizability (S)	0.374	0.239	0.619*	0.029	0.298	0.339
Global efficiency (E_{glob})	0.363	0.03	0.809**	−0.231	0.035	0.049
Cost efficiency (CE)	0.491	0.391	0.531	−0.18	0.229	0.272
Small-world (σ)	0.409	0.395	0.509	0.068	0.342	0.235
Assortativity (r)	−0.117	−0.005	0.239	−0.171	0.59*	0.38
Hierarchy (β)	0.289	0.021	0.603*	0.133	0.712**	0.421

The exact ICC values for all metrics and frequency bands during the 9 min of the n-back working memory task and the 2.25 min of resting state. * designates significant values ($P = 0.05$, FDR corrected) and ** marks ICC values that additionally pass the stricter criterion of being above 0.7 and having a coefficient of variation below 0.2.

general trend of increasing reliability for lower frequency bands, as shown in Fig. 2.

Reliability also varied somewhat between different metrics. We found this most clearly when we compared the reliabilities of average MI, first-order graph metrics (clustering, path length, efficiency, cost-efficiency and synchronizability), and second-order graph metrics (assortativity, hierarchy and small-worldness). Again using a repeated measures ANOVA model, we found that there was a significant effect of metric order ($F(2,6) = 9.91$, $P = 0.0126$). Posthoc comparisons showed that the reliabilities of the mutual information and of the first-order metrics were not significantly different from each other ($t(4) = 1.68$, $P = 0.14$), but both first-order graph metrics ($t(6) = 3.57$, $P = 0.01$) and average MI ($t(2) = 3.85$, $P = 0.008$) were significantly more reliable than the second-order metrics.

All the graph metrics were initially estimated in relatively sparse networks with low connection density: networks in frequency bands α, β, γ, and θ had 6–7% of all possible connections (with the exception of γ in the resting state data which had 13%). In δ and δ^{-1} networks, connection densities were higher, between 20% and 42%. The exact values can be found in Table 3. We subsequently explored the effects of connection density or cost on reliability by estimating all metrics in networks with connection densities 5% less and 5% greater than their initial values. We found no significant effect of connection density on reliability of graph metrics ($F(2,12) = 1.44$, $P = 0.28$); see also Fig. 3B.

In addition to these assessments of between-session reliability, we also evaluated the within-session reliability of all metrics. The intraclass correlations for two non-overlapping 4.5 minute segments from within a single session were generally quite high with a mean ICC of 0.77 ± 0.25(SD) over all frequency bands and metrics. The equivalent estimates of between-session reliability, comparing the metrics in the first 4.5 min of n-back data in each of the two experimental sessions, yielded a mean ICC of 0.54 ± 0.22(SD). A two-factor repeated measures ANOVA with frequency band as a repeated

measure and within versus between sessions as a group factor showed that the within-session ICCs were significantly higher than the between-session reliabilities ($F(1, 16) = 21.57$, $P = 0.0017$).

Effects of working memory task practice and performance

We were also interested to explore how reliability of network organization might be affected by practice on the n-back working memory task. To address this question, we estimated between-session intraclass correlations for each of three separate 2.25 min segments of the n-back experimental data: an early segment from 0 to 2.25 min, an intermediate segment from 3 to 5.25 min, and a late segment from 6 to 8.25 min. We found that reliability monotonically increased as a function of increasing practice on the task: the average ICC over all metrics was 0.50 ± 0.24(SD) for the early segment, 0.59 ± 0.22(SD) for the intermediate segment, and 0.63 ± 0.23(SD) for the late segment. A two-factor repeated measures ANOVA with task practice and frequency band as repeated measures showed a significant main effect of task practice ($F(2, 12) = 18.57$, $P = 0.00007$) and a significant main effect of frequency band ($F(5, 30) = 16.83$, $P < 0.00001$). Posthoc comparisons showed that the reliability was higher in the second practice session than the first ($F(1, 6) = 12.35$, $P = 0.0126$); the reliability was higher in the third practice session than the second ($F(1, 6) = 26.29$, $P = 0.0021$) and the first ($F(1, 6) = 22.13$, $P = 0.0033$), indicating a linear increase in reliability with practice.

There was also a significant interaction between frequency band and task practice ($F(10, 80) = 7.25$, $P = 0.00000$) indicating that task practice was not associated with increasing reliability in all frequency bands. As shown in Fig. 3A, this interaction is driven by the reduced reliability of network metrics in the β frequency interval in the last segment of the data. These practice-related changes in reliability of network measures were not reflected by changes in task accuracy: ANOVA modelling of task accuracy for the three segments with session and task practice as repeated measures revealed no significant main effects of session ($F(1, 14) = 3.46$, $P = 0.08$) or task practice ($F(2, 13) = 0.32$, $P = 0.73$) and no interaction of session and task practice ($F(2, 13) = 0.11$, $P = 0.89$), expressing clearly that accuracy was similar in both sessions and during different segments of the task. Similarly, ANOVA modelling of reaction time yielded no significant effect of task practice ($F(2,13) = 2.25$, $P = 0.15$), but revealed a significant effect of session ($F(2, 14) = 4.58$, $P = 0.05$) as reaction time in the second session was significantly smaller.

Additionally, we compared the reliability of global network metrics for resting state data to an equivalent period of the n-back working memory data. As can be seen clearly in Fig. 2, the reliability of the resting state data was generally lower for all frequency bands (mean ICC of 0.26 ± 0.25(SD)) except in the α band (mean ICC for $\alpha = 0.58 \pm 17$(SD)).

We used an ANOVA model to compare the ICCs for network metrics in resting networks to those estimated for the same metrics in the first 2.25 min of the n-back task. This comparison revealed a highly significant effect of task ($F(1,16) = 29.33$, $P = 0.000063$), due to the

Table 3
Connection densities.

Groups	γ	β	α	θ	δ	δ^{-1}
Rest~2.25 min	0.13	0.06	0.06	0.07	0.32	0.43
N-back~9 min	0.06	0.06	0.07	0.06	0.21	0.4
N-back~0–2.25 min	0.06	0.06	0.06	0.06	0.21	0.38
N-back~3–5.25 min	0.06	0.06	0.06	0.06	0.26	0.49
N-back~6–8.25 min	0.06	0.07	0.07	0.06	0.24	0.36
N-back within	0.06	0.06	0.06	0.06	0.22	0.42

The connection densities – or cost values – were used for thresholding for the different frequency bands and groups. The values are the smallest cost at which all graphs for participants and sessions would be fully connected. These cost values were set to at least 0.06 so that random networks used for the estimation of small-world would be fully connected.
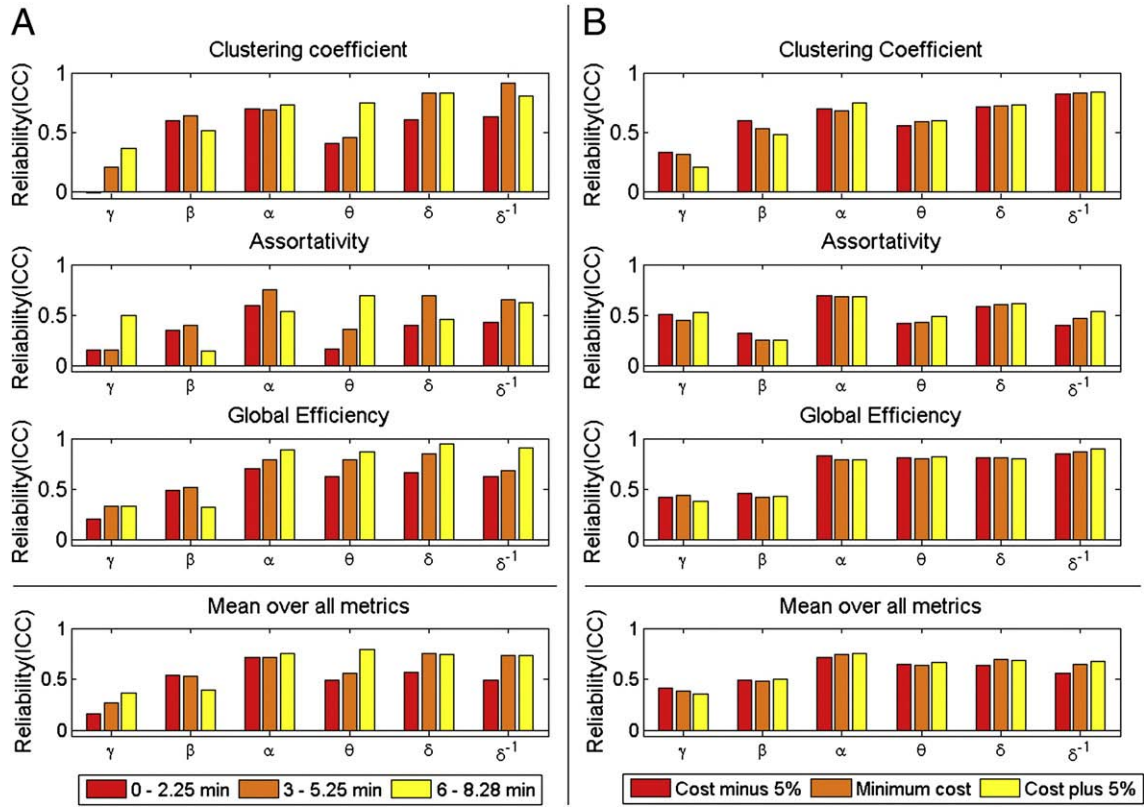
Fig. 3. Intraclass correlation (ICC) values for clustering coefficient, assortativity, global efficiency and the mean over all metrics over the six frequency bands considered. (A) For varying task practice in the n-back task; analysis shows a significant effect of task practice. (B) For cost values that were plus and minus 5% of the original cost value used; no significant effect of cost was found.
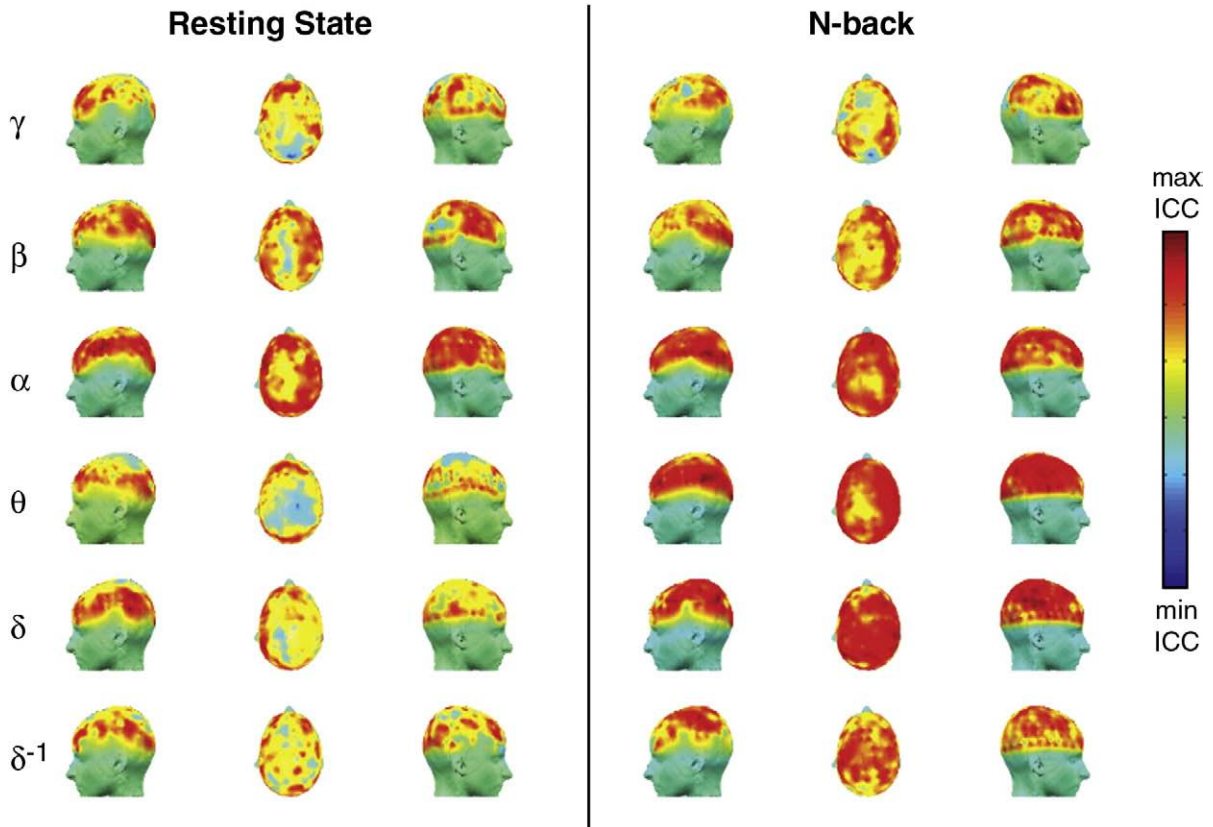

Fig. 4. Reliability (ICC) of network efficiency on a nodal (sensor) level, with sensors located on a scalp surface rendering. Results for the six different frequency bands are given from top to bottom, and categorized according to task into either resting state (left) or n-back (right).
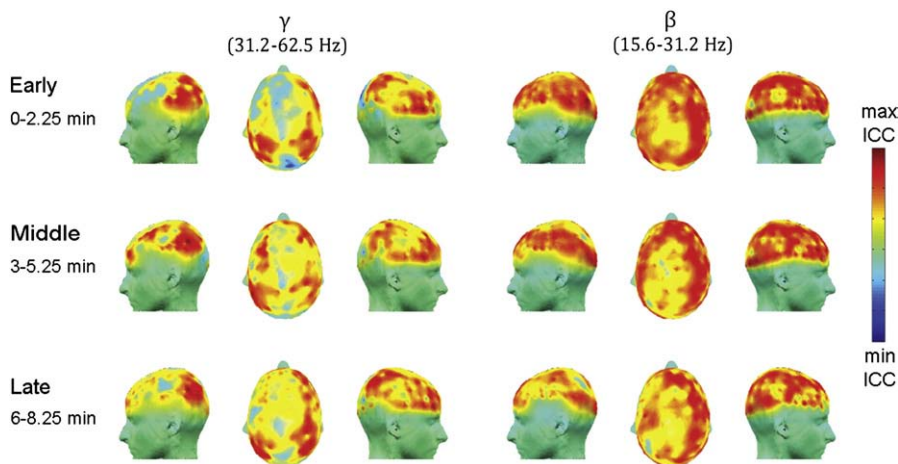
Fig. 5. Reliability of nodal efficiency changes with task practice. For early, intermediate and late segments of the n-back task, ICC values for efficiency at a nodal (sensor) level were mapped to the sensor locations on a scalp surface rendering, illustrating the effect of task practice. Results for γ-band are shown on the left side and for β-band on the right side.

greater reliability of the metrics in the n-back data. There was also a significant effect of frequency band ($F(5, 48) = 18.39$, $P < 0.00001$) and a significant interaction between task and frequency ($F(5, 40) = 8.38$, $P = 0.00001$).

These results reflect the observation (Fig. 2) that the reliability of the α-band network is distinctively greater than the reliabilities of all other frequency band networks in the resting state data; whereas, in the n-back data, the reliability of the α-band network is greater than that of higher frequency networks but approximately equivalent to that of lower frequency networks.

Reliability of nodal network metrics

Finally, we evaluated the reliability of selected network metrics at the level of individual network nodes (MEG sensors) rather than at the global level so far considered. For example, we estimated the between-session intraclass correlations for nodal efficiency estimated at each sensor in both n-back and resting state data and mapped these reliability estimates onto a surface rendering of the scalp (Fig. 4). This analysis demonstrated that reliability was not homogeneously distributed across all sensors. In the γ- and β-band networks, for which reliability of global network metrics was less than in lower frequency networks, we found that some sensors had highly reliable efficiency in both experimental conditions. Moreover, when we investigated the effects of n-back task practice on reliability of nodal efficiency, we found that the subset of highly reliable nodes became concentrated in right frontal and parietal regions of the scalp surface as a function of increasing practice on the task (Fig. 5). Collectively these observations indicate that reliability of local or nodal network metrics can be reasonably high even when reliability of global metrics is low; and that changes in experimental condition, including greater practice on the same experimental task, can be associated with differences in the spatial distribution of highly reliable nodal network properties.

Discussion

The primary objective of this study was to investigate the reliability of graph theoretical measures of human brain functional networks derived from experimental MEG data. Overall, we found that the reliability of graph metrics was reasonably good (mean ICC = 0.62), in comparison to the limited prior data available on reliability of other electrophysiological measurements. For example, EEG spectral parameters measured during a working memory task were found to be relatively stable over a period of up to 40 months, reflected in the fact that 35 out of 40 first and second sessions were

correctly matched to each other (Naepflin et al., 2008). Also Huang et al. (Huang et al., 2007) have examined the abnormal metabolic brain networks associated with Parkinson's disease in two PET sessions over 8 weeks and found them to be highly stable with an intraclass correlation coefficient of ICC = 0.89. However, we were also able to identify significant effects on reliability of a number of factors – including frequency band, type of metric and task conditions – which might help guide investigators on the most reliable applications of these metrics in future studies.

First, we noted that global brain network metrics tended to be more reliable at lower frequency intervals. This is perhaps not surprising in light of prior literature, (e.g. Honey et al., 2007), indicating that lower frequency coherent oscillations in neuroanatomically realistic computational models are more strongly dictated by structural constraints. More anatomically constrained systems would naturally be more reliable over the course of repeated measurements only a few weeks apart. In addition to the relationship between lower frequency functional connectivity and anatomical connections which may drive replicability, there is evidence for a basic replicability of resting state functional networks over a wide range of frequency bands based on high heritability (Smit et al., 2008). Indeed, univariate time series properties, such as the power spectrum, have also previously been shown to be highly heritable (van Beijsterveldt and van Baal, 2002) and this may be regarded as further evidence for a likely genetic effect on the high reliability of functional network metrics.

The higher frequency γ and β-band networks were less reliable at a global level. This is arguably consistent with prior work suggesting that higher frequency systems are rapidly reconfigurable in support of cognitive representations and perceptions. Dynamically nonstationary networks would be expected to have lower reliability on test–retest measurement over several weeks.

From a cognitive neuroscience perspective, it was notable that brain functional network metrics were generally more reliable when measured under experimentally controlled task conditions – the n-back working memory paradigm – than when measured in a relatively uncontrolled "resting" state. This is consistent with results reported in a recent study looking at the test–retest reliability of frequency components in EEG over an interval of 7 days (McEvoy et al., 2000). Reliability of these components was also found to be higher in a working memory task than in resting state, suggesting that the external direction of brain function necessitates specific functional configurations with little variance. The stability of event-related desynchronization and synchronization (ERD/ERS) for different frequency bands has also been investigated in 29 subjects over the course of two years (Neuper et al., 2005). These authors found

satisfactory stability (ICC>0.7) when subjects were performing a task; in the resting state, the α range was the most stable frequency band. These studies support our findings that reliability is frequency dependent and that the reliability of the α-band network is least affected by task.

The comparably low reliability in the resting state might be a consequence of the diverse nature of neural patterns found in resting state, as reported by Damoiseaux et al. in a test–retest fMRI study (Damoiseaux et al., 2006): 10 different resting state patterns, arguably related to different functions (e.g. memory) that may occur during resting state, were identified across subjects with a tensor probabilistic independent component analysis. Even though these patterns were consistent across subjects and sessions (9 of 10 patterns were found in both 40 minute sessions), this does not imply good reliability in our case, as different "patterns" may be active at different times, thereby reducing reliability, especially in the short 2.25 min recording used here. In planning a study where between-session reliability is important, it therefore seems sensible to use a task in the experimental paradigm to assure that subjects retain a similar time-dependent cognitive state in both sessions. It should be noted, however, that the resting state data used here was acquired while participants had their eyes open. It is probable that networks derived from resting state data with eyes closed may prove to be more reliable, since the signal-to-noise ratio of eyes-open data can be adversely impacted by alpha suppression and eye blink artefacts.

Moreover, we found that task practice was associated with greater reliability of global network metrics over time. As subjects rehearsed the task within each session their pattern of network configuration became more consistent over different recording sessions. This suggests that learning reliable behavior is related to emergence of more reliable brain network configurations: task training drives functional network selection to an asymptotic limit that is relatively stable over time — perhaps anatomically constrained. Task practice improved global network reliability in all frequency intervals except β-band. However, when we considered the task practice-related changes in the β-band network at a finer-grained nodal level of analysis, we found the emergence of a subset of highly reliable network nodes in frontal and parietal regions, apparently by a process of elimination on a larger initial set of reliable nodes. This could mean that during the beginning of the task, more brain areas are involved in task performance (and are therefore reliably reproducible) whereas later on, only a core set of brain regions is still concerned with the task, leaving the other brain areas to random fluctuation and poor reproducibility. There is evidence from functional imaging studies that with increasing task practice, activation decreases in some initially highly activated brain regions. For example, Raichle et al. found in a PET study that areas most active during naive performance were significantly less active during practiced performance in a verbal response selection task (Raichle et al., 1994). Practice produced decreases in activation were also found in a visuospatial working memory task (Garavan et al., 2000). This study also showed that fewer regions passed statistical criteria for activation at the end of practice as compared to the beginning of practice. These results are consistent with the idea that during the course of performing a task, underlying networks may become more focused. Moreover, the location of these more focused, reliable regions in frontal and parietal cortex matches regions that are found to be activated in the n-back task (Owen et al., 2005).

On a more methodological note, we found that there were differences in reliability between metrics. Generally the simpler metrics – such as average MI or first-order graph metrics (like path length) – were more reliable than second-order graph metrics (like small-worldness). It is not clear whether the worse reliability of second-order metrics is caused by added metric variance or increased sensitivity of these metrics to, for example, performance related changes in brain function. Our finding that the non-thresholded

measure of association or functional connectivity measure – average mutual information – did not prove to be more reliable than first-order metrics suggests that information important for reliability is retained in suprathreshold connections in an undirected graph. An alternative to thresholding is to use the full information of a coherence matrix in a weighted graph, as has been done, for example, in a study by Rubinov et al. (2009). It would be very interesting to investigate the reliability of graph metrics in weighted graphs, which unfortunately was outside the scope of this paper.

In conclusion, the results of this study support the potential use of graph metrics of brain functional network organization in longitudinal designs, e.g., cross-over studies where the effect of acute drug administration on network configuration is compared to the effects of placebo by repeated measurements on the same subjects. In such situations, we recommend that greater statistical power might be conferred by lower longitudinal variability of first-order metrics of global network properties at lower frequencies and under experimentally controlled or well-rehearsed task conditions. If it is hypothetically important to make longitudinal measurements on higher frequency networks then our results suggest that it will be more powerful to consider regional or nodal metrics rather than global network measures.

References

Achard, S., Bullmore, E., 2007. Efficiency and cost of economical brain functional networks. PLoS Comput. Biol. 3 (2), e17.

Barahona, M., Pecora, L.M., 2002. Synchronization in small-world systems. Phys. Rev. Lett. 89 (5), 054101.

Bartolomei, F., Bosma, I., Klein, M., Baayen, J.C., Reijneveld, J.C., Postma, T.J., Heimans, J.J., van Dijk, B.W., de Munck, J.C., de Jongh, A., Cover, K.S., Stam, C.J., 2006. Disturbed functional connectivity in brain tumour patients: evaluation by graph analysis of synchronization matrices. Clin. Neurophysiol. 117 (9), 2039–2049.

Bassett, D.S., Meyer-Lindenberg, A., Achard, S., Duke, T., Bullmore, E., 2006. Adaptive reconfiguration of fractal small-world human brain functional networks. Proc. Natl. Acad. Sci. U. S. A. 103 (51), 19518–19523.

Bullmore, E., Sporns, O., 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. Nat. Rev. Neurosci. 10 (3), 186–198.

Damoiseaux, J.S., Rombouts, S.A.R.B., Barkhof, F., Scheltens, P., Stam, C.J., Smith, S.M., Beckmann, C.F., 2006. Consistent resting-state networks across healthy subjects. Proc. Natl. Acad. Sci. U. S. A. 103 (37), 13848–13853.

David, O., Cosmelli, D., Friston, K.J., 2004. Evaluation of different measures of functional connectivity using a neural mass model. Neuroimage 21 (2), 659–673.

Garavan, H., Kelley, D., Rosen, A., Rao, S.M., Stein, E.A., 2000. Practice-related functional activation changes in a working memory task. Microsc. Res. Tech. 51 (1), 54–63.

Honey, C.J., Kotter, R., Breakspear, M., Sporns, O., 2007. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. Proc. Natl. Acad. Sci. U. S. A. 104 (24), 10240–10245.

Huang, C., Mattis, P., Tang, C., Perrine, K., Carbon, M., Eidelberg, D., 2007. Metabolic brain networks associated with cognitive function in Parkinson's disease. Neuroimage 34 (2), 714–723.

Humphries, M.D., Gurney, K., Prescott, T.J., 2006. The brainstem reticular formation is a small-world, not scale-free, network. Proc. R. Soc. B. 273 (1585), 503–511.

Latora, V., Marchiori, M., 2001. Efficient behavior of small-world networks. Phys. Rev. Lett. 87 (19), 198701.

McEvoy, L.K., Smith, M.E., Gevins, A., 2000. Test–retest reliability of cognitive EEG. Clin. Neurophysiol. 111 (3), 457–463.

Meunier, D., Achard, S., Morcom, A., Bullmore, E., 2009. Age-related changes in modular organization of human brain functional networks. Neuroimage 44, 715–723.

Micheloyannis, S., Pachou, E., Stam, C.J., Breakspear, M., Bitsios, P., Vourkas, M., Erimaki, S., Zervakis, M., 2006. Small-world networks and disturbed functional connectivity in schizophrenia. Schizophr. Res. 87 (1–3), 60–66.

Motter, A.E., Zhou, C.S., Kurths, J., 2005. Enhancing complex-network synchronization. Europhys. Lett. 69 (3), 334–340.

Naepflin, M., Wildi, M., Sarnthein, J., 2008. Test–retest reliability of EEG spectra during a working memory task. Neuroimage 43 (4), 687–693.

Neuper, C., Grabner, R.H., Fink, A., Neubauer, A.C., 2005. Long-term stability and consistency of EEG event-related (de-)synchronization across different cognitive tasks. Clin. Neurophysiol. 116 (7), 1681–1694.

Newman, M.E.J., 2002. Assortative mixing in networks. Phys. Rev. Lett. 89 (20), 208701.

Newman, M.E.J., 2004. Coauthorship networks and patterns of scientific collaboration. Proc. Natl. Acad. Sci. U. S. A. 101 (Suppl. 1), 5200–5205.

Owen, A.M., McMillan, K.M., Laird, A.R., Bullmore, E., 2005. N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. Hum. Brain Mapp. 25 (1), 46–59.

Percival, D.B., Walden, A.T., 2000. Wavelet Methods for Time Series Analysis. Cambridge University Press, Cambridge.

Raichle, M.E., Fiez, J.A., Videen, T.O., MacLeod, A.M.K., Pardo, J.V., Fox, P.T., Petersen, S.E., 1994. Practice-related changes in human brain functional anatomy during nonmotor learning. Cereb. Cortex 4 (1), 8–26.

Ravasz, E., Barabasi, A. -L., 2003. Hierarchical organization in complex networks. Phys. Rev. E. 67 (2 Pt 2), 026112.

Rubinov, M., Knock, S.A., Stam, C.J., Micheloyannis, S., Harris, A.W.F., Williams, L.M., Breakspear, M., 2009. Small-world properties of nonlinear brain activity in schizophrenia. Hum. Brain Mapp. 30 (2), 403–416.

Schank, T., Wagner, D., 2005. Approximating clustering-coefficient and transitivity. J. Graph. Algorithms Appl. 9, 265–275.

Smit, D.J.A., Stam, C.J., Posthuma, D., Boomsma, D.I., de Geus, E.J.C., 2008. Heritability of 'small-world' networks in the brain: a graph theoretical analysis of resting-state EEG functional connectivity. Hum. Brain Mapp. 29 (12), 1368–1378.

Stam, C.J., Jones, B.F., Nolte, G., Breakspear, M., Scheltens, P., 2007. Small-world networks and functional connectivity in Alzheimer's disease. Cereb. Cortex 17 (1), 92–99.

Taulu, S., Simola, J., Kajola, M., 2005. Applications of the signal space separation method. IEEE Trans. Signal Process. 53 (9), 3359–3372.

van Beijsterveldt, C.E.M., van Baal, G.C.M., 2002. Twin and family studies of the human electroencephalogram: a review and a meta-analysis. Biol. Psychol. 62, 111–138.

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of 'small-world' networks. Nature 393 (6684), 440–442.

Winterer, G., Coppola, R., Goldberg, T.E., Egan, M.F., Jones, D.W., Sanchez, C.E., Weinberger, D.R., 2004. Prefrontal broadband noise, working memory, and genetic risk for schizophrenia. Am. J. Psychiatry 161 (3), 490–500.