# Poincaré Embeddings for Learning Hierarchical Representations

Alicia Tsai

December 2018

**Abstract**

In this report, we discuss current word representation learning and investigate the work by Nickel and Kiela [9] where word embeddings are learned in hyperbolic space, a non-euclidean space. We first review the background of word representation, and introduce the geometric properties in hyperbolic space. Next, we explain why they are useful for learning hierarchies and present the experimental results to show the advantages. Finally, we discuss some limitations and extensions of the work in the last section.

## 1 Introduction

In machine learning applications, data representations generally influence the success of the algorithms. Different data representations may encode different information or explanatory factors of the data. For natural language such as text, distributed representations have proven to be effective and flexible for capturing prior knowledge. Distributed representations of words group similar words together in a vector space. Ideally, distributed representations embed words' semantic and syntactic relationship into low-dimensional vector spaces, called word embeddings.

This report is focused on one particular embedding method called Poincaré embeddings [9], which is used for learning latent hierarchical structures in the data. We start from the background behind distributed representations of words. Then we explain one embedding technique called Poincaré embedding, a new approach for learning hierarchical structures by embedding text into hyperbolic space. The report presents the results of our experiment to evaluate the capability of the Poincaré embeddings. Finally, we discuss the limitations of this method and its extensions.

## 2 Background

### 2.1 Word Representation Learning

Learning representation of words has becomes a central question in natural language processing. The ability to capture information from the data is the foundation for the learning

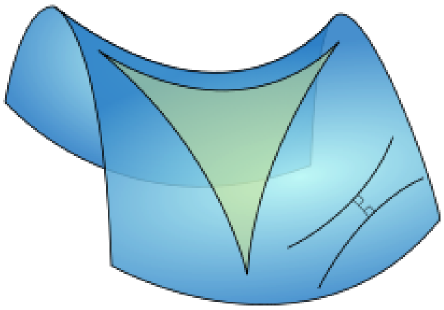

Figure 1: Negative curvature in hyperbolic space (hyperbolic triangle)
Source: Wikipedia

and generation of the downstream natural language processing tasks. In recent years, neural network based approaches utilize massive amounts of data to embed words into lower dimension vector spaces. Word embeddings are effective for many natural language processing tasks because they are flexible and encode valuable syntactic and semantic information. Word embeddings are motivated by the concept that semantic similarities between words are based on their distributional properties in the large amount of text. The idea of distributional properties is called distributional hypothesis [5], meaning that *linguistic items with similar distributions have similar meanings*.

Popular word embeddings such as *GloVe* [10], *Word2Vec* [8], and *FastText* [1] are widely used in various tasks and have shown great success. Although these embedding methods have proven successful, very few methods exist that are able to encode tree-like or graph-like hierarchical relationships of the data. Methods such as *Node2Vec* [4] and latent space approaches [6] are introduced to embed social networks into vector spaces.

## 2.2 Hyperbolic Geometry

In order to learn efficient representations for hierarchical relationships, researchers proposed to computed the embeddings in a non-euclidean space. Here, we introduce the hyperbolic space, space with negative curvature (see figure 1) that holds all the postulates of Euclid except the fifth one. In hyperbolic geometry, the fifth postulate, parallel postulate, of Euclidean geometry is replaced by its negation. In other words, in the hyperbolic space, "there exist a line $l$ and a point $P$ not on $l$ such that at least two distinct lines parallel to $l$ pass through $P$" (see figure 2).

There are several models for representing a hyperbolic plane. The approach discussed later is based on the Poincaré ball model. In two dimensions, all points are in the interior of the unit disk and in higher-dimensions, all points are in the interior of the unit ball. The hyperbolic geometry has certain properties that make it suitable to model hierarchical data.

In the Poincaré ball model, the set of points are denoted as $\mathcal{B}^d = \{x \in R^d, \|x\|_2 \leq 1\}$, where $\|x\|_2$ is the Euclidean norm. The hyperbolic distance between two points $u, v \in \mathcal{B}^d$

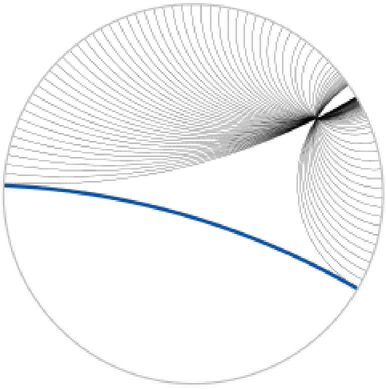

Figure 2: Poincaré disk hyperbolic parallel lines
Source: Wikipedia

has a particular form

$$d_H(u,v) = \text{arcosh}\left(1 + 2\frac{\|u-v\|^2}{(1-\|u\|^2)(1-\|v\|^2)}\right)$$

One property that makes it well-suited for hierarchical structure is that the hyperbolic disc area and circle length grow exponentially with its radius. For example, in two dimensions with curvature $K = 1$, the hyperbolic disc area is

$$2\pi(\cosh r - 1) = \pi\big((e^r + e^{-r}) - 2\big)$$
$$\cosh r = \frac{1}{2}(e^r + e^{-r})$$

and the hyperbolic circle length is

$$2\pi \sinh r = \pi(e^r - e^{-r})$$
$$\sinh r = \frac{1}{2}(e^r - e^{-r})$$

A normal tree-like structure that grows exponentially can now be easily placed in the hyperbolic circle with radius $r$ that is proportional to its height $h$. All leaf nodes at level $h$ are located at the sphere with $r$ radius and all internal nodes at level less than $h$ are located within the sphere.

In Euclidean space, the disc area $(2\pi r^2)$ and the circle length $(2\pi r)$ only grow quadratically and linearly with its radius $r$. Thus, we will need to increase the dimensionality to be able to model a single hierarchy. Due to the property discussed above, learning embeddings in hyperbolic space have captured the attention of some researchers.

In the following section, we discuss how to embed *WordNet* corpora in the hyperbolic space using the Poincaré ball model to capture the hypernymy, hyponymy relation. Even though a single hierarchy can be modeled in two dimensional hyperbolic space $(\mathcal{B}^2)$ as shown above, using a higher dimensional Poincaré ball $(\mathcal{B}^d)$ allows us to to model multiple latent hierarchies within the text corpora.

# 3 Poincaré Embeddings

## 3.1 WordNet Dataset

The main focus for Poincaré embeddings is its capability to embed data that exhibits latent hierarchical structures. Thus, we conduct the experiment using *WordNet* dataset [3]. *WordNet* is a large lexical database of the English language. It groups nouns, verbs, adjectives and adverbs into sets of cognitive synonyms (synsets).

We are interested in one of the relations of *WordNet* called hypernymy, hyponymy relation (super-subordinate relation or IS-A relation). This relation links more general synsets to increasingly specific ones. For example, `furniture` will be linked to `bed`. In this example, we say `bed` is a hyponymy of `furniture` and `furniture` is the hypernymy of `bed`.

Furthermore, we are also interested in the ***transitive closure*** of the hypernymy, hyponynm relation. The transitive closure states that the category `furniture` includes `bed` and the category of `bed` in turn includes `bunkbed`, which makes `bunkbed` a hyponynm of `furniture`. This is similar to a tree or a directed graph structure. Therefore, all noun hierarchies ultimately go up to the root node, which is `entity`.

The hypernymy, hyponynm relation is well suited for the Poincaré embeddings because it exhibits a clear latent hierarchical structure. In this experiment, we only use `mammal` subtree extracted from the noun synsets.

## 3.2 Experimental set-up

We use the implementation from the author of Poincaré embeddings [1] and the *gensim* [2] library for our experiment.

The transitive closure of the *WordNet* `mammal` hierarchy consists of 1181 words and 6541 hypernymy, hyponymy relations. The embeddings are learned from the set of observed hypernymy, hyponymy relations pair, denoted as $\mathcal{D} = \{(u, v)\}$. The embeddings are learned by minimizing the distance of related words and maximizing the distance of unrelated words in the embedding space.

## 3.3 Training Details

When embedding a structure into another space, we aim to preserve distances of the relationships and reflect the semantic similarity of the words. For the *WordNet* dataset, our goal is to preserve the hierarchical distance of the hypernynm, hyponynm relations. For example, if two hyponynm $x$ and $y$ are children of a hypernynm $z$, we can then place $z$ at the origin $O$ or as a root in a tree. Now, the hierarchical distance between $x$ and $y$ is the distance between origin $O$ and $x$ plus the distance between origin $O$ and $y$. We can then

[1] Github: `https://github.com/facebookresearch/poincare-embeddings`

[2] gensim library: `https://radimrehurek.com/gensim/index.html`

normalize the hierarchical distance and get a distance ratio of 1.

$$d(x, y) = d(x, O) + d(y, O)$$
$$\frac{d(x, y)}{d(x, O) + d(y, O)} = 1$$

However, in Euclidean space the distance ratio is a constant that is always smaller than 1 unless $x$, $y$ and $O$ are on the same line or plane, which contradicts the hierarchy assumption.

$$\frac{d_E(x, y)}{d_E(x, O) + d_E(y, O)} \leq 1$$

Furthermore, the distance ratio remains a constant as $x$ and $y$ moves further away from the origin. Figure 3 shows 7 pairs of two points, $A$ and $B$, moving further away from origin. The distance ratio for all pairs of $A$ and $B$ is the same. Figure 4 illustrates that as $x$ norm increases, the distance ratio in Euclidean space remain constant.

On the other hand, in hyperbolic space, the distance ratio approaches 1 as $x$ norm increases (see figure 4).

$$\frac{d_H(x, y)}{d_H(x, O) + d_H(y, O)} \approx 1$$

This means that if we place $x$ and $y$ close to the edge of the Poincaré ball, we can get a distance close to their original hierarchical distance.

### 3.3.1 Loss Function

Given a loss function $\mathcal{L}(\theta)$, we want to learn the embeddings $\Theta = \{\theta_i\}_{i=1}^{n}$, where $\theta_i \in \mathcal{B}^d$, in hyperbolic space that makes pairs of hypernynm and hyponynm close to each other according to their hyperbolic distance.

Given a set of observed mammal hypernymy, hyponymy relations pair, denoted as $\mathcal{D} = \{(u, v)\}$, we minimize the loss function

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

where $\mathcal{N}(u) = \{v' | (u, v') \notin \mathcal{D}\} \cup \{v\}$ is the set of negative examples that is not a hyponynm for $u$ plus the actual hyponynm $v$. This loss function minimizes the distance between related words and maximizes distance between words for which we didn't observe the hypernynm, hyponynm relationship.

## 4 Results and Evaluation

To evaluate the quality of the embeddings, we rank the distance between a pair of relations $d(u, v)$ among chosen the negative examples for $u$. After ranking all pairs of relations in

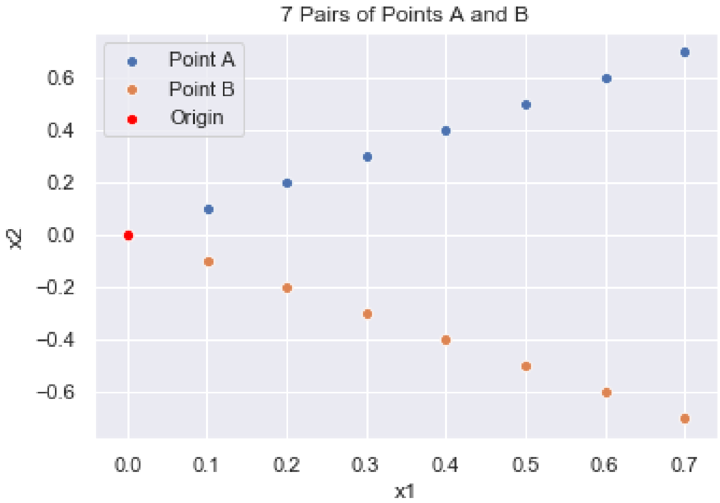

Figure 3: 7 pairs of point A and B in Euclidean space

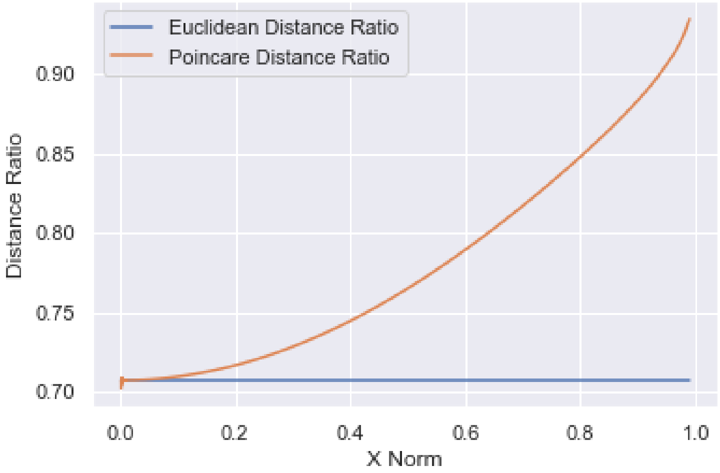

Figure 4: Poincaré and Euclidean distance ratio

| Poincaré Embedding | | | | | |
|---|---|---|---|---|---|
| Dimensions | 5 | 10 | 50 | 100 | 200 |
| MAP | 0.383 | 0.386 | 0.353 | 0.384 | 0.388 |
| Mean Rank | 4.0 | 3.9 | 3.9 | 3.9 | 3.9 |

Table 1: MAP and mean rank of Poincaré embedding

| Euclidean Embedding | | | | | |
|---|---|---|---|---|---|
| Dimensions | 5 | 10 | 50 | 100 | 200 |
| MAP | 0.019 | 0.038 | 0.305 | 0.325 | 0.341 |
| Mean Rank | 184.5 | 100.3 | 11.4 | 18.4 | 9.5 |

Table 2: MAP and mean rank of Euclidean embedding

the `mammal` subtree, we calculate the mean rank of the data set and the mean average precision (MAP) of the ranking. Table 1 and table 2 present the results of embedding `mammal` subtree in hyperbolic space and in euclidean space. It can be shown that Poincaré embeddings can achieve a better result with fewer dimensions.

Next, we visualize the Poincaré embedding by projecting them onto a two dimensional plane using t-SNE. The visualization shows that the learned embeddings form several clusters in hyperbolic space. We then color all hyponynms of a given word (its hyponynm subtree) to evaluate whether the embeddings capture the ***transitive closure*** of the relation. In figure 5, we color all hyponynms of `dog` in red. In figure 6 we color all hyponynms of `cat` in red. It is shown that hyponynms are placed close to each other in a given subtree in the embedding space. In figure 7, we randomly annotate some data points to understand what the clusters represent. We can see that there are several clusters including canidae (”dog-like” mammals), feliformia (”cat-like” mammals), primates, marine mammals, rodents, and bats. From the 3 figures, we can see that related words are placed closer and it forms a circle (ball) like structure that align with our assumption and the properties of hyperbolic geometry discussed earlier.

# 5 Discussion and Future Work

The focus of this report is evaluating the properties of hyperboblic geometry and its capability for embedding hierarchical relations. It is shown experimentally that embeddings learned in hyperbolic space require far fewer dimensions than embeddings learned in Euclidean space on the *WordNet* dataset. Furthermore, the simulated distance ratio shows that Poincaré distance can be used to approximate the true hierarchical distance of the tree-like structure data.

The embeddings are learned explicitly from text corpus that exhibit clear hierarchical relations. One aspect of future work is to evaluate their performance in the downstream tasks. Another aspect of future work is to learn word embeddings in hyperbolic space directly

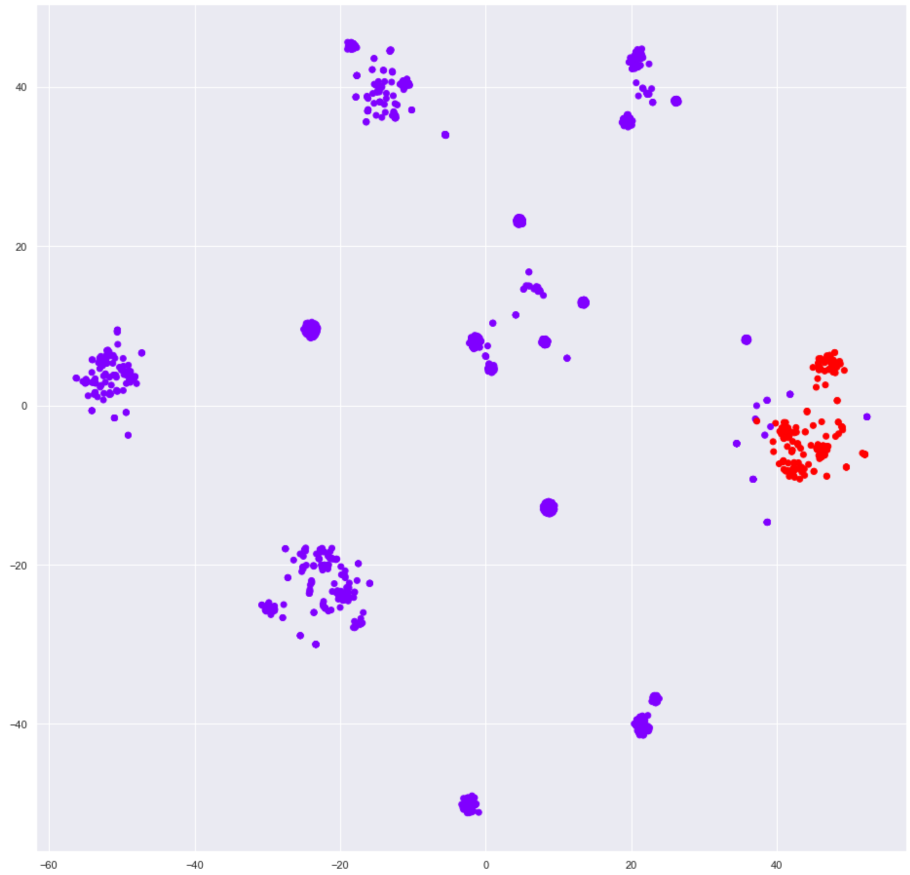

Figure 5: *dog* hyponyms subtree (highlighted in red) in a 25 dimensions Poincaré embeddings projected down to 2 dimensions

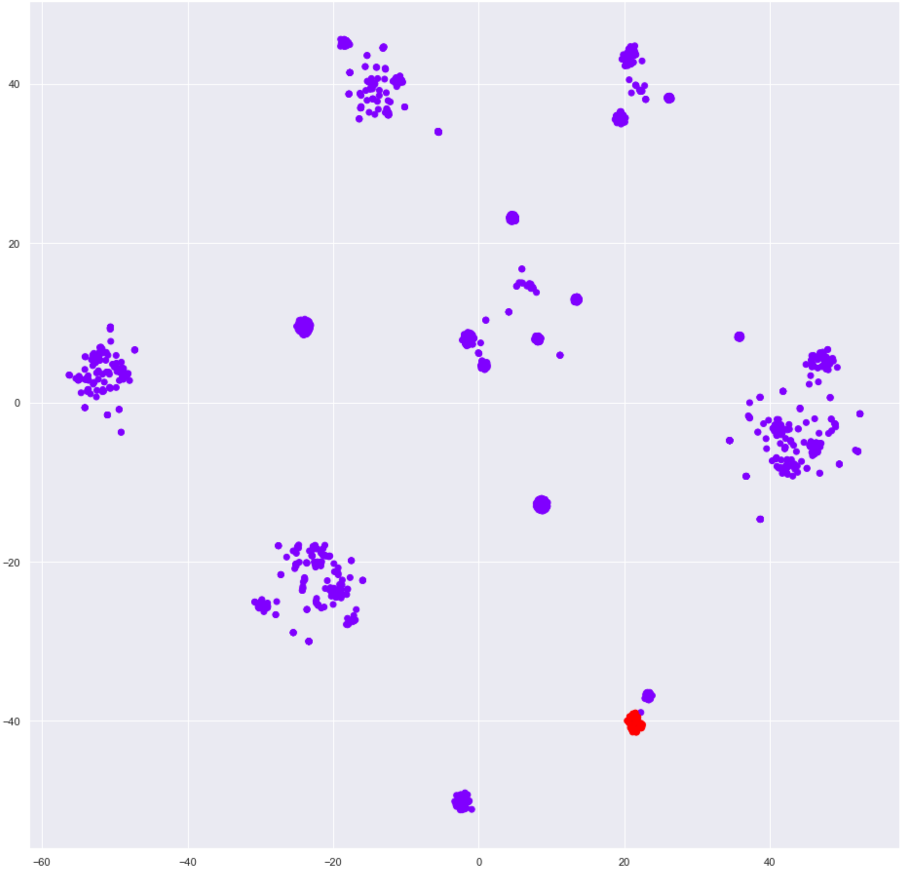

Figure 6: Poincaré Embeddings cat homonyms subtree within mammal subtree (25 dimensions)

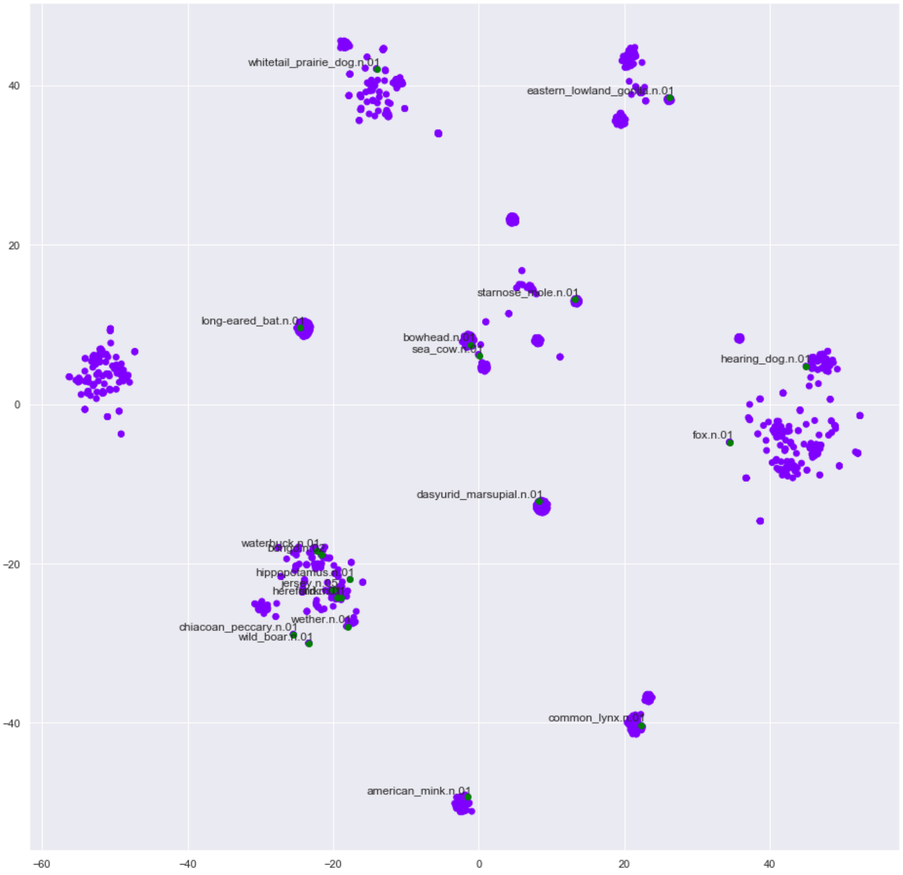

Figure 7: Poincaré Embeddings mammal subtree with labels (25 dimensions)

from free text corpus without exposing hierarchical relations explicitly.

Several recent research works have focused on this direction. Bhuwan et al. [2] extend the method from Nickel and Kiela [9] to allow learning embeddings with free text. Leimeister and Wilson [7] attempt to learn word embeddings in hyperbolic space with skip-gram architecture from *Word2Vec*. Alexandru et al. [11] adapts the *GloVe* algorithm to hyperbolic space.

These research works show evidence of improvements with tasks that exhibit intuitive hierarchy but not with all downstream tasks. Further investigation could focus on developing algorithms to generalize hyperbolic embeddings for more downstream tasks where hierarchical organizations may not be explicitly shown.

# References

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[2] Bhuwan Dhingra, Christopher J. Shallue, Mohammad Norouzi, Andrew M. Dai, and George E. Dahl. Embedding Text in Hyperbolic Spaces. *ArXiv e-prints*, page arXiv:1806.04313, June 2018.

[3] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[4] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. *CoRR*, abs/1607.00653, 2016.

[5] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[6] Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. Latent space approaches to social network analysis. *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION*, 97:1090–1098, 2001.

[7] Matthias Leimeister and Benjamin J. Wilson. Skip-gram word embeddings in hyperbolic space. *ArXiv e-prints*, page arXiv:1809.01498, August 2018.

[8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[9] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6338–6347. Curran Associates, Inc., 2017.

[10] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[11] Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincar\'e GloVe: Hyperbolic Word Embeddings. *ArXiv e-prints*, page arXiv:1810.06546, October 2018.