

Plongements hyperboliques de graphes hiérarchiques

Rapport de projet DU Mathématiques et Applications, Recherche et Interactions

Victor Le Guilloux et Claire Schnoebelen,
encadré par Emmanuel Opshtein

24 janvier 2023

Introduction

Ce projet s'inscrit dans le cadre de l'apprentissage de données symboliques, qui trouve, par exemple, des applications dans la détection des sentiments exprimés dans un texte, la prédiction de liens ou encore la détection de communautés dans un réseau social. Un jeu de données symboliques peut donc contenir des objets très divers sur lesquels on peut définir une notion de similarité. Nous étudierons ici le cas où les objets manipulés sont les sommets d'un graphe, que l'on dit similaires s'il sont reliés par une arête.

Une façon d'analyser un jeu de données symboliques est de le représenter par un ensemble de points dans un espace métrique, généralement un espace euclidien. L'idée de cette étape, appelée *plongement*, est d'organiser les concepts qui composent le jeu de données dans l'espace métrique de sorte à ce que les distances entre leurs représentants reflètent leurs similarités.

Les méthodes pour plonger des graphes dans des espaces euclidiens sont toutefois limitées par la dimension : plus le graphe considéré est complexe, plus il est nécessaire d'augmenter la dimension de l'espace des représentants. Ceci rend souvent impraticable l'analyse en raison des coûts de calcul qui deviennent exorbitants.

Dans [3] est proposé une méthode pour plonger des graphes dans un espace hyperbolique. Le but de ce projet est de comprendre cette méthode et de saisir l'intérêt de la géométrie hyperbolique dans l'étude des graphes. Pour ce dernier point, on s'appuie principalement sur [2].

Dans un premier temps, nous donnerons quelques éléments sur l'organisation d'un certain type de graphes, dits *hiérarchiques*, expliquant les difficultés rencontrées pour les plonger dans un espace euclidien de petite dimension. Nous présenterons ensuite certaines caractéristiques des espaces hyperboliques ainsi que les raisons qui en font des espaces adaptés au plongement de graphes. Cela fait, nous présenterons la méthode proposée dans [3] ainsi que les tests réalisés par nos soins sur deux graphes de collaboration scientifiques.

I Notion de hiérarchie dans un graphe

Plonger un graphe dans un espace métrique adapté permet de visualiser certaines de ses propriétés.

Une première idée consiste à plonger les graphes dans des espaces usuels. On peut penser dans un premier temps à plonger dans des espaces euclidiens, mais nous allons constater que ceux-ci ne sont pas adaptés aux plongements de certains graphes. Parmi ceux-ci, on peut trouver les arbres à facteur d'embranchement constant.

En effet, dans de tels arbres les cardinaux des boules et sphères sont des fonctions exponentielles du rayon : dans un arbre avec un facteur de branchement b , chaque sommet a exactement $(b+1)b^{r-1}$ sommets à distance r , et exactement $\frac{(b+1)b^r-2}{b-1}$ sommets à distance au plus r . Tous deux croissent comme b^r . En revanche dans un espace euclidien les mesures des boules et des sphères sont des fonctions polynomiales du rayon, quelle que soit la dimension de l'espace que l'on considère. Si l'on veut plonger de grands arbres dans de tels espaces, même s'ils ont un petit facteur de branchement, la dimension requise peut rendre tout calcul numérique irréalisable.

Nous nous pencherons plus spécifiquement sur une autre famille de graphes : les graphes *hiérarchiques*.

Définition I.1 Soit $G = (S, A)$ un graphe. La *distribution des degrés* de G est l'application :

$$\begin{aligned} k : \mathbf{N} &\longrightarrow \mathbf{N} \\ n &\longmapsto \#\{s \in S \mid \deg(s) = n\} \end{aligned}$$

S'il existe $\gamma > 1$ tel que $k(n) \propto n^{-\gamma}$, on dit que k est une *loi de puissance* (ou *power law*) d'exposant γ . On dit alors que G est un graphe *hiérarchique*.

Dans un graphe hiérarchique, il existe donc peu de sommets de haut degré, et de nombreux sommets de faible degré. Les premiers sont, en un certain sens, plus *universels* que les seconds.

Ex :

- Dans un arbre à facteur de branchement b , $k(n) = 0$ dès que $n \neq b+1$.
- Les graphes de nombreux réseaux présentent une structure hiérarchique.

En première approximation, on pourra considérer que la structure de ces graphes hiérarchiques est similaire à celle d'arbres (à facteur de branchement non constant), bien que le lien entre ces deux notions ne soit pas parfaitement clair. C'est avec cette idée en tête que nous essaierons de plonger ces graphes hiérarchiques dans des espaces adaptés aux plongements des arbres.

II Généralités sur les espaces hyperboliques

Les *espaces hyperboliques* constituent une famille d'espaces-modèles à courbure constante, et plus précisément à courbure constante négative. Les autres espaces-modèles à courbure constante sont les sphères (qui ont une courbure positive) et les espaces euclidiens (qui ont une courbure nulle). Les espaces hyperboliques disposent eux-même de plusieurs modèles. Celui que nous utiliserons dans la suite est le modèle de la *boule de Poincaré*.

Définition II.1 Soit $K = -\zeta^2 < 0$. Le modèle de la boule de Poincaré de l'espace hyperbolique de dimension d et de courbure K , noté $\mathbf{H}_\zeta^d$, est la boule unité de dimension d munie de la métrique riemannienne :

$$g_{hyp} = \frac{4}{\zeta^2} \frac{|\cdot|^2}{(1-r^2)^2},$$

où $|\cdot|^2 = dx_1^2 + \dots + dx_d^2$ est la métrique euclidienne standard.

On notera simplement $\mathbf{H}^d = \mathbf{H}_1^d$ la boule de Poincaré de courbure -1.

La sphère unité $\mathbf{S}^{d-1}$ est appelée le bord à l'infini de $\mathbf{H}_\zeta^d$ et est notée $\partial\mathbf{H}_\zeta^d$.

Remarque : La métrique hyperbolique sur le disque de Poincaré est conforme à la métrique euclidienne standard, c'est-à-dire qu'en tout point ces deux métriques sont proportionnelles. Cela a pour conséquence, entre autres, que les mesures des angles réalisées avec la métrique hyperbolique sont égales à celles réalisées avec la métrique euclidienne.

Les géodésiques des espaces hyperboliques sont bien connues, de même que certaines des propriétés métriques de ces espaces. Nous nous servirons en particulier de l'expression de la distance hyperbolique.

Proposition II.2 (i) Les géodésiques de $\mathbf{H}_\zeta^d$ sont les diamètres et les arcs de cercle orthogonaux à $\partial\mathbf{H}_\zeta^d$;

(ii) la longueur d'un cercle de rayon r et l'aire d'un disque de rayon r sont données respectivement par :

$$\frac{2\pi \sinh \zeta r}{\zeta} \text{ et } \frac{2\pi(\cosh(\zeta r) - 1)}{\zeta^2};$$

(iii) si $u, v \in \mathbf{H}_\zeta^d$, la distance hyperbolique séparant u et v est donnée par :

$$d(u, v) = \frac{1}{\zeta} \operatorname{argch} \left(1 + 2 \frac{|u - v|^2}{(1 - |u|^2)(1 - |v|^2)} \right).$$

Nous pouvons remarquer que dans le plan hyperbolique, l'aire des disques ainsi que la longueur des cercles croissent comme des fonctions exponentielles du rayon. Nous avons déjà observé un comportement similaire dans les arbres à facteur de branchement constant.

Cette similarité métrique nous conduit à considérer les plans hyperboliques comme des analogues continus d'arbres. Nous pouvons pousser l'analogie un peu plus loin en remarquant que la courbure de l'espace hyperbolique joue un rôle similaire à celui du facteur de branchement, puisque dans le plan de courbure $-\zeta^2$, les aires des disques et les longueurs des cercles croissent comme $e^{\zeta r}$, et dans un arbre à facteur de branchement b , elles croissent comme b^r . Dans cette analogie entre arbres réguliers et plans hyperboliques, e^ζ est un analogue du facteur de branchement.

Ainsi les propriétés métriques des espaces hyperboliques en font des candidats appropriés pour le plongement d'arbres.

Souvenons-nous qu'en première approximation, nous avons considéré que la structure hiérarchique d'un graphe se rapprochait d'une structure d'arbre à facteur de branchement variable. C'est pour cette raison qu'il semble raisonnable d'envisager de plonger ces graphes hiérarchiques dans des espaces hyperboliques.

III Plongement hyperbolique de graphes

III.1 Plongements de graphes

Plonger un graphe dans un espace métrique consiste à associer à chacun de ses nœuds un représentant dans cet espace. Un plongement intéressant doit associer à des sommets voisins des points proches et à des sommets non-voisins des points plus éloignés, de sorte que la distance dans l'espace des représentations soit cohérente avec la notion de similarité dont on dispose sur les données symboliques de départ. Ces plongements peuvent ensuite servir d'entrées à d'autres méthodes d'analyse comme des réseaux de neurones.

Les considérations précédentes incitent à voir les graphes, pour peu qu'ils aient une structure en power law, comme des équivalents discrets d'espaces hyperboliques. Le but de ce projet est d'étudier l'usage qui peut être fait de cette ressemblance dans la construction de plongements pour des graphes hiérarchiques.

Ce qui suit est un résumé de la méthode proposée dans [3] et une description des moyens mis en œuvre pour l'implémenter.

III.2 Formulation du problème d'optimisation

Soit $G = (S, A)$ un graphe connexe à n nœuds. À chaque nœud s_i , on souhaite associer un point x_i dans $\mathbf{H}^2$ de sorte à ce que deux sommets voisins s_i et s_j aient des représentants x_i et x_j proches dans $\mathbf{H}^2$. À l'inverse, la probabilité que deux points représentent des sommets voisins dans le graphe doit être de plus en plus forte à mesure que la distance hyperbolique qui les sépare est petite. Quand on dit que x_i et x_j sont proches dans $\mathbf{H}^2$, on entend que la distance (hyperbolique) entre ces deux points est plus petite que la distance qui les sépare respectivement des nœuds desquels ils ne sont pas voisins.

Un bon plongement $X = \{x_1, \dots, x_n\}$ réalise donc une faible valeur de la fonction

$$\mathcal{L}(X) = \sum_{(s_i, s_j) \in A} \log \left(\frac{e^{-d(x_i, x_j)}}{e^{-d(s_i, s_j)} + \sum_{s_k \in N(s_i)} e^{-d(s_i, s_k)}} \right),$$

où d est la distance de $\mathbf{H}^2$ et $N(s_i)$ l'ensemble des sommets qui ne sont pas voisins de s_i , autrement dit les s_k tels que $(s_i, s_k) \notin A$.

Pour trouver X , on cherchera donc à résoudre le problème de minimisation

$$\underset{X \in (\mathbf{H}^2)^n}{\operatorname{argmin}} \mathcal{L}(X). \quad (1)$$

III.3 Algorithme de descente de gradient

Pour trouver X , on effectue une descente de gradient sur $\mathbf{H}^2$. Dans $\mathbf{R}^m$, les itérés sont reliés par la formule

$$X_{t+1} = X_t - \gamma \nabla f(X_t),$$

où f est la fonction à minimiser et γ le learning rate choisi.

Cette dernière n'a pas de sens sur une variété puisqu'elle fait appel à la structure affine de $\mathbf{R}^m$. Un équivalent riemannien proposé dans [1] est donné par

$$X_{t+1} = \exp_{X_t}(-\gamma \nabla_{hyp} f(X_t)),$$

où $\nabla_{hyp} f(x)$ désigne le gradient de f en x pour la métrique riemannienne choisie sur la variété. En pratique, l'application exponentielle peut être difficile à calculer et on lui préférera une approximation R_x , appelée rétraction.

Dans le cas qui nous concerne, [3] propose de prendre

$$R_x(v) = \pi(x + v)$$

avec

$$\pi : x \mapsto \begin{cases} x & \text{si } |x| < 1, \\ \frac{x}{|x|+\epsilon} & \text{sinon,} \end{cases}$$

où ϵ est un réel positif et $|\cdot|$ la norme euclidienne. La somme $x + v$ dans la formule de R_x doit s'entendre comme l'action de v vu comme vecteur de $\mathbf{R}^2$ sur x vu comme point de l'espace affine $\mathbf{R}^2$.

III.4 Implémentation

Nous avons tenté d'implémenter cette méthode de plongement en utilisant TensorFlow. Nous avons choisi d'effectuer nos tests sur deux graphes, ERDOS992 et GRQC, disponible sur [4]. Chacun d'eux compte environ 5000 sommets représentant des auteurs scientifiques. Deux sommets sont liés par une arête si les scientifiques qu'ils représentent ont une publication commune. Le premier est le graphe d'Erdős tronqué tel qu'il était en 1992 : il contient tous les co-auteurs de Paul Erdős ainsi que leurs propres co-auteurs (autrement dit, il contient toutes les personnes ayant un nombre d'Erdős de 1 ou 2). Le second est un graphe de collaboration construit à partir des articles portant sur la relativité générale déposés sur arXiv.

Pour chacun de ces graphes, on extrait la plus grande des composantes connexes. Dans les deux cas, cette dernière contient en fait la quasi-totalité des sommets. Ensuite, on vérifie que ces deux graphes suivent bien une loi de puissance et entrent donc dans le cadre dans lequel a été construite la méthode. Cela est bien le cas, comme on peut le voir Figure 1.

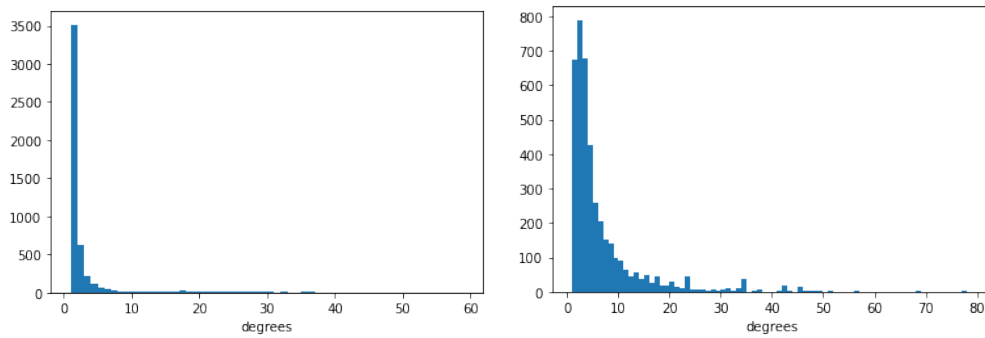


FIGURE 1 – Distribution des degrés sur les graphes d'Erdős (gauche) et GrQc (droite).

Tel quel, le problème d’optimisation est difficile à résoudre numériquement car le gradient explose rapidement. Dans [3], il est proposé de tirer aléatoirement dix exemples négatifs s_k par exemple positif s_j pour tous les s_i . Nous n’avons pas réussi à vectoriser cette étape, notamment parce que l’ensemble des indices sur lesquels il faut tirer ces exemples négatifs est différent pour chaque point s_i . Le calcul du gradient prend donc énormément de temps, ce qui a rendu la méthode impraticable pour nous.

Pour pouvoir produire un plongement, nous avons considéré la cross-entropie binaire, donnée par

$$\mathcal{L}_{bis}(X) = \sum_{i,j=1}^n \left((1 - N_{ij}) \log(1 - e^{-d(x_i, x_j)}) + N_{ij} \log(e^{-d(x_i, x_j)}) \right),$$

où N_{ij} vaut 1 si s_i et s_j sont voisins et 0 autrement. Avec cette loss également, on encourage les représentations des sommets voisins à être proches et celles des sommets non-voisins à être éloignés mais on perd l’aspect relatif de la première considération par rapport à la seconde.

On présente Figure 2 les plongements à des moments successifs de l’algorithme. Pour des questions de lisibilité de la figure, on ne représente pas les arêtes du graphe. La grosseur des points varie avec le degré et est d’autant plus importante que le graphe a de voisins. Ceci suffit pour voir qu’au fur et à mesure des itérations, les points se séparent et changent de position indépendamment de leur degré.

Un des intérêts de la méthode devrait être une mise en évidence de la hiérarchie : on voudrait voir, comme dans [3], les nœuds de plus hauts degrés placés près du centre du disque et les feuilles près du bord. Les observations que l’on vient de faire sur nos résultats montrent que ces derniers ne sont pas satisfaisants de ce point de vue mais cela peut se comprendre au vu de la loss que nous avons choisie.

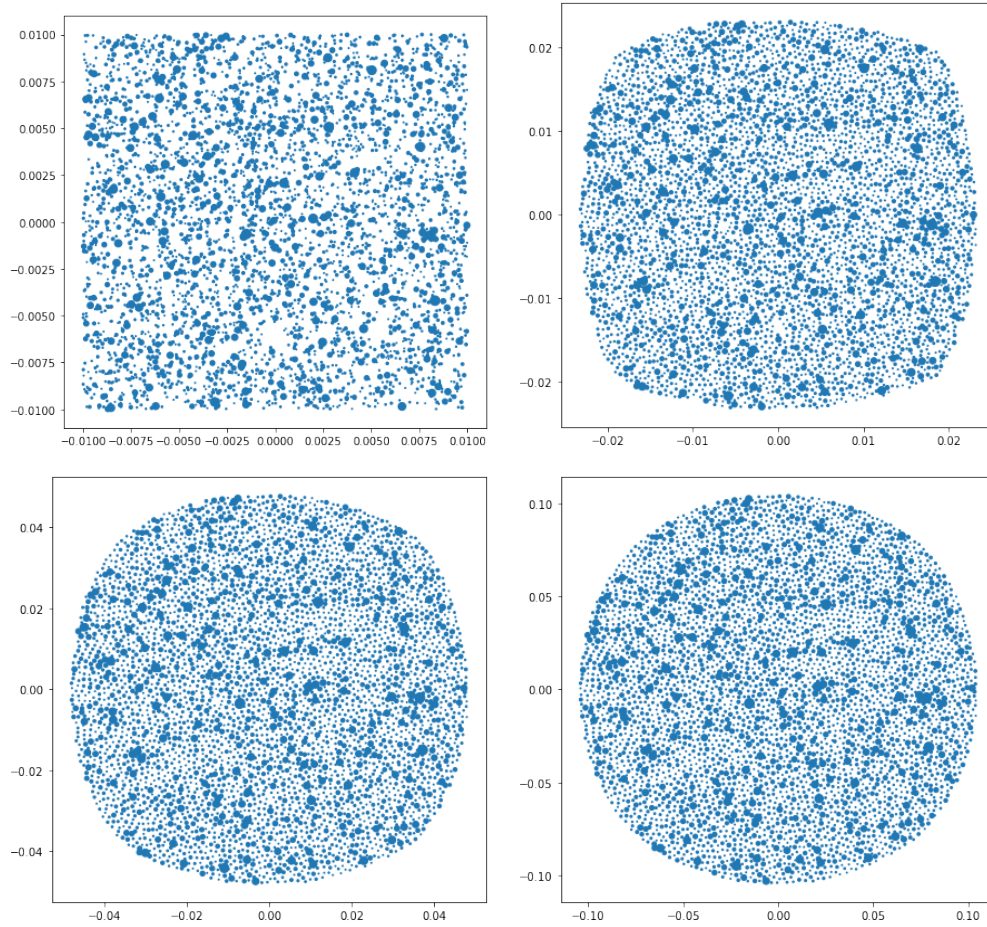


FIGURE 2 – Représentation des nœuds du graphe GrQc dans le disque de Poincaré après 0, 20, 100 et 500 itérations de l’algorithme de descente. Les points sont d’autant plus gros que leur degré est élevé.

On voit qu’à mesure que le nombre d’itérations croît, l’enveloppe convexe des x_i se déforme pour passer du carré $[-0.01, 0.01]^2$ à un disque de rayon croissant. Dans le même temps, on observe que les points tendent à se disposer régulièrement dans cette enveloppe, de façon à maximiser les distances entre plus proches voisins. Peut-être qu’en laissant tourner l’algorithme plus longtemps, on finirait par obtenir des clusters et par observer une distribution radiale en fonction du degré mais ce n’est a priori pas évident.

En jouant avec le learning rate et en le prenant très grand au début, en le diminuant progressivement et en l’augmentant brusquement pour quelques itérations à intervalle réguliers, on obtient le plongement présenté Figure 3. On y voit l’effet de la métrique hyperbolique : même si les points situés au bord sont proches pour la métrique euclidienne, ils ne le sont pas plus pour la métrique hyperbolique que ceux situés près du bord. La tendance qu’avaient les points à se répartir régulièrement, indépendamment de leur degré, aux premières itérations est donc toujours vraie.

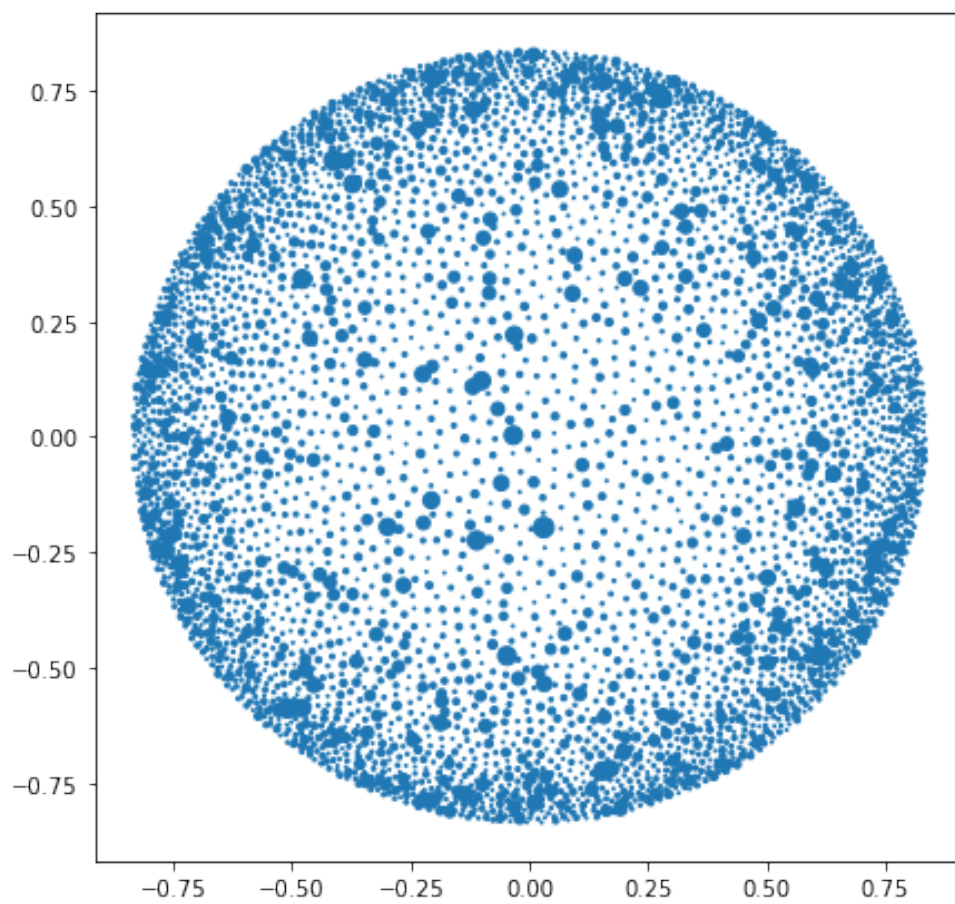


FIGURE 3 – Plongement après 1000 itérations

Conclusion

Ce projet nous a permis d’explorer les liens entre géométrie hyperbolique et graphes hiérarchiques. La notion de hiérarchie pour un graphe est assez vaguement définie. Toutefois il semble important, à la fois pour comprendre pourquoi il est impossible d’obtenir un plongement convenable dans un espace euclidien de petite dimension et pour saisir le lien avec la géométrie hyperbolique, que la distribution des degrés obéisse à une loi de puissance. Ces graphes s’apparentent alors à des arbres avec une croissance de plus en plus rapide du nombre de nœuds à chaque génération. C’est ce besoin exponentiellement croissant de place pour représenter les nœuds à chaque génération qui rend les espaces euclidiens peu adaptés aux questions de plongements de graphes hiérarchiques, alors qu’il s’agit d’une propriété métrique naturelle des espaces hyperboliques.

C’est en partant de ce constat que la méthode proposée dans [3] utilise le disque de Poincaré comme espace de représentation. Lors de nos tests, nous n’avons malheureusement pas réussi à minimiser la fonction objective proposée dans l’article. Nous avons toutefois mené des expériences avec une loss simplifiée, qui n’ont malheureusement pas abouti aux résultats espérés.

Références

- [1] Silvere BONNABEL. “Stochastic Gradient Descent on Riemannian Manifolds”. In : *IEEE Transactions on Automatic Control* 58.9 (2013), p. 2217-2229.
- [2] Dmitri KRIOUKOV et al. “Hyperbolic geometry of complex networks”. In : *Physical Review E* 82.3 (2010).
- [3] Maximilian NICKEL et Douwe KIELA. *Poincaré Embeddings for Learning Hierarchical Representations*. 2017. URL : <https://arxiv.org/abs/1705.08039>.
- [4] Ryan A. ROSSI et Nesreen K. AHMED. “The Network Data Repository with Interactive Graph Analytics and Visualization”. In : *AAAI*. 2015. URL : <https://networkrepository.com>.