2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW) | 979-8-3503-0261-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSPW59220.2023.10193544

# GENERATIVE MODELS FOR LARGE-SCALE SIMULATIONS OF CONNECTOME DEVELOPMENT

*Skylar J Brooks*[1,2], *Catherine Stamoulis*[1,3],

[1]Boston Children's Hospital, Department of Pediatrics, Boston, MA, USA
[2]University of California Berkeley, Helen Wills Neuroscience Institute, Berkeley, CA, USA
[3]Harvard Medical School, Department of Medicine, Boston, MA, USA

## ABSTRACT

Functional interactions and anatomic connections between brain regions form the connectome. Its mathematical representation in terms of a graph reflects the inherent neuroanatomical organization into structures and regions (nodes) that are interconnected through neural fiber tracts and/or interact functionally (edges). Without knowledge of the ground truth topology of the connectome, functional (directional or nondirectional) graphs represent estimates of signal correlations, from which underlying mechanisms and processes, such as development and aging, or neuropathologies, are difficult to unravel. Biologically meaningful simulations using synthetic graphs with controllable parameters can complement real data analyses and provide critical insights into mechanisms underlying the organization of the connectome. Generative models can be highly valuable tools for creating large datasets of synthetic graphs with known topological characteristics. However, for these graphs to be meaningful, the variation of model parameters needs to be driven by real data. This paper presents a novel, data-driven approach for tuning the parameters of the generative Lancichinetti-Fortunato-Radicchi (LFR) model, using a large dataset of connectomes (n = 5566) estimated from resting-state fMRI from early adolescents in the historically large Adolescent Brain Cognitive Development Study (ABCD). It also presents an application, i.e., simulations using the LFR, to generate large datasets of synthetic graphs representing brains at different stages of neural maturation, and gain insights into developmental changes in their topological organization.

***Index Terms***— Brain connectome, topology, generative models, development

## 1. INTRODUCTION

Coordinated brain activity, reflected in correlations between electroencephalographic (EEG), functional MRI (fMRI) or magnetoencephalographic (MEG) signals, is often represented by a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, with $V$ nodes, corresponding to brain regions, and $E$ edges representing correlation strength. This representation facilitates the investigation the brain's topological organization and properties [1]. Brain graphs are typically estimated using a wide range of methods, including time-domain cross-correlation, frequency-domain coherence, covariance, and probabilistic measures and their variants [2]. Despite invaluable insights gained by brain graph analyses, the absence of a 'ground truth' topology makes it difficult to elucidate the mechanisms underlying normal or pathological connectome changes. In turn, this limits our understanding of fundamental biological processes, such as neural maturation and degeneration, and the effects of neuropathologies on the organization of brain circuits. Simulations using synthetic brain graphs with known and controllable topological properties can complement real data analyses, facilitate perturbations of specific graph properties, and provide mechanistic insights into the effects of these perturbations on neural information processing and cognitive function.

Generative network models are valuable for simulating large datasets of brain graphs. An early generative model of the human connectome was proposed by [3] to describe the probability of edge formation between brain regions. It incorporated a function that favored connections between nodes sharing nearest neighbors, and was able to reproduce hallmark topological properties of the brain, such as global efficiency, clustering, and modularity. Similar models have been used to generate synthetic structural networks, and have shown that model parameters and fit are affected by age, with worse fit in older individuals [4]. The study by [5] modeled white matter networks, which were hypothesized to undergo developmental changes that are necessary to increase their controllability and decrease synchronizability. To test this hypothesis, structural networks of over 800 subjects, ages 8 -22 years, were analyzed. Network development was simulated using a generative model that optimized controllability. Results showed that simulated graphs had similar developmental trajectories as real brain circuits. Other types of models have also been used to describe the human connectome. The Weighted Stochastic Block Model (WSBM) has been used to study how community structure in brain networks

Research supported by the National Science Foundation, awards 1940096 and 2207733

changes across the lifespan [6]. Community structure is often estimated using modularity maximization approaches, such as the Newman method [7]. However, WSBM has been shown to be an effective alternative, resulting in synthetic networks with more realistic organization than those based on maximizing modularity. Exponential Random Graph Models (ERGM) have also been used to describe the the structural and functional connectomes and individual brain networks [8, 9, 10]. Finally, other types of network models, such as graph convolutional networks, are increasingly used to learn the organization of brain networks for classification and prediction purposes [11, 12].

When simulating the topology of the human connectomes, the resulting synthetic graphs need to be biologically realistic. Thus, real datasets play a critical role in deriving model parameters. If these datasets are small, derived model parameters may not be representative of the population. Large-scale studies, such as the Human Connectome Project (HCP) [13] and the Adolescent Brain Cognitive Development (ABCD) study [14], provide a unique opportunity to estimate generalizable model parameters, an realistic synthetic graphs with topological characteristics that are similar to those of real connectomes. In addition, the selected generative model needs to have biologically interpretable parameters that can be mapped onto the brain's topological properties.

We present a novel approach that leverages the historically large ABCD dataset to estimate resting-state connectomes of early adolescents and related property statistics. The latter are used as inputs to the Lancichinetti-Fortunato-Radicchi (LFR) model [15], to generate a large dataset of synthetic brain graphs with variable topologies. The LFR was chosen for its flexibility (compared to other models), to vary model parameters in a way that can be mapped onto biologically meaningful changes in topological graph properties. Tuning the parameters of the LFR model can be linked to specific changes in node degree, community size (and number of communities), and inter-community connectedness, which is not straightforward or even possible with other models. In some of these models, parameter tuning can lead to simultaneous changes in multiple topological properties that are difficult to disentangle. As an application, we use the generated dataset in simulations, to investigate systematical changes in graph topology that reflect developmental processes and reorganization of brain circuits.

## 2. METHODS

### 2.1. Neuroimaging Data

Resting-state (rs) fMRI data from 5566 early adolescents in the ABCD study (median age = 120.0 months, inter-quartile range (IQR) = 13.0 months) were analyzed, to estimate task-independent connectomes and their topological properties. These data were selected based on quality of fMRI signals (minimal contamination by motion-related and nonbiological artifacts). Data were analyzed in the custom-developed Next-Generation Neural Data Analysis (NGNDA) platform [16]. They were first preprocessed to register the fMRI to each participants structural MRI, map onto a common atlas, correct for motion, and suppress various cardiorespiratory and nonbiological artifacts. Voxel-level time series were then downsampled to a parcel-level resolution. For this purpose, a high-resolution cortical parcellation [17] and additional atlases for subcortical regions and the cerebellum were used. Connectivity was estimated as the peak cross-correlation between parcel time series. Each functional connectivity matrix was thresholded using bootstrapping of multiple statistical and percolation-based thresholds. The moderate outlying peak cross-correlation was used as a conservative but realistic threshold, to eliminate weak connections and retain relatively strong connections. For comparison, connectivity was also estimated using mutual information, to assess method-dependence of connectome topologies. Both methods yielded statistically similar connectivity patterns. The data processing and connectivity estimation are described in detail in [18].

### 2.2. Network Generation Algorithm

Synthetic networks were generated using the LFR algorithm and Python library Networkx [19]. The algorithm assumes a power law distribution for node degree and community size. The distribution can be controlled using power law exponents, $\tau_1$ and $\tau_2$, for degree and community size, respectively. The number of nodes in the graph is specified by parameter $n$. For each node, the fraction of its connections to a node outside its community is determined by the parameter $\mu$. Thus, $\mu = 0$ results in a graph where all edges are between nodes of the same community, whereas $\mu = 1$ leads to a network where all edges are between nodes from different communities. Other inputs, such as the average (median) degree, can be used to further constrain node degree and community size. The steps to generate a graph using the LFR algorithm are:

1. Assign a degree to a node, drawn from the power law distribution of $\tau_1$. If average degree is given as an input, the resulting degree sequence must have an average degree equal to that value.

2. Select community sizes, by drawing from the power law distribution for exponent $\tau_2$ until the sum of communities equals the number of graph nodes $n$.

3. Randomly assign each node $u$ to a community, under the condition that the assigned community contains at least $(1 - \mu) * (degree(u))$ nodes. If the community becomes too large, randomly select a node to be moved to a different community.

4. For each node $u$, generate $(1 - \mu) * (degree(u))$ edges

within and $(mu) * (degree(u))$ edges outside its community.

### 2.3. Data-Driven Simulations

Resting-state connectivity matrices were thresholded to obtain binary and weighted adjacency matrices. The Newman algorithm was used to identify the number of communities and their respective sizes. For each brain, median community size was then estimated. The $25^{th}$ and $75^{th}$ quartiles of the number of communities, median community size, and median inter-community edge ratio in n = 5566 brains were then estimated. Similar statistics were estimated for median degree, and are summarized in Table 1.

**Table 1**. Quartiles (Q) of graph properties estimated from resting-state brain networks.

| Property | $25^{th}$ Q | $75^{th}$ Q |
|---|---|---|
| Median Degree | 25 | 91 |
| Number of Communities | 4 | 9 |
| Community Size | 103 | 273 |
| Inter-Community Connectedness ($\mu$) | 0.13 | 0.34 |

These statistics were used to guide the model input. Given that the appropriate values for degree and community size were not a priori known, a wide range of starting values for $\tau_1$ and $\tau_2$ were used, within the estimated quartile bounds. Not all values $\tau_1$ and $\tau_2$ led to graphs with a biologically realistic number of communities. Only graphs with a relatively small number of communities ($< 20$) were used in the next set of simulations, which extended the range of values for average degree and $\mu$, to simulate developmental stages before and beyond early adolescence. Lower and upper bounds for average degree were set to 10 and 150, respectively, and for $\mu$, 0.05 to 0.70. The lower bound corresponds to developed brains, in which communities that are highly connected locally are linked to each other by strong but sparse long-range connections. The upper bound corresponds to highly underdeveloped brains (in early life), in which redundant connections have similar weights and communities are difficult to distinguish.

To create a large set of networks that replicated the heterogeneity of the real dataset, $\tau_1$, $\tau_2$, $\mu$ and average degree input were individually varied. Although every combination of parameters was attempted, some simulations did not converge, thus 2669 valid binary graphs were generated. Nonzero edges in these binary graphs were assigned weights by sampling from the distribution of peak cross-correlation values estimated from the real data.

To simulate realistic connectomes at different stages of neural maturation, the following approach was used. First, individual resting-state networks were identified in the real data, using the anatomical delineations in [20], and were categorized as fully-developed, partially developed, or underdeveloped based on a large body of prior work. Given that participants were in pre/early adolescence, visual networks were assumed to be fairly well developed, the somatomotor network to be partially developed, and the frontoparietal control, limbic, reward and default-mode networks to be underdeveloped. This categorization was necessary in order to estimate ranges of connectivity values that reflected differential stages of network development, and assign weights to the binary graphs based on distributions of correlation values in each of these developmental categories. Second, in each synthetic graph, *within-community connectivity* was assigned assuming a developmental stage of the community (fully, partially or underdeveloped). Each edge between nodes within that community was assigned a weight by randomly sampling from the distribution of correlation values for its corresponding category. Finally, *between-community connectivity* was assigned assuming low, medium, and high correlation ranges. The bottom 10% values in real weighted adjacency matrices were used as the range of low intra-community correlation. Medium correlation was drawn from values between the $40^{th}$ - $50^{th}$ percentiles (below the median). The top 10% values were used as the range of high intra-community correlation. Corresponding median values for the 3 categories were r = 0.54, 0.67 and 0.73. Since each of the 2669 binary graphs was associated with three weighted graphs (corresponding to graphs with weakly, moderately and highly correlated communities), a total of 8007 weighted graphs were analyzed. Topological properties were estimated using the Brain Connectivity Toolbox and custom-developed codes [18, 21], and included global efficiency, global clustering, small-worldness, modularity, topological robustness, and topological stability. All properties used the weighted graphs (thus results are based on 8007 graphs), except small-worldness, for which results are based on 2669 binary graphs.

## 3. RESULTS

Simulations first assessed the relationship between model parameters and topological properties. To examine the link between $\mu$ (reflecting inter-community connectedness) and topological properties, degree and number of communities were held constant. Parameter $\mu$ was inversely related with small-worldness, modularity, and global clustering, which monotonically decreased as a function of reaching a plateau at $\mu \sim$0.40. Global efficiency sharply increased from $\mu$ = 0.05 to 0.20, but at a slower rate after that. Both robustness and stability increased up to $\mu \sim$0.5, and then began decreasing. The results are shown in Figure 1. Each data point represents a median over all graphs with a particular $\mu$ value.

Next, the relationship between number of communities and topological graph properties was examined, holding $\mu$ and degree constant. A higher number of communities was

associated with higher modularity and small-worldness, but lower global efficiency, robustness, global clustering, and stability The results are shown in Figure 2.

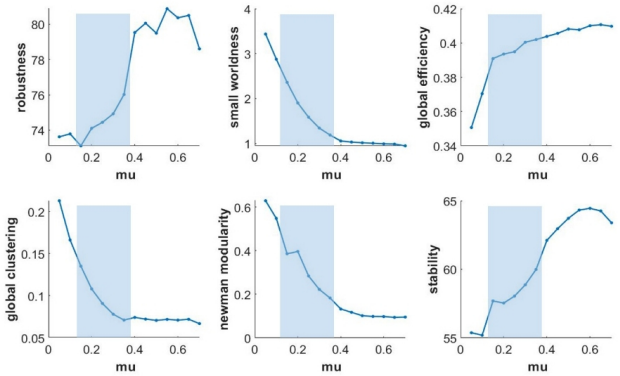


**Fig. 1**. Changes in topological properties as a function of $\mu$. Shaded areas reflect the range of values in the real data.

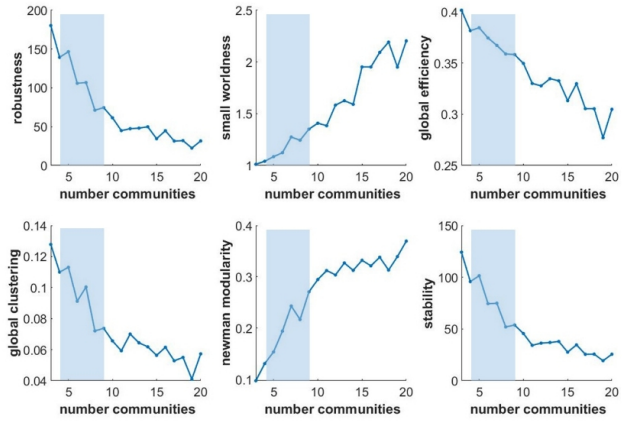


**Fig. 2**. Changes in topological properties as a function of number of communities. Shaded areas reflect the range of values in the real data.

Finally, the impact of varying intra- and inter-community connectivity was examined. Linear regression models assessed the relationship between median connectivity and graph properties (with the exception of small-worldness since it was calculated from binary graphs). Models were adjusted for number of communities, $\mu$ value, and average degree, and p-values were adjusted for the False Discovery Rate [22], over all network properties. Models also assessed the effect of inter-community connectivity, categorized as low (= 1), medium (= 2) or high (= 3). Median connectivity and inter-community connectivity were positively associated with global efficiency, stability, and clustering ($p < 0.04$). Both were negatively correlated with modularity. Median connectivity was linearly related with global efficiency and stability, but nonlinearly related with global clustering and modularity. The results are shown in Figure 3.

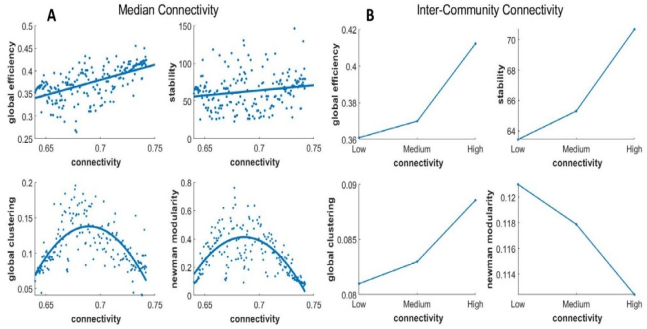


**Fig. 3**. Topological variations as a function of overall median connectivity (A), and inter-community connectivity (B).

## 4. CONCLUSION

We have presented a data-driven approach for generating large-scale synthetic brain graph datasets, using the flexible LFR model with realistic model parameters that reflect topological properties of real developing connectomes. We have outlined an approach for informing simulations of developmental topological variations and corresponding variations in model inputs, using statistics estimated from a large dataset of resting-state connectomes from early adolescents. As an application, we have used this generative graph framework to study the impact of normative development on the organization of the connectome. Together, the simulation results provide novel insights into topological changes across development. Early developmental stages (reflected in the choice of $\mu$ and lower median and inter-community connectivity) were associated with lower global clustering, efficiency and topological stability. Later developmental stages were associated with high small-worldness, stability and efficiency. However, increasing median connectivity was not monotonically related to modularity or clustering, with high median connectivity inversely proportional to global clustering and modularity. Instead, non-linear relationships between these parameters suggested a maximal optimization point as a function of changing connectivity. Increasing the number of network communities while holding $\mu$ to a range corresponding to partially developed connectomes was associated with lower efficiency, robustness, topological and global clustering. These results suggest that, in contrast to developed connectomes, an increasing community structure does not necessarily lead to more efficient and stable networks, when the underlying intra- and inter-community organization remains suboptimal (reflected in $\mu$). They also highlight the value of complementing real graph analyses with large-scale simulations to elucidate the effects of complex biological processes such human brain development.

## 5. REFERENCES


[1] O Sporns, *Networks of the Brain*, MIT Press, Cambridge, MA, 2010.

[2] J.T. Lizier, J. Heinzle, and A.and et al Horstmann, "Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fmri connectivity," *J Comp Neurosci*, vol. 30, pp. 85–107, 2010.

[3] P. E. Vertes, A. F. Alexander-Bloch, N. Gogtay, and et al, "Simple models of human brain functional networks," *Proceedings of the National Academy of Sciences*, vol. 109, pp. 58685873, 2012.

[4] R. F. Betzel, A. Avena-Koenigsberger, J. Goi, and et al, "Generative models of the human connectome," *NeuroImage*, vol. 124, pp. 10541064, 2016.

[5] E. Tang, C. Giusti, G. L. Baum, and et al, "Developmental increases in white matter network controllability support a growing diversity of brain dynamics," *Nature Communications*, vol. 8, pp. 1252, 2017.

[6] J Faskowitz, X Yan, X. N Zuo, and et al, "Weighted stochastic block models of the human connectome across the life span," *Scientific Reports*, vol. 8, pp. 12997, 2018.

[7] MEJ Newman, "Modularity and community structure in networks," *Proc Natl Acad Sci U S A*, vol. 103, pp. 85778582, 2006.

[8] S. L. Simpson, S. Hayasaka, and P. J Laurienti, "Exponential random graph modeling for complex brain networks," *PLoS ONE*, vol. 6, pp. e20039, 2011.

[9] M. R. Sinke, R. M. Dijkhuizen, A. Caimo, and et al, "Bayesian exponential random graph modeling of whole-brain structural networks across lifespan," *NeuroImage*, vol. 135, pp. 7991, 2016.

[10] P. E. Stillman, J. D. Wilson, M. J. Denny, and et al., "Statistical modeling of the default mode brain network reveals a segregated highway structure," *Scientific Reports*, vol. 7, pp. 11694, 2017.

[11] K.H. Oh, I.S. Oh, U. Tsogt, and et al, "Diagnosis of schizophrenia with functional connectome data: A graph-based convolutional neural network approach," *BMC Neurosci*, vol. 23, pp. 5, 2022.

[12] T.A. Song, S.R. Chowdhury, F. Yang, and et al., "Graph convolutional neural networks for alzheimers disease classification," *IEEE Int Symposium Biomed Imag*, vol. 1, pp. 414–417, 2019.

[13] "Hcp-development:," https://humanconnectome.org/study/hcp-lifespan-development/.

[14] B. Casey, T. Cannonier, Conley M.I., and et al, "The adolescent brain cognitive development (abcd) study: Imaging acquisition across 21 sites," *Developmental Cognitive Neuroscience*, vol. 3, pp. 4354, 2018.

[15] A. Lancichinetti, S. Fortunato, and F Radicchi, "Benchmark graphs for testing community detection algorithms," *Phys. Rev. E.*, vol. 78, pp. 046110, 2008.

[16] "Next-generation neural data analysis (ngnda) platform," http://github.com/cstamoulis1/Next-Generation-Neural-Data-Analysis-NGNDA/.

[17] A Schaefer, R Kong, Evan M Gordon, and et al, "Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri," *Cerebral cortex*, vol. 28, no. 9, pp. 3095–3114, 2018.

[18] SJ Brooks, SM Parks, and C Stamoulis, "Widespread positive direct and indirect effects of regular physical activity on the developing functional connectome in early adolescence," *Cereb Cortex*, vol. 31, pp. 4840–4852, 2021.

[19] "Networkx-generators-community-lfr-benchmark," https://networkx.org/documentation/stable/reference/generated/.

[20] BT Thomas Yeo, Fenna M Krienen, Jorge Sepulcre, and et al, "The organization of the human cerebral cortex estimated by intrinsic functional connectivity," *Journal of neurophysiology*, 2011.

[21] Rubinov M and Sporns O, "Complex network measures of brain connectivity: uses and interpretations," *Neuroimage*, vol. 52, pp. 10591069, 2010.

[22] Y Benjamini and Y Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, pp. 289–300, 1995.