Start Up Research

Multimodal Imaging Data in Neuroscience

April 6, 2017

Contents

1 Introduction

As John Tukey's claims: *"The best thing about being a statistician is that you get to play in everyone's backyard."* Many fields of research provide increasingly multidimensional and complex object-type data — e.g. networks, functions, tensors, and others — along with novel motivating applications, and new methodological questions. In this stimulating setting, the neuroscience field is a relevant example.

Modern magnetic resonance imaging (MRI) technologies, combined with state-of-the-art data pre-processing algorithms, has made it possible to reliably measure brain structure and function non-invasively in live humans. These data come from a wide variety of technologies, covering electroencephalography (EEG), functional magnetic resonance imaging (fMRI), and diffusion tensor imaging (DTI), among others, opening new avenues to study underlying activity and connectivity patterns within the brain, how they are distributed in the population and if this distribution changes across groups of subjects with different traits — e.g. behavior, abilities and neuropsychiatric diseases. See Craddock et al. (2013), Monti (2011), Sporns (2013) and Stam (2014) for relevant discussions on recent advances in the analysis of dynamic brain activity and functional synchronization monitored via fMRI, as well as brain anatomical connectivity measured via DTI.

An accurate answer to the above questions requires novel statistical methods which need to catch up with the fast improvements in image processing. For example, state-of-the-art neuroscience datasets are increasingly multidimensional and multi-domain, meaning that each individual in a study is monitored under different brain imaging technologies (e.g Biessmann et al., 2011; Calhoun and Sui, 2016). Each of these imaging technologies has technical limits and provides complementary views on neural activity and connectivity, thereby motivating novel statistical methods for joint modeling of complex and mixed-domain data — e.g. brain networks, functional activity and synchronization of brain regions — to learn underlying dependence structures, spatio-temporal patterns, shared wiring mechanisms, and changes in these architectures across subject-specific traits.

These data require novel methods for network analysis, spatial inference, graphical modeling, multiple testing, dynamic inference, joint density estimation, data fusion, tensor factorization, prediction and many others: a wonderful backyard to dream up new methodologies and contribute to the available statistical literature.

2 NKI1 Multimodal Imaging Dataset

The multimodal imaging dataset motivating your novel methodologies comes from a pilot study of the Enhanced Nathan Kline Institute-Rockland Sample project. This project aims at providing a large cross-sectional sample of publicly shared multimodal neuroimaging data and psychological information to support and motivate researchers in the relevant scientific goal of understanding the mechanisms underlying the complex brain system. A detailed description of the project, scopes, and technical aspects can be found at `http://fcon_1000.projects.nitrc.org/indi/enhanced/`.

The pilot `NKI1` study motivating this first edition of StartUp Research, comprises multimodal imaging data and subject-specific covariates for $n = 24$ subjects. Detailed information can be found at `http://fcon_1000.projects.nitrc.org/indi/CoRR/html/nki_1.html`. An appealing aspect of this pilot study — compared to the whole dataset — is that for a wide set of subjects scan–rescan imaging data are available, thereby allowing validation and inference also on subject-specific variability in brain functions and structures. Below we summarize the main sources of information available for each subject in the pilot study, and then describe them more in detail in the following subsections.

Main sources of information:

1. `MULTIMODAL NEUROIMAGING DATA:`

   (a) **Structural networks**: These data measure the anatomical interconnections — made by white matter fibers — among brain regions of interest, and are collected from DTI.

   (b) **Dynamic functional activity**: These data measure the dynamic activity of each brain region through changes in the blood-oxygen-level dependent (BOLD) signal during resting state fMRI (R-fMRI) sessions.

   (c) **Functional networks**: These data measure synchronization in brain activity for each pair of brain regions, and are obtained from the correlation in dynamic functional activity.

   The anatomical brain regions on which such data are collected, are always the same across the subjects and the different imaging methods. We provide

ready-to-use imaging data for the brain parcellation based on the Desikan atlas (Desikan et al., 2006). For such $V = 70$ regions we have additional information on 3-D spatial locations, hemisphere and lobe membership, which may be relevant in several analyses. One region — in the left and right hemisphere — is marked as `unknown`, and typical analyses consider only the 68 regions characterizing the Desikan atlas.

Researchers interested in exploring the above mentioned DTI and R-fMRI data for other parcellations, can find the `NKI1` imaging data for other brain atlases at `http://m2g.io`.

2. `BRAIN REGION AND SUBJECT-SPECIFIC INFORMATION:`

   (a) **Informations on brain regions**: For the brain regions characterizing the Desikan atlas, information on names, 3-D spatial locations, hemisphere and lobe membership are available.

   (b) **Informations on the subjects in the study**: For the subjects in the study, data on age, handedness and psychological traits are available.

Clearly the ready-to-use brain activity and connectivity datasets provided in such studies, are the output of innovative pre-processing pipelines which take as input the raw imaging data and perform different operations to ensure meaningful analyses. Some examples includes anatomical parcellation, regions alignment, motion correction, and others. Refer to Craddock et al. (2013), and Sporns (2013) for a general overview of these pipelines and the different pre-processing steps. More details on the different sources of information and how to access them, are provided in the following subsections.

2.1 Structural networks

Structural networks made by white matter fibers connecting the different brain regions — within each subject's brain — are derived from DTI scans. DTI maps the diffusion of water molecules across the biological brain tissues, thereby allowing reconstruction of the white matter fibers which act as highways for the directional diffusion of water within the brain. The available **structural networks** are obtained as the output of the pipeline `ndmg` described at `http://m2g.io`. For

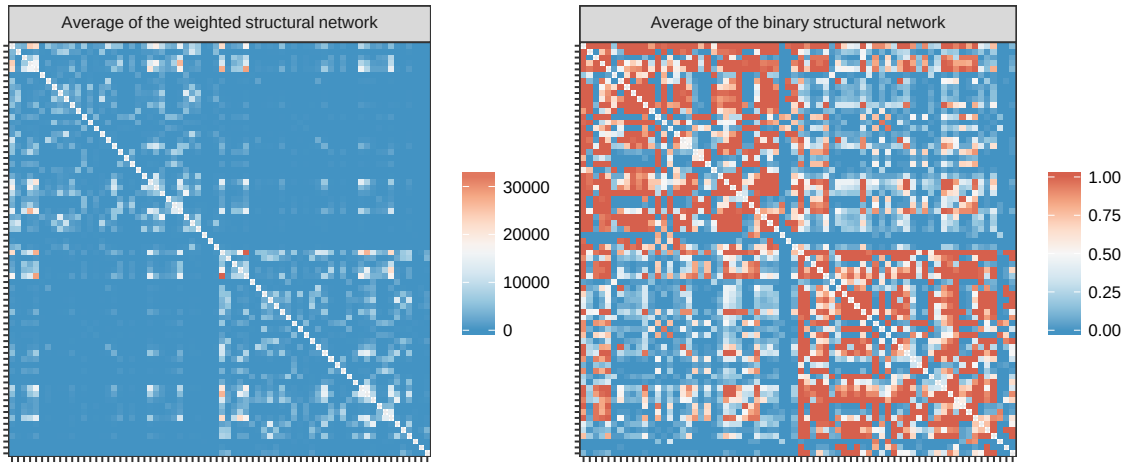

Figure 1: Left: Graphical representation of the structural network measuring fiber counts, averaged across subjects and scans. Right: Graphical representation of the dichotomized structural network measuring presence of at least one white matter fiber, averaged across subjects and scans.

citation, related softwares and details about the different steps characterizing this novel pipeline, refer to Gray Roncal et al. (2013) and Kiar et al. (2016). Further details on the data collection protocol are available at `http://fcon_1000.projects.nitrc.org/indi/CoRR/html/_static/scan_parameters/nki/nki_DTI.pdf`.

Structural networks — based on the Desikan atlas parcellation — can be found in the file `DTI-connectome.RData`. This dataset contains an array `D` of dimensions $70 \times 70 \times 24 \times 2$ comprising the 70×70 structural connectivity networks collected for the 24 subjects in each of the 2 scan-rescan imaging sessions. In particular `D[,,i,k]` is a 70×70 symmetric adjacency matrix measuring the total number of white matter fibers connecting each pair of brain regions in subject i, monitored during scan k, for every $i = 1, \ldots, 24$ and $k = 1, 2$. Based on this description, the element `D[v,u,i,k]` $\in \mathbb{N}$ is a discrete count characterizing how many white matter fibers are found to connect brain regions v and u in subject i, during scan k. Since self-relations are not of interest, the diagonal elements in each adjacency matrix are missing. Moreover, **structural networks** of 4 subjects are not available, and hence their corresponding adjacency matrices in `D` are empty.

As can be noticed from the left plot in Figure 1, these structural networks are sparse, with a moderate number of fibers counts being zero, and the others having

a wide range of variability. This has motivated a wide focus on binary structural networks measuring presence or absence of white matter fibers; see the right plot in Figure 1. However, there is still interest in weighted networks made by fiber counts, which may contain more information about brain connectivity.

2.2 Dynamic functional activity and synchronization

Dynamic functional activity — monitored for each brain region in the Desikan atlas — is obtained from resting state fMRI (R-fMRI). This imaging technology monitors brain functional activity at different regions via dynamic changes in blood flow creating a low frequency blood-oxygen-level dependent (BOLD) signal – when the subject is not performing an explicit task during the imaging session. In the present `NKI1` study, the subjects are simply asked to `stay awake with eyes open`. See Lee et al. (2013) for additional details about R-fMRI.

As for the structural networks, the raw R-fMRI scans are pre-processed to derive the time-series data for each brain region. In this case the `C-PAC` software has been used. For citation and details on the time-series extraction refer to documentation at `https://fcp-indi.github.io/` and `http://fcp-indi.github.io/docs/user/tse.html`, respectively.

Dynamic functional activity — based on the Desikan atlas parcellation — can be found in the file `fMRI-ROI-time-series.RData`. This dataset contains an

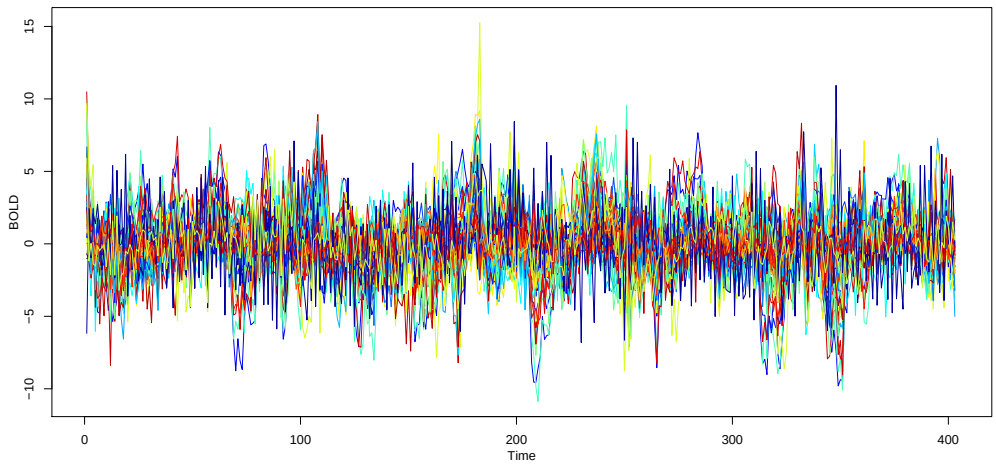

Figure 2: Dynamic activity time series of each brain region for a given subject.

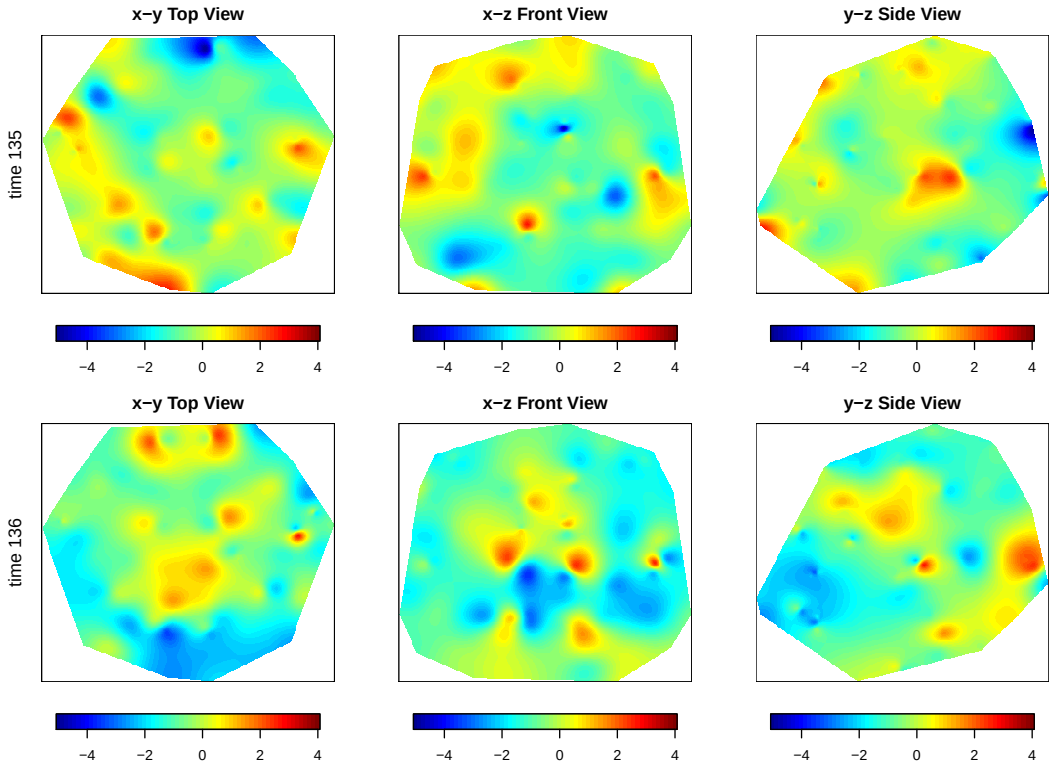

Figure 3: Spatial patterns in the dynamic activity during two consecutive times for a given subject. Data are shown under different 2-D views of the brain.

array `Y` of dimensions $70 \times 404 \times 24 \times 2$ comprising the 70×404 multivariate time-series activity data collected for the 24 subjects in each of the 2 scan-rescan imaging sessions. In particular `Y[,,i,k]` is a 70×404 matrix whose rows contain the dynamic activity data of the brain regions, collected at $T = 404$ equally spaced times, for subject i, monitored during scan k, for every $i = 1, \ldots, 24$ and $k = 1, 2$; see e.g. Figure 2. Based on this description, the element `Y[v,t,i,k]` $\in \Re$ is a real number measuring the activity of brain region v at time t, for subject i, during scan k. According to the data collection protocol at `http://fcon_1000.projects.nitrc.org/indi/CoRR/html/_static/scan_parameters/nki/nki_rest_1400.pdf`, the time lag between consecutive time observations is 1400 ms. **Dynamic functional activity** data for 2 subjects are not available, whereas for other 10 subjects only the first scan is observed. Therefore, their corresponding entries in `Y` are empty. Finally, during pre-processing, the spatial dependence has not been filtered out, and hence may play an important role in joint modeling of such data. Figure 3

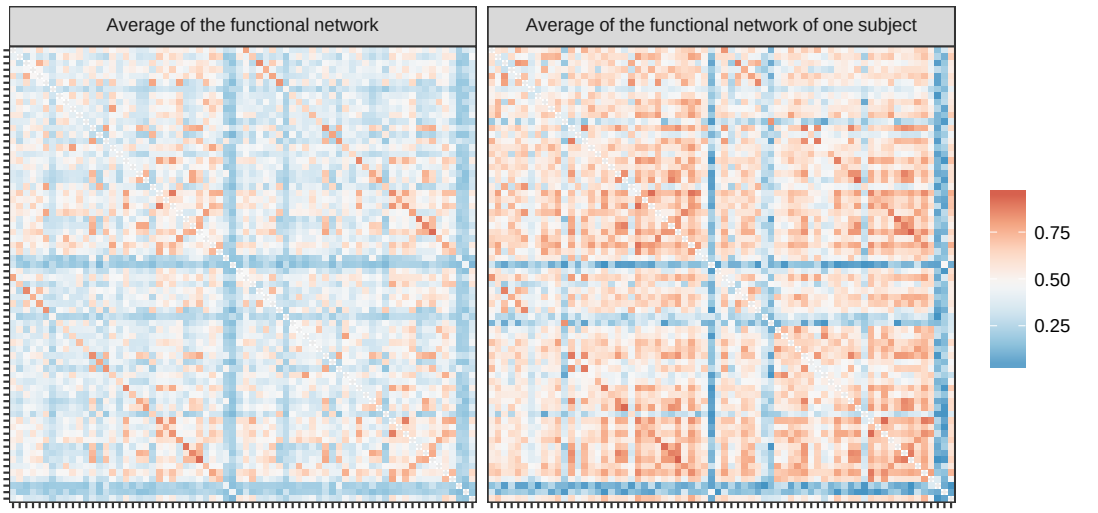

Figure 4: Left: Graphical representation of the functional network measuring pairwise correlation in brain activity between pairs of regions, averaged across subjects and scans. Right: Graphical representation of the functional network for one subject.

provides an example of the spatial patterns associated with the dynamic functional activity in two consecutive times for a given subject.

From the time-series data — characterizing **dynamic functional activity** — the **functional networks** — measuring synchronization in activity for each pair of brain regions — are also derived by simply computing the Pearson correlation coefficient for each pair of time-series data. These information are contained in the $70 \times 70 \times 24 \times 2$ array `W` which can be found in the file `fMRI-connectome.RData`. In particular `W[,,i,k]` is a 70×70 symmetric matrix measuring pairwise correlations for each pair of brain regions in subject i, monitored during scan k, for every $i = 1, \ldots, 24$ and $k = 1, 2$; see e.g. Figure 4. Based on this description, the element `W[v,u,i,k]` $\in [-1, 1]$ is simply equal to `cor(Y[v,,i,k],Y[u,,i,k])`. As can be noticed by a graphical analysis of these correlation matrices, some of them have an overall much higher functional correlation than others. This behavior has been noticed in the past by the computational neuroscientists who provided the data — with different pipelines — and their suggestion is to convert correlations to ranks. However, developing more consistent measures of functional synchronization or statistical models robust to this scaling effect may be relevant research directions.

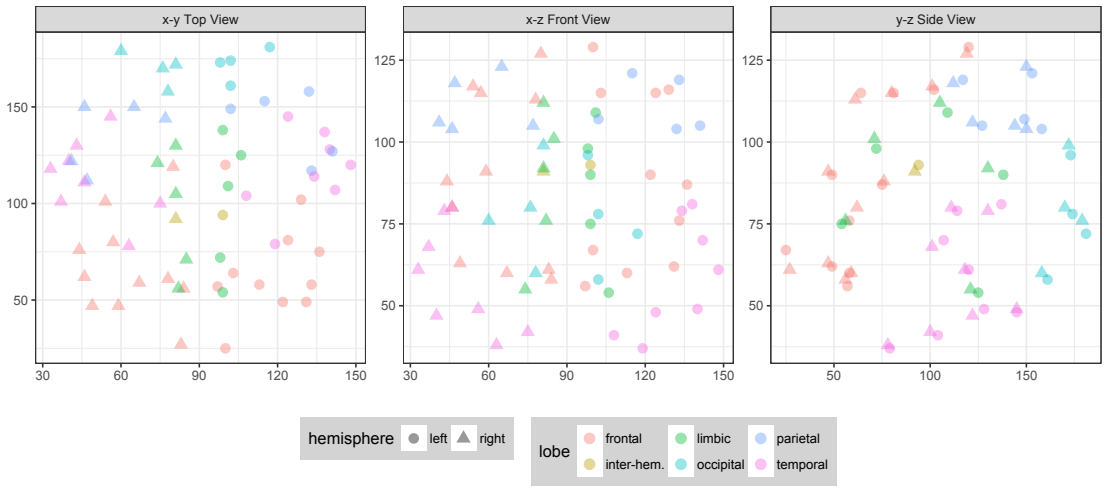

Figure 5: Graphical representation of the 3-D spatial coordinates for the centroids of the brain regions, under different 2-D views. Colors and shapes describe lobe and hemisphere membership, respectively.

2.3 Additional information on brain regions and subjects

Information on brain regions can be found in the file `ROI-covariates.txt`. In particular, for each brain region of interest characterizing the Desikan atlas we know its `name`, whether it belongs to the left or right `hemisphere`, the anatomical `lobe` membership — according to the classification of the Desikan atlas in anatomical lobes (Kang et al., 2012) — and the 3-D `spatial coordinates` of its centroid; refer to Figure 5 for a summary of the available covariates for the brain regions. These information have not been filtered out during the pre-processing, and it is of relevant interest to understand how the generative mechanism underlying the brain activity and connectivity data relates to such brain regions covariates.

Information on subjects can be found in the text file `SUBJ-covariates.txt`. In particular, for each subject in the study, we known her/his `id` — to match the statistical units with their corresponding imaging data — her/his `age` at the first scan, whether she/he is left-handed, right-handed or ambidextrous, and if the subject has been diagnosed a `current` and/or `lifetime` mental disorder. If present, the `type` of disorders are also described and separated by |. The description is based on the International Classification of Diseases. Refer to this link for more

information http://www.icd9data.com/2013/Volume1/290-319/default.htm.

These data clearly motivate several interesting directions of research. For example it is of relevant interest to develop flexible models which are able to infer global and local group differences in brain activity and connectivity across subjects with different traits — e.g. presence or absence of mental disorders, varying handedness or age. Improved prediction of mental disease — or other subject's traits — based on the information provided by the imaging data is also of interest and fundamental for early diagnosis. Finally, an appealing aspect of this dataset is that it provides replicated and multi-domain data on both brain anatomical connectivity and brain functional activity. Hence, statistical models for how structural links made by white matter fibers facilitate synchronization in brain dynamic activity, could have an important impact in the statistical and neuroscience literature.

Acknowledgements

These data are provided by Greg Kiar and Eric Bridgeford from NeuroData at Johns Hopkins University, who graciously pre-processed the raw DTI and R-fMRI imaging data available at http://fcon_1000.projects.nitrc.org/indi/CoRR/html/nki_1.html. If, as we hope, this experience will motivate some great contribution to the statistical literature, we kindly ask you to acknowledge Greg and Eric in your work and to cite the ndmg and C-PAC pipelines. Have fun!

References

Biessmann, F., Plis, S., Meinecke, F C., Eichele, T., and Muller, K. (2011). Analysis of multimodal neuroimaging data. *IEEE Reviews in Biomedical Engineering*, 4:26–58.

Calhoun, V. D., and Sui, J. (2016). Multimodal fusion of brain imaging data: A key to finding the missing link(s) in complex mental illness. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3:230–244.

Craddock, R. C., Jbabdi, S., Yan, C., Vogelstein, J. T., Castellanos, F. X., Di Martino, A., Kelly, C., Heberlein, K., Colcombe, S., and Milham, M. P. (2013). Imaging human connectomes at the macroscale. *Nature Methods*, 6:524–539.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., and Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31: 968–980.

Gray Roncal, W., Koterba, Z. H., Mhembere, D., Kleissas, D. M., Vogelstein, J. T., Burns, R., Bowles, A. R. Donavos, D. K., Ryman, S., and Jung, R. E (2013). MIGRAINE: MRI graph reliability analysis and inference for connectomics *Global Conference on Signal and Information Processing (IEEE)*, 313–316.

Kang, X., Herron, T. J., Cate, A. D., Yund, E. W., and Woods, D. L. (2012). Hemispherically-unified surface maps of human cerebral cortex: Reliability and hemispheric asymmetries. *PloS one*, 7: e45582.

Kiar, G., Gray Roncal, W., Mhembere, D., Bridgeford, E., Burns, R., Vogelstein, J. T. (2016). ndmg: NeuroData's MRI Graphs pipeline. *doi:10.5281/zenodo.60206*.

Lee, M. H., Smyser, C. D., and Shimony, J. S. (2013). Resting–state fMRI: a review of methods and clinical applications. *American Journal of Neuroradiology*, 34: 1866–1872.

Monti, M. M. (2011). Statistical analysis of fMRI time-series: a critical review of the GLM approach. *Frontiers in Human Neuroscience*, 5:1–13.

Sporns, O. (2013). Structure and function of complex brain networks. *Dialogues in Clinical Neuroscience*, 15:247–262.

Stam, C. J. (2013). Modern network science of neurological disorders. *Nature Reviews Neuroscience*, 15:683–695.