IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Curvature Motifs in Connectome Data: Pitfalls with Path Motifs

LIRON I. KARPATI[1], MENTOR: DR. JIE GAO[2]
[1]University of Maryland, College Park, MD 20742 USA (e-mail: lkarpati@terpmail.umd.edu)
[2]Department of Computer Science, Rutgers University, New Brunswick, NJ 08901 USA (e-mail: jg1555@rutgers.edu)

This research is funded by NSF HDR TRIPODS award CCF-1934924.

ABSTRACT Nervous systems are organized for efficient integration of information [1]. It has been shown that in neuronal connectomes of the C. Elegans worm, high degree nodes are highly connected to form what is called a "rich-club" structure [2]. This rich-club structure allows different neurons to reach one another in a relatively few number of hops. The rich-club organization is a global architectural feature of the C. Elegans connectome. It is yet to be explored how the local organization of the connectome is supporting integration integration. By using the notion of course Ricci curvature (a generalization of Ricci curvature) and a path motif analysis, we found that the C. Elegans connectome exhibits local weak-bridge structures. More importantly, our analysis highlights some of the critical pitfalls in hypothesis testing graph properties. These pitfalls motivate the need for a principled investigation into statistical methods for determining the significance of graph properties.

INDEX TERMS Connectomics, C. Elegans, Motif, Ricci Curvature, Neuroscience, Complex Network

I. INTRODUCTION

THE field of connectomics aims to understand the structure-functional relationship of the nervous by studying its graph theoretic and complex network theoretic properties. One particular line of research aims to understand how a connectome's structure facilitates inter-node communication. Early in the development of connectomics it was shown that, at both the micro and macro scale, structural connectomes have a rich club organization, meaning there are a few well interconnected hub nodes through which a significant number of shortest paths pass [1] [2] [3]. It is believed that this rich club structure supports the functional capacity for efficient integration of information across different brain regions since different regions can rapidly communicate through the rich club. This functional role of the rich club structure was supported using a path motif analysis [1]. Newer work in complex network analysis and connectomics has begun to look at notions of discrete Ricci curvature to analyze the structure of networks [4] [5] [6]. It has been shown that discrete Ricci curvature can be used for understanding important network structures like communities and bridges between communities [7]. In this paper we aim to understand whether curvature might help us understand how a neuronal connectome's organization might be supporting inter-neuron communication. Paralleling the path motif analysis done for rich clubs in [1], we put forth the notion of a curvature path motif and examine the curvature path motifs in the C. Elegans connectome. It will be seen that the usual methodology for doing motif analyses fails in our analysis. The reasons for this failure will be explored with ramifications for null hypothesis testing in connectomics more broadly.

II. BACKGROUND

A. CURVATURE ON GRAPHS

The notion of curvature comes from geometry. Loosely speaking, curvature is a measure of how much a surface deviates from being a plane. For example a sphere has positive curvature, a plane has zero curvature, and a saddle has negative curvature. One measure of curvature for Riemannian manifolds (which surfaces are the 2-dimensional case of) is Ricci curvature. We will restrict our attention to a particular generalization of Ricci curvature called *course Ricci curvature*, also known as *Ollivier's Ricci curvature* after Yann Ollivier who proposed it [8]. Course Ricci curvature allows us to generalize the notion of curvature to networks. We will first present the formal definition and then explain the intuition behind the definition in the context of networks.

Roughly speaking, course Ricci curvature is a measure comparing the distance between the neighborhoods of two points to the distance between the two points. Therefore, in order to give the definition we must define what distance between two neighborhoods means. This is given by the Wasserstein distance. The following definitions are taken

from [6].

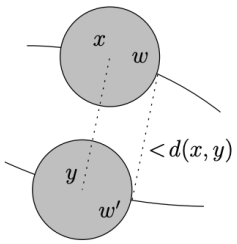
Wasserstein Distance: Let (X, d) be a metric space with two probability measures μ_1 and μ_2 each with mass 1. A transportation plan from μ_1 to μ_2 is a measure ξ on $X \times X$ that is mass preserving, meaning $\int_y d\xi(x, y) = d\mu_1(x)$ and $\int_x d\xi(x, y) = d\mu_2(y)$, where $d\xi(x, y)$ represents the amount of mass traveling from x to y. The *Wasserstein distance* between μ_1 and μ_2, denoted $W(\mu_1, \mu_2)$, is the minimum average traveling distance of any transportation plan:

$$W(\mu_1, \mu_2) = \inf_{\xi} \int \int d(x, y) d\xi(x, y)$$

We are now in the position to give the definition of course Ricci curvature.

Course Ricci curvature: Let (X, d) be a metric space and for each $x \in X$, let m_x be a probability distribution on X. Further let $x, y \in X$ be distinct points. The *course ricci curvature* along xy is given by $\kappa(x, y) = 1 - \frac{W(m_x, m_y)}{d(x,y)}$.

Intuitively, the course Ricci curvature is a measure comparing the Wasserstein distance (distance between neighborhoods) with the distance of the neighborhood centers. As can be verified from the formula $\kappa(x, y) = 1 - \frac{W(m_x, m_y)}{d(x,y)}$, if the neighborhoods are farther than their centers then the curvature is negative, if equidistant then curvature is zero, and if closer then curvature is positive. The reason this is a generalization of geometric curvature can be understood from the following picture which comes from Ollivier [8]:

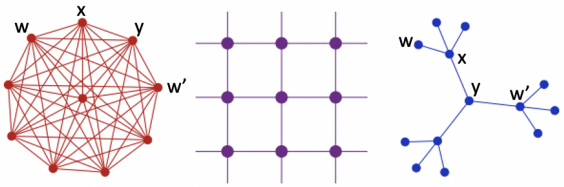


We observe two points x and y with small balls around them. Those balls are their neighborhood. w and w' are representative neighborhood points for illustrative purposes. If the points were on a flat Euclidean plane, the distance between x and y would be the same as the distance between w and w'. But, since the points are on a positively curved space where 'parallel' lines get closer and eventually intersect, we see that the distance between w and w' is actually less than the distance between x and y. On average, then, this will be true of any two points in the neighborhood. Therefore on positively curved surfaces, neighborhoods are 'closer' than their centers, on flat surfaces neighborhoods are 'equidistant' to their centers, and on positively curved surfaces, neighborhoods are 'further' than their centers. Course Ricci curvature generalizes this particular neighborhood property of curvature to discrete settings like networks.

We will now instantiate the definition of course Ricci curvature to define what we will call Ollivier-Ricci curvature on an undirected and unweighted network (it can be readily extended to directed and weighted networks). This instantiation comes from [6] and is the discrete curvature (modified for directed graphs) used in this paper. Let G be a directed network with vertices X. The distance metric on the graph is shortest path length. For a particular vertex x let $N(x) = \{x_1, ..., x_k\}$ be the neighbors of x. The corresponding probability distribution m_x^α, where $\alpha \in [0, 1]$ is a parameter, will be given by:

$$m_x^\alpha(x_i) = \begin{cases} \alpha & \text{if } x_i = x \\ (1-\alpha)/k & \text{if } x_i \in N(x) \\ \alpha & \text{else.} \end{cases}$$

We choose $\alpha = 0.5$ and only consider one-hop neighbors for the computations done in this paper. With the probability distributions of every point and the distance metric decided, we have all the necessary ingredients to establish a course Ricci curvature between any two vertices. In particular we calculate the curvature of vertices that share an edge which gives us *the Ollivier-Ricci curvature (ORC) of that edge*. For details on computing the ORC of an edge see [6]. From the definition, we can intuitively understand positive edge ORC as meaning the neighborhoods of endpoints overlap/ are well connected and negative curvature edge ORC as meaning the neighborhoods of endpoints are not well connected. The following picture from [9] (modified slightly) captures this intuition graphically.



(a) Clique (> 0) (b) Grid ($= 0$) (c) Tree (< 0)

xy is a representative edge and w, w' are representative neighbor vertices of x, y respectively. To summarize the picture: edges in a clique have positive curvature since the neighborhoods of the endpoints are very connected, edges in a grid are flat (0 curvature), and edges in a tree have negative curvature since the neighborhoods of the endpoints are only connected by xy (the neighborhoods are nearly disjoint).

III. METHODS

A. DATA SOURCE

The neuronal C. Elegans connectome used in our analysis comes from the open source connectome database NeuroData [10]. The C. Elegans connectome we use consists of 272 nodes and 4451 edges.

B. CLASSIFYING EDGE CURVATURE

We computed the the Ollivier-Ricci curvature (at an alpha value of 0.5) of each edge in the C. Elegans network using the networkX addon that can be found at [11]. Because Ollivier-Ricci curvature takes values between -1 and 1, we partitioned

$[-1, 1]$ into the three equally sized intervals $[-1, -0.33)$, $[-0.33, 0.33]$, $(0.33, 1]$ and then classified an edge as either having negative curvature (N), flat curvature (F), or positive curvature (P) when the value fell in the respective interval.

C. IDENTIFYING CURVATURE PATH MOTIFS

A *curvature path pattern* is a sequence of the characters N, F, and P which represents a possible sequence of curvature values along a path in our network. For example the pattern NFFP would correspond to paths in which there is an edge with negative curvature followed by an edge with flat curvature followed by an edge with flat curvature followed by an edge with positive curvature. The length of a path is the number of edges in the path. The length of a path pattern is the length of the string representing it.

Given a network, we can compute the number of paths that match a particular path pattern. We consider a path pattern a *motif* (or *anti-motif*) if the frequency of matches to the pattern in our network of interest significantly exceeds (or is significantly below) the frequency of matches to the pattern in a sample of random 'null' graphs. We computed a sample of 1000 random graphs which preserve the number of nodes, number of edges, and degree distribution of the C. Elegans network.

To test if a given path pattern is a motif (or anti-motif), for each random graph we counted the number of matches to that particular path pattern. This gives a null distribution for the frequency of such matches. We then test the null hypothesis that the C. Elegans network is not different from a random network (with respect to the number of matches to the curvature path pattern of interest) using a two-tailed z-test with an α-value of 0.01. If the p-value (Bonferroni corrected by the number of path patterns tested) is significant and with positive (or negative) z-score we call it a path motif (or path anti-motif) respectively.

Because there is still debate about whether it is sound to model neural information flow along shortest paths, by diffusion, or by navigation [3] we will do this analysis with respect to all possible paths and separately with respect to shortest paths. For all possible paths we consider only patterns of length three which appear in the C. Elegans network with a frequency of over 1% for computational efficiency purposes. For shortest paths we consider all possible patterns of length at least three which appear in the C. Elegans network with a frequency of over 1%.

All code used for the analysis in this paper can be found on GitHub at [12].

IV. RESULTS

A. C. ELEGANS ALL PATHS

It was found that only three patterns of length three match over 1% of all possible paths. These patterns were FFF, FNF, and FFN. We therefore ran three hypothesis tests to compare the number of matches of these patterns to their number of matches in a sample of 1000 degree preserving random graphs. The critical p-value is 0.00166 since we must divide the alpha value by two for a two-tailed test and then again by three for the Bonferonni correction. We summarize the results in the following table where each number is reported up to 5 significant digits:

Pattern	z-score	p-value
FFF	-320.53	0.0000
FNF	53556	0.0000
FFN	6962.8	0.0000

These z-scores are unreasonably far from zero. We will understand why this is the case in the discussion section of the paper. If we accept the results as shown we see that all the p-values are less than the critival p-value so we see that FFF is an anti-motif while FNF and FFN are motifs.

B. C. ELEGANS SHORTEST PATHS

It was further found that nine patterns of length at least three match over 1% of shortest paths. The critical p-value is 0.00056 since we must divide the alpha value by two for a two-tailed test and then again by nine for the Bonferonni correction. We summarize the results in the following table where each number is reported up to 5 significant digits:

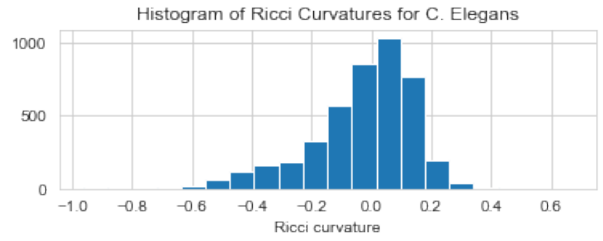
Pattern	z-score	p-value
FFF	-10.476	0.0000
FFN	-10.475	0.0000
FNF	1051.7	0.0000
NFF	-114.78	0.0000
FFFF	1136.7	0.0000
FFFN	229.72	0.0000
FFNF	1394.4	0.0000
FNFF	579.76	0.0000
FFFFF	14008	0.0000

Once again, these z-scores are unreasonably far from zero which will be addressed in the discussion section. If we take the results at face value then we see that every pattern that appears in over 1% of shortest paths in the C. Elegans network matches a significantly different number of shortest paths than in a corresponding random degree preserving network.

V. DISCUSSION

A. WHY SO MANY F'S?

By examining the different patterns that match more than 1% of paths we see that 'F' appears most frequently indicating the flat curvature edges are predominant on paths in the C. Elegans network. This is not surprising once we look at the distribution of curvature values.

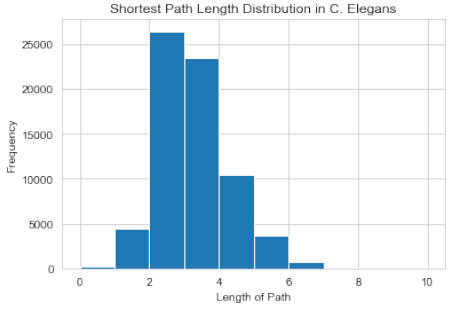


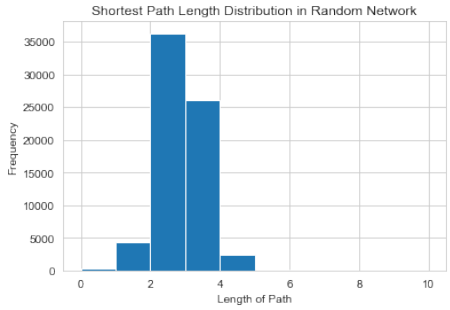
We can see that most of the curvature values fall between -0.33 and 0.33 so would be classified as F (flat). To be precise, there are 359 negative curvature edges, 4020 flat curvature edges, and 9 positive curvature edges in the C. Elegans network. In the random networks we also see that flat curvature is predominant. There are on average 238.455 negative, 4149.544 flat, and 0.001 positive curvature edges in the random networks. The predominance of flat curvature edges explains why all the patterns we examined (the ones that appear in at least 1% of paths) all contain at least two F's. The predominance of flat curvature edges also tells us that the C. Elegans network is overall a flat network. The particular lack of positive curvature edges suggests that there is not robust community/modular organization in the C. Elegans network.

B. UNDERSTANDING Z-SCORES AND LIMITATIONS

One should be skeptical when seeing such high z-scores as presented in this paper. In this subsection we explore why this occurred and what limitations such understanding places on the insights that can be drawn from the results.

An important observation is that there are patterns of length four and five in the shortest path analysis and all of them have positive z-scores. In particular FFFFF has the highest z-score meaning it is *extremely* over-represented in the C. Elegan connectome compared to a degree preserving random network. If we plot the distribution of path lengths in the C. Elegans and a representative random graph we begin to see why our z-scores take such extreme values.





There are simply more length 4 and 5 shortest paths in the C. Elegan network than in degree preserving random graphs. To be precise, the following table shows the number of length 3, 4, and 5 paths in the C. Elegans and the average across all the random graphs to 5 significant digits.

Length:	3	4	5
C. Elegans:	23476	10426	3647.0
Avg Random	25706	2222.8	48.756

We can now clearly observe that the C. Elegans network has around 4.7x the number of length four shortest paths and has around 75x the number of length five shortest paths as does an average random graph. Additionally we see that there are around 2000 less length 3 shortest paths in the C. Elegans network. Given these observations we now understand that the length 4 and 5 paths were over-represented (relative to random graphs) because there are simply many more of them. In retrospect, this tells us that the edge randomization procedure used to create the random sample shrinks the diameter of the C. Elegans graph. The extreme z-scores could perhaps have been mitigated if the randomization procedure preserved the path length distribution of the C. Elegans network. Future work would redo this analysis in such a way that the not only degree distribution but also path length distribution is preserved during randomization.

Our analysis highlights the need, in the field of connectomics, for a principled way to generate a reasonable random sample of null graphs when performing null hypothesis testing on graph properties. The current standard in the literature of randomizing the edges in such a way that preserves the degree distribution is not sufficient and can yield unreasonable results.

C. A CURVATURE PATH MOTIF

Without knowing the all-possible-paths length distributions (for C. Elegans and random graphs) it is difficult to contextualize the z-scores from subsection A of Section IV so we do not draw any insights from that analysis. We do have the shortest-path length distributions so we can say something about the results from subsection B of Section IV.

Because there are over 2000 fewer shortest-paths of length three in the C. Elegans network, we expect length three paths to be underrepresented in the C. Elegans network. However, the pattern FNF is extremely over represented in the C. Elegans network (since its z-score is very large and positive).

It is the only length three pattern over represented in the shortest path motif analysis. We can conclude at the very least that FNF is a path motif since it is over-represented when it is expected to be under-represented. A negative edge connects distinct neighborhoods so FNF being a motif means the C. Elegans has many local bridge like structures. If the pattern was PNP then we would have a bridge connecting two communities since positive curvature edges are usually found intra-communities [7]. Accordingly, we might want to call the FNF patterns 'weak bridges'. We can summarize this result by saying the C. Elegans connectome exhibits local weak bridge structures (when restricting attention to shortest paths).

VI. CONCLUSION

In this paper we used a path motif analysis to see what the curvature structure of a neuronal connectome can tell us about how network is organized for communication. It was found that the path motif analysis led to unreasonable results because the path length distribution of the C. Elegans network was not preserved in the random null sample. This motivates future work to figure out how to randomize a graph in such a way that preserves both degree and path length distributions. Then one can reconduct the curvature path analysis. The work also highlights the need for a rigorous treatment of null hypothesis testing for network properties so that there are clear guidelines of what should be preserved when generating a null sample.

REFERENCES

[1] van den Heuvel, M. P., Kahn, R. S., Goñi, J., & Sporns, O. (2012). High-cost, high-capacity backbone for global brain communication. Proceedings of the National Academy of Sciences of the United States of America, 109(28), 11372–11377. https://doi.org/10.1073/pnas.1203593109

[2] Towlson, E. K., Vértes, P. E., Ahnert, S. E., Schafer, W. R., & Bullmore, E. T. (2013). The rich club of the C. elegans neuronal connectome. The Journal of neuroscience : the official journal of the Society for Neuroscience, 33(15), 6380–6387. https://doi.org/10.1523/JNEUROSCI.3784-12.2013

[3] Fornito, Zalesky, A., & Bullmore, E. (2016). Fundamentals of brain network analysis. Elsevier.

[4] Elumalai, P., Yadav, Y., Williams, N., Saucan, E., Jost, J., & Samal, A. (2022). Graph Ricci curvatures reveal atypical functional connectivity in autism spectrum disorder. Scientific reports, 12(1), 8295. https://doi.org/10.1038/s41598-022-12171-y

[5] Weber, M., Stelzer, J., Saucan, E., Naitsat, A., Lohmann, G., & Jost, J. (2019). Curvature-based Methods for Brain Network Analysis (arXiv:1707.00180). arXiv. https://doi.org/10.48550/arXiv.1707.00180

[6] Ni, C.-C., Lin, Y.-Y., Gao, J., Gu, X. D., & Saucan, E. (2015). Ricci Curvature of the Internet Topology (arXiv:1501.04138). arXiv. http://arxiv.org/abs/1501.04138

[7] Ni, C.-C., Lin, Y.-Y., Luo, F., & Gao, J. (2019). Community Detection on Networks with Ricci Flow. Scientific Reports, 9(1), 9984. https://doi.org/10.1038/s41598-019-46380-9

[8] Ollivier, Y. (2012). A visual introduction to Riemannian curvatures and some discrete generalizations. https://doi.org/10.1090/crmp/056/08

[9] Topping, J., Di Giovanni, F., Chamberlain, B. P., Dong, X., & Bronstein, M. M. (2022). Understanding over-squashing and bottlenecks on graphs via curvature (arXiv:2111.14522). arXiv. https://doi.org/10.48550/arXiv.2111.14522

[10] Networks. (n.d.). Retrieved July 29, 2022, from https://neurodata.io/project/connectomes/

[11] Ni, C.-C. (2022). GraphRicciCurvature [Python]. https://github.com/saibalmars/GraphRicciCurvature (Original work published 2016)

[12] LIKarpa/DIMACS_REU: Research during my time at the DIMACS_REU. (n.d.). Retrieved July 29, 2022, from https://github.com/LIKarpa/DIMACS_REU