HYPERBOLIC EMBEDDING MODEL FOR A CLASS OF MICRORNA-DISEASE NETWORKS

Radu Angelescu[1] and Radu Dobrescu[2]

The main goal of this paper is to study the embedding of networks in a form that can be used in differential programming. We prove that certain networks have better embeddings in an appropriate metric space and provide a numerically stable embedding solution. We generate embeddings for the disease-microRNA network.

Keywords: network embedding, hyperbolic space, Poincare Disk, mathematical models, differential programming, microRNA

MSC2010: MSC2010: 68T99, 68U25, 92B20

1. Introduction

Network embedding is a method that generates low-dimensional representations of vertexes in networks, aiming to capture and preserve the network structure. Many real-world applications need to mine information from networks (recommendation systems [10], biological networks [34] [23], narrative network analysis [20]).

For large networks, such as those with thousands of nodes the traditional network representation poses several challenges for processing and analysis: high computational complexity, low parallelizability, inapplicability of differentiable programming methods. Network embeddings are used to solve these problems.

To be able to provide insight, the embedding is required to preserve the network structure. This is not trivial because network structures, are usually highly non-linear [13] and complex [11].

[1] Phd, Department of Automatic Control and Industrial Informatics, Faculty of Auto matic Control and Computer Science, University "Politehnica" of Bucharest, Romania, email: raduangelescu@gmail.com

[2] Professor, Department of Automatic Control and Industrial Informatics, Faculty of Auto matic Control and Computer Science, University "Politehnica" of Bucharest, Romania, email: rd_dobrescu@yahoo.com

We provide an embedding method in which we use random walks and Riemann optimization [17] in the Poincare ball [4]. We then convert a microRNA-disease network to a proper input form for differentiable programming algorithms. We evaluate by doing a network reconstruction and calculating its MAP. The results are presented in section 6.

2. Preliminaries

We denote a network G(V,E) where V is the vertex set and E is the edge set.

Definition 2.1. *As in [25, 3, 21] we define Mean Average Precision (MAP): As the average precision over all nodes:* $\frac{\sum_i AP(i)}{|V|}$ *where*

$AP(i) = \frac{\sum_k precision@k(i) \prod \{E_{pred_i}(k) \in E_{gt_i}\}}{|\{k : E_{pred_i}(k) \in E_{gt_i}\}|}$ *and*

$precision@k(i) = \frac{|E_{pred_i(1:k)} \cap E_{gt_i}|}{k}$ *where* E_{pred} *and* E_{gt} *are the predicted and ground truth edges respectively.*

Definition 2.2. *As in [22] we define a network embedding as a mapping of the network data into a low-dimensional latent space, where each vertex is represented as a low-dimensional vector and the network computing can be directly-realized.*

2.1. Poincare ball

In geometry, the Poincare disk model also called the conformal disk model hyperbolic geometry in which the points of the geometry are inside the unit disk, and the straight lines consist of all segments of circles contained within that disk that are orthogonal to the boundary of the disk, plus all diameters of the disk.

The Poincare ball model is the Poincare disk model in the n-dimensional unit ball $\mathfrak{B}^d = \{x \in \mathbb{R}^d \mid \|x\| < 1\}$ with metric $g_x = \left(\frac{2}{1-||x||^2}\right)^2 g^E$ where $x \in \mathfrak{B}^d$. We have the distance between two points: $arcosh\left(1 + 2\frac{||u-v||^2}{(1-||u||^2)(1-||v||^2)}\right)$

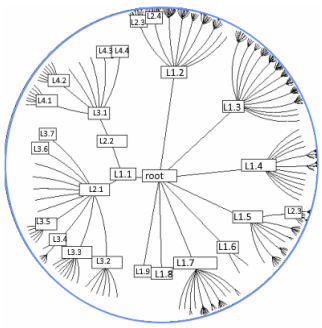


Fig.1. Tree embedding in Poincare Disk

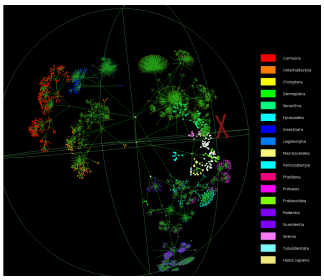


Fig. 2. Phylogenetic trees in three dimensional Poincare ball generated with the phylo3D[8] program

From Fig. 1 and [2] we can see that a tree has a natural representation in the Poincare disk. We may also note the existing visualization techniques in hyperbolic space [7] used for phylogenetic trees in Fig. 2.

3. **Building the network**

We downloaded and parsed the latest data from MESH [35] into our own pair representation of the disease network. The disease network is composed of 22 main disease categories, that have cross links (because diseases may be in multiple categories) so we can assume with no approximation that by combining these we are dealing with a complex network that is not a tree.

We downloaded data from mirWayDB [33] containing 66 diseases with experimentally validated associations between microRNAs and diseases and also the biological pathways in which they act as controllers. We construct a relationship network linking microRNAs through common pathways and diseases. The resulting network contains information about known microRNA-disease relationships.

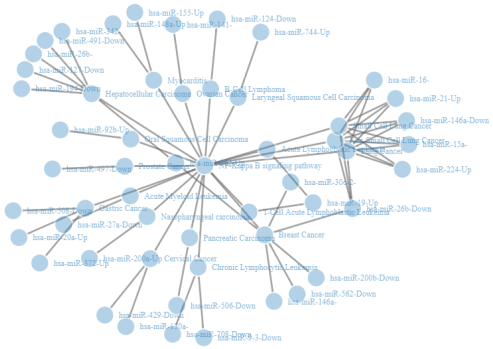


Fig. 3. Example of an subnetwork included in our final network

We determined the relationship between the diseases in MESH and mirWayDB and created a bigger network (figure 3) where the other two networks mentioned above are linked through common diseases.

4. Network Analysis

Trees can be embedded with arbitrarily low distortion in hyperbolic space [26],[14] but they cannot be embedded into euclidean space with arbitrarily low distortion for any number of dimensions [6]. We are dealing with a complex network but we can compare its hiperbolicity to that of a tree to gain insight.

Definition 4.1. *As defined in [1] a metric space (V, d) is $\delta-$ hyperbolic for a constant $\delta \geq 0$ if for every four points $w, x, y, z \in V$ that are ordered such that $d(w,x)+d(y,z)d(w,y)+d(x,z)d(w,z)+d(x,y)d(w,z)+d(x,y)d(w,y)+d(x,z)+\delta$ holds.*

In our network G we denote:
$a, b, c, d \in V$ (recall V as the set of nodes of the network G).

$$\begin{aligned} S_1(a,b,c,d) &= d_G(a,b)+d_G(d,c) \\ S_2(a,b,c,d) &= d_G(a,c)+d_G(b,d) \\ S_3(a,b,c,d) &= d_G(a,d)+d_G(b,c) \end{aligned} \tag{1}$$

Where d_G is the network distance (defined as the shortest path in the network)

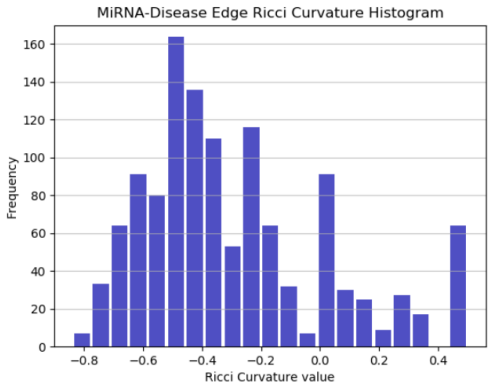
We calculate the hyperbolicity of our network G with the following formula:

$$\delta(G) = \frac{1}{2} max(|max(S1(a,b,c,d), S_2(a,b,c,d)) - \\ - max(S_2(a,b,c,d), S_3(a,b,c,d))|) \tag{2}$$

We found that for $\delta(G) = 2$ our network is hyperbolic.

We know that a tree is hyperbolic for $\delta = 0$, an NXN grid for $\delta = N-1$, and a N length cycles is hyperbolic for $\delta = \frac{N}{4}$.

As in [19] we also analyzed the network by calculating the Ricci curvature of the edges of our network.



For comparison, on a square lattice network each edge has a zero Ricci curvature. For a tree the Ricci curvature is negative for all edges except those

connecting leaves. On a complete network with at least two vertices the Ricci curvature is all positive. The average ricci curvature as defined on edges is -0.29 but it is also evident from the histogram that most edges have negative or 0 curvature. This indicates that doing a hyperbolic approximation should be better than an euclidean one, as negative curvature indicates that a negatively curved space is better than the 0 or positive curved space.

Unfortunately we can also see some positive curvature edges that will make our embedding require more than 2 dimensions. This was not a surprise as the network is not a tree.

5. **Finding the hyperbolic embeddings**

We generate a cost matrix FS on network $G(V, E)$ such that $FS[i, j] = d_G(i, j)$ where $i, j \in \{0..\mathbf{card}(V)\}$. Where $d_G(i, j)$ is the shortest path [12] between node i and node j. Our problem becomes approximating points for the network nodes in $\mathfrak{B}^d$ unit ball, so that by using the Poincare ball metric we may reconstruct the graph. To achieve this hyperbolic embedding space for our network we use the RSGD manifold optimization [15, 37, 16] on an adaptation of the skip-gram embedding loss function [9]

So we need to optimize the loss:

$$\tau(\phi) = \sum_{(u,v)\in D} log \frac{e^{-d(u,v)}}{\sum_{v'\in N(u)} e^{-d(u,v')}}$$

Where we have u, v point embeddings representing nodes that are directly linked in the network and v' nodes from the negative set (in our case, not linked with u).

For optimisations we actually random sample for the negative space, judging that our data is not fully connected but rather weakly connected, we picked a negative sample of 20. Intuitively the loss would go up when we have big distances between linked nodes or small distances between unlinked nodes. We then apply the basic formula for Riemann stochastic gradient optimization [16] on our Poincare disk (as our space is a Riemann manifold).

$$\theta' \leftarrow argmin_\phi \tau(\theta), \theta = \{\theta_i \mid \|\theta_i\| < 1\}$$

$$\theta_{t+1} \leftarrow \theta_t - \mu_t \nabla_R \tau(\theta_t)$$

Detailing the update procedure:

$$\theta_{t+1} = proj\left(\theta_t - \mu_t \frac{(1- \mid\mid \theta_t \mid\mid^2)^2}{4} \nabla E\right)$$

Where ∇E is the euclidean space gradient of the θ function and $\mu_t \in (0, 1]$ is the learning rate.

$$proj(\theta) = \begin{cases} \frac{\theta}{||\theta||} - \epsilon & \mid\mid \theta \mid\mid \geq 1 \\ \theta & otherwise \end{cases}$$

$$\nabla E = \frac{\partial \tau(\phi)}{\partial d(\phi, x)} \frac{\partial d(\phi, x)}{\partial \phi}$$

Furthermore for the final microRNA functional similarity network we create a new weighted fully connected network where the nodes are all microRNAs and edges represent the functional distance between them. The functional distance is calculated by finding the shortest path between microRNAs. This network is pruned with the average maximum distance and we generate network chains with weighted random walks[36].

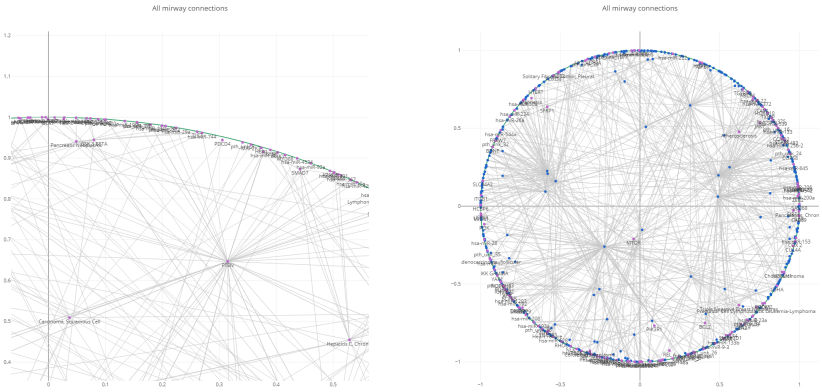
We generate 60 chains per node so we sample the most important neighbours and then feed the resulting links in to an "online" version of Poincare embedding generation. The negative edges are no longer sampled randomly (because we are dealing with a fully connected network) but picked from the pruned links. The resulting microRNA embeddings converge quicker, and the algorithm is now "online" (may be updated after generation) version of the previous one.

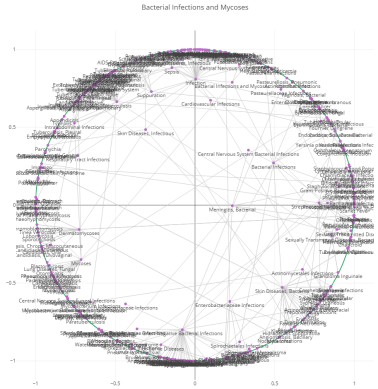
6. **Results and discussions**

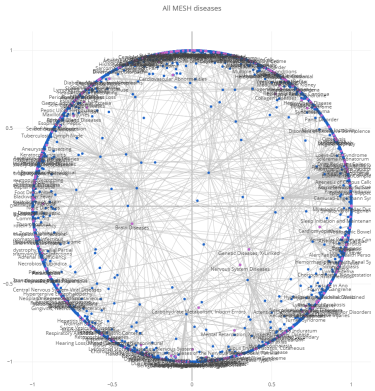
To have a base comparison we did an euclidean matrix factorization (single value decomposition) method first [18] on the big network. The results were very bad:

Vector embedding size	2	4	16	32	64	128
mAP	0.097	0.098	0.1	0.11	0.098	0.098

We first ran the experiments on different vector sizes, and for different epochs. We also did embeddings for trees, subtrees, the disease network, the microRNA network and finally the disease-network. The attached images illustrate our resulting embeddings in the Poincare disk (two dimensions) with the original links superimposed, for intuitive clarity over the obtained results.







The above images are plots of the embeddings in the Poincare disk, these have the following mAP values

network	mAP
all diseases (MESH)	0.55
bacterial infections and mycoses (sub DAG)	0.75
microRNA-disease from mirway	0.56
diseases (MESH) + microRNA-disease (mirway)	0.45

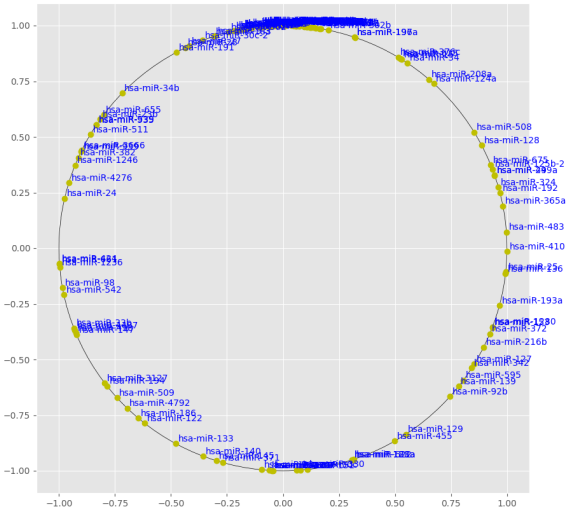
Embedding size 2

The above map values from the table implies that the resulting images are just to check our intuition but for better embeddings we need a bigger vector size. The best results we obtained were on the Poincare disk with vector size 16. These cannot be represented in pictures but you may check the attached table for microRNA to microRNA resulting distances.

network	mAP
all diseases (MESH)	0.7
bacterial infections and mycoses (sub DAG)	0.85
microRNA-disease from mirway	0.6
diseases (MESH) + microRNA-disease (mirway)	0.5

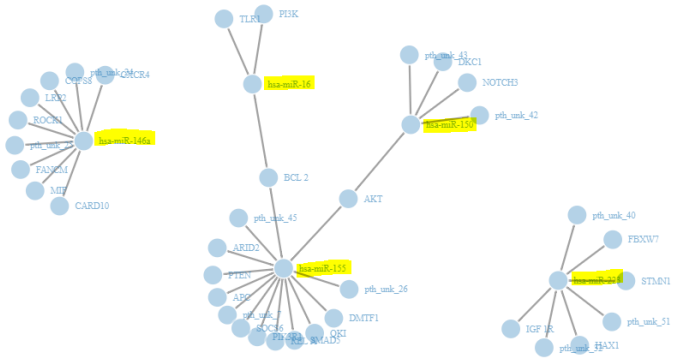
Embedding size 16

For a better understanding of the resulting 2d embeddings we did the interactive plots [29, 30, 31, 32] The disease-mirway network did not achieve a good representation (the mAP was very small) so we used the above described improved the algorithm.



Analyzing the results of small subnetwork embedding and retracing a network of neighbours based on distance cutoff we may observe particular results that correspond to biological reality:

We observe BCL2 and AKT, two pathways that are common to hsa-miR-150 and hsa-miR-155, both these connections are supported by paper [24, 5] , we also notice that these connections were not present in the investigating database but were inferred because of the small embedding distance. As future directions we may turn to neural hyperbolic networks for improving predictions but also add more information about microRNA structure itself to encompass the biological reality of microRNA - microRNA networks through common messenger RNA targets. A new improvement to the embeddings themselves could be done using the hyperboloid model [28] (Minkowski, Lorentz) that would result in an easier optimization problem because of the Minkowski bi-linear form



$$B((x_0 \dots x_n), (y_0 \dots y_n)) = x_0 y_0 - \sum_{i=1}^{n} x_i y_i$$

$$d(u, v) = arcosh(B(u, v))$$

that makes the hyperbolic distance easier and more numerically stable to compute.

7. Conclusions

The results of our method show that hiperbolic embeddings for microRNA networks generated from data are a good way of exploring and predicting underlying connections. These will provide a good fit for functional clustering of microRNAs but also a basis for disease prediction. The resulting distances from the more computationally expensive 16 dimensional hyperbolic embeddings are a good way of embedding our data into a form that is fit for differential programming. We may use these for disease-microRNA exploration (just calculating the hyperbolic distance between a disease and a microRNA) or for microRNA-microRNA distance calculation (for a sort of functional classification). The resulting embeddings could provide a good starting point for a hiperbolic machine learning algorithm for disease-prediction. The tool provided in this article may be used to exploit the already existing big databases of experiments and combine them with new experiments so to better explain certain phenomenon.

**Edit in proof: Similar efforts have been made in using Poincare embeddings to clarify phenotypes from hierarchical medical concepts. [27]*

References

[1] M Gromov. *Hyperbolic groups. In: Essays In Group Theory.* Ed. by New York Springer. 1987, pp. 75–263.

[2] John Lamping, Ramana Rao, and Peter Pirolli. "A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '95. Denver, Colorado, USA: ACM Press/Addison-Wesley Publishing Co., 1995, pp. 401–408. ISBN: 0-201-84705-1. DOI: 10.1145/223904.223956. URL: http://dx.doi.org/10.1145/223904.223956.

[3] Tefko Saracevic. "Evaluation of Evaluation in Information Retrieval". In: *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '95. Seattle, Washington, USA: ACM, 1995, pp. 138–146. ISBN: 0-89791-714-6. DOI: 10.1145/215206.215351. URL: http://doi.acm.org/10.1145/215206.215351.

[4] James Anderson. *Hyperbolic geometry*. Ed. by Springer. 1999, pp. 95–110.

[5] Hotchkiss RS1 Swanson PE Knudson CM Chang KC Cobb JP Osborne DF Zollner KM Buchman TG Korsmeyer SJ Karl IE. "Overexpression of Bcl-2 in transgenic mice decreases apoptosis and improves survival in sepsis". In: *Journal of Immunology* 31 (Apr. 1999).

[6] Jiří Matoušek. "On embedding trees into uniformly convex Banach spaces". In: *Israel Journal of Mathematics* 114 (Dec. 1999), pp. 221–237. DOI: 10.1007/BF02785579.

[7] Timothy Hughes, Young Hyun, and David Liberles. "Visualising very large phylogenetic trees in three dimensional hyperbolic space". In: *BMC bioinformatics* 5 (May 2004), p. 48. DOI: 10.1186/1471-2105-5-48.

[8] Timothy Hughes, Young Hyun, and David A. Liberles. "Visualising very large phylogenetic trees in three dimensional hyperbolic space". In: *BMC Bioinformatics* 5.1 (Apr. 2004), p. 48. ISSN: 1471-2105. DOI: 10.1186/1471-2105-5-48. URL: https://doi.org/10.1186/1471-2105-5-48.

[9] Cheng W. Greaves C. Warren M. "From n-gram to skipgram to concgram." In: International Journal of Corpus Linguistics 11 (2006), pp. 411–433.

[10] Massimiliano Zanin et al. "Complex networks in recommendation systems". In: (Jan. 2008).

[11] Blake Shaw and Tony Jebara. "Structure Preserving Embedding". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09. Montreal, Quebec, Canada: ACM, 2009, pp. 937–944. ISBN: 978-1-60558-516-1. DOI: 10.1145/1553374.1553494. URL: http://doi.acm.org/10.1145/1553374.1553494.

[12] DongKai Fan and Ping Shi. "Improvement of Dijkstra's algorithm and its application in route planning". In: vol. 13. Aug. 2010, pp. 1901–1904. DOI: 10.1109/FSKD.2010.5569452.

[13] Dijun Luo et al. "Cauchy Graph Embedding". In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11. Bellevue, Washington, USA: Omnipress, 2011, pp. 553–560. ISBN: 978-1-4503-0619-5. URL: http://dl.acm.org/citation.cfm?id=3104482.3104552.

[14] Rik Sarkar. "Low Distortion Delaunay Embedding of Trees in Hyperbolic Plane". In: *Graph Drawing*. Ed. by Marc van Kreveld and Bettina Speckmann. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 355–366. ISBN: 978-3-642-25878-7.

[15] A. Sossinsky. *The Poincare disk model of hyperbolic geometry*. Vol. The Student Mathematical Library. 2012, pp. 125–142.

[16] S. Bonnabel. "Stochastic Gradient Descent on Riemannian Manifolds." In: IEEE Trans. Automat. Contr. 58 (2013), pp. 2217–2229.

[17] Steven Smith. "Optimization Techniques on Riemannian Manifolds". In: *Hamiltonian and Gradient Flows, Algorithms and Control* 158 (July 2014).

[18] Shaosheng Cao, Wei Lu, and Qiongkai Xu. "GraRep: Learning Graph Representations with Global Structural Information". In: *Proceedings of the 24th ACM International*

on Conference on Information and Knowledge Management. CIKM '15. Melbourne, Australia: ACM, 2015, pp. 891–900. ISBN: 978-1-4503-3794-6. DOI: 10.1145/2806416.2806512. URL: http://doi.acm.org/10.1145/2806416.2806512.

[19] C. Ni et al. *Ricci curvature of the Internet topology*. Apr. 2015, pp. 2758–2766. DOI: 10.1109/INFOCOM.2015.7218668.

[20] Andrzej Jarynowski and Stephanie Boland. "Social Networks Analysis in Discovering the Narrative Structure of Literary Fiction". In: *CoRR* abs/1608.05982 (2016). arXiv: 1608.05982. URL: http://arxiv.org/abs/1608.05982.

[21] Daixin Wang, Peng Cui, and Wenwu Zhu. "Structural Deep Network Embedding". In: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, pp. 1225–1234. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939753. URL: http://doi.acm.org/10.1145/2939672.2939753.

[22] Daixin Wang, Peng Cui, and Wenwu Zhu. "Structural Deep Network Embedding". In: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, pp. 1225–1234. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939753. URL: http://doi.acm.org/10.1145/2939672.2939753.

[23] Emine Ezel Cilek, Hakime Ozturk, and Bala Gur Dedeoglu. "Construction of miRNA-miRNA networks revealing the complexity of miRNA-mediated mechanisms in trastuzumab treated breast cancer cell lines". In: *PLOS ONE* 12.10 (Oct. 2017), pp. 1–21. DOI: 10.1371/journal.pone.0185558. URL: https://doi.org/10.1371/journal.pone.0185558.

[24] Fatemeh Fattahi et al. "Complement-induced activation of MAPKs and Akt during sepsis: role in cardiac dysfunction". In: *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 31 (June 2017). DOI: 10.1096/fj.201700140R.

[25] Palash Goyal et al. "Embedding Networks with Edge Attributes". In: *Proceedings of the 29th on Hypertext and Social Media*. HT '18. Baltimore, MD, USA: ACM, 2018, pp. 38–42. ISBN: 978-1-4503-5427-1. DOI: 10.1145/3209542.3209571. URL: http://doi.acm.org/10.1145/3209542.3209571.

[26] MATTHIAS HAMANN. "On the tree-likeness of hyperbolic spaces". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 164.2 (2018), pp. 345–361. DOI: 10.1017/S0305004117000238.

[27] Brett K. Beaulieu-Jones, Isaac Kohane, and Andrew Beam. *Learning Contextual Hierarchical Structure of Medical Concepts with Poincairé Embeddings to Clarify Phenotypes*. Nov. 2018.

[28] Maximilian Nickel and Douwe Kiela. *Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry*. June 2018.

[29] Angelescu R. *miRNA plots 1*. http://raduangelescu.com/plots/just_mirnas_model_vec_sz2_ep1500.html. [Online;] 2018.

[30] Angelescu R. *miRNA plots 2*. http://raduangelescu.com/plots/Bacterial_Infections_and_Mycosis_model_vec_sz2_ep1500.html. [Online;] 2018.

[31] Angelescu R. *miRNA plots 3*. http://raduangelescu.com/plots/All_mirnas_vec_sz2_ep5000.html. [Online;] 2018.

[32] Angelescu R. *miRNA plots 4*. http://raduangelescu.com/plots/All_mesh_diseases.html. [Online;] 2018.

[33] Saha Nishant Chakravorty Sankha Subhra Das Pritam. "miRwayDB: a database for experimentally validated microRNA-pathway associations in pathophysiological conditions". In: *Database (2018) Vol. 2018* 2018 (2018).

[34] Carlo Cosentino. "Modeling Biological Networks". In: (Mar. 2019).

[35] U.S. National Library of Medicine. *MESH website*. 2019. URL: https://meshb.nlm.nih.gov/search.

[36] José L. Mateos A.P. Riascos. "Random walks on weighted networks: Exploring local and non-local navigation strategies". In: ().

[37] J. Deisenroth Chamberlain B. P. Clough. "Neural Embeddings of Graphs in Hyperbolic Space." In: arXiv ().