

# Neural Embeddings of Graphs in Hyperbolic Space

1  
2  
3  
4  
5  
6  
7  
8  
9 Benjamin Paul Chamberlain  
Department of Computing  
Imperial College London  
bpc14@imperial.ac.uk

James R Clough  
Centre for Complexity Science  
Imperial College London  
jrc309@imperial.ac.uk

Marc Peter Deisenroth  
Department of Computing  
Imperial College London  
m.deisenroth@imperial.ac.uk

## ABSTRACT

10 Neural embeddings have been used with great success in Natural  
11 Language Processing (NLP). They provide compact representations  
12 that encapsulate word similarity and attain state-of-the-art per-  
13 formance in a range of linguistic tasks. The success of neural  
14 embeddings has prompted significant amounts of research into  
15 applications in domains other than language. One such domain is  
16 graph-structured data, where embeddings of vertices can be learned  
17 that encapsulate vertex similarity and improve performance on  
18 tasks including edge prediction and vertex labelling. For both NLP  
19 and graph based tasks, embeddings have been learned in high-  
20 dimensional Euclidean spaces. However, recent work has shown  
21 that the appropriate isometric space for embedding complex net-  
22 works is not the flat Euclidean space, but negatively curved, hyper-  
23 bolic space. We present a new concept that exploits these recent  
24 insights and propose learning neural embeddings of graphs in hy-  
25 perbolic space. We provide experimental evidence that embedding  
26 graphs in their *natural* geometry significantly improves perfor-  
27 mance on downstream tasks for several real-world public datasets.

## KEYWORDS

neural networks, graph embeddings, complex networks, geometry

## 1 INTRODUCTION

35 Embedding (or vector space) methods find a lower-dimensional  
36 continuous space in which to represent high-dimensional complex  
37 data [4, 26]. The distance between objects in the lower-dimensional  
38 space gives a measure of their similarity. This is usually achieved  
39 by first postulating a low-dimensional vector space and then optimising  
40 an objective function of the vectors in that space. Vector  
41 space representations provide three principle benefits over sparse  
42 schemes: (1) They encapsulate similarity, (2) they are compact, (3)  
43 they perform better as inputs to machine learning models [27]. This  
44 is true of graph structured data where the native data format is  
45 the adjacency matrix, a typically large, sparse matrix of connection  
weights.

47 Neural embedding models are a flavour of embedding scheme  
48 where the vector space corresponds to a subset of the network  
49 weights, which are learned through backpropagation. Neural em-  
50 bedding models have been shown to improve performance in a  
51 large number of downstream tasks across multiple domains. These  
52 include word analogies [18, 20], machine translation [28], docu-  
53 ment comparison [16], missing edge prediction [12], vertex at-  
54 tribution [24], product recommendations [2, 10], customer value  
55 prediction [6, 14] and item categorisation [3]. In all cases the em-  
56 beddings are learned without labels (unsupervised) from a sequence  
57 of entities.

To the best of our knowledge, all previous work on neural embedding models either explicitly or implicitly (by using the Euclidean dot product) assumes that the vector space is Euclidean. Recent work from the field of complex networks has found that many interesting networks, such as the Internet [5] or academic citations [7, 8] can be well described by a framework with an underlying non-Euclidean hyperbolic geometry. Hyperbolic geometry provides a continuous analogue of tree-like graphs, and even infinite trees have nearly isometric embeddings in hyperbolic space [11]. Additionally, the defining features of complex networks, such as power-law degree distributions, strong clustering and hierarchical community structure, emerge naturally when random graphs are embedded in hyperbolic space [15].

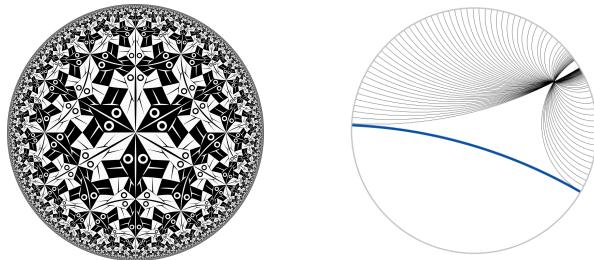
The starting point for our model is the celebrated word2vec Skipgram architecture, which is shown in Figure 3 [18, 19]. Skipgram is a shallow neural network with three layers: (1) An input projection layer that maps from a one-hot-encoded to a distributed representation, (2) a hidden layer, and (3) an output softmax layer. The network is necessarily simple for tractability as there are a very large number of output states (every word in a language). Skipgram is trained on a sequence of words that is decomposed into (input word, context word)-pairs. The model employs two separate vector representations, one for the input words and another for the context words, with the input representation comprising the learned embedding. The word pairs are generated by taking a sequence of words and running a sliding window (the context) over them. As an example the word sequence “chance favours the prepared mind” with a context window of size three would generate the following training data: (chance, favours), (chance, the), (favours, chance), ...}. Words are initially randomly allocated to vectors within the two vector spaces. Then, for each training pair, the vector representations of the observed input and context words are pushed towards each other and away from all other words (see Figure 2).

The concept can be extended from words to network structured data using random walks to create sequences of vertices. The vertices are then treated exactly analogously to words in the NLP formulation. This was originally proposed as DeepWalk [24]. Extensions varying the nature of the random walks have been explored in LINE [29] and Node2vec [12].

*Contribution.* In this paper, we introduce the new concept of neural embeddings in hyperbolic space. We formulate backpropagation in hyperbolic space and show that using the natural geometry of complex networks improves performance in vertex classification tasks across multiple networks.

## 2 HYPERBOLIC GEOMETRY

Hyperbolic geometry emerges from relaxing Euclid’s fifth postulate (the parallel postulate) of geometry. In hyperbolic space there is not



(a) “Circle Limit 1” by M.C. Escher illustrates the Poincaré disc model of hyperbolic space. Each tile is of constant area in hyperbolic space, but vanishes in Euclidean space at the boundary.

(b) A set of straight lines in the Poincaré disk that all pass through a given point and are all parallel to the blue (thicker) line.

**Figure 1: Illustrations of properties of hyperbolic space. a Tiles of constant area b Parallel lines.**

just one, but an infinite number of parallel lines that pass through a single point. This is illustrated in Figure 1b where every line is parallel to the bold, blue line and all pass through the same point. Hyperbolic space is one of only three types of isotropic spaces that can be defined entirely by their curvature. The most familiar is Euclidean, which is flat, having zero curvature. Space with uniform positive curvature has an elliptic geometry (e.g. the surface of a sphere), and space with uniform negative curvature is called hyperbolic, which is analogous to a saddle-like surface. As, unlike Euclidean space, in hyperbolic space even infinite trees have nearly isometric embeddings, it has been successfully used to model complex networks with hierarchical structure, power-law degree distributions and high clustering [15].

One of the defining characteristics of hyperbolic space is that it is in some sense *larger* than the more familiar Euclidean space; the area of a circle or volume of a sphere grows exponentially with its radius, rather than polynomially. This suggests that low-dimensional hyperbolic spaces may provide effective representations of data in ways that low-dimensional Euclidean spaces cannot. However this makes hyperbolic space hard to visualise as even the 2D hyperbolic plane can not be isometrically embedded into Euclidean space of any dimension, (unlike elliptic geometry where a 2-sphere can be embedded into 3D Euclidean space). For this reason there are many different ways of representing hyperbolic space, with each representation conserving some geometric properties, but distorting others. In the remainder of the paper we use the Poincaré disk model of hyperbolic space.

## 2.1 Poincaré Disk Model

The Poincaré disk models two-dimensional hyperbolic space where the infinite plane is represented as a unit disk. We work with the two-dimensional disk, but it is easily generalised to the  $d$ -dimensional Poincaré ball, where hyperbolic space is represented as a unit  $d$ -ball.

In this model hyperbolic distances grow exponentially towards the edge of the disk. The circle’s boundary represents infinitely distant points as the infinite hyperbolic plane is squashed inside the finite disk. This property is illustrated in Figure 1a where each tile is of constant area in hyperbolic space, but the tiles rapidly shrink to zero area in Euclidean space. Although volumes and distances are warped, the Poincaré disk model is **conformal**, meaning that Euclidean and hyperbolic angles between lines are equal. Straight lines in hyperbolic space intersect the boundary of the disk orthogonally and appear either as diameters of the disk, or arcs of a circle. Figure 1b shows a collection of straight hyperbolic lines in the Poincaré disk. Just as in spherical geometry, the shortest path from one place to another is a straight line, but appears as a curve on a flat map. Similarly, these straight lines show the shortest path (in terms of distance in the underlying hyperbolic space) from one point on the disk to another, but they appear curved. This is because it is quicker to move close to the centre of the disk, where distances are shorter, than nearer the edge. In our proposed approach, we will exploit both the conformal property and the circular symmetry of the Poincaré disk.

Overall, the geometric intuition motivating our approach is that vertices embedded near the middle of the disk can have more close neighbours than they could in Euclidean space, whilst vertices nearer the edge of the disk can still be very far from each other.

## 2.2 Inner Product, Angles, and Distances

The mathematics is considerably simplified if we exploit the symmetries of the model and describe points in the Poincaré disk using polar coordinates,  $x = (r_e, \theta)$ , with  $r_e \in [0, 1]$  and  $\theta \in [0, 2\pi)$ . To define similarities and distances, we require an inner product. In the Poincaré disk, the *inner product* of two vectors  $x = (r_x, \theta_x)$  and  $y = (r_y, \theta_y)$  is given by

$$\langle x, y \rangle = \|x\| \|y\| \cos(\theta_x - \theta_y) \quad (1)$$

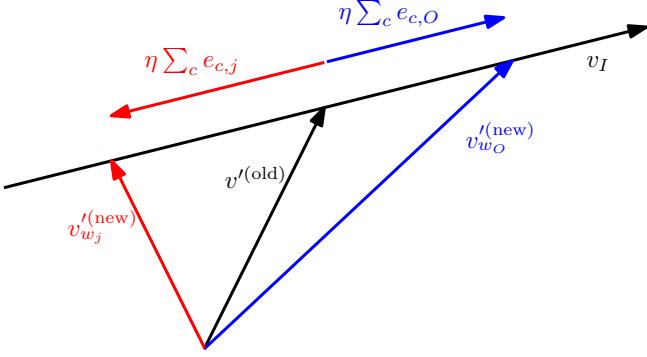
$$= 4 \operatorname{arctanh} r_x \operatorname{arctanh} r_y \cos(\theta_x - \theta_y) \quad (2)$$

The *distance* of  $x = (r_e, \theta)$  from the origin of the hyperbolic coordinate system is given by  $r_h = 2 \operatorname{arctanh} r_e$  and the circumference of a circle of hyperbolic radius  $R$  is  $C = 2\pi \sinh R$ .

## 3 NEURAL EMBEDDING IN HYPERBOLIC SPACE

We adopt the original notation of [18] whereby the input vertex is  $w_I$  and the output is  $w_O$ . Their corresponding vector representations are  $v_{w_I}$  and  $v'_{w_O}$ , which are elements of the two vector spaces shown in Figure 3,  $\mathbf{W}$  and  $\mathbf{W}'$  respectively. Skipgram has a geometric interpretation, which we visualise in Figure 2 for vectors in  $\mathbf{W}'$ . Updates to  $v'_{w_j}$  are performed by simply adding (if  $w_j$  is the observed output vertex) or subtracting (otherwise) an error-weighted portion of the input vector. Similar, though slightly more complicated, update rules apply to the vectors in  $\mathbf{W}$ . Given this interpretation, it is natural to look for alternative geometries in which to perform these updates.

To embed a graph in hyperbolic space we replace Skipgram’s two Euclidean vector spaces ( $\mathbf{W}$  and  $\mathbf{W}'$  in Figure 3) with two Poincaré disks. We learn embeddings by optimising an objective function that predicts output/context vertices from an input vertex,



**Figure 2: Geometric interpretation of the update equations in the Skipgram model.** The vector representation of the output vertex  $v'_{w_O}^{(\text{new})}$  is moved closer (blue) to the vector representation of the input vertex  $v_I$ , while all other vectors  $v'_{w_j}^{(\text{new})}$  move further away (red). The magnitude of the change is proportional to the prediction error.

but we replace the Euclidean dot products used in Skipgram with hyperbolic inner products. A softmax function is used for the conditional predictive distribution

$$p(w_O|w_I) = \frac{\exp(\langle v'_{w_O}, v_{w_I} \rangle)}{\sum_{i=1}^V \exp(\langle v'_{w_i}, v_{w_I} \rangle)}, \quad (3)$$

where  $v_{w_i}$  is the vector representation of the  $i^{th}$  vertex, primed indicates members of the output vector space (See Figure 3) and  $\langle \cdot, \cdot \rangle$  is the hyperbolic inner product. Directly optimising (3) is computationally demanding as the sum in the denominator extends over every vertex in the graph. Two commonly used techniques to make word2vec more efficient are (a) replacing the softmax with a hierarchical softmax [18, 21] and (b) negative sampling [20, 22]. We use negative sampling as it is faster.

### 3.1 Negative Sampling

Negative sampling is a form of Noise Contrastive Estimation (NCE) [13]. NCE is an estimation technique that is based on the assumption that a good model should be able to separate signal from noise using only logistic regression.

As we only care about generating good embeddings, the objective function does not need to produce a well-specified probability distribution. The negative log likelihood using negative sampling is

$$E = -\log \sigma(\langle v'_{w_O}, v_{w_I} \rangle) - \sum_{w_j \in W_{\text{neg}}} \log \sigma(-\langle v'_{w_j}, v_{w_I} \rangle) \quad (4)$$

$$= -\log \sigma(u_O) - \sum_{j=1}^K \mathbb{E}_{w_j \sim P_n} [\log \sigma(-u_j)] \quad (5)$$

where  $v_{w_I}$ ,  $v'_{w_O}$  are the vector representation of the input and output vertices,  $u_j = \langle v'_{w_j}, v_{w_I} \rangle$ ,  $W_{\text{neg}}$  is a set of samples drawn from the noise distribution,  $K$  is the number of samples and  $\sigma$  is the sigmoid function. The first term represents the observed data and

the second term the negative samples. To draw  $W_{\text{neg}}$ , we specify the noise distribution  $P_n$  to be unigrams raised to  $\frac{3}{4}$  as in [18].

### 3.2 Model Learning

We learn the model using backpropagation. To perform backpropagation it is easiest to work in natural hyperbolic co-ordinates on the disk and map back to Euclidean co-ordinates only at the end. In natural co-ordinates  $r \in (0, \infty)$ ,  $\theta \in (0, 2\pi]$  and  $u_j = r_j r_I \cos(\theta_I - \theta_j)$ . The major drawback of this co-ordinate system is that it introduces a singularity at the origin. To address the complexities that result from radii that are less than or equal to zero, we initialise all vectors to be in a patch of space that is small relative to its distance from the origin.

The gradient of the negative log-likelihood in (5) w.r.t.  $u_j$  is given by

$$\frac{\partial E}{\partial u_j} = \begin{cases} \sigma(u_j) - 1, & \text{if } w_j = w_O \\ \sigma(u_j), & \text{if } w_j = W_{\text{neg}} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

Taking the derivatives w.r.t. the components of vectors in  $\mathbf{W}'$  (in natural polar hyperbolic co-ordinates) yields

$$\frac{\partial E}{\partial (r'_j)_k} = \frac{\partial E}{\partial u_j} \frac{\partial u_j}{\partial (r'_j)_k} = \frac{\partial E}{\partial u_j} r_I \cos(\theta_I - \theta'_j) \quad (7)$$

$$\frac{\partial E}{\partial (\theta'_j)_k} = \frac{\partial E}{\partial u_j} r'_j r_I \sin(\theta_I - \theta'_j). \quad (8)$$

The Jacobian is then

$$\nabla_{\mathbf{r}} E = \frac{\partial E}{\partial r} \hat{\mathbf{f}} + \frac{1}{\sinh r} \frac{\partial E}{\partial \theta} \hat{\theta}, \quad (9)$$

which leads to

$$r'_j \text{new} = \begin{cases} r'_j \text{old} - \eta \epsilon_j r_I \cos(\theta_I - \theta'_j), & \text{if } w_j \in w_O \cup W_{\text{neg}} \\ r'_j \text{old}, & \text{otherwise} \end{cases} \quad (10)$$

$$\theta'_j \text{new} = \begin{cases} \theta'_j \text{old} - \eta \epsilon_j \frac{r_I r_j}{\sinh r_j} \sin(\theta_I - \theta'_j), & \text{if } w_j \in w_O \cup W_{\text{neg}} \\ \theta'_j \text{old}, & \text{otherwise} \end{cases} \quad (11)$$

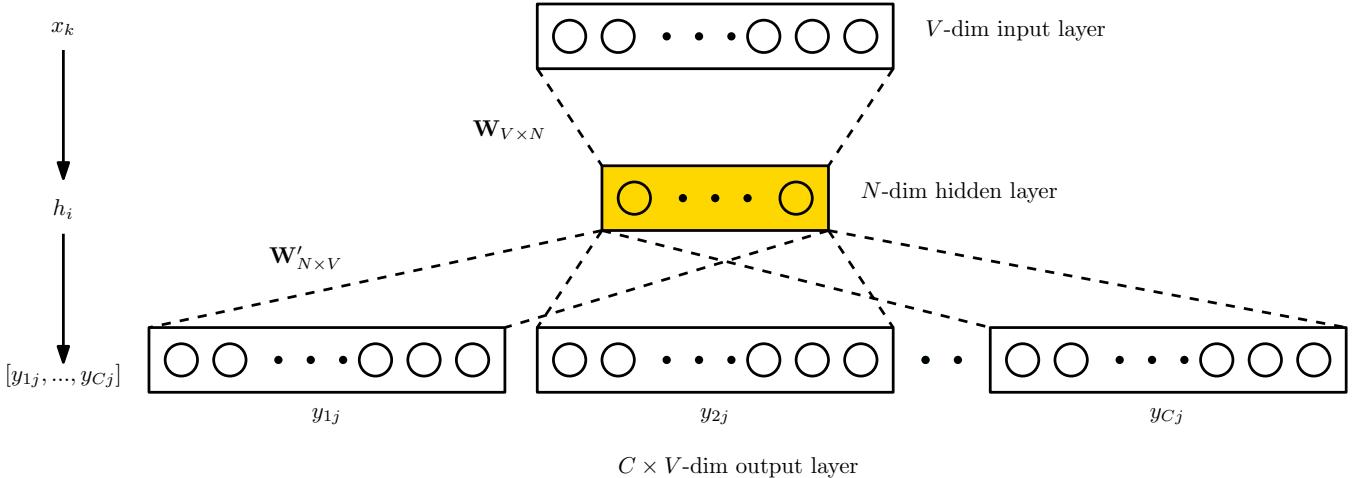
where  $\eta$  is the learning rate and  $\epsilon_j$  is the prediction error defined in Equation (6). Calculating the derivatives w.r.t. the input embedding follows the same pattern, and we obtain

$$\frac{\partial E}{\partial r_I} = \sum_{j:w_j \in w_O \cup W_{\text{neg}}} \frac{\partial E}{\partial u_j} \frac{\partial u_j}{\partial r_I} \quad (12)$$

$$= \sum_{j:w_j \in w_O \cup W_{\text{neg}}} \frac{\partial E}{\partial u_j} r'_j \cos(\theta_I - \theta'_j), \quad (13)$$

$$\frac{\partial E}{\partial \theta_I} = \sum_{j:w_j \in w_O \cup W_{\text{neg}}} \frac{\partial E}{\partial u_j} \frac{\partial u_j}{\partial \theta_I} \quad (14)$$

$$= \sum_{j:w_j \in w_O \cup W_{\text{neg}}} -\frac{\partial E}{\partial u_j} r_I r'_j \sin(\theta_I - \theta'_j). \quad (15)$$



**Figure 3: The skipgram model predicts the context vertices from a single input vertex.**

**Table 1: Description of experimental datasets. ‘Largest class’ gives the fraction of the dataset composed by the largest class and thereby provides the benchmark for random prediction accuracy.**

name	V	E	y	largest class	Labels
karate	34	77	2	0.53	Factions
polbooks	105	441	3	0.46	Affiliation
football	115	613	12	0.11	League
adjnoun	112	425	2	0.52	Part of Speech
polblogs	1,224	16,781	2	0.52	Affiliation

The corresponding update equations are

$$r_I^{new} = r_I^{old} - \eta \sum_{j:w_j \in w_O \cup W_{neg}} \epsilon_j r'_j \cos(\theta_I - \theta'_j), \quad (16)$$

$$\theta_I^{new} = \theta_I^{old} - \eta \sum_{j:w_j \in w_O \cup W_{neg}} \epsilon_j \frac{r_I r'_j}{\sinh r_I} \sin(\theta_I - \theta'_j), \quad (17)$$

where  $t_j$  is an indicator variable s.t.  $t_j = 1$  if and only if  $w_j = w_O$ , and  $t_j = 0$  otherwise. On completion of backpropagation, the vectors are mapped back to Euclidean co-ordinates on the Poincaré disk through  $\theta_h \rightarrow \theta_e$  and  $r_h \rightarrow \tanh \frac{r_h}{2}$ .

## 4 EXPERIMENTAL EVALUATION

In this section, we assess the quality of hyperbolic embeddings and compare them to embeddings in Euclidean spaces on a number of public benchmark networks.

### 4.1 Datasets

We report results on five publicly available network datasets for the problem of vertex attribution.

- (1) Karate: Zachary’s karate club contains 34 vertices divided into two factions [30].

- (2) Polbooks: A network of books about US politics published around the time of the 2004 presidential election and sold by the online bookseller Amazon.com. Edges between books represent frequent co-purchasing of books by the same buyers.
- (3) Football: A network of American football games between Division IA colleges during regular season Fall 2000 [9].
- (4) Adjnoun: Adjacency network of common adjectives and nouns in the novel David Copperfield by Charles Dickens [23].
- (5) Polblogs: A network of hyperlinks between weblogs on US politics, recorded in 2005 [1].

Statistics for these datasets are recorded in Table 1.

### 4.2 Visualising Embeddings

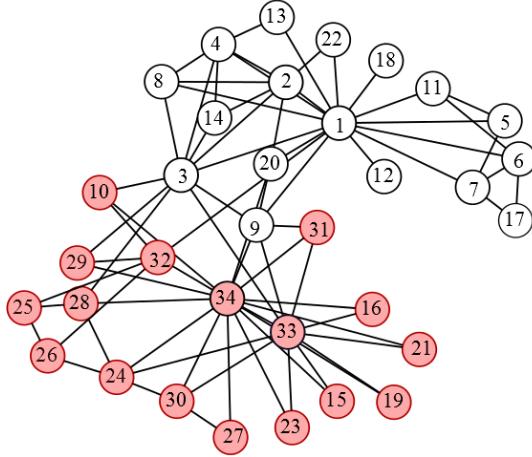
To illustrate the utility of hyperbolic embeddings we compare embeddings in the Poincaré disk to the two-dimensional deepwalk embeddings for the 34-vertex karate network with two factions. The results are shown in Figure 4. Both embeddings were generated by running for five epochs on an intermediate dataset of 34, ten step random walks, one originating at each vertex.

The figure clearly shows that the hyperbolic embedding is able to capture the community structure of the underlying network. When embedded in hyperbolic space, the two factions (black and white discs) of the underlying graph are linearly separable, while the Deepwalk embedding does not exhibit such an obvious structure.

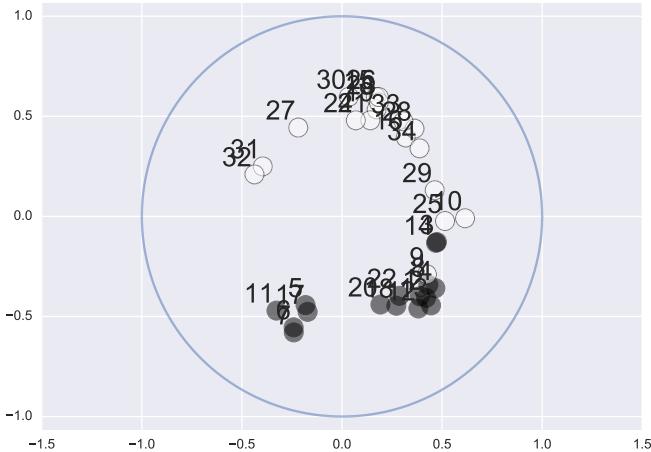
### 4.3 Vertex Attribute Prediction

We evaluate the success of neural embeddings in hyperbolic space by using the learned embeddings to predict held-out labels of vertices in networks. In our experiments, we compare our embedding to deepwalk [24] embeddings of dimensions 2, 4, 8, 16, 32, 64 and 128. To generate embeddings we first create an intermediate dataset by taking a series of random walks over the networks. For each network we use a ten-step random walk originating at each vertex.

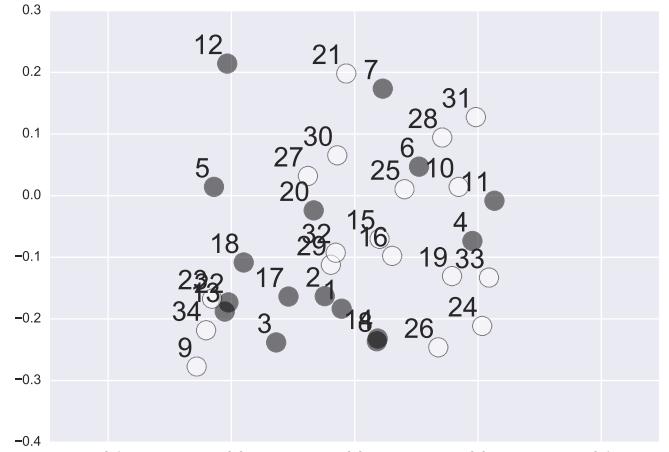
The embedding models are all trained using the same parameters and intermediate random walk dataset. For deepwalk, we use the gensim [25] python package, while our hyperbolic embeddings are



(a) Zachary's karate network. The network is split into two factions.



(b) Two-dimensional hyperbolic embedding of the karate network in the Poincaré disk.



(c) Two dimensional Deepwalk embedding of the karate network.

**Figure 4: The factions of the Zachary karate network are easily linearly separable when embedded in 2D hyperbolic space. This is not true when embedding in Euclidean space. Both embeddings were run for 5 epochs on the same random walks**

written in custom TensorFlow. In both cases, we use five training epochs, a window size of five and do not prune any vertices.

The results of our experiments are shown in Figure 5. The graphs show macro F1 scores against the percentage of labelled data used to train a logistic regression classifier. Here we follow the method for generating F1 scores when each test case can have multiple labels that is described in [17]. The error bars show one standard error from the mean over ten repetitions. The blue lines show hyperbolic embeddings while the red lines depict deepwalk embeddings at various dimensions. It is apparent that in all datasets hyperbolic embeddings significantly outperform deepwalk.

## 5 CONCLUSION

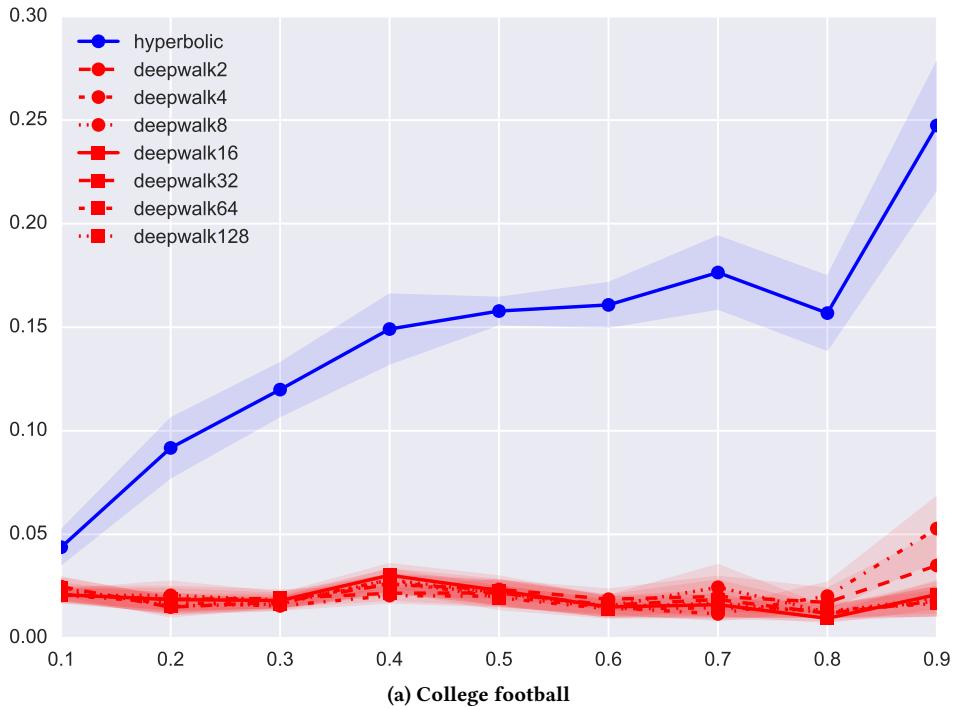
We have introduced the concept of neural embeddings in hyperbolic space. To the best of our knowledge, all previous embeddings models have assumed a flat Euclidean geometry. However, a flat

geometry is not the natural geometry of all data structures. A hyperbolic space has the property that power-law degree distributions, strong clustering and hierarchical community structure emerge naturally when random graphs are embedded in hyperbolic space. It is therefore logical to exploit the structure of the hyperbolic space for useful embeddings of complex networks. We have demonstrated that when applied to the task of classifying vertices of complex networks, hyperbolic space embeddings significantly outperform embeddings in Euclidean space.

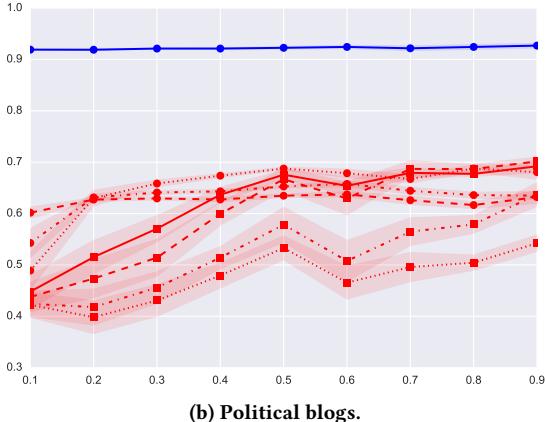
## REFERENCES

- [1] Lada A. Adamic and Natalie Glance. The political blogosphere and the 2004 U.S. election. *Proceedings of the 3rd international workshop on Link discovery - LinkKDD '05*, pages 36–43, 2005.
- [2] Ricardo Baeza-Yates and Diego Saez-Trumper. Wisdom of the Crowd or Wisdom of a Few? *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15*, pages 69–74, 2015.

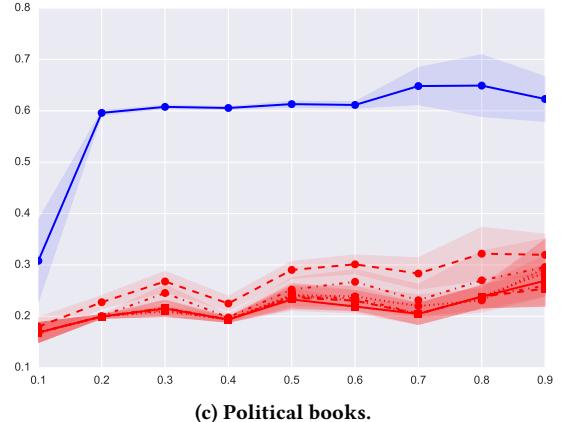
- [3] Oren Barkan and Noam Koenigstein. Item2Vec : Neural Item Embedding for Collaborative Filtering. *Arxiv*, pages 1–8, 2016.
- [4] Mikhail Belkin and Partha Niyogi. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. *Advances in neural information processing systems*, 14:585–591, 2001.
- [5] Marian Boguna, Fragkiskos Papadopoulos, and Dmitri Krioukov. Sustaining the Internet with Hyperbolic Mapping. *Nature Communications*, 1(62):62, 2010.
- [6] Benjamin P. Chamberlain, Angelo Cardoso, Chak H. Liu, Roberto Pagliari, and Marc P. Deisenroth. Customer Life Time Value Prediction Using Embeddings. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.
- [7] James R. Clough and Tim S. Evans. What is the dimension of citation space? *Physica A: Statistical Mechanics and its Applications*, 448:235–247, 2016.
- [8] James R. Clough, Jamie Gollings, Tamar V. Loach, and Tim S. Evans. Transitive reduction of citation networks. *Journal of Complex Networks*, 3(2):189–203, 2015.
- [9] Michelle Girvan and Mark E. J. Newman. Community structure in social and biological networks. In *Proceedings of the national academy of sciences*, 99:7821–7826, 2002.
- [10] Mihajlo Grbovic, Vladan Radosavljevic, Nemanja Djuric, Narayan Bhamidipati, Jaikit Savla, Varun Bhagwan, and Doug Sharp. E-commerce in Your Inbox: Product Recommendations at Scale Categories and Subject Descriptors. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1809–1818, 2015.
- [11] Mikhail Gromov. *Metric Structures for Riemannian and Non-riemannian Spaces*. Springer Science and Business Media, 2007.
- [12] Aditya Grover and Jure Leskovec. node2vec : Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864, 2016.
- [13] Michael U Gutmann. Noise-Contrastive Estimation of Unnormalized Statistical Models, with Applications to Natural Image Statistics. *Journal of Machine Learning Research*, 13:307–361, 2012.
- [14] Farshad Kooti, Mihajlo Grbovic, Luca Maria Aiello, Eric Bax, and Kristina Lerman. iPhone’s Digital Marketplace: Characterizing the Big Spenders. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, pages 13–21, 2017.
- [15] Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, and Amin Vahdat. Hyperbolic Geometry of Complex Networks. *Physical Review E* 82.3:036106, 2010.
- [16] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From Word Embeddings To Document Distances. In *Proceedings of The 32nd International Conference on Machine Learning*, 37:957–966, 2015.
- [17] Yi Liu, Rong Jin, and Liu Yang. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 6, pages 421–426, 2006.
- [18] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. *Advances in neural information processing systems*, pages 3111–3119, 2013.
- [19] Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations*, pages 1–12, 2013.
- [20] Andriy Mnih and Koray Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. *Advances in neural information processing systems*, pages 2265–2273, 2013.
- [21] Andriy Mnih and Geoffrey E. Hinton. A Scalable Hierarchical Distributed Language Model. *Advances in Neural Information Processing Systems*, pages 1081–1088, 2008.
- [22] Andriy Mnih and Yee Whye Teh. A Fast and Simple Algorithm for Training Neural Probabilistic Language Models. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1751–1758, 2012.
- [23] Mark E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 74(3):1–19, 2006.
- [24] Bryan Perozzi and Steven Skiena. DeepWalk : Online Learning of Social Representations. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pages 701–710, 2014.
- [25] Radim Rehurek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, 2010.
- [26] Sam T. Roweis and Lawrence K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science, New Series*, 290(5500):2323–2326, 2000.
- [27] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.
- [28] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to Sequence Learning with Neural Networks. *Advances in neural information processing systems*, pages 3104–3112, 2014.
- [29] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. LINE: Large-scale Information Network Embedding. *Proceedings of the 24th International Conference on World Wide Web*. ACM, , pages 1067–1077, 2015.
- [30] Wayne W. Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33:452–473, 1977.



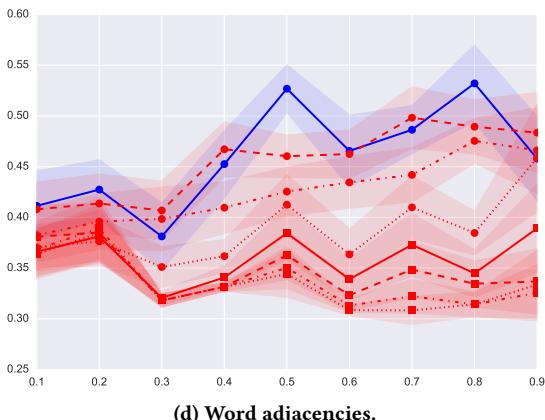
(a) College football



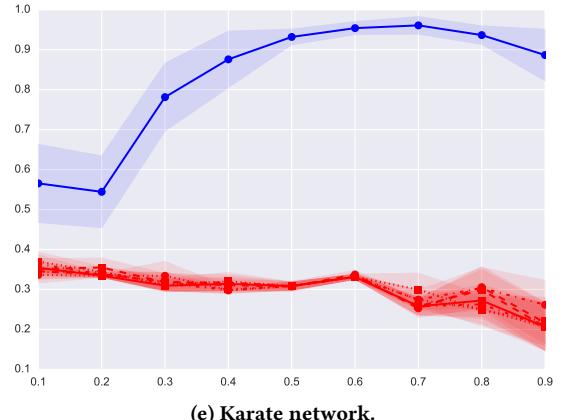
(b) Political blogs.



(c) Political books.



(d) Word adjacencies.



(e) Karate network.

**Figure 5: Macro F1 score (y-axis) against percentage of labelled vertices used for training (x-axis).** In all cases hyperbolic embeddings (blue) significantly outperform Euclidean deepwalk embeddings (red). Error bars show standard error from the mean over ten repetitions. The legend used in subfigure (a) applies to all subfigures. A consistent trend across the datasets is that an embedding into a 2D hyperbolic space outperforms deepwalk architectures with embeddings ranging from 2D to 128D.