IN PRESS, *COGNITIVE, AFFECTIVE, & BEHAVIORAL NEUROSCIENCE*

Reproducibility of functional network metrics and network structure: A comparison of task-related BOLD, resting ASL with BOLD contrast, and resting cerebral blood flow

Matthew J. Weber[1,2,*], John A. Detre[2,3], Sharon L. Thompson-Schill[1,2], & Brian B. Avants[4]

[1] Department of Psychology
[2] Center for Cognitive Neuroscience
[3] Department of Neurology, &
[4] Department of Radiology, University of Pennsylvania
[*] Corresponding author: mweb@psych.upenn.edu, 609 468 7599

Author note: We thank Branch Coslett, Roy Hamilton, and David Wolk for assistance with tDCS, John Pluta for assistance with data processing, and Sam Messing, Hengyi Rao, Marc Korczykowski, Tanya Kurtz, Jacqueline Meeks, Patricia O'Donnell, and Joan Sparano for assistance with scanning. Three anonymous reviewers provided invaluable insight and correction. Funding for the study was provided by NIH grants R01DC009209 and K24NS058386 and the University of Pennsylvania Center for Functional Neuroimaging. The authors declare no conflict of interest.

Abstract

Network analysis is an emerging approach to functional connectivity in which the brain is construed as a graph, and its connectivity and information processing estimated by mathematical characterizations of graphs. There has been little to no work examining the reproducibility of network metrics derived from different types of functional magnetic resonance imaging data (e.g., resting versus task-related, or pulse sequences other than standard BOLD), or of measures of network structure at levels other than summary statistics. Here we take up these questions, comparing the reproducibility of graphs derived from resting arterial spin-labeling (ASL) perfusion fMRI to those derived from BOLD scans collected while the participant was performing a task. We also examine the reproducibility of the anatomical connectivity implied by the graph by investigating test-retest consistency of the graphs' edges. We compare two measures of graph-edge consistency both within versus between subjects and across data types. We find a dissociation in the reproducibility of network metrics, with metrics from resting data most reproducible at lower frequencies and metrics from task-related data most reproducible at higher frequencies; that same dissociation is not recapitulated, however, in network structure, for which the task-related data is most consistent at all frequencies. Implications for the practice of network analysis are discussed.

Introduction

An emerging approach to neuroimaging data analysis is the construal of the brain's connections as a *graph*, in the mathematical sense of a set of nodes connected by edges, and the application of mathematical characterizations of graphs to the brain. For present purposes, we refer to this cognitive neuroscience method as *network analysis*, although network methods are widely used in other fields of the natural and social sciences. A four-dimensional fMRI dataset is generally converted into a graph by parceling it into regions, computing the correlation matrix of the time courses in those regions, and applying some transformation to those correlations so that some regions are connected to one another and others are not. (Non-binary approaches—i.e., examinations of graphs with edges that are signed and/or weighted—are possible but rare as yet.) The mathematical characterization of a graph can be regional—that is, applicable to a single node—or network-wide. Node metrics generally aim to characterize the *centrality* of a node, in the sense of the strength of its connectivity to other nodes. Network-wide metrics are somewhat more diverse. Here, we review six network metrics that are popular subjects of study in this emerging field.

Transitivity, *efficiency*, and *modularity* are, respectively, measures of clustering, efficiency of information transfer (derived from the mean path length between any given pair of nodes), and community structure. *Gamma* and *lambda* are measures of clustering and path length normalized relative to random graphs; their ratio quantifies a graph's *small-worldness*. Small-world networks are interesting in that they are highly clustered, yet few (<10) edges intervene between any given node and any other (Watts & Strogatz, 1998); such networks are robust to both targeted attack and random error (Achard et al., 2006). Several studies have reported that the brains of healthy adults appear to be organized in small-world networks (Eguiluz et al., 2005; Achard et al., 2006). Small-worldness and other network metrics have been reported as markers of individual traits, including age, intelligence, and language onset, and clinical status, including Alzheimer's disease, mild cognitive impairment, schizophrenia, traumatic brain injury, epilepsy and ADHD (for reviews, see Bullmore & Sporns, 2009; He and Evans, 2010; Wang, Zuo, & He, 2010; Bullmore & Sporns, 2012; see also Beckage

et al., 2012; Sato et al., 2013). They have also been documented as markers of noninvasive brain stimulation (Polanía et al., 2010, 2011, 2012). However, to be confident in the use of these metrics as measures of individual differences, we must be confident in their reliability, as reliability limits correlation (Nunnally, 1970).

A few investigators have already studied the reproducibility of various measures of graph structure and efficiency. Deuker et al. (2009) investigated test-retest magnetoencephalographic (MEG) data from participants at rest and engaged in a working memory task, with test and retest separated by 4–6 weeks. They found that whole-graph metrics including global efficiency, clustering coefficient, minimum path length, and small-worldness were highly reliable at lower frequencies (*a*, 7.8–15.6 Hz, through d^{-1}, 0.8–1.7 Hz) and during an *n*-back working memory task; metrics derived from resting-state graphs and at higher frequencies (*g*, 31.2–62.5 Hz, and *b*, 15.6–31.2 Hz) were less reliable. However, the resting-state data were collected from 2.25-minute recording sessions, whereas the task-related data were collected from 9-minute sessions.

Telesford et al. (2010) collected two 5-minute runs (154 images), separated by minutes, from healthy older adults performing an Eriksen flanker task. They examined graphs derived from correlations at a single frequency band, 0.009–0.08 Hz, and several thresholds for edge inclusion. They found that clustering coefficient, mean path length, and local efficiency were highly reliable across thresholds, but degree and global efficiency were not reliable at some thresholds; they also found that the location of high-degree nodes was more consistent both across and within subjects than that of low-degree nodes.

Vaessen et al. (2010) used diffusion tensor imaging (DTI) to examine the effects of image acquisition parameters on the reproducibility of small-world network metrics (mean degree, clustering coefficient, and mean path length) derived from white matter tractography. They quantified inter-regional connection strengths by counting the number of tracts connecting each pair of regions; they also analyzed binary versions of these connection matrices, where edges were present between any pair of regions that had an edge. They found that increasing directional resolution affected small-world network metrics but had no effect on reproducibility.

Bassett et al. (2011) also examined the reproducibility of network metrics derived from white matter tractography, using both DTI and diffusion spectrum imaging (DSI). Network metrics exhibited moderate to high reproducibility, with low coefficients of variation and intraclass correlation coefficients (ICCs) ranging from 0.4 to 0.8, with no consistent effects of imaging modality.

Schwarz and McGonigle (2011) examined graphs constructed from a publicly available reliability dataset, 25 healthy subjects with 197 volumes of resting BOLD data collected at each of three time points. Time points 1 and 2 were 5–16 months apart and time point 3 was collected on the same day as 2. Schwarz and McGonigle (2011) examined several different correlation thresholds and approaches to edge inclusion, with specific interest in converting correlations into positive weights on edges rather than the traditional approach of binarizing correlation matrices into unweighted graphs; they also examined the effect of including versus regressing out global signal, head motion parameters, and signal from white matter and CSF. They found that a "soft" mapping from correlations to positive weights with a very sharp power function (i.e., most edges set to a zero weight) recapitulated the properties found in binary networks. They also found that regressing global signal had a substantial effect on network metrics, in particular increasing gamma, the ratio between the observed clustering and clustering in a randomly wired network with the same degree distribution—a measure of small-world behavior.

Liang et al. (2012) examined the same dataset studied by Schwarz & McGonigle (2011), but focused on the effects of regressing out global signal and using partial rather than Pearson

correlation to define edges. They found that several network metrics were most reliable when defined using Pearson correlation, without global signal regressed out. They also found that network metrics derived from a higher frequency band (0.027–0.073 Hz) were more reliable than those derived from a lower band (0.01–0.027 Hz).

Most recently, Braun et al. (2012) collected 150-volume resting BOLD data from healthy young adults in two scans separated by two weeks, and examined graph metrics at two frequency ranges ("standard," 0.04–0.08 Hz, and "broad," 0.0083–0.15 Hz) and several edge densities, with and without global signal regressed out and/or truncation of the time series by 25 or 50 volumes. They found moderate reliability overall, with inclusion of more frequencies, more volumes, and fewer edges improving reliability, and global signal having little effect; all metrics tested (small-worldness, clustering coefficient, path length, local and global efficiency, hierarchy, assortativity, and modularity) were equally reliable.

It is evident that the approaches used in these studies are quite heterogeneous, a matter we pursue further in the Discussion. To motivate the present study, however, we note two things. First, the balance of the research on reproducibility of graph metrics focuses on BOLD fMRI (though other methods are also examined). This is understandable, as resting BOLD is a popular source of data for brain networks. However, it is well known that cerebral blood flow (CBF) computed from arterial spin-labeling (ASL) perfusion fMRI enjoys marked advantages in reliability and inter-subject consistency relative to BOLD fMRI, due in part to its attenuation of temporal autocorrelation and low-frequency noise (Aguirre, Detre, & Wang, 2005; Tjandra et al., 2005; Fernández-Seara et al., 2007; Leontiev and Buxton, 2007; Borogovac and Asllani, 2012; *inter alia*). This makes ASL potentially attractive as an imaging modality for constructing networks from brain activity. However, the details of network construction impose a countervailing statistical disadvantage: Network edges are constructed from inter-region correlations in time series, and the computation of the perfusion signal requires computing a difference image from each pair of labeled and unlabeled volumes, effectively decreasing the signal-to-noise ratio. It is an empirical question whether that statistical disadvantage will outweigh the known reproducibility advantages of ASL methods in the context of network construction.

Another respect in which the existing network reproducibility literature varies is the use of task-related versus resting data. Deuker et al. (2009) and Telesford et al. (2010) examine task-related data, whereas Schwarz & McGonigle (2011) and Braun et al. (2012) examine resting-state data. To our knowledge, no one has analyzed the reproducibility of network metrics derived from task-related BOLD in which task activity has been regressed out, or compared networks derived from resting and task-related data in the same subjects.

In this paper, we take a sample of data collected for another purpose and focus on several comparisons of network metrics hitherto unstudied in the same subjects. At two time points separated by about 30 minutes, our subjects each underwent standard BOLD fMRI scanning while they performed the Balloon Analog Risk Task (BART; Lejuez et al., 2002; Rao et al., 2008) as well as ASL scanning at rest. Between the scans, half the subjects received 15 minutes of bifrontal transcranial direct current stimulation (tDCS), and half received a sham tDCS protocol. From these data, we generated and examined six *data types* per subject. From the ASL data, we extracted BOLD data (designated ABC, ASL with BOLD contrast) and CBF, described above. In the BOLD data, we either left the task-related signal in (*BART*) or regressed it out (*residuals*); in addition, we either examined the whole 400-volume time series (*intact*) or reduced the number of volumes to 168 for parity with the resting data (*reduced*). Table 1 summarizes the data types and their key differences for reference. In graphs derived from these six data types, we examine the reproducibility of six standard network metrics, as

well as a hitherto unstudied metric, namely the similarity of network topology, operationalized by overlap between the edge vectors of two graphs. To our knowledge, this is the first study of network reproducibility in the same subjects using different functional imaging modalities, and one of very few to assess consistency of network topology as well as summary metrics. Although the many differences between our resting and task-related scans limit the generality of our observations (a matter on which we expand in the Discussion), we believe this type of work is necessary in order to develop maximally reliable approaches to network science in the brain.

Materials and Methods

Subjects. 22 subjects (9 female, ages 19–35) participated in the experiment. 11 (3 female, ages 19–31) underwent sham tDCS and the rest (6 female, ages 19–35) true tDCS, as described below. All participants had normal or corrected-to-normal vision, were right-handed and had no history of neurological or psychiatric disorders. Participants were informed of the experimental procedures and written consent was obtained from all participants according to the University of Pennsylvania Institutional Review Board. All research was performed in compliance with the Declaration of Helsinki. Each participant was paid $40 compensation for participating in the study.

Overview of experiment. Each subject was randomly assigned to receive true or sham tDCS and scanned immediately before and immediately after stimulation. Subjects were scanned with both conventional blood oxygen level-dependent (BOLD) fMRI and pulsed continuous ASL perfusion fMRI (Dai et al., 2008; Chen et al., 2011). The first scanning session consisted of an anatomical scan, during which the subject practiced the BART for approximately 6 minutes, followed by a resting ASL scan (6 minutes), a BOLD scan during which the subject performed the BART (11 minutes), and another resting ASL scan (6 minutes). The subject was then removed from the scanner for tDCS, which took place in the control room directly outside the scanner, and replaced in the scanner as soon as the tDCS procedure was complete. The second scanning session was identical to the first except without the anatomical and concomitant training. The delay between the end of tDCS and the beginning of the third ASL scan was approximately 3–5 minutes.

BART protocol. We adapted the BART protocol used by Rao et al. (2008), varying only some details of the risk and reward schedule. The task strongly engages the dorsolateral prefrontal cortex, anterior cingulate, anterior insula, basal ganglia, and thalamus (data not shown).

tDCS protocol. For all subjects, the anode over right DLPFC (F3 in the International 10-20 System) and the cathode over its left counterpart (F4). True stimulation lasted 15 minutes at 1.5 milliamps (mA), with an additional 30-second ramp-up time at the beginning of stimulation and a 30-second ramp-down at the end. Sham stimulation involved the same ramp-up and ramp-down periods with only 15 seconds of stimulation in between, after which subjects sat with the electrodes affixed for the remainder of the 15 minutes. Subjects were not told whether they had undergone true or sham stimulation. TDCS was conducted with a Magstim Eldith stimulator (Carmarthenshire, UK) and 5 × 5-cm rubber electrodes inside saline-soaked sponges, which were affixed to the scalp with rubber straps. Subjects were removed from the scanner for tDCS and replaced in the scanner immediately after tDCS was over.

Image acquisition. Imaging was conducted on a Siemens 3-Tesla Trio whole-body scanner (Siemens AG, Erlangen, Germany), using an 8-channel array coil. At the beginning of the first scanning session, high-resolution T1-weighted anatomical images were obtained using an MPRAGE sequence (TR = 1620 ms, TI = 950 ms, TE = 3 ms, flip angle = 15°, 160 contiguous slices of 1.0 mm thickness, in-plane resolution 1 mm ×

1 mm) while the subject performed training trials of the BART. A 400-volume sequence of conventional BOLD images was acquired at each time point using a standard echo-planar imaging sequence (TR = 1500 ms, TE = 30 ms, flip angle = 90°, 25 interleaved axial slices with 5 mm thickness, in-plane resolution 3.44 mm × 3.44 mm), identical to Rao et al. (2008). Two 84-volume resting ASL time series were acquired at each time point using a pseudocontinuous ASL sequence (Wu et al., 2007; Dai et al., 2008; Chen et al., 2011) with the following parameters: TR = 4 s, TE = 17 ms, flip angle = 90°, FOV = 22 cm, matrix = 64×64, labeling time = 1.5 s, post-labeling delay = 1.2s, 18 axial slices with 6 mm thickness and 1.2 mm gap, in-plane resolution 3.44 mm × 3.44 mm.

Image analysis. We used AFNI (Cox, 1996) to motion-correct and spatially normalize the functional images, and to skull-strip and spatially normalize the anatomical images. Functional and anatomical images were normalized to the Colin brain template in Talairach space at 2 × 2 × 2-mm resolution. The FMRIB Automated Segmentation Tool (FAST; Zhang et al., 2001) was used to segment the skull-stripped anatomical into grey matter, white matter, and cerebrospinal fluid (CSF); voxels were designated as a given tissue type if the FAST output indicated 100% confidence in that designation.

We concatenated the two 84-volume ASL time series collected at each time point into one 168-volume time series per time point. The PCASL sequence contains both tagged and untagged BOLD signal from which CBF measurements derive. Thus, it is possible to analyze either the underlying BOLD signal (ABC) or the time variation in the CBF signal which measures dynamic changes in cortical blood flow more directly than BOLD. ABC was computed from the ASL time series by regressing out a binary covariate representing the tagging. Difference images for the quantification of CBF from the ASL data were generated by simple subtraction (Aguirre et al., 2002); CBF was calculated from those difference images by the method of Wang et al. (2003b). For each of the four ASL scans administered to each subject, images with a global mean CBF more than 3 standard deviations away from that scan's mean were censored and a mean CBF image was calculated from the remaining images.

The intact BART time series required no further processing at this stage. For the intact residuals, we regressed out three regressors representing idealized task-related activation: two binary regressors encoding losses and wins and one parametric regressor encoding risk, each convolved with a canonical double gamma hemodynamic response function. Reduced BART and residual time series were generated by taking the middle 168 volumes of the 400-volume BOLD time series.

Graph construction. We used ANTs (Avants et al., 2011) to extract time series from the 90 cortical and subcortical ROIs defined in the Automated Anatomical Labeling (AAL) atlas of Tzourio-Mazoyer et al. (2002) and regress out variation associated with motion and global signal. The AAL atlas provides a standard label set frequently used in network analysis (see Wang, Zuo, & He (2010), Table 1). We then bandpass filtered each time series to yield fluctuations in three frequency bands: 0.01–0.03 Hz (the *low* band), 0.03–0.06 Hz (*middle*), and 0.06–0.11 Hz (*high*); we used Christiano-Fitzgerald filtering (Christiano & Fitzgerald, 1999) as implemented by the R function *cffilter* in package *mFilter*. We did not examine higher frequencies because the TR of the CBF time series was effectively 8 s, or 0.125 Hz. For each subject, data type, frequency band, and time point, we computed a 90 × 90 correlation matrix quantifying correlations between the ROI time series.

Graph analysis. To each correlation matrix, we applied several different sparsity thresholds to binarize the edges, such that the bottom 50% to 97.5% of the 4050 edges were discarded and the

remaining edges set to a common weight of 1. This approach maintains graph density across subjects and controls for inter-subject variation in base correlation levels (Liu et al., 2008; Power et al., 2011; Schwarz & McGonigle, 2011; Braun et al., 2012; Liang et al., 2012). On the binarized graphs, we computed six graph-wide metrics (transitivity, modularity, efficiency, clustering coefficient, characteristic path length, and small-worldness). These metrics measure efficiency of information transmission and community structure, and were calculated with the R package *brainwaver* (Achard et al., 2006). For each data type, frequency band, sparsity threshold, and metric, we calculated the intraclass correlation coefficient (ICC) between all subjects' scores at T1 and T2. We used ICC(1,1) in the terminology of Shrout & Fleiss (1979), a measure of absolute agreement. This procedure models the observations with a one-way analysis of variance (ANOVA) that incorporates the subject's true score and an error term that collapses a number of sources of error. The ICC score is given by

$$ICC(1,1) = \frac{BMS - WMS}{BMS + (k-1)WMS}$$

where *BMS* denotes the between-subject mean squared error of the ANOVA, *WMS* denotes the within-subject mean square of the same ANOVA, and *k* denotes the number of observations (two in our case). Further detail on the method can be found in Shrout & Fleiss (1979). We assessed the significance of each ICC by shuffling the labels of the T2 data 1000 times and comparing the observed ICC to the 1000 permuted ICCs. At each frequency band and metric, we then collapsed ICCs across sparsity thresholds and compared the mean ICC for each pair of data types via Wilcoxon rank-sum test.

We also computed two measures of similarity in network topology. These measures rely on the construal of the graph as a 4050-element *edge vector*; in this case, a given edge had a value of 0 (absent) or 1 (present). For each data type, frequency band, and sparsity threshold, we computed the 22 *test-retest* (TRT) Euclidean distances between each subject's edge vector at T1 and the same subject's edge vector at T2; we also computed the 462 *control* distances between each subject's edge vector at T1 and each *other* subject's edge vector at T2. Because Euclidean distances are obliged to go down with sparsity, we also calculated TRT and control Jaccard coefficients, which normalize for sparsity. The Jaccard coefficient is simply the intersection of the edge vectors (i.e., the number of edges appearing in both vectors) divided by the union (i.e., the number of edges appearing in either vector); edges present in neither vector are not considered. Note that high Jaccard coefficients denote high similarity, whereas high Euclidean distances denote low similarity. We compared TRT Jaccards to control Jaccards via Wilcoxon rank-sum test. Additionally, we compared within-subject distances (and Jaccards) across data types at each sparsity threshold and frequency band, quantifying those differences via Wilcoxon rank-sum test against zero.

Check for tDCS effects. Since half our subjects received a tDCS treatment between test and retest, we examined the groups for differences in network topology and metric ICCs. We assessed tDCS effects on reproducibility of network topology by Wilcoxon rank-sum test on TRT edge vector distances and Jaccards, and on reproducibility of network metrics by a test for the difference between independent correlations (implemented by the R function *paired.r* in package *psych*).

Results

All data types showed smaller TRT edge vector distances and greater TRT edge vector Jaccards relative to controls at almost all sparsity thresholds in almost all frequency bands. Distances and Jaccards showed the same pattern (Figures 1 and 2). Thus, most approaches to the data showed that subjects retained some network topology from scan to scan, although even reliable

differences between TRT and control edge vector distances were often small. Notable exceptions were CBF at high thresholds in the low and middle bands, and reduced BART at low thresholds in the same bands, suggesting that the statistical disadvantage of time series with fewer volumes harms reproducibility. Direct comparisons of TRT distances and Jaccards show a remarkably clear picture, with intact BART and reduced BART data showing consistent advantages in preserving network topology relative to the other data types at all frequency bands, and intact BART consistently outperforming reduced BART (Figure 3). As displayed in Figures 1 and 2, effects of tDCS were observed at several thresholds in the CBF data at low and high frequencies. All were in the expected direction: Subjects who had undergone true tDCS showed greater TRT distances and lower TRT Jaccards than those who had undergone sham tDCS. Thus, our measures of graph structure understate the reproducibility of CBF-derived graphs.

For graph metrics, the best reproducibility was found in the high frequency band in intact BART and task residuals, and less so in ABC (Figure 4). Reduced data and CBF did not exhibit good reproducibility in this band. ABC retained an advantage in reproducibility into the middle and lower frequency bands, especially at lower sparsities, whereas the BART and task residuals showed low ICCs in these bands. Figure 5 summarizes Figure 4 somewhat more compactly and specifically addresses comparisons among data types, comparing mean ICCs across sparsity thresholds at the different frequencies and metrics. Notable patterns include a consistent disadvantage for task residuals and reduced residuals at the low frequency band; in the middle, a somewhat consistent advantage for CBF and a consistent disadvantage for residuals; and, in the high frequency band, a consistent advantage for intact BART and disadvantage for ABC and CBF. TDCS effects on network metric reproducibility were rare and scattered (Figure 4).

Discussion

Summary of results. A few features of the results leap out. First, reproducibility of graph metrics is poor overall in the lowest frequency band, although metrics from CBF-derived graphs are somewhat more reliable—a remarkable finding, given the extreme statistical disadvantage imposed by the shorter time series. In the middle band, designated by Achard & Bullmore (2006) as the frequency range most strongly displaying small-world characteristics, reproducibility is quite low across the board. In the highest frequency band, several metrics were highly reproducible across a number of sparsity thresholds in intact BART and intact residuals, and those data types showed a strong advantage over the others across sparsity thresholds, with $-\log(p)$ ranging from 3 ($p<0.001$) to 10 ($p<10^{-10}$) (Figure 5); efficiency and mean path length were also highly reproducible in ABC data at low sparsities. However, graphs derived from reduced BOLD data and CBF yielded almost no highly reproducible metrics at any threshold in the highest frequency band. Reduced BART and reduced residuals performed nearly as poorly as ABC and CBF, suggesting that the statistical disadvantage of the shorter resting time series is highly consequential for reproducibility, although it is not clear at what point diminishing returns might set in—the 400 volumes of our BOLD time series might be much more or much less than the optimal acquisition length.

These results suggest a few interesting conclusions. First, since metrics from task-related data are most reproducible at the high frequency band and those from resting data are most reproducible in the lowest, it is possible that task-related coordination of brain activity occurs at higher frequencies than the coordinated activity underlying resting-state networks. However, given that our residuals show the same high-frequency advantage as the task-related BART data, that effect may also be attributable to the BOLD data's higher sampling rate (TR=1500 ms, versus 4000 ms for ABC and effectively 8000 for CBF) or simply their volume. This pattern of

results complicates the results of Braun et al (2012) showing that reducing the length of the time series substantially affects reproducibility of network metrics; although that appears to be true holding other things equal, it appears that imaging modality and TR may be able to compensate for the statistical disadvantage of shorter scan length in some circumstances. Future work should compare CBF and BOLD metrics more thoroughly, notably in longer time series, at varying test-retest intervals, compared to true resting BOLD as well as ABC, and (to state the obvious) without tDCS.

Low reproducibility in the CBF and ABC data may also be related to the tagging for CBF quantification; in general, ABC is less sensitive to task-related hemodynamic fluctuations than true BOLD data, and that lack of sensitivity may have led to correlation patterns that were difficult to reproduce (Aguirre, Detre, & Wang, 2005). Alternatively, the tDCS procedure may have negatively impacted the reproducibility of graph metrics in truly stimulated subjects by changing the topology of the graph; however, in previous analyses of these data, we saw no systematic effects of tDCS on global network metrics (though node centrality metrics were affected in a few instances). Low reproducibility for ABC and CBF is likely not due to the length of the ASL sequence, which at 168 volumes per subject is comparable to existing studies (154 volumes in Telesford et al. (2010), 197 in Schwarz & McGonigle (2011), and 150 in Braun et al. (2012)).

In most cases, TRT edge vector distances were lower and Jaccards higher than controls, suggesting that some network topology was retained over the test-retest interval. Network topology is more consistent in intact and reduced BART than the other data types, suggesting that task-related correlations confer an advantage in graph consistency over and above the statistical advantage of the 400-TR BOLD sequence over the 168-TR resting ASL sequence (since the 168-volume reduced BOLD graphs outperform graphs derived from the 400-TR intact residuals). It is also notable that, although Euclidean distances between test-retest graphs go down as the edge threshold becomes stricter, suggesting more test-retest consistency, Jaccard coefficients also go down, suggesting less. This suggests that the primary driver behind the shrinking of Euclidean distances is simply the fact that most edges are zero. The shrinking Jaccard coefficients tell us that an edge appearing at T1 is not more likely to appear at T2, or vice versa, in a strictly rather than a laxly thresholded graph; or, to put it differently, high correlations do not appear to be more reliable than modest ones. Likewise, Figure 5 shows that stricter thresholds tend to attenuate or erase differences in TRT distances and Jaccards across data types.

Fourth, and perhaps least obviously, network topology and network metrics paint a different picture of reproducibility. While intact and reduced BART show substantially more reproducible topology across frequencies than other data types, this advantage seems not to extend to network metrics, for which resting data types are more reproducible in the low frequencies. This may point to a basic lacuna in the field's interpretation of network metrics: We assume that similar summary statistics for a graph entail similar structure, but it appears that the test-retest correlation between network metrics varies at least somewhat independently of test-retest similarity in structure. (Imagine shuffling the rows/columns of the binarized correlation matrix that gives rise to the graph. All the summary statistics will remain the same, but the shuffled graph might share very few edges with the original.) Likewise, the tDCS effects we see on consistency of network structure do not appear to manifest in network metrics—perhaps because subjects are affected uniformly, leaving their rank ordering unchanged. Studies that show high reproducibility of graph metrics cannot, therefore, be assumed to show high reproducibility of graph topology; that question must be tested directly. Previous studies of network reproducibility have not consistently featured this level of inquiry; exceptions include Telesford et al. (2010), who

show that degree is much more consistent in network hubs than in low-degree nodes, and Bassett et al. (2011), who report high reproducibility of connectivity matrices (analogous to our measure of edge vector distance) derived from white matter tractography. The point is also foreshadowed by the work of Moussa et al. (2012), which compares the modular structure induced by graph-theoretic construals of brain networks with those extracted by spatial independent component analysis. Future work in this area would be well advised to investigate the reproducibility of topological measures as well as summary network metrics.

This result is foreshadowed by some recent inquiries into the properties of graph metrics. Zalesky, Fornito, and Bullmore (2012) have shown that small-worldness and its constituent measures (clustering coefficient and mean path length) are inflated by typical methods of null-network construction, such that even data constructed from noise can give rise to small-world networks if the transitivity of the underlying correlation matrix (see Table 1) is not held constant in the null benchmarks. In a similar vein, Davey et al. (2012) show that temporal filtering of fMRI time series inflates the variance in correlation coefficients, leading to spuriously high numbers of strong correlations simply by chance; this may account for our finding that graph structure (as quantified by TRT Jaccard; Figure 2) becomes less reliable as sparsity increases. Although neither of these results provides a precise account for our dissociation between reproducibility of graph metrics and that of graph structure, they underscore the general point that graph metrics represent a great reduction of the information found in graph structure and that details of the properties of correlation coefficients are highly consequential for network analysis, and functional connectivity analysis in general.

Relationship to existing work. The cited reports on reproducibility of network metrics take heterogeneous approaches. The fMRI studies examine white matter tractography, task-related BOLD, or resting BOLD, but only the tractography studies compare different scanning parameters. Different graph thresholding approaches are used—leaving aside the "soft" thresholding employed by Schwarz & McGonigle (2011) and restricting our attention to methods of binarizing correlations for unweighted graphs, we see thresholds chosen to minimize the number of edges while leaving all graphs fully connected with an equal number of edges (Deuker et al., 2009); thresholds set to equalize the quantity S = log(# nodes) / log (mean node degree), a measure of path length, across subjects (Telesford et al., 2010); thresholds either set to a minimum correlation coefficient or a target sparsity across subjects (Schwarz & McGonigle, 2011); thresholds only based on a target sparsity across subjects (Braun et al., 2012, Liang et al., 2012); and thresholds based on counting white matter tracts (Vaessen et al., 2010; Bassett et al., 2011). These approaches lead to varying ranges of sparsity under examination. Node definition is likewise variable; Deuker et al. (2009) take the natural approach of defining each of their 204 sensors as a node, whereas Telesford et al. (2010) use each grey matter voxel as a node, and the remaining studies create a 90-node graph from the AAL atlas (from which Bassett et al., 2011, go on to derive larger graphs through upsampling). Other variables are relevant to reproducibility as well, notably the inclusion or exclusion of global signal and the use of partial correlation for edge definition (Schwarz & McGonigle, 2011; Liang et al., 2012).

The present work takes a relatively conservative approach to these issues, using popular strategies for thresholding (sparsity, with binary edges), node definition (the AAL atlas), preprocessing (global signal is regressed out), and correlation metric (Pearson) in order to focus on the reproducibility of network structure and metrics in different data types. However, the relative novelty of the network approach to fMRI affords this subfield of imaging research a useful opportunity—namely, to identify best practices

early. As the cited papers on network reproducibility show, even this emerging and sparsely populated field already furnishes ample "investigator degrees of freedom" (Simmons, Nelson, & Simonsohn, 2011), and as the present work shows, the field's assumptions about the relationship between a graph's structure and its summary metrics may be meaningfully misguided. Recent concerns about reproducibility and validity in psychology and cognitive neuroscience should underscore the urgency of a standard procedure for network analysis.

Limitations. Our sample of 22 subjects is relatively small, limiting the reliability of intraclass correlation coefficients as a measure of reproducibility. Although our analyses indicated no strong effect of tDCS on network metrics, and no effect on topology in the BART and residual graphs, it may have had subtle effects that were difficult to detect in our sample. Note, however, that this should generally have caused us to understate reproducibility—although, since different data types might have been differentially affected, our conclusions about relative reproducibility should be viewed with caution.

Although the use of resting ASL allowed us to examine cerebral blood flow as a candidate measure for graph construction, ABC data is not quite as sensitive as true BOLD (Aguirre, Detre, & Wang, 2005), and our ASL scans had different parameters from our BART scans, most notably the TR (4000 ms in ASL versus 1500 ms in BART) and length (168 versus 400 volumes), but also voxel size. Comparison of resting BOLD to resting ABC and task-related BOLD will be of considerable interest in future work. On the other side, it might be interesting to examine task-related ABC and CBF data relative to task-related BOLD as well.

Our choice of task may be atypical in the extent to which it synchronizes neural activity across time and participants. The BART induces strong activation in a wide network of brain regions (Rao et al., 2008, 2010; Fukunaga et al., 2012; Schonberg et al., 2012). A task that engaged the brain less strongly or more locally might be expected to show lower reproducibility for reasons unrelated to scanning parameters, and possibly to retain more useful signal in the residuals. Likewise, the BART may have targeted our high frequency band more selectively than other tasks, or even versions of the BART with different timing, might. Thus, the generality of our observation that reproducibility varies across frequency and data type is necessarily limited.

Finally, it is worth noting that high reproducibility is not always desirable. State factors such as fatigue, mood, arousal, metabolism, and numerous others might be expected to induce systematic changes in information transfer in the brain, which might be reflected in network metrics or topology. Our work was motivated by the potential diagnostic applications of network methods to brain activation, which would ideally be quite resistant to state factors outside the clinician's control. However, such resistance may not be possible to achieve. In that case, understanding systematic variation in the features of brain networks across time will be at least as important as developing network measures that minimize such variation.

Conclusion. Both the statistical advantage of longer time series and the coordination of brain systems by a task appear to improve the reproducibility of network structure across frequencies, and of network metrics at relatively high frequencies (0.06–0.11 Hz). At lower frequencies, short resting scans yield more reproducible network metrics than do long scans on task, but do not yield a concomitant improvement in the reproducibility of network structure; this calls into question the relationship between network structure and metrics, an issue deserving of more and deeper investigation. Future work should continue to interrogate different data types and network structure as well as network metrics, in combination with other deviations from standard network analysis practices in preprocessing and node and edge definition.

Bibliography

Achard, S., Salvador, R., Whitcher, B., Suckling, J., & Bullmore, E. (2006). A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *The Journal of Neuroscience*, *26*(1), 63–72.

Aguirre, G. K., Detre, J. A., & Wang, J. (2005). Perfusion fMRI for functional neuroimaging. *International Review of Neurobiology*, *66*, 213–236.

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., & Gee, J. C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, *54*(3), 2033–2044.

Bassett, D. S., Brown, J. A., Deshpande, V., Carlson, J. M., & Grafton, S. T. (2011). Conserved and variable architecture of human white matter connectivity. *NeuroImage*, *54*(2), 1262–1279.

Beckage, N., Smith, L., & Hills, T. (2011). Small worlds and semantic network growth in typical and late talkers. *PLoS ONE*, *6*(5), e19348.

Borogovac, A., & Asllani, I. (2012). Arterial Spin Labeling (ASL) fMRI: Advantages, Theoretical Constrains and Experimental Challenges in Neurosciences. *International Journal of Biomedical Imaging*, *2012*.

Braun, U., Plichta, M. M., Esslinger, C., Sauer, C., Haddad, L., Grimm, O., Mier, D., Mohnke, S., Heinz, A., Erk, S., Walter, H., Seiferth, A., Kirsch, P., & Meyer-Lindenberg, A. (2012). Test–retest reliability of resting-state connectivity network characteristics using fMRI and graph theoretical measures. *NeuroImage*, *59*(2), 1404–1412.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, *10*(3), 186–198.

Bullmore, E., & Sporns, O. (2012). The economy of brain network organization. *Nature Reviews Neuroscience*, *13*(5), 336–349.

Chen, Y., Wang, D. J. J., & Detre, J. A. (2011). Test–retest reliability of arterial spin labeling with common labeling strategies. *Journal of Magnetic Resonance Imaging*, *33*(4), 940–949.

Christiano, L. J., & Fitzgerald, T. J. (1999). *The band pass filter*. National Bureau of Economic Research.

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*(3), 162–173.

Dai, W., Garcia, D., De Bazelaire, C., & Alsop, D. C. (2008). Continuous flow-driven inversion for arterial spin labeling using pulsed radio frequency and gradient fields. *Magnetic Resonance in Medicine*, *60*(6), 1488–1497.

Davey, C. E., Grayden, D. B., Egan, G. F., & Johnston, L. A. (2012). Filtering induces correlation in fMRI resting state data. *NeuroImage*.

Deuker, L., Bullmore, E. T., Smith, M., Christensen, S., Nathan, P. J., Rockstroh, B., & Bassett, D. S. (2009). Reproducibility of graph metrics of human brain functional networks. *NeuroImage*, *47*(4), 1460–1468.

Eguiluz, V. M., Chialvo, D. R., Cecchi, G. A., Baliki, M., & Apkarian, A. V. (2005). Scale-free brain functional networks. *Physical Review Letters*, *94*(1), 18102.

Fernández-Seara, M. A., Wang, J., Wang, Z., Korczykowski, M., Guenther, M., Feinberg, D. A., & Detre, J. A. (2007). Imaging mesial temporal lobe activation during scene encoding: comparison of fMRI using BOLD and arterial spin labeling. *Human Brain Mapping*, *28*(12), 1391–1400.

Fukunaga, R., Brown, J. W., & Bogg, T. (2012). Decision making in the Balloon Analogue Risk Task (BART): Anterior cingulate cortex signals loss aversion but not the infrequency

of risky choices. *Cognitive, Affective, & Behavioral Neuroscience*, *12*(3), 479–490.

He, Y., & Evans, A. (2010). Graph theoretical modeling of brain connectivity. *Current Opinion in Neurology*, *23*(4), 341–350.

Latora, V., & Marchiori, M. (2001). Efficient Behavior of Small-World Networks. *Physical Review Letters*, *87*(19), 1–4.

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., Strong, D. R., & Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: the Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, *8*(2), 75.

Leontiev, O., & Buxton, R. B. (2007). Reproducibility of BOLD, perfusion, and CMRO2 measurements with calibrated-BOLD fMRI. *NeuroImage*, *35*(1), 175.

Liang, X., Wang, J., Yan, C., Shu, N., Xu, K., Gong, G., & He, Y. (2012). Effects of different correlation metrics and preprocessing factors on small-world brain functional networks: A resting-state functional MRI study. *PLoS ONE*, *7*(3), e32766.

Liu, Y., Liang, M., Zhou, Y., He, Y., Hao, Y., Song, M., Yu, C., Liu, H., Liu, Z., & Jiang, T. (2008). Disrupted small-world networks in schizophrenia. *Brain*, *131*(4), 945–961.

Moussa, M. N., Steen, M. R., Laurienti, P. J., & Hayasaka, S. (2012). Consistency of Network Modules in Resting-State fMRI Connectome Data. (Y.-F. Zang, Ed.) *PLoS ONE*, 7(8), e44428.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113.

Nunnally, Jr., J. C. (1970). *Introduction to psychological measurement*. New York, NY: McGraw-Hill.

Polanía, R.,Nitsche, M. A., & Paulus, W. (2010). Modulating functional connectivity patterns and topological functional organization of the human brain with transcranial direct current stimulation. *Human Brain Mapping*, *32*(8), 1236–1249.

Polanía, R., Paulus, W., Antal, A., & Nitsche, M. A. (2011). Introducing graph theory to track for neuroplastic alterations in the resting human brain: A transcranial direct current stimulation study. *NeuroImage*, *54*, 2287–2296.

Polanía, R., Paulus, W., & Nitsche, M. A. (2012). Reorganizing the Intrinsic Functional Architecture of the Human Primary Motor Cortex during Rest with Non-Invasive Cortical Stimulation. *PLoS ONE*, *7*(1), e30971.

Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., & Petersen, S. E. (2011). Functional network organization of the human brain. *Neuron*, *72*(4), 665–678.

Rao, H., Korczykowski, M., Pluta, J., Hoang, A., & Detre, J. A. (2008). Neural correlates of voluntary and involuntary risk taking in the human brain: An fMRI Study of the Balloon Analog Risk Task (BART). *NeuroImage*, *42*, 902–910.

Rao, H., Mamikonyan, E., Detre, J. A., Siderowf, A. D., Stern, M. B., Potenza, M. N., & Weintraub, D. (2010). Decreased ventral striatal activity with impulse control disorders in Parkinson's disease. *Movement Disorders*, 25(11), 1660–1669.

Sato, J. R., Takahashi, D. Y., Hoexter, M. Q., Massirer, K. B., & Fujita, A. (2013). Measuring network's entropy in ADHD: A new approach to investigate neuropsychiatric disorders. *NeuroImage*, 77, 44–51.

Schonberg, T., Fox, C. R., Mumford, J. A., Congdon, E., Trepel, C., & Poldrack, R. A. (2012). Decreasing Ventromedial Prefrontal Cortex Activity During Sequential Risk-Taking: An fMRI Investigation of the Balloon Analog Risk Task. *Frontiers in Neuroscience*, 6.

Schwarz, A. J., & McGonigle, J. (2011). Negative edges and soft thresholding in complex network analysis of resting state functional

connectivity data. *NeuroImage*, *55*(3), 1132–1146.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(7), 420–428.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366.

Telesford, Q. K., Morgan, A. R., Hayasaka, S., Simpson, S. L., Barret, W., Kraft, R. A., Mozolic, J. L., & Laurienti, P. J. (2010). Reproducibility of graph metrics in fMRI networks. *Frontiers in Neuroinformatics*, *4*.

Tjandra, T., Brooks, J. C. W., Figueiredo, P., Wise, R., Matthews, P. M., & Tracey, I. (2005). Quantitative assessment of the reproducibility of functional activation measured with BOLD and MR perfusion imaging: implications for clinical trial design. *NeuroImage*, *27*(2), 393–401.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–289.

Vaessen, M. J., Hofman, P. A. ., Tijssen, H. N., Aldenkamp, A. P., Jansen, J. F. A., & Backes, W. H. (2010). The effect and reproducibility of different clinical DTI gradient sets on small world brain connectivity measures. *NeuroImage*, *51*(3), 1106–1116.

Wang, J., Zuo, X., & He, Y. (2010). Graph-based network analysis of resting-state functional MRI. *Frontiers in Systems Neuroscience*, *4*.

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

Wu, W.-C., Fernández-Seara, M., Detre, J. A., Wehrli, F. W., & Wang, J. (2007). A theoretical and empirical investigation of the tagging efficiency of pseudocontinuous arterial spin labeling. *Magnetic Resonance in Medicine*, *58*, 1020–1027.

Zalesky, A., Fornito, A., & Bullmore, E. (2012). On the use of correlation as a measure of network connectivity. *NeuroImage 60*(4), 2096–2106.

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *Medical Imaging, IEEE Transactions on*, *20*(1), 45–57.

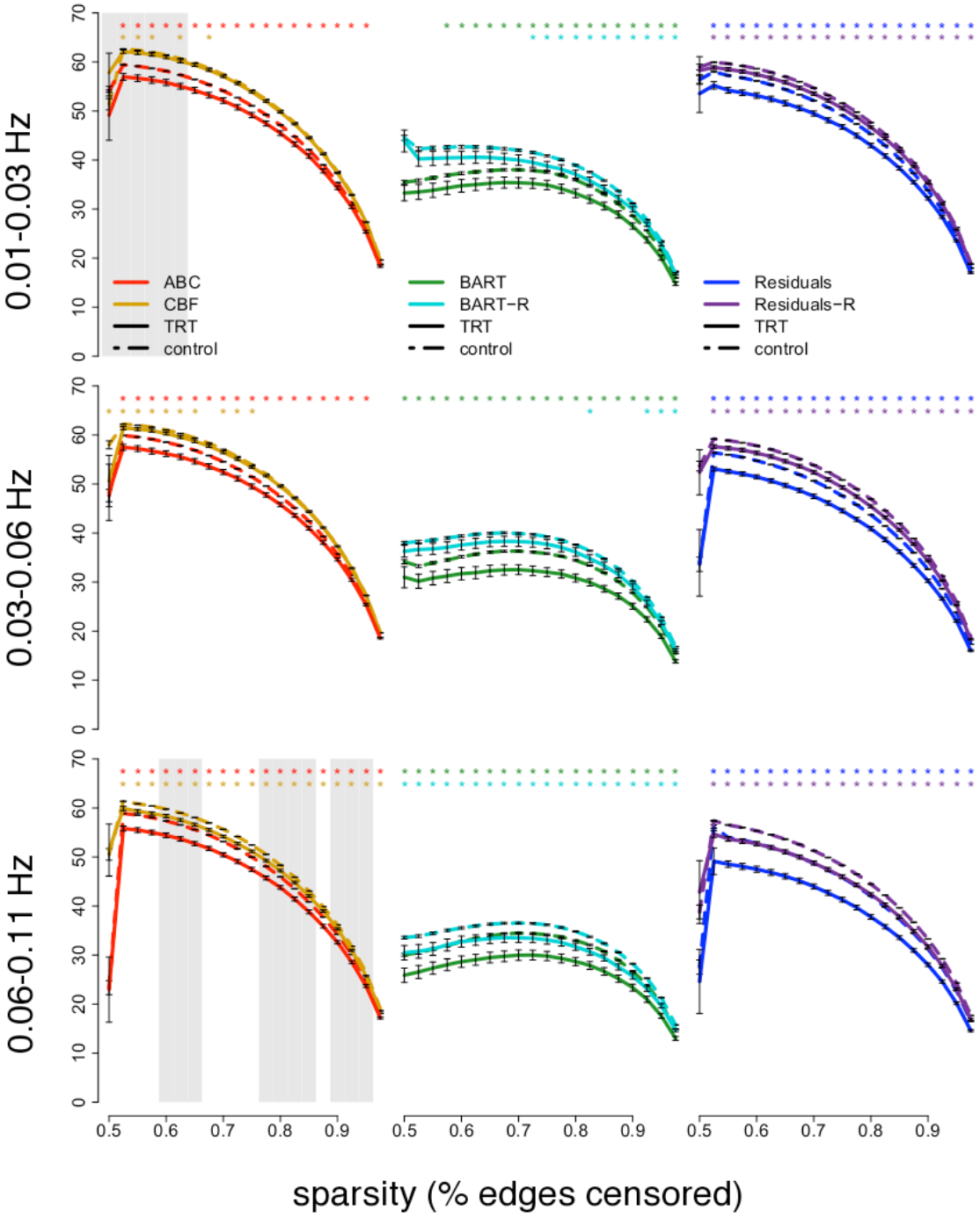

Figure 1: Edge vector distances as a function of network sparsity in three frequency bands. The leftmost graphs display ABC (red) and CBF (yellow); the middle graphs display task-related BOLD, intact at 400 TRs (green) and reduced at 168 (cyan); the rightmost graphs display BOLD with task activation regressed out, intact at 400 TRs (blue) and reduced at 168 (purple). Solid lines indicate "TRT" distances between the same subject at T1 and T2; dotted lines indicate "control" distances between a given subject at T1 and some other subject at T2. Error bars delimit ±1 standard error of the mean, and are smaller overall for control distances because there are many more control

distances than TRT distances. Stars (along the top edge of each graph) indicate a significant difference ($p<0.05$) between TRT and control distances by Wilcoxon rank-sum test at the given sparsity threshold. Thresholds shaded in grey denote that subjects who received true tDCS showed lower reproducibility (higher TRT distances) than subjects receiving sham tDCS; the opposite pattern was not observed.

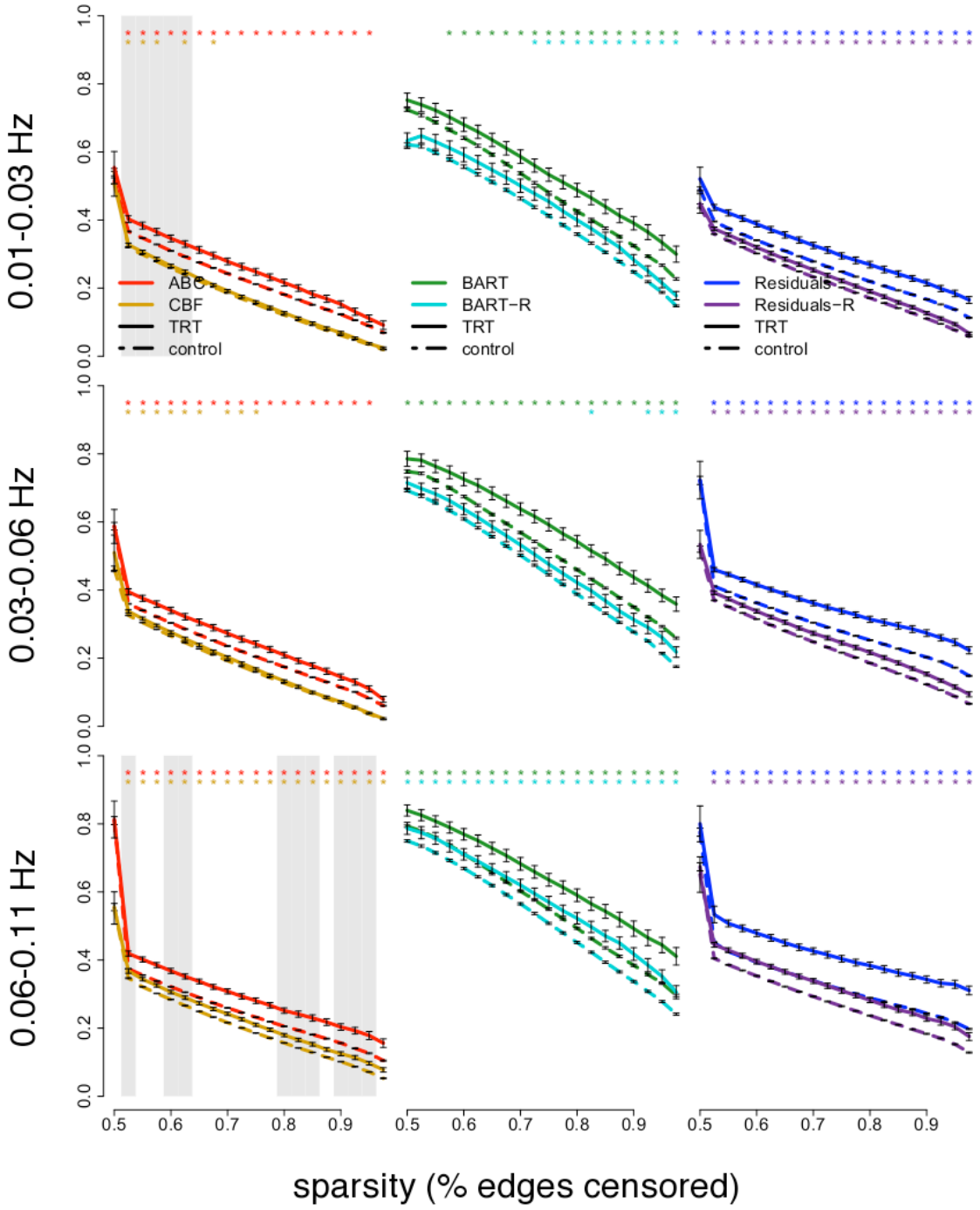

Figure 2: Jaccard coefficients as a function of network sparsity in three frequency bands. For calculation of Jaccard coefficients, see text; identical vectors have a Jaccard coefficient of 1. The leftmost graphs display ABC (red) and CBF (yellow); the middle graphs display task-related BOLD, intact at 400 TRs (green) and reduced at 168 (cyan); the rightmost graphs display BOLD with task activation regressed out, intact at 400 TRs (blue) and reduced at 168 (purple). Solid lines indicate "TRT" Jaccards between the same subject at T1 and T2; dotted lines indicate "control" Jaccards between a given subject at T1 and some other subject at T2. Error bars delimit ±1 standard error of the mean, and

are smaller overall for control Jaccards because there are many more control Jaccards than TRT Jaccards. Stars (along the top edge of each graph) indicate a significant difference ($p<0.05$) between TRT and control Jaccards by Wilcoxon rank-sum test at the given sparsity threshold. Thresholds shaded in grey denote that subjects who received true tDCS showed lower reproducibility (lower TRT Jaccards) than subjects receiving sham tDCS; the opposite pattern was not observed.

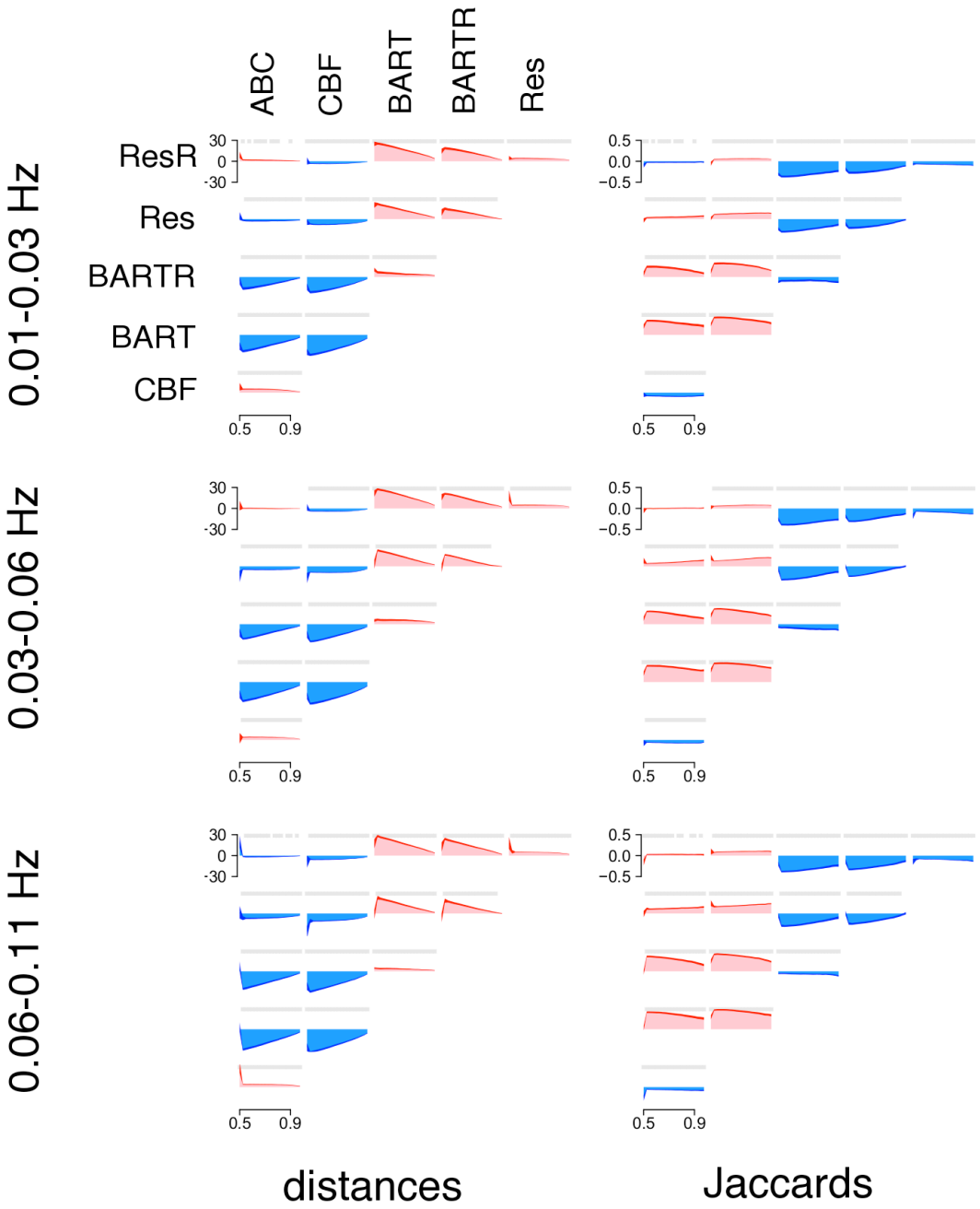

Figure 3. Differences in graph topology as a function of data type and sparsity. Red and dark blue plot the mean±SEM of the difference in a subject's TRT distances (left column) or Jaccards (right column) on graphs derived from two data types at a given sparsity threshold; red is used if the row result is greater than the column across sparsity thresholds, blue if the column result is greater than the row. Pink and light blue shading intervene between the mean±SEM and the x-axis. Grey squares along the top of the subgraph are present at thresholds where the difference is nonzero at $p<0.05$ by Wilcoxon rank-sum test (true in almost all cases). The difference is plotted as row minus column.

For example, the top left subgraph of the top left set of graphs (distances, 0.01–0.03 Hz) shows that ABC and reduced residuals have similar TRT distances across sparsity thresholds, with reduced residuals (the row) having slightly but significantly higher TRT distances than ABC (the column); this shows that network topology is slightly better preserved across the test-retest interval in ABC versus reduced residuals. Recall that *lower* TRT distances, but *higher* TRT Jaccards, correspond to higher test-retest similarity, hence the overall pattern of inversion from the left to the right column.

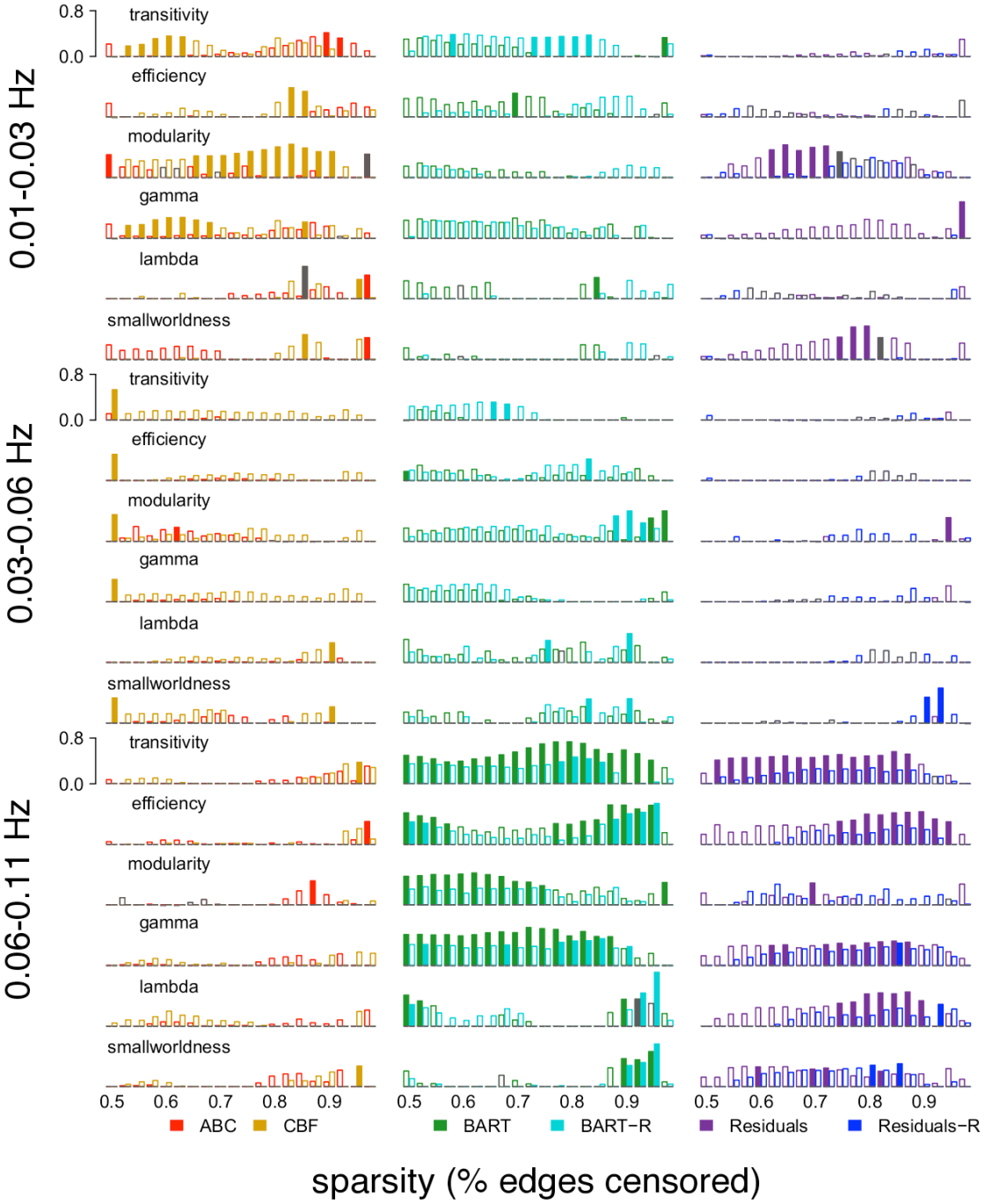

Figure 4. Reproducibility of graph metrics as a function of sparsity at three frequency bands. ICCs for six metrics are displayed for each frequency band, data type, and sparsity threshold. The left column displays ICCs for graphs derived from ABC (red) and CBF (yellow) data; the middle column displays ICCs for graphs derived from task-related BOLD, intact at 400 TRs (green) or reduced at 168 (cyan); the rightmost column displays ICCs for graphs derived from BOLD with task activation regressed out, intact at 400 TRs (blue) or reduced at 168 (purple). Filled bars are significant by permutation test at $p<0.05$; open bars are nonsignificant. Grey bars indicate that, at the given frequency,

metric, and threshold, subjects receiving true tDCS had ICCs significantly different from sham subjects as quantified by test for the difference between independent correlations.

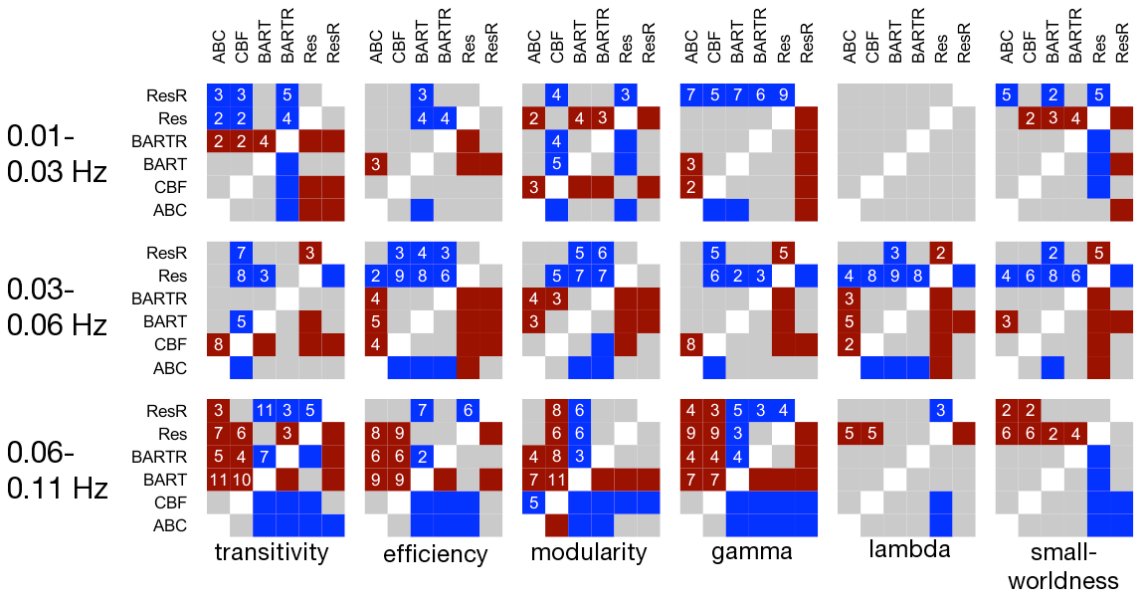

Figure 5. Comparisons of mean ICC across frequency band, graph metric, and data type. Each cell of each matrix codes a comparison between the mean ICC of a graph metric, at a given frequency band, across sparsity thresholds, for two data types. For example, the top left cell of the top left matrix codes the comparison of the mean ICC for transitivity, in the 0.01–0.03 Hz frequency band, for ABC versus reduced residuals. A cell is grey if the mean ICC does not differ by Wilcoxon rank-sum test at $p < 0.01$, red if the row had a greater ICC than the column, and blue if the row had a smaller ICC than the column; therefore, a red cell in the upper triangle will have a blue counterpart in the lower triangle and vice versa. Accordingly, a data type's overall performance in a frequency band can be read from the colors of its row, with numerous red cells denoting greater reproducibility than other approaches (e.g., 0.06–0.11 Hz, modularity, BART), and numerous blue cells denoting poorer reproducibility (e.g., CBF in the same matrix). Numbers code the rounded negative log of the p value of the comparison; thus, 2 codes a difference at $p<0.01$, 3 a difference at $p<0.001$, and so on. They are only displayed in the upper triangle of each matrix.

Data type	Source data	Resting?	Task signal?	# volumes
ABC	ASL	Yes	N/A	168
CBF	ASL	Yes	N/A	84
BART	BART	No	Yes	400
Reduced BART	BART	No	Yes	168
Residuals	BART	No	No	400
Reduced residuals	BART	No	No	168

Table 1. Summary of the six different data types.

Metric	Description
Transitivity	Also called *global clustering*. A measure of the extent to which the network is organized into highly connected clusters. Given by $\frac{\Sigma\tau_{\Delta}}{\Sigma\tau}$ where $\Sigma\tau_{\Delta}$ is the number of triangles in the graph (i.e., sets of three nodes all connected to one another) and $\Sigma\tau$ is the number of connected triples in the graph (i.e., sets of three nodes in which at least one is connected to all others).
Efficiency	Also called *global efficiency*. Mean shortest path length between each pair of nodes (Latora & Marchiori, 2001). Given by $\frac{1}{N(N-1)} \sum_{i \neq j \in G} \frac{1}{d_{ij}}$ where i and j are different nodes, G is the graph, d_{ij} is the shortest path from i to j, and N is the number of nodes in the graph.
Modularity	Extent to which the graph is divided into well-separated communities, as assessed by the edge betweenness method (Newman & Girvan, 2004). The betweenness of edge v is given by $\sum_{s \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$ where s and t are nodes in the graph, σ_{st} is the number of shortest paths from s to t, and $\sigma_{st}(v)$ is the number of such paths that include v. The community-finding algorithm works by iteratively assigning each edge a betweenness score, removing the edge with the highest score, and recalculating betweenness for all remaining edges. This procedure gives rise to a dendrogram showing the hierarchy of the graph's connections; cutting the dendrogram then gives rise to a community assignment. Given a community assignment, a graph's modularity is given by $Q = \sum_i (e_{ii} - a_i^2)$ where the i are community identifiers, e_{ii} is the fraction of edges connecting nodes in community i to one another, and $a_i = \Sigma_j e_{ij}$, the fraction of edges connecting nodes in community i to nodes in any community. In our work, the modularity of a graph is defined as its maximum modularity, i.e., the modularity of the community assignment that maximizes modularity.

Gamma	Given by C_G/C_{rand}. For graph X, C_X is the local clustering coefficient: the mean, over all nodes i, of $\frac{\Sigma\tau_{\Delta i}}{\Sigma\tau_i}$ where $\Sigma\tau_{\Delta i}$ is the number of triangles that include i, and $\Sigma\tau_i$ is the number of connected triples that include i. C_G is the local clustering coefficient of the observed graph, G; C_{rand} is the mean local clustering coefficient of 100 random graphs with the same number of nodes and edges.
Lambda	Given by L_G/L_{rand}. For graph X, L_X is the mean path length, the mean number of edges on the shortest paths between all pairs of nodes. L_G is the mean path length of the observed graph, G; L_{rand} is the mean path length of 100 random graphs with the same number of nodes and edges.
Small-worldness	Given by *gamma/lambda*. Small-world graphs are characterized by high clustering but low mean path length (relative to random graphs; Watts & Strogatz, 1998).

Table 2. Sketch of the six whole-graph metrics.